Jackie Gan

CIS 5526

Dr. Vucetic

Project Exploratory Data Analysis

In this project, we will predict the sales of thousands of products that are sold at Favorita store, a local grocery chain in Ecuador. From the Kaggle website, we were provided with a dataset that consist of train.csv, stores.csv, oil.csv, and holiday\_events.csv. Then we were provided with some additional notes, wages are paid every 2 weeks on 15th and last day of the month, we expect stores sales to increase throughout those time periods. Ecuador is a highly oil-dependent country, this might affect store sales due to higher item prices. Also, on April 16, 2016, there was a 7.8 earthquake in Ecuador, people rallied relief efforts of donating water and other essential need products, this might cause in drop of store sales.

Throughout the Exploratory Data Analysis, we first explored the oil and train dataset. In the oil dataset, we saw that the oil prices have dropped significantly from early 2013 at 110 dollars a barrel to as low as 26 dollars a barrel in early 2016 and it stabilized around 40 – 50 dollars around 2016 to late 2017. Chart, line chart

Description automatically generated

How does it compare to sales? In this graph, we can see that sales have increased year after year, but it did not have a drastic change. This means, certain categories of sales have gone up but not necessarily means people are spending more.Chart, line chart

Description automatically generated

How does it compare to transactions? The transaction has been stable over the last couple years. This shows that the amount of shopping has not changed.Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedWhat is the correlation? Even though Ecuador is an oil dependent country but the correlation between sales and transaction is weak, and we can conclude oil is not a significant feature for sales and transaction.

Table

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedNext, we will explore the family category. In this bar chart, we found that the top 5 category sold are Grocery, Beverages, Produce, Cleaning and Dairy. At the same time, we saw that the items that are promoted are the same that are the most sold. This implies promotion increases sales.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedNow, we will explore the sales by date and holidays. In this chart, we explored that most of the sales are made during end of the week. We expect the workers are free from work and can shop more often. In the sales by month chart, as the month increased the more spending is done throughout the year. The top spending year is December due to Christmas and New Years. In the holiday sales chart, we saw that most spending were bridge, transfer, and additional days. This is because they are all related to specific holidays, and they are usually observed holidays days for example Christmas eve is an additional day and bridge is an extension of holiday for the weekend. Lastly, we saw the sales has increases year after year.

Chart, bar chart

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Afterall, we can conclude Oil is not a reliable feature to conclude sales, however there are other features that are very important such as the date. For example, we saw a huge portion consumer spending on the weekends. They also spend a lot towards holidays such as Christmas and New Years, which shows a high spending record during the month of December. In this EDA, we also explored the categories what consumer typically spend on. We found, more than 50% of the sales consist of grocery and beverages. Next followed by produce, cleaning, and dairy as the next highest. So, the family category is an essential part of the project, we will emphasize heavily on holidays and family category.