

**Corporación Favorita Grocery Sales Forecasting**

Raksha Kaverappa

Emily Strong

Index

Abstract 4

Exploratory Data Analysis 6

3.1 <EDA1> 9

3.2<EDA> 9

3.3<EDA> 10

3.4 <EDA> 11

Feature Selection……………………………………………………………………………………………......11

Clustering……………..………...……………………………………………………………………………11

4.1 Unit Sales Clustering………………………………………………………………………………….11

4.2 Transactions Clustering………………………………………………………………………………..11

Time Series Forecast 11

5.1 Estimation and Forecasting using ARIMA 12

Azure Experiment-Unit Sales 14

4.1 Data collection 14

4.2 Data Cleaning 15

4.3 Compute the Events 22

4.4 Predict the ‘Events’ using dependent variable mentioned above. 25

4.5 Create a Web Service for the above model 26

Web Application 30

Reference 32

# **Abstract**

The dataset chosen is a Kaggle competition dataset hosted by Corporación Favorita. This is an Ecuadorian grocery chain with over 100 stored carrying over 200,000 products. Currently we are predicting the sales of just grocery, but this can be further extended to other classes of groceries and beauty products which you might find a super market.

The link to our dataset is <link>

The dataset has the following files and properties:

* Train.csv: Consists of train data with unit sales per iter per day.
* Stores.csv: Consists of all the stores, their location and their individual store numbers
* Items.csv: Consists all the items, their family, classes and the item number
* Holidays.csv: Consists of the holidays and events metadata.
* Oils.csv: Consists of Daily oil prices.

We are predicting the Unit Sales for the grocery items by clustering them based on the item classes. We have used Neural Networks to predict the Unit sales of the items.

We are also forecasting the future transactions of each store and studying the effect of oil prices on the transactions since Ecuador is an oil dependent country.

# **Exploratory Data Analysis**

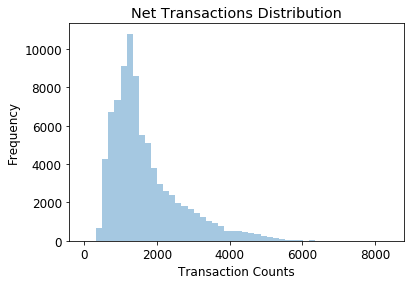
The Stores are distributed across Ecuador and we got the following plot for the same:

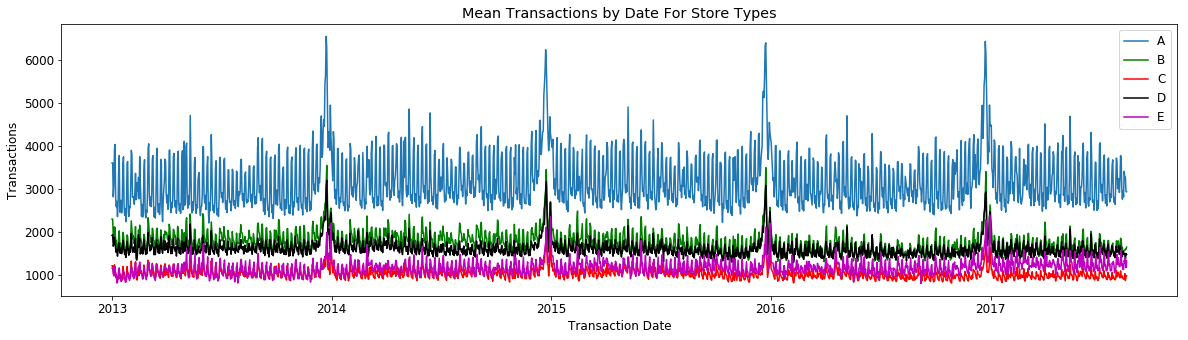


Figure 1: Density of Stores Per City. Legend: Black = 1 store, Blue = 2 stores, Purple = 3 stores,

Red = >5 stores, Orange = >10 stores

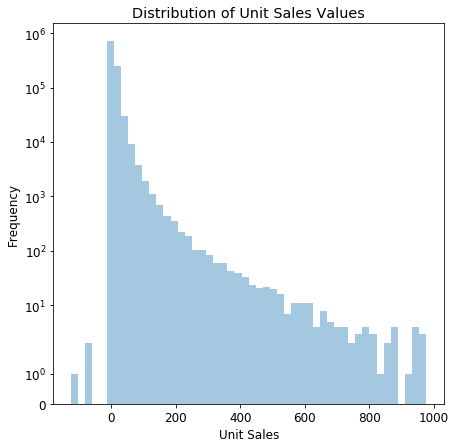
The frequency distribution of transactions and the mean transactions by date for each store type is as shown below:

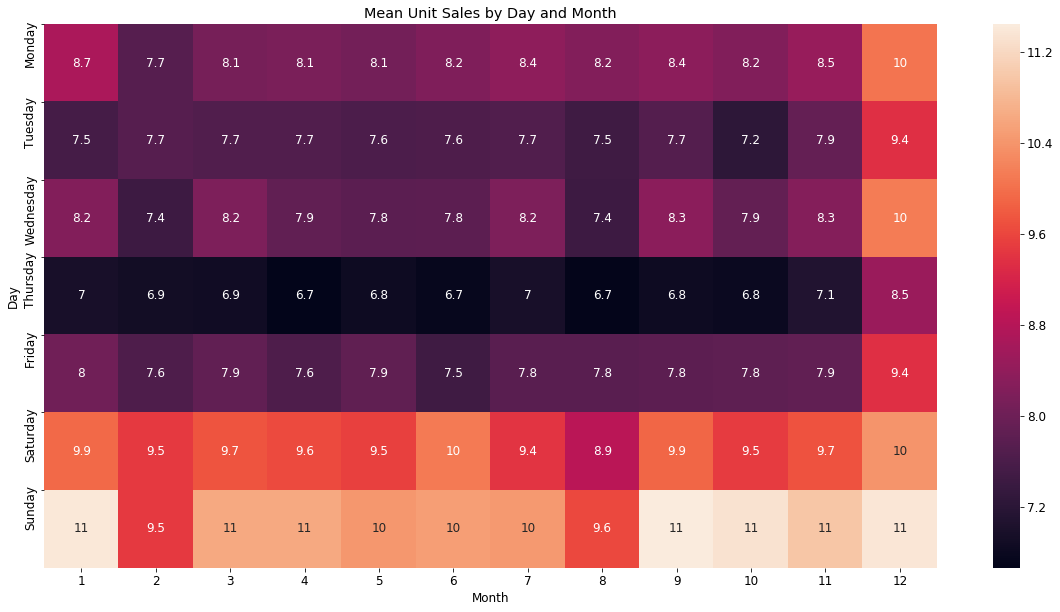




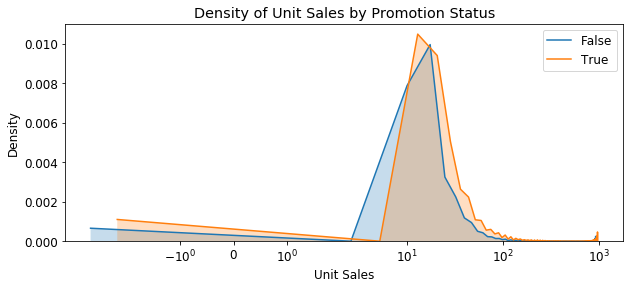
We can see that the transactions based on stores can we divided into 5 clusters. We have used this analysis for accurately clustering out Transactions data for time series forecasting.

The distribution of unit sales and the mean unit sales by day and month is as shown below:

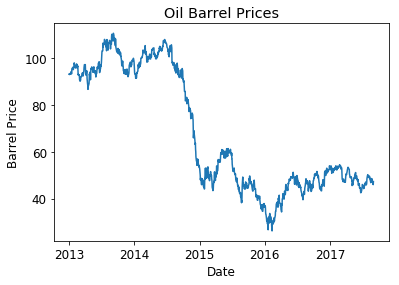




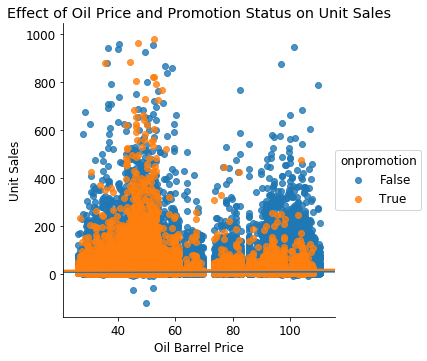
We can clearly see that the unit sales for every store was maximum on the weekends.



We also observed that there wasn’t much variation in unit sales when item were on promotion.



We also plotted the trend of oil prices by date. Further, we inspected the variation of unit sales based on oil prices and grouped them on the basis of promotion.



# **Clustering**