Gilbert_Nathaniel_329_Final_Project

2025-03-17

In this report we will be analyzing the hitters dataset from the ISLR library, we will be attempting to create a linear regression model in order to predict a hitters salary.

```r
head(Hitters)
```

```
##                    AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Andy Allanson       293   66     1   30  29    14     1    293    66      1
## -Alan Ashby          315   81     7   24  38    39    14   3449   835     69
## -Alvin Davis         479  130    18   66  72    76     3   1624   457     63
## -Andre Dawson        496  141    20   65  78    37    11   5628  1575    225
## -Andres Galarraga    321   87    10   39  42    30     2    396   101     12
## -Alfredo Griffin     594  169     4   74  51    35    11   4408  1133     19
##                    CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Andy Allanson        30   29     14      A        E     446      33     20
## -Alan Ashby          321  414    375      N        W     632      43     10
## -Alvin Davis         224  266    263      A        W     880      82     14
## -Andre Dawson        828  838    354      N        E     200      11      3
## -Andres Galarraga     48   46     33      N        E     805      40      4
## -Alfredo Griffin     501  336    194      A        W     282     421     25
##                    Salary NewLeague
## -Andy Allanson         NA         A
## -Alan Ashby         475.0         N
## -Alvin Davis        480.0         A
## -Andre Dawson       500.0         N
## -Andres Galarraga    91.5         N
## -Alfredo Griffin    750.0         A
```

The dataset features a number of baseball statistics, both for the previous year (1986) and career for a number of baseball playters, it also tells the salary for each player.

Nearly every variable is a discretre numerical varible, with the exception of league, division, and NewLeague which are categorival variables. Aditionally salary is a continuous numerical variable.

```r
is.na(Hitters)
```

The dataset does have null valies, all in the salary column. As that is our response variable we must remove any columns with a null value

```r
hitters= na.omit(Hitters)

str(hitters)
```

```
## 'data.frame':    263 obs. of  20 variables:
```

```
##  $ AtBat    : int  315 479 496 321 594 185 298 323 401 574 ...
##  $ Hits     : int  81 130 141 87 169 37 73 81 92 159 ...
##  $ HmRun    : int  7 18 20 10 4 1 0 6 17 21 ...
##  $ Runs     : int  24 66 65 39 74 23 24 26 49 107 ...
##  $ RBI      : int  38 72 78 42 51 8 24 32 66 75 ...
##  $ Walks    : int  39 76 37 30 35 21 7 8 65 59 ...
##  $ Years    : int  14 3 11 2 11 2 3 2 13 10 ...
##  $ CAtBat   : int  3449 1624 5628 396 4408 214 509 341 5206 4631 ...
##  $ CHits    : int  835 457 1575 101 1133 42 108 86 1332 1300 ...
##  $ CHmRun   : int  69 63 225 12 19 1 0 6 253 90 ...
##  $ CRuns    : int  321 224 828 48 501 30 41 32 784 702 ...
##  $ CRBI     : int  414 266 838 46 336 9 37 34 890 504 ...
##  $ CWalks   : int  375 263 354 33 194 24 12 8 866 488 ...
##  $ League   : Factor w/ 2 levels "A","N": 2 1 2 2 1 2 1 2 1 1 ...
##  $ Division : Factor w/ 2 levels "E","W": 2 2 1 1 2 1 2 2 1 1 ...
##  $ PutOuts  : int  632 880 200 805 282 76 121 143 0 238 ...
##  $ Assists  : int  43 82 11 40 421 127 283 290 0 445 ...
##  $ Errors   : int  10 14 3 4 25 7 9 19 0 22 ...
##  $ Salary   : num  475 480 500 91.5 750 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 2 1 2 2 1 1 1 2 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:59] 1 16 19 23 31 33 37 39 40 42 ...
##   ..- attr(*, "names")= chr [1:59] "-Andy Allanson" "-Billy Beane" "-Bruce Bochte" "-Bob Boone" ...
```

All of the variables are stored as the correct type.

Next we will do some numerical analysis on the data

```
summary(hitters)
```

```
##      AtBat            Hits          HmRun            Runs
##  Min.   : 19.0   Min.   :  1.0   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:282.5   1st Qu.: 71.5   1st Qu.: 5.00   1st Qu.: 33.50
##  Median :413.0   Median :103.0   Median : 9.00   Median : 52.00
##  Mean   :403.6   Mean   :107.8   Mean   :11.62   Mean   : 54.75
##  3rd Qu.:526.0   3rd Qu.:141.5   3rd Qu.:18.00   3rd Qu.: 73.00
##  Max.   :687.0   Max.   :238.0   Max.   :40.00   Max.   :130.00
##       RBI             Walks            Years            CAtBat
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0
##  1st Qu.: 30.00   1st Qu.: 23.00   1st Qu.: 4.000   1st Qu.:  842.5
##  Median : 47.00   Median : 37.00   Median : 6.000   Median : 1931.0
##  Mean   : 51.49   Mean   : 41.11   Mean   : 7.312   Mean   : 2657.5
##  3rd Qu.: 71.00   3rd Qu.: 57.00   3rd Qu.:10.000   3rd Qu.: 3890.5
##  Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0
##      CHits            CHmRun           CRuns             CRBI
##  Min.   :   4.0   Min.   :  0.00   Min.   :   2.0   Min.   :   3.0
##  1st Qu.: 212.0   1st Qu.: 15.00   1st Qu.: 105.5   1st Qu.:  95.0
##  Median : 516.0   Median : 40.00   Median : 250.0   Median : 230.0
##  Mean   : 722.2   Mean   : 69.24   Mean   : 361.2   Mean   : 330.4
##  3rd Qu.:1054.0   3rd Qu.: 92.50   3rd Qu.: 497.5   3rd Qu.: 424.5
##  Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.0
##      CWalks       League  Division    PutOuts          Assists
##  Min.   :   1.0   A:139   E:129    Min.   :   0.0   Min.   :  0.0
##  1st Qu.:  71.0   N:124   W:134    1st Qu.: 113.5   1st Qu.:  8.0
##  Median : 174.0                    Median : 224.0   Median : 45.0
##  Mean   : 260.3                    Mean   : 290.7   Mean   :118.8
##  3rd Qu.: 328.5                    3rd Qu.: 322.5   3rd Qu.:192.0
##  Max.   :1566.0                    Max.   :1377.0   Max.   :492.0
##      Errors          Salary       NewLeague
##  Min.   : 0.000   Min.   :  67.5   A:141
##  1st Qu.: 3.000   1st Qu.: 190.0   N:122
##  Median : 7.000   Median : 425.0
##  Mean   : 8.593   Mean   : 535.9
##  3rd Qu.:13.000   3rd Qu.: 750.0
##  Max.   :32.000   Max.   :2460.0
```

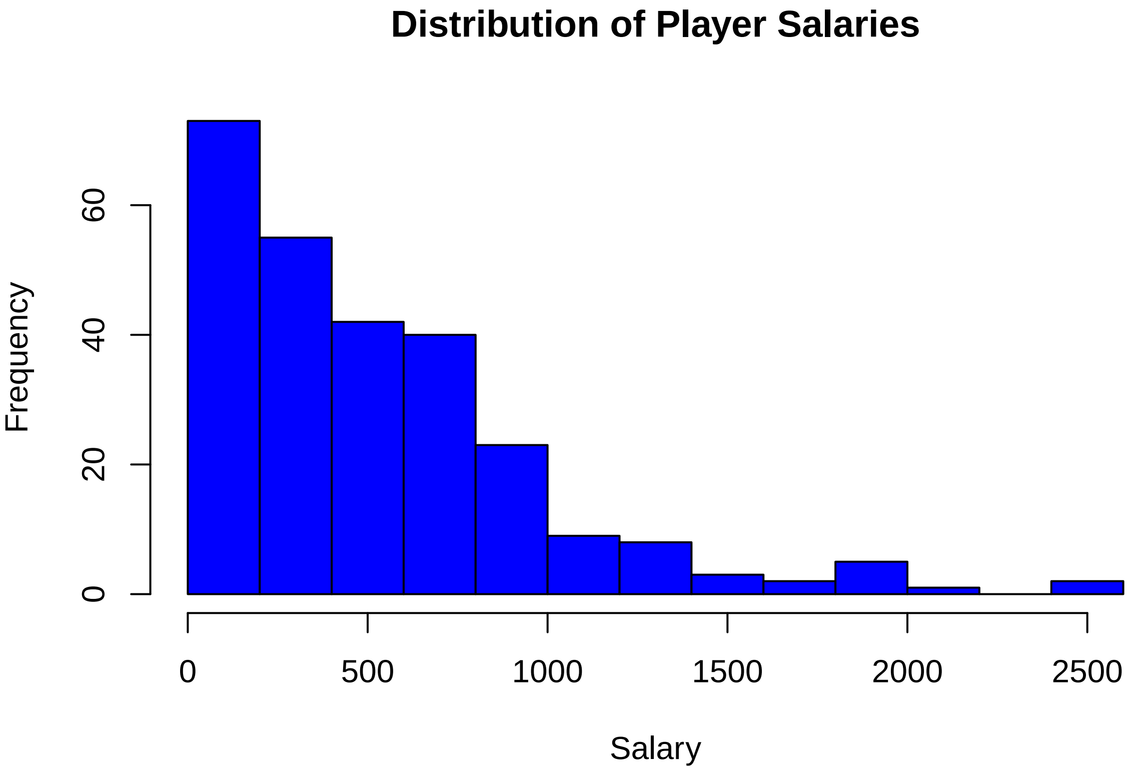
The summary shows us there is a very wide range for salaries, aditionnaly it shows there are roughly equal amount of players in each league and division.

```
numeric_hitters <-hitters[, sapply(hitters, is.numeric)]
cor(numeric_hitters)
```

```
##             AtBat       Hits        HmRun        Runs        RBI     Walks
## AtBat   1.0000000 0.96396913  0.555102154  0.89982910 0.79601539 0.6244481
## Hits    0.9639691 1.00000000  0.530627358  0.91063014 0.78847819 0.5873105
## HmRun   0.5551022 0.53062736  1.000000000  0.63107588 0.84910743 0.4404537
## Runs    0.8998291 0.91063014  0.631075883  1.00000000 0.77869235 0.6970151
## RBI     0.7960154 0.78847819  0.849107434  0.77869235 1.00000000 0.5695048
## Walks   0.6244481 0.58731051  0.440453717  0.69701510 0.56950476 1.0000000
## Years   0.0127255 0.01859809  0.113488420 -0.01197495 0.12966795 0.1347927
## CAtBat  0.2071663 0.20667761  0.217463613  0.17181080 0.27812591 0.2694500
## CHits   0.2253415 0.23560577  0.217495691  0.19132697 0.29213714 0.2707951
## CHmRun  0.2124215 0.18936425  0.492525845  0.22970104 0.44218969 0.3495822
## CRuns   0.2372778 0.23889610  0.258346846  0.23783121 0.30722616 0.3329766
## CRBI    0.2213932 0.21938423  0.349858379  0.20233548 0.38777657 0.3126968
## CWalks  0.1329257 0.12297073  0.227183183  0.16370021 0.23361884 0.4291399
## PutOuts 0.3096075 0.29968754  0.250931497  0.27115986 0.31206456 0.2808555
## Assists 0.3421174 0.30397495 -0.161601753  0.17925786 0.06290174 0.1025226
## Errors  0.3255770 0.27987618 -0.009743082  0.19260879 0.15015469 0.0819372
## Salary  0.3947709 0.43867474  0.343028078  0.41985856 0.44945709 0.4438673
##               Years       CAtBat       CHits      CHmRun       CRuns
## AtBat    0.01272550  0.207166254  0.22534146  0.21242155  0.23727777
## Hits     0.01859809  0.206677608  0.23560577  0.18936425  0.23889610
## HmRun    0.11348842  0.217463613  0.21749569  0.49252584  0.25834685
## Runs    -0.01197495  0.171810798  0.19132697  0.22970104  0.23783121
## RBI      0.12966795  0.278125914  0.29213714  0.44218969  0.30722616
## Walks    0.13479270  0.269449974  0.27079505  0.34958216  0.33297657
## Years    1.00000000  0.915680692  0.89784449  0.72237071  0.87664855
## CAtBat   0.91568069  1.000000000  0.99505681  0.80167609  0.98274694
## CHits    0.89784449  0.995056810  1.00000000  0.78665204  0.98454184
## CHmRun   0.72237071  0.801676089  0.78665204  1.00000000  0.82562483
## CRuns    0.87664855  0.982746941  0.98454184  0.82562483  1.00000000
## CRBI     0.86380936  0.950730141  0.94679739  0.92790264  0.94567701
## CWalks   0.83752373  0.906711655  0.89071842  0.81087827  0.92776846
## PutOuts -0.02001921  0.053392514  0.06734799  0.09382223  0.05908718
## Assists -0.08511772 -0.007897271 -0.01314420 -0.18888646 -0.03889509
## Errors  -0.15651196 -0.070477521 -0.06803583 -0.16536941 -0.09408054
## Salary   0.40065699  0.526135310  0.54890956  0.52493056  0.56267771
##                CRBI      CWalks     PutOuts      Assists       Errors
## AtBat    0.22139318  0.13292568  0.30960746  0.342117377  0.325576978
## Hits     0.21938423  0.12297073  0.29968754  0.303974950  0.279876183
## HmRun    0.34985838  0.22718318  0.25093150 -0.161601753 -0.009743082
## Runs     0.20233548  0.16370021  0.27115986  0.179257859  0.192608787
## RBI      0.38777657  0.23361884  0.31206456  0.062901737  0.150154692
## Walks    0.31269680  0.42913990  0.28085548  0.102522559  0.081937197
## Years    0.86380936  0.83752373 -0.02001921 -0.085117725 -0.156511957
## CAtBat   0.95073014  0.90671165  0.05339251 -0.007897271 -0.070477521
## CHits    0.94679739  0.89071842  0.06734799 -0.013144204 -0.068035829
## CHmRun   0.92790264  0.81087827  0.09382223 -0.188886464 -0.165369407
## CRuns    0.94567701  0.92776846  0.05908718 -0.038895093 -0.094080542
## CRBI     1.00000000  0.88913701  0.09537515 -0.096558877 -0.115316131
## CWalks   0.88913701  1.00000000  0.05816016 -0.066243445 -0.129935875
## PutOuts  0.09537515  0.05816016  1.00000000 -0.043390143  0.075305857
## Assists -0.09655888 -0.06624345 -0.04339014  1.000000000  0.703504693
## Errors  -0.11531613 -0.12993587  0.07530586  0.703504693  1.000000000
## Salary   0.56696569  0.48982204  0.30048036  0.025436136 -0.005400702
##               Salary
## AtBat    0.394770945
## Hits     0.438674738
## HmRun    0.343028078
## Runs     0.419858559
## RBI      0.449457088
## Walks    0.443867260
## Years    0.400656994
## CAtBat   0.526135310
## CHits    0.548909559
## CHmRun   0.524930560
## CRuns    0.562677711
## CRBI     0.566965686
## CWalks   0.489822036
## PutOuts  0.300480356
## Assists  0.025436136
## Errors  -0.005400702
## Salary   1.000000000
```

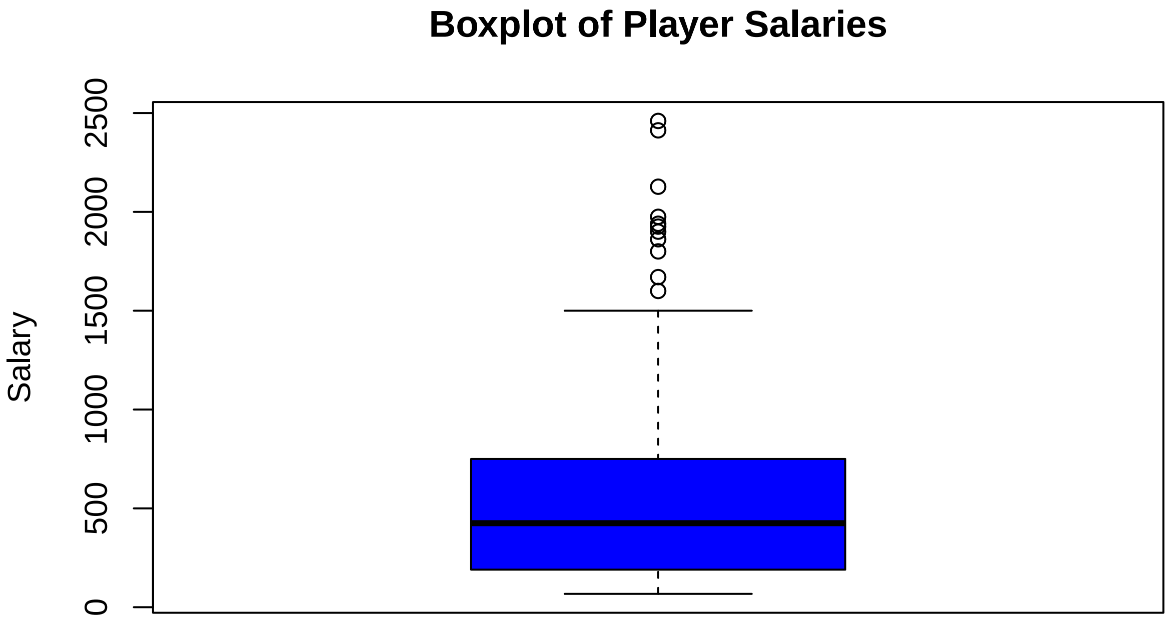
The correlation matrix shows how well all of the numeric variables are correlated. At Bats is very correlated with many offensive totals. Aditionally it seems Salary does not have a super strong correlation with any individual variable. Next we will do some visual analysis

```
hist(hitters$Salary, main="Distribution of Player Salaries", xlab="Salary", col="blue")
```



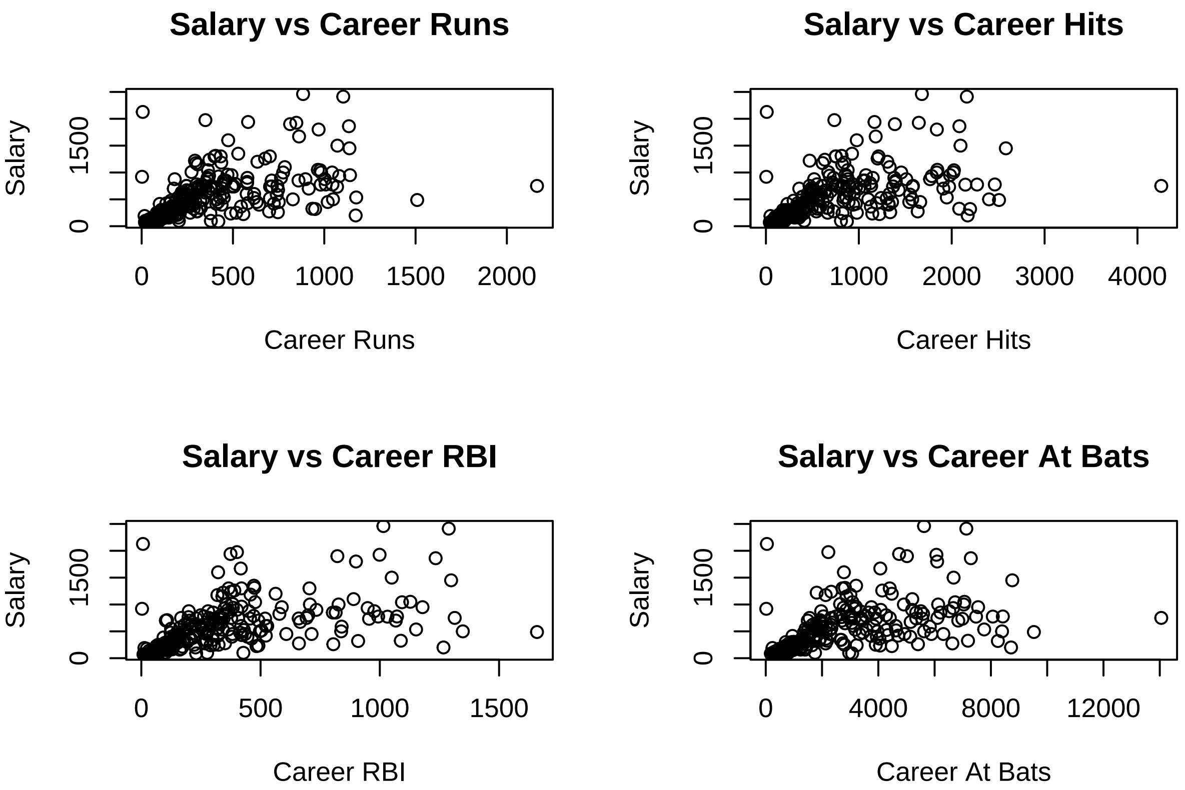
This graph shows that player salaries are heavily skewed right, and a majority of players had a salary under 500,000 dollars.

```
boxplot(hitters$Salary, main="Boxplot of Player Salaries", ylab="Salary", col="blue")
```



In fact, using a boxplot we can see there are an unually large number of outliers on the high end of salaries. We can also see the average is around 500,000 a year.

```
par(mfrow=c(2,2))
plot(hitters$CRuns, hitters$Salary, main="Salary vs Career Runs", xlab="Career Runs", ylab="Salary")
plot(hitters$CHits, hitters$Salary, main="Salary vs Career Hits", xlab="Career Hits", ylab="Salary")
plot(hitters$CRBI, hitters$Salary, main="Salary vs Career RBI", xlab="Career RBI", ylab="Salary")
plot(hitters$CAtBat, hitters$Salary, main="Salary vs Career At Bats", xlab="Career At Bats", ylab="Sala
```



Finally, we plot the salary in comparison to the variables that it had the 4 strongest correlation with. All of these graphs look very similar and clearly show a moderate positive correlation between the variable and salary.

Next, we will be training a regression model in order to predict a hitters salary.

```
model <- lm(Salary~.,data=hitters)
summary(model)
```

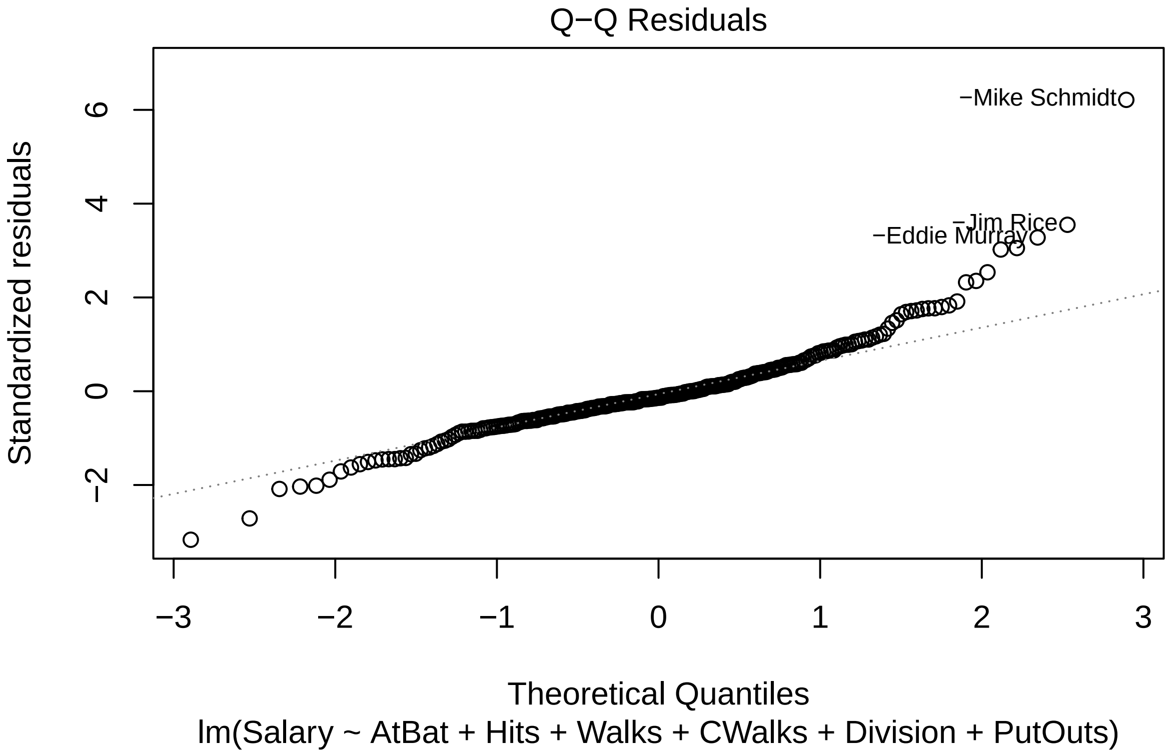
```
##
## Call:
## lm(formula = Salary ~ ., data = hitters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -907.62 -178.35  -31.11  139.09 1877.04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  163.10359   90.77854   1.797 0.073622 .
## AtBat         -1.97987    0.63398  -3.123 0.002008 **
## Hits           7.50077    2.37753   3.155 0.001808 **
## HmRun          4.33088    6.20145   0.698 0.485616
## Runs          -2.37621    2.98076  -0.797 0.426122
## RBI           -1.04496    2.60088  -0.402 0.688204
## Walks          6.23129    1.82850   3.408 0.000766 ***
## Years         -3.48905   12.41219  -0.281 0.778874
## CAtBat        -0.17134    0.13524  -1.267 0.206380
## CHits          0.13399    0.67455   0.199 0.842713
## CHmRun        -0.17286    1.61724  -0.107 0.914967
## CRuns          1.45430    0.75046   1.938 0.053795 .
## CRBI           0.80771    0.69262   1.166 0.244691
## CWalks        -0.81157    0.32808  -2.474 0.014057 *
## LeagueN       62.59942   79.26140   0.790 0.430424
## DivisionW   -116.84925   40.36695  -2.895 0.004141 **
## PutOuts        0.28189    0.07744   3.640 0.000333 ***
## Assists        0.37107    0.22120   1.678 0.094723 .
## Errors        -3.36076    4.39163  -0.765 0.444857
## NewLeagueN   -24.76233   79.00263  -0.313 0.754218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 315.6 on 243 degrees of freedom
## Multiple R-squared:  0.5461, Adjusted R-squared:  0.5106
## F-statistic: 15.39 on 19 and 243 DF,  p-value: < 2.2e-16
```

AtBat, Hits, Walks, CWalks, Division, and PutOuts are all significant at a 95% confidence level so these are the variables we will be using.

```
model <-lm(Salary~AtBat+Hits+Walks+CWalks+Division+PutOuts, data=hitters)
```

Next, we need to check the five assumptuions of linearity, we will do that using the diagnostic plots from our model. We will start with normality.

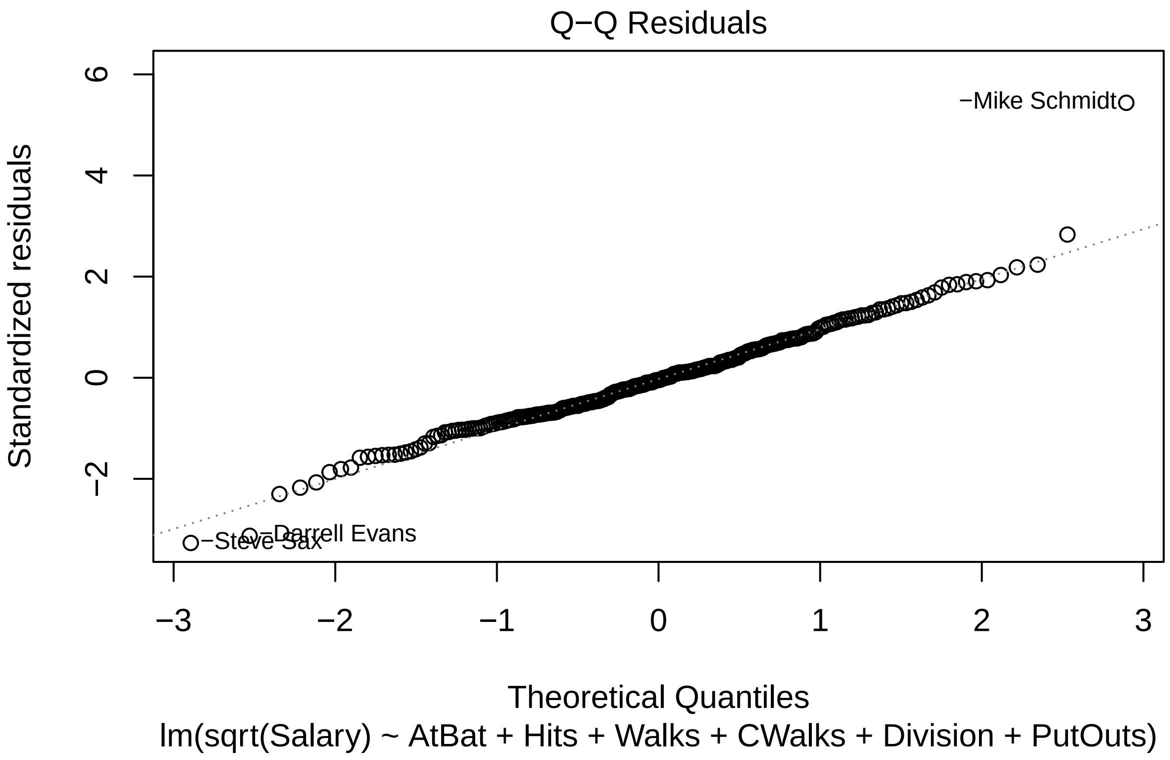
```
plot(model, which=2)
```



From

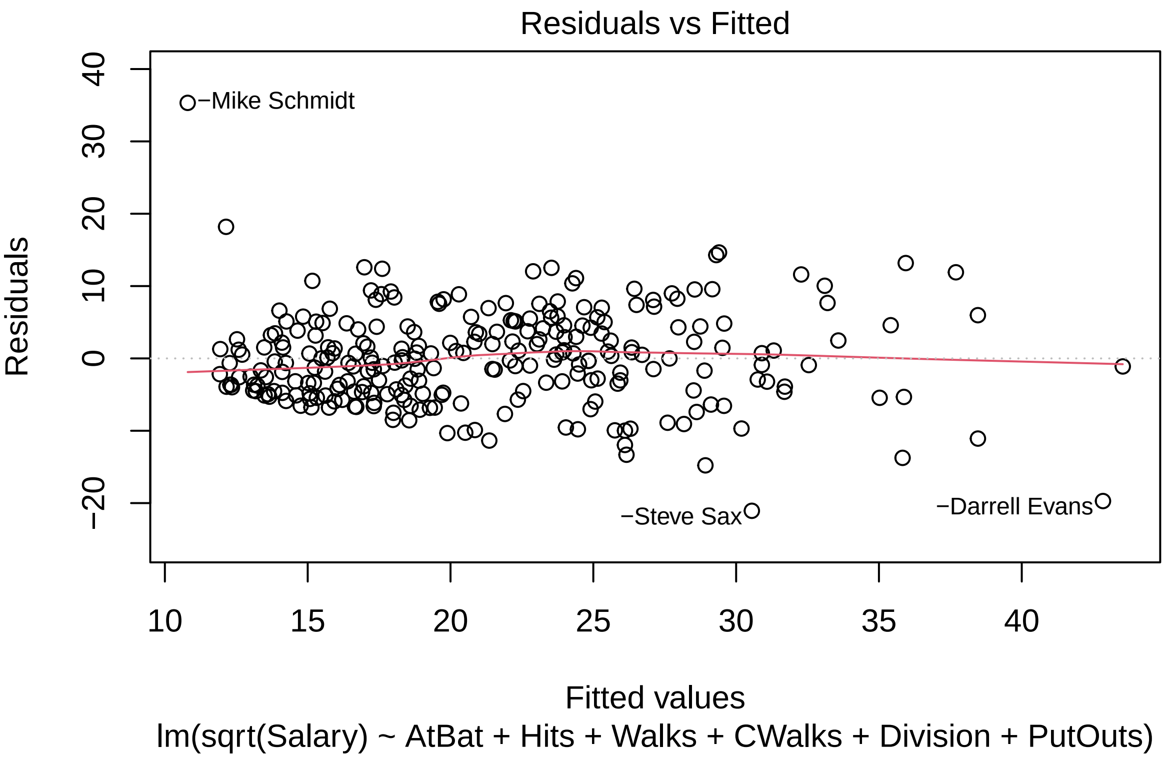
the Q-Q Residual plot we can see that the normality assumption of our model is violated, we can fix this by modifying the response variable.

```
model <-lm(sqrt(Salary)~AtBat+Hits+Walks+CWalks+Division+PutOuts, data=hitters)
plot(model, which=2)
```



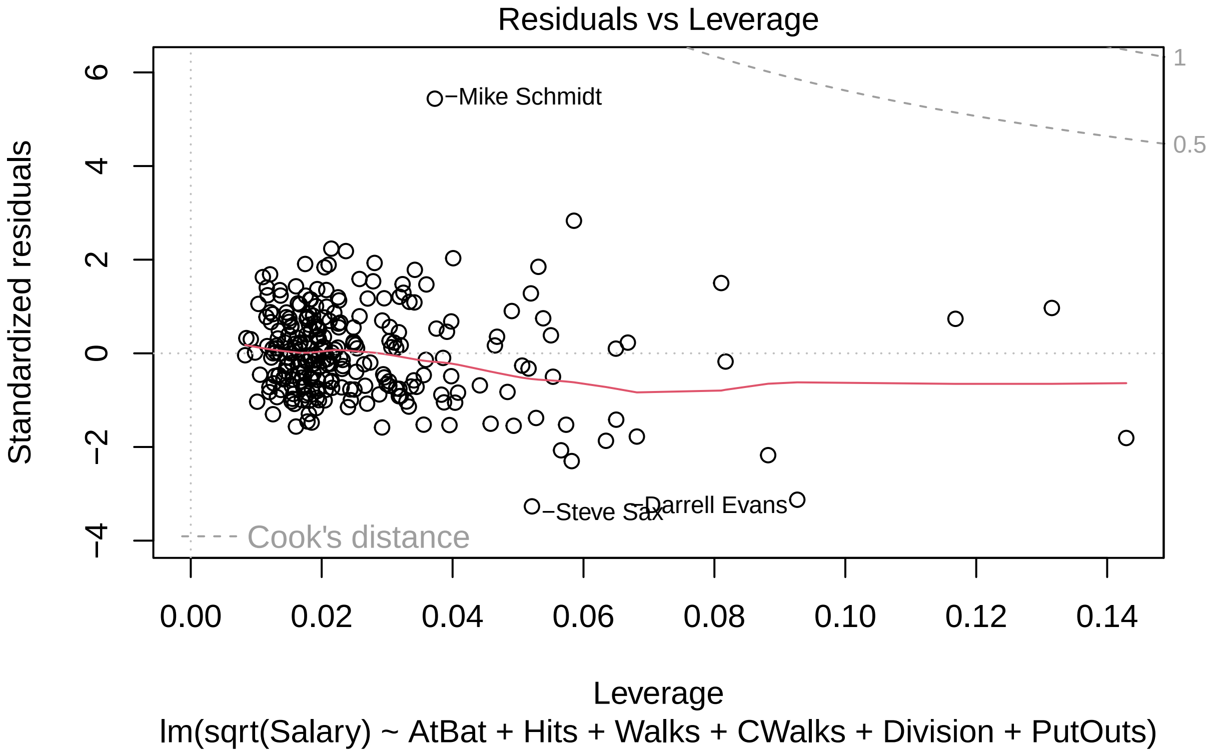
By taking the sqare root of Salary we fix the normality of our data, next we will see if the linearity of the data holds.

```
plot(model, which=1)
```



There does not seem to be a pattern in the residuals and the red line stays mostly flat through zero so our data does not violate the linearity assumption. Next we will check the data for outliers.

```
plot(model, which=5)
```



Mike Schmidt is a massive outlier (as his career stats in the dataset are incorrect), Steve Sax and Darrell Evans are also big outliers and all three should be removed.

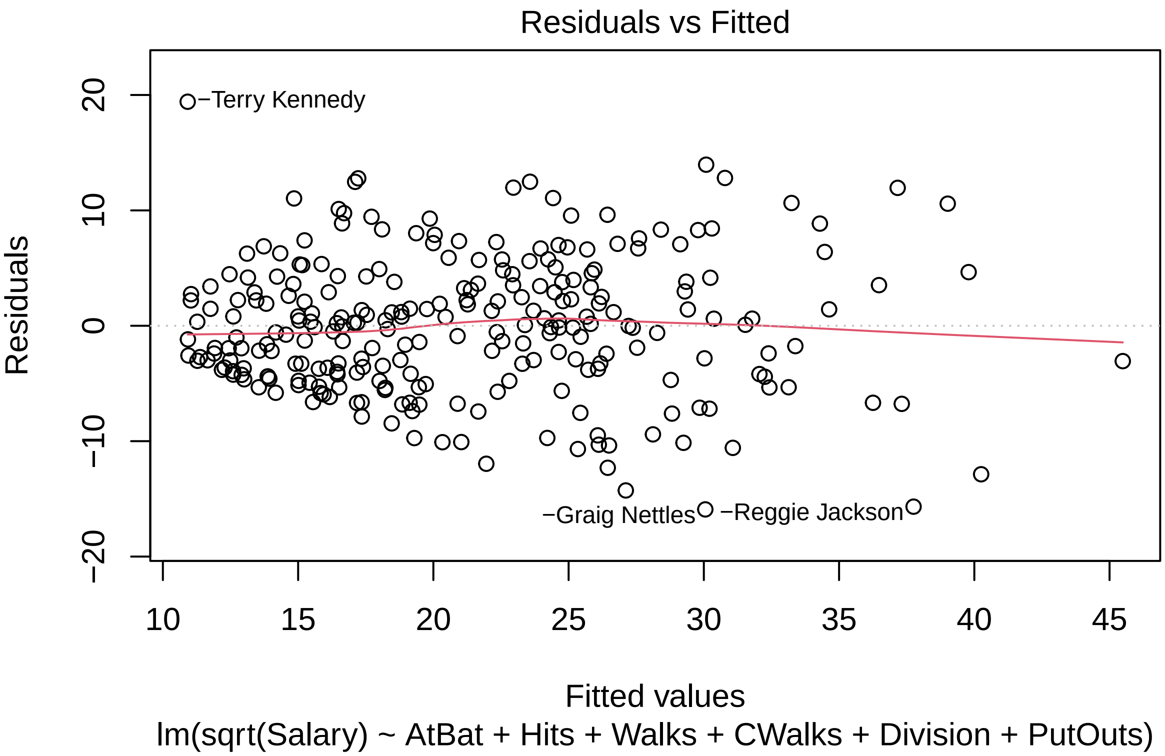
```
Schmidt <- which(hitters$AtBat == 20 & hitters$CRBI == 7)
Sax <-which(hitters$AtBat==633)
Evans <-which(hitters$AtBat==507)
remove <-c(Schmidt, Sax, Evans)
```

Since the dataset does not have row indices and uses player names(which are not numeric and therefore not compatable with slice) I found the index using the which function of the players I want to remove

```
hitters <- hitters %>% slice(-remove)
model <-lm(sqrt(Salary)~AtBat+Hits+Walks+CWalks+Division+PutOuts, data=hitters)
```

Finally, we have to check the homoscedasticity of the model.

```
plot(model, which=1)
```



The Residuals vs Fitted plot show that homoscedasity holds for the regression model.

So our final model is

```
model <-lm(sqrt(Salary)~AtBat+Hits+Walks+CWalks+Division+PutOuts, data=hitters)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = sqrt(Salary) ~ AtBat + Hits + Walks + CWalks + Division +
##     PutOuts, data = hitters)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -15.912  -4.083  -0.034   3.806  19.415 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8.825795   1.252646   7.046 1.74e-11 ***
## AtBat       -0.027482   0.010016  -2.744 0.006509 ** 
## Hits         0.162585   0.031677   5.133 5.70e-07 ***
## Walks        0.018217   0.024561   0.742 0.458953    
## CWalks       0.017788   0.001619  10.990  < 2e-16 ***
## DivisionW   -1.524899   0.749062  -2.036 0.042818 *  
## PutOuts      0.004987   0.001403   3.554 0.000452 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 5.966 on 253 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5608 
## F-statistic: 56.11 on 6 and 253 DF,  p-value: < 2.2e-16
```

With an Adjusted R Squared of 0.5608 show that our model does a moderately good job at correctly predicted player salaries. The intercept is 8.8258, while our coefficients are -0.028, 0.163,0.018, 0.018, -1.525, and 0.005.

Overall, we cleaned out the dataset by omitting null values, then preformed exploratory data analysis to try and find potential relationships within the data and get a better feel for the dataset. We then plotted the distribution of salaries and its relationship with othere variables to help visualize the salary variable. Then we created a regression model. First by finding the variables that had the strongest signifcance, then by checking the regression assumptions and modifying our model to ensure the assumptions are kept. One thing that seemed unusual while doing the project is the statistics for Mike Schmidt were incorrect, but the salary was correct causing him to be a massive outlier, it does call into question how accurate the rest of the dataset was.