Dear Team Leader,

Hi, my name is Surya Maniyan. I am a jr. Data Scientist. Here I have completed Exploratory Data Analysis in my fist data science project Gala Groceries supply chain issue. Groceries are highly perishable items. If you overstock, you are wasting money on excessive storage and waste, but if you understock, then you risk losing customers. They want to know how to better stock the items that they sell. Here are my findings and recommendations,

The distribution of unit\_price is positively skewed, that is, there are more sales of products with a low unit\_price compared to products with a high unit\_price.

The distribution of quantity is very different. We can see that only 4 unique values exist (1, 2, 3, and 4) and they are quite evenly distributed. It seems as though customers are buying in even quantities across 1 to 4 units.

The total follows a similar distribution to unit\_price. This you may expect, as the total is calculated as unit\_price x quantity.

However, this distribution is even more positively skewed. Once again, using intuition, this distribution makes sense. You'd expect customers at a grocery store to generally make more transactions of low value and only occasionally make a transaction of a very high value.

we can see that the product most frequently was sold within this dataset was ecac012c-1dec-41d4-9ebd-56fb7166f6d9, sold 114 times during the week. Whereas the product least sold was ec0bb9b5-45e3-4de8-963d-e92aa91a201e sold just 3 times.

There are 22 unique values for category, with fruit and vegetables being the 2 most frequently purchased product categories and spices and herbs being the least. Let's visualise this too.

From this sample of data, non-members appear to be the most frequent type of customers, closely followed by standard and premium customers.

Interestingly, cash seems to be the most frequently used method of payment from this sample of data, with debit cards being the least frequent.

From this we can see that the 11th, 16th and 18th hour of the day are the top 3 hours of the day for transactions being processed. This is interesting, this would suggest that their busiest times of day may be just before lunch, and as people are on the way home from work. Once again, this is a small sample of data, so we can't make assumptions on the population sample of data, but it gives us insights to go back to the business with.

This makes sense, you would expect a grocery store to sell more products that are cheap, and just a few products that are really expensive.

From this correlation matrix, we can see that the only columns that have a high correlation are unit\_price and total. This is understandable because total is calculated used unit\_price.

All the other correlations are close to 0, indicating that there is not a significant positive or negative correlation between the numeric variables.

We have completed an initial exploratory data analysis on the sample of data provided. We should now have a solid understanding of the data.

The client wants to know,

"How to better stock the items that they sell"

From this dataset, it is impossible to answer that question. In order to make the next step on this project with the client, it is clear that: We need more rows of data. The current sample is only from 1 store and 1 week worth of data. We need to frame the specific problem statement that we want to solve. The current business problem is too broad, we should narrow down the focus in order to deliver a valuable end product.

We need more features. Based on the problem statement that we move forward with, we need more columns (features) that may help us to understand the outcome that we're solving for.

Best Regards,

Surya Maniyan

Jr. Data Scientist