**Goal of the Competition**

In this competition, you’ll use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.

Specifically, you'll build a model that more accurately predicts the unit sales for thousands of items sold at different Favourite stores. You'll practice your machine learning skills with an approachable training dataset of dates, store, and item information, promotions, and unit sales.

**Context**

Forecasts aren’t just for meteorologists. Governments forecast economic growth. Scientists attempt to predict the future population. And businesses forecast product demand—a common task of professional data scientists. Forecasts are especially relevant to brick-and-mortar grocery stores, which must dance delicately with how much inventory to buy. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leading to lost revenue and upset customers. More accurate forecasting, thanks to machine learning, could help ensure retailers please customers by having just enough of the right products at the right time.

Current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing.

**Potential Impact**

If successful, you'll have flexed some new skills in a real world example. For grocery stores, more accurate forecasting can decrease food waste related to overstocking and improve customer satisfaction. The results of this ongoing competition, over time, might even ensure your local store has exactly what you need the next time you shop.

**Evolution**

The evaluation metric for this competition is **Root Mean Squared Logarithmic Error**.

The **RMSLE** is calculated as:  
[\sqrt{ \frac{1}{n} \sum\_{i=1}^n \left(\log (1 + \hat{y}\_i) - \log (1 + y\_i)\right)^2}]  
where:

* � is the total number of instances,
* �^� is the predicted value of the target for instance (i),
* �� is the actual value of the target for instance (i), and,
* log is the natural logarithm.

**Submission File**

For each **id** in the test set, you must predict a value for the **sales** variable. The file should contain a header and have the following format:

id,sales

3000888,0.0

3000889,0.0

3000890,0.0

3000891,0.0

3000892,0.0

etc.

## Dataset Description

In this competition, you will predict sales for the thousands of product families sold at Favorita stores located in Ecuador. The training data includes dates, store and product information, whether that item was being promoted, as well as the sales numbers. Additional files include supplementary information that may be useful in building your models.

## File Descriptions and Data Field Information

### train.csv

* The training data, comprising time series of features **store\_nbr**, **family**, and **onpromotion** as well as the target **sales**.
* **store\_nbr** identifies the store at which the products are sold.
* **family** identifies the type of product sold.
* **sales** give the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).
* **onpromotion** gives the total number of items in a product family that were being promoted at a store at a given date.

### test.csv

* The test data, having the same features as the training data. You will predict the target **sales** for the dates in this file.
* The dates in the test data are for the 15 days after the last date in the training data.

### sample\_submission.csv

* A sample submission file in the correct format.

### stores.csv

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

### oil.csv

* Daily oil price. Includes values during both the train and test data timeframes. (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.