**TIME SERIES FORECASTING**

**INTRODUCTION**

In this project, I will be dealing with Time-Series forecasting and I will build a model to predict the unit sales for thousands of items sold at different Favorita stores and provide answers to some important questions.

Time Series Forecasting is a technique that utilizes historical and current data to predict future values over a period or a specific point in the future.

It can be used in different fields for time-based predictions such as Weather Forecasting Models, Stock Market Predictions, Signal Processing, Engineering Domain – Control Systems and Communication Systems.

For this project,

**DATA**

The dataset used for this project was derived from Corporation Favorita, a large Ecuadorian-based grocery retailer and in this dataset, I worked with seven (7) datasets namely, oil.csv, test.csv, train.csv, stores.csv, holidays.csv, transactions and sample\_submission.csv.

**METHODOLOGY**

First, I started an Exploration Data Analysis (EDA) on each the given dataset to check out for relevant information on the dataset such the shape, distribution, datatypes, and missing values of the dataset.

After doing that, I realized the datasets weren’t too dirty, so I didn’t need to do much cleaning. I then merged the datasets using an outer join on the common columns resulting in total of 17 columns and 3054427 rows. Some columns had the name type x and type y which was renamed to holiday type and store type respectively.

In the Exploratory Data Analysis (EDA) stage, I was able to provide answers to the following questions:

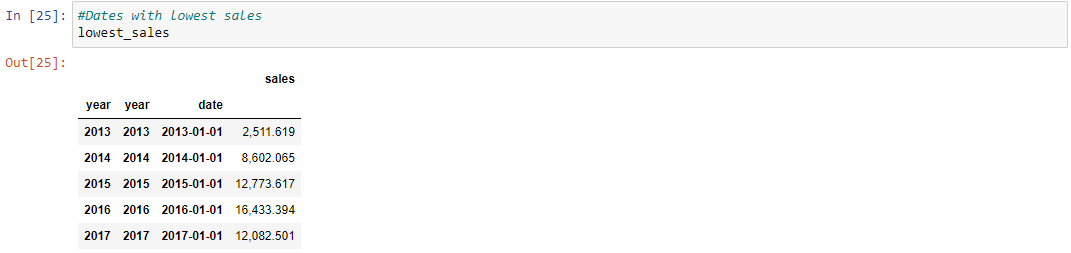
1. Is the train dataset complete (has all the required dates)?

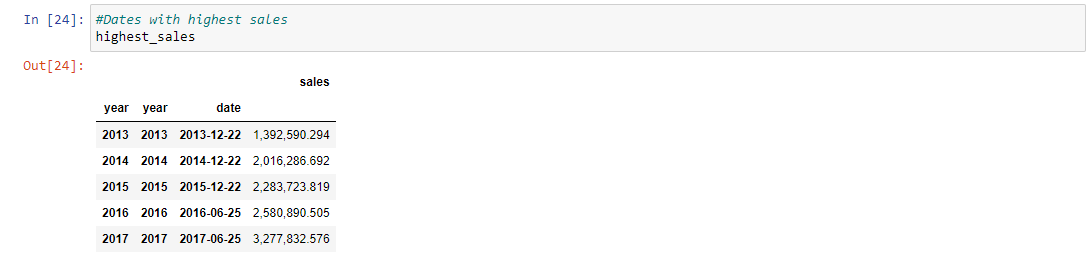
Using the isnull () function on the dataset, it shows that the dataset was complete. However, working with Time series, I decided to use the line of code below.

From the results, the missing dates are Christmas days for the year 2013, 2014, 2015, 2016 respectively.

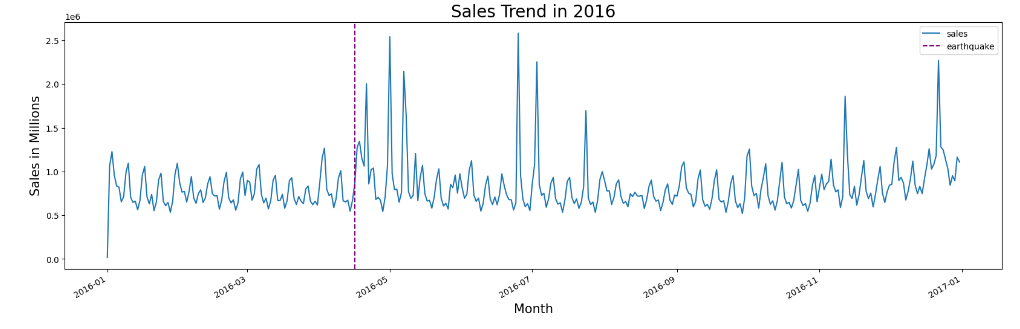
1. Which dates have the lowest and highest sales for each year?

For this question, using the grouped\_by function, I sorted for the lowest and highest sales for each year.

Below are the results. 



1. Did the earthquake impact sales?

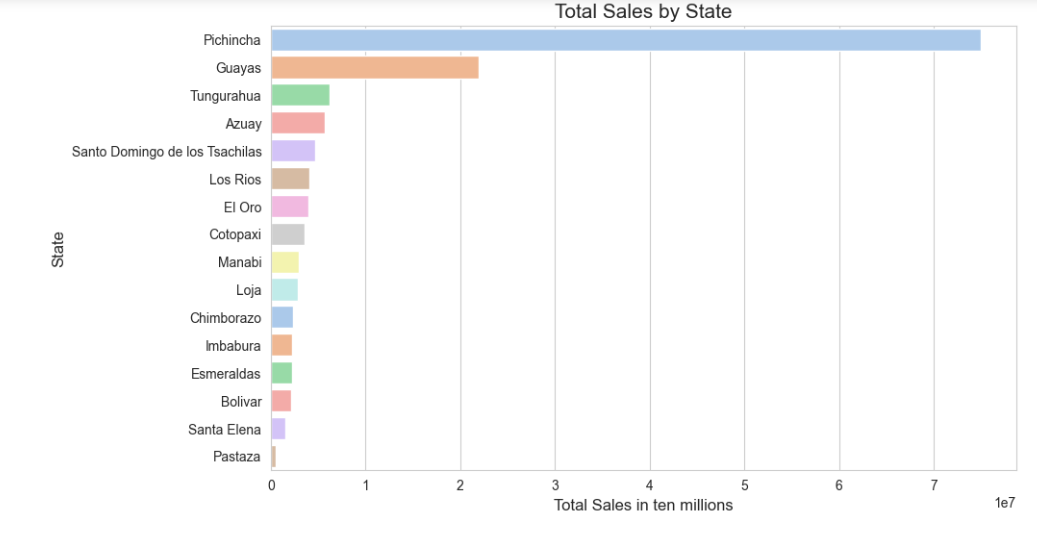


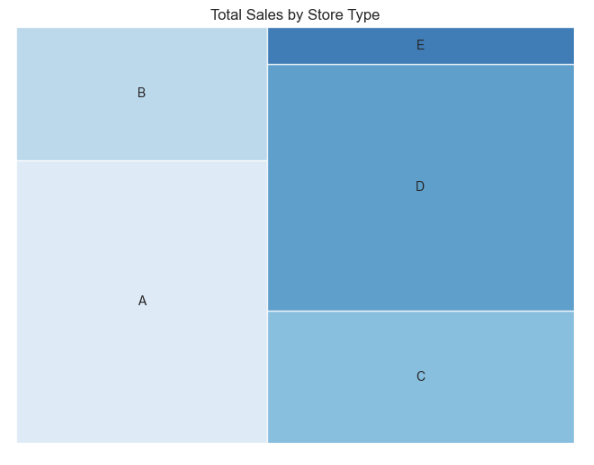
The purple dashed line has been used to mark the day of the earthquake however there was a rise in sales which lasted for several weeks.

1. Are certain groups of stores selling more products? (Cluster, city, state, type)

Chart, scatter chart

Description automatically generated





Also, for this project, I performed feature engineering, encode categorical variables with the LabelEncoder, drop unnecessary columns, scale columns with the StandardScaler and split the merged data into training and test dataset.

To predict sales, I used four different regression models, namely:

* Linear Regression
* Decision Tree Regression
* Random Forest Regression and
* XGBoost

Below are the evaluation results for each of the models used.

Text

Description automatically generated

A comparison table of evaluation results shows that the Random Forest model performed the best in terms of MSE, RMSE, and RMSLE. However, I had to reduce the number of estimators to 20 due to memory constraints. Therefore, in practical applications, we may need to choose a model that strikes a balance between accuracy and computational resources.

For mor details on this project, visit my

Special thanks to my teams members,

David Anifowoshie, Gideon Dadzie, Williams George and Efobi Godwin for their immense support and contributions.