**Descriptive Data Mining for NBA Position and Shot Choice Evolution**

Corey St Jean

Speed School of Engineering, University of Louisville

CSE-632: Data Mining

Dr. Mehmed Kantardzic

November 18, 2024

# Abstract:

Humans enjoy physical competition and strategy which is why there are many sports leagues across the world with millions of fans. One of the most popular sports leagues in the United States is the National Basketball Association (NBA), a league of 30 teams each made up of 15 players that compete through the sport of basketball. To gain a better understanding of the NBA’s evolution, descriptive data mining techniques like comparative analysis, clustering, and pattern discovery were applied to two datasets for the 2004 and 2024 seasons. Results gained from techniques show an increase in three-point shooting percentages for all positions as well as a merging of position clusters based on shooting trends. This shows that more positions are becoming skilled across the board in basketball. Centers had a huge increase in three-point shooting. Mid-range shots decreased, and close-range shots stayed consistent across the years.

# Problem Description:

The National Basketball Association (NBA), a league of 30 teams, has underwent significant changes over its storied history. It is very clear to see a stark difference between early NBA games and modern ones but tracking what specifically has changed is not so simple, especially over shorter time periods (20 years). To showcase the biggest changes that the NBA has experienced in its most recent 20 seasons, data mining can be employed. Using NBA statistics on shot data collected each season, we can see how positions, players, and the league in general has changed, but the specific source of the data as well as its preprocessing is critical to ensure that the revelations reflect the real changes that have occurred.

The data to be used for descriptive data mining comes from the NBA website but has been compiled together by another source and uploaded to GitHub. The GitHub contains data for every season from 2004 to the most recent 2024 season, but only the 2004 and 2024 seasons will be used for a comparative analysis. The datasets contain twenty-six features with over 250,000 samples per feature (Samangy, 2024). The self-explanatory features include TEAM\_NAME, PLAYER\_NAME, POSITION\_GROUP, POSITION, HOME\_TEAM, AWAY\_TEAM, these features contain strings that simply hold the information described in the feature title. There are four other categories that these features can be grouped into, these categories are numeric, categorical, binary, and temporal. The temporal attributes include GAME\_DATE, QUARTER, MINS\_LEFT, and SECS\_LEFT containing temporal data for date of game, as well as shot specific information about the time remaining in the quarter as well as the specific quarter. The binary attributes include SEASON\_1 and SEASON\_2 which indicate which season occurring during that specific calendar year the information is from, EVENT\_TYPE which indicates through characters if the shot was made or missed, SHOT\_MADE which is the same as EVENT\_TYPE but with Boolean values, and SHOT\_TYPE which can be 2PT or 3PT. The categorical features include features like ID’s (TEAM\_ID, PLAYER\_ID, GAME\_ID) as well as ACTION\_TYPE describing how the shot was attempted, BASIC\_ZONE indicating where the shot was taken, ZONE\_NAME describing in more detail where the shot was taken, ZONE\_ABB which is a simplified zone title, and ZONE\_RANGE containing ranges in feet from the basket. Finally, the numeric features include LOC\_X and LOC\_Y which are the coordinates that the shot was taken from ranging from 0 to 50 and SHOT\_DISTANCE which is the distance in feet that the shot was taken from. These datasets are stored as CSV (comma-separated values) files separated by year, meaning that two datasets are to be explored in total.

An example entry for the dataset could be a shot made by Bradley Beal. The entry would first store the date, season, team information, and player information for the shot which is shown below, then the event and action type indicate if the shot was made or missed as well as how the shot was performed which in this case would be “Made Shot” and “Jump Shot” respectively with SHOT\_MADE simply reflecting “Made Shot” but with True instead. Then specific information about the shot is also included like the zone, if it was a three or two-point shot, the range, X and Y coordinates, the distance and game time information like quarter and time left. This entry is included below as an example, but with some features excluded for readability:



The Git repository that contained the data also has an example of a possible use case in R. The R script included generates two graphs that show a basketball court with the various locations that shots were taken from graphed on said court. Both graphs show the same information but one for the Philadelphia 76ers in a specific game, and one for Joel Embiid, center for the 76ers, alone. The shot markers are also color coded with green indicating a made shot and red indicating a missed shot (Samangy, 2024).

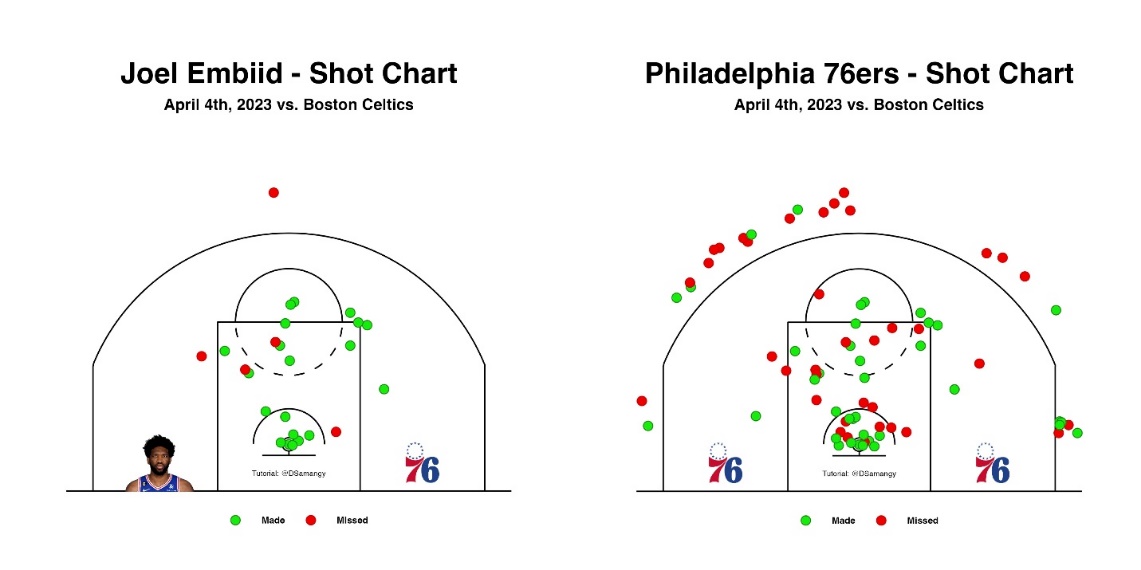


Figure II.1 Git Repository Visualization (Samangy, 2024)

Analysis of this data was also located on Kaggle where a user used machine learning models to predict whether a shot would be made based on the various features included in the datasets (dima806, 2024). They used CatBoost, a gradient boosting library that helps further define and distinguish decision tree nodes as well as the SHAP library to gain further information about machine learning models created from the data and how they can accurately predict information (dima806, 2024). They were able to use SHAP to figure out which features are most important for prediction, and how they impact results (negative or positive). SHAP values were also included for all values to see how each specific option for input affected output (SHOT\_MADE) which was able to graph which teams, players, locations, and times had the highest success on shots and impact on the machine learning models. Based on the decision tree model created and improved with the CatBoost library, the accuracy of the model was .6380 and the f-score was .5597 indicating that the machine learning model was around a 13% increase from randomly guessing (50%) which is not great (dima806, 2024). I believe this is due to the simplicity of the machine learning model chosen for prediction, decision trees, while simple to understand, do not do a great job at accurately predicting results, especially with huge datasets with lots of features. A better model would probably be random forest or even ANN given enough computing power and computing time. The upside of the decision tree model was that it was able to quickly obtain results.

# **Software Tools:**

For preprocessing and data mining, the R language was used along with specific libraries that add methods for data mining. R is a statistical and data processing language with various data mining and statistical libraries added to improve use (Robinson & Burns, 2024). The libraries used in this project include dplyr, tidyr, ggplot2, factoextra, Rtsne, cluster, and factoextra. The first library, dplyr, is a data manipulation library adding methods for easily adding variables from calculations and other functions. Tidyr is a data cleaning tool that sets a consistent structure for data allowing for easier manipulation of said data with other libraries. Ggplot2 is the library used for visualizations and graph creation as well as manipulating graph elements. Factoextra is a simplified visualization library that is used for elbow visualization and PCA (Principal Component Analysis) graphs in a manner that is easy to understand. Rtsne is used for t-SNE (t-distributed stochastic neighbor scaling) and finally cluster is used for clustering algorithms, specifically k-means clustering.

The t-SNE algorithm is a method for simplified visualization of feature dense data by combining probabilities for similar datapoints. T-SNE attempts to minimize the Kullback-Leibler divergence measurements between these joint probabilities as well as the original high dimensional data (*T-SNE,* n.d). PCA attempts to find the principal components which are described as the most important components that explain the most variance. So, the first principal component will explain the most variance and the second principal component will explain the most variance left after the first principal component (“A guide to principal component analysis…,” 2022). The elbow method is an algorithm used to calculate the optimal number of clusters to use in k-means clustering, it completes this process by calculating the Within-Cluster Sum of Squares (WCSS) for each value, k. The WCSS measures how the data is clustered around the centroid, which increases as the number of clusters increases until a certain point referred to as the elbow. The elbow point is the maximum number of clusters where the WCSS still increases, typically after the elbow point the WCSS only marginally increases (Gupta, 2024). Finally, the k-means clustering algorithm is used to group unlabeled data into clusters based on the distance from a centroid. This process is applied over the data for a set number of iterations, then the centroids that best cluster the data are used (GeeksforGeeks, 2024).

# **Preprocessing:**

To start preprocessing, we first must input the data via a read.csv command and the file path to the .csv containing the NBA dataset. After data is input, missing values can be dealt with, to see how much missing data is present in the datasets all missing values are graphed in a bar graph. For the 2004 dataset, there was no missing data, but for the 2024 dataset there were. To correct these missing values, the mean is calculated and used for numeric values while the median is used for categorical or character values. The visualizations included below show that there are no missing values for the 2004 NBA dataset and the before and after of outlier treatment via imputation for the 2024 NBA dataset.

A graph with text on it

Description automatically generated

Figure IV.1 Missing Values (2004)

A graph with text on it

Description automatically generatedA graph with white squares

Description automatically generated

Figure IV.2 Missing Values Before and After Imputation (2024)

After missing values are dealt with, outlier treatment can occur. To calculate outliers, the inter-quartile range (IQR) is calculated, then any values less than Quartile 1 minus 1.5 times the IQR are outliers below the standard and any values greater than Quartile 1 plus 1.5 time the IQR are outliers above the standard. After finding outlier values, treatment can then occur through imputation. These outlier values are replaced by the median value for that column. To not skew the data incorrectly, SHOT\_DISTANCE is the only feature where outlier treatment occurs, this is because all other columns are categorical with only a few standard values possible. While the coordinates are also numeric, the outliers within this column cannot accurate be treated due to each coordinate being contingent on the associated coordinate (X coordinates depend on Y coordinates and impact SHOT\_DISTANCE). So, outlier treatment was only applied to the SHOT\_DISTANCE feature and the before and after is shown below through a histogram.

A graph of a number of orange bars

Description automatically generatedA graph of a number of orange bars

Description automatically generated

Figure IV.3 Shot Distance Before and After Outlier Treatment (2004)

A graph of a number of orange bars

Description automatically generatedA graph of a number of orange bars

Description automatically generated

Figure IV.4 Shot Distance Before and After Outlier Treatment (2024)

Next, numeric values can be normalized, but as stated above this cannot be applied to coordinates. While the data will still be accurate on a set scale, this does not help us as coordinates will not be used for any of the data mining steps. Instead, these values are essentially restating through other columns like BASIC\_ZONE, ZONE\_NAME, and SHOT\_DISTANCE. So, normalization will instead only be applied to the SHOT\_DISTANCE column once again. Additionally, all other columns are not numeric so normalization will not apply. The before and after of normalizing the SHOT\_DISTANCE column is shown below.

A graph with purple bars

Description automatically generatedA graph with green bars

Description automatically generated

Figure IV.5 Shot Distance Before and After Normalization (2004)

A graph of a normalized shot distance

Description automatically generatedA graph with green bars

Description automatically generated

Figure IV.6 Shot Distance Before and After Normalization (2024)

Now that data has been properly cleaned and normalized, we can begin engineering specific features that will help in comparison of positions and shot types over a twenty-year period. First, the success rates of shots were determined based on the specific quarter the shots occurred in to see how the success of shots changes as the game continues. Then the success rate was calculated based on specific positions. Next, temporal attributes are converted into date and time format for better temporal analysis. Finally, the average shot distance based on position was calculated and graphed in a bar plot. More features are calculated including zone usage by position, shot success rate by position and shot type, average shot distance by position and court zone, distribution of shot distance by position group, and shot frequency by position and zone group. These visualizations are included in the results section and will be used for comparison later between the 2004 and 2024 seasons to see how positions and strategies have evolved over the year.

To help later data mining steps like clustering, we can reduce the dimensions of the data as well as figure out which components are most important for grouping players. T-SNE was used to first visualize the data to see which components are most important. But prior to T-SNE, features must be aggregated based on specific players to reduce the sample size of the dataset. Additionally, the previously calculated features like average shot distance and shot success rate as well as new calculated features like average time remaining, total shots, shots per quarter, and clutch shot rate were calculated. Most of these features are self-explanatory, but clutch shot rate contains the shot success percentage for shots made in the last two minutes of the game. These features were aggregated based on player for better T-SNE analysis. T-SNE was run on the dataset with perplexity set to twenty and the number of dimensions set to two. The results from T-SNE for the first two T-SNE factors (two most important) are included below.

A grid with many black dots

Description automatically generatedA graph showing a number of dots

Description automatically generated

Figure IV.7 T-SNE Projection for 2004 (Left) and 2024 (Right)

Principal component analysis (PCA) was then conducted to help with feature reduction prior to clustering. To start, data was aggregated once again based on player, but additional features were engineered for better analysis with reduced dimensions. Three-point percentage, success percentage, then shot success rates for each basic zone were also calculated (threes, restricted area, corner threes, and midrange). After data aggregation, PCA was run and the datasets created after PCA were used in clustering as well as the elbow method.

# **Data Mining:**

Two main data mining tasks were conducted, firstly pattern discovery was conducted previously during preprocessing through the visualizations created outlining various calculated features. These visualizations are created from methods included in the ggplot library and calculations of values visualized use methods from the stats and tidyr packages. Secondly, clustering is performed to see how certain positions compare to each other when it comes to the first two principal components as well as average shot distance and success rate using methods from the cluster library. Comparison of the visualizations from the 2004 and 2024 are included below based on visualizations created from the dataset. Only a few visualizations are included as all visualizations are shown in the results section below.

A graph of different colored squares

Description automatically generated

Figure V.1 Average Shot Distance by Position Group (2004)

A graph of different colored bars

Description automatically generatedA graph of a line graph

Description automatically generated

Figure V.2 Avg Shot Distance by Position and Zone (2004) Figure V.3 Distribution of Shot Distance by Position (2004)

After pattern discovery, the elbow method was utilized to gain information on what the optimal number of clusters would be. The fvis\_nbclust method was utilized from the factoextra library to visualize the elbow method which is included below.

A graph of a number of clusters

Description automatically generatedA graph of a number of clusters

Description automatically generated

Figure V.4 Optimal Number of Clusters (Elbow Method) (2004 Left) (2024 Right)

With the optimal number of clusters calculated (four), clustering can now occur. First, kmeans was calculated from the PCA data, which should give us centroid pairs to use for clustering. This cluster information was then added to a dataset that is then grouped by position groups. Then, the position group data was readded to the PCA data as it was removed during feature reduction. This allows for visualization to change the shape of points based on position group. Then two different visualizations are included for clustering, first the clustering based on the first two principal components, and second clustering based on average shot distance and shooting success rate. This is used for comparative analysis to see how principal components affect clustering readability.

A graph of different colored squares

Description automatically generatedA graph showing different colored dots

Description automatically generated

Figure V.5 Clusters by Avg Shot Distance and Success (Left) and Clusters by Principal Components (Right) (2004)

# **Results:**

To start with our result analysis, we begin with visualizations created from the 2004 and 2024 seasons for comparative analysis, starting with shot success rate based on the specific quarter the shot was taken in. In a typical NBA game, there are four quarters, but when a game is tied as the fourth quarter ends, additional time is added classified as overtime which is why there are up to seven quarters (four normal quarters and three overtime periods). In both cases the success rate of shots decreases as the game progresses, but the 2004 season had a spike during the third overtime period, which is typically unusual.

A graph showing a line

Description automatically generated with medium confidenceA graph showing a green line

Description automatically generated

Figure VI.1 Shot Success Rate by Quarter for 2004 (Left) and 2024 (Right)

The next visualization covers average shot distance by position group. The center position stayed consistent between the years, but forwards and guards both had an increase in average shot distance, showing that those positions are shooting from farther away (threes) more often than in 2004.

A graph of a group of colored squares

Description automatically generated with medium confidenceA graph of different colored squares

Description automatically generated

Figure VI.2 Average Shot Distance by Position Group for 2004 (Left) and 2024 (Right)

Next, the specific shot types per position group are graphed below. This shows, out of all shots taken by each position, the proportion of two-point to three-point shots. For all positions, the amount of three increased drastically, proving what was suggested from the previous visualization. Centers had the highest increase (almost 10x increase) followed by forwards (~2x increase) and lastly guards (~1.75x increase).

A graph of a group of green and orange bars

Description automatically generatedA graph of a group of green and orange squares

Description automatically generated

Figure VI.3 Shot Type Distribution by Position Group for 2004 (Left) and 2024 (Right)

Next, we look at how shot success rate (shots made) compares between the years. Surprisingly, this shows a decrease in almost all positions, but only marginally. The highest decrease was in the forward position group, then centers and guards had only a margin decrease if any at all. This does make sense though, because as these positions take more threes than typical and as shot distance increases, the probability of a shot being made decreases.

A graph of a rate by position

Description automatically generatedA graph of a bar chart

Description automatically generated with medium confidence

Figure VI.4 Shot Success Rate by Position for 2004 (Left) and 2024 (Right)

Now this next visualization is a combination of two of the previous graphs just to show their specific relationship and how they individually impact shot success rate. Centers had an increase in shot success for both twos and threes, showing that centers are becoming more versatile and skilled. This is also the case for forwards and guards as well, meaning that more shots in general are going in than in previous years. This could be due to a number of factors, some of which are more difficult to measure like defensive abilities and how much a defender can mentally affect a shooter shot success rate.

A graph of a bar graph

Description automatically generatedA graph of a bar graph

Description automatically generated with medium confidence

Figure VI.5 Shot Success Rate by Position and Shot Type for 2004 (Left) and 2024 (Right)

The next visualization takes the average shot distance visualization from earlier, and groups them by court zone. This was one of the most surprising results, showing a decrease in shot distance for most position groups and areas except for backcourt shots and above the break threes. This means that players are shooting from further where they can, which is typically above the break, but are shooting a bit closer when it comes to midrange, corner threes, restricted area, and in the paint. Once again, the exact reason why could be due to a number of reasons, but an educated guess would be that players are being guarded more heavily on the three-point line due to an increase in three-point shooting, while defense in the closer zones is less focused on, allowing players to shoot from closer.

A graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generated

Figure VI.6 Average Shot Distance by Position and Court Zone for 2004 (Left) and 2024 (Right)

The next visualization simply shows the distribution of shot distance based on position group. This graph is my opinion does a much better job at showing how much three-point shooting has increased over the past twenty years with an easily noticeable spike in shots around the three-point line and a slight decrease in shots right under the basket.

A graph of a line graph

Description automatically generatedA graph of a line graph

Description automatically generated

Figure VI.7 Distribution of Shot Distances by Position Group for 2004 (Left) and 2024 (Right)

This outlines next visualization, shot frequency by position and zone range, shows the percentage of each type of shot being taken based on the position group taking the shot. Once again, we can see that the center position has drastically increased the number of shots taken further away. The guard position surprisingly averaged out to all zones being consistent in the proportion taken by guards, and the forward position seems to be taking essentially the same amount, with a slight increase in further shots like the center position.

A graph of different colored squares

Description automatically generatedA graph of a number of different colored squares

Description automatically generated

Figure VI.8 Shot Frequency by Position and Zone Range for 2004 (Left) and 2024 (Right)

Our final visualization for comparative analysis of positions between the 2004 and 2024 seasons shows the shot zone usage by position group. This visualization is similar to the one above but uses specific court zones instead of shot distances. This shows that, like shown above, the number of threes taken by all positions has increased, and the area seemingly taking the hit from more three is the midrange. I think this could potentially be because players, if they are taking a ranged shot, prefer to get three points instead of two for a similar difficulty of shot. This also shows how much confidence in three-point shooting has increased since players are making this decision across the board, regardless of position group. Close shots (Restricted Area and In the Paint) are consistent.

A graph of a group of people

Description automatically generated with medium confidenceA graph of a group of colored squares

Description automatically generated with medium confidence

Figure VI.9 Shot Zone Usage by Position Group for 2004 (Left) and 2024 (Right)

Now, for our clusters, this gives us information about what players are like each other with the shape of the icon outlining the position group the player belongs to. These first two clusters show clustering based on average shot distance and success rate. These clusters will also be used for a comparative analysis between clustering based on PCA and based on features within the datasets. The first item of note is how much the success rate has increased for almost all players regardless of distance, but close shooters had a higher success rate than further shooters. Additionally, we can see that the most successful close shooters tend to be centers and the most successful far shooters tend to be guards, but this is only the case for the 2004 season. The 2024 season still had centers with the highest success rate with close shots, but the other clusters have no consistency when it comes to the positions making them up, with some centers even making up the rightmost cluster (far shot distance). We can additionally see that clusters based on normal dataset features are difficult to differentiate and tend to blend together, making it much harder to extract information from clusters created.

A graph of different colored squares

Description automatically generatedA graph of different colored triangles

Description automatically generated

Figure VI.10 Players Clusters by Avg Shot Distance and Success Rate for 2004 (Left) and 2024 (Right)

The next set of clusters is based on the first two principal components calculated during preprocessing. While it is harder to extract exact shooting trends when using principal components, it is much easier to differentiate clusters as they are spaced out better and have much less overlap. This is the biggest benefit of using PCA prior to clustering, as without it the clusters are almost unreadable for most. We can see though that clusters grouped together more during the 2024 season instead of the 2004 season. This is most likely because centers are becoming more skilled, like outlined previously, with more three-point shots. But the centers are still almost all grouped together. An interesting observation is that the location of the cluster containing the most centers seemed to change location from the rightmost (green cluster in 2004) to the leftmost (purple cluster in 2024). But, in both graphs the other clusters seem to be a fairly even mix of forwards and guards, with a few centers mixed in as well. This idea is also perpetrated through positions, specifically guards and forwards being much more similar with players often changing between guards and forwards based on the team’s current needs. A great example of this is the player Lebron James who has played every single position during at least one game during his NBA career. Although, it should be noted that Lebron James is a unique, generational player often considered one of the best of all time to play the sport.

A graph showing different colored dots

Description automatically generatedA diagram of different colored triangles

Description automatically generated

Figure VI.11 Player Clusters Based on Principal Components for 2004 (Left) and 2024 (Right)

To summarize some of the biggest findings we first begin with three-pointers. All positions seem to be taking more three-pointers during the 2024 season. This could be from several reasons, but some of the most plausible include an increase in overall skill within the NBA or from a change in strategy employed my many teams. In all honesty, both reasons are probably factoring along with other, unknown factors. Another finding is the drastic change in the role of the center. Centers used to take almost exclusively close shots where they can use their height and weight to get close to the basket, but now centers are hitting shots from almost every area including beyond the three-point line. The final major finding is how much players are getting more like other positions with many having the ability to switch between positions based on a team needs. This is shown in the PCA clusters, the clusters from the 2004 season are more distinct with certain positions and players utilizing specific strategies. But in the 2024 cluster, there seems to be a more homogenous feel, with almost all positions having a decent amount of overall between clusters. Some minor findings unable to be proven from these datasets could be a decrease in defensive skill for three pointers and an increase in defensive skill for midranges. Midrange shots had the highest proportional decrease to account for three-pointers increasing, which could simply be because players are much more likely to successfully defend and prevent shots from being taken in the midrange. Another, counter point could be that defense is getting worse at protecting against threes in general, leading to teams taking advantage of this diminishing defense by increasing the number of three-point shots taken.

# **Conclusions:**

Based on the preprocessing and data mining process, we can see that the NBA is a constantly changing and evolving league. Visualization showed that the NBA of twenty years ago is starkly different than the current NBA, but the specific reasons are unknown, at least from this dataset. To continue this study, additional future work could be conducted including an analysis of defensive stats, or an increase in the scope of the study to include every NBA season between these seasons to see how these features have changed when it comes to time. This could also show when this change occurred specifically, allowing for better and more accurate hypotheses to be made as to why. Additionally, exploration of this dataset could be conducted differently. One could change all the categorical attributes into numeric ones and use an artificial neural network to analyze shots. Another expansion of this study could be an analysis of shooting decisions to see how the specific location a player is in impacts shoot success via random forest or decision trees.

# **Appendix:**

*A guide to principal component analysis (PCA) for Machine Learning*. Keboola. (2022, April 2). https://www.keboola.com/blog/pca-machine-learning

dima806. (October 2024). *NBA made shot prediction CatBoost+SHAP.* [Computer Code and Analysis] Retrieved from   
https://www.kaggle.com/code/dima806/nba-made-shot-prediction-catboost-shap/notebook

GeeksforGeeks. (2024, August 29). *K means clustering - introduction*. https://www.geeksforgeeks.org/k-means-clustering-introduction/

Gupta, A. (2024, November 2). *Elbow method for optimal value of K in kmeans*. GeeksforGeeks. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/#what-is-the-elbow-method-in-kmeans-clustering

Robinson, S., & Burns, E. (2024, May 31). *What is the R programming language?: Definition from TechTarget*. TechTarget. https://www.techtarget.com/searchbusinessanalytics/definition/R-programming-language

Samangy, Dominic (May 2024). *NBA Shots 04 – 24* [Dataset] Retrieved from https://github.com/DomSamangy/NBA\_Shots\_04\_24

*T-sne*. scikit. (n.d.). https://scikit-learn.org/1.5/modules/generated/sklearn.manifold.TSNE.html