Sam Kenney
MAT 2400-02
May 11, 2023
Final Report

KMeans on NBA Player Data

1. Introduction

KMeans clustering is a method of grouping data together. It works by alternating between updating centroids- a random representative of data points, and partitions- the groups to which each data point belongs to. The K-means algorithm's goal is to reduce the Euclidean distance between the centroids and the data points in each cluster to the smallest possible number with the given *k* clusters. When running K-means, computer scientists or mathematicians choose *k* with overfitting and underfitting in mind. If K-means is overfitting, the number of clusters is too large to be meaningful. If K-means is underfitting, the number of clusters is too small, and there are, in actuality, more distinct groups than indicated. Therefore, many methods exist for estimating the optimal number of clusters, *k*. In my analysis, I chose to use the Elbow method, a common heuristic method for estimating the optimal *k* to choose. I applied K-means using Python to the 2022-23 regular season NBA player dataset created by Vivo Vinco. By selecting for certain traditional box score statistics, I created clusters of NBA players which are helpful for understanding the types of players in the NBA today.

2. Methodology

2.1 Data tools and selection

I used the Anaconda Python distribution and a Jupyter notebook, as recommended by the *Python Language Companion* by Jessica Leung and Dmytro Matsypura. The Anaconda distribution includes the scikit-learn library, from which I used the KMeans, the pandas library, which I used to read my NBA CSV data file, and matplotlib for plotting data. The original CSV file included duplicates of players because some players changed teams during the season, so I removed the duplicates and kept only the total statistics for duplicate players (e.g. if a player played for both Memphis and Boston, he would appear three times in the dataset- first as total stats, another as just his statistics in Memphis, and a third time as only his statistics in Boston. After I removed duplicates, only the first, total statistics version of that player remained). Then, I removed any player who played less than 41 games, which is equal to half of the games in the NBA season. I did this to reduce the number of players who were injured or players who had a limited role on their teams. I then selected which types of statistics to include for use in the KMeans clustering; player names were not useful, and several statistics bear essentially the same meaning, for example, Field Goals Made and Field Goals Attempted combined is the same as Field Goal Percentage.

2.2 Elbow Plot

To find the optimal number of clusters to use for executing KMeans on the NBA players dataset, I used the Elbow method. The elbow method attempts to identify an “elbow” at which preceding the elbow KMeans would be underfitting, or too few clusters, and following the elbow would be overfitting, or too many clusters. I used cluster inertia as a metric for determining how well each k clustering fit the data. Cluster inertia is defined as the sum of the squared Euclidean distances between each cluster’s centroid and each vector in the cluster:

$$\sum_{i=1}^{N} \quad dist(x_i, C_k)^2$$

Equation 1: Cluster Inertia, N is # of vectors (players) in the dataset and C is a cluster centroid

The point at which cluster inertia transitions between rapidly decreasing to decreasing at a linear rate is called the elbow. Since some datasets do not have an easily distinguishable elbow, the method is often used in conjunction with other methods, such as the Silhouette plot. However, our clustering only desires to find similarities and differences in NBA players, and a difference of one cluster from optimal will not heavily impact the results.

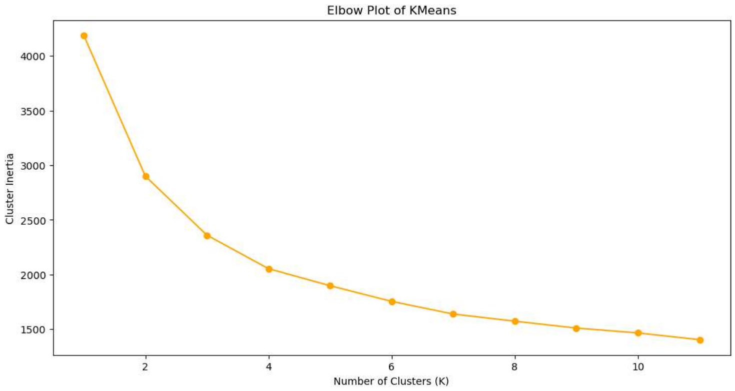

Figure 1: Elbow plot of KMeans on NBA player data for the 2022-2023 season

Based on Figure 1, I concluded that the elbow was at K=4 because the plot becomes nearly linear at that point.

3. Findings

3.1 General Findings

Upon executing KMeans on the NBA players dataset, I investigated the differences between the clusters. I first noticed that the clusters seemed to be highly correlated with minutes played. This makes sense, because each counting statistic (non percentage) should be highly correlated with minutes played- more minutes will generally equate to players putting up more points, rebounds, assists, etc. I graphed minutes played against several different statistics to help visualize the clusters (each counting stat is per game).

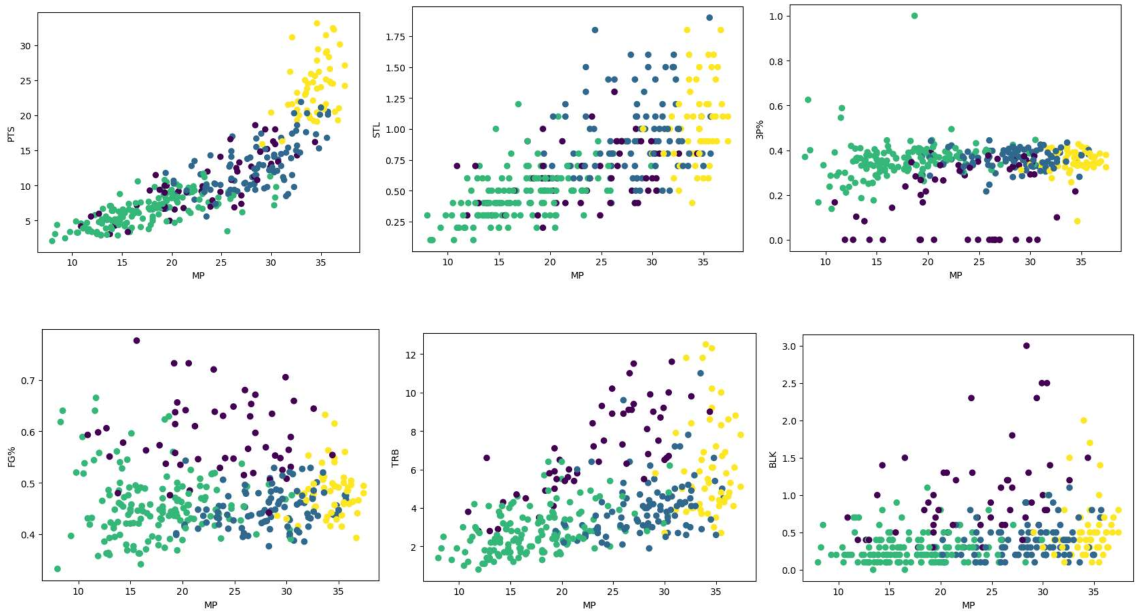

Figure 2: Minutes vs. points, steals, 3point%, field goal%, rebounds, and blocks

The takeaways here are that the yellow cluster tends to play the highest minutes, and thus gets the most steals and points. Blue plays almost as many minutes and has slightly lower statistics. Green plays the least minutes and gets the least of most stats.

3.2 Purple Cluster

The one unique cluster which does not correlate as highly with minutes played appears to be the purple cluster. Here’s a look at the purple cluster’s comparison to the average of all NBA players across all stats I considered:

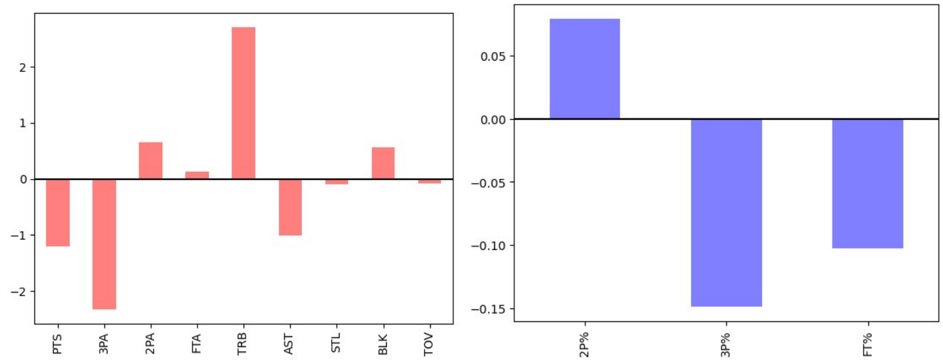

Figure 3: Purple cluster stats vs. the average NBA player in the 2022-23 regular season

The purple cluster gets the most rebounds and blocks and shoots the highest field goal percentage. They also tend to shoot the worst 3-point percentage. These trends are often what is thought of as a traditional center- players who spend most of their time very close to the basket, so their shots are easier and they are more likely to collect rebounds and blocks and less likely to take and hit threes. If we look at the first ten entries in this cluster, this checks out; there are eight centers and two power forwards, which is a very similar position to center.

3.3 Yellow, Green, and Blue Clusters

The next easiest to analyze is the yellow cluster; they played the most minutes and accumulated the best statistics. They scored the most points and were above average in every category except 2-point percentage. These are the NBA's stars. The green cluster was bad all around; their only above average statistic was their 3-point percentage. These are either 3-point specialists or developing prospects. The blue cluster was generally middle-of-the pack except for their exceptional 3-point percentage and free throw percentage. We can assume that these are role players, neither stars nor raw prospects, who are often utilized for their 3-point prowess.

4. Conclusion

KMeans clustering can tell us many useful facts about all kinds of datasets and help us to differentiate between groups in data. On the NBA data I used, I was able to accurately identify a position with one cluster and define clusters of star players, role players, and prospects. Grouping players together like this can help a fan identify what a player's role is and identify players that tend to get similar statistics. This was all done using the most basic box score metrics which are easily available to any curious basketball fan. Using different metrics, such as shot frequencies, could help to create groups of players that could be useful for both fans looking to get a greater insight into the game and teams looking for players that fit certain needs.

Bibliography

Aryanto, R. R. (2021, April 1). *Clustering NBA Player using K-Means*. Medium. https://medium.com/nerd-for-tech/clustering-nba-player-using-k-means-7b568830edfd

Keita, Z. (2023, February 24). *How to Perform KMeans Clustering using Python*. Medium. https://towardsdatascience.com/how-to-perform-kmeans-clustering-using-python-7cc296cec092

Leung J., and Matsypura D. (2019). *Python Language Companion to "Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares"*.

Vinco, V. (2023, April 23). *2022-2023 NBA Player Stats*. Kaggle. https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular?resource=download

Wikipedia contributors. (2023, February 22). Elbow method (clustering). In *Wikipedia, The Free Encyclopedia*. Retrieved 16:15, May 11, 2023, from https://en.wikipedia.org/w/index.php?title=Elbow_method_(clustering)&oldid=1140943172