Corporacion Favorita Grocery Sales Forecasting
Milestone Report 1

George Tang

1 Introduction

The importance of sales forecasting cannot be overstated for any brick and mortar retail business. Managers leveraging accurate forecasts can ensure their stores have the right products in stock at the right time. This in turn provides a positive shopping experience which is critical in maintaining customer satisfaction and their retention. Also, it helps businesses to control item inventory and improve shelf space usage, reduce perishable goods wastage, lower labor and transportation costs, and ultimately increase revenue and profit.

These benefits motivated Corporacion Favorita to host Kaggle's Corporacion Favorita Grocery Sales Forecasting competition in 2017. Corporacion Favorita is an Ecuadorian-based grocery retailer that operates hundreds of supermarkets with over 200,000 different products. At the time, the business relied on subjective forecasting methods backed by little data analytics. They used the competition to challenge data scientists to develop innovative, data-driven solutions that can ultimately improves their sales forecasts.

1.1 Objective

The objectives of this project are:

- To Identify features that affect sales, both short-term and long term
- To develop sales forecasting models and evaluate their performances
- To provide actionable recommendations

1.2 Significance

Through this project, we will identify important features that determine sales. The company can use the sales forecasts made by machine learning models to plan its logistics and inventories, which allows them to better utilize their resources and ultimately maximize profit.

1.3 Programming Code

The programming code can be found here.

2 Dataset

The data used in this project are obtained from the Kaggle competition's website, which contains 7 data tables. We will provide a description of each data table in this section.

2.1 Sales Data

The train dataset (train.csv) consists of the sales data for each item sold in each store. It consists of 125,497,040 rows and 6 columns namely, Id number, Date, Store Number, Item Number, Unit Sales and Promotion Status.

The test dataset (test.csv) is also provided for competition. Nonetheless, we will not use it in this project because the true target values (daily unit sales for each item at each store) are not provided, which makes model validation impossible.

2.2 Transaction Data

This dataset (transactions.csv) consists of the transaction information for each store. It contains 83,488 rows, one row for the number of transactions of the date at that store. The three columns are Date, Store Number and Number of Transactions.

2.3 Store Information

This dataset (stores.csv) consists of 54 rows, one for each of the stores owned and operated by the grocery store chain. The 5 columns are: Store Number, City, State, Type, and Cluster.

2.4 Item Information

This dataset (items.csv) consists of 4,000 rows, with each row consisting of information of one grocery item. Note that it is a subset of all products carried by this company. The 4 columns are: Item Number, Family, Class and Perishable Status.

2.5 Holiday Events

This dataset (holiday_events.csv) consists of holidays and events information of the country.

2.6 Oil Price

Since Ecuador is an oil-dependent country, Ecuadorians' spending can be tied to changes in oil prices. This dataset (oil.csv) contains two columns, namely Date and Oil Price.

2.7 Supplementary Information

Additionally, the competition's website specifically provided two pieces of information:

- Wages in the public sector are paid every two weeks, on the 15th and on the last day of the month;
- A magnitude 7.8 earthquake struck the country on April 16, 2016. People rallied in relief efforts donating water and other products which greatly affected supermarket sales for several weeks.

3 Data Cleaning and Wrangling

Since the data is provided for a Kaggle competition, all the datasets are relatively clean. Only the Sales dataset requires some data cleaning and wrangling steps which are summarized as the following:

- Change Boolean data type from string to Boolean
- Change datetime data type from string to datetime
- Create new columns of Year, Month, Day of Month and Day of Week based on the Date column
- Impute missing values for the Promotion Status column

We also merge datasets with common columns (e.g. Date, Item Number, Store Number, etc.) for data visualization and statistical analysis as needed.

4 Data Storytelling

4.1 Store Information

The number of stores in each city is shown in Figure 4-1(a). In total, the 54 grocery stores are spread across 22 cities in 16 states. Quito, the capital city of Ecuador, consists of the most stores with 18. Guayaquil, the largest city in the country, contains 8 stores.

Figure 4-1(b) shows the number of stores for each type. The five store types are Megamaxi, Supermaxi, Gran Akí, Super Akí and Akí, although their corresponding labels (A to E) are not provided. The most common store type is D (18), followed by C (15), while the least common type is E (3).

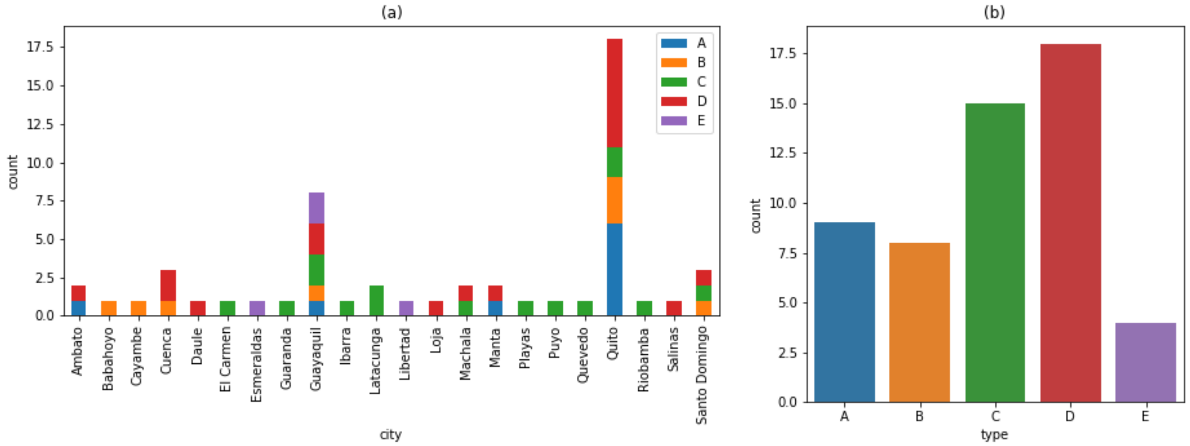

Figure 4-1 Store count (a) by city, (b) by type

Store clustering is based on store size, category sales volume and shopper purchase behavior. Figure XX shows that there are 17 clusters. Except for Cluster 10, all other clusters consist of only one store type. Cluster 3 consists of the most stores (7), followed by Clusters 6 and 10 (with 6 each). Each of Clusters 5, 12, 16 and 17 consists of only 1 store.

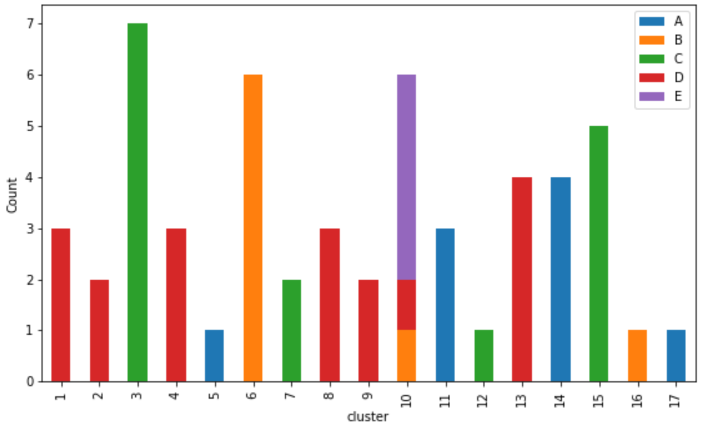

Figure 4-2 Store count by cluster

4.2 Item Information

Figure 4-3 shows the number of items for each product category. In total, there are 4,400 items in the dataset, which is a subset of the 200,000+ items carried by the company. Grocery I contains the highest number of items (1,314), followed by Beverages (613) and Cleaning (446). Baby care, Books and Home Appliances contain the fewest items with only one each.

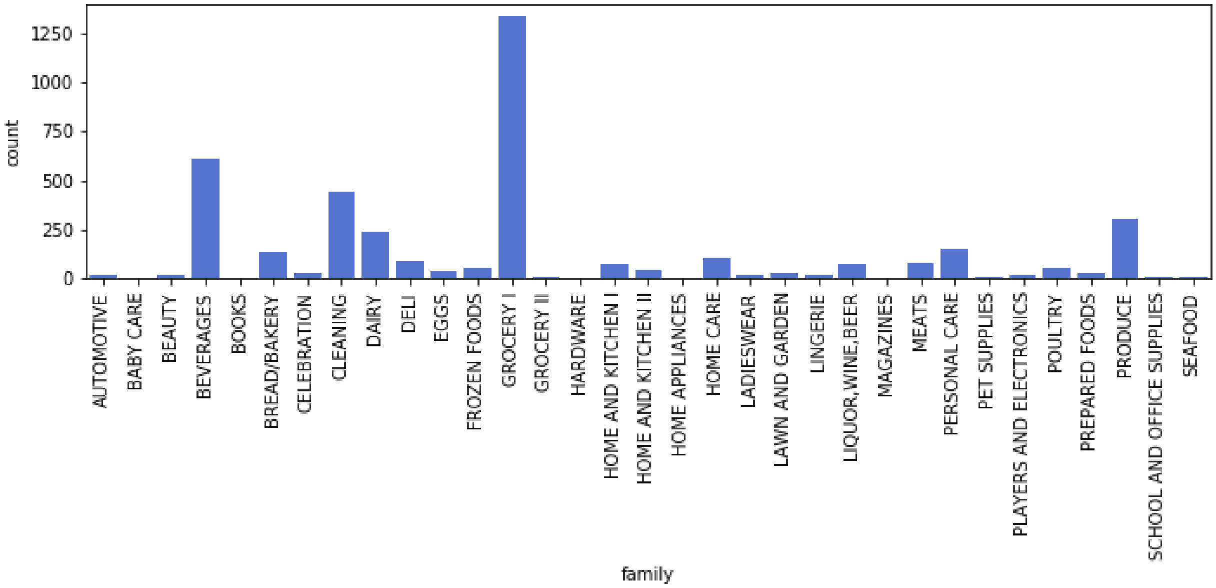

Figure 4-3 Item count by family

All items in each family are either perishable or non-perishable. Families of Bread/bakery, Deli, Poultry, Eggs, Dairy, Seafood, Prepared Foods, and Produce are perishable, which comprises of 986 items (24% of total items). This information is important to the company because unsold perishable products would go to waste and lead to a permanent loss in revenue.

4.3 Oil Price

Since Ecuador is an oil-dependent country, the nation’s consumer behavior may be affected by fluctuations in oil price. Daily oil price from 1/2013 to 8/2017 is shown in Figure 4-4. The figure shows that there is a substantial decline in oil price beginning in the end of 2014. Prior to that, oil price hovers around $100/barrel with maximum value of $110.62/barrel. After that, oil price hovers around $50/barrel with minimum value of $26.19/barrel on 2016-02-11.

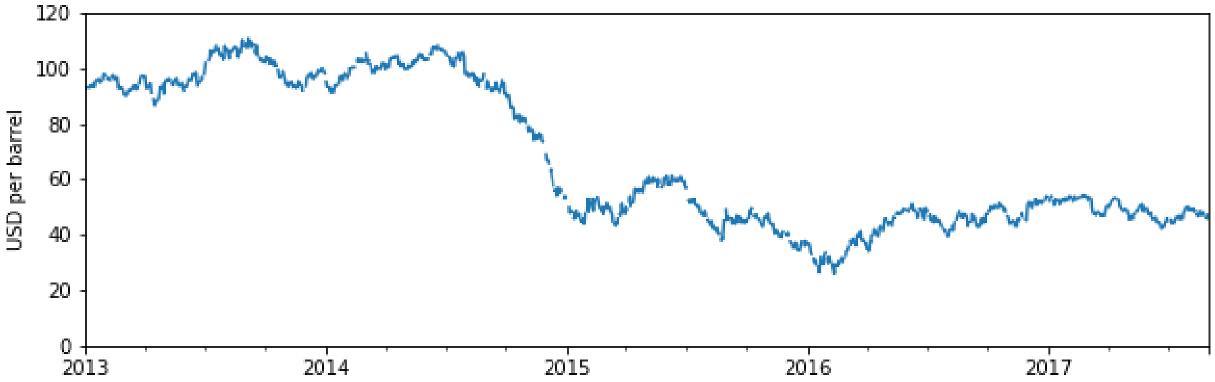

Figure 4-4 Daily oil price from January 2013 to August 2017

4.4 Holidays events

This dataset provides information on the holidays and special events in Ecuador. Some holidays and events are national while others are local. The most notable events are the World Cup in the Summer of 2014, and the earthquake event that happened in 2016-04-16.

4.5 Sales Data

The sales data consists of item sales information at each store from 2013-01-01 to 2017-08-15.
One objective of this project is to develop forecasting model to predict the unit sales for each item in every store. While it is appropriate to analyze the data at this level of detail (each item in each store), the vast number of store-item combination make this approach impractical for this report. As such, we will choose a few representative stores from one city, and a few item families for analysis.

We choose the city Quito because it consists of the highest number of stores, which means the sales records in that location are representative of the overall business of the company. It also contains stores of the four major types (A, B, C and D).

We choose the item families of “GROCERY I”, “LIQUOR, WINE, BEER”, and “PRODUCE” with the reasoning provided as follows.

- GROCERY I consists of the most items in the non-perishable category
- Intuitively, LIQUOR, WINE, BEER would be one of the item families affected by oil price fluctuations
- PRODUCE consists of the most items in the perishable category

Stores #1 (Type D), #3 (Type D), #9 (Type C), #10 (Type B), #45 (Type A) and #46 (Type A) are chosen for this study so that all store types available to the city are covered, with two stores for each of the most common store types (D and A).

4.5.1 GROCERY I

The time series plot of the sum of unit sales in the GROCERY I family for the different stores shown in Figure 4-5, with data aggregated weekly for better presentation. Note that the figure presents the sum of unit sales of all items regardless of their differences. While this approach is crude (e.g. adding boxes of napkins to packets of batteries etc.), it is useful in providing an overall summary of the sales data.

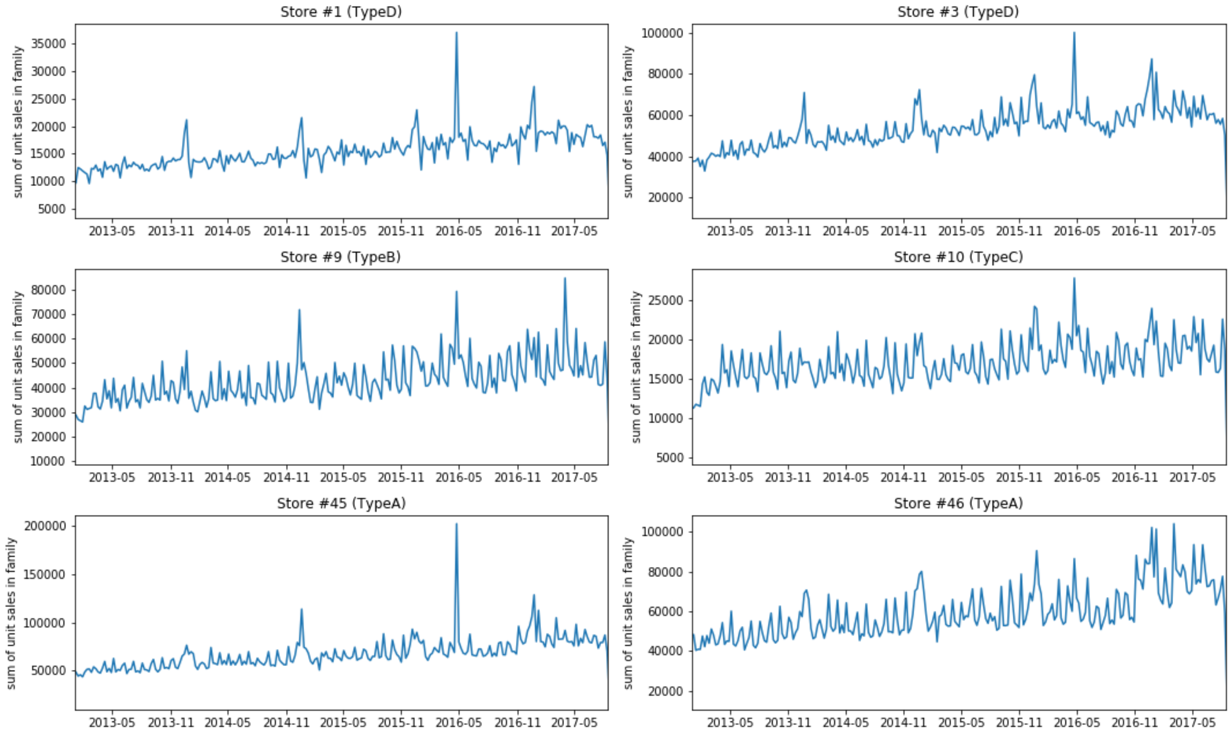

Figure 4-5 Sum of unit sales of GROCERY I items (aggregated weekly)

The figure shows that, even for stores in the same store type, the amount of unit sales can be vastly different. For example, the total unit sales of Store #3 are significantly higher than those of Store #1, even though both belong to Type D.

Based on the figure, it is difficult to identify any long-term trends. We can use statistical tests to validate the significance of any trends in Milestone Report #2.

4.5.1.1 Effect of Weekday

The effect of day of the week in sales is shown in Figure 4-6. All stores, except for Store #1, show increased sales during the weekends (Saturday and Sunday). For Store #1, sales are similar from Monday to Saturday, but show a significant dip on Sundays.

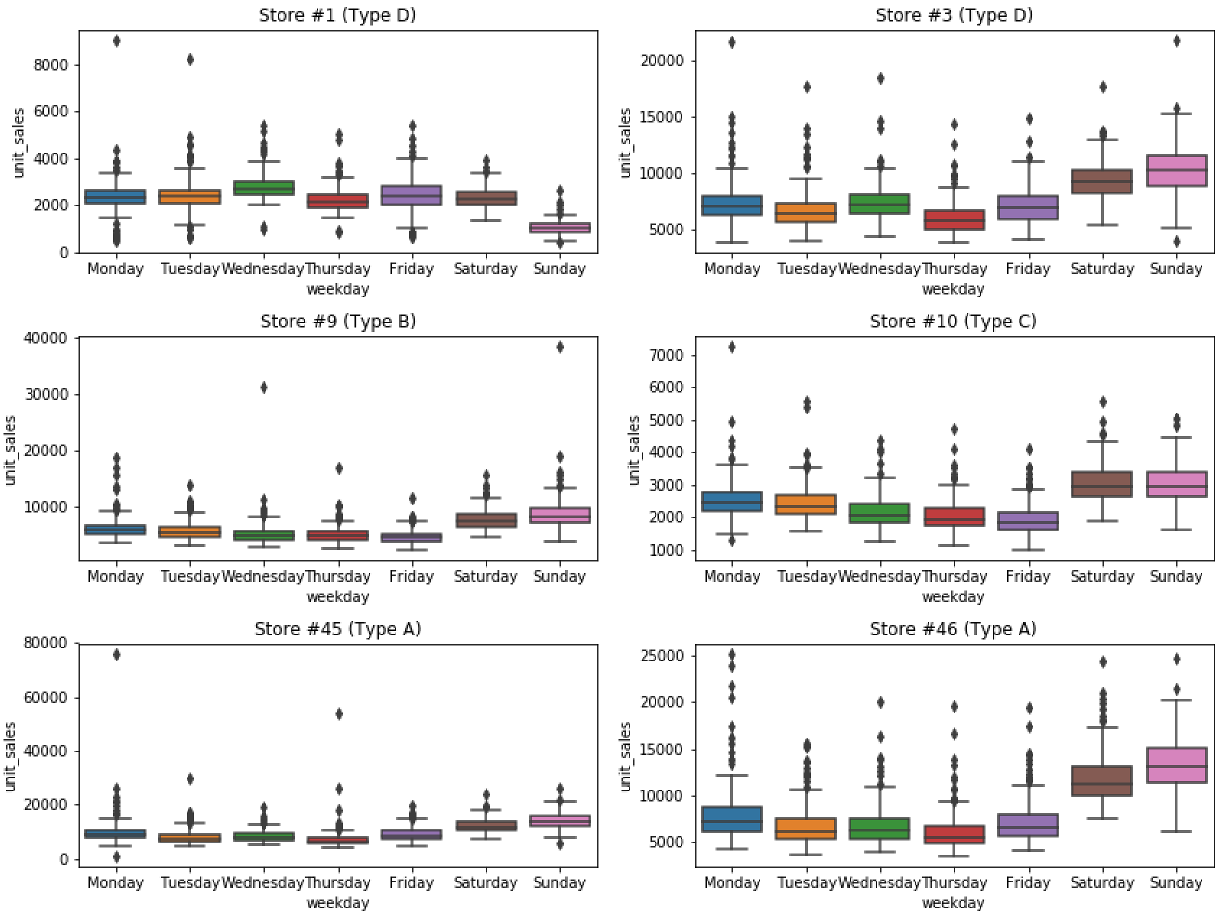

Figure 4-6 Sum of unit sales of GROCERY I items by day of the week

4.5.1.2 Effect of Holidays and Events

Figure 4-7 shows the daily sum of GROCERY I item unit sales for Store #3 between December 2016 and January 2017. We can see the seasonal pattern of sales which peaked on Sundays (e.g. December 4th, December 11th etc.). Increased in sales are observed from December 21st to 24th, as well as January 2nd, the day after New Year. The store was closed on both Christmas day and New Year day. Similar trends are also present in other years, as well as in Stores #1, #9 and #45.

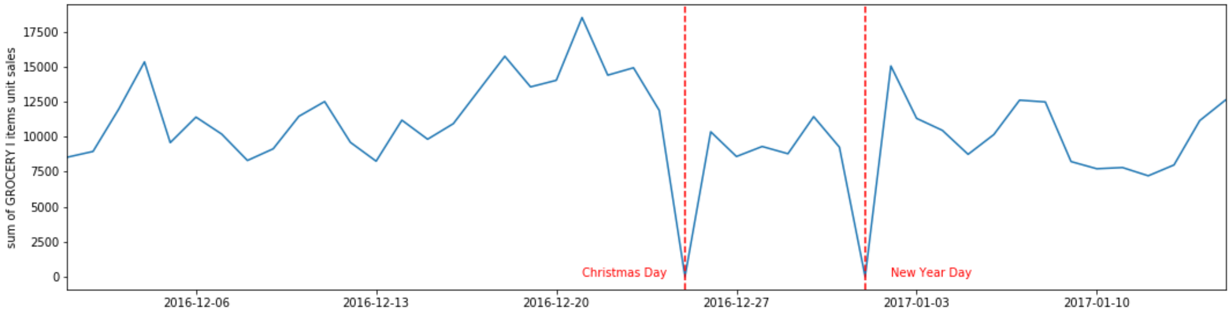

Figure 4-7 Sum of GROCERY I items unit sales in Store #3 between December 2016 and January 2017 (aggregated daily)

Figure 4-8 shows the daily sum of GROCERY I item unit sales for Store #3 during April 2016. A peak in sales is observed on April 17th and 18th, the days after the earthquake on April 16th. This corresponds to the national charity drive where people bought and donated supplies to the victims. Similar trends were also observed in some other stores.

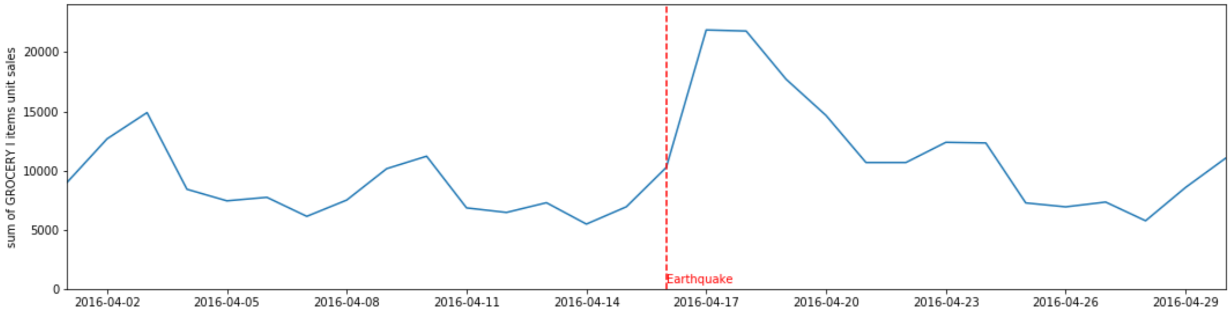

Figure 4-8 Sum of GROCERY I items unit sales in Store #3 during April 2016 (aggregated daily)

4.5.2 LIQUOR, WINE, BEER

The time series plot of the weekly-aggregated sum of unit sales in the GROCERY I family for the different stores shown in Figure 4-9. Intuitively, this item category will be affected by the drop of oil price around the end of 2014. However, there is no noticeable change in overall purchase for these six stores.

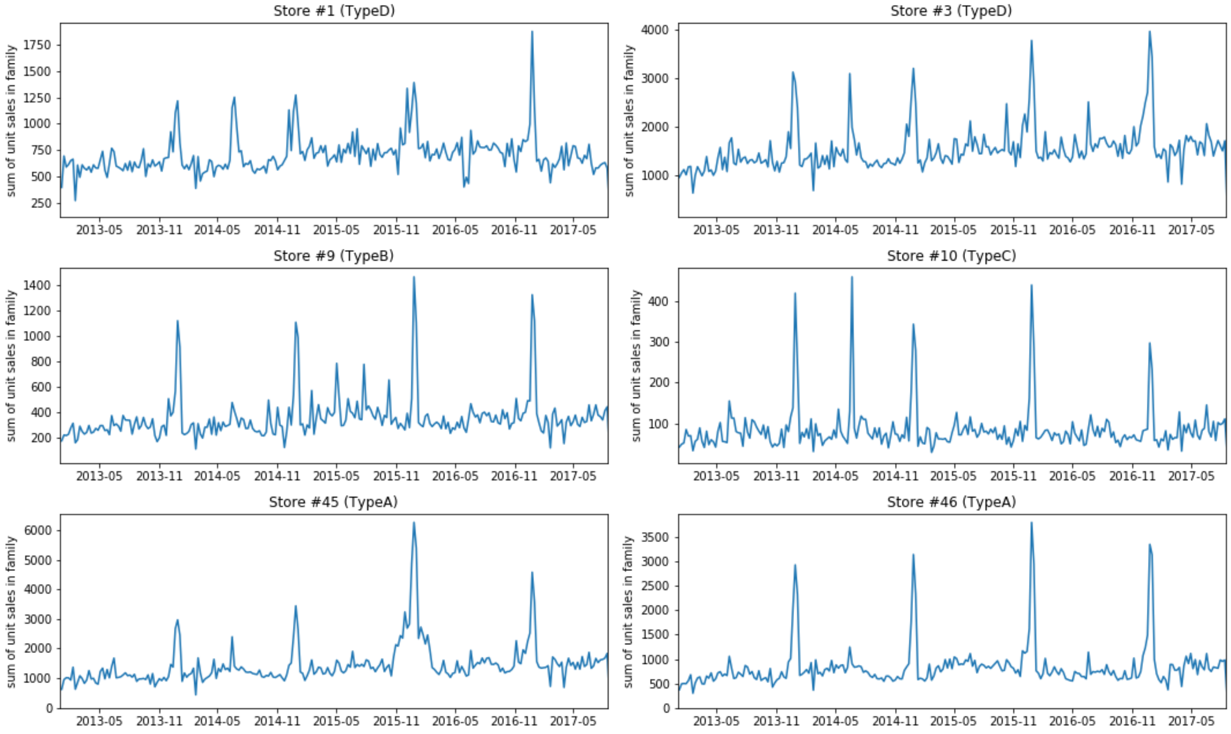

Figure 4-9 Sum of LIQUOR, WINE, BEER items unit sales (aggregated weekly)

4.5.2.1 *Effect of holidays and events*

Figure 4-10 shows the daily sum of LIQUOR, WINE, BEER item unit sales for Store #3 between December 2016 and January 2017. Increased in sales are observed prior to the two holidays. Similar trends are also present in other years, as well as in some other stores.

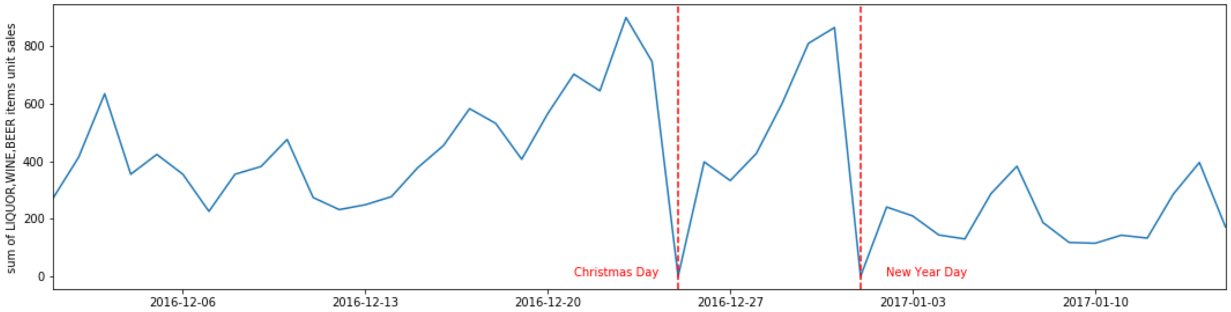

Figure 4-10 Sum of LIQUOR, WINE, BEER items unit sales in Store #3 between December 2016 and January 2017 (aggregated daily)

The Ecuador national team was part of the World Cup football tournament 2014, which was held in the summer of that year. Ecuadorians are famous for their passion for the sport. Figure 4-11 shows the daily sum of LIQUOR, WINE, BEER item unit sales for Store #3 in June 2014. We see the seasonal pattern of alcoholic beverage sales which peaked on Saturdays during the period. We can also see an increase in sales that corresponded to the dates when the team played. It seems that during this period, alcohol sale was prohibited on Sundays, and therefore the sales increase due to Game 1 occurred the day before. Similar trends are also present in some other stores.

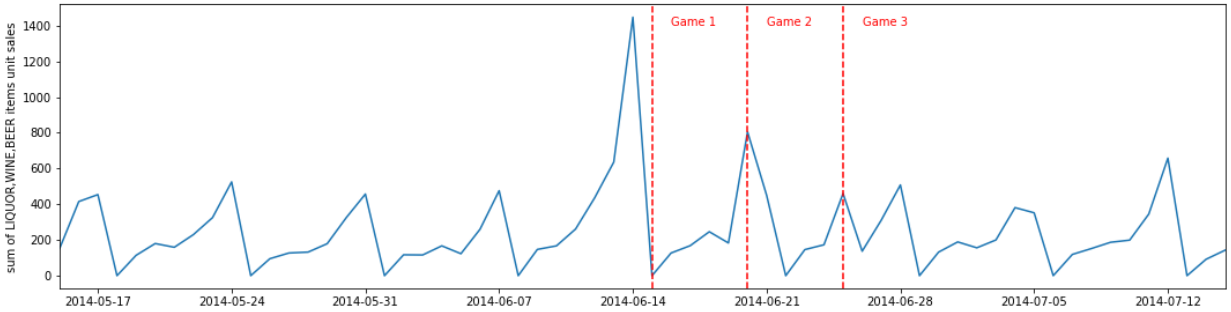

Figure 4-11 Sum of LIQUOR, WINE, BEER items unit sales in Store #3 during World Cup 2014 (aggregated daily)

4.5.3 PRODUCE

The time series plot of the weekly-aggregated sum of unit sales in the PRODUCE family for the different stores shown in Figure 4-12. We chose a begin date of 2015-08-01 because, prior to that, there is no sale for most days. There is no obvious trend for all six stores.

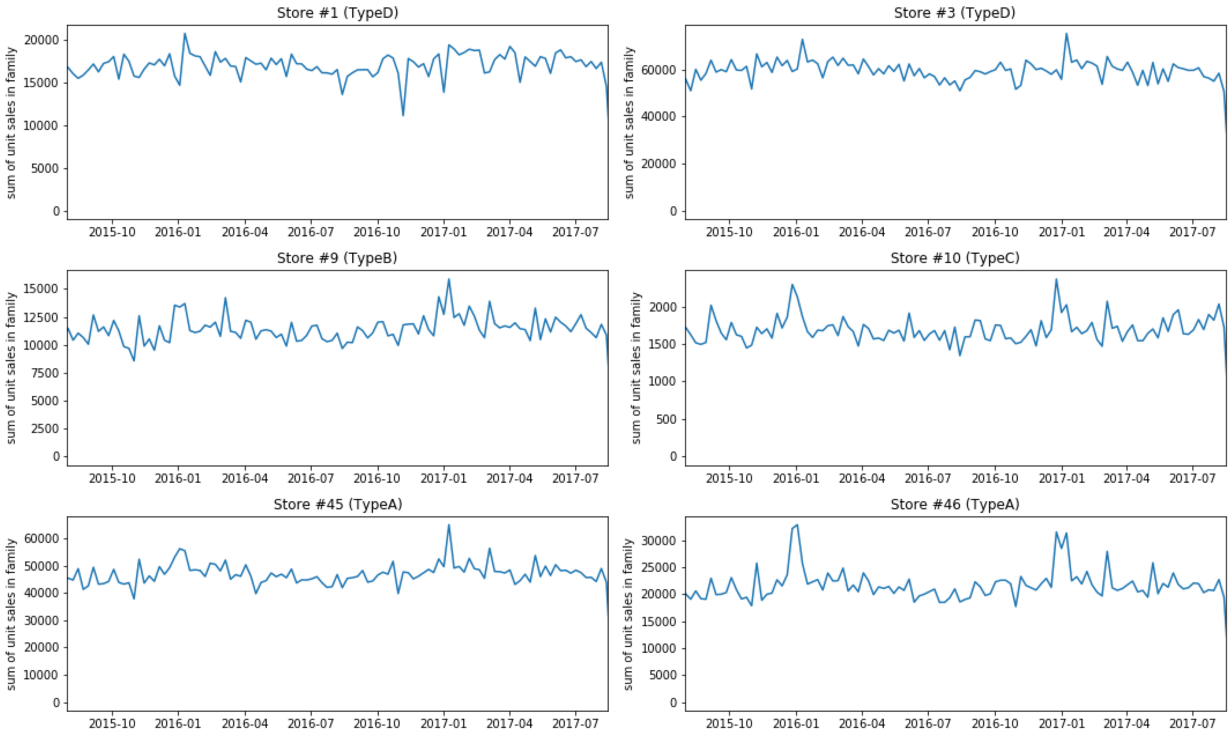

Figure 4-12 Sum of PRODUCE items unit sales (aggregated weekly)

4.5.3.1 Effect of holidays and events

Figure 4-13 shows the daily sum of PRODUCE item unit sales for Store #3 between December 2016 and January 2017. Interestingly, sales follow a seasonal trend, with sales typically peak on Wednesdays in this store. There is increased sale on December 23rd and 24th, days prior to the holiday.

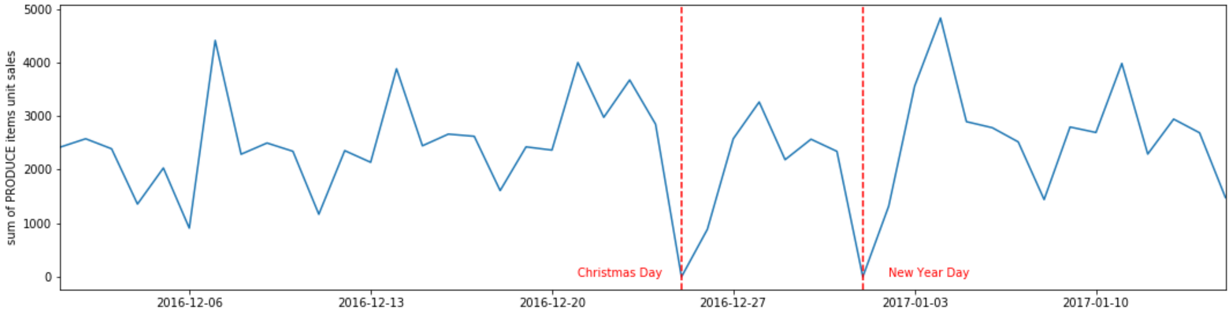

Figure 4-13 Sum of PORDUCE items unit sales in Store #3 between December 2016 and January 2017 (aggregated daily)

4.6 Statistical Inference

In this section, we conduct statistical hypothesis test to examine whether the unit sales (target variable) at different values of a categorical variable are from the same population or not. This way, we can identify features that may be helpful in predicting sales.

Here we consider the population to be a specific item in one store. Given the numerous store-item combination, we choose Store #3 because its overall unit sales are among the highest in all stores. We choose one of the top sellers in the three families examined above, namely, GROCERY I, LIQUOR,WINE,BEER and PRODUCE.

4.6.1 Effect of Day of Week on Unit Sales

Figure 4-14 shows the distribution of unit sales of the three chosen items between August 1, 2016 and August 15, 2017, the most recent 12-month period in the dataset. The figure shows that the effect of days of the week are different among the products. For GROCERY I Item #314384, unit sales are higher during the weekends and lower on Thursdays. For PRODUCE Item #1503844, unit sales are higher during the weekends and on Wednesdays. The high sale on Wednesdays is due to promotion which was on every Wednesday but not other days of the week. For LIQUOR,WINE,BEER Item #1004550, the highest unit sales occurred on Saturdays, followed by Fridays.

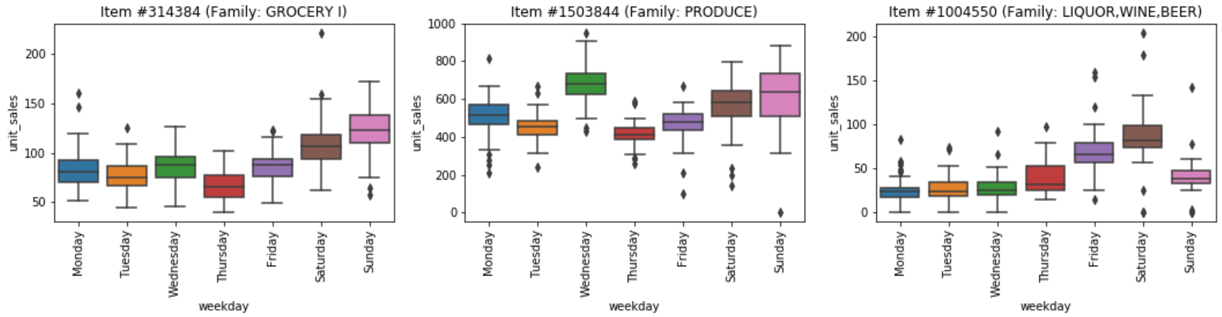

Figure 4-14 Unit sales distributions of selected items in Store #3 from August 1, 2016 to August 15, 2017

4.6.2 Effect of Promotion on Unit Sales

Figure 4-15 shows the effect of promotion of the three selected items. Among them, promotion status has the highest effect on PRODUCE Item #1503844.

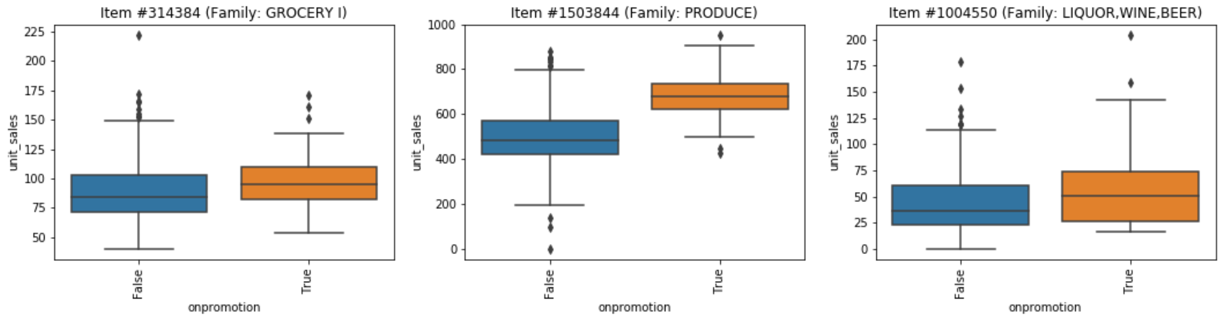

Figure 4-15 Unit sales distributions of selected items in Store #3 from August 1, 2016 to August 15, 2017

4.6.3 Statistical Analysis

In this section, we will conduct hypothesis tests on the two categorial features, namely days of the week, and promotion status. The null hypothesis is that the distributions of unit sales are the same among all categories within a feature, where the alternative hypothesis is that the distributions are not the same. To account for potential the non-normality of unit sales distribution, Mann Whitney U (MWU) test will be used for features with binary response (1/0, or t/f, etc), where Kruskal-Wallis (KW) test will be used for multi-category response. Both tests are non-parametric and do not assume any distribution of the samples. The hypothesis tested by both MWU test and KW test is whether the samples from the different categories are taken from the same population.

The results show that, for all features, the null hypothesis is rejected, i.e unit sales does not follow the same distribution among all categories in each feature. It means that the two features can be useful in predicting unit sales.

5 Summary

In this Report, we developed an understanding on the sales history for three item families in six stores. We find that:

- Sales pattern of different items can be different among the different stores;
- Sales typically follow a weekly seasonal pattern;
- Holidays and special events may impact sales of some item families;
- Unit sales in different days of the week and promotion status are different in some selected items.

In the next report (Milestone Report #2), we will analyze the data with statistical methods as well as develop and evaluate machine learning models for sales forecasting.