“SQL, T-tests, p values…Oh My!”
Project blog post (SQL/Northwind database project)
LaShanni Butler
3/22/19

Finally, I have completed this project! I must admit that Module 2 project has been the bane of my existence. From whatever reason, almost all aspects of SQL have yet to “click” for me. Thus, I find it incredibly frustrating having to do something that is clear as mud. For this project, we took an exploratory look at the fictional company created by Microsoft, Northwind. Based on our exploration, we had to generate several hypotheses (i.e. – use test statistics). Almost admittedly, stats are also another bane in my existence. But nonetheless, while this project was challenging, it was good exposure to working with SQL and refreshing my rusty memory when it comes to statistics.

So, we worked with the Northwind database, which is a fairly small database compared to actual companies’ databases. While “small” the database was quite robust with multiple schemas. While the schema was available to us, it was hard to ascertain what exactly Northwind did as a business. So, like with any dataset, it was necessary to do an exploratory data analysis (EDA). The EDA revealed that Northwind is a food distribution company, and their top products revenue sales are beverages and dairy. Further analyses showed that Northwind has offices in the United States and London. They have a worldwide customer base with 91 customers in total. Additionally, Northwind has 9 employees that work for the company, so it’s a relatively small company. The EDA helped me to get a better understanding of what this company is and what they do.

Using SQL, I loaded the Northwind database and incorporated it into the Pandas data frame. Luckily the Module 2 project already gave us our first hypothesis question to analyze and answer. Our first hypothesis wanted us to determine if discount amount has a statistically significant effect on the quantity of a product in an order? And if it did, at what level of discount? I formulated my null and alternative hypothesis and choose an alpha value of 0.05, which gave us a 95% confidence interval. With the pandas data frame, I looked at the Northwind’s order details to get a better sense of the quantities as well as the discounts offered. When I looked at the discount, I noticed very few discounts were given at 1%, 2%, 3%, 4% and 6% amount. Since they contributed very little to the dataset, I dropped these from the dataset and focused on the remaining discount amounts. Then I ran a t-test on the values, as well as a t-test on the non-discounted quantities to see if there was any statistical significance. After generating some graphs that looked at the total number of items ordered stratified by discount level (or lack thereof), I looked at the mean and standard deviation for each group. After coming up with the test statistic and p-value, I was able to know if the answer to my hypothesis question. My analysis showed that the null hypothesis was rejected. So, having a discount or no discount in place did not affect the price.

I did a similar type of analysis for the remaining 3 hypothesis questions I generated. Overall, I was curious about the following:

2. Is there a statistically significant difference in USA vs. London employee performance?
    - Null hypothesis: There is no difference in performance between the US and London employees
    - Alternative hypothesis: There is a difference in performance between the US and London employees

3. Is there a statistical significance between discounts given by the USA vs. London employees?
    - Null hypothesis: There is no difference in discounts given by from USA and UK employees
    - Alternative hypothesis: There is a difference in discounts given by from USA and UK employees

4. Do the USA or London employees have higher invoice totals?
    - Null hypothesis: London invoice totals are higher than USA invoices totals
    - Alternative hypothesis: London invoice totals are lower than USA invoice totals

Obviously, the null and alternative hypothesis were different for each question, but the process was similar for each. After identifying the null and alternative hypothesis, then it was onto the analysis by incorporating the necessary database via SQL. Then from there, I had to come up with the parameters for the statistical analyses. Then from there, I generated the mean and standard deviation. Lastly, I ran the test statistic and p values for each. I came up with the following conclusions:

2. Result: The null hypothesis showed that there is a statistically significant difference in employee performance between the USA and the London office

3. Result: The test shows that there is a statistically significant difference in discount amount between employees from USA and London (thus we reject the null hypothesis)

4. Result: Failure to reject the null hypothesis. Invoice totals from the London office are, on average, the same or higher than the invoice totals from the USA offices

After looking over all my hypotheses and determining if the null should be rejected or not, I learned that for the most part, all my hypothesis testing fell right in line with rejecting the null hypothesis. I was surprised and somehow, I felt that I had essentially failed in this challenge. But rejecting the null hypothesis isn't a bad thing. I think it just meant that I must figure out the parameters that determine a failure to reject the null hypothesis. So that most likely means that my sample size is probably limited (I think). Additionally, even though this company is fictional, I started brainstorming recommendations to help the company with their bottom line. To increase the revenue in sales, I thought it would be advantageous to offer more of the 5% discounts and less of the other discounts (since they did not generate much in terms of sales). Also, delving deeper into why Northwind customers respond to the 5% discount vs none or other amounts would be interesting to know. In terms of the Northwind employees, I would like to know why the USA and London employees on average had the same amount of invoices generated. I thought that since the USA was larger in terms of geography and regions covered, they would have the jump on London. But the data did not indicate that. Doing this exploration of Northwind’s database answered some questions, but it also raised more questions that need further examination. Maybe if we get assigned this dataset again, I can answer the additional questions this dataset generated.