1-1-2022

CHARACTERIZING GAME-PLAY STYLES IN NATIONAL BASKETBALL ASSOCIATION

Zian Zheng
University of Rhode Island, zian_zheng@uri.edu

CHARACTERIZING GAME-PLAY STYLES IN NATIONAL BASKETBALL ASSOCIATION

BY

ZIAN ZHENG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2022

MASTER OF SCIENCE THESIS

OF

ZIAN ZHENG

APPROVED:

Thesis Committee:

Major Professor Guangyu Zhu

Yichi Zhang

Feihong Xia

Brenton DeBoef

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2022

ABSTRACT

Data analysis provides important insights to inform decision-making by discovering patterns and trends in datasets. With the increasing collection and availability of large datasets, data science has the potential to revolutionize decision-making in many fields. Basketball is one of the most popular sports in the world, and the application of data science in basketball has attracted a large amount of attention in recent years. Our research explores the dissimilarity and similarity of game-play characteristics through recent years within the National Basketball Association. We focused on the data analysis of basketball teams and player performance, analysis, and prediction of competition results. We used multidimensional scaling and autoencoder to examine the dissimilarity and similarity of team performance. The research will show that data analysis can enable teams to increase their understanding of past performance and factors that drove success. Teams can also use analysis to make predictions of future success, and to inform decisions that lead to success.

ACKNOWLEDGMENTS

I would like to thank my major Prof. Guangyu Zhu for his dedicate support and guidance during my whole graduate years. Professor Zhu continuously provided encouragement and always willing and enthusiastic to assist all the way throughout my thesis research. Furthermore I would like to thank Prof. Yichi Zhang and Prof. Feihong Xia for providing suggestions to my research.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Table **Page**

CHAPTER 1

Introduction

1.1 Data Science in Sport

Many individuals consider sports to be vital activities. One explanation is that many of them play sports to exercise, improve their health, and lead healthier lives. Another factor is that watching and following professional sports is a popular pastime for both young people and adults. Sports are watched on television by people all around the world, many of them daily (Beck and Bosshart, 2003). Additionally, sports fans are frequently quite engaged, analyzing coaches' choices, contrasting players' metrics, and projecting game results as well as the ultimate standings of players and teams competing in contests. Regular sports sections are found in many newspapers, and both individual and team sports have their own television channels. The Olympic Games, the World Cup in soccer, the World Championships in basketball and swimming, as well as other major sporting events, are among the most watched events in the world (Leeds et al., 2018). The numerous facets of the sports industry cost billions of dollars, from the price of game tickets to the cost of broadcasting licenses, top players' wages, and advertising.

Data science applied to sports is growing rapidly, more and more coaches, players, scouts, and team managers recognize the value of data and data analytics to their teams (Karlis and Ntzoufras, 2003). Sports fans and commentators are also increasingly using sports data analysis, especially technical statistics, to evaluate games and athletes. Sports analytics is an emerging field because the domain's value for teams, players, and organizations is enormous. In recent years, it has been discovered that analytics and performance prediction in the sports domain is necessary for the evolution of any sport, team, and even the players (Song et al., 2020). The big organizations-teams have departments focused on

their and opponent team analytics, trying to optimize their playstyle and detect problems that staff, players, and coaches cannot see. For this reason, data has incredible value for the teams, and via different methods, collect as much data as possible for evaluation.

It is well known that Sports Analytics is an emerging field (O'donoghue, 2009). It is used by all big sports organizations and professional teams, helping develop the team, improving the results, and noticing problems that are hard to find. The technology improvement over the years has created new playstyles and tactics. Also, evaluating of the results with the help of analytics is a big deal for every kind of sport. Nowadays, the experience of a coach is not enough to be competitive at a professional level (Turner et al., 2015).

Years ago, when computers were not as capable tools as they are nowadays for gathering data and making analyses, collecting data was manual, handwritten, hard to observe, and time-consuming (Lees, 2002). For this reason, it is considered normal to have no statistical records for most sports games. While, chronologically, sports analytics appeared in the 19th century, when the idea of analyzing a player's play helps evaluate the player's skills comes up.

Since data from various sporting events have been routinely collected for many years and are frequently made available to the general public (entire games are recorded on video in addition to the quantitative data that are retained in datasets), the field of sport offers a particularly authentic setting for investigating research ideas. Sport is a great source for analytics and research, especially when it comes to certain aspects of human behavior (Castellano et al., 2008). This is in addition to the sport's big-data characteristic.

In sports, the rules of the games are well-defined and transparent. Professional athletes are regarded as subject-matter specialists who receive substantial

rewards for their greatest efforts. Varied sports have different rules (such as individual versus team sports), which result in a range of situations that each allow for the evaluation of various performance-related metrics. The fact that sports are popular everywhere makes it possible to do global analyses utilizing data from various regions of the world or by comparing countries. As universal factors that influence human conduct in general are frequently reflected in sports behavior as well, patterns of behavior in sports can frequently offer insights about all sorts of human behavior (Lees, 2002).

Many times, businesses that specialize in measuring and categorizing sporting performance give the data for sports research with the intention of selling specialized bundles of data to clubs, associations, broadcasters, and university researchers. For instance, these businesses might provide the team's scouting staff with thorough data on basketball or soccer players' accomplishments in any league that competes at the professional or semi-professional level worldwide (McCorey, 2021).

1.2 Basketball and NBA

History of Basketball In December 1891, James Naismith, a college teacher at Springfield in Massachusetts, invented Basketball (Rains, 2011). It all starts when his students run out of choices when they should play a game indoors. His choice then was to force them to play the already invented games, like Football or baseball. However, the circumstances exclude these options. Then he remembered a game called rock-tossing that he was playing as a child, and he proposed a game that players would throw a ball to a target. He used two peach baskets, nailed them on ten feet above the floor at each end of the gym and a football ball. That was the day, basketball was invented (Grissmer, 2008). Following years, Basketball spread quickly around the world, while Naismith's students helped by introducing the sport to new people. High Schools and colleges started to organize teams,

today's basketball ball was invented, and the rules changed, making the sport more entertaining (Naismith, 1996). Soon professional leagues and teams were formed, and the game became very popular around the world. Lastly, in 1936 Basketball became an Olympic sport, and ten years later, on the 6th of June in New York, The National Basketball Association (NBA) was created (DiFiori et al., 2018).

Basketball is a sport that entertains people in many ways, from watching it and supporting their favorite team to betting on it with plenty of choices (Win, points). However, a new opportunity has come up for people to become coaches and choose their teams in recent years. Their options are evaluated and rewarded based on the Fantasy Points their players will achieve in each performance (Lees, 2002).

Statistic analytics in Basketball The study of basketball sports analytics in competition has been used to identify variables that can distinguish between successful teams and players, which can lead to better sport results. Indeed, this field of research has recently become a subject of practical and scientific interest to coaches and sport scientists. Globally, available research has shown that basketball winning teams outperform losing teams in shooting field goals and securing defensive rebounds (Ibáñez et al., 2008). However, in specific game contexts such as closely contested games, other game-related statistics such as fouls and free-throws exhibit greater importance (Kozar et al., 1994). In these studies, other game-related statistics such as offensive rebounds, turnovers, steals, and assists have not been reported as consistently as discriminators between winning and losing teams. This suggests that winning teams' performances are based on the quality of player decision making and field-goal efficiency and efficacy within a well defined strategic and tactical team environment (Trninić et al., 2002). Concurrently, the defensive rebounds represent a team's ability to recover the ball after an opponent's missed shots. A successful defensive rebounding team has more opportunities to attempt

field goals, score points, and win games (Sampaio and Janeira, 2003). The analytics results would allow more specific strategic and tactical team preparation, from player recruitment to practice planning, execution, and control.

Nowadays, Basketball is a sport that is highly dependent on statistics and analytics. Teams use data for decision-making about players, playstyles, and game strategies. However, Sports Analytics is something new compared with other sports. The analytics in basketball happened when Lawrence Dean Oliver, an American statistician, published his book "Basketball on Paper" (Oliver, 2004). In 2004, Oliver introduced sports analytics in the Basketball world, while in the same year, and became the first full-time statistical analyst in the NBA. His carrier as a sports analyst starts from the Seattle SuperSonics team, and since then, he has worked with Denver Nuggets on NBA, with ESPN, Sacramento King, and as an assistant coach to Washington Wizards. A year after the publicity of "Basketball on Paper", the famous SportVU was created. GalOz and Miky Tamir from Israel develop a real-time optical tracking system that identifies the movements of every player on the pitch (Apostolou and Tjortjis, 2019; Sarlis and Tjortjis, 2020).

SportVU SportVU's introduction to the NBA began in the 2009-2010 season when Kopp convinced the Dallas Mavericks, Houston Rockets, Oklahoma City Thunder and San Antonio Spurs to purchase the technology. In the following season the Boston Celtics and Golden State Warriors also implemented this technology (Glockner, 2016). In the 2010-2011 season, four NBA teams started using SportVU, by the 2011-2012 season the Milwaukee Bucks, Minnesota Timberwolves, New York Knicks, Toronto Raptors, and Washington Wizards joined the mix. Lastly, since the 2013-2014 NBA season, all NBA arenas have installed the SportVU camera system, and their teams benefit from advanced statistics. SportVU offers plenty of innovative statistics based on speed,distance, player sep-

aration, and ball possession, and teams can benefit from their analysis with specific algorithms (Stephanos, 2021).

This move toward analytics has the potential to change the way fans and professionals look at games. The deal the NBA made allows fans to have access to the new tracking statistics. Fans and media will have access to the stats and the hope is that it will assist in a better understanding of why teams make the decisions they do that moves beyond the basic stats that have traditionally been shown to fans. Now the media and fans can see the statistics that the organizations do and can enjoy them beyond the traditional box score that has been the standard for so long (Rosenkrans, 2016).

Hot Hand Over three decades ago, Gilovich et al. (Gilovich et al., 1985) caused a stir by contradicting the generally held idea that a hitter's tendency to produce an increasing number of hits was true. There was no proof of a beneficial association between the results of successive shots in the data analysis. This study's main goal was not to evaluate basketball players' performances, but rather to show a widespread prejudice brought on by the representativeness heuristic. The Philadelphia 76ers' statistician gathered the information from records of 48 of the team's home games. According to the authors, records of a player's consecutive shots made during basketball games for other NBA clubs at the time were not available. Gilovich et al (Gilovich et al., 1985). hypothesized that the absence of evidence of streak shooting might also be explained by player decision-making and the defensive tactics employed by the opposing teams. However, they also came to the conclusion that there is no relationship between the results of successive attempts after looking at the Boston Celtics' free-throw shooting history and conducting a controlled shooting experiment with the male and female varsity teams at Cornell. The study by Gilovich et al (Gilovich et al., 1985). is an example of research that

uses information from the sports industry. It is one of the most noteworthy examples of how sports have served as testing grounds for significant psychological and economic theories. The literature on the hot hand is currently fairly extensive; for instance, this study has been mentioned more than 1300 times on Google Scholar (Huang and Li, 2021).

Data Analysis in NBA As one of the world's most significant large-scale basketball events, the NBA has a long-term data record and a long tradition of data analysis. This thesis focuses on the dataset from the NBA, the scope here is to produce character for the performance of any team participating in the NBA for the past 10 years. Basketball is selected because it offers plenty of statistics for players and teams (Koudoumas, 2021). Also, it can be considered a challenging domain of analysis and prediction-making because data should be appropriately selected and used, focusing on the target variable. The NBA league's understanding and application of basketball data analysis are not only reflected in the level of the game and the team but also lead to the development of the entire basketball game to a certain extent. Analyzing the evolution of game characteristics throughout the season can help coaches develop an informed cycle plan to maintain technical output for the next season (Morgulev et al., 2018).

NBA teams often focus on technical and tactical parameters throughout the season. Game-related technical stats include offensive and defensive rebounding, free throws, two- and three-point field goal percentages, passing skills, and turnovers. The passing skills and defense of players and teams in basketball are critical and support performance throughout the season. In addition, ball reversal, indirect screen, dribble breakthrough, and ball screen are the game's most commonly used offensive tactics. Different teams have different technical tactics and different performance styles (Huang and Li, 2021). This study needs to ex-

plore the time-varying patterns of the performance profiles of each team's multiple performance indicators, which fall under the dissimilarity and similarity analysis of the teams' non-linear behavioral patterns. This illustrates the game's evolution throughout the game and can provide sports managers with objective support around team adjustments (Huyghe et al., 2022).

1.3 Dimension Reduction

Over the last few decades, advances in data gathering and storage capacities have resulted in an information overload in most fields. Larger and larger observations and simulations are encountered daily by researchers working in fields as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions. These datasets provide novel difficulties in data analysis, as opposed to smaller, more typical datasets that have been researched extensively in the past. Traditional statistical techniques become ineffective as the number of observations rises, but mostly when the number of variables linked to each observation rises. The number of variables that are measured for each observation is the data's dimension.

High-dimensional datasets bring both opportunities and challenges in mathematics and are certain to lead to fresh theoretical ideas . One issue with high-dimensional datasets is that not all measured variables are always "essential" for comprehending the relevant underlying phenomena. Even while high-dimensional data can be used to create prediction models with high accuracy using some computationally intensive innovative methods, it is still desirable in many applications to lower the original data's dimension first (Wang et al., 2015).

Real world data, such as digital photographs usually has a high dimensionality. In order to handle real-world data adequately, its dimensionality needs to be reduced. Dimensionality reduction is the transformation of high-

dimensional data into a meaningful representation of reduced dimensionality (van der Maaten et al., 2007). Ideally, the reduced representation should have a dimensionality that corresponds to the intrinsic dimensionality of the data. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data (Cunningham, 2008). In addition to increasing processing efficiency, dimension reduction can also increase the accuracy of the analysis.

The main idea with dimension reduction is to reduce complexity of a data space to create a lower dimensional representation of that original space. Such a task makes data more accessible, patterns more intuitive, and thereby eases to task of detecting and interpreting natural structure (Waggoner, 2021). It maps the data to a lower dimensional space such that uninformative variance in the data is discarded, or such that a subspace in which the data lives is detected. Dimension reduction has a long history as a method for data visualization, and for extracting key low dimensional features (Burges et al., 2010). Apart from teaching us about the data, dimension reduction can also lead us to better models for inference. The need for dimension reduction also arises for other pressing reasons.

Since the mid-20th century, the overload of available data has stimulated a soar in the development of dimension reduction methods with inherent differences (Lee and Verleysen, 2007). Most of the methods fall into those aiming at distance and topology preservation. The distance-preserving methods have their basis in those preserving spatial distances in the dataset (e.g. the family of multidimensional scaling (MDS) methods), while topology-preserving methods have their basis in those preserving neighborhood relations in the dataset (Kohonen, 1982). Most differences in the quality of dimension reductions, as all structural information can impossibly be preserved in a lower dimension, derive from variations in pre-

served similarity relations, such as pairwise distances or topological relationships (Sarlin, 2015).

Nowadays some statisticians sometimes talk of problems that are "Big p Small n"; (Bernardo et al., 2003) these are extreme examples of situations where dimension reduction is necessary because the number of explanatory variables p exceeds (sometimes greatly exceeds) the number of samples n. From a statistical point of view it is desirable that the number of examples in the training set should significantly exceed the number of features used to describe those examples (Cunningham, 2008).

1.4 Structure

The rest the thesis is organized as four chapters: Chapter 2 provides the process of data achieving and preprocessing. Chapter 3 chapter is about the methods used to analysis the data. We analyzed the procedure of using the NMDS method in this chapter. Chapter 4, we apply autoencoder to

List of References

Apostolou, K. and Tjortjis, C. (2019). Sports analytics algorithms for performance prediction. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4. IEEE.

Beck, D. and Bosshart, L. (2003). Sports and media. *Communication Research Trends*, 22(4).

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742.

Burges, C. J. et al. (2010). Dimension reduction: A guided tour. *Foundations and Trends® in Machine Learning*, 2(4):275–365.

Castellano, J., Perea, A., Alday, L., and Hernández Mendo, A. (2008). The measuring and observation tool in sports. *Behavior Research Methods*, 40(3):898–905.

Cunningham, P. (2008). Dimension reduction. In *Machine learning techniques for multimedia*, pages 91–112. Springer.

DiFiori, J. P., Güllich, A., Brenner, J. S., Côté, J., Hainline, B., Ryan, E., and Malina, R. M. (2018). The nba and youth basketball: recommendations for promoting a healthy and positive experience. *Sports Medicine*, 48(9):2053–2065.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314.

Glockner, A. (2016). *Chasing perfection: a behind-the-scenes look at the high-stakes game of creating an NBA champion*. Da Capo Press.

Grissmer, J. (2008). *The Perfect Game: Jim Naismith Invents Basketball*. AuthorHouse.

Huang, M.-L. and Li, Y.-Z. (2021). Use of machine learning and deep learning to predict the outcomes of major league baseball matches. *Applied Sciences*, 11(10):4499.

Huyghe, T., Alcaraz, P. E., Calleja-González, J., and Bird, S. P. (2022). The underpinning factors of nba game-play performance: A systematic review (2001–2020). *The Physician and Sportsmedicine*, 50(2):94–122.

Ibáñez, S. J., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M. A., and Ortega, E. (2008). Basketball game-related statistics that discriminate between teams' season-long success. *European journal of sport science*, 8(6):369–372.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69.

Koudoumas, P. (2021). *Sports Analytics Algorithms for Performance Prediction*. PhD thesis. Accessed: December 10, 2022.

Kozar, B., Vaughn, R. E., Whitfield, K. E., Lord, R. H., and Dye, B. (1994). Importance of free-throws at various stages of basketball games. *Perceptual and Motor skills*, 78(1):243–248.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*, volume 1. Springer.

Leeds, M. A., Von Allmen, P., and Matheson, V. A. (2018). *The economics of sports*. Routledge.

Lees, A. (2002). Technique analysis in sports: a critical review. *Journal of sports sciences*, 20(10):813–828.

McCorey, J. M. (2021). *Forecasting Most Valuable Players of the National Basketball Association*. PhD thesis, The University of North Carolina at Charlotte.

Morgulev, E., Azar, O. H., and Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5(4):213–222.

Naismith, J. (1996). *Basketball: Its origin and development*. U of Nebraska Press.

O'donoghue, P. (2009). *Research methods for sports performance analysis*. Routledge.

Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc.

Rains, R. (2011). *James Naismith: The man who invented basketball*. Temple University Press.

Rosenkrans, S. (2016). Fragile nature of competitive advantage in an analytic driven market: Sportvu in the nba.

Sampaio, J. and Janeira, M. (2003). Statistical analyses of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions. *International Journal of Performance Analysis in Sport*, 3(1):40–49.

Sarlin, P. (2015). Data and dimension reduction for visual financial performance analysis. *Information Visualization*, 14(2):148–167.

Sarlis, V. and Tjortjis, C. (2020). Sports analytics—evaluation of basketball players and team performance. *Information Systems*, 93:101562.

Song, K., Zou, Q., and Shi, J. (2020). Modelling the scores and performance statistics of nba basketball games. *Communications in Statistics-Simulation and Computation*, 49(10):2604–2616.

Stephanos, D. (2021). Machine learning approaches to dribble hand-off action classification with sportvu nba player coordinate data.

Trninić, S., Dizdar, D., and Lukšić, E. (2002). Differences between winning and defeated top quality basketball teams in final tournaments of european club championship. *Collegium antropologicum*, 26(2):521–531.

Turner, A., Brazier, J., Bishop, C., Chavda, S., Cree, J., and Read, P. (2015). Data analysis for strength and conditioning coaches: Using excel to analyze reliability, differences, and relationships. *Strength & Conditioning Journal*, 37(1):76–83.

van der Maaten, L., Postma, E., and Herik, H. (2007). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research - JMLR*, 10.

Waggoner, P. D. (2021). Modern dimension reduction. *CoRR*, abs/2103.06885.

Wang, Y., Yao, H., Zhao, S., and Zheng, Y. (2015). Dimensionality reduction strategy based on auto-encoder. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, pages 1–4.

CHAPTER 2

Data Achieving and Prepossessing

2.1 Data Achieving

Our analysis relies on a wide range of publicly available NBA data. We used the web-scrape method to achieve data from the NBA's official website [1] and the reference website [2], which concludes basketball stats, history statistics, scores, and history for the NBA.

To achieve the data, we collect the game IDs for the regular season in the NBA season 2012 to season 2022, then we collect the team game logs per season and player game logs per season. Then we matched each game with the corresponding teams and players. After that, we have our initial data, which includes features such as team names, player positions, and related variables in a total of 24 variables. The processes are using the two useful libraries, BasketballAnalyzeR (Sandri et al., 2020) and nbastatR (Bresler, 2022).

The two libraries of BasketballAnalyzeR (Sandri et al., 2020) and nbastatR (Bresler, 2022) in the R language is mainly used in NBA analysis. nbastatR is an interface for professional basketball data in R. Data sources include, but are not limited to: NBA Stats API, Basketball Insiders, and Basketball-Reference. BasketballAnalyzeR is very professional as well. Its authors, Paola Zuccolotto and Marica Manisera have worked in sports data analysis for many years to advance sports statistics. They created the Big Data Analytics in Sports project [3] based on the Big and Open Data Innovation Lab (BODaI-Lab, https://bodai.unibs.it) at the University of Brescia, Italy. , which provides data analysis for multiple sports. BasketballAnalyzeR is one of the results of this project. BasketballAna-

[1] `https://www.nba.com/`

[2] `https://www.basketball-reference.com`

[3] `https://bodai.unibs.it/bdsports/`

lyzeR provides quite rich analysis methods, including not only descriptive analysis of common technical statistical data but also advanced data mining methods and statistical models to explore data patterns and connections. At the same time, BasketballAnalyzeR also integrates multiple R visualization tools based on ggplot2, providing rich and beautiful data display.

2.2 Data Preprocessing

Basketball is a sport with plenty of statistics. For this reason, several datasets with different types of statistics are needed. After achieving the data from the website, we need to do data preprocessing here.

First, we removed several unnecessary variables, for example, we removed variable position, because our analysis was based on the fact that these positions were not descriptive. While we had to clean and transform each dataset from unnecessary statistics that gave us no further information about the performance of the team, each type of dataset for players and teams had to be merged to proceed to characterization related to NBA points for each team. This gave us several sets of data for advanced metrics, per-possession metrics, and per-minute metrics.

Then we continue to process our data, testing for null values, duplicates, and noise. We checked our data and wanted to make sure there is no null values in each column. here are several missing data exist and for those several missing values, we manually added them based on the information from the reference website [4]. Then we worked on name consistency,it isn't uncommon anymore for professional sports teams to change their respective names. Some teams change their location yet stick with the name while others go for the opposite approach if a team changes its name, we take into account the totals after combining their data. For example, the New Orleans Pelicans is an American professional basketball team

[4] `https://www.basketball-reference.com`

based in New Orleans (Wikipedia, 2022). The Pelicans compete in the National Basketball Association (NBA) as a member of the league's Western Conference Southwest Division and play their home games at the Smoothie King Center. The Pelicans were established as the New Orleans Hornets in the NBA season 2002 to season 2003 when George Shinn, then owner of the Charlotte Hornets, relocated the franchise to New Orleans. Due to the damage caused by Hurricane Katrina in 2005, the team temporarily relocated to Oklahoma City, where they spent two seasons as the New Orleans/Oklahoma City Hornets before returning to New Orleans for the NBA season 2007 to season 2008. In the season 2013, the Hornets announced that they would change their name to the New Orleans Pelicans after the NBA season 2012 to 2013. In our analysis, we combined the data of the New Orleans Hornets and New Orleans Pelicans together and conclude its data under the name New Orleans Pelicans. We did the same process with the teams have similar situations as well.

The next phase was important, we focused on feature engineering and extraction because each dataset row should have historical data. First, the transformation of each dataset was necessary for merging Players', Opponents', and Teams' data into one single dataset. As we have got the game IDs and game log data by using the nbastatR for our analysis. Game IDs are unique IDs for each NBA game and are common in nearly every dataset available. Match log data refers to rows that contain player or team stats for each game of the season. The seasons chosen here are from the season 2012 to 2022. The obtained game IDs for the Regular Season and Playoffs are combined into all IDs. Then we retrieved game log data and created team data. We used the game log retrieved in the previous step and then created box scores for the entire season, seperated by the regular season and playoffs. Several datasets of box scores are created here, and the process followed

for each dataset is as follows:

(1) Select a game log table;

(2) Group by a target grouping variable, such as "season" and "team";

(3) A summary table calculates interesting competition indicators according to groups.

Finally, we have the details of the team tables, which separated by the regular season and playoffs. The team stats have the instances (rows) that are the teams analyzed, and the variables (columns) are the team's game results under consideration. We named it teamBox and it contains a set of stats for each team in the game.

Now we have the total number of games, the playing time of teams for each game, and other selected variables as the initial variables for our statistical analysis. Then we divided each variable in the matrix by the time of each game so that we can achieve the best effective value.

After all these steps, we have our statistical data of all 30 NBA teams containing 22 variables from the season 2012 to season 2022, and through dimension reduction, several indicators can be used to measure the comprehensive strength of a team. The data dictionaries for the datasets we used are in Table A.1.

List of References

Bresler, A. (2022). *nbastatR: R's interface to NBA data*. R package version 0.1.151.

Sandri, M., Zuccolotto, P., and Manisera, M. (2020). *BasketballAnalyzeR: Analysis and Visualization of Basketball Data*. R package version 0.5.0.

Wikipedia (2022). New Orleans Pelicans — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=New%20Orleans%20Pelicans&oldid=1121797579. [Online; accessed 21-November-2022].

CHAPTER 3

Data Analysis with MDS

3.1 Introduction to Multidimensional scaling

Similarity is a pervasive concept, one that is central to understanding and engaging in the world around us. The concept of similarity is widely used in almost every scientific field. A sense of 'sameness' between two things allows us to appropriately generalize and discriminate. Generally speaking, similarity is regarded as a numerical value which represents the equivalence between two objects, variables, items, or sets. Measures of similarity provide a numerical value which indicates the strength of associations between objects or variables. In data science, the similarity measure is a way of measuring how data samples are related or closed to each other.

Multidimensional scaling (MDS) (Torgerson, 1958; Kruskal, 1956) is a dimension-reduction treatment to discover the underlying structure of distance measures between objects or cases. MDS is a technique for condensing many variables into a small number of dimensions that retain the majority of the original variables' information. MDS finds the spatial map for objects given that the similarity or dissimilarity information between the objects is available (Hout et al., 2013). The goal is to place observations in a space based on similarity scores between them. This seems a little like eigen decomposition, Principal Components Analysis (PCA) or factor analysis

MDS does not depend on most common assumptions like linearity and normality which makes MDS is preferred over factor analysis (Borg and Groenen, 2005). The only assumption of MDS is that the number of dimensions should be one less than the number of points, which also means at least three variables must be entered in the model and at least two dimensions must be specified. Once the

proximities (dissimilarity information) are available, MDS can provide the lower dimension solution.

3.1.1 Classical Multidimensional scaling

Classical MDS (CMDS) (Torgerson, 1952) also known as Principal Coordinates Analysis (PCoA), Torgerson Scaling or Torgerson–Gower scaling, is a statistical technique originating in psychometrics. The data used for multidimensional scaling (MDS) are dissimilarities between pairs of objects.

In PCA, we start with n data points $\mathbf{x}_i \in R^p$, and then try to find a low dimensional projection of these points, e.g., $\mathbf{y}_1, \ldots, \mathbf{y}_n \in R^r$ with $r < p$, in such a way that they minimize the reconstruction error (or maximize the variance).

The focus of MDS is somewhat different. Instead of being given the data $\mathbf{X}$, our starting point is often a matrix of distances or dissimilarities between the data points, $\mathbf{D}$ (Borg et al., 2012). For example, if we have data on n different experimental units, then we would be given the distances d_{ij} between any pair of experimental units i and j. We compile these into a $n \times n$ distance matrix $\mathbf{D} = (d_{ij} : i, j = 1, \ldots, n)$.

The goal of MDS is to find a set of points in a low-dimensional Euclidean space R^r, usually R^2 or R^3, whose inter-point distances (or dissimilarities) are as close as possible to the d_{ij}. That is, we want to find $\mathbf{y}_1, \ldots, \mathbf{y}_n \in R^r$ whose distance matrix is approximately $\mathbf{D}$, i.e., for which

$$\text{distance}(\mathbf{y}_i, \mathbf{y}_j) \approx d_{ij}.$$

In other words, we are trying to create a spatial representation of the data, $\mathbf{y}_1, \ldots, \mathbf{y}_n$, from a distance matrix $\mathbf{D}$. The vectors $\mathbf{y}_i$ have no meaning by themselves, but by visualizing their spatial pattern we can hope to learn something about the dataset represented by $\mathbf{D}$. If we define the errors in terms of a square distance, then we can write the goal of MDS as the following optimization problem:

$$\text{Find} \quad \mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbf{R}^r$$
$$\text{to minimize} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij} - d(\mathbf{y}_i, \mathbf{y}_j))^2.$$

3.2 Results of MDS

Since there are 14 variables in the data set, in order to make the subsequent analysis more convenient, we need to condense the 14 indicators and simplify them, so we used the MDS analysis. It is used to extract important information from multivariate data tables, to descale the variable dimensions, and to represent the processed information as a new set of variables in each dimension.

Here we generated a scree plot in Figure 1 to represent the variance explained by each dimension and help us to determine how many dimensions need to be selected. In Figure 1, we find the first 4 main dimensions already account for about 80 percent of the variance, so we chose to keep the first four dimensions and these four dimensions will be important in our analysis. According to the MDS load diagram, the cumulative contribution rate of the first six dimensions exceeds 80 percent, which meets the analysis requirements. The figure of percentage of explained variances in each dimension is list below.

Here we created the graphs of each dimension to visualize and explore the contribution of variables. In Figure 2, we also created barplots to show the top ten contributed variables for each dimension. We also added a red reference dashed line to indicate the expected value if the contribution where uniform. In Figure 3, we created a table to show the correlation between each contributed variables for each dimension. We also used different color to indicate the correlation level. With the help of these figures, we can check the contributions for each dimension and interpret its meaning.

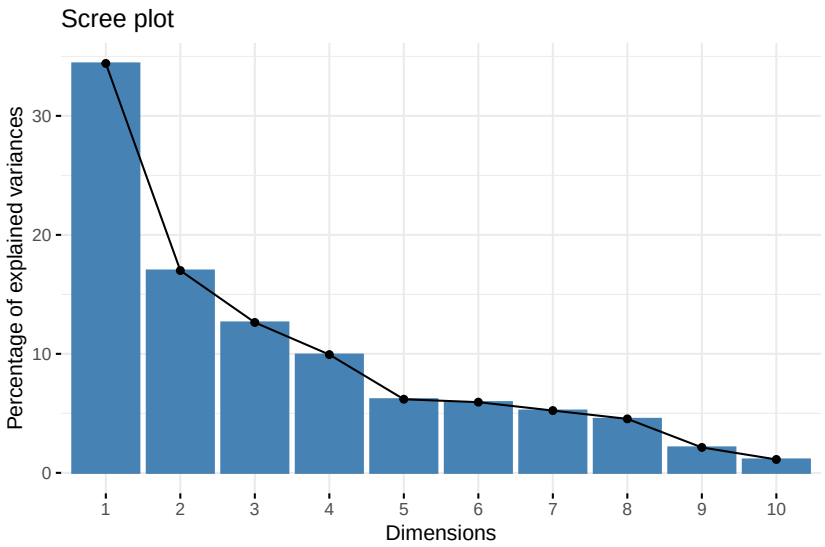

Figure 1. Scree plot for mds.

Analysis in Dimension1

we know that over 30 percent of variances is explained in the first scree from figure 1. Here we showed the contribution of variables to dimension 1. There are 6 variables more than 7 percent, which are P3M, P3A, PTS, P2A, DREB, and AST. The overall high level of P3A and P2A indicates that players are good at seizing opportunities to score. The three-point shot is a very important technical tactic, which can bring a lot of changes to the team's playing style.

The master of the three-point shot can greatly create individual scoring chances and increase the difficulty of the opponent's defense in the regular game. In this dimension, P3M contributes the most. P3A weighs more than P2A. DREB is a backcourt rebound. A backcourt rebound is a defensive rebound, that is, the rebound that the defensive team grabs after the offensive team does not score. Rebounding is an important means of conversion between offense and defense. It is an important way of controlling the ball and it has a direct impact on the outcome of the game. In dimension 1, the important variables are highly correlated to the overall action of scoring in a basketball game, it reflects the trend that the teams

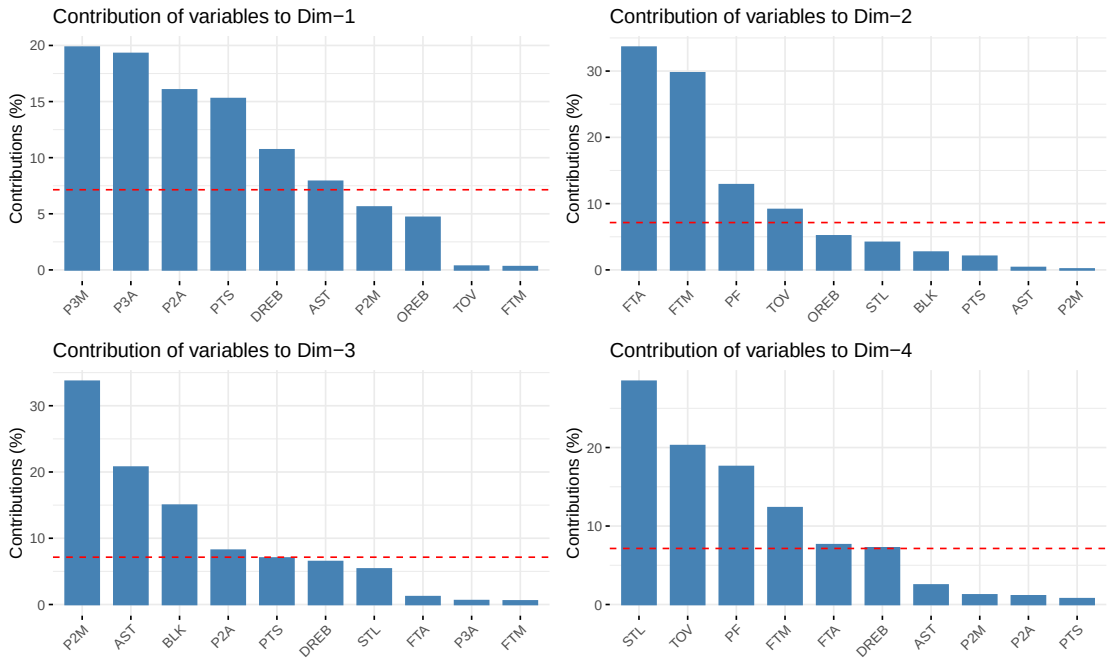

Figure 2. Contributed variables.

are focusing on the three-point shot and two-point shot, with higher points made as well.

Analysis in Dimension2

In dimension 2, we showed the contribution of variables to dimension 2, 4 variables contribute more than 7 percent, which is FTA, FTM, PF, and TOV. FTA is the number of free throws attempted and FTM is the number of free throws made. The free throw is a penalty imposed on the offending team after a foul is committed in a basketball game. The higher the number of free throws, the more free throws is executed. PF is the number of personal fouls a player commits on the basketball court. The more personal fouls a team receives, the more it is penalized, and the more it is at a disadvantage in the game. TOV is the number of turnovers. The more turnovers you make, the less likely you are to score, and the more chances the other team will score. The variables that make much contribution in dim 2 are more close to the error made by the players in the

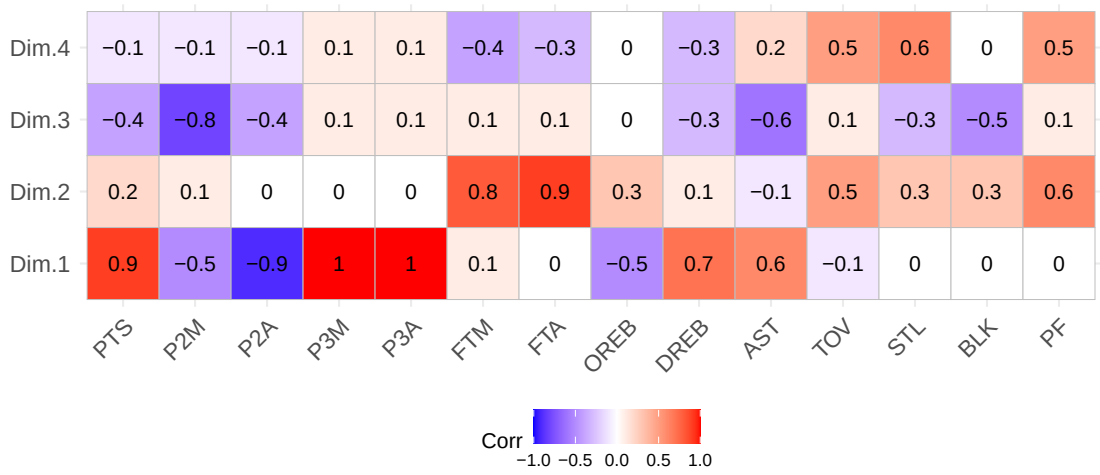

Figure 3. Correlation between contributed variables.

basketball game, including the mistakes and missteps. These are the points to be improved. It reflects the trend that players always made fouls while in the act of shooting, and attention to detail will take the team further.

Analysis in Dimension3

In dimension 3, 4 variables contribute more than 7 percent, which are P2M, PM, BLK, AST, and P2A. P2M is the number of two-pointers made. BLK, short for Block Shot, records the number of shots a player blocks on the basketball court. Many teams with a "block" master, will be in the expected location to receive a lot of ball from the blocker, because of the catch, from defense to attack can be launched immediately.P2A is the number of 2-point shots. AST is the number of assists, which is the number of times a player assists a teammate on the basketball court. These two variables have been introduced and explained in the lines before. The important variables in dim 3 are the most common behaviors in a basketball game. These behaviors are common and basic but important. Focusing on the fundamentals of the game and cementing them will lead the team to victory, as

shown in the figure.

Analysis in Dimension4

In dimension 4 shown in the following, 6 variables contribute more than 7 percent, STL, TOV, PF, FTM, AST, and FTA. To be more specific, STL which refers to steals, the total number of times a team or individual player or otherwise obtains possession of the ball after an opponent turns the ball over legally, contributes most to the game, more than 20 percent. Following this, it is the TOV, turnovers, making the second largest contribution. In addition, in dim 4, the power forward is significant. What is more, Free Throws Made are crucial, which refers to all attempts and shots made at the free throw line in response to either personal or technical fouls. In other words, it shows the higher the FTM percentage it is, the more possible the team wins the game. Furthermore, the team should pay attention to AST and FTA as well to increase the possibility of victory, as shown in the figure.

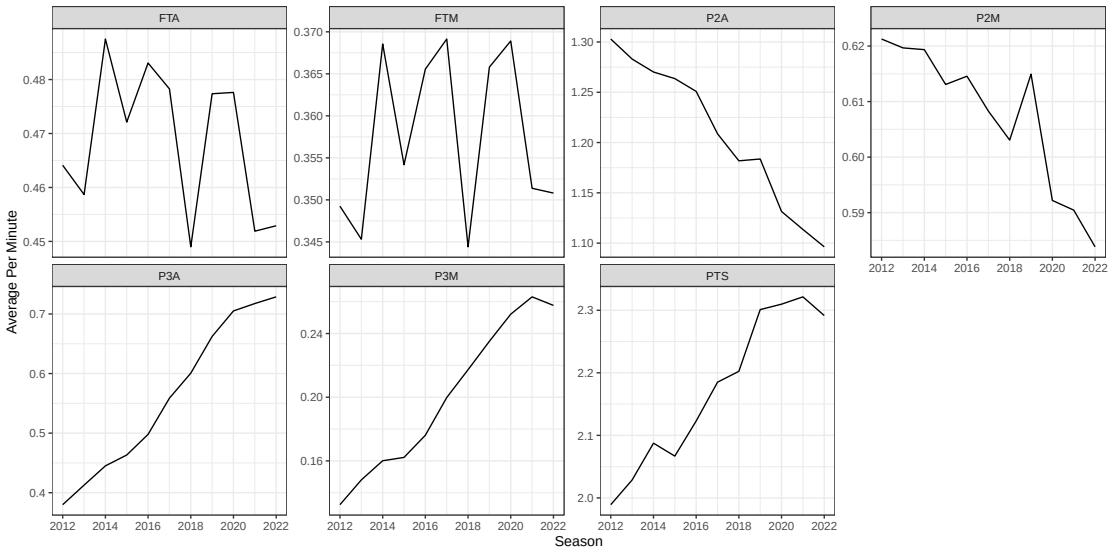

Figure 4. Average Points over teams for each season.

Score average for teams

Here we selected seven indicators that reflect the team's game for descriptive statistical analysis, which are FTA, FTM, P2A, P2M, P3A, P3M, and PTS. Figure 4 shows the average team points per minute in that category. Looking at the performance of each season between the decade of 2012 and 2022, the number of free throws made and free throw attempts have no significant upward or downward trend, but are oscillating slightly. And the overall number of free throws attempted in recent years is lower than the years at the beginning of the observation period, especially, there is a significant decrease in the number of free throw attempts in 2018. Personal free throws are considered to be one of the manifestations of player maturity, and the decrease in the number of free-throw attempts in recent years is also a reflection of the improvement of player maturity and the improvement of the overall tactical fluency of the team. And for all teams, the average metric of two point shooting performance showed a trend of decreasing two point shooting percentage, except for a significant standout in two-point shooting during 2019. But the number of three point shots attempted and made showed an upward trend over the decade. These results suggest that over the decade, players have progressively focused on the long-range aspect of their performance, with less attention paid to scoring performance from close range, and as a result, long-range shooting has had a greater impact on the dynamics of the game in recent years. The reason behind analyzing such a trend is twofold. On one hand, the presence of a strong three-point offense provides the opportunity to dish out breakdowns and is considered the key to a great offense. On the other hand, the option to shoot threes also creates space to operate in the post, forcing opposing guards to make the tough choice between the task at hand and helping others.

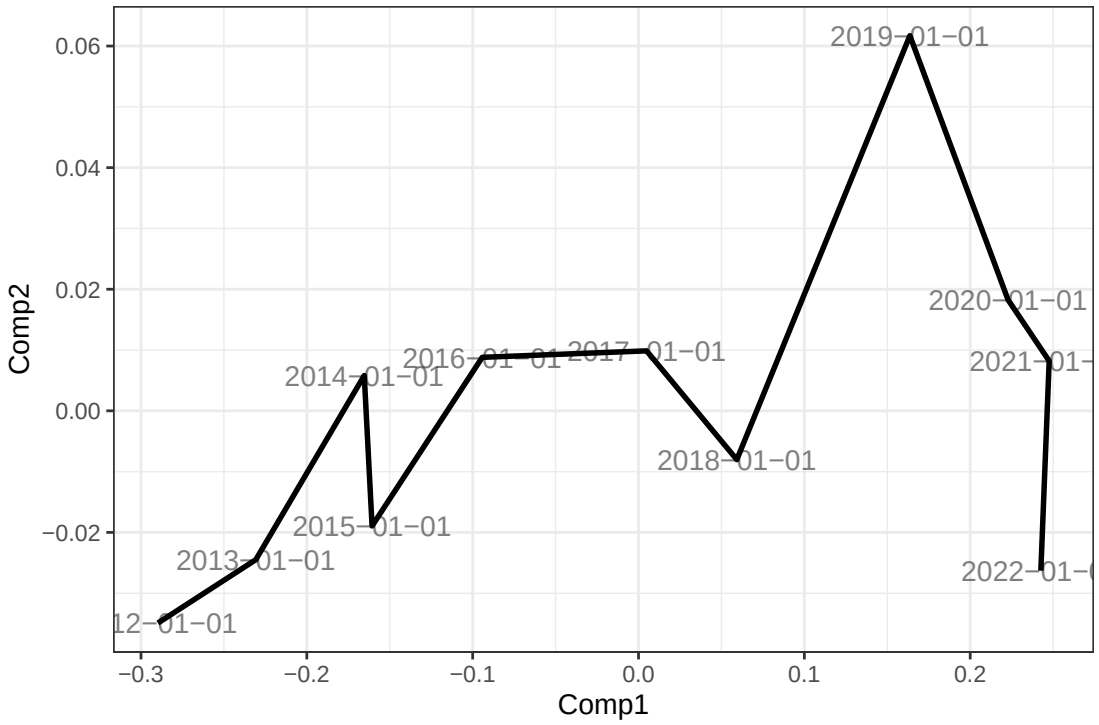

Figure 5. Score average for teams

Score for all teams

Figure 5 shows the scoring average for teams over the 2012-2022 decade in the two most critical dimensions. We explore the similarity of match characteristics across years. For dimension 1, there is a tendency for the scores also gradually to improve as the year advances. As for dimension 2, the scores are floating widely for all 10 years except for the season 2015 to 2021, in which 2020 to 2021 had a significantly higher score of 0.5 for dimension 2.

Since our purpose of this study was to explore the similarities and dissimilarities in the characteristics of the NBA game during the decade 2012 to 2022. Figure 6 focuses on the multivariate analysis based on dimension 1 and dimension 2 for the 25 teams. The multivariate analysis shows the similarity and dissimilarity of the individual team profiles during the observation period. It can be seen from the graph that each team has its own unique form of change, and the reason

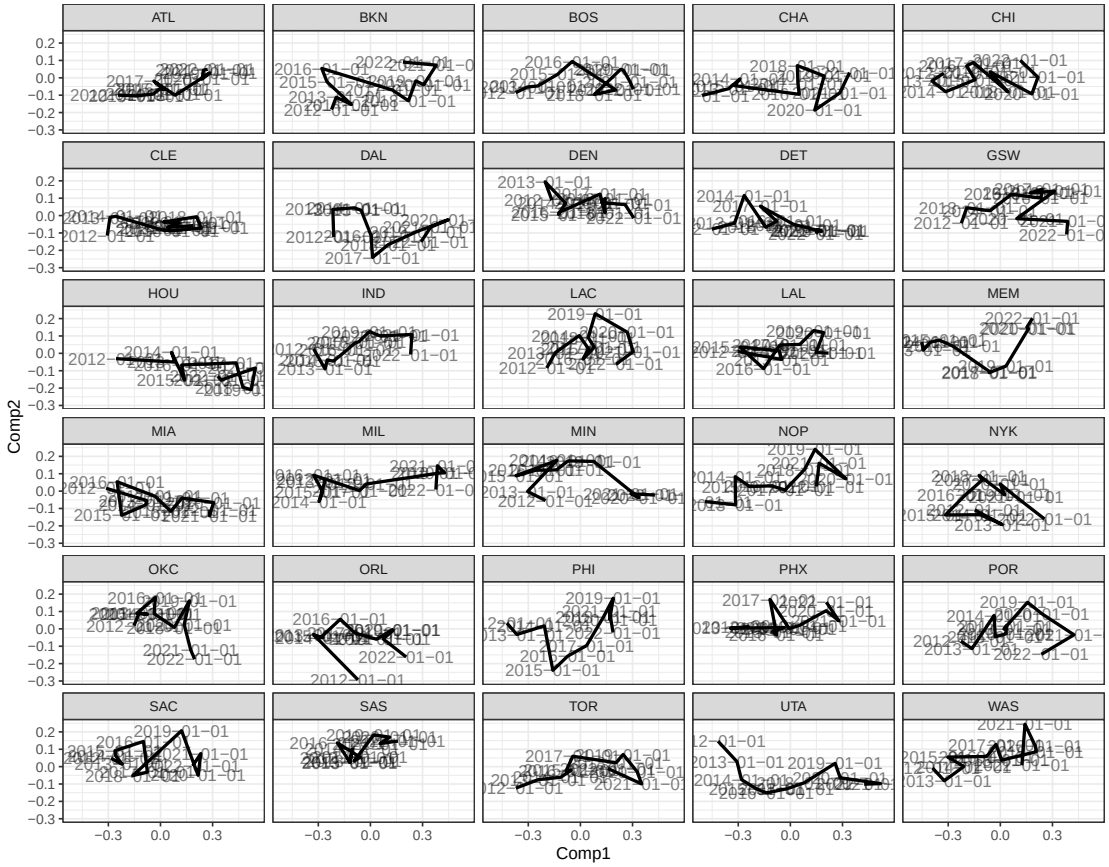

Figure 6. Score for all teams.

behind this may be that coaches and coaching staffs in each team have their own design and strategy for the game. And during the analysis period, changes in the personnel of the dominant team in the team may lead to the emergence of a new style of play and a unique path of change.

Using MDS, we characterized some game styles of the basketball game. Firstly, the game style of a basketball game can not be separated from the universal scoring behavior, including two point balls and three point shot, etc., which are the basic components of scoring. Secondly, we should pay attention to the details of the game, including free throws and mistakes, which really matter. Finally, we should pay attention to the basic ability of the team, including two-point goal, assist, etc.

3.2.1 MDS analysis of Golden State Warriors

In recent NBA history, there has been no team more dominant than the "Golden State Warriors". They won three championships in five years, during an incredible stretch from 2014 to 2019. Their style of play changed the league. According to the information from each season in the past 10 years of "Golden State Warriors", it is reflected in the two-dimensional space in the form of point and line connection. The distance between points determines the degree of difference between different seasons. In this way, the spatial positioning point map of the game is finally obtained as follows, as shown in figure 7.

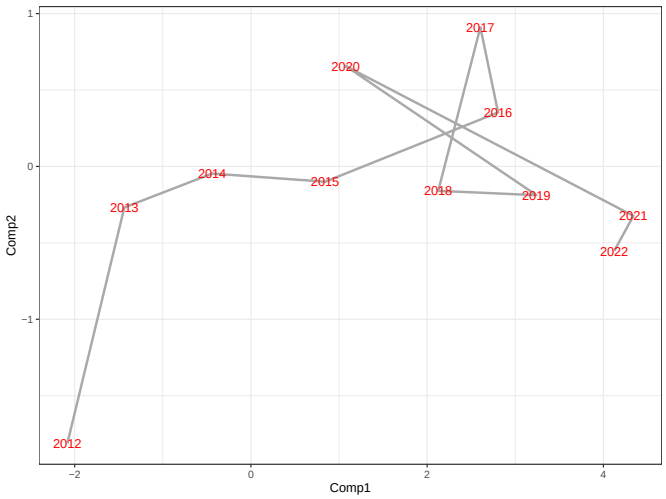

Figure 7. Multidimensional scaling plot for Golden State Warriors.

The graph shows that the "Golden State Warriors" had a dynamic road in these 10 years, which varies greatly over several years. Among them, the year-to-year differences are greatest from 2012 to 2013, 2015 to 2018, and 2019 to 2021, which have the largest point-to-point distance in two-dimensional space. The lesser differences are the technological route changes from 2013 to 2014, and 2018 to 2019, and the technological route has a dramatic change from 2015 to 2021.

The underlying idea behind the Golden States Warriors' play style is complete trust in data and avoiding emotional influence in their decision-making. As Joe Lacob planned to purchase the team, most issues could be boiled down to questions of data and probability. He stated in an interview that he "tore everything down and put it back up" when he took over the Warriors. That comes with the most significant change in Golden State Warriors. For a period of time starting in the 2013 season, the NBA was saturated with Warriors content, whether it was talking over their three-point happy playing style or their analytics-driven organizational philosophy.

With a continuously rising pace, the Warriors had increasing total points. Not like the rumor, the Warriors is not a three-point-shooting team, though they led the league in three points shots since the 2014 season and they have the best shooter of all time and the two great shooters in the league. Start from the 2015 season, the gap between the Warriors and the next team is greater than the gap between second place and the average non-Warriors squad. This is because of Curry and Thompson, two of the best shooters of all time are also great screen-setters. Their dominance continued in 2018 but due to some reasons including injury problems, in the 2019 season, they lose the final to the Raptors. Those lead to the dramatic change in figure 7.

List of References

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Borg, I., Groenen, P. J., and Mair, P. (2012). *Applied multidimensional scaling*. Springer Science & Business Media.

Hout, M. C., Papesh, M. H., and Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):93–103.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

Torgerson, W. S. (1958). Theory and methods of scaling.

CHAPTER 4

Data Analysis with Autoencoder

4.1 Introduction to Autoencoder

Autoencoders (Yann, 1987) is a neural network that to learn efficient representations of the input data. It learns how to efficiently compress and encode data then learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible. Autoencoder, by design, reduces data dimensions by learning how to ignore the noise in the data.

An autoencoder has a structure very similar to a feed-forward neural network which is also known as multi-layer perceptron (MLP). The primary difference when using in an unsupervised context is that the number of neurons in the output layer are equal to the number of inputs. Consequently, in its simplest form, an autoencoder is using hidden layers to try to reconstruct the inputs.

As shown in Figure 8, an autoencoder is a symmetrical neural network that may unsupervisedly learn features (Shin et al., 2012). We can describe this network structure in two parts: (1) an encoder function $Z = f(X)$ that maps X inputs to Z codings; and (2) a decoder function ($X' = g(Z)$) that maps the hidden representation back to produces a reconstruction of the inputs (X), which is expected to be as close as possible to the original input of the encoder. An autoencoder attempts to learn an approximation in the hidden layer such that the input data can be perfectly reconstructed in the output layer.

Undercomplete autoencoder The purpose of an autoencoder is to encode important information efficiently and create a reduced set of codings that adequately represents X. A common approach to achieve that is by creating a bottleneck,

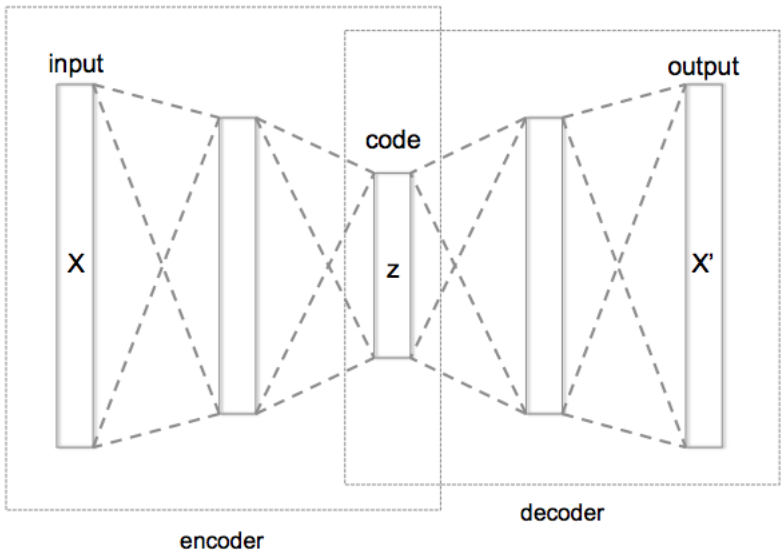

Figure 8. Schematic structure of an undercomplete autoencoder with three fully connected hidden layers .

which forces the model to preserve what's essential and discard unimportant bits. The bottleneck in the network forces a compressed knowledge representation of the original input. This compression of the hidden layers forces the autoencoder to capture the most dominant features of the input data and the representation of these signals are captured in the codings.

If the input features were each independent of one another, this compression and subsequent reconstruction would be a very difficult task. However, if some sort of structure exists in the data (ie. correlations between input features), this structure can be learned and consequently leveraged when forcing the input through the network's bottleneck.

$$\text{minimize} \quad \mathcal{L} = \|X, X'\|^2 \tag{1}$$

The reconstruction error in (1) measure how well the decoder is performing

and how close the output is to the original input. To learn the neuron weights and, thus the codings, the autoencoder seeks to minimize the reconstruction error (Irsoy and Alpaydın, 2017). The training then involves using back propagation in order to minimize the network's reconstruction loss.

By penalizing the network according to the reconstruction error, autoencoders can learn the most important attributes of the input data and how to best reconstruct the original input from an "encoded" state. Ideally, this encoding will learn and describe latent attributes of the input data.

Because neural networks are capable of learning nonlinear relationships, autoencoders can be thought of as a more powerful (nonlinear) generalization of PCA. Whereas PCA attempts to discover a lower dimensional hyperplane which describes the original data, autoencoders are capable of learning nonlinear manifolds (a manifold is defined in simple terms as a continuous, non-intersecting surface).

For higher dimensional data, autoencoders are capable of learning a complex representation of the data (manifold) which can be used to describe observations in a lower dimensionality and correspondingly decoded into the original input space.

4.2 Introductino to Sparse Autoencoder

A reliable autoencoder must make a trade-off between two important parts: (1) Sensitive enough to inputs so that it can accurately reconstruct input data; (2) Able to generalize well even when evaluated on unseen data. An undercomplete autoencoder has no explicit regularization term, the model is simply trained according to the reconstruction loss. This leads to a intrinsic problem of the autoencoder, it simply copying input layer to hidden layer. This intrinsic problem render undercomplete autoencoder ineffective to extract meaningful features even though its output can be a perfect recovery of the input data (Sun et al., 2016).

Given the fact that we'd like our model to discover latent attributes within our data, it's important to ensure that the autoencoder model is not simply learning an efficient way to memorize the training data. Similar to supervised learning problems, we can employ various forms of regularization to the network in order to encourage good generalization properties.

The Sparse Autoencoder (SAE), as an extension of the autoencoder, can learn relatively sparse features by introducing a sparse penalty term inspired by the sparse coding (Olshausen and Field, 1996) into the autoencoder. It can improve the performance of traditional autoencoder and exhibits more practical application values.

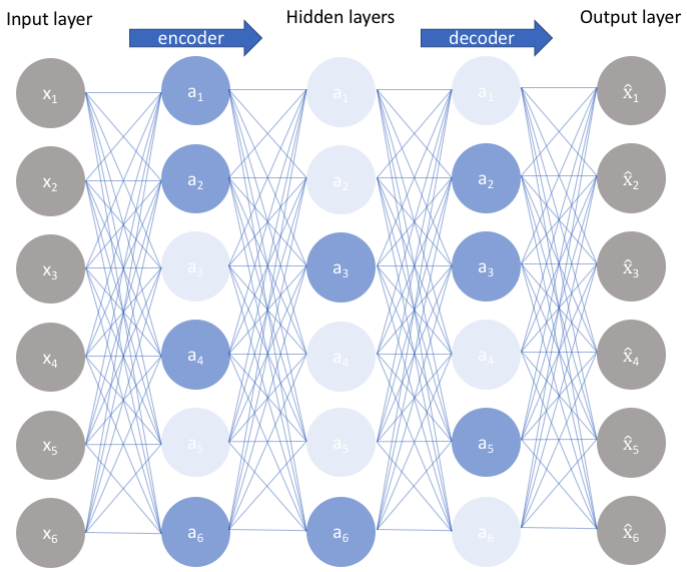

Figure 9. The Structure of Sparse Autoencoder.

SAE as shown is Figure 9 offer us an alternative method for introducing an information bottleneck without requiring a reduction in the number of nodes at

our hidden layers. Rather, we'll construct our loss function such that we penalize activations within a layer. For any given observation, we'll encourage our network to learn an encoding and decoding which only relies on activating a small number of neurons. It's worth noting that this is a different approach towards regularization, as we normally regularize the weights of a network, not the activations.

SAE are used to pull out the most influential feature representations. This is beneficial when trying to understand what are the most unique features of a data set. It's useful when using autoencoders as inputs to downstream supervised models as it helps to highlight the unique signals across the features. Incorporating sparsity forces more neurons to be inactive. This requires the autoencoder to represent each input as a combination of a smaller number of activations. In this project, we use SAE as our dimensional reduction tool.

4.3 Result of Autoencoder

The autoencoder classes are implemented in Tensorflow (Abadi et al., 2015), an open-source machine learning package written by Google Brain, with a dynamic graph-computation scheme that makes it very easy to implement unique structures like autoencoder.

To process our analysis, we have an encoder network that turns the input sample into a latent space. Then, a decoder network maps these latent space points back to the original input data. We map these sampled latent points back to reconstructed inputs. We train the model with a custom loss function, using Adam optimizer and mean squared error (Kingma and Ba, 2014).

Since our latent space is two-dimensional, some visualizations can be done at this point. Here we have a correlation table representing the relationship between contributed variables in each dimension. As shown in Figure 10, we used different colors to indicate the correlation level as well.

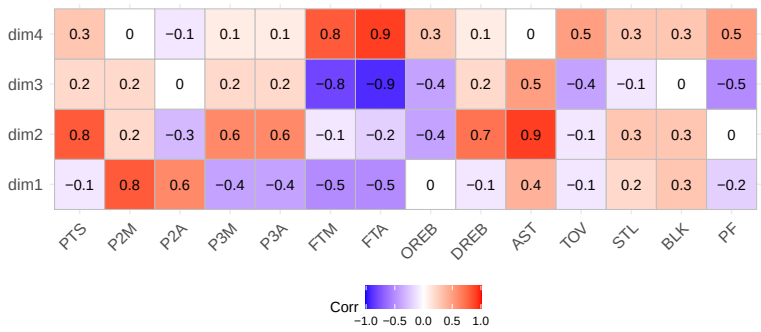

Figure 10. Correlation between contributed variables

The Figure 11 shows a multivariate matrix plot which explores the dissimilarity and similarity of the game characteristics of the NBA game from 2012 to 2022. It shows the evolution of NBA games from the past 10 years. During this decade, the technical tactics of many teams have varied a lot. It looks similar to the graph we made by MDS, those mainstream teams are performed in a similar way and the others are overlapped as well.

List of References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Irsoy, O. and Alpaydın, E. (2017). Unsupervised feature extraction with autoencoder trees. *Neurocomputing*, 258:63–73.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

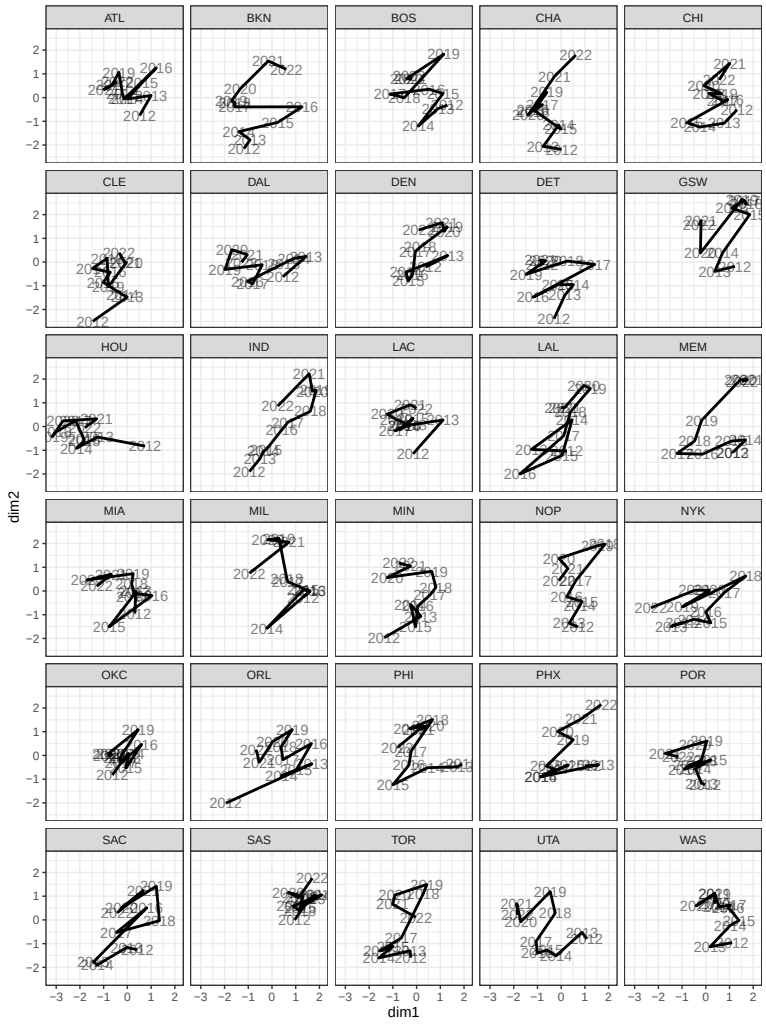

Figure 11. Autoencoder latent scores for each team over 2012-2022 seasons.

Shin, H.-C., Orton, M. R., Collins, D. J., Doran, S. J., and Leach, M. O. (2012). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1930–1943.

Sun, W., Shao, S., Zhao, R., Yan, R., Zhang, X., and Chen, X. (2016). A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement*, 89:171–178.

Yann, L. (1987). *Modeles connexionnistes de lapprentissage*. PhD thesis, These de Doctorat, Universite Paris.

CHAPTER 5

Conclusion

In this study, we investigate the use of space to fit a subspace of the 2012-22 NBA data space for each player and team. While the results are interesting, the progress through dimension reduction is also a good summary of the player and team character.

Free rows, two-point and three-point shooting, and rebounds are the significant characters of the best teams in the NBA today. Through the NBA, many different types and styles of play emanate over time. Some of them often define entire teams, like the SAN Antonio Spurs, known for their quick ball movement. We will describe the style or identity of the NBA teams. Popovich said not long ago that the style of play in the NBA is getting boring and every team is copying each other. The stats and charts prove that he was right about the style of offensive teams, even if the Houston Rockets and Golden State Warriors are different from the rest. Golden State is a red team, and even more special because, starting in 2015, the team has its own identity. The change is more dramatic, but for the same reasons as the orange team. Stephen Curry, Klay Thompson, and Kevin Durant, the last of the Warriors' three snipers since 2015, left the team, and the Splash brothers, Curry and Thompson were seriously injured after last year's playoffs. The team has changed so much that it hasn't played as well this year as it has in past years (D'Amour et al., 2015). Defensively, we found the fact that the Toronto Raptors head coach has made a lot of changes in the type of defense, whereas other teams use man-to-man defense most of the time. Also in this study we can find that steals and turnovers remained stable over time, in contrast to some common thoughts of winning teams in basketball tournaments.

Dimension reduction techniques are useful tools for demonstrating team evolution, combining statistical and analytical methods with the potential to provide greater modeling and classification/prediction of professional basketball performance.

List of References

D'Amour, A., Cervone, D., Bornn, L., and Goldsberry, K. (2015). Move or die: How ball movement creates open shots in the nba. In *Sloan Sports Analytics Conference*.

APPENDIX

Abbreviation

Table A.1. Variable Description.

Variables	Description	Type
Season	Season	numeric
Team	Analyzed team	character
Player	Analyzed player	character
GP	Games Played	numeric
MIN	Minutes Played	numeric
PTS	Points Made	numeric
W	Games won	numeric
L	Games lost	numeric
P2M	2-Point Field Goals (Made)	numeric
P2A	2-Point Field Goals (Attempted)	numeric
P2p	2-Point Field Goals (Percentage)	numeric
P3M	3-Point Field Goals (Made)	numeric
P3A	3-Point Field Goals (Attempted)	numeric
P3p	3-Point Field Goals (Percentage)	numeric
FTM	Free Throws (Made)	numeric
FTA	Free Throws (Attempted)	numeric
FTp	Free Throws (Percentage)	numeric
OREB	Offensive Rebounds	numeric
DREB	Defensive Rebounds	numeric
AST	Assists	numeric
TOV	Turnovers	numeric
STL	Steals	numeric
BLK	Blocks	numeric
PF	Personal Fouls	numeric
PM	Plus/Minus	numeric

Table A.2. Team Name Abbreviation

Team abbreviation	Team Full Name	character
ATL	Atlanta Hawks	character
BOS	Boston Celtics	character
CHA	Charlotte Hornets	character
CHI	Chicago Bulls	character
CLE	Cleveland Cavaliers	character
DAL	Dallas Mavericks	character
DEN	Denver Nuggets	character
DET	Detroit Pistons	character
GSW	Golden State Warriors	character
IND	Indiana Pacers	character
LAC	Los Angeles Clippers	character
LAL	Los Angeles Lakers	character
MEM	Memphis Grizzlies	character
MIA	Miami Heat	character
MIL	Milwaukee Bucks	character
MIN	Minnesota Timberwolves	character
NJN	Nets	character
NOP	New Orleans	character
NOH	New Orleans	character
NYK	New York Knicks	character
OKC	Oklahoma City Thunder	character
ORL	Orlando Magic	character
PHI	Philadelphia 76ers	character
PHX	Phoenix Suns	character
POR	Portland Trail Blazers	character
SAC	Sacramento Kings	character
SAS	San Antonio Spurs	character
TOR	Toronto Raptors	character
UTA	Utah Jazz	character
WAS	Washington Wizards	character
BKN	Brooklyn Nets	character

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

Apostolou, K. and Tjortjis, C., "Sports analytics algorithms for performance prediction," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2019, pp. 1–4.

Beck, D. and Bosshart, L., "Sports and media," *Communication Research Trends*, vol. 22, no. 4, 2003.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., "Bayesian factor regression models in the "large p, small n" paradigm," *Bayesian statistics*, vol. 7, pp. 733–742, 2003.

Borg, I. and Groenen, P. J., *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Borg, I., Groenen, P. J., and Mair, P., *Applied multidimensional scaling*. Springer Science & Business Media, 2012.

Bresler, A., *nbastatR: R's interface to NBA data*, 2022, r package version 0.1.151. [Online]. Available: https://github.com/abresler/nbastatR

Burges, C. J. *et al.*, "Dimension reduction: A guided tour," *Foundations and Trends® in Machine Learning*, vol. 2, no. 4, pp. 275–365, 2010.

Castellano, J., Perea, A., Alday, L., and Hernández Mendo, A., "The measuring and observation tool in sports," *Behavior Research Methods*, vol. 40, no. 3, pp. 898–905, 2008.

Cunningham, P., "Dimension reduction," in *Machine learning techniques for multimedia*. Springer, 2008, pp. 91–112.

DiFiori, J. P., Güllich, A., Brenner, J. S., Côté, J., Hainline, B., Ryan, E., and Malina, R. M., "The nba and youth basketball: recommendations for promoting a healthy and positive experience," *Sports Medicine*, vol. 48, no. 9, pp. 2053–2065, 2018.

D'Amour, A., Cervone, D., Bornn, L., and Goldsberry, K., "Move or die: How ball movement creates open shots in the nba," in *Sloan Sports Analytics Conference*, 2015.

Gilovich, T., Vallone, R., and Tversky, A., "The hot hand in basketball: On the misperception of random sequences," *Cognitive psychology*, vol. 17, no. 3, pp. 295–314, 1985.

Glockner, A., *Chasing perfection: a behind-the-scenes look at the high-stakes game of creating an NBA champion*. Da Capo Press, 2016.

Grissmer, J., *The Perfect Game: Jim Naismith Invents Basketball*. AuthorHouse, 2008.

Hout, M. C., Papesh, M. H., and Goldinger, S. D., "Multidimensional scaling," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 4, no. 1, pp. 93–103, 2013.

Huang, M.-L. and Li, Y.-Z., "Use of machine learning and deep learning to predict the outcomes of major league baseball matches," *Applied Sciences*, vol. 11, no. 10, p. 4499, 2021.

Huyghe, T., Alcaraz, P. E., Calleja-González, J., and Bird, S. P., "The underpinning factors of nba game-play performance: A systematic review (2001–2020)," *The Physician and Sportsmedicine*, vol. 50, no. 2, pp. 94–122, 2022.

Ibáñez, S. J., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M. A., and Ortega, E., "Basketball game-related statistics that discriminate between teams' season-long success," *European journal of sport science*, vol. 8, no. 6, pp. 369–372, 2008.

Irsoy, O. and Alpaydın, E., "Unsupervised feature extraction with autoencoder trees," *Neurocomputing*, vol. 258, pp. 63–73, 2017.

Karlis, D. and Ntzoufras, I., "Analysis of sports data by using bivariate poisson models," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 3, pp. 381–393, 2003.

Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Kohonen, T., "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

Koudoumas, P., "Sports analytics algorithms for performance prediction," Ph.D. dissertation, 2021, accessed: December 10, 2022. [Online]. Available: http://hdl.handle.net/11544/29775

Kozar, B., Vaughn, R. E., Whitfield, K. E., Lord, R. H., and Dye, B., "Importance of free-throws at various stages of basketball games," *Perceptual and Motor skills*, vol. 78, no. 1, pp. 243–248, 1994.

Kruskal, J. B., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.

Lee, J. A. and Verleysen, M., *Nonlinear dimensionality reduction*. Springer, 2007, vol. 1.

Leeds, M. A., Von Allmen, P., and Matheson, V. A., *The economics of sports*. Routledge, 2018.

Lees, A., "Technique analysis in sports: a critical review," *Journal of sports sciences*, vol. 20, no. 10, pp. 813–828, 2002.

McCorey, J. M., "Forecasting most valuable players of the national basketball association," Ph.D. dissertation, The University of North Carolina at Charlotte, 2021.

Morgulev, E., Azar, O. H., and Lidor, R., "Sports analytics and the big-data era," *International Journal of Data Science and Analytics*, vol. 5, no. 4, pp. 213–222, 2018.

Naismith, J., *Basketball: Its origin and development*. U of Nebraska Press, 1996.

O'donoghue, P., *Research methods for sports performance analysis*. Routledge, 2009.

Oliver, D., *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc., 2004.

Olshausen, B. A. and Field, D. J., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

Rains, R., *James Naismith: The man who invented basketball*. Temple University Press, 2011.

Rosenkrans, S., "Fragile nature of competitive advantage in an analytic driven market: Sportvu in the nba," 2016.

Sampaio, J. and Janeira, M., "Statistical analyses of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions," *International Journal of Performance Analysis in Sport*, vol. 3, no. 1, pp. 40–49, 2003.

Sandri, M., Zuccolotto, P., and Manisera, M., *BasketballAnalyzeR: Analysis and Visualization of Basketball Data*, 2020, r package version 0.5.0. [Online]. Available: https://CRAN.R-project.org/package=BasketballAnalyzeR

Sarlin, P., "Data and dimension reduction for visual financial performance analysis," *Information Visualization*, vol. 14, no. 2, pp. 148–167, 2015.

Sarlis, V. and Tjortjis, C., "Sports analytics—evaluation of basketball players and team performance," *Information Systems*, vol. 93, p. 101562, 2020.

Shin, H.-C., Orton, M. R., Collins, D. J., Doran, S. J., and Leach, M. O., "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1930–1943, 2012.

Song, K., Zou, Q., and Shi, J., "Modelling the scores and performance statistics of nba basketball games," *Communications in Statistics-Simulation and Computation*, vol. 49, no. 10, pp. 2604–2616, 2020.

Stephanos, D., "Machine learning approaches to dribble hand-off action classification with sportvu nba player coordinate data," 2021.

Sun, W., Shao, S., Zhao, R., Yan, R., Zhang, X., and Chen, X., "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, 2016.

Torgerson, W. S., "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

Torgerson, W. S., "Theory and methods of scaling." 1958.

Trninić, S., Dizdar, D., and Lukšić, E., "Differences between winning and defeated top quality basketball teams in final tournaments of european club championship," *Collegium antropologicum*, vol. 26, no. 2, pp. 521–531, 2002.

Turner, A., Brazier, J., Bishop, C., Chavda, S., Cree, J., and Read, P., "Data analysis for strength and conditioning coaches: Using excel to analyze reliability, differences, and relationships," *Strength & Conditioning Journal*, vol. 37, no. 1, pp. 76–83, 2015.

van der Maaten, L., Postma, E., and Herik, H., "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research - JMLR*, vol. 10, 01 2007.

Waggoner, P. D., "Modern dimension reduction," *CoRR*, vol. abs/2103.06885, 2021. [Online]. Available: https://arxiv.org/abs/2103.06885

Wang, Y., Yao, H., Zhao, S., and Zheng, Y., "Dimensionality reduction strategy based on auto-encoder," in *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, 2015, pp. 1–4.

Wikipedia, "New Orleans Pelicans — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=New%20Orleans%20Pelicans&oldid=1121797579, 2022, [Online; accessed 21-November-2022].

Yann, L., "Modeles connexionnistes de lapprentissage," Ph.D. dissertation, These de Doctorat, Universite Paris, 1987.