Feature	Value
Name	Moisey Alaev
SID	205208824
Kaggle Rank	21
KaggleR^2	0.55327
Number of Predictors	6
Number of Betas	9 (including intercept)
BIC	14068.13
AIC	14027.72
Complexity	$130 - 9 = 121$

Abstract

The intent of this project is to perform a multiple linear regression to predict NBA players' salaries using player statistics. To conduct this experiment, critical multiple regression techniques were utilized to select the best set of predictors for estimating a player's salary. These include, but are not limited to, stepwise regression, ANOVA interpretations, predictor vs Salary plots, creating new variables, and analyzing collinearity information. Our goal is to improve our Kaggle R^2 without overfitting our model with several predictors. Consequently, creating predictors that packed in a lot of information was critical.

In turn, my model was able to use only 3 numerical and 3 categorical predictors for a total of 6 predictors. Thus, the generated linear model was able to accomplish a $R^2$of 0.6895 and a $R^2_{adj}$of 0.6835 on the training data. The trivial difference in these R^2values indicate to us that the variables in our multiple linear regression are significant and likely won't be overfitting. In my best submission, this model obtained 21st ranking on the Kaggle leaderboards with an $R^2$of 0.55327. However, the model I mistakenly left checked as my final submission had a R^2 of 0.53964. This was an unintentional error and my intended final submission should have been my last submission with a $R^2$of 0.55327.

Introduction

The National Basketball Association (NBA) was founded in 1949 when the Basketball Association of America (BAA) and its competition, The National Basketball League (NBL) decided to merge (A&E Television Networks). The NBA is one of the major four sports leagues in the U.S alongside the MLB, NFL, and NHL, and It consists of 30 officially recognized teams from 29 cities in the U.S and 1 city in Canada (Toronto: Raptors). These teams can be bought, sold, and moved to new cities with potentially better markets or new fan bases. Traditionally, each team can have a maximum of 15 players on their roster and only 13 active players per game. This should translate to 450 maximum players. However, our training data consists of 420 players, and the testing consists of 180 players for a grad total of 600 players in the 2016/17 season. This could be a result of players who are free agents, (have no contract with a team) players who can migrate between teams, other players that have improved in the G-league and get drafted into the NBA, and some teams simply having more than 15 'official' players on their roster. Furthermore, players have 82 games to prove their worth in the regular season before they are either chosen for the playoffs (known as an "above 500 team") or have to wait until next season to contend for a championship title.

At the beginning of 1949 the average basketball player earned between $4,000 and $5,000 with a few exceptions (Bradley). This would equate to a range of $45,242.86 to $51,553.57 today, a respectable wage for a lower middle class earner (Webster). However, the modern NBA player's salary has skyrocketed to an average of 8.32 million dollars and a gross of 3.67 billion dollars for all players (Gough). This dramatic increase has been due to several factors that influenced the growth of the modern NBA over the course of the past 70 years. Notable changes include the growth of viewership and fandom, the rise of all-stars and superstars, and great leadership by NBA head offices.The most recent leadership being Adam Silver, the remarkable commissioner paving the way for a new era in the NBA.

As the NBA evolved to amass a vast viewership and immense funds, so did the game of basketball. Nowadays rules are defined to precision with referees expected to review all their calls postgame and attend conferences to agree on the future of regulated basketball. Players are now competing at the highest level imaginable with the use of cutting edge technological equipment for training and recovery. In addition, trainers and physical therapists are more educated about body dynamics than ever before, allowing them to help a player focus and improve on every aspect of their game. Lastly, data and statistics from each and every player in every game are analyzed and managed by large groups of statisticians who not only catalogue basic player stats (points, asists, rebounds, steals, etc.), but also utilize advanced player stats that have been developed just for basketball (win shares, box plus minus, defensive/offensive box plus minus, value over replacement player, etc.). The results of better statistics being available is that trainers and players know which aspect of their game they need to improve the most, scouts have a better understanding of what they are looking for in incoming players, teams can make better decisions in trades and drafts, and fans can definitively evaluate a player's success on the court. Above all, teams can better evaluate a player's worth when it comes to presenting players with contract offers. This would ideally result in being able to better estimate a players salary from their stats in games.

However, there is a level of uncertainty that cannot be easily modeled simply from player statistics. This may include if a player is the best at making 3- pointers–such as Stephen Curry–getting their team to the finals–like Lebron James–and dropping triple doubles on a nightly basis–like Russel Westbrook. Players like these are infamous for their contributions to their teams; Consequently, these players get an astronomical salary compared to the average NBA player. This is a strategic move on the part of the team franchise to maintain their image with this player as its face. Therefore, while teams lose out on tens of millions a year from keeping just one player, they make hundreds of millions more from the revenue that the player brings in. In this way, economics and politics interfere with data trends, making linear regression a difficult tool to model all the variance in salaries given dramatic outliers in our data.

The goal of this project is to extrapolate the salaries of NBA players given each player's statistics using multiple linear regression. We have 420 players and their respective stats and salaries as the training data set, and we have 180 players as the testing set. Using the training dataset, we devise a linear regression model that attempts to predict the salaries of the testing set from the testing sets' player stats. Practically, this project involved building a model that predicts a set of $\widehat{Y}$'s that are submitted on Kaggle. Kaggle tests these $\widehat{Y}$'s and returns the corresponding R^2 value for your submitted predictions. Your place on the leaderboards then is determined by your R^2value.

Methodology

Splitting our Training Data

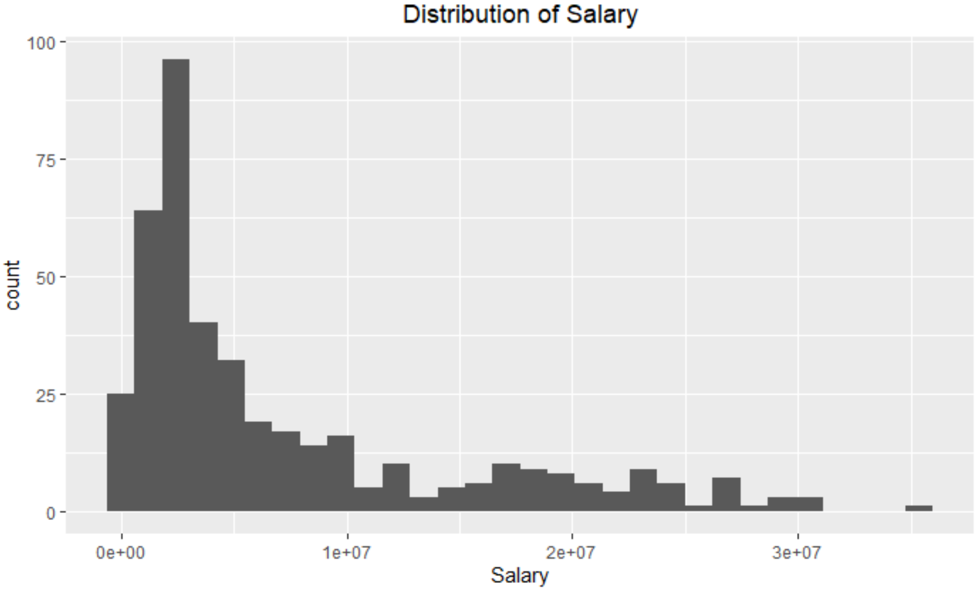
The objective of this project is to use the training data to create a model that can predict the salaries of NBA players in the testing data. Instead of going between my local model and testing the model on kaggle several times with the use of numerous submissions, I decided to take a more efficient approach. I decided to split the given training data into my own local training data set and testing data set. This could be accomplished by arbitrarily selecting ⅓ of the original training data set to be the new testing data set. The resulting dimensions are shown in the table below.

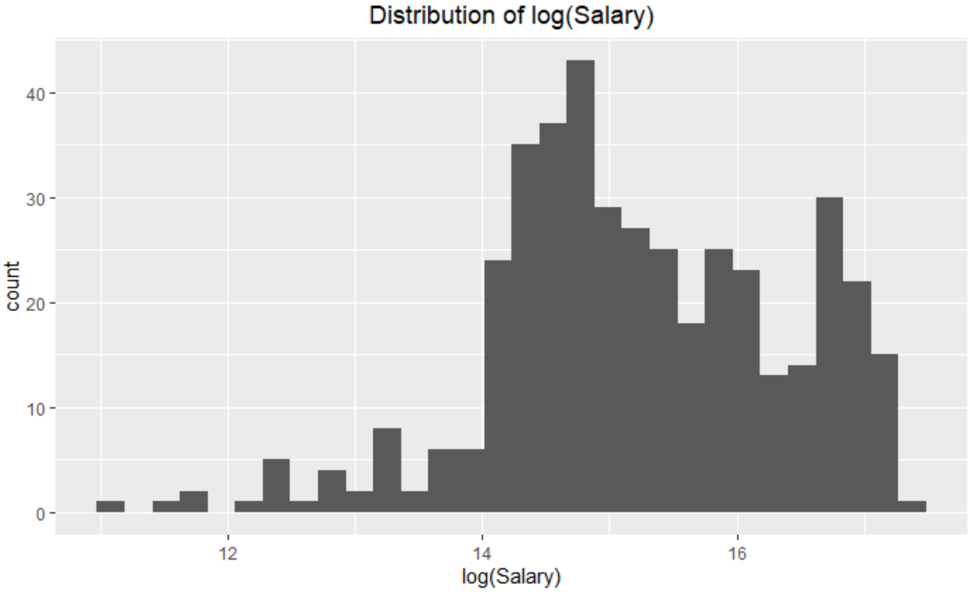
Data Set Dimensions

Data set	Dim(data set)
Original Training Set	[1] 420 69
New Training Set	[1] 294 69
New Testing Set	[1] 126 69

Response Variable: Salary

Before discussing the construction of the linear model, we must first do an analysis of our response variable. In particular, it may help to understand the distribution of NBA player salaries to help us understand the best way to develop our regression model. As we see from the first graph of distribution of salary, the data is very right skewed. This data deviates from the optimal normal distribution we would like to have our predictor at, so we attempt to normalize the data using a log() transformation on the response variable. The resulting distribution is shown below as the second graph. Here we can see the distribution is slightly better but is still somewhere in between being left skewed and normally distributed. Consequently, I would initially consider our final model to have a log() transformation on the Salary, as it seems to hurt more than help. Later, I will discuss why I removed this transformation, as I found it did not help our predictions.





Numerical Predictors

The next important step in creating the linear model is analyzing the relationships of the numerical variables with our response variable. Below we have a table of numerical predictors and their relative correlation coefficients with respect to the salary.

Predictor	**Correlation**
Ob	-0.056166846
G	0.297680506
MP	0.523499514
PER	0.384950280
TS.	0.215233447
X3PAr	-0.094733209
FTr	0.165224772
ORB.	-0.011577318
DRB.	0.240258043
TRB.	0.163207220
AST.	0.268287115
STL.	0.026993762
BLK.	0.096548167
TOV.	-0.042752252
USG.	0.295613111
OWS	0.611845192
DWS	0.570844552
WS	0.646419579
WS.48	0.263394925
OBPM	0.356879710
DBPM	0.209458161
VORP	0.626278500
Rk	-0.108736049
GS	0.582709367
FG	0.414049363
FGA	0.278722982
FG.	0.189564441
X3P	0.191617686
X3PA	0.113833021
X3P.	0.105766047
X2P	0.301289971
X2PA	0.204034976
X2P.	0.161019947
FT	0.518953561
FTA	0.477613417
FT.	0.183767266
ORB	-0.024290823
DRB	0.150610983
TRB	0.082656130
AST	0.245127577

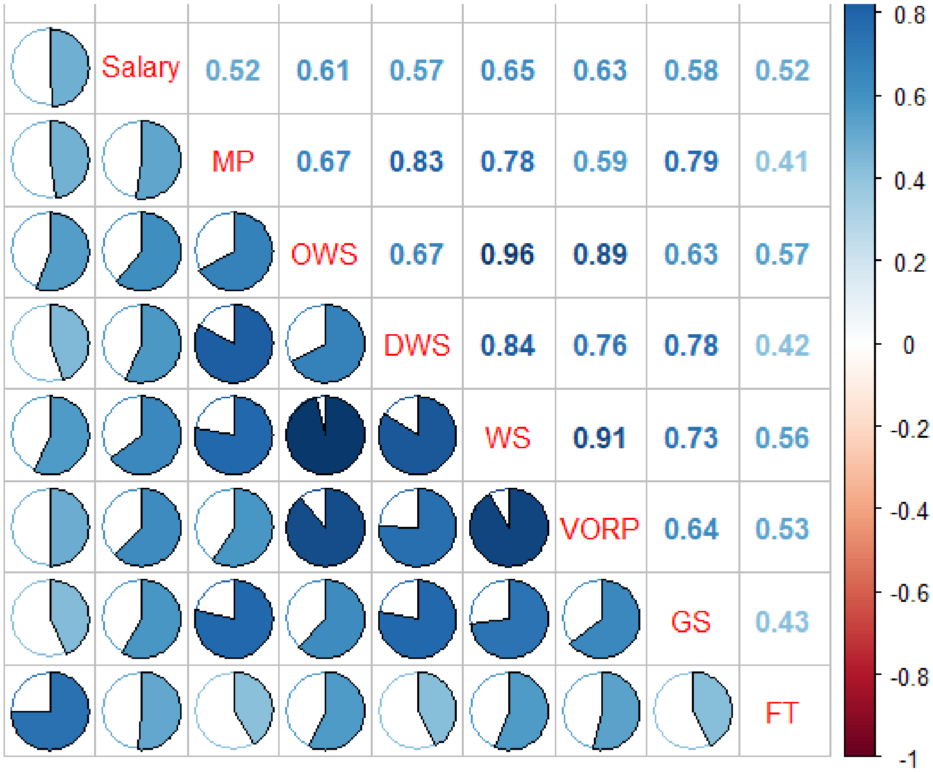
STL	-0.009284962
BLK	0.085045862
TOV	0.143860374
PF	-0.094834210
PTS	0.488528388
Ortg	0.251989286
DRtg	-0.081888825
Team.Rk	-0.064652428
T.W	0.090233782
T.L	-0.090233782
T.W.L.PERC	0.090115418
T.MOV	0.068237965
T.Ortg	0.073402954
T.DRtg	-0.029381132
NRtg	0.066360672
MOV.A	0.068368024
Ortg.A	0.072669383
DRtg.A	-0.025715821
NRtg.A	0.066462104

In this table we see many of the variables are not highly correlated with the salary. In fact, we can create a dramatically shorter list of predictors that have a correlation coefficient of at least 0.5 as follows:

Predictor	**Correlation**
MP	0.523499514
OWS	0.611845192
DWS	0.570844552
WS	0.646419579
VORP	0.626278500
GS	0.582709367
FT	0.518953561

This second table only contains 7 predictors that meet this criteria, these will be predictors we want to pay attention to in our final model since they have the strongest correlation with the response variable. However, with these variables there are pitfalls in their multicollinearity.

Correlation Plot



The final figure above shows the correlation plot of these 7 predictors against salary. Here we can see that most of these variables are not only highly correlated with the salary, but also highly correlated with the rest of the variables. For instance, WS and VORP have a correlation of 0.91, indicating that if we had both in our final regression model we would likely have a vif of greater than 5 in one or both variables. Consequently, we infer that using too many of these variables or the wrong subset of these variables would violate our multicollinearity assumption that the predictors are relatively independent. Regardless, we are likely to see some of these variables appear in our final model.

Categorical Predictors

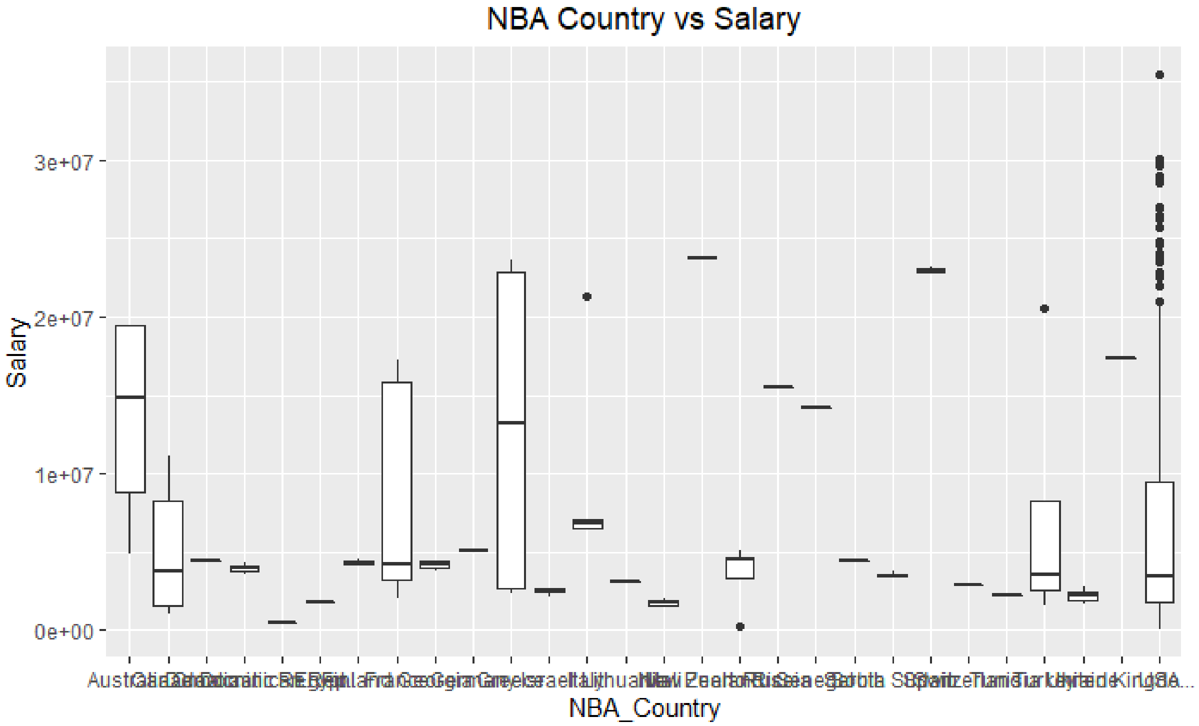
Given that we have established a good set of numerical variables to consider in our final model, we proceed to evaluate the categorical variables. Below we have created a table of categorical variables with the number of categories they contain.

Categorical Variables

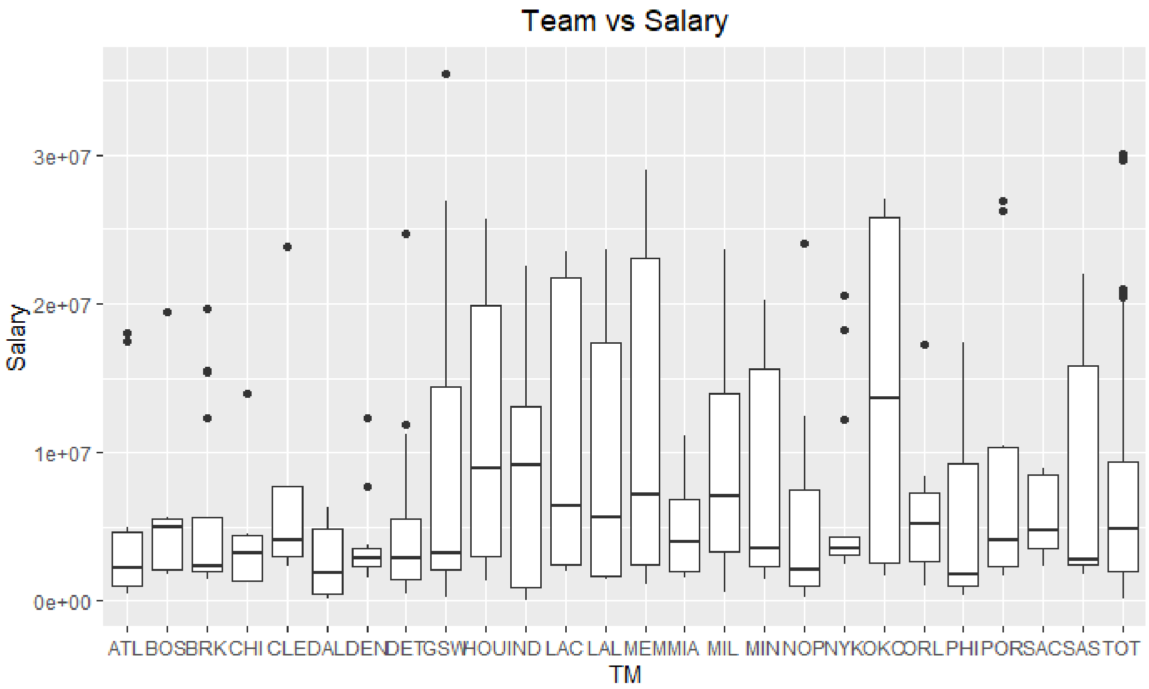
Predictor	**Available Categories**
NBA_Country	28

TM	26
Pos	7
T.Conf	2
T.Div	6

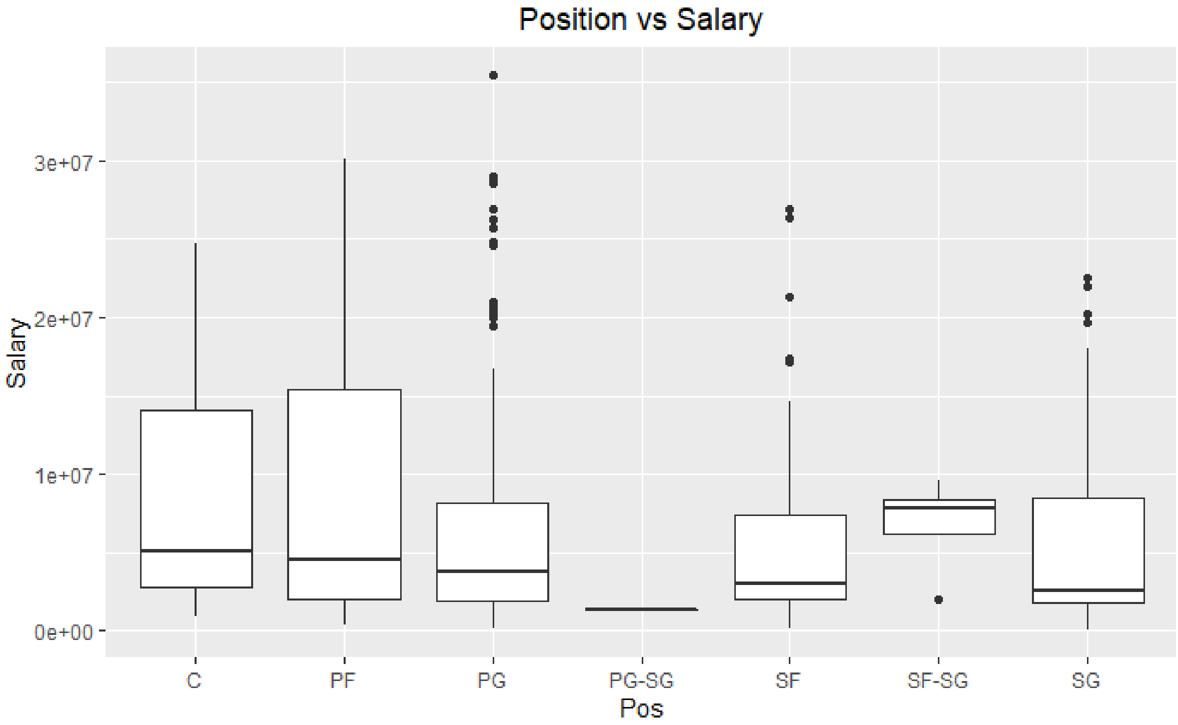
Now we will look at the relationship between these predictors and the salary. We will start with the variables with the most categories and work our way down to see if having fewer categories might indicate a better predictor. Therefore, let us inspect the relationship between NBA Country and Salary.



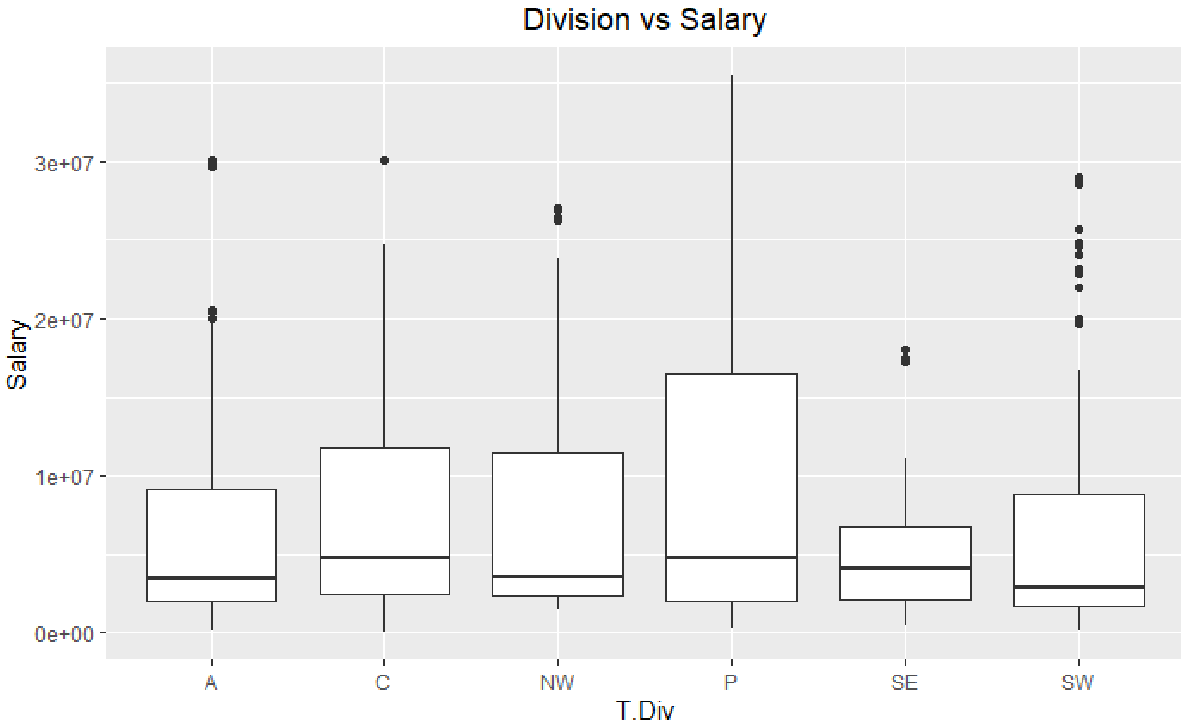
From the NBA Country vs Salary graph above, we note that the majority of the countries represented have either one or 2 players and of the other countries that are not the USA, the maximum number of players from any one country is 8. Since we have a small sample size from the remaining countries, it may be tough to use this data to predict unknown salaries in the testing data. However, we do note that the country with the highest number of top salaries is the usa. We can conclude that if we use this variable strategically, there is a reasonable amount of potential it can offer for our model.



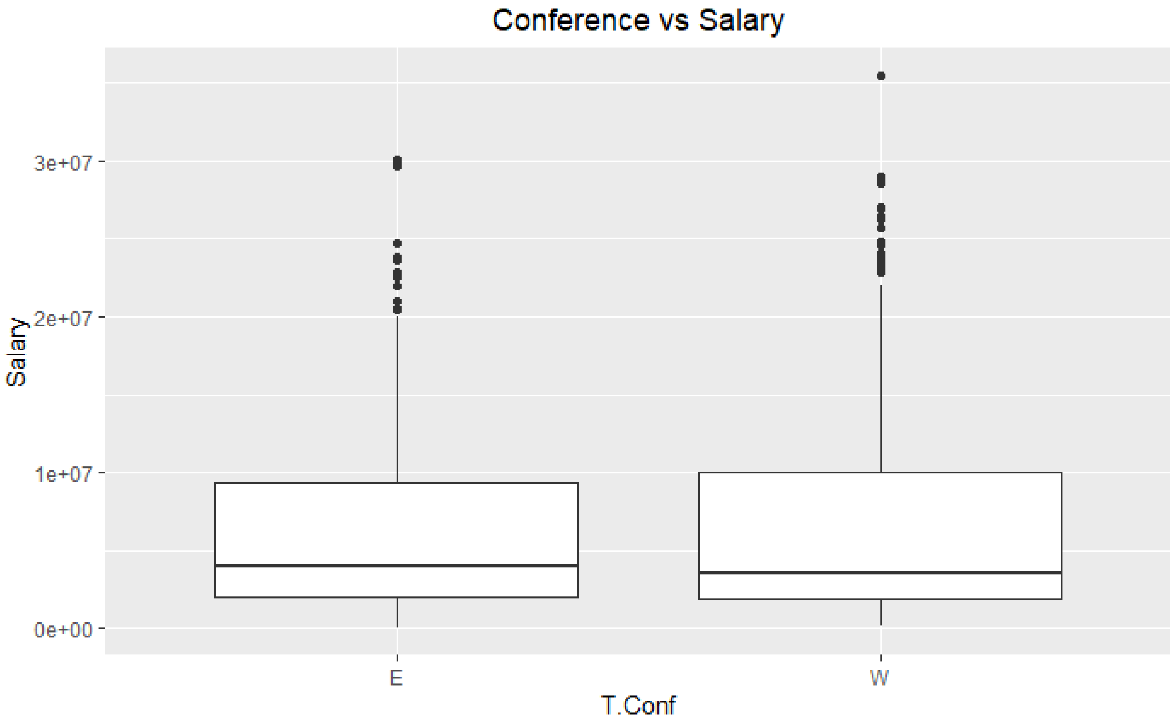
The above graph shows the distribution of salaries given the team. Clearly, there are a few teams like Oklahoma City Thunder, Houston Rockets, and Indiana Pacers that have higher average salaries than most other teams. However, there are relatively many teams with players that have extremely high salaries while the team average is relatively low. This indicates that Team may not be the most ideal categorical variable, but we will keep it in mind going forward.



From this Position vs Salary graph we see that there is only one player identified as “PG-SG” and only a few that are considered “SF-SG”, while the rest of the positions have numerous players identified as them. Moreover, the two hybrid positions have relatively low salaries in comparison to the non-hybrid positions. When it comes to the non-hybrid positions, the average salary is about the same and the only difference is in the wicks where the Centers and Power Forwards are paid slightly higher than Shooting Guards. All in all, it may be hard to use this variable to help predict salaries.



In the graph above, we can see that all the average salaries are within a narrow range indicating a poor predictor. However, we see that the 3rd quartile of SE is much lower than that of the rest, while the 3rd quartile of P is much higher than that of the rest. Again, we conclude that this variable may not be the best predictor but has some potential if utilized properly.



Finally, by analyzing the impact of conference on player salaries, we see that averages and quartiles are almost the same. The only difference is that there seems to be slightly more players in the West with higher salaries than in the East. However, this slight change is unlikely to compensate for how similar the salaries are for players in these respective conferences. Hence, this is likely a bad predictor for our model.

What we can conclude from the above analysis is that categorical variables are best if there are clear distinctions in salaries between categories. Further, we see that having fewer categories with similar size is better than having several categories each with 1 or 2 observations. Beyond this, there is not enough evidence to support that having fewer categories leads to better predictions to the salary of a player. However, it is true that having fewer categories leads to easier and better analysis of the categories, and how they can be utilized to predict salaries.

Creating Our Own Categorical Variables

In this section we explore the variables that were created in an attempt to improve upon the already existing variables. I began by choosing the existing categorical variables that have a lot of categories and trimming them down to only 2 or 3 categories. When investigating NBA__Country, I noticed that when I made it a factor and looked at its levels, I found that other than the USA the highest number of players that hail from one country is France –with only 8–compared to the USA with 344. Further, most of the other countries only had one player from there. This led me to modify this variable into two categories: “Forigen” and “Native”, a very self explanatory new variable. To test the validity of this new variable I created two simple linear regression models for the two different country variables and compared their R^2, these were the results:

R-Squared of SLR with NBA Country

Explanatory Variable	Multiple R-squared
Original NBA_Country	0.09658
New NBA_Country	0.0007706

First, we see that original variable's $R^2$was already very low but this new variable made our R^2even lower, nearing zero. This is a clear indication that this new variable is not a good consideration for our model. My second hypothesis was splitting the countries variables into categories based on continent; however, I ran into issues in which there were countries represented in my testing set that were not in my training set. This, alongside the fact that our starting $R^2$for this variable was fairly low, made it easy for me to not consider changing this variable at all. It would make sense to leave out this variable in our final model. as in the U.S it is illegal to pay someone less solely based on where they are from.

Next, I attempted to redefine the TM variable because it is known that different teams exist in different "markets" of different sizes, and consequently, the total player salaries of these teams would be different. Therefore, I attempted two approaches on categorizing teams based on total player salaries (an indirect measurement of market size). In the first attempt, I divided the 30 teams into three groups of 10: the 10 highest total player salaries were in the bigMarket group, the middle were in the medMarket group, and the last 10 were in the smallMarket group. In my second attempt, I divided the teams into two groups of 15: one for bigMarket teams and one for smallMarket teams. These were my results from SLR using the original TM variable and then the two new team variables:

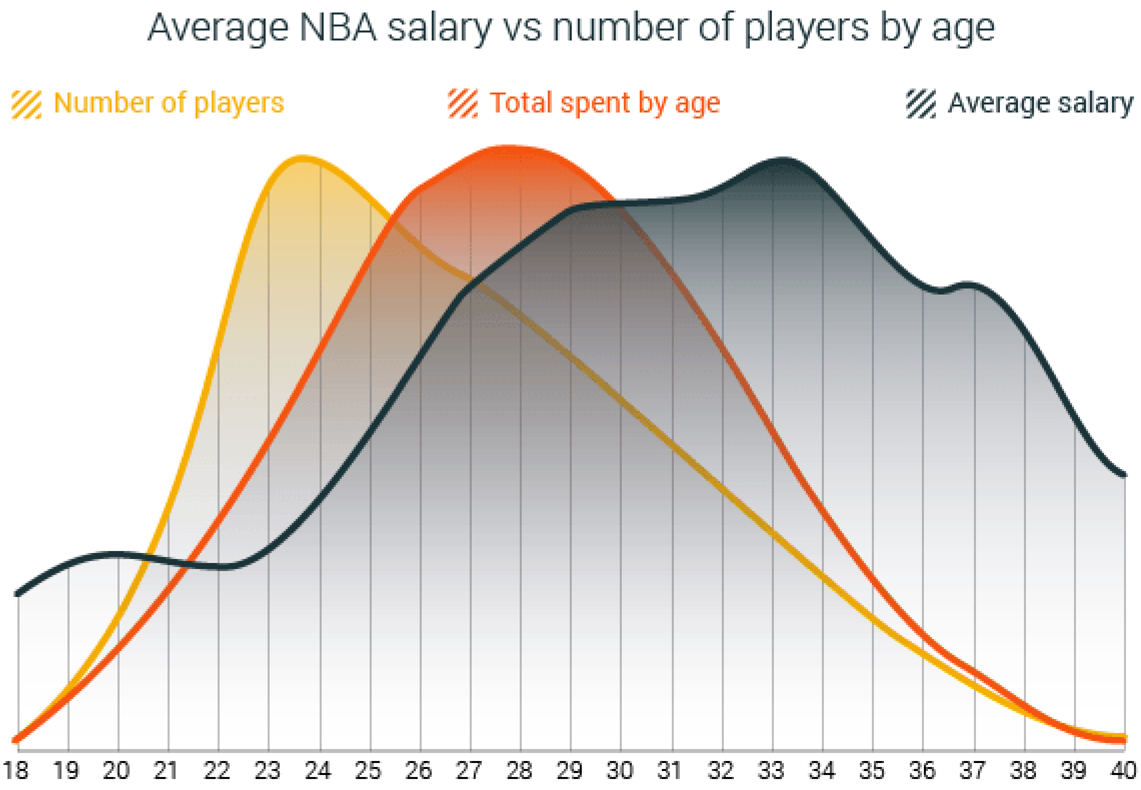
R-Squared of SLR with TM

Explanatory Variable	Multiple R-squared
Original TM	0.09874
3-Categorical TM	0.006775
2-Categorical TM	0.0001211

Again we see the shortcomings of our new variables as their $R^2$are much less (nearly zero) than the original R^2. I hypothesize that the loss is a result of using the table provided in the project details to make my groups. Hence, I have no hard evidence that these numbers actually reflect our dataset. Therefore, I attempt to make another variable based on the evidence we do have. Earlier, when examining the Team vs Salary graph, we saw that the Oklahoma City Thunder have the highest average salary with the Houston Rockets and the Indiana Pacers seemingly tied for second and third place. Our new teams variable will have four categories, the first three being the aforementioned teams and the last being "others" for the remaining teams.

The resulting R^2in our SLR is 0.03397, which is not nearly as high as our original variable, but again not nearly as low as the other two variables. I decided to keep this variable open for consideration since, while it had a lower R^2, it still had only 4 categories–and 4 betas–as compared to 26 betas using the teams variable.

I started considering creating categorical variables out of numerical variables. The first variable that I considered was Age. I was led to conclude this would be a good variable to manipulate, because in sports there are usually ranges for rookies who have not yet reached their full potential, players in their prime, and older players who are approaching the sport's retirement age. To find the intervals for these categories, I would need to do some research. In my research, I came upon a source that had a great graph to illustrate the cutoffs for these groups (Curcic).



From the average salary curve, we can see that the rookie range at any age before the salary increase rate seems to plateau at about 26. The prime years would be from 26 to 33, leaving the old aged players in the range of 33 to 40 or more. However, when testing these cutoffs I found that the best results come from using the following values: rookies are younger than 24, players in their prime are 24 to 28, and old players are 28 and beyond. These are the SLR results using the new variable in comparison to the original variable.

R-Squared of SLR with Age

Explanatory Variable	Multiple R-squared
Original Age	0.08776
New Age	0.1353

As evident from the table, our new age variable produces an $R^2$of 1.5 times that of theR^2 produced using the original age variable. Hence, we will replace the old age variable with the new age variable in our final model if we decide to keep age (which ends up being the case).

Next we will try to make GS into a categorical variable. The intent is to identify the starters, bench players, and backup bench players. I first attempted this with MP and found roughly the same results using the same procedure and categories. I only included my process for GS, because it had better to interpret numbers. First, I determined the cutoffs by analyzing the summary of GS.

Summary(GS)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	1.00	7.00	21.74	41.50	82.00

Using this table we allow the starting players to be any player that started in 41 games or more. Then the bench players would be those who started in at least 7 games (but less than 41). Finally, The backup bench would be any player who started in fewer than 7 games. My resulting R^2's from the SLRs are shown below.

R-Squared of SLR with GS

Explanatory Variable	Multiple R-squared
Original GS	0.3396
New GS	0.2942

The table shows that the original GS variable is a better predictor. However, since these cutoffs were set using an educated guess, I attempted to try different numbers within a reasonable range of our original cutoffs to determine which produces the best new predictor. As a result, I found that the best upper limit is 49 and the best lower limit is 11. These new intervals give us a $R^2$of 0.3364, seemingly identical to our original GS variable; but, by making it a categorical variable we have increased the number of betas from one to three.Thus, we will likely leave this new variable out of our final model, but not forget that it exists as an option.

Next is BPM, I choose this because it is an advanced player statistic that is proven to be a good indicator of a players output per game: it is often referred to in discussions of players' contribution on the court. From past experiences as a viewer of NBA games, I know that a BPM of +4 is considered All-Star status while a BPM of 0.0 is considered average, so any negative BPM is considered bad. I used these cut off values to create my new BPM variable grouping players into one of three categories: “GOAT”, “Good”, or “Meh". After creating my initial SLR with these cutoffs, I found that I can improve my $R^2$by just shifting my requirement for GOAT to be a BPM of +2. We are left to test our new variable using a SLR as we have done before.

R-Squared of SLR with BPM

Explanatory Variable	Multiple R-squared
Original BPM	0.1205

New BPM	0.2933

From the table above we can easily conclude that our new variable is very good, since the new $R^2$is almost 3 times that of the R^2using the original variable. Hence, we will redefine BPM as this new categorical variable and intend to use it in our final MLR.

Finally, the last variable created is called TopBracket that very cleverly determines if a player is or is not in the top 25% of earners in the NBA. This was done by performing a logistic regression to determine if a player was in the top earners bracket based on predictors of my best model at the time (FT + GS + VORP + USG. + FG + G + Age + MP + T.Div + Pos + BPM). If the logistic regression finds–based on other predictors–that one is in the top bracket of salary earners, then it outputs "yes," otherwise "no". This turned out to be a fantastic predictor, as can be seen when we perform SLR with TopBracket as the explanatory variable. Our result is a model with a $R^2 = 0.7788$. This makes sense, since we are indirectly using the training data's salaries to place players in the appropriate bracket.

Step Functions

Now that we have done all the preliminary steps of understanding our response and explanatory variables, we can begin to build our model. To begin, use the forward and backward step functions with both AIC and BIC as measurements of the model's "goodness." Here are the resulting R_{adj}^{2} for these models:

R-Squared for Given Model Type

	AIC	**BIC**
Forward	0.7675	0.6877
Backward	0.7743	0.6992

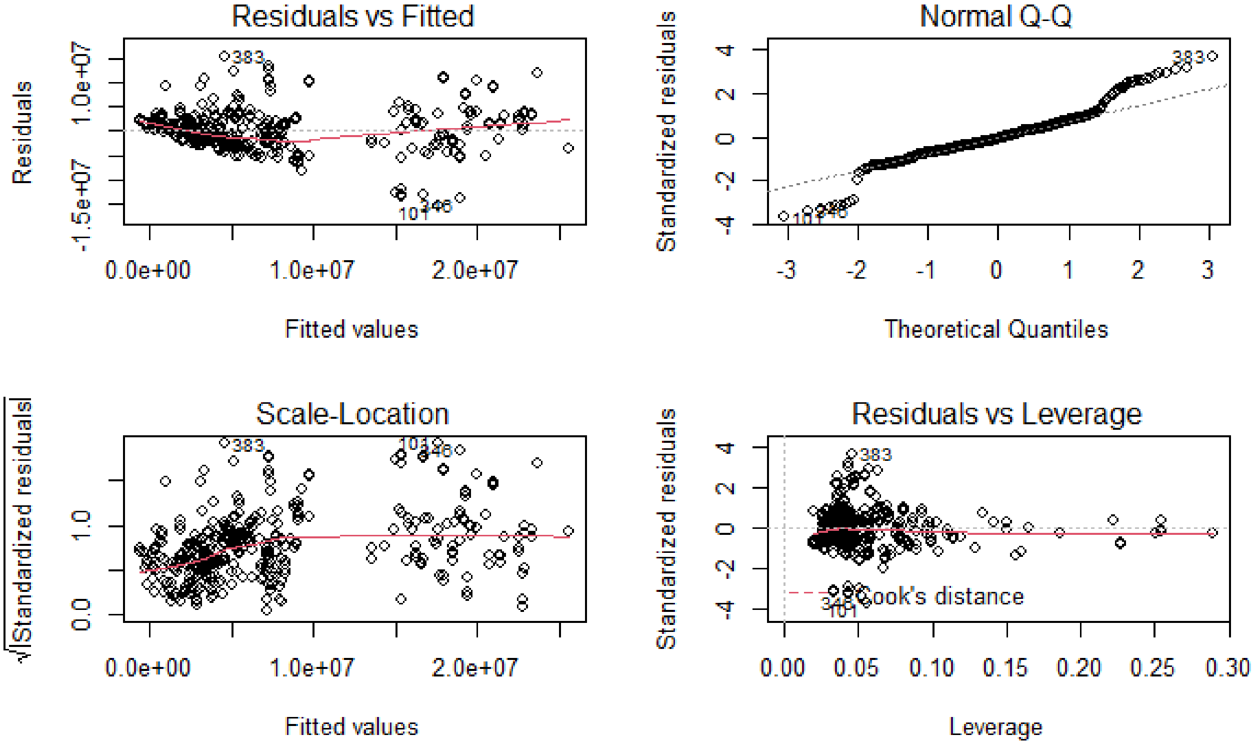
From the above graph, it is very easy to conclude that we should lead with the backward step function in both cases of measurement of goodness. However, this is a critical factor to consider when following through with this conclusion, and it is the resulting degrees of freedom. Since forward step functions start at nothing and pick the best predictors to improve their AIC/BIC, they tend to end up with fewer predictors than step functions that start with the full model and work their way back. This concept is depicted well by the table below:

Degrees of Freedom for Given Model Type

	AIC	**BIC**
Forward	344	410

Backward	340	403

We see that the forward step functions give us better degrees of freedom than their backward counterparts. Of course, we see that the BIC models produce much better degrees of freedom than their AIC counterparts since there is more penalty in BIC models for using more predictors. Our goal is to balance maximizing the R_{adj}^{2} with the minimum number of predictors as possible. For my initial model, I choose to use the forward AIC model. I noticed that I can compensate for the horrible degree of freedom by removing the NBA_Country and TM variables since they have 27 and 26 betas respectively. As we noted earlier, their R^{2}in a SLR with Salary was not all that good. Thus, our initial model is Salary ~ TopBracket + WS + Age + USG. + GS + FTA + DBPM + BLK. + Pos + X2P + G + STL. + DRB. + Ortg + DRB + Rk + OWS with $R_{adj}^{2} = 0.6872$on 396 degrees of freedom. This model gave us a $R^{2} = 0.48558$on Kaggel which is a borderline average start. The diagnostic plots for this multiple linear regression can be seen below. Here there are several violations to what we assume for linear regression; however, we will deal with these issues in our final model.



Improving Initial Model

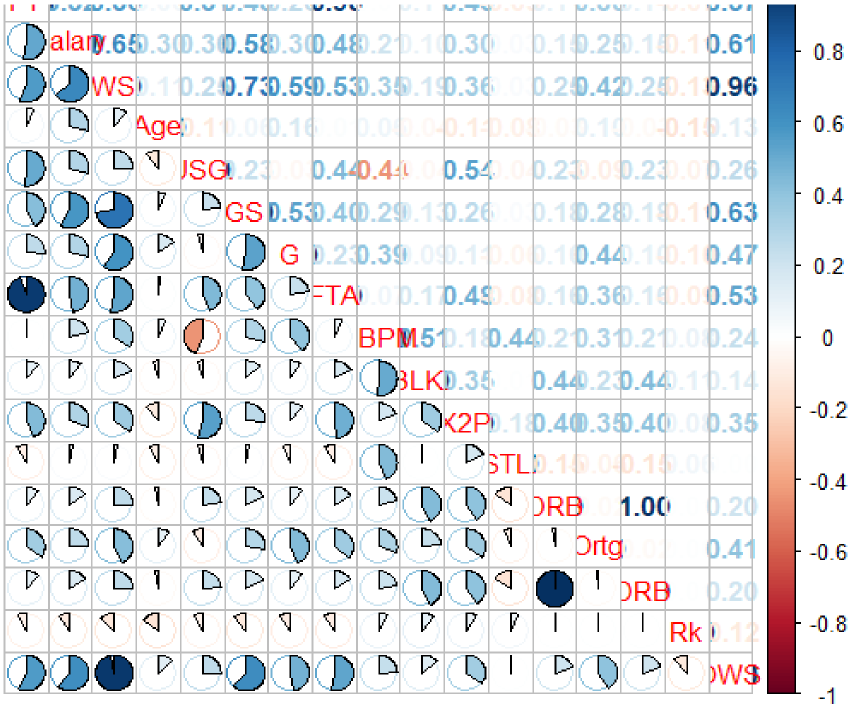
Now that we have a starting point, we will continue to improve our model by removing the least significant variables and replacing them with better variables. The first step of pruning our variables is analyzing the multicollinearity of the model.

vif(currentModel)

```
                 GVIF   Df    GVIF^(1/(2*Df))
TopBracket   2.397053  1       1.548242
  WS        43.227802  1      6.574785
  Age        1.602994  2      1.125208
  USG.       5.633605  1       2.373522
  GS         3.105582  1      1.762266
  FTA        2.137445  1      1.462000
  DBPM        8.813604  1      2.968772
  BLK.       2.405703  1       1.551033
  Pos        4.213374  6      1.127334
  X2P        3.556427  1      1.885849
  G          2.640019  1      1.624813
  STL.       2.560724  1      1.600226
  DRB.       6.773244  1      2.602546
  Ortg       2.174400  1      1.474585
  DRB        3.975580  1      1.993886
  Rk         1.232692  1      1.110267
  OWS       28.402492  1      5.329399
```

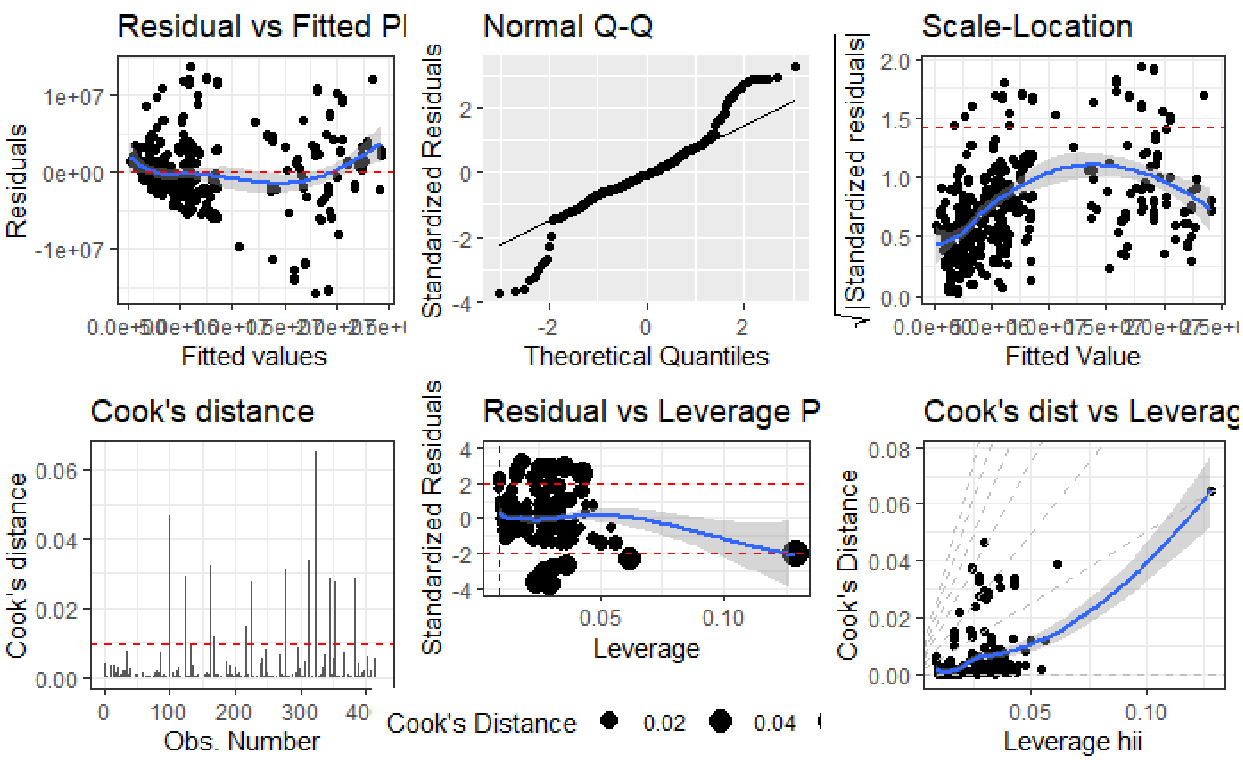
Here we see that WS and OWS violate the maximum allowed VIF of 5 we have for a predictor. Therefore, we will investigate the correlation plot of these variables and remove those that have high correlation with many of the other predictors.

Correlation Plot



Using the above graph, we remove as many variables as possible that are highly correlated with other variables, while keeping those that are low to moderately correlated but have a decent correlation with Salary. The remaining model we get is Salary ~ FT + GS + VORP + USG. + FG + G + Age + MP + T.Div + Pos + BPM + TopBracket with $R_{adj}^2 = 0.6863$on 396 degrees of freedom. This gave us a Kaggle score of $R^2 = 0.53164$, a large improvement over the initial model.

Now to get the final model, we would run backward and forward AIC/BIC models as we did before and choose one that gave the best $R^2$with the best dF (degrees of freedom). The resulting model was Salary ~ FT + G + Age + MP + BPM + TopBracket with a R_{adj}^2 =0.6835 on 411 dF. Using only 6 Predictors (9 betas overall including intercept), we were able to get an $R^2 = 0.55327$on Kaggle. The final model without any transformations had the following diagnostics:



Without needing to view these plots individually, we can see clear violations in all the assumptions for linear regression. The errors are non constant and as evident from Residual vs Fitted Value plot. The normality assumption of our data is evidently violated in our Normal Q-Q graph. Finally, we can see we have several outliers in the Residual v Leverage plot. Further using this plot, we can see we have numerous leverages. In fact, the Cook's Distance plot gives us a good idea as to the many bad leverages we have (these will be quantified later after the discussion on transformations). We will attempt to get rid of these violations with transformations on our variables in the following section.

Transformations

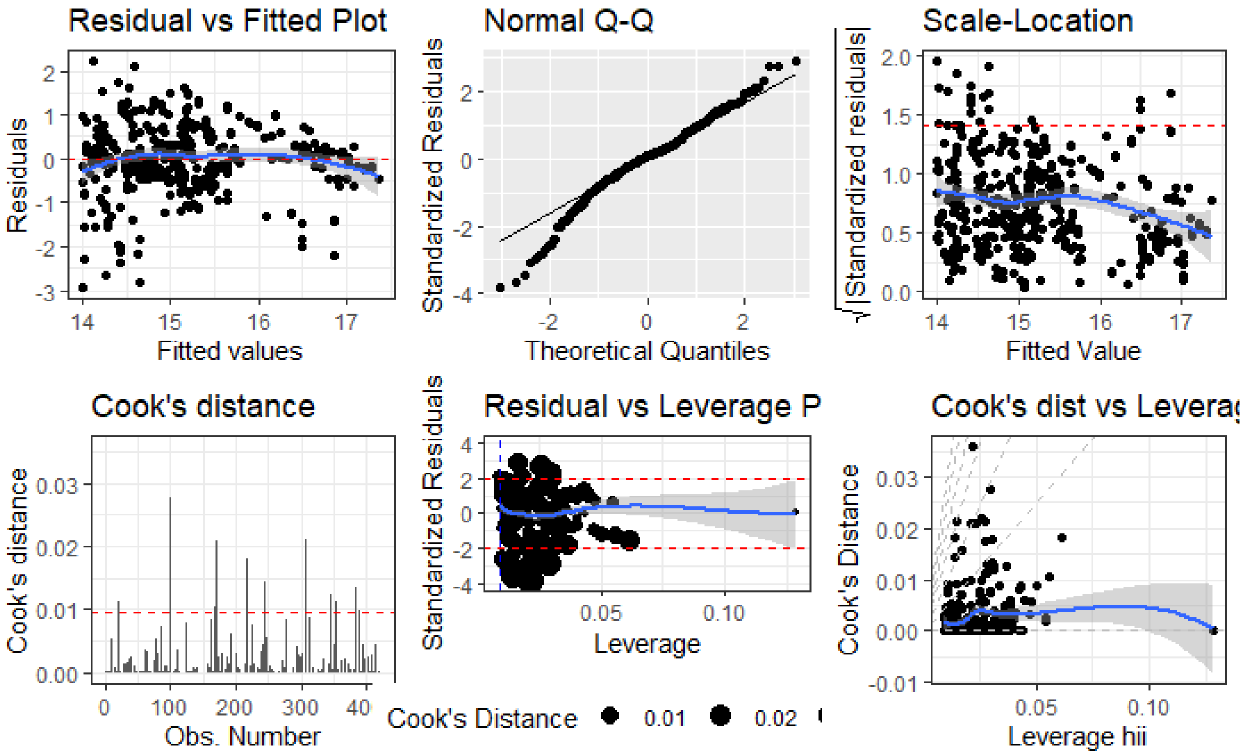
To deal with the violations in our diagnostics and to attempt to improve our model using only our current predictors, we will apply transformations to our variables. As we discussed in the response variable section, we will investigate the effects of performing a log() on salary. Below we compare the MLR with an untransformed response variable and the MLR with log(salary).

Comparing Fit of log(Salary) with Salary

Response variable of MLR	Adjusted R-squared

Salary	0.6835
log(Salary)	0.5389

Here we see that performing a log on our model results in a R_{adj}^{2} reduction of 0.1446, which is a significant loss in the fitness of our model. But, does this new model fix any of our diagnostic plots?



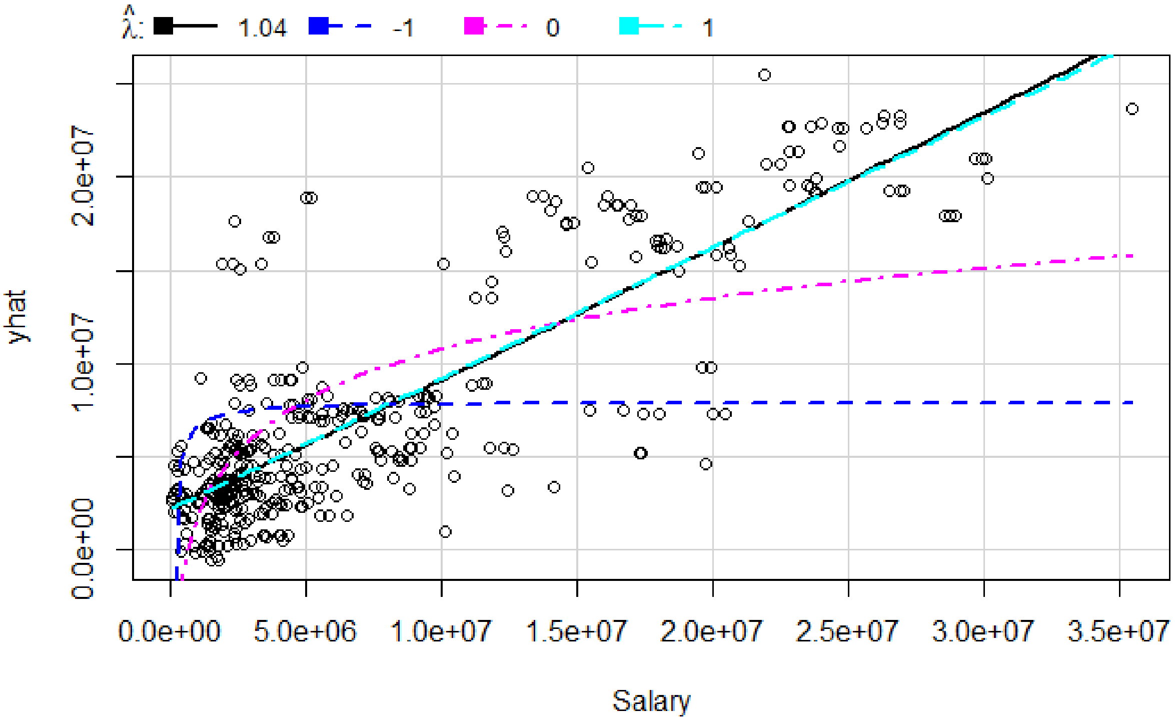
We see that our new model has overall improved our diagnostic plots. In particular, the variation in our errors is nearly constant as is represented by the blue line of our Residual vs. Fitted Value graph. Further, the normality of our data has improved, but we still face some violation at the ends of the normal Q-Q graph. Lastly, we see that we have fewer bad leverage points from the Cook's distance graph. With these improved diagnostics, we must consider if we are willing to sacrifice 0.1446 in our R_{adj}^{2} for some improvement in our diagnostics. At this moment, it is best that we do not discount the log(salary) model but proceed with the original non transformed model as our final model.

Next, we will apply inverseResponsePlot() to our final model and see if it offers a log as the best transformation or if it suggests a better power to improve our model's fitness and/or diagnostics.

inverseResponsePlot(finalModel)

Lambda	**RSS**

1.036853	5.030333e+15
-1.000000	1.600909e+16
0.000000	8.392487e+15
1.000000	5.032742e+15



In the above graph, we see the suggested lambda is 1.04, which improves our RSS by an order of $10^{16}$from the original model. This suggested transformation is not worth the computational effort it would take to find $Salary^{1.04}$,because the RSS would virtually stay the same. Further, we notice that a lambda of 0.00, indicating a log transformation, would decrease our RSS by $3.36 * 10^{15}$. This, in combination with the decreased $R_{adj}^{2}$of 0.1446, gives us enough evidence to dismiss using the log(salary) model.

Now, we will analyze the effect of transforming our predictors using powerTransform(). The following table depicts the resulting lambdas for each numerical predictor when using powerTransform().

Suggested Lambdas for Predictors

Predictor	**Lambda**

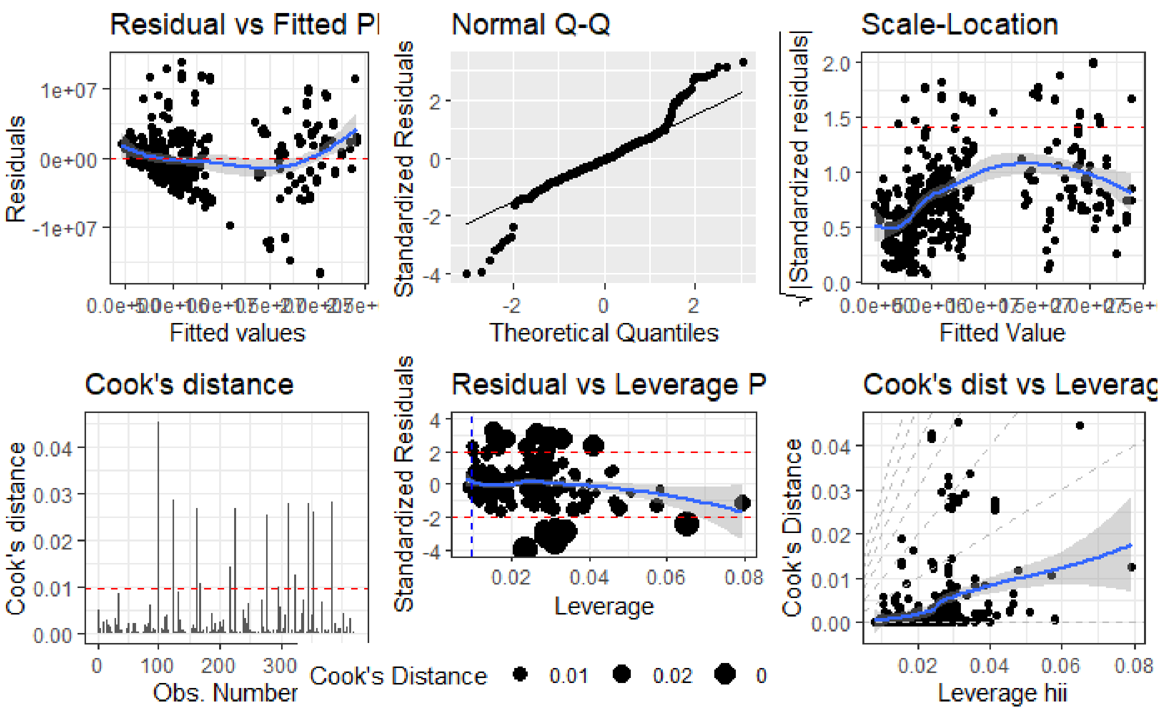
FT	0.42 ⇒ sqrt()
G	1
MP	0.5 ⇔ sqrt()

Using these lambdas, we will illustrate the $R_{adj}^{2}$of the MLR of the model with transformed predictors as it compares to the original non transformed model. In addition, we provide the $R_{adj}^{2}$of the log(salary) model and the log(salary) model with transformed predictors

Comparing Fit of Transformed Models

Multiple Linear Regression Model	Adjusted R-squared
Salary ~ $FT + G + Age + MP + BPM + TopBracket$	0.6835
log(Salary) ~ $FT + G + Age + MP + BPM + TopBracket$	0.5389
Salary ~ $\sqrt{FT} + G + \sqrt{MP} + Age + BPM + TopBracket$	0.6902
log(Salary) ~ $\sqrt{FT} + G + \sqrt{MP} + Age + BPM + TopBracket$	0.5494

From the table, we see that the model that produces the best fit is the one with transformations only on the predictor. Notice that this only increases the R_{adj}^{2} by .0067. However, when we try to submit this model on Kaggle, we get $R^{2} = 0.53964$,which is a loss of 0.0136 from our original model. Let us analyze the diagnostics of this model to see if it is worth the loss in our R^{2} .



Notice that these plots nearly identically reflect the plots of our original model. Therefore, we conclude that since the diagnostics are not improved and the R^2 is worsened on Kaggle with the transformed predictors model, the ordinal model is our best model. After some discussion with classmates, I found that I was not the only one who found that transformations were ineffective in improving the diagnostics or the fitness of their model. There were some instances that people used some transformations on their predictors, but as we saw, our best model does not modify the predictors. Hence, we have arrived at our final model: Salary ~ $FT + G + Age + MP + BPM + TopBracket$.

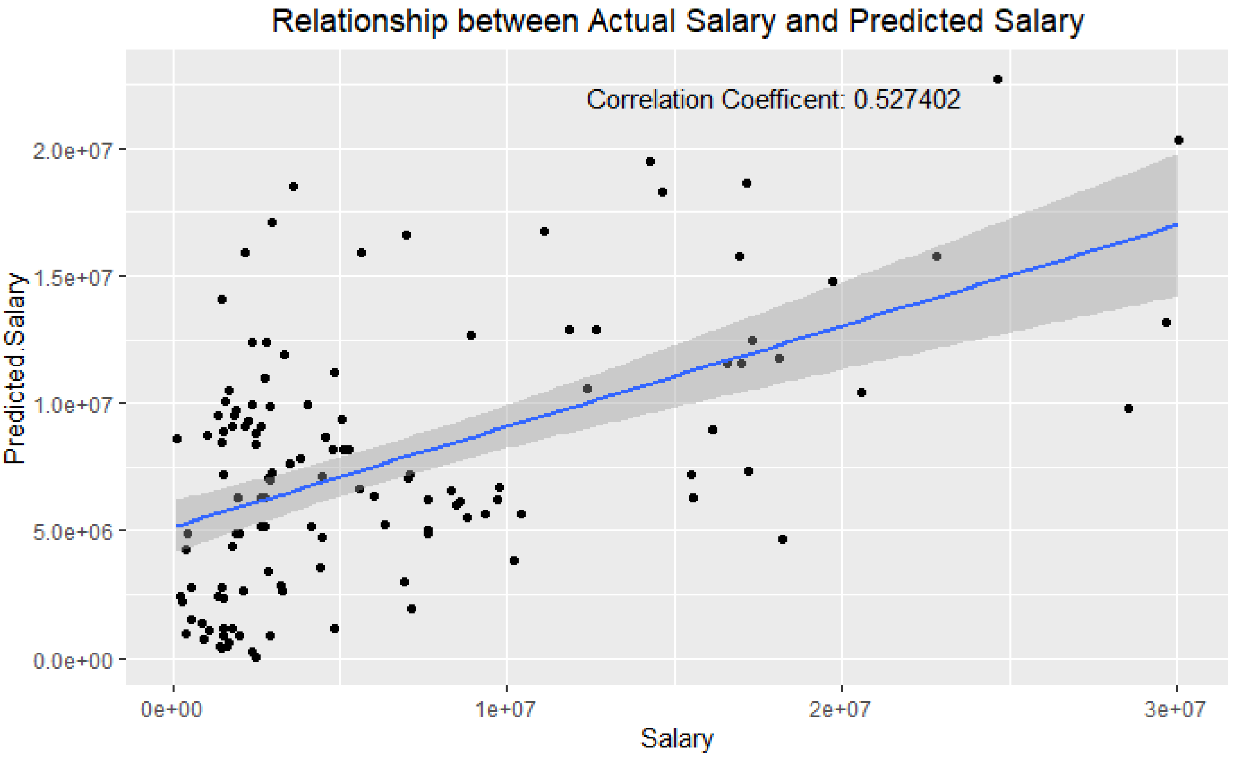
Results and Discussion

After several attempts at trying to reach a final model, we arrived at a simple model that uses only 6 predictors translating to 9 betas, including the intercept. The final model with all respective coefficients is given below:

$$Salary = 8469551.9 + 410758.1 * FT - 63695.1 * G - 2931013.8 * AgePrime - 3837816.6 *$$

$$+ 3411.0 * MP - 3398724.2 * BPMGood - 3081170.5 * BPMMeh + 9254655.0 * TopBracketYe$$

To test the accuracy of this model without simply getting out Kaggle score, we use our split data set to give us a good idea on how we did with predicting the test data. Below we see the graph between the actual and the predicted salaries with a label for the correlation coefficient.



As expected from previous discussions, the correlation is relatively average: 0.527402. This value is slightly lower than the Kaggle R-squared value, likely because the training data set used above is smaller than the training data set we used for the predictions that were submitted on Kaggle. This plot gives us a rough idea of where along the data our model predicts the correct salaries. For instance, we see that for salaries up to $\$(1.0 * 10^{7})$,our model does relatively well when compared to much larger salaries. This is a result of there being more data points in the smaller salary range, making it easier for our model to predict their actual values. However, in my comparison there are fewer top salary earners, so it is harder for our linear model to accurately predict those values. Below we will take a look at our training and testing model’s and compare their fitness and the significance of their predictors.

Training Model Summary:

```
                    Residuals:
      Min       1Q    Median      3Q       Max
-14732674  -2271492  -173750  2139294  13113364

                   Coefficients:
              Estimate     Std. Error   t value    Pr(>|t|)
(Intercept)   9361013.3    1353214.2    6.918      3.03e-11 ***
```

```
FT            518365.7      158562.8    3.269      0.001211 **
G             -74772.8       24491.8   -3.053      0.002480 **
AgePrime    -3668876.3      649023.4   -5.653      3.83e-08 ***
AgeYoung    -4409300.9      710068.8   -6.210      1.87e-09 ***
MP              4297.9         851.9    5.045      8.09e-07 ***
BPMGood     -3734619.7     1064016.6   -3.510      0.000521 ***
BPMMeh      -4027316.2     1008155.8   -3.995      8.25e-05 ***
TopBracketYes 7127807.2     917739.5    7.767      1.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error: 4528000 on 285 degrees of freedom
Multiple R-squared:  0.6838,  Adjusted R-squared:  0.6749
F-statistic: 77.02 on 8 and 285 DF,  p-value: < 2.2e-16
```

Testing Model Summary:

```
Residuals:
     Min       1Q   Median      3Q      Max
-12085738 -2981092  -276155  2950166 15961324

Coefficients:
               Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)    12737661    2210036     5.764    6.81e-08 ***
FT               748660     286189     2.616    0.010070 *
G               -148916      39761    -3.745    0.000281 ***
AgePrime       -2409303    1212641    -1.987    0.049279 *
AgeYoung       -5166167    1179700    -4.379    2.61e-05 ***
MP                 5897       1425     4.139    6.61e-05 ***
BPMGood        -3586901    1803997    -1.988    0.049112 *
BPMMeh         -5968009    1564915    -3.814    0.000220 ***
TopBracketYes   -258714    1029328    -0.251    0.801990
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error: 4873000 on 117 degrees of freedom
Multiple R-squared:  0.5006,  Adjusted R-squared:  0.4664
F-statistic: 14.66 on 8 and 117 DF,  p-value: 1.116e-14
```

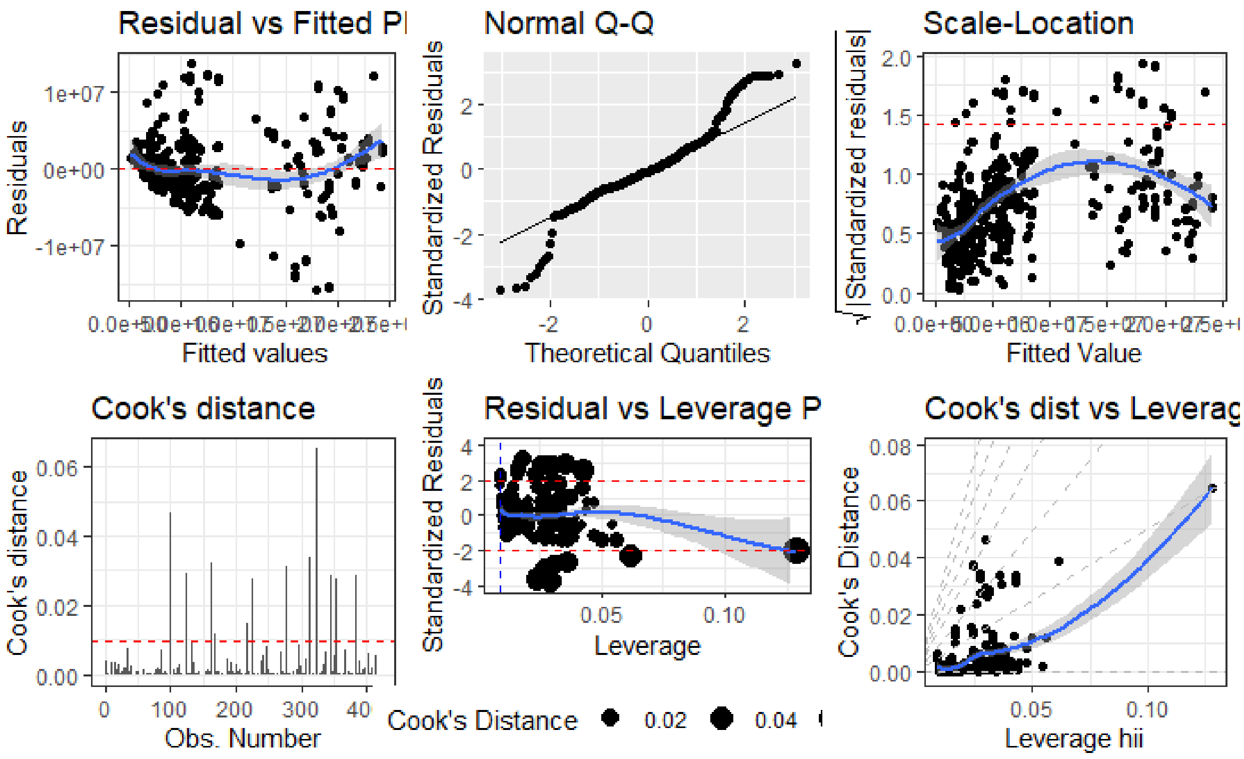
Above we see the model applied to the training and testing data separately. In the training data we get an adjusted-R-squared of 0.6749, while in the testing model our adjusted-R-squared falls off to 0.4664. Analyzing the significance of our variables, this difference in fitness starts to make some sense. We see that the predictors in the training model are all the highest degree of significance, so they all are maximally important to modeling our training salary. However, out of 9 betas in the testing model, we see that only 5 are significant to the highest degree, while 3 are slightly significant, and one is not significant. This could lead us to want to explore how we can make our betas be significant for both models. Accomplishing this would require trial and error in redefining some of our variables or altogether removing consistently insignificant ones with better variables.

Below we will confirm that we have no violations in our multicollinearity assumption by examining the output of vif.

```
                  vif(finalModel)
                    GVIF     Df     GVIF^(1/(2*Df))
FT               1.498865    1      1.224281
G                5.435769    1      2.331474
Age              1.242098    2      1.055696
MP               6.767064    1      2.601358
BPM              1.641074    2      1.131832
TopBracket       2.337119    1      1.528764
```

From the results we see that all our vif values are less than 5 and, in fact, at no greater than 3, indicating that we have relative independence in our predictors. This indicates our model is using good predictors for multiple linear regression, as they do not share much of the same information.

Next we will take a look at diagnostic plots of our final model:



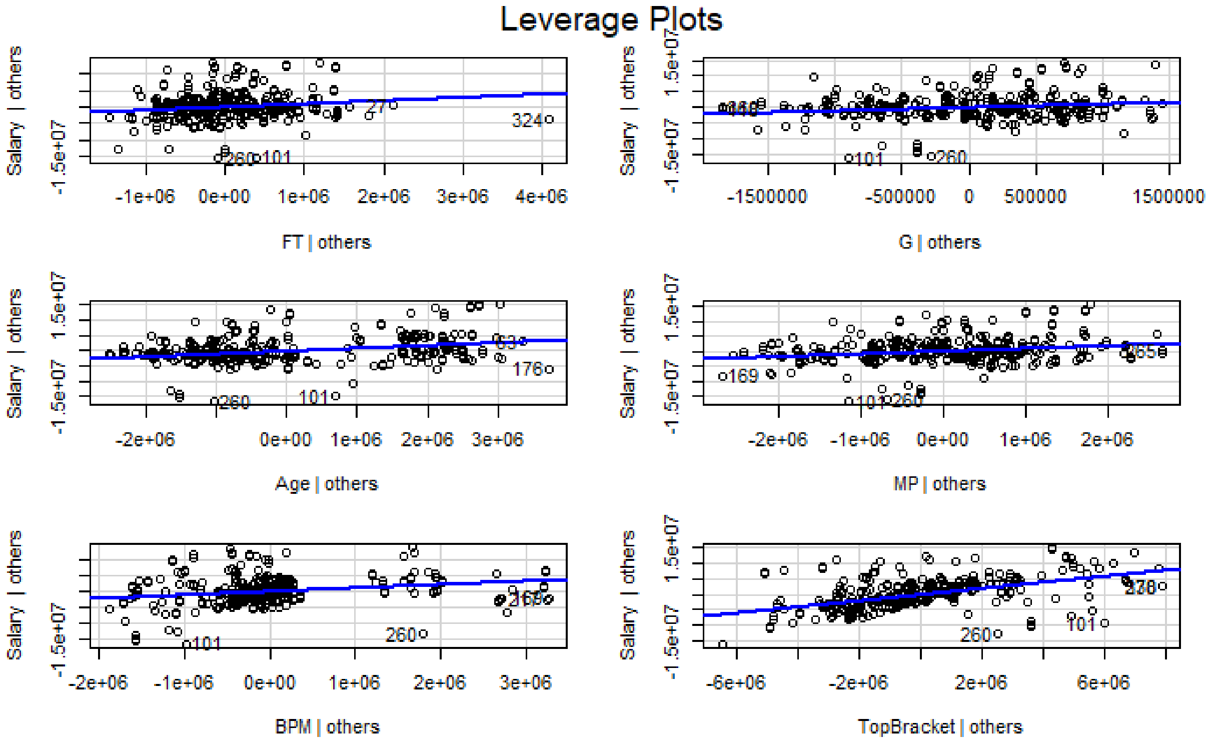
As formally discussed, the diagnostics plots show violations in most of our assumptions for linear regression. Firstly, from the Residual vs Fitted plot, we see that the assumption of constant variance in our residuals is violated; the blue curve is not close enough to being a horizontal line. Further, the Normal Q-Q plot illustrates that we have a normality violation at the ends of our data. Finally, we see that the errors are not randomly distributed, as they have a megaphone-like pattern in the first plot. We noted earlier that performing a log() transformation on salary would improve these diagnostics, but it would come at a detrimental cost to our model. Furthermore, we discussed that others attempting to build their own models had a similar issue in trying to fix the diagnostics plots but failed because the fixes (if there even were any in their case) would hurt their model significantly. These reasons led us to accept our final model despite the violations in the diagnostics.

Now we will examine our leverage points and outliers to see how our model performed in this category. Below is a 2-way table with leverages being described column wise and outliers being shown row wise.

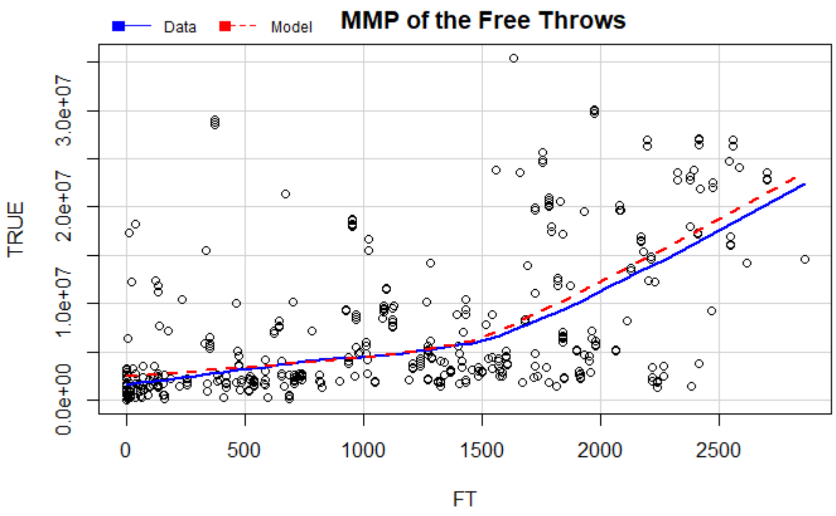
2-way table of Leverages and Outliers

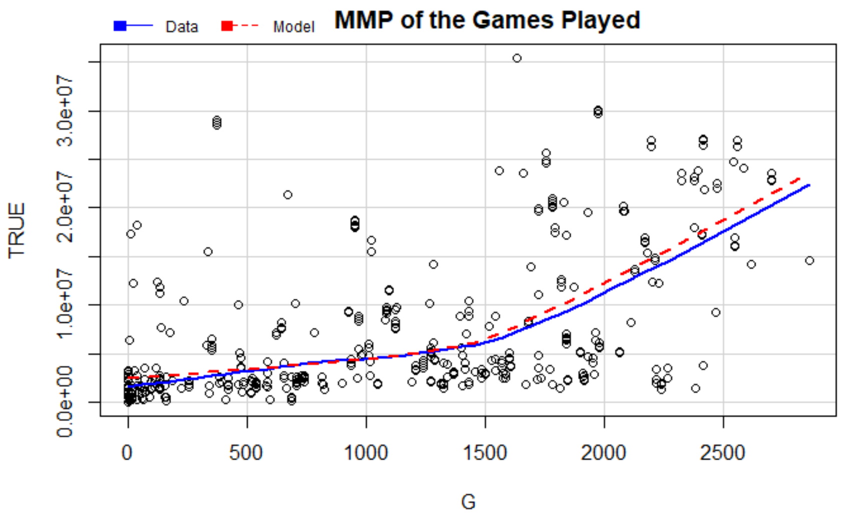
Leverages / Outliers	**No**	**Yes**
No	301	87
Yes	18	14

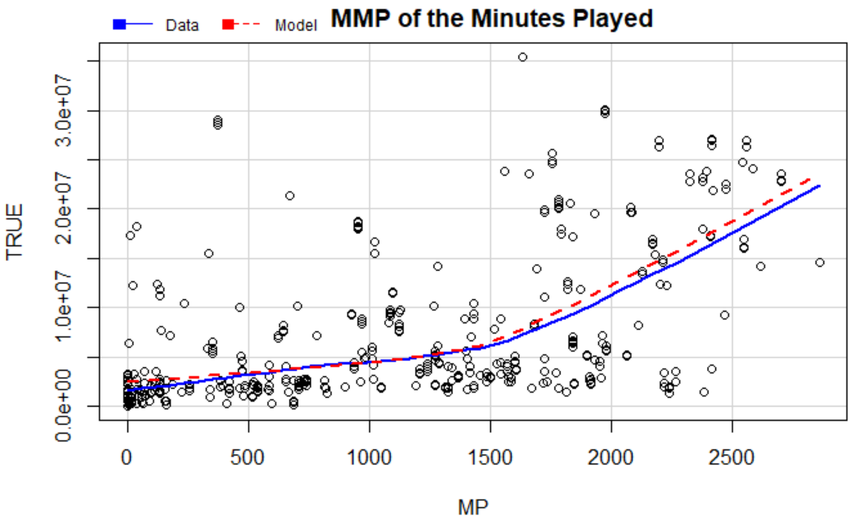
From this graph, we see that we have 301 points that are neither leverages nor outliers, so they are “normal” points. Further, we extract from the table that there are 87 outliers that are not leverages. Lastly, we note that we have 18 good leverage points and 14 bad leverage points. A pro from these results is that we have more good leverages than bad leverages; however, we still have 14 bad leverage points which is more than we like. These bad leverage points are likely what is hurting our model in estimating the top salary earners. Below we have provided the leverage plots for any additional information about the leverage trends.



Additionally, we will take a look at our mixed marginal plots for each of our numerical predictors:







These plots depict how the model maps the data using each predictor. In summary, all of these plots have a model that is very close to the ideal trend of the data. Therefore, our predictors do not require any manipulation to perform accurate predictions in our data.

Lastly, the final anova table of our model is provided below.

```
                 Analysis of Variance Table
                     Response: Salary
              Df     Sum Sq     Mean Sq F value    Pr(>F)
FT             1 6.5089e+15 6.5089e+15 356.536 < 2.2e-16 ***
G              1 6.9847e+14 6.9847e+14  38.260 1.497e-09 ***
Age            2 1.9897e+15 9.9487e+14  54.496 < 2.2e-16 ***
MP             1 3.8810e+15 3.8810e+15 212.588 < 2.2e-16 ***
BPM            2 1.0593e+15 5.2967e+14  29.014 1.634e-12 ***
TopBracket     1 2.5279e+15 2.5279e+15 138.469 < 2.2e-16 ***
```

```
Residuals  411 7.5032e+15 1.8256e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Here we have more evidence of the significance of our predictors instead of having to look at each individual beta as we had when analyzing the summary of our final model. Additionally, we have information on the RSS of each variable and their partial F-scores.

Limitations and Conclusions

As noted in previous sections, the glaring issue we have in our model is the erroneous diagnostic plots. From these plots we can determine we have violations in the normality of our data, randomness of our residuals, and constant variance in our errors. From this we conclude that it might be best to try another model form, perhaps one that allows for curves to better represent the trend of our data. Further, it can be said that since the log(salary) model produced better diagnostics, then we should reconsider using this model. This would be an immediate solution but would require a lot of work in trying to improve the adjusted R-squared.

In addition, we note that our R-squared value on kaggle is only 0.55 when the top score is 0.77. While this top score is still considered a decent model at best, it is still much better than our current model. Hence, we conclude that our model is not reaching its max potential and can likely improve by adding new good predictors while removing/modifying current predictors that have low significance in the model for the testing data. Nevertheless, we found that we can get an average R-squared with very few predictors and in turn very few betas. The lack of complexity in the number of predictors and transformations helps compensate for the overall fitness of the model.

All in all, our final model’s greatest strengths are simplicity and using inventive categorical predictors, while its major weaknesses are fitness and proper diagnostics. Moving forward, we would like to improve upon our weaknesses by finding more good predictors; this should include interaction predictors. Further, we need to investigate the potential of working with the log(salary) model for improved diagnostics. However, these efforts for improvement need to be balanced by what is already good about our model: its simplicity and creative categorical variables. By applying what we have learned in this project with these concepts, we hope to achieve an even better model that is still balanced.

Works Cited

A&E Television Networks. (2009, November 16). *NBA is born*. History.com. https://www.history.com/this-day-in-history/NBA-is-born.

Bradley, R. (n.d.). The History of NBA Labor. http://apbr.org/labor.html.

Curcic, D. (1970, March 2). *The Ultimate Analysis of NBA Salaries [1991-2019]*. Athletic shoe reviews. https://runrepeat.com/salary-analysis-in-the-NBA-1991-2019.

Gough, C. (2021, May 25). *Annual wages in the NBA 2019/20*. Statista. https://www.statista.com/statistics/1120257/annual-salaries-NBA/.

Webster, I. (n.d.). *Inflation Rate between 1949-2021: Inflation Calculator*. $5,000 in 1949 → 2021 | Inflation Calculator. https://www.in2013dollars.com/us/inflation/1949?amount=5000#:~:text=Value%20of%20%245%2C000%20from%201949,cumulative%20price%20increase%20of%201%2C031.07%25.