# Week 5 Assignment – Two Way ANOVA

By: Zach Adair

Regis University

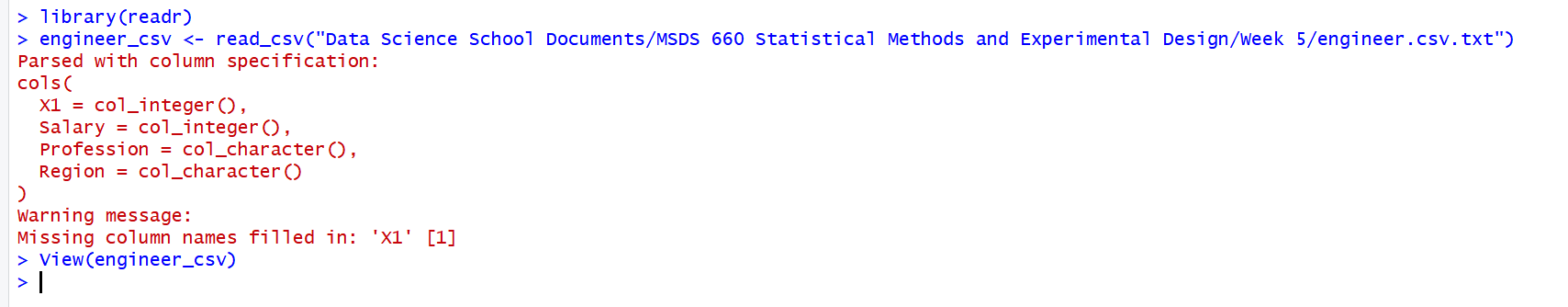
MSDS 660 – Statistical Methods & Experimental Design

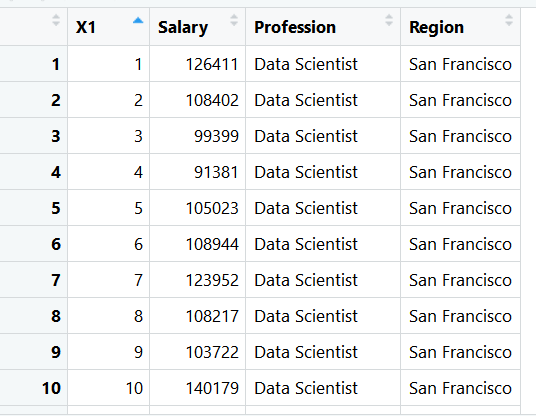
## Introduction

For this week’s assignment I will be exploring the two-way ANOVA and how it can be done in RStudio. Two-way ANOVA is used to evaluate the effect of two grouping variables simultaneously upon a dependent variable. The data used will be a sample of 180 people from New York, Seattle and New York and comparing the different salaries in those regions by the professions of BI engineer, Data Scientist and Software Engineer.

## Load the Data

Before doing any analyzing you must first need to load your data.





I wanted to get a view of the data just to see what it came in like after loading it into RStudio.

## Hypothesis

**Interaction**

Null Hypothesis: There is no interaction effect between the variables region and profession.

Alternative Hypothesis: There is a an interaction effect between region and profession.

**Region**

Null Hypothesis: The region variable has no effect on the salary of the workers.

Alternative Hypothesis: The region variable does have an effect on the salary of the workers.

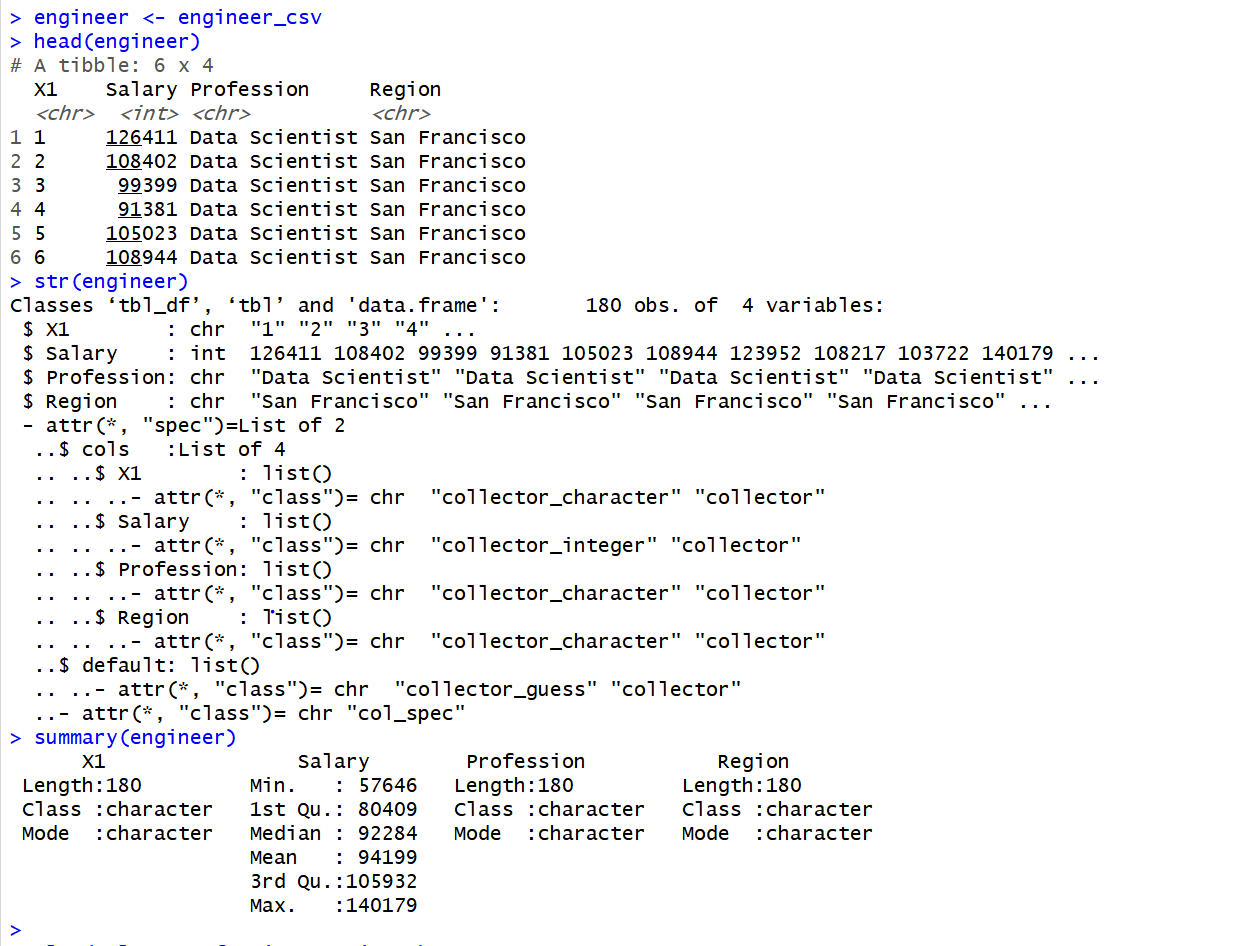
**Profession**

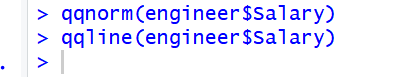
Null Hypothesis: The profession variable has no effect on the salary of the workers.

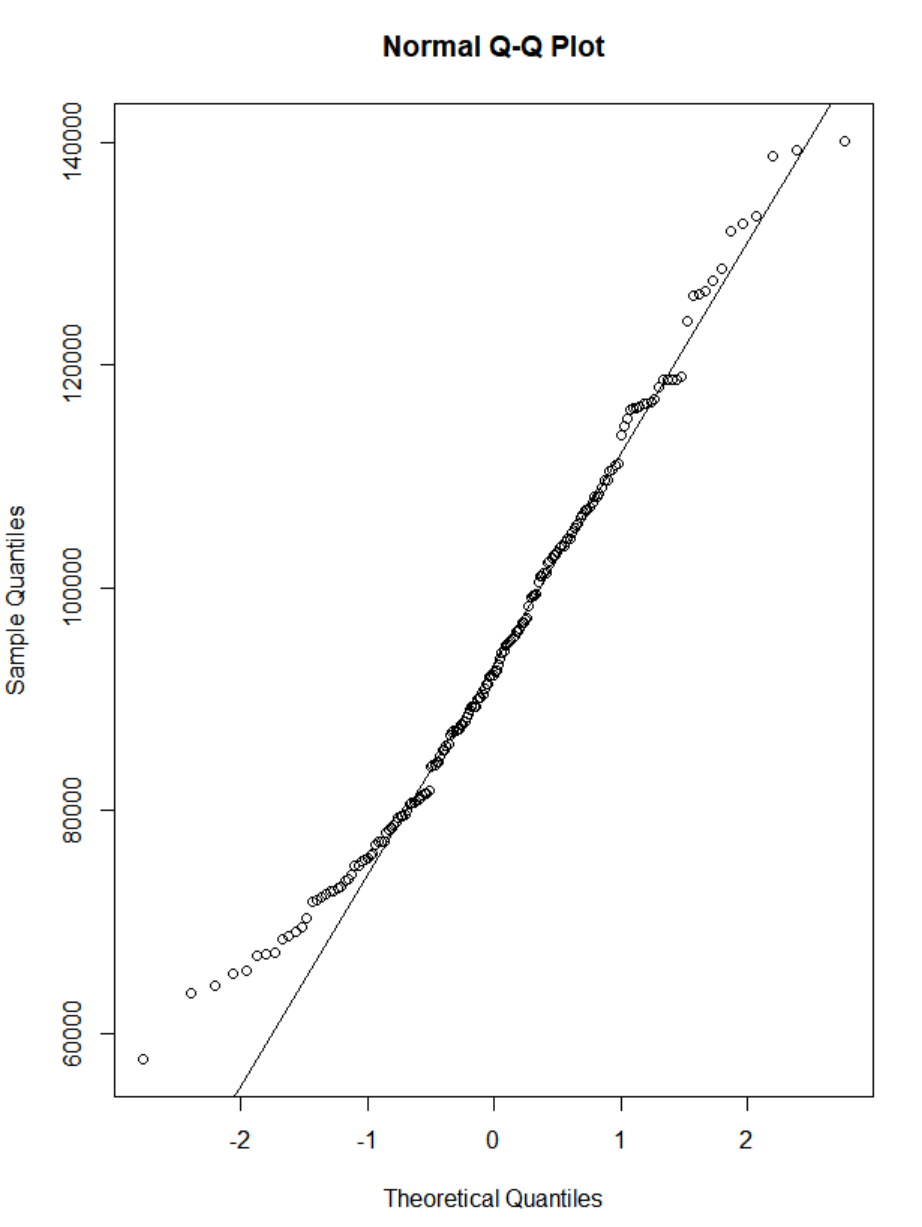
Alternative Hypothesis: The profession variable does have an effect on the salary of the workers.

## Data Exploration

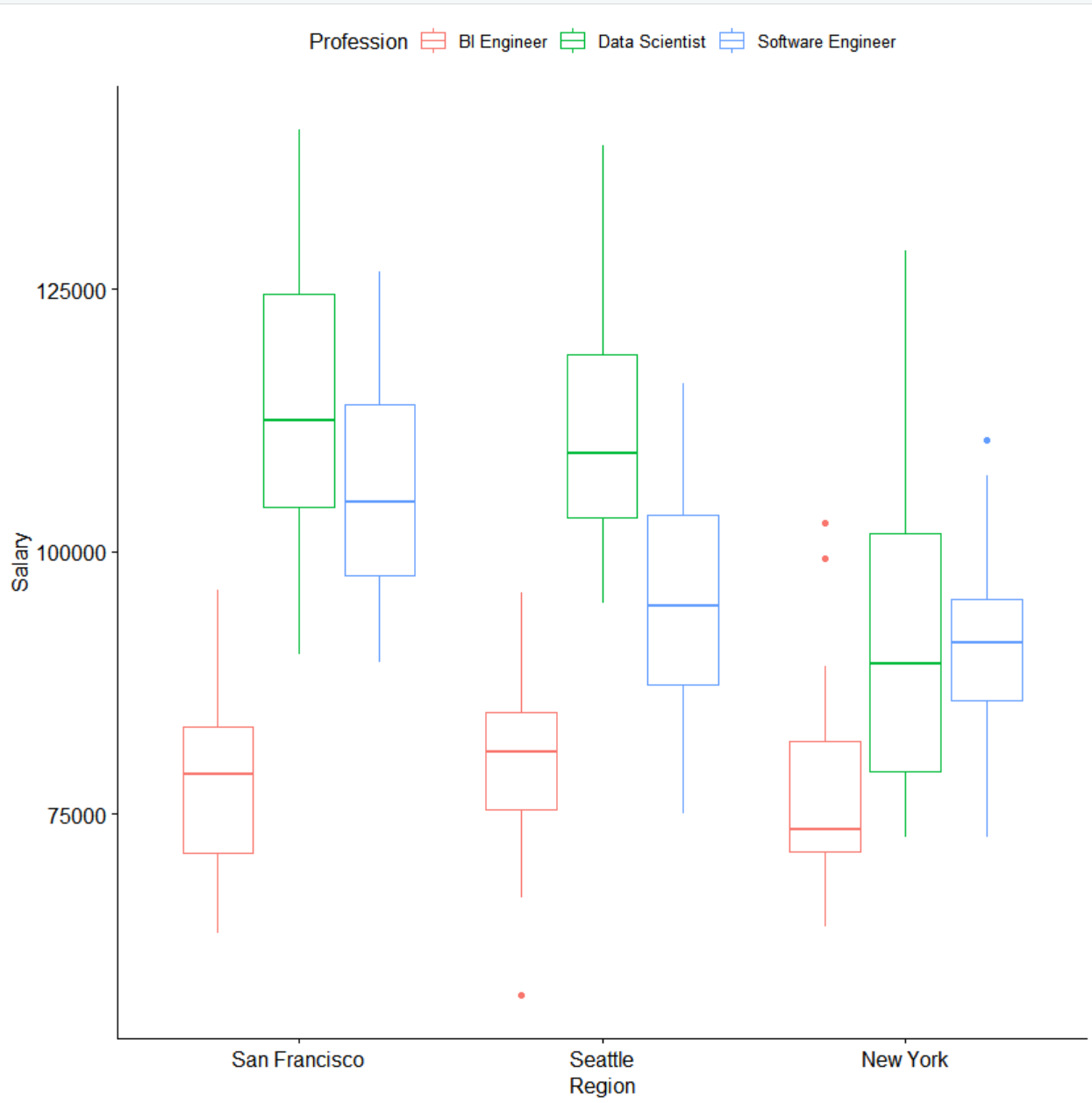
After bringing in the data it is always smart to explore the data and see what is there and potentially find any noticeable trends within the data.



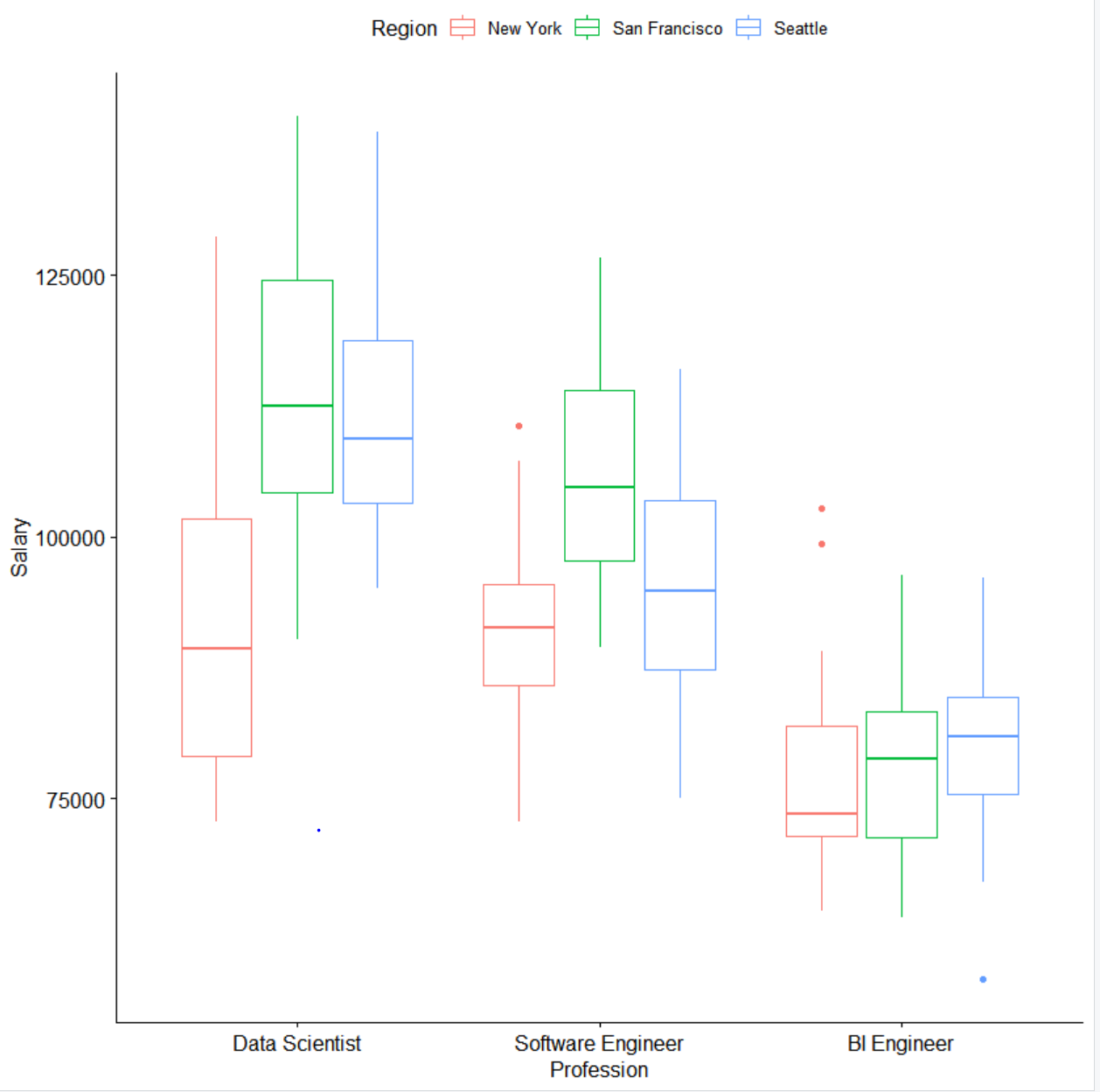


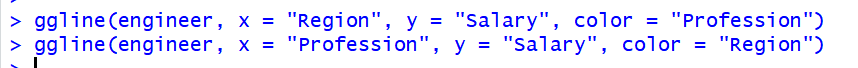


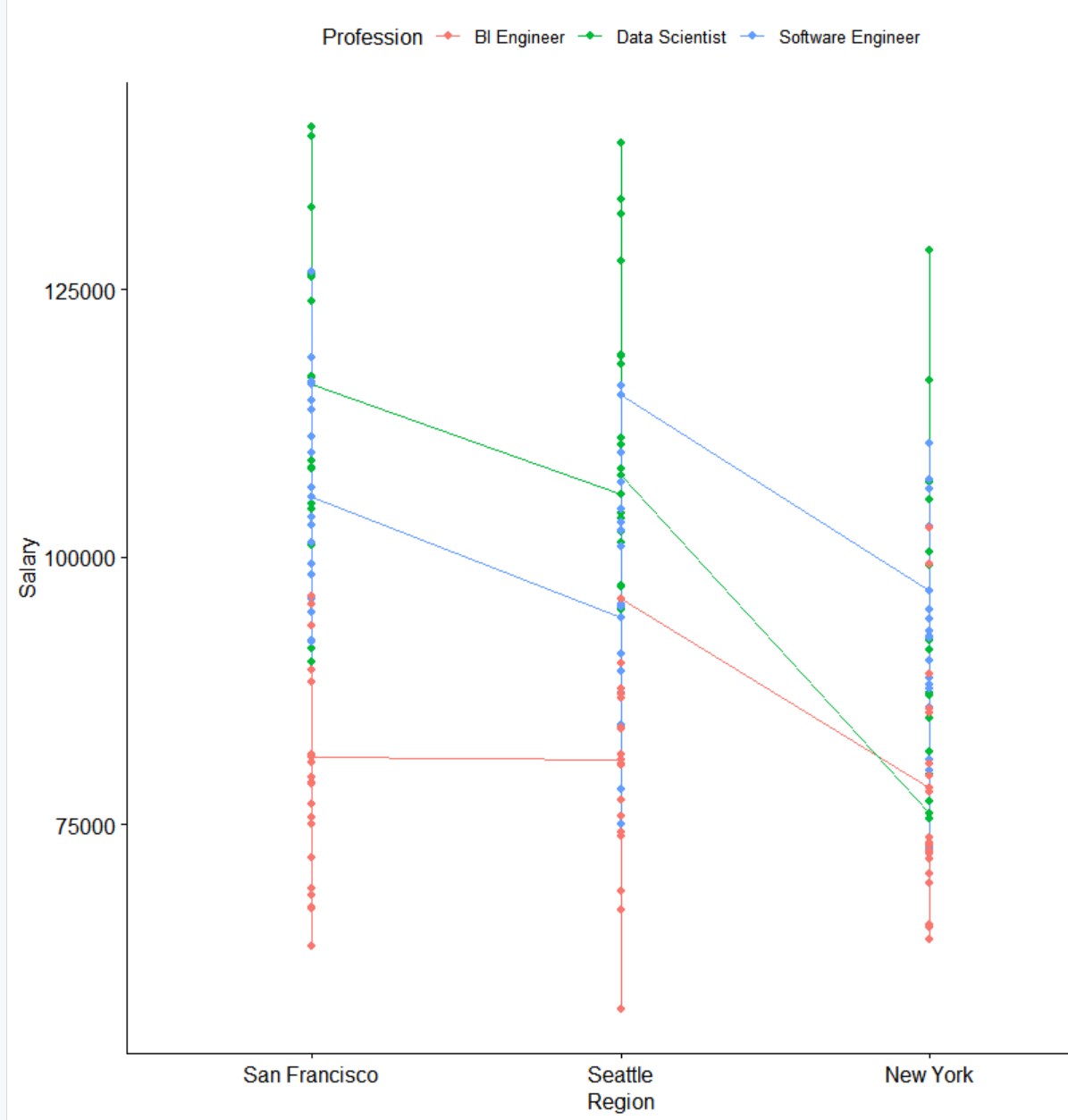


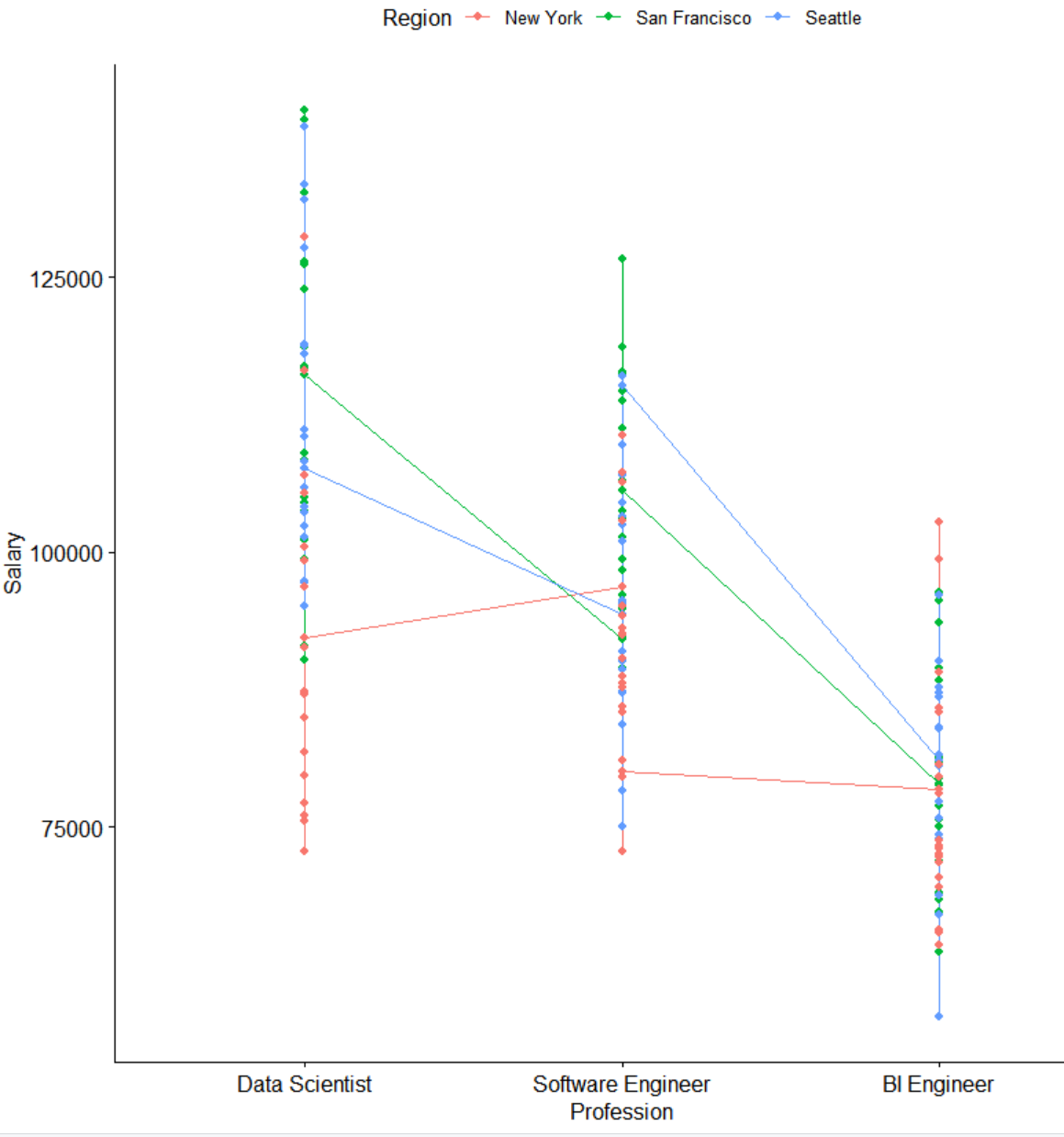








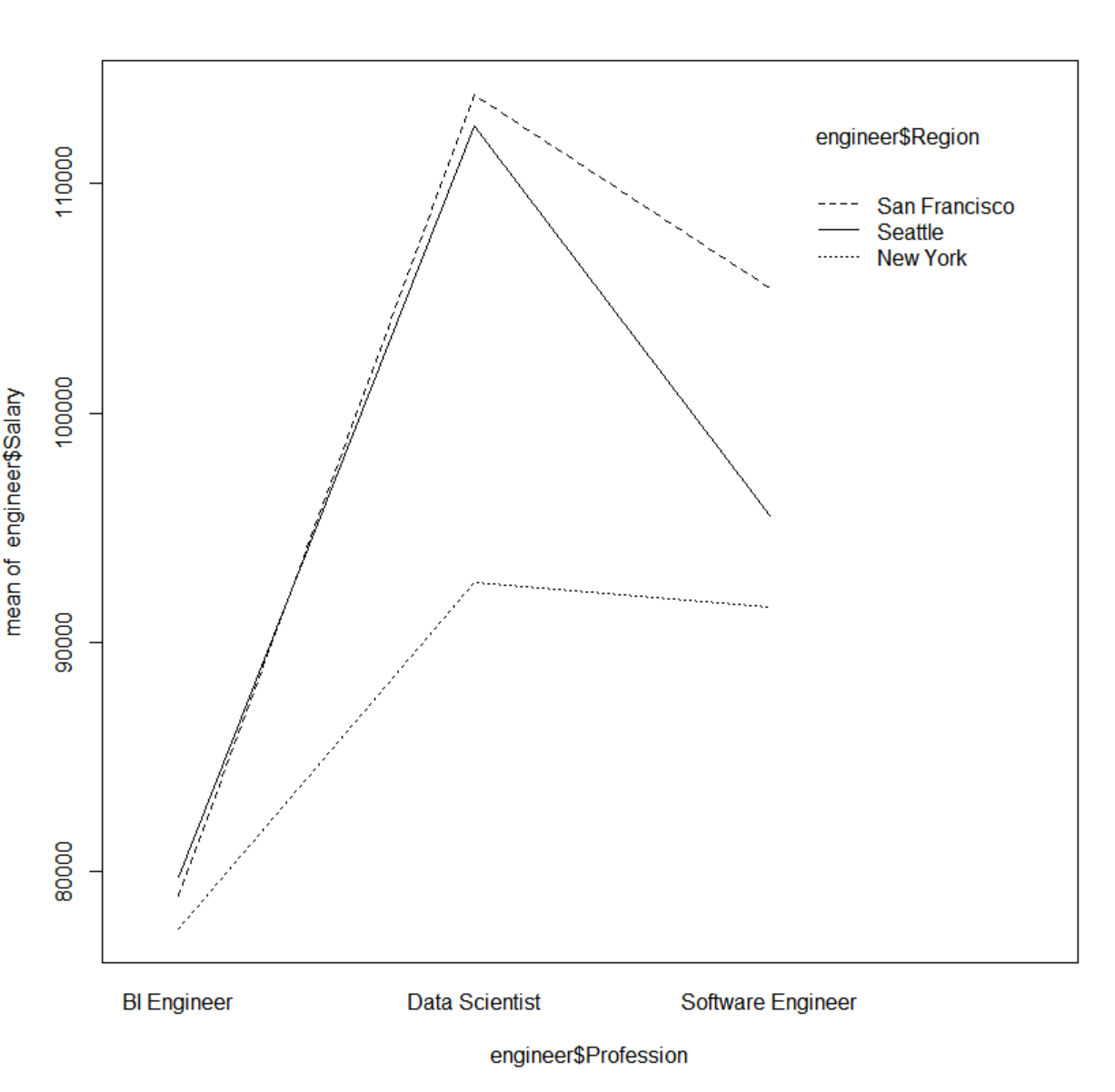




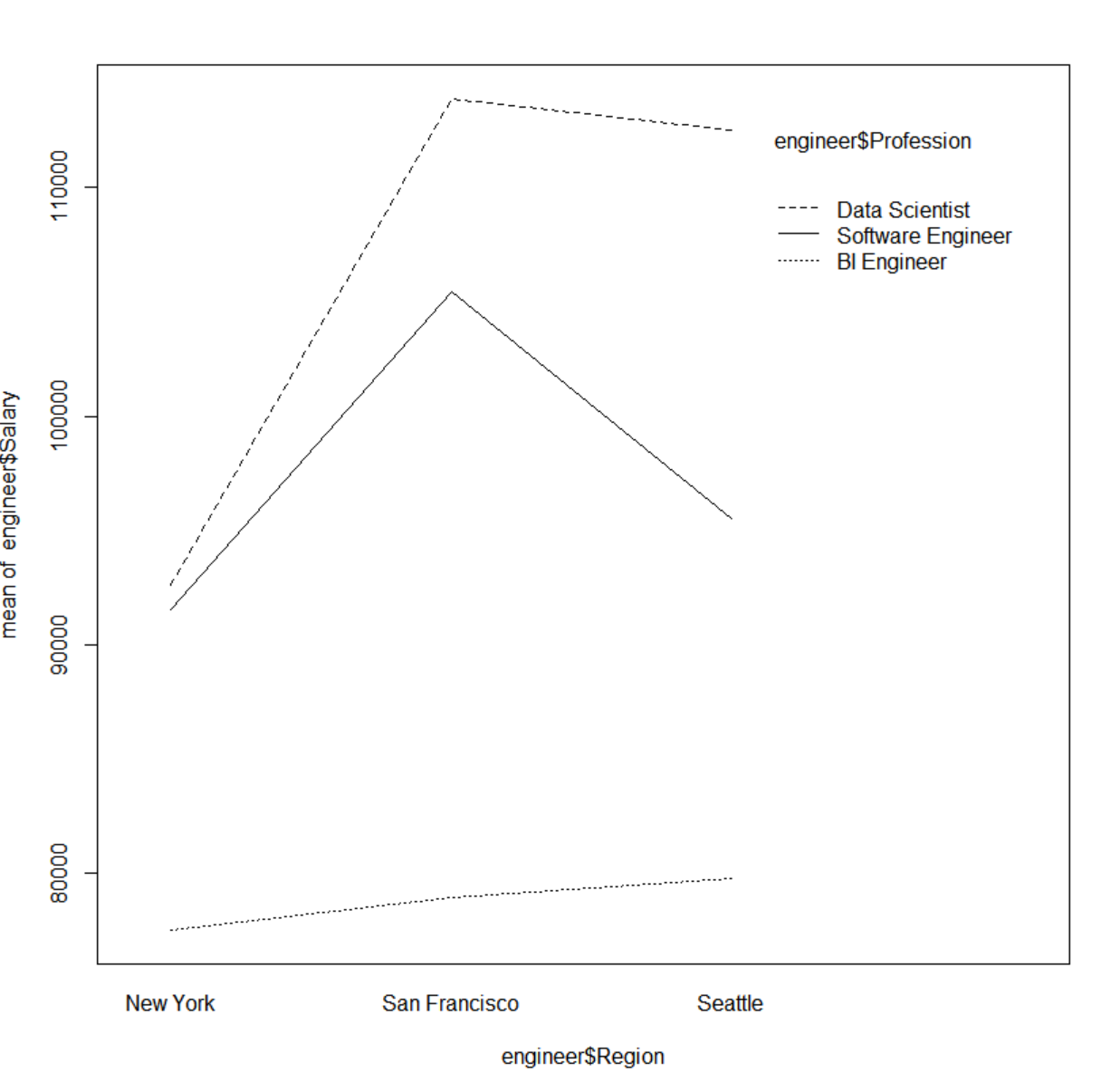
From the commands ran in RStudio we can make some assertions about the data and how it played out from our data exploration. First, it would appear that San Francisco is the region who pays the most and New York is the region that pays the least. When looking at the professions, it would appear that Data Science is the highest paid of the three professions and BI specialist is paid the least. With the data exploration done here I think I can confidentially construct a hypothesis and run a two-way ANOVA.

Before I get to my two-way ANOVA I will be running an interaction plot on the data as well, this should prove my thoughts about both region and profession that I stated above.





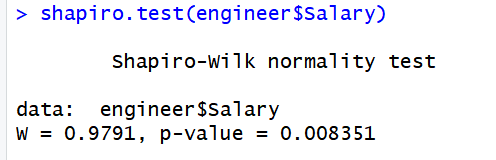




## Checking Assumptions

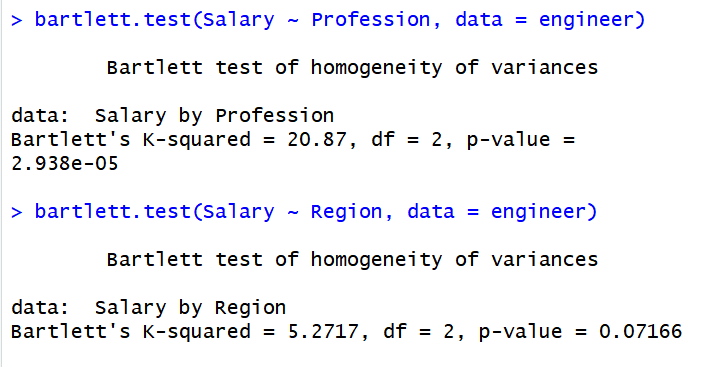
Before running our two-way ANOVA, assumptions must be checked. To do this we will check normality, independence and homogeneity.

I’ll start with checking for normality, to do this I will use the Shapiro-Wilk Test by testing the Salary data. The Shapiro-Wilk Test is a test where if the p-value is less than 0.05 we will reject the null hypothesis that the data is normal in favor for the alternative which would state that the data is normal.



From our result, it would appear our data is not normal, which means our Salary numbers skew in some way.

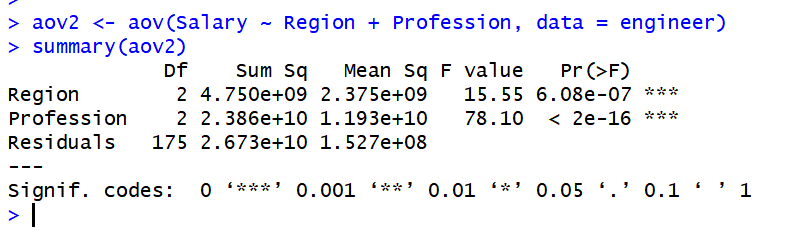
Next I will be doing the Bartlett Test to test the homogeneity of the data.



From the two Bartlett Tests run it is apparent that profession is not a homogenous variable but region is. Just like the Shapiro-Wilk test we are judging this based on the p-value, with the p-value being lower than 0.05 for profession we can conclude that it is not homogenous but with the p-value being greater than 0.05 for region we can conclude that region is a homogenous variable. From the results of the Shapiro-Wilk and the Bartlett Tests as well as the data exploration it is becoming apparent that the null hypothesis will most likely be reject in favor of the alternative hypothesis and thus stating that there is an interaction effect and that interaction effect may come from either region or profession and it’s possible it could come from both.

## Two-Way ANOVA

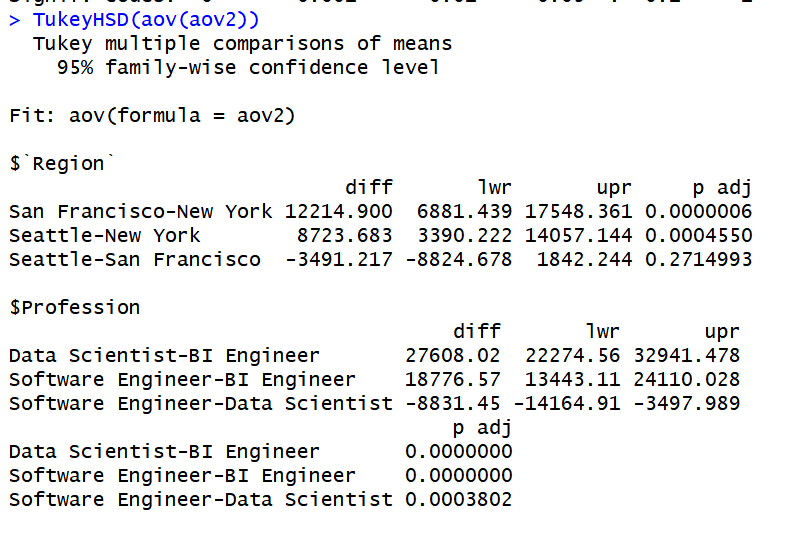
Even though I’m starting to doubt the original null hypothesis I believe clearing this up best will come from running the ANOVA test, even if the null hypothesis needs to be rejected it won’t be made as clear until I run a post hoc test after.



Just like I was beginning to assume, the ANOVA test resulted in extremely low p-values for both the region and profession variables. Thus concluding the null hypothesis should be rejected and that there is an interaction effect on salary and the variables profession and region have an effect on it. Regardless that the null hypothesis was rejected the next steps here are to run a post hoc test to check the distance between salary and the independent variables of region and profession.

## Post Hoc Analysis

The post hoc test I will be running on this data will be Tukey’s HSD test. This will let me know if the lengths are equal by the p-values but since we know they will not be it will help distinguish where those differences lie between the different variables.



From Tukey’s we can distinguish the differences between the region and profession variable and which caused an interaction effect on Salary. For region, it would appear that the difference between San Francisco and Seattle compared to New York were drastic enough to have an overall effect on salary. For profession, the difference found between Data Scientist is substantial for compared to BI Engineer and Software Engineer but for BI Engineer the difference is even greater because there is a substantial difference between it and Software Engineer.

After this analysis we can officially reject the null hypotheses and conclude that there is an interaction effect within our data and it is caused by both the region and the profession, primarily the regions of Seattle and San Francisco and then from the professions overwhelmingly Data Scientist had an effect but so too did the Software Engineer profession.

## Summary

In this project I worked with two-way ANOVA and creating an analysis with it in RStudio. Last week getting to work with one-way ANOVA helped prepare a little bit about what to expect and how to read each of the tests but the differences were still easy to spot and it was fun to learn about those differences. Even the differences in setting up the two-way ANOVA vs the one-way was interesting to see and play around with through the assignment.