Predicting NBA Players' Salaries

Avani Kanungo

Name:	Avani Kanungo
SID:	705308996
Kaggle Name:	Avani Kanungo
Kaggle Rank:	13
Kaggle R^2:	0.60395
Total Number of Predictors:	10
Total Number of Betas:	22
BIC:	12920.18
Complexity Grade:	108

ABSTRACT

The purpose of this project was to create a multiple linear regression modeling an NBA player's salary based on their playing data. Our task was to use a training set of NBA data to "predict the players' salaries" (NBA Project Spring 2021) to determine the strength of the relationship between a player's salary and their playing data. Using this training data, a model was built using 10 predictors and 22 coefficients (Betas) to predict a player's salary.

On the training data, the model had an R^2 score of 0.7099. The model was evaluated on the test data through Kaggle, an online statistics competition platform. The Kaggle R^2 score was 0.60395, ranking 13th out of 52 models.

INTRODUCTION

A set of training data was given to us to build our model. The training data contained 68 predictors and 420 observations of NBA players' data (NBA Project Spring 2021) to use to predict a player's salary. Using these predictors, a model was built and submitted to Kaggle to predict the Salary of players based on the test data set, which contained 180 observations (NBA Project Spring 2021). Below is the list of predictors provided in the training data set:

Types of Predictors

Variable	Type	Variable	Type	Variable	Type
NBA_Country	Categorical	BPM	Numerical	BLK	Numerical

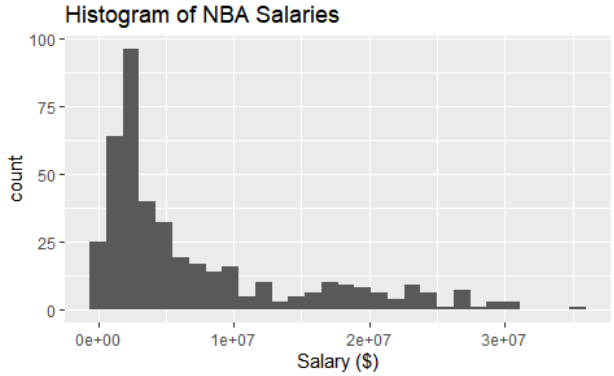
Age	Numerical	VORP	Numerical	TOV	Numerical
TM	Categorical	Rk	Numerical	PF	Numerical
G	Numerical	Pos	Categorical	PTS	Numerical
MP	Numerical	GS	Numerical	Ortg	Numerical
PER	Numerical	FG	Numerical	DRtg	Numerical
TS.	Numerical	FGA	Numerical	Team.Rk	Numerical
X3PAr	Numerical	FG.	Numerical	Team	Categorical
FTr	Numerical	X3P	Numerical	T.Conf	Categorical
ORB.	Numerical	X3PA	Numerical	T.Div	Categorical
DRB.	Numerical	X3P.	Numerical	T.W	Numerical
TRB.	Numerical	X2P	Numerical	T.L	Numerical
AST.	Numerical	X2PA	Numerical	T.W.L.PERC	Numerical
STL.	Numerical	X2P.	Numerical	T.MOV	Numerical
BLK.	Numerical	FT	Numerical	T.Ortg	Numerical
TOV.	Numerical	FTA	Numerical	T.DRtg	Numerical
USG.	Numerical	FT.	Numerical	NRtg	Numerical
OWS	Numerical	ORB	Numerical	MOV.A	Numerical
DWS	Numerical	DRB	Numerical	Ortg.A	Numerical
WS	Numerical	TRB	Numerical	DRtg.A	Numerical
WS.48	Numerical	AST	Numerical	NRtg.A	Numerical
OBPM	Numerical	STL	Numerical	Salary	Numerical

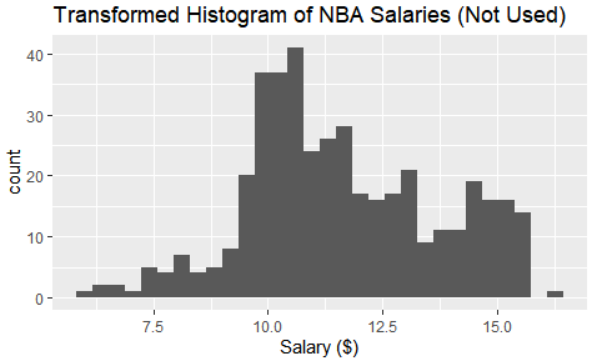
METHODOLOGY

The Response Variable: Salary

First, the normality of the variable Salary was evaluated. As seen in the histogram below, the values of Salary are right-skewed, suggesting that the data should be transformed. Using the R

function powerTransform, a lambda value of 0.16 was suggested to transform the data. However, upon exploring the data further, transforming Salary consistently created worse models that had much lower R^2 values than when Salary was left as is. Additionally, the outliers in the data were a point of interest in the data, as the purpose was to evaluate how correlated a player's data and their salary are. Therefore, variable Salary was left untransformed while building models.

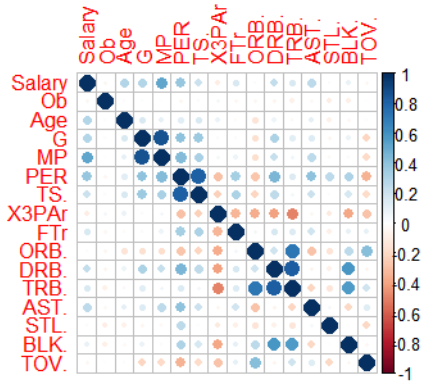


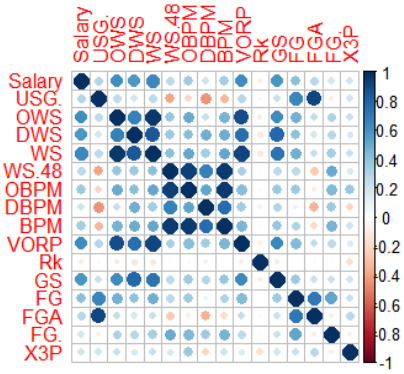


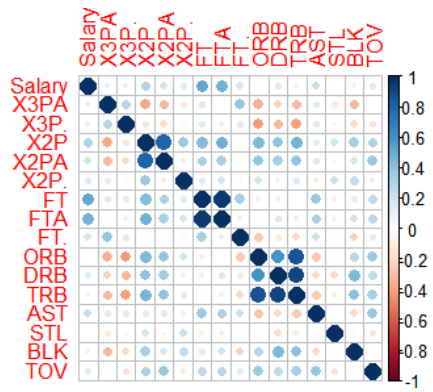
The Numerical Predictors

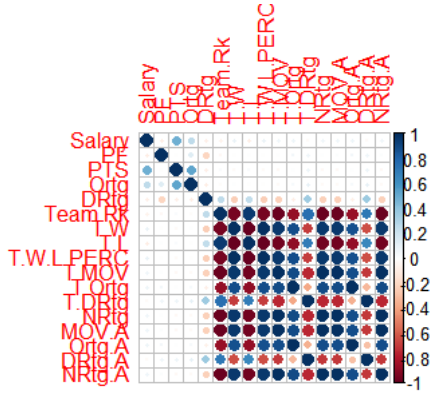
Next, correlation plots of all the numerical variables and Salary were made to determine which variables would help the most in creating the linear model. Since there were so many numerical variables, the data set was split into groups of 15/16 variables + Salary to make the correlation plots easier to read.

Correlation Plots:



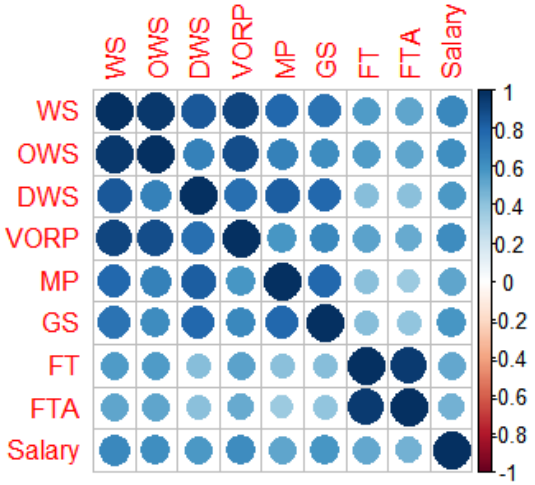






Based on these correlation plots, the strongest predictors for Salary appear to be WS (win shares), VORP (value over replacement player), GS (games started), OWS (offensive win shares), DWS (defensive win shares), MP (minutes played), FT (free throws) and FTA (free throws attempted). Taking a look at the correlation plot of these stronger variables (below), multicollinearity was a major concern as all 8 of these variables are correlated with each other to some degree.

Reduced Correlation Plot:



The concerns about multicollinearity were validated by looking at the output of a linear model created with only these variables and the VIF of that model.

Using only these highly significant variables, an initial model was built.

Summary of Initial Model:

```
Call:
lm(formula = Salary ~ WS + OWS + DWS + VORP + MP + GS + FT + 
    FTA, data = NBAnumeric)

Residuals:
      Min        1Q    Median        3Q       Max 
-13940489  -3126451   -875407   2123162  20214520 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1974714.7   638526.1   3.093 0.002119 ** 
WS          -6819391.2  5605089.6  -1.217 0.224438    
OWS          7250743.0  5612743.7   1.292 0.197140    
DWS          7464521.5  5632471.1   1.325 0.185819    
VORP         1364728.9   669134.6   2.040 0.042035 *  
MP              -296.5      836.4  -0.355 0.723144    
GS             68509.4    17704.5   3.870 0.000127 ***
FT           1414505.3   468330.0   3.020 0.002683 ** 
FTA          -495743.0   369320.8  -1.342 0.180236    
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5475000 on 411 degrees of freedom
Multiple R-squared:  0.4902,    Adjusted R-squared:  0.4803 
F-statistic: 49.41 on 8 and 411 DF,  p-value: < 2.2e-16
```

VIF of Initial Model:

```
         WS          OWS          DWS         VORP           MP           GS           FT          FTA 
3087.971600 1612.264019  404.145924    8.998425    6.046597    3.203990   11.958108   11.288947 
```

Only MP, GS, and FT are considered significant in that model, and only GS has a VIF of less than 5, suggesting that these variables cannot be used as they are in a model. Following up on these diagnostics, the model was reduced to just MP, GS, and FT.

Summary of Reduced Initial Model:

```
Call:
lm(formula = Salary ~ MP + GS + FT, data = NBAnumeric)

Residuals:
      Min        1Q    Median        3Q       Max 
-14447492  -3356987   -507706   2971230  18834456 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  327003.8   587818.8   0.556   0.5783    
MP             1018.3      583.8   1.744   0.0819 .  
GS           101928.5    17097.7   5.962 5.34e-09 ***
FT          1232981.8   158470.9   7.780 5.76e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5739000 on 416 degrees of freedom
Multiple R-squared:  0.4331,    Adjusted R-squared:  0.429 
F-statistic: 105.9 on 3 and 416 DF,  p-value: < 2.2e-16
```

VIF of Reduced Initial Model:

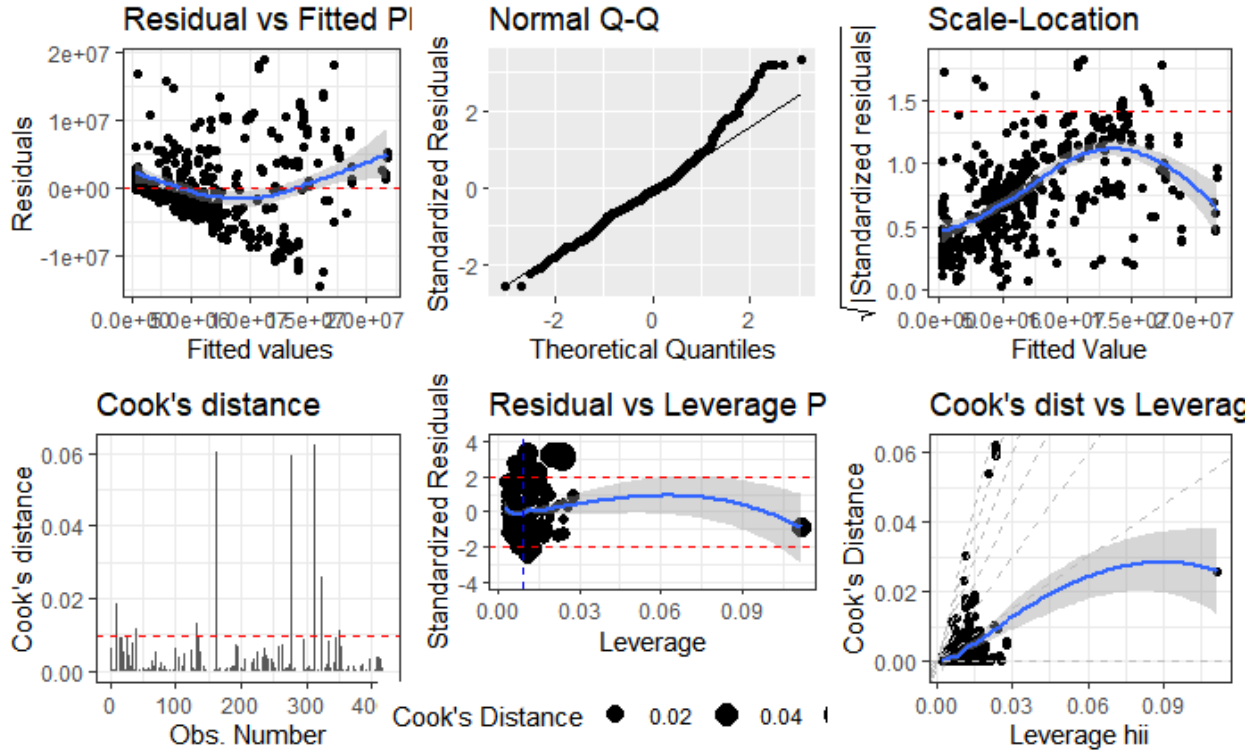
```
       MP        GS        FT 
2.681447 2.719559 1.246100 
```

The reduced model does have a lower R^2 score, but multicollinearity is removed from the model, so it is a considerable improvement on the initial model.

Using Professor Almhalwas's diagnostic plots code from Chapter 5(Chapter-5 Winter-2020), the following diagnostic plots were created. There is some concern looking at the Scale-Location

plot, but otherwise, the reduced initial model doesn't have any major violations and is a good place to start building the model.

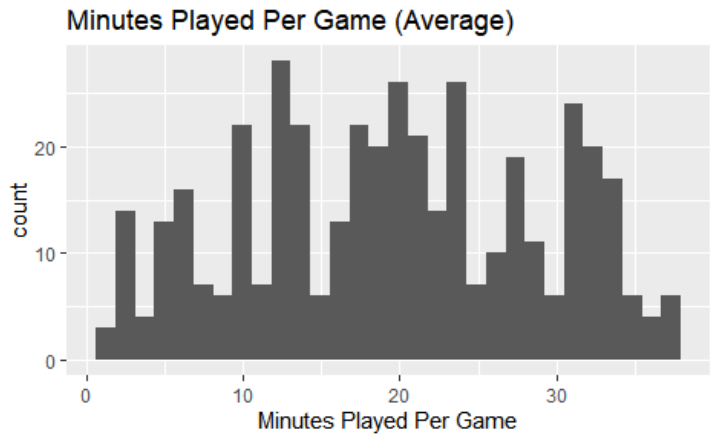
Diagnostic Plots:

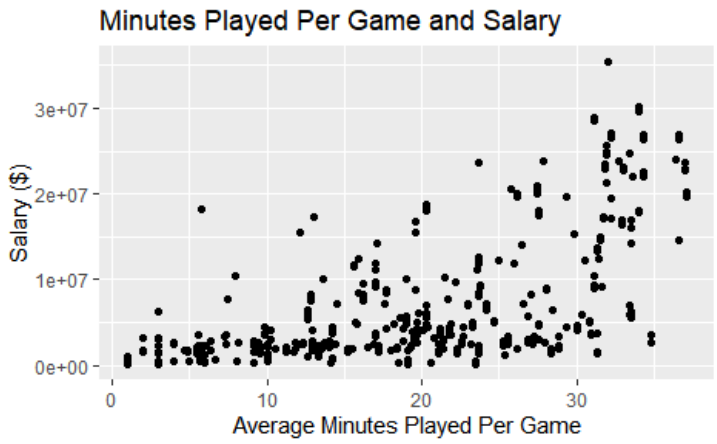


Creating Numerical Variables

MP and G to MPG

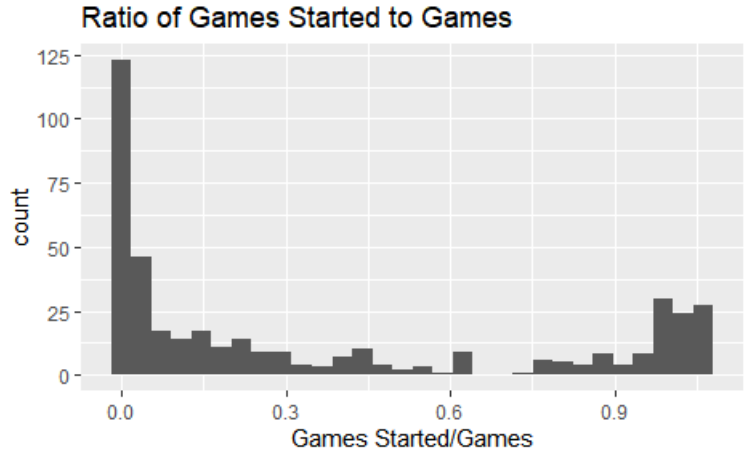
Using the variables MP and G, I created the variable MPG, which is the average number of minutes the player plays in a single game. The histogram looks approximately normal, so I didn't transform the variable. The resulting variable is decently correlated with Salary, with an R^2 score of 0.3793 in a simple linear regression with Salary.

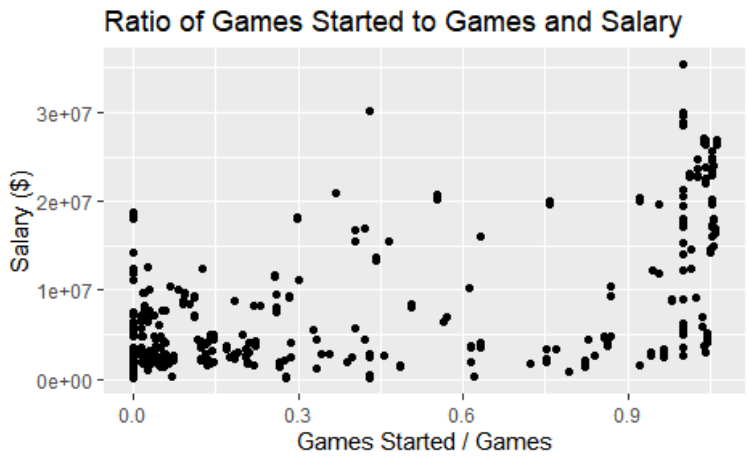




G (Games Played) and GS to GGS

Using the variable G and GS, I used the ratio of the two GS/G to create the new variable GGS. The histogram of this ratio definitely wasn't normal, but this variable could be used in interactions to see if players who started more games or finished more games had differences in salaries. The resulting variable had a vague relationship with Salary, but the R^2 value was 0.3867, so it still was a variable to consider for my model.





The Categorical Variables

In addition to the many numerical variables provided, there were six categorical variables in the data set to utilize. Since several of the variables have too many categories to show the correlation in a diagram, the variables' correlation with Salary was calculated using simple linear regression models between each variable and Salary.

Categorical Variables:

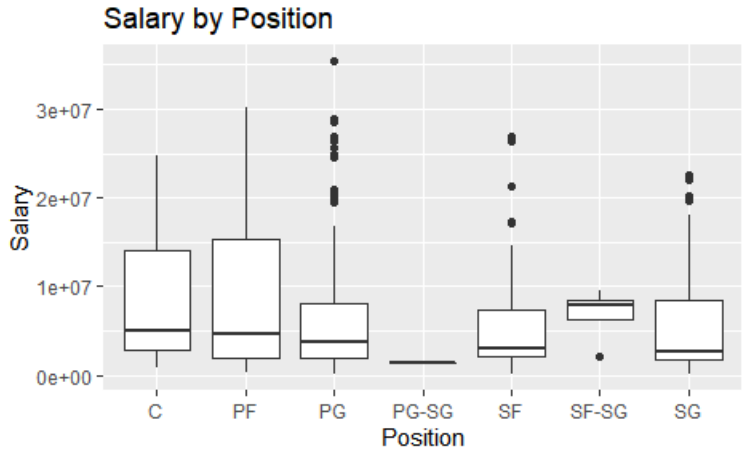
Variable	Number of Categories	Correlation with Salary (R^2)
NBA_Country	28	0.09658
TM	26	0.09874
Pos	7	0.03413
Team	26	0.0836
T.Conf	2	0.003221
T.Div	6	0.01732

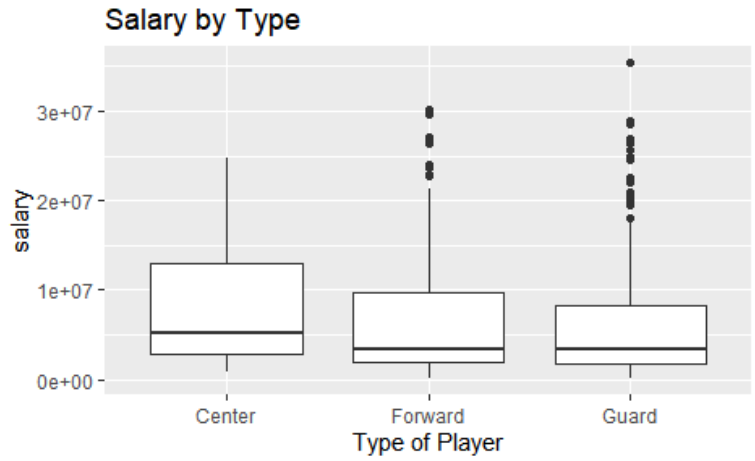
Although the R^2 values for all of these categorical variables were very low, I was intrigued by using interactions between categorical and numerical variables, so I created simplified categorical variables from the existing categorical variables.

Pos to Type

Variable Pos describes the position of the player and has 7 categories: C (center), PF (power forward), PG (point guard), PG-SG (point guard - shooting guard), SF (small forward), SF-SG (small forward - shooting guard), SG (shooting guard). PG-SG only had one observation, and SF-SG only had 4, so I decided to consolidate Pos into "Type" of player: Guard, Forward, or Center. The consolidation

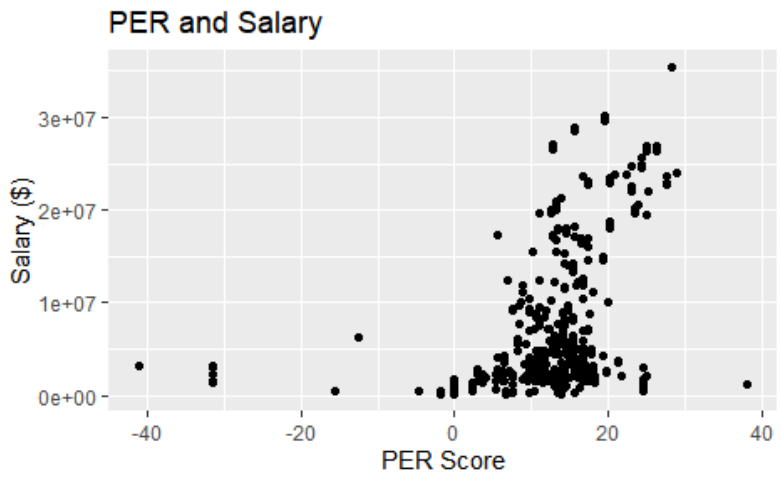
of the variable reduces the R^2 score from 0.03413 to 0.012 in simple linear regressions against Salary, but I believe the reduction was necessary.

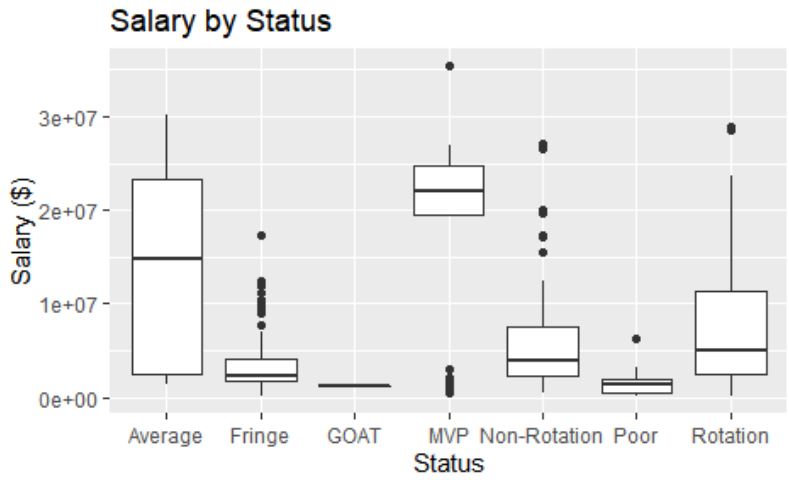




PER to Status

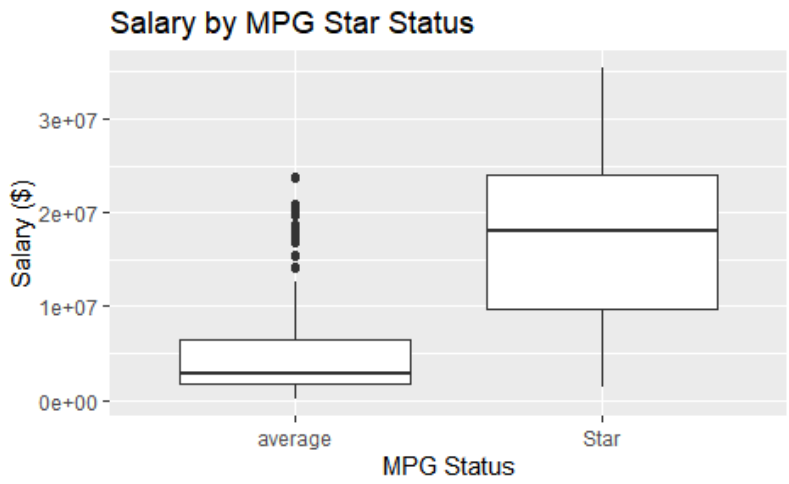
While learning about the different statistics in the data set, I discovered that PER is an advanced statistic that attempts to measure a player's performance in one statistic (Wikimedia Foundation, 2021). The Wikipedia article also included a reference guide, evaluating a player based on what range their PER score falls into. The guide contains 13 categories, but I simplified it to 7 categories as to not overcomplicate the new variable. The 7 categories group the article's evaluations by similar descriptors for adjacent groups, ex. Combining "Weak MPV Candidate," "Strong MVP Candidate," and "Runway MPV Candidate" to simply "MVP." The 7 resulting categories are Poor (PER ≤ 0), Fringe (PER $\leq$ 10), Non-Rotation (PER ≤ 13), Rotation (PER ≤ 18), Average (PER ≤ 22.5), MVP (PER ≤ 35), and GOAT (PER > 35). I found that these categories were better predictors for Salary than PER as a numerical variable. Changing PER from a numerical variable to a categorical variable improved the R^2 value from 0.1482 to 0.2783 in a simple linear regression against Salary, greatly helping my model.





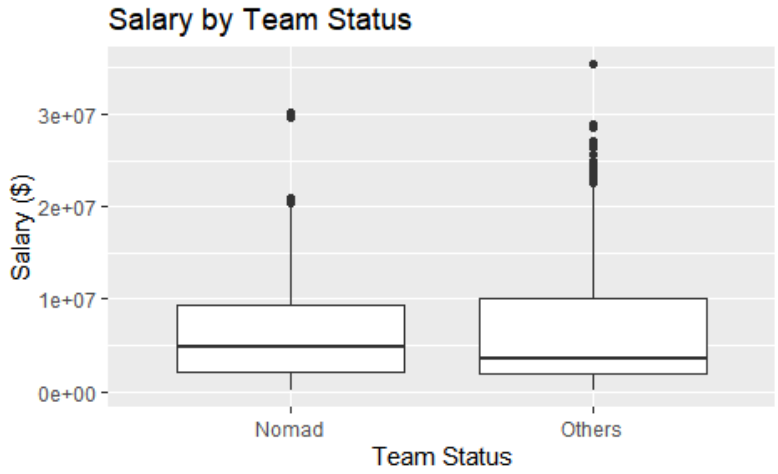
MPG to MPG2

Using the variable I created, MPG, which measures a player's average minutes played per game, I created the variable MPG2, splitting the players into two categories based on their average MPG. For the players who averaged over 30 minutes played per game, I put them in the "Star" category, as they are strong enough players that they are kept on the court for most of the game. All other players were put in the "Other" category. My rationale was that players who play for longer are more skilled and contribute more to the game, and therefore get paid more. As seen in the histogram below, there is a clear difference in salary between the "Star" and "Average" groups. Additionally, in a simple linear regression against Salary, MPG2 has an R^2 value of 0.4129, the strongest relationship with Salary of any of the categorical values, and an improvement of an R^2 value of 0.3793 with MPG.



TM (Team) to TM.Status

In the variable TM, "TOT" means that that player was traded during the season, and therefore played for multiple teams (NBA Project Spring 2021). I created the variable TM.Status to separate the traded players from the players who stayed with one team the whole season, calling them "Nomad" and everyone else "Other." My rationale was that if a player was traded, they aren't very valuable to the team and therefore would be paid less. This variable ended up not being the best predictor, as seen in the boxplot below - there wasn't much of a difference in Salary between those who were traded and those who stayed with one team.

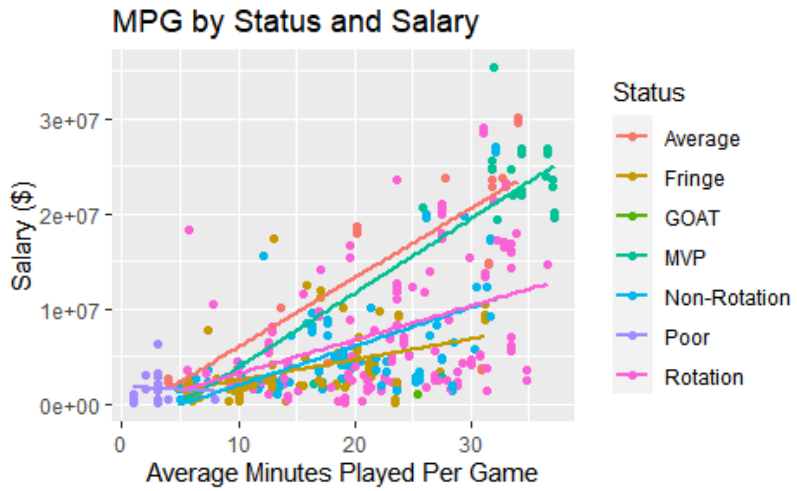


Interactions

With the categorical variables, I was very interested in using them to make seemingly less correlated variables useful to my model using interactions. For example, some of the player statistics like FT would be varied by type of player, as guards don't usually score. I investigated the interactions between many variables over the three weeks spent on this project, and the ones I'm showcasing in this paper are some of the ones that ended up being the most useful in my models.

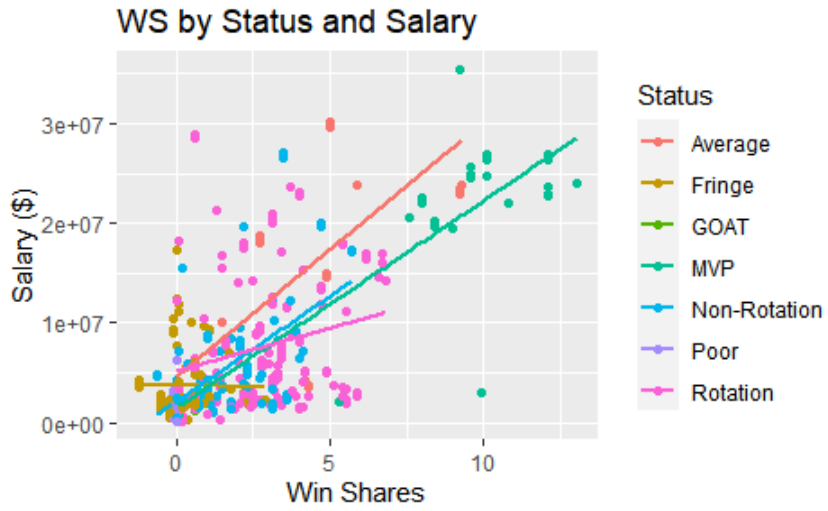
MPG and Status

In making my model, I evaluated the interaction between my strongest numerical and categorical variables - MPG and Status. There definitely were different regression lines for different Status groups, so I deemed this a useful interaction for my model. The linear regression of just this interaction had an R^2 value of 0.5457.



WS and Status

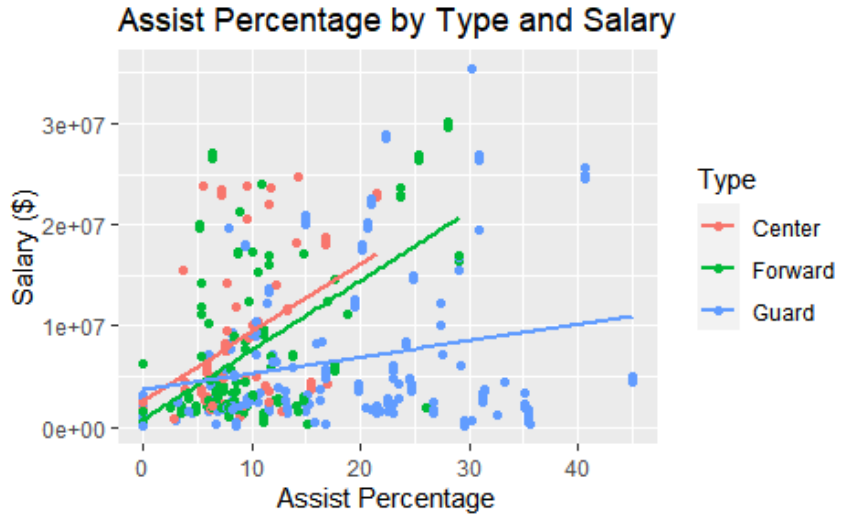
I looked at the interaction between WS and Status, another pair of variables with higher correlations to Salary. There definitely was a difference in slopes between the regression lines for Status groups, so I deemed this a useful interaction for my model. This interaction had an R^2 value of 0.4674, which wasn't as strong as MPG:Status, but still respectable.



Type and AST. (Assist Percentage)

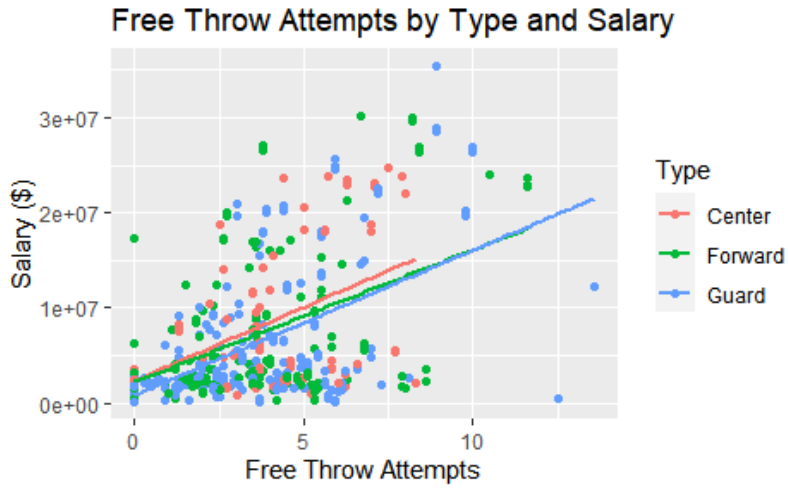
One of the main reasons I created the Type variable was to see if the interaction between Type and some distinct responsibilities of each Type of player would be helpful to my model. Assist statistics are considered important in the value of a player, so I investigated the interaction between Type and

AST. There are distinct slopes for Salary for each Type of player, making this a useful interaction. The R^2 value for this interaction was 0.1892, which wasn't much but definitely was an improvement on the correlation of either variable alone with Salary.



Type and FTA

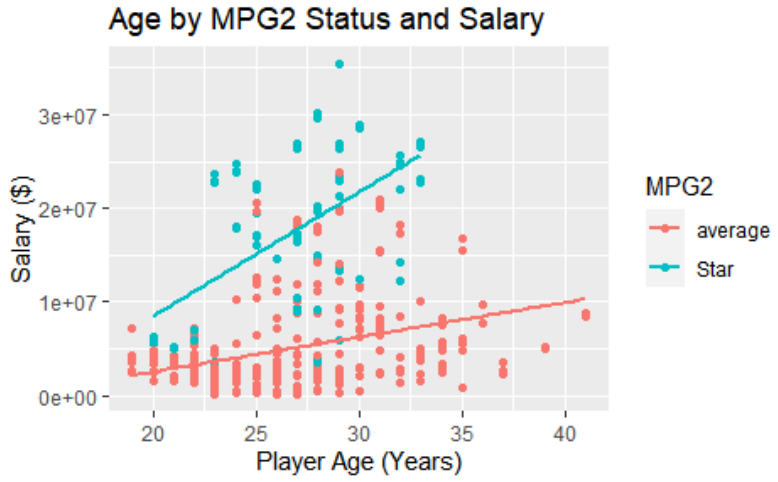
Another variable that I thought would have an interesting interaction with Type was FTA. My rationale was that if a player is attempting more free throws, then they are considered a reliable scorer and perhaps would be paid more. There isn't much difference in the slope between the Types, but this interaction had an R^2 value of 0.2329, which kept it in my list of interactions to consider for the model.



Age and MGP2

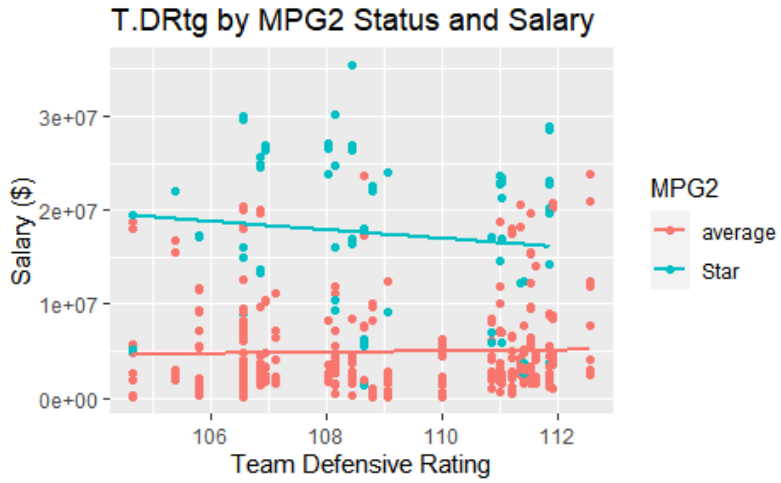
A variable that I found to be a surprisingly good predictor was Age. Alone, it only had an R^2 value of 0.08776, but I decided to explore interactions between Age and my categorical variables. I was very surprised to see how much the interaction between Age and MPG2 status mattered - there was a major difference in slope between the two groups, suggesting that Stars have their salary increase

more per year than Average players. This interaction had an R^2 value of 0.5159, which definitely made it a key interaction in my models.



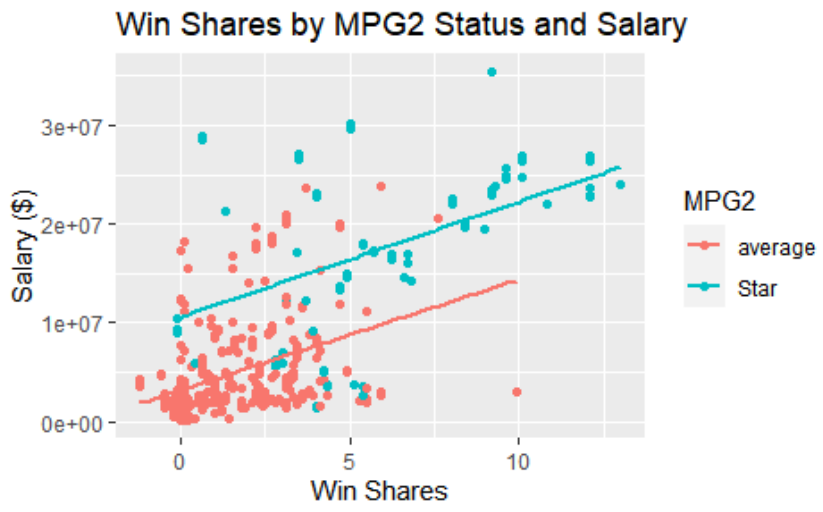
MPG2 and T.DRtg (Team Defensive Rating)

Most of the variables I had been investigating were individual statistics, so I cycled through interactions with team statistics to see if any had stronger R^2 scores. There were separate slopes for the Star and Average groups, showing the Salary of Average players slightly increasing as their Team Defensive Rating increased, while the Stars Salaries decreased. MPG2 and T.DRtg had an R^2 score of 0.4114, which was surprising and made it an important interaction to consider because it was less likely to have multicollinearity issues with individual statistics in my model.



WS and MPG2

Another interaction that I thought might be important was between WS and MPG2. Stars have more time on the court, and I thought that having a higher WS would make them more valuable players, and thus paid more. There were separate slopes for the two groups, suggesting that Stars with higher WS values get paid more than Average players with the same WS value. This interaction had an R^2 value of 0.4688, making it valuable for my models.



Building Models

Over the course of three weeks, I built over 135 models, using various different sets of variables and interactions. In this paper, I explained only my best model (lower BIC score) and my runner-up model, which had the highest R^2 score and a different approach than my best model. I utilized the variables that were highly correlated with Salary, the numerical and categorical variables I made, and many interactions.

Runner-Up Model (Best R^2)

MODEL:
Salary~MPG2:Age+MPG:Status+WS:MPG:Status+Type:AST.+Type:FTA+MPG2:T.DRtg+MPG2:WS

This model was mostly built with the interactions I showed earlier in this paper, as they worked well together. The final model has 27 coefficients, using 10 predictors. It had a training R^2 score of 0.7119 and a test R^2 score of 0.60472. At least one category was significant for each of the categorical variables.

Since there were many interactions in this model, I couldn't run VIF on the model itself, but I ran the function on a model that only had the predictors and no interactions. Since Status has 7 categories, GVIF was used to evaluate multicollinearity instead of VIF, and since none of the variables had a GVIF value over 5, there were no multicollinearity issues. Additionally, this model had a BIC score of 12941.51.

I used Professor Almohalwas's code (Chapter-5 Winter-2020) to create diagnostic plots. There were some concerns with the diagnostic plots, as the Residual vs Fitted Plot and Scale-Location Plot had concerning shapes. Once again, all attempts to normalize the data worsened the model, and I considered extreme values points of interest for many variables, so I accepted the violations and this temporarily became my best model. This model is my "final" Kaggle model, but

only because of the slightly higher R^2 value. I consider the next model to be my best model due to the lower BIC score.

Summary of Final Model:

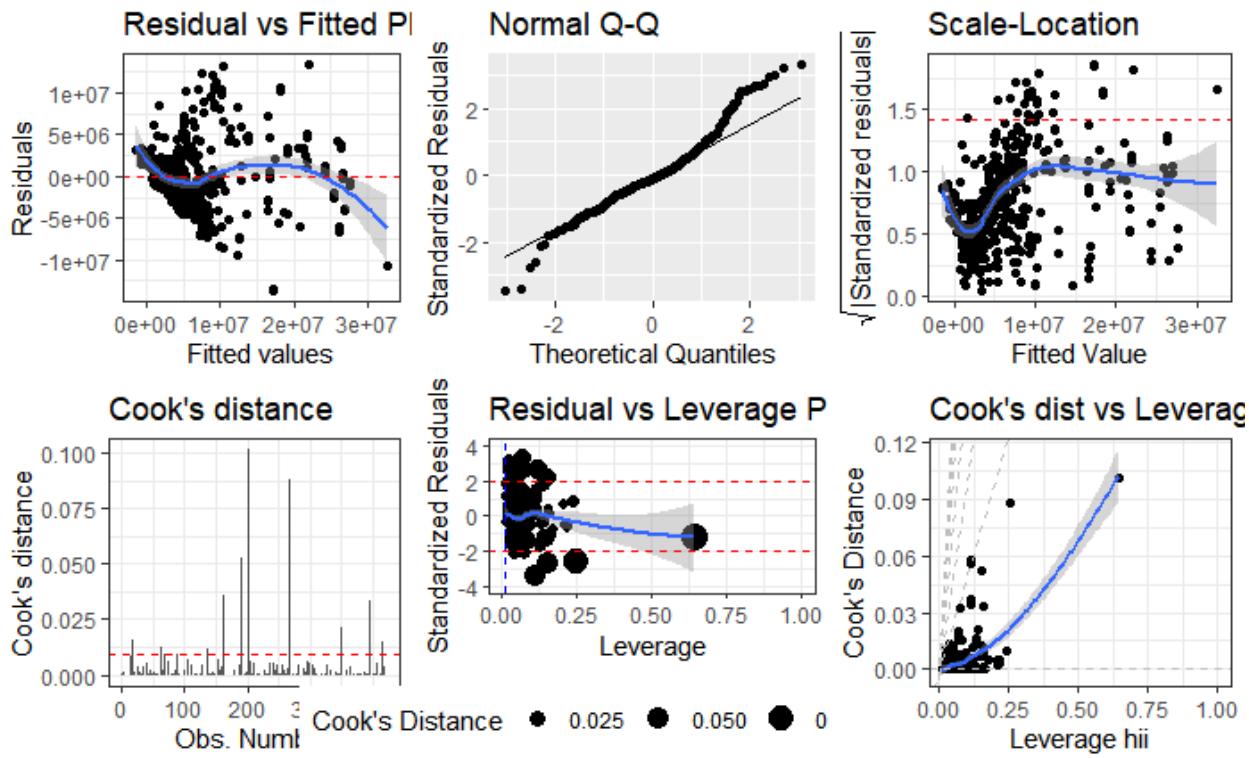
```
Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                9351917   10480333   0.892 0.372761
MPG2average:Age             324720      52740   6.157 1.83e-09 ***
MPG2Star:Age               1354604     147349   9.193  < 2e-16 ***
MPG:StatusAverage           378393     115371   3.280 0.001131 **
MPG:StatusFringe            189911      61224   3.102 0.002061 **
MPG:StatusGOAT               79102     173524   0.456 0.648747
MPG:StatusMVP               452583     175180   2.584 0.010139 *
MPG:StatusNon-Rotation      184199      63934   2.881 0.004180 **
MPG:StatusPoor             -183206     449322  -0.408 0.683687
MPG:StatusRotation          341303      61346   5.564 4.89e-08 ***
TypeCenter:AST.             311287     102124   3.048 0.002458 **
TypeForward:AST.            157788      65095   2.424 0.015801 *
TypeGuard:AST.              -108412      32996  -3.286 0.001109 **
TypeCenter:FTA                3284     234205   0.014 0.988820
TypeForward:FTA            -111464     174997  -0.637 0.524530
TypeGuard:FTA               419605     141921   2.957 0.003298 **
MPG2average:T.DRtg          -154861      96078  -1.612 0.107800
MPG2Star:T.DRtg             -387382     106053  -3.653 0.000295 ***
MPG2average:WS             -2198267     607702  -3.617 0.000336 ***
MPG2Star:WS                -2235831     833356  -2.683 0.007605 **
MPG:StatusAverage:WS          89204      29204   3.055 0.002407 **
MPG:StatusFringe:WS           45356      40222   1.128 0.260159
MPG:StatusGOAT:WS                NA         NA      NA       NA
MPG:StatusMVP:WS              71826      26192   2.742 0.006379 **
MPG:StatusNon-Rotation:WS     97748      27396   3.568 0.000404 ***
MPG:StatusPoor:WS          -5526810    6239311  -0.886 0.376263
MPG:StatusRotation:WS         54716      24494   2.234 0.026055 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4204000 on 394 degrees of freedom
Multiple R-squared:  0.7119,    Adjusted R-squared:  0.6936
F-statistic: 38.94 on 25 and 394 DF,  p-value: < 2.2e-16
```

VIF of Final Model:

```
           GVIF Df GVIF^(1/(2*Df))
Status 6.550903  6        1.169567
MPG    5.021666  1        2.240907
Age    1.063730  1        1.031373
MPG2   2.491539  1        1.578461
Type   2.011952  2        1.190980
AST.   2.101464  1        1.449643
FTA    1.687088  1        1.298879
T.DRtg 1.095376  1        1.046602
WS     5.037638  1        2.244468
```

Diagnostic Plots:



Best Final Model (Lower BIC)

MODEL:
Salary~Status:MPG+Age:MPG+MPG2+Age:MPG2+MPG:GGS+Type:AST.+Type:FTA+TM.Status:MPG2
+WS:MPG2

This model was made after my runner-up model, and I was trying a different approach to see if a similar but different set of variables and interactions would create a stronger model. This model has 22 coefficients, and had a training R^2 score of 0.7099 and a test R^2 score of 0.60395, which isn't much less than my runner-up model had with five fewer betas. At least one category was significant for each of the categorical variables.

Since there were many interactions in this model, I couldn't run VIF on the model itself, but I ran the function on a model that only had the predictors and no interactions. Since Status has 7 categories, GVIF was used to evaluate multicollinearity instead of VIF, and since none of the variables had a GVIF value over 5, there were no multicollinearity issues. Additionally, this model had a BIC score of 12920.18, less than my runner-up model's BIC score of 12941.51.

I used Professor Almohalwas's code (Chapter-5 Winter-2020) to create diagnostic plots. There were some concerns with the diagnostic plots. The Cook's Distance vs Leverage plot was a parabola, and the Residual vs Fitted Plot and Scale-Location Plot also had concerning shapes. However, all attempts to normalize the data worsened the model, and I considered extreme values points of interest for many variables, so I accepted the violations and made this model my final model due to the smaller BIC score and fewer coefficients, ignoring the slight loss in R^2.

Summary of Best Model:

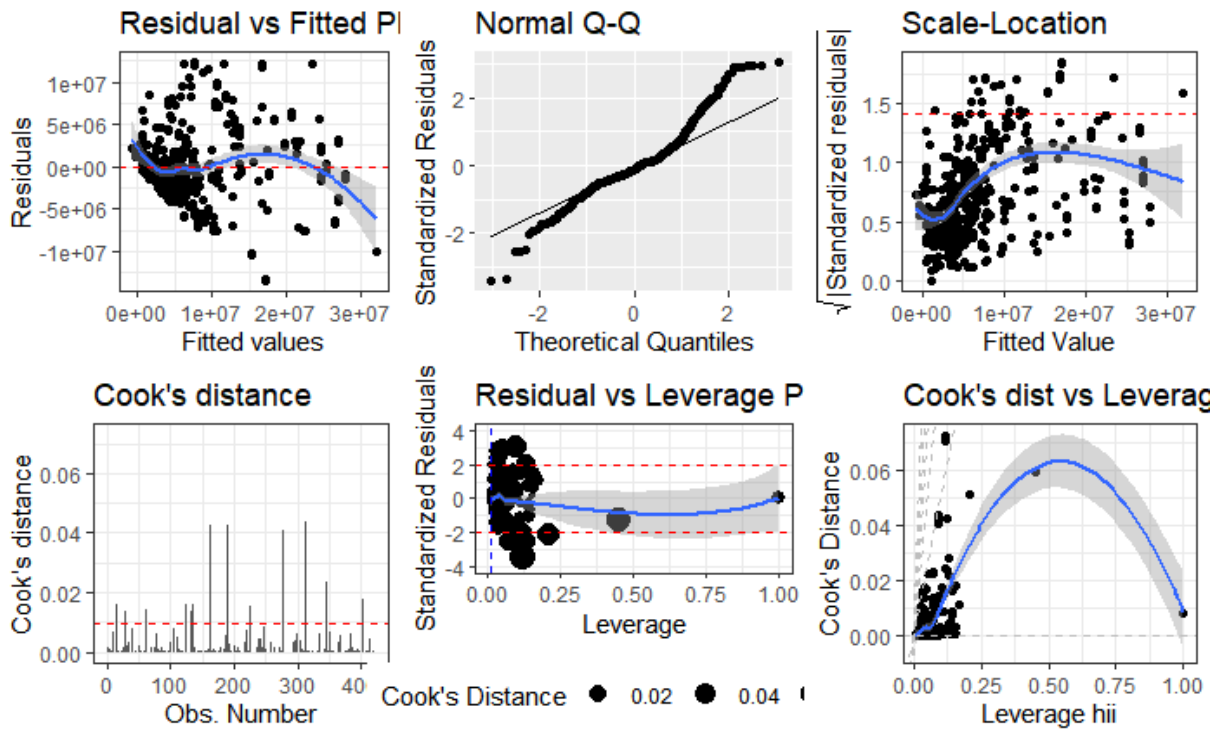
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1390054     906245   1.534 0.125859
MPG2Star                     -17069159    5053990  -3.377 0.000804 ***
StatusAverage:MPG               -89067     107387  -0.829 0.407374
StatusFringe:MPG               -356297      94161  -3.784 0.000178 ***
StatusGOAT:MPG                 -523323     192612  -2.717 0.006876 **
StatusMVP:MPG                   -63313     117086  -0.541 0.588994
StatusNon-Rotation:MPG         -291696      93937  -3.105 0.002037 **
StatusPoor:MPG                 -437492     382440  -1.144 0.253333
StatusRotation:MPG             -262171      96916  -2.705 0.007121 **
MPG:Age                         19140       2956   6.474 2.81e-10 ***
Age:MPG2Star                   664705     172534   3.853 0.000136 ***
MPG:GGS                        130912      37269   3.513 0.000495 ***
TypeCenter:AST.                227665      94772   2.402 0.016753 *
TypeForward:AST.               126056      63454   1.987 0.047655 *
TypeGuard:AST.                 -95995      32494  -2.954 0.003321 **
TypeCenter:FTA                 111417     222196   0.501 0.616341
TypeForward:FTA                -24189     168104  -0.144 0.885657
TypeGuard:FTA                  363091     140200   2.590 0.009955 **
MPG2average:TM.StatusOthers   -865083     538644  -1.606 0.109059
MPG2Star:TM.StatusOthers     -2460345    1504192  -1.636 0.102701
MPG2average:WS                -612586     247537  -2.475 0.013750 *
MPG2Star:WS                    210074     254740   0.825 0.410059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4197000 on 398 degrees of freedom
Multiple R-squared:  0.7099,    Adjusted R-squared:  0.6946
F-statistic: 46.39 on 21 and 398 DF,  p-value: < 2.2e-16
```

VIF of Best Model:

```
              GVIF Df GVIF^(1/(2*Df))
Status    7.109649  6        1.177572
MPG       6.717433  1        2.591801
Age       1.103447  1        1.050451
MPG2      2.621385  1        1.619069
GGS       3.608288  1        1.899549
Type      2.100360  2        1.203853
AST.      2.082609  1        1.443125
FTA       1.676353  1        1.294740
TM.Status 1.270395  1        1.127118
WS        5.034631  1        2.243798
```

Diagnostic Plots of Best Model:



Evaluating BIC Scores:

```
[1] "Runner-Up Model BIC" "26"                "12941.5123737971"
[1] "Best Model BIC: " "22"                 "12920.1775844536"
```

RESULTS AND DISCUSSION

Final Model

$\widehat{Salary}$ =-17069159(MPG2Star) -89067(MPG:StatusAverage) -356297(MPG:StatusFringe) -523323(MPG:StatusGOAT) - 63313(MPG:StatusMVP) -291696(MPG:StatusNon-Rotation) -437492(MPG:StatusPoor) -262171(MPG:StatusRotation) +19140(MPG:Age) +664705(Age:MPG2Star) +130912(MPG:GGS) +227665(AST.:TypeCenter) +126056(AST.:TypeForward) -95995(AST.:TypeGuard) +11417(FTA:TypeCenter) -24189(FTA:TypeForward) +363091(FTA:TypeGuard) -865083(MPG2Average:TM.StatusOthers) -2469345(MPG2Star:TM.StatusOthers) -612586(WS:MPG2Average) +210074(WS:MPG2Star)

Considering the training and test R^2 scores, BIC score, and significance of all variables used in the model, I chose my "Best Model" to be my final model. I didn't split my training data into sub-training and -testing datasets. Almost all of the predictors are significant, at least one from each

interaction with a categorical variable. The R^2 value is 0.7099, a moderately strong correlation between predicted and actual values.

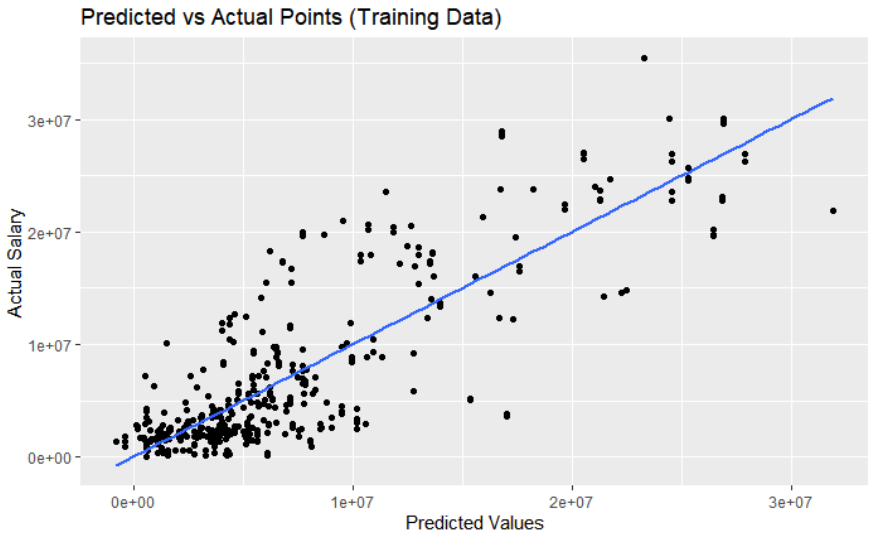
Final Diagnostics

Leverage vs Outliers:

```
       rii
hii     FALSE TRUE
  FALSE   262   16
  TRUE    131   11
```

There are 11 bad leverage points.

Actual vs Predicted Salary On Training Data



R^2 is 0.7099.

Summary of Final Model:

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1390054     906245   1.534 0.125859
MPG2Star                     -17069159    5053990  -3.377 0.000804 ***
StatusAverage:MPG               -89067     107387  -0.829 0.407374
StatusFringe:MPG               -356297      94161  -3.784 0.000178 ***
StatusGOAT:MPG                 -523323     192612  -2.717 0.006876 **
StatusMVP:MPG                   -63313     117086  -0.541 0.588994
StatusNon-Rotation:MPG         -291696      93937  -3.105 0.002037 **
StatusPoor:MPG                 -437492     382440  -1.144 0.253333
StatusRotation:MPG             -262171      96916  -2.705 0.007121 **
MPG:Age                          19140       2956   6.474 2.81e-10 ***
Age:MPG2Star                    664705     172534   3.853 0.000136 ***
MPG:GGS                         130912      37269   3.513 0.000495 ***
TypeCenter:AST.                 227665      94772   2.402 0.016753 *
TypeForward:AST.                126056      63454   1.987 0.047655 *
TypeGuard:AST.                  -95995      32494  -2.954 0.003321 **
TypeCenter:FTA                  111417     222196   0.501 0.616341
TypeForward:FTA                 -24189     168104  -0.144 0.885657
TypeGuard:FTA                   363091     140200   2.590 0.009955 **
MPG2average:TM.StatusOthers    -865083     538644  -1.606 0.109059
MPG2Star:TM.StatusOthers      -2460345    1504192  -1.636 0.102701
MPG2average:WS                 -612586     247537  -2.475 0.013750 *
MPG2Star:WS                     210074     254740   0.825 0.410059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4197000 on 398 degrees of freedom
Multiple R-squared:  0.7099,    Adjusted R-squared:  0.6946
F-statistic: 46.39 on 21 and 398 DF,  p-value: < 2.2e-16
```

LIMITATIONS AND CONCLUSIONS

Overall, the biggest limitation of the method would be overfitting the model to the training data. The R^2 value for the training data was 0.7099, whereas the R^2 for the testing data was 0.60396, a very large drop. I did my best to use variables that did well with both the training and testing data, but I still ended up with a 0.1 drop in R^2 score in my best model.

The accuracy of the model is another limitation - I placed 13th in the competition, and the highest R^2 score was 0.77793, showing that my model had a lot of room for improvement. I tried to improve my R^2 score in 138 models, but my Runner-Up model had the highest R^2 score I was able to achieve (0.60472).

The validity of the model is of concern, as discussed in methodology, looking at the behavior in the Residual vs Fitted Plot and Scale-Location Plot. I was unable to normalize the plots, even with normalizing the individual variables or using different variables entirely. There are also 11 bad leverage points that I could have removed from the data, but since I wanted to keep extreme values, this may have hurt the validity of my model too.

Personally, I am unfamiliar with basketball, so perhaps having more in-depth knowledge of the sport and its statistics would have allowed me to make a stronger, more valid model.

REFERENCES

Almohalwas, Akram. 2020. *Chapter-5 Winter-2020*.

Almohalwas, Akram. 2021. *NBA Project Spring 2021*.
Wikimedia Foundation. (2021, May 26). *Player efficiency rating*. Wikipedia. https://en.wikipedia.org/wiki/Player_efficiency_rating.