Using KMeans Clustering to Identify Potential 2023 NBA Champions

Jorden Anfinson

February 2023

Abstract

K-Means clustering is traditionally used to identify patterns within unlabeled data, assembling similar observations together into groups known as clusters. K-Means is useful for many different tasks, such as pattern recognition and fraud detection, however, it is seldom thought of as a method for prediction. This paper details how K-Means clustering can be used for prediction, specifically, how it can be used to identify potential future 2023 NBA Champions using NBA Team Data from the past 13 NBA seasons. We run K-Means clustering 1000 times to cluster the data of teams over the past 13 seasons, including the data of the past 13 NBA Champions, with the data of teams currently playing in the 2023 NBA Season. We do so to see which currently playing teams most frequently cluster with past NBA champions, noting that the teams that most frequently do so are the most statistically similar to the past NBA champions. We identify the **Milwaukee Bucks** as the team that best fits the statistical profile of the past 13 NBA champions.

1 Introduction

Unsupervised learning can be described as a process in which patterns are discovered within unlabeled data. Algorithms that fall under the umbrella of unsupervised learning have become essential to any field involved in large scale data analysis.

A common unsupervised learning technique is called clustering. Clustering is the act of organizing segments of data into groups of data with similar characteristics. K-Means Clustering is a specific type of clustering algorithm that organizes data points into some 'k' number of clusters, while trying to minimize the distance between each data point and its assigned cluster center. As with any clustering algorithm, K-Means can be leveraged to find patterns within data that normally wouldn't be readily apparent. This makes it a powerful tool in the analysis of sports data.

Analysts within sports organizations often use clustering algorithms to identify potential replacements for players who may decide to leave during free agency. For example, if Lebron James decided to leave the Los Angeles Lakers, analysts within the Lakers organization would likely use clustering algorithms to attempt to find a player that fits James' statistical profile and player archetype.

Clustering can also be used to identify and cluster statistically similar teams, that is: given a team's statistics, K-Means clustering can be used to see which other teams have similar statistics, as teams with similar statistics will cluster together. In this paper, we use K-Means clustering to cluster active teams (teams currently playing in the 2023 season) with the past 13 NBA champions. We accomplish this by using NBA team data from the current 2023 season and from the past 13 seasons. Ultimately, we run K-Means 1000 times on this data, and find the currently active teams that most frequently cluster with the past 13 NBA Champions across all iterations of K-Means.

2 Data

We utilize a dataset that contains the overall season stats of every team since the founding of the NBA in 1946. Containing 54 features in total, it contains over 40 continuous features such as "w" (wins), "l" (losses), "sos" (strength of schedule). The majority of these continuous features represent per-game statistics, such as "points per game" and "rebounds per game". Statistics such as these represent the season averages of a team.

The dataset also contains many descriptive features that detail the team's name, whether they made the playoffs, and whether they were an NBA Champion. A preview of the data used can be seen below. Additionally, a table describing the majority of the features within the dataset is enclosed in the appendix.

Season	lg	Team	abbreviation	playoffs	champion	g	w	l	sos	...
2023	NBA	Atlanta Hawks	ATL	FALSE	FALSE	59.0	29.0	30.0	-0.22	...
2023	NBA	Boston Celtics	BOS	FALSE	FALSE	59.0	42.0	17.0	-0.35	...
2023	NBA	Brooklyn Nets	BRK	FALSE	FALSE	59.0	34.0	24.0	0.05	...
2023	NBA	Chicago Bulls	CHI	FALSE	FALSE	59.0	26.0	33.0	0.33	...
...	...	...	...	...	...	...	...	...	...	...

Table 1: Preview of Dataset

3 Methodology

We look to identify potential 2023 NBA Champions through the use of K-Means clustering. As stated in the introduction, K-Means clustering works by grouping data points within a dataset in a way that minimizes the mean-squared error (MSE) of the data. The algorithm takes in one parameter: K, which is equivalent to the number of clusters that will be used to group the data. For our purposes, K-Means clusters the data of currently playing NBA teams (teams actively playing in the 2023 season) with the data of NBA champions from the past 13 seasons.

Before K-Means can be implemented, alterations to the data must be made. As many features within our data have different ranges, we first standardize the data to ensure that no statistic has a larger impact on the clustering than another.

Feature	Min	Median	Max
fg_per_game	34.500000	39.300000	44.700000
sos	-0.840000	-0.110000	0.830000

Table 2: Example Feature Ranges Before Standardization

Feature	Min	Median	Max
fg_per_game	-2.224	$-9.06e^{-0.2}$	2.31
sos	-2.1105	$1.59e^{-2}$	2.754

Table 3: Example Feature Ranges After Standardization

We also create two subsections of our original 76 season dataset. The first subsection is made up of the data of all teams currently playing in the 2023 season. We refer to these teams as ”active teams” throughout the rest of this paper. As the 2023 season has not ended as of the writing of this report, active team data is not as complete as the other data within the entire 77-season dataset, as these teams have played fewer games.

season	lg	team	abbreviation	playoffs	champion	g	...
2023	NBA	Atlanta Hawks	ATL	FALSE	FALSE	59.0	...
2023	NBA	Boston Celtics	BOS	FALSE	FALSE	59.0	...
2023	NBA	Brooklyn Nets	BRK	FALSE	FALSE	59.0	...
...	...	...	...	...	...	...	...

Table 4: Preview of Active Team Data

The second subsection includes the data of all teams (including the data of each year’s NBA Champion) since the 2010 season. A brief preview of this selection of data is shown in Table 5 below.

We only select data from the past 13 seasons so as to avoid clustering by time. Basketball has changed over time, and the statistics reflect this. The graph below in Figure 1 depicts how the average number of possessions per game has changed over the past 40 seasons.

season	lg	team	abbreviation	playoffs	champion	games	...
...	...	...	...	...	...	...	...
2010	NBA	Philadelphia 76ers	PHI	FALSE	FALSE	82.0	...
2010	NBA	Phoenix Suns	PHO	TRUE	FALSE	82.0	...
2010	NBA	Portland Trail Blazers	POR	TRUE	FALSE	82.0	...

Table 5: Preview of Team Data From the Past 13 Seasons

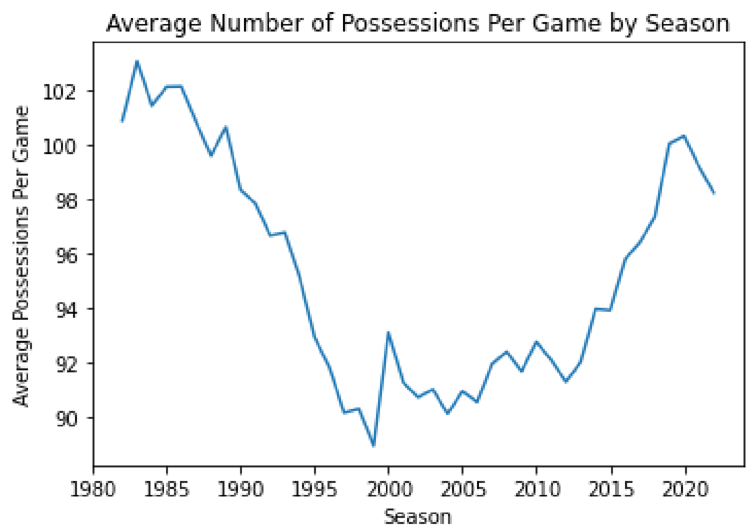


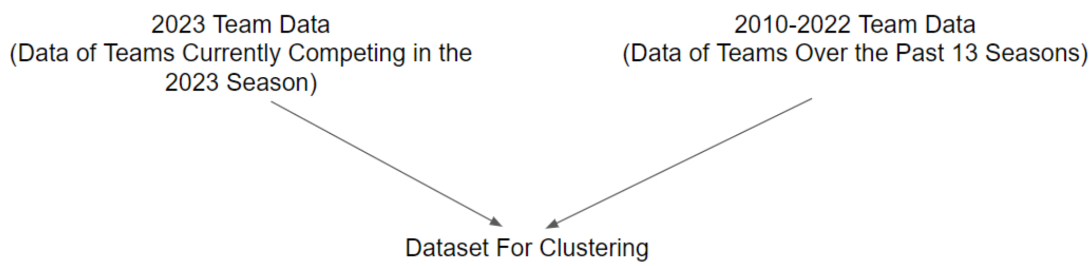
Figure 1: Average Number of Possessions Per Game by Season

If we included teams from the 1990s in our clustering dataset, their statistics would differ from the data of active teams purely because basketball was played differently in the 90s. This could lead to teams from the 90s exclusively clustering together, away from teams from other decades. While there still likely will be some clustering due to time, by restricting the data to the past 13 seasons, we reduce the potential that the data will be clustered by time rather than pure statistics, while also providing enough additional data to evaluate how active teams cluster with former NBA Champions.

Additionally, although we are mainly interested in how active teams cluster with past champions, we include the data of teams who did not win the championship in their respective season to introduce more noise. If teams more frequently cluster with more teams that did not win the championship in their respective season rather than past NBA Champions, they do not closely fit the statistical profile of an NBA champion. One could say these teams are less likely to win an NBA championship.

Illustrated in Figure 1, we merge the two subsections of data (active teams and teams that played from 2010-2022) to create the full clustering data set K-Means is implemented on.

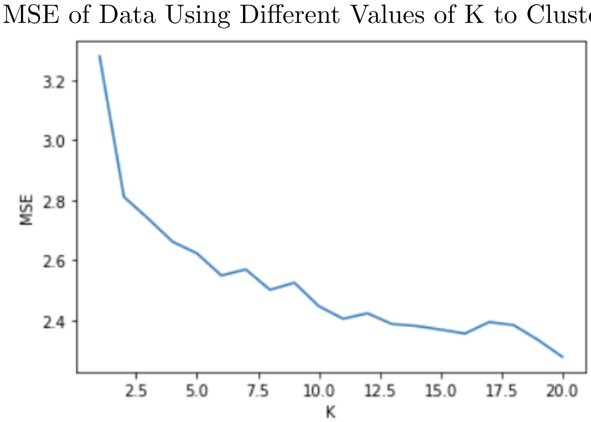
Figure 2: Data Combination Procedure



One more step must occur before clustering: we remove any feature that is indicative of a team's success. We are interested in how the data will cluster solely using the team's stats, not measures of their success such as wins, losses, or whether they won a championship. After making copies of each, we remove the following features: "w" (wins), "l" (losses), "g" (games), "playoffs", and "champion", as we do not want clustering to be affected by these values.

We determine the optimal value of K by using the graph in Figure 2, which depicts how K-Means performs using different values of K. The graph in Figure 2 indicates that a K value of 4 can be used to cluster the data well.

Figure 3: MSE of Data Using Different Values of K to Cluster the Data



After identifying the optimal value of K, we run K-Means 1000 times on our clustering dataset using a K of 4.

K-Means is initialized by positioning K random cluster centers throughout the data. Because of this, the final location of cluster centers will often times vary. By running the algorithm 1000 times, we reduce the randomness caused by the random initialization of cluster centers and can better quantify how similar each active team is to past NBA Champions.

After each iteration of K-Means we record how many past NBA Champions each active team clusters with. We use a dictionary to do so, using each active team's name as a key and the total number of champions each team clustered with across all 1000 iterations of K-Means as a value.

For example, if 4 former NBA Champions clustered with the Milwaukee Bucks during the first iteration of K-Means, and 5 in the second, the value in the dictionary assigned to Milwaukee after two iterations of K-Means would be 9, and we would repeat this procedure 998 more times. Pseudocode for this procedure is included below.

Algorithm 1 Counting the number of former champions each active team clustered with over 1000 iterations of K-Means

```
teamDict =                          ▷ Dictionary for total past champions clustered with
firstRun = True                     ▷ True if first iteration of K-Means
for i in range(1000) do
    Run K-Means                     ▷ Run K-Means to assign clusters to each team
    for TeamA in activeTeams do
        numChampions = 0
        cluster = TeamA.cluster
        for TeamB in cluster do
            if TeamB.champ == True then
                numChampions +=1
            end if
        end for
        if firstRun == True then    ▷ initializing key for TeamA
            teamDict[TeamA] = numChampions
        end if
        if firstRun != True then
            teamDict[TeamA] += numChampions
        end if
    end for
    firstRun = False
end for
```

4 Results

Team	Number of Former Champions Clustered With
Milwaukee Bucks	10572
Brooklyn Nets	10172
Boston Celtics	9924
Washington Wizards	9980
Denver Nuggets	9924
Chicago Bulls	9732
Philadelphia 76ers	9656
Los Angeles Clippers	9372
Cleveland Cavaliers	9368
Miami Heat	8604

Table 6: The 10 Most Similar Teams to the Past 13 NBA Champions

A table depicting the results for all 30 NBA teams is included in the appendix.

Table 6 shows the 10 teams that cluster with the most former NBA Champions across all 1000 iterations of K-Means.

We can see that the Milwaukee Bucks cluster with 10572 former NBA Champions, across all 1000 iterations of K-Means. Because Milwaukee clusters with the most former NBA Champions, we can say they are likely the most statistically similar to the past 13 NBA Champions out of the active teams.

The Brooklyn Nets, Boston Celtics, Washington Wizards, Denver Nuggets. Chicago Bulls, Philadelphia 76ers, Los Angeles Clippers, Cleveland Cavaliers, and Miami Heat cluster with the next highest numbers of former NBA champions over all iterations of K-Means. Interestingly, each of these teams (minus Washington and Chicago) are within the top eight teams in their respective conferences as of the writing of this paper.

Position	Team	W	L
1	Milwaukee Bucks	45	17
2	Boston Celtics	45	19
3	Philadelphia 76ers	40	22
4	Cleveland Cavaliers	39	26
5	New York Knicks	38	27
6	Brooklyn Nets	35	28
7	Miami Heat	33	31
8	Atlanta Hawks	32	31

Table 7: Current Eastern Conference Playoff Standings

Position	Team	W	L
1	Denver Nuggets	44	19
2	Memphis Grizzlies	38	23
3	Sacramento Kings	36	25
4	Phoenix Suns	35	29
5	Golden State Warrior	33	30
6	Dallas Mavericks	33	31
7	Los Angeles Clippers	33	32
8	Minnesota Timberwolves	32	32

Table 8: Current Western Conference Playoff Standings

5 Conclusion

We showed that K-Means can be used to identify potential future NBA champions by counting the number of former champions each active team clusters with over repeated runs of the K-Means algorithm. In doing so, we identified the teams that are the most statistically similar to past NBA Champions.

We identified the **Milwaukee Bucks** as the most statistically similar team to the past 13 NBA Champions, and predict that Milwaukee will go on to win their second NBA title in three seasons.

6 Appendix

The full table of results is presented below.

Team	Number of Former Champions Clustered With
Milwaukee Bucks	10572
Brooklyn Nets	10172
Boston Celtics	9924
Washington Wizards	9980
Denver Nuggets	9924
Chicago Bulls	9732
Philadelphia 76ers	9656
Los Angeles Clippers	9372
Cleveland Cavaliers	9368
Miami Heat	8604
New York Knicks	7878
Dallas Mavericks	7836
Pheonix Suns	7745
Sacramento Kings	6803
Portland Trail Blazers	6785
Golden State Warriors	6453
Orlando Magic	6153
Los Angeles Lakers	6047
New Orleans Pelicans	6047
Atlanta Hawks	6003
Memphis Grizzlies	5881
Utah Jazz	5847
Toronto Raptors	5549
Minnesota Timberwolves	5545
Oklahoma City Thunder	5029
Indiana Pacers	4959
Charlotte Hornets	4931
San Antonio Spurs	4911
Houston Rockets	4781
Detroit Pistons	4729

Table 9: Former NBA Champions Clustered Over 1000 Iterations of K-Means

Season	Team Name
2022	Golden State Warriors
2021	Milwaukee Bucks
2020	Los Angeles Lakers
2019	Toronto Raptors
2018	Golden State Warriors
2017	Golden State Warriors
2016	Cleveland Cavaliers
2015	Golden State Warriors
2014	San Antonio Spurs
2013	Miami Heat
2012	Miami Heat
2011	Dallas Mavericks
2010	Los Angeles Lakers

Table 10: Past 13 NBA Champions

Feature Name	Meaning
Season	The Season the Team Played In
lg	The League in Which the Season Took Place
team	The Name of the Team
abbreviation	The Typical Abbreviation of the Team Name
playoffs	True if the Team Made the Playoffs
champion	True if the Team Won the Championship
g	The Number of Games the Team Played
w	The Number of Games the Team Won
l	The Number of Games the Team Lost
sos	Strength of Schedule
pace	The Number of Possession a Team Had Per Game
mp_per_game	Minutes Played Per Game
fg_per_game	Field Goals (Made Shots) Per Game
fga_per_game	Field Goals Attempted Per Game
fg_percent	Field Goal Percentage
x3p_per_game	3Pointers Made Per Game
x3pa_per_game	3Pointers Attempted Per Game
x3p_percent	Percentage of 3Pointers Made
x2p_per_game	2Pointers Made Per Game
x2pa_per_game	2Pointers Attempted Per Game
x2p_percent	Percentage of 2Pointers Made Per Game
ft_per_game	Free Throws Made Per Game
fta_per_game	Free Throws Attempted Per Game
ft_percent	Free Throw Percentage
orb_per_game	Offensive Rebounds Per Game
drb_per_game	Defensive Rebounds Per Game
trb_per_game	Total Rebounds Per Game
ast_per_game	Assists Per Game
stl_per_game	Steals Per Game
blk_per_game	Blocks Per Game
tov_per_game	Turnovers Per Game
pf_per_game	Personal Fouls Per Game
pts_per_game	Points Scored Per Game

Table 11: Many of the Features Within the Dataset