PATTERNS, OR NO PATTERNS, THAT IS THE QUESTION

Mehrsa Khaleghikarahrodi

University of Rhode Island, mehrsakh@my.uri.edu

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

Recommended Citation
https://digitalcommons.uri.edu/theses/1900

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@etal.uri.edu.
PATTERNS, OR NO PATTERNS, THAT IS THE QUESTION

BY

MEHRSA KHALEGHIKARAHRODI

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

UNIVERSITY OF RHODE ISLAND

2020
MASTER OF SCIENCE THESIS

OF

MEHRSA KHALEGHIKARAHRODI

APPROVED:

Thesis Committee:

Major Professor - Gretchen A. Macht

Inside Member - Valerie Maier-Speredelozzi

Outside Member - Noah Daniels

Dean of Graduate School - Brenton DeBoef

UNIVERSITY OF RHODE ISLAND

2020
ABSTRACT

To reduce the threshold of EV adoption, the electric vehicle supply equipment (EVSE) infrastructure needs to fit the EV systems to its users to satisfy their expectations. There needs to be a reflection of user behavior variability in the EVSE infrastructure to improve its functionality. Individual users’ charging actions, over time, construct a charging behavioral pattern that distinguishes the users from one another. This research analyses the EV users’ unique behavioral charging patterns. Therefore, profiles of users’ charging actions are clustered to investigate the unique behavioral patterns. An unsupervised clustering algorithm (i.e., K-means) is implemented through three distance metrics (i.e., Hamming, Wasserstein, and Manhattan) to classify the Rhode Island public charging stations’ frequent users. Frequent users of these stations were established using the Pareto Principle, in which 608 users (20% of users) contributed to 89% of charging events. The results indicated five clusters through the Wasserstein distance metric, which proposed different EV charging behavior patterns. The five achieved clusters represent: 39% of anxious or opportunistic users (1); 27% of users with consistent charging regardless of their EVs state of charge (2); while 21% of users having sporadic charging behavior revealing no pattern (3); and then approximately 5% of users with procrastination tendencies with only charging when practically out of charge (4) and 8% of users who experience that rush toward running out of charge (procrastinators) but exhibit some opportunistic early charging (5). Knowing these users’ unique behaviors helps both public and private stakeholders target EV market growth through user-centric EVSE location placement.
ACKNOWLEDGMENTS

I would like to thank my academic advisor, Dr. Gretchen A. Macht, for all her support and guidance during the completion of this research. She has been always helpful and encouraging. It was a pleasure working with her and the amazing SIS Lab. Their infinite support was the major key to the success of this thesis.

I would like to thank Dr. Noah Daniels. His advanced classes and help during this research, exposed me to Machine Learning and shaped an interest of mine to further pursue in my PhD program.

I would like to thank my committee members, Dr. Valerie Speredelozzi and Dr. Christopher Hunter for being flexible with the timing of this thesis work.

I would like to thank Dominique Engome Tchupo, James Houghton, and Nicholas Bernardo. They have been sharing their knowledge and have been by my side during this challenging time.

I would like to thank Rhode Island Office of Energy Resources and ChargePoint® for collecting and sharing this data.

Most importantly, I would like to thank my family and close friends for their company, understanding, and support during this journey. They have always encouraged and helped me to set my priorities and stick to them until achieved.
PREFACE

This manuscript is prepared for submission to the *Transportation F: Traffic Psychology and Behavior* journal. Formatting requirements, based on the international version of the journal, dictate that British English is required in the keywords, while the main document is in American English. For the submission, the keywords would be as follow:

**Keywords:** electric vehicle, charging behaviour, charging infrastructure, cluster analysis, pattern recognition
TABLE OF CONTENTS

ABSTRACT ........................................................................................................... ii

ACKNOWLEDGMENTS ....................................................................................... iii

PREFACE ............................................................................................................. iv

TABLE OF CONTENTS .......................................................................................... v

LIST OF FIGURES ............................................................................................... vii

LIST OF EQUATIONS .......................................................................................... viii

1. Introduction ..................................................................................................... 1

2. Literature Review ........................................................................................... 3

   2.1. Resource Allocation Models ..................................................................... 3

   2.2. Behavior Recognition Approaches ......................................................... 4

   2.3. Behavioral Patterns Objective ................................................................. 7

3. Method ............................................................................................................ 9

   3.1. Data Description and Processing ............................................................. 9

      3.1.1. Data Preparation ............................................................................. 9

      3.1.2. Frequent Users Selection ............................................................... 11

      3.1.3. Users’ Profiles ............................................................................ 12

   3.2. Distance Metrics ....................................................................................... 13

      3.2.1. Hamming Distance: binary attributes ............................................. 13

      3.2.2. Wasserstein Distance: histogram ................................................. 14

      3.2.3. Minkowski Distance: metric attributes ......................................... 15

3.3. K-means Clustering .................................................................................... 16

4. Results ........................................................................................................... 17

5. Discussion ..................................................................................................... 20
5.1. Charging Behavior.................................................................20
  5.1.1. Anxious or Opportunistic Users .......................................21
  5.1.2. Consistent Users ..........................................................23
  5.1.3. Sporadic Behavior .........................................................25
  5.1.4. Procrastination ..............................................................26
  5.1.5. Combination of Procrastination and Opportunistic ..............28
5.2. Comparison of Distance Metrics ..........................................29
6. Conclusion..................................................................................30
7. References..................................................................................33

Appendix A - Supplementing Materials ........................................45
Appendix B - R Markdown..............................................................49
LIST OF FIGURES

Figure 1: Distance Matrices Scree Plot ................................................................. 19
Figure 2: Distance Matrices Clusters of Five.......................................................... 19
Figure 3: Distance Matrices Heatmaps.................................................................. 20
Figure 4: Distance Matrices Silhouette Values........................................................ 20
Figure 5: Cluster 1 Boxplot and Violin Plot ............................................................ 22
Figure 6: Cluster 2 Boxplot and Violin Plot ............................................................ 24
Figure 7: Cluster 3 Boxplot and Violin Plot ............................................................ 25
Figure 8: Cluster 4 Boxplot and Violin Plot ............................................................ 27
Figure 9: Cluster 5 Boxplot and Violin Plot ............................................................ 28
LIST OF EQUATIONS

Equation 1: Charge Received ................................................................. 12
Equation 2: Hamming Distance ............................................................. 14
Equation 3: Wasserstein Distance ......................................................... 15
Equation 4: Minkowski Distance ............................................................ 16
1. Introduction

Electric Vehicles (EVs) in the transportation sector play a significant role in achieving a future, low-carbon transportation system (Smith, 2010). Hence, many governments promote EVs and encourage the public to shift from conventional internal combustion engines to EVs (Wee et al., 2012). Along with the encouragement and promotion, an electric vehicle supply equipment (EVSE) infrastructure needs to evolve around EV users. An in-depth understanding of users’ needs, expectations, and behaviors are valuable to public and private stakeholders (i.e., energy producers, distributors, and government bodies) who are targeting to grow the EV market by improving the EVSE functionality (Wee et al., 2012).

Knowledge of how users charge their EVs is vital to creating an equitable and user-centric EV charging station infrastructure. Despite the growth in EV sales, the existing EVSE infrastructure is challenging the users. The EVSE users are experiencing challenges due to limited availability, shortage, and spatial imbalance distribution of charging stations (Xu et al., 2018). Associated issues with EVSE networks can be addressed as the absence of comparing charging stations’ growth trends with users’ growth trends periodically and noticing individual differences towards charging usage.

EV resource allocation models focus on locating the charging stations to maximize EV drivers’ use and minimize infrastructure investment cost (Upchurch et al., 2009; Xi et al., 2013; Gong et al., 2019; Qin et al., 2019), with little to no consideration of actual users’ influence. Sun (2016) believes a certain number of charging stations are necessary to facilitate the industrialization of transportation e-mobility and satisfy the EV users.
However, the quantity of charging stations does not guarantee users’ satisfaction nor necessarily reduce thresholds to adoption. Although many simulation models have been utilized to present optimal locations for deploying EVSE infrastructure (Gong et al., 2019; Qin et al., 2019; Topic et al., 2019), none of these models consider human behavior variability in charging usage. This gap is not necessarily due to a lack of desire, but a lack of information, even though it is in a domain where data is omnipresent. Therefore, a study on charging behavior is beneficial to conduct and reflect through structuring the EVSE infrastructure.

The variability in human behavior can identify distinct behavioral patterns (Korda et al., 2015). Understanding behavior over time enables the construction of a profile of individual actions that establish a unique pattern. A behavioral pattern is a trace of people’s performance or preference influenced by the environment (Tang, 2016; Zhang et al., 2013). Accordingly, the charging behavior pattern can be interpreted as the relationship between individual charging performance and the influence of charging stations’ distribution. Collectively understanding how charging infrastructure is being used to recognize EV users’ charging behavior patterns can accurately forecast future designs (Lee et al., 2020). Applying the quantified usage patterns to stations’ allocation settings can significantly improve the future placement of stations. Additionally, it can reduce EVSE installation costs by making informed, strategic, and economic decisions on stations’ locations. Thus, understanding usage patterns can support sustainable, user-centric charging stations’ distribution trends.

This study investigates the behavioral patterns of the collected EV charging events data with a primary focus on the users’ charging patterns. Unsupervised clustering will
explore the individual users’ charging behavior patterns for their existence within a community-level system as their charge is received, given multiple charging events throughout time in that system. The charging behavior exploration will be able to answer the following research questions: (1) Do unique charging patterns’ groups exist? (2) If so, how many unique charging patterns’ groups exist? (3) What do these groups tell us about the current users’ charging approach?

2. Literature Review

2.1. Resource Allocation Models

An insufficient number of charging stations significantly limits the development of the EV industry (Gong et al., 2019). Qin et al. (2019) stated that the relationship between the number of charging stations and their locations limits EVs’ development. Developing a location model was critical to establishing enough charging stations to maximize the utilization and to minimize infrastructure costs. The appropriate charging station locations are deployed through a parallel comparison of EV public charging stations distribution and EV drivers distribution. Additionally, the potential sites for charging stations have been detected in remote areas with less demand (Gong et al., 2019) to support the EV users with the higher driven miles on their trips. Before proper location selection, the capacitated flow refueling location model helped define a capacity or a limited number of users, each charging facility can provide service (Upchurch et al., 2009). Xi et al. (2013) designed a model to explicitly maximize the charging stations’ service level and increase the acceptance rate of stations for EV drivers arriving and unwilling to wait. To optimize infrastructure deployment with cost-effective solutions, the resource allocation models
utilized different charging stations with varying construction costs per need per locale (Qing et al. 2019). These models acknowledged the problem in the EVSE infrastructure and attempted to propose a valid solution to resolve charging stations’ shortage and limited availability.

Although sensitivity analyses validated these described models, none of them accounted for the impact of infrastructure on users’ actual charging usage, corresponding behavior, and generated charging behavioral patterns. Simulation-optimization models are quantity- and cost-oriented; they all lead to answering the question, “How many EVs are served?”, whereas barely any answer, “How well are EVs being served?”. In other words, what good is a charging station if people will not use it, and what good is an EV if there is no charging station where someone needs it? There needs to be a reflection of human behavior into EVSE infrastructure for improving organizational efficiency and effectiveness while responding to users’ expectations and needs (Kim et al., 2016; Kowalsky, 2017).

2.2. Behavior Recognition Approaches

Understanding users’ driving and charging behavior, in various platforms, is critical for the rollout of EVSE infrastructure. The exploration of these behaviors is attempted on fleet EVs due to their accessibility to conduct test drives and track them (Speidel & Bräunl, 2014; Weldon et al., 2016). Speidel and Bräunl (2014) collected EV data on 11 fleet EVs at 23 charging stations, over two years, to understand their driving and charging behavior's impact on the EVSE infrastructure. Although the assessment metrics to conduct this study were valid, the number of drives was limited, and thus, not considered representative to
develop a comprehensive resource allocation strategy. Weldon et al. (2016) explored this idea by significantly increasing the sample size through analyzing 5,838 fleet EV charging events to obtain valuable information on trip events characteristics, driver preferences, timing, and energy consumption patterns of charging events. Weldon et al. (2016) indicate that generally, there is an early morning peak in charging and that users like to charge more frequently and at shorter distances in their trip. Both studies are based on fleet EVs which operate differently from privately-owned EVs (Voss, 2018). Hence, their derived behavior can only account for the effects of EV integration into the fleet vehicles, and might not accurately reflect the entire impact of EVSE infrastructure distribution.

Moreover, conclusions on driving and charging behaviors have been derived from collected plug-in hybrid and battery electric vehicles’ GPS data (Hardman et al., 2018). Analyzing the GPS data and, more specifically, the users’ driving patterns, highlighted their preference to charge at home, workplace, and public charging stations, respectively (Björnsson & Karlsson, 2015; Jakobsson et al., 2016). A German case study devotes 84% of its charging events to private or public charging stations at home or residential areas (Franke & Krems, 2013b). The users demand to charge closer to their origin due to their limited battery sizes. GPS data mainly addresses the importance of battery sizes in changing the usage patterns. It indicates that extended battery sizes need to be promoted because they profoundly affect the charging patterns (Björnsson & Karlsson, 2015; Kullingsjö et al., 2013). These discoveries give direction to the potential locations for charging stations to improve the EVSE infrastructure. Still, these discoveries are not merely dependent on EVSE distribution, and variation in battery sizes play a significant role in users’ usage patterns.
In addition to the previous studies, the differences in driving and charging behaviors have been captured through conducting surveys (Kelly et al., 2012). Alshahrani et al. (2019) administered EV surveys for explaining the users’ charging preference in Singapore. The results implied that the users prefer to avoid the morning peak times in charging and instead charge during off-peak times, which expresses dissimilarity between survey outcome and the mentioned result from fleet EVs. However, Wong (2018) suggests Singapore as a relatively underdeveloped EV market. Thus, this study’s findings might not be reliable enough to make assumptions about users’ charging behavior to further consider the EVSE infrastructure improvement process. On the other hand, China’s EV market growth is ahead of all other countries (IEA, 2019). As a result, Lee et al. (2020) administered a seven-day survey in China to obtain generalized behavior on 7,979 plug-in electric vehicles based on their travel patterns, preferences, and infrastructure access. This study indicates a midnight home charging trend due to users facing daily demand challenges accessing the nearest available charging station. Contrary to previous studies, there are four peaks which best describe users’ charging patterns: early morning (7 to 8 a.m.), late morning (10 to 11 a.m.), afternoon (1 to 2 p.m.), and evening (5 to 6 p.m.) (Hardman et al., 2018; Schäuble et al. 2017). Although it can be expected that users’ intentions and actions can differ; they can perceive their charging, driving, and traveling differently from how they truly act due to investigated deficiencies in EVSE infrastructure, limited battery size, and driving range (Huang & Liaw, 2005; Tal et al., 2014).

Several researchers have investigated EVSE data demand by station port type (Jochem et al., 2015), consumed energy (Fan et al., 2015; Speidel & Bräunl, 2014), charging time (Ensslen et al., 2017), idle time (Hardman et al. 2018; Lee et al., 2020), start
time (Yang et al. 2016), state of charge (SOC) before and after charging (Speidel & Bräunl, 2014), or charging location (Neaimeh et al., 2015). There has been no actual study on EVSE data investigating the users charging behavior and how they manage their mobility needs and resources. Such studies need to be administered to support the efficient utilization of charging stations. For this purpose, Jackobsson et al. (2016) designed a three-month EV trial to observe the participants’ charging patterns in a closer view. The outcomes show low changes in daily driving distance and convenient access of participants to charging stations when needed. Franke and Krems (2013b) likewise leased 40 EVs for a sample of 79 participants to conduct a field study on their charging for six months. Participants experienced convenient charging and relatively convenient interface with the charging stations and charging at a similar SOC throughout the study. The limited number of participants and even studies’ duration lead to stable charging experiences for users. For reflecting the embedded pros and cons in EVSE infrastructure, a longstanding study including a significant number of EV users will comprehensively present a realistic interaction of EV drivers with the existing infrastructure and their charging experiences.

2.3. Behavioral Patterns Objective

The individual differences in charging, driving, or traveling can relate to psychological variables and how they affect each user’s behavioral pattern (Franke et al., 2012; Franke & Krems, 2013a, 2013b). Therefore, it is useful to understand the variables underlying the differences in users’ utilization of EVSE infrastructure. Their charging patterns need to be explored and classified into common behaviors to identify the users’ motivations and influential factors in charging behavior (Philipsen et al., 2018). Philipsen et al. (2018) believed that differentiation in behavioral patterns could identify distinct types
of users in terms of charging behavior. The study’s research tool was a questionnaire based on their previous interviews and research on the prompt. With the clustering analysis, which included the frequencies and quantities of the recharging events, this study identified seven types of users among 279 respondents. Admitting the differences between groups of users can help develop a sustainable distribution of charging stations, which other studies are missing. Additionally, this study had the metrics independent from influential factors on charging behavior, such as time of the day, battery size, and driving range. Most importantly, it targeted studying the users’ actual usage of EVSE infrastructure (Philipsen et al., 2018). Surveys have made valuable progress in mirroring charging behaviors, but individuals’ actions versus intentions are a gap to discuss further. Instead, this paper attempts to expand the similar work, independent from any external features, and based on the EVSE data to run statistical analysis and quantitatively approach the charging behavior study.

The integration of the users’ charging behaviors into the EVSE infrastructure planning and modeling is crucial because the users and their decisions regarding the charging, the refilling quantities, and frequencies play a decisive role in the positioning and dimensioning of charging infrastructure (Philipsen et al., 2018). The information accumulated through this review reveals the gap that needs to be filled in this research field. GPS data due to its dependence on battery sizes and surveys due to their inconsistency of actions and intentions will lead this research to be conducted on EVSE data, including charging events. Since short-term data is not representative of obstacles EV users face daily, a long time collected EVSE data with a significant number of frequent users is needed. The differences in EV users’ charging behavior will be explored, and users with
similar charging behavioral patterns will be analyzed. This research aims to assist the EVSE infrastructure to appropriately locate the charging stations by realizing distinct groups of users in charging behavior. Selecting the proper charging stations location not only alleviates the users’ charging related concerns, but it reduces the negative impact on the grid (Azadfar et al., 2015; Schäuble et al., 2017).

3. Method

3.1. Data Description and Processing

The statistical analysis of this data is the core of this research. To conduct a user behavior study, have a closer look on individual users, apply various tests to derive conclusions, and have a graphical representation of findings, the coding for this research is executed with the statistical program R within R-Markdown (Appendix B). Also, Visual Basic for Applications (VBA) in Microsoft Excel for time adjustments of charging events to Coordinated Universal Time (UTC) is partially used.

3.1.1. Data Preparation

This study analyzed public charging stations in the State of Rhode Island (RI) and Providence Plantations as a representative overview of EV users’ actual usage from publicly available charging stations. The data was provided by the Rhode Island Office of Energy Resources with access to 68 Level-2 ChargePoint® public charging stations across the state. Providing public charging stations are critical for EV market growth (He et al., 2013). EV charging stations have three types of Level-1, Level-2, and DC Fast Chargers (Falvo et al., 2014). Level-1 chargers are typically dedicated to at home charges which
charges for three to five miles in an hour and are counted as public charging stations; along with DC Fast it claims to provide 80% charge in 20-30 minutes (Brodd, 2017). The Level-1 chargers directly affect the usage of public charging stations. The users’ charging behavior could be better investigated if home charger’s data was available (Voss, 2018); however, there is limited access to this type of data. This research focuses on the provided public data, Level-2 charging stations, which are free from charge in the State of RI after they initiated the Driving Rhode Island to Vehicle Electrification (DRIVE) to facilitate the adoption of EVs by local consumers. During this program, the state residents who were interested in purchasing or leasing an EV were able to apply. Additionally, RI assigned the charging stations free from cost to promote and encourage the public to enable filling the gap for the state clean energy portfolio (RI OER, 2017). Recently, the charging stations management are changing with time thresholds and costs added to a proportion of stations. Since RI is second most densely populated state, before New Jersey, the provided data was worthy to further understand the usage of EV users from the public charging stations.

The data represents over six years of charging (2013 Q1 to 2019 Q3) with 65,041 charging events. The data is comprised of information related to stations (e.g., station name, address, geographical dimensions, model), users (e.g., identification number, postal code), and events themselves (e.g., locale date and time, plugged-in duration, charging time, energy received, greenhouse gasses and gasoline savings, termination reason). There are 16 possible reasons for a charging event to end, which can either occur by the station due to technical issues/inquiries (e.g., station deactivated) or by the user due to preference (e.g., plug out at vehicle) or a fully charged EV (e.g., timeout) (Table A1). In order to process and fulfill a charging behavior study, the data was narrowed down into absolute
users-related charging events that were ended by the users themselves. This approach left 64,890 charging events to find frequent users and further investigate their charging actions throughout the data and time.

3.1.2. Frequent Users Selection

Exploring the accessed dataset showed that not all users were taking advantage of RI EVSE infrastructure in a continuous or periodic order. There exist users who used the charging stations only a few times; the reasons are unknown but can range from traveling purposes to at-home charging. Assumptions are that each user identification number, belongs to a single user and it is not shared with multiple users nor multiple vehicles within the timeframe. Therefore, a subset of frequent users’ charging events is defined in order to proceed with the relevant illustration of charging behavior affected by the existing EVSE infrastructure. Next, the Pareto Principle is implemented to identify the frequent users through the data with higher counts in charging events that have been in the system earlier than the others.

The Pareto Principle explains that 80% of consequences come from 20% of the causes (Kim et al., 2017; Brynjolfsson et al., 2011). The extension of this principle has been observed in numerous datasets, such as marketing (Kim et al., 2011) and library studies (Zhu & Xiang, 2016). The application of the Pareto Principle indicates that 89% of charging events are caused by 20% of the users. Since this principle applies to the users and their charging events, a subset of frequent users (i.e., 20% of users) is generated. In conclusion, 608 users are selected for charging behavior exploration and 58,273 charging events are studied deliberately. The results indicated the Pareto Principle is followed within
the data; thus, a comprehensive empirical study is conducted on the subset of achieved frequent users.

3.1.3. Users’ Profiles

Tracking the frequent users’ charging actions over time enabled the construction of users’ profiles. The users’ profiles are structured based on the amount of energy received or added at the stations. According to the data, a higher amount of energy (kWh) was received when the users stayed at the charging stations for an extended period, and their SOC was closer to zero. The following equation (Equation 1) was applied to compute the proportion of charge received by each user during their charging events:

\[
\text{Charge Received} = \left( \frac{\text{Energy Received at Charging Event}}{\text{Individual's Maximum Received Energy}} \right) \times 100 \quad (1)
\]

As a result, the CR ranging from 0 to 100% was obtained to create the users’ profiles. A histogram was made for each user, consisting of 10 continuous and equally-sized bins starting from 0% and ending at 100%. For example, the first bin contains CR from [0%, 10%), including the lower bound of 0%, while being less than 10% for the next upper bound, repeating until (90%, 100%]. Each bin represents how often users received charge within that specified range. Users’ collected frequencies were divided by their total number of charging events to create standardized and scaled users’ profiles in order to compare the different users. Through this method, a profile of individual users \((n = 608)\) describing their recharging quantities and frequencies was created to simplify user comparisons.
3.2. Distance Metrics

A distance metric was necessary to find the similarity or regulatory between users’ profiles and group them with similar charging behavioral patterns (Singh et al., 2013). In essence, the distance metric was the main idea behind the clustering (Akalin, 2020). To develop robust clustering, the distance metrics are used to find the similarity and dissimilarity within data objects (Singh et al., 2013). Before applying any distances, users’ profiles were normalized to avoid the unequal scale of profiles and compare users to one another accurately. Using various distance metrics and selecting the metric that appropriately explains similarity and dissimilarity within the data, is crucial in pattern recognition (Zhai et al., 2018). To construct distinct charging behavior groups, the k-means clustering was deployed with Hamming, Wasserstein, and Minkowski distances to measure the similarity between users’ scaled profiles.

3.2.1. Hamming Distance: binary attributes

The Hamming distance is a natural similarity measure between binary codes, defined as the number of bits that differ in the given sets (Zhai et al., 2018). The Hamming distance provides an appropriate measure of discrimination by achieving a minimum number of errors while converting strings/vectors to one another (Taheri et al., 2020). Pandit & Gupta (2011) applied this distance metric to two equal strings in length. The distance between the given strings is the number of positions that corresponding symbols are different. This distance metric is referred to as the number of bits that any given set of strings need to be changed to turn one string in the set into the other (Wang et al., 2015). Taheri et al. (2020) applied this distance metric between the same length vectors to estimate
the difference between injected elements in the vectors. Considering each user’s profile as a vector with ten inserted numbers \((n)\), this distance metric (Equation 2) is applicable for every profile of users \((a \text{ and } b)\) to achieve the distance matrix for all frequent users.

\[
\text{Dist} (a, b) = \sum_{k=0}^{n} \frac{|a(x) - b(x)|}{n(n-1)}
\]  

(2)

3.2.2. Wasserstein Distance: histogram

The Wasserstein distance metric is introduced to measure the dissimilarity between histograms diversity. This distance was initially described as the earth mover’s distance, with the given problem of moving a pile of dirt with a minimum amount of work required into a whole with the same volume. Rubner et al. (2000) introduced the earth mover’s distance probability distribution dissimilarity when histograms are normalized to one another. Thus, the Wasserstein distance is known as earth mover’s distance with a similar function in distance recognition of histograms (Chan et al., 2007). For clustering the histograms the Wasserstein distance is the best-proposed metric (Rubner et al., 2000; Irpino et al., 2014); which allows expressing the variability of a set of histograms in their means (variability of average) and their dispersion (shape and size) (Irpino et al., 2014). Pele & Werman (2009) computed this distance on a large histogram database considering the minimum cost to be paid to move one histogram to another. The Wasserstein distance can be performed between two equalized \(n\)-bins histograms, either high or low dimensional (Shirdhonkar & Jacobs, 2008). The profile of users in this study was practically obtained from their histogram of charge received with 10-bins histogram. The nature of this distance and the formation of users’ profiles were synced and coming from the same approach, which ideally, should work best for this study. It is expected to perform the Wasserstein
distance among users’ profiles which are univariate since they are only contain and depend on the frequencies of charge received for an individual (Panaretos & Zemel, 2019; Schuhmacher, 2020). For this purpose, the Wasserstein distance is applied between every two-standardized-univariate profile of users. Considering $a$ and $b$ two given vectors to compare through this distance, the following equation (Equation 3) is applied in order to obtain a distance matrix for all 608 users with their similarity or dissimilarity values to further execute the k-means clustering analysis.

$$W_1(a, b) = \int_{-\infty}^{\infty} |a(x) - b(x)| dx$$  \hspace{1cm} (3)

3.2.3. Minkowski Distance: metric attributes

The Minkowski distance is a generalized distance metric for both ordinal and quantitative variables (Singh et al., 2013). The distance between any two given vectors is an $n$-dimensional space, is well performed through Minkowski distance (Shirkhorshidi et al. 2015). This distance is commonly used in fuzzy clustering with partitioning clusters for overlapping purposes to understand further of differences within the data (Groenen et al., 2007; Srivastava et al., 2012). This metric is a distance measure for numeric attributes (Saxena et al., 2017). Wilson & Martinez et al. (1997) modified this distance for continuous features and calculated the distance between counts for nominal attributes. Shirkhorshidi et al. (2015) suggested this distance for having an isolated and compacted set of clusters. Applying any clustering algorithms is favorable in order to achieve isolated groups; hence, this study applied this distance metric to users’ charging behavior profiles to assess the users and related charging approaches with similar behavioral patterns. To obtain a similarity/ dissimilarity distance matrix which represents clear variation among the users
in shape, size, and density, this study users $p = 1$ which turns the Minkowski distance into city block distance or Manhattan distance (Singh et al. 2013). With the selected value of $p$ for this equation (Equation 4), the Manhattan distance will compute the similarity and dissimilarity of users to one another.

$$\text{Dist}(a, b) = \left( \sum_{1}^{n} |a(x) - b(x)|^{\frac{1}{p}} \right)^{p} \quad (4)$$

3.3. K-means Clustering

The clustering algorithms are utilized as an explanatory tool to integrate the information contained in a large number of data (Wang et al., 2015). Clustering is a ubiquitous procedure in any field that deals with high-dimensional data to categorize the data into groups (Singh et al., 2013; Sharma, 2020). K-means is an unsupervised clustering algorithm which is a heuristic technique of partitioning the set of an object into $k$ clusters considering the sum of squared distance in each cluster is minimized, and the objects in each cluster are similar to each other (Irpinio et al., 2011; Pandit & Gupta, 2011; Singh et al., 2013). This unsupervised clustering approach is used to assess how related users are to each other; if related, the users with similar user behavioral charging patterns are combined into groups and further investigated to understand the type of behavior they are following.

The number of clusters for each distance metric was selected through the scree plot and elbow technique (Zhan et al., 2016). The most appropriate number of clusters was shown on the scree plot as an elbow shape with a minimum difference within the sum of squares after the momentum drop in the plot (Sahmer et al., 2006). Based on the application domain, the distances were compared, and the relevant distance with a better expression of
users was selected. The Silhouette validation score aids in the interpretation and validation of the relevant distance matrices with the higher cluster engagement, given the profile of the users. The Silhouette value is favorable to be closer to one since it states a higher chance of connection between the selected distance and the selected number of clusters (Rousseeuw, 1987). Additionally, a heatmap for distance matrices was prepared to visualize the distances for thoughtful distance decision making. Once the decision upon the proper distance was made, a box plot and violin plot of each cluster was depicted. To further investigate each cluster’s differences, a Mann-Whitney Wilcoxon (MWW) test with a 95% confidence level was performed.

4. Results

The distance matrices of users’ profiles are obtained with dimensions of 608 by 608, in which the values within the matrices represent the similarity or dissimilarity of users. The platform for the application of clustering analysis was prepared. The distance metrics were expected to exhibit the minimum distance between objects within a cluster and maximum distance between objects within different clusters (Pandit & Gupta, 2011). Accordingly, the selected number of clusters is achieved by the scree plot and the elbow technique (Zhang et al., 2016). The Hamming distance (Figure 1a), the Wasserstein distance (Figure 1b), and Manhattan distances (Figure 1c) all indicated the five clusters based on their scree plot.

Pandit & Gupta (2011) mentioned that the distance matrix is expected to exhibit minimum distance between the objects within each cluster and maximum distance between objects of different clusters. Comparing the clusters of five for all the desired distance
matrices showed that the Wasserstein (Figure 2b) distance was following the same criteria, among others (Figure 2a & c), and this distance best represented the users’ profiles with common behaviors.

Moreover, the distance matrices’ heatmaps were prepared to visualize their differences and select the well-justified distance metric. As depicted in Figure 3, Wasserstein (Figure 3b) showed an insightful existence of clusters by the different box-like patterns; whereas, Hamming (Figure 3a) showed that all of the users highly correlated with large sweeping singular colors, and similarly Manhattan (Figure 3c) showed less clustering compared to Wasserstein.

Additionally, to validate how well each distance matrices is representing the users within each cluster, the silhouette clustering validation is performed. This value ranges from -1 (i.e., the objects in the clusters are not well-defined) to 1 (i.e., the objects in the cluster are well defined). The comparison of silhouette values for all three distances and clusters of five, validated the Wasserstein as the most engaged distance matrix within the given set of users’ profiles since this distance was representing a silhouette value closer to 1; Hamming, Wasserstein, and Manhattan’s silhouette values are 0.18, 0.36, and 0.28, respectively. Figure 4 shows for each distance, within each cluster there are some objects that might not fit exactly in the cluster and they are leaning towards -1. However, the objects in Wasserstein distance (Figure 4b) are closer to the zero when comparing the three distances; whereas other distances exhibit higher tendency to -1. In conclusion, Wasserstein has better demonstration for the objects that are not fitting the assigned cluster as depicted in Figure 4b.
After carrying out the clustering analysis by Wasserstein distance, results showed that users were partitioned into five clusters which represent: 39% of anxious or opportunistic users (1); 27% of users with consistent charging regardless of their EVs SOC (2); while 21% of users having sporadic charging behavior revealing no real pattern (3); and then approximately 5% of users with procrastination tendencies are only charging when practically out of charge (4) and 8% of users who experience that rush toward running out of charge (procrastinators) but exhibit some early opportunity in charging (5).

Figure 1
Distance Matrices Scree Plot

Figure 2
Distance Matrices Clusters of Five
5. Discussion

5.1. Charging Behavior

Visualization of each cluster charging performance assisted in assigning a proper identification and label to each group based on their charging behavior. The different types of charging behaviors observed from these clusters are anxious or opportunity, consistent,
sporadic, procrastination, and procrastination with opportunistic tendencies. Each of these will be explained in detail in the following sections.

5.1.1. Anxious or Opportunistic Users

The first cluster observed in this study showed that almost 39% of frequent users (240 out of 608) tend to charge their battery by 20% to 50% during their charging events (Figure 4). The MWW test was conducted to assess users’ tendencies to receive charge each of the different ranges. The results indicate that users were as likely to receive a charge in the last four ranges (60-70%, 70-80%, 80-90%, and 90-100%) as in the first range (0% to 10%), and as likely to receive a charge in the second range as the sixth range (p-value > 0.05). More importantly, the test results showed that users are equally likely to receive a charge within the three densely populated ranges (20-30%, 30-40%, and 40-50%). Hence, the conclusion from the test and plotted graphs (Figure 4) indicated the CR range from 20% to 50% was significantly used compared to the other ranges. This high frequency leads to two assumptions: (i) That this CR range is the threshold to alleviate any anxiety caused by a low battery, and (ii) that users’ time availability and geographic location allowed them to charge more frequently within those ranges.
Some users managed their charging occurrences to mitigate the influence of unexpected events. As such, these users attempted to receive a certain level of charge from having enough coverage in case of uncertainties. Alsabbagh et al. (2020) described the concept of range anxiety based on the users who prioritize their CR and experience some degrees of stress caused by the limited amount of range available (Franke & Krems, 2013b; Neubauer & Wood, 2014). Indeed, from the combination of the MWW test and plots for clusters, one could assume that the users were managing to receive a 20% to 50% charge more frequently to avoid experiencing anxiety caused by the low battery.

Another possibility is that users were frequently receiving 20% to 50% of charge because they had regular access to the EVSE infrastructure. In this opportunistic charging scenario, the users continuously received charge and charged longer within a given event (Zoep et al., 2013). The access or the opportunity was most likely given to them at their work and/or home, where they regularly spent an extended amount of time (Rangaraju et al., 2015). The Philipsen et al. (2018) study found that a distinct group of EV users exhibit
opportunistic behavior when considering their battery and wanting to immediately compensate for their travel at an instant charging event. Therefore, this work is unclear as to how it directly supports Philipsen et al.’s (2018) claim but it does start to put a quantitative proportion of the population who might actually behave in that way. Thus, users in this cluster could either have an anxious behavior towards their charging with managing their mobility needs and resources or they could have an opportunistic behavior to take advantage of accessible charging stations around their work and/or home area. However, further investigation is needed to detect the elements defining this charging pattern.

5.1.2. Consistent Users

27% of the frequent users (164 out of 608) at different SOCs are arriving at the charging stations and receiving equal amounts of charge frequently throughout their recorded charging events. They appeared to have a consistent utilization of charging stations, which potentially establishes a consistent charging behavior. To ascertain the consistency in this classification of users, the MWW test is performed. Except for the 0% to 10% and 80% to 90% CR ranges, the rest were found to have an equal population (p-value > 0.05). The test results and prepared graphs (Figure 5) for this cluster implied that the users are equally likely to receive charge within any of the CR ranges except the two aforementioned ones (0-10% and 80-90%). The adverse impact of limited electric battery and driven range required the users to be consistent with their charging decisions to be able to complete their daily commute). This consistent behavior swapped the charging events mostly at workplace charging stations, residential areas, and public places respectively (Chakraborty et al., 2019). The users have limited amount of time and limited availability to
charging stations and so they are tending to frequently charge at different charge received with a consistency in action and as a result in behavior (Franke & Krems, 2013b). This increases the user’s ability to meet their daily miles traveled and complete their trips confidently. This cluster provides evidence that these users represent less than 30% of the frequent users in this study. Consistent utilization of charging facilities minimizes all EVs’ aggregated waiting time that need to be charged (Mousavi & Flynn, 2016; Gusrialdi et al., 2017). Gusrialdi et al. (2017) believed the consistent utilization of the service stations, results in reducing traffic congestion and bottlenecks at charging stations. Furthermore, investigating the causes of this behavior can assist with developing strategies for service stations to be used consistently and, consequently, enhance the EVSE infrastructure deployment.

**Figure 6**
*Cluster 5 Boxplot and Violin Plot*
5.1.3. Sporadic Behavior

21% of frequent users (127 out of 608) did not establish a pattern in charge received. The MWW test did not detect a certain order on how ranges related to the population. However, Figure 6 might provide an explanation for these results. It can be assumed that the lack of clear pattern originates from the users’ sporadic behavior. Ferreira et al. (2011) did a study on human battery interface with smartphones during which a proportion of users demonstrated erratic behavior. Their charging was taking place at any time with no consideration of their battery percentage and with random interruptions during the charging process. EV studies also showed that users charge throughout the day, at any available charging station, with random charge received frequencies (Hardman et al., 2018). This approach of using the charging stations, results in sporadic charging patterns which in the long run affects the lifespan of EVs’ batteries (Ferreira et al., 2011). It can also be a sign that these users are not planning their charging events and make sudden decisions to charge at downtown, commercial, and open-space districts (Voss, 2018). This cluster behavior provides the opportunity for the EVSE infrastructure to implement a timer and a charge received threshold, when utilizing smart charging stations, in order to make sure that all the users at least have the possibility to take advantage of charging stations equally.
5.1.4. Procrastination

Five-percent of the frequent users (29 out of 608) showed a distinct behavior when they only had limited driving range left; thus, their SOC was closer to zero upon arrival. According to the MWW test, the first seven CR ranges (0% to 70%) have an equal population (p-value > 0.05) with similar density, as shown in Figure 7. The last three CR ranges (70 to 100%) are expanding in density; although, the last CR range showed the highest density and frequency in CR. This shows that a limited number of charging events took place when the EVs were considered to have enough battery, and most of the charging was performed when their battery was nearing empty. These users are likely to procrastinate their charging events and accept the risk of performing charging instantly when it was absolutely necessary to recharge. The Philipsen et al. (2018) study found that a distinct group of EV users drive their car until the battery is as empty as possible and refill for a full battery at charging event. Kim et al. (2018) described individuals who
discount taking actions immediately and rather rely on the future as the procrastinators (Konig & Kleinmann, 2005). This cluster of users similarly appears to discount the immediate availability of charging stations and prefer the rush to charge in the far future and when it is imperative. This work directly supports Philipsen et al.’s (2018) claim that there are these types of users but also identifies a quantitative proportion of the frequent users who actually behave in that way.

The procrastination behavior can be a predictor of EVSE infrastructure reflection on EV users’ performance. The procrastination can be viewed as the result of availability of opportunities and resources, availability of information, and skills and abilities (Brito & Laan, 2010). The established causes of procrastination are either having the opportunity and access to charging stations at different SOCs, expect lower range of SOC or zero; or, having enough experience with the existing infrastructure to avoid the anxiety associated with the limited availability of charging stations (Steel et al. 2001). Although, very little proportion of users, 5% of frequent users are experiencing as such, but maybe that is because they charge conveniently through their selected charging stations which can be close to their home or workplace origin. On the other hand, Nguyen et al. (2013) mentioned income and gender, can be detective variables in procrastination behavior. This may indicate that the users’ motivation can be other variables that EVSE availability and their SOC. Uncovering the layers of this behavior will expose more information about the infrastructure target users and it directly helps the infrastructure with making an informed strategic decisions regarding the charging stations deployment.
5.1.5. Combination of Procrastination and Opportunistic

The last cluster consists of eight-percent of the frequent users (48 out of 608) who tend to procrastinate charging with some charge received as shown earlier on their pattern (Figure 8). The MWW test showed the last CR range (90-100%) does not have an equal population with the other ranges (p-value > 0.05). The statistical analysis indicated that procrastination is the dominant behavior in this group of users. Despite that, the graphed plots showed significant frequency in two ranges of 10-20% and 20-30% respectively. This is similar to cluster 1: Anxious or Opportunistic behavior. Merging the idea from Procrastination and what was observed in this cluster’s pattern, it can be assumed that there was an opportunity for the users to receive charge at those ranges (10% to 30%) but they preferred (based on their frequency) to charge when empty. They showed some frequencies earlier on the CR but the procrastination mostly shaped this pattern. The described behavior for this proportion of users was associated with a combination of procrastination and opportunistic.
5.2. Comparison of Distance Metrics

Initially, three distance metrics were proposed to conduct a clustering analysis. The Wasserstein Silhouette value was greater than the others (i.e., Hamming and Manhattan) and closer to one. The clustering analysis of this research was mainly on the Wasserstein distance matrix. However, to examine if other distance metrics are providing common patterns, Hamming and Manhattan distance with clusters of five were graphed, the CR ranges of each cluster were evaluated with the MWW test and then compared to patterns of other distances. Hamming and Wasserstein displayed patterns similar to anxiousness or opportunistic behavior (Figure A1), Procrastination behavior (Figure A2), and Consistent behavior (Figure A3). Manhattan and Wasserstein had Anxiousness or Opportunistic behavior (Figure A4) and Procrastination (Figure A5) in common as well. Accuracy of clusters might indicate which best represents the data, but the similarity between different distances in an empirical evaluation is common to reveal the same pattern among all (Gupta & Chandra, 2019). This research study is confident to state the existence
of users with *Anxiousness* or *Opportunistic* behavior, *Procrastination* behavior, and *Consistent* behavior since distance metrics, both graphically and statistically, demonstrated similarly.

**6. Conclusion**

EV users are experiencing challenges in EVSE infrastructure due to a lack of periodic comparison of charging stations’ growth with users’ growth. Many simulation-optimization models are defined to maximize the service rate, the acceptance rate at charging facilities and minimize the infrastructure cost with little to no consideration of users’ needs and expectations. This paper suggested an in-depth understanding of users charging behavior to reflect the users’ differences through EVSE infrastructure deployment. The integration of users’ behavior in EVSE infrastructure assists with reducing the EV adoption. Thus, this paper attempted to investigate unique charging behavioral patterns among Rhode Island public charging stations users. The results indicated that the users formed charging patterns in five distinct groups. Quantitative analysis of the behavioral charging patterns further demonstrated how users in each group approach the EVSE infrastructure for charging purposes. To discover more about the charging behaviors, the underlying reasons for their established charging patterns need to be examined. The analysis of each charging behavior cluster will expose the users’ psychological variables and motivations for the performed charging behavior patterns. A proportion of underlying variables could be dedicated to the population demographics (i.e., age, gender, income, and gender), which can assist with justifying the obtained behaviors versus attitudes. In addition to EVSE data, being exposed to EVs’ related information (i.e.,
vehicle make, model, year, battery capacity, state of charge) would benefit future studies’ accuracy.

It is useful to understand the variables underlying the differences in users’ charging behaviors as they could be integrated into charging management, EVSE infrastructure deployment, increasing the satisfaction of existing EV users, and targeting new users. However, to strengthen these research findings, a mixed-methods study, incorporating interviews and surveys, on the same population of the users can be implemented. The conducted study was an analysis of users’ reflection on Rhode Island electric vehicle supply equipment infrastructure, which includes charging stations free from charge; however, the price can be highly influential in users’ charging decision-making. The expansion of this study to other regions and considering the influential factor (e.g., cost of charging), accumulates a top-down analysis of charging stations while considering potential motivations for derived charging behaviors.

In the future, the Electric Vehicle Infrastructure Projection (EVI-Pro) tool can undoubtedly assist with designing a strategic EVSE infrastructure. This tool is developed through the collaboration of the National Renewable Energy Laboratory and the California Energy Commission. This tool is projecting the demand for EV charging infrastructure by providing detailed data on personal vehicle travel pattern (e.g., GPS data), EV attributes (e.g., plug-in hybrid electric vehicles, battery electric vehicles), and charging station characteristics (e.g., home/work/public, Level-1/Level-2/DCFC). EVI-Pro estimates the quantity and type of charging infrastructure necessary to support the EV adoption in different regions. So far, this tool has been used for infrastructure planning in Massachusetts, Columbus, California, Maryland, and Seattle (U.S. Department of Energy,
2018). Accessing to this tool to make future predictions on a sufficient number of charging stations needed in RI, the second most densely populated state in the U.S., or any other regions, can also assist with the deployment of EVSE infrastructure.
7. References


Transactions on Intelligent Transportation Systems, 18(10), 2713–2727.
https://doi.org/10.1109/tits.2017.2661958


https://doi.org/10.1016/j.chb.2004.02.020


https://doi.org/10.1146/annurev-statistics-030718-104938


https://doi.org/10.1109/iccv.2009.5459199


https://doi.org/10.1088/1755-1315/252/3/032164


http://www.drive.ri.gov/about-drive/.


https://doi.org/10.1016/j.rser.2014.07.177

https://doi.org/10.1007/s12652-012-0161-8

https://doi.org/10.1016/s0191-8869(00)00013-1


Appendix A - Supplementing Materials

Table A1
*Charging Events Termination Reasons*

<table>
<thead>
<tr>
<th>Reason</th>
<th>Ended by</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CPS Serve</td>
<td>Station</td>
</tr>
<tr>
<td>2 Customer</td>
<td>User</td>
</tr>
<tr>
<td>3 Final GFCI Tripped</td>
<td>Station</td>
</tr>
<tr>
<td>4 Ghost_Session</td>
<td>Station</td>
</tr>
<tr>
<td>5 Holter Plugin</td>
<td>User</td>
</tr>
<tr>
<td>6 Outlet Unreachable</td>
<td>Station</td>
</tr>
<tr>
<td>7 Plugout at Vehicle</td>
<td>User</td>
</tr>
<tr>
<td>8 Plug Removed while Rebooting</td>
<td>User</td>
</tr>
<tr>
<td>9 Relay Stuck Closed</td>
<td>Station</td>
</tr>
<tr>
<td>10 Relay Stuck Open</td>
<td>Station</td>
</tr>
<tr>
<td>11 Station Deactivated</td>
<td>Station</td>
</tr>
<tr>
<td>12 Station Offline</td>
<td>Station</td>
</tr>
<tr>
<td>13 Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>14 Timeout</td>
<td>User</td>
</tr>
<tr>
<td>15 Ventilation Fault</td>
<td>Station</td>
</tr>
<tr>
<td>16 Blanks</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
### Table A2

**Mann-Whitney Wilcoxon Test Results for Wasserstein Cluster 1**

<table>
<thead>
<tr>
<th></th>
<th>0-10% (B1)</th>
<th>10-20% (B2)</th>
<th>20-30% (B3)</th>
<th>30-40% (B4)</th>
<th>40-50% (B5)</th>
<th>50-60% (B6)</th>
<th>60-70% (B7)</th>
<th>70-80% (B8)</th>
<th>80-90% (B9)</th>
<th>90-100% (B10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10% (B1)</td>
<td>0.131</td>
<td>0.638</td>
<td>0.071</td>
<td>0.078</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-20% (B2)</td>
<td></td>
<td>0.906</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.085</td>
</tr>
<tr>
<td>20-30% (B3)</td>
<td>0.574</td>
<td>0.780</td>
<td>0.312</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-40% (B4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-50% (B5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-60% (B6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.937</td>
</tr>
<tr>
<td>60-70% (B7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.177</td>
</tr>
<tr>
<td>70-80% (B8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-90% (B9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90-100% (B10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance Code: p-value < 0.05 (‘*’)

### Figure A1

*Hamming Cluster 1 Boxplot and Violin Plot*
Figure A2
*Hamming Cluster 5 Boxplot and Violin Plot*

Figure A3
*Hamming Cluster 4 Boxplot and Violin Plot*
Figure A4
*Manhattan Cluster 4 Boxplot and Violin Plot*

![Boxplot and Violin Plot for Cluster 4](image1)

Figure A5
*Manhattan Cluster 4 Boxplot and Violin Plot*

![Boxplot and Violin Plot for Cluster 1](image2)
Appendix B - R Markdown

Patterns or no Patterns
Mehrsa Khaleghikarahrodi
7/31/2020

Upload Packages

The following packages are downloaded for running the analysis on the data.

```r
knitr::opts_chunk$set(echo = TRUE)
library(plyr)
library(dplyr)

## Attaching package: 'dplyr'
## The following objects are masked from 'package:plyr':
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

## The following objects are masked from 'package:stats':
##   filter, lag

## The following objects are masked from 'package:base':
##   intersect, setdiff, setequal, union

library(maptools)

## Loading required package: sp

## Checking rgeos availability: TRUE

library(tidyverse)

## Attaching packages -------------------------------------- tidyverse 1.3.0 ------

## v ggplot2 3.3.0   v purrr  0.3.4
## v tibble  3.0.1   v stringr 1.4.0
## v tidyr  1.1.0   v forcats 0.5.0
## v readr   1.3.1

## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()  masks plyr::compact()
## x dplyr::count()    masks plyr::count()
## x dplyr::failwith() masks plyr::failwith()
```
## x dplyr::filter() masks stats::filter()
## x dplyr::id() masks plyr::id()
## x dplyr::lag() masks stats::lag()
## x dplyr::mutate() masks plyr::mutate()
## x dplyr::rename() masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()

library(maps)

## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
## map

## The following object is masked from 'package:plyr':
##
## ozone

library(viridis)

## Loading required package: viridisLite

library(ggthemes)
library(readr)
library(mapdata)
library(gridExtra)

## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
## combine

library(grid)
library(cumplyr)
library(RColorBrewer)
library(ggpubr)

## Attaching package: 'ggpubr'

## The following object is masked from 'package:plyr':
##
## mutate

library(stats)
library(forecast)
## Registered S3 method overwritten by 'quantmod':
##   method    from
##   as.zoo.data.frame zoo

## Attaching package: 'forecast'

## The following object is masked from 'package:ggpubr':
##
##   gghistogram

library(colorspace)
library(tseries)
library(survival)
library(coin)
library(astsa)

## Attaching package: 'astsa'

## The following object is masked from 'package:forecast':
##
##   gas

## The following object is masked from 'package:maps':
##
##   unemp

library(tidyr)
library(ggplot2)
library(maps)
library(tibble)
library(lubridate)

## Attaching package: 'lubridate'

## The following objects are masked from 'package:dplyr':
##
##   intersect, setdiff, union

## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union

library(animation)

## Warning: package 'animation' was built under R version 4.0.2

library(gganimate)
library(ggpubr)
library(skbmr)
library(lattice)
library(caret)

## Warning: package 'caret' was built under R version 4.0.2

## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##   cluster

## The following object is masked from 'package:purrr':
##   lift

library(e1071)

## Warning: package 'e1071' was built under R version 4.0.2

library(emdist)
library(cluster)

## Attaching package: 'cluster'

## The following object is masked from 'package:maps':
##   votes.repub

library(tidyLPA)

## You can use the function citation('tidyLPA') to create a citation for the use of {tidyLPA}. Mplus is not installed. Use only package = 'mclust' when calling estimate_profiles().

library(poLCA)

## Warning: package 'poLCA' was built under R version 4.0.2

## Loading required package: scatterplot3d

## Loading required package: MASS

## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##   select

library(HistDAWass)

## Warning: package 'HistDAWass' was built under R version 4.0.2

library(transport)
## Warning: package 'transport' was built under R version 4.0.2

```r
library(data.table)
```

## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year

## The following object is masked from 'package:purrr':
##
##     transpose

## The following objects are masked from 'package:plyr':
##
##     between, first, last

```r
library(mclust)
```

## Package 'mclust' version 5.4.6
## Type 'citation("mclust")' for citing this R package in publications.

## Attaching package: 'mclust'

## The following object is masked from 'package:maps':
##
##     map

## The following object is masked from 'package:purrr':
##
##     map

```r
library(dtw)
```

## Warning: package 'dtw' was built under R version 4.0.2

## Loading required package: proxy

## Warning: package 'proxy' was built under R version 4.0.2

## Attaching package: 'proxy'

## The following objects are masked from 'package:stats':
##
##     as.dist, dist

## The following object is masked from 'package:base':
##
##     as.matrix
## Loade v1.21-3. See ?dtw for help, citation("dtw") for use in publication.

library(fields)

## Warning: package 'fields' was built under R version 4.0.2
## Loading required package: spam
## Warning: package 'spam' was built under R version 4.0.2
## Loading required package: dotCall64
## Warning: package 'dotCall64' was built under R version 4.0.2
## Spam version 2.5-1 (2019-12-12) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

## Attaching package: 'spam'

## The following objects are masked from 'package:base':
## backsolve, forwardsolve

## See https://github.com/NCAR/Fields for
## an extensive vignette, other supplements and source code

library(htmltools)

## Attaching package: 'htmltools'

## The following object is masked from 'package:mclust':
## em

library(cluster)
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.0.2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(ggridges)

## Warning: package 'ggridges' was built under R version 4.0.2

library(forcats)
library(fpc)

## Warning: package 'fpc' was built under R version 4.0.2
library(vioplot)

## Warning: package 'vioplot' was built under R version 4.0.2
## Loading required package: sm
## Warning: package 'sm' was built under R version 4.0.2
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
## Attaching package: 'sm'
## The following object is masked from 'package:MASS':
##   muscle
## The following object is masked from 'package:astsa':
##   birth
## Loading required package: zoo
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##   as.Date, as.Date.numeric

Read and Clean the Data

The original data is cleaned from any unknown or blank information. Dr. Macht’s test drives are removed because the associated charging events were designed for the test drives.

chargeData = read.csv('RI-ChargingData-2019-11.csv', stringsAsFactors = TRUE)

print(paste("Raw data has ", nrow(chargeData), " records")) #73947 events

## [1] "Raw data has  73947  records"

clean = which(chargeData$User.ID != "" &
               chargeData$User.ID != "0" &
               chargeData$User.ID != "457609" & # Dr. Macht
               chargeData$End.Date.Time != "" &
               chargeData$Start.Date.Time != "" &
               chargeData$Total.Duration..min. != "" &
               chargeData$Charging.Time..min. != "" &
               chargeData$Charging.Time..min. != "0" &
               chargeData$Start.Time.Zone != "" &
               chargeData$End.Time.Zone != "" &
               chargeData$Total.Duration..min. > chargeData$Charging.Time..min. &
               chargeData$Charging.Time..min. != "0" &
               chargeData$Total.Duration..min. != ""

chargeData$Energy.kWh.!="0"

chargeData = chargeData[!clean,]

print(paste("Editing for various columns leave ", nrow(chargeData), " records"))

## [1] "Editing for various columns leave  65041  records"

Clean from Ended.By (User cause System cause)

They are 15 possible reasons for a charging session to end. The ones that were caused by the users are kept in the data since this research is a charging behavior study.

count(chargeData, Ended.By)

## # A tibble: 13 x 2
## #  Ended.By                           n
## 1 ""                               103
## 2 "CPS Server"                     2621
## 3 "Customer"                      18996
## 4 "Holster Plugin"                  11
## 5 "Outlet Unreachable"              24
## 6 "Plug Out at Vehicle"            43229
## 7 "Plug Removed While Rebooting"    27
## 8 "Relay Stuck Closed"              1
## 9 "Station Deactivated"             6
##10 "Station Offline"                 5
##11 "Timeout"                        6
##12 "Unknown"                        11
##13 "Ventilation Fault"              1


chargeData=chargeData[clean.end,]

count(chargeData, Ended.By)

## # A tibble: 6 x 2
## #  Ended.By                        n
##1 CPS Server                       2621
## 2 Customer 18996
## 3 Holster Plugin 11
## 4 Plug Out at Vehicle 43229
## 5 Plug Removed While Rebooting 27
## 6 Timeout 6

print(paste("Editing resulted in ", nrow(chargeData), " records"))

## [1] "Editing resulted in 64890 records"

*Convert Start & End Date Time timezones to UTC*

Picked UTC to avoid all changes and to uniformize the charging events occurred time
This section has been done in Excel, Visual Basics

### Start Timezone

```r
startTime = as.POSIXct(chargeData$Start.Date.Time, format="$m/%d/%Y %H:%M")

# EDT
edt = which(chargeData$Start.Time.Zone == "EDT")
startTime[edt] = startTime[edt] - 4*60*60
defined

# EST
est = which(chargeData$Start.Time.Zone == "EST")
startTime[est] = startTime[est] - 5*60*60
```

### End Timezone

```r
endTime = as.POSIXct(chargeData$End.Date.Time, format="$m/%d/%Y %H:%M:%S")

# EDT
edt = which(chargeData$End.Time.Zone == "EDT")
endTime[edt] = endTime[edt] - 4*60*60

# EST
est = which(chargeData$End.Time.Zone == "EST")
endTime[est] = endTime[est] - 5*60*60
```

#Generated another column in Excel, names Start Date Time which is the converted (cleaned) format. Original one is being name Start Date Time 0

*Merge Stations*

The charging stations names can slightly change due to ownership modifications in the system. Minor changes occurred which is fixed here!

There are repeats in the stations due to formatting: taken out the #, ’, and spaces to merge the same stations

```r
chargeData$Station.Name = gsub("[# ]", "", chargeData$Station.Name)

print(paste("Editing for various columns leave ", nrow(chargeData), " records"))
```
Report Clean Data

Just to make sure all the numbers are right. Another code for reporting the data is written.

#Unique users and stations

```r
numVertices = 0
textNames = c()
textNumbers = c()
```

#Set up the vertices

```r
for (stationName in unique(chargeData$Station.Name)) {
    numVertices = numVertices + 1
    textNumbers[stationName] = numVertices
    textNames[numVertices] = stationName
}
numStations = numVertices
for (userId in unique(chargeData$User.ID)) {
    numVertices = numVertices + 1
    textNumbers[userId] = numVertices
    textNames[numVertices] = userId
}
numUsers = numVertices - numStations
```

#Edge list

```r
total = matrix(0, nrow = nrow(chargeData), ncol = 2)
for (index in 1:nrow(chargeData)) {
    total[index, 1] = vertexNumbers[chargeData$User.ID[index]]
    total[index, 2] = vertexNumbers[chargeData$Station.Name[index]]
}
```

```r
print(paste("Found ",numUsers," users and ", numStations," Stations"))
```

## [1] "Found 3044 users and 68 Stations"

## 80 20 Principle/Rule for Users

Pareto Principle is applied on the users to find out a subset of frequent users which this research continues on.

#Ranked from lowest to highest frequency

```r
TUsers = count(chargeData, User.ID)
Ufqnc = as.numeric(TUsers$n)
TUsers = TUsers[order(unlist(Ufqnc)),]
```

```r
URule = TUsers[order(TUsers$n),]
```

```r
URule80 = sum(URule[1:2435, 2])
URule20 = sum(URule[2436:3044, 2])
Utotal = sum(URule[, 2])
```
print(paste("Charging events total is", Utotal))

## [1] "Charging events total is 64890"

URule80/ Utotal

## [1] 0.1069502

URule20/ Utotal

## [1] 0.8930498

print(paste("Are 20 percent of the users contributed to 80 percent of the charging events? 
20% of the users contributed to %", round((URule20/ Utotal)*100) , "of the charging events"))

## [1] "Are 20 percent of the users contributed to 80 percent of the charging events? \n20% of the users contributed to % 89 of the charging events"

#Proved % 80 of charging events happened by % 20 of users

Subset top %20 Users

The frequent users are selected and the analysis were conducted on the subset of frequent users.

Users=TUsers[2436:3044,]
chargeSub=subset(chargeData,User.ID %in% Users$ User.ID)

#check to see if there are same users
print(paste("% 20 of the users contributed to %", round(nrow(chargeSub)/nrow(chargeData)*100) , "of the charging events. Is it greater than % 80 ?"))

## [1] "% 20 of the users contributed to % 89 of the charging events. Is it greater than % 80 ?"

#unique(chargeSub$User.ID) #3149 users
subUsers=count(chargeSub, User.ID)

subfqnc = as.numeric(subUsers$n)
subUsers = subUsers[order(unlist(subfqnc)),]

chargeSub=chargeSub[order(chargeSub$User.ID),]

##Tested the overall normality

ggdensity(chargeSub$Charging.Time..min.)
ggqqplot(chargeSub$Charging.Time..min.)
```r
# shapiro.test(chargeSub$Charging.Time..min.)

x = chargeSub$Charging.Time..min.
normtest = ks.test(x, "pnorm")

## Warning in ks.test(x, "pnorm"): ties should not be present for the Kolmogorov-Smirnov test

normtest$p.value

## [1] 0

if (normtest$p.value < 0.05) {
  print("Reject the null hypothesis the data is not normal")
} else {
  print("Fail to reject the null hypothesis the data is normal ")
}

## NULL

## Extracting the Subset file##

Created a .csv file from the subset of data.

# write.csv(chargeSub, "Charging Data-Subset.csv")
# SubData <- read.csv('Charging Data-Subset.csv',stringsAsFactors = TRUE)

## Approximate state of charge
```
Augmented the %SOC for creating the profile of users based on their received energy at each charging event.

```r
#SOCData <- SubData %>% data.frame %>%
#  group_by(User.ID) %>%
#  mutate(SOC=(Energy..kWh./max(Energy..kWh.))*100) %>%
#  mutate(R.SOC=round(SOC)) %>%
#  mutate(Flp.SOC=(1-(Energy..kWh./max(Energy..kWh.)))*100)
```

**Extracting SOCData & Read**

Created a .csv file from the existing data + augment SOCs received.

```r
#write.csv(SOCData, "Charging Data-SOC.csv")
SOCData <- read.csv( 'Charging Data-SOC.csv', stringsAsFactors = TRUE)
nrow(SOCData)
## [1] 58273
```

## Vector histogram for each user

Created a profile of individual based on their frequencies at different state of charges. Scaled their values (dividing by counts) to be comparable for distances calculations.

```r
#Approach I
bins=c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
list.ids = c()
for (i in unique(SOCData$User.ID)) {
  list.ids[[as.character(i)]] =as.vector(hist(SOCData[SOCData$User.ID == i, 'SOC'], breaks=bins)$counts / sum(hist(SOCData[SOCData$User.ID == i, 'SOC'], breaks=bins)$count) * 100)
}
length(list.ids)
## [1] 608
class(list.ids[[i]])
## [1] "numeric"

#Approach II
#lapply(split(SOCData, SOCData$User.ID), function(x) hist(x$SOC, breaks=bins)$counts / sum(hist(x$SOC, breaks = 10)$count) * 100)

## Hamming distance

Hamming distance computation.

```r
Mfreq=do.call(rbind, list.ids)
Hdist=hamming.distance(Mfreq)
View(head(Hdist))
class(Hdist)
```
## [1] "matrix" "array"

dim(Hdist)
## [1] 608 608

#test
no.1 = c(3.703704, 6.172840, 8.641975, 24.69136, 39.50617, 4.938272, 4.938272, 3.703704, 1.234568, 2.469136)
no.2 = c(4.347826, 8.695652, 8.695652, 21.73913, 26.08696, 4.347826, 13.043478, 0.000000, 8.695652, 4.347826)

hamming.distance(no.1, no.2)
## [1] 10

### Hamming K-means clustering

Screeplot a range number of clusters. The knee in the plot is the optimal number of clusters.

```r
sse = 0
for (i in 1:30) {
  km.out <- kmeans(Hdist, centers = i, nstart = 25)
  sse[i] <- km.out$tot.withinss
}

# Plot total within sum of squares vs. number of clusters
plot(1:30, sse, type = "b",
     xlab = "Hamming Number of Clusters",
     ylab = "Within Groups Sum of Squares")
```
Examined the knee on the screeplot, one above and under.

# Approach I
# cluster for 5
# ?kmeans
# Hk5=kmeans(Hdist, 5, nstart=25)
# Hk5.sz=kmeans(Hdist, 5, nstart = 25)$size

# Graphing each
# clusplot(Hdist, Hk4$cluster)

# Representing the Clusters Hamming

Approach II: Examined the clusters of 4, 5, 6, 7, and 8 for this distance. 1. Each cluster is actually implemented 2. A boxplot of each is created to better understand the related behavior within each cluster. 3. OPTIONAL: the Violin Plot is graphed to understand the variation of each bin within each cluster.

Hk4C <- eclust(Hdist, "kmeans", k = 4, nstart = 25, graph = FALSE)
Hk4Cg=fviz_cluster(Hk4C, geom = "point", ellipse.type = "norm", palette = "jco", ggtheme = theme_minimal())

Mfrq.df=as.data.frame(Mfrq)
Hk4C.df=as.data.frame(Hk4C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Hk4C.df <- tibble::rownames_to_column(Hk4C.df, "User.ID")
SOCclust <- merge(Hk4C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Hk4C$cluster")]<- "Cluster"

SOCsplit4H <- split(SOCclust, SOCclust$Cluster)

Hk4C$size
## [1] 331 89 57 131

par(mfrow=c(2,2))
boxplot(SOCsplit4H$1[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit4H$2[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit4H$3[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit4H$4[,3:12], main="Cluster 4", ylim=c(0,100))

nrow(SOCsplit4H$4)
## [1] 131

nrow(SOCsplit4H$3)
## [1] 57

nrow(SOCsplit4H$2)
## [1] 89

nrow(SOCsplit4H$1)
```r
## [1] 331

Hk5C <- eclust(Hdist, "kmeans", k = 5, nstart = 25, graph = FALSE)
Hk5Cg=fviz_cluster(Hk5C, geom = "point", ellipse.type = "norm",
palette = "jco", ggtheme = theme_minimal())

Mfrq.df=as.data.frame(Mfrq)
Hk5C.df=as.data.frame(Hk5C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Hk5C.df <- tibble::rownames_to_column(Hk5C.df, "User.ID")

SOCclust <- merge(Hk5C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Hk5C$cluster")]<- "Cluster"

SOCsplit5H <- split(SOCclust, SOCclust$Cluster)

Hk5C$size
## [1]  57 121  95  57 278

par(mfrow=c(3,2))
boxplot(SOCsplit5H$`1`, main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit5H$`2`, main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit5H$`3`, main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit5H$`4`, main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit5H$`5`, main="Cluster 5", ylim=c(0,100))

#Wk5$size
nrow(SOCsplit5H$`5`)
## [1] 278
nrow(SOCsplit5H$`4`)
## [1] 57
nrow(SOCsplit5H$`3`)
## [1] 95
nrow(SOCsplit5H$`2`)
## [1] 121
nrow(SOCsplit5H$`1`)
## [1] 57
```

par(mfrow=c(3, 2))
vioplot(SOCsplit5HS$1[,3:12], col="grey", main="Cluster 1", ylim=(0,100))
vioplot(SOCsplit5HS$2[,3:12], col="grey", main="Cluster 2", ylim=(0,100))
vioplot(SOCsplit5HS$3[,3:12], col="grey", main="Cluster 3", ylim=(0,100))
vioplot(SOCsplit5HS$4[,3:12], col="grey", main="Cluster 4", ylim=(0,100))
vioplot(SOCsplit5HS$5[,3:12], col="grey", main="Cluster 5", ylim=(0,100))
Hk6C <- eclust(Hdist, "kmeans", k = 6, nstart = 25, graph = FALSE)
Hk6Cg = fviz_cluster(Hk6C, geom = "point", ellipse.type = "norm",
 palette = "jco", ggtheme = theme_minimal())

Mfrq.df = as.data.frame(Mfrq)
Hk6C.df = as.data.frame(Hk6C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Hk6C.df <- tibble::rownames_to_column(Hk6C.df, "User.ID")

SOCclust <- merge(Hk6C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Hk6C$cluster")]<- "Cluster"

SOCsplit6H <- split(SOCclust, SOCclust$Cluster)

Hk6C$size
## [1]  72  55  91  63 273  54
par(mfrow = c(3, 2))
boxplot(SOCsplit6H$i[, 3:12], main = "Cluster 1", ylim = c(0, 100))
boxplot(SOCsplit6H$i[, 3:12], main = "Cluster 2", ylim = c(0, 100))
boxplot(SOCsplit6H$i[, 3:12], main = "Cluster 3", ylim = c(0, 100))
boxplot(SOCsplit6H$i[, 3:12], main = "Cluster 4", ylim = c(0, 100))
boxplot(SOCsplit6H$i[, 3:12], main = "Cluster 5", ylim = c(0, 100))
boxplot(SOCsplit6H$i[, 3:12], main = "Cluster 6", ylim = c(0, 100))
```r
#Wk5$size
nrow(SOCsplit6H$`6`)  # [1] 54
nrow(SOCsplit6H$`5`)  # [1] 273
nrow(SOCsplit6H$`4`)  # [1] 63
nrow(SOCsplit6H$`3`)  # [1] 91
nrow(SOCsplit6H$`2`)  # [1] 55
nrow(SOCsplit6H$`1`)  # [1] 72

Hk7C <- eclust(Hdist, "kmeans", k = 7, nstart = 25, graph = FALSE)
Hk7Cg = fviz_cluster(Hk7C, geom = "point", ellipse.type = "norm",
                      palette = "jco", ggtheme = theme_minimal())
```
Mfrq.df <- as.data.frame(Mfrq)
Hk7C.df <- as.data.frame(Hk7C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq, "User.ID")
Hk7C.df <- tibble::rownames_to_column(Hk7C.df, "User.ID")

SOCclust <- merge(Hk7C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Hk7C$cluster")]<- "Cluster"

SOCsplit7H <- split(SOCclust, SOCclust$Cluster)

Hk7C$size

## [1]  39 262  71  27  80  57  72

par(mfrow=c(3,3))
boxplot(SOCsplit7H$`1`[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit7H$`2`[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit7H$`3`[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit7H$`4`[,3:12], main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit7H$`5`[,3:12], main="Cluster 5", ylim=c(0,100))
boxplot(SOCsplit7H$`6`[,3:12], main="Cluster 6", ylim=c(0,100))
boxplot(SOCsplit7H$`7`[,3:12], main="Cluster 7", ylim=c(0,100))

nrow(SOCsplit7H$`7`)  
## [1] 72
nrow(SOCsplit7H$`6`)  
## [1] 57
nrow(SOCsplit7H$`5`)  
## [1] 80
nrow(SOCsplit7H$`4`)  
## [1] 27
nrow(SOCsplit7H$`3`)  
## [1] 71
nrow(SOCsplit7H$`2`)  
## [1] 262
nrow(SOCsplit7H$`1`)  
## [1] 39
Hk8C <- eclust(Hdist, "kmeans", k = 8, nstart = 25, graph = FALSE)
Hk8Cg=viz_cluster(Hk8C, geom = "point", ellipse.type = "norm",
palette = "jco", ggtheme = theme_minimal())

Mfrq.df=as.data.frame(Mfrq)
Hk8C.df=as.data.frame(Hk8C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Hk8C.df <- tibble::rownames_to_column(Hk8C.df, "User.ID")

SOCclust <- merge(Hk8C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Hk8C$cluster")]<-"Cluster"

SOCsplit8H <- split(SOCclust, SOCclust$Cluster)

Hk8C$size
## [1] 262 47 27 55 71 36 39 71

par(mfrow=c(3,3))
boxplot(SOCsplit8H$`1`[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit8H$`2`[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit8H$`3`[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit8H$`4`[,3:12], main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit8H$`5`[,3:12], main="Cluster 5", ylim=c(0,100))
boxplot(SOCsplit8H$`6`[,3:12], main="Cluster 6", ylim=c(0,100))
boxplot(SOCsplit8H$`7`[,3:12], main="Cluster 7", ylim=c(0,100))
boxplot(SOCsplit8H$`8`[,3:12], main="Cluster 8", ylim=c(0,100))
\begin{verbatim}
nrow(SOCsplit8H$^8$)
## [1] 71
nrow(SOCsplit8H$^7$)
## [1] 39
nrow(SOCsplit8H$^6$)
## [1] 36
nrow(SOCsplit8H$^5$)
## [1] 71
nrow(SOCsplit8H$^4$)
## [1] 55
nrow(SOCsplit8H$^3$)
## [1] 27
nrow(SOCsplit8H$^2$)
## [1] 47
nrow(SOCsplit8H$^1$)
## [1] 262
\end{verbatim}
#Wasserstein Distance

Wasserstein distance computation.

\[
Wdist = \text{matrix}(0, 608, 608)
\]

for (i in 1:nrow(Mfrq)) {
  for (j in 1:nrow(Mfrq)) {
    Wdist[i,j] <- wasserstein1d(Mfrq[i,], Mfrq[j,])
  }
}

#Wasserstein K-means Clustering

Screeplot a range number of clusters. The knee in the plot is the optimal number of clusters.

\[
nse = 0
\]

for (i in 1:20) {
  km.out <- kmeans(Wdist, centers = i, nstart = 25)
  sse[i] <- km.out$tot.withinss
}

# Plot total within sum of squares vs. number of clusters

plot(1:20, sse, type = 'b',
     xlab = "Wasserstein Number of Clusters",
     ylab = "Within Groups Sum of Squares")


Representing the Clusters Wasserstein

Examined the clusters of 4, 5, and 6 for this distance. 1. Each cluster is actually implemented. 2. A boxplot of each is created to better understand the related behavior within each cluster. 3. OPTIONAL: the Violin Plot is graphed to understand the variation of each bin within each cluster.

```r
Wk4C <- eclust(Wdist, "kmeans", k = 4, nstart = 25, graph = FALSE)
Wk4Cg = fviz_cluster(Wk4C, geom = "point", ellipse.type = "norm",
palette = "jco", ggtheme = theme_minimal())

Mfrq.df = as.data.frame(Mfrq)
Wk4C.df = as.data.frame(Wk4C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Wk4C.df = cbind(Mfrq.df$User.ID, Wk4C.df)
colnames(Wk4C.df)[which(names(Wk4C.df) == "Mfrq.df$User.ID")]
<- "User.ID"

SOCclust <- merge(Wk4C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Wk4C$cluster")]
<- "Cluster"

SOCsplit4W <- split(SOCclust, SOCclust$Cluster)

Wk4C$size
## [1] 229 31 278 70
```
par(mfrow=c(2,2))
boxplot(SOCsplit4WS$1[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit4WS$2[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit4WS$3[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit4WS$4[,3:12], main="Cluster 4", ylim=c(0,100))

Wk5C <- eclust(Wdist, "kmeans", k = 5, nstart = 25, graph = FALSE)
Wk5Cg=fviz_cluster(Wk5C, geom = "point", ellipse.type = "norm",
    palette = "jco", ggtheme = theme_minimal())

Mfrq.df=as.data.frame(Mfrq)
Wk5C.df=as.data.frame(Wk5C$cluster)
Mfrq.df <- tibble::rownames_to_column(Mfrq.df, 'User.ID')
Wk5C.df=cbind(Mfrq.df$User.ID, Wk5C.df)
colnames(Wk5C.df)[which(names(Wk5C.df) == 'Mfrq.df$User.ID')] <- 'User.ID'
SOCclust <- merge(Wk5C.df, Mfrq.df, by='User.ID')
colnames(SOCclust)[which(names(SOCclust) == 'Wk5C$cluster')] <- 'Cluster'
SOCsplit5W <- split(SOCclust, SOCclust$Cluster)
MSOC=merge(Wk5C.df, SOCData, by='User.ID')
colnames(MSOC)[which(names(MSOC) == 'Wk5C$cluster')] <- 'Cluster'
MsSOC=splIt(MSOC, MSOC$Cluster)
summary(MsSOC$`$Charging.Time..min.)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.05   75.77  128.04  132.69  191.00  792.62

Wk5C$size
## [1] 240  48 127  29 164

par(mfrow=c(3,2))
boxplot(SOCsplit5W$`1`[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit5W$`2`[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit5W$`3`[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit5W$`4`[,3:12], main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit5W$`5`[,3:12], main="Cluster 5", ylim=c(0,100))

nrow(SOCsplit5W$`5`)
## [1] 164

nrow(SOCsplit5W$`4`)
## [1] 29

nrow(SOCsplit5W$`3`)
## [1] 127

nrow(SOCsplit5W$`2`)
## [1] 48

nrow(SOCsplit5W$`1`)
## [1] 240
```r
par(mfrow=c(3,2))
vioplot(SOCsplit5WS1[,3:12], col="grey", main="Cluster 1", ylim=c(0,100))
vioplot(SOCsplit5WS2[,3:12], col="grey", main="Cluster 2", ylim=c(0,100))
vioplot(SOCsplit5WS3[,3:12], col="grey", main="Cluster 3", ylim=c(0,100))
vioplot(SOCsplit5WS4[,3:12], col="grey", main="Cluster 4", ylim=c(0,100))
vioplot(SOCsplit5WS5[,3:12], col="grey", main="Cluster 5", ylim=c(0,100))
```
Wk6C <- eclust(Wdist, "kmeans", k = 6, nstart = 25, graph = FALSE)
Wk6Cg = fviz_cluster(Wk6C, geom = "point", ellipse.type = "norm";
palette = "jco", ggtheme = theme_minimal())

Mfrq.df = as.data.frame(Mfrq)
Wk6C.df = as.data.frame(Wk6C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq, "User.ID")
Wk6C.df = cbind(Mfrq.df$User.ID, Wk6C.df)
colnames(Wk6C.df)[which(names(Wk6C.df) == "Mfrq.df$User.ID")]<-"User.ID"

SOCclust <- merge(Wk6C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Wk6C$cluster")]<-"Cluster"

SOCsplit6W <- split(SOCclust, SOCclust$Cluster)
Wk6C$size
## [1] 45 103 29 165 178 88

par(mfrow=c(3,2))
boxplot(SOCsplit6W$V1[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit6W$V2[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit6W$V3[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit6W$V4[,3:12], main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit6W$V5[,3:12], main="Cluster 5", ylim=c(0,100))
boxplot(SOCsplit6W$V6[,3:12], main="Cluster 6", ylim=c(0,100))
Mann-Whitney Wilcoxon Test

This test has been performed with each cluster to better make decisions on charging behaviors.
W6C1 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6WS$1 [, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
  for (j in 1:ncol(SOCsplit6V)) {
    W6C1[i, j] <- wilcox.test(SOCsplit6V[, i], SOCsplit6V[, j])$p.value
  }
}
View(W6C1)

W6C2 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6WS$2 [, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
  for (j in 1:ncol(SOCsplit6V)) {
    W6C2[i, j] <- wilcox.test(SOCsplit6V[, i], SOCsplit6V[, j])$p.value
  }
}
#wilcox.test(SOCsplit6S$2[,3], SOCsplit6S$2[,7])

W6C3 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6WS$3 [, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
  for (j in 1:ncol(SOCsplit6V)) {
    W6C3[i, j] <- wilcox.test(SOCsplit6V[, i], SOCsplit6V[, j])$p.value
  }
}
View(W6C3)

W6C4 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6WS$4 [, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
  for (j in 1:ncol(SOCsplit6V)) {
    W6C4[i, j] <- wilcox.test(SOCsplit6V[, i], SOCsplit6V[, j])$p.value
  }
}
View(W6C4)

W6C5 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6WS$5 [, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
  for (j in 1:ncol(SOCsplit6V)) {
    W6C5[i, j] <- wilcox.test(SOCsplit6V[, i], SOCsplit6V[, j])$p.value
  }
}
View(W6C5)

W6C6 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6WS$6 [, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
  for (j in 1:ncol(SOCsplit6V)) {
    W6C6[i, j] <- wilcox.test(SOCsplit6V[, i], SOCsplit6V[, j])$p.value
  }
}
# Minkowski Distance

Minkowski distance computation.

```r
Kowdist = as.matrix(dist(Mfrq, method = "minkowski", p = 2))
```

# Minkowski K-means clustering

Screeplot a range number of clusters. The knee in the plot is the optimal number of clusters.

```r
sse = 0
for (i in 1:30) {
  km.out <- kmeans(Kowdist, centers = i, nstart = 25)
  sse[i] <- km.out$tot.withinss
}
```

# Plot total within sum of squares vs. number of clusters

```r
plot(1:30, sse, type = "b",
     xlab = "Minkowski Number of Clusters",
     ylab = "Within Groups Sum of Squares")
```

# Representing the Clusters Minkowski
Examined the clusters of 4, 5, and 6 for this distance. 1. Each cluster is actually implemented. 2. A boxplot of each is created to better understand the related behavior within each cluster. 3. OPTIONAL: the Violin Plot is graphed to understand the variation of each bin within each cluster.

```r
Kk4C <- eclus(Kowdist, "kmeans", k = 4, nstart = 25, graph = FALSE)
Kk4Cg = fviz_cluster(Kk4C, geom = "point", ellipse.type = "norm", palette = "jco", ggtheme = theme_minimal())
Mfreq.df = as.data.frame(Mfreq)
Kk4C.df = as.data.frame(Kk4C$cluster)
Mfreq.df <- tibble::rownames_to_column(Mfreq.df, "User.ID")
Kk4C.df <- tibble::rownames_to_column(Kk4C.df, "User.ID")
SOCclust <- merge(Kk4C.df, Mfreq.df, by = "User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Kk4C$cluster")]<-"Cluster"
SOCsplit4K <- split(SOCclust, SOCclust$Cluster)
Kk4Cssize
## [1] 319 147 104 38
par(mfrow = c(2, 2))
boxplot(SOCsplit4K$`1`[,3:12], main = "Cluster 1", ylim = c(0, 100))
boxplot(SOCsplit4K$`2`[,3:12], main = "Cluster 2", ylim = c(0, 100))
boxplot(SOCsplit4K$`3`[,3:12], main = "Cluster 3", ylim = c(0, 100))
boxplot(SOCsplit4K$`4`[,3:12], main = "Cluster 4", ylim = c(0, 100))
```
nrow(SOCsplit4K$'4')
## [1] 38
nrow(SOCsplit4K$'3')
## [1] 104
nrow(SOCsplit4K$'2')
## [1] 147
nrow(SOCsplit4K$'1')
## [1] 319

Kk5C <- eclust(Kowdist, "kmeans", k = 5, nstart = 25, graph = FALSE)
Kk5Cg=fviz_cluster(Kk5C, geom = "point", ellipse.type = "norm",
palette = "jco", ggtheme = theme_minimal())

Mfrq.df=as.data.frame(Mfrq)
Kk5C.df=as.data.frame(Kk5C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Kk5C.df <- tibble::rownames_to_column(Kk5C.df, "User.ID")

SOCclust <- merge(Kk5C.df, Mfrq.df, by="User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Kk5C$cluster")]<-"Cluster"
SOCsplit5K <- split(SOCclust, SOCclust$Cluster)
Kk5C$size
## [1]  33 201 218  47  10

par(mfrow=c(3,2))
boxplot(SOCsplit5K$`1`[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit5K$`2`[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit5K$`3`[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit5K$`4`[,3:12], main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit5K$`5`[,3:12], main="Cluster 5", ylim=c(0,100))

nrow(SOCsplit5K$`5``)
## [1] 109
nrow(SOCsplit5K$`4`)
## [1] 47
nrow(SOCsplit5K$`3`)
## [1] 218
nrow(SOCsplit5K$`2`)
## [1] 201
nrow(SOCsplit5K$`1`)
## [1] 33
```r
par(mfrow=c(3, 2))
vioplot(SOCsplit5KS$1[,3:12], col="grey", main="Cluster 1", ylim=c(0,100))
vioplot(SOCsplit5KS$2[,3:12], col="grey", main="Cluster 2", ylim=c(0,100))
vioplot(SOCsplit5KS$3[,3:12], col="grey", main="Cluster 3", ylim=c(0,100))
vioplot(SOCsplit5KS$4[,3:12], col="grey", main="Cluster 4", ylim=c(0,100))
vioplot(SOCsplit5KS$5[,3:12], col="grey", main="Cluster 5", ylim=c(0,100))
```
Kk6C <- eclust(Kowdist, "kmeans", k = 6, nstart = 25, graph = FALSE)
Kk6Cg = fviz_cluster(Kk6C, geom = "point", ellipse.type = "norm", palette = "jco", ggtheme = theme_minimal())

Mfrq.df = as.data.frame(Mfrq)
Kk6C.df = as.data.frame(Kk6C$cluster)

Mfrq.df <- tibble::rownames_to_column(Mfrq.df, "User.ID")
Kk6C.df <- tibble::rownames_to_column(Kk6C.df, "User.ID")

SOCclust <- merge(Kk6C.df, Mfrq.df, by = "User.ID")
colnames(SOCclust)[which(names(SOCclust) == "Kk6C$cluster") ] <- "Cluster"

SOCsplit6K <- split(SOCclust, SOCclust$Cluster)

Kk6C$size
## [1] 146 42 28 126 54 212

par(mfrow=c(3,2))
boxplot(SOCsplit6K$1[,3:12], main="Cluster 1", ylim=c(0,100))
boxplot(SOCsplit6K$2[,3:12], main="Cluster 2", ylim=c(0,100))
boxplot(SOCsplit6K$3[,3:12], main="Cluster 3", ylim=c(0,100))
boxplot(SOCsplit6K$4[,3:12], main="Cluster 4", ylim=c(0,100))
boxplot(SOCsplit6K$5[,3:12], main="Cluster 5", ylim=c(0,100))
boxplot(SOCsplit6K$6[,3:12], main="Cluster 6", ylim=c(0,100))
# Mann-Whitney Wilcoxon Test

This test has been performed with each clusters to better make decisions on charging behaviors.
K6C1 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6K$'1'[, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
    for (j in 1:ncol(SOCsplit6V)) {
        K6C1[i,j] <- wilcox.test(SOCsplit6V[,i], SOCsplit6V[,j])$p.value
    }
}
View(K6C1)

K6C2 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6K$'2'[, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
    for (j in 1:ncol(SOCsplit6V)) {
        K6C2[i,j] <- wilcox.test(SOCsplit6V[,i], SOCsplit6V[,j])$p.value
    }
}
View(K6C2)

K6C3 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6K$'4'[, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
    for (j in 1:ncol(SOCsplit6V)) {
        K6C3[i,j] <- wilcox.test(SOCsplit6V[,i], SOCsplit6V[,j])$p.value
    }
}
View(K6C3)

K6C4 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6K$'5'[, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
    for (j in 1:ncol(SOCsplit6V)) {
        K6C4[i,j] <- wilcox.test(SOCsplit6V[,i], SOCsplit6V[,j])$p.value
    }
}
View(K6C4)

K6C5 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6K$'6'[, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
    for (j in 1:ncol(SOCsplit6V)) {
        K6C5[i,j] <- wilcox.test(SOCsplit6V[,i], SOCsplit6V[,j])$p.value
    }
}
View(K6C5)

K6C6 = matrix(0, 10, 10)
SOCsplit6V = as.matrix(SOCsplit6K$'6'[, 3:12])
for (i in 1:ncol(SOCsplit6V)) {
    for (j in 1:ncol(SOCsplit6V)) {
        K6C6[i,j] <- wilcox.test(SOCsplit6V[,i], SOCsplit6V[,j])$p.value
    }
}
View(K6C6)

# Silhouette Validation

Decision making between the distance metrics.
```r
eHk4 <- eclust(Hdist, "kmeans", k = 4, nstart = 25, graph = FALSE)
sil=silhouette(eHk4$cluster, dist(Hdist))
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1  331          0.24
## 2       2   89          0.23
## 3       3   57          0.24
## 4       4  131          0.07
```

**Clusters silhouette plot**

**Average silhouette width: 0.2**

```
eHk5 <- eclust(Hdist, "kmeans", k = 5, nstart = 25, graph = FALSE)
sil=silhouette(eHk5$cluster, dist(Hdist))
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1   57          0.15
## 2       2  121          0.07
## 3       3   95          0.07
## 4       4   57          0.22
## 5       5  278          0.26
```
Clusters silhouette plot
Average silhouette width: 0.18

eHk6 <- eclus(Hdist, "kmeans", k = 6, nstart = 25, graph = FALSE)
sil <- silhouette(eHk6$cluster, dist(Hdist))
fviz_silhouette(sil)

##   cluster size  ave.sil.width
## 1       1   72          0.11
## 2       2   55          0.16
## 3       3   91          0.07
## 4       4   63          0.09
## 5       5  273          0.26
## 6       6   54          0.23
eHk7 <- eclust(Hdist, "kmeans", k = 7, nstart = 25, graph = FALSE)
sil = silhouette(eHk7$cluster, dist(Hdist))
fviz_silhouette(sil)

##   cluster size ave.sil.width
## 1       1  39          0.19
## 2       2 262          0.23
## 3       3  71          0.10
## 4       4  27          0.17
## 5       5  80          0.05
## 6       6  57          0.07
## 7       7  72          0.14
Clusters silhouette plot
Average silhouette width: 0.16

eHk8 <- eclust(Hdist, "kmeans", k = 8, nstart = 25, graph = FALSE)
sil = silhouette(eHk8$cluster, dist(Hdist))
fviz_silhouette(sil)

##   cluster size ave.sil.width
## 1       1  262          0.23
## 2       2   47          0.16
## 3       3   27          0.17
## 4       4   55          0.08
## 5       5   71          0.12
## 6       6   36          0.06
## 7       7   39          0.19
## 8       8   71          0.09
Clusters silhouette plot
Average silhouette width: 0.17

```
eWk4 <- eclust(Wdist, "kmeans", k = 4, nstart = 25, graph = FALSE)
sil = silhouette(eWk4$cluster, dist(Wdist))
fviz_silhouette(sil)
```

<table>
<thead>
<tr>
<th>cluster</th>
<th>size</th>
<th>ave.sil.width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>229</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>278</td>
<td>0.43</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Clusters silhouette plot
Average silhouette width: 0.39

```
eWk5 <- eclus(Wdist, "kmeans", k = 5, nstart = 25, graph = FALSE)
sil=silhouette(eWk5$cluster, dist(Wdist))
fviz_silhouette(sil)
```

<table>
<thead>
<tr>
<th>cluster</th>
<th>size</th>
<th>ave.sil.width</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1</td>
<td>0.31</td>
</tr>
<tr>
<td>#2</td>
<td>2</td>
<td>0.39</td>
</tr>
<tr>
<td>#3</td>
<td>3</td>
<td>0.32</td>
</tr>
<tr>
<td>#4</td>
<td>4</td>
<td>0.64</td>
</tr>
<tr>
<td>#5</td>
<td>5</td>
<td>0.38</td>
</tr>
</tbody>
</table>
eWk6 <- eclus(Wdist, "kmeans", k = 6, nstart = 25, graph = FALSE)
sil = silhouette(eWk6$cluster, dist(Wdist))
fviz_silhouette(sil)

##   cluster size ave.sil.width
## 1       1   45          0.38
## 2       2  103          0.30
## 3       3   29          0.63
## 4       4  165          0.23
## 5       5  178          0.33
## 6       6   88          0.36
eKk4 <- eclus(Kowdist, "kmeans", k = 4, nstart = 25, graph = FALSE)
sil=silhouette(eKk4$cluster, dist(Kowdist))
fviz_silhouette(sil)

##   cluster size ave.sil.width
## 1       1  319          0.36
## 2       2  147          0.28
## 3       3  104          0.18
## 4       4   38          0.47
\texttt{eKk5 <- eclust(Kowdist, "kmeans", k = 5, nstart = 25, graph = FALSE)}
\texttt{sil = silhouette(eKk5$cluster, dist(Kowdist))}
\texttt{fviz_silhouette(sil)}

<table>
<thead>
<tr>
<th></th>
<th>cluster</th>
<th>size</th>
<th>ave.sil.width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>33</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>201</td>
<td>0.24</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>218</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>47</td>
<td>0.19</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>109</td>
<td>0.26</td>
</tr>
</tbody>
</table>
eKk6 <- eclust(Kowdist, "kmeans", k = 6, nstart = 25, graph = FALSE)
sil=silhouette(eKk6$cluster, dist(Kowdist))
fviz_silhouette(sil)

<table>
<thead>
<tr>
<th>cluster</th>
<th>size</th>
<th>ave.sil.width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>146</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>126</td>
<td>0.19</td>
</tr>
<tr>
<td>5</td>
<td>54</td>
<td>0.31</td>
</tr>
<tr>
<td>6</td>
<td>212</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Clusters silhouette plot
Average silhouette width: 0.28
#Heatmap Distances

Visualized the distance metrics. Another approach for deciding between distance metrics.

`heatmap(Hdist)`