This is my final project for Introduction to Scientific Computing, I will be doing a full linear regression analysis on the Hitters dataset from the ISLR package.

Exploratory Analysis:

Overview:

```
#loading necessary libraries
library(ISLR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)

#First look at the data
data(Hitters)
summary(Hitters)
```

```
##      AtBat            Hits         HmRun            Runs
##  Min.   : 16.0   Min.   :  1   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   1st Qu.: 30.25
##  Median :379.5   Median : 96   Median : 8.00   Median : 48.00
##  Mean   :380.9   Mean   :101   Mean   :10.77   Mean   : 50.91
##  3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00   3rd Qu.: 69.00
##  Max.   :687.0   Max.   :238   Max.   :40.00   Max.   :130.00
##
##       RBI             Walks            Years            CAtBat
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0
##  1st Qu.: 28.00   1st Qu.: 22.00   1st Qu.: 4.000   1st Qu.:  816.8
##  Median : 44.00   Median : 35.00   Median : 6.000   Median : 1928.0
##  Mean   : 48.03   Mean   : 38.74   Mean   : 7.444   Mean   : 2648.7
##  3rd Qu.: 64.75   3rd Qu.: 53.00   3rd Qu.:11.000   3rd Qu.: 3924.2
##  Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0
##
##      CHits            CHmRun           CRuns             CRBI
##  Min.   :   4.0   Min.   :  0.00   Min.   :   1.0   Min.   :   0.00
##  1st Qu.: 209.0   1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.:  88.75
##  Median : 508.0   Median : 37.50   Median : 247.0   Median : 220.50
##  Mean   : 717.6   Mean   : 69.49   Mean   : 358.8   Mean   : 330.12
##  3rd Qu.:1059.2   3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25
##  Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.00
##
```

```
##      CWalks        League  Division    PutOuts          Assists
##  Min.   :   0.00   A:175   E:157    Min.   :   0.0   Min.   :  0.0
##  1st Qu.:  67.25   N:147   W:165    1st Qu.: 109.2   1st Qu.:  7.0
##  Median : 170.50                    Median : 212.0   Median : 39.5
##  Mean   : 260.24                    Mean   : 288.9   Mean   :106.9
##  3rd Qu.: 339.25                    3rd Qu.: 325.0   3rd Qu.:166.0
##  Max.   :1566.00                    Max.   :1378.0   Max.   :492.0
##
##      Errors          Salary       NewLeague
##  Min.   : 0.00   Min.   :  67.5   A:176
##  1st Qu.: 3.00   1st Qu.: 190.0   N:146
##  Median : 6.00   Median : 425.0
##  Mean   : 8.04   Mean   : 535.9
##  3rd Qu.:11.00   3rd Qu.: 750.0
##  Max.   :32.00   Max.   :2460.0
##                  NA's   :59
```

```
help(Hitters) #information about the dataset
```

The first thing is looking at and analyzing the data given to us. The data is structured as a data frame with 322 rows of observations of major league players from the 1986 and 1987 seasons. There are 20 variables: AtBat, Hits, HmRun, Runs, RBI, Walks, Years, CAtBat, CHits, CHmRun, CRuns, CRBI, CWalks, League, Division, PutOuts, Assists, Errors, Salary, and NewLeague. AtBat measures the number of times at bat in 1986, Hits measures the number of hits in 1986, HmRun measures the number of home runs in 1986, Runs measures the number of runs in 1986, RBI measures the number of runs batted in in 1986, Walks measures the number of walks in 1986, Years measures the number of years in the major leagues, CAtBat measures the number of times at bat during their career, CHits measures the number of hits during their career, CHmRun measures the number of home runs during their career, CRuns measures the number of runs during their career, CRBI measures the number of runs batted in during their career, CWalks measures the number of walks during their career, League measures the player's league at the end of 1986, Division measures the player's division at the end of 1986, PutOuts measures the number of put outs in 1986, Assists measures the number of assists in 1986, Errors measures the number of errors in 1986, Salary measures the 1987 annual salary on opening day in thousands of dollars, and NewLeague measures the player's league at the beginning of 1987. League, Division, and NewLeague are categorical variables, and the rest are continuous.

Data Cleaning:

```
sum(is.na(Hitters)) #checking for missing values
```

```
## [1] 59
```

```
colSums(is.na(Hitters)) #checking which column they are in
```

```
##     AtBat      Hits     HmRun      Runs       RBI     Walks     Years    CAtBat
##         0         0         0         0         0         0         0         0
##     CHits    CHmRun     CRuns      CRBI    CWalks    League  Division   PutOuts
##         0         0         0         0         0         0         0         0
##   Assists    Errors    Salary NewLeague
##         0         0        59         0
```

The first line of code told me that there are 59 missing values, and the second told me that these are all from the salary column. Usually this may be cause to delete the column and continue with our analysis, however since it is the response variable, I will instead delete the rows with this missing data.

```
Hitters1 <- na.omit(Hitters) #creating a copy of the data set with the null values removed
nrow(Hitters1)
```

```
## [1] 263
```

Next, we have to make sure all the data is in the correct form and if not we have to fix them.

```
sapply(Hitters1, class) #finding the class of each column
```

```
##     AtBat      Hits     HmRun      Runs       RBI     Walks     Years    CAtBat
## "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
##     CHits    CHmRun     CRuns      CRBI    CWalks    League  Division   PutOuts
## "integer" "integer" "integer" "integer" "integer"  "factor"  "factor" "integer"
##   Assists    Errors    Salary NewLeague
## "integer" "integer" "numeric"  "factor"
```

What needs to get changed is integer should be numeric to account for any non whole numbers in the data.

```
Hitters1 <- Hitters1 %>% #changing integer to numeric for the necessary columns
mutate(across(c(AtBat, Hits, HmRun, Runs, RBI, Walks, Years, CAtBat, CHits, CHmRun, CRuns, CRBI, CWalks
sapply(Hitters1, class) #checking to make sure it worked
```

```
##     AtBat      Hits     HmRun      Runs       RBI     Walks     Years    CAtBat
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##     CHits    CHmRun     CRuns      CRBI    CWalks    League  Division   PutOuts
## "numeric" "numeric" "numeric" "numeric" "numeric"  "factor"  "factor" "numeric"
##   Assists    Errors    Salary NewLeague
## "numeric" "numeric" "numeric"  "factor"
```

Now that our data is all cleaned we can do a full numeric analysis on it. Starting off with a summary of all the numeric data.

Numeric Analysis:

```
numeric_vars <- select(Hitters1, where(is.numeric)) #creating a subset of only the numeric variables
summary(numeric_vars) #gathering info
```

```
##      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          PutOuts          Assists          Errors
##  Min.   :   1.0   Min.   :   0.0   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:  71.0   1st Qu.: 113.5   1st Qu.:  8.0   1st Qu.: 3.000
##  Median : 174.0   Median : 224.0   Median : 45.0   Median : 7.000
##  Mean   : 260.3   Mean   : 290.7   Mean   :118.8   Mean   : 8.593
##  3rd Qu.: 328.5   3rd Qu.: 322.5   3rd Qu.:192.0   3rd Qu.:13.000
##  Max.   :1566.0   Max.   :1377.0   Max.   :492.0   Max.   :32.000
##      Salary
##  Min.   :  67.5
##  1st Qu.: 190.0
##  Median : 425.0
##  Mean   : 535.9
##  3rd Qu.: 750.0
##  Max.   :2460.0
```

Next, the range and spread (aka standard deviation and variance) of the data.

```
apply(numeric_vars, 2, function(x) c(Range = max(x) - min(x), SD = sd(x), Variance = var(x)))
```

```
##                AtBat       Hits     HmRun      Runs       RBI     Walks
## Range      668.0000  237.00000 40.000000 130.00000 121.00000 105.00000
## SD         147.3072   45.12533  8.757108  25.53982  25.88271  21.71806
## Variance 21699.4138 2036.29504 76.686936 652.28218 669.91490 471.67396
##                Years      CAtBat       CHits     CHmRun       CRuns        CRBI
## Range     23.000000   14034.000   4252.0000  548.00000   2163.0000   1656.0000
## SD         4.793616    2286.583    648.1996   82.19758    331.1986    323.3677
## Variance 22.978754 5228461.493 420162.7781 6756.44240 109692.4932 104566.6488
##               CWalks    PutOuts    Assists    Errors      Salary
## Range     1565.0000  1377.0000   492.0000 32.000000   2392.5000
## SD         264.0559   279.9346   145.0806  6.606574    451.1187
## Variance 69725.5014 78363.3666 21048.3737 43.646823 203508.0641
```

Then a correlation analysis to see which variables have the most to do with our response variable Salary

```
cor_salary <- cor(numeric_vars$Salary, numeric_vars)
cor_salary
```

```
##           AtBat      Hits     HmRun      Runs       RBI     Walks    Years
## [1,] 0.3947709 0.4386747 0.3430281 0.4198586 0.4494571 0.4438673 0.400657
##          CAtBat     CHits    CHmRun     CRuns      CRBI   CWalks   PutOuts
## [1,] 0.5261353 0.5489096 0.5249306 0.5626777 0.5669657 0.489822 0.3004804
##          Assists       Errors Salary
## [1,] 0.02543614 -0.005400702      1
```

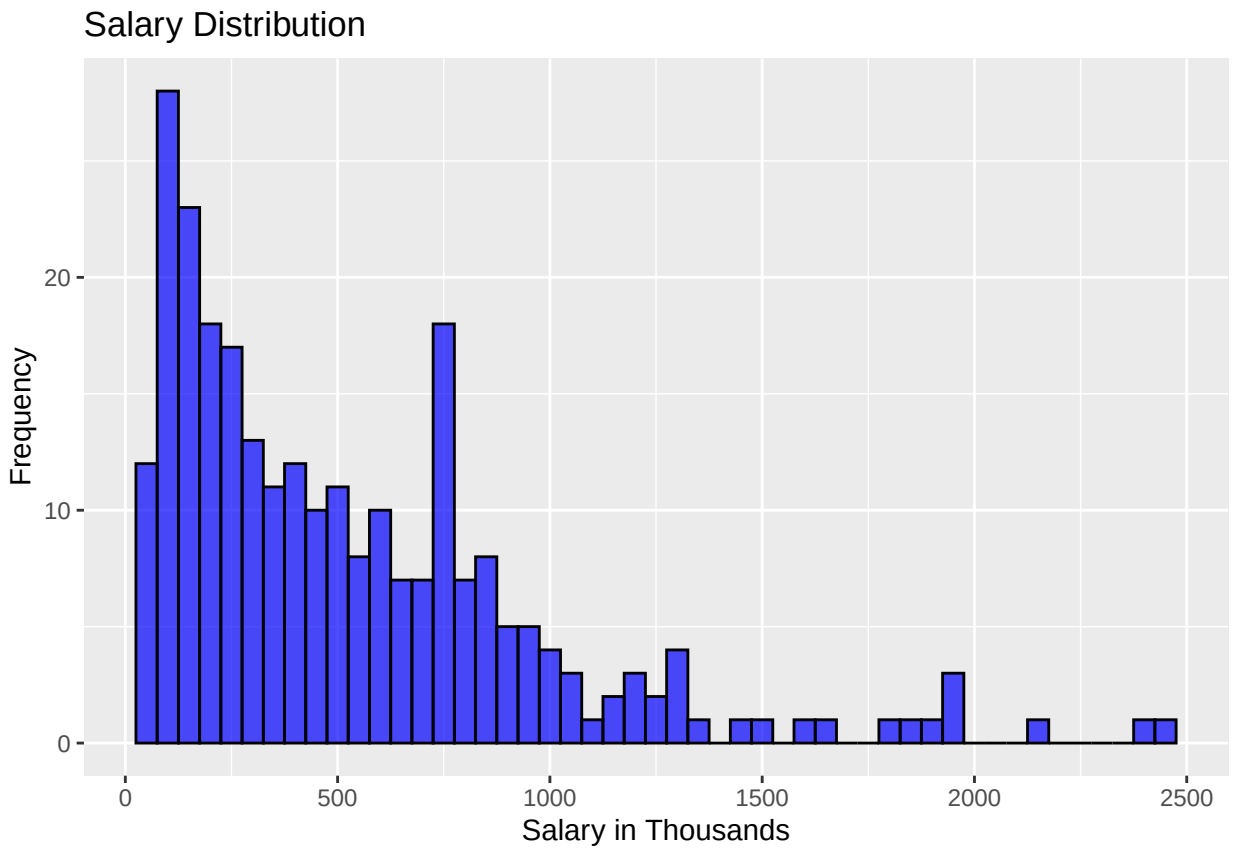
Here we can see that the most correlated variables with Salary are CRBI, CRuns, CHits, CAtBat, and CHmRun, which shows that career statistics are strong predictors of Salary. In contrast, Errors, Assists, and

PutOuts are the least correlated variables, suggesting that these variables are not meaningful predictors of Salary.

Next, we are going to do some visualizations. Using our response variable salary, we are going to make a few plots. The first is a histogram for salary.

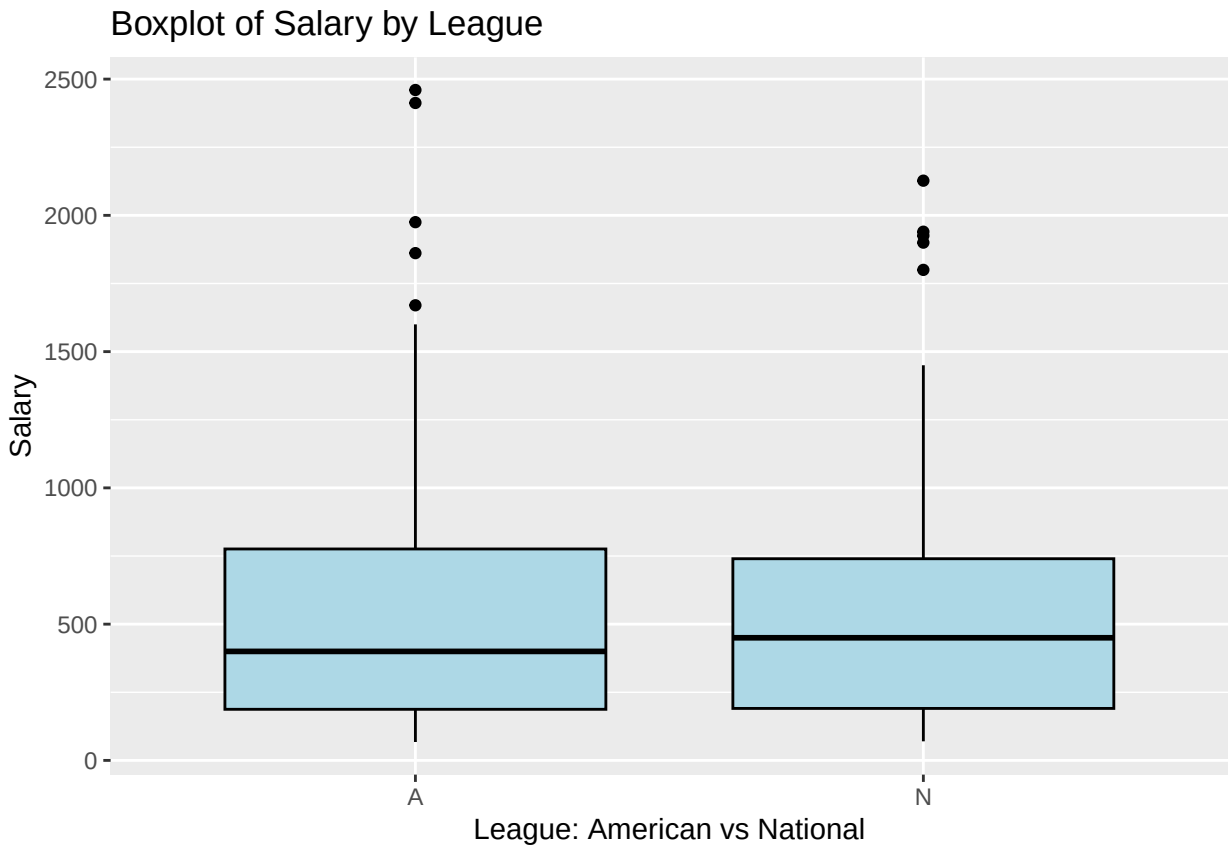
Visual Analysis:

```
ggplot(Hitters1, aes(x = Salary)) +
  geom_histogram(binwidth = 50, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Salary Distribution", x = "Salary in Thousands", y = "Frequency")
```



We can see here that most players made in the 0 to 1,000,000 million range for the 1987 season.

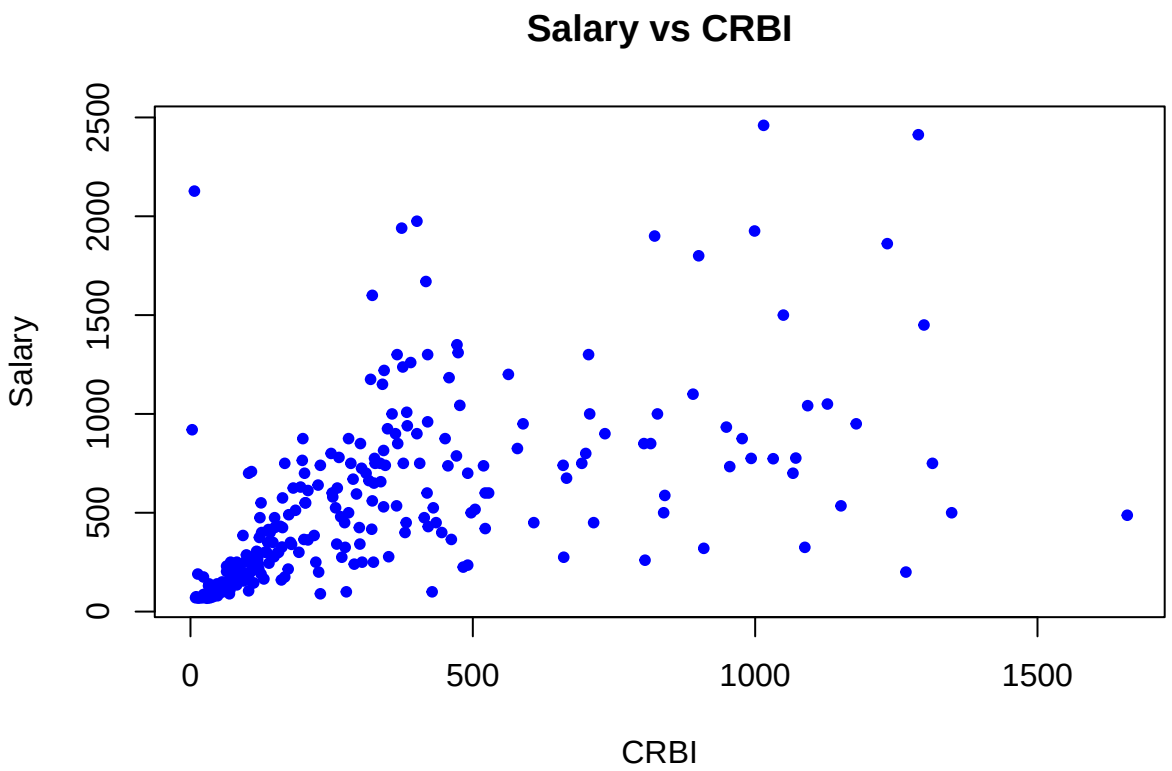
```
ggplot(Hitters1, aes(x = League, y = Salary)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Boxplot of Salary by League", x = "League: American vs National", y = "Salary")
```



Here we can see that while the American league had a larger spread of salaries, the National League had a higher median salary.

Next, a scatter plot between the most correlated variable to salary.

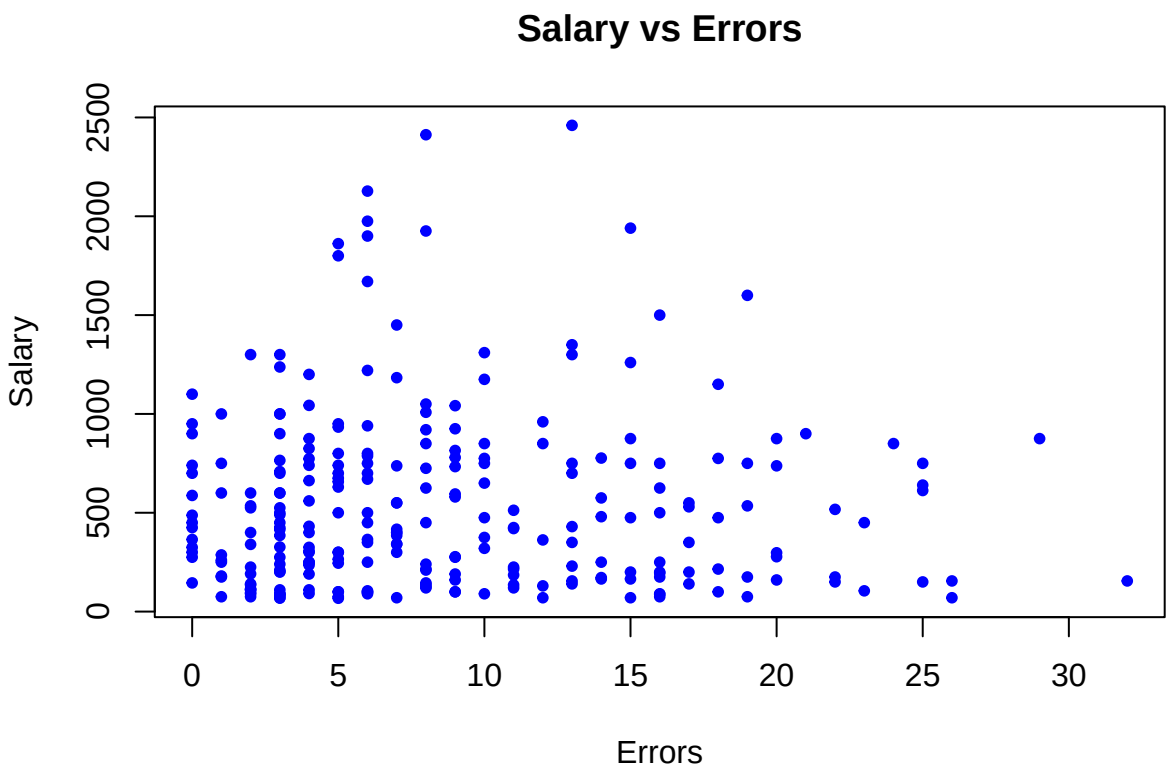
```
plot(Hitters1$CRBI, Hitters1$Salary,
     main = "Salary vs CRBI",
     xlab = "CRBI",
     ylab = "Salary",
     pch = 20,
     col = "blue")
```



Here we can a positive correlation between the number of runs batted in during his career and the salary they recieved. Showcasing that a higher number of runs batted means a higher salary.

Next, a scatterplot between the least correlated variable to salary.

```
plot(Hitters1$Errors, Hitters1$Salary,
     main = "Salary vs Errors",
     xlab = "Errors",
     ylab = "Salary",
     pch = 20,
     col = "blue")
```

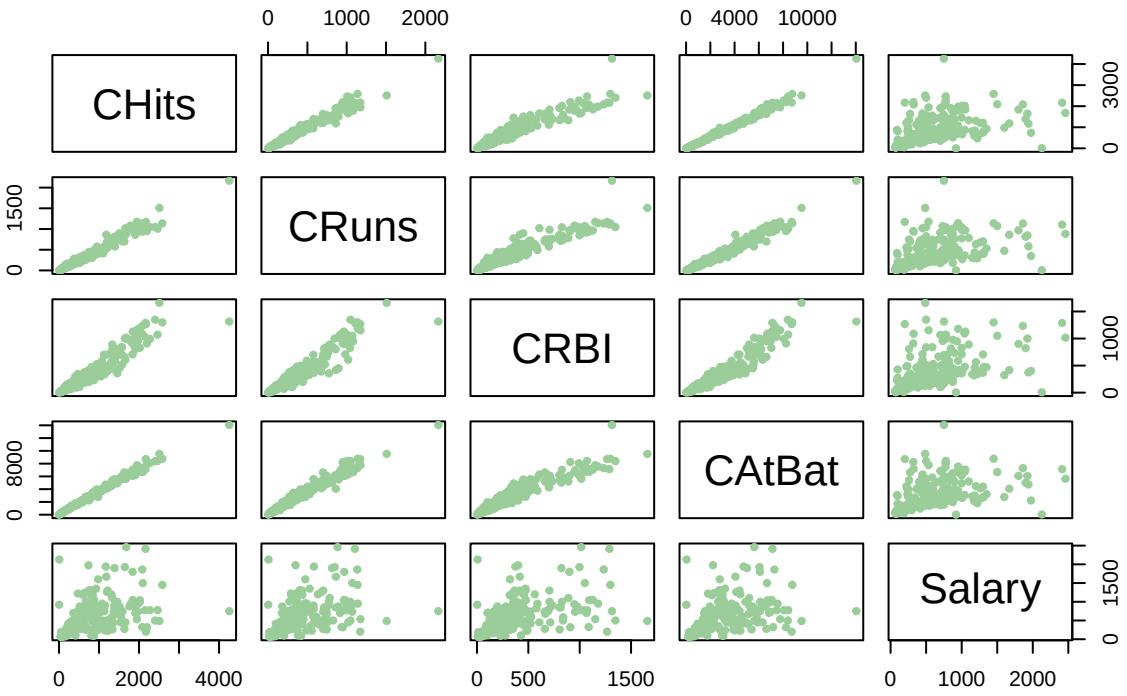


Here we can see there is not correlation between the two values showcasing that error has very little impact on salary.

Finally, a pairplot with the most correlated variables to salary.

```
selected_vars <- Hitters1[, c("CHits", "CRuns", "CRBI", "CAtBat", "Salary")]

pairs(selected_vars,
      main = "Pairwise Plot: CHits, CRuns, CRBI, CAtBat vs Salary",
      col = "darkseagreen3",
      pch = 20,
      labels = c("CHits", "CRuns", "CRBI", "CAtBat", "Salary"))
```



Here we can see that most of the graphs show a positive correlation meaning as they increase, so does the predicted salary of the player.

Regression Analysis:

Now onto the regression analysis. First we are going to decide which predictors to use in our model.

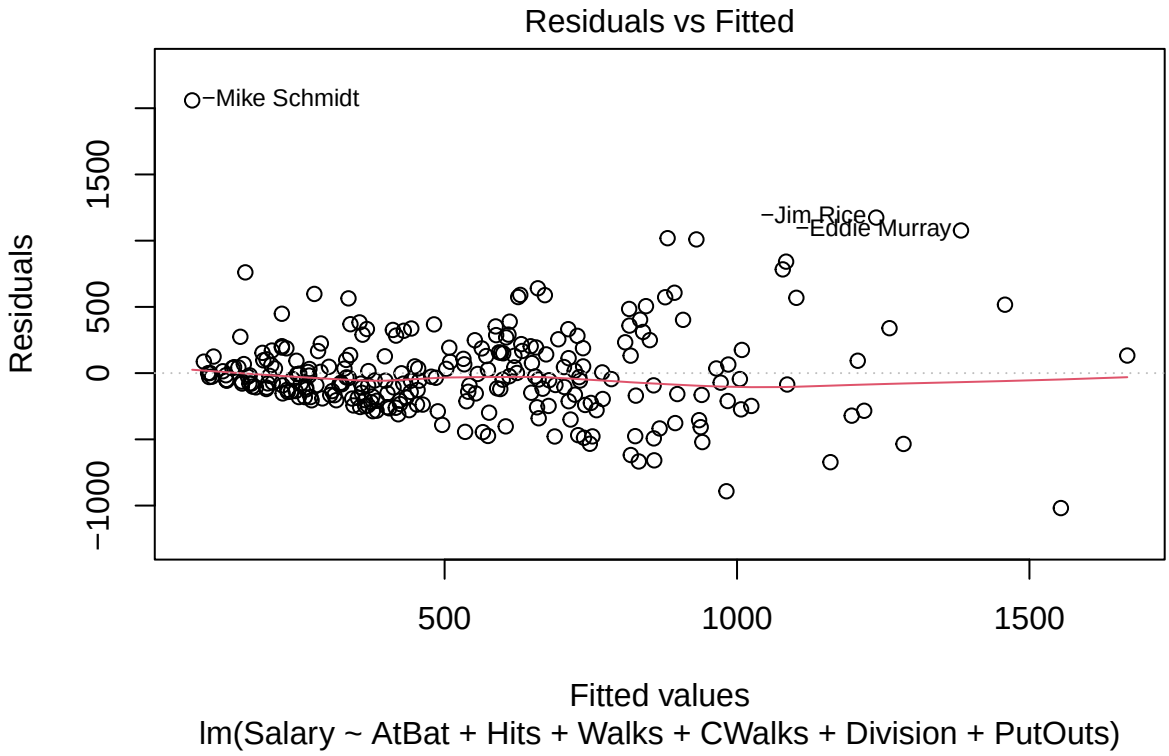
```
full_model <- lm(Salary ~ ., data = Hitters1)
summary(full_model)
```

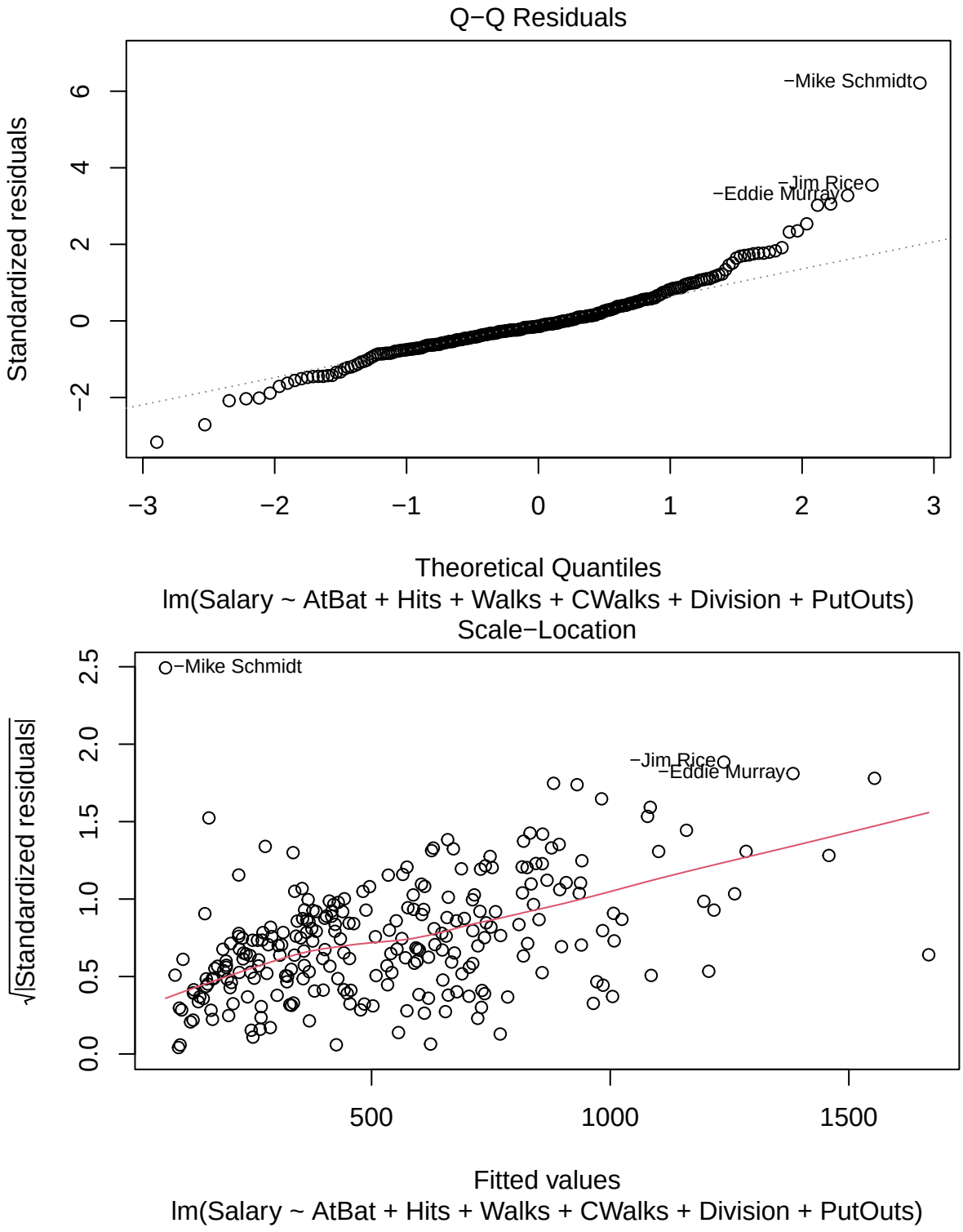
```
##
## Call:
## lm(formula = Salary ~ ., data = Hitters1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -907.62 -178.35  -31.11  139.09 1877.04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  163.10359   90.77854   1.797 0.073622 .
## AtBat         -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
```

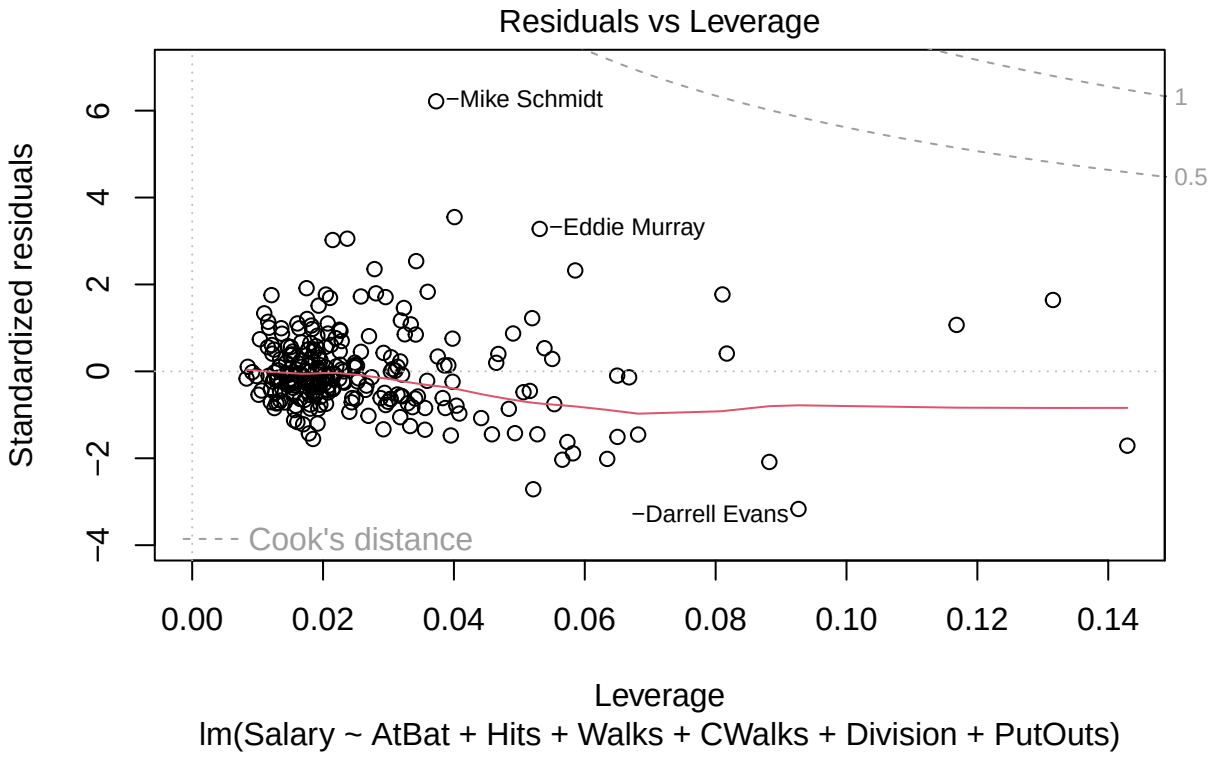
```
## CHmRun        -0.17286    1.61724  -0.107 0.914967
## CRuns          1.45430    0.75046   1.938 0.053795 .
## CRBI           0.80771    0.69262   1.166 0.244691
## CWalks        -0.81157    0.32808  -2.474 0.014057 *
## LeagueN       62.59942   79.26140   0.790 0.430424
## DivisionW   -116.84925   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
```

Based on p-values, we can see that the significant variables are AtBat, Hits, Walks, CWalks, Division, PutOuts.

```
model <- lm(Salary ~ AtBat + Hits + Walks + CWalks + Division + PutOuts, data = Hitters1)
plot(model)
```



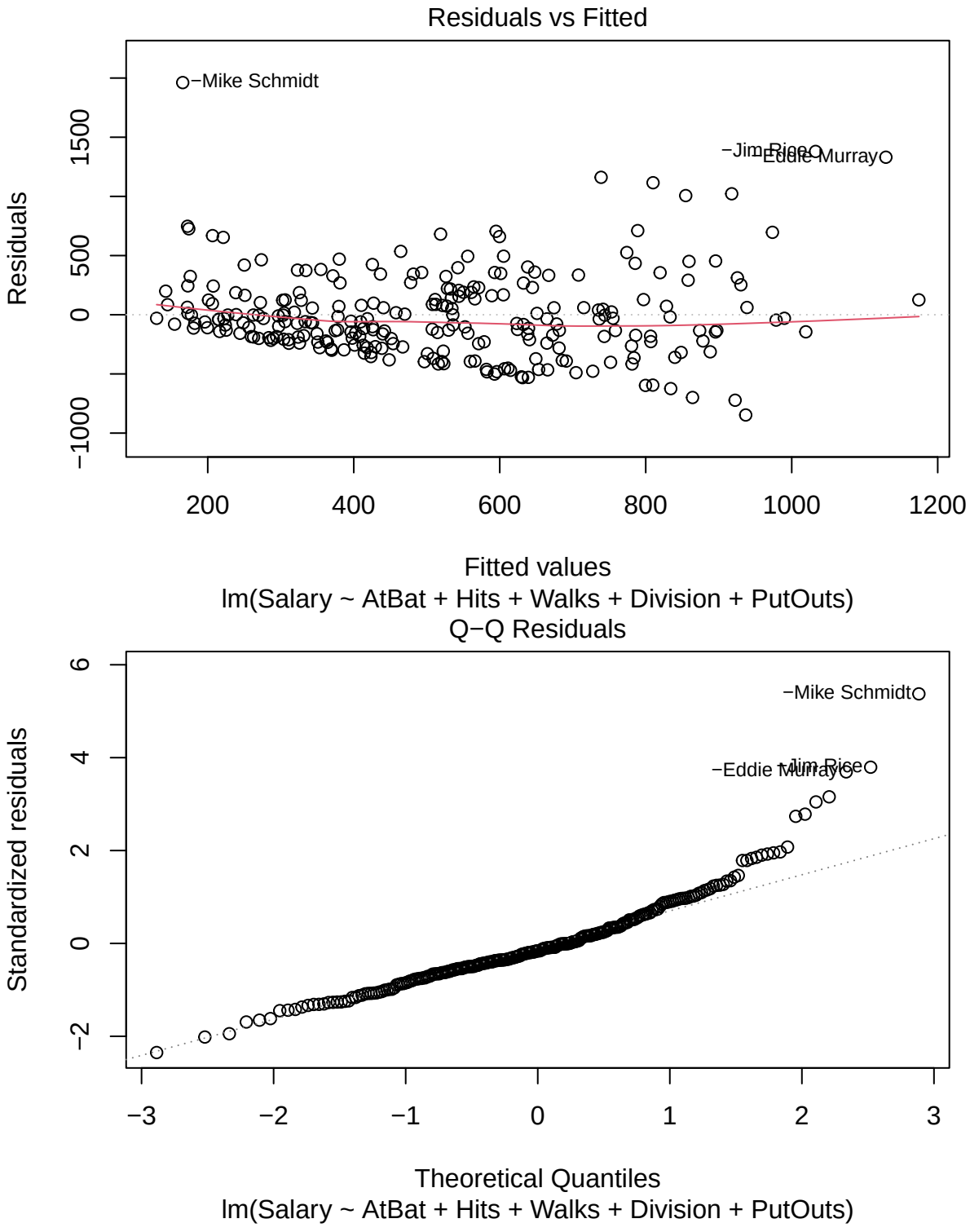

Q−Q Residuals
Standardized residuals
−Mike Schmidt
−Jim Rice
−Eddie Murray
−3
−2
−1
0
1
2
3
6
4
2
0
−2
Theoretical Quantiles
lm(Salary ~ AtBat + Hits + Walks + CWalks + Division + PutOuts)
Scale−Location
√|Standardized residuals|
−Mike Schmidt
−Jim Rice
−Eddie Murray
0.0
0.5
1.0
1.5
2.0
2.5
500
1000
1500
Fitted values
lm(Salary ~ AtBat + Hits + Walks + CWalks + Division + PutOuts)

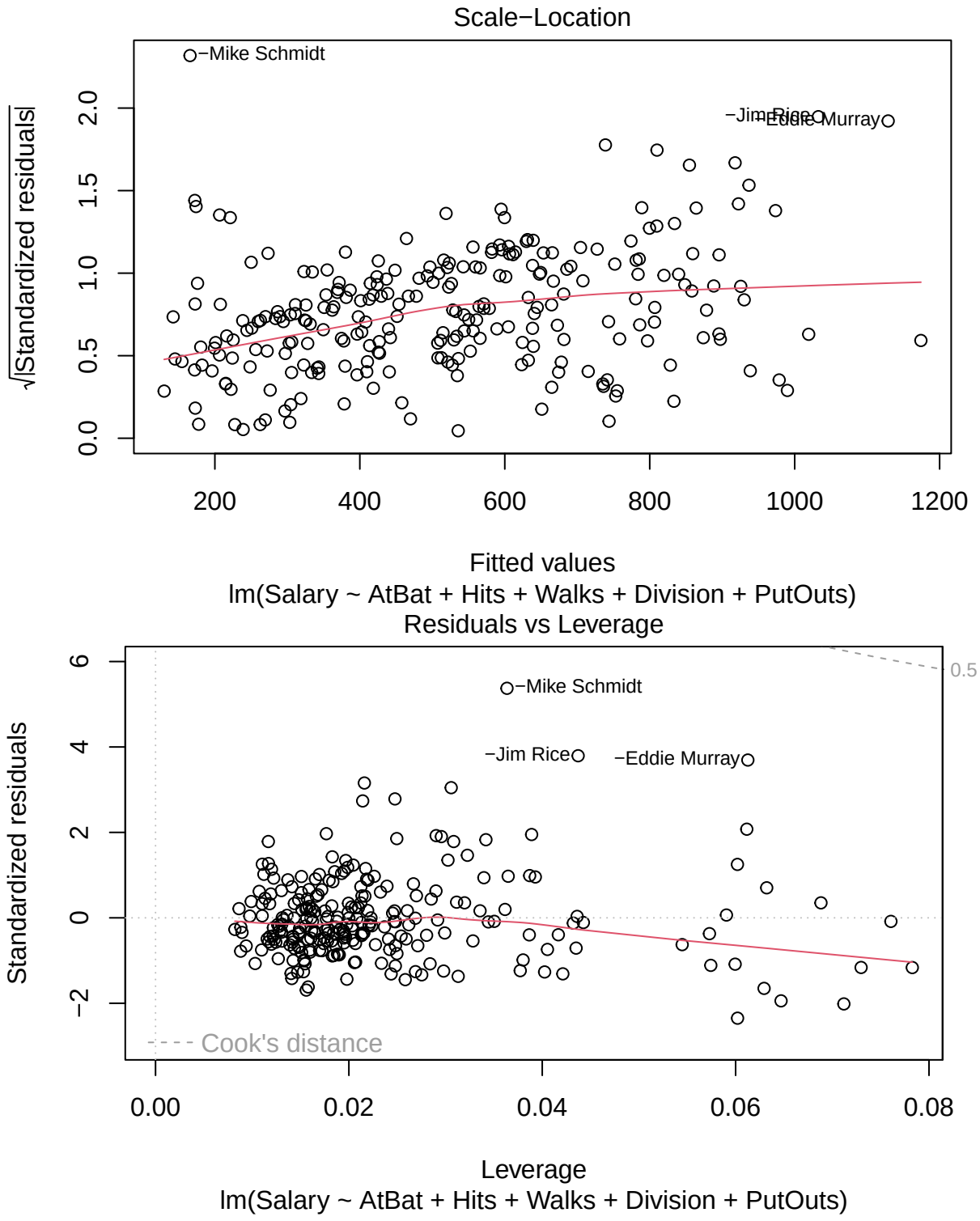


For residual vs fitted the linearity assumption does not seems to be violated so I wouldn't really change anything. While the data does not follow the line exactly, the line is relatively straight. For the residual vs leverage plot there are some outliers, I will try to fix it by removing them and then recomputing the model. For the scale-location plot I would say there is a discernible pattern meaning, the homoscedasticity assumption appears to be violated, I am going to fix the other issues first and then see if it fixes itself. For Q-Q residuals, the data does not fall on the line y=x but does closely follow the reference line so I am going to say it is normally distributed.

Fixing residual vs leverage plot by removing the outliers. I noticed that most of the outliers were after 0.07 on the leverage scale so by removing all the points with a leverage greater than .07 I effectively remove all the outliers:

```
leverage_values <- hatvalues(model)
outliers_leverage <- which(leverage_values > 0.07)
Hitters_clean <- Hitters1[-outliers_leverage, ]
model_clean <- lm(Salary ~ AtBat + Hits + Walks + Division + PutOuts, data = Hitters_clean)
plot(model_clean)
```


Residuals vs Fitted
Mike Schmidt
Jim Rice
Eddie Murray
Residuals
1500
500
0
−1000
200
400
600
800
1000
1200
Fitted values
lm(Salary ~ AtBat + Hits + Walks + Division + PutOuts)
Q−Q Residuals
Mike Schmidt
Jim Rice
Eddie Murray
Standardized residuals
6
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Salary ~ AtBat + Hits + Walks + Division + PutOuts)



By removing the outliers I fixed both the residual vs leverage plot by getting rid of the outliers so now the data is all close to each other and the scale-location plot by having the data no discernable pattern in the spread of the points because the outliers aren't pulling the data in one direction.

```
summary(model_clean)
```

```
##
## Call:
## lm(formula = Salary ~ AtBat + Hits + Walks + Division + PutOuts,
##     data = Hitters_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -846.98 -221.54  -59.08  165.16 1961.60
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  180.03125   75.37988   2.388 0.017668 *
## AtBat         -1.71959    0.66061  -2.603 0.009793 **
## Hits           7.10075    2.09953   3.382 0.000835 ***
## Walks          7.10114    1.47150   4.826 2.43e-06 ***
## DivisionW   -125.80286   46.70680  -2.693 0.007550 **
## PutOuts        0.16656    0.09439   1.765 0.078851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 371.9 on 250 degrees of freedom
## Multiple R-squared:  0.2728, Adjusted R-squared:  0.2582
## F-statistic: 18.75 on 5 and 250 DF,  p-value: 8.026e-16
```

I would say by final model is not the best. Going off the R-squared value, my model does not fit the data very well. I think this could be because of a few reasons, one that I am using a subset of a larger dataset meaning the whole picture isn't being shown, the second is that the relationship between the variables and salary might not be exactly linear, and lastly, I could have not cleaned the data properly enough during the model assumptions.

Conclusion:

In my work, I analyzed the Hitters dataset from the ISLR package and was able to do a few things. I learned about the data and did some necessary cleaning to remove empty rows and fix the column labels so they were accurate to the data. Then I created a smaller subset with just numeric data so I could run analysis on it. Using our response variable I learned about which factors contributed most to salary, and which contributed least. With things like player statistics contribute more and game statistics less. I then created some graphs to showcase both how individual variables look in the dataset and also how they relate to other ones. Then I completed a regression analysis and cleaned the data even more to create my final model. Overall this process taught me a lot about being meticulous and making sure to double check all my code. Something I kept forgetting was that I made a copy of the dataset to do work on which is what was cleaned and when the graphs would look weird or the code wouldn't work I would have to remind myself to go back and check to make sure I used the right dataset. I think what was unusual about this project was that the all of the data was in a one or two year period over different leagues, not one league over time so sometimes it was hard to think about what questions I wanted to answer with my analysis. I did enjoy doing this and I like how completed it feels and how accomplished I feel.