# Multi-ANOVA project_Chi Nguyen

2022-06-20

## Introduction:

In this project, I analyse the "engineer.csv" data. This data is about salary of different engineer profession in different regions of the US.

The dependent variable is the "salary", and the 2 independent categorical variables are "Profession", "Region".

I will do a multi-ANOVA analysis to have an understanding about the data and the interaction inside it.

## The Analysis:

Firstly, I load the libraries:

```
library(devtools)
```

```
## Loading required package: usethis
```

```
library(qdata)
library(data.table)
library(dplyr)
```

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

## The following objects are masked from 'package:data.table':
##
##     between, first, last

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

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

```
library(ggplot2)
```

Load 'engineer.csv' data set:

```r
setwd("~/Documents/DATA SCIENCE/MSDS/03. MSMS 660/06. Week 6/In class")
engineerdt <- read.csv(file = 'engineer.csv',sep=",", header=T)
```

Check structure of dt:

```r
dim(engineerdt)
```

```
## [1] 180   4
```

The data has 4 columns and 180 rows.

Check the class of each variables:

```r
str(engineerdt)
```

```
## 'data.frame':    180 obs. of  4 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Salary    : int  126411 108402 99399 91381 105023 108944 123952 108217 103722 140179 ...
##  $ Profession: chr  "Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ...
##  $ Region    : chr  "San Francisco" "San Francisco" "San Francisco" "San Francisco" ...
```

The class looks good. But the "X" column has no meaning to the analysis since it's jus the number order. So, I will remove it.

```r
engineerdt <- engineerdt[-c(1)]
dim(engineerdt)
```

```
## [1] 180   3
```

```r
str(engineerdt)
```

```
## 'data.frame':    180 obs. of  3 variables:
##  $ Salary    : int  126411 108402 99399 91381 105023 108944 123952 108217 103722 140179 ...
##  $ Profession: chr  "Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ...
##  $ Region    : chr  "San Francisco" "San Francisco" "San Francisco" "San Francisco" ...
```

The data looks good now and is ready for the analysis.

Convert the 2 independent variables (Profession, Region) to factors:

```r
engineerdt$Profession <- as.factor(engineerdt$Profession)
engineerdt$Region <- as.factor(engineerdt$Region)
```

Double check the class of 2 those variables:

```r
str(engineerdt)
```

```
## 'data.frame':    180 obs. of  3 variables:
##  $ Salary    : int  126411 108402 99399 91381 105023 108944 123952 108217 103722 140179 ...
##  $ Profession: Factor w/ 3 levels "BI Engineer",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Region    : Factor w/ 3 levels "New York","San Francisco",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Now, let's check on which Profession and City that have the highest salary:

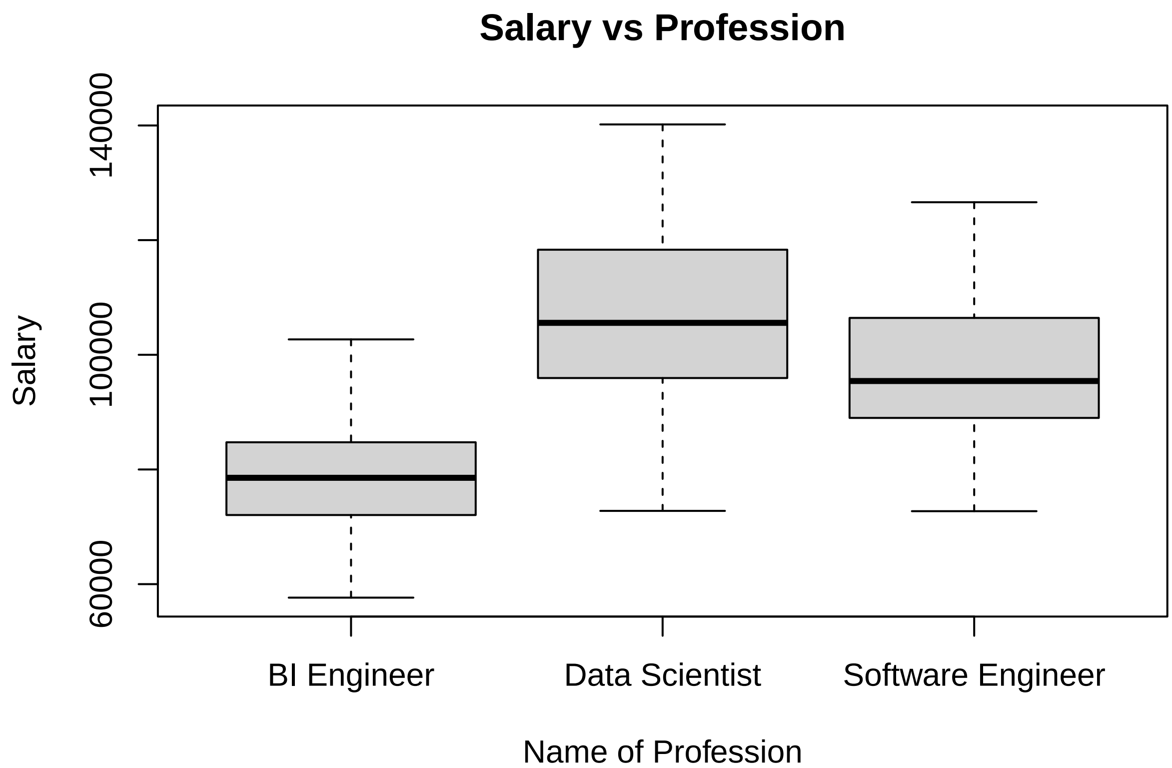
But first, plot histogram of Salary to have a surfing view on the Salary data distribution:

```
hist(engineerdt$Salary)
```


According to the plot, most of people's salary are in the range from 70k to 120k.

Plot Salary vs the 2 other factors:

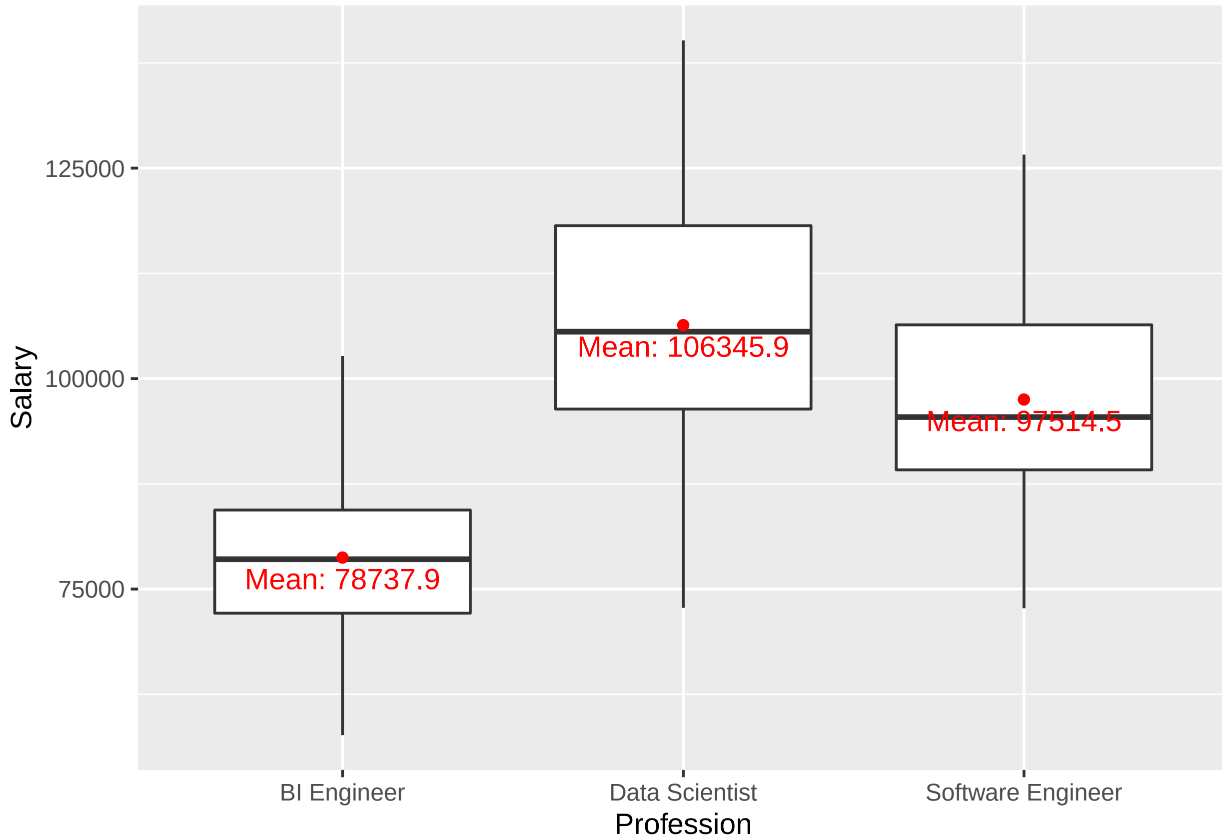
```
boxplot(Salary ~ Profession,data=engineerdt, main="Salary vs Profession",
        xlab="Name of Profession", ylab="Salary")
```



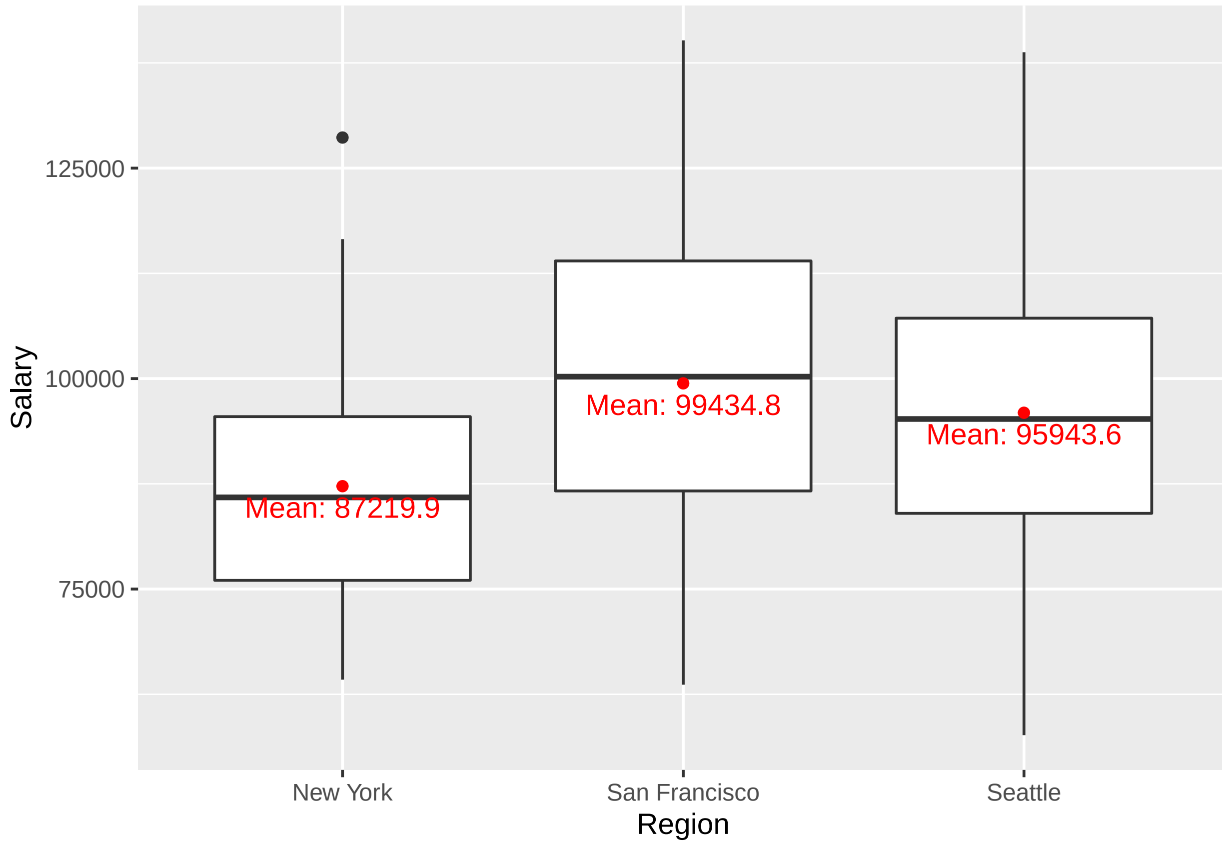

```
boxplot(Salary ~ Region,data=engineerdt, main="Salary vs Region",
        xlab="Name of Region", ylab="Salary")
```


Plot Individual Boxplots with means on them:

```
ggplot(engineerdt, aes(x = Profession, y = Salary)) +
  geom_boxplot() +
  stat_summary(fun = mean, geom = "point", col = "red") +  # Add points to plot
  stat_summary(fun = mean, geom = "text", col = "red",     # Add text to plot
               vjust = 1.5, aes(label = paste("Mean:", round(..y.., digits = 1))))
```




```
ggplot(engineerdt, aes(x = Region, y = Salary)) +
  geom_boxplot() +
  stat_summary(fun = mean, geom = "point", col = "red") +  # Add points to plot
  stat_summary(fun = mean, geom = "text", col = "red",     # Add text to plot
               vjust = 1.5, aes(label = paste("Mean:", round(..y.., digits = 1))))
```

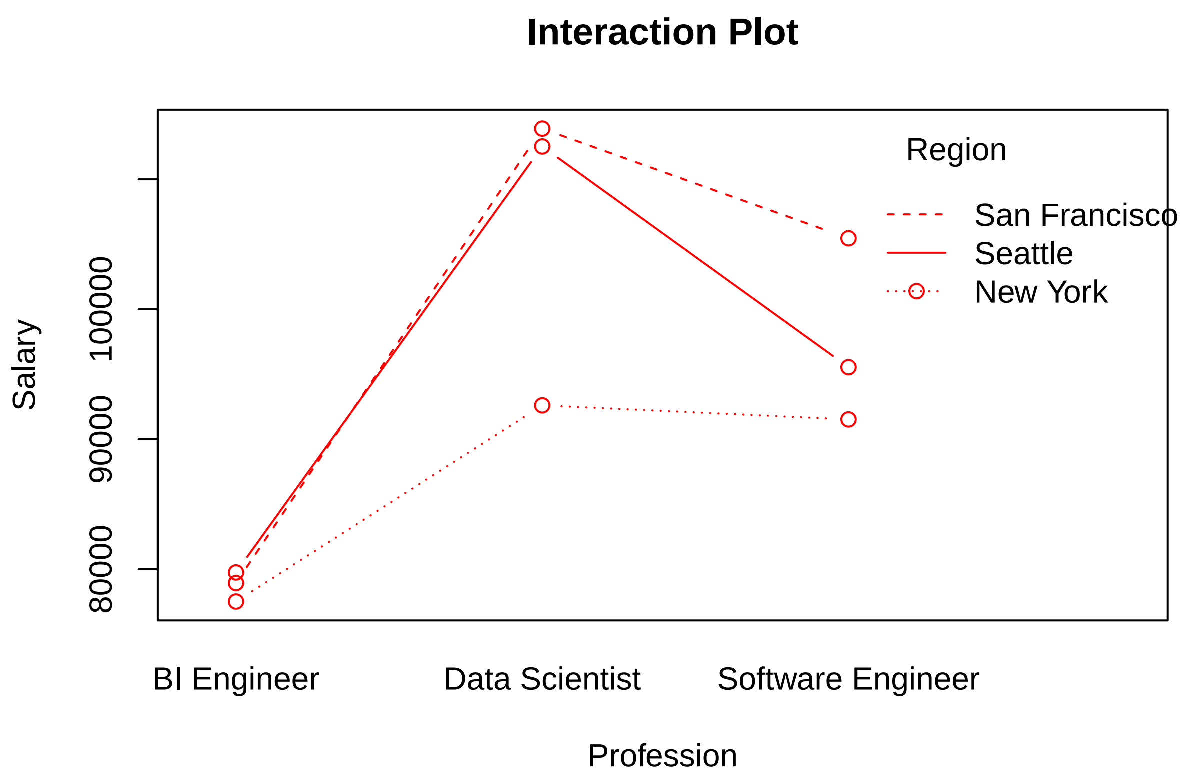



According to the boxplots:

- We can see that there's probably a significant difference between average salary of different professions but not really significant between different regions.
- In terms of Profession, the average salary of Data Scientist is the highest comparing to the 2 others profession.
- In terms of Region, engineers living in San Francisco have the highest average salary, but not much higher than Seattle.

Create interaction plot looking at Profession and Region:

```
interaction.plot(x.factor = engineerdt$Profession,
                 trace.factor = engineerdt$Region,
                 response = engineerdt$Salary,
                 fun = mean,
                 type = "b",  # shows each point
                 main = "Interaction Plot",
                 legend = TRUE,
                 trace.label = "Region",
                 xlab = "Profession",
                 ylab="Salary",
                 pch=c(1),
                 col = c("Red"))
```




There are two lines intersect, hence we can indicate that there's a considerable interaction between Profession and Region in terms of Salary.

Now, I will double check that interaction by ANOVA:

```
model <- aov(Salary ~ Profession * Region, data = engineerdt)
summary(model)
```

```
##                      Df    Sum Sq   Mean Sq F value   Pr(>F)    
## Profession            2 2.386e+10 1.193e+10  86.098  < 2e-16 ***
## Region                2 4.750e+09 2.375e+09  17.143 1.64e-07 ***
## Profession:Region     4 3.037e+09 7.593e+08   5.481 0.000355 ***
## Residuals           171 2.369e+10 1.385e+08                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the p-values and a significance level of 0.05, the model tell us key things:

- The p-value of Profession, and Region are <2e-16 and 1.64e-07, which indicate that the different Profession or Region are associated with Salary. In other words, salary of different profession or different region are not the same.
- The p-value of "Profession:Region" is 0.000355, much smaller than 0.05 as expected, hence, there's a significant interaction effect between Profession and Region in terms of Salary. In other words, those 2 factors together interact and affect people's salary.

### TukeyHSD

The p-value has just showed the significant interaction between the 2 factos, now, I will perform TukeyHSD post hoc test to check more into the details:

```
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Salary ~ Profession * Region, data = engineerdt)
##
## $Profession
##                                     diff       lwr       upr     p adj
## Data Scientist-BI Engineer       27608.02  22527.33 32688.707 0.0000000
## Software Engineer-BI Engineer    18776.57  13695.88 23857.257 0.0000000
## Software Engineer-Data Scientist -8831.45 -13912.14 -3750.759 0.0001807
##
## $Region
##                              diff       lwr       upr     p adj
## San Francisco-New York 12214.900  7134.209 17295.591 0.0000002
## Seattle-New York        8723.683  3642.993 13804.374 0.0002197
## Seattle-San Francisco  -3491.217 -8571.907  1589.474 0.2380471
##
## $`Profession:Region`
##                                                              diff
## Data Scientist:New York-BI Engineer:New York             15092.65
## Software Engineer:New York-BI Engineer:New York          14010.80
## BI Engineer:San Francisco-BI Engineer:New York            1421.35
## Data Scientist:San Francisco-BI Engineer:New York        36380.45
## Software Engineer:San Francisco-BI Engineer:New York     27946.35
## BI Engineer:Seattle-BI Engineer:New York                  2236.10
## Data Scientist:Seattle-BI Engineer:New York              35008.40
## Software Engineer:Seattle-BI Engineer:New York           18030.00
## Software Engineer:New York-Data Scientist:New York       -1081.85
## BI Engineer:San Francisco-Data Scientist:New York       -13671.30
## Data Scientist:San Francisco-Data Scientist:New York     21287.80
## Software Engineer:San Francisco-Data Scientist:New York  12853.70
## BI Engineer:Seattle-Data Scientist:New York             -12856.55
## Data Scientist:Seattle-Data Scientist:New York           19915.75
## Software Engineer:Seattle-Data Scientist:New York         2937.35
## BI Engineer:San Francisco-Software Engineer:New York    -12589.45
## Data Scientist:San Francisco-Software Engineer:New York  22369.65
## Software Engineer:San Francisco-Software Engineer:New York 13935.55
## BI Engineer:Seattle-Software Engineer:New York          -11774.70
## Data Scientist:Seattle-Software Engineer:New York        20997.60
## Software Engineer:Seattle-Software Engineer:New York      4019.20
## Data Scientist:San Francisco-BI Engineer:San Francisco   34959.10
## Software Engineer:San Francisco-BI Engineer:San Francisco 26525.00
## BI Engineer:Seattle-BI Engineer:San Francisco              814.75
## Data Scientist:Seattle-BI Engineer:San Francisco         33587.05
## Software Engineer:Seattle-BI Engineer:San Francisco      16608.65
## Software Engineer:San Francisco-Data Scientist:San Francisco  -8434.10
## BI Engineer:Seattle-Data Scientist:San Francisco        -34144.35
```

```
## Data Scientist:Seattle-Data Scientist:San Francisco          -1372.05
## Software Engineer:Seattle-Data Scientist:San Francisco      -18350.45
## BI Engineer:Seattle-Software Engineer:San Francisco         -25710.25
## Data Scientist:Seattle-Software Engineer:San Francisco        7062.05
## Software Engineer:Seattle-Software Engineer:San Francisco    -9916.35
## Data Scientist:Seattle-BI Engineer:Seattle                   32772.30
## Software Engineer:Seattle-BI Engineer:Seattle                15793.90
## Software Engineer:Seattle-Data Scientist:Seattle            -16978.40
##                                                                    lwr
## Data Scientist:New York-BI Engineer:New York                  3398.181
## Software Engineer:New York-BI Engineer:New York               2316.331
## BI Engineer:San Francisco-BI Engineer:New York              -10273.119
## Data Scientist:San Francisco-BI Engineer:New York            24685.981
## Software Engineer:San Francisco-BI Engineer:New York         16251.881
## BI Engineer:Seattle-BI Engineer:New York                     -9458.369
## Data Scientist:Seattle-BI Engineer:New York                  23313.931
## Software Engineer:Seattle-BI Engineer:New York                6335.531
## Software Engineer:New York-Data Scientist:New York          -12776.319
## BI Engineer:San Francisco-Data Scientist:New York           -25365.769
## Data Scientist:San Francisco-Data Scientist:New York          9593.331
## Software Engineer:San Francisco-Data Scientist:New York       1159.231
## BI Engineer:Seattle-Data Scientist:New York                 -24551.019
## Data Scientist:Seattle-Data Scientist:New York                8221.281
## Software Engineer:Seattle-Data Scientist:New York            -8757.119
## BI Engineer:San Francisco-Software Engineer:New York        -24283.919
## Data Scientist:San Francisco-Software Engineer:New York      10675.181
## Software Engineer:San Francisco-Software Engineer:New York    2241.081
## BI Engineer:Seattle-Software Engineer:New York              -23469.169
## Data Scientist:Seattle-Software Engineer:New York             9303.131
## Software Engineer:Seattle-Software Engineer:New York         -7675.269
## Data Scientist:San Francisco-BI Engineer:San Francisco       23264.631
## Software Engineer:San Francisco-BI Engineer:San Francisco    14830.531
## BI Engineer:Seattle-BI Engineer:San Francisco               -10879.719
## Data Scientist:Seattle-BI Engineer:San Francisco             21892.581
## Software Engineer:Seattle-BI Engineer:San Francisco           4914.181
## Software Engineer:San Francisco-Data Scientist:San Francisco -20128.569
## BI Engineer:Seattle-Data Scientist:San Francisco            -45838.819
## Data Scientist:Seattle-Data Scientist:San Francisco         -13066.519
## Software Engineer:Seattle-Data Scientist:San Francisco      -30044.919
## BI Engineer:Seattle-Software Engineer:San Francisco         -37404.719
## Data Scientist:Seattle-Software Engineer:San Francisco       -4632.419
## Software Engineer:Seattle-Software Engineer:San Francisco   -21610.819
## Data Scientist:Seattle-BI Engineer:Seattle                   21077.831
## Software Engineer:Seattle-BI Engineer:Seattle                 4099.431
## Software Engineer:Seattle-Data Scientist:Seattle            -28672.869
##                                                                      upr
## Data Scientist:New York-BI Engineer:New York                 26787.11898
## Software Engineer:New York-BI Engineer:New York              25705.26898
## BI Engineer:San Francisco-BI Engineer:New York               13115.81898
## Data Scientist:San Francisco-BI Engineer:New York            48074.91898
## Software Engineer:San Francisco-BI Engineer:New York         39640.81898
## BI Engineer:Seattle-BI Engineer:New York                     13930.56898
## Data Scientist:Seattle-BI Engineer:New York                  46702.86898
## Software Engineer:Seattle-BI Engineer:New York               29724.46898
```

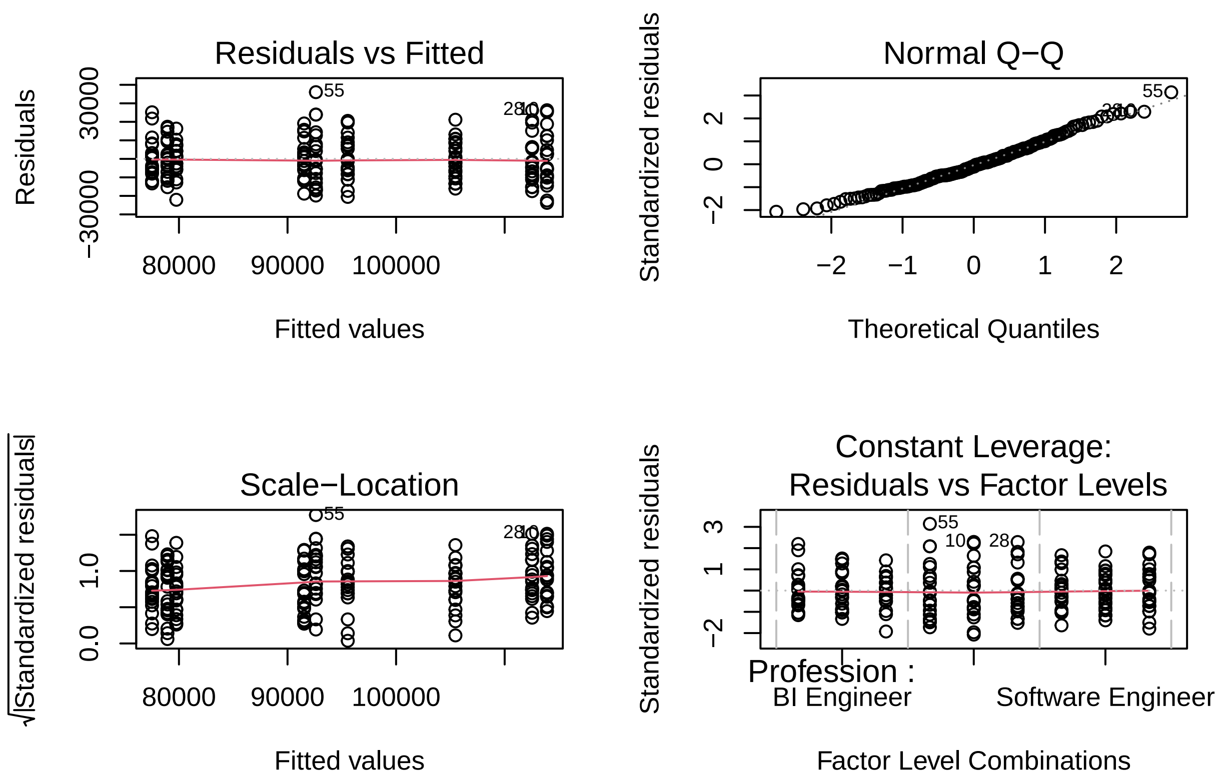
```
## Software Engineer:New York-Data Scientist:New York            10612.61898
## BI Engineer:San Francisco-Data Scientist:New York            -1976.83102
## Data Scientist:San Francisco-Data Scientist:New York         32982.26898
## Software Engineer:San Francisco-Data Scientist:New York      24548.16898
## BI Engineer:Seattle-Data Scientist:New York                  -1162.08102
## Data Scientist:Seattle-Data Scientist:New York               31610.21898
## Software Engineer:Seattle-Data Scientist:New York            14631.81898
## BI Engineer:San Francisco-Software Engineer:New York          -894.98102
## Data Scientist:San Francisco-Software Engineer:New York      34064.11898
## Software Engineer:San Francisco-Software Engineer:New York   25630.01898
## BI Engineer:Seattle-Software Engineer:New York                 -80.23102
## Data Scientist:Seattle-Software Engineer:New York            32692.06898
## Software Engineer:Seattle-Software Engineer:New York         15713.66898
## Data Scientist:San Francisco-BI Engineer:San Francisco       46653.56898
## Software Engineer:San Francisco-BI Engineer:San Francisco    38219.46898
## BI Engineer:Seattle-BI Engineer:San Francisco                12509.21898
## Data Scientist:Seattle-BI Engineer:San Francisco             45281.51898
## Software Engineer:Seattle-BI Engineer:San Francisco          28303.11898
## Software Engineer:San Francisco-Data Scientist:San Francisco   3260.36898
## BI Engineer:Seattle-Data Scientist:San Francisco            -22449.88102
## Data Scientist:Seattle-Data Scientist:San Francisco          10322.41898
## Software Engineer:Seattle-Data Scientist:San Francisco       -6655.98102
## BI Engineer:Seattle-Software Engineer:San Francisco         -14015.78102
## Data Scientist:Seattle-Software Engineer:San Francisco       18756.51898
## Software Engineer:Seattle-Software Engineer:San Francisco     1778.11898
## Data Scientist:Seattle-BI Engineer:Seattle                   44466.76898
## Software Engineer:Seattle-BI Engineer:Seattle                27488.36898
## Software Engineer:Seattle-Data Scientist:Seattle             -5283.93102
##                                                                 p adj
## Data Scientist:New York-BI Engineer:New York                0.0024207
## Software Engineer:New York-BI Engineer:New York             0.0069368
## BI Engineer:San Francisco-BI Engineer:New York              0.9999868
## Data Scientist:San Francisco-BI Engineer:New York           0.0000000
## Software Engineer:San Francisco-BI Engineer:New York        0.0000000
## BI Engineer:Seattle-BI Engineer:New York                    0.9995865
## Data Scientist:Seattle-BI Engineer:New York                 0.0000000
## Software Engineer:Seattle-BI Engineer:New York              0.0000975
## Software Engineer:New York-Data Scientist:New York          0.9999984
## BI Engineer:San Francisco-Data Scientist:New York           0.0094978
## Data Scientist:San Francisco-Data Scientist:New York        0.0000017
## Software Engineer:San Francisco-Data Scientist:New York     0.0195719
## BI Engineer:Seattle-Data Scientist:New York                 0.0195243
## Data Scientist:Seattle-Data Scientist:New York              0.0000098
## Software Engineer:Seattle-Data Scientist:New York           0.9970431
## BI Engineer:San Francisco-Software Engineer:New York        0.0244634
## Data Scientist:San Francisco-Software Engineer:New York     0.0000004
## Software Engineer:San Francisco-Software Engineer:New York  0.0074423
## BI Engineer:Seattle-Software Engineer:New York              0.0470207
## Data Scientist:Seattle-Software Engineer:New York           0.0000024
## Software Engineer:Seattle-Software Engineer:New York        0.9764101
## Data Scientist:San Francisco-BI Engineer:San Francisco      0.0000000
## Software Engineer:San Francisco-BI Engineer:San Francisco   0.0000000
## BI Engineer:Seattle-BI Engineer:San Francisco               0.9999998
## Data Scientist:Seattle-BI Engineer:San Francisco            0.0000000
```

```
## Software Engineer:Seattle-BI Engineer:San Francisco          0.0004900
## Software Engineer:San Francisco-Data Scientist:San Francisco 0.3687205
## BI Engineer:Seattle-Data Scientist:San Francisco             0.0000000
## Data Scientist:Seattle-Data Scientist:San Francisco          0.9999900
## Software Engineer:Seattle-Data Scientist:San Francisco       0.0000667
## BI Engineer:Seattle-Software Engineer:San Francisco          0.0000000
## Data Scientist:Seattle-Software Engineer:San Francisco       0.6165068
## Software Engineer:Seattle-Software Engineer:San Francisco    0.1687988
## Data Scientist:Seattle-BI Engineer:Seattle                   0.0000000
## Software Engineer:Seattle-BI Engineer:Seattle                0.0011759
## Software Engineer:Seattle-Data Scientist:Seattle             0.0003253
```

Looking at the p-values, we clearly see that there's many interactive pairs of "Profession" and "Region", but some of them are not interacted. This explain why there's one line that not intersect others in the interaction plot.

Plot the residuals of the fit:

```
par(mfrow = c(2,2))
plot(model)
```




- Accroding to the Residual vs Fitted plot, we can see that the data is linear since there's no clear pattern here.
- Normal Q-Q plot shows a normal distribution of the errors with some outliers.

- In the Scale-Location plot, the residuals are not randomly scattered around the red line, it means that the model probably does not fit the data well.

Perform Shapiro test to see if residuals are normaly distributed:

```
shapiro.test(engineerdt$Salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  engineerdt$Salary
## W = 0.9791, p-value = 0.008351
```

From the output obtained we can assume normality. The p-value is greater than 0.05. Hence, the distribution of the given data is not different from normal distribution significantly. In other words, the variable"Salary" may be normally distributed as expected, and this information can be used to decide to use a parametric test on this data set.

## Summary:

Firstly, I imported "engineer.csv" data for the analysis about salary of different engineer profession in different regions of the US.

Then I did some cleaning action for the data: checked the structure, changed the class, removed unused column.

Next, I plotted a histogram to have a look at the salary data. According to the plot, most of people's salary are in the range from 70k to 120k.

Next, I plotted boxplots of Salary with each 2 factors (Profession, Region) to check the distribution and the means. There's probably a significant difference between average salary of different professions but not really significant between different regions.

Next, I checked the interaction between 2 factors using the interaction plot. There are two lines intersect, hence we can indicate that there's a considerable interaction between Profession and Region in terms of Salary.

After that, I double checked the result by ANOVA. The p=value results indicate that there's a significant interaction effect between Profession and Region in terms of Salary. Salary of different profession or different region are not the same.

Next, I performed the TukeyHSD post hoc test to check that above result in details. The test shows that there are many interactive pairs of "Profession" and "Region", but some of them are not interacted. This explains why there's one line that not intersect others in the interaction plot.

Finally, I check the residuals of the fit model and then double check the distribution of the residuals by Shapiro test. And the output indicates that variable"Salary" may be normally distributed as expected, and this information can be used to decide to use a parametric test on this data set.