Analyzing and Predicting Offensive Performance in the NBA

Vagmi Bhagavathula
Jake Friend
Jordan Wecler
Kyle Lusignan
Bishoy Soliman
Danilo Ruberti

Abstract:

Our project centered around predicting and analyzing offensive performance in the National Basketball Association (NBA). Our first statistical approach was to train a regression model with the three-point percentage as the independent variable and total points as the dependent variable. Our second statistical method was to develop a decision tree that determines whether a team scored above 100 points for the game, depending on box score statistics. Both our statistical approaches achieved a 70 percent threshold in training data and the remaining 30 percent for testing data.

Background:

In the NBA, three-point shooting has become a necessity in recent times. Driven by analytics, teams have been encouraged to shoot the ball behind the arc. The average points per shot increases despite the few feet difference from the point of the jump. To test whether this theory holds, we took the datasets of the Golden State Warriors from 2015-2016. The dataset contains the game log of all 82 games, their corresponding box score statistics (points, rebounds, assists, etc.), and efficiency from each part of the court.

2015 was a year where the three-point revolution garnered mass media attention. We decided to move forward with game logs from that season because prominent players like Stephen Curry and Klay Thompson initiated the movement. Our team was also curious about what other features had an underlying relationship with the total points scored in a game, let alone predict their outcome. We measured a random team like the Boston Celtics to predict their offensive performance from diverse quant.

Approaches:

With our first goal being predicting how many points a team will score based on their 3p%, it was clear-cut that linear regression was a good fit. Because we're looking to measure the correlation between 2 variables, this was a solid opportunity to use one to attempt to predict the other. To test this, we took the Golden State Warriors 2015-16 year statistics for points scored per game and 3p% per game. Using the first half of the season, we put the data into a scatter plot and gain a line of best fit.

However, our second goal was predicting a team's total points scored in a game based on several variables, which we knew was more complex than a linear regression line. After conducting PCA analysis, we determined shooting splits, such as free throw percentage, field goal percentage, and three-point percentage, had the strongest relationship with points scored.

We determined these approaches best because they observed various offensive metrics as features under different models. A regression model fits far better than a classification model because we measured the trend in points scored based on quantitative variables. We considered other approaches and datasets to pursue when testing this theory. We initially attempted to download a library from GitHub containing every box score from each game leading since the 1976 league's official institution. We also attempted to produce probabilistic values for winning or losing depending on box score statistics. However, these approaches failed because downloading the library directly from GitHub created an overflow in data; determining the probability of winning and losses yielded inconsistent visualizations.

Before we proceeded with our statistical approaches, we had to ensure the data was cleaned and restructured, such as converting categorical values to a binary value (0 or 1). Considering every game out of all 82 rows was played, there was little reason to remove certain rows. However, we removed unnecessary attributes within the tuples to simplify our models as much as possible.

Results:

Our results bolstered our initial hypothesis. It provided us more insight into the importance of the three-point shot toward total points scored. Our visualizations highlighted a positive relationship between the three-point percentage and the total points scored.

Linear Regression Approach

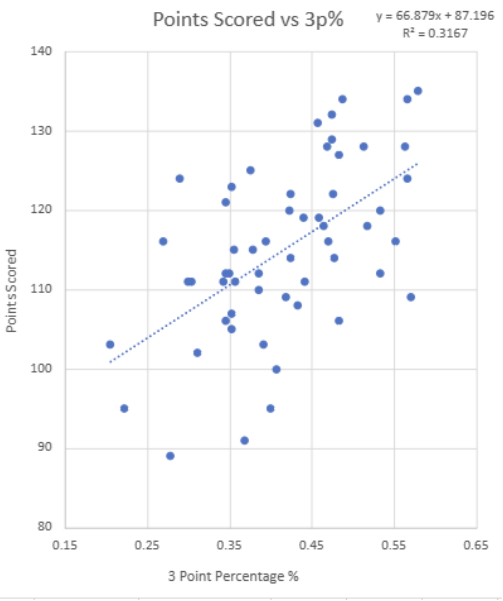

Figure 1: We provided the model with 70% of our dataset (56 games out of 82) to train. The figure shows the positive relationship between the three-point percentage and points scored.

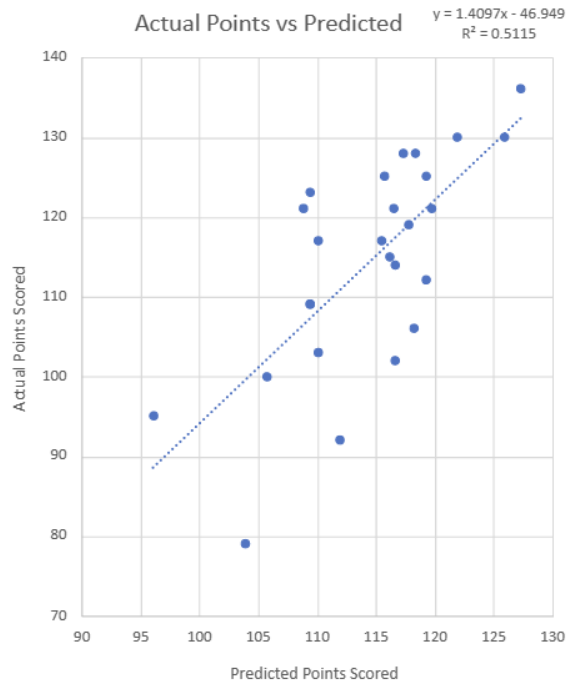

Figure 2: Here's an attached graph of the predicted points scored vs the actual points scored. The values deviate from each other, but the predicted values are not far off.

PCA Approach:

As per the PCA approach, we normalized and scaled the datasets for the Celtics and the Warriors. We retrieved the correlation matrices (figs 3 & 8) for each dataset and looked at the cumulative proportion by applying PCA analysis. As per the principal component, we included more visualizations to understand which variables were important. Such plots we implemented were the Scree plot, Biplot of the attributes, and Biplot combined with cos2 (value attributed as the importance of a principal component for a given observation). We produced the regression tree for the selected predictors. However, we understood the inconsistency that came with the regression trees, had we based winning or losing as the response variable. So, we elected team points as our response variable. We also decided to see the X-relative error in the model so we can have a better judgment on our work.

Celtics Visuals:

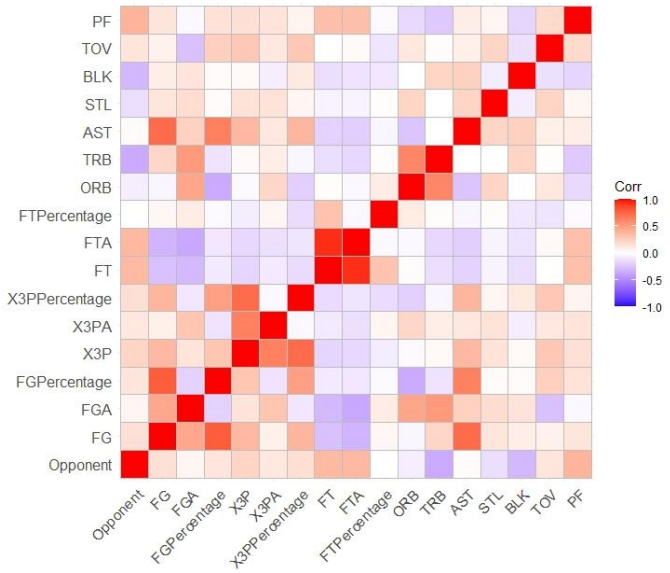

Figure 3: Correlation matrix for the Celtics dataset.

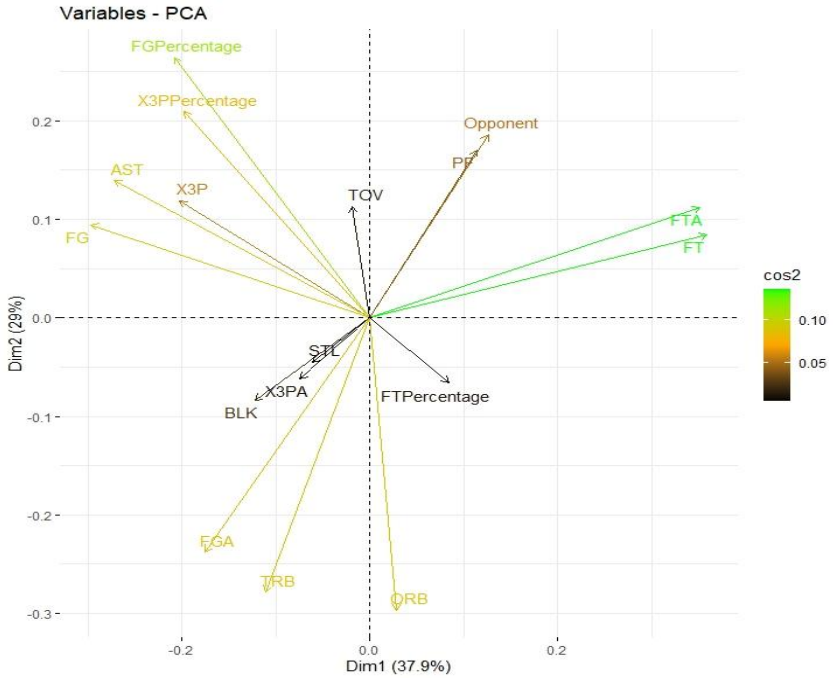

Figure 4: PCA Biplot for the Celtics that highlights the attributes and their corresponding cos2 value. The graph suggests attributes such as field goal percentage, free throws, and free throw attempts are of the most importance

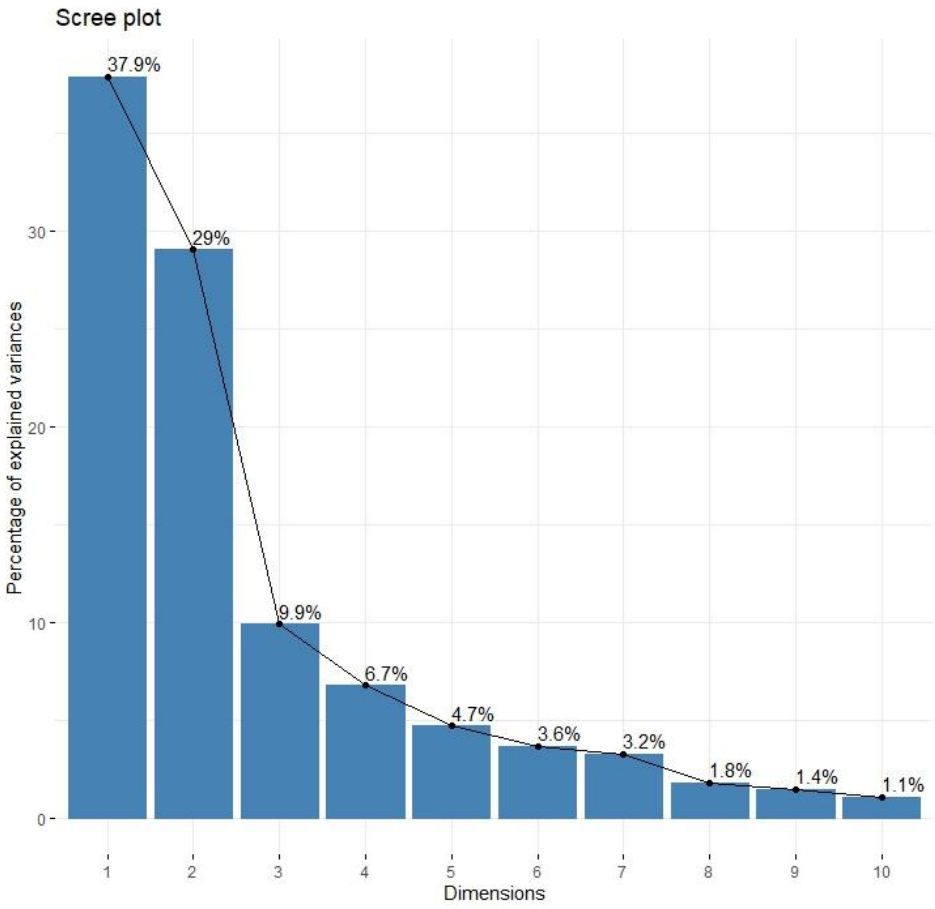

Figure 5: Scree plot that observes the principle components of variables and their percentages

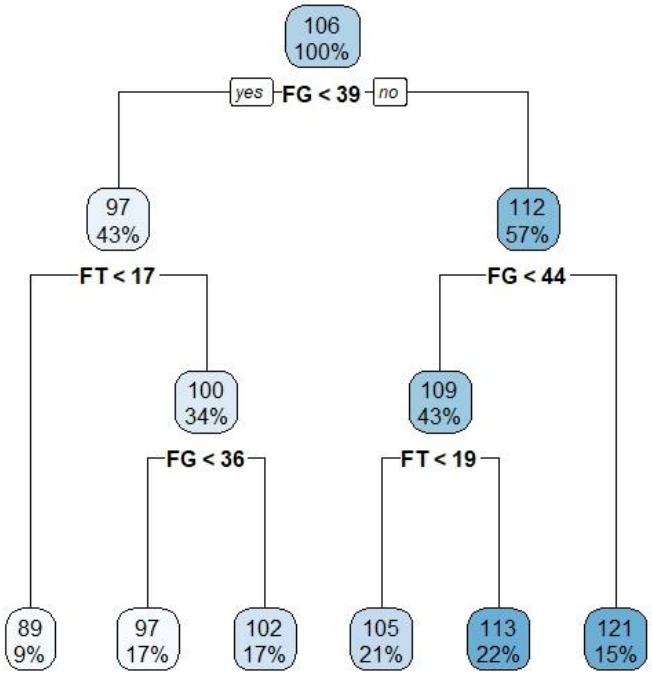

Figure 6: Regression tree of the Celtics that provides the probability of how many points scored depending on field goals, field goal percentage, free throws, etc.

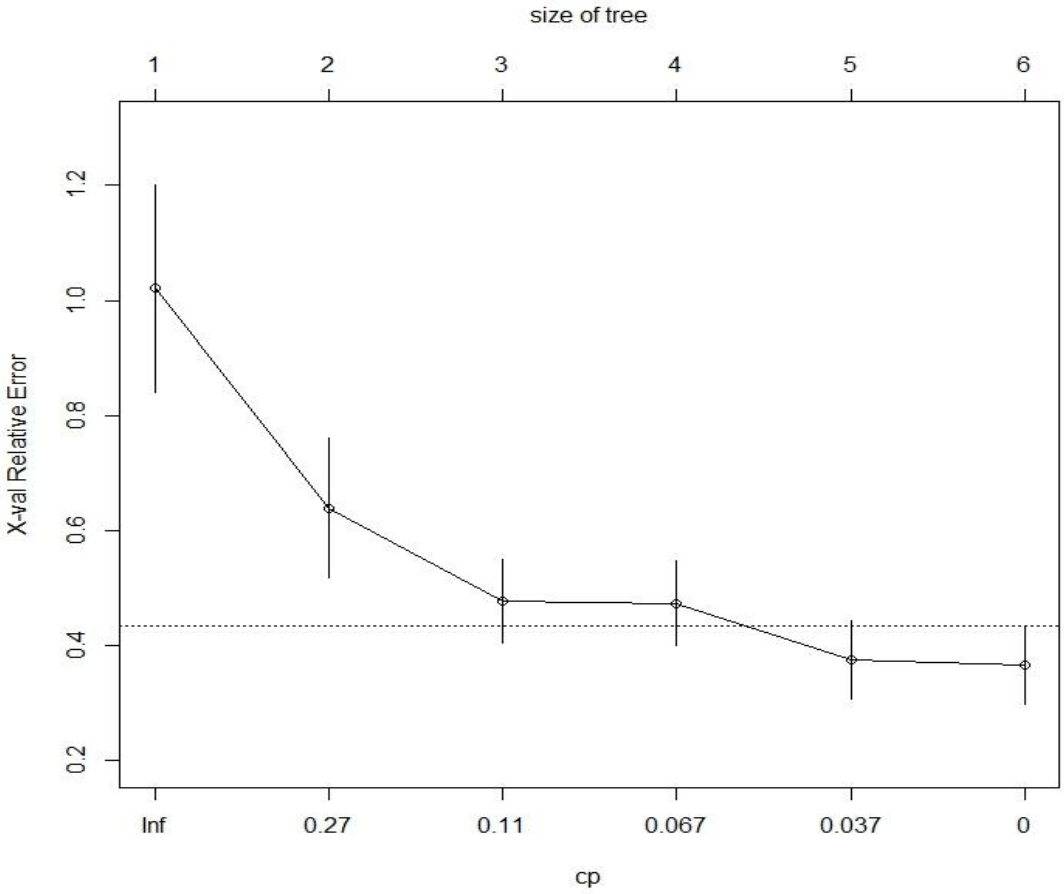

Figure 7: X-relative error plotted for the Celtics dataset. The dashed line is set at the minimum xerror + xstd. The top axis shows the number of splits in the tree. And the model shows at what level we should prune the tree.

Warriors Visuals:

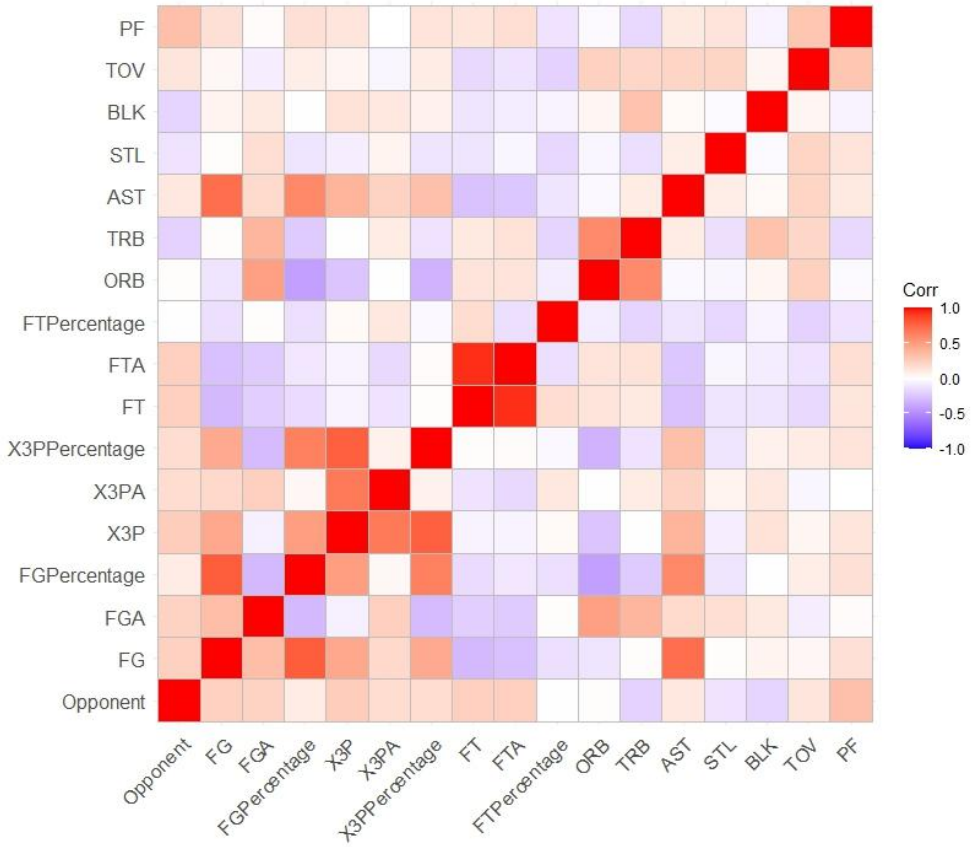

Figure 8: Correlation matrix for the Warriors dataset.

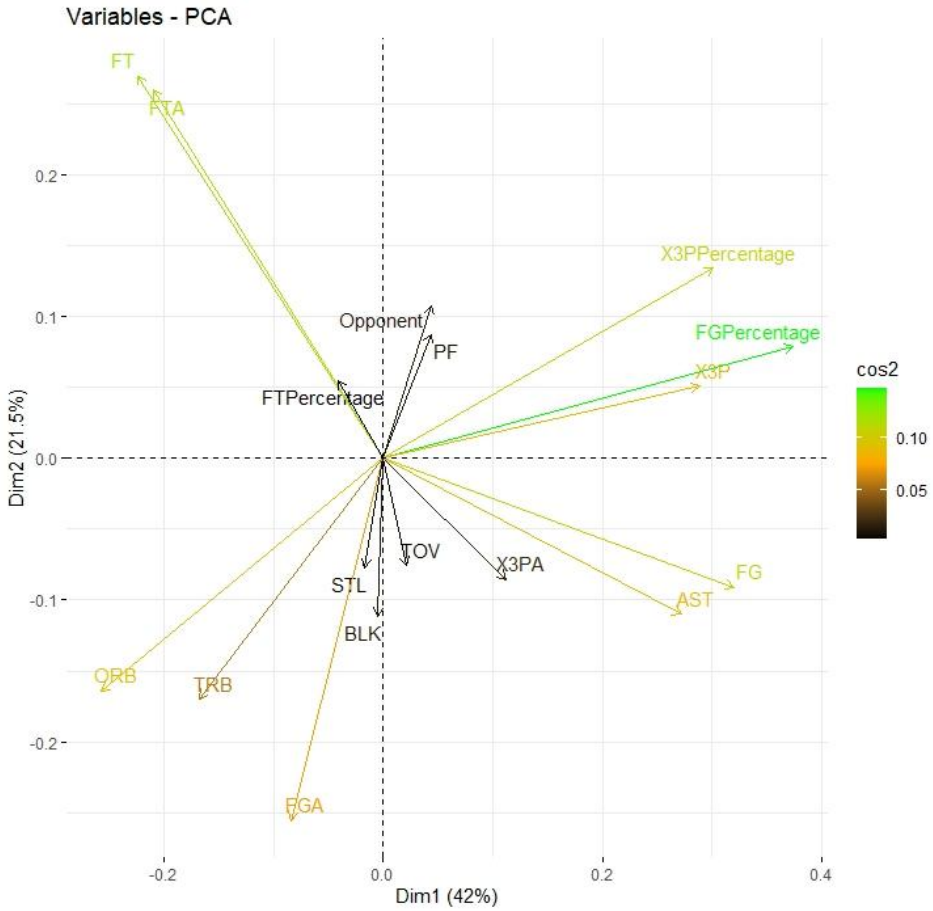

Figure 9: PCA Biplot for the Warriors that highlights the attributes and their corresponding cos2 value. The graph suggests field goal percentage, free throws, free throw attempts, etc. as the most important

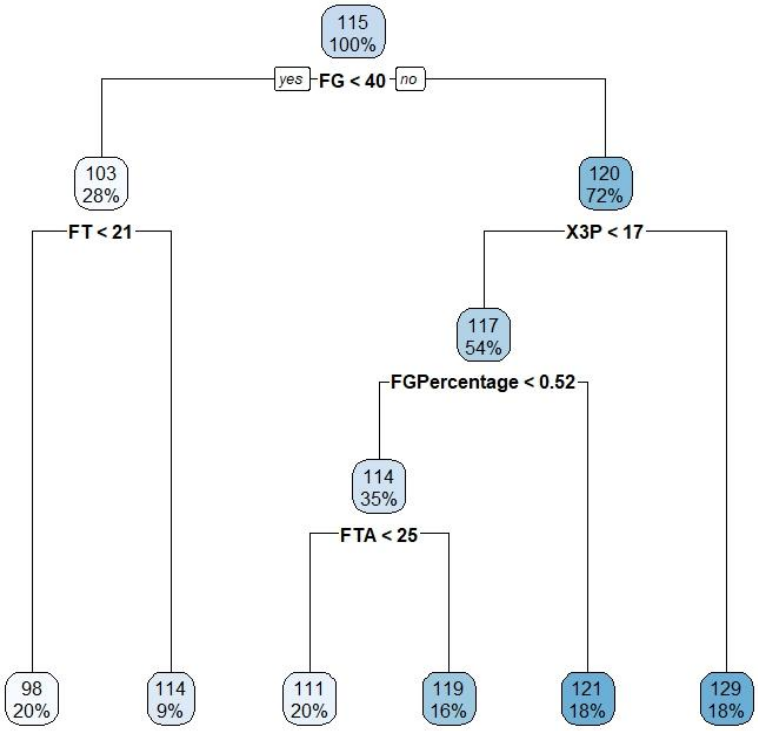

Figure 10: Regression tree of the Warriors that provides the probability of how many points scored depending on field goals, field goal percentage, free throws, etc.

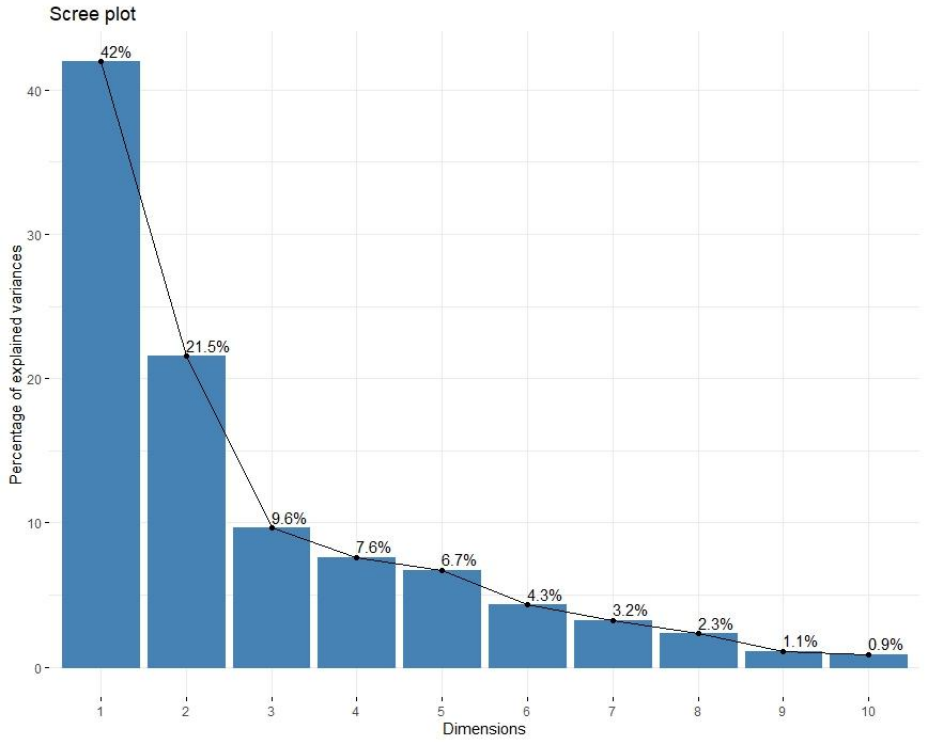

Figure 11: Scree plot of the Warriors dataset that observes the principle components of variables and their percentages

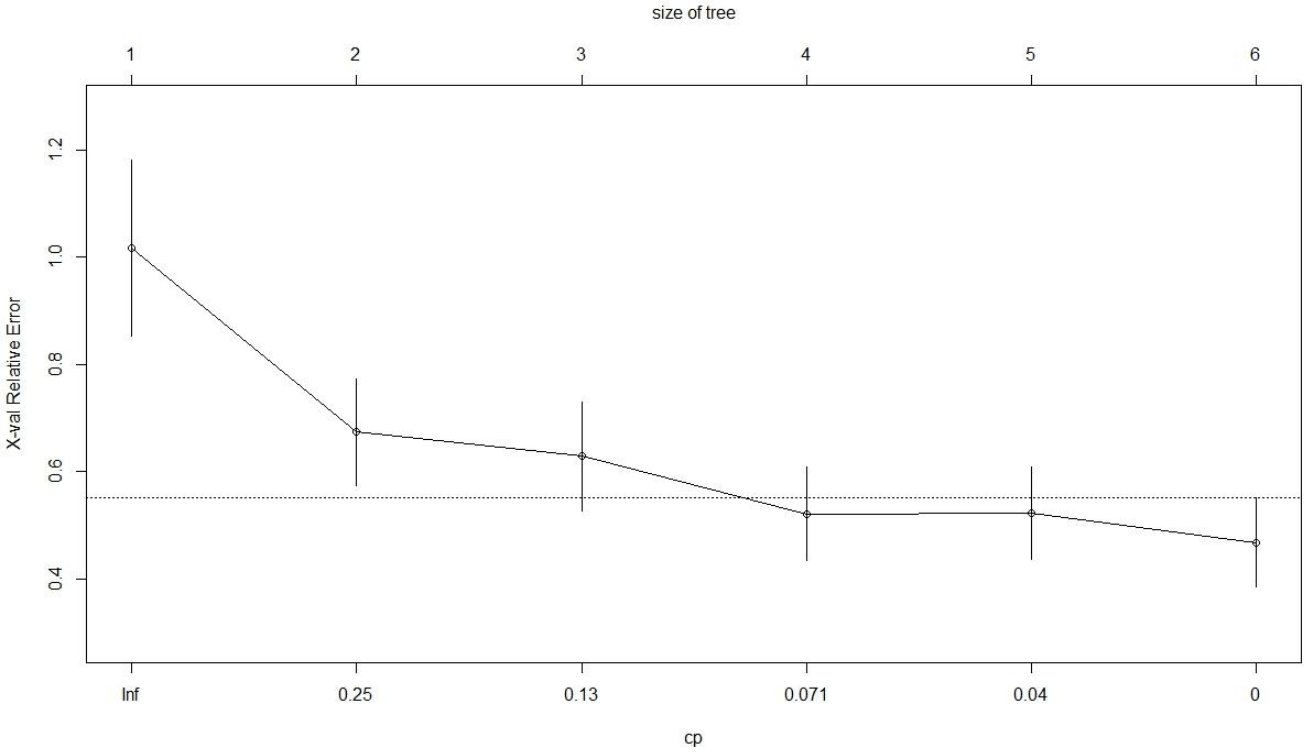

Figure 12: X-relative error plotted for the Warriors dataset. The dashed line is set at the minimum xerror + xstd. The top axis shows the number of splits in the tree. And the model shows at what level we should prune the tree.

Conclusion:

In conclusion, we determined the total points scored to have a positive relationship with a three-point percentage. Efficiency plays a role in predicting total points. After developing the regression tree, we found that we can predict the number of points acquired based on certain decisions or factors for each dataset. After conducting both statistical approaches, we detected both benefits and drawbacks to developing either model. While linear regression models are simpler to train than a regression tree, they are more susceptible to noise and variability. Also, linear regression models assume a direct linear relationship between the independent and

dependent variables, which is inaccurate given the several attributes within this data set. Regression trees produce more precise values than linear regression models but are hard to develop. We struggled to train a regression tree to output the probability of a win or loss, considering it took in several factors and created inconsistent graphs. Our research was very informative, as we applied datasets to prove the scoring revolution currently present in the NBA.