**Data Science**

**Project Protocol**

**Authors:**

**Ariel Shafir**

**Tzvika Senderov**

**Introduction**

[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.

We tried in our project to **predict the unit sales** for thousands of items sold per store at different Favorita stores located in Ecuador.

We will try to answer the following questions

* Dose the payment day influence the buying behave? (I.e. when you get the salary do you buy more?) the payment day happens twice a month, at the 15th and at the end of month
* Dose the holydays day influence the buying behave? (I.e. buy more before the holyday and less right after?)
* Is there a difference types of items preferred by region?
* Dose the oil price influence the buying behave?( Ecuador economy based on oil sale)

The data includes store and item information about the number of items sailed by day, whether that item was being promoted, as well as the unit sales, supplementary information as local and national holydays, payment days, oil cost for each date and the population of the cities that we assume will be useful in building our models.

**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.

**Example :**

Each training sample consists of N data entries – the first N contain the following features: date and hour, season…

weather is a categorical variable with 4 levels representing different weather conditions, holiday is a binary indicator variable representing whether the particular day was a holiday, and working-day is a binary indicator variable signifying whether the day was a non-holiday weekday…

The training data had to be extracted and transformed from the database tables.  Categorical features were converted into equivalent 1/‐1 values in order to be compatible with the SVM algorithm requirements:

