Abstract:

A program that determines what variables and statistics are factored in to determine an NFL Quarterback's base salary. This project aspires to give insight to optimal spending on a team's Quarterback and getting value from what they spend. Every Quarterback's personal and team statistics from years 2005 to 2018 will be used to analyze how much value they had to the team and will be compared to their salary to determine their value. Based on the findings, projected value of the current (2019-20) Quarterback's contract can be found as well as the optimal spending for a new NFL team.

Problem and Motivation:

The data set was collected from multiple sources; the official NFL website, Spotrac, and football outsiders. The NFL website provided the Quarterback's statistics, spotrac provided the Quarterback's salary, and football outsiders provided every team's offensive and defensive rankings. The conclusions drawn from this project can provide NFL teams a guideline and strategy to get another quarterback, or an amount to offer their current one.

Data:

1. Player (the player's name)
2. Team (the player's team)
3. Comp (the number of completed passes)
4. Att (the number of attempted passes)
5. Yds (the number of passing yards)
6. TD (the number of touchdowns)
7. Int (the number of intercepted passes)
8. Rate (the quarterback rating)
9. Year (the year of the season they played)
10. Following Year (the following year of the season they played)
11. Following Salary (the player's base salary of the following year)
12. Offensive Rank (the team's offensive ranking for which the player plays for)

Questions of Interest:

Which Quarterback statistics are important to consider when determining their salary for the following year? What type of Quarterbacks are getting overpaid or underpaid?

NFL QB Salary Data Analysis

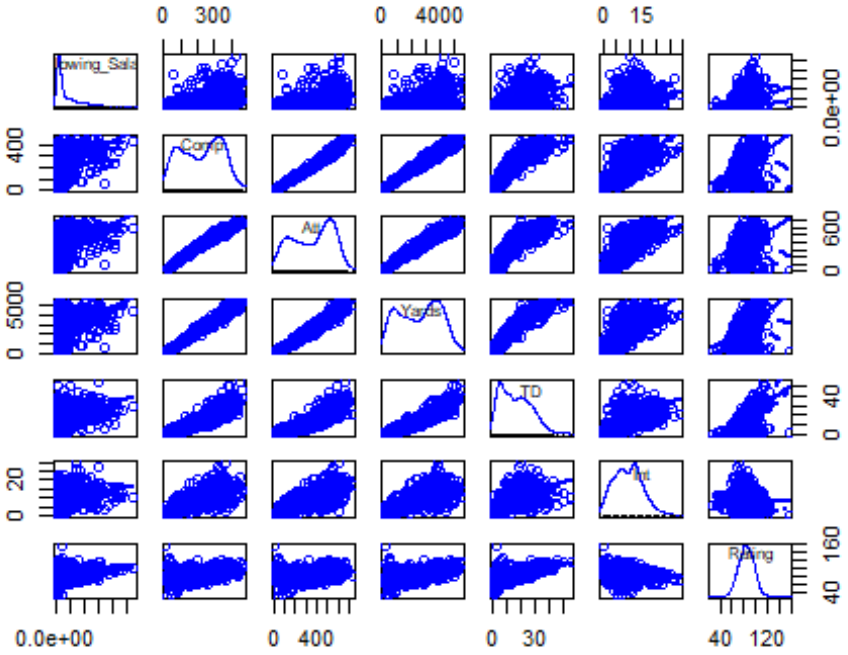
Preliminary Expectations

Before fitting the model, I expect the following relationships to occur within the model:

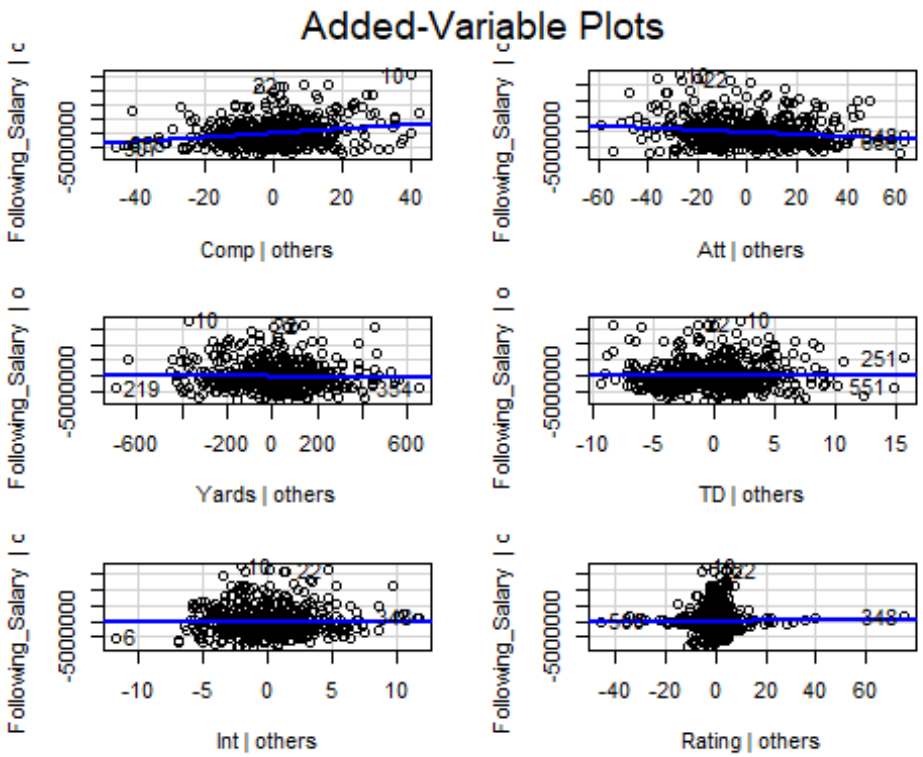
1. A positive linear relationship between the number of completed passes and the player's base salary of the following year since the quarterback's primary role is to complete passes.
2. A positive linear relationship between the number of attempted passes and the player's base salary of the following year due to the volume of which the quarterback was used.
3. A positive linear relationship between the yards, rating, and touchdowns and the player's base salary of the following year due to these variables correlating with the effectiveness of the quarterback.
4. A negative linear relationship between the number of intercepted passes and the player's base salary of the following year due to the negative effects intercepted passes have for a team.

Exploratory Analysis

```
fit <- lm(Following_Salary~Comp+Att+Yards+TD+Int+Rating)
scatterplotMatrix(~Following_Salary+Comp+Att+Yards+TD+Int+Rating)
```




```
avPlots(fit)
```



Observation 10 seems to be an outlier.

```
outlierTest(fit)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 10  4.448561         1.0313e-05    0.0062395
## 22  4.078148         5.1552e-05    0.0311890
## 103 4.063827         5.4730e-05    0.0331120
## 57  4.032229         6.2412e-05    0.0377590
## 39  4.006258         6.9481e-05    0.0420360
```

Based on the calculated P-Values, observation 10, 22, 103, 57 and 39 are significant outliers.

```
summary(fit)
```

```
##
## Call:
## lm(formula = Following_Salary ~ Comp + Att + Yards + TD + Int +
##     Rating)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8115686 -2301632  -644821  1062328 17430849
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -213998.6  1708214.4  -0.125 0.900347
## Comp          73459.4    13265.2   5.538 4.59e-08 ***
## Att          -29649.4     8574.7  -3.458 0.000583 ***
## Yards         -1073.0      931.5  -1.152 0.249821
## TD            15344.3    47055.9   0.326 0.744473
## Int           18900.7    55945.8   0.338 0.735602
## Rating         9957.1    20796.3   0.479 0.632260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4034000 on 598 degrees of freedom
##   (95 observations deleted due to missingness)
## Multiple R-squared:  0.2628, Adjusted R-squared:  0.2554
## F-statistic: 35.53 on 6 and 598 DF,  p-value: < 2.2e-16
```

The fitted linear model for the Following Salary of a quarterback based on the Completions, Attempts, Yards, Touchdowns, Interceptions and Rating is:

Following Salary =-213998.6+73459.4(Completions)-29649.4(Attempts)-1073(Yards)+15344.3(Touchdowns)+18900.7(Interceptions)+9957.1(Rating).

```
comptest <- lm(Following_Salary~Comp)
atttest <- lm(Following_Salary~Att)
yardstest <- lm(Following_Salary~Yards)
tdtest <- lm(Following_Salary~TD)
inttest <- lm(Following_Salary~Int)
ratingtest <- lm(Following_Salary~Rating)
dfComp <- c("Comp SLR Coefficient", (comptest$coefficients[2]))
dfAtt <- c("Att SLR Coefficient", (atttest$coefficients[2]))
dfYards <- c("Yards SLR Coefficient", (yardstest$coefficients[2]))
dfTD <- c("TD SLR Coefficient", (tdtest$coefficients[2]))
dfInt <- c("Int SLR Coefficient", (inttest$coefficients[2]))
dfRating <- c("Rating SLR Coefficient", (ratingtest$coefficients[2]))
df <- data.frame(dfComp, dfAtt, dfYards, dfTD, dfInt, dfRating)
print.data.frame(df, row.names = FALSE)

##                dfComp               dfAtt               dfYards
##  Comp SLR Coefficient Att SLR Coefficient Yards SLR Coefficient
##      18461.0310404323     11336.847173946      1532.28661132708
##                dfTD               dfInt               dfRating
```

```
##  TD SLR Coefficient Int SLR Coefficient Rating SLR Coefficient
##    194442.617550415      237892.1215764       100040.975627026
```

Completions has a positive relationship with Following Salary, Attempts has a positive relationship with Following Salary, Yards has a positive relationship with Following Salary, Touchdowns has a positive relationship with Following Salary. Surprisingly, Interceptions has a positive relationship with Following Salary and Rating has a positive relationship with Following Salary.

The positive relationship between Interceptions and Following Salary is probably due to the fact that Quarterback's get paid more for how significant they are to the team, and a quarterback with more attempted passes tend to throw more interceptions.

Tests for Significance of Predictors in Model:

Completions: Null Hypothesis $H_0: \beta_1 = 0$, Alternative Hypothesis $H_A: \beta_1 \neq 0$.

```
#T-Value
T=fit$coefficients[2]/13265.2
#Critical Value
CV=qt(0.995, 601)
#Test
abs(T)>CV
```

```
## Comp
## TRUE
```

Since the absolute value of the T value is greater than the critical value at level $\alpha = 0.01$, the null hypothesis is rejected and it is determined that $\beta_1 \neq 0$. Therefore Completions is a significant predictor in the model.

Attempts: Null Hypothesis $H_0: \beta_2 = 0$, Alternative Hypothesis $H_A: \beta_2 \neq 0$.

```
#T-Value
T=fit$coefficients[3]/8574.7
#Critical Value
CV=qt(0.995, 601)
#Test
abs(T)>CV
```

```
##  Att
## TRUE
```

Since the absolute value of the T value is greater than the critical value at level $\alpha = 0.01$, the null hypothesis is rejected and it is determined that $\beta_2 \neq 0$. Therefore Attempts is a significant predictor in the model.

Yards: Null Hypothesis $H_0: \beta_3 = 0$, Alternative Hypothesis $H_A: \beta_3 \neq 0$.

```
#T-Value
T=fit$coefficients[4]/931.5
#Critical Value
CV=qt(0.995,601)
#Test
abs(T)>CV
```

```
## Yards
## FALSE
```

Since the absolute value of the T value is less than the critical value at level $\alpha = 0.01$, the null hypothesis is not rejected and it is determined that $\beta_3 = 0$. Therefore Yards is not a significant predictor in the model.

Touchdowns: Null Hypothesis $H_0: \beta_4 = 0$, Alternative Hypothesis $H_A: \beta_4 \neq 0$.

```
#T-Value
T=fit$coefficients[5]/47055.9
#Critical Value
CV=qt(0.995,601)
#Test
abs(T)>CV
```

```
##    TD
## FALSE
```

Since the absolute value of the T value is less than the critical value at level $\alpha = 0.01$, the null hypothesis is not rejected and it is determined that $\beta_4 = 0$. Therefore Touchdowns is not a significant predictor in the model.

Interceptions: Null Hypothesis $H_0: \beta_5 = 0$, Alternative Hypothesis $H_A: \beta_5 \neq 0$.

```
#T-Value
T=fit$coefficients[6]/55945.8
#Critical Value
CV=qt(0.995,601)
#Test
abs(T)>CV
```

```
##   Int
## FALSE
```

Since the absolute value of the T value is less than the critical value at level $\alpha = 0.01$, the null hypothesis is not rejected and it is determined that $\beta_5 = 0$. Therefore Interceptions is not a significant predictor in the model.

Rating: Null Hypothesis $H_0: \beta_6 = 0$, Alternative Hypothesis $H_A: \beta_6 \neq 0$.

```
#T-Value
T=fit$coefficients[7]/20796.3
#Critical Value
CV=qt(0.995,601)
#Test
abs(T)>CV

## Rating
##  FALSE
```

Since the absolute value of the T value is less than the critical value at level $\alpha = 0.01$, the null hypothesis is not rejected and it is determined that $\beta_6 = 0$. Therefore Rating is not a significant predictor in the model.

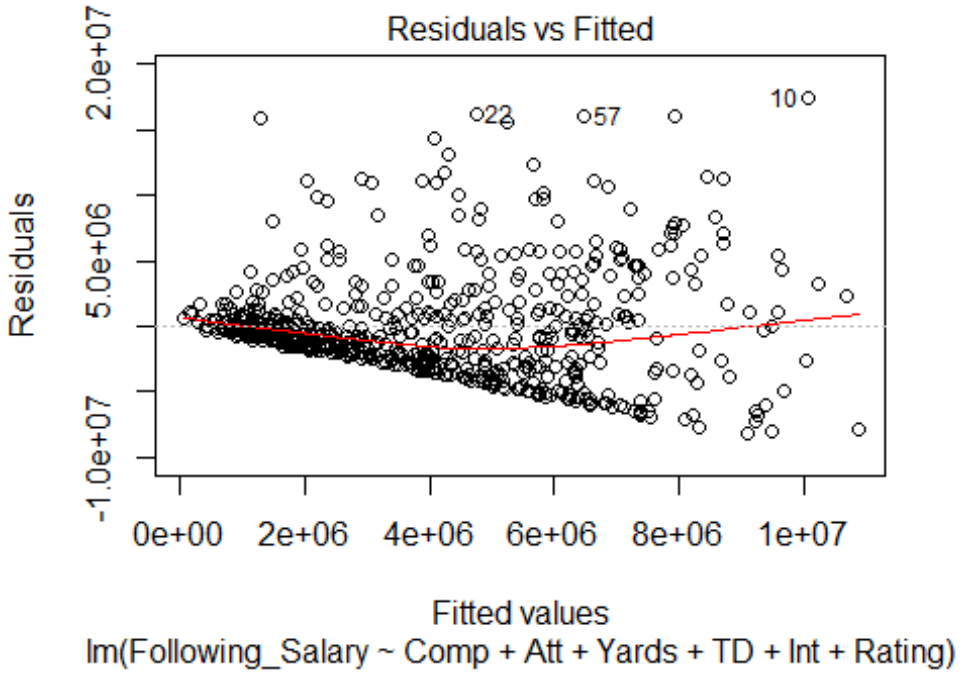
Adding Player Name

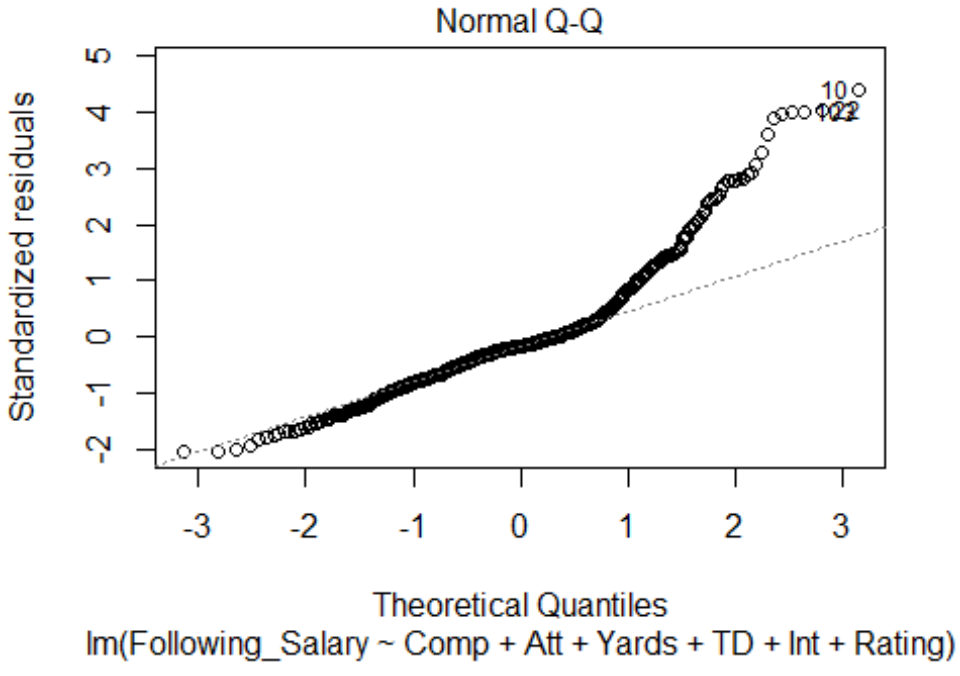
The following test will determine if the model is improved enough with the inclusion of the predictor Player Name to warrant the increased complextiy of an added predictor. Null Hypothesis H_0: The model without Player Name is sufficient, Alternative Hypothesis H_A: Player Name is a significant addition to the model given the predictors.

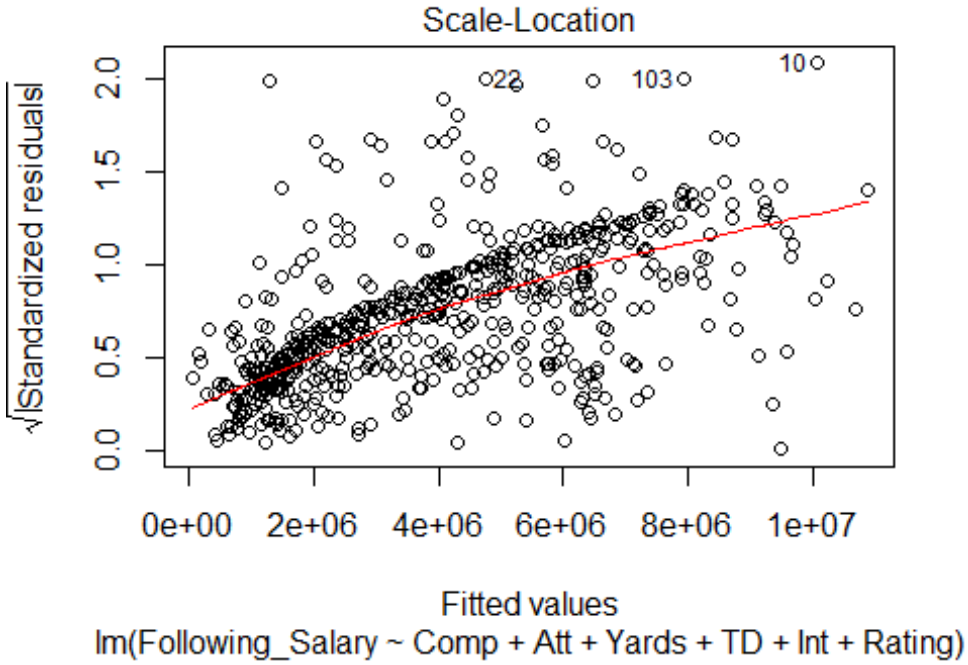
```
fullmodel <- lm(Following_Salary~Comp+Att+Yards+TD+Int+Rating+Player_Name)
anova(fit,fullmodel)

## Analysis of Variance Table
##
## Model 1: Following_Salary ~ Comp + Att + Yards + TD + Int + Rating
## Model 2: Following_Salary ~ Comp + Att + Yards + TD + Int + Rating + Playe
r_Name
##   Res.Df        RSS  Df  Sum of Sq      F  Pr(>F)
## 1    598 9.7312e+15
## 2    450 6.5769e+15 148 3.1543e+15 1.4582 0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The value of the test statistic is 1.4582, the null distribution of the test statistic is $F_{1,450}$. The P-Value is 0.00175, which is less than $\alpha = 0.05$. Therefore the null hypothesis is rejected and it is determined that Player Name is a significant addition to the model.

```
plot(fit, which = 1:3)
```

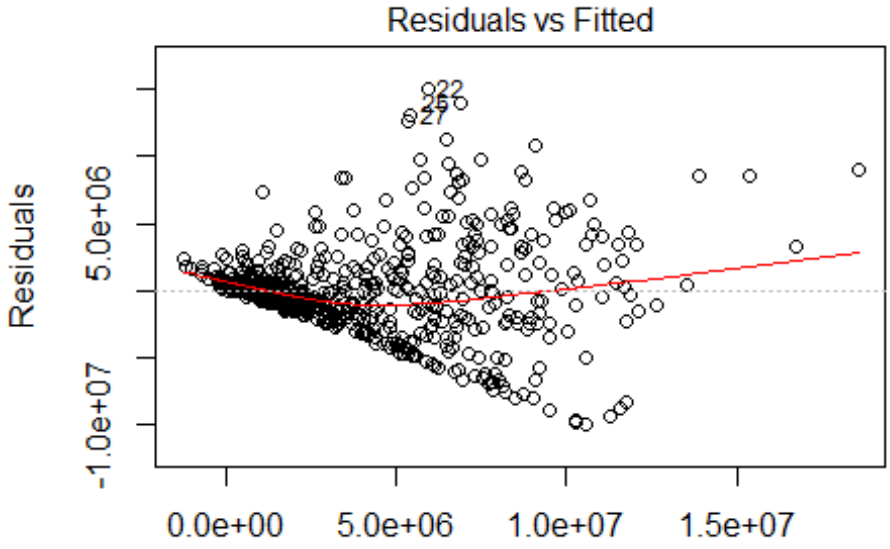

Residuals vs Fitted
Residuals
Fitted values
lm(Following_Salary ~ Comp + Att + Yards + TD + Int + Rating)
22
57
10
2.0e+07
5.0e+06
-1.0e+07
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Following_Salary ~ Comp + Att + Yards + TD + Int + Rating)
10
-2
-1
0
1
2
3
4
5
-3

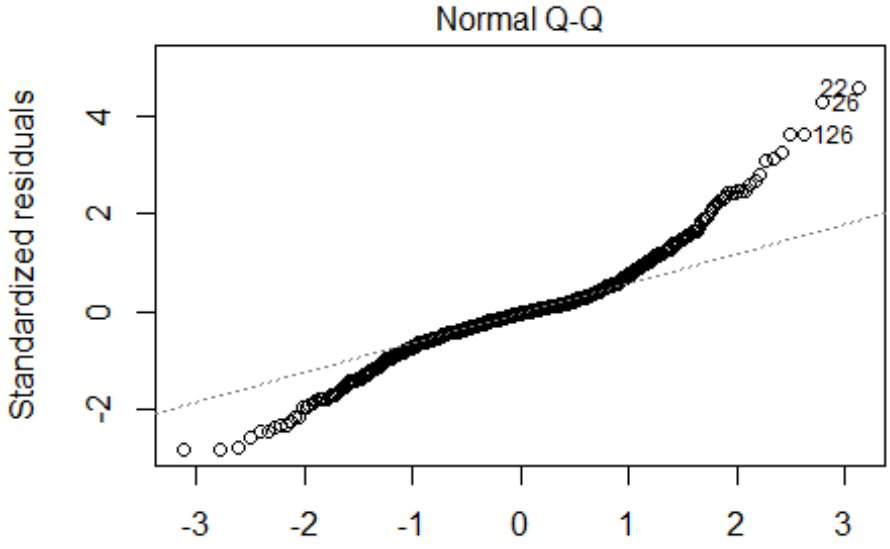



```
plot(fullmodel, which = 1:3)
```

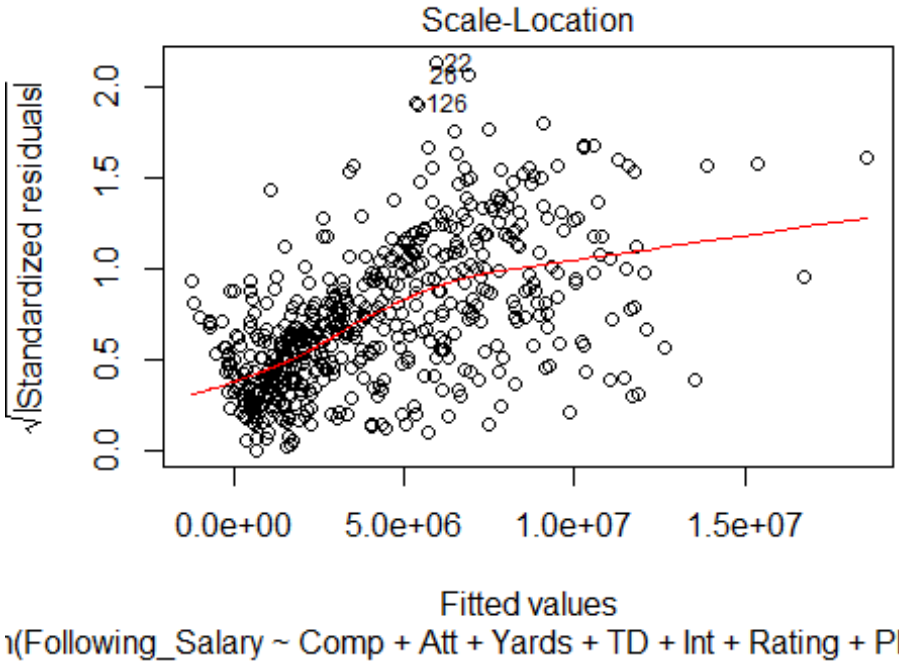
```
## Warning: not plotting observations with leverage one:
##   17, 23, 29, 30, 32, 36, 37, 44, 45, 73, 85, 87, 124, 128, 171, 205, 258,
346, 348, 382, 395, 398, 439, 468, 509, 517, 521, 523, 525, 555, 565, 567, 57
6, 586, 595, 604
```



ı(Following_Salary ~ Comp + Att + Yards + TD + Int + Rating + Player_



ı(Following_Salary ~ Comp + Att + Yards + TD + Int + Rating + Player_

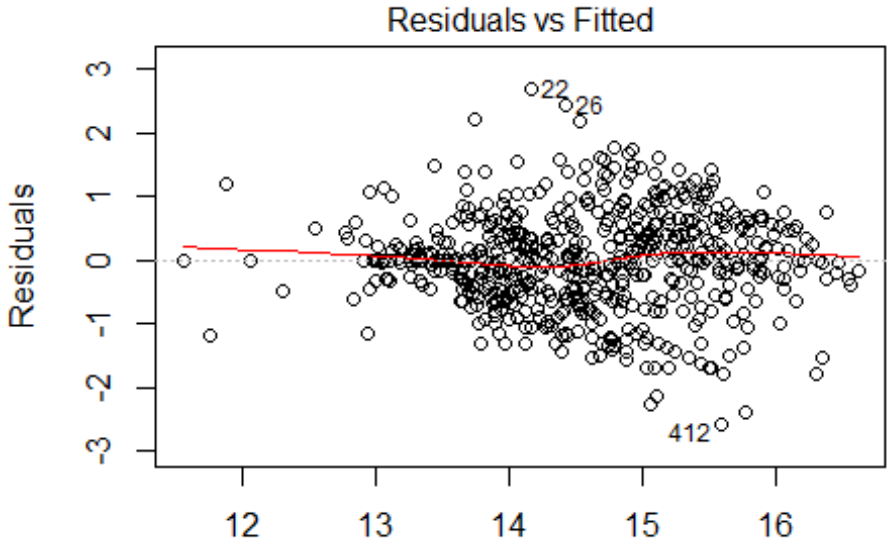



```
c(summary(fit)$r.squared, summary(fullmodel)$r.squared)

## [1] 0.2628142 0.5017674

c(AIC(fit), AIC(fullmodel))

## [1] 20130.29 20189.27
```

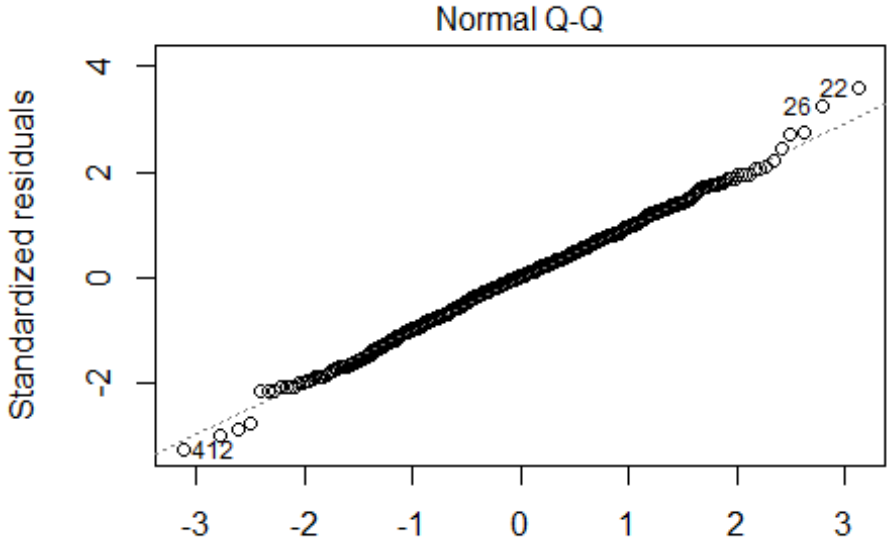
It is clear that the model with Player Name is a better fit as it has a higher R squared value although the AIC rose slightly.

Transfroming the Response

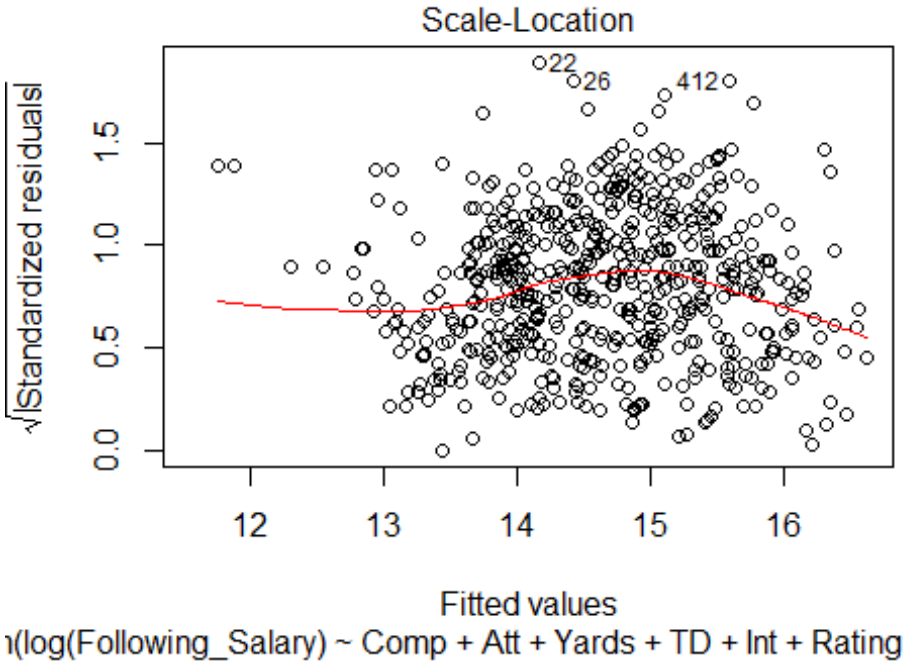
```
logtransform <- lm(log(Following_Salary)~Comp+Att+Yards+TD+Int+Rating+Player_
Name)
plot(logtransform, which = 1:3)

## Warning: not plotting observations with leverage one:
##   17, 23, 29, 30, 32, 36, 37, 44, 45, 73, 85, 87, 124, 128, 171, 205, 258,
346, 348, 382, 395, 398, 439, 468, 509, 517, 521, 523, 525, 555, 565, 567, 57
6, 586, 595, 604
```



ı(log(Following_Salary) ~ Comp + Att + Yards + TD + Int + Rating + Pla



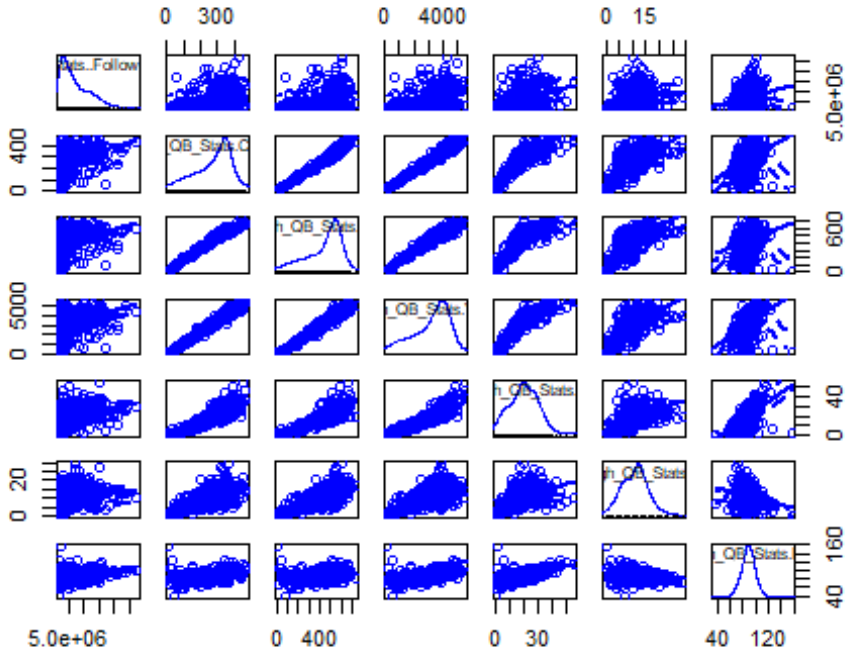
ı(log(Following_Salary) ~ Comp + Att + Yards + TD + Int + Rating + Pla



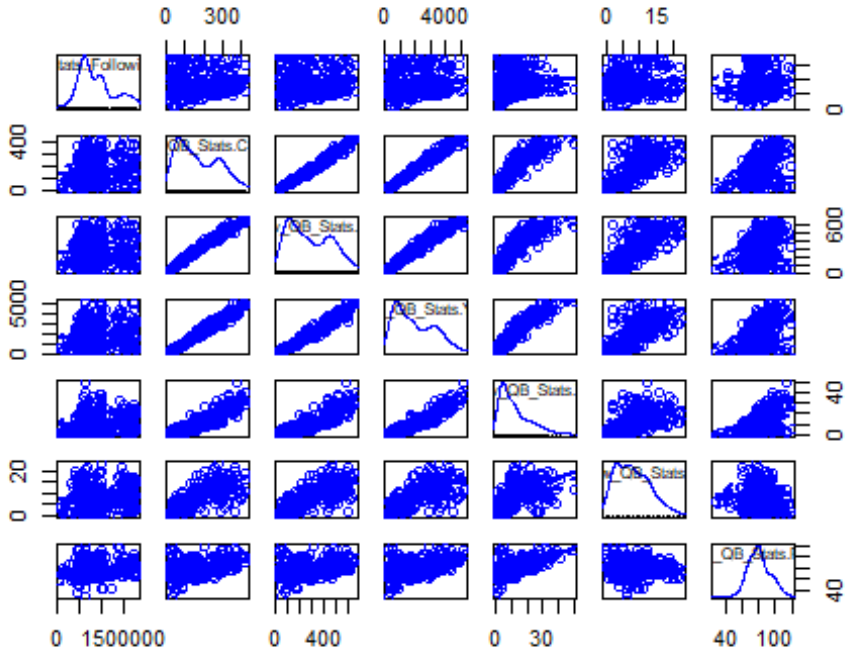

```
summary(logtransform)$r.squared

## [1] 0.5799462

AIC(logtransform)

## [1] 1683.422
```

The logarithmically transformed response improved the model's adherence to the normality of errors assumption and is the best fit out of the models created.

```
New_QB_Stats <- na.omit(QB_Stats)
median(New_QB_Stats$`Following Salary`)

## [1] 1800000

High_QB_Stats <-subset(QB_Stats ,Following_Salary > 1800000)
High_QB_lm <- lm(High_QB_Stats$`Following Salary`~High_QB_Stats$Comp+High_QB_
Stats$Att+High_QB_Stats$Yds+High_QB_Stats$TD+High_QB_Stats$Int+High_QB_Stats$
Rate+High_QB_Stats$Player)
scatterplotMatrix(~High_QB_Stats$`Following Salary`+High_QB_Stats$Comp+High_Q
B_Stats$Att+High_QB_Stats$Yds+High_QB_Stats$TD+High_QB_Stats$Int+High_QB_Stat
s$Rate)
```

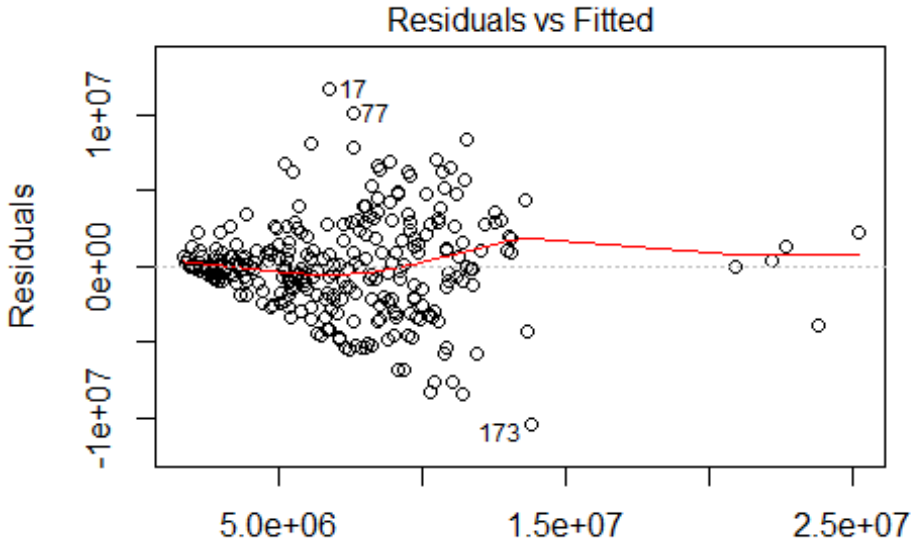



```
Low_QB_Stats <- subset(QB_Stats, Following_Salary <= 1800000)
Low_QB_lm <- lm(Low_QB_Stats$`Following Salary`~Low_QB_Stats$Comp+Low_QB_Stat
s$Att+Low_QB_Stats$Yds+Low_QB_Stats$TD+Low_QB_Stats$Int+Low_QB_Stats$Rate+Low
_QB_Stats$Player)
scatterplotMatrix(~Low_QB_Stats$`Following Salary`+Low_QB_Stats$Comp+Low_QB_S
tats$Att+Low_QB_Stats$Yds+Low_QB_Stats$TD+Low_QB_Stats$Int+Low_QB_Stats$Rate)
```

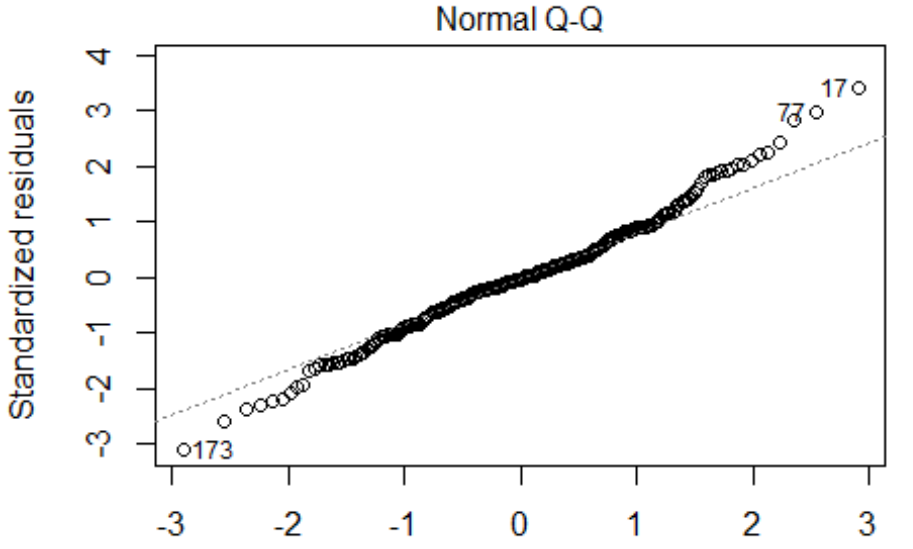



```
plot(High_QB_lm, which = 1:3)
```

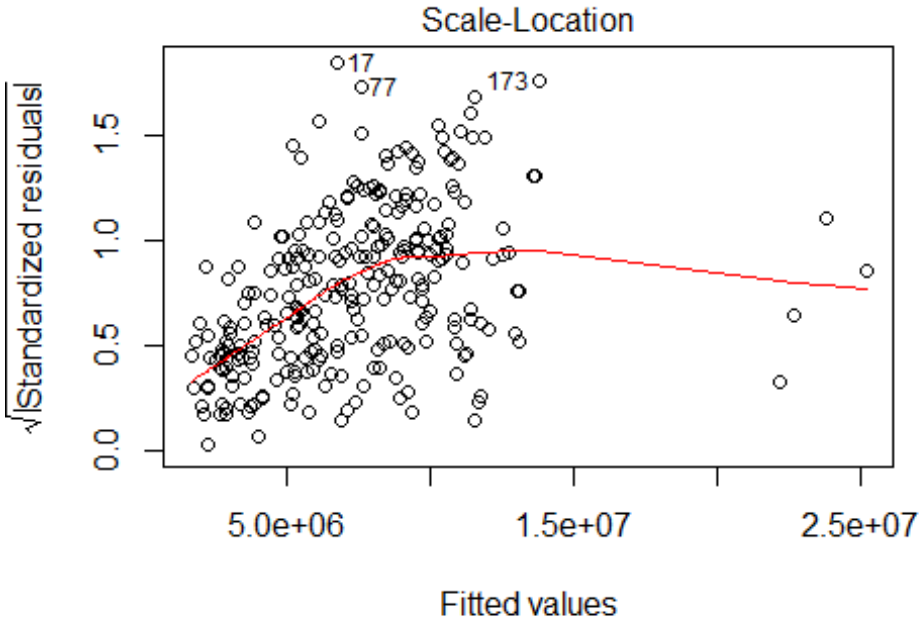
```
## Warning: not plotting observations with leverage one:
##   10, 14, 16, 22, 42, 67, 88, 110, 133, 134, 171, 194, 196, 197, 198, 237,
243, 245, 247, 268, 269, 279, 280, 283, 292, 298
```



h_QB_Stats$`Following Salary` ~ High_QB_Stats$Comp + High_QB_

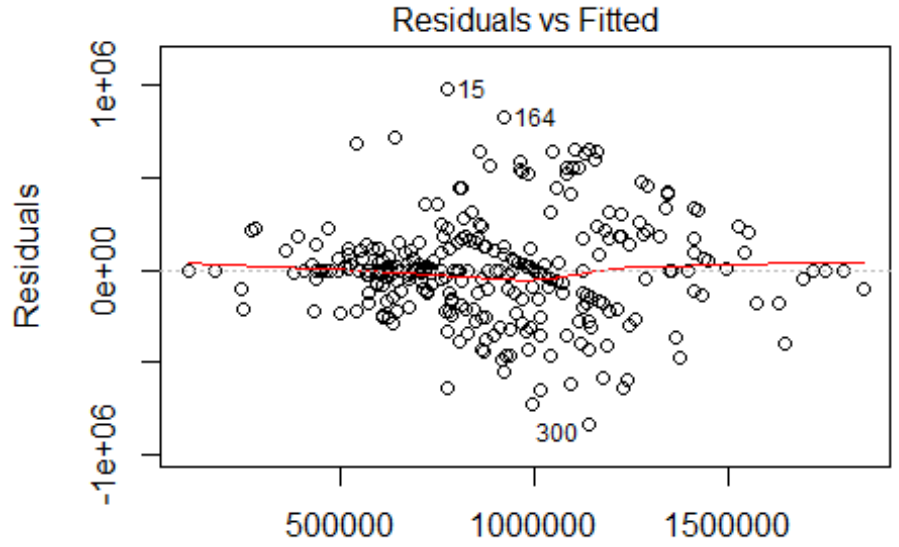


h_QB_Stats$`Following Salary` ~ High_QB_Stats$Comp + High_QB_

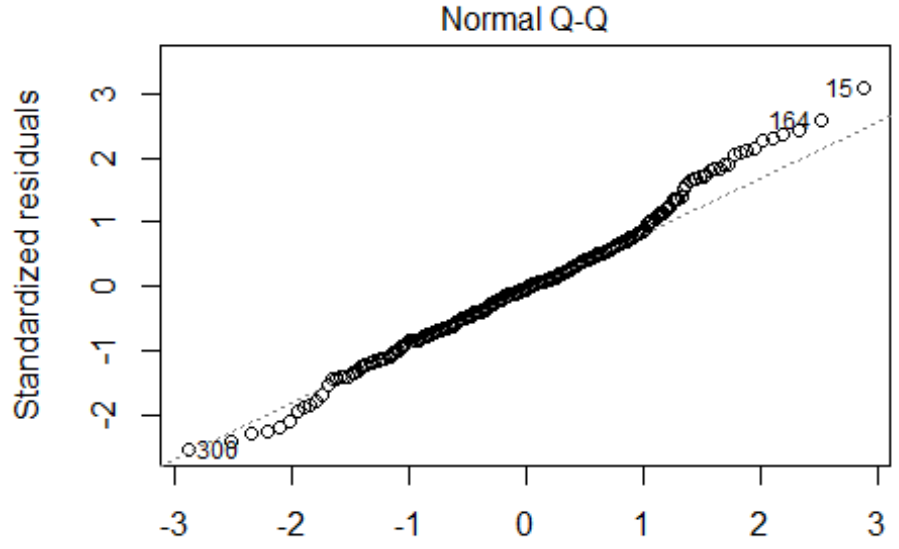



```
plot(Low_QB_lm, which = 1:3)
```

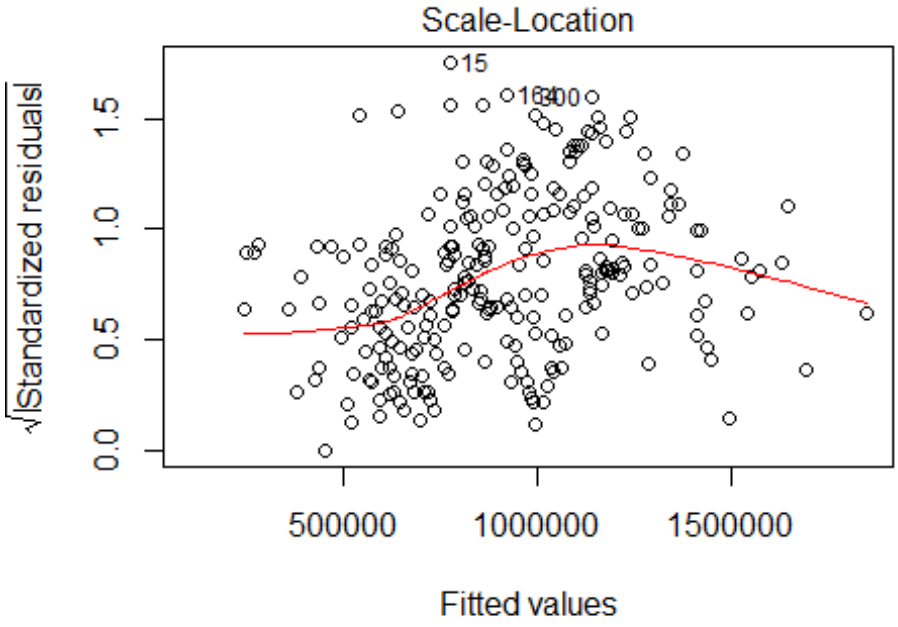
```
## Warning: not plotting observations with leverage one:
##   6, 9, 12, 13, 14, 17, 18, 22, 23, 29, 31, 40, 42, 49, 56, 58, 60, 61, 79
, 82, 96, 101, 126, 134, 140, 176, 178, 181, 182, 189, 198, 201, 215, 216, 22
5, 243, 249, 251, 252, 254, 270, 277, 279, 286, 288, 294, 298, 301, 303
```



_QB_Stats$`Following Salary` ~ Low_QB_Stats$Comp + Low_QB_S



_QB_Stats$`Following Salary` ~ Low_QB_Stats$Comp + Low_QB_S



_QB_Stats$`Following Salary` ~ Low_QB_Stats$Comp + Low_QB_S

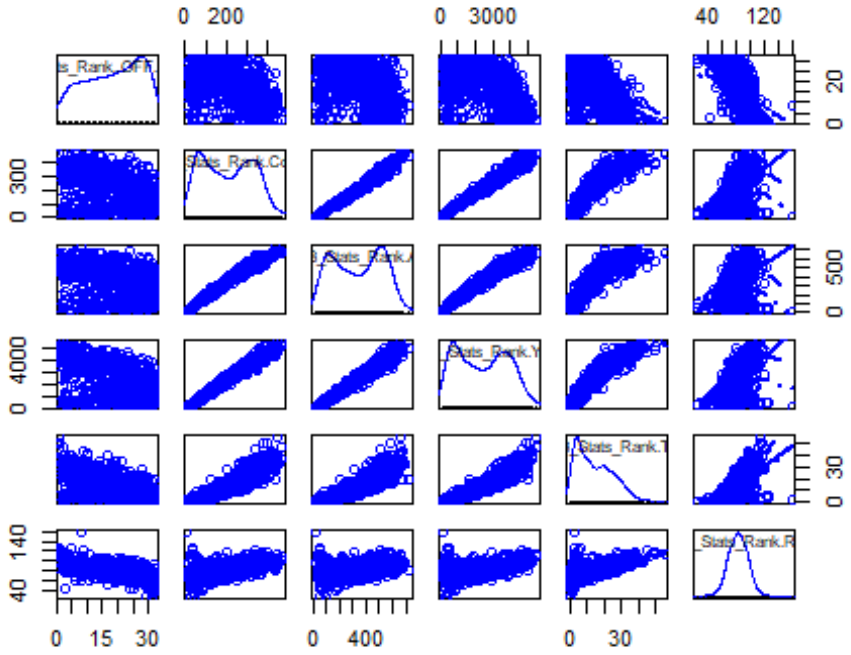
```
summary(High_QB_lm)$r.squared

## [1] 0.5989564

summary(Low_QB_lm)$r.squared

## [1] 0.5683152
```

Because of the high R squared value for the model using the dataset of Quarterbacks following salary above the median (of this dataset), we can transform the response of this model to get a stronger fit.

```
logtransform_high <- lm(log(High_QB_Stats$`Following Salary`)~High_QB_Stats$C
omp+High_QB_Stats$Att+High_QB_Stats$Yds+High_QB_Stats$TD+High_QB_Stats$Int+Hi
gh_QB_Stats$Rate+High_QB_Stats$Player)
summary(logtransform_high)$r.squared

## [1] 0.5950035
```

Since the R squared value is only slightly higher than the log transformed response of the full model which includes all the Quarterbacks in the dataset, we will use that model to estimate a Quarterback's value.

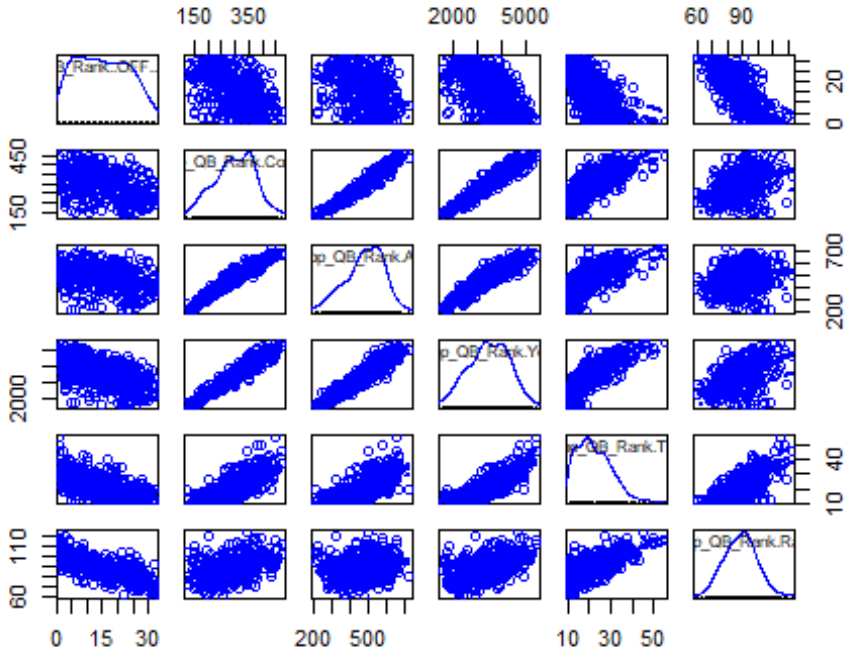
```
scatterplotMatrix(~QB_Stats_Rank$`OFF. RANK`+QB_Stats_Rank$Comp+QB_Stats_Rank
$Att+QB_Stats_Rank$Yds+QB_Stats_Rank$TD+QB_Stats_Rank$Rate)
```



There seems to be some correlation between the Team's offensive rank and the number of Touchdowns, as well as a correlation between the Team's offensive rank and Quarterback rating.

By eliminating Quarterbacks who have thrown less than 10 touchdowns, we can check if the correlation is improved since the visual above includes 2nd or even 3rd string quarterbacks.

```
Top_QB_Rank <- subset(QB_Stats_Rank, QB_Stats_Rank$TD > 10)
scatterplotMatrix(~Top_QB_Rank$`OFF. RANK`+Top_QB_Rank$Comp+Top_QB_Rank$Att+T
op_QB_Rank$Yds+Top_QB_Rank$TD+Top_QB_Rank$Rate)
```



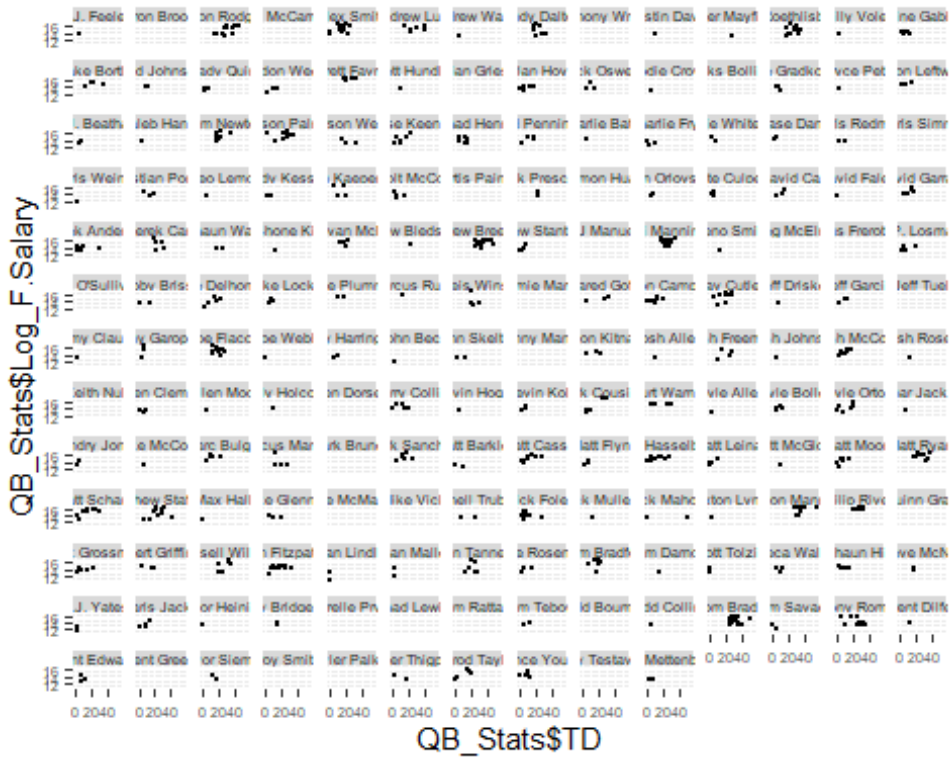

```
Top_QB_TD_lm <- lm(Top_QB_Rank$`OFF. RANK`~Top_QB_Rank$TD)
All_QB_TD_lm <- lm(QB_Stats_Rank$`OFF. RANK`~QB_Stats_Rank$TD)
Top_QB_Rate_lm <- lm(Top_QB_Rank$`OFF. RANK`~Top_QB_Rank$Rate)
All_QB_Rate_lm <- lm(QB_Stats_Rank$`OFF. RANK`~QB_Stats_Rank$Rate)
summary(Top_QB_TD_lm)$r.squared

## [1] 0.3998975

summary(All_QB_TD_lm)$r.squared

## [1] 0.342086

summary(Top_QB_Rate_lm)$r.squared

## [1] 0.5289898

summary(All_QB_Rate_lm)$r.squared

## [1] 0.42583
```

As you can see, the offensive rank of the Quarterbacks who throw more than 10 Touchdowns regressed on their team's offensive rank has the highest correlation.

```
summary(Top_QB_Rate_lm)

##
## Call:
## lm(formula = Top_QB_Rank$`OFF. RANK` ~ Top_QB_Rank$Rate)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6408  -4.2869  -0.5725   4.1536  18.5191
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.04931    2.44884   26.97   <2e-16 ***
## Top_QB_Rank$Rate -0.57452    0.02735  -21.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.987 on 393 degrees of freedom
## Multiple R-squared:  0.529,  Adjusted R-squared:  0.5278
## F-statistic: 441.4 on 1 and 393 DF,  p-value: < 2.2e-16
```

```
ggplot(QB_Stats, aes(x = QB_Stats$TD, y = QB_Stats$Log_F.Salary))+
        geom_point(colour = "black", size = .5) + facet_wrap(~QB_Stats$Playe
r)+
  theme(axis.text=element_text(size=5), strip.text = element_text(size = 5, m
argin = margin()))

## Warning: Removed 95 rows containing missing values (geom_point).
```



Conclusion:

Based on my findings, the most important variables to consider when determining a Quarterback's base salary for the future are completed passes, attempted passes, yards, touchdowns and rating. Since the quarterback rating uses these variables as well as interceptions to be measured, rating is a strong predictor to use. The base salary is difficult to measure how valuable a quarterback is because most contracts are for multiple years and there are many factors such as injury and retirement that can affect the quarterback's performance. By looking at individual cases for each player, they generally follow that more touchdowns in a previous season leads to a higher base salary. When a team is looking for a new quarterback, they should be looking for someone who would accept a low base salary with the potential to have high touchdowns. These players are generally rookies who have yet to prove their value but have the potential to perform much higher than what they are paid for. In the end, the teams with the higher offensive ranks are not those who pay the most for their quarterback but those who pay well under for what their quarterback is worth.

Works Cited

NFL QB Regular Season Statistics. Retrieved throughout 2005.

(http://www.nfl.com/stats/categorystats?tabSeq=1&season=2018&seasonType=REG&d-447263-n=1&d-447263-o=2&d-447263-p=1&statisticPositionCategory=QUARTERBACK&d-447263-s=PASSING_YARDS&qualified=true)

Spotrac NFL QB Base Salary Rankings. Retrieved throughout 2005.

(https://www.spotrac.com/nfl/rankings/base/quarterback/)

Football Outsiders Team Efficiency Ratings. Retrieved throughout 2005.

(https://www.footballoutsiders.com/stats/teameff/2010)