Time series analysis- store sales

## Introduction

Machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data. In most of the situations we want to have a machine learning system to make **predictions**, so we have several categories of machine learning tasks depending on the type of prediction needed: **Classification, Regression, Clustering, Generation**, etc.

**Classification** is the task whose goal is the prediction of the label of the class to which the input belongs (e.g., Classification of images in two classes: cats and dogs). **Regression** is the task whose goal is the prediction of numerical value(s) related to the input (e.g., House rent prediction, Estimated time of arrival ). **Generation** is the task whose goal is the creation of something new related to the input (e.g., Text translation, Audio beat generation, Image denoising ). **Clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other **clusters** (e.g., Clients clustering).

In machine learning, there are learning paradigms that relate to one aspect of the dataset: **the presence of the label to be predicted**. **Supervised Learning** is the paradigm of learning that is applied when the dataset has the label variables to be predicted, known as y variables. **Unsupervised Learning** is the paradigm of learning that is applied when the dataset does not have the label variables to be predicted. **Self-supervised Learning** is the paradigm of learning that is applied when part of the X dataset is considered as the label to be predicted (e.g., the Dataset is made of texts and the model tries to predict the next word of each sentence).

## Aim:

Our aim was to analyze the provided data sets, answer some questions about the data and train machine learning models that can predict retail store sales based on data given to them.

## Method:

This work was done based on the CRISP-DM algorithm and this article shall be focused mainly on explaining these steps which involves the following;

* Business understanding.
* Data understanding.
* Data preparation.
* Modeling.
* Evaluation.
* Deployment.

## Business Understanding

Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices. We were also made aware that a magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied on relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake. Wages in the public sector are paid every two weeks on the 15th and on the last day of the month.

Being aware of all these information, we were set to answer the following questions;

1. Is the train dataset complete (has all the required dates)?
2. Which dates have the lowest and highest sales for each year?
3. Did the earthquake impact sales?
4. Are certain groups of stores selling more products? (Cluster, city, state, type)
5. Are sales affected by promotions, oil prices and holidays?

## Data Understanding

The data sets were provided with the following descriptions;

### **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** gives 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. We predicted 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.

### **transaction.csv**

* Contains date, store\_nbr and transaction made on that specific date.

### **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** which includes values during both the train and test data timeframes.

### **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 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 pay back the Bridge.

* Additional holidays are days added to a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

## Data Preparation

We imported all the libraries and packages we needed then read the data into pandas using the .read\_csv()

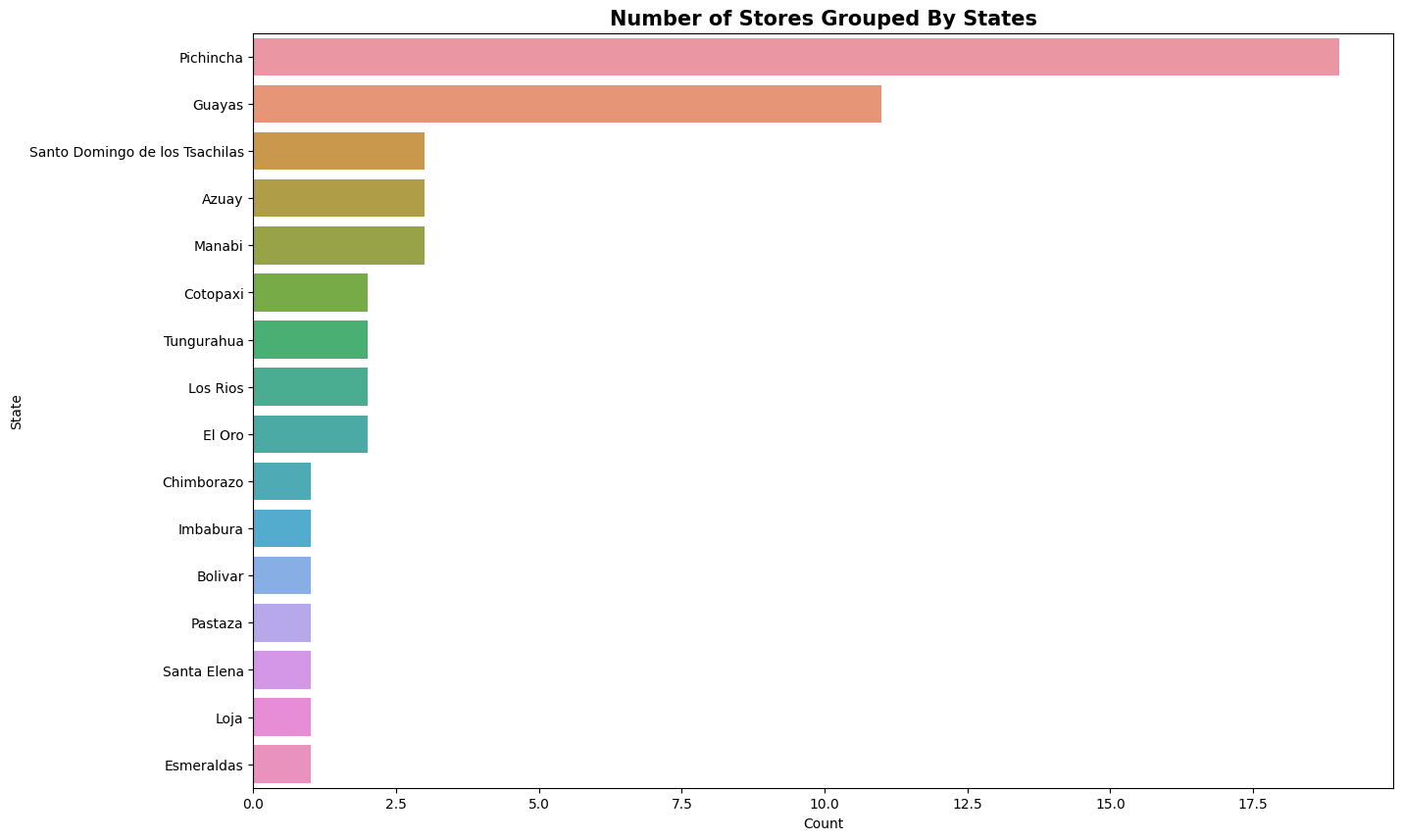
We used the train data set for the training of the model but a look at the data set showed that it did not have all the information needed to answer the questions we were to answer and it did not contain all the features to give the models all the necessary information to make good prediction so me merged it with other data sets using the pandas .merge() method. We used different types of merge as the relationship between the datasets was not the same.

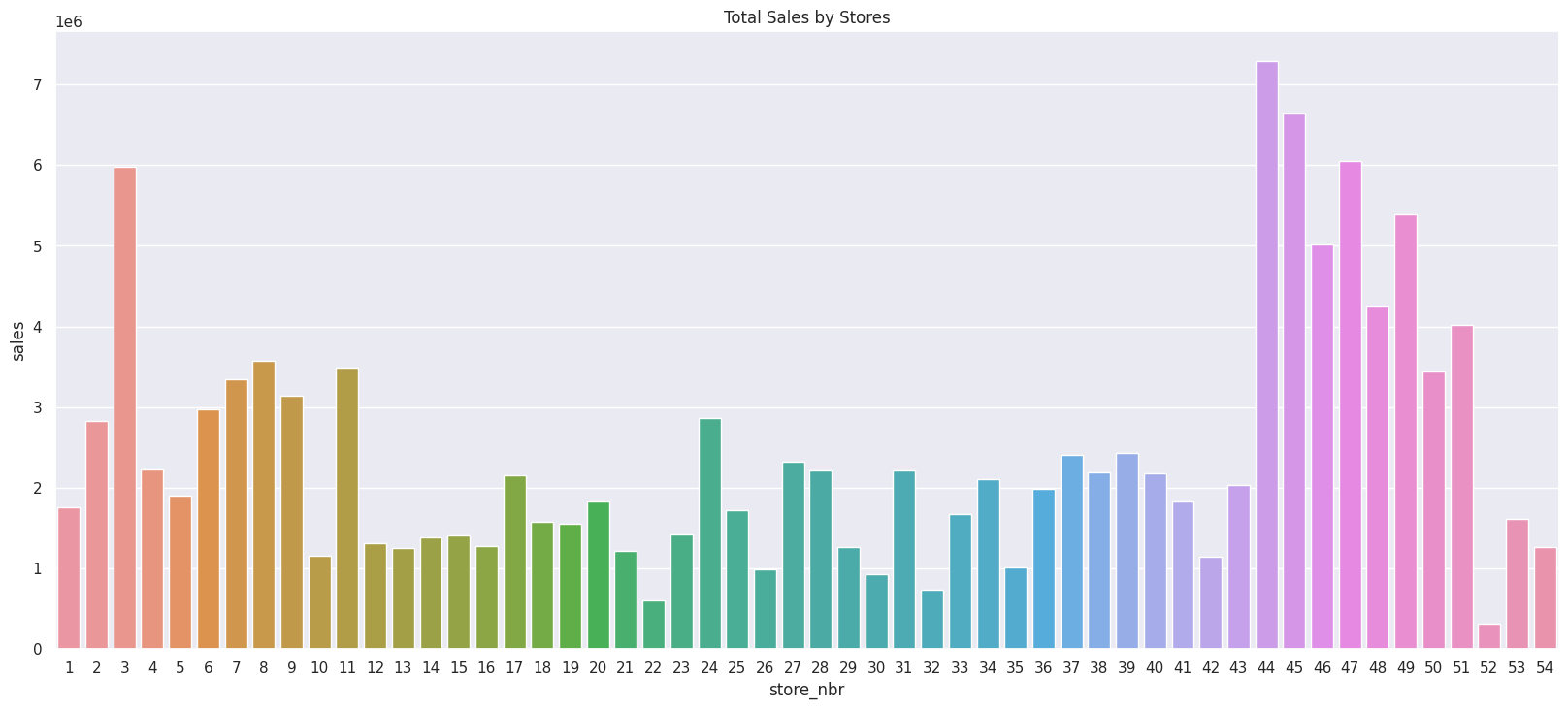
* The train data set itself had no null values and no duplicates
* We dropped the id column
* We converted the date from object data type to datetime data type
* We merged it with the store dataset on store number
* The store data type had no null values and no duplicates
* We renamed the type column in the store dataset to store type
* The holidays data set did not contain any null values
* We renamed the type column in the holidays dataset to holiday type.
* We converted the date from object data type to date time
* We merged the holidays dataset with the train dataset on the date column
* The oil dataset had some null values, we filled them with the mode.
* We merged the oil that set with the train dataset on the date column
* The transactions dataset had no null values,
* We merged it with the train dataset on date and store-number column
* We extracted date features(months, weeks and days)

### Analysis

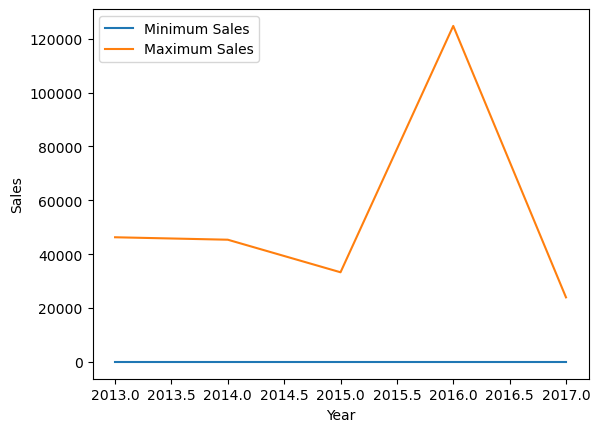
From the analysis we did, we saw that;

1. The business comprises 54 stores distributed in Ecuador with Pichincha having 19 stores, the highest number.

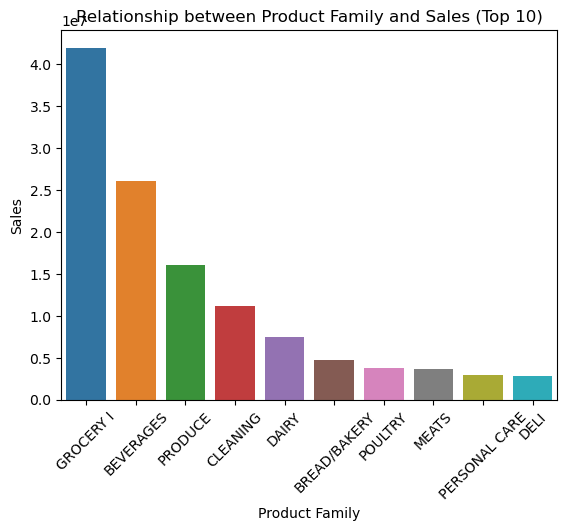


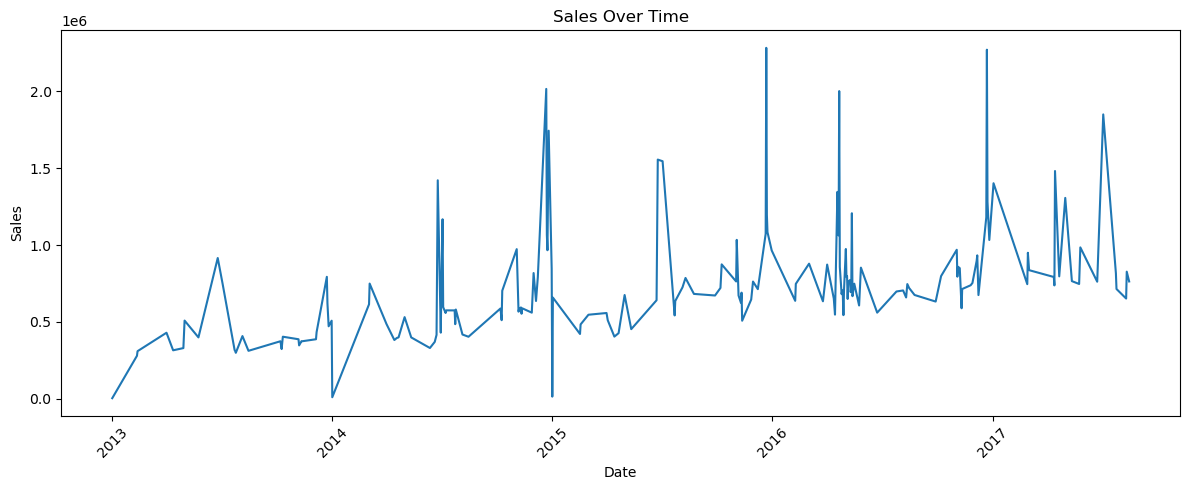


From the above chart we also saw that store number 44 made the highest sales between 2013 to 2017 while store number 52 made the list sales.



The above chart shows sales distribution over time





Trend of sales over time.