7/8/2018

TEAM INNVENTERA

2017

Capstone

NFL Salary Data Analysis in R

Skilled Offensive Players

Contents

[Abstract 4](#_Toc518374145)

[Hypothesis 4](#_Toc518374146)

[Initial Model(s) 5](#_Toc518374147)

[All Player Model 5](#_Toc518374148)

[QB Model 6](#_Toc518374149)

[WR Model 7](#_Toc518374150)

[Final Models 8](#_Toc518374151)

[All Player Final Model 8](#_Toc518374152)

[QB Final Model 9](#_Toc518374153)

[model coefficients 9](#_Toc518374154)

[WR Final Model 9](#_Toc518374155)

[Model coefficients 10](#_Toc518374156)

[Summary / Conclusion 10](#_Toc518374157)

[Salary Equations for all models: 11](#_Toc518374158)

[Future Modeling 12](#_Toc518374159)

[Feedback 12](#_Toc518374160)

[APPENDIX 12](#_Toc518374161)

[Statistical Analysis (All Offensive Skilled Players) 12](#_Toc518374162)

[Fitting the Model 13](#_Toc518374163)

[*New Correlation and Scatter Plot Matric* 13](#_Toc518374164)

[Regression Tree results 13](#_Toc518374165)

[Model 1 (with independent variables choose from regression tree and matrix plot) 15](#_Toc518374166)

[CIs for model parameters .95 17](#_Toc518374167)

[Variable Selection (library(MASS)) 18](#_Toc518374168)

[Diagnostic Plots 20](#_Toc518374169)

[Variable Relative Importance 21](#_Toc518374170)

[Final Model 24](#_Toc518374171)

[Show summary results 24](#_Toc518374172)

[CIs for model parameters Level = .95 24](#_Toc518374173)

[ANOVA table 24](#_Toc518374174)

[Statistical Analysis (by Position QB) 25](#_Toc518374175)

[Sub setting data (sub setting QB position from new data table based upon POS variable) 25](#_Toc518374176)

[Fitting the QB Model 25](#_Toc518374177)

[Regression Tree results 26](#_Toc518374178)

[QB Model 1 (with independent variables choose from regression tree and matrix plot) 28](#_Toc518374179)

[model coefficients 29](#_Toc518374180)

[QB Variable Selection (library(MASS)) 30](#_Toc518374181)

[QB Diagnostic Plots / Normality 34](#_Toc518374182)

[QB Final Model 37](#_Toc518374183)

[Statistical Analysis (WR) 39](#_Toc518374184)

[Fitting the Model 39](#_Toc518374185)

[Regression Tree results 40](#_Toc518374186)

[Model 1 (with independent variables choose from regression tree and matrix plot) 42](#_Toc518374187)

[Variable Selection (library(MASS)) 44](#_Toc518374188)

[WR Diagnostic Plots / Normality 47](#_Toc518374189)

[Outlier 47](#_Toc518374190)

[WR Final Model 49](#_Toc518374191)

[Description of the Dataset 51](#_Toc518374192)

[Primary Source website 51](#_Toc518374193)

[QB Source Data 51](#_Toc518374194)

[WR Source Data 51](#_Toc518374195)

[TE Source Data 52](#_Toc518374196)

[RB Source Data 52](#_Toc518374197)

[FB Source Data 52](#_Toc518374198)

[Data Issue 53](#_Toc518374199)

[Data Formatting / Importing 53](#_Toc518374200)

[R Code to import dataset 53](#_Toc518374201)

[Description of Dataset Variables: 54](#_Toc518374202)

[Summary of Data 54](#_Toc518374203)

[Pre-Normalization Correlation matrix and Scatterplot 54](#_Toc518374204)

[Basic Correlation Matrix 54](#_Toc518374205)

[Basic Scatterplot Matrix 56](#_Toc518374206)

[Normalizing data 57](#_Toc518374207)

[Tableau for Visualization 59](#_Toc518374208)

[Problems in the Data 60](#_Toc518374209)

Data and Files

github.com/illusionic

# Abstract

I was assigned this project to look at NFL player Salary data, so right off we used player Salary for 2017 NFL Season for all active players as the defined X = Dependent Variable. Now the issue of determining all the Y = Independent Variables. We had an issue with how Stats for each player position was tracked, because offense and defensive measurement were completely different we had to separate those into two different categories and break down each category into sub groups based on position. This data was sourced from [www.spotrac.com](http://www.spotrac.com) and crossed referenced against the [www.epsn.com](http://www.epsn.com) and [www.nfl.com](http://www.nfl.com). The initial model was to prove or dis-prove that TD, Yards, Games Started absent of position, would be the decided factor within determine Salary. After building out the model we were able to dis-prove this hypothesis and in fact determine multiple variables outside of those listed actually determine salary depending on which regression analysis was used and based on which sample data was being analysis.

# Hypothesis

This study was initially going to see what variable within the statistic collected determined which player was paid the most, immediately I was wanted to see if position played a major factor and based on an initial analysis of Salary versus Position it was easily determined yes this was true and was the most important factor. That in fact Quarter Back position was the highest paid position within the league on average by a large margin. Thus, I decided to look at what variable within each position group are the most important variables in determine the who gets paid most within that sub group.

I believed that TD, Rank, Yards and Completion % will be the most important variable with determining who gets paid most per position.

# Initial Model(s)

## All Player Model

All player initial model ended up with 8 independent variables with 5 of them showing as not significant or almost not significant with p-value greater than alpha .05 or very close to it. The model Adjusted R Squared is significant at .48 but will need to tweak the model to make sure all variable is significant to the model.

Call:

lm(formula = sqsalary ~ GP + GS + nlatt + nlattcom + nlcompPer +

nlyds + YDS.ATT + nltd, data = Norm\_OFF\_Salary)

Residuals:

Min 1Q Median 3Q Max

-2.20672 -0.37096 -0.06661 0.25072 2.88607

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.580124 0.107970 5.373 1.16e-07 \*\*\*

GP -0.007507 0.007442 -1.009 0.3136

GS 0.063286 0.008105 7.809 3.11e-14 \*\*\*

nlatt 0.215411 0.129073 1.669 0.0957 .

nlattcom -0.231415 0.125292 -1.847 0.0653 .

nlcompPer -0.003942 0.043068 -0.092 0.9271

nlyds 0.092056 0.117698 0.782 0.4345

YDS.ATT -0.024021 0.014043 -1.711 0.0877 .

nltd 0.309286 0.050216 6.159 1.45e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6752 on 530 degrees of freedom

Multiple R-squared: 0.4913, Adjusted R-squared: 0.4836

F-statistic: 63.98 on 8 and 530 DF, p-value: < 2.2e-16

## QB Model

QB initial model ended up with 8 independent variables with all 8 of them showing as not significant or almost not significant with p-value greater than alpha .05 or very close to it. We must find out why none of the variables are showing any real significance, we may have possible co-linearity.

Call:

lm(formula = sqsalary ~ GP + GS + nlatt + nlattcom + nlcompPer +

nltd + nlyds + YDS.ATT, data = QB\_Data)

Residuals:

Min 1Q Median 3Q Max

-224.284 -56.698 6.499 67.331 178.343

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 167.04261 47.38608 3.525 0.00081 \*\*\*

GP 3.04546 3.86260 0.788 0.43349

GS 2.21236 4.96315 0.446 0.65735

nlatt -5.04034 3.23387 -1.559 0.12426

nlattcom 5.19632 2.69146 1.931 0.05818 .

nlcompPer -0.36416 0.23646 -1.540 0.12872

nltd 6.19797 3.59704 1.723 0.08994 .

nlyds 0.51463 1.00138 0.514 0.60916

YDS.ATT 0.08127 0.28488 0.285 0.77641

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

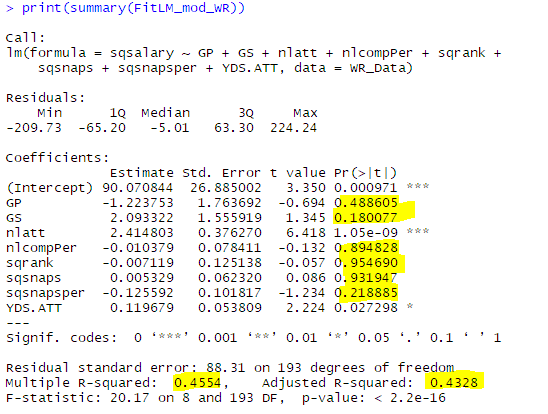
Residual standard error: 94.14 on 61 degrees of freedom

Multiple R-squared: 0.4715, Adjusted R-squared: 0.4022

F-statistic: 6.804 on 8 and 61 DF, p-value: 2.379e-06

## WR Model

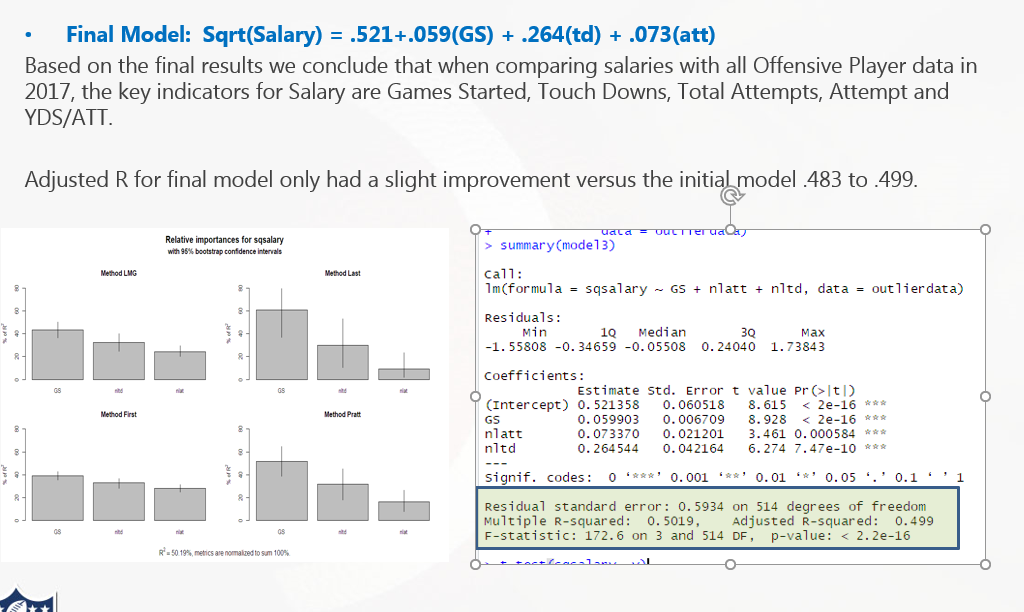
WR initial model started out with 8 independent variables similar to the other two model and shows two variable out of 6 are significant, we will need to check for co-linearity and variable fit to try to get a better model.



# Final Models

## All Player Final Model

Final model has a slightly better Adjusted r squared value and all variables are now significant



## QB [Final Model](#_Toc515016535)

As you can see the final model for QB deviate drastically from our hypothesis, the only variable that shows any significance to Salary are TD, this was an interesting find and should be looked into when we have a larger dataset.

> print(summary(QBmod3)) # Show summary results

Call:

lm(formula = sqsalary ~ nltd, data = QBOutlier)

Residuals:

Min 1Q Median 3Q Max

-174.50 -59.22 9.07 70.53 185.18

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 155.86 15.79 9.87 1.8e-14 \*\*\*

nltd 9.74 1.26 7.74 9.2e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.1 on 64 degrees of freedom

Multiple R-squared: 0.484, Adjusted R-squared: 0.476

F-statistic: 59.9 on 1 and 64 DF, p-value: 9.21e-11

### model coefficients

> print(coefficients(QBmod3)) # model coefficients

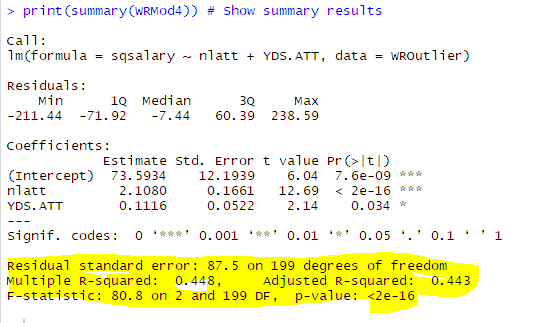
(Intercept) nltd

155.86 9.74

865

## WR [Final Model](#_Toc516215852)

WR final model also rejects the initial hypothesis that TD, Yards, and Games started will be key factors in WR salary, in fact during the initial modeling process those variable showed no significance in the initial model.



### [Model coefficients](#_Toc516215854)

> print(coefficients(WRMod4)) # model coefficients

(Intercept) nlatt YDS.ATT

73.593 2.108 0.112

# Summary / Conclusion

Our goal was to figure out if we could determine the key variables to predict player salary in the NFL, we ran into a few issue with cleaning and sorting our data, such as none of the data was normalized and the datasets were smaller than anticipated. We resolved the issues with normalization of data by using natural log, log10, square root functions to normalize the data and we used multiple modeling techniques to accuracy create a data model to choose the proper variables to predict Salary.

We started with an assumption that for Offensive players the variable widely used on TV and on ESPN and NFL.com websites Touch Downs, Yards and Games Started would be the key factors in determining salary.

With this approach we sorted and created cluster to break down the different offensive position into 5 cluster and created regression models for top two position QB and WR, we also created a overall model minus position to base line salary for overall players minus position.

The final outcome concluded that in fact our assumption for key variables was wrong when we broke out the players by position that in those variables for the most part because insignificant for predicting salary, they may be key statistics for ranking the player but when deciding the importance on Salary those variables proved inaccurate.

Important Variables: QB = TD; WR = Attempts and Yard Per Attempt

When comparing the model against all players the original hypothesis was only wrong when stating yards would be a key deciding factor, that instead of yards it was Attempts that proved significant.

Important variables for all players: Games Started, Touch Down, Attempts.

Once was able to see the new variables that showed significance in predicting salary they started to make sense, if you are only looking at Offensive skilled player. You have understood these are the players that move the ball down field and score point.

So, for the all player model the most significant measure of GS games stared, this makes sense because if you are not in the game you can’t score TD and move the ball, also most importantly the best players on the field are surely starting at each position and that would suggest they are also the highest paid players on the team for each position.

Breaking down Touch downs as a measure is important because Touch downs is the sole goal for the offensive skill players such as QB they want to win the game and they need to score TD to win the game, so the importance of this measure for both QB and all Players is a key measure

For WR position they finding that Yards per Attempt and total Attempts (receptions) does make sense because other than score points the key reason you have WR is to gain yards, this position is on the field to make huge chuck plays, so a WR with high average yard per catch would be more important than a WR with lower Yard per Catch number and if you’re not catching the ball that means you’re not score touch downs or gaining yards.

Overall the initially the results surprised me but after looking deeper into the output the analysis I can understand and agree with the assessment.

## Salary Equations for all models:

* All player (Salary) = .521+.059(GS) + .264(td) + .073(att)
* (QB) Salary = 155.86 + 9.74(TD)
* WR(Salary) = 73.59 + 2.108(ATT) + .111(Yards per Attempt)

# Future Modeling

The major issue we had with accuracy of the data modeling for all players, QB and WR was the limited data sample size. This was due to the initial focus on only using 2017 player information outlined in the initial project. For better modeling I would like to increase the sample size to include the player and salary information for the past 10 years. I would also perform modeling for both Offensive and Defensive players and later do a comparison on average salary depending on which side of the ball our play on.

The major benefit of getting a larger data sample would be generating better accuracy for our modeling, I would also be able to find the average variance different per year and create a forecast model to predict future salary for all player and by position.

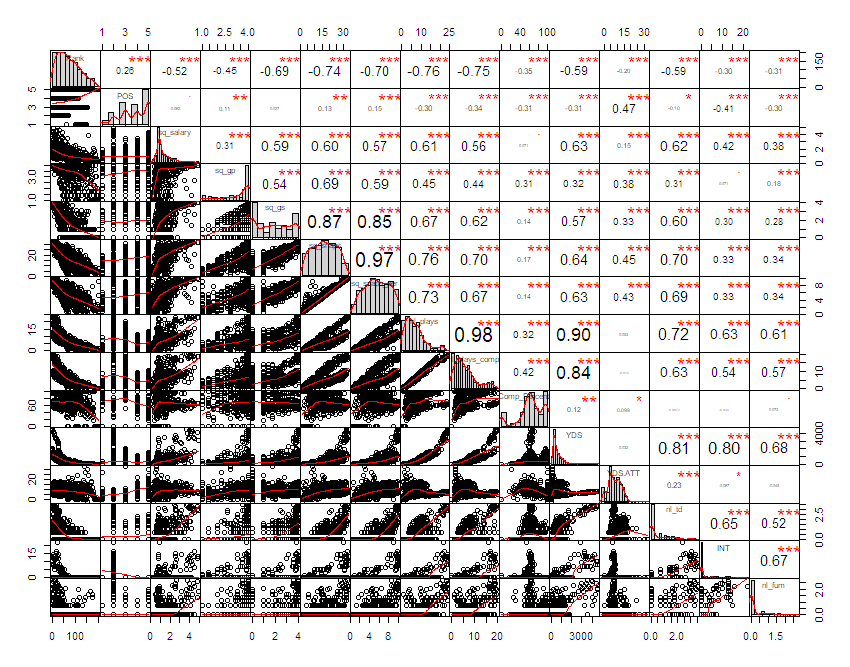
# Feedback

# APPENDIX

## Statistical Analysis (All Offensive Skilled Players)

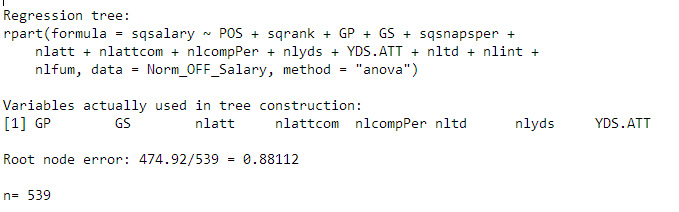
### Fitting the Model

#### New Correlation and Scatter Plot Matric with Normalized data



### Regression Tree results

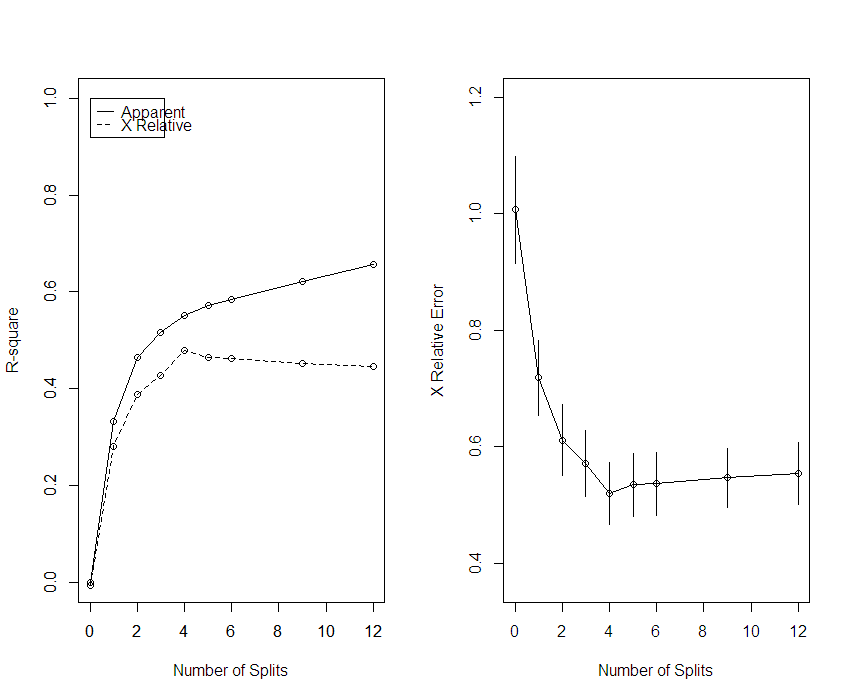
#### Regression tree:



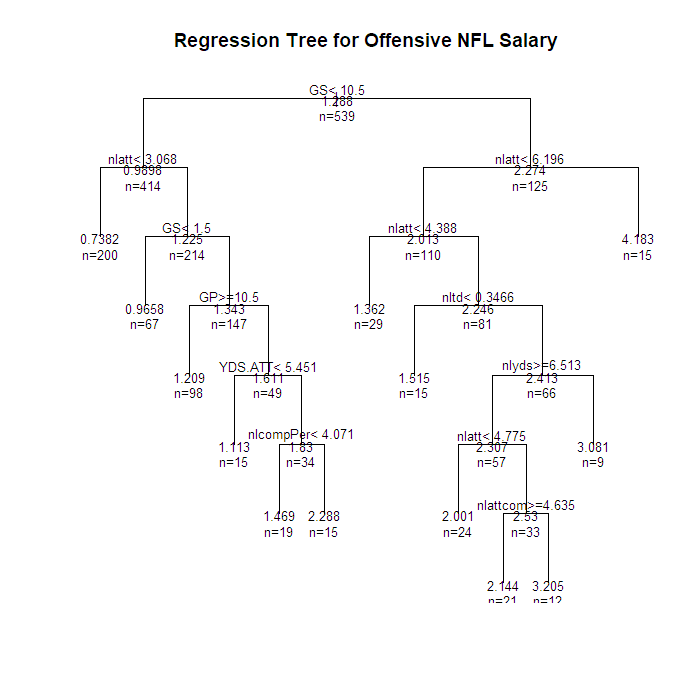
Full text output of the regression tree results

Based on the regression tree analysis and matrix plot I will build my initial model.

#### visualize cross-validation results – number of splits



#### Regression tree



### Model 1 (with independent variables choose from regression tree and matrix plot)

>

> ## Created lm Model 1

> fit\_Off\_NFL\_Salary = lm(

+ sqsalary ~ GP+GS+nlatt +nlattcom +

+ nlcompPer+nlyds+YDS.ATT+nltd,

+ data = Norm\_OFF\_Salary

+ )

> sink(

+ "School/DA 485/lm\_Summary\_AOV.txt",

+ type = "output",

+ append = FALSE,

+ split = TRUE

+ )

> print(summary(fit\_Off\_NFL\_Salary))

Call:

lm(formula = sqsalary ~ GP + GS + nlatt + nlattcom + nlcompPer +

nlyds + YDS.ATT + nltd, data = Norm\_OFF\_Salary)

Residuals:

Min 1Q Median 3Q Max

-2.20672 -0.37096 -0.06661 0.25072 2.88607

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.580124 0.107970 5.373 1.16e-07 \*\*\*

GP -0.007507 0.007442 -1.009 0.3136

GS 0.063286 0.008105 7.809 3.11e-14 \*\*\*

nlatt 0.215411 0.129073 1.669 0.0957 .

nlattcom -0.231415 0.125292 -1.847 0.0653 .

nlcompPer -0.003942 0.043068 -0.092 0.9271

nlyds 0.092056 0.117698 0.782 0.4345

YDS.ATT -0.024021 0.014043 -1.711 0.0877 .

nltd 0.309286 0.050216 6.159 1.45e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6752 on 530 degrees of freedom

Multiple R-squared: 0.4913, Adjusted R-squared: 0.4836

F-statistic: 63.98 on 8 and 530 DF, p-value: < 2.2e-16

> print(anova(fit\_Off\_NFL\_Salary))

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

GP 1 47.261 47.261 103.6731 < 2.2e-16 \*\*\*

GS 1 143.317 143.317 314.3862 < 2.2e-16 \*\*\*

nlatt 1 19.512 19.512 42.8016 1.430e-10 \*\*\*

nlattcom 1 1.362 1.362 2.9878 0.08448 .

nlcompPer 1 1.621 1.621 3.5550 0.05991 .

nlyds 1 0.142 0.142 0.3110 0.57732

YDS.ATT 1 2.807 2.807 6.1574 0.01339 \*

nltd 1 17.293 17.293 37.9350 1.447e-09 \*\*\*

Residuals 530 241.608 0.456

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> print(Confint(fit\_Off\_NFL\_Salary))

Estimate 2.5 % 97.5 %

(Intercept) 0.580124017 0.36802326 0.792224778

GP -0.007507203 -0.02212750 0.007113089

GS 0.063286437 0.04736533 0.079207545

nlatt 0.215410562 -0.03814684 0.468967962

nlattcom -0.231414876 -0.47754489 0.014715142

nlcompPer -0.003941907 -0.08854656 0.080662747

nlyds 0.092055828 -0.13915565 0.323267311

YDS.ATT -0.024021255 -0.05160764 0.003565126

nltd 0.309286229 0.21063973 0.407932731

> print(exp(coef(fit\_Off\_NFL\_Salary)))

(Intercept) GP GS nlatt nlattcom nlcompPer nlyds YDS.ATT nltd

1.7862599 0.9925209 1.0653319 1.2403710 0.7934102 0.9960659 1.0964260 0.9762650 1.3624523

#### Summary of model

> print(summary(fit\_Off\_NFL\_Salary))

Call:

lm(formula = sqsalary ~ GP + GS + nlatt + nlattcom + nlcompPer +

nlyds + YDS.ATT + nltd, data = Norm\_OFF\_Salary)

Residuals:

Min 1Q Median 3Q Max

-2.20672 -0.37096 -0.06661 0.25072 2.88607

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.580124 0.107970 5.373 1.16e-07 \*\*\*

GP -0.007507 0.007442 -1.009 0.3136

GS 0.063286 0.008105 7.809 3.11e-14 \*\*\*

nlatt 0.215411 0.129073 1.669 0.0957 .

nlattcom -0.231415 0.125292 -1.847 0.0653 .

nlcompPer -0.003942 0.043068 -0.092 0.9271

nlyds 0.092056 0.117698 0.782 0.4345

YDS.ATT -0.024021 0.014043 -1.711 0.0877 .

nltd 0.309286 0.050216 6.159 1.45e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6752 on 530 degrees of freedom

Multiple R-squared: 0.4913, Adjusted R-squared: 0.4836

F-statistic: 63.98 on 8 and 530 DF, p-value: < 2.2e-16

#### model coefficients

> print(coefficients(fit\_Off\_NFL\_Salary) )# model coefficients

(Intercept) GP GS nlatt nlattcom nlcompPer nlyds YDS.ATT

0.580124017 -0.007507203 0.063286437 0.215410562 -0.231414876 -0.003941907 0.092055828 -0.024021255

nltd

0.309286229

### CIs for model parameters .95

> print(confint(fit\_Off\_NFL\_Salary, level=0.95)) # CIs for model parameters

2.5 % 97.5 %

(Intercept) 0.36802326 0.792224778

GP -0.02212750 0.007113089

GS 0.04736533 0.079207545

nlatt -0.03814684 0.468967962

nlattcom -0.47754489 0.014715142

nlcompPer -0.08854656 0.080662747

nlyds -0.13915565 0.323267311

YDS.ATT -0.05160764 0.003565126

nltd 0.21063973 0.407932731

#### ANOVA table

> print(anova(fit\_Off\_NFL\_Salary)) # anova table

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

GP 1 47.261 47.261 103.6731 < 2.2e-16 \*\*\*

GS 1 143.317 143.317 314.3862 < 2.2e-16 \*\*\*

nlatt 1 19.512 19.512 42.8016 1.430e-10 \*\*\*

nlattcom 1 1.362 1.362 2.9878 0.08448 .

nlcompPer 1 1.621 1.621 3.5550 0.05991 .

nlyds 1 0.142 0.142 0.3110 0.57732

YDS.ATT 1 2.807 2.807 6.1574 0.01339 \*

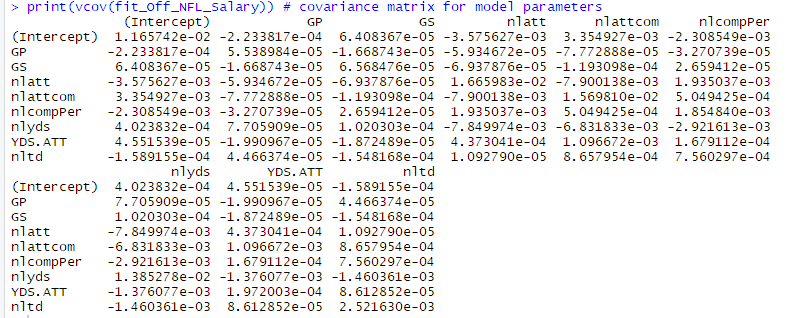
nltd 1 17.293 17.293 37.9350 1.447e-09 \*\*\*

Residuals 530 241.608 0.456

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#### covariance matrix for model parameters



### Variable Selection (library(MASS))

#### AIC Stepwise - Both

> aic\_both\_nfl\_Salary = stepAIC(fit\_Off\_NFL\_Salary, direction = "both")

Start: AIC=-414.49

sqsalary ~ GP + GS + nlatt + nlattcom + nlcompPer + nlyds + YDS.ATT +

nltd

Df Sum of Sq RSS AIC

- nlcompPer 1 0.0038 241.61 -416.48

- nlyds 1 0.2789 241.89 -415.87

- GP 1 0.4638 242.07 -415.46

<none> 241.61 -414.49

- nlatt 1 1.2697 242.88 -413.67

- YDS.ATT 1 1.3339 242.94 -413.53

- nlattcom 1 1.5551 243.16 -413.03

- nltd 1 17.2932 258.90 -379.23

- GS 1 27.7966 269.40 -357.80

Step: AIC=-416.48

sqsalary ~ GP + GS + nlatt + nlattcom + nlyds + YDS.ATT + nltd

Df Sum of Sq RSS AIC

- nlyds 1 0.3632 241.97 -417.67

- GP 1 0.4774 242.09 -417.42

<none> 241.61 -416.48

- YDS.ATT 1 1.4027 243.01 -415.36

- nlatt 1 1.5004 243.11 -415.15

- nlattcom 1 1.5544 243.17 -415.03

+ nlcompPer 1 0.0038 241.61 -414.49

- nltd 1 19.9060 261.52 -375.81

- GS 1 28.0089 269.62 -359.36

Step: AIC=-417.67

sqsalary ~ GP + GS + nlatt + nlattcom + YDS.ATT + nltd

Df Sum of Sq RSS AIC

- GP 1 0.5084 242.48 -418.54

<none> 241.97 -417.67

- nlattcom 1 1.1920 243.17 -417.03

+ nlyds 1 0.3632 241.61 -416.48

- YDS.ATT 1 1.6773 243.65 -415.95

+ nlcompPer 1 0.0881 241.89 -415.87

- nlatt 1 2.6170 244.59 -413.88

- nltd 1 20.2995 262.27 -376.25

- GS 1 27.7932 269.77 -361.07

Step: AIC=-418.54

sqsalary ~ GS + nlatt + nlattcom + YDS.ATT + nltd

Df Sum of Sq RSS AIC

<none> 242.48 -418.54

+ GP 1 0.5084 241.97 -417.67

- nlattcom 1 1.3016 243.78 -417.66

+ nlyds 1 0.3942 242.09 -417.42

+ nlcompPer 1 0.0631 242.42 -416.68

- YDS.ATT 1 2.3855 244.87 -415.27

- nlatt 1 2.5838 245.07 -414.83

- nltd 1 22.0297 264.51 -373.67

- GS 1 27.9423 270.43 -361.76

#### output: for stepwise AIC

>

> sink(

+ "stepwise\_aicstepwise.txt",

+ type = "output",

+ append = FALSE,

+ split = TRUE

+ )

> print(step\_both\_nfl\_Salary)

> print(aic\_both\_nfl\_Salary)

Call:

lm(formula = sqsalary ~ GS + nlatt + nlattcom + YDS.ATT + nltd,

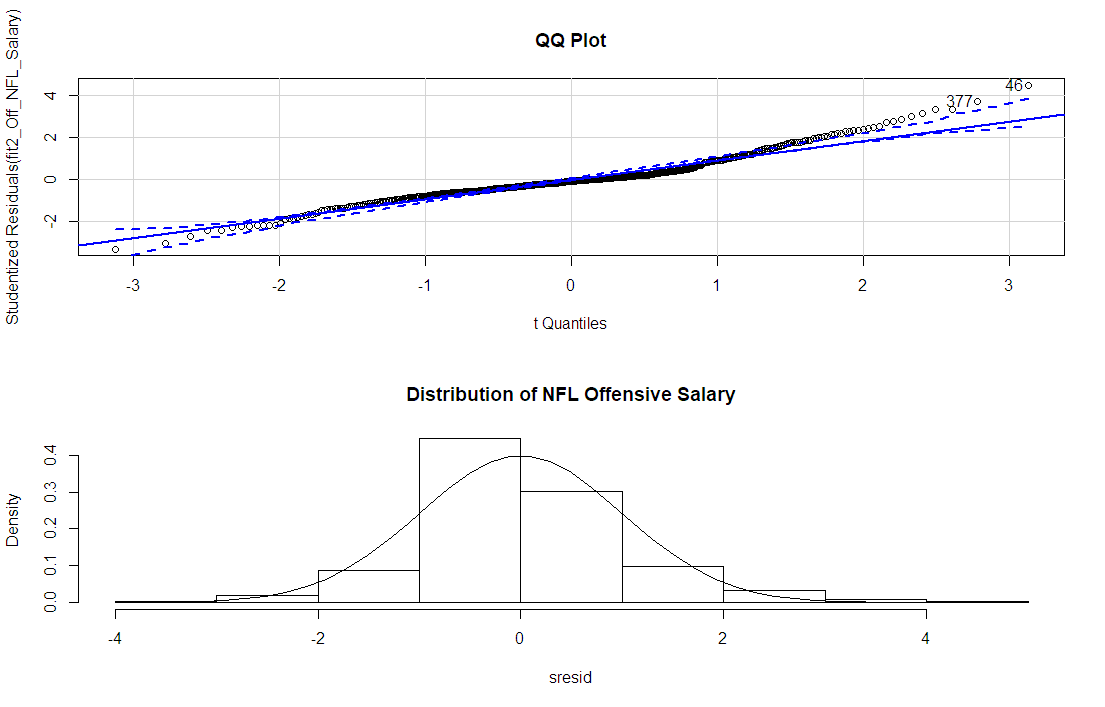
data = Norm\_OFF\_Salary)

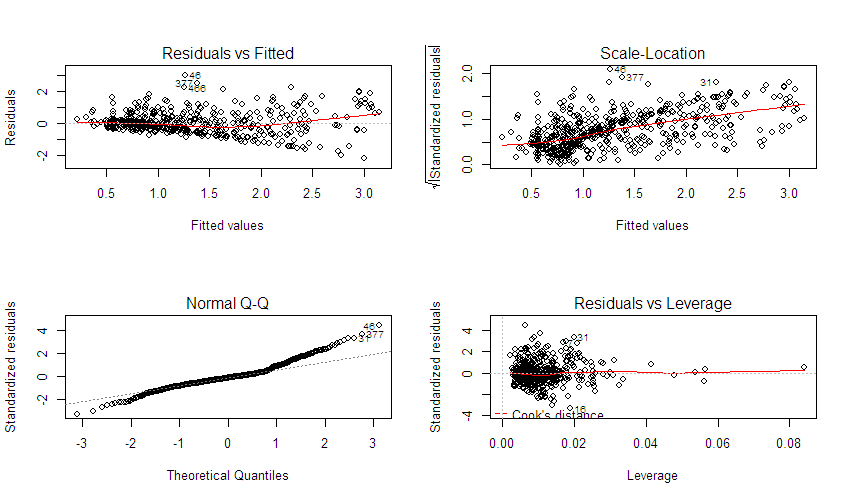
Coefficients:

(Intercept) GS nlatt nlattcom YDS.ATT nltd

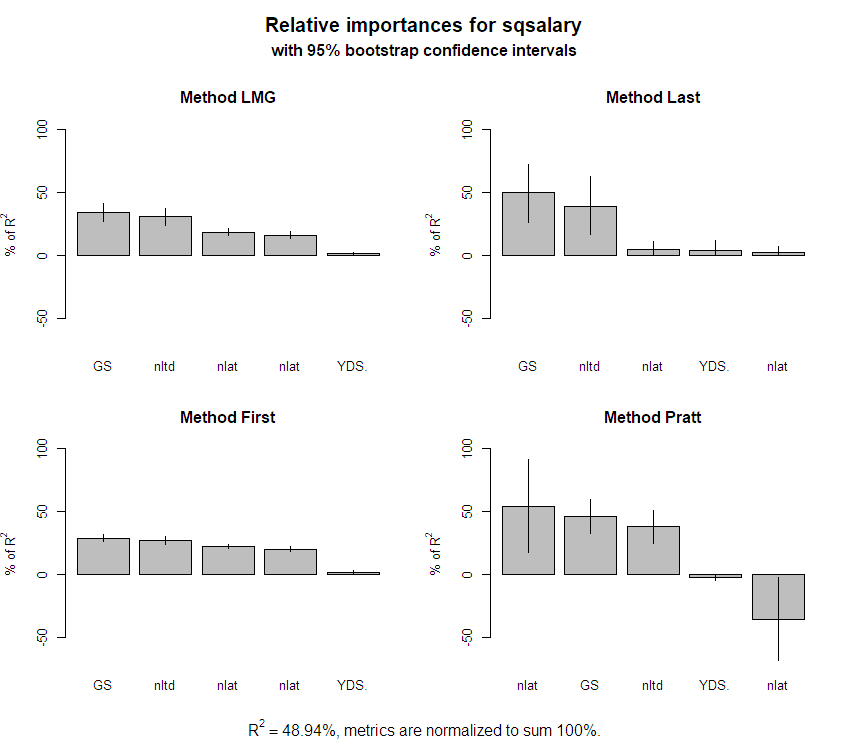
0.56881 0.05964 0.26238 -0.18177 -0.01533 0.32178

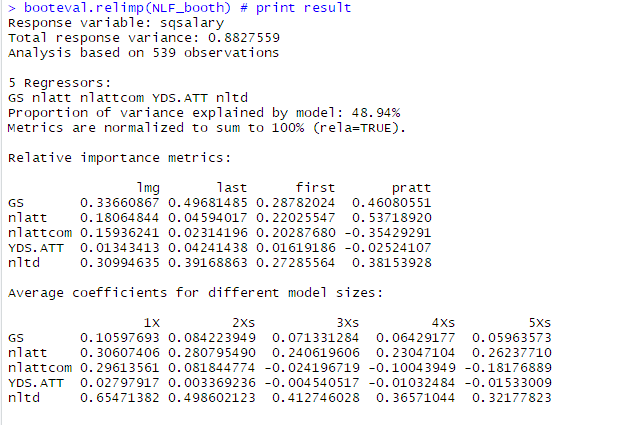
### Diagnostic Plots





### Variable Relative Importance





Confidence interval information ( 1000 bootstrap replicates, bty= perc ):

Relative Contributions with confidence intervals:

Lower Upper

percentage 0.95 0.95 0.95

GS.lmg 0.3366 AB\_\_\_ 0.2726 0.4079

nlatt.lmg 0.1806 \_\_C\_\_ 0.1577 0.2099

nlattcom.lmg 0.1594 \_\_\_D\_ 0.1383 0.1883

YDS.ATT.lmg 0.0134 \_\_\_\_E 0.0079 0.0263

nltd.lmg 0.3099 AB\_\_\_ 0.2409 0.3759

GS.last 0.4968 AB\_\_\_ 0.2581 0.7224

nlatt.last 0.0459 \_\_CD\_ 0.0054 0.1129

nlattcom.last 0.0231 \_\_\_DE 0.0002 0.0717

YDS.ATT.last 0.0424 \_\_CDE 0.0018 0.1152

nltd.last 0.3917 AB\_\_\_ 0.1678 0.6269

GS.first 0.2878 AB\_\_\_ 0.2601 0.3207

nlatt.first 0.2203 \_\_C\_\_ 0.2044 0.2382

nlattcom.first 0.2029 \_\_\_D\_ 0.1839 0.2231

YDS.ATT.first 0.0162 \_\_\_\_E 0.0039 0.0374

nltd.first 0.2729 AB\_\_\_ 0.2381 0.3062

GS.pratt 0.4608 ABC\_\_ 0.3249 0.5960

nlatt.pratt 0.5372 ABC\_\_ 0.1767 0.9123

nlattcom.pratt -0.3543 \_\_\_\_E -0.6748 -0.0230

YDS.ATT.pratt -0.0252 \_\_\_D\_ -0.0450 -0.0068

nltd.pratt 0.3815 ABC\_\_ 0.2457 0.5112

Letters indicate the ranks covered by bootstrap CIs.

(Rank bootstrap confidence intervals always obtained by percentile method)

CAUTION: Bootstrap confidence intervals can be somewhat liberal.

Differences between Relative Contributions:

Lower Upper

difference 0.95 0.95 0.95

GS-nlatt.lmg 0.1560 \* 0.0801 0.2423

GS-nlattcom.lmg 0.1772 \* 0.0992 0.2614

GS-YDS.ATT.lmg 0.3232 \* 0.2589 0.3938

GS-nltd.lmg 0.0267 -0.0968 0.1632

nlatt-nlattcom.lmg 0.0213 \* 0.0114 0.0311

nlatt-YDS.ATT.lmg 0.1672 \* 0.1398 0.1954

nlatt-nltd.lmg -0.1293 \* -0.2082 -0.0454

nlattcom-YDS.ATT.lmg 0.1459 \* 0.1184 0.1747

nlattcom-nltd.lmg -0.1506 \* -0.2284 -0.0674

YDS.ATT-nltd.lmg -0.2965 \* -0.3624 -0.2220

GS-nlatt.last 0.4509 \* 0.1985 0.6922

GS-nlattcom.last 0.4737 \* 0.2288 0.7032

GS-YDS.ATT.last 0.4544 \* 0.1968 0.7028

GS-nltd.last 0.1051 -0.3551 0.5380

nlatt-nlattcom.last 0.0228 \* 0.0044 0.0477

nlatt-YDS.ATT.last 0.0035 -0.0564 0.0663

nlatt-nltd.last -0.3457 \* -0.6015 -0.0963

nlattcom-YDS.ATT.last -0.0193 -0.0829 0.0343

nlattcom-nltd.last -0.3685 \* -0.6165 -0.1292

YDS.ATT-nltd.last -0.3493 \* -0.6028 -0.1119

GS-nlatt.first 0.0676 \* 0.0252 0.1099

GS-nlattcom.first 0.0849 \* 0.0402 0.1280

GS-YDS.ATT.first 0.2716 \* 0.2393 0.3085

GS-nltd.first 0.0150 -0.0384 0.0763

nlatt-nlattcom.first 0.0174 \* 0.0101 0.0249

nlatt-YDS.ATT.first 0.2041 \* 0.1735 0.2268

nlatt-nltd.first -0.0526 \* -0.0944 -0.0025

nlattcom-YDS.ATT.first 0.1867 \* 0.1526 0.2133

nlattcom-nltd.first -0.0700 \* -0.1154 -0.0185

YDS.ATT-nltd.first -0.2567 \* -0.2935 -0.2121

GS-nlatt.pratt -0.0764 -0.4736 0.2978

GS-nlattcom.pratt 0.8151 \* 0.4494 1.1710

GS-YDS.ATT.pratt 0.4860 \* 0.3474 0.6246

GS-nltd.pratt 0.0793 -0.1707 0.3466

nlatt-nlattcom.pratt 0.8915 \* 0.2047 1.5756

nlatt-YDS.ATT.pratt 0.5624 \* 0.1878 0.9492

nlatt-nltd.pratt 0.1556 -0.2346 0.5798

nlattcom-YDS.ATT.pratt -0.3291 \* -0.6436 -0.0080

nlattcom-nltd.pratt -0.7358 \* -1.0606 -0.3892

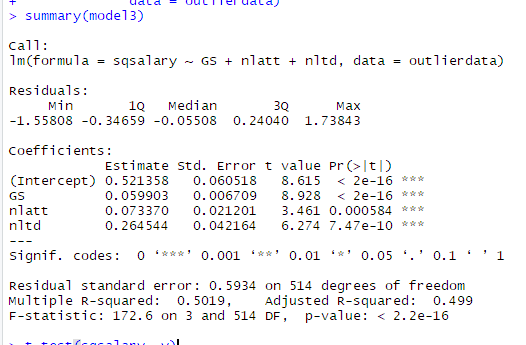
YDS.ATT-nltd.pratt -0.4068 \* -0.5385 -0.2656

## Final Model

### Show summary results

Final model has a slightly better Adjusted r squared value and all variables are now significant

> print(summary(fit2\_Off\_NFL\_Salary)) # Show summary results



### CIs for model parameters Level = .95

> print(confint(fit2\_Off\_NFL\_Salary, level = 0.95))# CIs for model parameters

2.5 % 97.5 %

(Intercept) 0.41102507 0.726589409

GS 0.04468756 0.074583897

nlatt 0.04610105 0.478653150

nltd 0.23094077 0.412615700

### ANOVA table

> print(anova(fit2\_Off\_NFL\_Salary))# anova table

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

GS 1 190.341 190.341 418.3857 < 2.2e-16 \*\*\*

nlatt 1 16.476 16.476 36.2147 3.289e-09 \*\*\*

nltd 1 22.030 22.030 48.4233 1.012e-11 \*\*\*

Residuals 533 242.483 0.455

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## [Statistical Analysis (by Position QB)](#_Toc515016519)

### Sub setting data (sub setting QB position from new data table based upon POS variable)

QB\_Data<- subset(NFLOffSalary, POS == 'QB')

View(QB\_Data)

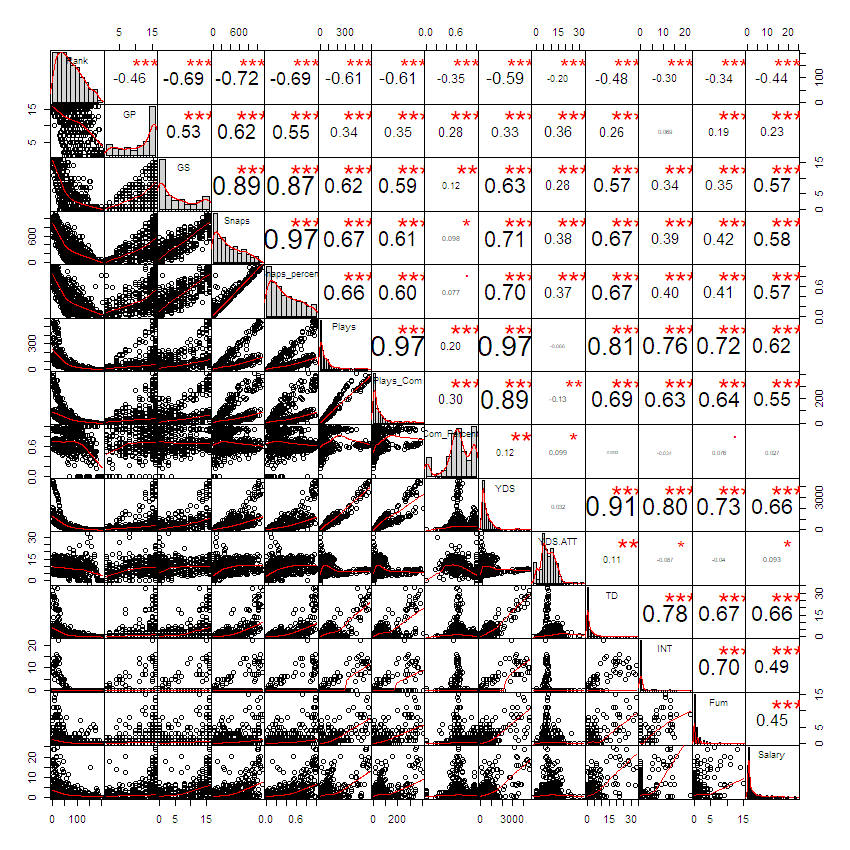
QB\_Data = QB\_Data[, 2:15]

QB\_Data = data.frame(QB\_Data)

head(QB\_Data)

### [Fitting the QB Model](#_Toc515016520)

#### Correlation Matrix Plot



### [Regression Tree results](#_Toc515016521)

# grow tree

names(QB\_Data)

QBtreefit <-

rpart(

sqsalary ~ sqrank + GP + GS + sqsnaps + sqsnapsper + nlatt + nlattcom +

nlcompPer + nlyds + nltd + nlint + nlfum ,

method = "anova",

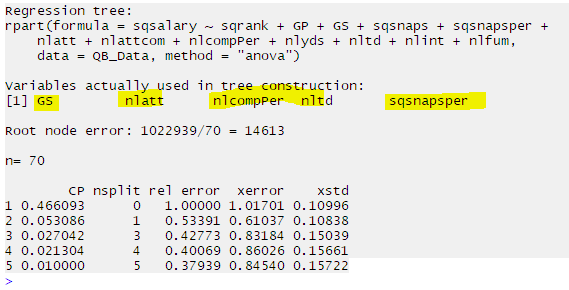
data = QB\_Data

)

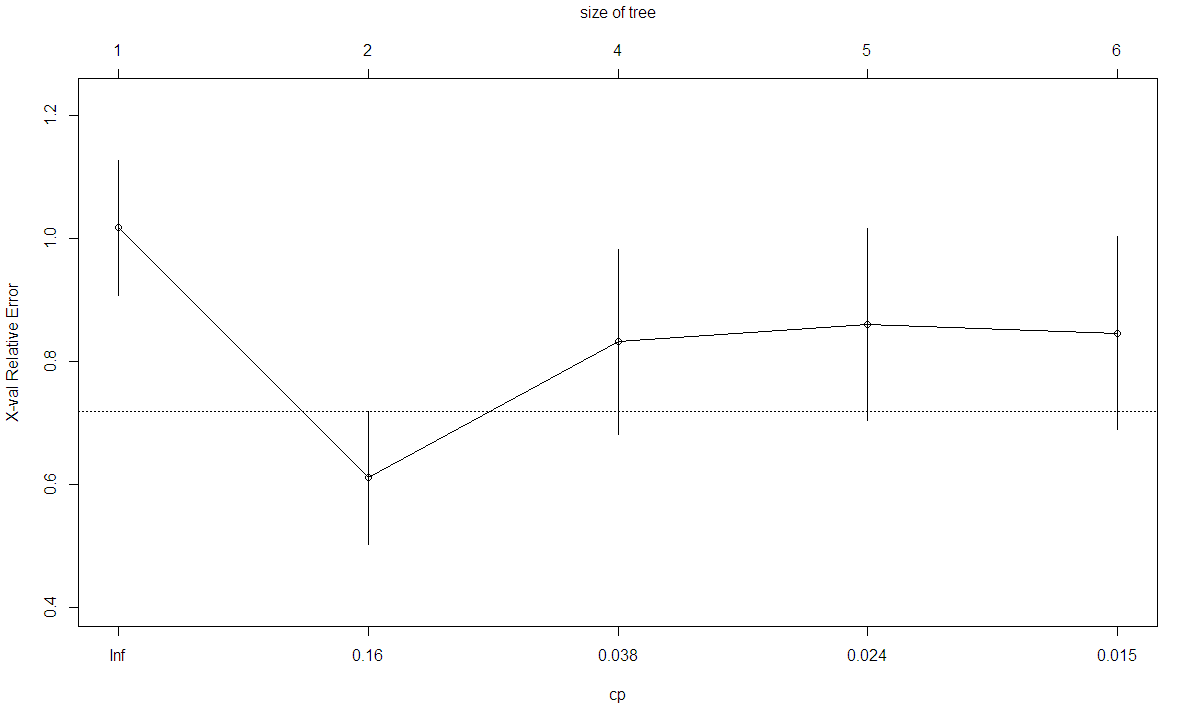
# create additional plots

par(mfrow = c(1, 2)) # two plots on one page

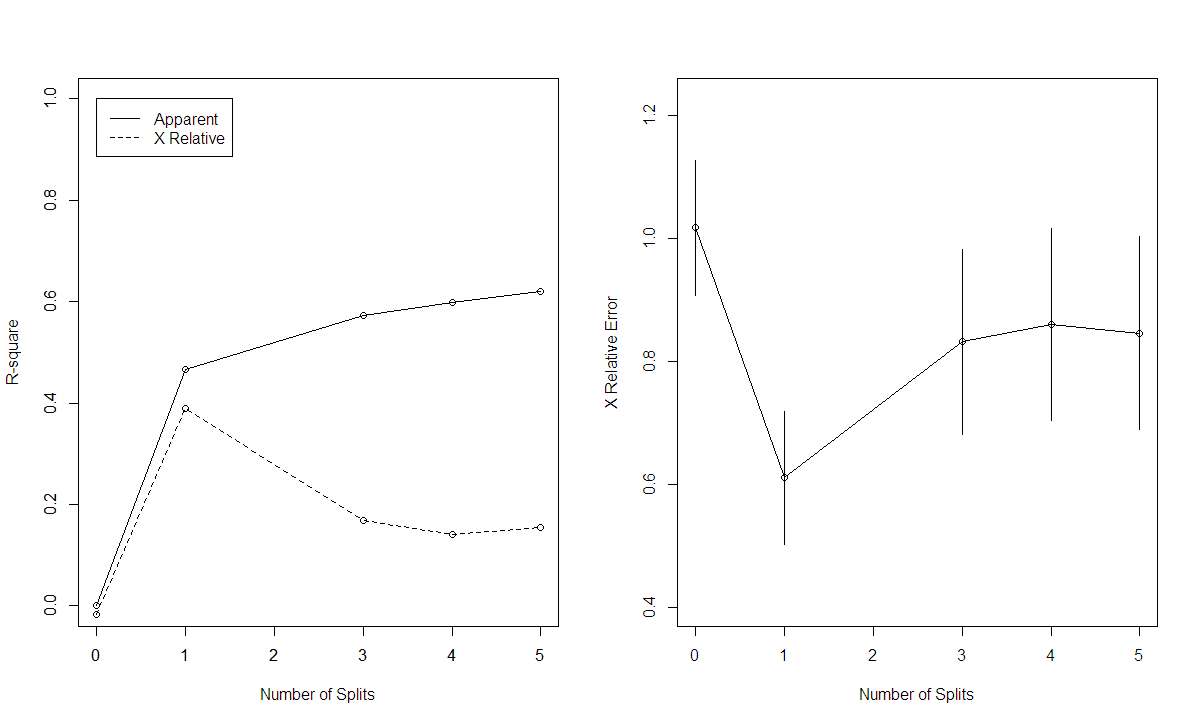
rsq.rpart(QBtreefit) # visualize cross-validation results



#### Variable importance



#### [visualize cross-validation results – number of splits](#_Toc515016522)



#### [Regression tree](#_Toc515016523)

# plot tree

par(mfrow = c(1, 1))

plot(QBtreefit, uniform = TRUE,

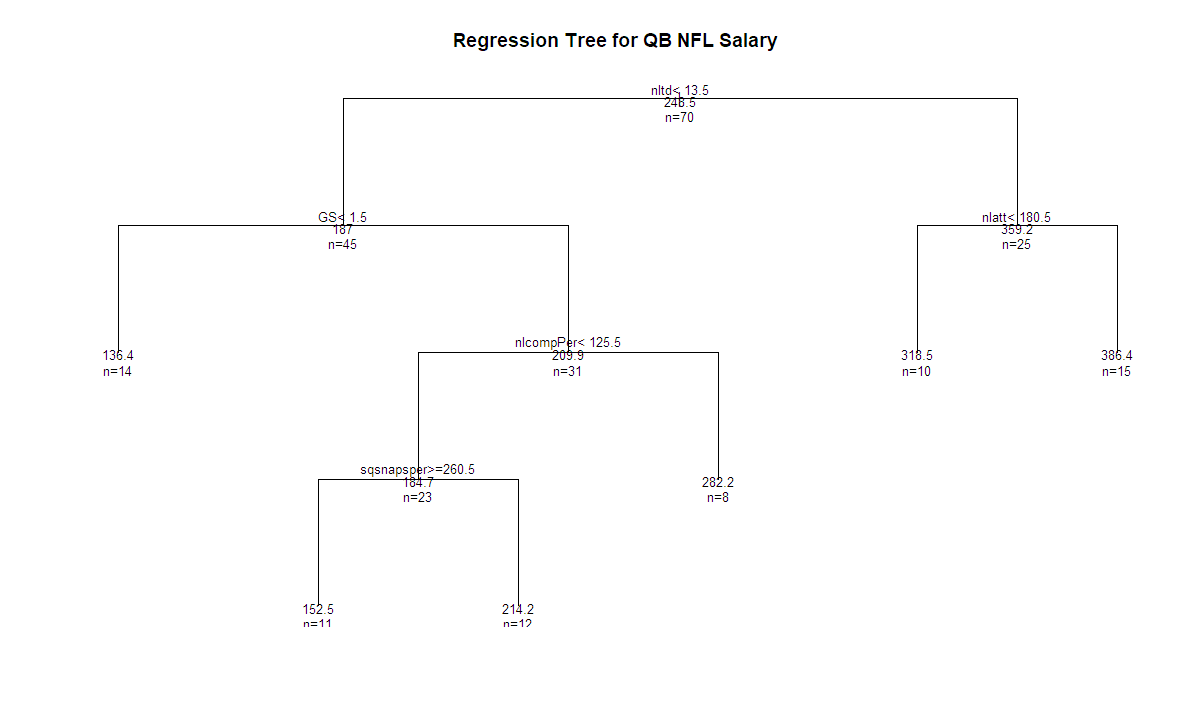
main = "Regression Tree for QB NFL Salary ")

text(QBtreefit,

use.n = TRUE,

all = TRUE,

cex = .8)



### QB [Model 1 (with independent variables choose from regression tree and matrix plot)](#_Toc515016524)

> QBmod1 = lm(

+ sqsalary ~ GS + nlatt + nlcompPer+nltd+ sqsnapsper,

+ data = QB\_Data)

> print(summary(QBmod1))

#### [Summary of model](#_Toc515016525)

Call:

lm(formula = sqsalary ~ GS + nlatt + nlcompPer + nltd + sqsnapsper,

data = QB\_Data)

Residuals:

Min 1Q Median 3Q Max

-202.80 -68.43 17.59 78.65 207.66

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 176.32310 30.13445 5.851 1.82e-07 \*\*\*

GS 0.08058 4.39300 0.018 0.98542

nlatt 0.15142 1.10567 0.137 0.89150

nlcompPer -0.08396 0.18962 -0.443 0.65943

nltd 9.89165 3.22359 3.069 0.00315 \*\*

sqsnapsper -0.11007 0.42254 -0.260 0.79533

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.73 on 64 degrees of freedom

Multiple R-squared: 0.4266, Adjusted R-squared: 0.3818

F-statistic: 9.523 on 5 and 64 DF, p-value: 7.956e-07

The initial model has low adjusted R value and shows many of the variables as insignificant to the Salary and model, will need to use stepwise to determine best fit variable for the model.

### [model coefficients](#_Toc515016526)

> coefficients(QBmod1) # model coefficients

(Intercept) GS nlatt nlcompPer nltd sqsnapsper

176.32310210 0.08058106 0.15141733 -0.08395561 9.89165021 -0.11006583

#### [CIs for model parameters .95](#_Toc515016527)

> confint(QBmod1, level = 0.95) # CIs for model parameters

2.5 % 97.5 %

(Intercept) 116.1226185 236.5235857

GS -8.6954437 8.8566058

nlatt -2.0574153 2.3602500

nlcompPer -0.4627570 0.2948458

nltd 3.4517930 16.3315074

sqsnapsper -0.9541876 0.734056

#### [ANOVA table](#_Toc515016528)

> anova(QBmod1) # anova table

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

GS 1 66074 66074 7.2096 0.009224 \*\*

nlatt 1 266956 266956 29.1286 1.055e-06 \*\*\*

nlcompPer 1 84 84 0.0092 0.923883

nltd 1 102660 102660 11.2016 0.001372 \*\*

sqsnapsper 1 622 622 0.0679 0.795325

Residuals 64 586543 9165

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

nltd 1 24727 24727 2.7903 0.099963 .

nlyds 1 14630 14630 1.6509 0.203700

YDS.ATT 1 721 721 0.0814 0.776412

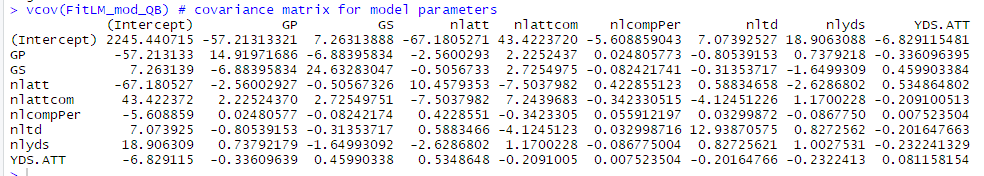
Residuals 61 540573 8862

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Items in red show not significant

#### [covariance matrix for model parameters](#_Toc515016529)



### QB [Variable Selection (library(MASS))](#_Toc515016530)

# Stepwise Regression AIC LMM

print(QBStep\_LM <- stepAIC(FitLM\_mod\_QB, direction = "both"))

print(QBStep\_LM$anova) # display results

sink()

#### [AIC Stepwise - Both](#_Toc515016531)

> print(QBStep\_LM$anova) # display results

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

sqsalary ~ GS + nlatt + nlcompPer + nltd + sqsnapsper

Final Model:

sqsalary ~ nltd

Step Df Deviance Resid. Df Resid. Dev AIC

1 64 586543.1 644.3454

2 - GS 1 3.08364 65 586546.2 642.3458

3 - nlatt 1 365.33700 66 586911.6 640.3894

4 - sqsnapsper 1 871.82543 67 587783.4 638.4933

5 - nlcompPer 1 2018.83896 68 589802.2 636.7333

> sink()

Warning message:

In sink() : no sink to remove

> # Stepwise Regression AIC LMM

> print(QBStep\_LM <- stepAIC(QBmod1, direction = "both"))

Start: AIC=644.35

sqsalary ~ GS + nlatt + nlcompPer + nltd + sqsnapsper

Df Sum of Sq RSS AIC

- GS 1 3 586546 642.35

- nlatt 1 172 586715 642.37

- sqsnapsper 1 622 587165 642.42

- nlcompPer 1 1797 588340 642.56

<none> 586543 644.35

- nltd 1 86294 672837 651.95

Step: AIC=642.35

sqsalary ~ nlatt + nlcompPer + nltd + sqsnapsper

Df Sum of Sq RSS AIC

- nlatt 1 365 586912 640.39

- sqsnapsper 1 831 587377 640.44

- nlcompPer 1 1794 588340 640.56

<none> 586546 642.35

+ GS 1 3 586543 644.35

- nltd 1 88254 674800 650.16

Step: AIC=640.39

sqsalary ~ nlcompPer + nltd + sqsnapsper

Df Sum of Sq RSS AIC

- sqsnapsper 1 872 587783 638.49

- nlcompPer 1 2039 588950 638.63

<none> 586912 640.39

+ nlatt 1 365 586546 642.35

+ GS 1 197 586715 642.37

- nltd 1 94384 681295 648.83

Step: AIC=638.49

sqsalary ~ nlcompPer + nltd

Df Sum of Sq RSS AIC

- nlcompPer 1 2019 589802 636.73

<none> 587783 638.49

+ sqsnapsper 1 872 586912 640.39

+ nlatt 1 406 587377 640.44

+ GS 1 16 587768 640.49

- nltd 1 411121 998904 673.61

Step: AIC=636.73

sqsalary ~ nltd

Df Sum of Sq RSS AIC

<none> 589802 636.73

+ nlcompPer 1 2019 587783 638.49

+ sqsnapsper 1 852 588950 638.63

+ nlatt 1 311 589491 638.70

+ GS 1 2 589800 638.73

- nltd 1 433137 1022939 673.28

Call:

lm(formula = sqsalary ~ nltd, data = QB\_Data)

Coefficients:

(Intercept) nltd

164.184 8.656

> print(QBStep\_LM$anova) # display results

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

sqsalary ~ GS + nlatt + nlcompPer + nltd + sqsnapsper

Final Model:

sqsalary ~ nltd

Step Df Deviance Resid. Df Resid. Dev AIC

1 64 586543.1 644.3454

2 - GS 1 3.08364 65 586546.2 642.3458

3 - nlatt 1 365.33700 66 586911.6 640.3894

4 - sqsnapsper 1 871.82543 67 587783.4 638.4933

5 - nlcompPer 1 2018.83896 68 589802.2 636.7333

#### Check for Outliers

#### Model 2 Summary

> print(summary(QBmod2)) # Show summary results

Call:

lm(formula = sqsalary ~ nltd, data = QB\_Data)

Residuals:

Min 1Q Median 3Q Max

-206.30 -64.99 15.66 76.45 198.85

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 164.184 16.319 10.061 4.26e-15 \*\*\*

nltd 8.656 1.225 7.067 1.08e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.13 on 68 degrees of freedom

Multiple R-squared: 0.4234, Adjusted R-squared: 0.4149

F-statistic: 49.94 on 1 and 68 DF, p-value: 1.076e-09

Adjusted R value has gone up we are now checking for outliers before we finalize the model.

> ##OUTLIERS

> # Influential Observations

> par(mfrow = c(1, 1))

> # added variable plots

> av.Plots(QBmod2)

> # Cook's D plot

> # identify D values > 4/(n-k-1)

> cutoff <- 4 / ((nrow(QBmod2) - length(QBmod2$coefficients) - 2))

> plot(QBmod2, which = 4, cook.levels = cutoff)

> # Influence Plot

> influencePlot(QBmod2,

+ id.method = "identify",

+ main = "Influence Plot",

+ sub = "Circle size is proportial to Cook's Distance")

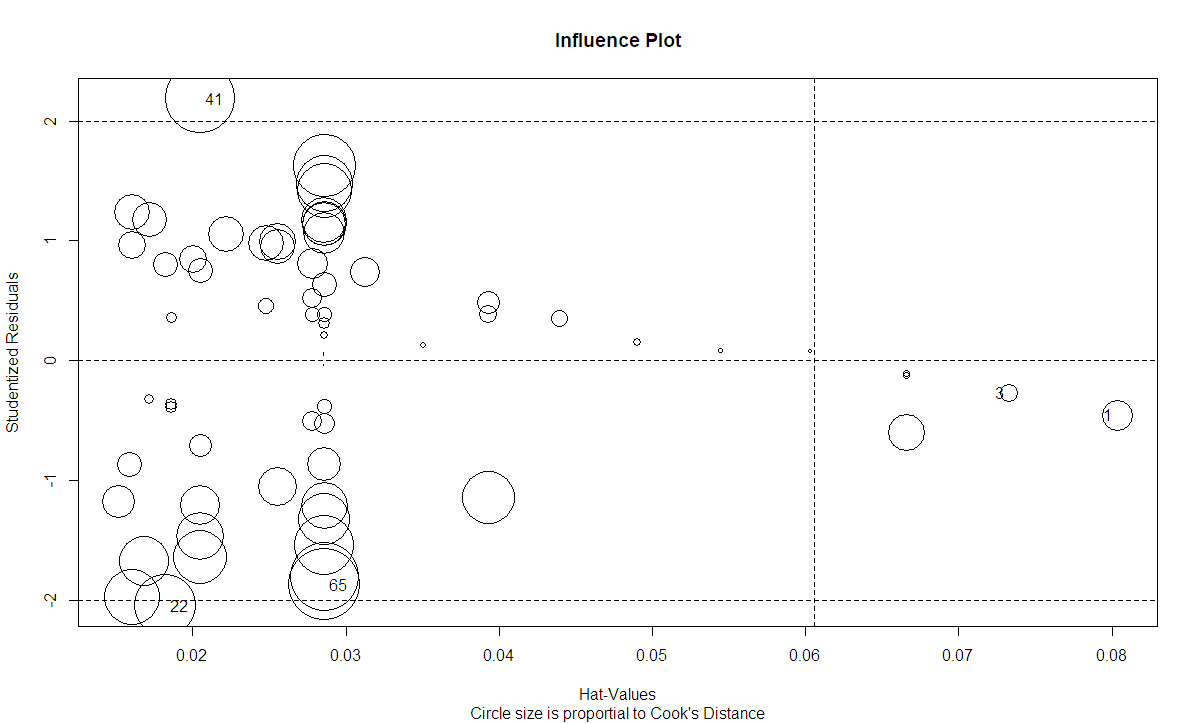
StudRes Hat CookD

9 -0.3800746 0.07842924 0.006225236

16 -2.3236798 0.03248364 0.085133875

19 -0.6944123 0.07194042 0.018833091

46 2.2207232 0.02214996 0.052801794



After running outlier again the below variable did not improve adjusted r and will not need to remove them from model.

StudRes Hat CookD

1 -0.4573519 0.08037271 0.009254812

3 -0.2708379 0.07328082 0.002942830

22 -2.0484035 0.01822128 0.037085407

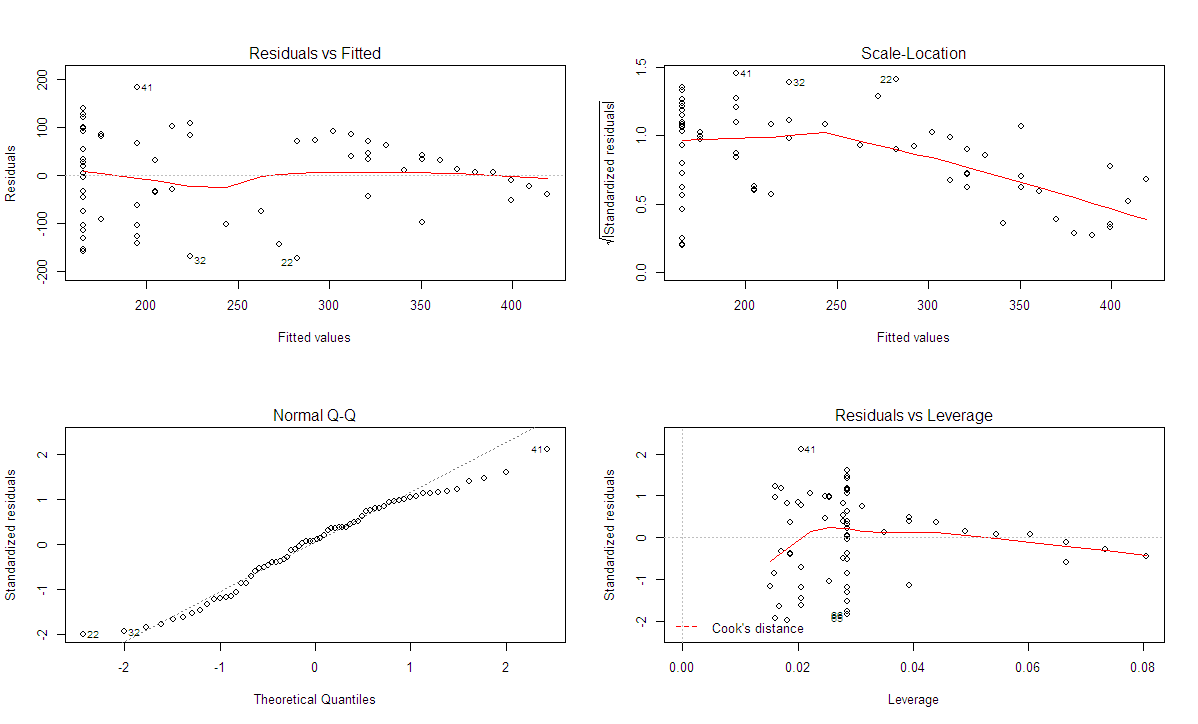
41 2.1856834 0.02050280 0.047211824

65 -1.8738928 0.02860870 0.049756162

### QB [Diagnostic Plots](#_Toc515016533) / Normality

> layout(matrix(c(1, 2, 3, 4), 2, 2)) # optional 4 graphs/page

> plot(QB\_mod2\_fit)



#### Normality of Residuals

> # qq plot for studentized resid

> qqPlot(QB\_mod2\_fit, main = "QQ Plot")

[1] 16 41

> # distribution of studentized residuals

> library(MASS)

> sresid <- studres(QB\_mod2\_fit)

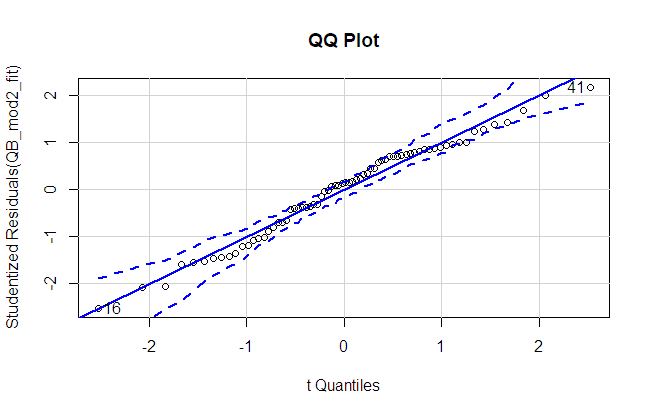
> hist(sresid, freq = FALSE,

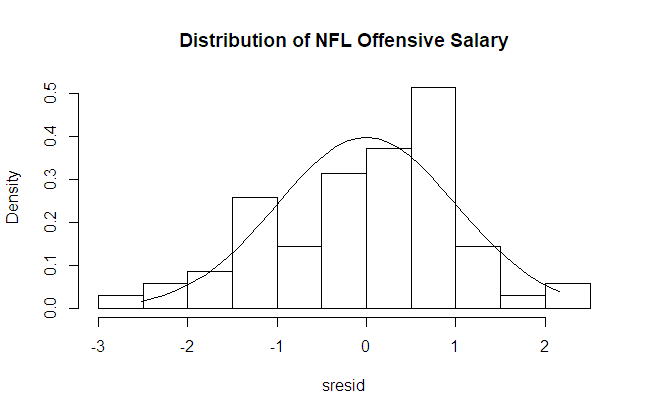
+ main = "Distribution of NFL Offensive Salary")

> xfit <- seq(min(sresid), max(sresid), length = 40)

> yfit <- dnorm(xfit)

> lines(xfit, yfit)





> booteval.relimp(QB\_booth) # print result

Response variable: sqsalary

Total response variance: 14825

Analysis based on 70 observations

2 Regressors:

nltd nlcompPer

Proportion of variance explained by model: 42.5%

Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

lmg last first pratt

nltd 0.9701 0.99511 0.9474 1.0168

nlcompPer 0.0299 0.00489 0.0526 -0.0168

Average coefficients for different model sizes:

1X 2Xs

nltd 8.66 8.8419

nlcompPer 0.29 -0.0882

Confidence interval information ( 1000 bootstrap replicates, bty= perc ):

Relative Contributions with confidence intervals:

Lower Upper

percentage 0.95 0.95 0.95

nltd.lmg 0.9701 A\_ 0.7216 0.9930

nlcompPer.lmg 0.0299 \_B 0.0068 0.2780

nltd.last 0.9951 A\_ 0.8239 1.0000

nlcompPer.last 0.0049 \_B 0.0000 0.1760

nltd.first 0.9474 A\_ 0.6555 1.0000

nlcompPer.first 0.0526 \_B 0.0001 0.3440

nltd.pratt 1.0168 A\_ 0.7773 1.0520

nlcompPer.pratt -0.0168 \_B -0.0518 0.2230

Letters indicate the ranks covered by bootstrap CIs.

(Rank bootstrap confidence intervals always obtained by percentile method)

CAUTION: Bootstrap confidence intervals can be somewhat liberal.

Differences between Relative Contributions:

Lower Upper

difference 0.95 0.95 0.95

nltd-nlcompPer.lmg 0.9401 \* 0.4433 0.9865

nltd-nlcompPer.last 0.9902 \* 0.6479 0.9999

nltd-nlcompPer.first 0.8949 \* 0.3111 0.9998

nltd-nlcompPer.pratt 1.0336 \* 0.5545 1.1036

\* indicates that CI for difference does not include 0.

CAUTION: Bootstrap confidence intervals can be somewhat liberal

.

### QB [Final Model](#_Toc515016535)

> print(summary(QBmod3)) # Show summary results

Call:

lm(formula = sqsalary ~ nltd, data = QBOutlier)

Residuals:

Min 1Q Median 3Q Max

-174.50 -59.22 9.07 70.53 185.18

#### Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 155.86 15.79 9.87 1.8e-14 \*\*\*

nltd 9.74 1.26 7.74 9.2e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.1 on 64 degrees of freedom

Multiple R-squared: 0.484, Adjusted R-squared: 0.476

F-statistic: 59.9 on 1 and 64 DF, p-value: 9.21e-11

#### model coefficients

> print(coefficients(QBmod3)) # model coefficients

(Intercept) nltd

155.86 9.74

865

#### CIs for model parameters level =.95

> print(confint(QBmod3, level = 0.95))# CIs for model parameters

2.5 % 97.5 %

(Intercept) 124.31 187.4

nltd 7.23 12.3

#### ANOVA table

> print(anova(QBmod3))# anova table

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

nltd 1 465091 465091 59.9 9.2e-11 \*\*\*

Residuals 64 496679 7761

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#### Covariance matrix for model parameters

> print(vcov(QB\_mod2\_fit)) # covariance matrix for model parameters

(Intercept) nlatt nlattcom nlcompPer nltd nlyds

(Intercept) 1091.59 -21.172 30.457 -4.4314 -19.9035 -1.7072

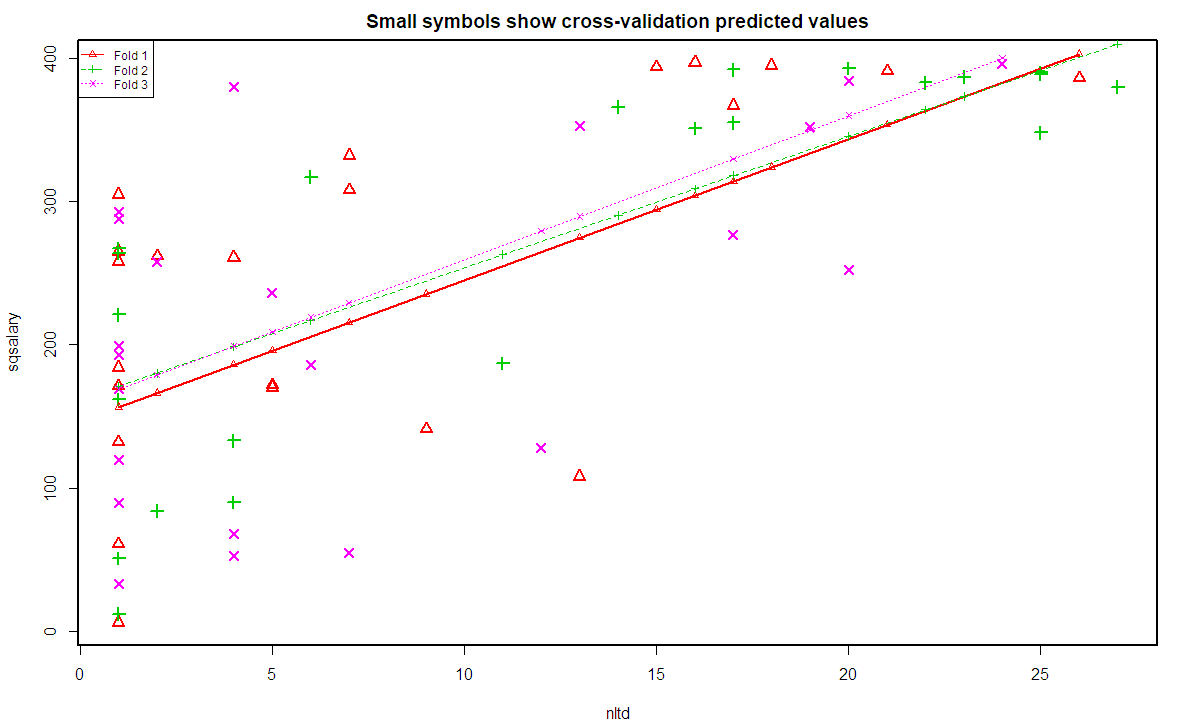
nlatt -21.17 6.100 -5.156 0.3467 1.8676 -1.1493

nlattcom 30.46 -5.156 5.463 -0.2980 -4.4132 0.6676

nlcompPer -4.43 0.347 -0.298 0.0531 0.0581 -0.0648

nltd -19.90 1.868 -4.413 0.0581 11.9249 0.2219

nlyds -1.71 -1.149 0.668 -0.0648 0.2219 0.3176



## [Statistical Analysis (WR)](#_Toc516215835)

> ## subsetting WR postion from new data table based upon POS varaible

>

> WR\_Data <- subset(NFLOffSalary, V2 == 'WR')

> WR\_Data <- WR\_Data[c(-2)]

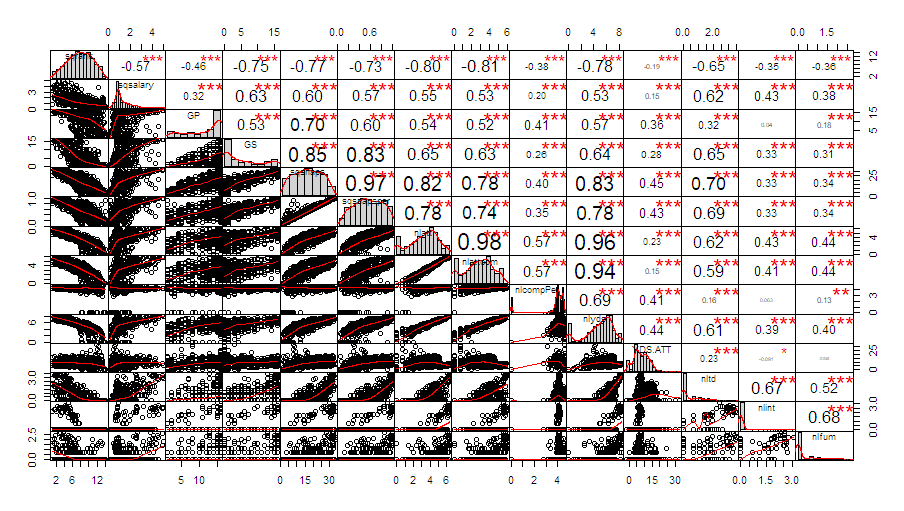
> View(WR\_Data)

> WR\_Data = data.matrix(WR\_Data)

> WR\_Data = data.frame(WR\_Data)

### [Fitting the Model](#_Toc516215836)

#### [Correlation and Scatter Plot Matric with Normalized data](#_Toc516215837)



### [Regression Tree results](#_Toc516215838)

> print(printcp(WRtreefit)) # display the results

Regression tree:

Variables actually used in tree construction:

[1] GP GS nlatt nlcompPer sqrank sqsnaps sqsnapsper YDS.ATT

Root node error: 2763670/202 = 13682

n= 202

CP nsplit rel error xerror xstd

1 0.348695 0 1.00000 1.00684 0.062338

2 0.123200 1 0.65131 0.78363 0.060230

3 0.030987 2 0.52811 0.62599 0.058469

4 0.026567 3 0.49712 0.64485 0.059659

5 0.018371 5 0.44398 0.62486 0.060699

6 0.018086 7 0.40724 0.62871 0.063715

7 0.014299 8 0.38916 0.63187 0.064060

8 0.014058 9 0.37486 0.64539 0.066338

9 0.013197 10 0.36080 0.65355 0.066719

10 0.010435 11 0.34760 0.64149 0.066404

11 0.010000 12 0.33717 0.66569 0.068303

CP nsplit rel error xerror xstd

1 0.34869485 0 1.0000000 1.0068392 0.06233842

2 0.12319971 1 0.6513052 0.7836333 0.06022995

3 0.03098652 2 0.5281054 0.6259925 0.05846910

4 0.02656721 3 0.4971189 0.6448508 0.05965861

5 0.01837070 5 0.4439845 0.6248551 0.06069851

6 0.01808617 7 0.4072431 0.6287138 0.06371546

7 0.01429931 8 0.3891569 0.6318660 0.06405986

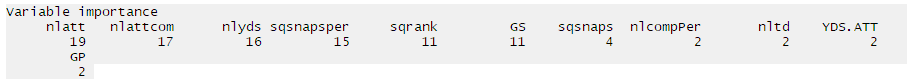
8 0.01405824 9 0.3748576 0.6453919 0.06633838

9 0.01319679 10 0.3607994 0.6535459 0.06671905

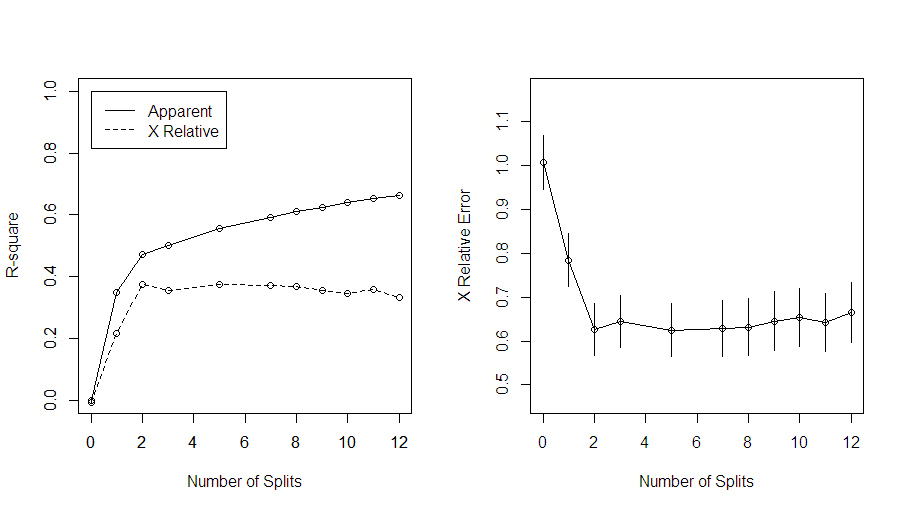
10 0.01043518 11 0.3476026 0.6414907 0.06640429

11 0.01000000 12 0.3371674 0.6656902 0.06830296

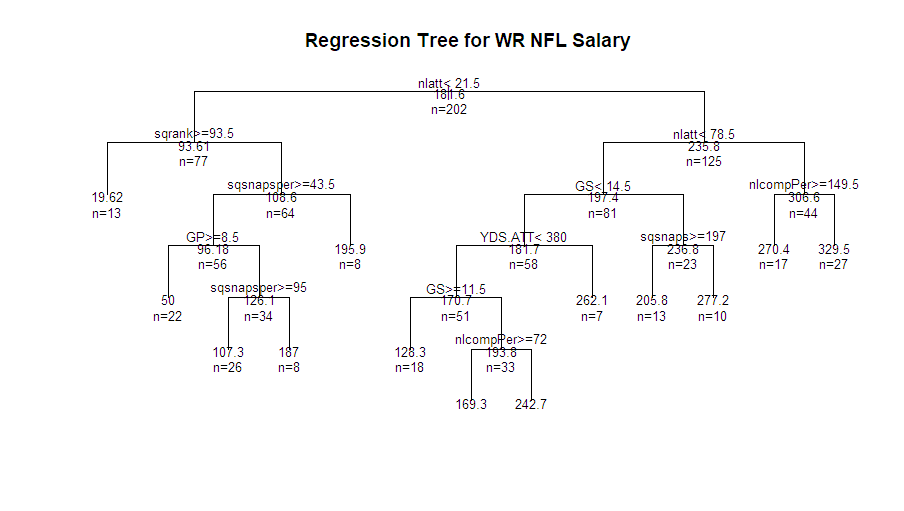
#### Variable importance



#### [visualize cross-validation results – number of splits](#_Toc516215839)



#### [Regression tree](#_Toc516215840)



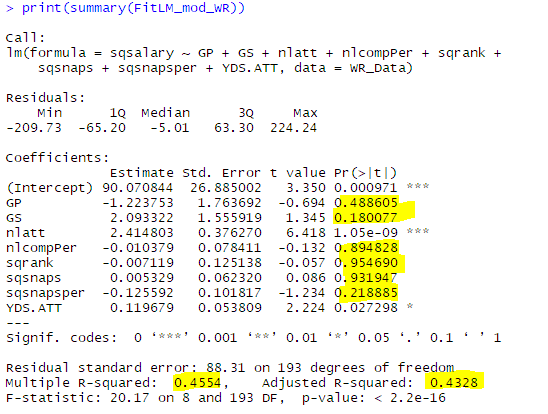
### [Model 1 (with independent variables choose from regression tree and matrix plot)](#_Toc516215841)

> FitLM\_mod\_WR = lm(sqsalary ~ GP + GS + nlatt+nlcompPer+sqrank +sqsnaps+sqsnapsper +

+ YDS.ATT,

+ data = WR\_Data)

#### [Summary of model](#_Toc516215842)



#### [model coefficients](#_Toc516215843)

> coefficients(FitLM\_mod\_WR) # model coefficients

(Intercept) GP GS nlatt nlcompPer sqrank sqsnaps sqsnapsper YDS.ATT

90.070844241 -1.223753493 2.093321899 2.414802662 -0.010379364 -0.007119395 0.005328826 -0.125592446 0.119679064

#### [CIs for model parameters .95](#_Toc516215844)

> confint(FitLM\_mod\_WR, level=0.95) # CIs for model parameters

2.5 % 97.5 %

(Intercept) 37.04470287 143.09698562

GP -4.70233830 2.25483131

GS -0.97546663 5.16211043

nlatt 1.67267423 3.15693110

nlcompPer -0.16503096 0.14427223

sqrank -0.25393397 0.23969518

sqsnaps -0.11758705 0.12824470

sqsnapsper -0.32640929 0.07522439

YDS.ATT 0.01354984 0.22580829

#### [ANOVA table](#_Toc516215845)

> anova(FitLM\_mod\_WR) # anova table

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

GP 1 9103 9103 1.1673 0.2813

GS 1 328714 328714 42.1496 6.979e-10 \*\*\*

nlatt 1 871732 871732 111.7786 < 2.2e-16 \*\*\*

nlcompPer 1 41 41 0.0052 0.9424

sqrank 1 1619 1619 0.2077 0.6491

sqsnaps 1 30 30 0.0039 0.9504

sqsnapsper 1 8696 8696 1.1150 0.2923

YDS.ATT 1 38579 38579 4.9468 0.0273 \*

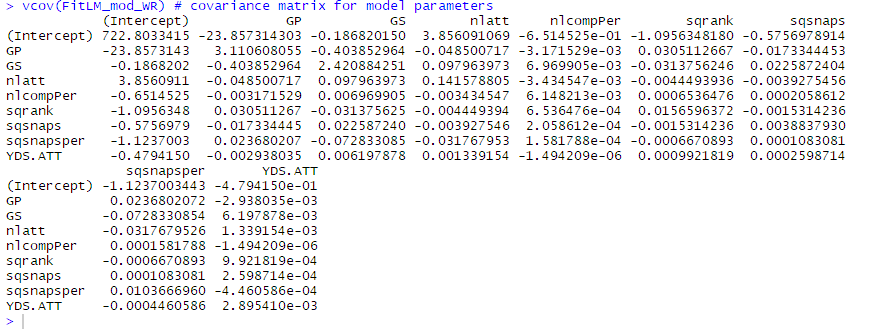
Residuals 193 1505156 7799

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#### [covariance matrix for model parameters](#_Toc516215846)

> vcov(FitLM\_mod\_WR) # covariance matrix for model parameters



### [Variable Selection (library(MASS))](#_Toc516215847)

# Stepwise Regression AIC LMM

print(WRStep\_LM <- stepAIC(FitLM\_mod\_WR, direction = "both"))

print(WRStep\_LM$anova) # display results

#### [AIC Stepwise - Both](#_Toc516215848)

> print(WRStep\_LM <- stepAIC(FitLM\_mod\_WR, direction = "both"))

Start: AIC=1819.06

sqsalary ~ GP + GS + nlatt + nlcompPer + sqrank + sqsnaps + sqsnapsper +

YDS.ATT

Df Sum of Sq RSS AIC

- sqrank 1 25 1505181 1817.1

- sqsnaps 1 57 1505213 1817.1

- nlcompPer 1 137 1505293 1817.1

- GP 1 3755 1508911 1817.6

- sqsnapsper 1 11866 1517022 1818.7

- GS 1 14116 1519273 1819.0

<none> 1505156 1819.1

- YDS.ATT 1 38579 1543735 1822.2

- nlatt 1 321210 1826366 1856.1

Step: AIC=1817.06

sqsalary ~ GP + GS + nlatt + nlcompPer + sqsnaps + sqsnapsper +

YDS.ATT

Df Sum of Sq RSS AIC

- sqsnaps 1 45 1505226 1815.1

- nlcompPer 1 130 1505311 1815.1

- GP 1 3741 1508923 1815.6

- sqsnapsper 1 11956 1517138 1816.7

- GS 1 14296 1519477 1817.0

<none> 1505181 1817.1

+ sqrank 1 25 1505156 1819.1

- YDS.ATT 1 39733 1544914 1820.3

- nlatt 1 323561 1828743 1854.4

Step: AIC=1815.07

sqsalary ~ GP + GS + nlatt + nlcompPer + sqsnapsper + YDS.ATT

Df Sum of Sq RSS AIC

- nlcompPer 1 139 1505365 1813.1

- GP 1 3699 1508925 1813.6

- sqsnapsper 1 11967 1517193 1814.7

- GS 1 14596 1519823 1815.0

<none> 1505226 1815.1

+ sqsnaps 1 45 1505181 1817.1

+ sqrank 1 13 1505213 1817.1

- YDS.ATT 1 39921 1545148 1818.4

- nlatt 1 337267 1842493 1853.9

Step: AIC=1813.09

sqsalary ~ GP + GS + nlatt + sqsnapsper + YDS.ATT

Df Sum of Sq RSS AIC

- GP 1 3738 1509103 1811.6

- sqsnapsper 1 11914 1517279 1812.7

- GS 1 14814 1520179 1813.1

<none> 1505365 1813.1

+ nlcompPer 1 139 1505226 1815.1

+ sqsnaps 1 54 1505311 1815.1

+ sqrank 1 7 1505358 1815.1

- YDS.ATT 1 39854 1545219 1816.4

- nlatt 1 339411 1844776 1852.2

Step: AIC=1811.59

sqsalary ~ GS + nlatt + sqsnapsper + YDS.ATT

Df Sum of Sq RSS AIC

- sqsnapsper 1 10287 1519390 1811.0

- GS 1 13479 1522581 1811.4

<none> 1509103 1811.6

+ GP 1 3738 1505365 1813.1

+ nlcompPer 1 177 1508925 1813.6

+ sqrank 1 18 1509085 1813.6

+ sqsnaps 1 1 1509102 1813.6

- YDS.ATT 1 38974 1548077 1814.7

- nlatt 1 335731 1844834 1850.2

Step: AIC=1810.96

sqsalary ~ GS + nlatt + YDS.ATT

Df Sum of Sq RSS AIC

- GS 1 5881 1525271 1809.7

<none> 1519390 1811.0

+ sqsnapsper 1 10287 1509103 1811.6

+ GP 1 2111 1517279 1812.7

+ nlcompPer 1 112 1519278 1813.0

+ sqrank 1 7 1519382 1813.0

+ sqsnaps 1 3 1519387 1813.0

- YDS.ATT 1 36376 1555765 1813.7

- nlatt 1 902504 2421894 1903.1

Step: AIC=1809.74

sqsalary ~ nlatt + YDS.ATT

Df Sum of Sq RSS AIC

<none> 1525271 1809.7

+ GS 1 5881 1519390 1811.0

+ sqsnapsper 1 2689 1522581 1811.4

+ GP 1 1859 1523411 1811.5

+ nlcompPer 1 275 1524995 1811.7

+ sqsnaps 1 256 1525015 1811.7

+ sqrank 1 100 1525170 1811.7

- YDS.ATT 1 35033 1560304 1812.3

- nlatt 1 1234586 2759857 1927.5

Call:

lm(formula = sqsalary ~ nlatt + YDS.ATT, data = WR\_Data)

Coefficients:

(Intercept) nlatt YDS.ATT

73.5934 2.1080 0.1116

> print(WRStep\_LM$anova) # display results

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

sqsalary ~ GP + GS + nlatt + nlcompPer + sqrank + sqsnaps + sqsnapsper +

YDS.ATT

Final Model:

sqsalary ~ nlatt + YDS.ATT

Step Df Deviance Resid. Df Resid. Dev AIC

1 193 1505156 1819.060

2 - sqrank 1 25.24229 194 1505181 1817.064

3 - sqsnaps 1 44.82231 195 1505226 1815.070

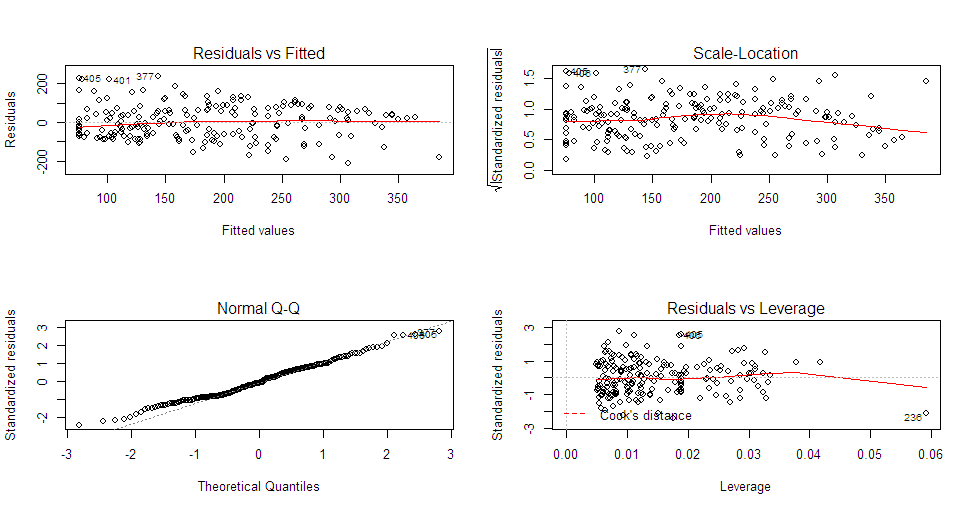
4 - nlcompPer 1 138.69764 196 1505365 1813.088

5 - GP 1 3737.63032 197 1509103 1811.589

6 - sqsnapsper 1 10287.30581 198 1519390 1810.961

7 - GS 1 5880.82032 199 1525271 1809.742

### WR [Diagnostic Plots](#_Toc516215850) / Normality



### Outlier

> # Cook's D plot

> # identify D values > 4/(n-k-1)

> cutoff <- 4/((nrow(WRMod3)-length(WRMod3$coefficients)-2))

> plot(WRMod3, which=4, cook.levels=cutoff)

> # Influence Plot

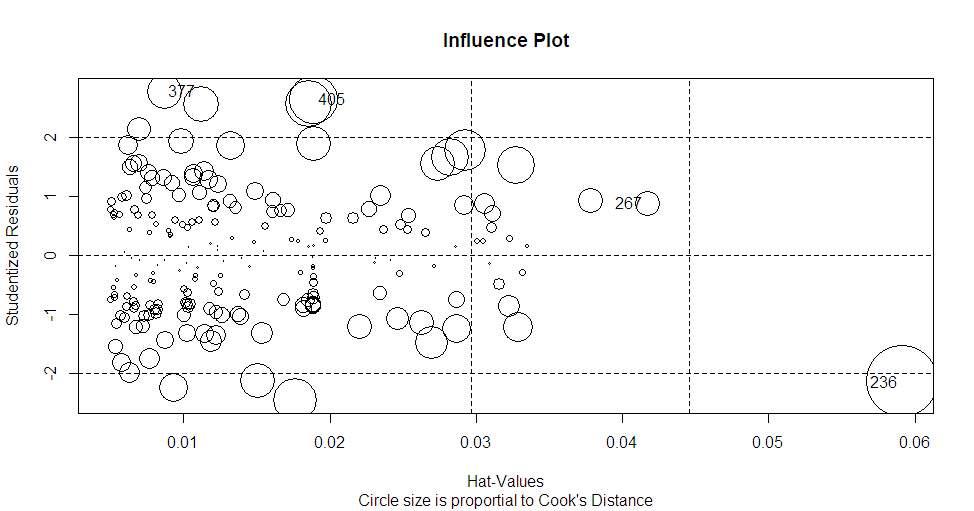
> influencePlot(WRMod3, id.method="identify", main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )

StudRes Hat CookD

236 -2.1473338 0.059106025 0.09483287

267 0.8827446 0.041692961 0.01131329

377 2.7831607 0.008650967 0.02179290



#### [Variable Relative Importance](#_Toc516215851)

> # Calculate Relative Importance for Each Predictor

> #library(relaimpo)

> calc.relimp(

+ WRMod4,

+ type = c("lmg", "last", "first", "pratt"),

+ rela = TRUE

+ )

Response variable: sqsalary

Total response variance: 13750

Analysis based on 202 observations

2 Regressors:

nlatt YDS.ATT

Proportion of variance explained by model: 44.8%

Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

lmg last first pratt

nlatt 0.9843 0.9724 0.99684 0.99061

YDS.ATT 0.0157 0.0276 0.00316 0.00939

Average coefficients for different model sizes:

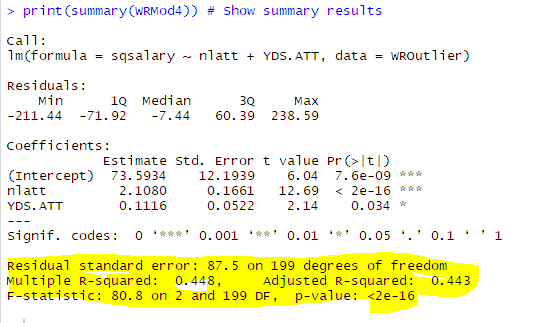
1X 2Xs

nlatt 2.0678 2.108

YDS.ATT 0.0366 0.112

### WR [Final Model](#_Toc516215852)

#### [Show summary results](#_Toc516215853)



#### [Model coefficients](#_Toc516215854)

> print(coefficients(WRMod4)) # model coefficients

(Intercept) nlatt YDS.ATT

73.593 2.108 0.112

#### [CIs for model parameters Level = .95](#_Toc516215855)

> print(confint(WRMod4, level = 0.95))# CIs for model parameters

2.5 % 97.5 %

(Intercept) 49.54749 97.639

nlatt 1.78046 2.436

YDS.ATT 0.00866 0.21

#### [ANOVA table](#_Toc516215856)

> print(anova(WRMod4))# anova table

Analysis of Variance Table

Response: sqsalary

Df Sum Sq Mean Sq F value Pr(>F)

nlatt 1 1203366 1203366 157.00 <2e-16 \*\*\*

YDS.ATT 1 35033 35033 4.57 0.034 \*

Residuals 199 1525271 7665

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#### [Covariance matrix for model parameters](#_Toc516215857)

> print(vcov(WRMod4)) # covariance matrix for model parameters

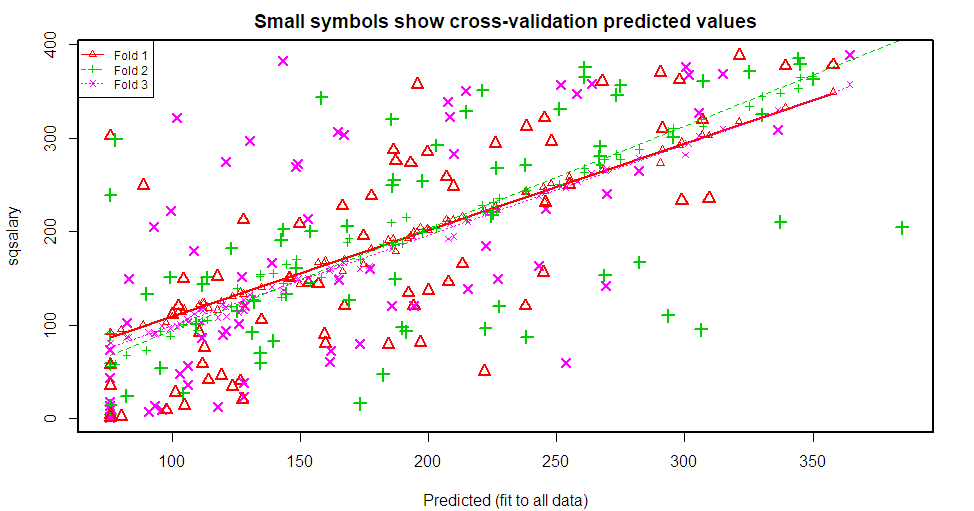
(Intercept) nlatt YDS.ATT

(Intercept) 148.692 -1.352106 -0.394030

nlatt -1.352 0.027587 0.000982

YDS.ATT -0.394 0.000982 0.002725

#### K-fold cross-validation with 3 folds



# Description of the Dataset

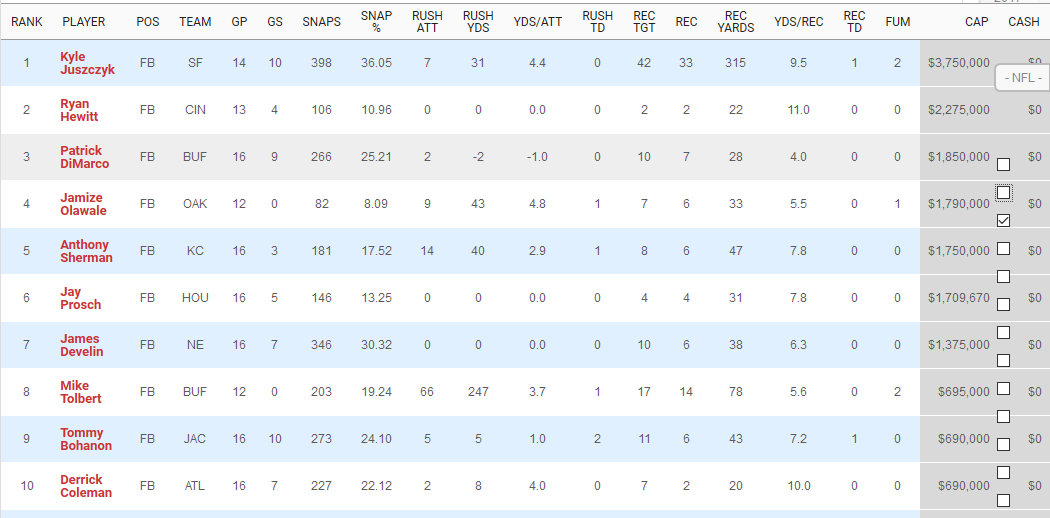
Source Data spotrac website, we paid $30 dollars for access to all NFL and the other 5 major sports in the US statistical information that are relevant for each sport. We validated this information by check other major sports website, do confirm statistical information was correct.



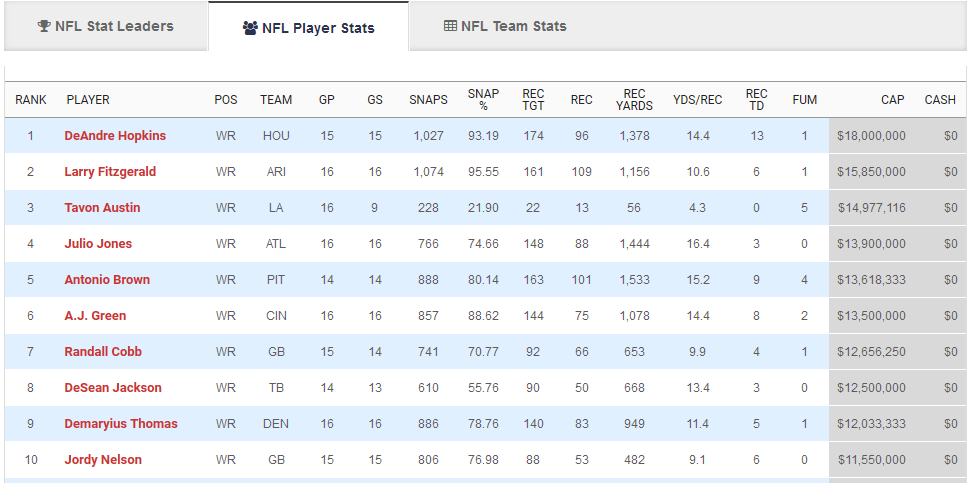
## Primary Source website

* <http://www.spotrac.com> – Main (Landing Page)
* <http://www.spotrac.com/nfl/statistics/> - NFL Stats
* <http://www.spotrac.com/nfl/statistics/player/>

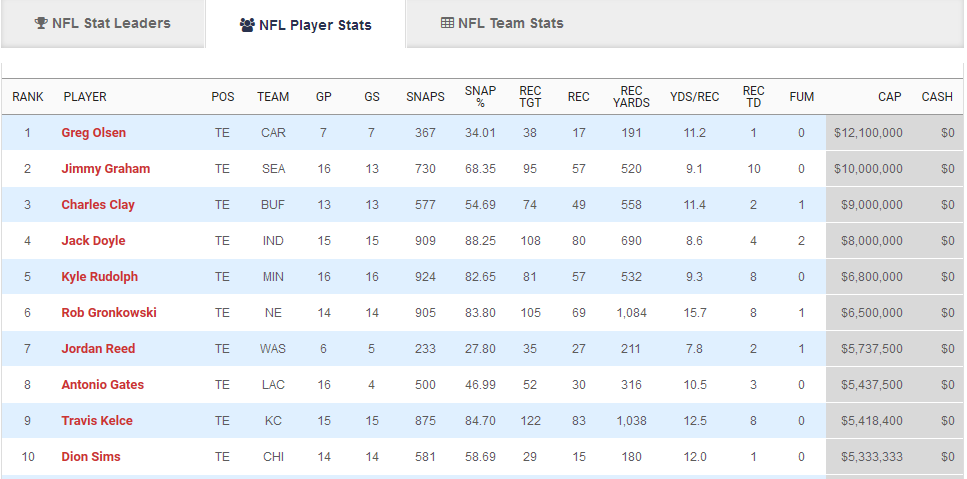
## QB Source Data



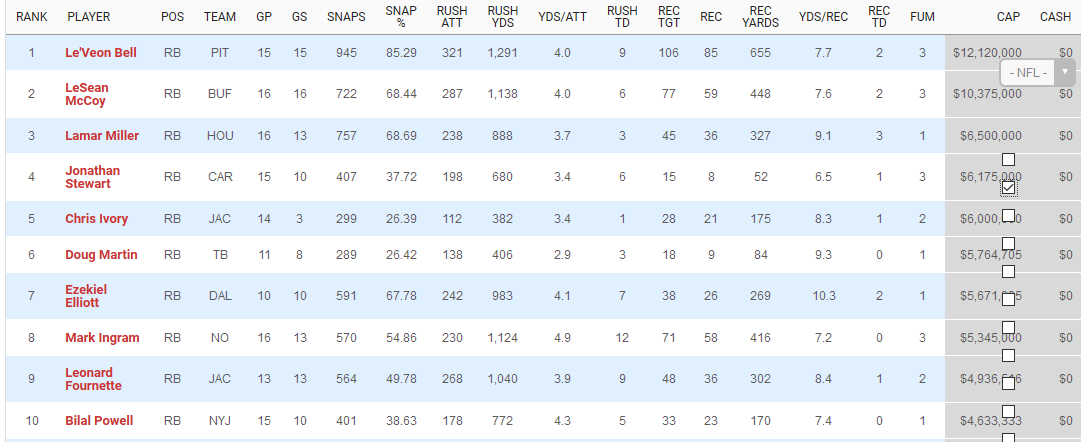
## WR Source Data



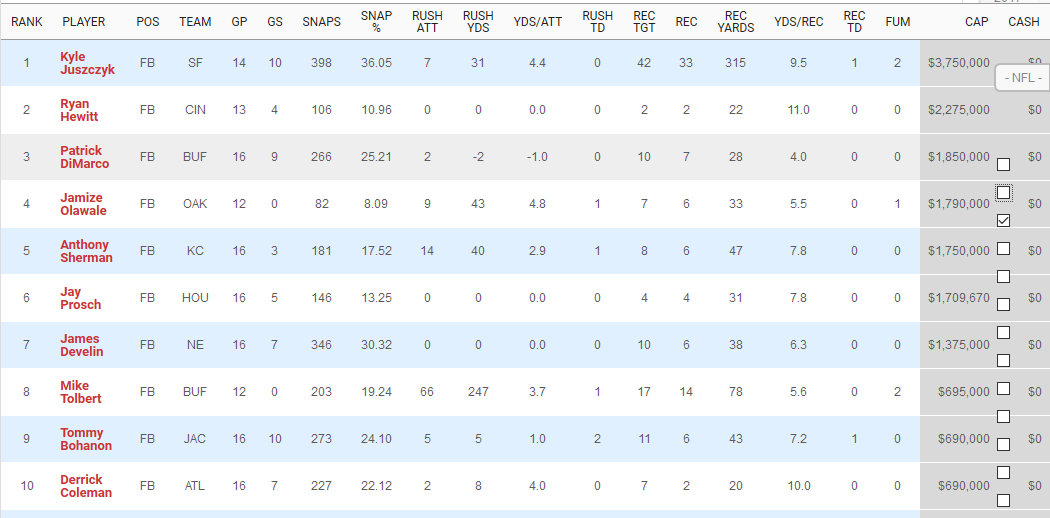
## TE Source Data



## RB Source Data



## FB Source Data



## Data Issue

The only major issue I had with the data was how they tracked statistical information for each position in the Football (NFL), each player depending on what side of the ball and position had different statistical categories for tracking performance on the field. We ended up separating the data into three primary dataset, Offensive players, Defensive Players and special teams. Then I had to break down each category into different subsets again based on position they played on the field. I ended up breaking each category into these sub categories.

* Offense – Skilled Players
  + QB, WR, TE, RB FB
* Offense – Linemen
  + G,LT,RT, C
* Defense – Players
  + OLB, ILB,CB, FS, SS, DT, DE,
* Special Teams
  + K, P

## Data Formatting / Importing

Due to the website not having an export option, we ended up copy and pasting all data into excel. At that point was able to manually create different datasheet based on position and players. I had to format the within excel and create new fields for yards and completion to merge the skilled players statistical yards, and updated the column names to reflect all yards, attempts, completions.

## R Code to import dataset

> library(readxl)

> OffNFLSalary <-

+ read\_excel("NFLSalaryOff.xlsx",

+ sheet = "2017OffensivePlayerSalary\_New")

> OffNFLSalary = data.frame(OffNFLSalary)

> attach(OffNFLSalary)

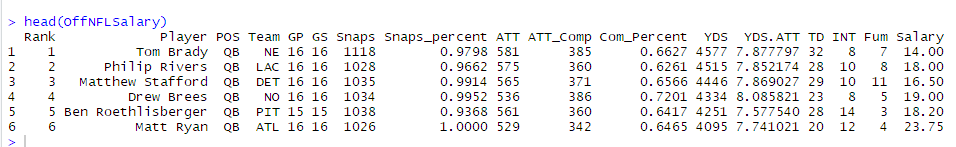
# Convert data set to data.frame#

OffNFLSalary = data.frame(OffNFLSalary)

attach(OffNFLSalary)

head(OffNFLSalary)

> head(OffNFLSalary)

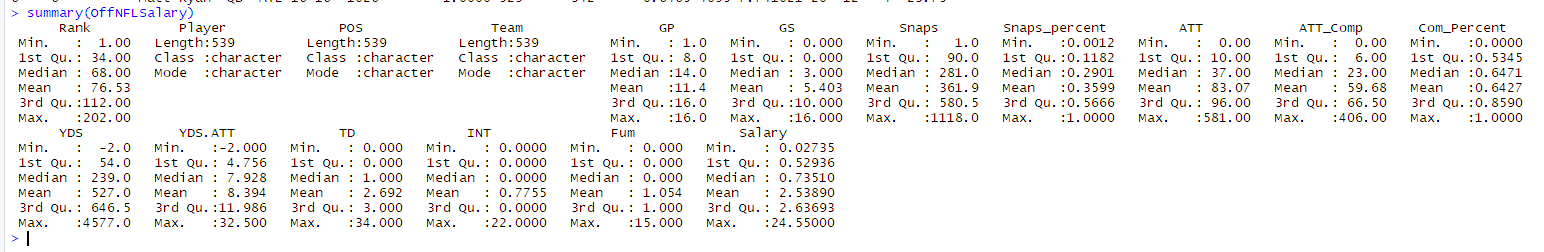


Description of the variables in the dataset

## Description of Dataset Variables:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Rank | NFL players ranking within his position group |
| Player | NFL Players Name |
| POS | The position NFL Player has on the field |
| Team | NFL players sports Team he plays for |
| GP | Games Played |
| GS | Games Started |
| Snaps | Total amount of Snaps |
| Snaps\_percent | Snap Percentage |
| ATT | Total Active plays player is directly involved |
| ATT\_Com | Total Plays completed |
| Com\_Percent | Completion Percentage |
| YDS | Total Yards (passing, receiving, rushing) |
| YDS/ATT | Yards / total attempts |
| TD | Touch Down |
| INT | Interception |
| FUM | Fumble |
| Salary | NFL Salary for 2017 |

### Summary of Data



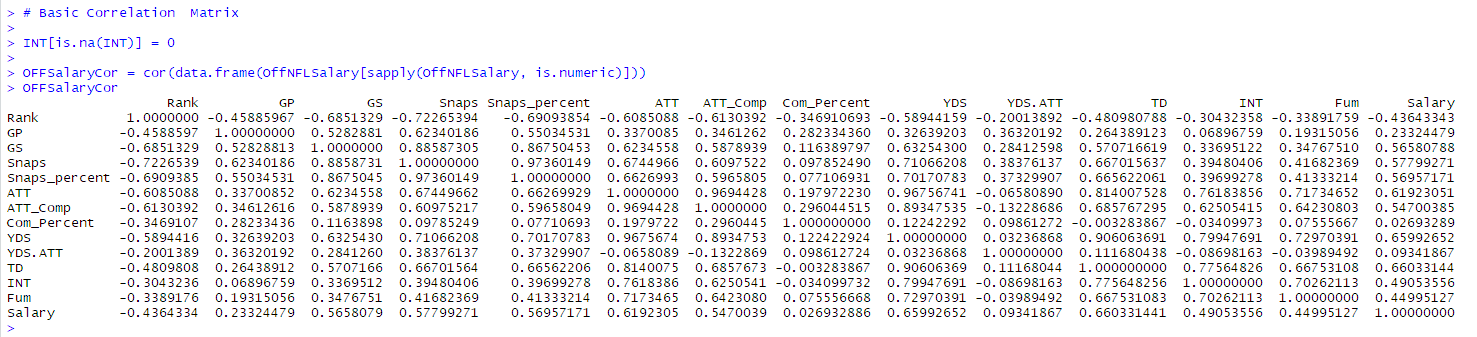
# Pre-Normalization Correlation matrix and Scatterplot

## Basic Correlation Matrix

INT[is.na(INT)] = 0

OFFSalaryCor = cor(data.frame(OffNFLSalary[sapply(OffNFLSalary, is.numeric)]))

OFFSalaryCor



## Basic Scatterplot Matrix

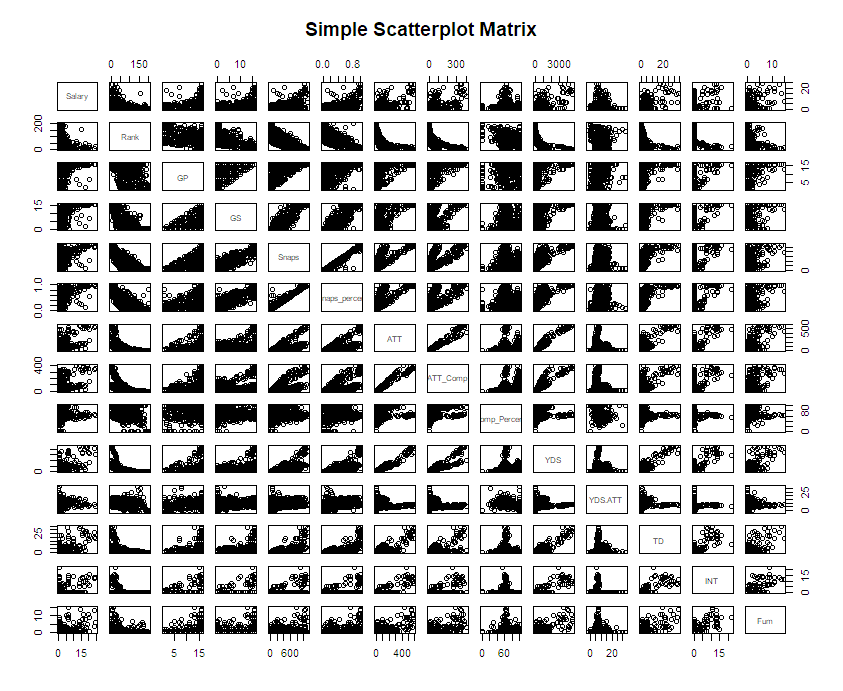
> pairs(

+ ~ Salary + Rank + GP + GS + Snaps + Snaps\_percent + ATT + ATT\_Comp +

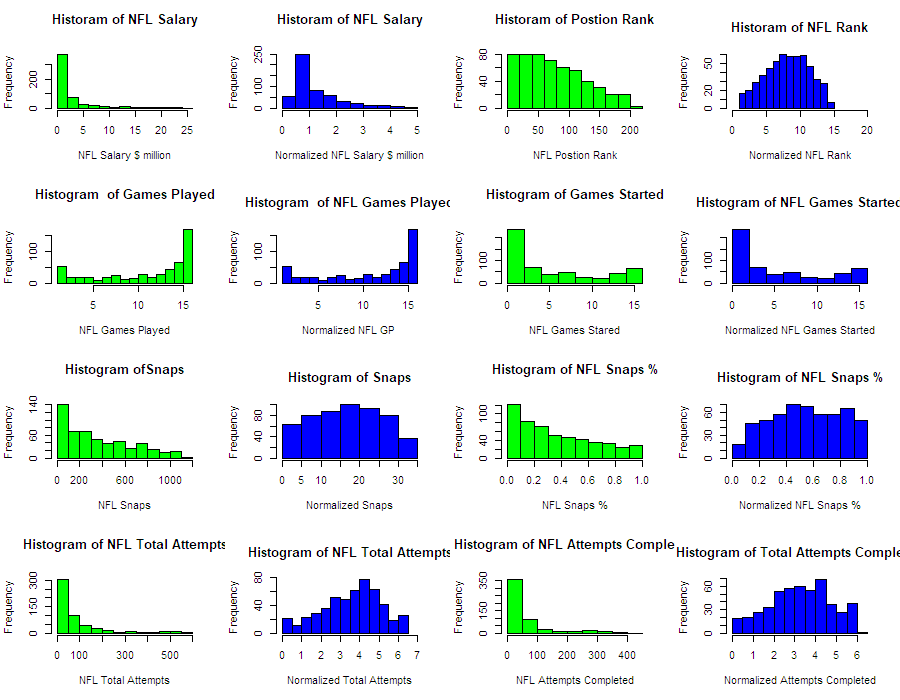
+ Comp\_Percent + YDS + YDS.ATT + TD + INT + Fum,

+ data = OffNFLSalary,

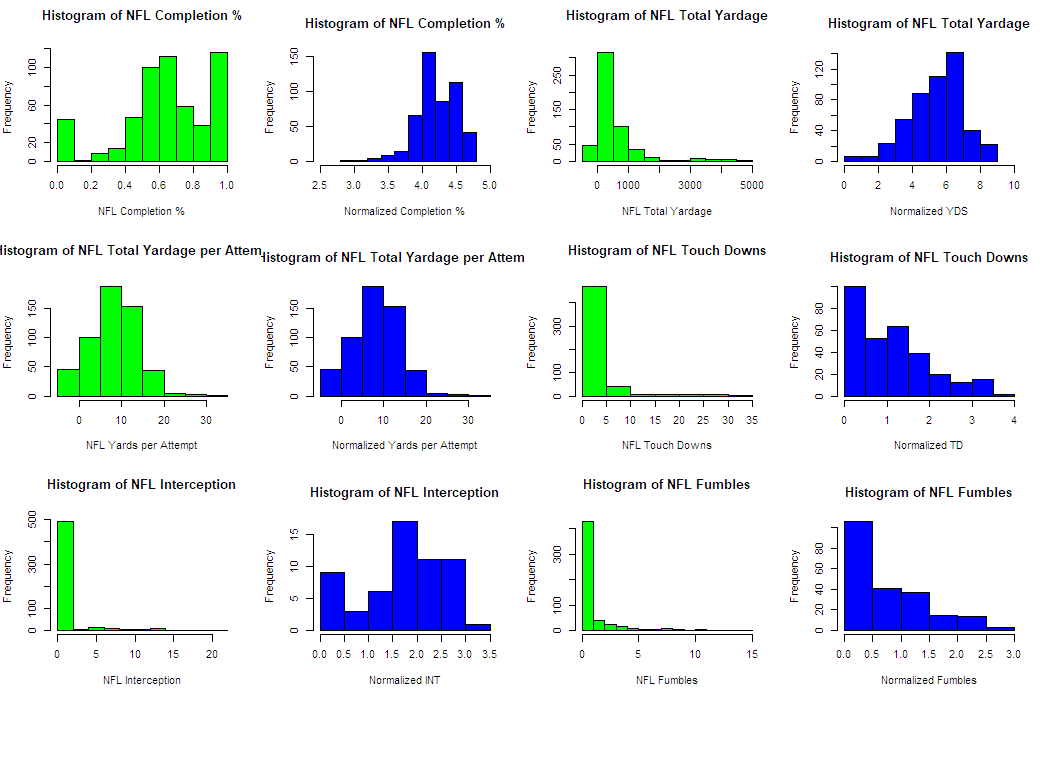
+ main = "Simple Scatterplot Matrix



# Normalizing data

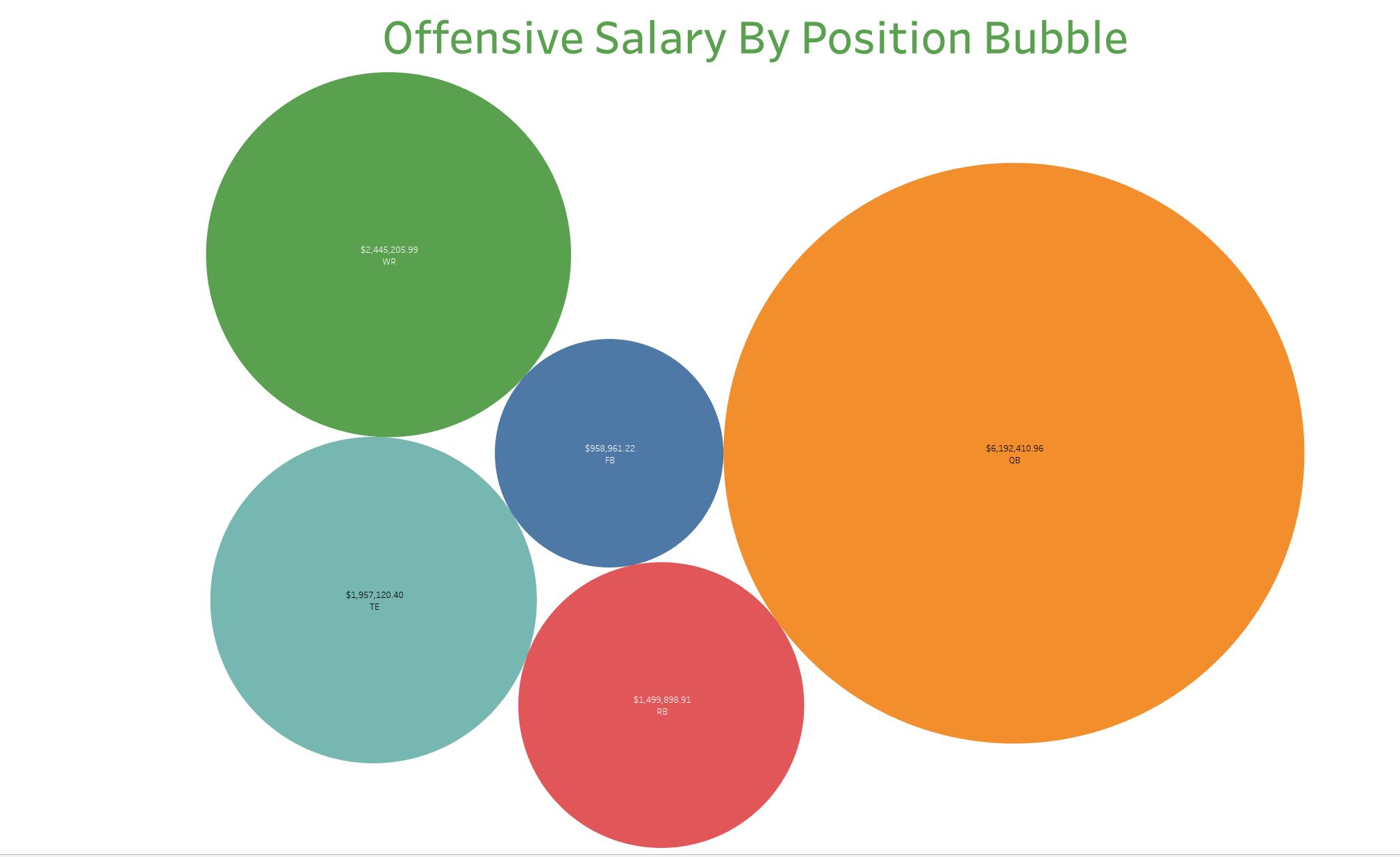


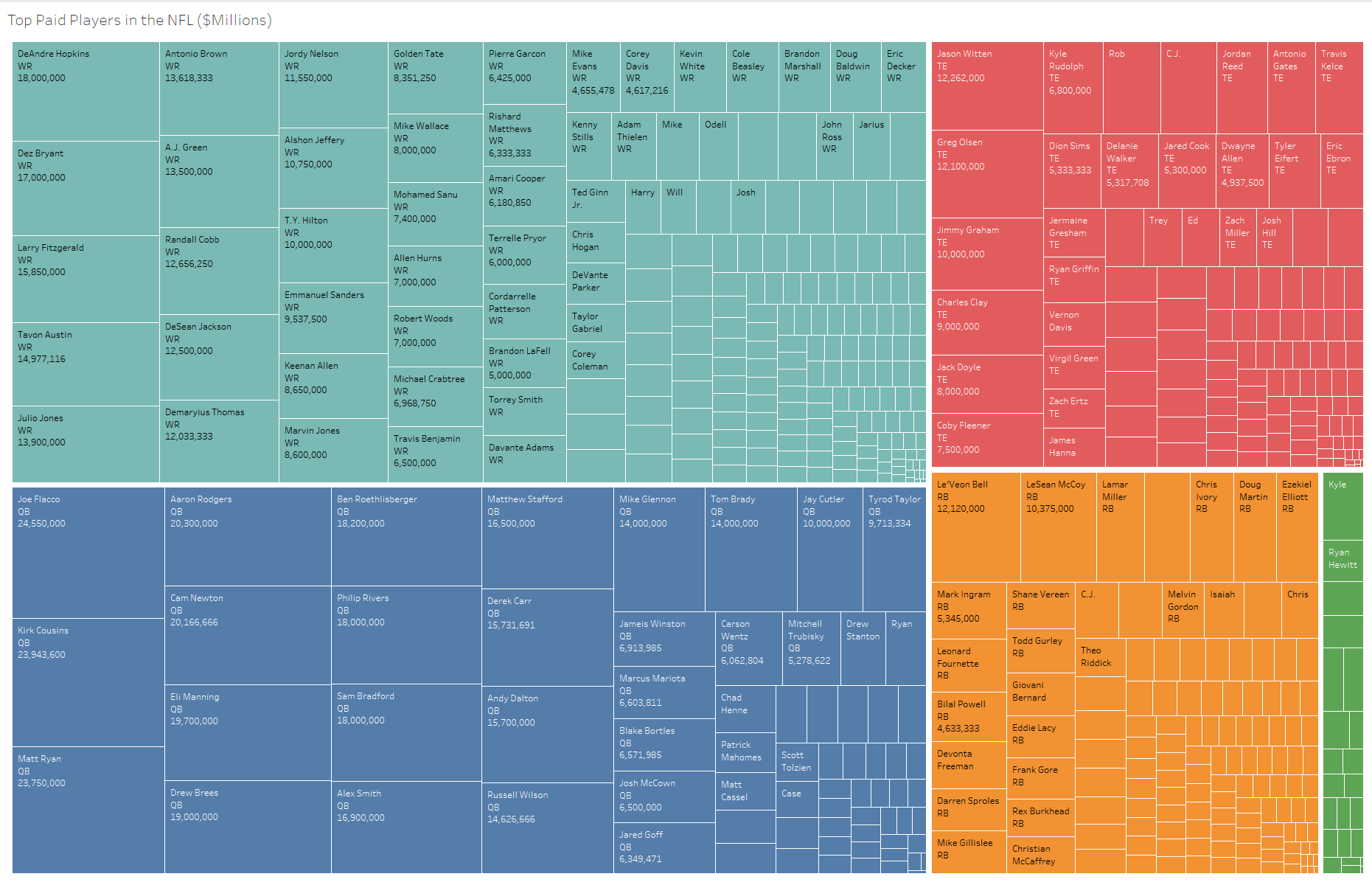
All listed variables used either nl (Natural Log), Log10, sqrt (square root) except games started and games played those variable used source data due to being the most normalized.

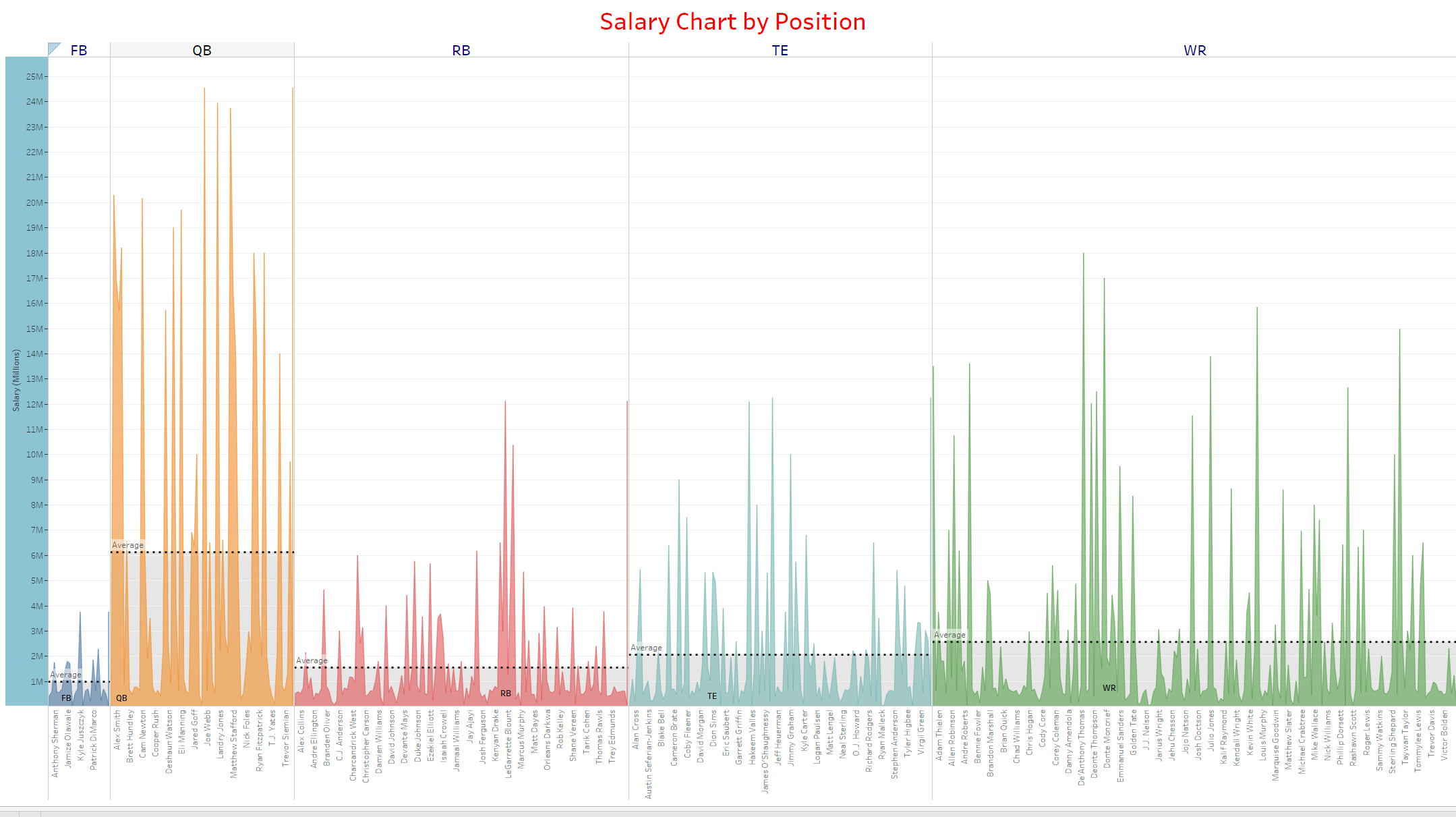


All normalization for following variables used either nl (natural log), Log10, sqrt(square root) to normalize the data.

# Tableau for Visualization







## Problems in the Data

Ran into an issue after normalizing the data we had NaN, -inf within the dataset, Used the following code to clear out the na, NaN, -inf and insert 0.

# removed NA value column

nlatt[is.infinite(nlatt)] = 0

nlattcom[is.infinite(nlattcom)] = 0

nlcompPer[is.infinite(nlcompPer)] = 0

nlyds[is.infinite(nlyds)] = 0

nltd[is.infinite(nltd)] = 0

nlint[is.infinite(nlint)] = 0

nlfum[is.infinite(nlfum)] = 0

nlatt[is.na(nlatt)] = 0

nlattcom[is.na(nlattcom)] = 0

nlcompPer[is.infinite(nlcompPer)] = 0

nlyds[is.na(nlyds)] = 0

nltd[is.na(nltd)] = 0

nlint[is.na(nlint)] = 0

nlfum[is.na(nlfum)] = 0