## horizontal line



[Corporación Favorita Grocery Sales Forecasting](#_au51mny0sx6)

April 09,2019

**─**

Ramakrishnan Sundararaman

Springboard Data Science Career Track - Capstone Project 1 - Proposal

New York, NY

# 

# Overview

Brick and mortar / traditional grocery stores ( especially nationwide chain stores) have always had the issue of balancing purchasing and sales forecasting. This activity gains significant importance when it comes to perishable goods with limited shelf-life.If the quantities purchased are less the stores run the risk of dissatisfied customers. On the other hand if quantities purchased are more than requirement then it leads to financial losses. This project is an attempt to develop an analytical model that can be used by the purchasing department(s) to forecast sales and use the prediction methodology to procure the required items.

# Goals

1. To use Exploratory data analysis, data visualization and other techniques to analyse the dataset
2. To develop a predictive model that can be used to forecast what items are expected to be sold in what quantity and at store /group of store levels to enable purchasing departments make data driven decisions

# Data:

This was a competition hosted by kaggle.com. The client is a group of grocery stores in Ecuador and the data set has been provided by the client through kaggle.com. I will be using the various data sets provided by the client(s) to develop the analytical model using Python and Jupyter notebooks.

## File Descriptions and Data Field Information

### train.csv

* Training data, which includes the target unit\_sales by date, store\_nbr, and item\_nbr and a unique id to label rows.
* The target unit\_sales can be integer (e.g., a bag of chips) or float (e.g., 1.5 kg of cheese).
* Negative values of unit\_sales represent returns of that particular item.
* The onpromotion column tells whether that item\_nbr was on promotion for a specified date and store\_nbr.
* Approximately 16% of the onpromotion values in this file are NaN.
* NOTE: The training data does not include rows for items that had zero unit\_sales for a store/date combination. There is no information as to whether or not the item was in stock for the store on the date, and teams will need to decide the best way to handle that situation. Also, there are a small number of items seen in the training data that aren't seen in the test data.

### test.csv

* Test data, with the date, store\_nbr, item\_nbr combinations that are to be predicted, along with the onpromotion information.
* NOTE: The test data has a small number of items that are not contained in the training data. Part of the exercise will be to predict a new item sales based on similar products..
* The public / private leaderboard split is based on time. All items in the public split are also included in the private split.

### sample\_submission.csv

* A sample submission file in the correct format.
* *It is highly recommend you zip your submission file before uploading!*

### stores.csv

* Store metadata, including city, state, type, and cluster.
* cluster is a grouping of similar stores.

### items.csv

* Item metadata, including family, class, and perishable.
* NOTE: Items marked as perishable have a score weight of 1.25; otherwise, the weight is 1.0.

### transactions.csv

* The count of sales transactions for each date, store\_nbr combination. Only included for the training data timeframe.

### oil.csv

* Daily oil price. Includes values during both the train *and* test data timeframe. (Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices.)

### holidays\_events.csv

* Holidays and Events, with metadata
* NOTE: Pay special attention to the transferred column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday *Independencia de Guayaquil* was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.
* Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

## Additional Notes

* Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
* A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

# Approach

1. Analyze and understand if items have correlation to seasonality factor, economic or geographical factors
2. Use various visualization tools/techniques (scatter plots, pair plots etc to ) understand the distribution of data - how consumption of items are varying from one store to other , if there is a relation to the region/group of stores.
3. Generate various graphs and relevant code so that the research/analysis can be reproduced in a consistent manner.

# Milestones

1. Review and approval of proposal
2. Initial data analysis, data scrubs - code
3. Regression and visualization models