Ravinit Chand

Predicting Salary In The NBA Using Statistical Models

Abstract:

This project aims to predict the salary for National Basketball Association (NBA) players based on five of the most basic yet popular statistics in basketball: points, rebounds, assists, steals, and blocks. A Kaggle dataset from the 2022-2023 NBA season is used, with over 50 explanatory variables. The models in this project involve three Ordinary Least Squares (OLS) methods, Least Absolute Shrinkage And Selection Operator (LASSO), and a regression tree. Using test MSE as the measurement, the model with the lowest test MSE will be the most optimal, and will also be tested using the median and mean values to see how far the prediction for salary is from the actual median and mean salary value. The significance of this project revolves around a better understanding of the methods that are the best for predicting the salary for NBA players, as players of diverse skill sets make a wide range of money in the NBA.

Introduction:

Athletes are amongst the highest-paid professionals in the world, with players in soccer, football, and baseball easily earning millions of dollars per season. NBA players make a significant amount as well, with player archetypes ranging from small point guards to tall centers making millions of dollars per season. The billion-dollar TV deals the NBA gets along with sponsorships have allowed some players to make hundreds of millions during their playing career. With the diversity of talent in the NBA, it's hard to find an objective number as to how much annually a player should receive for their contributions. In the NBA, the five most popular

statistics are points, rebounds, assists, steals, and blocks. Using these statistics, this project will look to answer the question "What statistical model is the best for predicting salary in the NBA?"

Similar research articles have also utilized statistical methodologies to compare NBA players to their salaries. One research article titled "The Prediction of NBA Player's Salary" looked into the R-squared between LASSO, Ridge, Elastic Net, and Random Forest models for each of the five positions in the NBA. R-squared is the amount of variance in the response variable that is explained by the explanatory variables, and a higher R-squared implies that more variability is explained in the explanatory variables. This paper is similar due to how comparing different models was conducted for research about the NBA, with the result being that Ridge and Elastic Net had predictions that were more accurate due to a higher R-squared (Lu, 2024). Here, R-squared was the measurement that was implemented to see what model was the optimal choice, which is what another article did as well.

In this article titled "A Non-Linear Approach to Predict the Salary of NBA Athletes using Machine Learning Technique," RMSE was also used along with R-squared to compare how Multiple Linear Regression, Decision Tree, XGBoost, and Random Forest did when various explanatory variables were used to predict salary. RMSE is just the square root of MSE, which is a metric that compares the squared differences between the observed and predicted values, with a smaller RMSE indicating that the gap between these two values is smaller. In this context, Random Forest had both the lowest RMSE and highest R-squared, showing this method outperformed the other (Jain, Jain, Neelu M, George, 2012).

The final research article that followed the same pattern of predicting salaries and comparing models with metrics was titled "National Basketball Association Player Salary Prediction Using Supervised Machine Learning Methods," and employed Multiple Linear

Regression, Lasso, Ridge Regression, and Bagging to compare what methodologies were the best for predicting salary. This was conducted by way of MSE, with values ranging from 0.35 to 0.71 (Özbalta, Yavuz & Kaya, 2021). These three papers outline how there is lots of evidence of comparing methods, followed up by metrics to determine what is the optimal performance method when predicting salary in the NBA using explanatory variables.

In addition to finding out what model is the best by comparing them, this project aims to see the correlation between explanatory variables such as points, rebounds, assists, steals, and blocks to the response variable, which is the salary for players in millions of dollars. A paper by Mohan Cao titled "Predicting NBA Player's Salary Based on Statistics from the Game Using Linear Regression" researched something very similar, where they found out that assists and points had a positive relationship with salary while turnovers had a negative relationship with salary (Cao, 2024).

Another paper titled "Model Prediction of Factors Influencing NBA Players' Salaries Based on Multiple Linear Regression" had a similar goal, where they compared fourteen explanatory variables to salary as the response variable using multiple linear regression to see the correlation between the variables. In this paper, it was found that age along with value over replacement had a strong relationship with salary, while team-related statistics didn't have a significant impact on salaries (Zhao, 2022). Both of these papers highlight how there have been research papers conducted to highlight the degree of importance between explanatory variables and salary in the NBA.

Data:

This dataset uses various statistics coming from the 2022-2023 NBA season. The original dataset is from Kaggle called "NBA Player Salaries (2022-23 Season)," which is a reliable website useful for finding datasets of several types. The owner of the dataset used Hooshype to web-scrape out the players' salaries, along with Basketball Reference to web scrape out the other basketball statistics. These are two of the most popular and accurate websites for finding basketball statistics. Within the dataset, there are a total of 53 columns, highlighting the versatility of this dataset. Additionally, there are 467 rows to spotlight the players in the dataset. More specifically, there's information about each player such as their age, position, and team that is categorical, along with more complex statistics like offensive win shares, defensive win shares, and overall win shares. With so many variables, lots of these measurements could be used to find the correlations with players' salaries; however, this project aims to use five continuous traditional statistics in basketball which are points, rebounds, assists, steals, and blocks to predict salaries for players in the NBA.

When handling the data, the salary was divided for each player by 1,000,000 to make the numbers in the dataset more simple and the regression tree easier to read. Additionally none of the salary in millions or the five explanatory variables had N/A values, which made modeling the data easier.

When running some basic summary statistics on the dataset, it was found that the average number of points from the 467 players during the 2022-2023 season was 9.13, while the 50th percentile was at 7.10, implying that points are skewed right. This was run for all five of the explanatory variables, which can be seen in Table 1 in the Appendix, along with the minimum and maximum values, and the 25th percentile and 75th percentile values.

Methodology

Within this project, various types of models will be run to see what is the most optimal in terms of predicting salary in millions using the five explanatory variables mentioned previously. The measurement used to compare the models will be the test MSE. The test MSE is formed by the data getting split into a training and a test set, with the training set used to build the model and make predictions on the test set. Afterward, the predicted values are subtracted from the actual test values, and the differences are squared to ensure all values are positive and to penalize large errors more heavily. This process is repeated for each of the inputs in the test set, and the results are summed up and then divided by the number of observations in the test data to form the test MSE. This process will be referred to as the "test MSE process" and will be constantly referenced throughout the paper as this process is applied consistently during the models.

The first model that will be run is OLS regression, which is useful for estimating the relationship between the explanatory variables and response variables by fitting a line that minimizes the sum of the squared distance between the predicted and actual values. The multiple linear regression lines follow the standard regression formula, where beta one hat through beta five hats represent the estimated coefficients for each explanatory variable, and the five X variables correspond to the five explanatory variables. Once the OLS regression model is built, the test MSE process will be conducted to get the test MSE for the OLS regression model. OLS Regression is the optimal choice over logistic regression due to the response variable here being continuous, which is suitable for OLS regression.

The second model that will be conducted is OLS regression with polynomials, which is similar to the traditional OLS regression model, but instead of just linear terms in the regression formula, the polynomial model includes exponential terms in the model. The exponential degree

for the explanatory variables is found by having a range of values, which is two through five in this project, and then constructing the OLS regression, just with the addition of exponents in this model. The usual test MSE process occurs for all four of the models, and the model with the lowest test MSE will be used as the official OLS regression with the polynomial model. For example in the exponential two model, the fitted model would only include the linear and squared terms for the explanatory variables, followed by the fitted model getting compared to the actual values in the test set. The main advantage of this model is that the polynomial model captures non-linear relationships by having quadratic and cubic variables, as opposed to the linear model which doesn't capture these relationships.

The final OLS model will include step functions, which are suitable when the relationship between the explanatory and response variables has sudden shifts, leading to different relationships at different points in the data. Here, the models vary based on the different number of steps, as models with two through five steps were tested with OLS regression. The usual test MSE process occurs for each of the four-step models, and the model with the lowest test MSE will be the OLS regression with steps model. As an example, in the step two model, the fitted model would include a model with only two steps to make predictions on the test data.

Along with the three OLS models, Least Absolute Shrinkage Selector Operator (LASSO) will be implemented as well. LASSO is a type of shrinking method that shrinks variables that have a weaker relationship with the response variable, even to the point where the coefficients become zero, making variable selection occur. This model will look to see if only including the most important variables will lead to the best model. Finally, a decision tree will be applied, which takes into account the explanatory variables, and then splits the model into an upside-down tree-like graph. Here, based on whether the players are over or under various

statistics, the model will predict how much annual earnings the player will make. Just like with the OLS models, both LASSO and decision trees follow the test MSE process, just with specific instructions for both of the models that will be explained in detail later in the project. In the end, a comparison between the test MSE for the five models will be made to see what is the optimal model to utilize. A lower test MSE is the goal, as this showcases the smallest gap between the actual and predicted values.

Results

For this project, 80% of the dataset was used as the training set, with this being 373 out of the 467 players, while 20% was the test set, which was the remaining 94 players. Before looking into the models, the correlations between the explanatory and response variables were a point of interest. Therefore, a heatmap was run to see the correlation between the five explanatory variables and the salary in millions. This is denoted by Plot 1 in the Appendix and is interpreted by a more dark red image interpreting to a more positive correlation, while a more light red image corresponds to a less positive correlation. When examining the correlation between the explanatory variables and the salary (in millions of dollars), the heatmap shows how all five of the variables have a positive correlation with the response variable by the blue color. This is intuitive because increasing any of your points, rebounds, assists, steals, or blocks in the NBA is seen as a performance boost, so having this lead to more earnings is logical. Additionally points appear to have the highest correlation with salary, followed up by assists, rebounds, steals, and blocks.

The first model run was the traditional OLS regression model, which yielded an MSE of 48.78, which implies that the squared average difference between the actual and predicted values for salary is 48.78 million dollars squared.

When running OLS with polynomial regression, the degree that had the lowest test MSE was the second degree. This led to the second-degree polynomial getting used, as the lowest test MSE implies the closest gap between the predicted test values and actual test values, which was a test MSE of 46.39. Similarly for the OLS with stepwise regression, the number of steps that had the lowest MSE was at step five. In this model, there were five different steps and four different breaks to separate the steps, which had a test MSE of 52.8.

LASSO had a couple more steps, specifically with choosing the tuning parameter lambda. For getting the tuning parameter, the dataset that was initially split into a training and test set was split even more for the X and Y. This led to the training set for X to be the five explanatory variables for only the training inputs, and therefore the test set for X to be the five explanatory variables for only the test inputs. The training set for Y was the salary in millions for only the training inputs, while the testing set for Y was the salary in millions for only the testing inputs. Additionally, to ensure that LASSO is getting used instead of Ridge Regression, an L1 penalty was utilized, with k-fold cross-validation with five folds. Cross-validation is required here to find the best lambda that will balance the complexity of the model along with the prediction accuracy. Cross-validation is a common resampling method, with k-fold cross-validation specifically being used in this project. Here, the data is split into five equal or roughly equal folds. In each iteration, four folds are used for training, while the remaining fold is used as the validation set. The four folds will be used for training, and the predictions will be made on the validation set, with this process occurring five times to ensure that each set is a validation set. Additionally, the

cross-validation error rate is found by averaging the sum of squared differences between the actual and predicted values across all folds. The optimal lambda is found by finding out which lambda has the lowest cross-validation MSE, which is used along with the training set to form the LASSO model. The higher the lambda value is, the more of a penalty is added to the less useful variables. The LASSO model will be formed using this best lambda value, along with the L1 penalty and training dataset to make the predictions. These predictions will be compared to the actual test values, with the usual test MSE process occurring subsequently. LASSO led to a best lambda value of 0.07 and a test MSE of 50.99, which was a test MSE that was higher than two of the three OLS models.

Finally for decision trees, a regression tree was fitted with the training data, with the regression tree getting used to make predictions on the test dataset. After that, the predictions are compared to the actual results, followed up by the usual test MSE process. Finally, for decision trees, the test MSE was 68.89, which was by far the highest MSE out of the five models.

Looking at the decision tree, which was labeled by Plot 2, there are multiple decision nodes, which are points in which the decision tree is split up. The first decision node splits the data based on points, based on whether a player averages less than or more than 13.2 points per game. Within that, more decision nodes occur, with the first decision node for players averaging less than 13.2 points per game being over the assists statistic, specifically if a player averages less than or more than 5 assists per game. On the opposite side, for players averaging more than 13.2 points per game, the first decision node was over if they averaged more than or less than 22.5 points per game. An interesting decision path was the one that followed less than 13 points, more than 5 assists, and less than 6.15 assists per game, which produced 27. This implies that for players that averaged between 5 and 6.15 assists and less than 13 points, they averaged $27

million. This seemed abnormally high, so after filtering out the data for players that qualify for these statistics, there were players whose performance during the 22-23 season didn't align with their pay. For example, Ben Simmons averaged 6 points and 6 assists that season, but got paid $35 million, and John Wall averaged 11 points and 5.2 assists but got paid $47 million; both of whom were extremely overpaid. The rest of the decision paths also output values that represented how much in millions of dollars they would average satisfying specific criterias. The highest value was 39.4, which implies that for players averaging more than 26.85 points per game, they made an average of $39.4 million during the 22-23 NBA season.

As mentioned, mean squared error is the squared distance between the predicted and actual values for the test observations, so having the lowest MSE possible implies that the gap between the predicted and actual values was the smallest. Due to this, the prediction model that works the best is the OLS regression model with polynomials, due to the MSE of 46.39 having the lowest value. Table 2 shows the test MSE for each of the five models.

After concluding that OLS with polynomial regression was the best model to run, the model's performance was tested using the mean and median values for the explanatory variables. As mentioned, due to the second-degree polynomial leading to the smallest error, it was chosen as the model under polynomial regression. Therefore to run the test, the mean and median values of the five explanatory variables, along with their squared values, were calculated respectively. After that, the polynomial model (which included the linear and squared terms) now used means and medians of the five explanatory variables along with their squared values to form a prediction of what the salary value is. The prediction yielded a value of 7.37 for the mean, which implies that if an NBA player was average in all five of the statistics mentioned in this study (points, rebounds, assists, steals, blocks), they would make $7.3 million. Additionally, the

prediction was 5.16 for the median, implying that if an NBA player was in the 50% percentile in all five of the explanatory variable statistics, the player would make $5.16 million.

This predicted value for mean does seem reasonable, as Plot 3 in the Appendix shows a scatterplot of players' salaries with the salary as the X variable and count as the Y variable. It's clear in the figure that the data is right-tailed, as a majority of the players in the NBA made less than $10 million during the 22-23 season. More specifically, over 73% of the players made less than $10 million, so it makes sense that if a player is average at everything, the amount they would make in millions of dollars would be small since the data ranges from 0-50 in the millions scale.

Additionally, when examining Plot 4, which employs a boxplot of the dataset for specifically the salary in millions of dollars, there was also evidence that the range of 0-50 for the millions scale is rather misleading, as the median value was between 0-10. The right whisker only went up until roughly 25, with values higher than this being outliers. This makes sense due to only 48 values, or a bit more than 10% of the players in the dataset making more than $25 million, so making over $25 million in the NBA is seen as a rarity. Due to this, along with knowing that a majority of players make less than $10 million in the NBA, a median value of $5.16 is reasonable.

Upon examining the salary more closely, the mean salary value was $8.4 million during the 22-23 season. While this isn't the same concept as the predicted value, it is logical to expect the player with the average salary in the NBA to have skills that are close to average in the five explanatory variables. Therefore since the $8.4 million mean salary value is relatively close to the $7.3 predicted salary, and considering that the scale of the data is from 0-50, the mean prediction made by the polynomial model is logical.

Additionally, the median salary value was $3.72 for the 22-23 season, and again while this isn't exactly $5.16, this is still a logical estimate as the values in the data range from 0-50 in terms of millions. Additionally, since there are clear outliers with that being players making more than $25, the mean should be higher than the median, which is the case in both the predicted salary by the polynomial regression model and the actual data.

Finally, the principal component analysis (PCA) was conducted on the explanatory variables, which reduces the dimensionality while keeping the most valuable information. PCA is useful when dealing with high-dimensional data, which reduces complexity reduction. In regards to PCA, a scree plot was run which is seen in Plot 5 in the Appendix. A scree plot is a visualization representation of how much variance is explained by each principal component, and the elbow within the scree plot is when the principal components start to level off. In the plot, it appears to happen after component two, which implies that the first two principal components are the most important, as they capture a majority of the variance in the explanatory variables. Due to this, if this project were to only focus on the dimension reduction data, only the first two principal components would be employed.

Conclusion

This project aimed to find the best method that predicted salary given five explanatory variables, which were points, rebounds, assists, steals, and blocks. The five methods were Ordinary Least Squares (OLS) regression, OLS with polynomial regression, OLS with stepwise regression, LASSO regression, and decision trees, and the estimated out-of-sample performance of comparison was test MSE. This was conducted with 80% of the NBA dataset being the training set, which was used to make predictions on the 20% training set, and the squared

average gap between the values of the actual and predicted values of the test set is the test MSE value.

The best method found by the lowest test MSE, which implies the closest gap between the predicted test values and actual test values was the polynomial regression with degree two. This was reasonable as the relationships in the data are not linear, so having a degree term is more useful. Using the best method, a test was run given the mean and median value of each explanatory variable to see if the value was reasonable.

Additionally, correlations were made to see how the strength of the relationship was between five explanatory variables and the response variable. The result was that all five variables had a positive correlation with salary in millions of dollars, with points having the highest correlation.

A limitation of this project was that all the players in this dataset were used, despite some of them playing just one game during the season. Due to that very small sample size, their statistics in that one game might not have provided an accurate representation of their true value for the five explanatory variables. A way to address this is to have a cutoff of at least thirty games played, as that is a good amount of games out of the eighty-two game NBA season, so a player's true value is usually shown thirty games in.

Further research could look into more advanced basketball statistics provided in this dataset like defensive and offensive win shares to see if the models have a better representation of predicting salary. Overall the NBA is a data-driven league and statistics are used every day, whether it's friends talking in school discussing who's a better player or millionaires on ESPN discussing who's the league's MVP. This project looked into the salary discourse of basketball, to see what type of model led to the smallest test mean squared error.

References:

Yang, Z. (2022, December 31). Atlantic Press. https://www.atlantis-press.com/proceedings/icedbc-22/125983656

Jain, A., Jain, S., Pancinovia, N. M., & George, J. P. (2022). *A Non-Linear Approach to Predict the Salary of NBA Athletes using Machine Learning Technique*. 1–5. https://doi.org/10.1109/tqcebt54229.2022.10041664

Cao, M. (2024). *Predicting NBA Player's Salary Based on Statistics from the Game Using Linear Regression*. 476–480. https://doi.org/10.5220/0012823900004547

Emirhan Özbalta, Yavuz, M., & Kaya, T. (2021). National Basketball Association Player Salary Prediction Using Supervised Machine Learning Methods. *Lecture Notes in Networks and Systems*, 189–196. https://doi.org/10.1007/978-3-030-85577-2_22

Lu, Y. (2024). The Prediction of NBA Players Salary. *Advances in Economics, Management and Political Sciences*, *57*(1), 196–203. https://doi.org/10.54254/2754-1169/57/20230732

Appendix

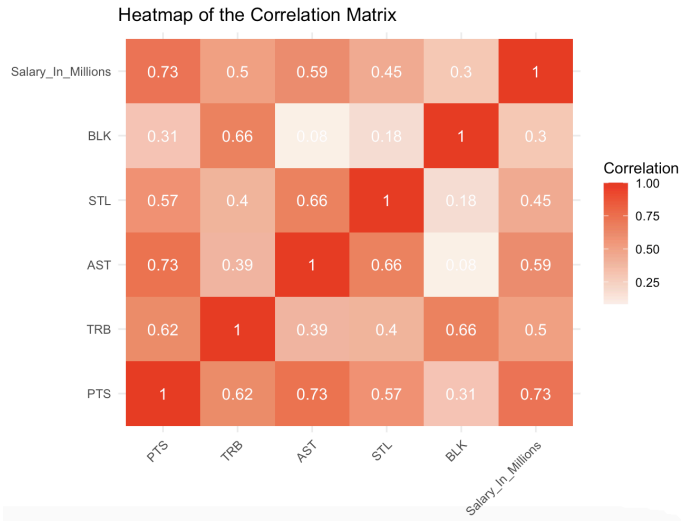
Table 1: Summary Statistics For All 5 Statistics

	PTS	TRB	AST	STL	BLK
Min	0	0	0	0	0
1st Quartile	4.1	1.9	0.8	0.3	0.1
Median	7.1	3	1.4	0.6	0.3
Mean	9.13	3.5	2.1	0.61	0.37
3rd Quartile	11.7	4.5	2.9	0.8	0.5
Max	33.10	12.5	10.7	3	2.5

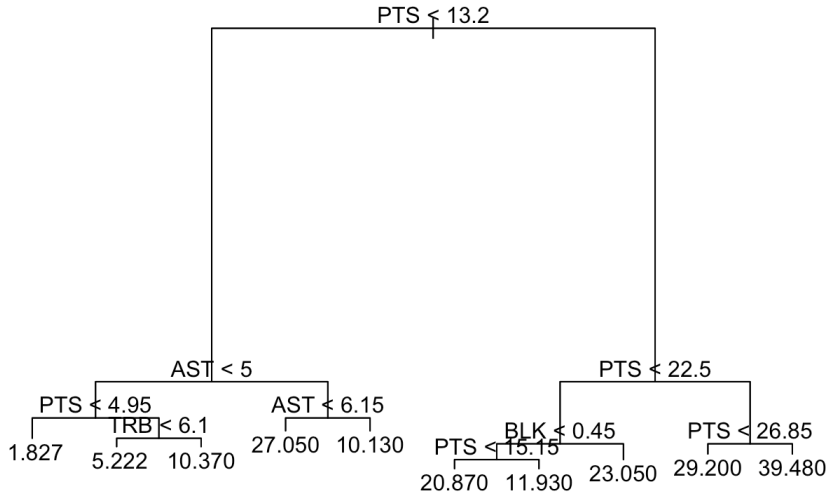
Table 2: Test MSE's Across All 5 Methods

OLS	Poly	Step	Lasso	Tree	
48.78	46.39	52.8	50.99	68.80	**Test MSE**

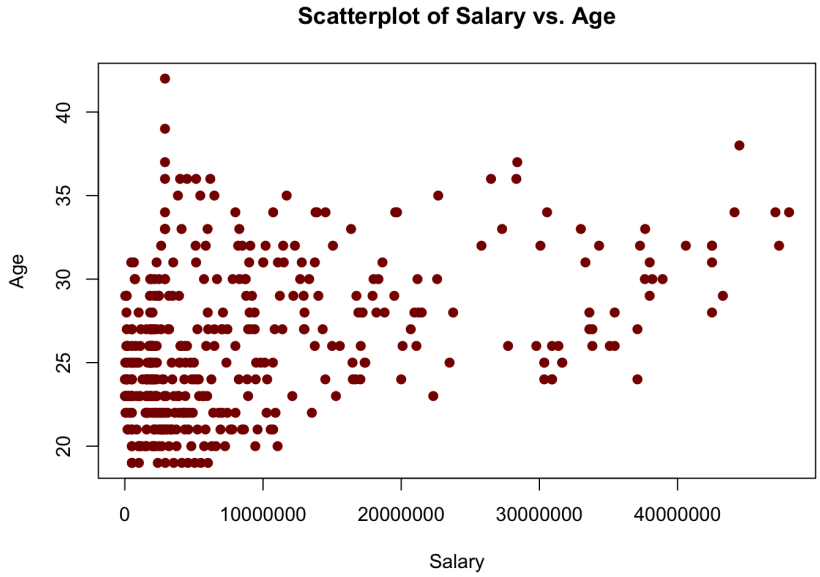
Plot 1: Heatmap Of Variables



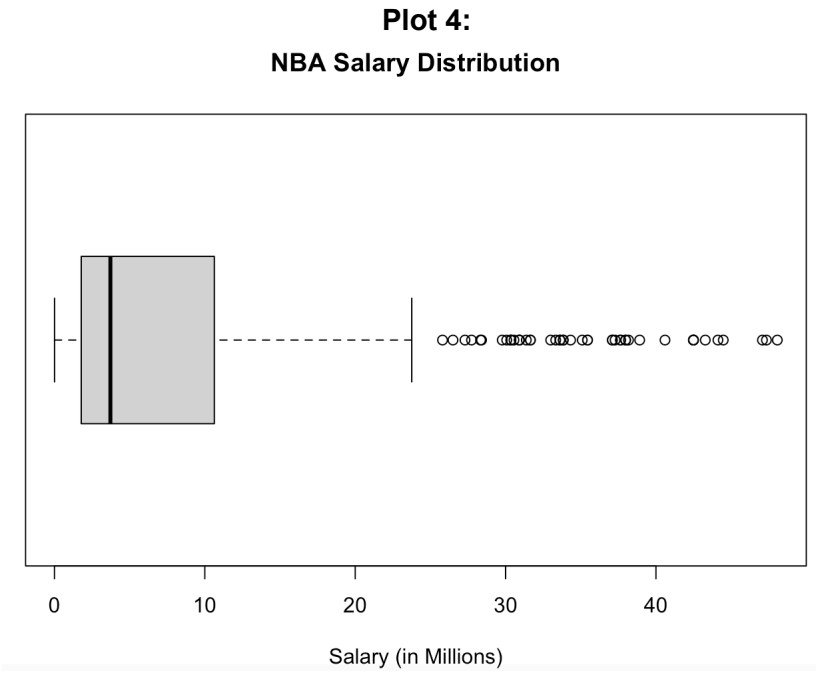
Plot 2: Decision Tree Graph



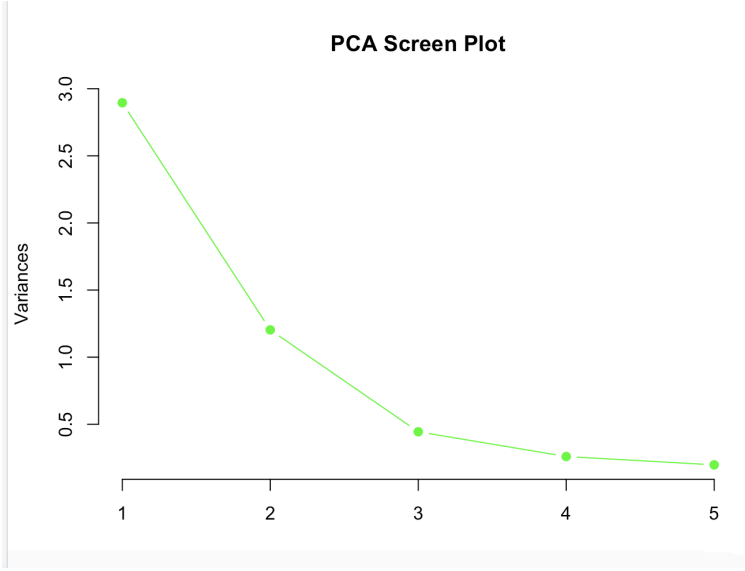
Plot 3: Scatterplot



Plot 4:


NBA Salary Distribution
0
10
20
30
40
Salary (in Millions)

Plot 5:


PCA Screen Plot
Variances
0.5
1.0
1.5
2.0
2.5
3.0
1
2
3
4
5

Dataset Link:

https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season