**Data Science**

**FAVORITA** **Project Protocol**

**Authors:**

**Ariel Shafir & Tzvika Senderov**

**Jan 2019**

**Introduction**

Grocery stores have always faced the challenge of forecasting procurement and sales of consumable products. High forecasting will result in excess inventory being thrown into the trash. If the prediction is low, it will result in a loss of income mainly in products that are sold a lot (Popular).

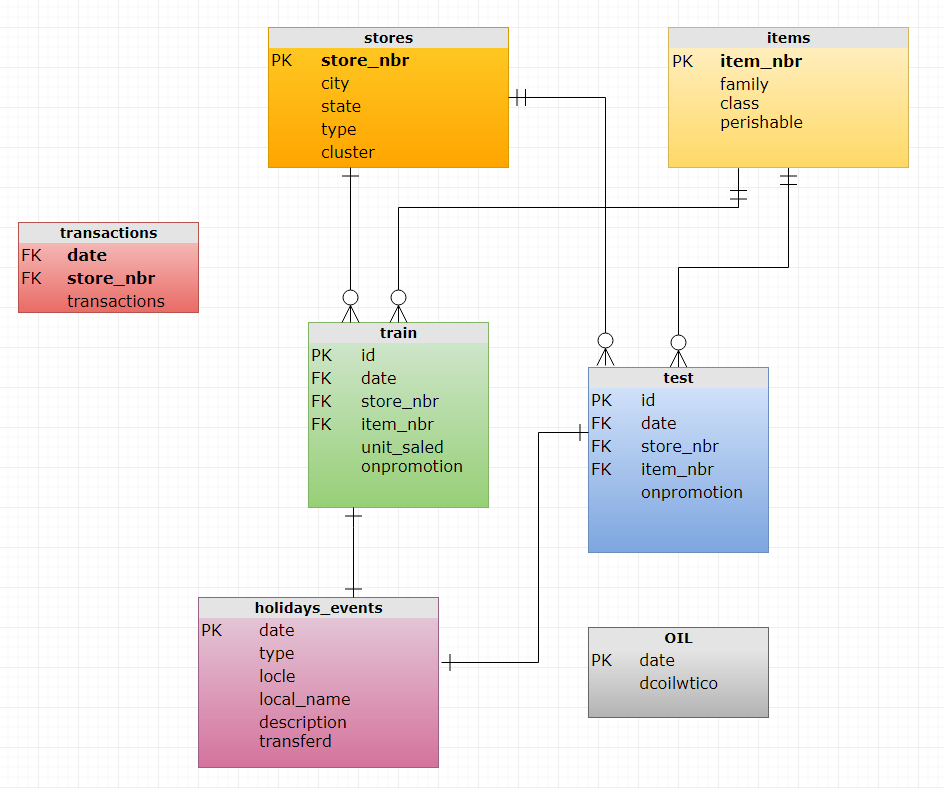
[Corporación Favorita](http://www.corporacionfavorita.com/) is a large Ecuadorian-based grocery retailer.

They operate hundreds of supermarkets, with over 200,000 different products on their shelves tackles this challenge on a daily basis.

The difficulty is greater when it comes to new products and seasonal products, which should be prepared accordingly according to the period of the year, which sometimes also takes into consideration unique storage conditions,

We want to see how using ML machine models will predict the quantity of products sold at a daily level in each store.

The data source is KAGGLE. We will use the DATASET defined by the corporation.



**Mythology (project design)**

**Data**

The data source was taken from Kaggle Competition – *"Corporación Favorita Grocery Sales Forecasting, Can you accurately predict sales for a large grocery chain?"*

This was the csv we used

### train.csv

         Training data, which includes the target unit\_sales by date, store\_nbr, and item\_nbr and a unique id to label rows onpromotion column tells whether that item\_nbr was on promotion for a specified date and store\_nbr   Negative values of unit\_sales represent returns of that particular item.



### stores.csv

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

### items.csv

         Item metadata, including family, class, and perishabl - : Items marked as perishable have a score weight of 1.25.

### 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 , Bridge  and Events

From Google we took information about the population <https://www.worldatlas.com/articles/biggest-cities-in-ecuador.html>

We build a function that calculate the payment day and payment day +1

The train dataset includes dates for one year 4/2015 until 4/2016 and all the information as in the train dataset

The dev dataset includes dates for one year 4/2015 until 4/2016 and all the information as in the train dataset

The test dataset includes dates for one year 4/2016 until 4/2017 and all the information as in the train dataset

The original dataset include more than 4000 items and more than 120,000,000 transaction. In our datasets we took the 250 most items sold and partitioned the data by years so we handled about 300,000 transaction per dataset.