NBA Analysis

Benjamin Fishman, Hamza Farooq, Juan Serrano Nino

ABSTRACT

NBA salaries range from $50,000 (practice players) to over $40,000,000 for star players. This report will look at statistics from the Basketball Reference database for the 2020-2021 NBA season. The data analyzed includes NBA salary, conference, position, player's age, games played this season, minutes played per game, field goals made per game, 3-pointers made per game, rebounds per game, assists per game, steals per game, turnovers per game, and points per game. Our goal for this analysis is to predict the salary and determine which factors are significant. After conducting a regression analysis, we found that the primary variables for predicting NBA Salary were **Age, Games, Mins, FG, Rebound, Assist, Steal, and Point**. In our analysis, we found no statistical significance between different positions and salary.

INTRODUCTION

The NBA is an extremely profitable enterprise which is why athletes are paid, on average, millions of dollars per season. The more a player contributes to his team's success, the higher his salary is expected to be. By looking at an extensive list of statistics, a general manager can expect to pay their respective athlete an equitable amount for their contributions to the team. Our motivation for choosing this topic stems from the amount of money NBA players make compared to a regular job. There are many recorded statistics that contribute to the amount of money a player makes, and we were determined to find out which were the most significant with the hopes of being able to correctly predict the salary a player should be making.

By conducting a data analysis, we will be able to determine which predictor variables are the most significant towards influencing a player's NBA salary. This data set is composed of many variables, but we will be analyzing the following predictors:

Salary – per season in $ (response variable)

Conf – Conference (East or West)

Pos – Position (PG, SG, SF, PF, C)

Age – NBA player's current age (years)

Games – total games played this season

Mins – average number of minutes played per game

FG – average number of field goals made per game

Three – average number of three-pointers made per game

Rebound – average number of rebounds per game

Assist – average number of assists per game

Steal – average number of steals per game

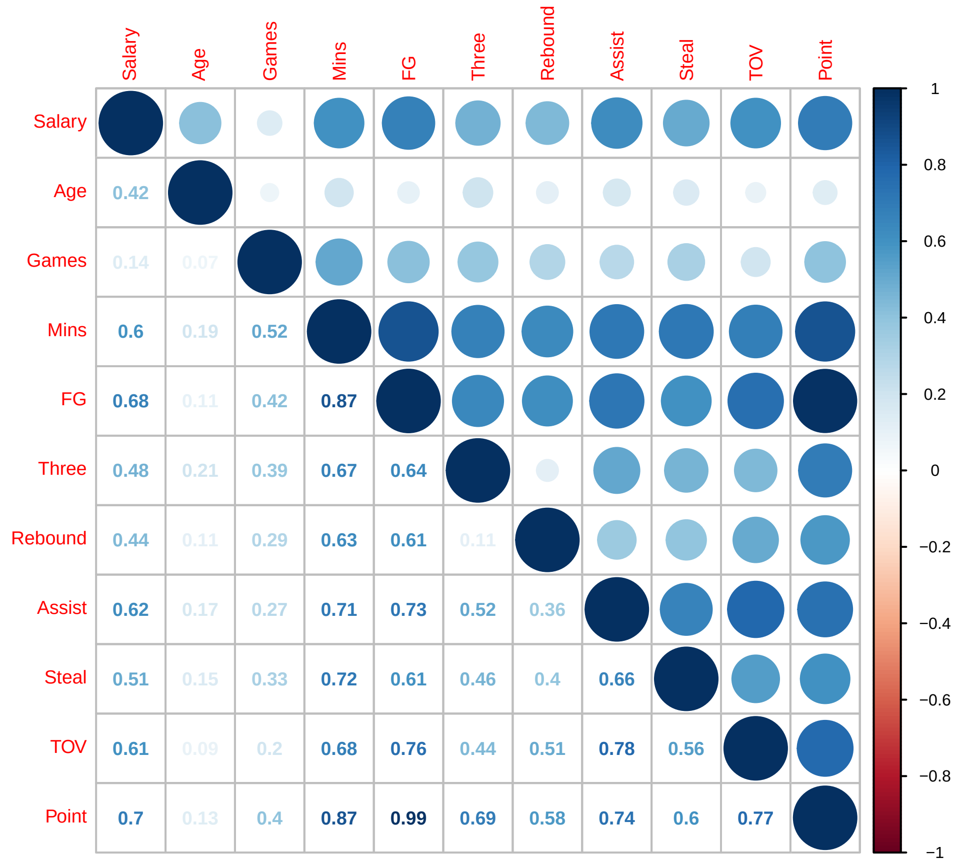
TOV – average number of turnovers per game

Point – average number of points scored per game

```
##
## Call:
```

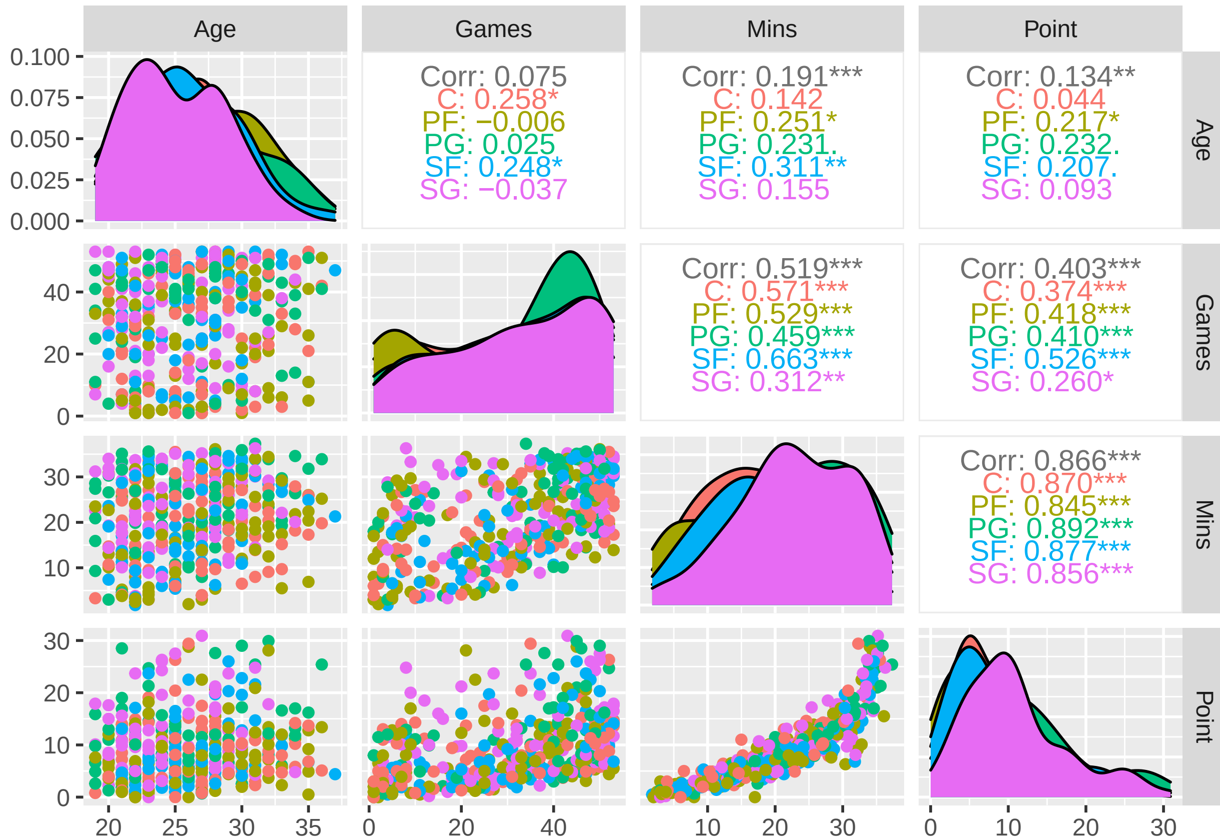
```
## lm(formula = Salary ~ ., data = NBA)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -18165360  -3638839   -524211   3133247  20561102
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16339628    2292411  -7.128 5.36e-12 ***
## ConfW            3571     591849   0.006 0.995189
## PosPF         1063446     995141   1.069 0.285927
## PosPG        -1036934    1374035  -0.755 0.450930
## PosSF          514190    1127064   0.456 0.648498
## PosSG        -1583933    1232791  -1.285 0.199651
## Age            672985      76096   8.844  < 2e-16 ***
## Games          -78367      21904  -3.578 0.000393 ***
## Mins          -176780      93970  -1.881 0.060722 .
## FG           -1373990    1052779  -1.305 0.192665
## Three           87395     631188   0.138 0.889951
## Rebound        296229     246586   1.201 0.230391
## Assist         815357     325651   2.504 0.012716 *
## Steal         3007008    1125895   2.671 0.007901 **
## TOV            363134     649194   0.559 0.576252
## Point         1365832     405196   3.371 0.000828 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5681000 on 371 degrees of freedom
## Multiple R-squared:  0.6541, Adjusted R-squared:  0.6401
## F-statistic: 46.77 on 15 and 371 DF,  p-value: < 2.2e-16
```

We began our analysis of variables by creating a correlation plot. This plot consists of the 12 predictor variables that our dataset provided. Our interest was to see how these variables correlated with one another both numerically and by the color scale on the right side.



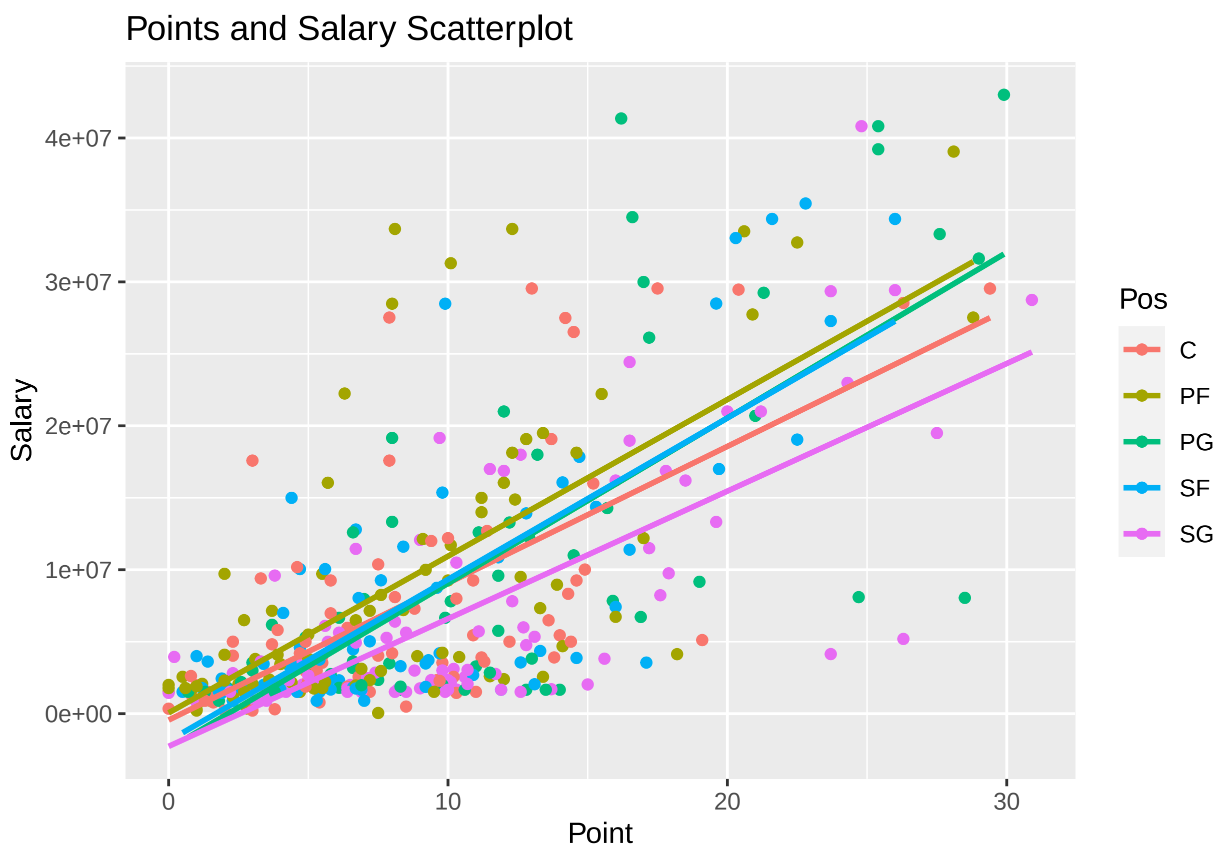
The figure above gives us a visual representation of the correlation between each variable pair. Our response variable, Salary, has a very weak correlation with Games, while every other variable has at least a semi-strong correlation with our response variable (greater than 0.40). High correlation between the independent variables hints at multicollinearity. Additionally, the correlation between Age and Games is very weak, which would suggest that they do not rely on one another. We want to explore these relationships more in future figures, but it supports our theory that more game time (minutes played) relates to higher player statistics.

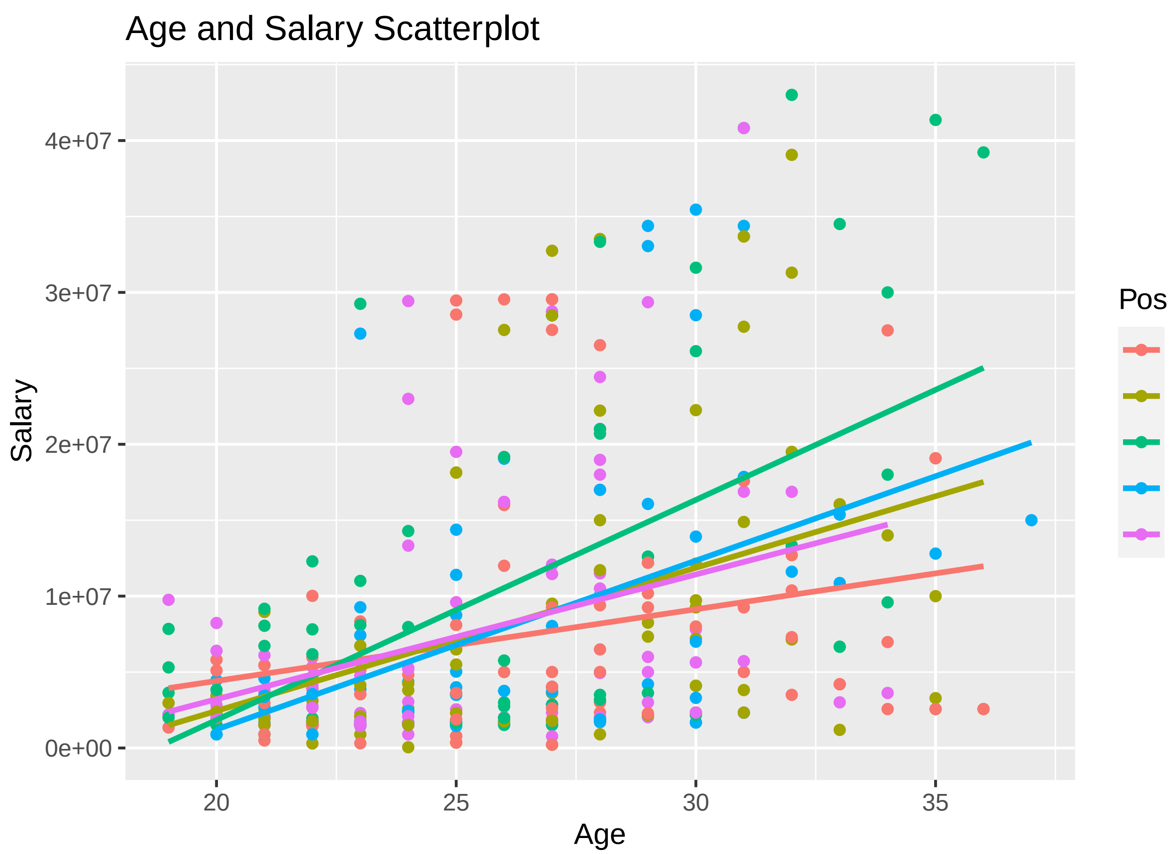
We inserted a correlation matrix to help our initial analysis of the database. Coloring the variables by Position proved to be the most useful.



In the figure above, red, gold, green, blue, and purple represent the center, power forward, point guard, small forward, and shooting guard positions respectively. We expected each position to have different values for age, games played, minutes per game, and average points scored per game. By looking at the minutes against minutes plot, we see that shooting guards have a higher peak in average minutes played per game compared to other positions. In the point-by-point model, shooting guards have a median points per game around 10 while small forward and center are closer to 5. The correlation between minutes and games overall is 0.519. If we look at the independent correlations of positions, shooting guard is the lowest with a 0.312 while center is the highest at 0.571. Overall, the scatterplot of minutes and points shows us that all positions follow a similar pathway with a slight exponential curve.

We decided to create two scatterplot models to further investigate our earlier hypotheses that both Position and Age are related to Salary.


Points and Salary Scatterplot
Salary
4e+07
3e+07
2e+07
1e+07
0e+00
0
10
20
30
Point
Pos
C
PF
PG
SF
SG



From these scatterplots above, we can visualize the relationship between the number of points scored per game and Salary on one figure, and the relation between Age and Salary on the other. For 30-point scorers, the shooting guard position gets paid the least (at about $25,000,000) while the center and point guard are getting paid over $30,000,000. In the Age graph, the older the point guard, the higher they are paid while the oldest centers have a salary more than $10,000,000 less than point guards.

ANALYSIS

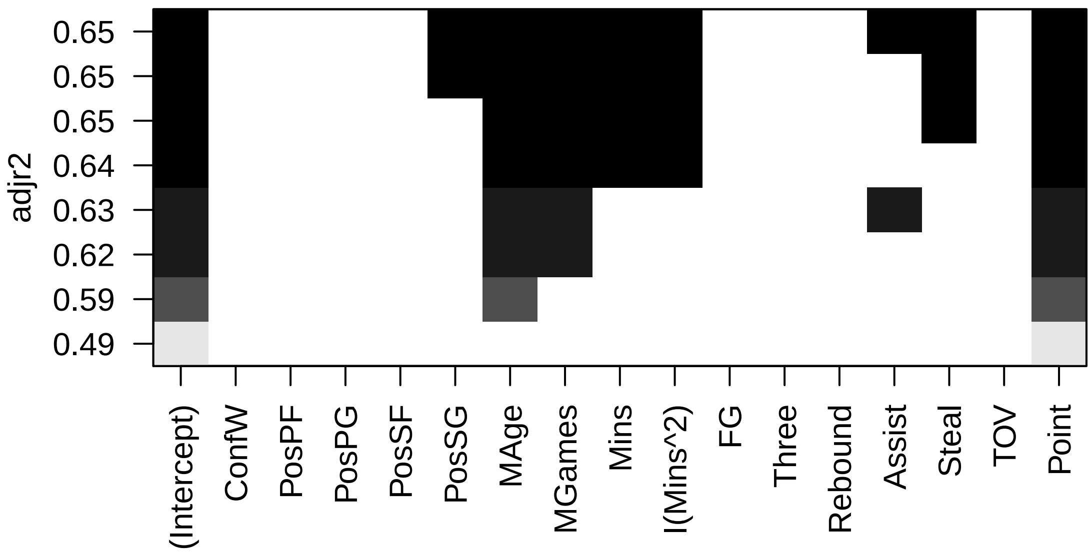
Once we conducted an initial analysis of the dataset, we noticed that the model's intercept was -16,440,628 which was very interesting to us as it was a massive negative number. To make sense of this intercept, we performed mean-centering on both the Age and Games variables while also inserting a squared Mins term. This process then led to an interpretable intercept of 3,073,394, which can be understood as the base salary for an NBA player who is of an average age (26.1 years old) and games played this season (31), while other variables are held at 0.

```
##
## Call:
## lm(formula = Salary ~ Conf + Pos + MAge + MGames + Mins + I(Mins^2) +
##     FG + Three + Rebound + Assist + Steal + TOV + Point, data = NBA)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17749835  -3269261   -557824   3026245  21277211
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3073394    1669565   1.841 0.066446 .
## ConfW         -90398     582390  -0.155 0.876733
```

```
## PosPF         826464     980384   0.843 0.399773
## PosPG        -512974    1358112  -0.378 0.705862
## PosSF         390881    1108504   0.353 0.724574
## PosSG       -1229671    1215676  -1.012 0.312432
## MAge          716942      75735   9.467  < 2e-16 ***
## MGames        -72959      21583  -3.380 0.000801 ***
## Mins         -686041     165023  -4.157 4.01e-05 ***
## I(Mins^2)      15839       4253   3.724 0.000226 ***
## FG          -1207798    1035942  -1.166 0.244408
## Three          24993     620743   0.040 0.967905
## Rebound       349403     242837   1.439 0.151042
## Assist        535944     328819   1.630 0.103973
## Steal        3030605    1106878   2.738 0.006481 **
## TOV           333327     638268   0.522 0.601818
## Point        1154510     402367   2.869 0.004350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5585000 on 370 degrees of freedom
## Multiple R-squared:  0.6666, Adjusted R-squared:  0.6522
## F-statistic: 46.24 on 16 and 370 DF,  p-value: < 2.2e-16
```

At this point, our current regression model has 13 variables (not including 4 dummy variables for position), and our goal is to make a parsimonious model. We ran a backwards elimination regression model which resulted in 9 predictor variables, all of which are significant except for FG. This process yielded an adjusted R-squared of 0.6537. Because this is a heuristics method, it does not guarantee the optimal model, so we ran a best subsets regression on this data, which produced an adjusted R-squared of 0.652, which is almost identical to the backwards elimination model. Nevertheless, we will use the backwards elimination model for our future analysis because of the higher adjusted R-squared.

We created an adjusted R square plot below to verify our best subsets regression significant variable conclusion.



MODELING

```
##
## Call:
## lm(formula = Salary ~ MAge + MGames + Mins + I(Mins^2) + FG +
##     Rebound + Assist + Steal + Point, data = NBA)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17467698  -3349598   -480205   3065657  22107372
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3494131    1450872   2.408 0.016506 *
## MAge          741278      71866  10.315  < 2e-16 ***
## MGames        -76459      20774  -3.681 0.000267 ***
## Mins         -756403     156695  -4.827 2.02e-06 ***
## I(Mins^2)      17586       4139   4.249 2.71e-05 ***
## FG          -1427128     942044  -1.515 0.130629
## Rebound       515223     168296   3.061 0.002361 **
## Assist        526895     258790   2.036 0.042449 *
## Steal        3008377    1084754   2.773 0.005824 **
## Point        1213258     344925   3.517 0.000489 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5572000 on 377 degrees of freedom
```

```
## Multiple R-squared:  0.6618, Adjusted R-squared:  0.6537
## F-statistic: 81.96 on 9 and 377 DF,  p-value: < 2.2e-16
```

The interpretation of the NBA parsimonious model above, found through a backwards elimination regression method (and checked with a best subset regression), is the following:

The salary of an NBA player who is of mean age (26.1 years old) is increased by $741,278, given all other variables are constant.

The salary of an NBA player who has played the mean number of games (31) this season is decreased by $76,459, given all other variables are constant.

For every minute played, on average, the salary of an NBA player decreases by $756,403, given all other variables are constant.

For every minute squared, on average, the salary of an NBA player increases by $17,586, given all other variables are constant.

For every unit increase in FG made per game, salary will decrease by $1,427,128, given all other variables are constant. (this is the only non-significant variable in the model)

For every unit increase in rebounds per game, salary will increase by $515,223, given all other variables are constant.

For every unit increase in assists per game, salary will increase by $526,895, given all other variables are constant.

For every unit increase in steals per game, salary will increase by $3,008,377, given all other variables are constant.

For every unit increase in points per game, salary will increase by $1,213,258, given all other variables are constant.

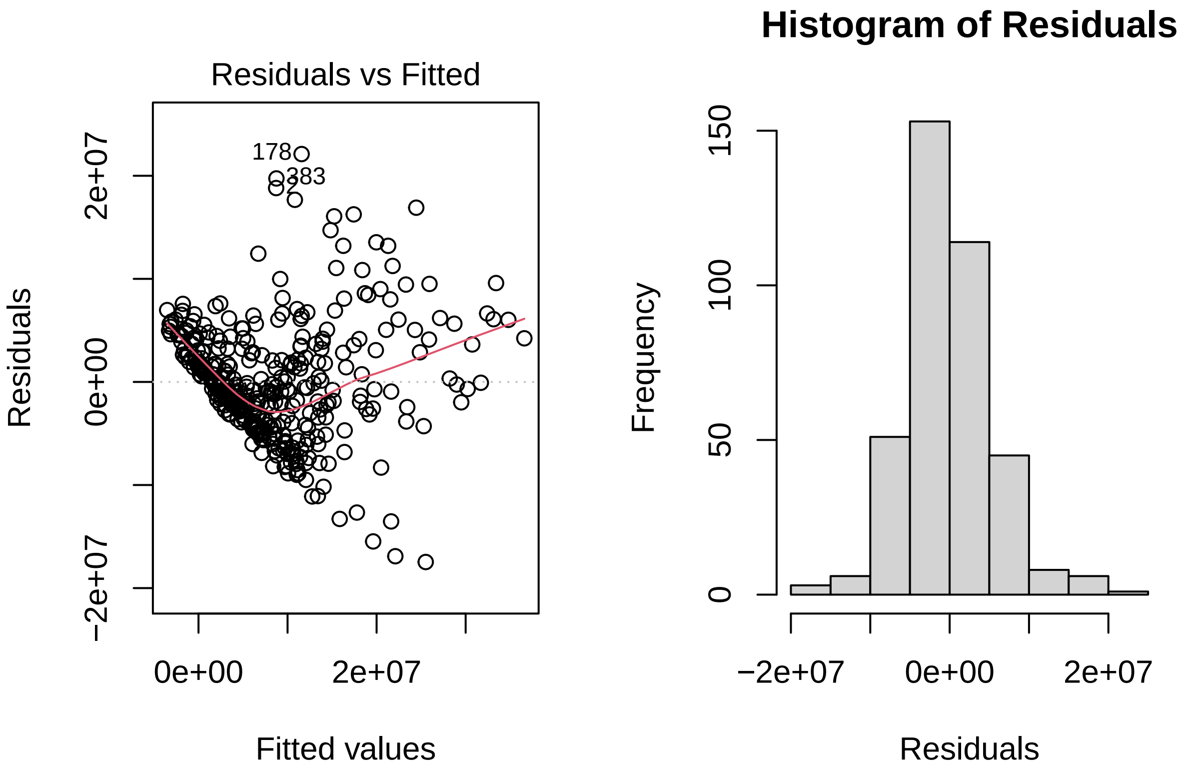
The plots below were designed to verify the assumptions of multiple linear regression:

The relationship between X and Y is linear

```
Each error is independent
```

```
The error is random and normally distributed
```

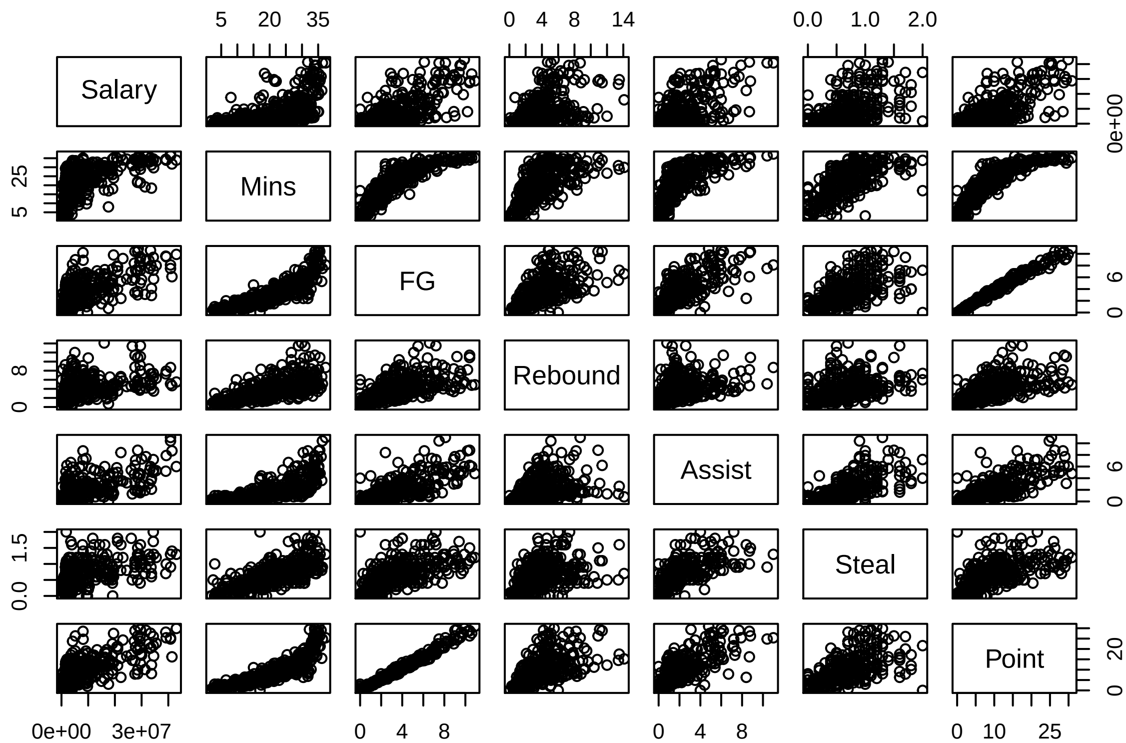
```
The variable of the error is constant
```



The scatterplot on the left indicates there is slight heteroscedasticity as many data points are clustered on the left side. We attempted to log and square root the response variable (Salary) after running a boxCox diagnostic (which suggested a log transformation, but a square root transformation was also a possibility), but it did not improve the dataset, in fact it decreased our adjusted R square value significantly. We applied a log transformation to Mins, as well as Rebounds, but once again, there was no improvement to this scatterplot. There is clearly a linear relation between X and Y, as well as each error being independent. The histogram suggests normality and randomness for the residuals, which suggests our dataset is valid.

Even though our scatterplot does not exhibit perfect homoscedasticity, our histogram plot indicates normality, which is an assumption that allows us to utilize this dataset for our conclusions.

The scatterplot matrix below shows the relations between all variables. We want to make sure that there is a linear correlation between Salary and each predictor variable.



The graphs in the Salary row that are not perfectly linear are Mins and Rebounds, which is why we applied transformations to them, but they did not improve the data set. This scatterplot matrix lead us to create a squared Mins term in our parsimonious model.

QUESTION 1

Done above in the Introduction and Modeling sections of this report.

QUESTION 2

Your teacher's favorite basketball player is Lonzo Ball and his stats (on the New Orleans Pelicans) are found through this link: https://www.basketball-reference.com/players/r/rosede01.html. Is Lonzo Ball's salary reasonable?

We were curious to see if our parsimonious model could accurately predict any NBA player's salary using a 95% prediction interval. We imported Lonzo Ball's 2020-2021 NBA season stats and predicted that his salary should be $29,614,878. This result is significant because it was predicted using our regression model.

We turned this question into a hypothesis test, where the null is the salary should be $29,614,878 while the alternative hypothesis is that it should not be $29,614,878.

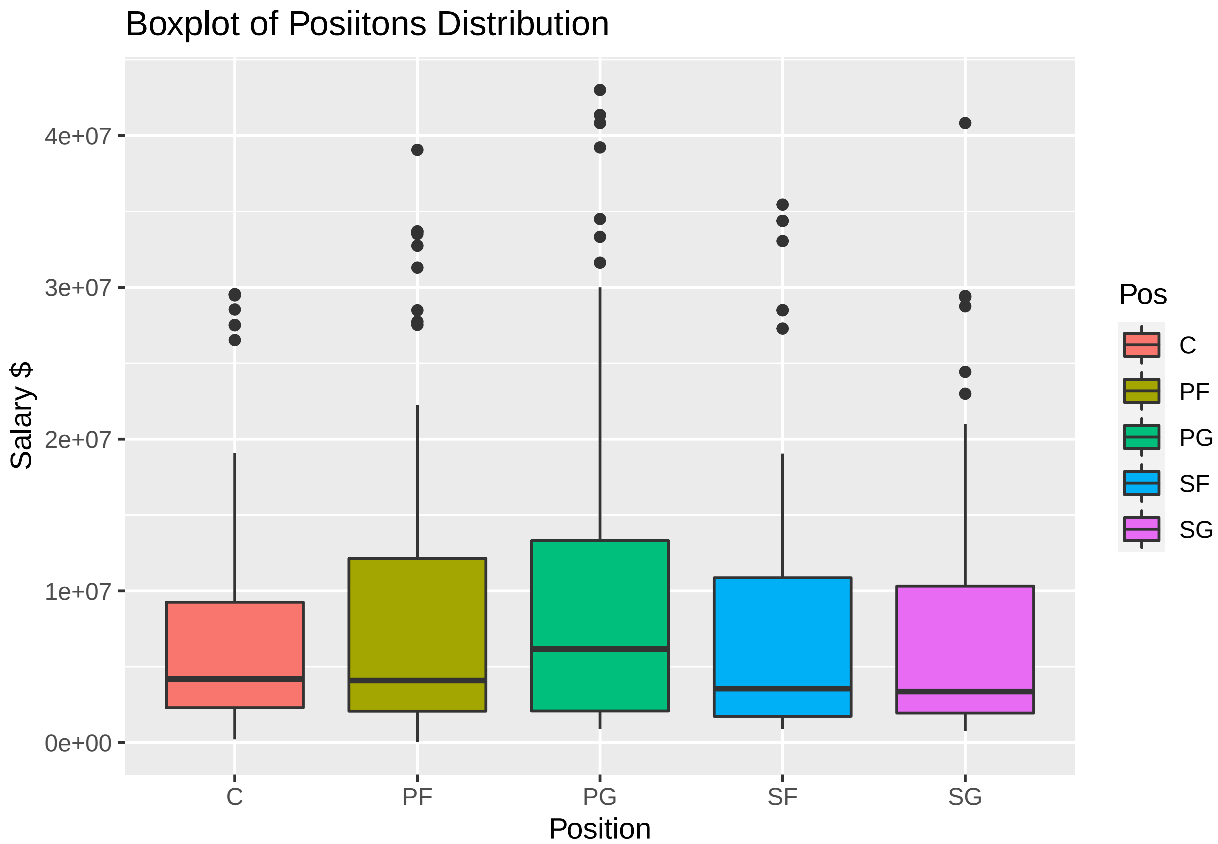
$$H_0 : \beta_1 = \$29,614,878$$

$$H_1 : \beta_1 \neq \$29,614,878$$

Our 95% prediction interval yielded a range from $2,590,330 to $24,753,414. Because our predicted value is outside of this prediction interval range, we reject the null hypothesis based on 95% significance, however, the value we found is not far from being within the range.

Question 3

The NBA Commissioner believes that all 5 positions have the same mean salary. Can you refute his claim?



The above figure is a boxplot distribution split on the 5 positions in the NBA. An initial analysis shows that the PG position has the highest mean salary by a large margin, while also having the highest outliers of all positions. We are going to create a hypothesis test to analyze the commissioner's claim.

Our null hypothesis will be this: salaries across positions are the same and therefore are not statistically significant

Null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Our alternative hypothesis will be this: salaries across positions are NOT the same and therefore is statistically significant.

Alternative Hypothesis

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

To conclude whether or not position is statistically significant, we ran the new regression model below that shows the relation between Salary and Position.

```
##
## Call:
## lm(formula = Salary ~ Pos, data = NBA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -9897372 -6034806 -4099694  2538659 33370675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7478532    1055879   7.083 6.84e-12 ***
## PosPF        1327824    1471115   0.903   0.3673
## PosPG        3317150    1590787   2.085   0.0377 *
## PosSF         382247    1528616   0.250   0.8027
## PosSG         -25207    1466962  -0.017   0.9863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9444000 on 382 degrees of freedom
## Multiple R-squared:  0.01565,    Adjusted R-squared:  0.005341
## F-statistic: 1.518 on 4 and 382 DF,  p-value: 0.1961
```

After running the linear regression model where we predicted Salary based on Position (dummy variables were automatically made within R), we found that C (center) was our baseline reference. This summary concludes that Position is NOT a statistically significant predictor of Salary. The p-value is very high, 0.1961, while our adjusted R square is very low, 0.005341, so compared to the baseline reference of C, the other positions are not statistically different. There is a very large amount of variability within this data set as our Residual Standard Error is 9,444,000.

Conclusion

After completing our analysis for this project, we created two different regression models to help us find the significant variables for our designated questions. These models allowed us to understand the relationship between response variable (Salary) and predictor variables. Our model to predict Salary has an adjusted R square of 0.6537 which is a relatively strong linear relationship between our response and predictor variables.

We wanted to put our parsimonious model to the test by predicting an NBA player's salary based upon their current 2020-2021 stats. We chose one of the most hyped up players of our generation, Lonzo Ball. After conducting a 95% prediction interval, we found that his predicted salary fell outside of this interval which had an upper range of about $25,000,000.

Our regression model to explore the relationship between Salary and Position yielded a large p-value, 0.1961, which is larger than our alpha of 0.05, leading us to believe that there is no statistical significance between Position and Salary – which was a surprise for us because we believed that Position would influence Salary as depicted in the boxplot above. This make sense because players are paid before a season starts (sometimes many years in advance) which allows them to make more money than they deserve based upon their actual statistical performance.