August 14, 2019

Shaurya Mehta

NOrtheastern University

Summer 2019

ALY6070 Final project

by: Shrikant Palmer

For the final project, we were asked to find several insights from a Costco data set. In this paper, we will cover those questions and how we came to our insights. The first part was asking us to find three exploratory analysis from the data that we found interesting. Below is the result of our data cleaning.

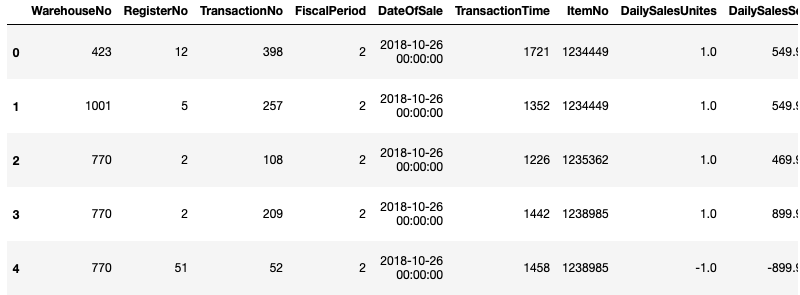


Figure 1.0. Data after the initial clean up.

Starting out, we had to clean the data. We had over 200,000 rows that were missing values across all columns except the member number and we believe it must have been from a typo error. Regardless, since most of the data was missing, we removed these rows from our data. We also changed some values in the columns from floats to integers and changed the date to the correct format to be able to further analysis. Also, the column names were renamed to prevent confusion. We found the column names were hard to remember and interpret so we changed them to what is shown above so that we could clearly understand what we were working on.

Our first analysis was revolving around finding which city had the largest number of members. Initially, we found that our answer to this question was Issaquah. But after the presentation, we were made painfully aware that our insight was incorrect. The reason was because we did not consider the E-commerce region which obviously did not have any physical stores. As a result, our graph was heavily skewed and showed Issaquah having over 200,000 members and over $222 million in sales. This was not correct so we removed the EC region so it could no longer have a bias in our data. As a result, we now have a new graph which showed that San Diego has the largest number of members.

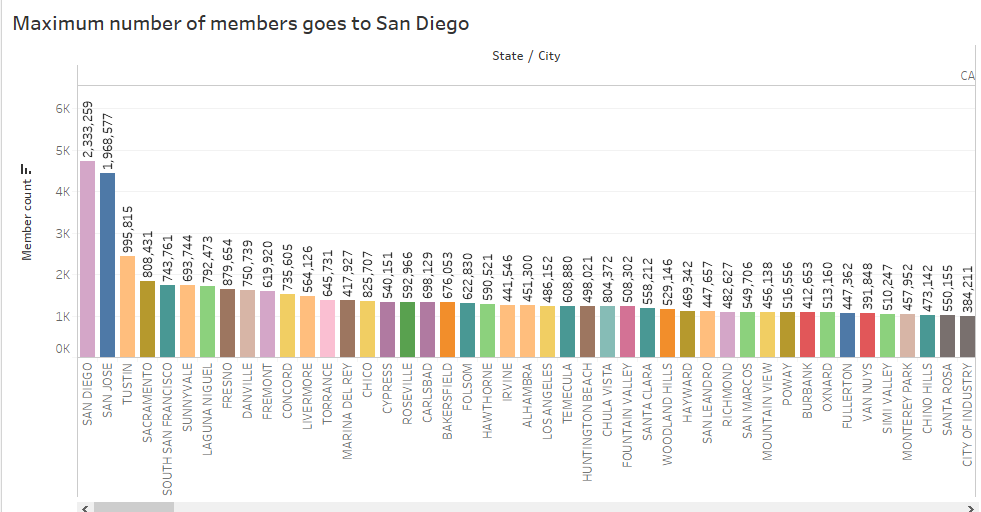


Figure 1.1. Graph showing San Diego as maximum numbers.

The second exploratory analysis was determining which fiscal period was the best for sales. Our data set was in seven fiscal periods which started with September. We assumed that the best fiscal periods were going to be those in November and December since those are the months with major holidays.

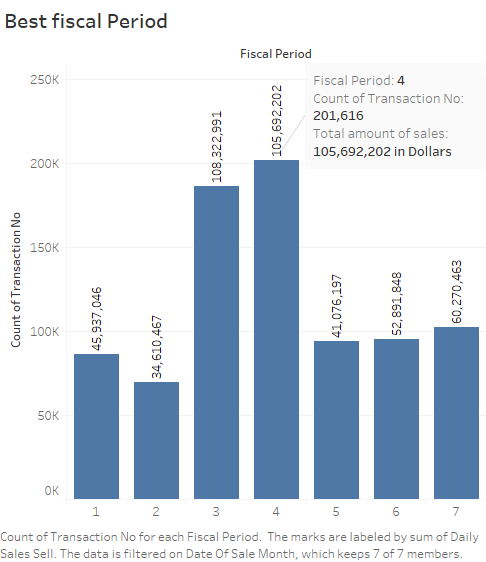


Figure 1.2. Fiscal Periods in the data set.

The figure above shows that as was our assumption, the best fiscal periods were 3 and 4 which correspond to November and December respectively. We wanted to find the best fiscal periods because that is usually how businesses go by in terms of timeline. Our last exploratory analysis involved trying to find the account type which had the highest number of members. The data set had several account types. They were the following: first was goldstar, the second was business members, third was employee, fourth was complementary, fifth was supply.

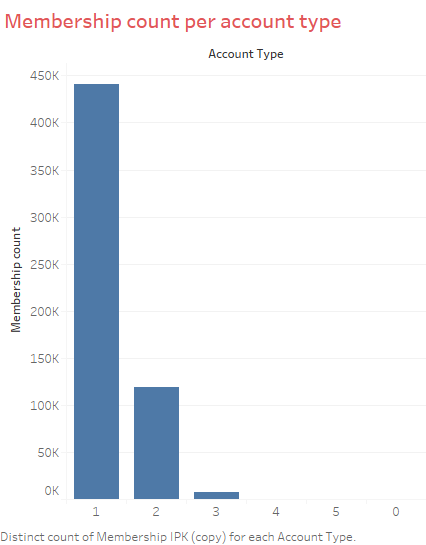


Figure 1.3. Showing the member count per account type.

With this analysis, we also assumed that the goldstar and business members would make up the bulk of our customers. As the graph above shows, goldstar members made up a significant amount of from the total membership count. Next was business members followed by Employee and the last 2 barely registered any members on the scale. So, our assumption was correct and the data backs it up.

The second part of the project involved answering the questions that were asked. The first question asked Which items should we stock more in which region, during different months of the year, based on the sales? This question has multiple facets so there was a lot of different aspects of the data we looked at. We needed a combination of regional data, different time periods of the year, and sales data.

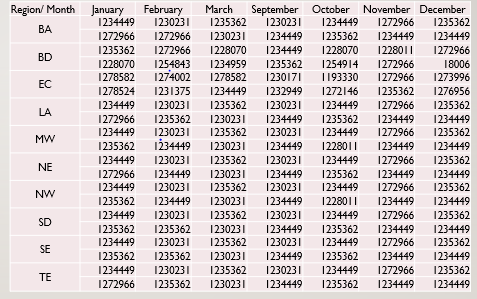


Figure 1.4. Showing the region sales per item in each month.

During the presentation, we found that this was complex for the class to understand and for us to explain as well. To make the graph simpler to understand, we made some changes and attempted to create a second graph in Tableau that would address the issue from the first. This was a very complicated to question to answer as well and we had to really think about how to rearrange the data in order to make it more presentable. The graph below is what we decided upon for the final iteration.

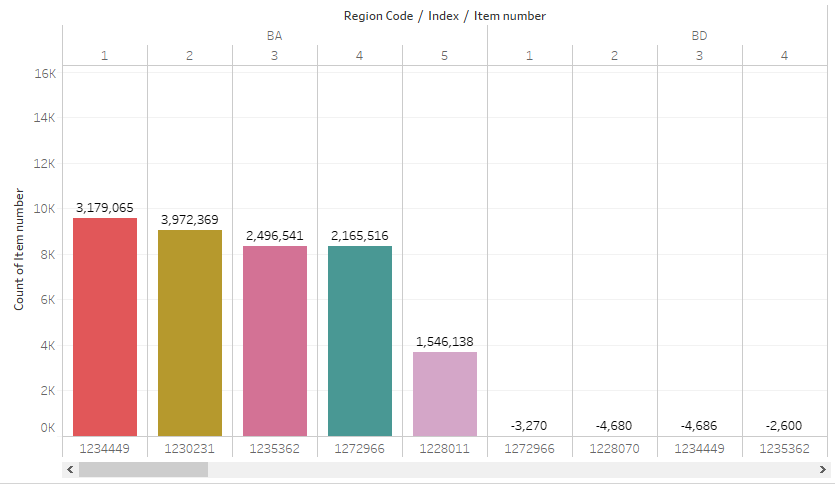


Figure 1.5. The second graph for part 1.

In this graph, the region is shown with the count of the item number. It also shows just the top five selling items in that region. Moving onto the second question of the report. Can we rank warehouses based on the sales of items in terms of quality as well as dollars? In order to answer this question, we initially started by taking each item and calculated the cost per unit and then divided by the total number of units. We then took this number and filtered it by each warehouse to come with a ranking for each warehouse. We also had some warehouses that were only in the negatives. There was also an issue with this graph as the negative scale could become confusing. So, we also made changes to this graph to make it simpler.

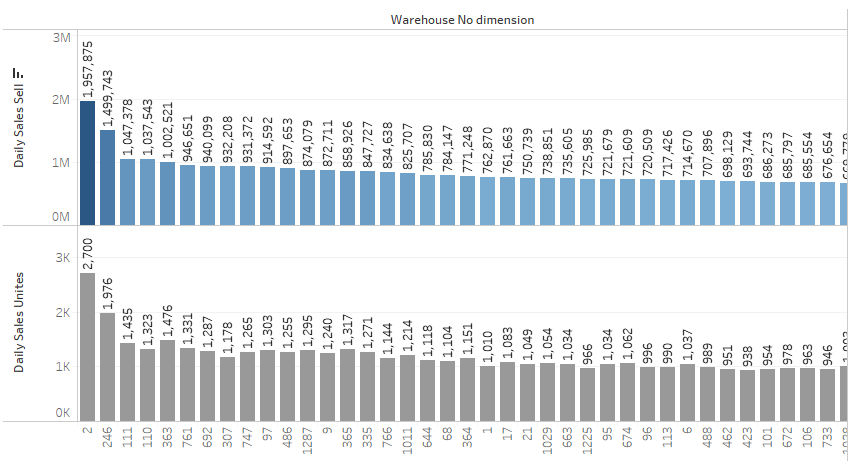


Figure 1.6. Daily sales per warehouse.

To make the data easier to visualize, we took the data and split it up in sales per unit and number of units sold. We also removed the negative scale so that it would not lead to misinterpretation of the data. The resulting graphs are a lot easier to interpret and make decisions from.

The last question for the project was Is there a correlation of items to account type of the members? This was possibly the most complicated question to among all three questions. Within the data set, we used the item number and account type and tried to create some sort of ranking to determine if there was a correlation. This was not an easy feat to do by any means and we went through many iterations of this question as well. For example, we tried to present this information with a heat map but found that it looked much too complicated. Not only was it hard on the eyes but we knew it would be even more difficult for the class to understand.

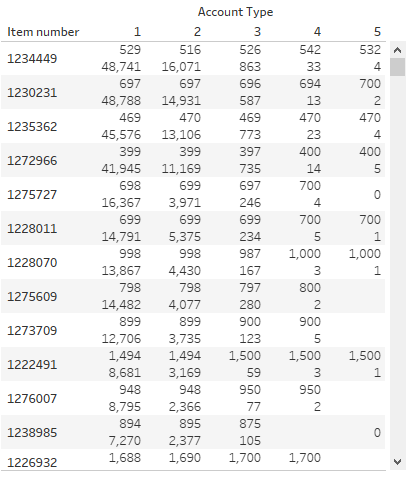


Figure 1.7. Graph for part C.

We decided this was going to our final iteration for this question. The heat map has been removed to just have the values in place. Above, the columns represent each account type and the rows are the item numbers. What we did is take the total number of units for each account which is represented by the 48,741 for item 1234449 as an example. Then we took this and divided it by the total number of sales for that item number and reached the 529 which represents the per unit cost for that account number divided up by the different account types. There is a difference in each account type and while we do not have a concrete reason for why that is, we believe it might be due to different prices throughout each region because normally the price should not change by account type.