SPATIO-TEMPORAL ANALYSIS OF USAGE PATTERNS IN A PUBLIC BIKE SHARING SYSTEM DURING A PANDEMIC

Vincent Marcus Garcia
University of Rhode Island, vincent1996@live.de

Follow this and additional works at: https://digitalcommons.uri.edu/theses

Recommended Citation
https://digitalcommons.uri.edu/theses/1990

This Thesis is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Master's Theses by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.
SPATIO-TEMPORAL ANALYSIS OF USAGE PATTERNS IN A PUBLIC BIKE
SHARING SYSTEM DURING A PANDEMIC

BY

VINCENT MARCUS GARCIA

A MASTER'S THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
SYSTEMS ENGINEERING

UNIVERSITY OF RHODE ISLAND
2021
MASTER OF SCIENCE
OF
VINCENT GARCIA

APPROVED:

Thesis Committee:

Major Professor    Valerie Maier-Speredelozzi
Thomas Wettergren
Norbert Mundorf
Brenton DeBoef
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2021
ABSTRACT

Bicycle sharing quickly became a popular mode of transportation in urban environments in the United States until the COVID-19 pandemic was officially announced. System providers have faced a severe loss in the number of individuals who used bicycle sharing systems (BSS), although cycling was recommended to engage social distancing. While other BSS met the ridership level of the previous year in the short term, the BSS Capital Bikeshare (CaBi) still can not build on prior-year numbers. Because today’s bicycle-sharing business is fiercely competitive, a high service level and customer satisfaction are keys to achieving a sustainable operation. This thesis aims to understand how usage patterns of the CaBi system changed in spatial and temporal aspects before and after the first twelve months of the COVID-19 pandemic outbreak on March 7th, 2020, which may affect the BSS’s processes.

A literature review was conducted to accomplish that goal, and data analysis approaches and clustering algorithms were applied to the time series data from March 2019 till March 2021. First, based on the literature review, the raw data was analyzed with basic statistical metrics to distinguish the number of trips and trip duration for both periods. Second, the member type of riders was analyzed, as well as the temporal distribution of trips. Applying the k-means clustering algorithm, stations with similar rental activities corresponding to the hour of the day and trip duration were categorized in the third step.

Overall, it can be said that a significant change in temporal usage patterns of the stations has been noticed, whereas the spatial clusters have not changed during the pandemic. Bicycle-sharing operators and urban planners can use those insights and the
aggregated list of factors influencing BSS use to enhance inventory management, dynamic bicycle relocation, dynamic demand management, locating new stations, or planning bicycle networks for cyclists and bicycle sharing riders.
ACKNOWLEDGMENTS

Desperate times call for extraordinary measures: In March 2020, a coronavirus, which turned later into a global pandemic, kept the world in suspense. Due to international travel restrictions, my plans to travel to the U.S. to participate in the exchange program of the University of Rhode Island (URI), U.S., and Technische Universität Braunschweig, Germany, for a Dual-Degree program were postponed. That is the reason why I want to thank the professors and organizers of the program in general for all of their understanding, patience, confidence, and flexibility they showed us at first.

Nonetheless, this thesis would not have been possible without the support of many people. Many thanks to my advisor Dr. Valerie Maier-Speredelozzi, who provided me invaluable with feedback on my analysis and framing, at times responding to E-Mails late at night or early in the morning.

Also, thank you to my committee members, Dr. Norbert Mundorf, Dr. Thomas Wettergren, and Professor Carolyn Thornber. Your encouraging words and thoughtful feedback in the proposal process have been very important to me and helped me develop a critical problem statement.

At the Technische Universität Braunschweig, I want to thank Dr. Christian Thiess and Dr. Thomas Spengler for assuming the role of the advisor and thus complement the dual-degree committee.

Lastly, and most importantly, I am grateful for my family´s and friend´s unconditional, unequivocal, and loving support.
TABLE OF CONTENTS

ABSTRACT .................................................................................................................. ii
ACKNOWLEDGMENTS ............................................................................................... iv
TABLE OF CONTENTS ................................................................................................. v
LIST OF TABLES ........................................................................................................ viii
LIST OF FIGURES ....................................................................................................... ix
LIST OF EQUATIONS ................................................................................................. xii
LIST OF ABBREVIATIONS ......................................................................................... xiii
CHAPTER 1 – Introduction ......................................................................................... 1
  1.1 Significance of Study ......................................................................................... 1

  1.2 Aim and Objectives ......................................................................................... 3

  1.3 Research Questions ......................................................................................... 4

CHAPTER 2 – Literature Review .............................................................................. 5
  2.1 Mobility ............................................................................................................ 5

  2.2 Transportation Activities in the United States ............................................ 6

  2.3 Bicycle Behavior ........................................................................................... 10

    2.3.1 Why Cycling? .......................................................................................... 11

    2.3.2 Factors Influencing Bicycle Use ............................................................ 12

    2.3.3 Travel Characteristics of Cyclists .......................................................... 16

  2.4 Bicycle Sharing ............................................................................................. 19

    2.4.1 History of Bike-Sharing .......................................................................... 19

    2.4.2 Differences of Public Bicycle Sharing Users and Cyclists ..................... 21

    2.4.3 Operational Difficulties for BSS ............................................................. 24

    2.4.4 Bike Sharing Systems in the United States .......................................... 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.5</td>
<td>Capital Bikeshare in Washington D.C.</td>
<td>27</td>
</tr>
<tr>
<td>2.5</td>
<td>COVID-19 Pandemic</td>
<td>30</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Development</td>
<td>31</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Human Mobility Dynamic During the COVID-19 Pandemic</td>
<td>33</td>
</tr>
<tr>
<td>2.6</td>
<td>Related Work</td>
<td>35</td>
</tr>
<tr>
<td>3.1</td>
<td>Process Model</td>
<td>38</td>
</tr>
<tr>
<td>3.2</td>
<td>Descriptive Statistics</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>Correlation Analysis</td>
<td>42</td>
</tr>
<tr>
<td>3.4</td>
<td>Unsupervised Learning</td>
<td>45</td>
</tr>
<tr>
<td>3.5</td>
<td>Cluster Algorithms</td>
<td>46</td>
</tr>
<tr>
<td>3.5.1</td>
<td>K-means Clustering</td>
<td>50</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Density-Based Clustering (DBSCAN)</td>
<td>51</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Finding Optimal Number of Clusters</td>
<td>53</td>
</tr>
<tr>
<td>3.6</td>
<td>Data Analysis Plan</td>
<td>54</td>
</tr>
<tr>
<td>3.7</td>
<td>Programming Language and Environment</td>
<td>56</td>
</tr>
<tr>
<td>4.1</td>
<td>Business Understanding</td>
<td>59</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Understanding</td>
<td>60</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Preliminary Raw Data Analysis</td>
<td>61</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Data Cleaning, Concatenating, and Preparation</td>
<td>63</td>
</tr>
</tbody>
</table>
CHAPTER 5 - Data Analysis

5.1 Descriptive Analysis

5.2 Correlation Analysis Preliminary Data Analysis

5.3 Evaluation of the Descriptive and Correlation Analysis

5.4 Modelling Cluster Algorithms

5.5 Analysis of Clusters

5.6 Evaluation of the Usage Patterns

CHAPTER 6 – Significance of the Results

6.1 Evaluation of the Research Questions

6.2 Bike Sharing Systems

6.3 University of Rhode Island

6.4 Discussion

6.5 Conclusion

APPENDICES

CODE

BIBLIOGRAPHY
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1: Proportions, Distance, and Duration of U.S. Bicycling and Bike Sharing Trips by Purpose</td>
<td>18</td>
</tr>
<tr>
<td>Table 2: Descriptive Statistics for P1 &amp; P2</td>
<td>63</td>
</tr>
<tr>
<td>Table 3: Attributes of the Trip History Data</td>
<td>65</td>
</tr>
<tr>
<td>Table 4: Summary Table for Trip Frequency Correlation</td>
<td>75</td>
</tr>
<tr>
<td>Table 5: Summary Table for Trip Duration Correlation</td>
<td>75</td>
</tr>
<tr>
<td>Table 6: Summary Table for Trip Frequency Regression Analysis According to Member Type</td>
<td>76</td>
</tr>
<tr>
<td>Table 7: Result plots of the k-means Cluster Analysis</td>
<td>89</td>
</tr>
<tr>
<td>Table 8: Capital Bikeshare Quick Facts (1/2)</td>
<td>105</td>
</tr>
<tr>
<td>Table 9: Capital Bikeshare Quick Facts (2/2)</td>
<td>106</td>
</tr>
<tr>
<td>Table 10: Factors Correlated with the Use of Bike-Sharing Stations (1/3)</td>
<td>107</td>
</tr>
<tr>
<td>Table 11: Factors Correlated with the Use of Bike-Sharing Stations (2/3)</td>
<td>108</td>
</tr>
<tr>
<td>Table 12: Factors Correlated with the Use of Bike-Sharing Stations (3/3)</td>
<td>109</td>
</tr>
<tr>
<td>Table 13: Most Used Origin-Destination Paths in P1</td>
<td>110</td>
</tr>
<tr>
<td>Table 14: Most Used Origin Stations in P1</td>
<td>110</td>
</tr>
<tr>
<td>Table 15: Most Used Destination Stations in P1</td>
<td>110</td>
</tr>
<tr>
<td>Table 16: Most Used Origin-Destination Paths in P2</td>
<td>111</td>
</tr>
<tr>
<td>Table 17: Most Used Origin Stations in P2</td>
<td>111</td>
</tr>
<tr>
<td>Table 18: Most Used Destination Stations in P2</td>
<td>111</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1: General framework of factory explaining bicycle use by Riedtield and Daniel (2004)</td>
<td>13</td>
</tr>
<tr>
<td>Figure 2: Comparison of Demographics of Washington D.C., area cyclists and short-term and annual CaBi members</td>
<td>22</td>
</tr>
<tr>
<td>Figure 3: Docking Stations of CaBi’s BSS</td>
<td>27</td>
</tr>
<tr>
<td>Figure 4: CaBis Trip Purposes By Home Jurisdiction</td>
<td>29</td>
</tr>
<tr>
<td>Figure 5: Phases of the CRISP-DM reference model</td>
<td>40</td>
</tr>
<tr>
<td>Figure 6: An Overview of Clustering Taxonomy</td>
<td>47</td>
</tr>
<tr>
<td>Figure 7: Methodological Approach of the Master’s Thesis</td>
<td>55</td>
</tr>
<tr>
<td>Figure 8: COVID-19 Cases and Monthly Ridership Level for P1 &amp; P2</td>
<td>62</td>
</tr>
<tr>
<td>Figure 9: Distribution of Duration Considering Member Type in P1</td>
<td>69</td>
</tr>
<tr>
<td>Figure 10: Distribution of Duration Considering Member Type in P2</td>
<td>70</td>
</tr>
<tr>
<td>Figure 11: Number of Rides per Month for P1 &amp; P2</td>
<td>71</td>
</tr>
<tr>
<td>Figure 12: Number of Rides Per Day for P1 and P2</td>
<td>72</td>
</tr>
<tr>
<td>Figure 13: Average Duration Per Day for P1 and P2</td>
<td>73</td>
</tr>
<tr>
<td>Figure 14: Hourly Number of Rides per Hour in P1 &amp; P2</td>
<td>74</td>
</tr>
<tr>
<td>Figure 15: Geospatial Clustering Map of P1</td>
<td>79</td>
</tr>
<tr>
<td>Figure 16: Geospatial Clustering Map of P2</td>
<td>81</td>
</tr>
<tr>
<td>Figure 17: Geospatial Map of Stations in C1 in P1</td>
<td>83</td>
</tr>
<tr>
<td>Figure 18: Geospatial Map of Stations in C2 in P1</td>
<td>84</td>
</tr>
<tr>
<td>Figure 19: Geospatial Map of Stations in C3 in P1</td>
<td>85</td>
</tr>
</tbody>
</table>
Figure 20: Geospatial Map of Stations in C1 in P2 ......................................................... 86
Figure 21: Geospatial Map of Stations in C2 in P2 ......................................................... 87
Figure 22: Geospatial Map of Stations in C3 in P2 ......................................................... 88
Figure 23: Population Density Washington D.C. (129) ................................................. 112
Figure 24: Employment Density Washington D.C. (129) .............................................. 113
Figure 25: Central Employment Area Washington D.C. (129) ....................................... 114
Figure 26: Great Streets and Transit Stations in Washington D.C. (129) ................. 115
Figure 27: Misplaced Bicycle Storage Rack at Keaney Road (Boss Ice Arena) ...... 116
Figure 28: Blocked Bicycle Storage Rack at Greenhouse Road............................... 117
Figure 29: Upper College Road: Lanterns, Power Lines, Parked Vehicles and Cracks
in the Asphalt Make Bicycle Traffic Problematic..................................................... 118
Figure 30: Closed Access to the Bike Path at Brookside Residence Hall (W Alumni
Avenue)..................................................................................................................... 119
Figure 31: Code for Concatenation (1/3) ................................................................. 120
Figure 32: Code for Concatenation (2/3) ................................................................. 121
Figure 33: Code for Concatenation (3/3) ................................................................. 122
Figure 34: Code for Descriptive Analysis (1/2)....................................................... 123
Figure 35: Code for Descriptive Analysis (2/2)....................................................... 124
Figure 36: Code for Correlation Analysis (1/5)....................................................... 125
Figure 37: Code for Correlation Analysis (2/5)....................................................... 126
Figure 38: Code for Correlation Analysis (3/5)....................................................... 127
Figure 39: Code for Correlation Analysis (4/5)....................................................... 128
Figure 40: Code for Correlation Analysis (5/5)....................................................... 129
Figure 41: Code for Cluster Analysis (1/7)................................................................. 130
Figure 42: Code for Cluster Analysis (2/7)................................................................. 131
Figure 43: Code for Cluster Analysis (3/7)................................................................. 132
Figure 44: Code for Cluster Analysis (4/7)................................................................. 133
Figure 45: Code for Cluster Analysis (5/7)................................................................. 134
Figure 46: Code for Cluster Analysis (6/7)................................................................. 135
Figure 47: Code for Cluster Analysis (7/7)................................................................. 136
# LIST OF EQUATIONS

<table>
<thead>
<tr>
<th>EQUATION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equation 1: Pearson´s Correlation Coefficient</td>
<td>43</td>
</tr>
<tr>
<td>Equation 2: Spearman´s Rho Coefficient</td>
<td>44</td>
</tr>
<tr>
<td>Equation 3: Kendal´s Tau</td>
<td>44</td>
</tr>
<tr>
<td>Equation 4: Kendal´s Tau Concordant and Discordant Pairs</td>
<td>45</td>
</tr>
<tr>
<td>Equation 5: k-means Approach</td>
<td>50</td>
</tr>
<tr>
<td>Equation 6: k-means Measuring Value</td>
<td>50</td>
</tr>
<tr>
<td>Equation 7: k-means Minimizing Centorids Distance</td>
<td>51</td>
</tr>
<tr>
<td>Equation 8: k-means Arithmetic Mean</td>
<td>51</td>
</tr>
<tr>
<td>Equation 9: DBSCAN Number of Samples Constraint</td>
<td>52</td>
</tr>
<tr>
<td>Equation 10: Optimal Number of Clusters - Silhouette Coefficient</td>
<td>54</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTS</td>
<td>Bureau of Transportations Statistics</td>
</tr>
<tr>
<td>CAD</td>
<td>Canadian Dollar</td>
</tr>
<tr>
<td>CRISP-DM</td>
<td>cross-industry standard process for data mining</td>
</tr>
<tr>
<td>CSV</td>
<td>comma separated values</td>
</tr>
<tr>
<td>E.G.</td>
<td>exempli gratia</td>
</tr>
<tr>
<td>EDA</td>
<td>explorative data analysis</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System Mapping</td>
</tr>
<tr>
<td>KTC</td>
<td>Kendall’s Tau coefficient</td>
</tr>
<tr>
<td>MGWR</td>
<td>Multi-scale Geographically Weighted Regression</td>
</tr>
<tr>
<td>NHTS</td>
<td>National Household Travel Survey</td>
</tr>
<tr>
<td>PCC</td>
<td>Pearson’s correlation coefficient</td>
</tr>
<tr>
<td>URI</td>
<td>University of Rhode Island</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>SOC</td>
<td>State-of-the-Commute</td>
</tr>
<tr>
<td>SPC</td>
<td>Spearman’s Rho coefficient</td>
</tr>
<tr>
<td>SEMMA</td>
<td>Sample, Explore, Modify, Model, Assess</td>
</tr>
<tr>
<td>TPMP</td>
<td>Transportation and Parking Master Plan</td>
</tr>
<tr>
<td>XLSX</td>
<td>XML Spreadsheet</td>
</tr>
</tbody>
</table>
CHAPTER 1 – Introduction

In this chapter information about the significance of the study, the aim and objective, and the research questions for the Master’s Thesis with the title *Spatio-Temporal Analysis of Usage Patterns in a Public Bike Sharing System* is provided.

1.1 Significance of Study

In recent years, the importance of mobility in different life aspects has increased and plays a impressive role in human beings’ daily lives. According to modern human evolution, the homo sapiens began 300,000 years ago, moving from southern to eastern Africa. Those moving patterns were evoked by factors such as inhospitable landscapes, rivalry, and food scarcities (I). In contrast to the prehistoric patterns, today's mobility is often driven by socio-economic factors such as wage imbalance, globalization, and dissimilarity in welfare and living conditions. This particular type of movement occurs daily and is also set in motion to complete leisure and social activities.

Scientists are working relentlessly to evolve sustainable solutions to meet human mobility needs in the long term. In the last decade, the scientific focus has been in most instances on electromobility, autonomous vehicles (A.V.), and Mobility as a Service (MaaS). Unlike the mechanisms of a combustion engine, alternative fuel and an electric vehicle emit minor harmful greenhouse gases and do not consume finite raw materials such as petroleum. Research, tests, and investigation on A.V.’s hope to improve traffic safety, accessibility, environmental impact, and at the same time aim to decrease costs in both the short- and long-term. This research focuses on transportation modes that are always available, on-demand, and tailored to each person's requirement, defined as
Mobility as a Service. By shifting from privately owned modes of transportation to service providers, customers of those platforms can use services like ride-hailing, ride-pooling, scooters, or even bike-sharing (2). Under the circumstances that 68 percent of the world's population is projected to live in an urbanized environment by the year 2050 (3), urban planners are challenged to meet the need for sustainable transportation and the increased population's mobility (4).

Therefore, it is not surprising that the so-called Bike Sharing-Systems (BSS) are of great interest to researchers and urban planners. Besides, public transportation systems benefit from implementing BSS docks next to stations to provide an excellent overall mobility system and solve the last mile problem (5). In this context, the last mile problem is to be understood so that the first and last "mile" of a public transport trip is one of the essential components of the overall quality experience by a transit user (6). The distance "mile" is not seen as a metric unit but as the first or last part of a journey, e.g., the distance to/ from the closest bus stop. The BSS can be used as an individual means of transport for short distances (5). While BSS were initially spatially embedded at urban public transport networks to solve the last mile in public transportation, they are found with a densely distributed station system network in larger cities (7). Moreover, BSS's provide citizens with access to a variety of activities like education, markets, employment, recreation, health care, and other services (4).

To ensure the best possible transportation mode, service providers are faced with crucial operational challenges. For instance, according to different decision-making levels, on an operational level, e.g., inventory management, dynamic bicycle relocation, and dynamic demand management notably impact a BSS’s success (8). In this context,
regular monitoring and predictive algorithms are needed for granular analysis of bicycle movements. Nevertheless, each BSS has unique spatio-temporal patterns, making generalization difficult between different cities (9). As a result, urban planners and systems engineers can use this knowledge for sustainable urban development and transportation services to make operational processes more efficient (4).

From these few examples, it should be clear that mobility has an enormous impact on human societies and the environment. An accurate quantitative description of bicycle usage patterns is fundamental to economically and sustainably operate BSS processes. Consequently, data analysis and clustering algorithms can gather insights into spatio-temporal usage patterns of BSS biking behavior.

1.2 Aim and Objectives

One such BSS in Washington D.C. is called *Capital Bikeshare* (CaBi). Whereas a general bicycle boom was provoked by the pandemic (10), and other BSS like the CITI Bike rebounded to pre-pandemic ridership levels quickly (11), CaBi still shows sparse usage numbers, 12 months after the American outbreak of COVID-19. Therefore, it should be of operational and scientific relevance to analyze how the pandemic impacted cycling behavior according to long-term changes in spatio-temporal patterns in Washington D.C.. Such time-series data is provided by the CaBi for research purposes, making this study possible. By obtaining insights into the system's usage patterns, it is intended to identify usage clusters, inequities, and utilization peaks. Depending on the stimulus, those behavioral changes in the system network may affect forecast models, relocation of bicycles, and/or ensure the availability of bikes, causing the decrease of
the usage numbers. A general statistical analysis, correlation analysis, and clustering algorithm will be applied to derive insights in the context of BSS providers to improve the station network or boost the general ridership level in the future.

1.3 Research Questions

First, it is hypothesized that the global COVID-19 pandemic may have caused changes in spatio-temporal behavioral patterns, affecting the utilization ratio in the context of the ridership level compared to previous years time period.

Second, based on the governmental proclamations, like the governor’s stay-at-home act and its resulting increase of remote work, the number of trips to high density employment areas has decreased and impacted the system’s network.

Third, the purpose of using BSS has shifted from those high dense employment areas to more recreational-friendly ones like parks and the suburbs.

Fourth, concerning the methodological approach, spatial and temporal changes in the usage behavior can be uncovered using clustering techniques like k-means or DBSCAN.

Lastly, a larger amount of clusters with more noise is expected during the COVID-19 phase, impeding the generation and cluster interpretation.

Generally, to make the comparison reasonable, it is assumed that external factors, like seasonality, weather impact, and social changes in the population, did not cause a bias when observing the two periods.
CHAPTER 2 – Literature Review

In this chapter, comprehensive insights into the general mobility patterns, transportation systems, bicycle behavior, bicycle-sharing systems, and the development of the COVID-19 pandemic are described.

2.1 Mobility

The Mode Share of the United States (U.S.) demonstrates the past nationwide development from a traditional, compact, and walkable cities toward more dispersed, automobile-oriented communities (12). Thereby, 59.5 percent of vehicle trips were five miles (8.1 kilometers) or less in length (13). This trip length tends to be a reasonable distance to see the bicycle as an alternative to the car (14). However, less than one percent of the overall number of trips people made were on bicycles in 2017 (13). This raises the question of whether, and to what extent, people are willing to ride the bicycle as their mode of travel. This chapter aims to provide a general understanding of transportation and review the existing research on cycling and critical bike-sharing characteristics, factors that affect the mode choice of cycling, motivations, and behavior in an urban environment. In the context of mobility, a lack of detailed studies on everyday mobility data was identified by Schonfelder and Axhausen (2016) (15). More specifically, Kuzmyak and Dill (2012) claim that non-motorized volume data are missing, like walking and biking and their varieties. The authors argue that "transportation agencies have well-established procedures for collecting, summarizing, and disseminating motor vehicle traffic volumes, but these procedures do not generally include systemwide pedestrian and bicycle volume data" (12).
In contrast to motorized transportation, there are only limited regional studies on non-motorized transportation. In the context of this research, it was decided to make deliberate use of grey literature. This non-peer-reviewed literature comprises theses, dissertations, reports, conference proceedings, working papers, leaflets, media reports, and newspapers. It may cause potential bias, which will be used with caution in the literature review.

2.2 Transportation Activities in the United States

The U.S. nationwide survey database about public travel behavior is the National Household Travel Survey (NHTS). It is conducted by the Bureau of Transportation Statistics (B.T.S.) every six years and analyzes trends in personal and household travel behavior. This travel behavior includes trips made by all modes of travel and all purposes to explore topics on "traffic safety, congestion, the environment, energy consumption, demographic trends, bicycle and pedestrian studies, the mobility sharing economy, and transit planning for planning and policy applications" (13). In 2017, the B.T.S. added specific variables addressing the needs of changing travel patterns and behavior, mainly for walking and bike-sharing transportation modes (13).

According to the latest NHTS in 2017, Americans took 411 billion daily trips per year. Applied to one person, 4.1 trips per person per day were made on average. Relatively, the total length of one person's daily trips was 40 miles (63.37 kilometers) on average. In total, the nationwide travel distance amounts to 11 billion miles (17.70 kilometers) each day. To cover this distance, drivers spent, on average, 55 minutes behind the wheel. Most of those trips were made on Fridays (16 percent), whereas on
Sundays, the fewest daily trips were made (12 percent). Transportation is the movement of a person or cargo from one point to another. Hence, daily trips are the number of transportation modes a person chooses in a single day to move spatially (16).

In addition to trips per day, the modal split is used in literature to describe the percentage of how many people use a particular type of transportation to make trips. Chiefly, the modes of a private vehicle, carpool, public transit, pedestrian, and cycling are differentiated. Using the modal split indicator for different temporal dimensions, distinctions can be made on the purpose of the travel (e.g., commute, school, recreation, and personal/family trips) (17).

As stated in the NHTS, in the U.S., 83 percent of all daily trips in 2017 were taken by car, 11 percent by walking, 5 percent by public transportation, 0.9 percent by bicycle, and 0.2 percent by long-distance trains (13). In the context of the modal split, differences in travel behavior derive considerable from public policy differences, variations in transit subsidies, land use controls, and housing programs. Even with similar income, technology, and urbanization, the transportation behavior among a country or region can vary widely (17), so the spatial research area of this Master's Thesis will be limited in the following section.

On a state-base, the State-of-the-Commute (S.O.C.) survey provides information about commuting patterns and prevalent attitudes about transportation services every three years. The S.O.C. was completed with a total of 8,246 interviews for the Commuter Connections program of the Metropolitan Washington Council of Governments, most recently in 2019 (18).
Additionally, the Regional Travel Survey (R.T.S.) of the Metropolitan Washington Council of Governments maintains a database of the daily travel patterns of citizens in Washington D.C. and adjacent areas from 2017. The survey is conducted once in a decade and collects travel and demographic data from 15,000 randomly selected representative samples of households. The available data is primarily used by stakeholders to estimate, align, and validate regional travel demand models. It was also used in the past to forecast travel demand and analyze the air quality conformity of the long-range transportation plan (19).

Looking at the metropolitan area of Washington D.C., the distribution of travel modes deviates from the national average. According to the S.O.C. Survey, the share of weekly commuting trips were made by 57.2 percent driving alone, 18.2 percent taking the train, 5.9 percent taking the bus, 4.6 using Carpool/Vanpool, 3.3 percent walking or biking, 1.1 percent taking taxi or ride-hail and 9.7 percent compressing work schedule or telework (18). In contrast to this, the results of the R.T.S. indicate that 41 percent drove alone, 38 percent accompanied others, 3.6 percent used rail transit, 2.0 percent used the bus, 9.3 percent walked, 1.3 percent cycled, 1.0 percent used taxi/ ride-hail, 3.5 percent used the school bus and 0.4 percent used other modes in 2017/18 (19).

Accumulated in the S.O.C. Survey report, the results represent that the average travel time has increased continuously, even though the average travel distance has remained approximately equal. In 2019, the average one-way travel time lasted 43 minutes for all transportation modes, compared to 39 minutes in 2016 and 36 minutes in 2012. Contemporaneously, the average travel distance remained at 17.1 miles (27.5 kilometers), almost the same (18).
Comparing the TPB’s R.T.S. Data of 2017/2018 to the data collected ten years earlier, four findings should be highlighted in the context of this study: First, reflecting a nationwide trend, the number of weekday trips had already decreased before COVID-19 for the TPB research area in Washington with a daily trip rate of 8.3 compared to 8.9 in 2007/2008. It is argued that the rise of online shopping and delivery services partly replaces trips to stores and restaurants. Second, the car share of commute trips has decreased from 78.1 percent in 2007/2008 to 72.3 percent in 2017/2018. On the other hand, an increase in commuting trips of the modes taking the bus, walking, biking, and taking the taxi/ride-hail trips was observed. Third, the share of trips by rail has declined since 2007/2008. Both changes have occurred due to variations and development in land use and investment in transit and walk/bike infrastructure. Lastly, the share of all trips of bicycling has doubled regionwide since 2007/2008. This increase of trips is seen in investments of the greater Washington region in bicycle infrastructure, implementing bike paths, connecting multi-use trains, and launching the bicycle sharing service Capital Bikeshare in 2010 (19).

Despite the differences in the mode-shares, it should be kept in mind that the interviews of the surveys were conducted in different manners and years. Likewise, Broach et al. (2009) discussed that the data acquisition by travel behavior surveys might bias survey participants and under- or overestimate their behavior. Additionally, respondents may relate their behavior with the given choices or think that the outcome may influence the policy and cause a bias in the dataset (20).

Furthermore, in the different research regions of Washington D.C. (core region, inner suburb, and outer suburb), differences in mode shares became evident in the R.T.S.
Whereas in the outer suburb, 0.3 percent of all trips were made by bicycle, in the core region, 3.8 percent were covered by the same mode in 2017/18 (19). This exemplifies the statement of Pucher (1988) that the travel behavior and mode choice vary among a city. For this reason, the findings of bicycling behavior will be reviewed in the following.

2.3 Bicycle Behavior

"[...] it would be naïve to suppose that all practitioners of the same sport (or any other practice) confer the same meaning on their practice or even, strictly speaking, that they are practicing the same practice."

(Pierre Bourdieu, 1984 page 209-211)

From the social science perspective, riding a bicycle for practice is often connotated by people's own experiences. Either cycling anecdotes, stories, fears, or theories make universality difficult for researchers in complexity and diversity. As a result, Rosen et al. claim that bicycles and cycling are ubiquitous in our daily social life but relatively unthought from the social science perspective. The authors identified four categories for cycling in the academic context. First, from a historical perspective, cycling has been explored by analyzing the international history in cycle technology and its manufacturers, cycle sport, and cycling innovators. Second, sociologists of sport inspected traditional cycling, e.g., track and road racing, to more recent developments such as mountain biking or triathlons. Third, engineers, designers, and planners have also taken an interest in cycling. Most studies focused on increasing cycling levels as a mode of transportation in the built environment and design aspects. Fourth, attention has come from the medical approaches to cycling regarding the positive health effect
and the analyses of accident data, health promotion, and protective headgear (22). The following sub-section addresses the factors influencing cycling behavior as a mode of transportation.

2.3.1 Why Cycling?

Pucher and Buehler (2012) recognize that cycling bears many good reasons as an everyday transportation mode. They state that riding the bike is viewed - despite all the different perspectives - as an inexpensive transportation mode, which comes with benefits for personal health and wellbeing. Furthermore, it is seen by the majority as a solution for climate change, among other arguments. Analyzing German, Danish, and Dutch cycling behavior characteristics, the authors show that cycling is a healthy, practical, convenient, attractive, and safe means of transportation (23). Other literature examined that there is scientific evidence that the health benefits exceed the traffic dangers of cycling and yield economic benefits for individuals and society (24, 25).

Further, cyclists are faster than motorized alternatives in the rush hour at a distance of fewer than 5 miles in urban areas (14, 24, 26). While the average speed of motorized vehicles considering traffic and all other factors was 10-15 kilometers per hour, in the study by Jensen et al. (2013), bicyclists were riding at an average velocity of above 15 kilometers per hour according to the G.P.S. data from a BSS (27). This illustrates the practical and convenient feasibility of reducing driving through cycling. Additionally, the majority benefits from the cost and health advantages.

Overall, climate change challenges and achieving a sustainable and healthy way of living demonstrate that cycling as an everyday transportation mode can alleviate the negative impact which derives from motorized transportation modes (23).
2.3.2 Factors Influencing Bicycle Use

In the reviewed literature, multiple factors like activity, purpose and time, population demographic, geographic location, topography, climate, weather, facilities, urban design, safety, politics and transport policies were revealed to influence cycling behavior \((22, 23, 25, 28–30)\). Those factors could be categorized into either quantitative (objective) or qualitative (subjective) factors. Because this empirical research focuses on spatial and temporal behavior patterns, both categories are further discussed in the following section.

A methodological approach to categorize the various factors is provided by the journal article of Rietfeld and Daniel (2004). In their framework, to determine cycling behavior from the inter-municipally perspective (c.f. Figure 1), the authors present potentially thematic areas that influence bicycle use. While determining factors that influence cycling behavior, they notice that subjective factors may also affect the mode choice. For example, the authors explain that the municipal differences in mode shares exist due to the various perceptions of the bicycle as a mode of transport. Next to individual factors, the researchers also group social-cultural factors like age, gender, income, activity, ethnicity, and political lineage in one theme. In addition, the authors consider the cost of transportation as another themed area. Here, they distinguish between cost factors for cycling and those for other modes of transport. Lastly, they focus on a theme area for initiatives and policies at the local authority level \((30)\).
Based on the framework of Rietveld and Daniel, among other literature, the authors Fernández-Heredia and Monzón (2010) re-categorize the factors for an online survey at the University of Madrid (n = 3,000) into three major groups. First, objective factors can be analyzed without any direct interaction with the cyclist because Geographic Information System Mapping (G.I.S.) comprises structural and environmental factors. To further differentiate objective factors, the authors divide the group into sub-factors that are personal in nature or collective and associated with the

Figure 1: General framework of factory explaining bicycle use by Riedtfeld and Daniel (2004) (30)
direct environment. Second, subjective and evaluative factors are named, including perceptions of cyclists and other intangible variables, which are measurable to some extent. Last, individual features such as age or income level related to socio-demographic characteristics are considered (31). Based on previous studies, it is evident that the influence of subjective factors in human activities' spatial existence is equally essential to the objective measures of human temporal behavior (32).

In addition, factors have been identified in the literature that can be measured objectively on the one hand but perceived subjectively on the other. Factors like journey purpose, time, distance, mode of transport, weather conditions, cycle network, and topography are factors that influence cycling behavior from an individual perspective. The authors Fernández-Heredia and Monzón (2010) affirm that the purpose of the travel destination and travel time is essential when there is a decision to make in the context of the travel mode. Whereas the flexibility offered by bicycles about timetables and frequencies in public transportation is a plus, long-distance trips may be covered faster by motorized mode shares (31). Rietveld and Daniel (2004) exhibit that travel time and the context of the route decision affect bicycle behavior. They give the example that directness of routes and heavy traffic volume affect the bicycle behavior route choice.

Additionally, the purpose of the cyclist's journey and physical condition also impact the route choice. Heinen et al. (2009) established evidence from an internet survey in the Netherlands (n = 4,299) which shows the frequency of workdays, locations, hours, and possessing a driver's license did not influence bicycle mode choice that the car availability did. Furthermore, they examined how part-time, full-time, and non-cyclists as groups are affected by work-related factors like working time, kind of
clothing worn, and colleagues' perception (33). The study of Stinson and Bhat (2004) reports the difficulties in implementing car use reduction policies in the North American context and suggests why the U.S. has such a low level of cycling uptake (34).

Other objective factors identified in the literature are topographic and meteorological characteristics (12). While steep terrains harm cycling behavior, moderate hilly environments were preferred by both non-commuter cyclists (81 percent) and commuter cyclists (63 percent) over a flat topography in the study of Sener et al. (2009) (n= 1621 respondents) (35). Moreover, Cerero and Duncan (2003) stated that slopes, in general, are essential for the non-motorized mode choice in the context of stated preferences (36). In contrast to those studies, G.P.S. data from cyclists in Portland, Oregon, indicate that the typical cyclist would travel 27 percent farther to avoid each 1 percent of additional average upslope (20). Even if the definition of hilliness was not clearly indicated in all studies, the results would reiterate the essence of considering the purpose of bicycle trips and their research. Next to the topographic characteristics, Rietveld and Daniel (2004) reveal that wind affects the effort, pleasure, and comfort of cycling. It is more important than the factor rain due to the spatial variability at coastal areas in the study area (30). Nevertheless, the study of Kuzmyak and Dill (2012) claims that the most volatile changes in bicycle activities are due to acute weather events (12).

Another set of factors that is considered in the research is the transport infrastructure. From research around motorized modes, it is well acknowledged that physical infrastructure is an essential requirement. Mackett and Brown (2011) claim in their literature review that the accessibility to work increases by adequate physical infrastructure, such as a transport network along with other services, and thereby health
benefits and improved community reconciliation will be achieved (37). In their journal article, Pucher and Buehler compared in their journal article the U.S. walking and bicycling experience with that in Europe. The authors identified high-quality and coordinated travel networks for cyclists to positively impact the high European non-motorized modes. A high level of connectivity network, in conjunction with a mixed-use design of communities to allow direct and convenient non-motorized transportation, is evident. Likewise, public policies and attitudes that support cycling were identified as a practical approach to popularize cycling (38).

Furthermore, higher rates for non-motorized modes were identified in environments with compact, mixed-use area settings. Kuzmyak and Dill (2012) argue that households in mixed-land-use areas own fewer vehicles per person, make more trips to nearby destinations, and are more likely to use transit for trips farther away. Additionally, a compact mixed-land-use area as a destination is more likely to attract non-motorized traffic modes (12).

Despite all of the above discussed studies, it is not easy to highlight any particular factor because, in consensus, the mix is crucial. For a better understanding, the travel characteristics of cyclists will be reviewed.

2.3.3 Travel Characteristics of Cyclists

At the international scale, data about the mode share of cycling is available for most economically prosperous societies (22). According to the latest short distance passenger mobility survey in Europe, NHTS, and Statistics Canada, the mode share by cycling varies from very low (< 5 percent) in Australia, Canada, France, New Zealand, Spain, the United Kingdom, and the U.S., to low (6-10 percent) in Austria, Finland, and
Germany, to moderate (11-25 percent) in Denmark, Sweden, Switzerland, to relatively high (>26 percent) in the Netherlands (13, 39, 40).

The NHTS revealed that the most frequent travelers by non-motorized modes were between 40 and 64 years old (34 percent). An additional 24 percent were between the ages 25 and 34. Non-motorized travelers ages 5 to 15 (16 percent), 65+ (14 percent), and 16 to 24 (12 percent) were the minor groups (13).

Among other things, walking, biking, and their variations are considered as non-motorized traffic and include trips with exercise, recreation, shopping, and commuting as their purpose (41). Unlike other non-motorized mode shares, a gender difference across all age groups was identified in cycling. Males are two to four times more likely to ride the bike than females, depending on age (13). According to Kuzmyak et al., cycling is mainly consistent across income rates in the U.S., with the highest rate of 1.3 percent of daily person trips in the $20,000 to $40,000 range, 0.9 in the $75,000 to $99,000 range, and 1.1 percent in all other groups.

Regarding the education level of cyclists, the authors state that the highest rate of daily person trips by bike are among people without a high school diploma and people with professional or graduate degrees (both 1 percent), and the lowest rates among people with a high school diploma, G.E.D., a college or associate degree (12). Households in which the number of licensed drivers outnumbered vehicles available tend to use non-motorized modes more often in their daily trips. As reported by the short trip analysis of the League of American Bicyclists, 50 percent of bicycle trips are 3 miles or less, 13 percent between 3.1 and 5 miles, and 37 percent more than 5 miles (42). Based on the NHTS in 2009, social and other recreational activities, including exercise,
playing sports, going out for entertainment, visiting a public place, social event, and
getting or eating a meal, coffee, or snack, were the most prominent purposes for taking
the bicycle. Additionally, commuting to or from work had a percentage of 10.9 percent
and was associated with the most extended trip length of 3.8 miles, whereas shopping
(9.9 percent of trips) and visiting friends or relatives (13.0 percent) had the shortest
distance with 1.3 miles and 1 mile, respectively. The proportions, length, and duration
of American bicycling trips are summarized in Table 1.

Table 1: Proportions, Distance, and Duration of U.S. Bicycling and Bike Sharing Trips
by Purpose

<table>
<thead>
<tr>
<th>Trip purposes:</th>
<th>Bicycle</th>
<th></th>
<th>Bike Sharing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent of Trips</td>
<td>Average Trip Length (miles)</td>
<td>Average Travel Time (minutes)</td>
<td>Percent of Trips</td>
</tr>
<tr>
<td>To or from work</td>
<td>10.9*</td>
<td>3.8*</td>
<td>21.2*</td>
<td>38**</td>
</tr>
<tr>
<td>Work-related business</td>
<td>1.8*</td>
<td>3.3*</td>
<td>21.7*</td>
<td>6**</td>
</tr>
<tr>
<td>School or Church</td>
<td>6.0*</td>
<td>1.6*</td>
<td>15.2*</td>
<td>N.A.**</td>
</tr>
<tr>
<td>Shopping</td>
<td>9.8*</td>
<td>1.3*</td>
<td>14.0*</td>
<td>4**</td>
</tr>
<tr>
<td>Other family or personal business/social</td>
<td>8.2*</td>
<td>1.4*</td>
<td>15.5*</td>
<td>19**</td>
</tr>
<tr>
<td>Medical or dental</td>
<td>0.2*</td>
<td>2.2*</td>
<td>26.0*</td>
<td>N.A.**</td>
</tr>
<tr>
<td>Vacation</td>
<td>2.1*</td>
<td>2.4*</td>
<td>21.0*</td>
<td>N.A.**</td>
</tr>
<tr>
<td>Visit friends or relatives</td>
<td>13*</td>
<td>1.0*</td>
<td>13.9*</td>
<td>N.A.**</td>
</tr>
<tr>
<td>Other social creational/ recreation</td>
<td>47.3*</td>
<td>2.6*</td>
<td>22.5*</td>
<td>28**</td>
</tr>
<tr>
<td>Other</td>
<td>0.1*</td>
<td>2.3*</td>
<td>16.0*</td>
<td>5**</td>
</tr>
<tr>
<td>Refused or not available</td>
<td>0.8*</td>
<td>2.7*</td>
<td>25.7*</td>
<td>N.A.**</td>
</tr>
<tr>
<td>All purposes</td>
<td>100.0*</td>
<td>2.3*</td>
<td>19.4*</td>
<td>100.0**</td>
</tr>
</tbody>
</table>

Data retrieved from: * Kuzmyak and Dill (2012) (12); **Shaheen et al. (2012) (43); ***Kou, Cai (2019)(44)
The authors, Moudon et al., claim that bicycling is one of the underutilized modes of travel. The high percentages of higher distances suggest that bicycling is commonly used as an exercise or recreational activity among males and younger adults, which are already physically active (26).

2.4 Bicycle Sharing

In the last decade, bike-sharing – networks of unattended locations for short-term bicycle rentals - has become more popular. The bike-share consultancy PBSC tracked 2,110 different bike-sharing systems (BSS) with a total of 17,792,000 bicycles in their fleets worldwide as of the year 2020 (45). As a result, BSS has also become of great interest to researchers around the globe. This sub-section begins with a review of the historical development of BSS, followed by a presentation of the American systems, the strategic, tactical, and operational difficulties of operating, and finally, a description of the BSS Capital Bikeshare in Washington D.C..

2.4.1 History of Bike-Sharing

Over the past 60 years, there have been four generations of bike-sharing systems (BSS). These differ from each other in terms of technical maturity, logistical organization, conceptual orientation, and transport policy relevance.

The earliest version of BSS was implemented in Copenhagen, Denmark, in 1965. The Witte Fietsenplan (Danish for white bikes) were a dozen white-painted bicycles left in public for everyone to use. The idea contributor Luud Schimmelpennik considered the system not to be a business offering but a political statement against the rising number of cars in the city and the air pollution that came with it. Because most white
bikes were stolen, vandalized, or confiscated by the police, the system was shut down within days (46).

Problems raised by the previous generation let the government and the City Bike Foundation of Copenhagen, Denmark launch their system, so-called Bycyklen. This system is seen in literature as the starting point for the second generation because the bike used in the BSS were equipped with solid rubber tires and wheels with advertising plates for intense utilitarian use. Further, they could be picked up and returned at a specific location, so-called stations, in the central city with a coin deposit, which was refunded upon bicycle return. Nowadays, the system still exists and is called City Bike. Coin deposit and bicycle stations created a much more reliable network. Nevertheless, such an infrastructure resulted in a high administrative effort, anonymity, and low cost, leading people to steal bicycles (47).

The beginning of the third generation is represented by the newly developed IT-based technology and more modern design of bicycles. Information-based systems integrated automatic bike locks, RFID smart cards, and GPS tracking devices, linking them to web-based applications. Users were now required to complete a registration before starting a ride, either through credit card details or personal information. In addition, the rental bikes were equipped with a user interface to automate the rental process and make it more user-friendly. Driven by the localization of the rental bikes in real-time, a dock-less variant was also able to establish itself in the BSSs (45).

From 2013, the fourth generation of public BSSs was characterized by tariff integration into the public transport system. BSSs that were integrated into the public transport system could now be rented with a shared RFID card and at the same time
gave the right to use the public transport system. One example is the Canadian BIXI system, which integrated BSSs with public transport and car-sharing and allowed them to be used via a single card (47). Next to that integration, DeMaio (2009) identifies a technical advancement at the rented bicycle. So-called pedelecs (electrical pedal assistance) were more often integrated into the system. Additionally, the technological enhancement enabled a free-floating system, in which users were able to return the bicycle anywhere in a predefined geofence, and renting stations were made unnecessary (46). However, this system led to theft and the destruction of the bikes (48). Another problem for municipalities was that the users had not parked the bicycle following the rules. Due to the disorderly parked bicycles, which often blocked the way, several cities banned this type of BSS (49).

2.4.2 Differences of Public Bicycle Sharing Users and Cyclists

Even though the number of BSSs has grown exponentially worldwide, there is a lack of literature in which user behavior insights associated with regular cyclists can be transferred to bicycle-sharing users (50). Whereas some researchers generalize that there are slight or no differences between the two (43, 51), other studies have shown that BSS users tend to differ from regular cyclists (50, 51). The following section highlights observations and hypotheses in the literature that deviate from previous findings in chapter 2.3.2 regarding demographics, purposes, and influencing factors.

First, the researchers Buck et al. (2013) identified different socio-economic characteristics in their travel survey of area cyclists in Washington D.C., short-term Capital Bikeshare (CaBi) users, and annual CaBi members. The data was originated from the regional household travel survey of cyclists in 2007-2008, the Capital
**Bikeshare Casual User Survey** in 2011, and the **Capital Bikeshare Casual User Member Survey** in 2011. The results are summarized in Figure 2. The study indicates that CaBi short-term users and members are more likely to be women, be younger, have lower income, and are less likely to own a bicycle or motorized vehicle. Additionally, the author suggests that the general CaBi user is more likely to make utilitarian trips than the area cyclists of Washington D.C. (50).

![Comparison of Demographic Variables of Area Cyclists and CaBi's Short-term and Long-term users](image)

**Figure 2: Comparison of Demographics of Washington D.C., area cyclists and short-term and annual CaBi members** (50)

Second, researchers have documented that BSS users' primary trip purpose is commuting to either work or school (51), while area cyclists cycle for recreational purposes (12). It should be noted that the bicycle survey is justified by data from the NHTS (entire U.S.) in 2009, and the BSS user survey is related to results from CaBi's study in Washington D.C. in 2011. However, the second and third most common trips in BSS were for social and recreational purposes. Other trip purposes seem relatively
less important (51). Besides that, BSS commuter users are likely to ride less distance than area cyclist commuters (c.f. Table 1) (44). Besides the trip length, the trip type also varies. Next to round-trips, users of BSS also tend to use the system for point-to-point trips (43). This reinforces the hypothesis that bike-sharing provides an essential connection to the public transportation network and complements the system's deficits in infrastructure (4).

Third, because factors like social norms have a strong influence on the mode share related to cycling (52), bike share programs help normalize the positive image of cycling (53). In this context, the spatial stations are particularly conspicuous and contribute to other variables influencing the use of BSS (51).

In order to attract the most substantial number of users that may rent a bike, stations should be located in the closest possible locations to potential users to reach maximum coverage. In both activities, returning and renting a bicycle, access to the nearest station is desirable, according to Shaheen et al. (2011).

The researchers also identified a willingness of users to walk a distance of approximately 300 meters maximum to still be likely to use a BSS (43). Other researchers surveyed an adequate distance of willingness to walk to the next station of less than 200 meters (54, 55) to 800 meters (56). The willingness to walk to the next station is a much discussed topic in the literature.

A list of variables correlated with BSS use has been summarized in the appendices in Table 10 - Table 12.
2.4.3 Operational Difficulties for BSS

With the continuous expansion, bicycle-sharing service planning problems (BSSPP) have also risen. Hence, the number of publications facing design, operation, and management issues has grown. According to planning decision-making levels, the authors Shui and Szeto (2020) identified and classified eight BSSPP in their literature review. First, on a strategic level, the authors typify long-term decisions related to the BSS infrastructure in both non- and existing networks and the total bicycle inventories. Decisions about a bikeway network, bicycle station, and fleet-sized design must be made on that level. Next, medium-term decisions, which help maintain the performance and efficient use of existing resources in the BSS, are listed at the tactical decision level. That includes static bicycle relocation like vehicle routing problems and static demand management like demand regulation strategies. Lastly, on an operational level, short-term daily operation decisions, such as inventory level management, dynamic bicycle relocation, and dynamic demand management, are made (8).

A common denominator that can be deduced from the described decision levels is the general availability of bicycles. According to Vogel et al. (2011), the overall objective should therefore be for a BSS provider to have as many bicycles as possible available close to the customer (57). In another study by Vogel et al. (2011), the researchers claim that the strategic network design and the operational repositioning of bikes are the main logistical measures for antagonizing imbalances (58). In general, spatio-temporal analyses exposed in literature dependencies in dynamic movements and revealed different usage patterns from the analysis of ridership data (59–61). Next to the design and repositioning process, insights into trip history have been used to
overcome imbalances in the system and/or to support location decisions (62). As a result, investigating the usage patterns of BSS users is of key importance to understanding cyclists' behavior and, consequently, optimizing BSS. Additionally, the researchers Hu et al. (2019) claim that the knowledge of customer patterns is crucial for assuring the sustainable operation of a BSS in general (63).

2.4.4 Bike Sharing Systems in the United States

In 2008, Washington D.C. was the first jurisdiction in the United States to implement an IT-based BSS. The BSS, called Smartbike DC, started with 120 bicycles at ten different stations and was replaced by BSS Capital Bikeshare (CaBi) in 2010 (64). In the same year, Denver, Colorado, Minneapolis, and Minnesota followed with public-private partnership-owned BSS. Since then, according to Shaheen, bike-sharing has accomplished popularity partially due to the increased concern about environmental sustainability, together with the success of previous systems (43).

Based on the data gathered in the report Bicycle-Sharing Systems across the United States of America, by the end of 2018, 248 municipalities had implemented active BSS in their jurisdiction. In total, 145 were identified as station-based or hybrid and 103 as dock-less systems. However, according to the authors, there can be more than one BSS in a municipality (either cities or counties). On the one hand, they cite the example of Washington D.C., in which, in addition to CaBi, other varieties of BSS from different operators exist. On the other hand, some BSSs, like the system in New York, cover multiple neighboring cities and counties (65). Thus, the absolute number of BSSs may vary.
Alongside the report of the Pan American Health Organization, the author Paul de Maio began to document the global growth of BSS in a Blog in 2007. He states that "[…] the task soon became overwhelming as the number of new bike-share services that had been launching at a frequency of once every couple of weeks grew to once each day" (45).

The Blog BEYOND DC implemented a similar tracking system, which tracked the number of BSSs once per year and ranked every city with a bike-sharing network in order of the number of stations. The database was last updated in 2016 and recorded 119 different cities with BSS and a nationwide total of 4,789 stations/hubs. At the time, CityBike in Manhattan, New York, was the largest BSS with 645 stations in the U.S., followed by Chicago (581 stations) and Washington D.C. (437 stations) (66).

In contrast to those figures, the Bureau of Transportation Statistics monitored 103 docked BSS in the U.S. in 2019. On the one hand, from 2015 to 2019, the number of docked BSS has nearly doubled. On the other hand, it declined from 103 BSS in 2019 to 66 in 2020, as many BSS had to close permanently following the temporary suspension of enterprise due to COVID-19, according to the BTS (11).

Besides the extensive tracking of BSS, the literature claims that most systems are built on lessons learned from Europe and Canada due to the young modern American BSS program history. Despite the lack of valid data tracking and characterizing the BSS program in the U.S., an increased number of governmental guidance reports of practice and implementation was identified (23, 25, 43, 65).
2.4.5 Capital Bikeshare in Washington D.C.

As expressed before, BSS CaBi replaced SmartBike in the year 2010. Hence, Washington D.C. was the first jurisdiction that implemented a third-generation public BSS with a cooperative agreement with Clear Channel Outdoor’s advertising company in the U.S. (64). The program started with 400 bicycles at 49 stations and grew until the end of 2010 to 1000 bicycles at 114 stations (67). Today, the system includes more than 4,500 bicycles at 527 stations and serves an area of 140 square miles (68).

![Docking Stations of CaBi’s BSS](image)

*Figure 3: Docking Stations of CaBi’s BSS*

According to the *American Community Survey* in 2012, CaBi increased the share of bicycle commuters in Washington D.C. from 3.1 percent in 2010 up to 4.1 percent in 2012 (67). More recent data from the *Washington Post-Scholar School Poll* indicates that the percentage of bicycle trips of CaBi users is 16 percent of the share of bicycle trips in Washington D.C. (69). As of May 2019, 23.6 million trips were generated, the
busiest day being April 14th, 2018, with 19,703 trips. At the time, the three-speed bicycle fleet included more than 4,500 bikes. In the CaBi system, bikes can be taken from and returned at any station. The system is available to the riders 365 days a year, 24 hours a day, and seven days a week. To participate in the system, registration is required, in which a differentiation is made between casual riders and registered members. The distinction is noticeable at different price scales, nonetheless, trips less than 30 minutes are included in basic fees. For trips longer than 30 minutes, usage fees increase incrementally with the trip duration (68).

Next to the trip history data, CaBi provides reports of their biannual member surveys. With these surveys, CaBi assesses how people use the system and the impact of the BSS on their lives and the community (70). Contrary to the trip history data, which still gets updated, the member survey has not been continued or published since 2016.

Nonetheless, looking at the Capital Bikeshare Member Survey Report (n = 5,646) of the year 2016, two-thirds (65 percent) of the respondents reported that the primary purpose of the BSS was for commuting to work. Another 55 percent of respondents said social/entertainment was a primary trip purpose like personal appointments (42 percent) and shopping/errands (40 percent). In this survey, multiple responses were permitted. As a result, it is difficult to compare the report findings with the survey of Shaheen et al. (2012). Nevertheless, the report illustrates that the BSS serves both work-related and personal travel needs and differs in the different jurisdictions (c.f. Figure 4).
Figure 4: CaBis Trip Purposes By Home Jurisdiction (71)

Next to the purpose of use, 56 percent of the respondents claim that they used the CaBi system because it was either faster or an easier way to get to their destination. Additionally, another 57 percent answered the question of their motive for joining CaBis BSS because it enables one-way travel options, and another 69 percent because it is a fun way to travel.

In respect to the user demographics and membership characteristics, compared to the SOC from the same year, members were on average, younger (under 35 years), slightly more likely to be male (58 percent male; 42 percent female), much more likely to be white (80 percent), less wealthy than the regional employee population, and much more likely to live and work in the urban core region of Washington D.C., Arlington County, Virginia, or Alexandria, Virginia. According to the report's use distribution, 14
trips per user per month were made on average. It should be noted that the share of annual member respondents in this survey was 82 percent, 9 percent casual riders, and another 9 percent former customers, which may have caused a bias in the survey results.

Lastly, according to the report, nine in ten CaBi members would increase their BSS usage if the service were expanded and/or other service enhancements would be made (71).

More recently, the provider of CaBi was forced with the spread of the global COVID-19 pandemic. According to the Bureau of Transportation Statistics, the ridership declined by 56 percent in the COVID-19 months of March through May 2020 (72). While other BSSs are already matching and far surpassing last year’s numbers, CaBi is still experiencing a declining number of trips generated, compared to the previous year, before COVID-19 (11). The outdated understanding of the usage patterns, new database and the situation situation of the COVID-19 pandemic highlights the relevance of this research so that CaBi may come back stronger with a better user understanding.

2.5 COVID-19 Pandemic

The global outbreak and spread of the COVID-19 pandemic has had an enormous impact on travel behavior in most parts of the world (73–77). Researchers already found evidence that the pandemic affected transportation-related behavior of non- (75) and motorized transportation modes (78) and public transportation systems (73) based on human mobility data or by surveys. To provide a brief overview, a description of the development process of the pandemic outbreak is outlined. The focus
is set on governmental interventions such as contact restriction and business closures that may affect public transport and mobility of individuals, followed by a critical review of journals published in the context of COVID-19 and non-motorized transportation.

2.5.1 Development

More than a year ago, on January 30th, 2020, the World Health Organization (WHO) declared an outbreak of a global emergency because of the disease caused by a novel coronavirus (79). On the next day, the former acting president followed the declaration of the WHO and declared a public health emergency in the U.S. that temporarily barred foreign nations from the U.S. Along with naming it COVID-19, the novel coronavirus was categorized as a pandemic on March 11th, 2020 (80). Before that, on March 5th, 2020, Maryland authorities announced the first confirmed case of the coronavirus in the broader Washington D.C. region (81). A few days later, on March 13th, 2020, D.C.'s government announced that schools would be closed until the end of the month (82). Because the number of cases continued to climb, the D.C. Health Department forced movie theaters, health clubs, spas, massage parlors, bars, restaurants, and other large businesses to close on March 16th, 2020.

Additionally, all shopping malls and entertainment venues had to close and gatherings were limited to 10 people or less (83). Because the ridership had fallen by 90 percent, the Metro transit system announced service reduction and closed seventeen stations indefinitely on March 24th, 2020 (84). By March 30th, 2020, residents of Washington D.C. were directed to stay in their homes except for trips to essential work, medical care, the grocery store, and outside recreation based on the social distancing
guidelines phrased by the formal "stay-at-home orders". Those measures were initially set to expire on April 24th, 2020. By May 29th, 2020, the district's "stay-at-home orders" were lifted, and a three-step recovery plan was executed. During the first phase, restaurants were able to re-open outdoor dining, and non-essential retailers were allowed to offer curbside pickup at their stores. On June 22nd, 2020, the district moved into the second phase, enabling indoor dining at 50 percent capacity, host activities, indoor shopping at non-essential businesses, and re-opening parks and gyms with safety restrictions.

Nonetheless, universities and schools remained closed. In August 2020, Washington D.C. announced that the state would remain in a modified Phase 2 as of late December 2020. On December 17th, traveling was limited to essential purposes again, issued by a new order (83).

With the emergency rollout of a vaccine on December 11th, a massive vaccination campaign was started, which was seen as the beginning of the outbreak's end. The distribution of the vaccine followed a predetermined order. First, the vaccine was made accessible to the state's oldest residents and other categories of essential workers (85). On February 14th and 16th, 2021, limits on outdoor dining, among other restrictions, were lifted, and the first students returned to in-person classes (83). By the end of April 2021, no noteworthy modifications were identified in the literature. As a result, the district of Washington D.C. has remained in phase two of recovery ever since, limiting indoor capacity, doing outdoor sport with social distance, and encouraging residents to work remotely (86).
2.5.2 Human Mobility Dynamic During the COVID-19 Pandemic

In order to control the spread of COVID-19, the government recommended and imposed various measures that consequential influenced travel behavior (78). However, as Chapter 2.2 presented, people have different travel needs varying from grocery shopping to commuting. The researchers Abdullah et al. (2020) explain that mode choices, distances traveled, and frequencies of trips have changed before and during the pandemic. In an online survey, the majority of participants indicated that shopping was their primary trip purpose during the pandemic. A shift from public transportation to private-owned non- and motorized mode share was detected. Also, pandemic-related health concerns affected the mode share decision (78).

Looking at the non-motorized mode shares in North America in detail, the study of Zhang and Fricker (2021) reveals, based on pedestrian and bicycle count data from five different cities, that COVID-19 led to a decrease in non-motorized activity in urban areas, but an increase in rural areas. However, in two cities, those trends reversed about 10-20 days after the first confirmed disease case of COVID-19 (75). This finding reinforces the importance of spatially differentiating the impact of COVID-19 on travel behavior.

Buehler and Pucher (2021) analyzed the impacts of COVID-19 from the year 2019 to 2020 on cycling in various cities and countries in Europe, the U.S., and Australia to characterize travel trends and time variations. Unlike the study of Zhang and Fricker (2021), the researchers identify an increase in the number of cycling trips across their research area. Based on the comparison between both years and the automatic bicycle trip counter system Eco-Counter data, the number of trips is validated (76). Although
COVID-19 first occurred in March 2020 (88), the researchers analyzed the entire year, which may have caused a bias.

Teixeira and Lopes (2020) researched the link between bike-sharing and subway use during COVID-19 in New York city. The researchers provide evidence that BSSs are more resilient than subway systems, with a ridership drop of just 71 percent instead of 90 percent. Additionally, the ridership ratio has decreased about 50 percent on the one hand. Still, it has increased in the average duration of a trip from 13 minutes to 19 minutes per trip on the other hand. Next to it, the study highlights that a modal transfer from subway users to the BSS took place (87).

According to the general linear model with random intercepts and random slopes of the researcher Tokey (2020), bike usage decreased for April 2020 for BSSs in Washington D.C., Boston, Pittsburgh, Columbus, and Portland compared to the ridership one year before. After April, the researcher identified a significant increase in ridership in Washington D.C., Pittsburgh, and Columbus, which had dropped, with the rise of the second wave of COVID-19 in June and July of 2020. Until August, which was when the period of examination finished, BSS experienced a slow growth again. By deploying advanced modeling techniques at a local level, such as land use, closure, and re-opening of different services and transit agency policies, the author suggests several future research concepts (74).

Retrieving the latest release schedule of the Bikeshare Ridership column by July 2020, the ridership of the CaBi BSS has dropped 56 percent for the same days of the week from March through May in 2019. According to the U.S. Department of Transportation’s Bureau of Transportation Statistics data, CaBi experienced the second-
largest decline after San Francisco's BSS Bay Wheels, with ridership down by 60 percent. Whereas the trip frequency of docked BSSs like New York and Chicago has already increased back to nearly pre-pandemic levels in May 2020, CaBi experienced a continuation of lower ridership than the pre-pandemic metrics (11).

2.6 Related Work

In the past, the CaBi BSS has attracted substantial interest from researchers, which is expressed in various journal articles, and thus the relatively low levels are astounding and need to be analyzed. Besides the already mentioned studies about socio-demographic and ridership characteristics, based on customer survey data, analyses attempting to understand the feasibility and potential of BSS were presented. Furthermore, according to El-Assi (2017), the third category of studies can be identified using real-time or historical bicycle ridership data from systems in operation to forecast and hypothesize the location, number, and distribution of stations and bicycles (88).

Gebhart and Noland (2014) used real-time bicycle ridership data and weather data to examine the impact of weather variables and the proximity of BSS stations and metro stations on hourly ridership levels. The researchers claim that reduced ridership is correlated with rain, high humidity levels, and cold temperatures (89).

Also, using real-time bicycle ridership data, Buch und Buehler (2012) investigated the effect of bicycle infrastructure, population density, and land use mix around stations on the yearly ridership level. Moreover, the study examined the number of households without a car using CaBi’s BSS. As stated in the journal article, BSS stations near bicycle lanes may increase the ridership (50).
In the same year, the researcher Daddio (2012) performed a regression analysis based on trip history data of October. The study concludes that stations away from the center of Washington D.C. are correlated to reduced ridership. At the same time, the nearness to retail outlets and the metro rail tend to increase the number of persons who used the rental system. In addition, socio-demographic factors were investigated. The results highlight that White, middle-aged riders predominately used CaBi (90) and thus do not deviate from the general BSS socio-demographic aspects which are named in Chapter 2.4.2.

More recent studies, like the study of Mc Kenzie (2020), analyzed the spatio-temporal changes in usage patterns of CaBi, also using ridership data of June 13th through October 23rd, 2018. This study identifies differences and similarities of BSS patterns with the scooter sharing system in Washington D.C.. One important finding is that BSS was primarily used for commuting purposes, while scooters are not.

The authors Wergin and Buehler (2017) analyzed the routes and activities undertaken in-between the renting and returning processes for the bicycles of the CaBi system. Therefore 94 GPS trackers were mounted on the bikes in spring 2015. The results show that both member types avoid roadways without bicycle infrastructure. A statistical significant difference was found in the user activities for the annual and casual members. Trips generated by casual members were slower in velocity but longer in time and distance and centered around the National Hall. Contrary, annual members were more likely to be found in the north around the neighborhood of Columbia Heights, an area in the north of Washington D.C.
The body of literature facing differences in the interest usage patterns based on trip history data of CaBi’s BSS is even small, there exist many remarkable insights. Nevertheless, these studies, some outdated or refer to an older network of stations, cannot explain any associations with the changes in behavior caused by the COVID-19 pandemic. This study will focus on how riders’ usage patterns changed concerning a spatio-temporal analysis using time series data.
CHAPTER 3 – Methodology

To establish a transparent methodological approach, process models and procedures are discussed in the context of data mining projects. A critical review of methodological approaches, like statistic metrics and algorithms, is essential to ensure the clearness and reproducibility of the research work. First, different process models are discussed, followed by selecting modeling techniques to ensure the assessment of the research questions.

3.1 Process Model

A process model is a set of sequential steps for software engineers to make repeatable, manageable, and measurable processes. It is a set of tasks that must be performed to gather new insights from databases (91).

Over the past years, different standard models for process models have become, according to Shafique and Qasier (2014), prominent. In their comparative study, the process models of Knowledge Discovery Databases (KDD), Cross-industry standard process for data mining (CRISP-DM), and Sample, Explore, Modify, Model, Assess (SEMMA) model are elaborated (92).

The KDD model process is an iterative process and interactive model. The first methodological framework was proposed by Fayyad et al. in 1996 (93). Since then, several distinct approaches have been developed in both academia and industry. Kurgan and Musilek (2006) provide a detailed comparison of different KDD process models in their survey (94). Following the predefined approach, KDD provides a basic framework for extracting hidden information from databases and emphasizing a high level of
specific data mining methods. For such a project, relevant prior knowledge about the project, domain, and goals is necessary to apply the process correctly (92).

The five-step SEMMA process model was developed by the SAS Institute Inc. in 1997 and consists of the steps named Sample, Explore, Modify, Model, and Assess. Because this model is incorporated into the commercial knowledge development software platform SAS Enterprise Miner (94), it is not available for this study and will not be described in more detail.

The CRISP-DM process model was first proposed in 1996 by a consortium of the companies SPSS, NCR, Daimler Chrysler, and OHRA and officially released in 2000 with a six-step approach as seen in Figure 5: Phases of the CRISP-DM reference model (95). The presented approach contains all phases of a project, their corresponding tasks, and their relationships (95).

Although the presented list of process models is not complete, additional process models and their evolution of data mining and knowledge discovery processes can be found in the article of Marbán et al. (2009) (96). Nonetheless, the CRISP-DM model was identified in the literature as the most used methodological approach for data-mining processes and can be seen as the “de facto standard” in both research and industry (91, 92, 94–96) and will also be used as a framework on this study.

The initial phase of the CRISP-DM model is the business understanding of project objectives and requirements from a business perspective. The data analyst's knowledge and function are to assign a problem definition and plan to achieve those goals.
During the data understanding phase, data is collected: At the same time, the researcher becomes familiar with the data, identifies quality problems, discovers first insights, and/or exposes subsets to form hypotheses involving hidden information. At the data preparation phase, the final dataset is constructed according to the modeling tools defined before. Also, tables, records, attribute selections, transformations, and the cleaning of the modeling tools are likely to be performed several times. Modeling techniques are selected and applied in the modeling phase. Also, the corresponding parameters are calibrated to optimal values. Because several schemes for the same data mining problem exist, different forms of data are required. In the evaluation phase, a model with high quality data analysis already exists, which is to be critically reviewed. The deployment phase requires a written report with the methodological approach of
the analysis in an understandable way. It is essential for the project and the deployment of the findings that the data analyst and the customer both understand the data analysis process and applied methodology (95).

3.2 Descriptive Statistics

Generally speaking, descriptive statistics summarize data without concluding specific assumptions about the data. It characterizes values like mean or diagrams like histograms (97). Standard statistics used to measure the central tendency are the mode, median, and mean metrics. The mode is a numerical value with the most considerable frequency. The middle score of an ordered distribution in the form of a rank is the median. However, the mean represents the average score. The frequency distribution range, interquartile range, and standard deviation can be used to validate dispersion in datasets. The number of cases per category represents the frequency distribution. The range is defined as the distance between the lowest and the highest score. An interquartile range is a range within which 50 percent of the scores fall. The average difference of each score to the mean is called the standard deviation (98).

Besides handling basic statistics, the literature also discusses the proper handling of the data cleaning process, which ensures the quality of the data. Further, because most of the articles start with a preprocessing data cleaning procedure, it highlights the order and the importance of this step. In the BSS studies of Vogel et al. (2011) (57) and Xu and Li (2018) (99), it is suggested that trips with a duration of less than 1 minute are removed from a database to avoid a bias of the data. The CaBi system declares at their website for the trip history data that those trips are potentially falsely generated or that
users tried to re-dock a bicycle to secure the appropriate returning process (70). Nonetheless, the researcher Yahya (2017) was able to identify usage patterns of a BSS, even without deleting short time duration trips, by using average performance measures to reveal the metrics of effective bike use (100).

Wand and Wang (2020) surveyed the context of time series data cleaning and discovered four difficulties. First, erroneous data can occur due to a variety of circumstances. The researchers name the sensor acquisition and associated source of error as a problem for a considerable amount of data and its error rate. Second, due to the complexity of the environment and its variations, false data may be gathered due to inaccuracy of the sensors or transmission errors. Third, errors can be caused by the wrong data cleaning methods or challenges facing the data type. Thus, time-series datasets differ from traditional relational data. Fourth, even if smooth filtering algorithms, such as moving average or interpolation, are most widely used for missing values, they change the original data so that the loss of information may occur.

As a result, the authors recommend data cleaning tools that avoid changing the information level of the original data. Because the smaller the change, the better the information quality, data cleaning processes for time series data should be based on the minimum modification principle (101).

3.3 Correlation Analysis

Scatter plots visualize the distribution of values for one, two, or more attributes and dependencies, and correlation analysis can compute the correlation of two features to confirm the expected dependencies (97). A correlation analysis is necessary to
measure the impact that the pandemic may have had on the ridership level of both periods of the analysis.

It is assumed that both samples have a linear relationship to each other. Hence, Pearson’s correlation coefficient is adequate to use. The correlation of two numerical attributes is defined as:

**Equation 1: Pearson’s Correlation Coefficient**

\[ r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} \]  

where \( x \) and \( y \) are the mean values of \( X \) and \( Y \) attributes, respectively, and \( s_x \) and \( s_y \) are the corresponding standard deviations. The yield values of Pearson’s correlation coefficient are between \(-1\) and \(1\). The larger the absolute value, the stronger is the linear relationship between the attributes. For \(|r_{xy}| = 1\), values of the features are located on a line. In the case of \( r_{xy} = 1 \), the line has a positive slope, whereas \( r_{xy} = -1 \) indicates a negative one.

Since Pearson’s correlation coefficient measures linear correlation, it is not appliable for nonlinear functions but monotone. In that case, the coefficient will not be \(-1\) or \(1\) and can be even farther away from these values. The rank correlation coefficients avoid these problems by ignoring the exact numerical values of the attributes. Additionally, it considers just the ordering of the values. Thus, it intends to measure monotonous correlations between features, in which the monotonous function has not to be linear.
An example for such a correlation is the Spearman´s rank correlation coefficient or Spearman´s rho, which measures the sum of quadratic distances of ranks and scales the outcome value to the interval $[-1, 1]$. It is defined by:

**Equation 2: Spearman´s Rho Coefficient**

$$\rho = 1 - 6 \frac{\sum_{i=1}^{n} (r(x_i) - r(y_i))^2}{n(n^2 - 1)},$$

where $r(x_i)$ is the rank of value $x_i$ when the list $(x_1, \ldots, x_n)$; $r(y_i)$ is defined analogously.

If the rankings of the attributes of the x- and y-values are in the same order, Spearman´s Rho will yield the value 1. The opposite is the case, if they are in reverse order, Spearman´s rho would yield the value -1.

If two or more values coincide, their rank could not be defined. Because *Spearman´s Rho* assumes that there are no ties, the rank $r(x_i)$ is defined as the mean value of all ranks of consecutive coinciding values in the sorted list if there is a tie.

Kendall´s tau rank correlation coefficient can be used to compare the order of pairs of values to keep this from happening. In case that two pairs are in the same order, and $x_i < x_j$, the pairs of $(x_i,x_j)$ and $(y_i,y_j)$ are called concordant if $y_i < y_j$. When the two pairs are in reverse order, they are called discordant, in case of $y_i > y_j$. The Kendall´s Tau is computed as:

**Equation 3: Kendal´s Tau**

$$\tau_a = \frac{C - D}{\frac{1}{2}n(n - 1)},$$

where C and D represent the numbers of concordant and discordant pairs, respectively.
In contrast to Pearson’s correlation coefficient, Rank Correlations are more robust against extreme outliers because they depend only on the value ranks. But, they can handle numerical attributes. For categorical features, an independency test like $\chi^2$ must be performed (97).

3.4 Unsupervised Learning

Along with supervised and reinforcement learning, unsupervised learning is one of the three different types of machine learning. Reinforcement learning is used to develop a system that learns a series of data to improve its performance based on interaction with the environment. On the other hand, supervised learning aims to learn to predict an outcome or the future based on labeled data, while unsupervised learning discovers hidden structures in unlabeled data. Because, this study seeks to identify usage patterns in unlabeled trip history data, unsupervised learning is suitable for the methodology approach and is described more in detail.

According to the authors, Raschka and Mirjalili (2019), unsupervised learning techniques enable the programmer to explore the data and retrieve profound information without supporting a known outcome variable or reward function (102). Conversely, the model is proposed to learn some fragments of data without any formal indication. Hence, an unsupervised algorithm is typically designed to discover similarities and
patterns among samples in the data or reproducing an input distribution based on a set of vectors derived from it.\textsuperscript{(103)}

The most common categories of unsupervised learning are cluster analysis, generative models, and association rules. Association rules are considered to find the most important associations of samples and are used widely to highlight relations that account for strategic and logistic purposes. For example, an online shop could employ an association rule algorithm to identify products that are frequently bought together and promote these products efficiently. Another unsupervised learning approach is based on generative models, aiming to model a parameterized distribution and optimize its parameters. As a result, the distance between candidate distribution and the data generating process is diminished. Lastly, cluster analysis aims to increase the homogeneity of sample data to set similar measures and proximity criteria in one cluster. Correspondingly, unlabeled samples are clustered so that data points within the same cluster are similar to each other, and the data point in different cluster groups is dissimilar.\textsuperscript{(103)} Because the goal of clustering approaches is to create groups that have low inter-cluster similarity and high intra-cluster similarity, it was used in many studies to cluster temporal and spatial usage behavior in BSS (5, 9, 57, 99, 104–107) and is briefly described in the following.

3.5 Cluster Algorithms

One of the essential techniques in unsupervised learning is the clustering problem. As described before, cluster algorithms aim to find structure in a collection of unlabeled data. Differentiation is made according to the types of clustering. First, in hard clustering
approaches, each data point is entirely part of one cluster or not. Second, contrary to hard clusters, data points in soft cluster methods are assigned with a probability or likelihood to be in those clusters.

Next to the differentiation of subgroups, the clustering algorithms can also be categorized into different categories. Since there exist plenty of varying clustering algorithms, due to its subjective task, an ordered framework, presented by Fahad et al. (2014), is described briefly (see Figure 6).

![Figure 6: An Overview of Clustering Taxonomy (108)](image)

First, in partitioning-based algorithms, all clusters are determined immediately. Data objects are divided into a defined number of partitions, which represent the clusters. Consequently, each division must contain at least one data point, and each data point belongs to exactly one group. Both rules have to be met so that the algorithm is valid.

Second, in hierarchical algorithms, the data points are divided into different levels to form a dendrogram. Thus, there are two approaches: bottom-up aggregation and top-down splitting strategy to cluster the data set. Because the iterative merging or splitting steps are performed and can not be undone, it is a critical drawback of that category.
Third, density-based algorithms separate data points based on their region of density, connectivity, and boundary. For example, a cluster is classified as a connected factor that grows in any direction the density leads to. The algorithm isolates several different density regions and groups all data points in these regions in the same cluster. The DBSCAN algorithm is in this category because it filters out noises in the dataset and discovers clusters of arbitrary shapes.

Fourth, in grid-based algorithms, the space of the data points is divided into grids. One main advantage of that approach is that it is fast in terms of processing time and performing the clustering on the grid instead of the data directly. However, it is not suitable for highly irregular data distributions because of its lack of clustering quality.

Fifth, in model-based algorithms, additional mathematical models are involved in optimizing the fit between the data. Statistical and neural networks are two major model-based approaches, which should be named (108).

Based on the variety of clustering algorithms, the common clustering approaches in the context of spatial and/or temporal analysis of time series data of BSS in the literature are presented.

Vogel et al. (2011) applied unsupervised techniques to analyze the typical activity patterns at the BSS Citybike Wien, Austria. Furthermore, the researchers express, based on the clusters identified, a new approach to model and forecast the bike demand based on the activity patterns (9).

Etienne and Latifa (2014) used a Poisson mixture model to combine an expectation-maximization algorithm to identify clusters of stations with similar usage
profiles. The researchers identified 8 clusters at the French BSS V'elib’, Paris (109) in the study context.

Three years later, the researchers Feng et al. (2017) also clustered the BSS of V'elib’, Paris, in the context of operational process improvements like controlling and re-designing the docking stations. For this purpose, they have used the k-means as well as the hierarchical algorithm. Based on the quality of clusters, the researchers claim that the k-means algorithm is more appropriate to cluster station behavior patterns (105).

Next to the k-means and hierarchical clustering algorithms, the scientists Cao et al. (2019) used a singular value decomposition to identify flows of the origin and destination areas of bicycle-sharing trips and then clustered the results (107). In another study focusing on spatio-temporal activity patterns, the researchers proved that the k-means clustering algorithms had the best results. The clustering quality was best by using k-means, compared to hierarchical and expectation-maximization clustering algorithms for the case of a BSS in Ningbo, China (106).

Moreover, the researcher Xue (2018) argues that k-means and DBSCAN are appropriate clustering algorithms to reflect the user's social and economic activities at different times and locations (99).

The algorithms k-means and DBSCAN will be presented in their theoretical approach, and the k-means will be applied in the context of this thesis based on the results presented. Because the data basis and the focus of analysis are similar to the studies that majorly used k-means, this method was selected. Nonetheless, both cluster methods will be described below to ensure a comprehensive understanding of unsupervised learning algorithms.
3.5.1 K-means Clustering

According to the *Hands-on Unsupervised Learning with Python* guidebook, the k-means is one of the most straightforward implementations of the maximum separation principle and maximum internal cohesion. By using a dataset \( X \in \mathbb{R}^{M \times N} \), which is \( M \times N \)-dimensional, samples are clustered in k clusters. The set of K centroids correspond to the means of the samples allocated to each Cluster \( K_j \):

*Equation 5: k-means Approach*

\[
M^{(0)} = \{ \bar{\mu}_0^{(0)}, \bar{\mu}_1^{(0)}, \ldots, \bar{\mu}_K^{(0)} \} \text{ where } \bar{\mu}_i^{(t)} \in \mathbb{R}^N
\]  

Here, the set M and the centroids have a different index, indicating the iterative step. \( M(0) \) defines the initial guess, and the algorithm aims to minimize the objective function. This inertia is the total average distance of an intra-cluster between the data points assigned to a cluster \( K_j \) and its centroid \( \mu_j \):

*Equation 6: k-means Measuring Value*

\[
S(t) = \sum_{k=1}^{K} \sum_{\bar{x}_i \in K_j} \| \bar{x}_i - \bar{\mu}_k^{(t)} \|^2
\]

Because the variance of the data points can highly influence the \( S(t) \) value, it cannot be considered an absolute measure. In the case of \( S(t+1) < S(t) \), the centroids are moving closer to an optimal position, in which a cluster may have the smallest possible distance to the corresponding centroid. Hence, Lloyd’s algorithm starts by initializing \( M(0) \) with random values.
The next step of the k-means algorithm is the assignment of each data point \( x_i \in X \) to a cluster, where the centroids have the smallest distance from \( x_i \):

**Equation 7: k-means Minimizing Centroids Distance**

\[
 c(\bar{x}_i; M^{(t)}) = \arg\min_j^{(t)} d(\bar{x}_i, \bar{\mu}_j^{(t)})
\]  

New centroids are recomputed as arithmetic means, once all tasks have been completed:

**Equation 8: k-means Arithmetic Mean**

\[
 \bar{\mu}_j^{(t)} = \frac{1}{N_{Kj}} \sum_{\bar{x}_i \in K_j} \bar{x}_i = \langle \bar{x}_j \rangle_{K_j}
\]

The process is iteratively repeated until the centroids stop changing because the sequence \( S(0) > S(1) > ... > S(\text{end}) \) has achieved the optimal configuration. In this process, the initial guess \( M^{(0)} \) can influence the computational time if \( M^{(0)} \) is close to \( M^{(\text{end})} \). Conversely, based on \( M^{(0)} \) randomization, an inefficient initial choice probability is high.

### 3.5.2 Density-Based Clustering (DBSCAN)

Based on density estimation, DBSCAN is another clustering method. Its fundamental idea is that a neighborhood radius must cover at least a minimum of points for each data point in a cluster. Contrary to the mean shift in k-means, it does not refer directly to the data generation process. Instead, the algorithm builds the relationship between data points with bottom-up analysis. It starts from a general assumption, in which \( X \) is made up of high-density regions divided by low-density data points. As a result, it is only suitable for well-separated clusters, whereas partitioning algorithms
work for spherical-shaped or convex clusters. Furthermore, DBSCAN does not allow stating the number of clusters because of the structure of $X$.

In particular, the algorithm requires two fundamental parameters: $\varepsilon$, which is defined as the maximum distance between two data points which are to be considered as neighboring points and the minimum number ($n_{\text{min}}$) of datapoints a dense region should contain. The outline of the cluster represents a cloud, which is defined by the datapoints $x_i$ and their distance to the core point of the minimum number of neighboring points. The ball, which the radius covers from the center of the cluster, is considered as $B_\varepsilon(x_i)$. A datapoint is regarded as a core point if it has the minimum number of samples at an epsilon distance. Only, if $n_{\text{min}}$ is met, the sample $x_i$ belongs to a cluster.

Formally, given a function $N(\cdot)$ that can count the number of data points contained in a set, a sample $x_i \in X$, if the following condition is met:

\[ N(B_\varepsilon(x_i)) \geq n_{\text{min}} \]  \hspace{1cm} (9)

Data that does not meet the constraint or is a border point is considered as noise or an outlier. Further, it could be possible that a data point has less than the minimum number of data points defined but has at least one core point in the neighborhood.

The four-step approach of the algorithm can be characterized by the definition of $\varepsilon$ and $n_{\text{min}}$ first. Next, distances have to be calculated, and the samples have to be classified in either neighbor of $X$, noise, or core points. Third, for each core point, which is not assigned to a cluster, a new cluster is created recursively until all of its neighboring
points are identified and assigned to the same cluster as the core point. Fourth, these steps are iteratively repeated until all unvisited points are covered (103).

3.5.3 Finding Optimal Number of Clusters

One drawback of both the k-means and DBSCAN clustering algorithms is the definition of the number of clusters. Because the data is unlabeled and the results uncertain, additional effort is necessary to define the appropriate number of clusters. Berthold et al. (2020) name three different approaches to plot the result for the optimal number of clusters, verified by finding an extreme.

First, a top-down approach is presented, with divisive clusters, by starting with a small number of clusters. Then, if the clusters do not fit the associated data, a further cluster split would be necessary. Second, a bottom-up strategy for an agglomerative clustering approach is presented. For this, the programmer should start with an overestimation of the number of clusters. Then, by merging similar data, the number of clusters is narrowed down to an optimum. Third, an algorithm can be executed for a full array of a possible number of clusters. As a result, each number of clusters can be evaluated concerning the goodness of fit to the data (97).

Thus, to quantify the quality of clustering results of k-means and DBSCAN, intrinsic metrics, such as a the within-cluster sum of squared errors (SSE), can be applied to compare the performance of different clusterings. For that reason, the different values of k are plotted. Because if k increases the average distortion will decrease, each cluster will have fewer constituent samples, and the samples will be closer to their respective centroids. Hence, if the number of k-clusters increases, the improvements in average distortion will decline. This turning point is named the elbow.
point because the decrease stops sharply. The number of clusters at which the progress of distortion declines is when the programmer should stop dividing the data into additional clusters. \(102\)

To evaluate the results of the elbow method, the authors Berthold et al. (2020) recommend the implementation of a so-called validity measure. For this purpose, they recommend establishing the silhouette coefficient of a cluster \(C\), which characterizes the average of silhouette coefficients \(s(x)\) of its members \(x \in C\):

Equation 10: Optimal Number of Clusters - Silhouette Coefficient

\[
s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \in [-1, 1]
\]

Where \(a(x) = d(x, C)\) represents the average distance of \(x\) to representatives of the same cluster \(C\), and \(b(x)\) is the average distance to the representatives of the nearest Cluster \(C'\) other than \(C\) \((b(x) = \min_{C'\neq C} d(x, C'))\). Based on the fact that a well-clustered data \(x\) is adjacent to the members of its cluster (small \(a(x)\)) but far away from participants of other clusters (large \(b(x)\)), the silhouette coefficient is close to 1. Accordingly, good clusters exhibit a high average silhouette coefficient from their participants \(97\).

3.6 Data Analysis Plan

Based on the presented methodological approach and the formulated research hypothesis, the methodological procedure of the thesis was elaborated and is summarized in Figure 7. The six-step process is inspired by the structure of the CRISP-DM process model and involves each crucial step with different analyses.
First of all, on the topic of understanding the project, the research questions should be pointed out and defined. Then, to make results measurable, factors and data mining goals are formulated, which can assess the usage patterns. This part of the process model is covered in the introduction (Chapter 1.1) and the literature review (Chapter 2.1).

![Figure 7: Methodological Approach of the Master’s Thesis](image)

Second, in the context of the data understanding phase, the database of the CaBi BSS is retrieved to find appropriate data for the time-series clustering approach in the context of the spatio-temporal aspect. Subsequently, the accuracy of the data has to be determined, starting with fundamental statistical analysis, such as standard deviation, mean, and quartile, which are applied in Chapter 4.2. Next to the feature observation, the data structure is visualized in the form of figures and tables to characterize the BSS data and provide a sense of the number individuals use the BSS. Furthermore, concerning the research hypothesis, the first insights and conclusions are summarized based on descriptive statistics and correlation analysis.
Third, based on the quality data, databases must be cleaned and concatenated to achieve a clustering analysis. Based on the data quality, the modeling techniques are to be selected. A model was selected and presented in terms of the algorithms and publicly available data presented in Chapter 3.1.

Fourth, the implementation and evaluation of the results take place in Chapter 5.5 and complements the modeling phase of the CRISP-DM process. Here, the focus is set on the comparison of long-term behavior changes before and during the pandemic.

Fifth, in the evaluation phase, the question of whether both cluster algorithms were appropriate for the type of analysis will be answered and evaluated. The main differences in the clustering results are displayed and placed in the context of the spatial characteristics of cluster centroids. Finally, cluster results are plotted on a map to ensure the comparability to different factors affecting the bicycling behavior to answer one of the last hypotheses.

3.7 Programming Language and Environment

The field of data science research and machine learning has exponentially grown over the last decade, with unnumerable research papers and complete tools. According to Raschka and Mirjalili (2019), the programming language Python has contributed to this through its use efficiency, elegance in coding, and compactness of application, to create a complete scientific ecosystem (102). Part of that scientific ecosystem are packages, or so-called libraries, which include production-ready algorithms. Those features are easy to learn and applied effectively in a short time and represent the backbone of many research papers.
Next to Python, other accessible data science languages are available, like R, Java, or Scala. Java and Scala also offer production-ready libraries, but Java is not as compact and easy to use as Python from the perspective of the author Bonaccorso (2019). Furthermore, in the case of R, the author identified a lack of supporting frameworks, which are essential to building a complete application. Conversely, although Scala has gained popularity in the big data aspects due to its properties and frameworks, Python is compared to be a better programming language in the context of data mining processes (103). For such reasons, Python was selected to create and train the complex model necessary for this thesis with the open-source software *Jupyter Notebooks*. Due to the variety of libraries offered, the following is a brief description of packages used in the program code (see Figure 31 through Figure 47):

NumPy adds support for large-scale, multi-dimensional array and matrices, next to many mathematical functions to operate in arrays (110). NumPy, for instance, was used to create the data array of the trip history data. The library pandas is used to manipulate and analyze large datasets. It can handle data structures and operations for controlling numerical tables and time series (111). An application example in this work is formatting the source .xlsx data into the .csv format, which other libraries can use.

Additionally, Matplotlib, with its function to create, animate and visualize graphs, was used. Next to the plotting role, it extends the NumPy functions by numerical mathematics (112). Finally, the library was used to provide the graphic design of the figures. Also, the Matplotlib extension, seaborn, was used to display high-level interfaces for displaying statistical graphics (113), like the distribution of frequency graph considering the member type.
Next, the package folium was used to visualize data on an interactive leaflet map (114). This library was necessary to display the clusters according to latitudes and longitudes of the station. Lastly, scikit-learn features various clustering, regression, and classification algorithms. It was used for the feature extraction, normalization, and application of the k-means algorithm.

All of the mentioned libraries are open-source tools in the context of machine learning projects. On behalf of the visualization, the colorblind color scheme of seaborn was used, which span a range of different luminance and saturation values.
CHAPTER 4 – Project Understanding

Examining the number of trips at the BSS CaBi, since the beginning of the COVID-19 outbreak in March 2020, a decline in the yearly ridership level can be witnessed. Because the efficiency of operations of a BSS has a substantial impact on its success, a general change of usage patterns should be understood first before redesigning forecast models, rebalancing strategies, and station allocations. This chapter on the project understanding provides a brief overview of the service provider CaBi and then proceeds with the data handling procedure. In order to provide a proper business understanding, the data source and study area are described concisely. According to the CRISP-DM process model, the steps and approaches of the project understanding, data understanding, and data preparation are indicated in this chapter. Additionally, the results of the descriptive statistics are visualized and described in the chapter of the preliminary data analysis.

4.1 Business Understanding

The third-generation public BSS CaBi was confronted with a decrease of the monthly ridership level of 56 percent in the first months of the pandemic (72). Because no long-term scientific analysis was performed to validate that the COVID-19 outbreak has impacted usage behavior, this thesis is dedicated to addressing that subject. To provide the reader a brief business understanding, Table 8 provides quick facts about the CaBi system in the appendix. In addition, further information about bicycle sharing in general and in the region of Washington D.C. is provided in Chapter 2.4 and is intended to assure a comprehensive business understanding at this point.
4.2 Data Understanding

The public BSS CaBi provides the trip history data of this analysis. The system data gets published quarterly in the comma-separated value (.csv) format and includes, according to the website, the following seven variables:

- Duration – duration of the trip in seconds
- Start Date – start date and time (YYYY-MM-DD hh:mm:ss)
- End Date – end date and time (YYYY-MM-DD hh:mm:ss)
- Start Station – Station number and name
- End Station – Station number and name
- Bike number – ID number of bike used for the trip
- Member Type – indicates whether the user was a registered annual member or casual rider (70).

In the context of this study, the data for the period of March 2019 – March 2021 was retrieved on April 17th, 2021. Two twelve-month analysis periods were defined to ensure the comparability of the phases before and during COVID-19. Due to the same start and end of the periods, it is assumed that external factors, such as weather effects, have not influenced the results, enabling an unbiased comparison. The first period (P1), from March 7th, 2019 - March 6th, 2020, is defined as pre-pandemic, and the second period, from March 7th, 2020 - March 6th, 2021, represents the time during the pandemic. The date of the first officially confirmed COVID-19 case in Washington D.C. serves as the start date of the second period. Because the long-term changes of the usage patterns are the focus of this study, 12 months was set as the analysis period.
Additional information that provided insights into the user groups and business understanding was obtained from a member survey, which was last published in 2016 (71).

Next to the historical trip data, which was retrieved from the CaBi Webpage, a COVID-19 dataset was retrieved on May 17th, 2021, from the official governmental website of Washington D.C., updated on an irregular basis every few days (115). The downloaded file (.xlsx format) provides information about several connections like total cases by ward, race, as well as lives lost by race, sex, age, ward, and community cases tested. The number of total daily cases was retrieved from that list and adjusted to the analysis period.

This chapter is divided into the essential statistical characteristics of the ridership level for both periods and the data cleaning procedure for the modeling approach to ensure a holistic understanding of the data and its structure.

4.2.1 Preliminary Raw Data Analysis

Commencing with the database of the COVID-19 cases, the first officially confirmed case was on March 7th, 2020. This date represents the start day of the analysis. Since March 2020, the infection numbers increased steeply and were at their peak feature in May 2020, with a total of 4,199 cases in one month. Since May, cases fell until June 2020, after which they began to increase slightly again in July 2020. After July, there was a further decrease in cases, which started to rise from September 2020 onwards, reaching its peak in November 2020. The most striking number of cases per day was attained on December 26th, 2020, with 493 confirmed cases, followed by January 11th and 15th, 2021, with 430 and 397 cases, respectively. The numbers were
consistently high until January, with a total case number of 4,316 in November 2020, 7,567 in December 2020, and 7,756 in January 2021. In February and March 2021, a decrease can be observed to the average case number of 3,676 and 3,972 (115).

The course of the pandemic is described in the media in symbolic waves (116). If this depiction is applied to the graphic shown, a first wave with a peak starting in April 2020, another roughly around July 2020, and the third wave in beginning in December 2020 can be identified.

Regarding the data quality, it should be mentioned that the database did not report any cases for December 25th, 2020, probably because of the public holiday and, therefore, no test performance. Further, a negative numerical value of COVID-19 cases was reported on February 23rd, 2021, which may be related to a database update or adjustment of actual values.

Figure 8: COVID-19 Cases and Monthly Ridership Level for P1 & P2
Progressing with the trip history dataset, a total number of samples of 3,130,856 trips in P1 and 1,696,210 in P2 was retrieved, and the presented basic descriptive statistics in chapter 3.2 applied. The results for trip duration are displayed in Table 2.

Table 2: Descriptive Statistics for P1 & P2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>3,295,128</td>
<td>1,058.52</td>
<td>2,128.71</td>
<td>60.0</td>
<td>400.0</td>
<td>683.0</td>
<td>1,150.0</td>
<td>86,365.0</td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>1,962,941</td>
<td>2,901.34</td>
<td>9,703.10</td>
<td>-4,000.0</td>
<td>476.0</td>
<td>872.0</td>
<td>1,525.0</td>
<td>2,005,616</td>
<td></td>
</tr>
</tbody>
</table>

If each count is considered a trip, there is a decline in the ridership level of 45.82 percent from P1 to P2. Looking at the seasonality effect, which was described in chapter 2.6, P1 (2019) exhibits such an effect, whereas P2 (2020) has a bulbous pattern over the year with a constant ridership duration (Figure 8),

It is worth mentioning that contrary to CaBi’s statement that the datasets are already cleaned, the files of the second period contain not only values below 60 seconds but also negative trip durations. These negative values were detected on December 25, 2020 and February 23rd, 2021. Handling wrong and missing values will be described in the next chapter.

4.2.2 Data Cleaning, Concatenating, and Preparation

As identified in the literature review of the methodological approach, data preparation is one of the most important steps to ensure high-quality data mining analysis.
First, for each month, one file with a representative dataset is provided by CaBi in a comma-separated value (.csv) format (70). According to the provider, trips lasting less than 60 seconds (chapter 4.2.) and trips performed by staff for the service and inspection system or test purposes of stations at the warehouse were already processed to remove. Contrary to the statement, 187,147 samples were identified during the analysis which did not meet the time minimum. Because most researchers characterized trips less than 60 seconds as outliers (57, 99), the data points were removed from the data-set for the clustering approach. Remarkably, a high number of wrong values appeared only in the second period. To ensure that the outliers do not affect the analysis results, it was decided to remove these data points using an interquartile approach (Figure 41). It is critical to note that the second period showed a high scatter of data, which is probably due to erroneous values or wrong return of bicycles caused by users. In addition, from April 2020 onwards, the BSS was expanded by electronically assisted bicycles. These e-Bikes could be borrowed and returned at the roadside, independent of the station system offered by CaBi. Since this study focuses on the characterization of station-based BSSs, the trips of e-bikes were also removed. Nonetheless, the fact that riders could use for one way station-based bicycles and to get back back e-bikes, may cause a bias in the clustering approach.

In addition to these erroneous values, negative lending times were noted for Dec. 25, 2020. Since this traceability is impossible, these samples were manually removed from the raw data and have no impact on any analysis.

Besides the data understanding, problems resulted in the context of concatenating multiple .csv files. Indeed, the data files from March 2019 until March 2020 exhibit
seven attributes in the data structure presented in Chapter 4.2. From April 2020 onwards, the dataset had a different data structure with 13 features. However, because March 2020 was an essential part of the second period and should not be neglected in this analysis, a minimum standard was developed for factors of both periods to combine files and to ensure the comparability of the datasets (c.f. Table 3).

**Table 3: Attributes of the Trip History Data**

<table>
<thead>
<tr>
<th>Attributes until February 2020 (P1):</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>Start date</td>
</tr>
<tr>
<td>Attributes since March 2020 (P2):</td>
<td></td>
</tr>
<tr>
<td>Ride_id</td>
<td>Rideable_type</td>
</tr>
<tr>
<td>Concatenated Attributes for Analysis:</td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>Start_date</td>
</tr>
</tbody>
</table>
To proceed with the spatio-temporal analysis on a holistic BSS, GPS coordinates were added in the form of longitude and latitude attributes for both start and end stations of all trips for the data until March 2020 (Figure 31 - Figure 33). Furthermore, using feature engineering, the trip duration in seconds was added for the second period. Moreover, additional data frames were created to describe the time series data according to hourly and monthly aspects, which can be retrieved from the appendix. Finally, to find further incorrect data, the null function of the library pandas was used and implemented in each analysis, with no interpretable results.

Lastly, the attributes had to be adapted to the requirements of the libraries. To name an example, the time was transformed into a nominal attribute using the alteration function of the library pandas (Figure 32).

In addition to the trip history sample data of the BSS, the COVID-19 cases in Washington D.C. were implemented in a column of both datasets. The raw file was retrieved from the official COVID-19 surveillance website of the District of Columbia (115), and unnecessary information was removed so that just the overall daily COVID-19 cases and total cases persisted. The daily reported new cases were added to the file for the second period for further analysis in order to provide data for the descriptive and correlation analysis. The wrong value of -1 cases was manually corrected to 0 for February 23rd, 2021.

After all conducted processes, it can be said that the initial number of 3,295,128 data points in P1 and 1,962,941 in P2 was corrected to 3,077,308 and 1,703,359, respectively. The error-proneness in the second period and the generally lower level of ridership can influence the analysis results. On the one hand, the data quality mainly
impacts the exactness of an analysis. Because differences in the context of data quality was observed in between the periods, the analysis may be more accurate for one period than the other period. On the other hand, by deleting false values, the number of datapoints in the dataset decreases. A minor number of datapoints make the classification of the clustering algorithm more difficult in a dataset. Besides those difficulties, a descriptive analysis and the clustering approach are applied in the next chapter of the Data Analysis, with the awareness of those problems.
CHAPTER 5 - Data Analysis

Generally speaking, based on the preliminary analysis, seasonal variations in the yearly ridership level can be seen in for each month in Figure 8 for P1. As expected, people were biking less during the pandemic and colder months. Moreover, as the number of cases increased, the frequency of bike-share trips decreased.

This chapter establishes a context for the described relation of ridership level, COVID-19, and the research questions. A descriptive analysis, correlation analysis, and cluster algorithm are applied to the cleaned and formatted dataset of P1 and P2. At the end of this chapter, the findings are evaluated in the context of the presented research questions. Those patterning techniques represent the modeling and evaluation phase of the CRISP-DM model.

For the reason of visualization of the data, the colorblind scheme of the library seaborn is used. To provide consistency over the variety of graphs, a color code for member types (annual = light blue; casual = grey), Periods (P1 = brown, P2 = pink), and Clusters (C1 = dark blue, C2 = purple, and C3 = orange) was defined to provide the most distinction in saturation and luminance values.

5.1 Descriptive Analysis

The sample data can be grouped according to the member type, starting the descriptive analysis for the overall trip duration. Here, the annual member shapes for both the frequency and the duration of trips are displayed in Figure 9 and Figure 10 according to the member type.
Figure 9: Distribution of Duration Considering Member Type in P1

While the uncleaned average trip duration was 1,058 seconds (ca. 18 minutes), it is clear from the graph that most annual members (light blue) rented the bike between 200 seconds (ca. 3 minutes) and 800 seconds (ca. 13 minutes) in P1. Casual members (grey), on the other hand, borrowed the bikes for a more extended time. Most of them rented a bike between 500 seconds (8 minutes) and 1800 seconds (30 minutes). After 30 minutes, or 1,800 seconds, a separate fee is charged in addition to the flat rate per loan (red arrow in Figure 9), at which point, a smooth drop in the number of rentals for annual members was identified. Of the overall 3,077,308 trips, 90.33 percent were made by annual members (2,779,857), and the remaining 297,451 by casual members.

Compared to P1, P2 can be characterized by a larger number of casual members using the BSS. While the total number of ridership levels for the extent of the analysis has decreased by 45.8 percent, the number of trips generated by casual members has increased from 287,451 trips to 570,858. The trip duration of annual members can be
considered about the same, but here the trip frequency has decreased by 66.01 percent from 2,779,857 to 944,843 so that these represented 60.01 percent of the total trips. Furthermore, after the free of charge time of 30 minutes, there are more trips than in P1.

![Ridership Level According to Frequency and Duration P2](image)

**Figure 10: Distribution of Duration Considering Member Type in P2**

In terms of spatial use of the BSS, the top 10 stations for origin and destination and the most popular junction routes were summarized in the appendix (Figure 35). In this context, it is noteworthy that both the ridership level and the start and finish hierarchy have changed over the analysis period, reinforcing the indication that since COVID-19, the usage behavior has changed in a spatial and temporal context. Next to a remarkable decrease in the monthly ridership level, only eight of the beginning and end stations in P1 are represented in P2. Furthermore, only four of the most popular junction routes from P1 are included in P2, strengthening the need for further spatial analysis in the clustering analysis.
In consideration of descriptive analysis for the temporal differences of the periods, Figure 11 was elaborated for the comparison of the number of rides per month, Figure 12 for the numbers of rides per day, and Figure 13 for the average duration of trips according to the day of the week and number of rides from the raw data, excluding negative duration times.

![Number of Rides Per Month](image)

**Figure 11: Number of Rides per Month for P1 & P2**

Considering the monthly ridership levels of the CaBi BSS in Figure 11, the influence of the spread of COVID-19 is seen. Based on the late stay-at-home proclamation (Chapter 2.5) in Washington D.C., there were still a high number of rentals in March. Because case numbers increased and the declaration came into force on March 26\(^{th}\), 2020, the ridership level each month followed seasonal patterns, by increasing in warmer months and decreasing during the winter. In April 2020, the number of trips decreased remarkable by 78.6 percent, compared to the previous year. Comparing each month of P1 with P2 for the ridership level, the most significant decline occurred in
April. In the following month, the ridership level recovered and remained at a level of about 180,000 trips per month. A decrease in numbers is noted, beginning with October, which may be caused by seasonal patterns or increasing COVID-19 cases. This association is examined in Chapter 5.3 concerning correlation analysis.

Considering the ridership levels by weekdays, Figure 12: Number of Rides Per Day for P1 and P2 was plotted.

![Figure 12: Number of Rides Per Day for P1 and P2](image)

Before the pandemic, the BSS was embraced by annual members, in which the casual riders accounted for only a minor part of the ridership. Taking the total number of rides into account, in P1, the most trips took place on Wednesdays (529,748), followed by Friday (520,428), Thursday (514,772), Tuesday (513,419), Saturday (503,593), Monday (481,467), and Sunday (415,714). Changes in usage behavior in terms of days of the week were observed during the pandemic. A marginal increase of casual riders was occurring. In general, over the analysis period of P2, most people were
using the BSS on weekends. According to the data, the most trips were on Saturdays (355,878), followed by Sunday (278,582), Friday (237,330), Tuesday (211,682), Wednesday (210,616), Monday (201,692), and Thursday (200,014).

Considering the average trip durations per day, especially in times of COVID-19, the average duration per day increased for casual riders considerable and slightly for annual riders on the weekend, compared to pre-COVID-19 values (Figure 13).

![Average Duration Per Day for P1 and P2](image)

*Figure 13: Average Duration Per Day for P1 and P2*

Inspecting the ridership level according to the hours of the day, P1 can be characterized with striking peaks around the rush hours at 7 am and 5 pm, with a lower level during the day for annual members. On the other hand, the casual members continuously increase according to the ridership level over the day, reaching the maximum level at 5 pm (see Figure 14).
5.2 Correlation Analysis Preliminary Data Analysis

As part of the correlation analysis, a data frame had to be generated that aggregated the daily trips of the ridership level. For this purpose, the `period-2-final_tripdata.csv` was used as an input file to create the output data frame `bike` by applying the count function of the `pandas` library. Concatenating with the `COVID-19_cases_Washington_D.C._new.csv` values, the data frame `bycount` was established. In this context, duration was used for the counting mechanism and kept in the header as an indicator for the trip frequency.

Following the application of Pearson's correlation coefficient (PCC), Spearman’s Rho coefficient (SPC), and Kendall’s Tau coefficient (KTC) function from the library `seaborn` as described in chapter 3.3, the result matrix in Table 4 was generated.
Table 4: Summary Table for Trip Frequency Correlation

<table>
<thead>
<tr>
<th></th>
<th>Trip Frequency</th>
<th>Total Positives</th>
<th>New Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>SRC</td>
<td>KTC</td>
</tr>
<tr>
<td>Trip Frequency</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total Positives</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>New Cases</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

It can be stated that the correlation of the feature trip frequency to the number of total positives and new cases is inverse. While the total number of cases tends to have a minor negative impact on the daily ridership level, the correlation value for trip frequency and new cases is near-significant at -0.5. Only the KTC has a lower value than 0.5.

Since the usage behavior in the context of the member type differed in P1 and P2 in the descriptive analysis (as in Chapter 4.2.1), it was decided to perform another correlation analysis regarding the trip duration time (Table 5).

Table 5: Summary Table for Trip Duration Correlation

<table>
<thead>
<tr>
<th></th>
<th>Trip Duration</th>
<th>Total Positives</th>
<th>New Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>SRC</td>
<td>KTC</td>
</tr>
<tr>
<td>Trip Duration</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total Positives</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>New Cases</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The results support the previous hypothesis that fewer trips were made during the pandemic but lasted longer. Although negative correlations can be identified at this point, they are not as significant as the trip frequency. This may be because the generally lower willingness to ride a bicycle affected the correlation and the longer driving times reduced that value.
Because the descriptive analysis revealed a difference between the member types in terms of daily ridership levels from before and during the pandemic, the statement that more casual users took rides at CaBi was validated by regression analysis. For this, the *seaborn* library was used, and the results in Table 6 were obtained.

### Table 6: Summary Table for Trip Frequency Regression Analysis According to Member Type

<table>
<thead>
<tr>
<th></th>
<th>Trip Frequency</th>
<th>Total Positives</th>
<th>New Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>Annual</td>
<td>Casual</td>
</tr>
<tr>
<td><strong>Trip Frequency</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total Positives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>New Cases</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tables 4, 5, and 6 from the correlation analysis confirm that the pandemic had a higher negative impact on the ridership level of annual members than for casual members. However, since the respective regression values for casual members are closer to zero, the pandemic affected them, but not in a statistically significantly manner.

### 5.3 Evaluation of the Descriptive and Correlation Analysis

The descriptive analysis indicated that in the context of the yearly, monthly, daily and hourly ridership level and its temporal distribution, differences in P1 and P2 could be identified. Generally, the periodic ridership level decreased since the beginning of the pandemic and has not met the previous year's trip numbers ever since. Changes in the general usage behavior in terms of days of the week were noted during the pandemic, shifting from a concentrated weekday utilization to a primarily weekend orientated use (c.f. Chapter 5.1). Also, higher utilization hours of the day shifted from two striking
peaks at the rush hours to a constantly increasing the hourly ridership level starting at 7 am, reaching its peak with the most extended trip durations at 6 pm.

Considering the member type of the BSS users, a remarkable increase of 91.91 percent of casual rider trips occurred, whereas the number of annual member rider trips decreased by 66 percent. Nevertheless, the proportion of annual members still outweighs that of casual members. However, while this proportion was 90.04 percent in P1, it was only 37.66 percent in P2. Looking at the graphs of Figure 9 and Figure 10, it becomes evident that the hypothesis that casual members have longer travel times than annual members is confirmed. Also, this finding supports the previous hypothesis that fewer trips were made during the pandemic but lasted longer. Because no association with the COVID-19 pandemic can be made from the descriptive analysis, the results of the correlation analysis are now used.

Taking the results of the correlation analysis into account, the trip frequency is significantly negatively correlated with the COVID-19 cases according to PCC and SRC. Only the KTC does not show any significant correlation, which may be due to the different valuation methods applied in Chapter 5.2. Looking at the trip duration and new cases, a slightly negative correlation was identified with the SRC. However, considering the member type and applying the PCC, a difference according to the trip frequency becomes evident. Whereas the PCC for the annual members is -0.53, it is just -0.39 for casual members, supporting the previous statement that the general number of trips has been reduced during P2. At this point, the hypothesis can be verified that the general COVID-19 cases with upcoming legislative restrictions harmed the yearly ridership level of cyclists at CaBi’s BSS, despite a clear indicated data scatter.
Nonetheless, a significant negative correlation was found for the annual members. Also, the presented figures corroborate the conjecture that although fewer casual members have taken the bicycle, they have traveled more long distances. As a result, the cycling behavior has changed in temporal aspects and frequency and trip duration from P1 to P2.

5.4 Modelling Cluster Algorithms

This analysis is based on the concatenated trip history data, which was separated according to the periods P1 and P2.

With a focus on the pre-COVID-19 period, the basic statistics of the data frame were retrieved to evaluate the data quality. In the process, many outliers were identified, resulting from the high differentiation of the minimum value and the 25th percentile and the maximum value and the 75th percentile. Beneficial to an optimized model, the outliers have been removed. In the process of eliminating outliers, trips less than 60 seconds have been removed. The number of data points decreased by 6.61% from an initial 3,295,128 to 3,077,308 counts. Additionally, column names were corrected, and the data type of date columns has been changed from an object to a date-time data format.

Furthermore, because Latitudes and Longitudes are beneficial for creating geospatial visualization, the data type and structure had to be transferred so that the library folium could print the coordinates of the entire system on a Leaflet map (Figure 3).

In consideration of spatio-temporal usage patterns at the BSS CaBi, the trips were aggregated according to stations and time and the duration of trips. As a result, a data frame with 567 rows (for 567 different stations used in P1) and 25 columns (for 24
hours of the day and one for the station name) was created for the start stations. Each cell’s attributes represent the sum of all trip durations starting in the particular hour at the station.

Employing the *sklearn* library, the k-means model was trained with the feature scaled data using a random number of k-clusters. The optimal number of clusters was diminished based on the heuristic elbow method and its resulting plot for the explained variation (see APPENDICES). Because the value had to be either three or four, the silhouette value was calculated. The silhouette score was higher for $k = 3$ clusters (0.5556), than for $k = 4$ (0.5390). As a result, $k = 3$ was defined as the optimal number for P1. The geospatial map of stations according to their clusters is displayed in Figure 15.

![Figure 15: Geospatial Clustering Map of P1](image)
The first cluster (C1; n = 387; blue) contains the most remarkable number of stations but the shortest trip duration. Peaks can be identified around 8 am and 5 pm, with decreased hourly ridership levels in-between (Table 7). The second cluster (C2; n = 152; pink) grouped a total of 152 stations. The overall number of users is higher compared to C1. A noteworthy rise of peaks can be seen at 8 am and 7 pm, as in C1 (Table 7). The last cluster (C3; n = 27; yellow) is characterized by the lowest number of stations. Contrary to C1 and C2, C3 has a first peak around 8 am, but the number of users and its trip duration increase continuously over the day, reaching its peak at 5 pm.

Based on the before presented approach of k-means, the same procedure was implemented for P2. By doing so, the data size was decreased from 1,702,848 to 1,515,700 by 10.9 percent. Caused by the new data layout, the latitude and longitude values have not met the standards of the library and showed a higher number of total stations (n = 641). The coordinate values have been corrected and added by a separate data frame to the list. Stations like on 10th & E St NW, or Wisconsin Ave & O St NW represented no rentals at certain hours in P2 but were kept in the analysis. As before, the feature scaling was applied, and the optimal number of clusters (n=3) determined.
Figure 16: Geospatial Clustering Map of P2

As a result, three new clusters can be characterized for P2 (Figure 16), which differ from the previous clusters in P1. For P2, the first cluster groups 461 stations together. Again, the first cluster (C1; n = 461; blue) has the lowest duration numbers. However, the course of the curve differs from that in the first period. The first cluster indicates a continuous increase beyond the day until 6 pm.

Contrary to the other gradients, no striking peak can be detected. Next, the second cluster (C2; n = 154; purple) has the most frequent trips during COVID-19. In this graph, a stagnation of the number of trips can be seen around 8 am, which increases again from 10 am onwards. The maximum number of trips is reached at 5 pm, followed in time by a remarkable decrease. Lastly, a third cluster (C3; n = 26; yellow) is typified with n = 26 stations. It can be characterized by a medium number of trips as well as a medium time duration.
An assessment of the clusters, categorization within the Washington D.C. region, and evaluation will be in the analysis chapter.

5.5 Analysis of Clusters

After the temporal classification of the usage patterns, the spatial consideration and connections will now be discussed. To make the comparison feasible, the assumptions are made that seasonality, weather, and socio-demographic changes have not affected the usage behavior or were equally intense for both P1 and P2. Geographical maps concerning the built environment are used for further analysis, which can be found in the appendices (Figure 23 - Figure 26).

Considering the first cluster of P1 (see Figure 17), it is remarkable that it is very noisy in its spatial distribution. The 387 stations are located outside the town center of Washington D.C., focusing around the Arlington region. According to the temporal analysis, there are light peaks at the office hours 7 am and 5 pm, with below-average trip durations. In the context of land-use characteristics, especially for the city of Washington D.C., most start stations are located in medium dense areas of residents and with average employment density (Figure 24). In general, this cluster can be characterized by the highest number of stations but with a low hourly-ridership level and intense varying duration times.
Examining the second cluster (see Figure 18), it is noticeable in the urban core area of central Washington D.C.. Looking into the geographic map of Washington, those stations are located in the most important employment area, which is also represented by a lower density-residential area. Additionally, most transit stations (Figure 26) and federal lands, which are also used for recreational purposes like parks and open spaces (dotted areas in Figure 16), are located around the stations. It possesses the second-highest duration time, with peaks at times which can be considered as rush hours. Because the duration time of trips is longer than in the other clusters, it is assumed that tourists and short-distance commuters most likely cause the trips.

**Figure 17: Geospatial Map of Stations in C1 in P1**

Examining the second cluster (see Figure 18), it is noticeable in the urban core area of central Washington D.C.. Looking into the geographic map of Washington, those stations are located in the most important employment area, which is also represented by a lower density-residential area. Additionally, most transit stations (Figure 26) and federal lands, which are also used for recreational purposes like parks and open spaces (dotted areas in Figure 16), are located around the stations. It possesses the second-highest duration time, with peaks at times which can be considered as rush hours. Because the duration time of trips is longer than in the other clusters, it is assumed that tourists and short-distance commuters most likely cause the trips.
Figure 18: Geospatial Map of Stations in C2 in P1

Lastly, C3 P1 is the third cluster (see Figure 19), with its clustered stations around the city center. It has the highest number of users, especially around the rush hour of 7 am and 5 pm. Considering the land-use map, those stations are located in medium- to high-density residential areas, which feature a combined land use characteristic (Figure 23). Furthermore, those stations are well located in great streets, transit stations, and medium-dense employment areas. Due to the temporal and spatial distribution of the cluster, it is assumed that these are mainly commuter trips.
Figure 19: Geospatial Map of Stations in C3 in P1

Just as P1, the first cluster of P2 (see Figure 20) is characterized by a vast spread in the station network in and around Washington. Caused by the changes made for the GPS coordinates in P2, the number of the station is higher than before, with 464 including stations. The difference, however, can be seen in the timing of usage. Instead of the usual peaks around rush hour, the number of durations is continuously increasing, but with lower values than P1. Accordingly, bicycles in this cluster tend to be rented for more extended periods in the afternoon rather than during office hours (Table 7: Result plots of the k-means Cluster Analysis). It is assumed that next to the sparsely developed station network, people used CaBi for longer trips, especially for recreational purposes.
In the context of the second cluster in P2 (see Figure 21), a general decline in the number of trips as well as in the time of duration was observed. Regarding the time distribution over the day, a continuous increase of trips was noted after the first-morning peak, which grow until 5 pm. In general, the heights of the graph are smoother than in the previous period. As a result, it can be assumed that fewer commuters used the BSS during the pandemic.
The most considerable difference can be found in the third cluster in P2 (see Figure 22). Contrary to the other clusters, the travel time starting from the 26 stations has increased notable. Next, the associated stations are similar to P1 C3 in the highly populated areas, which can be found north of the central employment area of Washington D.C.

By analyzing the plotted maps from the modeling approach of k-means, changes of stations, swapping from one to another cluster is barely viewable by the eye. That was why the labeled stations' data from the cluster labeled data frame was retrieved and analyzed. In total, 133 stations have changed clusters or were added to the system compared to P1. Generally speaking, 74 stations were added in P2 or did not appear in P1 due to bad data quality because no matching station was identified in the comparison.
Figure 22: Geospatial Map of Stations in C3 in P2

On the other hand, a total of 59 stations changed the cluster category. Major changes were identified with 22 stations, switching from the vast spread cluster to be more likely in the suburban area. On the other hand, 17 stations also changed vice versa. Furthermore, 17 stations differed from the central cluster to the suburban cluster, located north of the main employment area. Also, eight stations swapped the category of the same clusters. Lastly, just one or none have changed from the vast spread cluster to the center or vice versa. Because of the balanced changes of stations from one cluster to another, it can be assumed that the spatial usage patterns have not changed observably by a visual analysis.
Table 7: Result plots of the k-means Cluster Analysis

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td><img src="image1.png" alt="Graph for C1 P1" /></td>
<td><img src="image2.png" alt="Graph for C1 P2" /></td>
</tr>
<tr>
<td></td>
<td><img src="image3.png" alt="Graph for C2 P1" /></td>
<td><img src="image4.png" alt="Graph for C2 P2" /></td>
</tr>
<tr>
<td></td>
<td><img src="image5.png" alt="Graph for C3 P1" /></td>
<td><img src="image6.png" alt="Graph for C3 P2" /></td>
</tr>
</tbody>
</table>
5.6 Evaluation of the Usage Patterns

According to the chapter of related work, evaluating the clusters shows that occasional users use the first cluster for P1 over a more extended period of time due to further distances and a less dense station network of CaBi.

As mentioned in the research of Wergin and Buehler (2017), tourists cause longer trip durations (117). The second cluster in P1 is probably mainly used by tourists as casual members. Its centroid is next to several museums, monuments, memorials, the Capitol, and the White House, reinforcing the statement that tourists rent bikes, stop, visit sightseeing locations, and cause longer trip duration in high-density areas with mixed land use. Lastly, as Daddio (2012) revealed, the stations close to high density populated areas experience the most frequent trips (90). This pattern can be exemplified by regular users who use the BSS for short trips, such as commuting to work in the third cluster. The collected results for P1 agree with the described results in the literature.

After clustering the time-series data in P1, the usage behavior of the CaBi BSS has shifted from a rush-hour orientated system to a plateau-like structure peaking in the evening according to the trip duration for all clusters in P2. In general, the heights of the graph are smoother than in the previous period. As a result, fewer individuals used the BSS, showing a lower periodic ridership level. Especially the annual members with the purpose of commuting declined the most and affected the BSS.

With the focus on the spatial change of the BSS usage patterns, no remarkable changes were identified in the analysis of the clusters. A few swaps of stations in the clusters existed during the pandemic but had no serious impact on the cluster location. Because individuals did not move during the pandemic, following the stay-at-home
declaration, working remotely, and making fewer trips, the usage behavior in the context of spatial usage behavior stayed the same. Nonetheless, it can be assumed that due to the increase of trip duration, people used the CaBi for different spatial activities, which took place mainly in the afternoon hours between 2 pm and 6 pm. Furthermore, the days of use shifted from a weekday-oriented system in P1 to a weekend-oriented usage pattern. The lack of commuters and the increase of recreational trips on weekends by casual members would explain that variation.

Additionally, the lists of the most used origin and destination stations have more common stations in P2 than in P1. As a result, individuals used CaBi more likely for round trips, starting and ending at the same stations, rather than for one-way trips.
CHAPTER 6 – Significance of the Results

This chapter displays the steps of Evaluation and Deployment of the CRISP-DM process model. First, the results of the analysis are discussed in the context of the research questions and the BSS CaBi. How the changes affect the BSS provider and its decision-making will be described in detail and extensions of how URI can benefit from the provided analysis are proposed, followed by the discussion and conclusion.

6.1 Evaluation of the Research Questions

First, it was hypothesized that the global COVID-19 pandemic might have caused changes in spatio-temporal behavioral patterns. Albeit a global event, there was still a local impact, which did indeed impact the usage patterns of CaBi´s riders in Washington D.C. comparing to both periods. Although spatial changes were not considerable in the analysis of the k-means approach, temporal aspects shifted from a rush hour-oriented ridership to late afternoon and evening peaks in the descriptive analysis.

Based on the result that individuals tend to use the BSS more likely in the afternoon with a more extent trip duration, it can be concluded that also the trip purpose has changed. Whereas studies (71, 104, 117), which were published before the pandemic, confirm the result of P1 that short trip durations during the rush hour peaks (around 7 - 9 am and 4 - 6 pm) were generated by primarily annual users, and P2 was defined by longer trip durations from casual members. As a result, the decrease of trips caused by annual members can be considered as a consequence of the governmental proclamation and the resulting increase of remote-work, validating the second research question.
Taking the spatial changes into consideration, no distinct change of the usage patterns could be identified. In the context of the third research question that the purpose of using BSS has shifted from those high dense employment areas to more recreational-friendly ones like parks and the suburbs, a change could not be identified due the approached methodology. Neither the descriptive statistics, correlation analysis, nor the k-means approach were suitable for analyzing the trip purpose or the destination and origin selection of the BSS users. The statement, therefore, remains unanswered.

Relating to the methodological approach, spatial and temporal changes in the usage behavior were revealed by using the clustering technique of k-means, correlation analyses, and descriptive statistics. Established on the trip history data and the COVID-19 case number data, deviations were noticed by comparing both periods, validating the fourth research question.

Contrary to the last statement, the number of clusters did not increase in the second period. Taking the heuristic Elbow method and the exact method of the Silhouette Score into account, the same number of clusters were defined for both analysis periods. On closer examination, there were only a few changes of station belonging to different clusters comparing P1 and P2.

6.2 Bike Sharing Systems

The study results provide new insights into the spatio-temporal use of a BSS by individuals before and during the COVID-19 pandemic and can be used by BSS managers and urban planners to optimize systems, locate new stations, and explore the feasibility of new strategies.
Caused by the identified time change of usage peaks in the evenings, some system operations should be reconsidered. While lower ridership has generally been noted, commuters and annual members will likely return to the system once the governmental restrictions are loosened. Then, the system would be faced with a high daily ridership level both on weekdays and weekends. Whereas trips in the morning rush hour would remain short but in high frequency, the later it was in the day, longer trip durations occur. Based on the comparison of the periods and their frequency and duration, system imbalance could probably occur. Based on the results of this study, CaBi should investigate if the forecast models and relocation strategies of the bicycles would handle those changes in the usage patterns. Furthermore, the flattened peaks of the hourly ridership level may alleviate the problem of relocation of the bicycles and their maintenance because bikes are not rented as frequently as before the Pandemic.

In the context of the clustering approach and exploratory techniques, the research confirms that they can be used as a first step towards the initial applicative objectives to plan urban areas, locate firm locations, and analyse fleet management. The results can be particularly used to calibrate simulation tools with realistic values on the operational decision level, optimize the bike distribution policies, and enhance the BSS system for repositioning, sizing stations, or general bike fleet size. Furthermore, crossing that methodological approach with socio-economical and geographical variables may provide evidence on the essential aspects of the city, explaining the demand variation. Therefore it may help design new predictive demand models used by system engineers to position and dimension the BSS stations. Eventually, the spatial analysis of the discovered clusters may be helpful to understand a complex
metropolitan mobility dynamic and benefit from a variety of applications, such as location choice for a business, advertisements, and social recommendations.

Because the trip duration during the pandemic tended to be longer than before, it is assumed that individuals use the BSS for recreational purposes and long-distance commuting, as more people look for socially distant ways to commute or exercise. Furthermore, public transportation was available only during limited hours of use during the pandemic, boosting trip numbers at the end of the service.

Next, analyzing the rental numbers of each station, stations with low frequency should be assessed for relocation from the provider. Table 10 through Table 12 show the influence of cycling behavior compiled in this research could establish new evaluation criteria. This may help plan future stations and develop incentives to rebalance the service performance better.

The evidence that the spatial distribution has remained mostly the same or has shifted only slightly, whereas the temporal distribution has changed significantly, indicates that the k-means is an adequate methodology approach to characterize a bicycle rental system, trip generation, and complexity. It is assumed that the spatial patterns just changed marginally because individuals remained in their homes, for the most part, to engage in social distancing.

Overall, the presented results will allow bicycle-sharing system operators to better plan their services by examining COVID-19 and the impact of travel frequency, distance, land use, and the built environment before and during the pandemic. The insights can be translated into guidelines to increase bike share activities in urban areas or extensions for URI can be made.
6.3 University of Rhode Island

In 2018, the consultant team Traverse Landscape Architects stated already in the Transportation Parking Master Plan (TPMP) that the University of Rhode Island (URI) lacks a bicycle system and network. In the report, a distinction was made between short term, intermediate, and long term improvement recommendations (118), which from today's perspective have only been partially implemented or not at all. Indeed, the university celebrated the two-mile branch to connect the Kingston Campus to the public bicycle path in 2019 (119). However, from the cycling perspective, there are still a lot of improvements necessary to enhance the bicycle traffic at URI. In the following, the literature review results and the conducted analysis, which can be generalized from the author's perspective to the URI campus site, are described.

First, as presented in the literature review, several factors affect the mode choice, especially for cyclists (Table 10 - Table 12). According to the tables, the built environment has a substantial impact on the level of cyclists. For example, a bicycle network with bicycle paths, which can be categorized as a factor of the built environment, is barely available at the URI campus. As highlighted in TPMP, the campus roads are primarily oriented for motorized vehicles (118), and besides, the branch connection was non-existent in 2018. There are several pedestrian walkaways next to roads, which could be used, but they are not consistently implemented or do not offer enough space for cyclists. Looking at Upper College Road, which is one of the main access roads to the campus, street lights or mailboxes obstruct the way and make it impossible to pass this way with a wheelchair or a bicycle. Should the cyclist switch
to the road, the rider would risk getting stuck in the asphalt crack, which spreads over the entire length of the street, with the rim (Figure 29). In order to encourage people to use the bicycle, effort should be made to establish the complete street principle, including bike lanes, landscape buffers, and brandings to accommodate all road users. The fundamental characteristics of this approach, with visual illustrations, can be derived from the TPMP (118). With the completion of this thesis, it should be noted that at Upper College Road, construction work has started. Whether and to what extent the principle of the complete street will be implemented remains to be seen.

Second, looking at the existing bicycle infrastructure next to the bicycle network, a number of bicycle storage racks can be identified by walking around the Campus. The black-painted bicycle racks are often located in close proximity to the teaching buildings, student housings, fraternities, and athletic fields. Some are firmly bolted to the ground, while others are movable. Nonetheless, some bicycle racks were identified from which conflicts with other traffic participants could arise and make parking more difficult. For example, at the site at the Keaney Road (in front of the Boss Ice Arena), cyclists would either block the pedestrian sidewalk or the disabled parking spots (Figure 27).

In contrast, another bicycle storage rack on Greenhouse Road is in front of a state fleet parking lot. At this location, cyclists are deprived of the opportunity to park the bike because a car blocks the closer area around the rack permanently (Figure 28). In order to increase bicycle-friendliness and attractiveness, such positions of bicycle stands should be avoided in the distribution policy, and identified misplaced racks should be relocated. To prevent bicycle storage from being blocked by cars, educational work on both sides is recommended to guarantee a respectful interaction among all road users.
Third, as identified in this research paper, people tend to use the bike for longer recreational trips. Again, the connection to the local bike path offers a high potential to encourage people to ride the bicycle, leading to other parks, green spaces, and bodies of water. Here, the Keaney parking lot could be used on weekends by bicycle enthusiasts as a park and ride site for day trips from people around the region. Offering an adequate infrastructure by letting people park and rest both the car and the bike may increase the general ridership level of the bike path and further improve the image of the URI campus. For this, a visitor parking lot must be implemented on the weekends.

Overall, the Kingston Campus is essentially in need of enhancements in bicycle friendliness before more bicycle traffic can be anticipated or an on-campus bicycle rental system could be established. Research revealed that a first-generation bicycle rental system called URIde was already in place at URI in September 2003. Here, 60 reconditioned bicycles were made freely available (120). Since there was no press release of an end of the URIde BSS, the author contacted the initiators of the system. Here, Dr. David E. Fastovsky, the Tuesday night bike repair workshop organizer, was reached. On the inquiry to what had happened to the BSS, he made the still timely statement that the “[… ]URI was not ready for a BikeShare program […]” in the context of the built environment and individuals behavior. For the BSS of the first generation, highlighted in Chapter 2.4.1, the bicycles were stolen, damaged, or hidden beside the street. If URI and the department for sustainability are seriously considered implementing a BSS at the URI, the built environment should be improved for cyclists in general.
On the one hand, the built-environment factor offers the possibility for cyclists to participate safely with traffic. On the other hand, it would alert attention to the mode share of cycling. Second, based on the learnings of the past, a station-based BSS connected to an IT-based technology should be taken into account to track bicycles and prevent the system's mistreatment.
6.4 Discussion

Although CaBi is one of the largest BSS in the U.S., when compared to other American BSS providers, it was the only one that was not recovered from COVID-19, whether in the short-term or the long-term periodic ridership level. This may have been caused by stricter governmental restrictions but could also be provoked by displeased service. Since this analysis only used trip history data, it is highly recommended that the service provider survey general customer satisfaction, as the previous study is from 2016.

Furthermore, because the methodological approach was applied to trip history data of the greater Washington D.C. area, the results may not necessarily be applicable to other cities or general cycling behavior. As mentioned in chapter 2.3.2, cycling and its behavior strongly depend on a variety of factors, which are dependent on the built environment, demographic metrics, or weather. Those differences were already identified in the literature before the pandemic and make it even more challenging to relate the results to other American BSSs.

Furthermore, this analysis was limited to a cluster analysis based on trip history data. For more information on how and where individuals are using CaBi, continuous GPS tracking of the bicycles should be enabled for scientific purposes. Similar to the study of Wergin and Buehler (2017) (117), a long-term study on more bikes can support urban planners in their decisions and help the service operator plan, implement and execute the mobility service efficiently.

Next, the conclusion made in this thesis can be further executed by multi-scale geographically weighted regression (MGWR) (121). This approach could validate the
impacts of COVID-19, concerning the factors identified in Table 10 - Table 12, affecting the bicycling behavior precisely. A virtual diameter could be plotted around the stations to characterize the factors around them and identify new characteristics, which may affect the usage of the BSS and improve the location of rental stations. Using the willingness to walk the distance to a station, BSS providers could use advertisement measures to draw attention to the system with the right incentives to increase the ridership. Because the mobility provider market is highly competitive, doing marketing to promote the system is more important than ever. Next, other mobility sharing providers are located in the greater Washington D.C. area, offering car, moped, scooter, and e-bike sharing systems, which could all interact with the CaBi BSS as part of one connected transportation system.

Nonetheless, all those analysis approaches require detailed, high-quality data. Already the data provided by CaBi had some inconsistencies in its features and changes, like different total numbers of stations in each period, as described in Chapter 4.2.2. Of course, it was tried to remove or replace those values as precisely as possible in the analysis process. Still, different approaches to handling missing values could lead to different results, which should be questioned. Additionally, the provided data about the land use was from 2005 but was accepted because the spatial changes were peripheral.

The COVID-19 pandemic did indeed impact the usage patterns of CaBi’s riders in Washington D.C. compared to the previous year. Although spatial changes were not considerable in the analysis of the k-means approach, temporal aspects shifted from a rush hour-oriented ridership to late afternoon and evening peaks in the descriptive analysis. Especially the stations in the third cluster show in the afternoon the most
frequent rentals with the most extended trip durations compared to the first period. Overall, the trip duration increased for both annual and casual member types, confronting the system provider with new challenges in its operational processes. Furthermore, a correlation of the increase of COVID-19 cases and the decrease of the ridership level was identified. Although the general ridership level decreased for both annual and casual members, the rise in the trip duration of the casual members formed new usage patterns, causing a new user behavior in the CaBi system. It is assumed that more people participated in the BSS to practice social distancing in their mode share and recreational purposes. Against the hypothesis that the number of clusters will increase and the distribution of stations will be noisy, no crucial changes in the spatial aspects of the number were identified. After all, it is not surprising that those usage patterns have not changed because individuals have lived in the same place despite the pandemic. It is assumed that residents found in CaBi a new way of transportation and/or changed their usage purpose, which would explain the longer trip durations and a higher percentage of casual members in the ridership level.

Despite this, CaBi should continuously monitor and question the rental behavior of its riders. Based on the findings of this thesis, resulting in the increased demand for bicycles in the afternoon for longer durations, system processes should be adjusted to match, if not improve on, the previous year's ridership level. Regardless of the pandemic, cycling remains a healthy and sustainable mode of transportation.
6.5 Conclusion

The study results provide new insights into the spatio-temporal use of a BSS by individuals before and during the COVID-19 pandemic and can be used by BSS managers and urban planners to optimize systems, locate new stations, and explore the feasibility of new strategies.

Aside from an increase in recreational usage, bike-sharing may have picked up some typical transit riders who have shifted away due to service cuts or out of a desire to socially distance more effectively. Municipalities where CaBi has coverage should closely look at this dataset, as it reveals where people are still moving in an era of greatly reduced mobility. These trips are displaying patterns not typically seen and can provide better insight on what locations people consider essential, and what facilities need greater support.

The descriptive analysis indicated that in the context of the ridership level and temporal distribution, differences in P1 and P2 were identified. Generally, the ridership level decreased since the beginning of the pandemic. Changes in the general usage behavior in terms of days of the week were noted during the pandemic, shifting from a concentrated weekday utilization to a primarily weekend orientated use. Also, higher utilization hours of the day shifted from two peaks at the rush hours to a constantly increasing ridership level starting at 7 am, reaching its peak with the most extended trip durations at 6 pm.

Taking the results of the correlation analysis into account, the trip frequency is significantly negatively correlated with the COVID-19 cases according to PCC and SRC. Only the KTC does not show any significant correlation, which may be due to the
different valuation methods applied. Looking at the trip duration and new cases, a slightly negative correlation was identified with the SRC. At this point, the hypothesis can be verified that the general COVID-19 cases and legislative restrictions harmed the ridership level of cyclists at CaBi´s BSS, despite a clear indicated data scatter. Also, the presented figures corroborate the conjecture that although fewer individuals are riding, they have traveled more long distances. As a result, the cycling behavior has changed in temporal aspects, frequency and trip duration from pre-pandemic to post-pandemic times.
### Table 8: Capital Bikeshare Quick Facts (1/2)

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Alexandria, Virginia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arlington, Virginia</td>
</tr>
<tr>
<td></td>
<td>Fairfax County, Virginia</td>
</tr>
<tr>
<td></td>
<td>Falls Church, Virginia</td>
</tr>
<tr>
<td></td>
<td>Montgomery County, Maryland</td>
</tr>
<tr>
<td></td>
<td>Prince George's County, Maryland</td>
</tr>
<tr>
<td></td>
<td>Washington, D.C. (71)</td>
</tr>
<tr>
<td>Opening date</td>
<td>September 20, 2010 (64)</td>
</tr>
<tr>
<td>Website</td>
<td>capitalbikeshare.com (71)</td>
</tr>
<tr>
<td>Size</td>
<td>140 square miles (68)</td>
</tr>
<tr>
<td>Service Area</td>
<td>4.09 stations per square mile (68)</td>
</tr>
<tr>
<td>Station Density</td>
<td>1100/ &lt; 4,500 (68)</td>
</tr>
<tr>
<td>Bikes (Start/Current)</td>
<td>114/ 527(68)</td>
</tr>
<tr>
<td>Stations (Start/Current)</td>
<td>11 to 39 (30)</td>
</tr>
<tr>
<td>Docks per Station range</td>
<td>24/7 year- round (71)</td>
</tr>
<tr>
<td>Operation</td>
<td></td>
</tr>
<tr>
<td>Ridership Metrics</td>
<td></td>
</tr>
<tr>
<td>Annual members</td>
<td>29,843 (68)</td>
</tr>
<tr>
<td>Casual members</td>
<td>35,100 (71)</td>
</tr>
<tr>
<td>Total Trips as of May 10th 2019</td>
<td>26,600,000 (68)</td>
</tr>
<tr>
<td>Annual trips (2019)</td>
<td>3,402,525 (29)</td>
</tr>
<tr>
<td>Service Area demographics (per sq. mi)</td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>5,010 jobs</td>
</tr>
<tr>
<td>Median Household Income</td>
<td>$ 66,508</td>
</tr>
<tr>
<td>Housing Density</td>
<td>6,344 units (64)</td>
</tr>
<tr>
<td>Equipment ownership</td>
<td>Jurisdiction (64)</td>
</tr>
<tr>
<td>Operator name:</td>
<td>Motivate (71)</td>
</tr>
<tr>
<td>Equipment Provider</td>
<td>PBSC Urban Solutions (64)</td>
</tr>
<tr>
<td>Business model</td>
<td>Jurisdiction owned and managed (64)</td>
</tr>
<tr>
<td>Funding source</td>
<td>Federal: CMAQ</td>
</tr>
<tr>
<td></td>
<td>Local: vehicle decal fee, commissions from transit fare media sales</td>
</tr>
<tr>
<td></td>
<td>Private: business sponsorship, Membership and usage fees (64)</td>
</tr>
<tr>
<td>City’s denomination (League of American Bicyclist)</td>
<td>Gold - Washington D.C.</td>
</tr>
<tr>
<td></td>
<td>Silver - Virginia</td>
</tr>
<tr>
<td></td>
<td>Bronze - Maryland (42)</td>
</tr>
</tbody>
</table>
### Table 9: Capital Bikeshare Quick Facts (2/2)

<table>
<thead>
<tr>
<th>Reported Bike Thefts (in 2012)</th>
<th>7 (64)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported bike-share crashes (in 2012)</td>
<td>1 (64)</td>
</tr>
<tr>
<td>Bike facilities characteristics</td>
<td>120.6 miles of marked bike lines in Washington D.C.; Continously growing network of bike lanes, signed bike routes, and trails planned (J22)</td>
</tr>
<tr>
<td>Membership and usage fees</td>
<td>$75 annual; $25 30 days; $15 3 days; $7 24 hours. No fee for first 30 min; $1.50/$2.00 annual/casual members 30-60 min; $4.50/$6.00 for annual/casual members 60-90 minutes; $6/$8 for annual/casual members for every half-hour after that (50)</td>
</tr>
<tr>
<td>Category</td>
<td>Variable</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td><strong>Socio-demographic</strong></td>
<td>Population density</td>
</tr>
<tr>
<td></td>
<td>Employment Density</td>
</tr>
<tr>
<td></td>
<td>Racial structure</td>
</tr>
<tr>
<td></td>
<td>Age</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Education</td>
</tr>
<tr>
<td></td>
<td>Income</td>
</tr>
<tr>
<td><strong>Built environment/land use</strong></td>
<td>Intersections</td>
</tr>
<tr>
<td></td>
<td>Bike Infrastructure</td>
</tr>
<tr>
<td></td>
<td>Stations</td>
</tr>
<tr>
<td></td>
<td>Transit</td>
</tr>
<tr>
<td></td>
<td>University</td>
</tr>
<tr>
<td>Category</td>
<td>Variable</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td><strong>Built environment/land use</strong></td>
<td><strong>Level of active travel (land use mix)</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Entropy</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Employment</strong></td>
</tr>
<tr>
<td><strong>Road and facility variables</strong></td>
<td><strong>Principal</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Arterial</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Collector</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Local</strong></td>
</tr>
<tr>
<td></td>
<td><strong>On-Street</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Off-Street</strong></td>
</tr>
<tr>
<td><strong>Recreation</strong></td>
<td><strong>Restaurant</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Lake</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Water</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Park</strong></td>
</tr>
</tbody>
</table>

Table 11: Factors Correlated with the Use of Bike-Sharing Stations (2/3)
<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Definition</th>
<th>Unit</th>
<th>Authors</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>Weekend</td>
<td>Reduced number of commuter trips, but more recreational</td>
<td>NA</td>
<td>(51, 123)</td>
<td>- / +</td>
</tr>
<tr>
<td></td>
<td>Season</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spring</td>
<td></td>
<td>NA</td>
<td>(123, 124)</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Summer</td>
<td></td>
<td>NA</td>
<td>(123, 124)</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Fall</td>
<td></td>
<td>NA</td>
<td>(123, 124)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Winter</td>
<td></td>
<td>NA</td>
<td>(123, 124)</td>
<td>-</td>
</tr>
<tr>
<td>Weather</td>
<td>Temperature</td>
<td>Deviation of daily average temperature from the long-term average, in Fahrenheit</td>
<td>°F</td>
<td>(51, 122–124)</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Humidity</td>
<td>Relative humidity</td>
<td>%</td>
<td>(122–124)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Precipitation</td>
<td>Amount of precipitation</td>
<td>Mm</td>
<td>(122–124)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sun</td>
<td>Deviation of daily percentage sunshine from the long-term average</td>
<td>%</td>
<td>(122–124)</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Rain</td>
<td>Precipitation per square feet</td>
<td>Inch/square feet</td>
<td>(122–124)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Snow</td>
<td>Amount of snow on the ground</td>
<td>Cm</td>
<td>(123, 124)</td>
<td>-</td>
</tr>
<tr>
<td>Others</td>
<td>Crime rate</td>
<td>Number of violent crimes per 100,000 persons</td>
<td>Crimes / 100,000</td>
<td>(51, 122)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Number of bicycle thefts</td>
<td></td>
<td>NA</td>
<td>(126)</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Owning of bicycle</td>
<td>The proportion of the population owning one or more bicycle(s)</td>
<td>%</td>
<td>(126)</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 13: Most Used Origin-Destination Paths in P1

<table>
<thead>
<tr>
<th>Path Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW//Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW</td>
<td>4491</td>
</tr>
<tr>
<td>Columbus Circle / Union Station//6th &amp; H St NE</td>
<td>4305</td>
</tr>
<tr>
<td>6th &amp; H St NE/Columbus Circle / Union Station</td>
<td>3612</td>
</tr>
<tr>
<td>Columbus Circle / Union Station//8th &amp; F St NE</td>
<td>3413</td>
</tr>
<tr>
<td>Jefferson Dr &amp; 14th St SW//Jefferson Dr &amp; 14th St SW</td>
<td>3259</td>
</tr>
<tr>
<td>4th St &amp; Madison Dr NW//4th St &amp; Madison Dr NW</td>
<td>3226</td>
</tr>
<tr>
<td>Lincoln Memorial/Jefferson Memorial</td>
<td>3072</td>
</tr>
<tr>
<td>8th &amp; F St NE/Columbus Circle / Union Station</td>
<td>2929</td>
</tr>
<tr>
<td>Jefferson Dr &amp; 14th St SW//Lincoln Memorial</td>
<td>2925</td>
</tr>
<tr>
<td>Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW//Lincoln Memorial</td>
<td>2789</td>
</tr>
<tr>
<td>15th St &amp; Constitution Ave NW//15th St &amp; Constitution Ave NW</td>
<td>2722</td>
</tr>
<tr>
<td>17th St &amp; Independence Ave SW//Lincoln Memorial</td>
<td>2632</td>
</tr>
<tr>
<td>Henry Bacon Dr &amp; Lincoln Memorial Circle NW//Henry Bacon Dr &amp; Lincoln Memorial Circle NW</td>
<td>2480</td>
</tr>
<tr>
<td>14th &amp; Irving St NW//11th &amp; Girard St NW</td>
<td>2469</td>
</tr>
<tr>
<td>Lincoln Memorial/Lincoln Memorial</td>
<td>2456</td>
</tr>
<tr>
<td>Lincoln Memorial/Jefferson Dr &amp; 14th St SW</td>
<td>2411</td>
</tr>
<tr>
<td>Gravelly Point/Gravelly Point</td>
<td>2378</td>
</tr>
<tr>
<td>13th &amp; H St NE/Columbus Circle / Union Station</td>
<td>2363</td>
</tr>
<tr>
<td>17th St &amp; Independence Ave SW//17th St &amp; Independence Ave SW</td>
<td>2309</td>
</tr>
<tr>
<td>Maryland Ave &amp; E St NE/Columbus Circle / Union Station</td>
<td>2282</td>
</tr>
</tbody>
</table>

### Table 14: Most Used Origin Stations in P1

<table>
<thead>
<tr>
<th>Station Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbus Circle / Union Station</td>
<td>56425</td>
</tr>
<tr>
<td>15th &amp; P St NW</td>
<td>33985</td>
</tr>
<tr>
<td>14th &amp; Irving St NW</td>
<td>33498</td>
</tr>
<tr>
<td>Lincoln Memorial</td>
<td>33380</td>
</tr>
<tr>
<td>New Hampshire Ave &amp; T St NW</td>
<td>32171</td>
</tr>
<tr>
<td>Henry Bacon Dr &amp; Lincoln Memorial Circle NW</td>
<td>31904</td>
</tr>
<tr>
<td>4th St &amp; Madison Dr NW</td>
<td>30722</td>
</tr>
<tr>
<td>Jefferson Dr &amp; 14th St SW</td>
<td>30554</td>
</tr>
<tr>
<td>1st &amp; M St NE</td>
<td>30340</td>
</tr>
<tr>
<td>Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW</td>
<td>29763</td>
</tr>
</tbody>
</table>

### Table 15: Most Used Destination Stations in P1

<table>
<thead>
<tr>
<th>Station Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbus Circle / Union Station</td>
<td>60725</td>
</tr>
<tr>
<td>15th &amp; P St NW</td>
<td>36330</td>
</tr>
<tr>
<td>Lincoln Memorial</td>
<td>32905</td>
</tr>
<tr>
<td>Henry Bacon Dr &amp; Lincoln Memorial Circle NW</td>
<td>32064</td>
</tr>
<tr>
<td>Jefferson Dr &amp; 14th St SW</td>
<td>32057</td>
</tr>
<tr>
<td>Massachusetts Ave &amp; Dupont Circle NW</td>
<td>31781</td>
</tr>
<tr>
<td>Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW</td>
<td>31104</td>
</tr>
<tr>
<td>4th St &amp; Madison Dr NW</td>
<td>31036</td>
</tr>
<tr>
<td>1st &amp; M St NE</td>
<td>31028</td>
</tr>
<tr>
<td>14th &amp; V St NW</td>
<td>30200</td>
</tr>
</tbody>
</table>
### Table 16: Most Used Origin-Destination Paths in P2

<table>
<thead>
<tr>
<th>Path Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jefferson Dr &amp; 14th St SW//Jefferson Dr &amp; 14th St SW</td>
<td>4340</td>
</tr>
<tr>
<td>15th St &amp; Constitution Ave NW//15th St &amp; Constitution Ave NW</td>
<td>3873</td>
</tr>
<tr>
<td>Hains Pt/Buckeye &amp; Ohio Dr SW//Hains Pt/Buckeye &amp; Ohio Dr SW</td>
<td>3661</td>
</tr>
<tr>
<td>Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW//Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW</td>
<td>3368</td>
</tr>
<tr>
<td>Anacostia Park//Anacostia Park</td>
<td>3366</td>
</tr>
<tr>
<td>4th St &amp; Madison Dr NW//4th St &amp; Madison Dr NW</td>
<td>3319</td>
</tr>
<tr>
<td>Gravelly Point//Gravelly Point</td>
<td>3227</td>
</tr>
<tr>
<td>34th &amp; Water St NW//34th &amp; Water St NW</td>
<td>2732</td>
</tr>
<tr>
<td>National Harbor Carousel//National Harbor Carousel</td>
<td>2394</td>
</tr>
<tr>
<td>Lincoln Memorial//Lincoln Memorial</td>
<td>2268</td>
</tr>
<tr>
<td>Ohio Dr &amp; West Basin Dr SW / MLK &amp; FDR Memorials / Ohio Dr &amp; West Basin Dr SW / MLK &amp; FDR Memorials</td>
<td>2100</td>
</tr>
<tr>
<td>1st &amp; M St NE//1st &amp; M St NE</td>
<td>1980</td>
</tr>
<tr>
<td>Roosevelt Island//Roosevelt Island</td>
<td>1975</td>
</tr>
<tr>
<td>17th St &amp; Independence Ave SW//17th St &amp; Independence Ave SW</td>
<td>1931</td>
</tr>
<tr>
<td>Henry Bacon Dr &amp; Lincoln Memorial Circle NW//Henry Bacon Dr &amp; Lincoln Memorial Circle NW</td>
<td>1926</td>
</tr>
<tr>
<td>4th &amp; M St SW//4th &amp; M St SW</td>
<td>1820</td>
</tr>
<tr>
<td>19th St &amp; Constitution Ave NW//19th St &amp; Constitution Ave NW</td>
<td>1668</td>
</tr>
<tr>
<td>Anacostia Ave &amp; Benning Rd NE / River Terrace //Anacostia Ave &amp; Benning Rd NE / River Terrace</td>
<td>1428</td>
</tr>
<tr>
<td>Lincoln Memorial//Jefferson Dr &amp; 14th St SW</td>
<td>1318</td>
</tr>
<tr>
<td>Prince St &amp; Union St//Prince St &amp; Union St</td>
<td>1307</td>
</tr>
</tbody>
</table>

### Table 17: Most Used Origin Stations in P2

<table>
<thead>
<tr>
<th>Station Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lincoln Memorial</td>
<td>21434</td>
</tr>
<tr>
<td>New Hampshire Ave &amp; T St NW</td>
<td>18814</td>
</tr>
<tr>
<td>Jefferson Dr &amp; 14th St SW</td>
<td>18775</td>
</tr>
<tr>
<td>4th St &amp; Madison Dr NW</td>
<td>18602</td>
</tr>
<tr>
<td>1st &amp; M St NE</td>
<td>17442</td>
</tr>
<tr>
<td>15th &amp; P St NW</td>
<td>17179</td>
</tr>
<tr>
<td>Henry Bacon Dr &amp; Lincoln Memorial Circle NW</td>
<td>16127</td>
</tr>
<tr>
<td>11th &amp; M St NW</td>
<td>15264</td>
</tr>
<tr>
<td>4th &amp; M St SW</td>
<td>15109</td>
</tr>
<tr>
<td>Smithsonian-National Mall / Jefferson Dr &amp; 12th St SW</td>
<td>14256</td>
</tr>
</tbody>
</table>

### Table 18: Most Used Destination Stations in P2

<table>
<thead>
<tr>
<th>Path Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>15th &amp; P St NW</td>
<td>20583</td>
</tr>
<tr>
<td>Lincoln Memorial</td>
<td>19651</td>
</tr>
<tr>
<td>1st &amp; M St NE</td>
<td>19016</td>
</tr>
<tr>
<td>Jefferson Dr &amp; 14th St SW</td>
<td>18882</td>
</tr>
<tr>
<td>New Hampshire Ave &amp; T St NW</td>
<td>18696</td>
</tr>
<tr>
<td>4th St &amp; Madison Dr NW</td>
<td>18011</td>
</tr>
<tr>
<td>14th &amp; V St NW</td>
<td>17322</td>
</tr>
<tr>
<td>4th &amp; M St SW</td>
<td>16098</td>
</tr>
<tr>
<td>Henry Bacon Dr &amp; Lincoln Memorial Circle NW</td>
<td>15954</td>
</tr>
<tr>
<td>15th &amp; W St NW</td>
<td>14156</td>
</tr>
</tbody>
</table>
Figure 23: Population Density Washington D.C. (123)
Figure 24: Employment Density Washington D.C. (123)
Figure 25: Central Employment Area Washington D.C. (123)
Figure 26: Great Streets and Transit Stations in Washington D.C. (123)
Figure 27: Misplaced Bicycle Storage Rack at Keaney Road (Boss Ice Arena)
Figure 28: Blocked Bicycle Storage Rack at Greenhouse Road
Figure 29: Upper College Road: Lanterns, Power Lines, Parked Vehicles and Cracks in the Asphalt Make Bicycle Traffic Problematic
Figure 30: Closed Access to the Bike Path at Brookside Residence Hall (W Alumni Avenue)
CODE

Concatenating and Feature Engineering

Import Libraries

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.api.types import CategoricalDtype
from matplotlib.cycler import import_register_formatter

import warnings
warnings.filterwarnings('ignore')
```

Period 1 (P1)

Import of Raw Data for P1 Dataset

```python
mar2019 = pd.read_csv('2019-capitalbikeshare-tripdata-march.csv')
apr2019 = pd.read_csv('2019-capitalbikeshare-tripdata-april.csv')
may2019 = pd.read_csv('2019-capitalbikeshare-tripdata-may.csv')
jun2019 = pd.read_csv('2019-capitalbikeshare-tripdata-june.csv')
Jul2019 = pd.read_csv('2019-capitalbikeshare-tripdata-july.csv')
Aug2019 = pd.read_csv('2019-capitalbikeshare-tripdata-august.csv')
Sep2019 = pd.read_csv('2019-capitalbikeshare-tripdata-september.csv')
Oct2019 = pd.read_csv('2019-capitalbikeshare-tripdata-october.csv')
Nov2019 = pd.read_csv('2019-capitalbikeshare-tripdata-november.csv')
Dec2019 = pd.read_csv('2019-capitalbikeshare-tripdata-december.csv')
Jan2020 = pd.read_csv('2020-capitalbikeshare-tripdata-january.csv')
Feb2020 = pd.read_csv('2020-capitalbikeshare-tripdata-february.csv')
Mar2020 = pd.read_csv('2020-capitalbikeshare-tripdata-march.csv')
```

Concatenating P1 Data

```python
                     nov2019, dec2019, jan2020, feb2020, mar2020])
year2019.shape
```

Data Transformation P1

```python
year2019.to_csv('period-1_tripdata.csv', index=False)
trip_data_pl = pd.read_csv('period-1_tripdata.csv')
trip_data_pl.head()
```

```python
trip_data_pl.drop(trip_data_pl.columns[[3,5,7]], axis=1, inplace = True)
trip_data_pl
```

Checkig for Missing Data

```python
trip_data_pl.isnull().sum()
```

Creating Column for COVID Cases

Adding Station Locations

```python
location_df = pd.read_csv('CapitalBikeshare_Stations.csv')
location_df.head()
```

```python
data_merged_period1 = pd.merge(trip_data_pl, location_df, on = 'Start station', how='inner')
data_merged_period1
```

```python
location_end_df = pd.read_csv('CapitalBikeshare_Stations_End.csv')
location_end_df.head()
```

```python
data_merged_period2 = pd.merge(data_merged_period1, location_end_df, on = 'End station', how='inner')
data_merged_period2
```

Figure 31: Code for Concatenation (1/3)
Print Merged Data for P1 into .csv file for Analysis

```
data_merged_period1.to_csv('period-1_tripdata.csv', index=False)
```

General Descriptive Analysis for P1

```
data_merged_period1.describe().T
```

Flow Stations Statistic for P1 (Same Start and End Station)

```
sum_column = data_merged_period1["Start station"]."/""+data_merged_period1["End station"]
```

```
sum_column.value_counts().head(10)
```

Creating New Station List for End Stations in P1

```
data_merged_period1["Start station"] .to_csv("StartStation.csv")
```

Frequency of End Stations in P1

```
end_station_count_pl = data_merged_period1["End station"]=value_counts()
```

Frequency of Start Stations in P1

```
start_station_count_pl = data_merged_period1["Start station"]=value_counts()
```

Period 2 (P2)

Import of Raw Data for P2

```
mar2020 = pd.read_csv("2020-2-capitalbikeshare-tripdata.csv")
apr2020 = pd.read_csv("2004-capitalbikeshare-tripdata.csv")
may2020 = pd.read_csv("2009-capitalbikeshare-tripdata.csv")
jun2020 = pd.read_csv("2007-capitalbikeshare-tripdata.csv")
```

```
 aug2020 = pd.read_csv("2009-capitalbikeshare-tripdata.csv")
sep2020 = pd.read_csv("2009-capitalbikeshare-tripdata.csv")
```

```
 oct2020 = pd.read_csv("2010-capitalbikeshare-tripdata.csv")
nov2020 = pd.read_csv("2011-capitalbikeshare-tripdata.csv")
dec2020 = pd.read_csv("2012-capitalbikeshare-tripdata.csv")
```

```
jan2021 = pd.read_csv("2011-capitalbikeshare-tripdata.csv")
```

```
```

```
year2020 .shape
```

```
year2020 .head(5)
```

Data Transformation

Deleting Electric Bikes from the Dataset

```
year2020 .drop(year2020.index[year2020["rideable_type"] == 'electric_bike'], inplace=True)
```

```
year2020 .to_csv('period-2_tripdata.csv', index=False)
```

```
trip_data_p2 = pd.read_csv('period-2_tripdata.csv')
```

```
trip_data_p2 .head()
```

**Figure 32: Code for Concatenation (2/3)**
Figure 33: Code for Concatenation (3/3)
Descriptive Analysis

Import Libraries

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
sns.set_style("whitegrid")
```

Defined Colorcodes of Colorblind Scheme Provided by Seaborn

```python
colors_member = ["#FF4444", "#8F4E4E"]
c_period = ["#9A001B", "#F59E0B"]
c_cluster = ["#007B8B", "#AD936C", "#6E3B92"]
c_covid = ["#DEE0D5"]
```

Load Datasets

```python
df1 = pd.read_csv("period-1_tripdata.csv")
df2 = pd.read_csv("period-2_tripdata.csv")
```

Period 1 (P1)

Illustrate Dataframe

```python
df1.head()
```

Check the Datatype of the Dataset

```python
df1.info()
```

Check of Null Values

```python
df1.isnull().sum()
```

Descriptive Analysis

```python
df1.describe()
```

Elimination of All Negative Values

```python
df1 = df1[df1["Duration"] > 0]
```

Change of Datatype from Object to Datetime

```python
df1["Start_date"] = pd.to_datetime(df1["Start_date"])
df1["End_date"] = pd.to_datetime(df1["End_date"])
```

Create a Day, Hour, Year, Month Column for Start Date

```python
df1["Day"] = df1["Start_date"].dt.day
df1["Month"] = df1["Start_date"].dt.month
df1["Year"] = df1["Start_date"].dt.year
df1["Hour"] = df1["Start_date"].dt.hour
```

Figure 34: Code for Descriptive Analysis (1/2)
For the procedure of the P 2, please retrieve the electronic version.

Figure 35: Code for Descriptive Analysis (2/2)
Correlation Analysis

Import of Libraries

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import folium
import seaborn as sns
import emoji

pd.set_option('display.float_format', lambda x: '%.5f' % x)
```

Create COVID-19 Dataframe

```python
dataframe = pd.read_csv("COVID-19 Cases_Washington_D.C_.nav.csv", delimiter = ";")
df_covid = dataframe.copy()
df_covid
```

Display Datatype of COVID-19 Dataframe

```python
df_covid.dtypes
```

Descriptive Statistics

```python
df_covid.describe()
```

Check for Null Values

```python
df_covid.isnull().sum()
```

Create Dataframe for Trips History Data

```python
df = pd.read_csv("period-2-final_tripdata.csv")
df_bike = df.copy()
df_bike
```

Check Datatype

```python
df_bike.dtypes
```

Descriptive Statistics

```python
df_bike.describe()
```

Check Datatype

```python
df_bike.dtypes
```

Descriptive Statistics

```python
df_bike.describe()
```

Check for Null Values

```python
df_bike.isnull().sum()
```
Clean the Dataframe

Correct Column Names

```python
df_bike.columns = ['Start Time', 'NULL', 'End Date', 'Start Station', 'End Station', 'Latitude', 'Longitude', 'End Latitude', 'Start Time', 'Member Type', 'Start Date', 'Total Positives', 'New Cases']
df_bike
```

Change Datatypes of String to Date Time Object

```python
df_bike['Start Date'] = df_bike['Start Date'].apply(np.datetime64)
df_bike['End Date'] = df_bike['End Date'].apply(np.datetime64)
df_bike['Date'] = df_covid['Date'].apply(np.datetime64)
df_covid['Date']
```

Create Duration Column

```python
df_bike['Duration'] = (df_bike['End Date'] - df_bike['Start Date']).dt.total_seconds()
df_bike
```

Correlation Analysis by Count

```python
df_bycount = df_bike[['Start Date', 'Duration']]
df_bycount['Start Date'] = df_bycount['Start Date'].dt.date
for key in df_bycount.columns:
    df_bycount = df_bycount.groupby([key]).count()
df_bycount
```

```python
df_bycount['Total Positives'] = df_covid['Total Positives'].values
df_bycount['New Cases'] = df_covid['New Cases'].values
```

Create Pair Plot

```python
plt.figure(figsize=(20,10))
g = sns.pairplot(df_bycount, diag_kind='kde', height=4)
g.map_diag(plt.hist)
g.map_upper(sns.kdeplot, levels=4, color='.2')
```

Pearson’s R Correlation Matrix

```python
df_bycount.corr()
```

Spearman’s RHO Correlation Matrix

For Duration and Total Positives

```python
td = scipy.stats.spearmanr(df_bycount['Total Positives'], df_bycount['Duration'])[0]
td
```

For Duration and New Cases

```python
nd = scipy.stats.spearmanr(df_bycount['New Cases'], df_bycount['Duration'])[0]
nd
```

For Total Positives and New Cases

```python
nt = scipy.stats.spearmanr(df_bycount['New Cases'], df_bycount['Total Positives'])[0]
nr
```

**Figure 37: Code for Correlation Analysis (2/5)**
Create Summary Dataframe

dictionary = {"Duration": [1, td, nd], "Total Positives": [td, l, nt], "New Cases": [nd, nt, l]}
df_bysum = pd.DataFrame(dictionary)
corr = df_bysum.corr()

By Sum

df_bysum = df_bysum["Start Date", "Duration"]
df_bysum.loc["Start Date"] = df_bysum.loc["Start Date"].dt.date.copy()
df_bysum = df_bysum.groupby(["Start Date"]).sum()
df_bysum

df_bysum["Total Positives"] = df_covid["Total Positives"].values
df_bysum["New Cases"] = df_covid["New Cases"].values

df_bysum

Create Pair Plot

plt.figure(figsize=(20,10))
g = sns.pairplot(df_bysum, diag_kind="kde", height = 4)
g.map_diag(plt.hist)
g.map_upper(sns.kdeplot, levels=4, color=".2")

Pearson’s R Correlation Matrix

df_bysum.corr()

Spearman’s RHO Correlation Matrix

For Duration and Total Positives

td = scipy.stats.spearmanr(df_bysum["Total Positives"], df_bysum["Duration"])[0]
td

For Duration and New Cases

nd = scipy.stats.spearmanr(df_bysum["New Cases"], df_bysum["Duration"])[0]
nd

For Total Positives and New Cases

nt = scipy.stats.spearmanr(df_bysum["New Cases"], df_bysum["Total Positives"])[0]
nt

Create Dataframe Summary

dictionary = {"Duration": [1, td, nd], "Total Positives": [td, l, nt], "New Cases": [nd, nt, l]}
corr = pd.DataFrame(dictionary)
corr

Kendal’s Tau Correlation Matrix

For Duration and Total Positives

td = scipy.stats.kendalltau(df_bysum["Total Positives"], df_bysum["Duration"])[0]
td

For Duration and New Cases

nd = scipy.stats.kendalltau(df_bysum["New Cases"], df_bysum["Duration"])[0]

Figure 38: Code for Correlation Analysis (3/5)
For Total Positives and New Cases

```python
nt = scipy.stats.kendalltau(df_bysum("New Cases"), df_bycount("Total Positives"))[0]
```

Create Dataframe Summary

```python
dictionary = {"Duration" : dl, ts, nd, "Total Positives" : [td, 1, nt], "New Cases" : [ind, nt, 11]}
corr = pd.DataFrame(dictionary)
corr
```

Regression Plots

For Overall Frequency vs Case Number

```python
overall = df_bycount
overall
```

For Overall Frequency vs Total Positives

```python
ax = sns.scatterplot(x = "Duration", y = "Total Positives", data = overall, height = 8, line_kws=./
ax.set(ylabel="Frequency")
```

For Overall Frequency vs New Cases

```python
ax = sns.scatterplot(x = "Duration", y = "New Cases", data = overall, height = 8, line_kws=('color': 'red'))
ax.set(ylabel="Frequency")
```

Analysis According to Membertype

For Annual Members

```python
member = df_hike[df_hike("Member Type") == "member"]
member = member[\["Start Date", "Duration"]]
member.loc[\["Start Date"] = member.loc[\["Start Date"].dt.date.copy()]
member = member.groupby("Start Date").count()
member[\["Total Positives"] = df_covid[\["Total Positives"]].values
member[\["New Cases"] = df_covid[\["New Cases"]].values
```

Regression Plot for Frequency vs Total Positives

```python
ax = sns.scatterplot(x = "Duration", y = "Total Positives", data = member, height = 8, line_kws=./
ax.set(ylabel="Frequency")
```

Figure 39: Code for Correlation Analysis (4/5)
Figure 40: Code for Correlation Analysis (5/5)
Cluster Analysis

Importing Libraries

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import folium
from sklearn import datasets
from sklearn.preprocessing import StandardScaler
pd.set_option('display.float_format', lambda x: '%.5f' % x)
colors_namer = ['#f0e4b8', '#f59948']
```

Read Datasets According to Periods

```python
df_2019 = pd.read_csv("period-1-final_tripdata.csv", header = "infer", infer_datetime_format = True)
df_2020 = pd.read_csv("period-2-final_tripdata.csv", header = "infer", infer_datetime_format = True)
```

Period 1 (P1)

Create Copy of Dataset

```python
df = df_2019.copy()
df.head()
```

Display Descriptive Statistics

```python
df.describe().T
print("Rows:", df.shape[0])
print("Columns:", df.shape[1])
```

Check Data Types

```python
df.dtypes
```

Cleaning Data

1) Correcting Column Names:

```python
df.columns = ['Duration', 'Start Date', 'End Date', 'Start Station', 'End Station', 'Member Type', 'COVID-19 Cases', 'Latitude Start', 'Longitude Start', 'Latitude End', 'Longitude End']
df
```

2) Transform Data Types

```python
df["Start Date"] = df["Start Date"].apply(np.datetime64)
df["End Date"] = df["End Date"].apply(np.datetime64)
df.dtypes
```

`# Transform Latitudes and Longitudes in Appropriate Format for Library`

Figure 41: Code for Cluster Analysis (1/7)
```python
def value_change(x):
    lst = list(x)
    if x > 0:
        return float("%.2f" % (lst[2] / lst[3]))
    else:
        return float("%.2f" % (lst[2] / lst[3]))

df["Longitude Start"] = df["Longitude Start"].apply(value_change)
df["Latitude Start"] = df["Latitude Start"].apply(value_change)
df["Longitude End"] = df["Longitude End"].apply(value_change)
df["Latitude End"] = df["Latitude End"].apply(value_change)
df["Longitude Start", "Latitude Start", "Longitude End", "Latitude End"]

# Remove Data Less Than 60 seconds and Outliers

df = df[(df["Duration"] >= 60)]
q1 = df["Duration"].quantile(0.25)
q3 = df["Duration"].quantile(0.75)
interquartile_range = q3 - q1
df = df[~((df["Duration"] < (q1 - 1.5 * interquartile_range)) | (df["Duration"] > (q3 + 1.5 * interquartile_range)))].reset_index(drop=True)

Transforming Data

1) Create Day, Departure and Arrival Column

df["Day"] = df["Start Date"].dt.day_name()
df["Departure"] = df["Start Date"].dt.hour
df["Arrival"] = df["End Date"].dt.hour
df.head()

2) Separate Days into Weekdays and Weekends

def days(x):
    if x in ["Monday", "Tuesday", "Wednesday", "Thursday", "Friday"]:
        return "Week-Day"
    else:
        return "Week-End"
df["Day Type"] = df.Day.apply(days)
df_graph_1019 = df.copy()
df.head()

3) Creating a New Dataframe for Visualization

df_geo = df[["Start Station", "Latitude Start", "Longitude Start"]].drop_duplicates()
print["Total number of Stations:", df_geo.shape[0]]
df_geo

Visualization

Plot Geospatial Data of Bicycle Stations before Clustering for P1 in Map

latitude = 38.9072
longitude = -77.0365

washington_map = folium.Map(location=[latitude, longitude], zoom_start=12)

for lat, lng, label in zip(df_geo["Latitude Start"], df_geo["Longitude Start"], df_geo["Start Station"]):
    folium.Marker([lat, lng],
                  icon=folium.Icon(color='black', fill='true',
                                   popup=label, fill_color='white', fill_opacity=0.6)).add_to(washington_map)

washington_map
```

Figure 42: Code for Cluster Analysis (2/7)
Figure 43: Code for Cluster Analysis (3/7)
Search for Null Values

```
df_bystation_unstack.T.isnull().sum()
```

Replace NaN with Zeros

```
df_bystation_unstack.fillna(0, inplace = True)
df_bystation_unstack.T.isnull().sum()
```

```
df_normal = df_bystation_unstack.copy()
```

Adding COVID-Cases Column

```
df_bystation_unstack["Covid"] = df.groupby(["Start Station"]).sum()["COVID-19 Cases"].sort_index()
df_bystation_unstack
```

```
columns = df_bystation_unstack.columns
index = df_bystation_unstack.index
```

Feature Scaling

```
s_s = StandardScaler()
df_bystation_unstack = s_s.fit_transform(df_bystation_unstack)
df_bystation_unstack = pd.DataFrame(df_bystation_unstack, index = index, columns = columns)
df_bystation_unstack
```

Train k-Means Using Random Number of Clusters

```
from sklearn.cluster import KMeans
km = KMeans(n_clusters = 6, init = 'k-means++', max_iter=1000)
km.fit(df_bystation_unstack)
kmeans_inertia_
```

Optimize hyperparameter Using Elbow-Method

```
from sklearn.cluster import KMeans
vcss = []
for i in range(1, 10):
    kmeans = KMeans(n_clusters = i, init = 'k-means++')
    kmeans.fit(df_bystation_unstack)
    vcss.append(kmeans.inertia_)
plt.figure(figsize = (20,10))
plt.plot(range(1,10), vcss, marker = 'o')
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('Distance')
plt.annotate("Elbow", xy = (5.75,5500), xytext = (3.65,7550), arrowprops = dict(facecolor = "red", shrink = 0.01))
plt.show()
```

Apply Silhouette Score

```
clusters = []
im = []
for k in range(1, 10, 1):
    model = KMeans(n_clusters = k, init = 'k-means++')
    model.fit(df_bystation_unstack)
    clusters.append(model)
    im.append(model.inertia_)
```

```
from sklearn.metrics import silhouette_score
for i in range(1,10,1):
    print("-----------------------------")
    print(clusters[i])
    print("Silhouette score:")(df_bystation_unstack, clusters[i].predict(df_bystation_unstack))
```

**Figure 44: Code for Cluster Analysis (4/7)**
Train k-Means Based on Result of Silhouette Score

```r
km = kmeans(n_clusters = 5, init = "k-means++", max_iter=1000)
kfit <- km.fit(df_bystation_unstack)
set(km.labels_)
```

Add Label (Cluster) Column

```r
km.labels_ <-
df_normal.sort_index(inplace = True)
df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
```

Add Latitude and Longitude Coordinates

```r
df_bystation_unstack['Latitude'] = df_geo.sort_index('Start Station').sort_index(['Latitude Start'])
df_bystation_unstack['Longitude'] = df_geo.sort_index('Start Station').sort_index(['Longitude Start'])
df_bystation_unstack['Period'] = df_period_unstack['Period']
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
df_bystation_unstack['Covid'] = km.labels_
df_bystation_unstack['Covid'] = df_bystation_unstack.drop('Covid', axis = 1, inplace = True)
df_bystation_unstack['Label'] = km.labels_
```

Create Geospatial Map of Clusters

```r
latitude = 38.9072
longitude = -77.0365
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=15)
rainbow = ['#78e68d', '#cc790c', '#f21e30']
markers_colors = []
for lat, lon, pos, cluster in zip(df_bystation_unstack['Latitude'],
                                 df_bystation_unstack['Longitude'],
                                 df_bystation_unstack['index'],
                                 df_bystation_unstack['Label']):
    label = folium.Popup(str(pos) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=3,
        popup=label,
        color=rainbow[cluster],
        fill=True,
        fill_color=rainbow[cluster],
        fill_opacity=1).add_to(map_clusters)
```

Create Interactive Map

```r
washington_map = folium.Map(location=[latitude, longitude], zoom_start=5)
incidents = interactive(washington_cluster_show, cluster=range(n_clusters))
add_to(washington_map)
rainbow = ['#0f7feb', '#ff5733', '#f21e30']
for lat, lng, label, cluster in zip(df_bystation_unstack['Latitude'],
                                    df_bystation_unstack['Longitude'],
                                    df_bystation_unstack['index'],
                                    df_bystation_unstack['Label']):
    label = folium.Popup(str(pos) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=0,
        popup=label,
        color=rainbow[cluster],
        fill=True,
        fill_color=rainbow[cluster],
        fill_opacity=0.7).add_to(incidents)

# Display map
washington_map
```

Figure 45: Code for Cluster Analysis (5/7)
Display Cluster 0 in P1

c = df_normal[df_normal['Label'] == 0]
c = cluster_0.drop(labels = "Label", axis = 1).T
c = cluster_0

c = cluster_0['Mean'] = cluster_0.mean(numeric_only=True, axis=1)
c = cluster_0

fig, ax = plt.subplots(figsize=(7,7))
c = cluster_0.plot(ax = ax, color = "red", alpha = 0.08)
ax.get_legend().remove()
plt.ylabel('Duration')
plt.xlabel('Hour')
plt.ticklabel_format(style='plain')
plt.ylim(0, 1299999)

latitude = 58.9872
longitude = -77.0569
map_clusters = folium.Map(location=[latitude, longitude],tiles="Stamen Toner", zoom_start=12)

rainbow = ['g017Br', 'g0c7Bm', 'g0e139']
markers_colors = {}

for lat, lon, poi, cluster in zip(df_bystation_unstack['Latitude'],
                                 df_bystation_unstack['Longitude'],
                                 df_bystation_unstack['Label'],
                                 df_bystation_unstack['Label'],
                                 df_bystation_unstack['Label']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker([lat, lon],
                        radius=3,
                        popup=label,
                        color=rainbow[cluster],
                        fill=True,
                        fill_color=rainbow[cluster],
                        fill_opacity=1).add_to(map_clusters)

map_clusters

Display Cluster 1 in P1

c = df_normal[df_normal['Label'] == 1]
c = cluster_1.drop(labels = "Label", axis = 1).T
c = cluster_1

c = cluster_1['Mean'] = cluster_1.mean(numeric_only=True, axis=1)
c = cluster_1

fig, ax = plt.subplots(figsize=(7,7))
c = cluster_1.plot(ax = ax, color = "red", alpha = 0.08)
ax.get_legend().remove()
plt.ylabel('Duration')
plt.xlabel('Hour')
plt.ticklabel_format(style='plain')
plt.ylim(0, 1299999)

Figure 46: Code for Cluster Analysis (6/7)
For the procedure of the P 2, please retrieve the electronic version.

*Figure 47: Code for Cluster Analysis (7/7)*
BIBLIOGRAPHY


120. McLeish, and Todd. *URI Launches Bike Sharing Program on Campus.*

