Two Way ANOVA

Ahmad Alqurashi

9/26/2021

Introduction

As week 4 was about one way ANOVA, we continue with Analysis of variances. This project uses Two Way ANOVA. Similar to one-way ANOVA, two way ANOVA is analysis of the variances, meaning that it is an analysis of the effects of more than one factors on a response. Moreover, factors can have interactions with each other which means that the effect of one factor relies on the level of the other factor. an interaction can be proven when p-value is less than 0.05. This project is using two-way ANOVA to study the effects of professions and regions on salaries.

Dataset

The dataset we going to use shows the salary, profession, and regoin. Engineer dataset contains 4 variables and 180 observations. It contains yearly salary for three professions in three different cities. the professions are Business Intelligence engineer, data scientist, and software engineer. those jobs are located in San Francisco, Seattle, and New York. There is a variable called V1 which sort observations in numbers, so it will not be in the model.

Methods and results

- Loading required libraries

```
# Load the libraries
library(data.table)
library(ggpubr)
```

```
## Loading required package: ggplot2
```

- Loading data into data table format

```
# Load 'engineer.csv' data set
dt <- fread("C:\\Users\\Ahmad\\Desktop\\MSDS\\MSDS660\\week 5\\assignment\\Engineer.csv")
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   180 obs. of  4 variables:
##  $ V1        : 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" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

- Data cleaning:
- Since our factors are stored as characters, we need to convert them to factors, so that we can fit them into ANOVA model.
- Also, dataset has V1 column, which is incremental number to ID the observations, so we are going to remove it entirely.

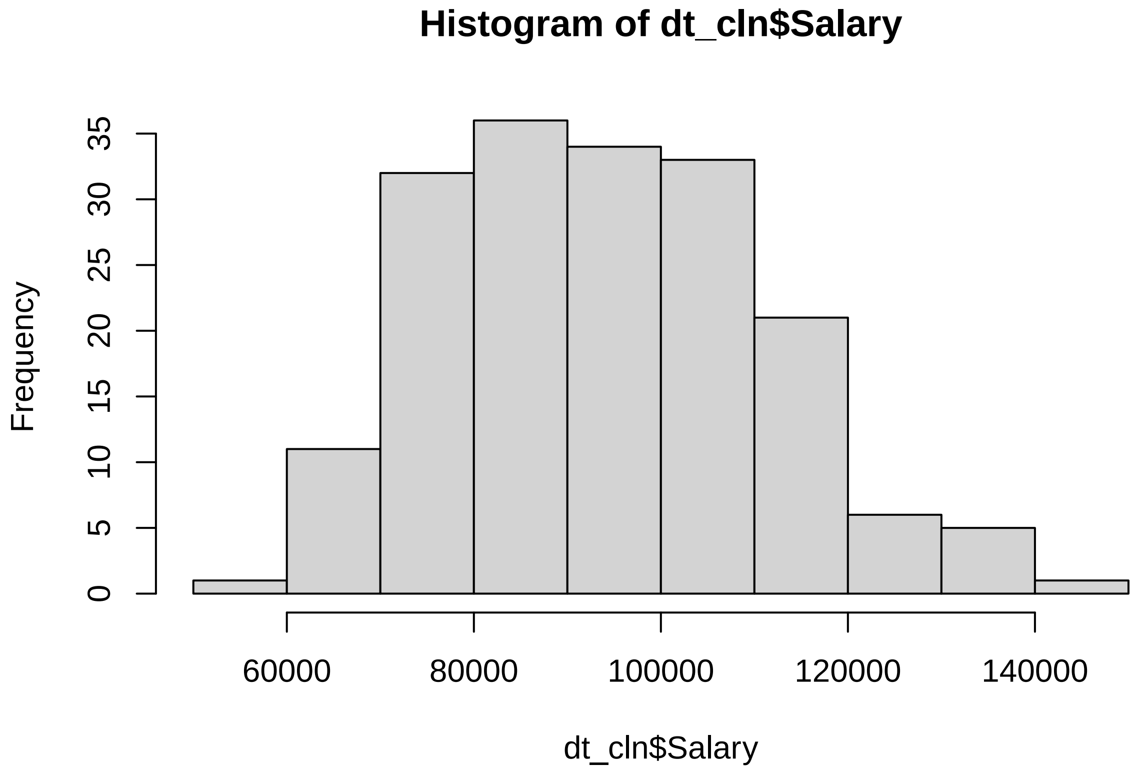
```
# Convert Profession and Region to factors
dt$Profession = as.factor(dt$Profession)
dt$Region = as.factor(dt$Region)
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   180 obs. of  4 variables:
##  $ V1        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
dt_cln <- dt[,-c(1)]
```

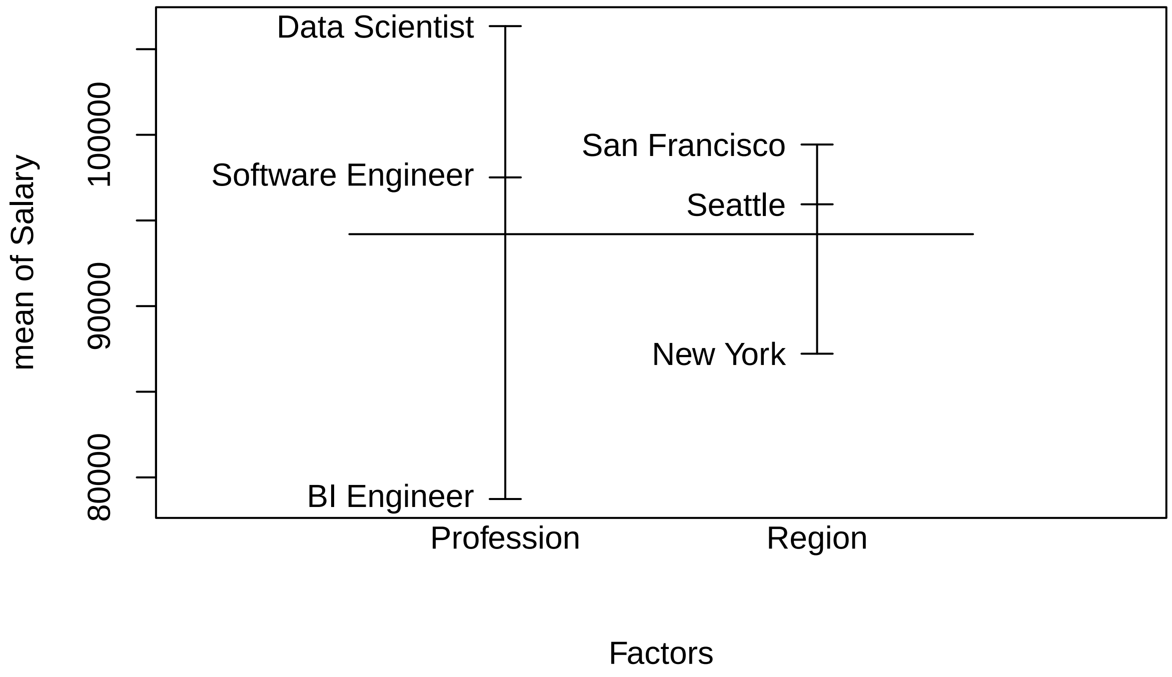
- The histogram plot for the response below shows that data is distributed normally. Although, it seems data is slightly skewed. However, it will not effect the result.

```
# Plot histogram of Salary
hist(dt_cln$Salary)
```



The figure below shows the difference in salaries based on the job and the city. we can see that Data Scientists are the highest paid jobs while Business intelligence engineers are the lowest. In terms of cities, San Francisco is the most city that has highest paid jobs. Also, it seems that Seattle and San Francisco are colse to each other in terms of salaries while New York falls behind by a lot. in the same sense, Data Scientists and Software Engineers are also close but BI Engineers is much lower.

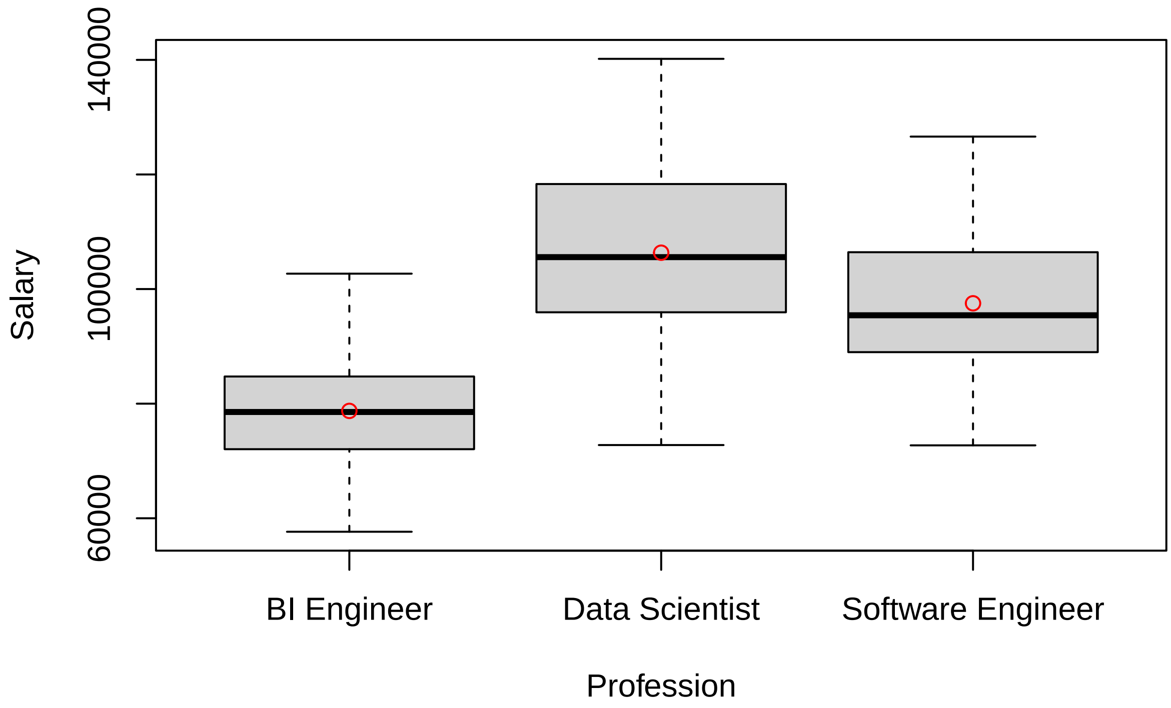
```r
# Plot Salary vs the 2 other factors
plot.design(Salary ~ ., data = dt_cln)
```



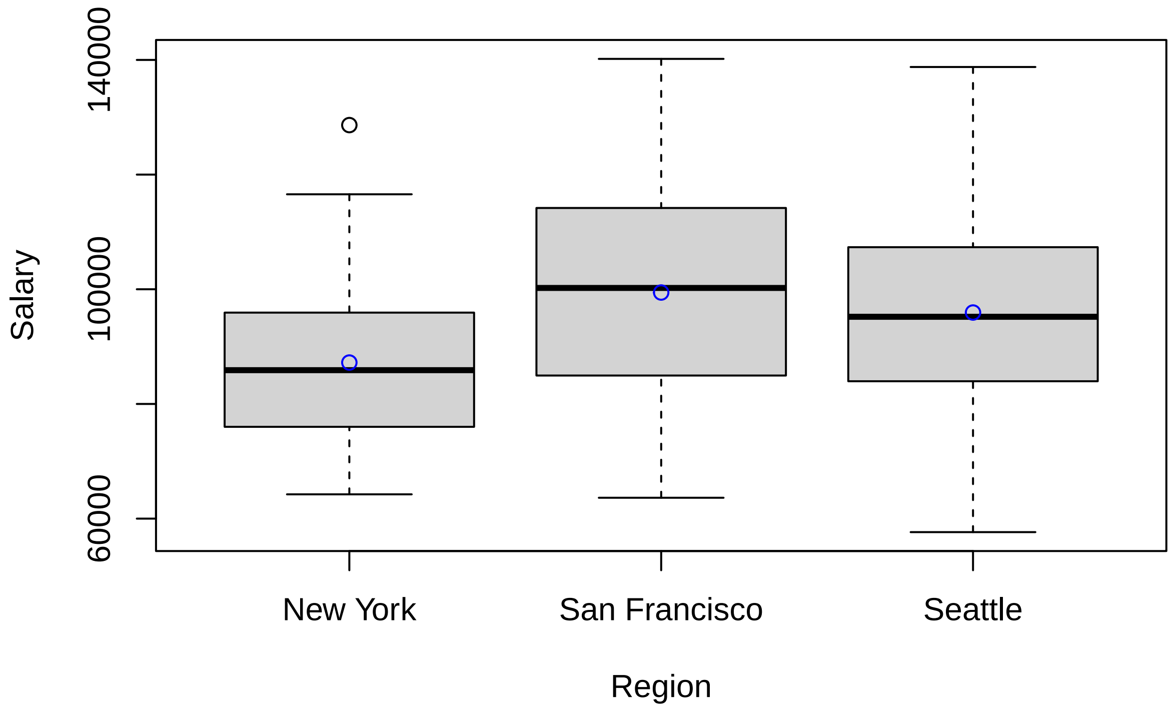
Boxplots below confirm that New York and BI Engineers are the lowest, while the other cities and professions are higher and close to each other. In regards to outliers, boxplots do not show any outliers except for New York city. which the salary at 130000. In my opinion, this outlier will not effect the analysis.

```r
# Plot Individual Boxplots with means

boxplot(Salary ~ Profession, data = dt_cln, ylab = 'Salary', xlab = 'Profession')
points(dt_cln[, mean(Salary), by=Profession], col = 'Red')
```



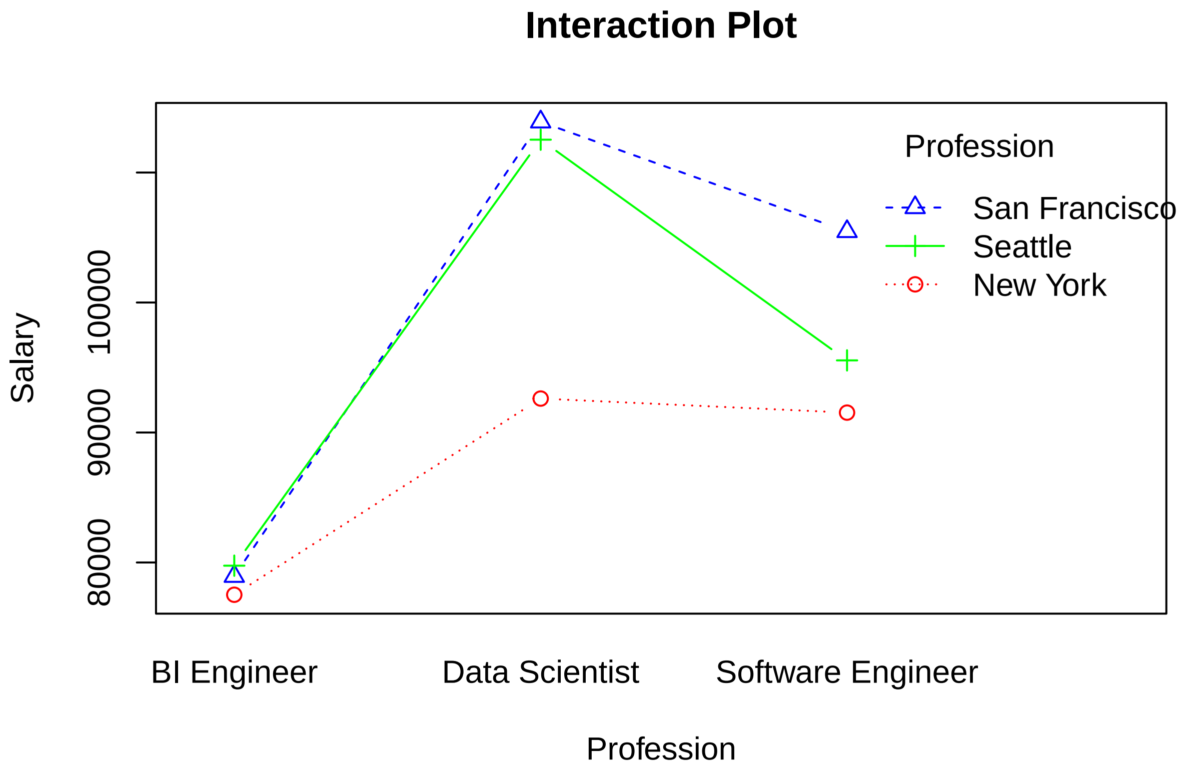
```
boxplot(Salary ~ Region, data = dt_cln, ylab = 'Salary', xlab = 'Region')
points(dt_cln[, mean(Salary), by=Region], col = 'Blue')
```



Interaction plot below shows interactions between region and profession factors, it shows strong interaction between San Francisco and Seattle where the jobs are BI Engineers and Data scientists. Moreover, it shows weak interaction for New York city. it Also shows that San Francisco and Seattle are close and New York.

```
# Create interaction plot looking at Region and Profession

interaction.plot(x.factor = dt$Profession,
                 trace.factor = dt$Region,
                 response = dt$Salary,
                 fun = mean,
                 type = "b",  # shows each point
                 main = "Interaction Plot",
                 legend = TRUE,
                 trace.label = "Profession",
                 xlab = "Profession",
                 ylab="Salary",
                 pch=c(1, 2, 3, 4),
                 col = c("Red", "Blue", "Green","Black"))
```



Below is two way ANOVA model where we fit both of our factors in the model. The model Shows that profession, region are significant, Also, the interaction profession and region is Also significant. Furthermore, the degree of freedom for profession and region interaction is more less than number of observations, so no need to modify the model.

```
fit <- aov(Salary ~ Profession * Region, data = dt)
summary(fit)
```

```
##                      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
```

TukeyHSD post hoc test below shows that Profession and region interaction is significant because of the adjusted p-value is less than 0.05 for most of the professions and regions. for example, Software Engineer in Seattle and BI Engineer in New York have adjusted p-value of 0.0000975 which is less than 0.05.

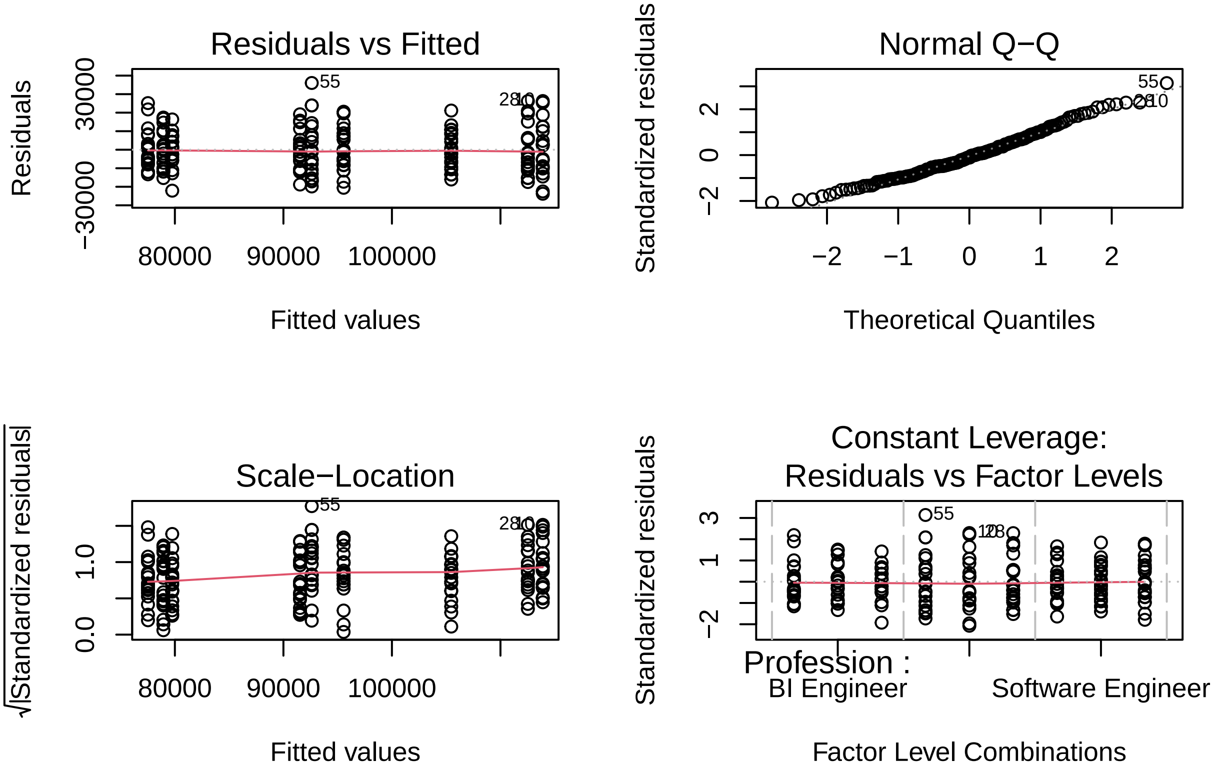
```
# Perform TukeyHSD to check if which interactions have a significant difference
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = Salary ~ Profession * Region, data = dt)
##
## $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
```

Below plots show distribution of the residuals, residuals vs fitted plot looks normal. scale location plot does not look good but we an accept it. Normal Q-Q shows that residuals are normally distributed. Although, it shows that there are outliers for observations number 55 and 28. lastly, leverage plot looks normal.

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



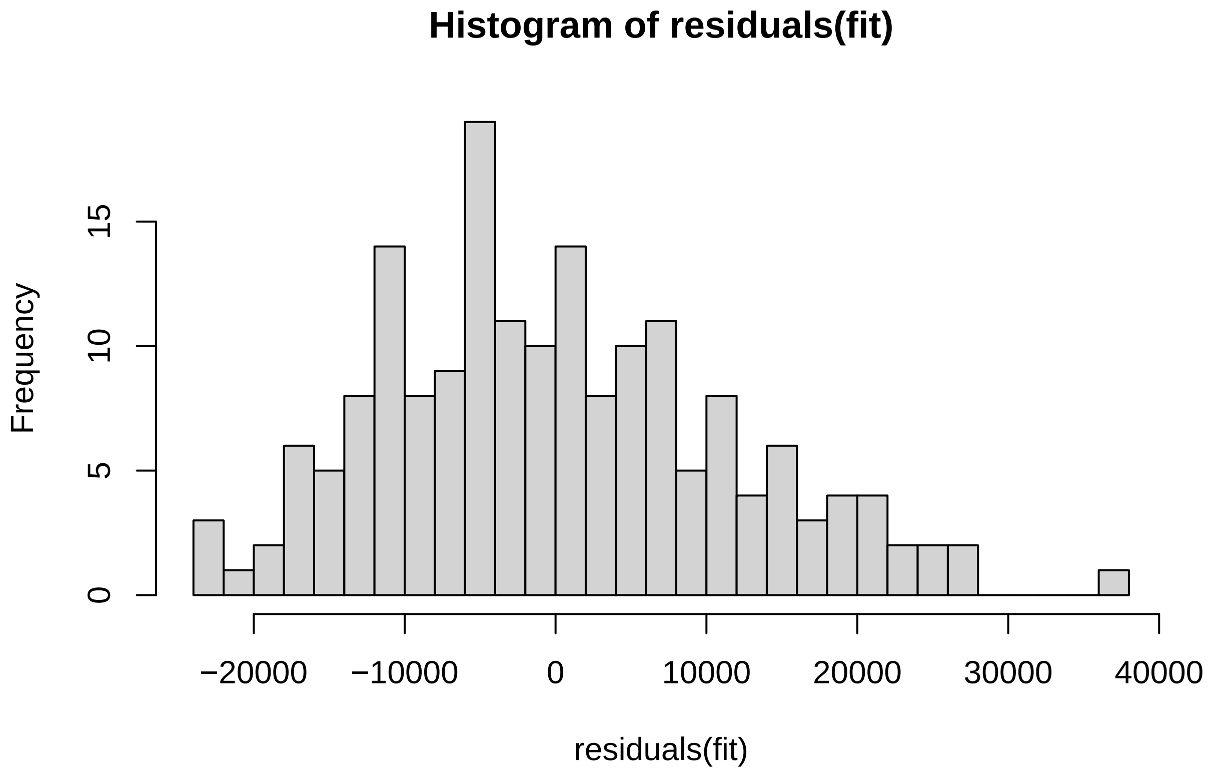
Shapiro test of residuals and the histogram of residuals show that residuals are normally distributed.

```
# Perform Shapiro test to see if residuals are normally distributed.

shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.98346, p-value = 0.03161
```

```
hist(residuals(fit), breaks=40)
```



Conclusion

Lastly, Null hypothesis is that there is no difference in means with any factors, which is rejected because both factors are significantly different in means, along side with the interaction of the factors. Alternative hypothesis is accepted because we have at least one factors that is significantly different. generally, as this test proved it, salaries are effected by location and profession, Also different location and different professions are strong factors on salaries.

References

Two way ANOVA. (n.d.). Retrieved from From the Expert.

Two Way ANOVA - MSDS660. (2021). Denver, CO, USA.