**Capstone II: Project Report**

**Problem Statement**

In this data set we wanted to find out how a player’s salary related to a player’s statistics. We also wanted to find out which statistics played more of a role in their salary than others. Also, by using this data set, we are able to track a players stats, as well as their pay as they progress in the NBA. We are also able to see how a team can use this information to negotiate a player’s salary and stay within the salary cap.

**Data Wrangling and Cleaning**

This data set was loaded in by using pgadmin to set up an ODBC connection and first loading the csv file into a dataframe in pgadmin with PostgreSQL. The data set contained 467 rows and 32 columns. Fortunately, this data set did not require too much data wrangling. Two columns needed to be removed, namely the index column and the player-additional column. The only other issue we had were a couple of null values that needed to be dropped.

**Exploratory Data Analysis (EDA)**

There were many steps to our EDA. First, we needed to plot our statistics by means of a boxplot to see if we had any outliers in our data set. You can see that our data did have some outliers, but this was due to the high performance of some players in the NBA (Figure 1). We then created a box plot of the players salary range. Just like in our last box plot you can see that there were some outliers, and this is due to the performance of better players in the league (Figure 2).

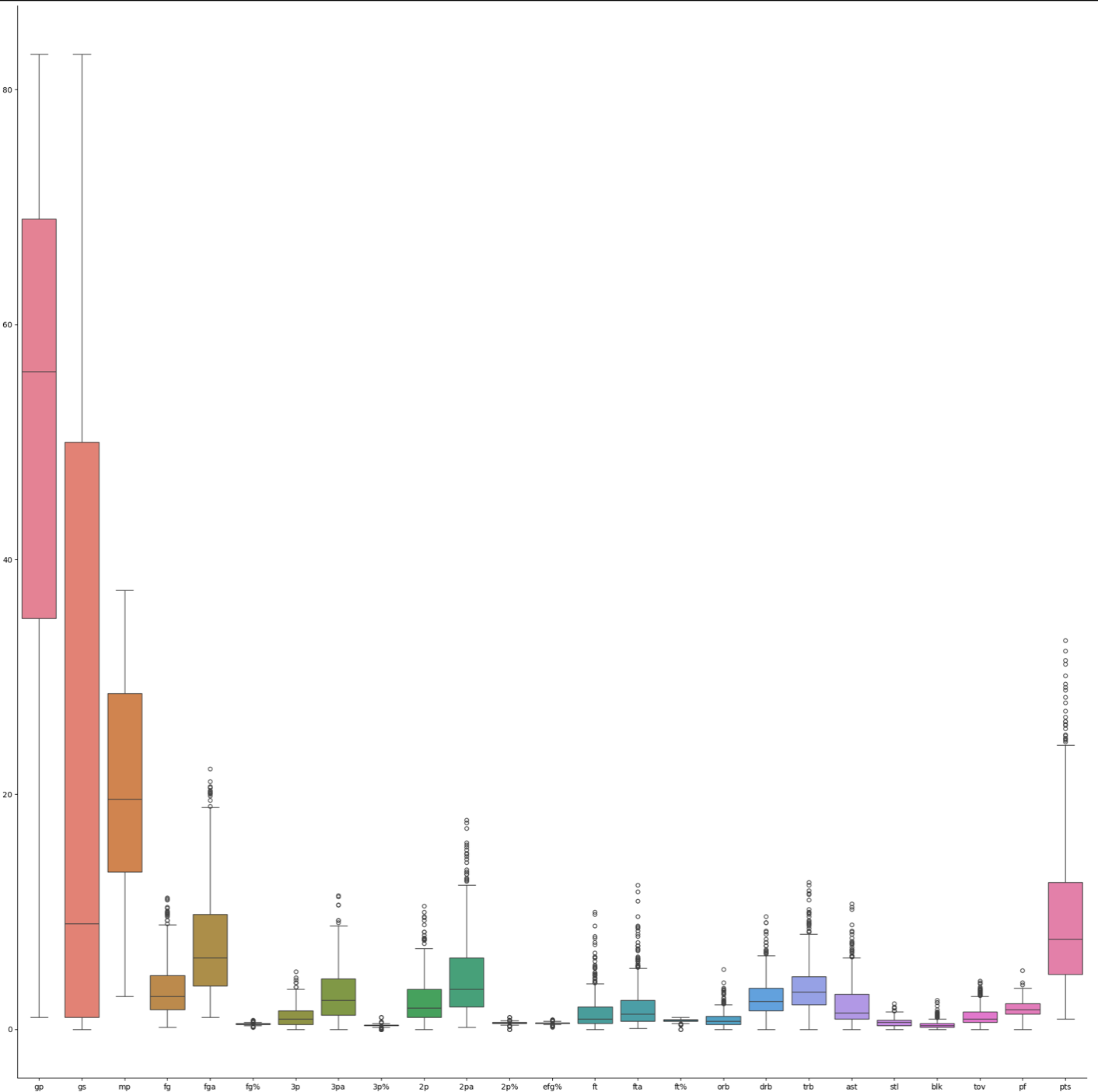


Figure : Box Plot of Player Statistics

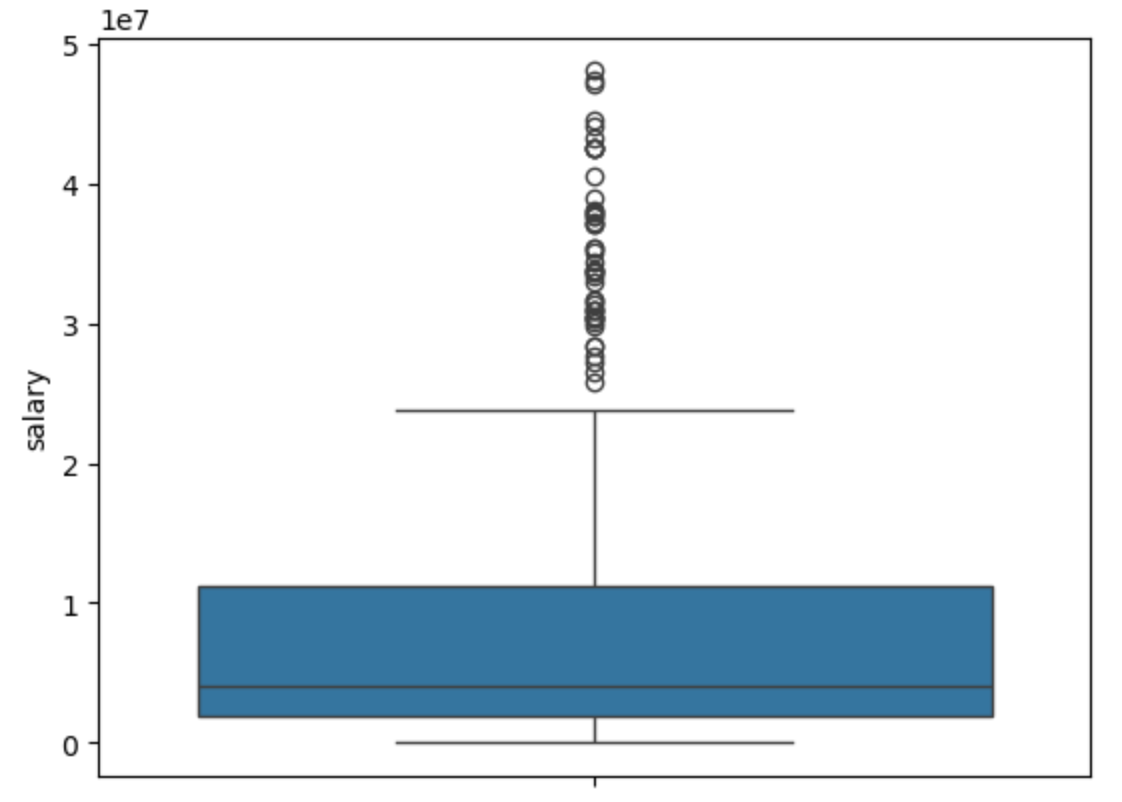


Figure : Box Plot of Player Salary Range

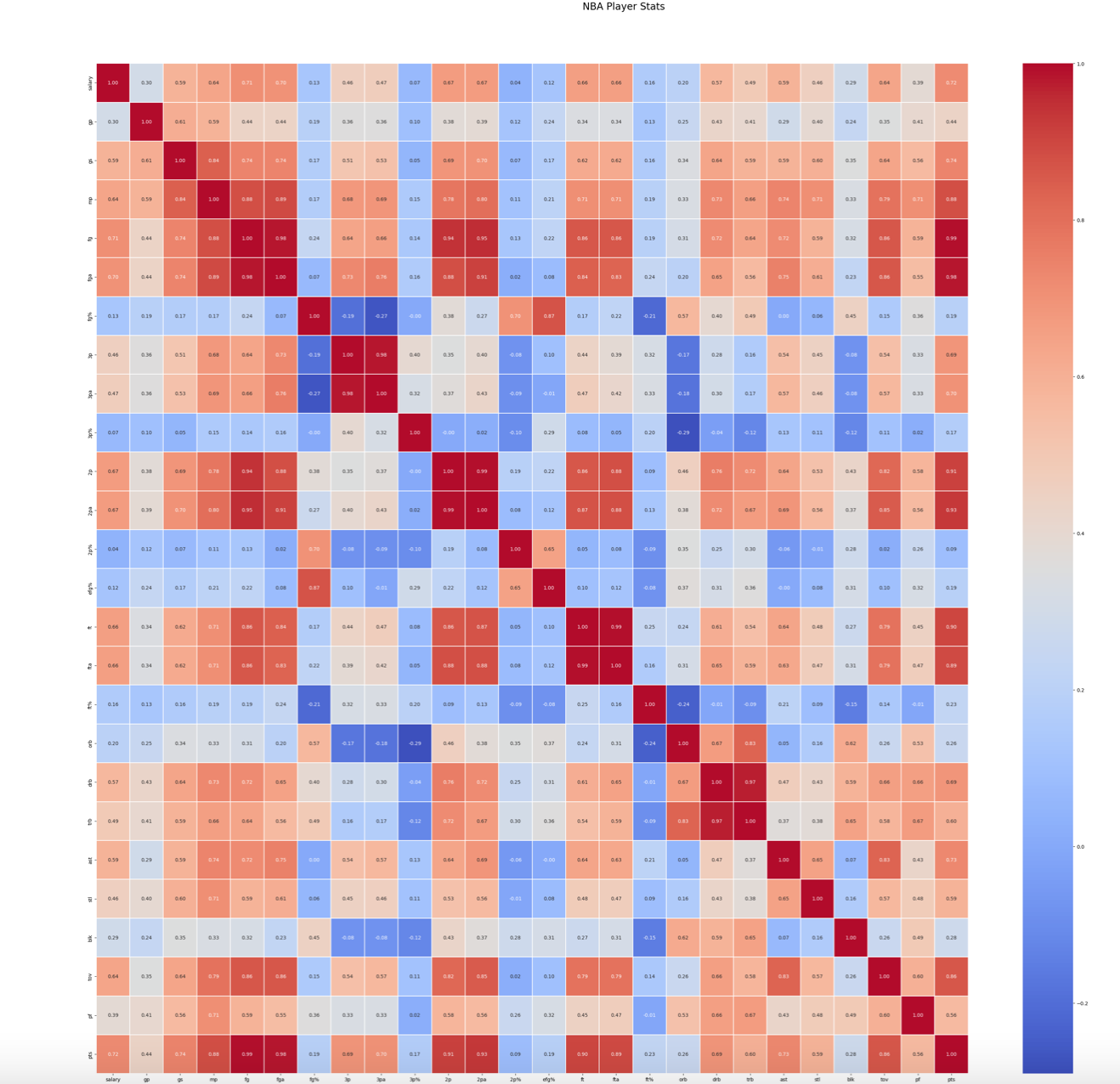
Next, we needed to see how our data correlated. In order to do this we created a heat map for our data set. This gave us the ability to find out which statistics correlated the most with our salary column. Below you can see the heat map and the associated correlation between the data set (Figure 3). 

Figure : Heat Map

From the heat map we were able to find out that the three highest correlated statistics with salary were FG, FGA, and PTS. We then plotted the relationship between these stats with the salary column to see if the data was positively correlated. First, we plotted the FG statistic which had a correlation of 0.71 (Figure 4). Next, we plotted the FGA statistic which had a correlation of 0.70 (Figure 5). Finally, we plotted the PTS statistic which had a correlation of 0.72 (Figure 6).

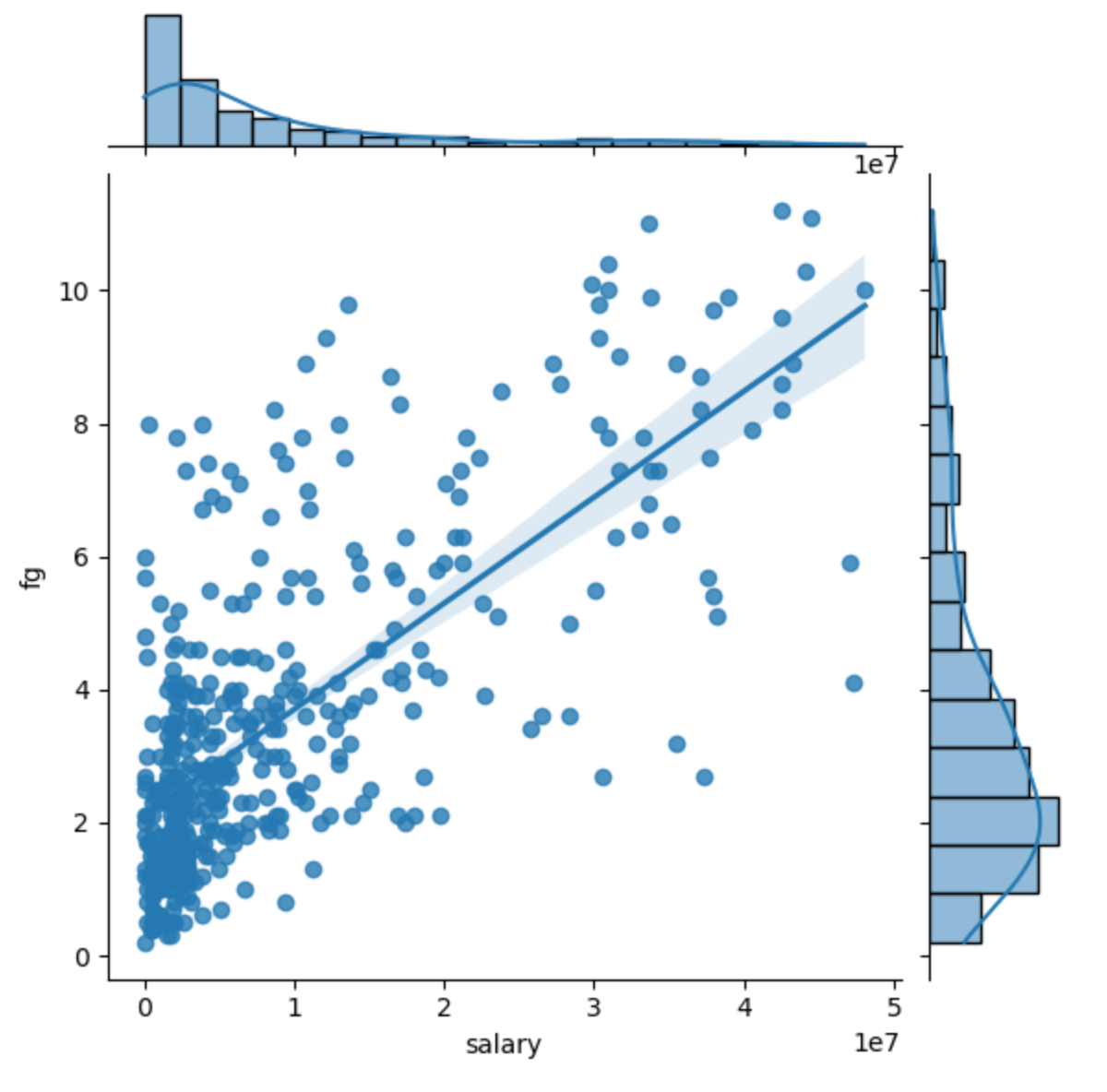


Figure : FG Statistic

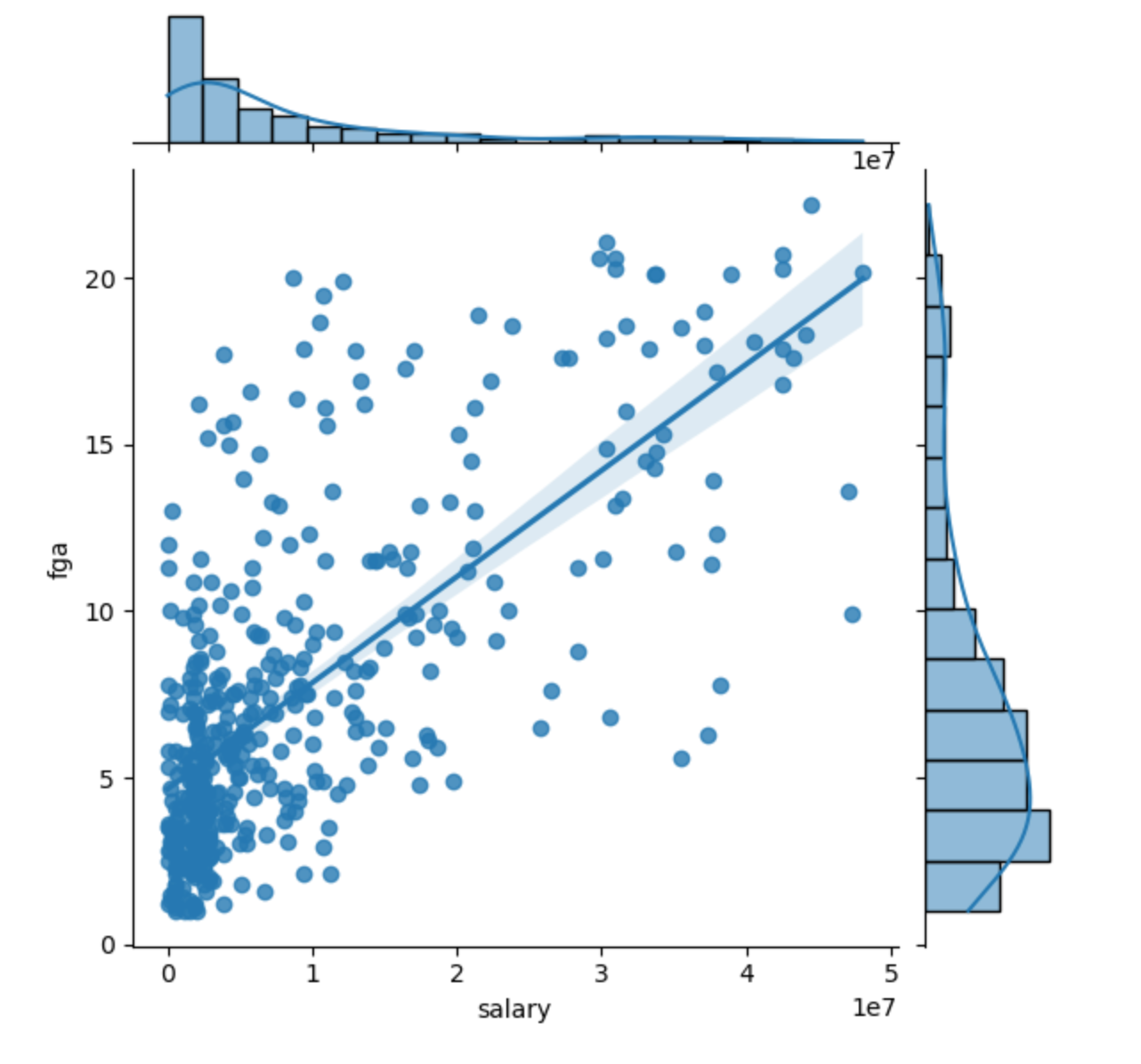


Figure : FGA Statistic

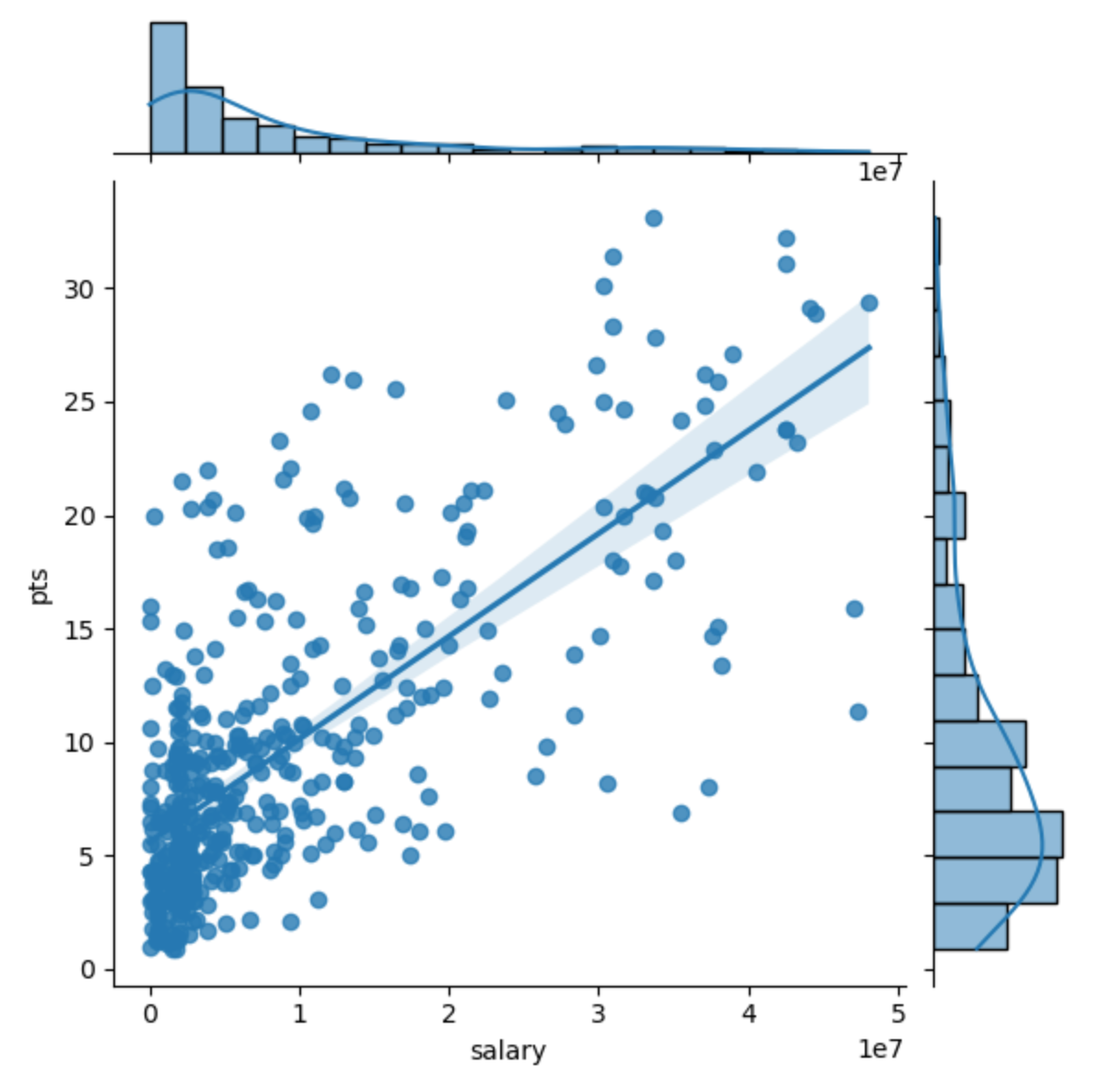


Figure : PTS Statistic

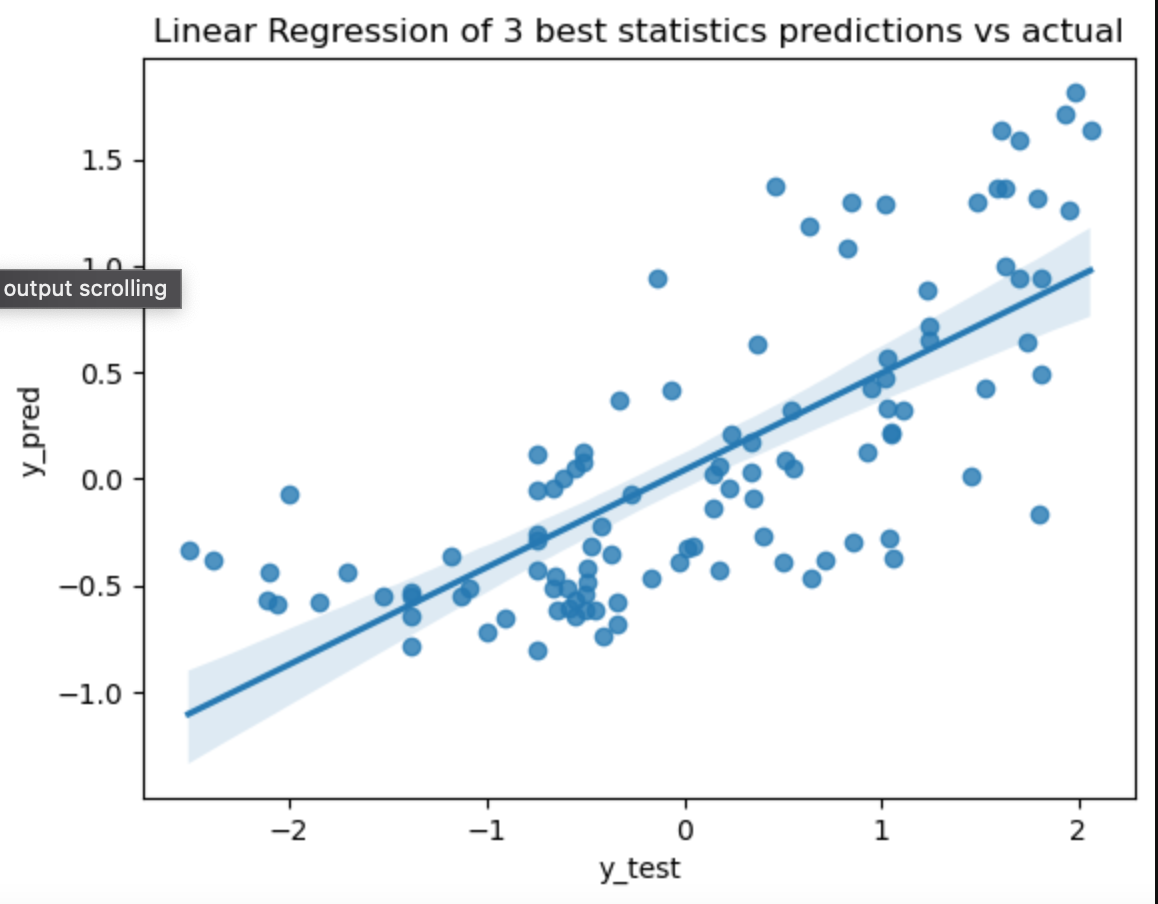
Modeling

Three separate models were chosen for this analysis which included linear regression, ordinary least squares, and random forest regression. The first model that we chose to start with was a simple linear regression. We first had to standardize the dependent variable before performing the linear regression. Once the dependent variable was standardized, the linear regression was performed on our three highest correlated features.

The three highest correlated features performance had a root mean squared error (RMSE) of 0.76 and the prediction versus test data did show positive correlation, but this would not be our best model (Figure 7). We then chose to go with a linear regression with our independent variable being all columns excluding the salary column and our dependent variable being the standardized salary column, and we then performed a linear regression. With this model we were able to obtain a RMSE of 0.49 and our graph of the predicted vs actual was very positively correlated (Figure 8).

For our second model, we chose to go with an ordinary least squares (OLS). With this model, we did as before in our linear regression model, by keeping the independent variable to be all columns excluding the salary column, and our dependent variable to the standardized salary column. With this model we were able to obtain a RMSE of 0.51. Our graph of the predicted versus actual for the OLS model did have some positive correlation (Figure 9).

Our final model we chose to go with a random forest and to perform hyper parameter tuning on the model. We first performed a random forest regression to see how the model would perform and we obtained an RMSE of 5886562.26. This was significantly worse than our other models, but we still chose to perform hyper parameter tuning to see if it would increase the models performance. Once hyper parameter tuning was performed on our model we calculated an RMSE of 5845684.85. This was only slightly better than our first random forest regression model. The graph of the predicted versus actual did show some signs of positive correlation (Figure 10).

  
Figure Three Best Statistics

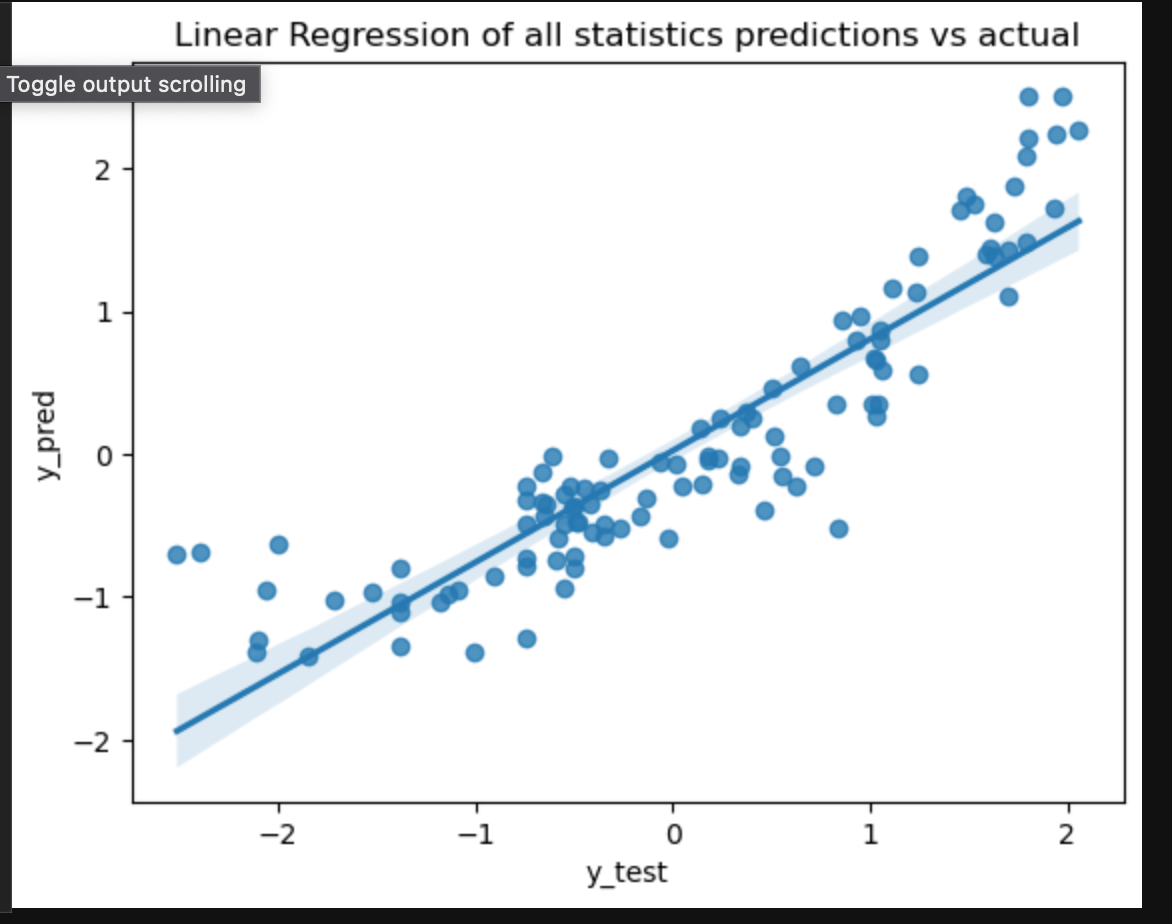


Figure : All statistics

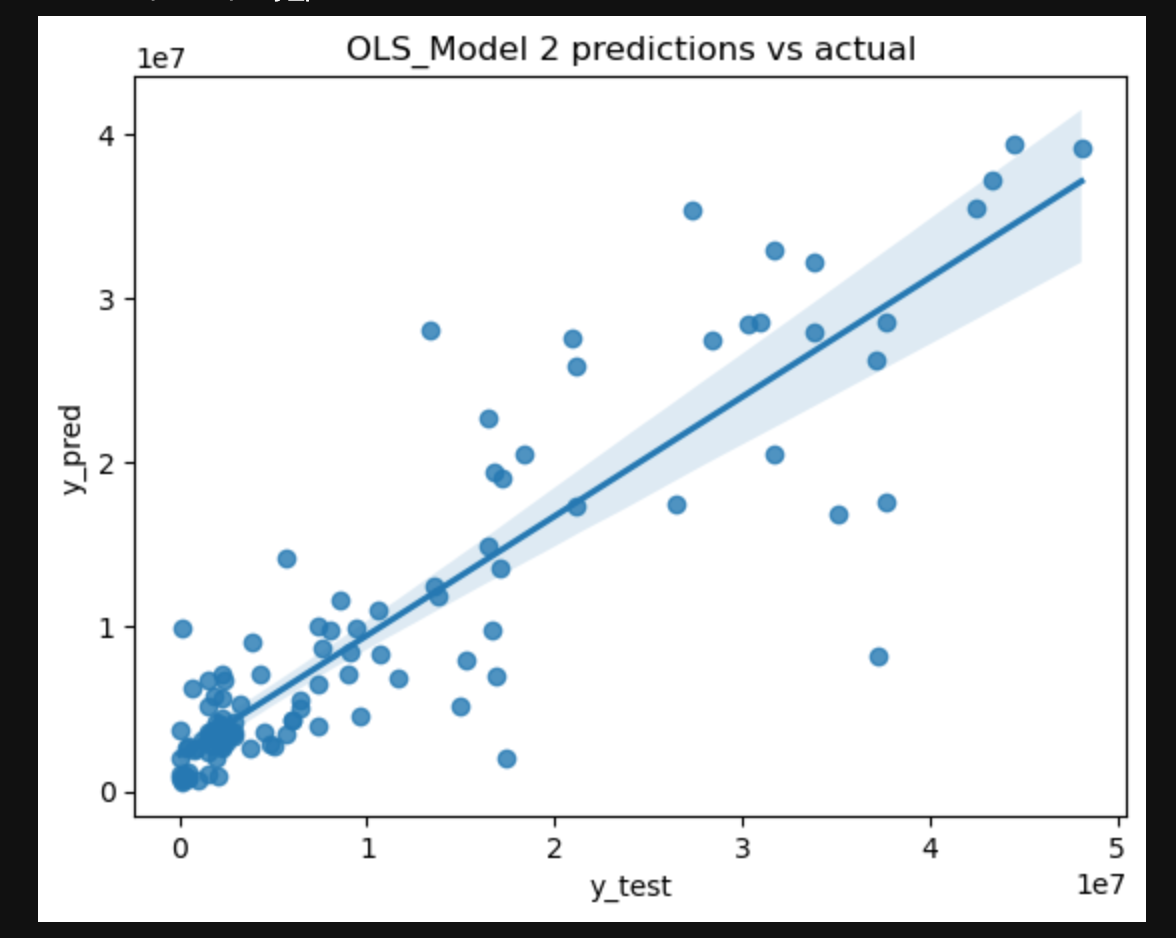


Figure : Ordinary Least Squares

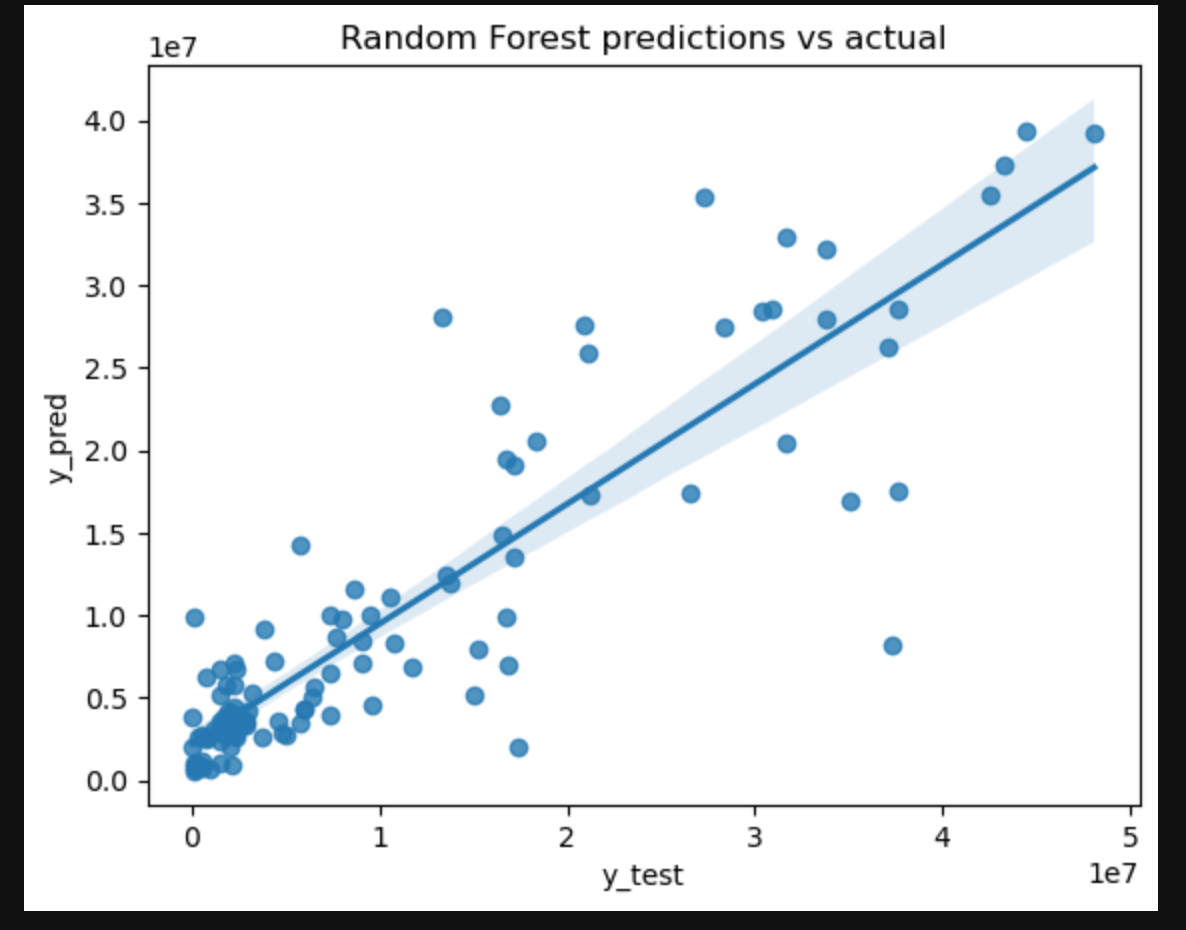


Figure : Random Forest

Final Thoughts

From the above modeling we can see that linear regression performed the best with having an RMSE of 0.49 for our data set. This is great due to the fact of the simplicity of a linear regression and how easy it is to model and the fact that we get such low errors with running a simple linear regression. Even if more data was added to our data set, the simplicity of running a linear regression means that we don’t have to run complicated models and perform hyper parameter tuning to our model every time just to get good results out of our model.

Future Use

There are three takeaways from this data set for future use. First, this analysis will allow a team to track a players progression over time as the player continues to play in the NBA. Next, this analysis will allow teams to calculate a player’s salary based on the data from this analysis. Finally, this analysis will allow teams to be able to stay within the salary cap for each player by using this analysis to calculate a player and team salary.