Two-way ANOVA

Assignment

Investigate salary by region (San Francisco, Seattle, New York) and Profession (Data Scientist, Software Engineer, BI Engineer) using a sample of 180 people combining regions and profession.

Research question:

Were there any differences in mean salary between professions and regions?

First, I prepared the environment for this assignment and imported the data using the following commands:

```
> rm(list=ls()) #Clear the environment
> setwd("YOUR_PATH") #Set working directory for the assignment
> getwd() #Check working directory

[1] "YOUR_PATH"
>

> #######Input data from a csv file
>
> salary<-read.csv("engineer.csv", header=TRUE)
```

I checked that the data imported correctly:

```
Console   Terminal ×
E:/Dropbox/RU DataScience/MSDS660/Week5/Assignment/
> salary<-read.csv("E:/Dropbox/RU DataScience/MSDS660/Week5/Assignment/engineer.csv", header=TRUE)
> salary
    X Salary       Profession        Region
1   1 126411   Data Scientist San Francisco
2   2 108402   Data Scientist San Francisco
3   3  99399   Data Scientist San Francisco
4   4  91381   Data Scientist San Francisco
5   5 105023   Data Scientist San Francisco
6   6 108944   Data Scientist San Francisco
7   7 123952   Data Scientist San Francisco
8   8 108217   Data Scientist San Francisco
9   9 103722   Data Scientist San Francisco
10 10 140179   Data Scientist San Francisco
11 11 116892   Data Scientist San Francisco
12 12 118661   Data Scientist San Francisco
13 13 116750   Data Scientist San Francisco
14 14 132715   Data Scientist San Francisco
15 15 139308   Data Scientist San Francisco
16 16 104449   Data Scientist San Francisco
17 17 126215   Data Scientist San Francisco
18 18  90123   Data Scientist San Francisco
```

The resulting table salary contains 180 observations of 4 variables.

I then checked the structure of the table to display the variable types:

```
> str(salary) #Display internal table structure
'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: 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 ...
```

```
179 179  89069       BI Engineer      New York
180 180  80685       BI Engineer      New York
> str(salary) #Display internal table structure
'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: 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 ...
> 
```

The table contains two categorical variables – factor **Profession** with 3 levels and factor **Region** also with 3 levels, that can be used as predictors in ANOVA model. The dependent variable **Salary** has the type int.

The next step is to explore the data and to check for the ANOVA model assumptions – independence, normality and homogeneity of variance.

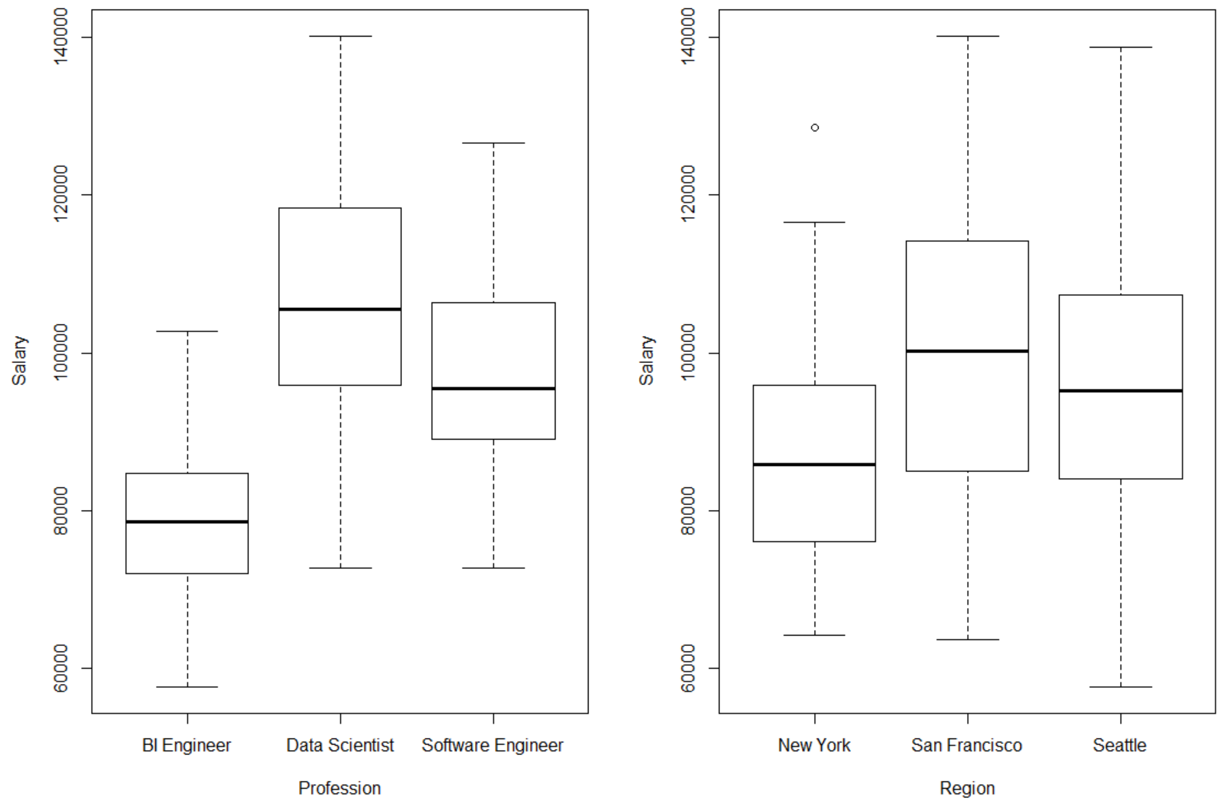
According to the assignment, a sample of 180 people combining region and profession are collected. There are no indications of how this sample was collected, but we are going to assume that the independence requirement is met for this dataset.

Next, I created side-by-side boxplots for Salary vs. Profession and Region for each treatment group using the following code:

```
> #Make side-by-side boxplots
>
> par(mfrow=c(1,2)) #Display 2 graphs in a row
> plot(Salary ~ Profession + Region, data=salary)
```

The resulting plots are presented below:

Two-way ANOVA



I visually inspected the boxplots for outliers, skewness and unequal variance.

For either of the boxplots there are no separated points that would indicated outliers. The boxes for groups BI Engineer, Data Scientist are roughly symmetrical, the box for Software Engineer suggests slightly higher values in the third quantile. The boxes for New York, San Francisco and Seattle are also roughly symmetrical. The extended whiskers for the Seattle boxplot suggest wider range of values for Seattle. On the Profession vs. Salary graph BI Engineer box is slightly smaller in size compared to Data Scientist and Software Engineer boxes. The boxes for New York and Seattle are equal in sizes, with the San Francisco box just a little bigger. So, there are no big discrepancies in box sizes indicating approximately equal variances.

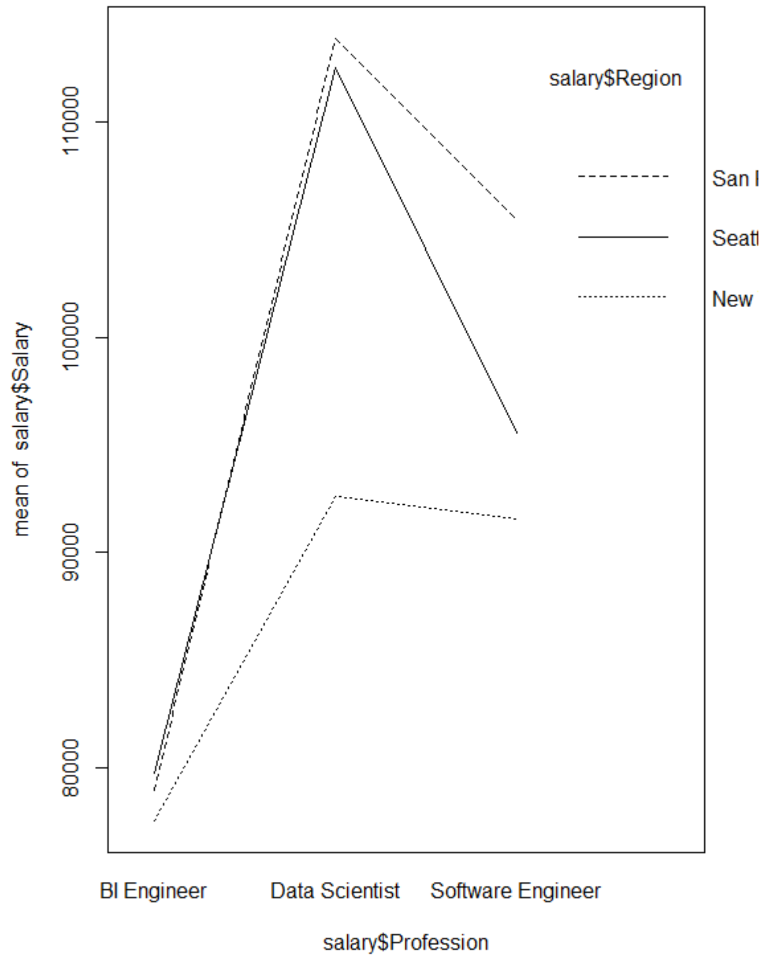
Overall, visual analysis in this case did not reveal any obvious problems or counterindications for using the two-way ANOVA model. So, I proceeded with additional testing of possible interaction.

I used the following code to create interaction plots:

```
> #Display interaction plots
> interaction.plot(salary$Profession, salary$Region, salary$Salary)
```

With the following output:

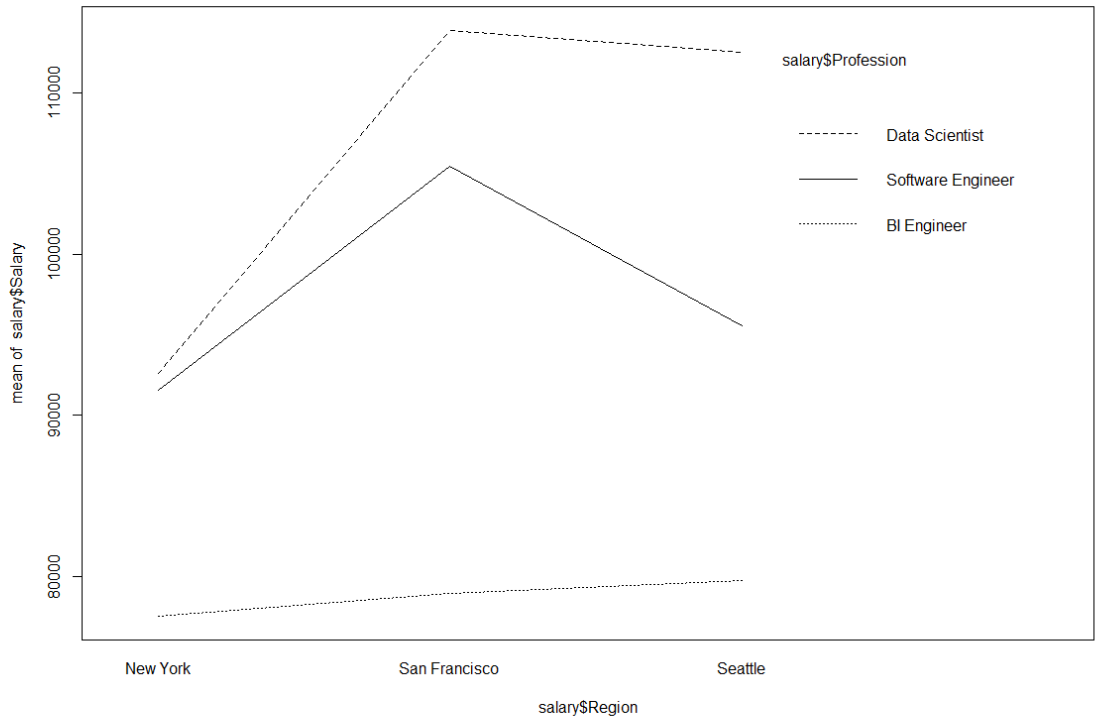
Two-way ANOVA



Reversed factors:

```
#Display interaction plots, reverse factors
> interaction.plot(salary$Region, salary$Profession, salary$Salary)
```

Two-way ANOVA



+

Since lines on the graphs are not parallel, both plots suggest presence of significant interaction between two factors (profession and region). It means that we need to include interaction factor in our ANOVA model along with the two main factors.

Hypothesis for the two-way ANOVA model:

H0:

- There is no difference in salary means for different levels of factor Profession;
- There is not difference salary means for different levels of the factor Region;
- There is no significant interaction between the two factors (Profession and Region).

Ha:

- There is difference in mean salary for different combinations of Profession and Region factors.

The next step is to fit the model.

```
> ####Fit two-way ANOVA model with interactions and display the results
>
> #Fit the model
> salarymodel<-lm(salary$Salary ~ salary$Profession + salary$Region + salary$Profession * salary$Reg
ion, data=salary)
> anova(salarymodel) #Display the ANOVA table for the model
```

Two-way ANOVA

```
Analysis of Variance Table

Response: salary$Salary
                                  Df     Sum Sq    Mean Sq F value    Pr(>F)    
salary$Profession                  2 2.3855e+10 1.1928e+10 86.0975 < 2.2e-16 ***
salary$Region                      2 4.7499e+09 2.3750e+09 17.1433 1.638e-07 ***
salary$Profession:salary$Region    4 3.0372e+09 7.5929e+08  5.4809 0.0003555 ***
Residuals                        171 2.3690e+10 1.3854e+08                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> ####Fit two-way ANOVA model with interactions and display the results
>
> #Fit the model
> salarymodel<-lm(salary$Salary ~ salary$Profession + salary$Region + salary$Profession * salary$Region, data=salary)
> anova(salarymodel) #Display the ANOVA table for the model
Analysis of Variance Table

Response: salary$Salary
                                  Df     Sum Sq    Mean Sq F value    Pr(>F)    
salary$Profession                  2 2.3855e+10 1.1928e+10 86.0975 < 2.2e-16 ***
salary$Region                      2 4.7499e+09 2.3750e+09 17.1433 1.638e-07 ***
salary$Profession:salary$Region    4 3.0372e+09 7.5929e+08  5.4809 0.0003555 ***
Residuals                        171 2.3690e+10 1.3854e+08                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

The above output of the model shows that there is a significant interaction effect between the two factors (F=5.4809, p-value = 0.0003555). P=value < 0.05 indicates that there is a statistically significant interaction effect between the two factors (Profession and Region) for at least one group.

The test for the main effect shows that both factors have significant effect on the dependent variable (Salary). The test for the main effect of Profession (F- statistic =86.0975 and p-value = 2.2e-16) returned a p-value considerably smaller than the significance level of 0.05 which indicates that there is a significant effect of Profession on Salary. The test for the main effect of Region (F-statistic = 17.1433, p-value = 1.638e-07) shows us that there is a significant effect of Region on resulting Salary.

Then I displayed more detailed summary of the model:

```
> #Display the detailed summary for the model
> summary (salarymodel)

Call:
lm(formula = salary$Salary ~ salary$Profession + salary$Region + 
    salary$Profession * salary$Region, data = salary)

Residuals:
   Min     1Q Median     3Q    Max 
-23776  -8369  -1215   7426  36023 

Coefficients:
                                                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                                      77519       2632  29.454  < 2e-16 **
*
salary$ProfessionData Scientist                                  15093       3722   4.055 7.61e-05 **
*
salary$ProfessionSoftware Engineer                               14011       3722   3.764 0.000229 **
*
```

```
salary$RegionSan Francisco                                     1421        3722    0.382 0.703029
salary$RegionSeattle                                           2236        3722    0.601 0.548786
salary$ProfessionData Scientist:salary$RegionSan Francisco    19866        5264    3.774 0.000221 **
*
salary$ProfessionSoftware Engineer:salary$RegionSan Francisco 12514        5264    2.377 0.018538 *
salary$ProfessionData Scientist:salary$RegionSeattle          17680        5264    3.359 0.000965 **
*
salary$ProfessionSoftware Engineer:salary$RegionSeattle        1783        5264    0.339 0.735213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11770 on 171 degrees of freedom
Multiple R-squared:  0.5719, Adjusted R-squared:  0.5518
F-statistic: 28.55 on 8 and 171 DF,  p-value: < 2.2e-16
```

```
Console   Terminal
E:/Dropbox/RU DataScience/MSDS660/Week5/Assignment/
> #Display the detailed summary for the model
> summary (salarymodel)

Call:
lm(formula = salary$Salary ~ salary$Profession + salary$Region +
    salary$Profession * salary$Region, data = salary)

Residuals:
   Min     1Q Median     3Q    Max
-23776  -8369  -1215   7426  36023

Coefficients:
                                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                      77519       2632  29.454  < 2e-16 ***
salary$ProfessionData Scientist                                  15093       3722   4.055 7.61e-05 ***
salary$ProfessionSoftware Engineer                               14011       3722   3.764 0.000229 ***
salary$RegionSan Francisco                                        1421       3722   0.382 0.703029
salary$RegionSeattle                                              2236       3722   0.601 0.548786
salary$ProfessionData Scientist:salary$RegionSan Francisco       19866       5264   3.774 0.000221 ***
salary$ProfessionSoftware Engineer:salary$RegionSan Francisco    12514       5264   2.377 0.018538 *
salary$ProfessionData Scientist:salary$RegionSeattle             17680       5264   3.359 0.000965 ***
salary$ProfessionSoftware Engineer:salary$RegionSeattle           1783       5264   0.339 0.735213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11770 on 171 degrees of freedom
Multiple R-squared:  0.5719,	Adjusted R-squared:  0.5518
F-statistic: 28.55 on 8 and 171 DF,  p-value: < 2.2e-16
```

Overall, the F-statistic for the model id 28.55 on 8 and 171 degrees of freedom with p-value of less than 2.2e—16. The resulting p-value is much smaller than the significance level of 0.05. It means that we need to reject the null hypothesis in favor of the alternative hypothesis, stating that not all means are equal and interaction effect is present.

Close look at the results indicates that there are statistically significant differences in means detected by the model for Data Scientists, Software Engineers and those working in San Francisco and Seattle. However, the t-statistic and p-values calculated by the ANOVA model can be unreliable in presence of the significant interaction effect. Therefore, we need to proceed with some additional post-hoc testing.

Post-Hoc testing

I used the following code for Tukey Honest Significance test in order to conduct pairwise comparisons.

```
> #Post-Hoc Testing
> #TukeyHSD test
> TukeyHSD(aov(salary$Salary ~ salary$Profession  + salary$Region, data=salary))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salary$Salary ~ salary$Profession + salary$Region, data = salary)

$`salary$Profession`
                                    diff       lwr       upr     p adj
Data Scientist-BI Engineer       27608.02  22274.56 32941.478 0.0000000
Software Engineer-BI Engineer    18776.57  13443.11 24110.028 0.0000000
Software Engineer-Data Scientist -8831.45 -14164.91 -3497.989 0.0003802

$`salary$Region`
                             diff       lwr       upr     p adj
San Francisco-New York 12214.900  6881.439 17548.361 0.0000006
Seattle-New York        8723.683  3390.222 14057.144 0.0004550
Seattle-San Francisco  -3491.217 -8824.678  1842.244 0.2714993
```

Below is the screenshot of the output:

E:/Dropbox/RU DataScience/MSDS660/Week5/Assignment/

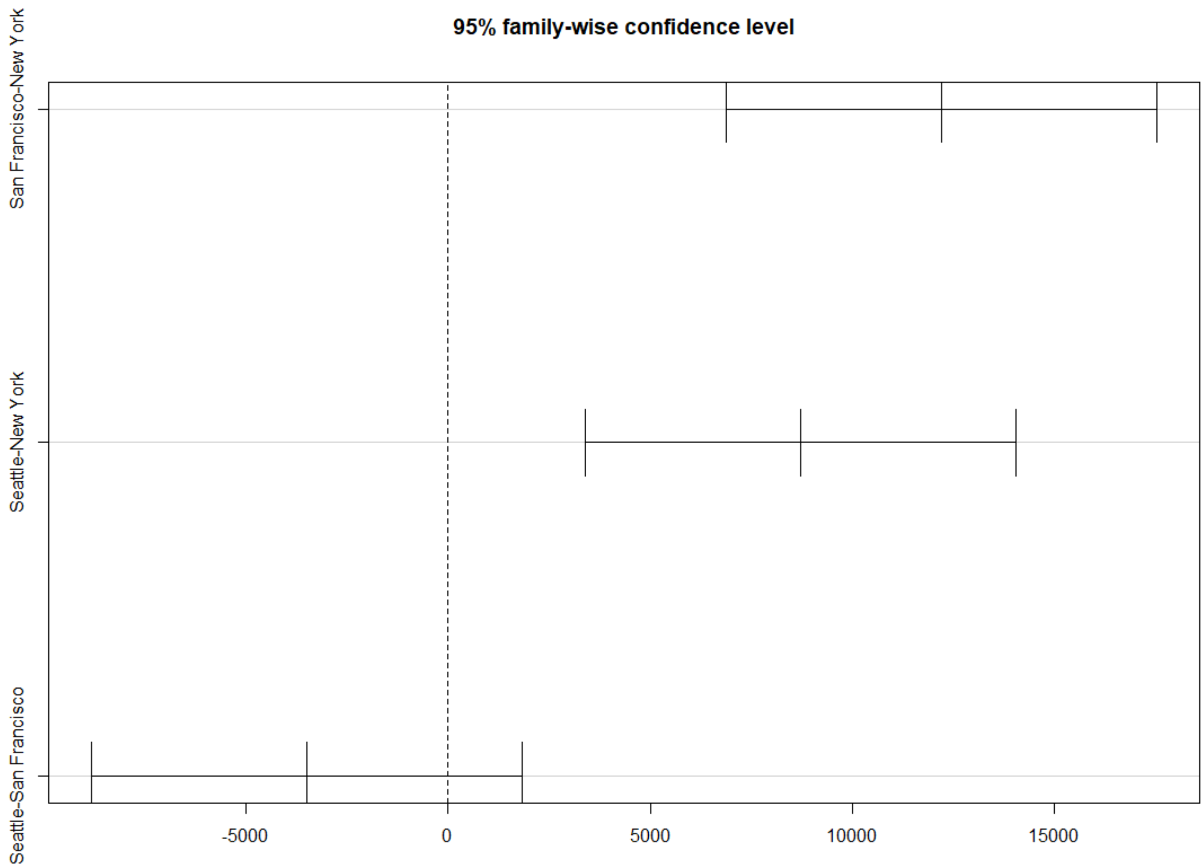

```
> #Post-Hoc Testing
> #TukeyHSD test
> TukeyHSD(aov(salary$Salary ~ salary$Profession  + salary$Region, data=salary))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salary$Salary ~ salary$Profession + salary$Region, data = salary)

$`salary$Profession`
                                    diff       lwr       upr     p adj
Data Scientist-BI Engineer       27608.02  22274.56 32941.478 0.0000000
Software Engineer-BI Engineer    18776.57  13443.11 24110.028 0.0000000
Software Engineer-Data Scientist -8831.45 -14164.91 -3497.989 0.0003802

$`salary$Region`
                             diff       lwr       upr     p adj
San Francisco-New York 12214.900  6881.439 17548.361 0.0000006
Seattle-New York        8723.683  3390.222 14057.144 0.0004550
Seattle-San Francisco  -3491.217 -8824.678  1842.244 0.2714993

> 
```


The above results indicate that there is a statistical difference for five of the pairwise comparisons. The p-values below the significance level of 0.05 indicate that there are differences in mean salary between the Data Scientists and BI Engineers group (p-value adjusted for multiple comparisons is close to 0), between Software Engineers and BI Engineers (p-value close to 0), between Software Engineers and Data Scientists (adjusted p-value of 0.0003802), between mean values for San Francisco and New York (very low p-value of 0.0000006), and Seattle and New-York (p value of 0.0004550). All these pair had p-values significantly lower than the 0.05. The only pair-wise comparison that did not indicate any statistically significant difference is comparison of means for Seattle and San Francisco – p-value of about 0.27.

I also used plot() command to visualize results of the Tukey HSD test:

```
> #visualise Tukey HSD test results
> plot(TukeyHSD(aov(salary$Salary ~ salary$Profession  + salary$Region, data=salary)))
```

As the graph below shows, Seattle -San Francisco is the only pair with confidence interval crossing zero, meaning no significant differences. All five remaining intervals do not cross zero which implies statistically significant differences in group means.



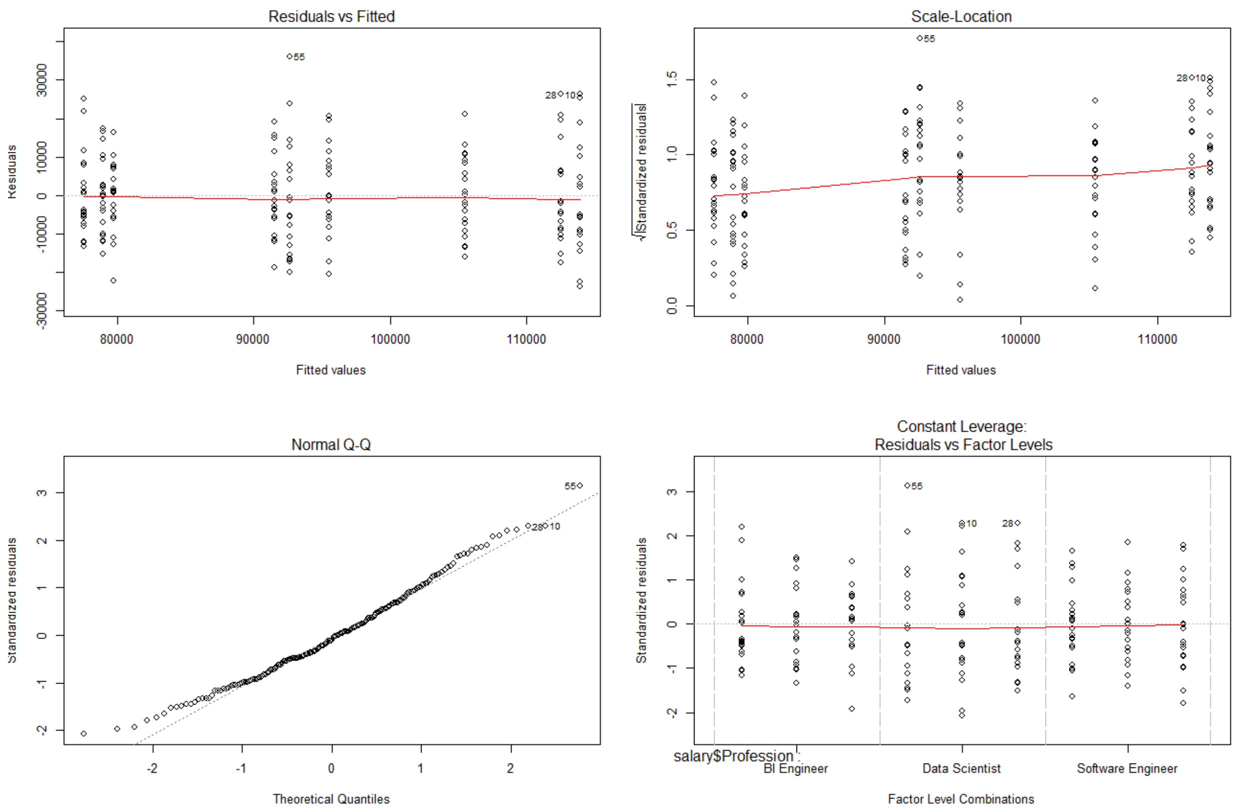
Model Diagnostic

In order to assess validity of the model's results, I proceeded with the diagnostic and constructed the following four plots for the model to visually assess normal distribution of the residuals and equal variance of the residuals.

```
> #####Model Diagnostic
> #check assumptions using graphs
>
> #diagnostic plots for salarymodel
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(salarymodel)
```

With the following output:

Two-way ANOVA



According to the Residuals vs. Fitted graph, the residuals variate around 0 and the red trend line is very close to horizontal. There are no real outliers that could have significantly distorted the models results.

Normal Q-Q plot for the residual distribution is very close to the normal line with some deviation in the first and fourth quadrants. So, I used Shapiro-Wilk test to check for normality of the distribution of the residuals.

Ho: The model residuals are normally distributed

H1: The model's residuals are not normally distributed.

I used the following code to run the test:

```
> shapiro.test(residuals(salarymodel))

        Shapiro-Wilk normality test

data:  residuals(salarymodel)
W = 0.98346, p-value = 0.03161
```

Two-way ANOVA

```
> shapiro.test(residuals(salarymodel))

        Shapiro-Wilk normality test

data:  residuals(salarymodel)
W = 0.98346, p-value = 0.03161
```

The results of the test indicate a p-value of 0.03 with is smaller than significance level of 0.05. It means that we have to reject the null hypothesis, the residuals are not completely following normal distribution.

To test the assumption of the homogeneity of variance for the residuals versus each of the factors I used the Levene Test with the following hypothesis

Ho: The

I used the following code:

```
> #Test homogenity of variance
> library(car)
> leveneTest(salarymodel$residuals ~ salary$Profession)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)   
group   2  5.8742 0.003388 **
      177                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
> #Test homogenity of variance
> leveneTest(salarymodel$residuals ~ salary$Region)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.1015 0.9035
      177               
> 
> #Test homogenity of variance
> leveneTest(salarymodel$residuals ~ salary$Profession *salary$Region)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)  
group   8  1.7669 0.08667 .
      171                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the following results:

```
> #Test homogenity of variance
> library(car)
> leveneTest(salarymodel$residuals ~ salary$Profession)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   2  5.8742 0.003388 **
      177
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #Test homogenity of variance
> leveneTest(salarymodel$residuals ~ salary$Region)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.1015 0.9035
      177
>
> #Test homogenity of variance
> leveneTest(salarymodel$residuals ~ salary$Profession *salary$Region)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   8  1.7669 0.08667 .
      171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The p-values above the significance level of 0.05 for Residuals vs. Region and Residuals vs. interaction of two factors do not allow us to reject the hypothesis of homogeneity However, for the variation of the Residuals vs. Profession p-value is approximately 0.034, which is smaller than the significance level of 0.05 meaning that the null hypothesis has to be rejected in favor of the alternative hypothesis (non-homogeneous variance).

<u>Conclusions:</u>

- Two-way ANOVA testing provide sufficient evidence allowing us to reject the null hypothesis, meaning that there is significant difference between salary means for al least one group and there is significant interaction effect.
- Since we rejected the null hypothesis, post-hoc testing was necessary to establish what pairs show significant difference in means. Tukey HSD test showed that Seattle-San Francisco was the only pair that did not show statistically significant difference in mean salary, all other five pairwise comparisons demonstrated significant differences.
- ANOVA model diagnostic showed some deviations from the model assumptions. The model residuals are not completely normal, but still loosely follow a normal distribution. While the requirements of homogeneity of the residuals mostly holds, the Levene test for the Residuals vs. Profession indicated non-homogeneous variance.