Blog Post - Mod2 Project

Lucy Hayes

INTRODUCTION

The purpose of this project was to connect to a database using SQL, extract data and join tables, in order to see if there were any significant findings along the way

PROCESS:

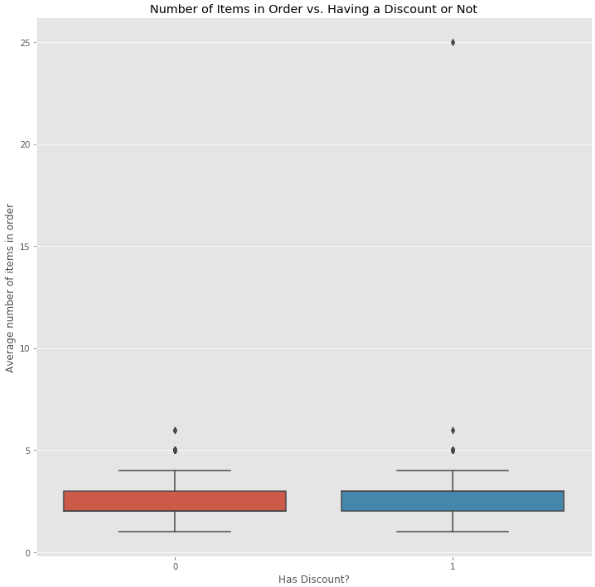
First step was to study the database schema of the Northwinds database we were given. Next, using sqlite, I initiated a session and inspected the real data to see if it matched up to that of the EDA. The first noticeable difference was that the Table names are listed plural on the EDA, but are not in reality. Made a mental note. Next, I looked at the columns within the tables of the database, to see if there were any differences in naming conventions, and to get a general feel of the data.

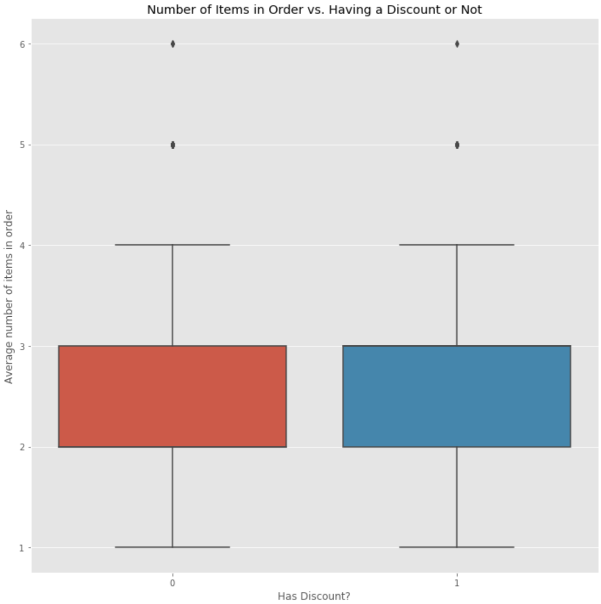
Once this was completed, I initiated a connection and tried a simple query to make sure everything was working accordingly. I repeated this simple query to get a better view of individual tables a few times throughout the process, before hypothesis testing and before pulling more complex queries. (A fun finding was that there were only 9 employees, and they had hire dates in the future!).

Next was to start exploring four hypotheses, which I have grouped below by their general themes

DISCOUNTS

Here, I wanted to look into if discounts and an effect on the number of products customers ordered. Since all of this information was stored in the OrderDetails table, the query was relatively simple: I pulled the Order ID, the Product ID, and the Discount. After viewing the data frame, it seemed like there were multiple discounts applied to a single order, sometimes of different values. To account for this, I took the unique orderids, and created a dictionary of key value pairs that corresponded to the unique order id, and the maximum discount at the order number. I then created a new data frame based off this dictionary, where there was the unique order Id, the quantity of items in the order, and the maximum discount. My first step was to classify the data as having a discount or not. Upon visualizing, I found there was a major outlier, which I removed from the dataset .



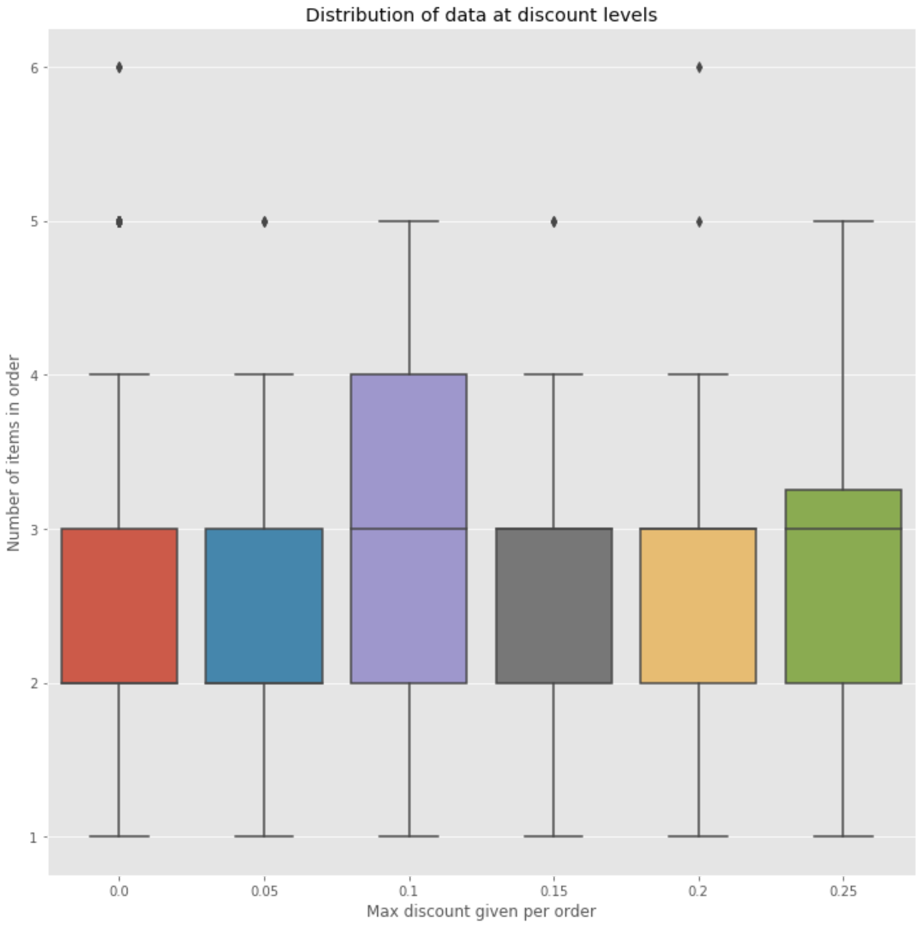


BEFORE Outliers……

AFTER Outliers

Time to Run some stats:
Null Hypothesis: discounts have no effect on the number of products ordered by customers
Alternate Hypothesis: discounts have a significant effect at alpha = .05 on the number of products ordered by customers

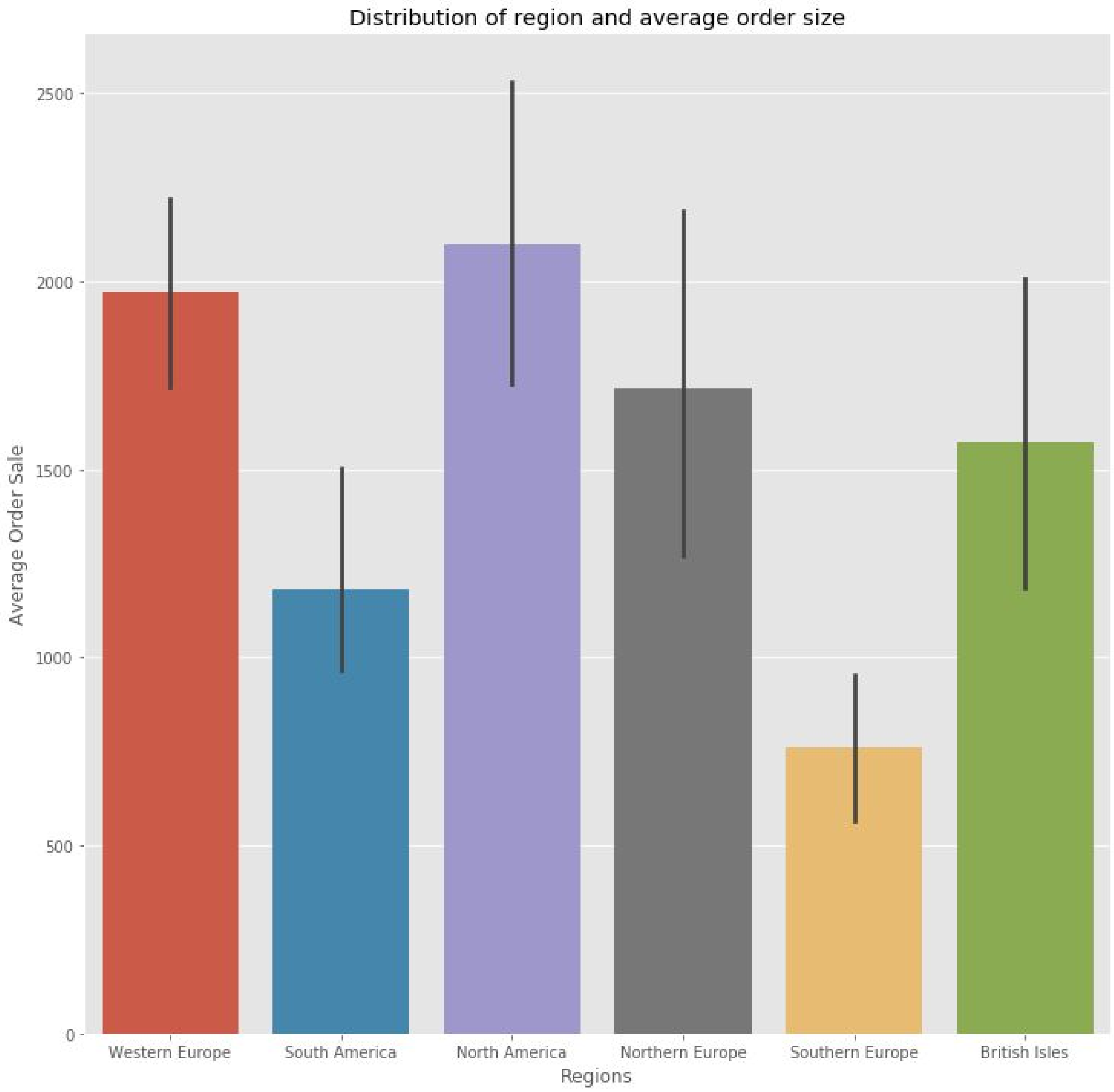
Since I was comparing two means, I used a T-Test and a MannWhitney U test (since I didn’t test for normality). Both were highly significant, meaning we could reject the null hypothesis. But now that we could see there was an effect, I wanted to go further and see at which level of discount the effect was occurring. When visualizing the data, it immediately seemed like .1 and .25 had the largest spread.



Since I was comparing multiple means, I used an ANOVA Test, and found that the results were statistically significant (p(F) = .01, p < alpha). The individual largest effect is seen at .1 and .25, which were the only two individual levels with a p < .05 value. Thus we can reject the Null Hypothesis that discounts had no effect on the number of products ordered by customers.

REGIONALITY CUSTOMERS

Here I looked into if the regions where customers ordered from had a significant effect on the average sales order price. In order to do this, I needed to connect the order details table to the orders table, to the customers table, all using Ids. I then grouped by order ID to sum up to the individual order, since we were no longer concerned with line item discounts. My exported Dataframe contained the orderID, the total amount for that order, and the region which the order would be delivered to. I then checked to see if there were any NAs in the regions, of which there were 29, so I dropped them from the dataframe. Next I checked to see if there were any regions that did not have enough data to run analysis, which left us with 6 total regions. Next I visualized the data:

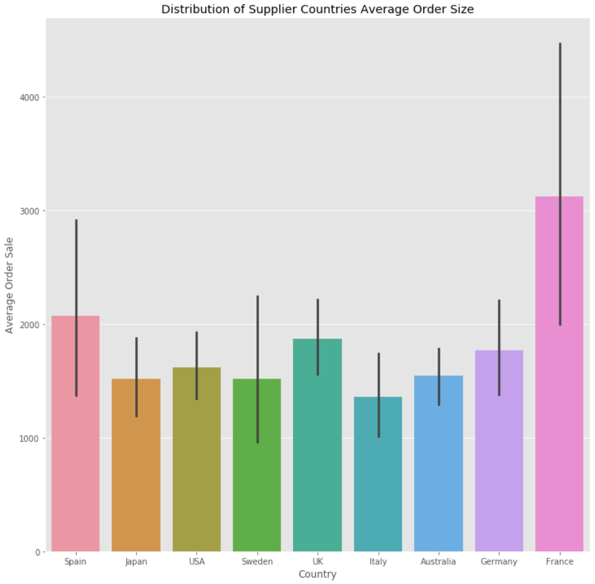


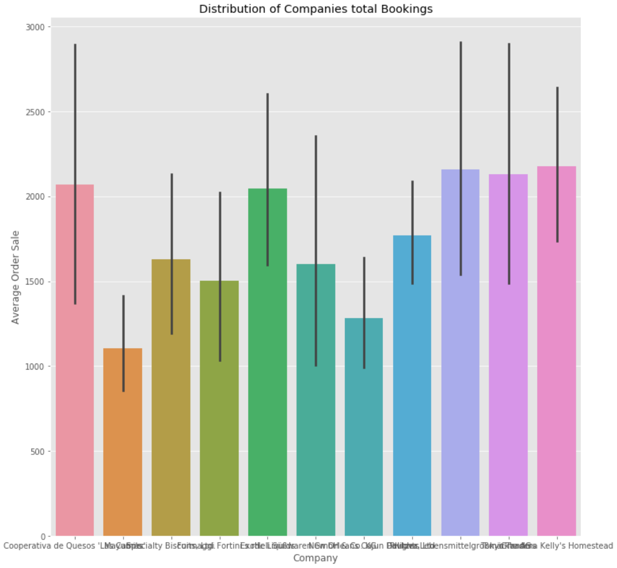
At a glance, it seems like some regions have a much higher on average order price than others. I used an ANOVA test to compare across regions with the two hypotheses:
Null Hypothesis: regions have no effect on the average sales price of the order
Alternate Hypothesis: regions have a significant effect at alpha = .05 on average sales price of the order

The results of the ANOVA were highly significant, so I used a TukeyHSD test to delve a little further to see if there were any regions in particular that were significant. I found that we can reject the null at the pairings of North America and South America, North America and Southern Europe, Southern Europe and Western Europe, and South America and Western Europe. This is showing that there is significance between the lowest on average sales price (Southern Europe and South America), and the two largest (North America and Western Europe).

REGIONALITY SUPPLIER

Here I wanted to look into if there were any countries of suppliers that were generating larger average order prices than others. To do so, I needed to link the OrderDetail table to the Product table, to the Supplier table using key IDs. I also extrapolated the supplier company name, the total bookings per order (quantity*unitprice), and the country of the company. There were 16 countries, and 29 companies. I did not find any NAs in the data, however there were a few countries and companies with less than 30 datapoints that needed to be removed from the dataset. We could visualize the data as follows:





At a glance, it seems like some countries and companies have a much higher or lower on average order price than others. I used an ANOVA test to compare across regions with the two hypotheses:

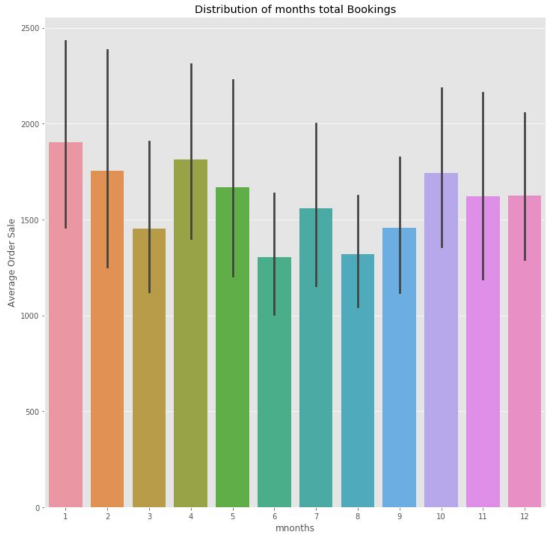
Null Hypothesis: countries/companies have no effect on the average sales price of the order
Alternate Hypothesis: countries/companies have a significant effect at alpha = .05 on average sales price of the order

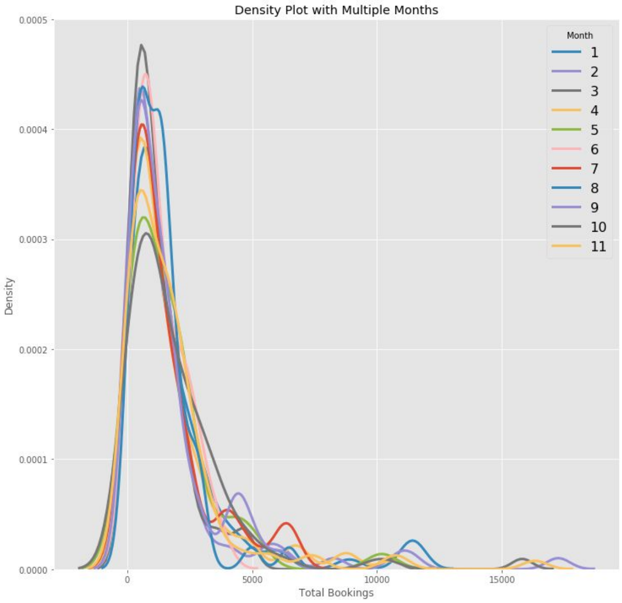
For countries, the results of the ANOVA were significant, so I used a Tukey HSD Test to check further. The results showed that we could reject the null at France between every country except for Germany. There were no other pairings of countries that had an effect, which confirms what we saw in the graph that France has the largest order size of any country on the supplier side.

For companies, after running the ANOVA, I found there to be no significant effect of the supplier company on average order sales price.

SEASONALITY OF SALES

This was my personal favorite hypotheses of the project! Here, I wanted to see if there was any seasonal effect of the average order size on weekday, month, and quarter. I queried the data by connecting the order details table to the order table in order to obtain the date of the order. Next, I converted the date into a datetime object, and created new columns for the month of year,, the day of the week, and the quarter. First I looked into months and their data distributions:





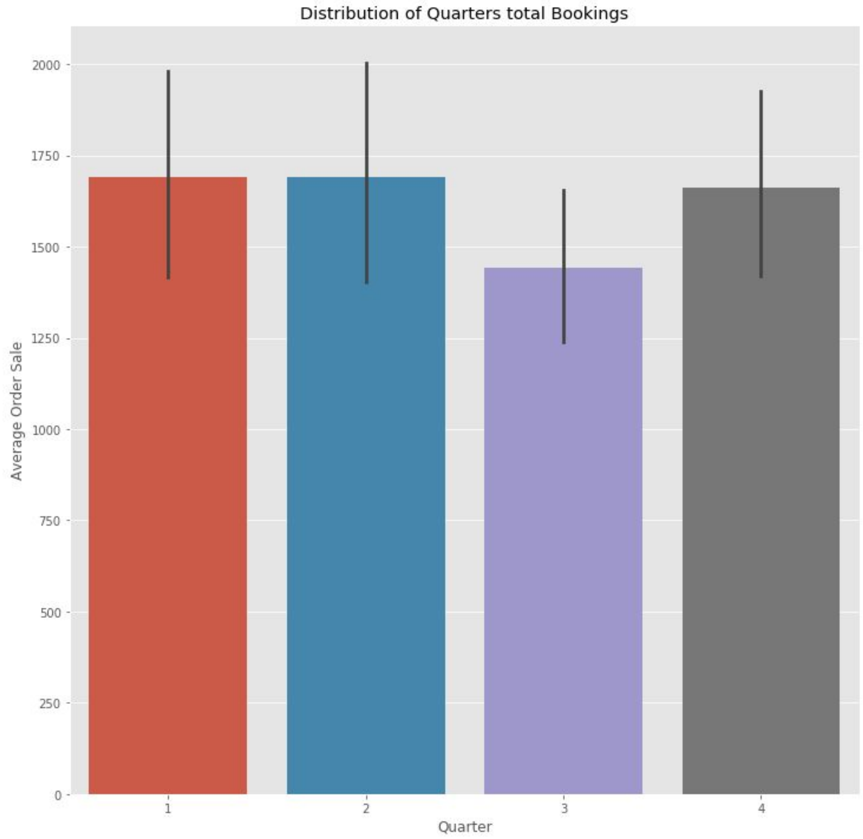
At first glance, it seemed like the months showed a seasonality effect, with lower than average sales prices during the summer months, so I ran an ANOVA with the hypotheses:

Null Hypothesis: Month of order date has no significant effect on the average order size

Alternate Hypothesis: Month of order date has a significant effect on the average order size at a p level of .05

I found there was no significant effect of month on the average order price. So next I looked into quarters, since usually sales goals are set quarterly.

Next I looked at Quarters:

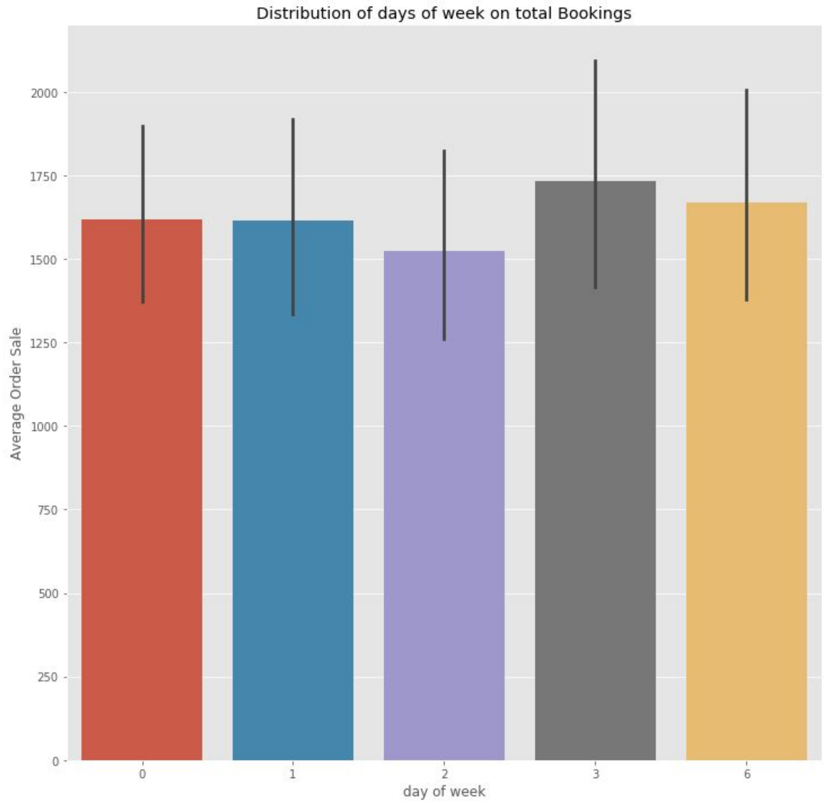


At first glance it looks like Q3 had a lower than average sales price, however the error bars are concerning. I ran another ANOVA test with the following hypotheses:

Null Hypothesis: Changes in quarter has no significant effect on the average order price

Alternate Hypothesis: Changes in quarter has a significant effect on the average order price at alpha level .05.

But, once again found this to be insignificant. The last trend I wanted to look into was to see if there was any trend across the day of the week on sales. To do this, I converted the date into the day of the week, but interestingly saw there were missing values for all Fridays and Saturdays. Since this was a sample database, I made the assumption that this was meant to refer to Saturdays and Sundays, and continued with the analysis only using the weekdays (since likely no sales are placed on the weekends).



Woof. Looking at the graph, I already could tell what the results would be. I ran the ANOVA with the following hypotheses:

Null Hypothesis: There is no significant difference between the days of the week and the average order size

Alternative Hypothesis: the day of the week has a significant effect on the average order size

And as predicted, found no significance.

CONCLUSION:

Across four different hypotheses themes, we had some interesting results! First, we found that discounts had a significant effect on the number of items in an order, specifically at the 10% and 25% discount level. Next, we found that North America and Western Europe had significantly larger order sizes than South America and Southern Europe on the consumer side. On the supplier side, we found that country had a significant effect on the average order size, specifically France, but there was no effect seen of a particular supplier company on average order size. Lastly, across multiple levels of seasonality, we found there was no effect of the timing of a sale on the average order size.