University of Colorado Boulder

Principal Component-Based Clustering of NBA Team Data: Investigating Championship Features

Gregor Tzinov

Advanced Data Science: CSCI 4022

Zachary Mullen

5/4/2022

I: Introduction

The data-driven analytics era of sports has arrived. As more and more data are collected, new approaches to analyzing sports arise. Player and team evaluations have changed to use more advanced data-hungry statistics, and team strategies are evolving to reflect what is found from modern and powerful data science techniques. As a basketball fanatic, I've always had an interest in applying my Data Science skills to uncover patterns in the game that can't be seen by just watching. The following is an investigation into what separates the basketball teams that have what it takes to win it all and those who do not. From the previous NBA championship teams, each team has had their own recipe to winning the championship. There are teams that have relied on high powered offenses to carry them, teams who have had gritty defenses that breaks down opposing teams, and teams that focus on playing fundamental basketball for 48 minutes and focus on low mistake basketball. There also have been many teams that have often gotten close to winning it, but never were able to. Are there certain ways a team should play that leads to increased championship potential? Can we statistically separate the teams that win it all from those that don't based off their regular season data? The goal is to perform exploratory analysis and see whether we can uncover structure in our data that separates our championship teams, and if so, are there strategies a team can adhere to to increase their chances of becoming a championship team.

II: Data Sources

The data that is used was collected from Basketball Reference [1]. Two different data sets were compiled together. The first is data of the past 39 championship seasons, ranging from the 1980-1981 season to the 2019-2020 season. I'll refer to this data set as data set A. For each season, the championship team was taken, and for each team, there are the following set of statistics.

- Margin of Victory (MOV): the average margin of victory in points for a team
- Strength of Schedule (SOS): the average opposing team's win percentage
- Offensive Rating (ORtg): the number of points a team scores, normalized to the amount calculated for 100 possessions, not a single game necessarily
- Defensive Rating (DRtg): similar to above, but for number of points per 100 possessions for the opposing team
- Effective Field Goal Percentage (eFG%): the percentage of shot attempts that go in, altered to account for 3 pointers being worth more points.
- Turnover percentage (TOV%): the number of possessions that lead to a turnover
- FT/FGA ratio: the number of free throws made per field goal attempt, reflects how reliant a team is on free throws for points
- Opponent field goal percentage (oppFG%): an average for the field goal percentage of the opposing team
- Opponent turnover percentage (oppTOV%): an average for the turnover percentage for the opposing team

The other data set that was compiled (data set B) was the same set of variables, but for each of the past 39 seasons, an additional 4 teams were taken. Two of these teams were randomly selected from the pool of teams that had won more games than they lost, excluding the championship team for the season, and the other two teams were selected from the pool of teams that lost more games than they had lost. Therefore, dataset A has dimensions 39x9, while dataset B has dimensions 195x9.

III: Current Championship Data Analysis

With the aforementioned coined "data-driven analytics era" of sports, there's been a great increase in the number of analyses being performed on NBA data. Ultimately, there aren't any specific well known championship team analyses that have game changing. There are also no academic papers on the subject. However, a search for NBA championship team analytics reveals many articles present for various ways to use NBA team data to try to predict playoff performance. A few of them [2,3] use a variation of regression models to try to predict a single winner directly, and it was interesting to note similar predictors were used and PCA was used similarly as well. There are a few others using tree-based methods to predict playoff and championship teams [4]. Outside of individualized articles, I saw more and more evidence that despite no specific research papers or methods for how NBA team data is analyzed, professional basketball teams are beginning to prioritize their data analytics staff. Part of the success of the Toronto Raptors championship run in 2019 was credited to their partnership with IBM for more advanced data-driven decisions [5]. Also, in 2019, the NBA hosted a hackathon for college students to solve business and basketball analytics problems that the NBA had at the time [6]. There isn't a single methodology to analyzing basketball data, but since ultimately the NBA is a business, if there are more specific successful methods, they'll be kept internal to an organization. However, as tracking has increased, more decisions are being made off data. This specific report will be aiming to examine past NBA team data with less supervised techniques and aim to investigate structures in past championship and non-championship data.

IV: Data Exploration and Preparation of Datasets 'A' and 'B'

Trend Detection

One of my first thoughts when thinking contextually of the data set for what would need to be done to answer my guiding questions were whether a trend existed in the variables, and to remove it to ensure the analysis is just, since the game of basketball has significantly changed overtime. Figure 4.1 shows the trends of 3 variables in the data set. Between the 40 years of data included, there does not appear to be any significant trend over time. There are perhaps minute trends on smaller scales, however these are more results of the play styles of each championship team. Overall, there isn't a major trend up or trend down in any of the features shown in these 40 years. If we were to take data that dated back longer, perhaps those intuitions of the game changing over time would show. Additional discussion on this is done in the conclusion.

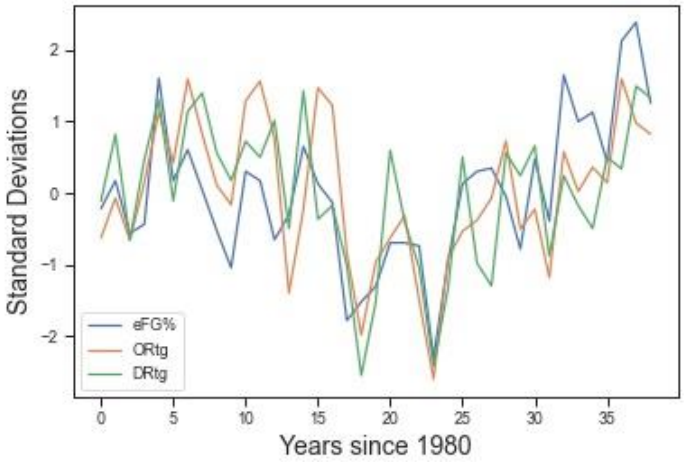


Figure 4.1: Trends of 3 prominent features

Pairwise Correlations

To get a better understanding of the data, some initial data explorations and visualizations were performed. Figure 4.2 shows the pairwise correlation plots with both data sets.

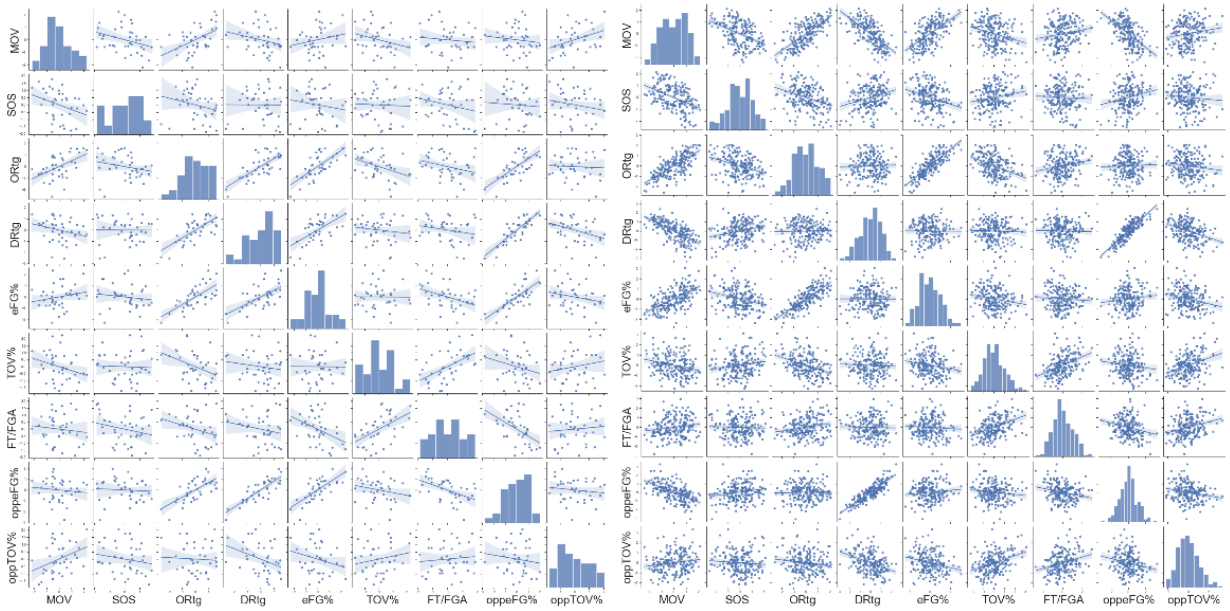


Figure 4.2: Pairwise Correlation Plots for Data Sets A, B

The two plots together show lots of correlation between the variables. There are primarily intuitive ones in the larger Data Set B, such as opponent field goal percentage and defensive rating, and margin of victory with almost all others. In Data Set A, there are some interesting weaker correlations such as between defensive rating and offensive rating (a stronger offensive team being correlated with a weaker defensive team). The biggest conclusion out of the plots is that there are many variables that explain the same

direction, and that some dimension reduction and/or feature extraction could be helpful.

Principal Component Analysis

One of the first major steps done with this data set was to perform Principal Component Analysis to reduce the data set into a lower dimension. This was done before other analysis to ensure that each "feature" we include in future analysis adds its own direction of data that is significant. The data was first centered and normalized, and then an SVD decomposition was performed to obtain the U, Sigma and V transpose matrices to be able to examine the energy of the system with the varying sigma values, and extract the left eigenvectors to perform the rotation of our data set into the coordinate system of the eigenvectors. The first 6 Principal Components were kept based off the energy plot seen in Figure 4.3 (which was nearly identical for both data sets), and consequently our data was then transformed into a 6-column matrix, with number of observations staying the same as before for Data Set A and B. The rest of the analysis was performed using these resulting data sets.

Outlier Analysis

Since the data was carefully chosen, there are no outliers in the data coming from poor recording, poor entry, or other logistical sources. However, an outlier analysis done using Isolation Forests can give valuable preliminary insights of the data. Using sklearn's IsolationForest library, an Isolation Forest algorithm was run using 20,000 trees to convincingly reach convergence, contamination level automatically set since there is no intuition for how many "outliers" there could be and included all features of the PCA-reduced data set[1]. For our Data Set A, the 1995-1996 Chicago Bulls team was the largest outlier, while the 2006-2007 Spurs team was the largest inlier. Off the bat, it seems like these two teams can perhaps be separated into their own groups. The 95-96 Bulls team were one of the most dominant basketball teams of all time, with very high offensive firepower and a stingy defense (as also seen with their ORtg and DRtg), which lead to the 2nd greatest regular season performance of all time and a decisive finals run in which they lost 3 games in the entire post season [1]. The Spurs on the other hand are known for playing traditional fundamental basketball, in which they do a bit of everything well but nothing particularly exceptional, leading to a good intuition for while they represent the most "typical" championship team. Performing the Isolation Forest analysis on Data Set B, we get some overlap of outliers from Data Set A, such as the 95-96 Bulls Team and the 98-99 Spurs Team, however the largest outliers were teams that were very bad (Knicks, Raptors and Hawks teams that were a whole standard deviation under for Margin of Victory. The difference between the weakest teams in the

[1] This will create 20,000 random trees of our data, where for each split of the tree, the algorithm will take a random column, split the data based off a random value chosen in between the minimum and maximum values of the column, and continue for each branch of the resultant tree to obtain how many "cuts" is needed for each observation to be isolated. By performing 20,000 iterations, we obtain an average number of cuts for each observation, and a subsequent anomaly score that is correlated to the number of cuts. The lower number of cuts needed, the more indicative it is that a particular is an outlier.

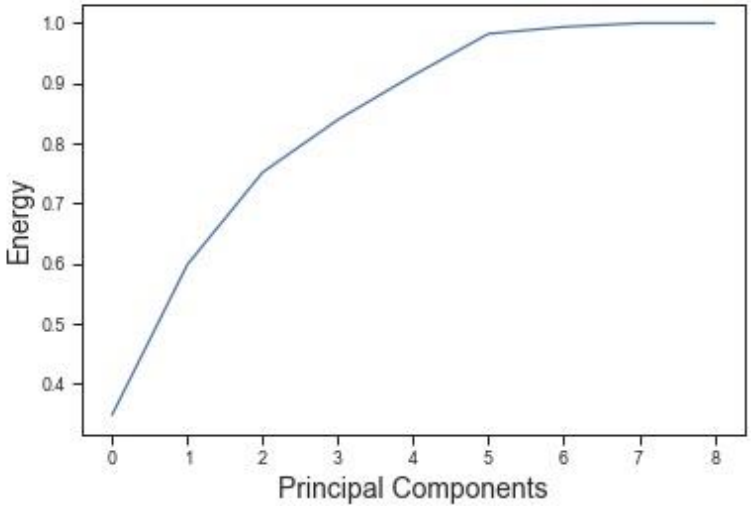


Figure 4.3: Cumulative Energy of System

NBA and the average team is perhaps greater than the average team and a championship team, as even by not including every team from each season, the three largest outliers were losing teams, and not necessarily the worst in the league that season.

V: Numerous Clustering Results using K-Means Clustering

After performing various initial data explorations, I performed a few different runs of K-Means clustering with slightly different features and options for *k*. Performing Principal Component Analysis to reduce my data set gave me better convergence for my K-Means runs and gave me more confidence to cluster based off only two or three features since I know they captured a decent proportion of the variability of the data. Additionally, seeing the outlier analysis results gave me the intuition for how to interpret the results, as it was evident that there was some structure in the statistics for each team that matched the basketball intuition. Since I have two primary data sets, I'll give each their own section. For both data sets, I ran K-Means using both a handwritten algorithm and using sklearn's implementation of K-Means to ensure I get consistent results [2].

[2] K-Means was performed using random initialization for the centroid beginnings and ran according to how K-Means clustering algorithms is defined. For each observation, we assign each point to the nearest centroid by computing the distance (with distance function being chosen as Euclidian Distance) from the observation to each centroid. After performing this with each observation, we recompute a centroid's center point by taking all of the points assigned to that centroid and find the new center point to represent a given centroid. Following this step, we have the same number of centroids as initialized in the beginning as a k value, however they will have been most likely shifted. This repeats and continues until the total reconstruction error (the distance between each observation and its assigned centroid) changes less than a given tolerance error defined before running the algorithm. The algorithm is often repeated several times, and the run with the lowest reconstruction error by the end is chosen as the way the points are clustered.

Clustering on the Championship-Only Data

To get good plotting ability, I began the clustering analysis by using the first two principal components. To get the ideal number of centroids, I ran K-Means 50 times with k ranging from 2-10. The plot of the reconstruction error is seen in Figure 5.1. We see that the error decreases when the number of centroids increases, which will always be the case since we allow the points to have an increasingly closer centroid to them. However, the aim is to choose the point where there's a balance between minimizing the error, and not choosing too many centroids. This is typically the point where there's a sharper angle in the line, and with this example, this is seen clearly at a k value of two. This is the sort of methodology done for choosing the number of centroids with K-Means, and was performed many times in the subsequent iterations, so I will focus on discussing just the results for the remaining analysis.

The resulting clustering is seen in Figure 5.2. This is of the cleanest clusters we see in this report. By using the first two principal components, we see a very natural split into two clusters. We see lots of teams squished together, and then a handful of teams that are seemingly in their own space. We see all 6 of Michael Jordan's Bulls Teams, the two most dominant Golden State teams, and Kobe Bryant's 2008-2009 Lakers team.

Upon deeper dive, it's interesting to note that not all of Jordan's teams were necessarily dominant in the regular season, as in the 97-98 season, their MOV was average. From the 9 teams selected, the columns that were consistently abnormal were the turnover percentage column, in which all 9 teams were around a standard deviation lower than average, the ratio between free throws made to field goal attempts, as it was under average or 8 of the 9 teams, indicating these teams relied less on the free throw

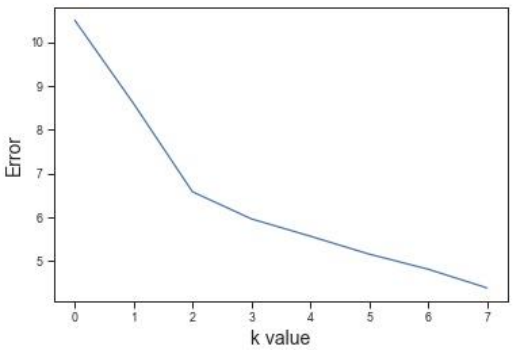


Figure 5.1: Reconstruction Error vs Number of Clusters

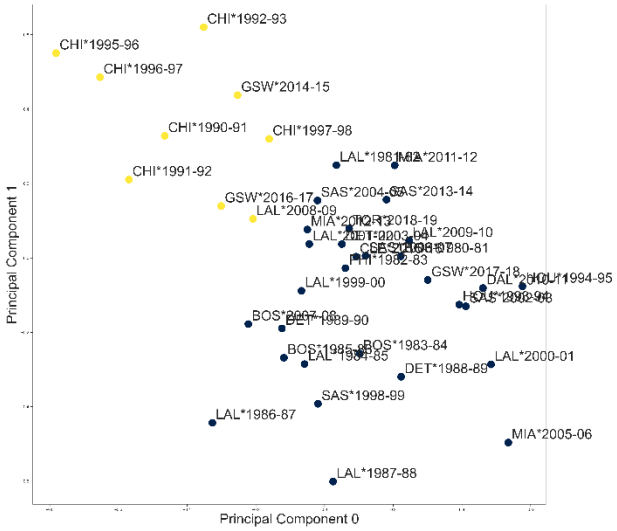


Figure 5.2: Clustering Results Using Principal Components and k=2

line, and finally the ORtg was around a standard deviation or two higher for 8 of the 9 teas as well. This subset of teams represents of the most dominant in history, and they separate themselves with a very high powered offense and low turnovers, indicating that offensive efficiency in the regular season is important for overall success. Defensive Rating varied in these teams, as some teams were under average, while others were over, and typically not by much. A high-powered offense with decent defense seems to separate the greatest teams.

When including 3 Principal Components and getting a resultant elbow plot that gives reasoning to use 3 clusters, we get Figure 5.3. What emerges is a new cluster that represents a different brand of basketball. We get teams like the "Bad Boy" Pistons back in 1989 and the revamped version of them in 2003, where both teams were known for playing physical defense, and similarly the David Robinson-Tim Duncan era of the San Antonio Spurs. The teams in the cluster typically have exceptionally low DRtgs (meaning opponent's points scored per 100 possession was low) with them tending to be nearly two standard deviations below average, and relied more on the free throw line. Ultimately, the most natural clusterings based on two and three principal components captured natural structure in the past championships. Unsurprisingly, these teams mentioned had high anomaly scores in the preliminary analysis. Our

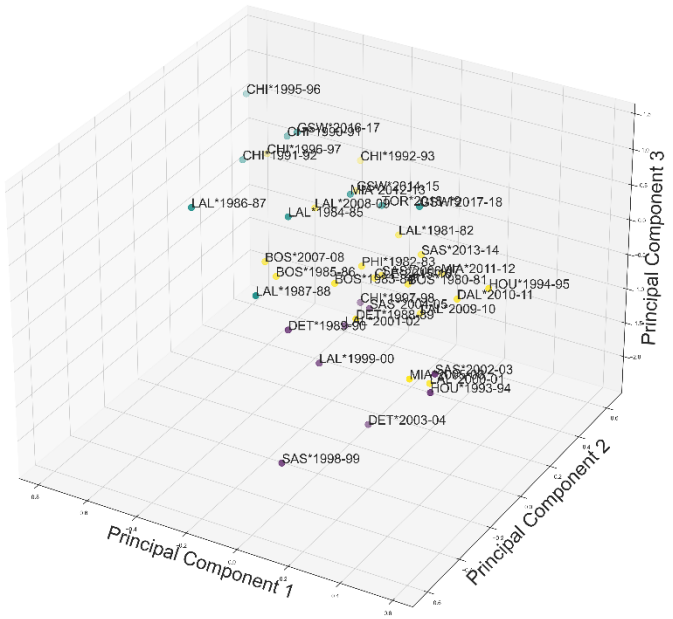


Figure 5.3: Clustering Results using 3 Principal Components, and k=3

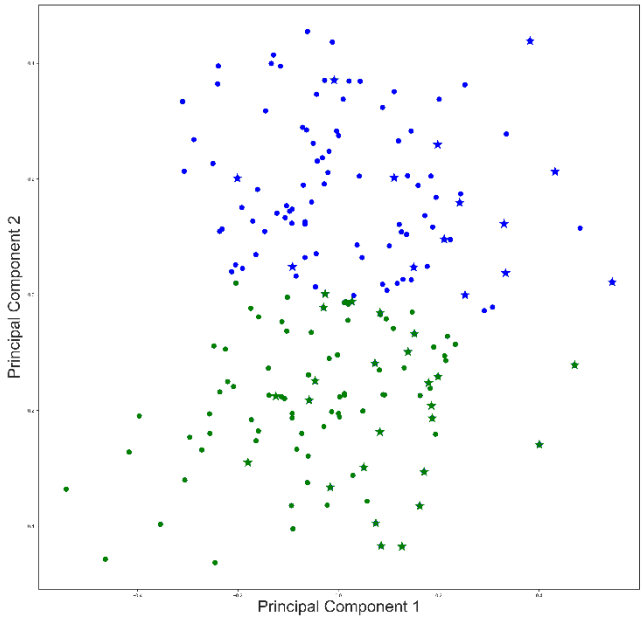
clusters tended to represent the overall dominant teams with efficient offensive and defenses strong enough for the playoffs, the very physical and strong defensive teams, and then "the rest" of the teams that were more average championship winners.

Clustering on the Entire Data Set

I first wanted to discover the question whether championship teams are such statistical anomalies that if we ran K-Means using k=2, whether they would nearly all be in their own cluster. As seen in Figure 5.4 and 5.5, this partition was not achieved. We see about an equal number of points in each centroid, and while one has more championship observations than the other, due to the amount of points in each centroid, there doesn't seem to be a natural partition using two principal components.

If we forgive the ability to have a 2-dimensional plot to show our clustering, we can see what the ideal number of principal components would be to have the combination of dimension reduction and coverage of variability in our data. Looking at Figure 4.3, we see we keep most of the variability of the data in only 6 Principal Components.

By running K-Means using 6 Principal Components, and again use an elbow plot to find the ideal number of clusters (k=4), we get Figures 5.6 and 5.7.



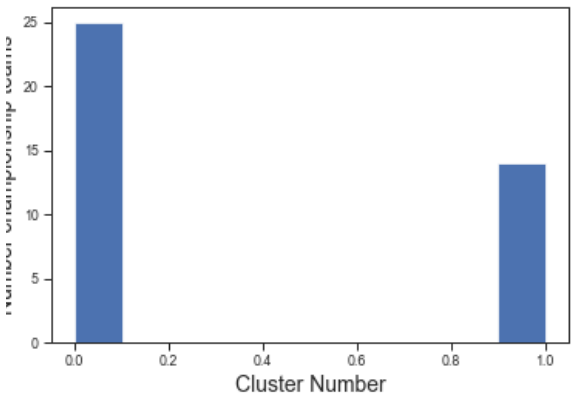
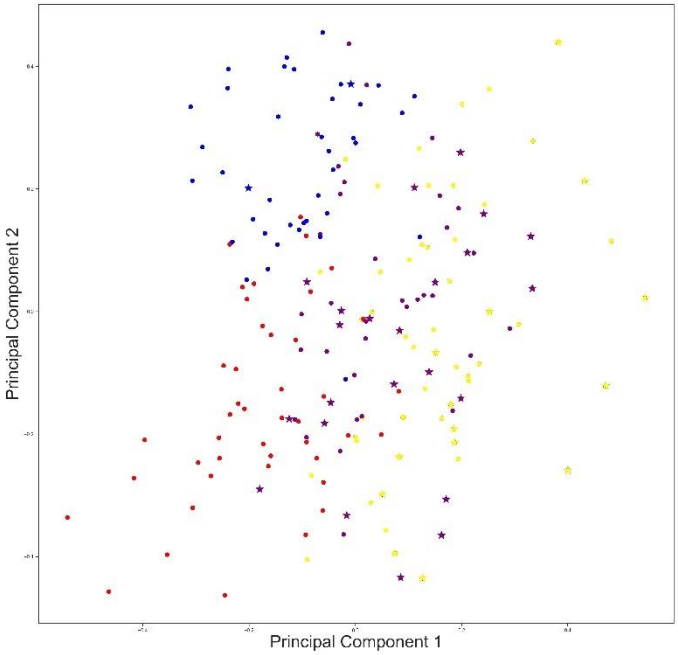


Figure 5.4 and 5.5: All Data Clustering Using Two Principal Components and k=2



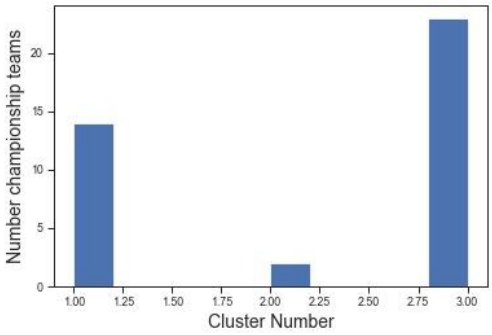


Figure 5.6 and 5.7: All Data Clustering Using Six Principal Components and k=4

We see a slightly different answer to our question of whether we can separate the championship teams from non-championship teams using clustering. In Figure 5.6, we are plotting our points in the first two principal components directions, but coloring based off the 4 different clusterings, and labeling as stars the championship teams. When we look closer, we see that most purple observations are indeed stars, and most of the championship teams are in cluster 3, which is the purple coloring. We get slightly closer to having a "championship" centroid, in which a clustering result separates the championship teams, and doesn't include too many non-championship teams.

Additional Analysis: Can we Predict Using the Psuedo-Championship Centroid?

While our cluster 2 from before doesn't entirely encapsulate all championship teams and only them, we do get most championship teams in cluster 2 and not many non championship teams. From this result, I decided to take modern NBA season data (2020-2021 season data and 2021-2022 season data) and see whether a K-Nearest-Neighbor algorithm could give us the closest neighbors to this centroid. With each season's data, I converted it to the coordinate system generated from the SVD composition of the entire data set by cleaning the season data as did with the original data, and matrix multiplying our "test" data by the left eigenvectors to rotate it into the same space. Our result was a 30x6 data set for each season. Sklearn's KNN algorithm was trained on this new data set, and I fed it to calculate the closest neighbors to the championship centroid. For the 2020-2021 season, the 5 closest neighbors were the

Celtics, Pacers, Mavericks, Suns, and Heat. The result of the season was the Suns meeting the Bucks in the finals, and ultimately losing to them, while the other teams were barely playoff teams, so not great performance. For the 2021-2022 that just passed and is currently in the post season, the 5 closest neighbors were the Cavs, Mavericks, Nets, Lakers and Heat, two of which are dominant teams and have a chance to win it, but the other 3 being pretty off.

Conclusion and Further Analysis

The analysis done on the NBA team data revealed some interesting structure to the data. Sub structure was found on the championship data set. Using 2 and 3 Principal Components, there was some grouping of the outlier teams, and there was consistency in the groupings despite observations being placed in an abstract lower dimensional space. Looking at the full data set with all of the teams, the structure found wasn't as clear cut, and further analysis on the sub structures found from the ideal clusterings would be beneficial to further examine the data, such as looking at characteristics of teams in each centroid, and why teams that were championships and weren't in our championship centroid were placed in a different cluster, and for those teams that were in the championship cluster but didn't win, what may have been the reason.

After performing PCA to reduce the data set, I realized that using more features in the initial set may have been valuable in adding the information we had for each observation since I was going to be using PCA to procedurally reduce my dataset. Additionally, in the second half of the analysis, I should have used the full season data for each year and had only towards the end of the project realized this wouldn't have been difficult as I thought. Full season data would also give us more robust data to analyze whether trends of features existed in the 40 years included.

As a follow up to this to where I left off on this project, I would love to perform more classical prediction using the data I have. The clustering methods used are great ways to explore the data, but the prediction method with KNN used isn't as robust, which is evident by the results I got. Using a larger data set with more observations, with labels for non-playoff teams, playoff teams and championship teams, and a larger initial feature set that is reduced down using PCA, prediction and classification could be done using a neural network architecture, as they excel at performing such tasks. This sort of methodology can be used to train using lots of data, and test using training samples held out from the training sample to see whether it could predict championship teams well. Hyperparameters and specific architectures could be adjusted to maximize our performance. However, the goal of this project was to see how Outlier Detection, K-Means Clustering, and KNN, can be used on a data set that was of great interest to me, and overall, they were worked nicely together to learn more about past championship teams.

References

1. *Basketball statistics and history*. Basketball. (n.d.). Retrieved May 4, 2022, from https://www.basketball-reference.com/
2. Yoon, T. (2020, July 6). *Predicting the 2020 NBA champion with Machine Learning*. Medium. Retrieved May 4, 2022, from https://towardsdatascience.com/predicting-the-2020-nba-champion-with-machine-learning-32100f6b253d
3. Big, D. (n.d.). *2019-2020 NBA playoffs prediction*. Advanced Data Science Final Project. Retrieved May 4, 2022, from https://advds71x.github.io/NBAproj/
4. *Using decision tree algorithms to test the accuracy of NBA playoff predictions*. Samford University. (n.d.). Retrieved May 4, 2022, from https://www.samford.edu/sports-analytics/fans/2022/Using-Decision-Tree-Algorithms-to-Test-the-Accuracy-of-NBA-Playoff-Predictions
5. *Toronto Raptors use data-driven command center on path to NBA finals*. THINK Blog. (2019, June 17). Retrieved May 4, 2022, from https://www.ibm.com/blogs/think/2019/06/toronto-raptors-use-data-driven-command-center-on-path-to-nba-finals/
6. *NBA Hackathon Overview*. NBA Basketball Analytics Hackathon. (n.d.). Retrieved May 4, 2022, from https://hackathon.nba.com/