**Executive Summary**

The NBA is continuously looking at player salary based on their PER, points-per-game, and turnovers- per-game. Analyst can use RStudio to create predictions for future salaries of players based on these statistics and give general managers an idea of what it would cost to maintain a certain player. The predictions will help general managers decide who to draft that will keep them under the salary cap but will help their team win games and get into the championship. We used regression to model the relation between a player’s salary and their PER and another variable such as points-per-game. Correlation can be used to measure the degree of the relationship between two variables such as measuring such as salary and PER. Prediction salary is not always accurate as in the case of Phoenix Sun Superstar Devin Booker for example. His prediction salary and actual salary had a $13-million-dollar variance. We can take these findings and compare them to the actual salaries of the players for the following season and see the accuracy of the prediction salary. General managers can then start to use prediction salary to draft players that will benefit the team and keep them under the salary cap.

**Introduction**

The NBA is one of the most successful sports mediums in the world today. It features players from all regions of the world, ethnic and social backgrounds, skills, and talent levels. They are becoming global icons; competing with polarizing athletes from Soccer and the NFL. Accordingly, there is significant interest from NBA stakeholders to maximize productive and profitability. Players train year-round to perfect their crafts, knowing that large monetary payoffs are available for those whose statistical metrics are superior to their peers. Contrastingly, team owners and general managers want to produce a competitive product that also earns substantial monetary returns.

In 2017-18 season, the salary cap and the luxury tax of NBA was $99 million. The luxury tax, which is the amount over the salary cap that a team can go before becoming penalized, was $119 million. This puts team owners and management of NBA teams in a position to balance a teams’ success and the amount of money they pay for the product. Data analyst are relied on to assist in this decision making. Models are built and relied upon to demonstrate which player statistics correlate to higher salaries. Particularly, the NBA\_season1718\_salary, player\_data, Players, and Season\_Stats datasets from Kaggle are used in this analysis.

This report attempts to answer the question of whether there is a strong correlation between a players’ salary and other performance metrics, such as Points Per Game (PPG), Rebounds Per Game (RPG), and Turnovers Per Game (TOPG). Moreover, if this correlation is enough to predict which players will be wise contract options for teams.

**Description of Data (**Data Preprocessing, Descriptive Statistics and Outliers**)**

For our project, we will be using the NBA\_season1718\_salary, player\_data, Players, and Season\_Stats datasets from https://www.kaggle.com/koki25ando/nba-salary-prediction-using-multiple-regression/data. These datasets contain over 400 records of NBA player salaries and production metrics from the 2017 – 2018 season. To make this more manageable, we applied various methods of data preprocessing, such as merging the NBA\_season1718\_salary and Season\_stats into one dataset named stats\_salary. These datasets were merged on the keys of Player in the salary and stats datasets. Stats were filtered to include only values from 2017. Metrics, including Minutes, Points, Assists, Rebounds, Turnovers, Blocks, and Steals were divided by total games played for each player to give per-game averages.

**Some variables of interest:**

**Salary** – Monetary figure made per individual in dataset.

Summary statistics of Salary are:

|  |
| --- |
| > summary(stats\_salary$salary)      Min.  1st Qu.   Median     Mean 3rd Qu.     Max.     17224  1518316  3519283  7106258 10812701 34682550 |
|  |
|  |

**PPG** – Points per game of individuals in dataset.

Summary statistics of Age are:

|  |
| --- |
| > summary(stats\_salary$PPG)     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   0.4444  4.9280  7.5227  9.1396 12.2229 31.5802 |
|  |
|  |

**Age** – Age, in years, of individuals in dataset.

Summary statistics of Age are:

|  |
| --- |
| > summary(stats\_salary$Age)     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    19.00   23.00   26.00   26.19   29.00   40.00 |
|  |
|  |

**G** – Total games played in regular season. Maximum of 82

Summary statistics of Games are:

|  |
| --- |
| > summary(stats\_salary$G)     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     2.00   42.25   66.00   57.91   76.00   82.00 |
|  |
|  |

**MPG**– Minutes played per game, per individual, in regular season.

Summary statistics of Minutes Played are:

|  |
| --- |
| > summary(stats\_salary$MPG)     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     2.60   15.18   21.25   21.27   28.01   37.76 |
|  |
|  |

**PER** – PER strives to measure a player's per-minute performance, while adjusting for pace. A league-average PER is always 15.00, which permits comparisons of player performance across seasons. Primarily an offensive measure of skills.

|  |  |
| --- | --- |
| All-time great season | 35.0+ |
| Runaway MVP candidate | 30.0-35.0 |
| Strong MVP candidate | 27.5-30.0 |
| Weak MVP candidate | 25.0-27.5 |
| Definite All-Star | 22.5-25.0 |
| Borderline All-Star | 20.0-22.5 |
| Second offensive option | 18.0-20.0 |
| Third offensive option | 16.5-18.0 |
| Slightly above-average player | 15.0-16.5 |
| Rotation player | 13.0-15.0 |
| Non-rotation player | 11.0-13.0 |
| Fringe roster player | 9.0-11.0 |
| Player who won't stick in the league | 0-9.0 |

Summary statistics of PER:

|  |
| --- |
| > summary(stats\_salary$PER)     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    -2.10   10.20   13.00   13.64   16.30   30.80 |
|  |
|  |

The above metrics provide insight about the quartile ranges, which identifies the mean and outliers. For example, Miles Plumlee has a salary of $12500000, which is above the 3rd quartile, and PPG of 5.04 which below the 1st quartile. Moreover, Plumlee has a PER of 8.4. according to the above table, 8.4 represents a player who will not stick in the league. 8.4 also is below the 1st quartile range, indicating a player who should not have a large salary.

The information below further details descriptive statistics:

Call:

lm(formula = salary ~ PPG + RPG + TOPG + PER, data = stats\_salary)

Residuals:

      Min        1Q    Median        3Q       Max

-18961132  -3075039   -155232   2723798  19018569

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) -2149623     709884  -3.028  0.00261 \*\*

PPG           739551      83687   8.837  < 2e-16 \*\*\*

RPG           796706     130325   6.113 2.17e-09 \*\*\*

TOPG         -244278     556410  -0.439  0.66086

PER           -15876      68313  -0.232  0.81633

---

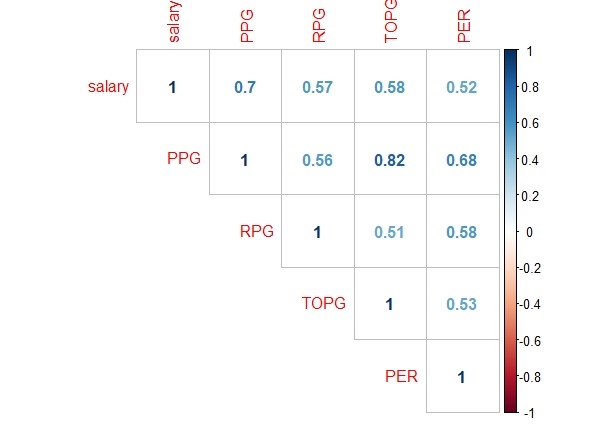
Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5174000 on 437 degrees of freedom

Multiple R-squared:  0.5379,Adjusted R-squared:  0.5336

F-statistic: 127.2 on 4 and 437 DF,  p-value: < 2.2e-16

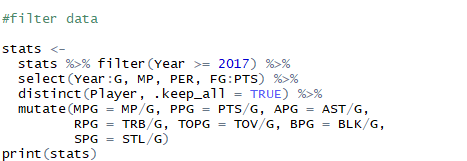
PPG, RPG, TOPG, and PER explain .5379 of the variability in the model when predicting Salary. According to the linear model, when predicting salary, PPG and RPG are significant at the 99% level. The following figure shows the correlation of the five selected variables. Moreover, it demonstrates how correlation and significance can vary.



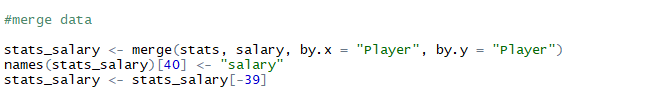
Although there is a stronger correlation for salary with TOPG than RPG, the significance of the beta coefficient (RPG) more accurately predicts salary.

**Steps of Analysis**

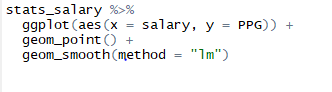
We loaded our two datasets into R Studio then filtered and mutated some of the data. We filtered the data to make it the year 2017 and then we mutated the variables that we thought that would correlate to Salary the most.

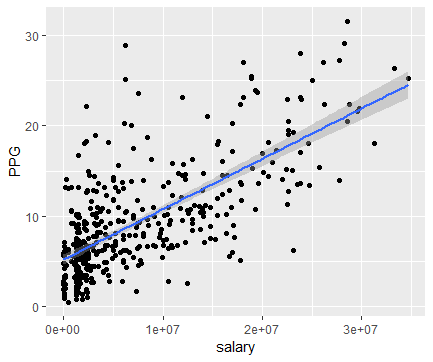


The two sets were then merged together to create one dataset that we will pull the rest of our data for the project.

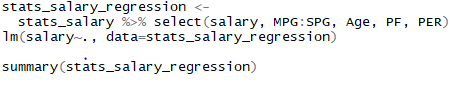


 The data that was pulled was used to create plots to show the relationships between certain variables i.e., salary and points-per-game.

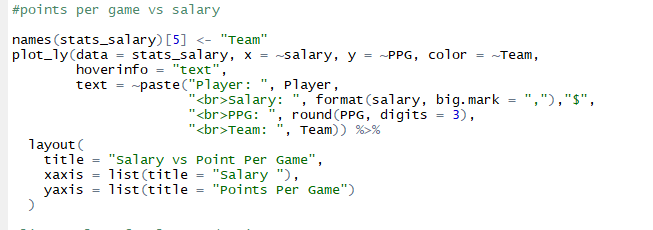


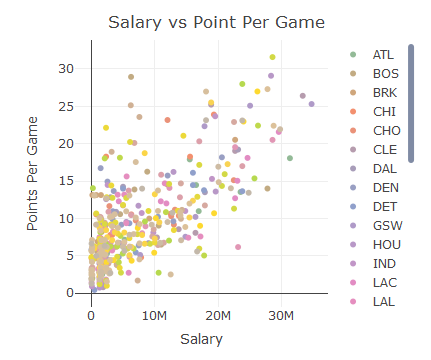


We then created coefficients to show how a player’s statistics can affect their salaries. The coefficients represent the amount of money a player loses or gains during a game or a season and can help shed a little light on explaining why a player’s salary can be low or high. For example, a NBA player makes $988,339 rebound per game in a NBA season.

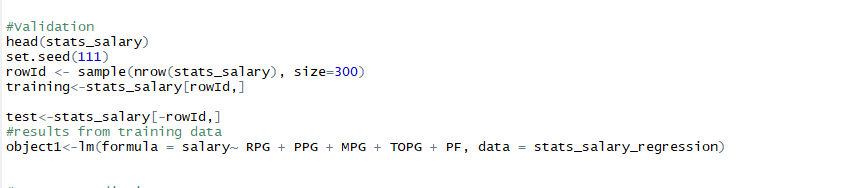


Lastly, we created a salary prediction method to predict the salary of a couple test players using rebounds, points-per-game, minutes played, turnovers, and personal fouls as the variables to be included in the prediction. We were getting a high Salary prediction so we had to offset the data with a negative correlation to Salary which would be personal fouls.  Theses variables can also provide some insight as to the amount of salary a player can earn. The main ones are points per game and salary from our coefficients. So, let's make one more linear model with salary and ppg to better understand the market for NBA players.

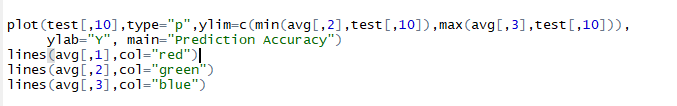


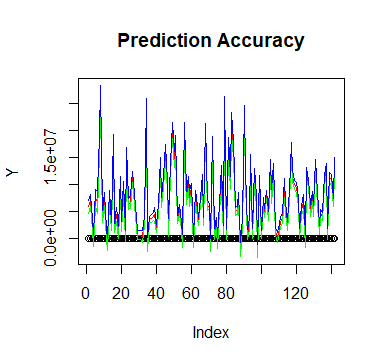


**Validation Analysis**

For validation we want to test if our regression model is accurate and to see if our regression has error in it. So, we made a training set of 300 rows and a test set of our data to see what we could

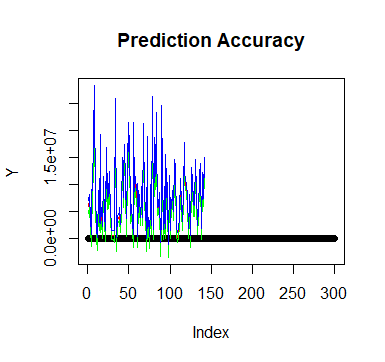
find out.





In the above we wanted to see the accuracy of our regression model and to make sure we were getting predictions accurately. Now we just used the test data, so it looks like our data is not that much skewed. Now if we use the training data set, it will be skewed quite a bit more. We do this

by using the same code as we did above but just by replacing the test data set with the training data set.

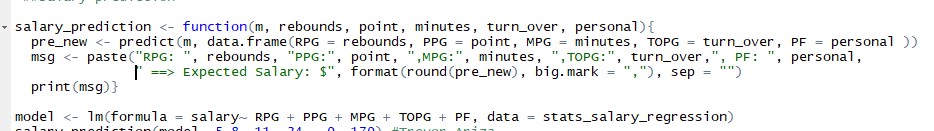


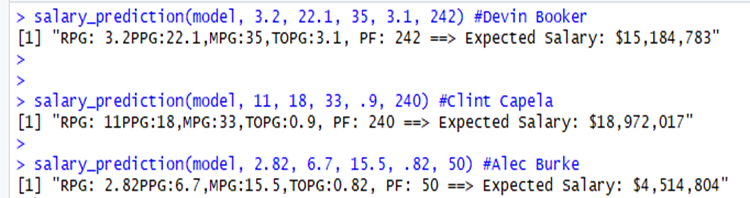
As you could tell by the graph, the data is skewed a little bit more than the test data set due to there are more observations in the training data set. So, we can assume if we did all our observations from stats salary data set it would look similar to the training dataset. Even though our data is skewed decently, the graph still has a good trend and our predictions are accurate. You can also see this when we use error in R.

**Model and Choice of Evaluation**

The model we chose to predict the salary of various NBA players was a linear regression model that took account of points, minutes, turnovers, and personal fouls per game.

With our regression model, we were able to find and display the expected salary of the NBA players: Devin Booker, Clint Capela, and Alec Burke.





We then compared the expected salary of the three to their actual salary and saw a slight variance. Once we compared the linear prediction model with the training dataset, it was clear that with a few outliers, our prediction analysis is a valid option for concluding a professional player’s salary in the NBA.

     The one that fits are salary prediction to a tee is Clint Capela whose NBA stats deserved the money. His new contract is 18 million a year which is close to our salary prediction. The other guys, are young basketball players who you could say are outliers due to their age. For example, Devin Booker is 20 years old and is scoring the ball at a high rate on the Phoenix Suns which are a bad team. So, the Suns desperately need Booker. The result of that is that the Suns overpaid him with $20 million a year for a new contract.

**Conclusion**



NBA owners and executives are always looking for the next superstar, but some of those once-in-a-lifetime players don’t fit inside their budget within the salary cap. The problem these general managers face is creating a team, whether that be through a draft pick, contract extension, or free agency acquisitions, that yields the best performance and the greatest chance to win the championship by spending the least amount possible, or at least staying within their cap. We have taken the NBA statistics from the 2017 -18\_NBA\_salary.csv data set for the 2017-2018 NBA season. We chose the Phoenix Sun superstar Devin Booker’s stats as the sample data to compare our prediction model to his salary in real life. Using our model, looking at points per game, minutes per game, turnovers per game, rebounds per game, and personal fouls per game we found that Booker’s annual pay should be $15,184,783 compared to his actual 2017-2018 salary of $2,319,360 which is roughly a $13 million variance. This shows how statistics and real-life application may be different which could be described as an outlier. Outliers in the NBA include statistics of players who are considered draft busts, those who experience sudden injury, traumatic experience, and, quite frankly, time, which can be equated to old age, or amount of years in the league.  We discussed the correlation between each category and the strength of each relationship, revealing salary has the greatest odds to predict an NBA player’s salary, and that points per game (PPG) and minutes per game (MPG) have the strongest relationship of all the given stats. We used data preprocessing by sorting and filtering the data sets we combined to merge them so we can access the data in a manner that allows easy access. We then created a linear model that shows the coefficients and how they affect a player’s salary in a negative or positive way. Our analysis breaks down how much money a player would earn per point per game, or on the other end, how much each turnover per game would cost a player- relative to their projected salary. The coefficient for PPG is $840,485 meaning depending on the players salary, it would increase by m increment of that every time the individual scored a basket. Our prediction model gives a statistical analysis of what the player should make based solely from

stats, intangibles not included. Before we implemented the training model, our salary projection

model showed a solid prediction that passed the eye test; After comparing our data to the training data set, our prediction shows slightly less accuracy and a slight skew. Overall, the prediction model we used has a strong enough correlation to the training data to be able to predict an NBA player’s salary.