Two-Way ANOVA

Taylor Shrode

4/12/2020

MSDS 660

Two-way ANOVA is an extension of one-way ANOVA. One-way ANOVA has one independent variable, that has 2 levels, that affects a dependent variable. Whereas, two-way ANOVA has two independent (nominal) variables, that can have multiple levels, that affects one dependent (measurement) variable (Anova test, n.d.). *Levels* are different groups within the same nominal variable (Anova test, n.d.). The nominal variables are often called *factors* or *main effects*, and each of these nominal variables is found in combination with each value of the other variable (McDonald, 2014). Thus, to find the number of treatment groups, we multiply the number of levels in Factor A (*a*) by the number of levels in Factor B (*b*). In other words, there are *ab* treatment combinations in a two-factor experiment (WorldClass FTE).

Recall, one-way ANOVA calculates a significant or insignificant result, meaning that either the two means from the two independent variables are equal, or they are not (Anova test, n.d.). Two-way ANOVA calculates a main effect and an interaction effect. The main effect is like the results in one-way ANOVA where each factor is considered separately (Anova test, n.d.). An interaction is the effect one independent variable has on the other independent variables (Jones, 2020). The interaction effect is when all factors are considered at the same time and are easier to test when there is more than one observation in each cell (Anova test, n.d.).

The procedure for two-way ANOVA follows the steps below (WorldClass FTE):

1. Check two-way ANOVA assumptions.
2. Test for interactions and perform a F-test for interaction effect.
    a. If there is an interaction, use an interaction plot to visualize the graph. Then, create a combination of factors and apply Tukey's HSD to examine which groups are different.
    b. If there is no interaction, do main effects testing on each factor. Compare these individual levels by applying t-test with Bonferroni correction for the number of comparisons made.

For this assignment, we will be using a dataset containing a sample of 180 people combining region and profession. These samples are used to investigate salary by region (San Francisco, Seattle, and New York) and profession (Data Scientist, Software Engineer, and BI Engineer). Using two-way ANOVA, we explore the effect region and profession has on salary. Both of our factors have three levels, meaning there are nine treatment combinations. To begin our analysis, we need to upload our data into R. This can be done several different ways, but we will import the data into R by reading the file from the internet (Reading data from txt|csv files, n.d.).

```
> my_data <- read.csv("https://raw.githubusercontent.com/ywchiu/rcookbook/master/chapt
er5/engineer.csv")
> View(my_data)
>
```

To ensure that our data was uploaded properly, we use the **View(<dataset>)** command.

	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

Now, we can use the **summary()** function to view a summary of our data and use the **str()** function to view the structure of our dataset.

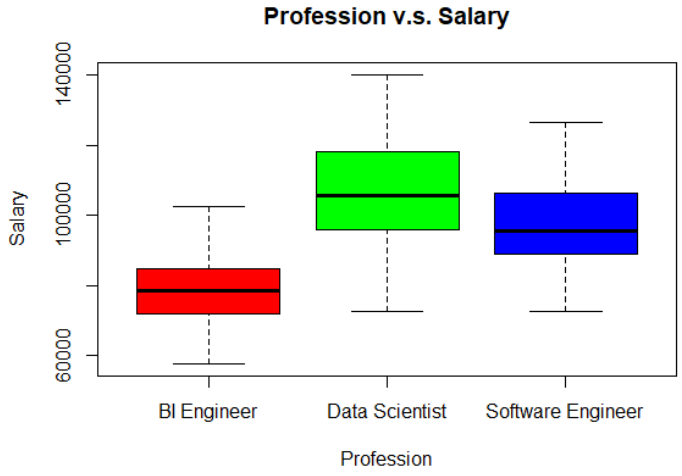
```
> summary(my_data)
       X                Salary                   Profession           Region
 Min.   :  1.00   Min.   : 57646   BI Engineer      :60   New York     :60
 1st Qu.: 45.75   1st Qu.: 80409   Data Scientist   :60   San Francisco:60
 Median : 90.50   Median : 92284   Software Engineer:60   Seattle      :60
 Mean   : 90.50   Mean   : 94199
 3rd Qu.:135.25   3rd Qu.:105932
 Max.   :180.00   Max.   :140179
> str(my_data)
'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 14017
9 ...
 $ 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
 ...
>
```

Notice that our 180-sample dataset is equally split between each factor level. We can further investigate to determine whether our experiment has a balanced or unbalanced design. We can do this by generating a frequency table.

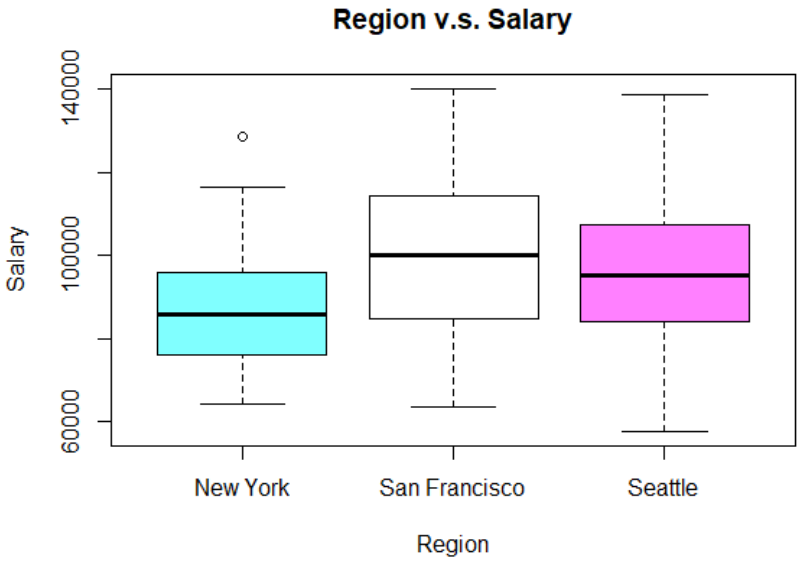
```
> table(my_data$Region, my_data$Profession)

               BI Engineer Data Scientist Software Engineer
  New York              20             20                20
  San Francisco         20             20                20
  Seattle               20             20                20
>
```

Our experiment has a balanced design because there is an equal number of observations for all possible level combinations (*Two-way anova test in r*, n.d.).

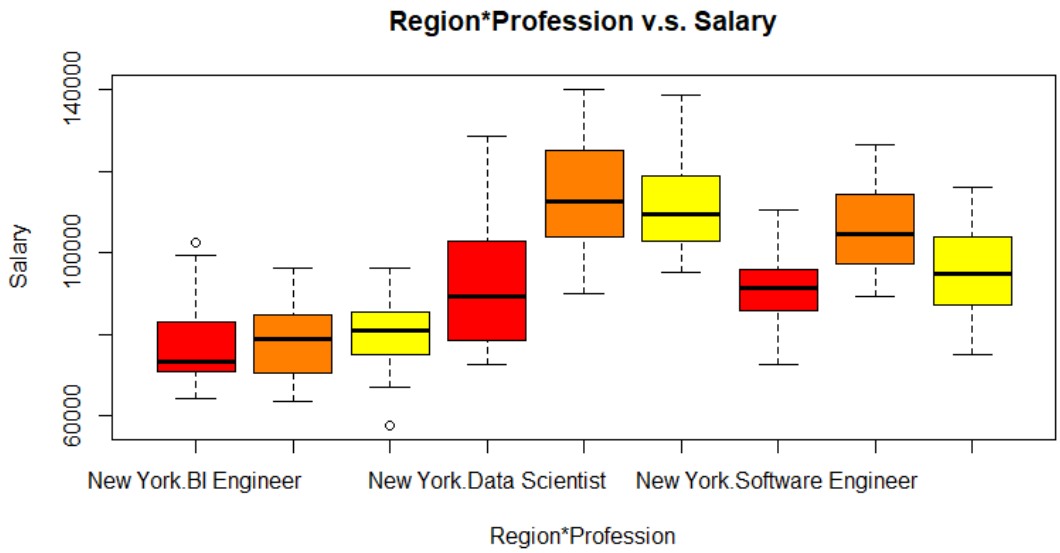
Now, we can better visualize our data by plotting boxplots of the salary factor in regard to profession and region. First, we plot salary against profession using the command **boxplot(Salary~Profession, data = my_data, xlab='Profession', ylab = "Salary", main="Profession v.s. Salary", col = rainbow(3)).**



From this, we can gather that the profession “BI Engineer” has the smallest range of salary while “Data Scientist” has the widest range of salary. We can also see that the medians do not overlap. Next, we plot salary against region using the command **boxplot(Salary~Region, data = my_data, xlab='Region', ylab = "Salary", main="Region v.s. Salary", col = cm.colors(3)).**



We can see that the range of salaries in "Seattle" are greater than the other two regions. We can also produce a boxplot of Salary versus the combination of Region and Profession.

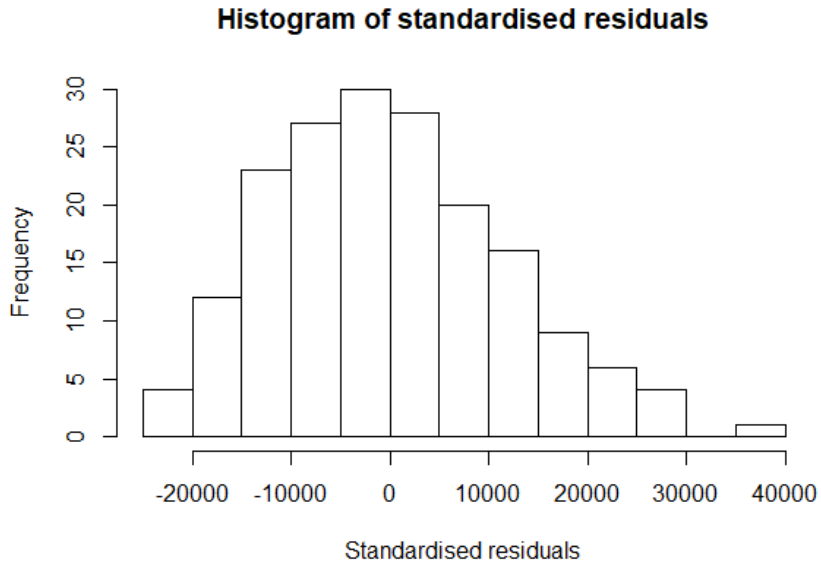


This plot implies that Data Scientists have a higher salary than the other professions. Now, we can begin the two-way ANOVA procedure. First, we need to check the assumptions of two-way ANOVA. The assumptions that need to be met with two-way ANOVA are like the ones of one-way ANOVA. The assumptions for two-way ANOVA are (Jones, 2020):

1. The populations from which the samples were obtained must be normally distributed.

2. The samples are independent.
3. Variances of the populations are equal.

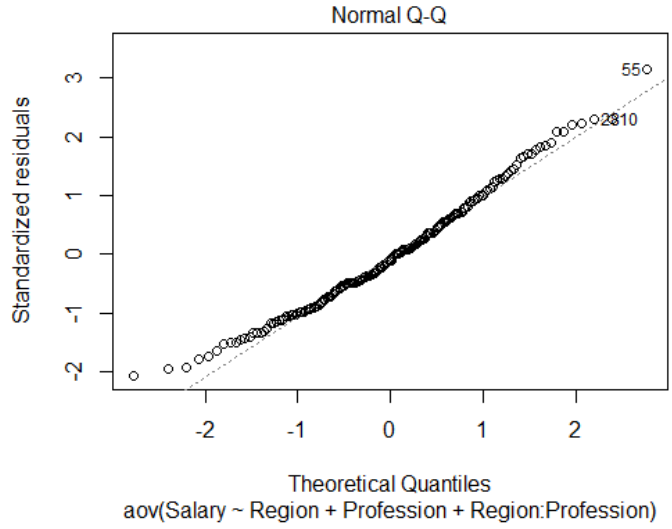
We begin by checking for normality. This can be done by creating a histogram of the residuals and conducting a Shapiro-Wilk test. The histogram of the residuals can be found below.



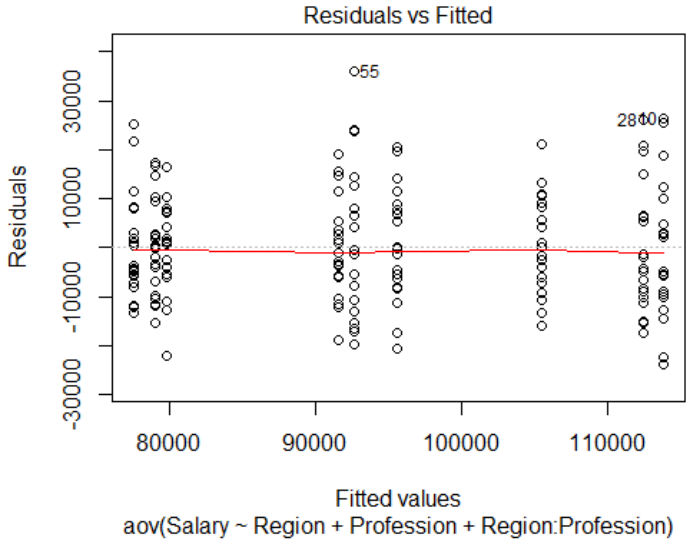
The histogram is slightly skewed to the right, indicating that normality may be violated. We can confirm this with the Shapiro-Wilk test. By using the **shapiro.test()** function in R, we can test for normality.

```
> shapiro.test(residuals)

        Shapiro-Wilk normality test

data:  residuals
W = 0.98346, p-value = 0.03161

>
```

The Shapiro-Wilk test gives a p-value less than our level of significance. This indicates that we do not have normality. In addition to these methods, we can view the Normal Q-Q plot of our model.



The normality probability plot of the residuals follows a straight line, which indicates normality. Now, we test for homogeneity of variances, which can be done by examining the residuals versus fits plot and using Levene's test.



The residuals versus fits plot indicates no relationship between the residuals and the fitted values, so we can assume homogeneity of variances. We can verify this conclusion with Levene's test.

```
> library(car)
Loading required package: carData
> leveneTest(interaction_aov)
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
> 
```

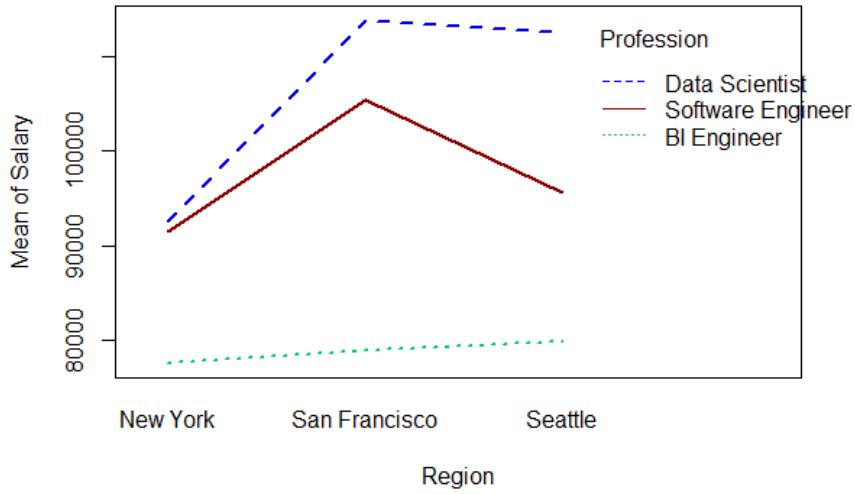
The p-value is not less than our level of significance, which means there is no evidence that the variance across groups is statistically significantly different.

Next, we test for interactions, and this is done by using interaction plots. An interaction plot is a plot of means of the response against the levels of one factor, with lines from the other factor (Two-way interaction plot, n.d.). The built-in **interaction.plot()** function in R is one way to create these plots. For this experiment, there are two possible plots:

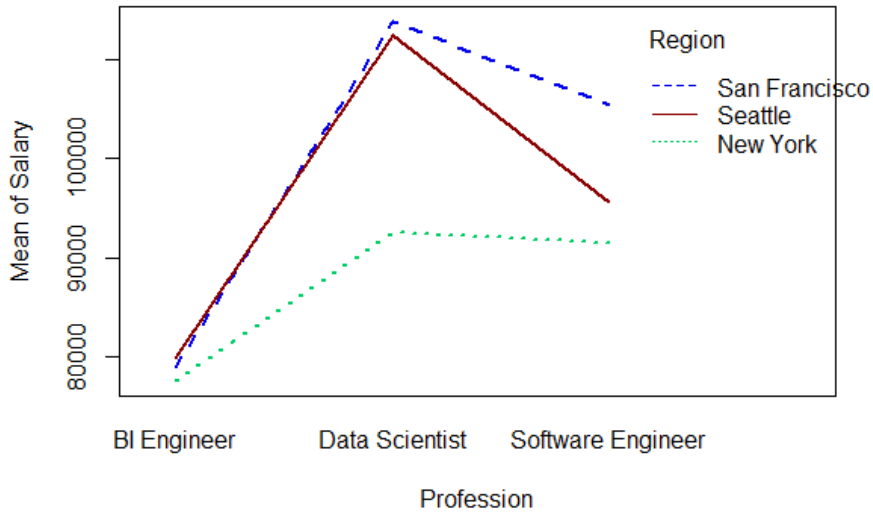
1. Salary versus Region, with different lines representing different Professions.
2. Salary versus Profession, with different lines representing different Regions.

Interaction plots that are parallel, or near parallel, imply that the effect of one factor is the same at all levels of the other factor (WorldClass FTE). In other words, there is no significant interaction. Lines that are not parallel imply that there is an interaction. First, we will plot Salary versus Region, with the lines representing Profession. This is done with this command **interaction.plot(x.factor = my_data$Region, trace.factor = my_data$Profession, response = my_data$Salary, fun = mean, type = 'l', col = c("springgreen3", "blue2", "darkred"), xlab = "Region", ylab = "Mean of Salary", lwd = 2, trace.label = "Profession").** The descriptions of the arguments are described below (Two-way interaction plot, n.d.):

1. x.factor: Factor to plot on the x-axis
2. trace.factor: Factor form the lines (traces)
3. Response: Variable to plot on the y-axis
4. fun: Function to compute summary statistic for response (single value)
5. type: type of plot; “l” for lines
6. col: Colors for traces
7. xlab, ylab: x and y axis labels
8. lwd: Line width
9. trace.label: Label for legend



This plot indicates that there is an interaction between the levels of Region and Profession. Similarly, to above, we plot Salary versus Profession, with the lines representing Region.



This plot suggests that there is an interaction between the levels of Profession and Region since the lines are not parallel and the distance between the means across the levels are not the same. Now that we have determined that we have an interaction, we can create our two-way ANOVA model. There are three sets of hypotheses for two-way ANOVA (McDonald, 2014):

1. Tests for an interaction effect.
    a. H_{01}: *There is no interaction between the two independent variables*
    b. H_{A1}: *There is an interaction between the two independent variables*
2. Tests the main effect for the first independent variable.

   a. H_{02}: *The means of all the first independent variable groups are equal*
   b. H_{A2}: *The means of at least one first independent variable groups are different*
3. Tests the main effect for the second independent variable.
   a. H_{03}: *The means of all the second independent variable groups are equal*
   b. H_{A3}: *The means of at least one second independent variable groups are different*

The hypotheses for main effects need to be tested if we find that the interaction between Region and Profession is not significant. Thus, we will first test the hypotheses for an interaction effect.

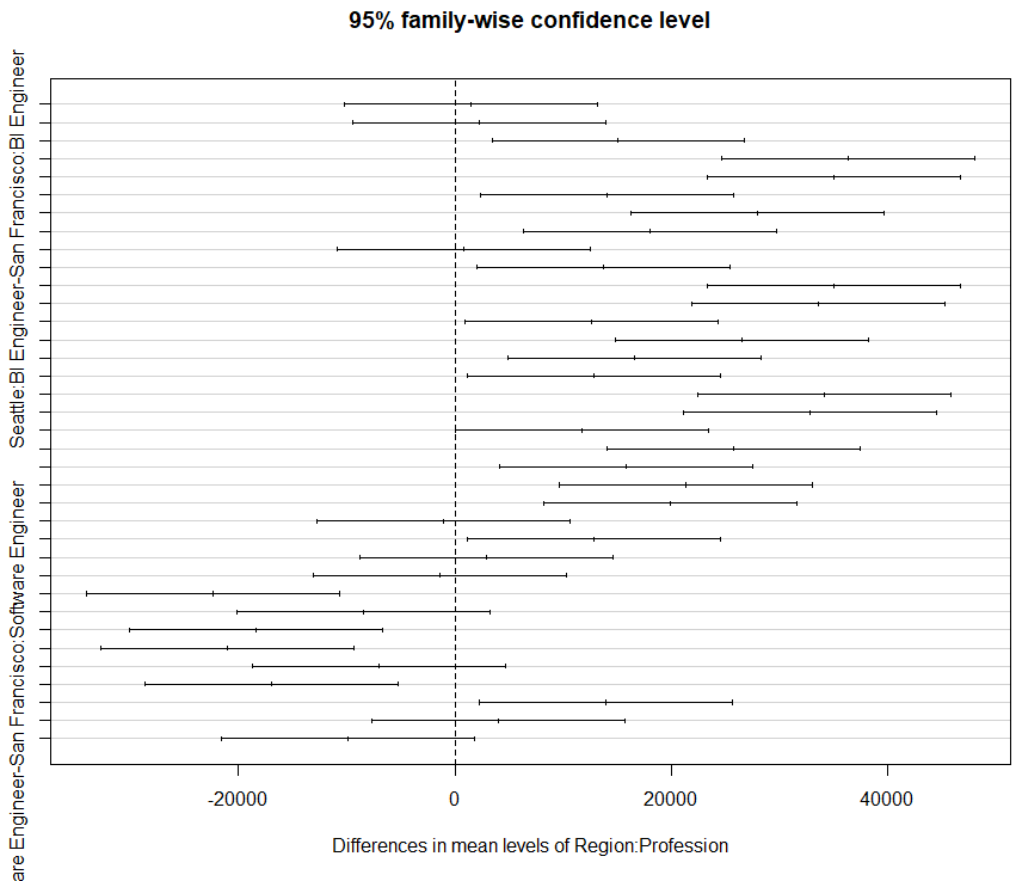
```
> interaction_aov <- aov(Salary ~ Region + Profession + Region:Profession, my_data)
> summary(interaction_aov)
                   Df    Sum Sq   Mean Sq F value   Pr(>F)
Region              2 4.750e+09 2.375e+09  17.143 1.64e-07 ***
Profession          2 2.386e+10 1.193e+10  86.098  < 2e-16 ***
Region:Profession   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
>
```

Using our level of significance of 0.05, we can see that our main effect Region and Profession are both statistically significant with the interaction also being significant because all the p-values are less than 0.05. Consulting our table of critical values for the F distribution, we obtain our critical F value, which is less than 2.47 (Table of critical values for the F distribution, n.d.). Our ANOVA calculated a F value of 5.481 which is greater than 2.47, which indicates that our interaction is significant.

Similarly, to one-way ANOVA, we run a post hoc test with our results of the two-way ANOVA model above using the function **TukeyHSD()** function. This post hoc test will tell us which levels of our independent variables and which combinations of the levels are the most significant.

```
> tukey_interaction <- TukeyHSD(interaction_aov, conf.level = 0.95, which = 'Region:Profession')
> tukey_interaction
  Tukey multiple comparisons of means
    95% family-wise confidence level

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

$`Region:Profession`
                                                                diff          lwr         upr     p adj
San Francisco:BI Engineer-New York:BI Engineer                 1421.35 -10273.11898  13115.819 0.9999868
Seattle:BI Engineer-New York:BI Engineer                       2236.10  -9458.36898  13930.569 0.9995865
New York:Data Scientist-New York:BI Engineer                  15092.65   3398.18102  26787.119 0.0024207
San Francisco:Data Scientist-New York:BI Engineer             36380.45  24685.98102  48074.919 0.0000000
Seattle:Data Scientist-New York:BI Engineer                   35008.40  23313.93102  46702.869 0.0000000
New York:Software Engineer-New York:BI Engineer               14010.80   2316.33102  25705.269 0.0069368
San Francisco:Software Engineer-New York:BI Engineer          27946.35  16251.88102  39640.819 0.0000000
Seattle:Software Engineer-New York:BI Engineer                18030.00   6335.53102  29724.469 0.0000975
Seattle:BI Engineer-San Francisco:BI Engineer                   814.75 -10879.71898  12509.219 0.9999998
New York:Data Scientist-San Francisco:BI Engineer             13671.30   1976.83102  25365.769 0.0094978
San Francisco:Data Scientist-San Francisco:BI Engineer        34959.10  23264.63102  46653.569 0.0000000
Seattle:Data Scientist-San Francisco:BI Engineer              33587.05  21892.58102  45281.519 0.0000000
New York:Software Engineer-San Francisco:BI Engineer          12589.45    894.98102  24283.919 0.0244634
San Francisco:Software Engineer-San Francisco:BI Engineer     26525.00  14830.53102  38219.469 0.0000000
Seattle:Software Engineer-San Francisco:BI Engineer           16608.65   4914.18102  28303.119 0.0004900
New York:Data Scientist-Seattle:BI Engineer                   12856.55   1162.08102  24551.019 0.0195243
San Francisco:Data Scientist-Seattle:BI Engineer              34144.35  22449.88102  45838.819 0.0000000
Seattle:Data Scientist-Seattle:BI Engineer                    32772.30  21077.83102  44466.769 0.0000000
New York:Software Engineer-Seattle:BI Engineer                11774.70     80.23102  23469.169 0.0470207
San Francisco:Software Engineer-Seattle:BI Engineer           25710.25  14015.78102  37404.719 0.0000000
Seattle:Software Engineer-Seattle:BI Engineer                 15793.90   4099.43102  27488.369 0.0011759
San Francisco:Data Scientist-New York:Data Scientist          21287.80   9593.33102  32982.269 0.0000017
Seattle:Data Scientist-New York:Data Scientist                19915.75   8221.28102  31610.219 0.0000098
New York:Software Engineer-New York:Data Scientist            -1081.85 -12776.31898  10612.619 0.9999984
San Francisco:Software Engineer-New York:Data Scientist       12853.70   1159.23102  24548.169 0.0195719
Seattle:Software Engineer-New York:Data Scientist              2937.35  -8757.11898  14631.819 0.9970431
Seattle:Data Scientist-San Francisco:Data Scientist           -1372.05 -13066.51898  10322.419 0.9999900
New York:Software Engineer-San Francisco:Data Scientist      -22369.65 -34064.11898 -10675.181 0.0000004
San Francisco:Software Engineer-San Francisco:Data Scientist  -8434.10 -20128.56898   3260.369 0.3687205
Seattle:Software Engineer-San Francisco:Data Scientist       -18350.45 -30044.91898  -6655.981 0.0000667
New York:Software Engineer-Seattle:Data Scientist            -20997.60 -32692.06898  -9303.131 0.0000024
San Francisco:Software Engineer-Seattle:Data Scientist        -7062.05 -18756.51898   4632.419 0.6165068
Seattle:Software Engineer-Seattle:Data Scientist             -16978.40 -28672.86898  -5283.931 0.0003253
San Francisco:Software Engineer-New York:Software Engineer    13935.55   2241.08102  25630.019 0.0074423
Seattle:Software Engineer-New York:Software Engineer          4019.20  -7675.26898  15713.669 0.9764101
Seattle:Software Engineer-San Francisco:Software Engineer     -9916.35 -21610.81898   1778.119 0.1687988

> plot(tukey_interaction)
>
```

The columns we are most interested in is **diff** (difference in means) and **p adj** (adjusted p-value). Any p-value less than 0.05 is significant. Most of the insignificant pairs occur within comparisons of the same profession. Now, we can plot the results.



Recall that the significant pairs are the ones that do not cross the "zero" line in the plot. From the plot, we gather that most of the pairs are significant.

We performed two-way ANOVA to examine the effect that Profession and Region had on Salary. Boxplots were used first to examine the means of Salary by Region, Profession, and Region combined with Profession. Then, interaction plots were used to visualize the relationship between Region and Salary as the level of Profession changes and visualize the relationship between Profession and Salary as the level of Region changes. These plots showed that Region and Profession have an effect on Salary. Thus, we performed two-way ANOVA to identify which terms were statistically significant. We found that both main effects and the interaction effect were significant, thus rejecting our null hypotheses. To examine which combinations were significant and which have the largest differences, Tukey's HSD post hoc test was used. This revealed that a Data Scientist from San Francisco has a much higher salary than a BI Engineer in New York. It also revealed that a Software Engineer in New York makes much less than a Data Scientist in San Francisco.

Resources

Anova test. (n.d.). Statistics How To. Retrieved April 8, 2020, from https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/

Jones, J. (2020, January 11). *Stats: Two-way anova*. https://people.richland.edu/james/lecture/m170/ch13-2wy.html

McDonald, J. H. (2014, July 20). *Two-way anova*. http://www.biostathandbook.com/twowayanova.html

Reading data from txt|csv files. (n.d.). Retrieved April 9, 2020, from http://www.sthda.com/english/wiki/reading-data-from-txt-csv-files-r-base-functions

Table of critical values for the F distribution. (n.d.). http://users.sussex.ac.uk/~grahamh/RM1web/F-ratio%20table%202005.pdf

Two-way anova test in r. (n.d.). Retrieved April 10, 2020, from http://www.sthda.com/english/wiki/two-way-anova-test-in-r

Two-way interaction plot. (n.d.). Retrieved April 9, 2020, from https://rdrr.io/r/stats/interaction.plot.html