Preliminary Analysis of Final Project

By Karan Daiya

NUID: 001083274

Northeastern University: College of Professional Studies

Introduction:

Here, I have selected the dataset of Men's Fashion Stores. There are total 400 observations in the whole dataset. These observations are from Netherland country and a cross-section from 1990. This report will analyze the operations performed in R Notebook.

Format:

tsales :annual sales in Dutch guilders sales

sales: per square meter

margin: gross-profit-margin

nown : number of owners (managers)

nfull : number of full-timers

npart :number of part-timers

naux :number of helpers (temporary workers)

hoursw : total number of hours worked

hourspw: number of hours worked per worker

inv1: investment in shop-premises

inv2: investment in automation.

Ssize: sales floorspace of the store (in m^2).

Start: year start of business

Analysis:

Part 1:

```
setwd("/Users/karan/Downloads")

attention <- read.csv('Clothing (1).csv')

View(attention)
```

Here, setwd is used to set the working directory and View() function is used to display the whole dataset.

Part 2:

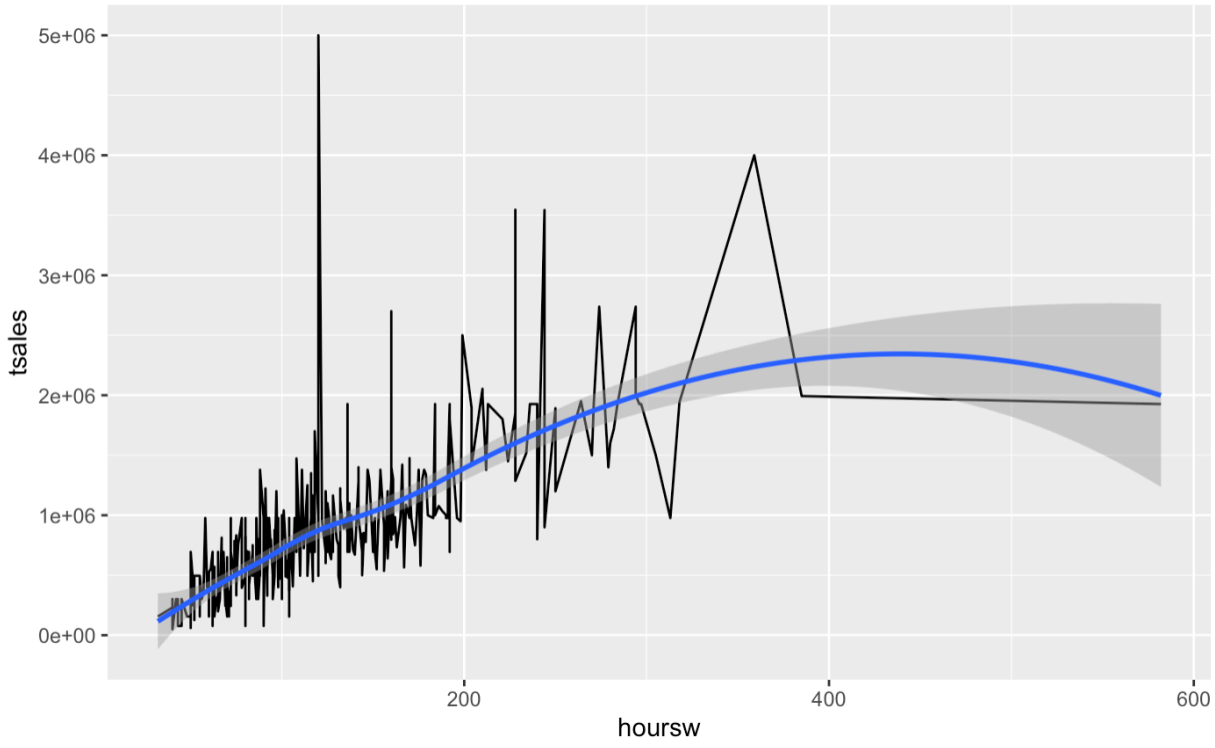
```
summary(attention)
```

```
       X             tsales            sales           margin          nown        
 Min.   :  1.0   Min.   :  50000   Min.   :  300   Min.   :16.00   Min.   : 1.000  
 1st Qu.:100.8   1st Qu.: 495340   1st Qu.: 3904   1st Qu.:37.00   1st Qu.: 1.000  
 Median :200.5   Median : 694227   Median : 5279   Median :39.00   Median : 1.000  
 Mean   :200.5   Mean   : 833584   Mean   : 6335   Mean   :38.77   Mean   : 1.284  
 3rd Qu.:300.2   3rd Qu.: 976817   3rd Qu.: 7740   3rd Qu.:41.00   3rd Qu.: 1.295  
 Max.   :400.0   Max.   :5000000   Max.   :27000   Max.   :66.00   Max.   :10.000  
     nfull           npart            naux           hoursw         hourspw      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 32.0   Min.   : 5.708  
 1st Qu.:1.923   1st Qu.:1.283   1st Qu.:1.333   1st Qu.: 80.0   1st Qu.:13.541  
 Median :1.956   Median :1.283   Median :1.367   Median :104.0   Median :17.745  
 Mean   :2.069   Mean   :1.566   Mean   :1.390   Mean   :121.1   Mean   :18.955  
 3rd Qu.:2.066   3rd Qu.:2.000   3rd Qu.:1.367   3rd Qu.:145.2   3rd Qu.:24.303  
 Max.   :8.000   Max.   :9.000   Max.   :4.000   Max.   :582.0   Max.   :43.326  
      inv1             inv2           ssize           start       
 Min.   :   1000   Min.   :   350   Min.   :  16.0   Min.   :16.00  
 1st Qu.:  20000   1st Qu.: 10000   1st Qu.:  80.0   1st Qu.:37.00  
 Median :  22207   Median : 22860   Median : 120.0   Median :40.00  
 Mean   :  58257   Mean   : 27829   Mean   : 151.1   Mean   :42.81  
 3rd Qu.:  62269   3rd Qu.: 22860   3rd Qu.: 190.0   3rd Qu.:42.00  
 Max.   :1500000   Max.   :400000   Max.   :1214.0   Max.   :90.00  
```

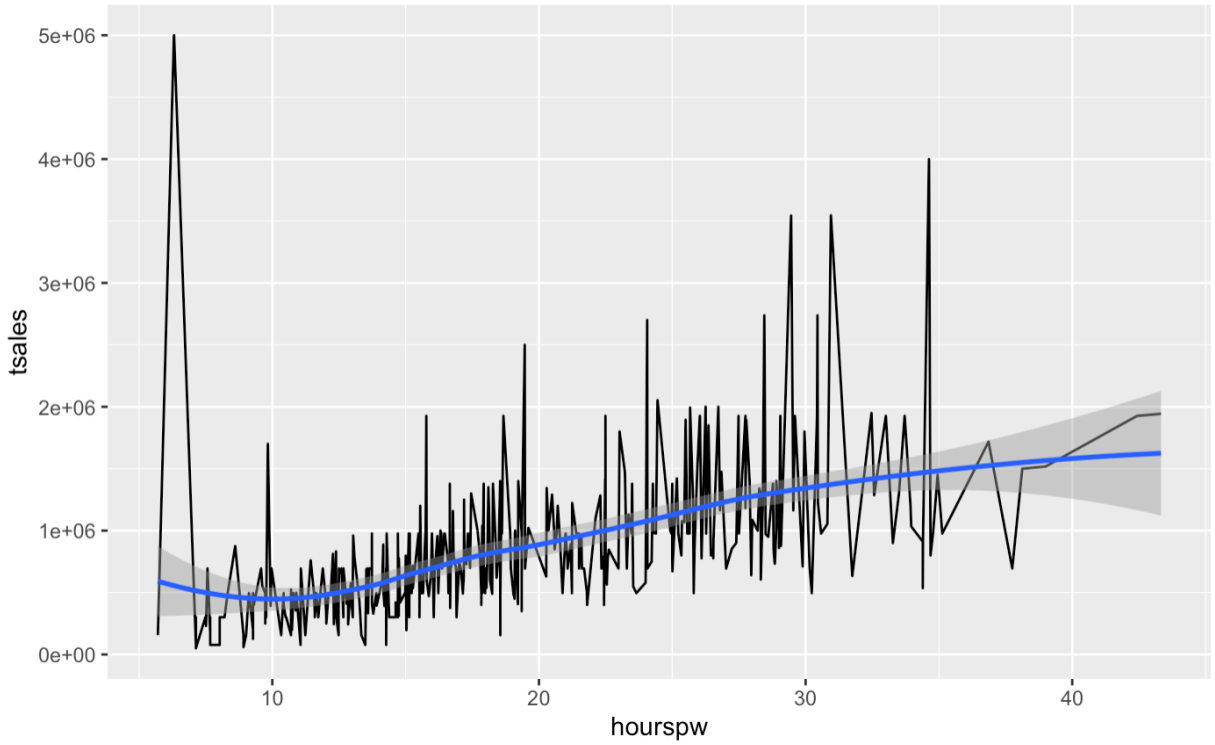
Part 3:

```{r Splitting annual sales by total hours worked}

library(ggplot2)

par(mfrow = c(1, 2))

ggplot(data = attention, aes(x= hoursw, y= tsales)) + geom_line() + geom_smooth()

ggplot(data = attention, aes(x= hourspw, y= tsales)) + geom_line() + geom_smooth()

```

Output:



The above graph shows the line chart of total hours worked vs total annual sales of showrooms. Geom_smooth() aids the eye in seeing patterns in the presence of overplotting.

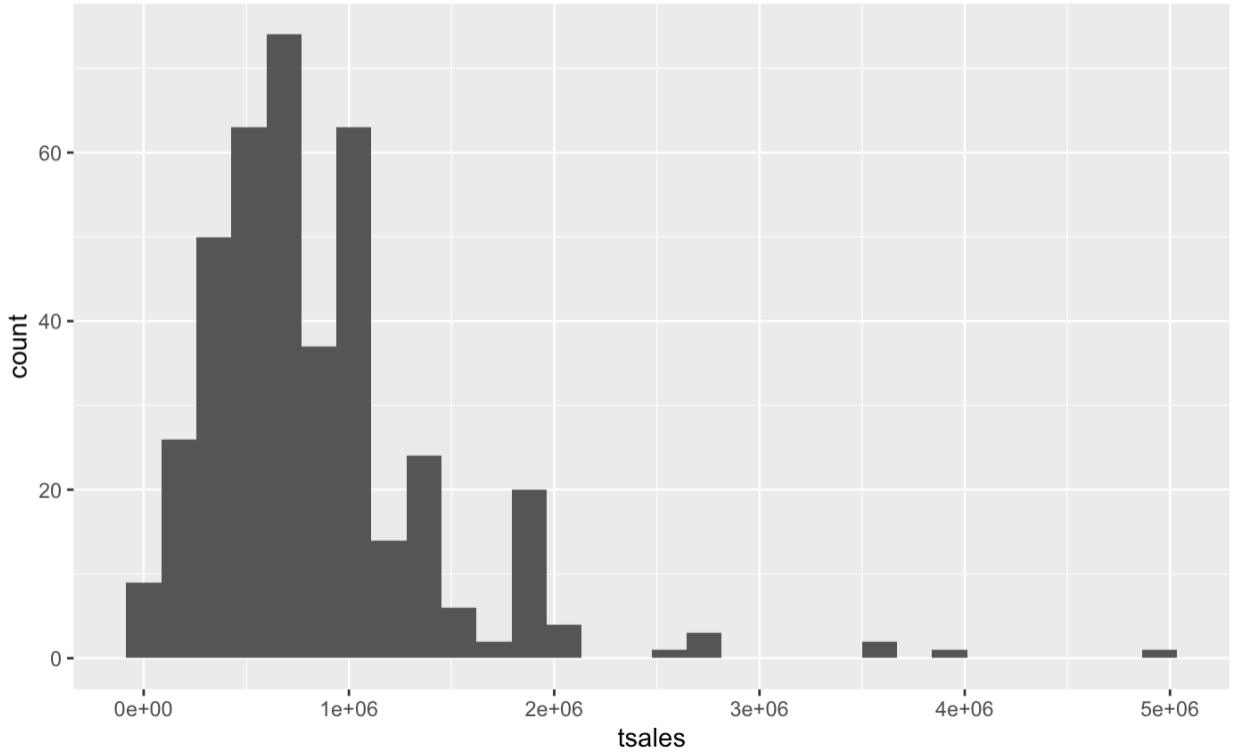


The above graph shows the line chart of total hours worked per worker vs total annual sales of showrooms.

Part 4:

```{r Histogram of total sales}

ggplot(data = attention, aes(x= tsales)) + geom_histogram()

```

Output:

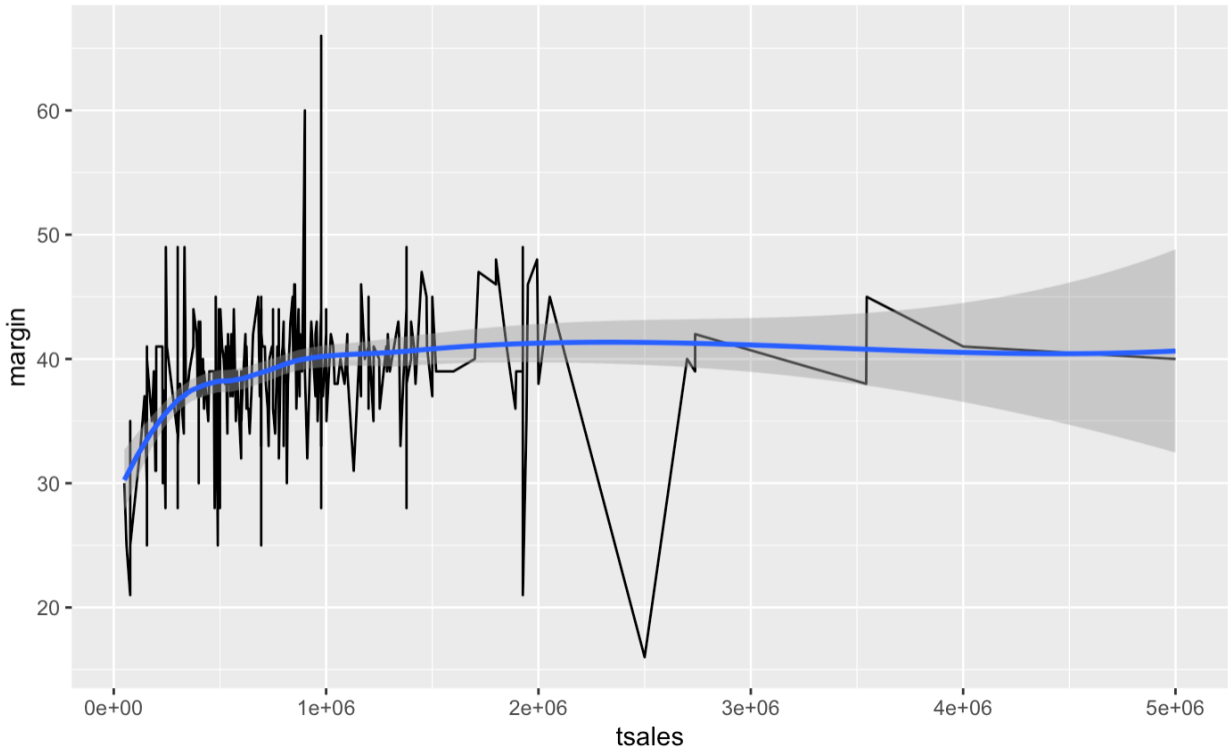


Here by creating the histogram of total annual sales, we can say that maximum number of hours worked is in the range of 0e+00 to 2e+06 of tsales.

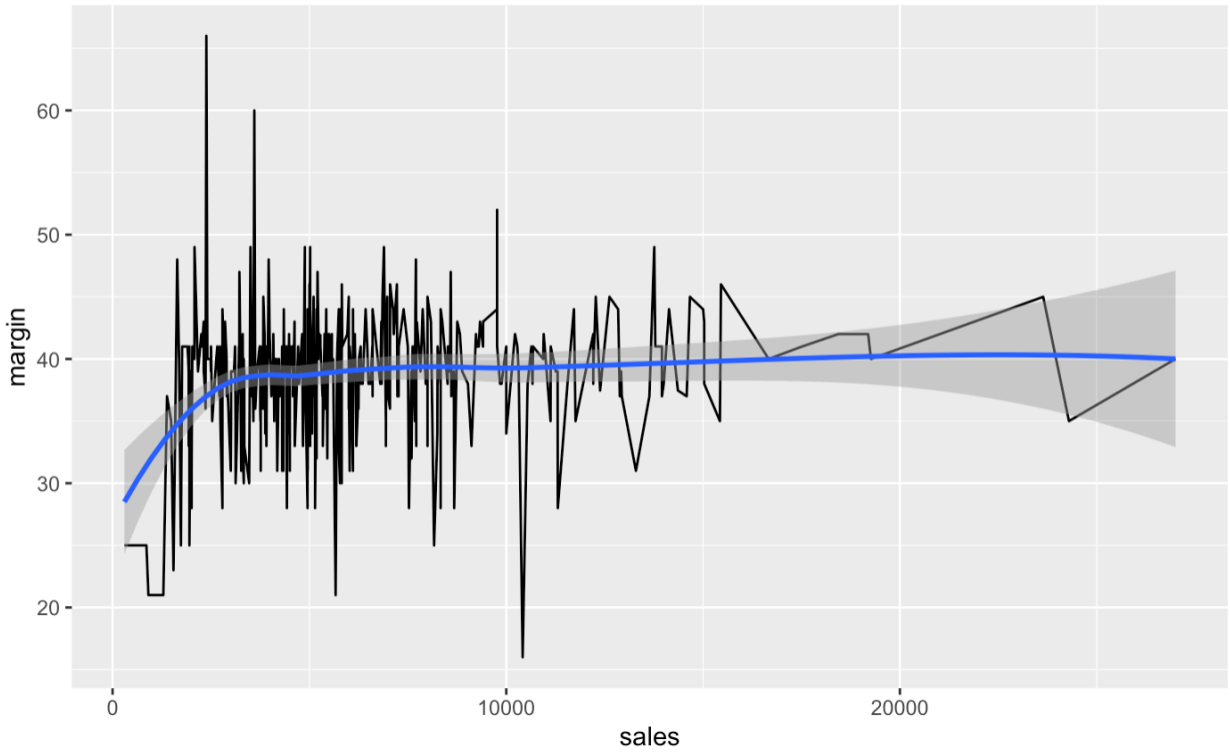
Part 5:

```{r Splitting Margin Vs tsales and sales}

par(mfrow = c(1, 2))

ggplot(data = attention, aes(x= tsales, y= margin)) + geom_line() + geom_smooth()

ggplot(data = attention, aes(x= sales, y= margin)) + geom_line() + geom_smooth()

```

Output:



The above line chart shows the correlation between margin and the total annual sales. We can say that after a specific point, the sales are not much affected by the margin variable.

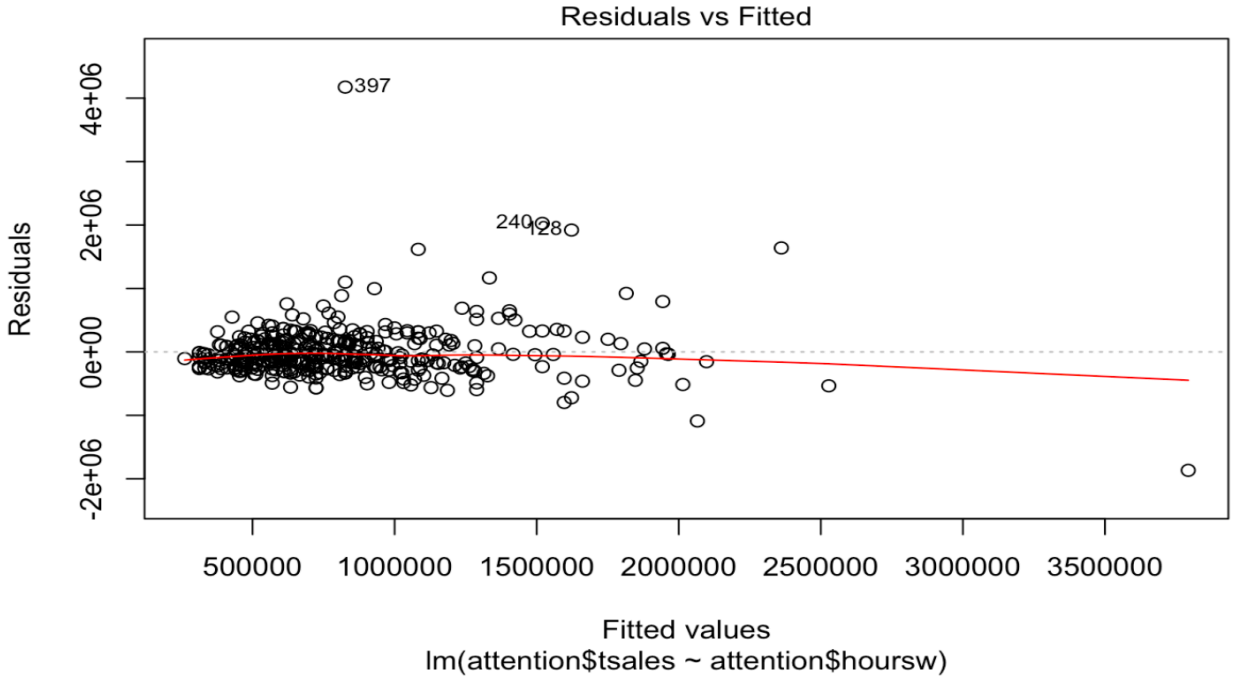


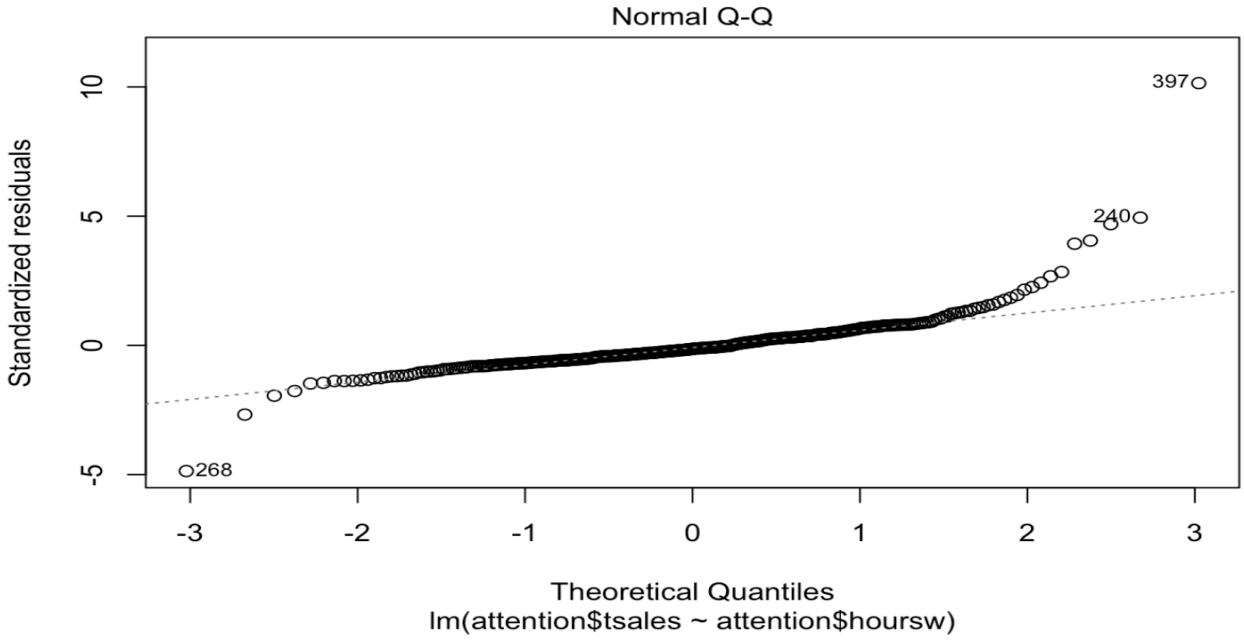
The above line chart shows the correlation between margin and the sales per square meter. We can say that after a specific point, the sales are not that much affected by the margin variable.

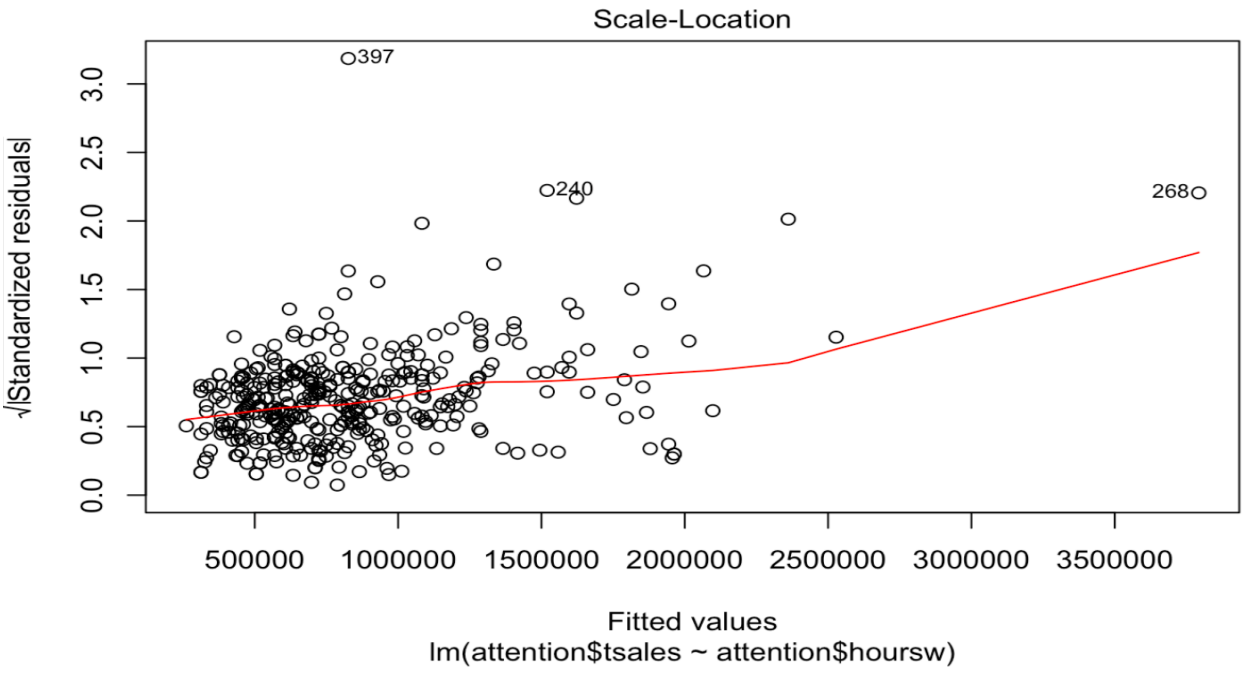
Part 6:

```{r}
reg_TSvHrs <- lm(attention$tsales~attention$hoursw, )
plot(reg_TSvHrs)
```

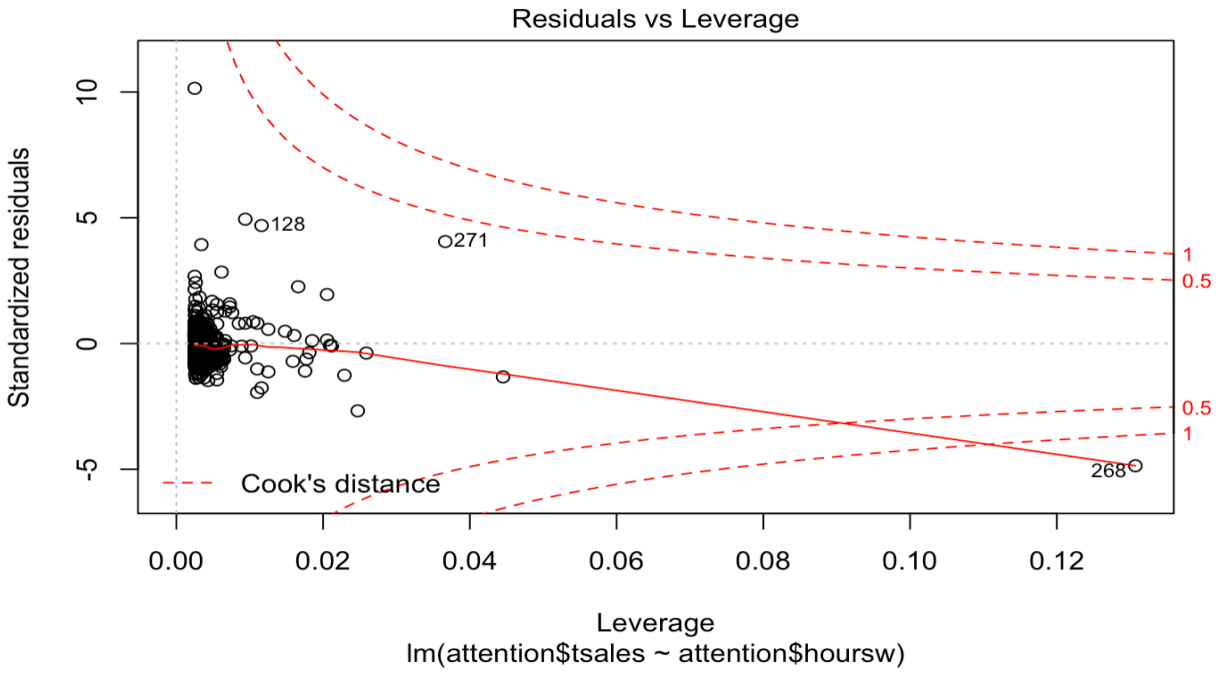
Output:


Residuals vs Fitted
Residuals
Fitted values
lm(attention$tsales ~ attention$hoursw)
397
240
128
-2e+06
0e+00
2e+06
4e+06
500000
1000000
1500000
2000000
2500000
3000000
3500000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(attention$tsales ~ attention$hoursw)
397
240
268
-5
0
5
10
-3
-2
-1
0
1
2
3



The line in above Scale-location chart represents the regression line for my dataset.



Conclusion:

After focusing on the data and applying the preliminary analysis to my dataset, we can conclude from these 400 observations that maximum number of hours a worker did was between 10 to 35 hours. Lastly I have done regression on my dataset and above are given the snapshots of regression models.