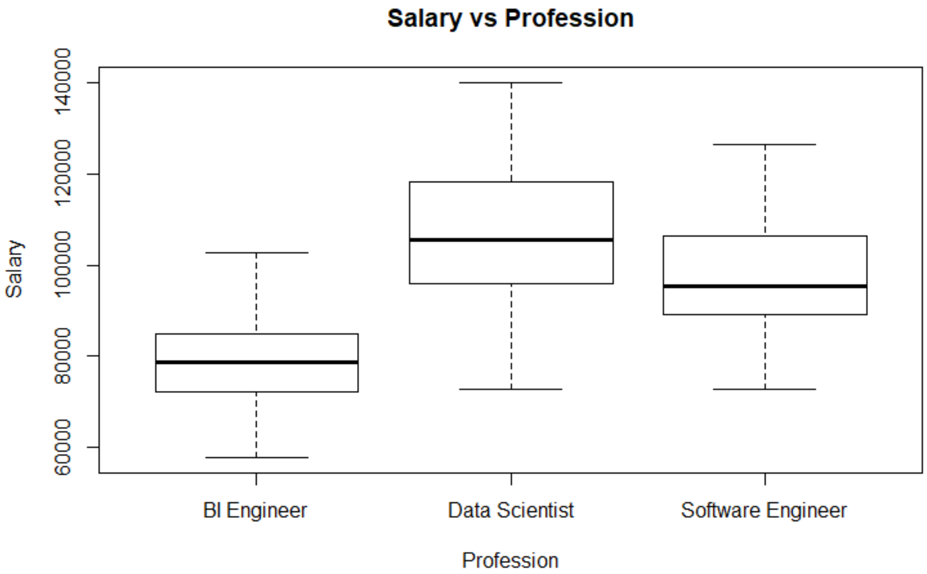
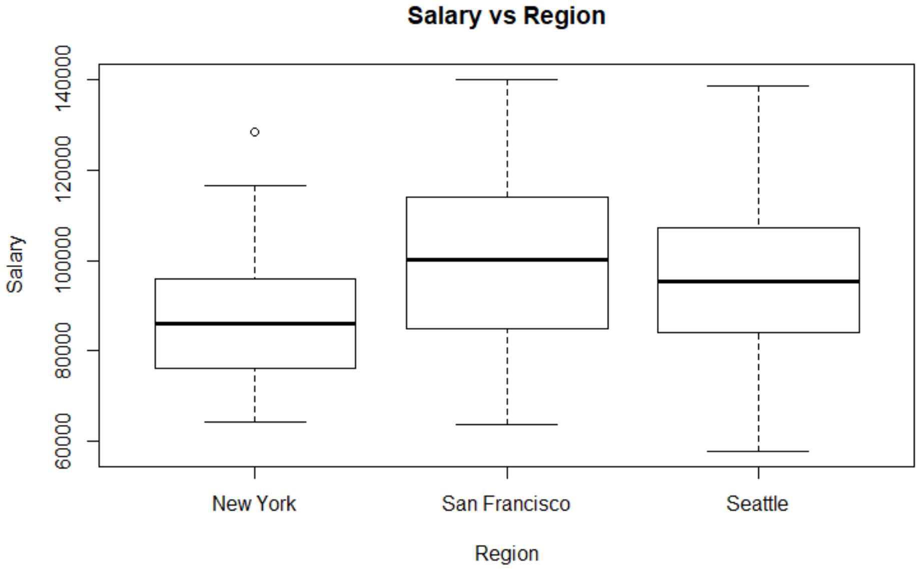
Alla Topp
MSDS 660
Project. Two-way ANOVA

A researcher wants to investigate salary by region (San Francisco, Seattle, New York) and Profession (Data Scientist, Software Engineer, BI engineer). A sample of 180 people combining region and profession are examined.

1. **the given data do the data exploration such as box plot of salary VS profession, and salary VS region, etc.,**





After examining the mean of Salary by Region and Profession by plotting two boxplot we can see that Data scientists make more money than Software and BI Engineers and that highest salaries are in San Francisco area in comparison with New York City and Seattle.

2. **State the hypotheses (in the form of H0: H1:)**

Alla Topp
MSDS 660
Project. Two-way ANOVA

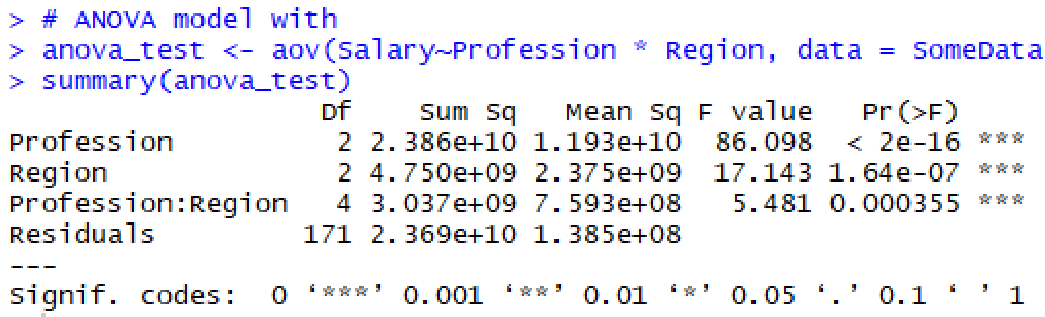
1) **H0:** There is no difference in the means of factor A (Profession).
   H1: the means are not equal.
2) **H0:** There is no difference in means of factor B (Region).
   H1: the means are not equal.
3) **H0:** There is no interaction between factor A (Profession) and factor B (Region).
   H1: There is an interaction between Profession and Region.

3. Construct an ANOVA table

alpha (significance level) = 0.05

Source	DF	Sum of Squares	Mean Square	F-Value	P-Value
Profession	2	2.386e+10	1.193e+10	86.098	02e-16
Region	2	4.750e+09	2.375e+09	17.143	1.64e-7
Profession * Region	4	3.037e+09	7.593e+08	5.481	0.000355
Error	171	2.369e+10	1.385e+08		
Total	179	5.5337e+10			

4. Do the complete analysis and summarize your findings using significance level at 0.05 (95% confidence level)?



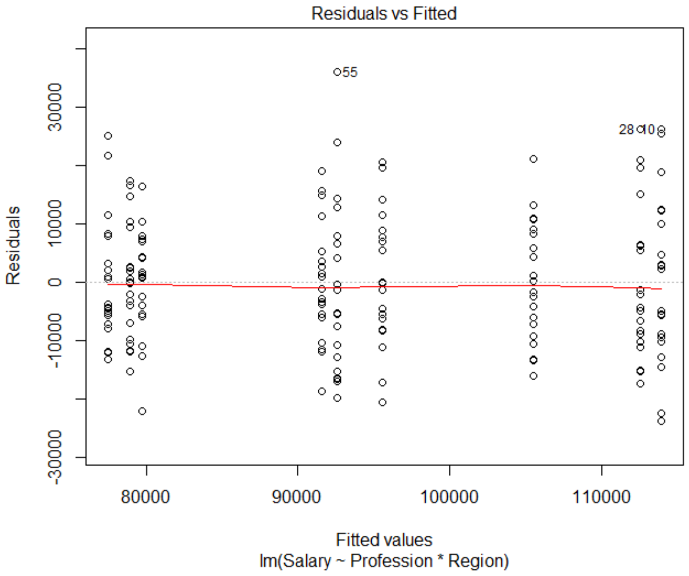
```
> # ANOVA model with
> anova_test <- aov(Salary~Profession * Region, data = SomeDataSet)
> summary(anova_test)
                   Df    Sum Sq   Mean Sq F value   Pr(>F)    
Profession          2 2.386e+10 1.193e+10  86.098  < 2e-16 ***
Region              2 4.750e+09 2.375e+09  17.143 1.64e-07 ***
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
```

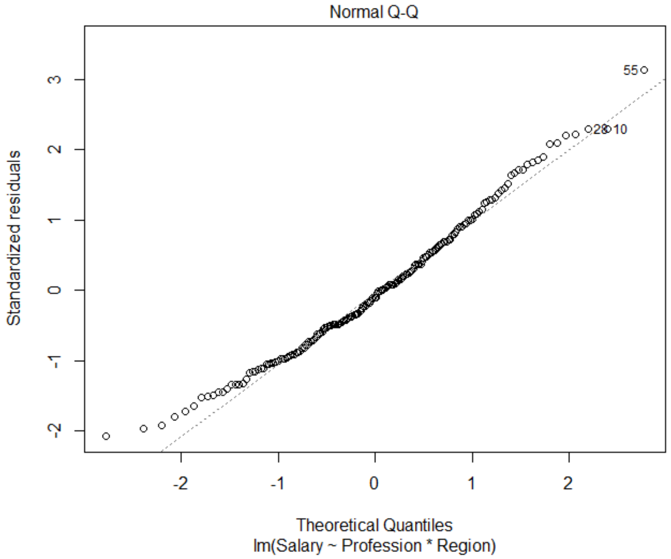
We performed a two-way ANOVA with the AOV function to examine the influence of the independent variables, Profession and Region, on dependent variable Salary. The output shows the p-value of factor Profession and Region, and the combination of these two factors rejects the null hypothesis. It can be seen that the two main effects (profession and region) are statistically significant, as well as their interaction (profession: region).

Since we rejected null hypothesis it would be a good idea to run post hoc comparison test (Tukey Test) to see the differences between professions and differences between regions.

ANOVA assumes that the data are normally distributed and the variance across groups are homogeneous. We can check that with some diagnostic plots.

Alla Topp
MSDS 660
Project. Two-way ANOVA

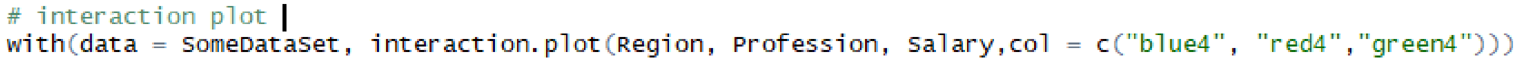




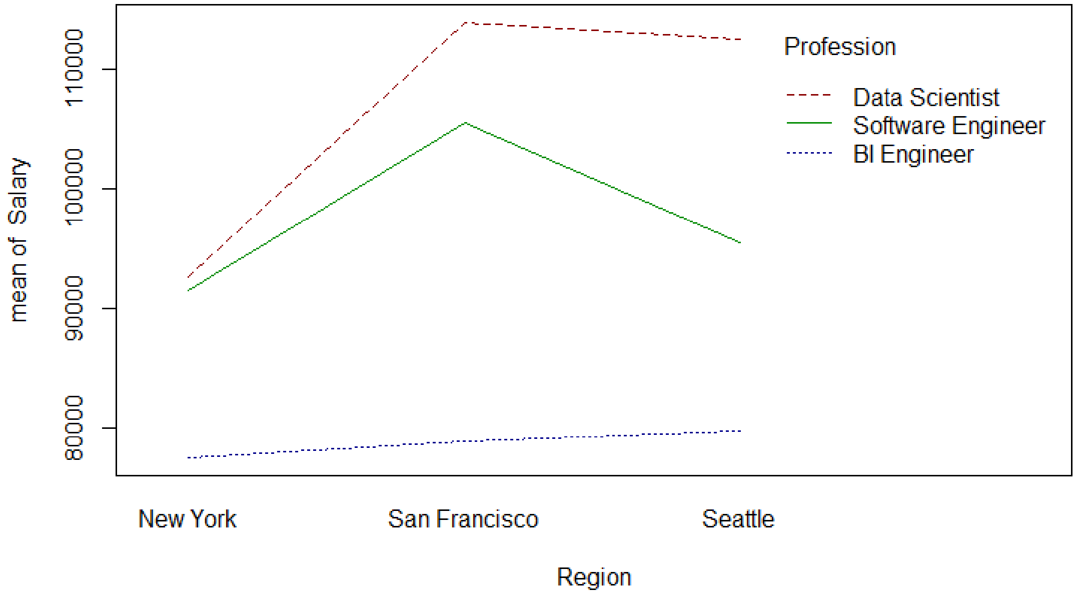
Based on the first plot, there is no evident relationships between residuals and fitted values (the mean of each groups), which is good. So, we can assume the homogeneity of variances.

Looking at the second plot, as all the points fall approximately along this reference line, we can assume normality.

5. **Indicate which effects are significant, if any. Show your plots (e.g. interaction effect) and analyze them.**

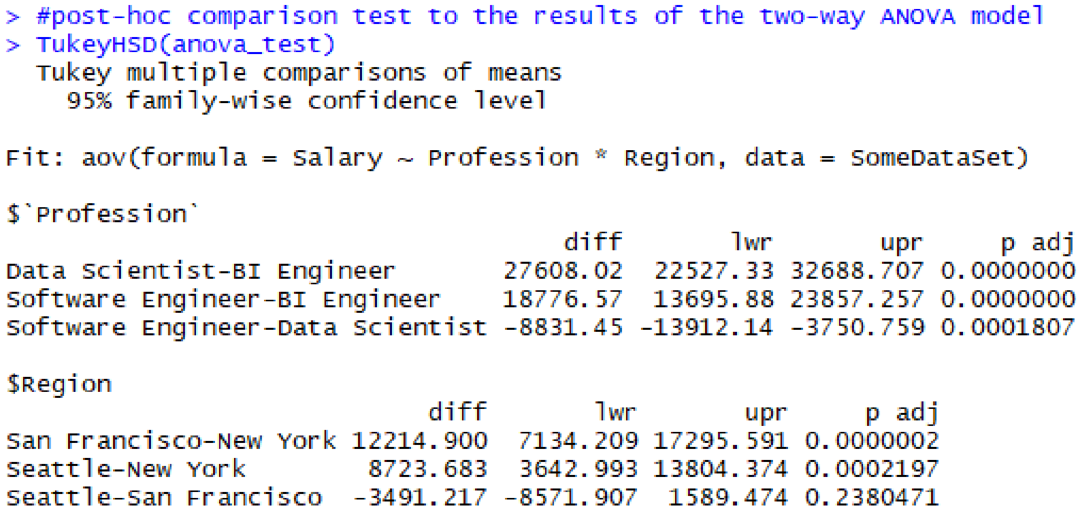


```
# interaction plot
with(data = SomeDataSet, interaction.plot(Region, Profession, Salary,col = c("blue4", "red4","green4")))
```



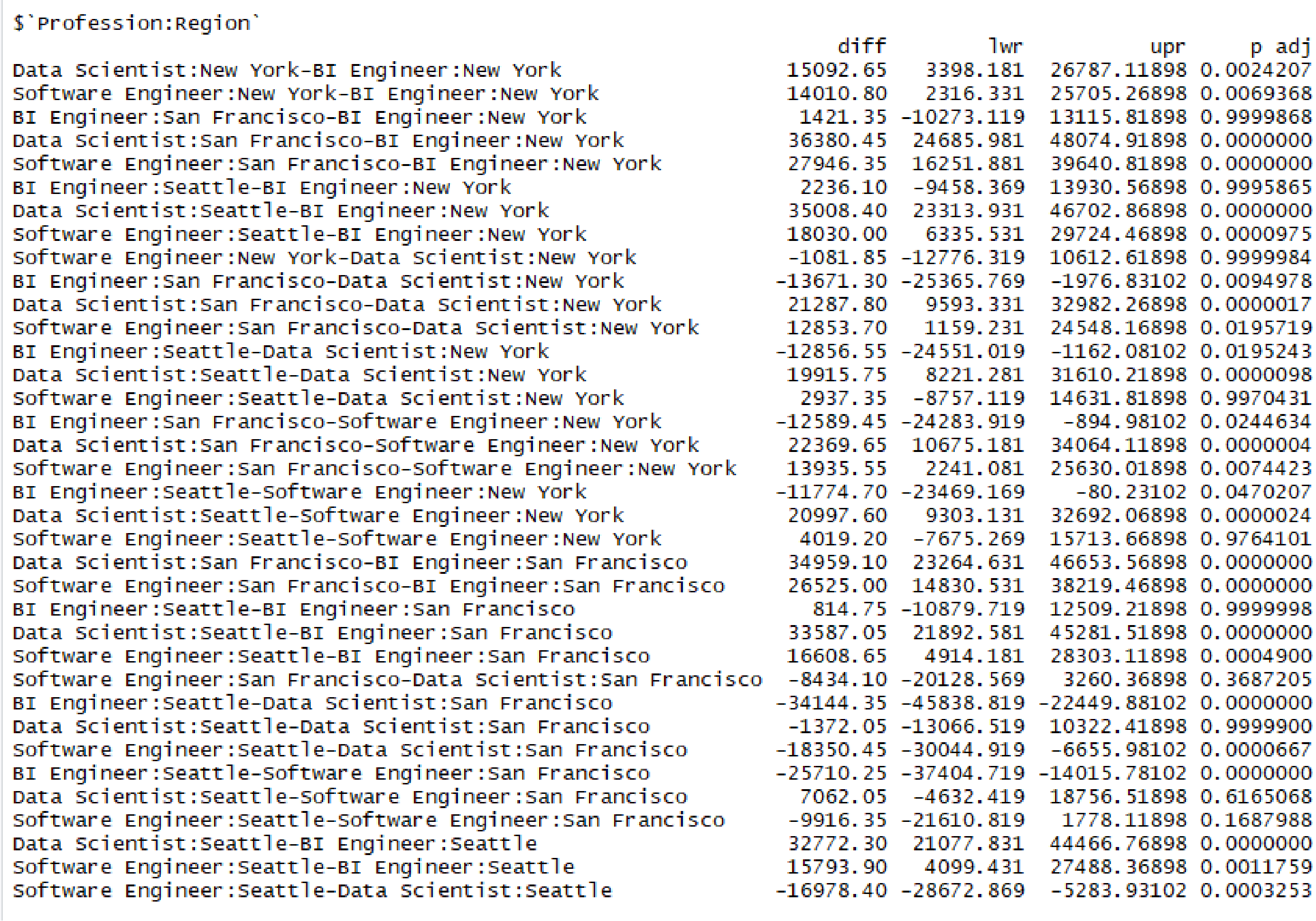
Alla Topp
MSDS 660
Project. Two-way ANOVA

We applied an interaction plot to visualize the change of salary in regard to different regions and professions. plot shows us that Profession and Region do have an effect on the mean of Salary. In other words, both Profession and Region affect the average salary of an engineer.



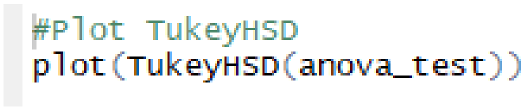
```
> #post-hoc comparison test to the results of the two-way ANOVA model
> TukeyHSD(anova_test)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Salary ~ Profession * Region, data = SomeDataSet)

$`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
```

To examine which two populations have the largest differences, we performed a post-hoc analysis, which revealed that a data scientist from San Francisco has a much higher salary than a BI engineer in New York.

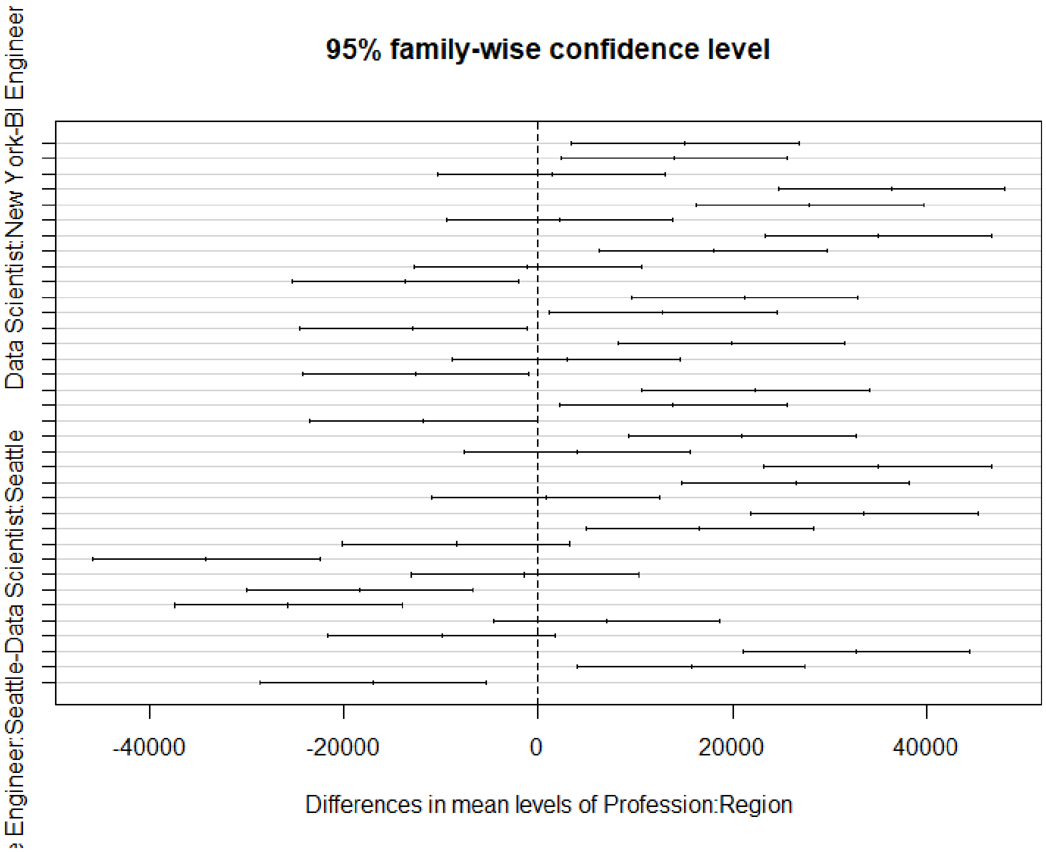


```
$`Profession:Region`
                                                          diff        lwr         upr     p adj
Data Scientist:New York-BI Engineer:New York          15092.65   3398.181  26787.11898 0.0024207
Software Engineer:New York-BI Engineer:New York       14010.80   2316.331  25705.26898 0.0069368
BI Engineer:San Francisco-BI Engineer:New York         1421.35 -10273.119  13115.81898 0.9999868
Data Scientist:San Francisco-BI Engineer:New York     36380.45  24685.981  48074.91898 0.0000000
Software Engineer:San Francisco-BI Engineer:New York  27946.35  16251.881  39640.81898 0.0000000
BI Engineer:Seattle-BI Engineer:New York               2236.10  -9458.369  13930.56898 0.9995865
Data Scientist:Seattle-BI Engineer:New York           35008.40  23313.931  46702.86898 0.0000000
Software Engineer:Seattle-BI Engineer:New York        18030.00   6335.531  29724.46898 0.0000975
Software Engineer:New York-Data Scientist:New York    -1081.85 -12776.319  10612.61898 0.9999984
BI Engineer:San Francisco-Data Scientist:New York    -13671.30 -25365.769  -1976.83102 0.0094978
Data Scientist:San Francisco-Data Scientist:New York  21287.80   9593.331  32982.26898 0.0000017
Software Engineer:San Francisco-Data Scientist:New York 12853.70   1159.231  24548.16898 0.0195719
BI Engineer:Seattle-Data Scientist:New York          -12856.55 -24551.019  -1162.08102 0.0195243
Data Scientist:Seattle-Data Scientist:New York        19915.75   8221.281  31610.21898 0.0000098
Software Engineer:Seattle-Data Scientist:New York      2937.35  -8757.119  14631.81898 0.9970431
BI Engineer:San Francisco-Software Engineer:New York -12589.45 -24283.919   -894.98102 0.0244634
Data Scientist:San Francisco-Software Engineer:New York 22369.65  10675.181  34064.11898 0.0000004
Software Engineer:San Francisco-Software Engineer:New York 13935.55   2241.081  25630.01898 0.0074423
BI Engineer:Seattle-Software Engineer:New York       -11774.70 -23469.169    -80.23102 0.0470207
Data Scientist:Seattle-Software Engineer:New York     20997.60   9303.131  32692.06898 0.0000024
Software Engineer:Seattle-Software Engineer:New York   4019.20  -7675.269  15713.66898 0.9764101
Data Scientist:San Francisco-BI Engineer:San Francisco 34959.10  23264.631  46653.56898 0.0000000
Software Engineer:San Francisco-BI Engineer:San Francisco 26525.00  14830.531  38219.46898 0.0000000
BI Engineer:Seattle-BI Engineer:San Francisco           814.75 -10879.719  12509.21898 0.9999998
Data Scientist:Seattle-BI Engineer:San Francisco      33587.05  21892.581  45281.51898 0.0000000
Software Engineer:Seattle-BI Engineer:San Francisco   16608.65   4914.181  28303.11898 0.0004900
Software Engineer:San Francisco-Data Scientist:San Francisco  -8434.10 -20128.569   3260.36898 0.3687205
BI Engineer:Seattle-Data Scientist:San Francisco     -34144.35 -45838.819 -22449.88102 0.0000000
Data Scientist:Seattle-Data Scientist:San Francisco   -1372.05 -13066.519  10322.41898 0.9999900
Software Engineer:Seattle-Data Scientist:San Francisco -18350.45 -30044.919  -6655.98102 0.0000667
BI Engineer:Seattle-Software Engineer:San Francisco  -25710.25 -37404.719 -14015.78102 0.0000000
Data Scientist:Seattle-Software Engineer:San Francisco  7062.05  -4632.419  18756.51898 0.6165068
Software Engineer:Seattle-Software Engineer:San Francisco  -9916.35 -21610.819   1778.11898 0.1687988
Data Scientist:Seattle-BI Engineer:Seattle            32772.30  21077.831  44466.76898 0.0000000
Software Engineer:Seattle-BI Engineer:Seattle         15793.90   4099.431  27488.36898 0.0011759
Software Engineer:Seattle-Data Scientist:Seattle     -16978.40 -28672.869  -5283.93102 0.0003253
```

Alla Topp
MSDS 660
Project. Two-way ANOVA



```
#Plot TukeyHSD
plot(TukeyHSD(anova_test))
```


95% family-wise confidence level
e Engineer:Seattle-Data Scientist:Seattle
Data Scientist:New York-BI Engineer
Differences in mean levels of Profession:Region