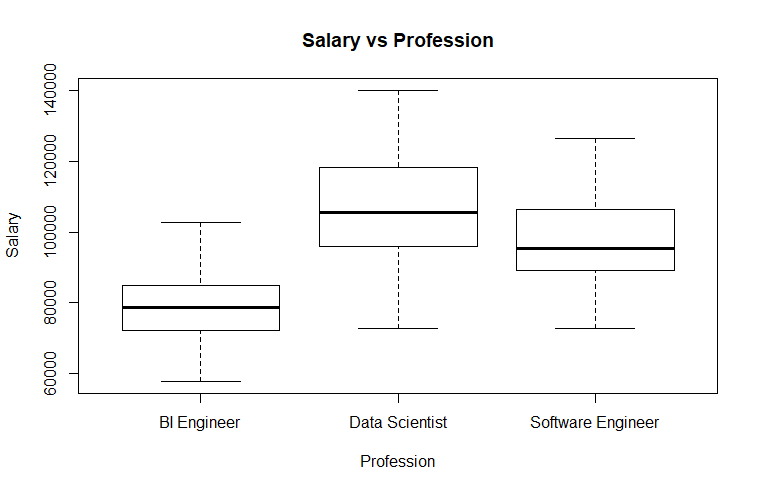
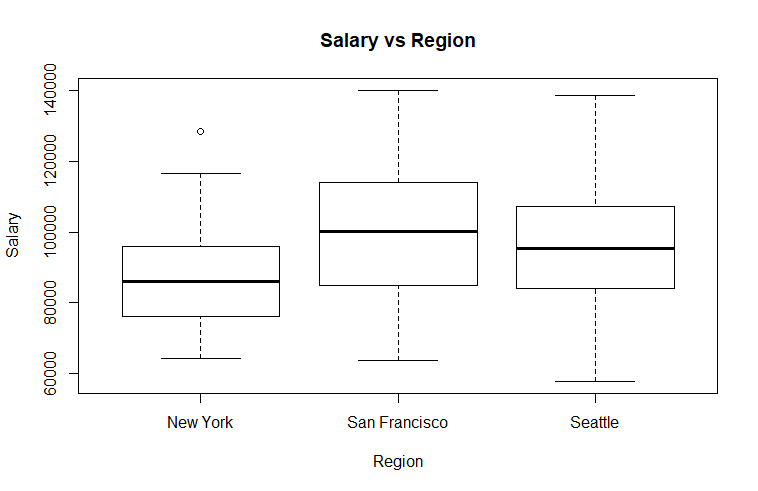
A researcher wants to investigate salary by region (San Francisco, Seattle, New York) and Profession (Data Scientist, Software Engineer, BI engineer). A sample of 180 people combining region and profession are examined.

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.

1. **State the hypotheses (in the form of H0: H1:)**
2. **H0:** There is no difference in the means of factor A (Profession).

**H1:** the means are not equal.

1. **H0:** There is no difference in means of factor B (Region).

**H1:** the means are not equal.

1. **H0:** There is no interaction between factor A (Profession) and factor B (Region).

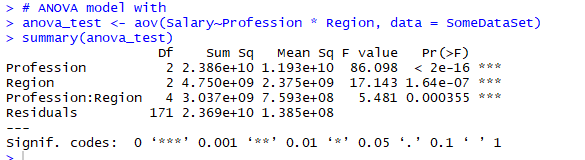
**H1:** There is an interaction between Profession and Region.

1. **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 |  |  |  | |

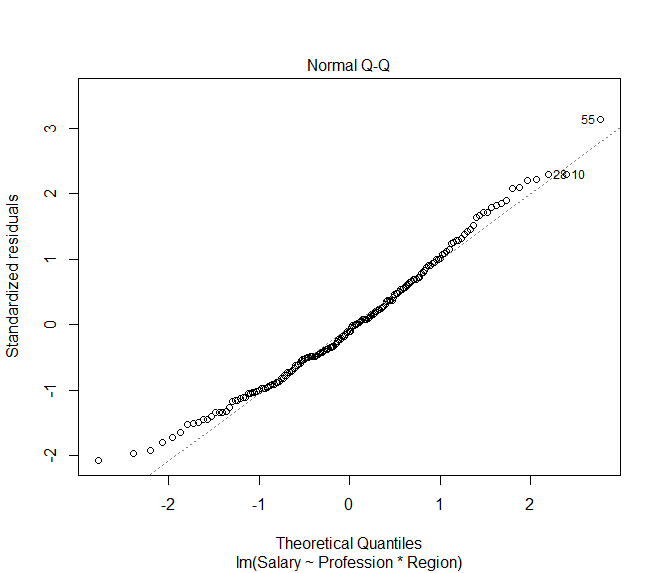
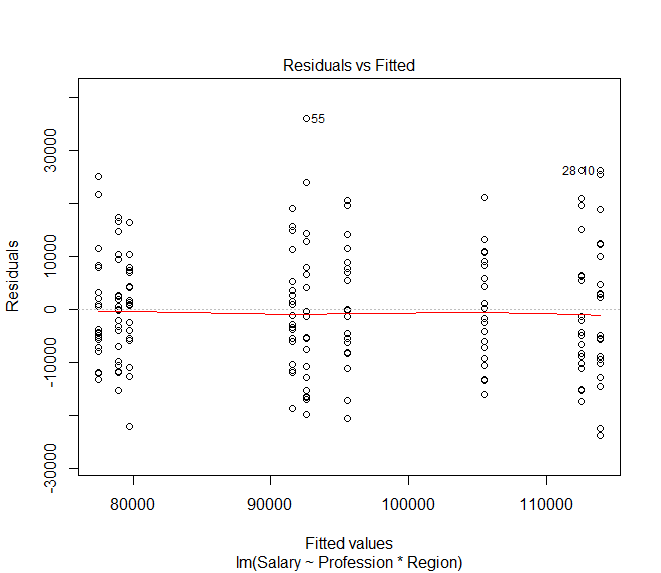
1. **Do the complete analysis and summarize your findings using significance level at 0.05 (95% confidence level).**



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.

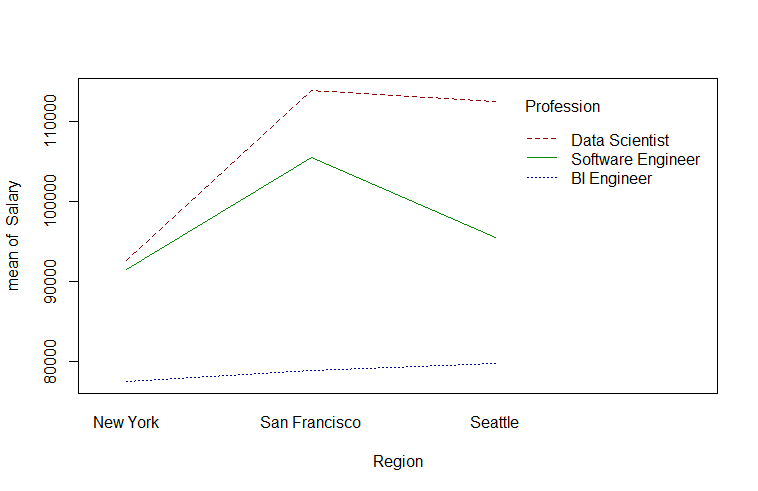


Based on the first plot, there is no evident relationships between residuals and fitted values (the mean of each groups), which is good. So, we can assume the homogeneity of variances.

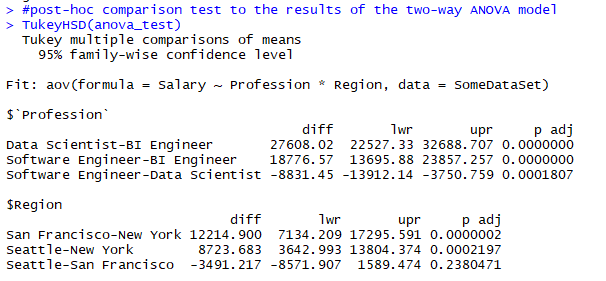
Looking at the second plot, as all the points fall approximately along this reference line, we can assume normality.

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





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.



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.

