REGRESSION PROJECT (STORE SALES -- TIME SERIES FORECASTING

AMSTERDAM

Oh-so-famous central canal, rightly dubbed a UNESCO World Heritage Site in 2010. Add to that

swathe of green spaces, storied red-brick facades, and museums filled with Van Gogh paintings, and you have yourself one of Europe's most gorgeous culture epicenters

Project Title: Predictive Analytics for Grocery Sales Forecasting: A Case Study of Favorita Stores.

Introduction

This is a time series forecasting problem. In this project, you will predict store sales on data from Corporation Favorita, a large Ecuadorian-based grocery retailer.

Specifically, we will build a machine learning model that accurately predicts the unit sales for thousands of items sold at different Favorita stores.

The Favorita Grocery Sales Forecasting dataset is a fascinating collection of data that provides a great opportunity for analysis and prediction. In this article, we will take a deep dive into the dataset, explore its various attributes, and analyze the sales patterns to build a robust sales forecasting model and answer some pertinent question on the dataset.

Project Description

The objective of this project is to analyze the sales data of a store and build a regression model to predict future sales. The data for this project is obtained from a retail store that sells various products, such as food, clothing, electronics, and home appliances.

Overview of the Dataset

The dataset contains information about the sales of various products in different stores belonging to the Favorita chain of grocery stores in Ecuador. The original train dataset has a total of 3,000,888 rows and 5 columns. The columns in the train dataset include the date, store number, sales, family, and on promotion. The data in the dataset ranges from January 2013 to August

2017. Additional files include supplementary information that may be useful in building the models.

File Descriptions and Data Field Information

Here, we briefly describe the dataset. The supplied dataset has 7 csv files. It is explained as follows

1. train.csv

The training data, comprising time series of features store_nbr, family, and onpromotion as well as the target sales.

a. **store_nbr:** identifies the store at which the products are sold.
b. **Family:** identifies the type of product sold.
c. **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.
d. **Onpromotion:** gives the total number of items in a product family that were being promoted at a store at a given date.

2. 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.

3. transaction.csv

Contains date, store_nbr and transaction made on that specific date.

4. sample_submission.csv

A sample submission file in the correct format.

5. stores.csv

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

6. oil.csv

Daily oil price: which includes values during both the train and test data timeframes. (Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices.)

7. holidays_events.csv

Holidays and Events, with metadata

Business Questions

1. Which dates have the lowest and highest sales for each year?
2. Did the earthquake impact sales?
3. Are certain groups of stores selling more products? (Cluster, city, state, type)
4. Are sales affected by promotions, oil prices and holidays?
5. What analysis can we get from the date and its extractable features?
6. What is the difference between RMSLE, RMSE, MSE (or why is the MAE greater than all of them)?

Hypothesis

NULL: Earthquake has significant impact on sales

ALTERNATIVE: Earthquake has no significant impact on sales

Data Exploration and data cleaning

Data cleaning

The Favorita grocery sales forecasting dataset consists of multiple files containing information about store locations, product categories and transactions. Before performing any analysis or modeling, it is essential to clean and prepare the data to ensure its accuracy and completeness.

The following are the steps involved in the data cleaning process, including merging other files into the train dataset:

Handling missing values

The first step is to identify and handle missing values in the dataset. Here we dropped some rows and columns with missing values and imputed appropriate values depending on the level of that column or row significance to our target variable.

Data type conversion

We converted our date column to datetime format for our analysis. The column sales was also converted to its appropriate data type better handling and analysis.

Removing duplicates

We Identified and removed any duplicates in the dataset to ensure that the data is not overrepresented.

Merging additional datasets

Merged other datasets such as transactions, sales, onpromotion, oil and store dataset into the main train dataset to enhance the analysis. For example, merging the store data and transaction data into the train dataset.

Dealing with outliers

Here we Identified and removed any outliers in the data to avoid them from skewing the analysis results.

Feature engineering

Create new variables or features that can be useful in the analysis. For example, extracting the month, day, and year from the date variable for better analysis. Extracting these features helped us to plot the trend analysis of our transactions and sales fields.

Scaling and normalization

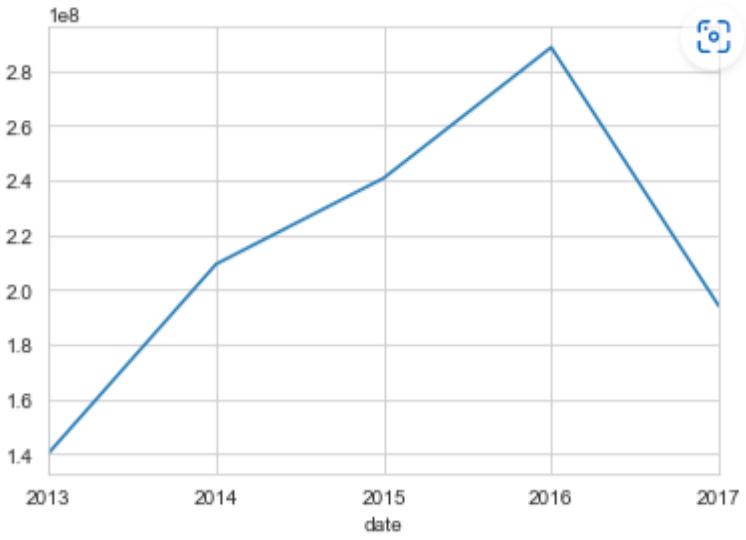
Scale and normalize the variables in the dataset to ensure that they are on the same scale and are comparable.

By performing these steps, the Favorita grocery sales forecasting dataset was made cleaned and ready for analysis and modeling.

Data exploration (EDA)

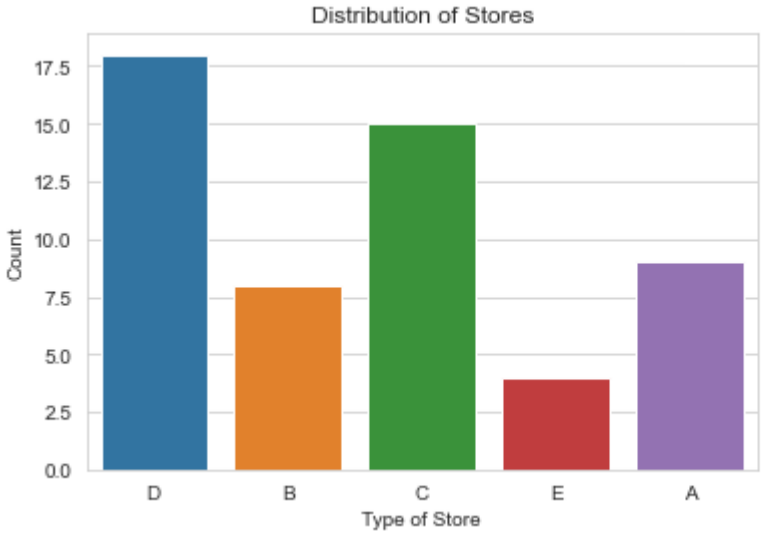
Before we start building a sales forecasting model, we need to explore the data and analyze the patterns in the sales data. We start by importing the dataset and analyzing its different attributes. We can then use various visualization techniques to understand the trends in the data.

One interesting visualization is to plot the sales data against time. This gives us a clear idea of how sales have changed over the years. We can see that sales have been increasing gradually over time, with a few sharp peaks and drops.

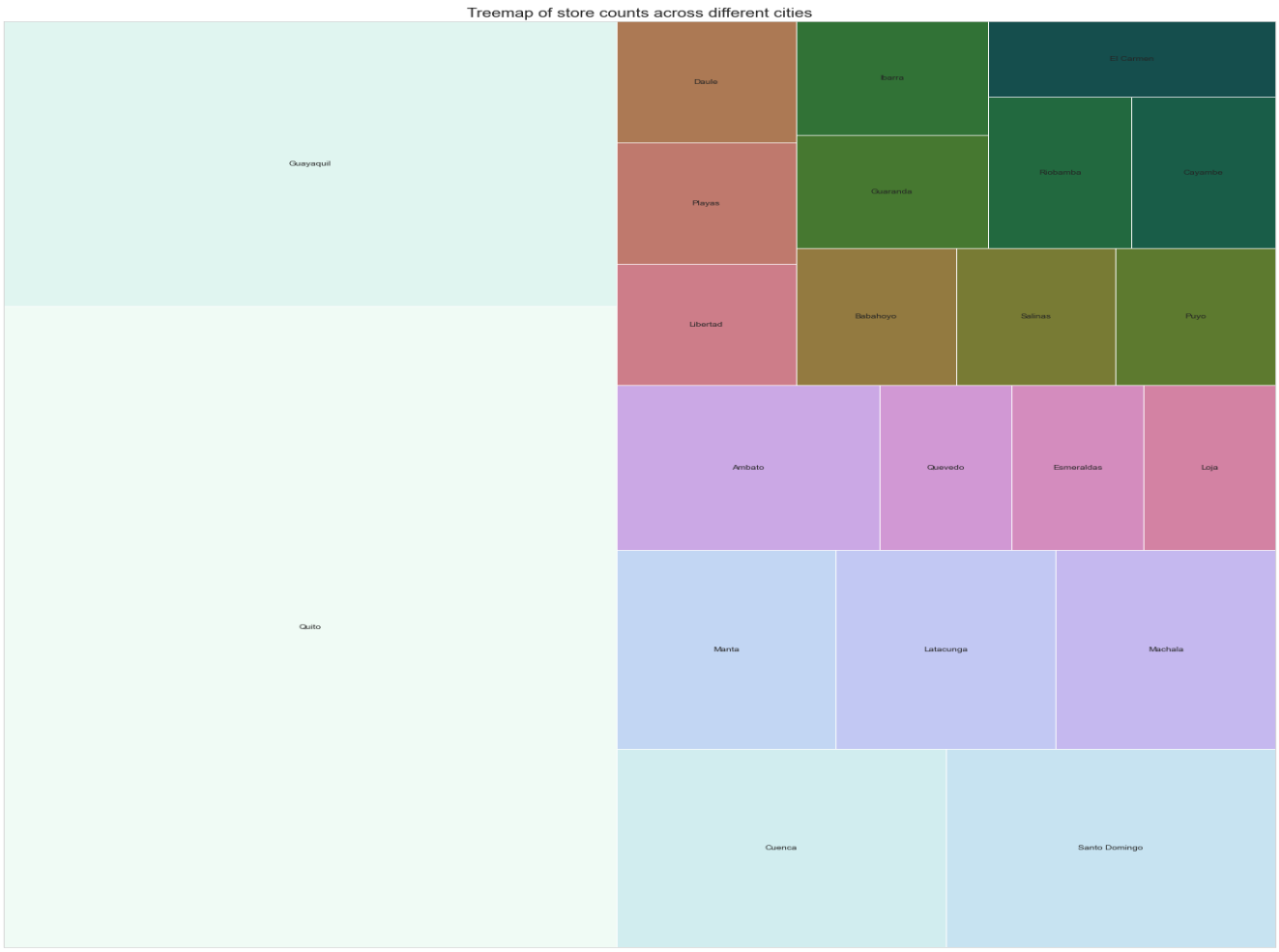


Sales over time

From the chart above we can say that, sales have increased over time from 2013 to 2016. However, there was a sharp decline in sales from 2016 to 2017. This can be as a result of earthquake that occurred on 16th April, 2016.



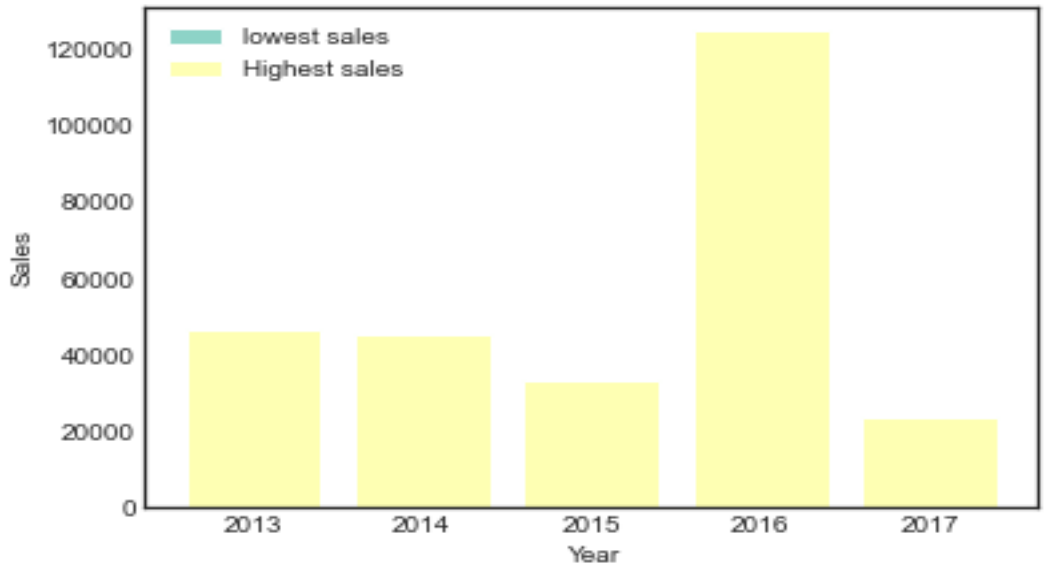
Here, we plotted distribution of each store type and the number of stores in that category. it was observed that store type D had the majority share, followed by type C with A, B and E in that order.



Guayaquil and Quito are two cities that stand out in terms of the range of retail kinds available. These are unsurprising given that Quito is Ecuador's capital and Guayaquil is the country's largest and most populated metropolis. As a result, one might expect Corporacion Favorita to target these major cities with the most diverse store types, as evidenced by the highest counts of store nbrs attributed to those two cities.

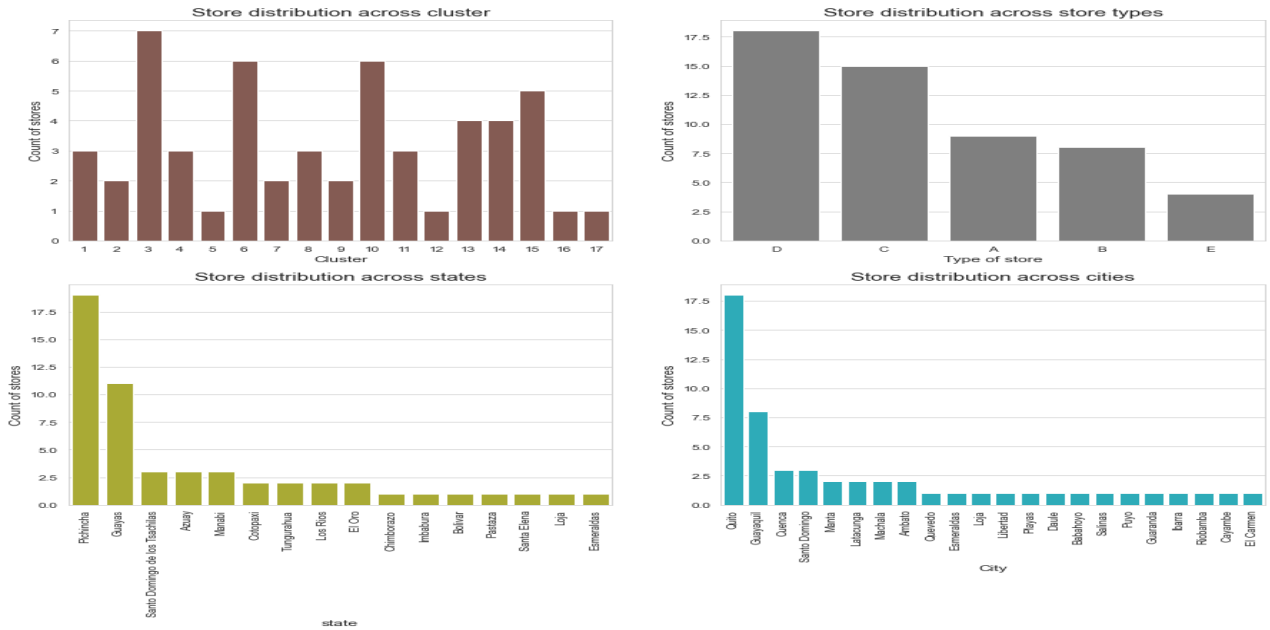
Answering the questions and testing hypothesis

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

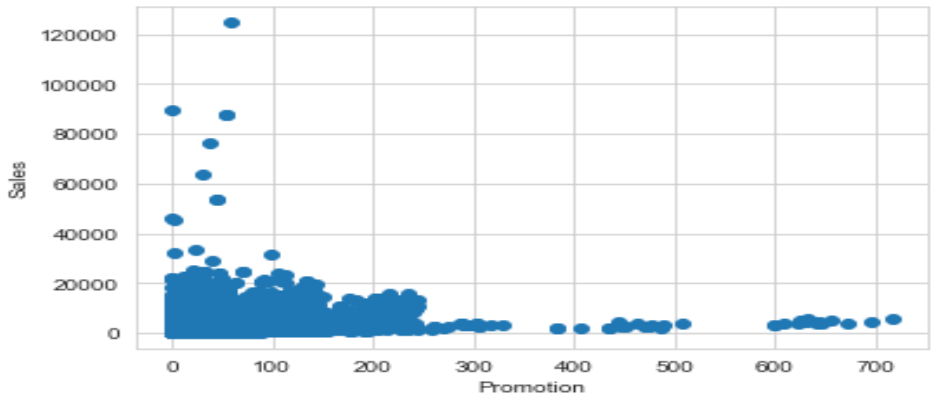


- It can be concluded here that each year recorded 0 as their lowest sales.
- It can also be said that 2016 had the highest sales, followed by 2013 it is also followed closely by 2014 with almost the same sales
- It can also be noted that 2015 and 2017 recorded the lowest highest score among the years

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

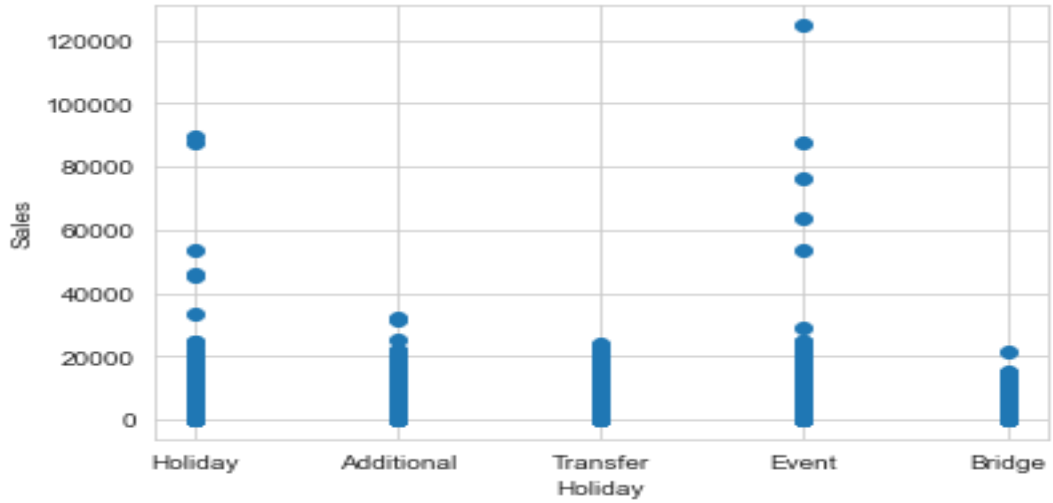


3. Are sales affected by promotions, oil prices and holidays?



Yes, sales can be affected by promotion. When a product is promoted, it can increase consumer awareness and interest in the product, which can lead to an increase in sales. The effectiveness of the promotion will depend on a variety of factors, such as the type of promotion, the target audience and the product itself.

We can however conclude that; promotion did not have any significant impact on sales as seen in the above scatter plot. The reason could be that the promotion did not reach the targeted audience or the product is not good enough to catch people eye.

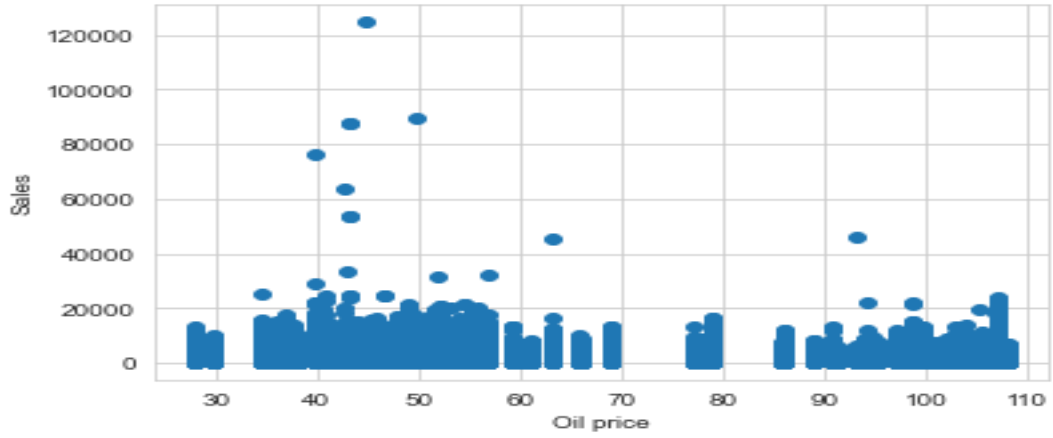


Yes, sales can be affected by Holidays. Holidays can lead to increased consumer spending due to factors such as gift-giving, travel and seasonal purchases. Additionally, holidays can create a sense of urgency for consumers to make purchases, as they may be seeking to take advantage of sales or to meet gift-giving deadlines The impact of holidays on sales will depend on various factors, such as the type of holiday, the products being sold, and the target audience. For example, a holiday like Christmas is typically associated with gift-giving, which may increase sales for retailers selling items such as toys, electronics, and clothing. Similarly, holidays like

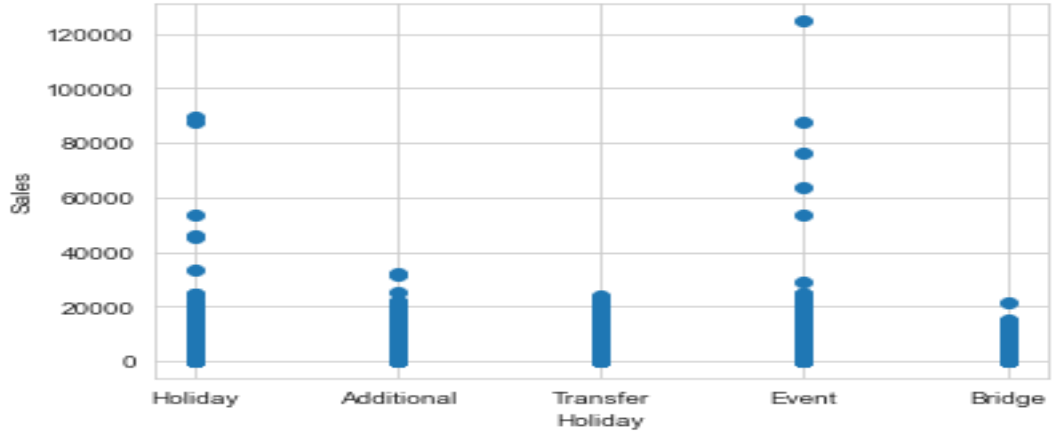
Memorial Day and Labor day may increase sales for retailers selling items such as outdoor gear and home goods

We can conclude by saying that,

1. Holiday has significant impact on sales, from the above scatter plot we can on holidays sales increased, however during Events sales increased more as compare to Holidays this can be attributed to a lot factors a. it could be because it's a holiday people really want to rest rather than to go shopping b. during events people actually have to prepare some foods or other stuffs to suite the occasions.



It can be observed from the scatter plot above that as prices of oil goes up, sales also reduce. So, we can say that oil prices have negative correlation with sales or inversely proportional to sales



Event holiday had

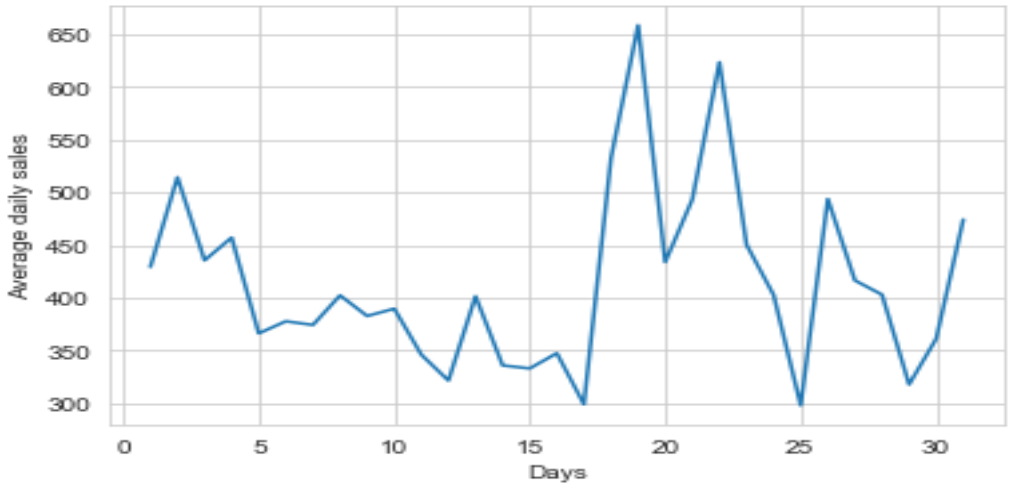
4. What analysis can we get from the date and its extractable features?

Date and its extractive features can provide valuable insights in data analysis. Here are some examples of analysis that can performed using date and its extractive features

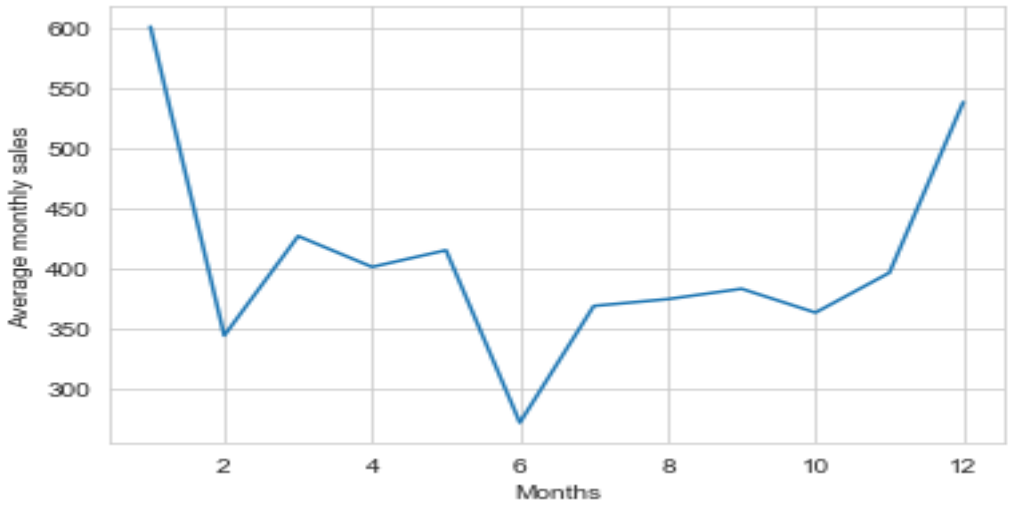
1. Trend analysis: By extracting the year, month or day from a date, you can analyze trend over time. For Example, you can plot the number of sales per month over several years to see if there are any patterns or trends.

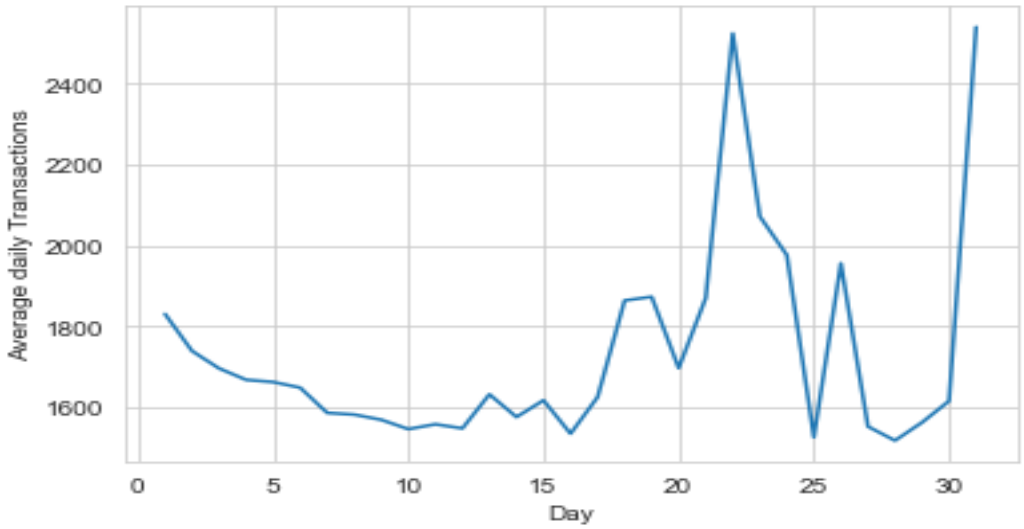
2. Seasonal analysis: Extracting the month or quarter from the date can help you understand seasonal trends in your data. For example, you might find that sales of certain it m increases in a particular period while sales of different item may also increase in a different period

Overall, date and its extractive features are essential for understanding trends, seasonality and patterns in data. By extracting and analyzing these features, you can gain valuable insights that can help you make data-driven decisions.

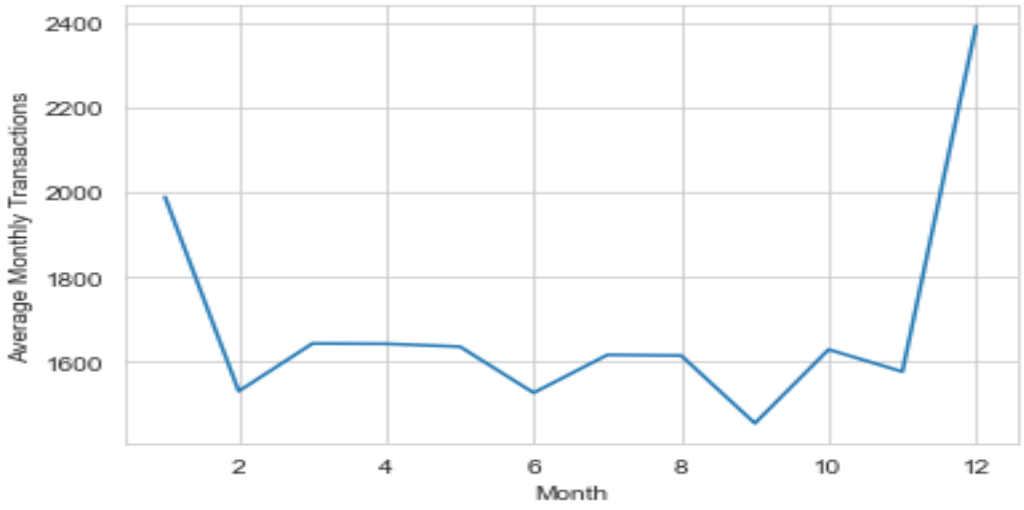


From the commentary from the dataset, we were made aware that salaries are paid twice very each. First payment is on the 15th of the month and second is paid on a day before the month ends. From the graph above, we can say sales increased after 15th which is a true reflection of the reality. As people received salaries it is normal to spend. Sales dropped on 20th and saw a raise on the 30th.

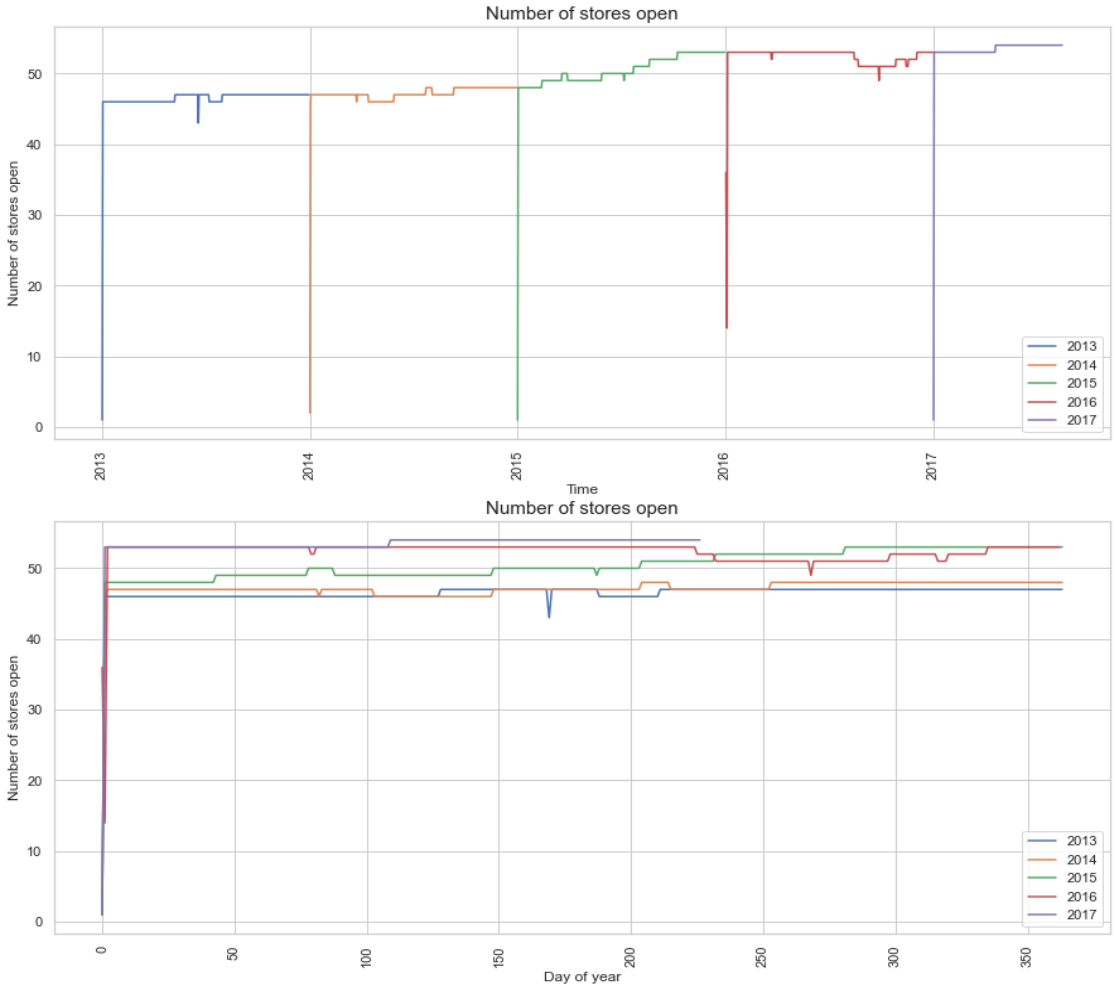




We can have observed a decline from beginning of day 1 to 15 then right from 15th we see a sharp increase. As salaries are paid on 15th its normal to see increase in transactions from 15th. From 22nd from 25th saw a sharp decline in transactions. Transaction also increased sharply from 30th to 31st.



There is sharp decline in transactions right from Jan to Feb. On the contrary transactions increased sharply from Nov toward Dec. This can be attributed to so many factors. Some of these factors maybe because is Christmas and they need to buy so may items



5. What is the difference between RMSLE, RMSE, MSE (or why is the MAE greater than all of them?)

RMSLE, MSE, RMSE and MAE are all error metrics used to evaluate the performance of regression models. MSE (Mean Squared Error) measures the average squared difference between the predicted and actual values. It is calculated as the average of the squared differences between the predicted and actual values. MSE gives ore weight to larger errors and is sensitive to outliers.

1. RMSE (Root Mean Squared Error) is the square root of the MSE. It measures the average distance between the predicted and actual values, but since it is the square root of the MSE, it has the same units as the dependent variable. RMSE is sensitive to outliers and gives more weight to larger errors.
2. RMSLE (Root Mean Squared Logarithmic Error) is similar to RMSE, but takes the logarithm of the predicted and actual values before calculating the error. RMSLE is used when the dependent variable has a wide range of values and is skewed towards large values. It is less sensitive to outliers than RMSE.

3. MAE (Mean Absolute Error) measures the average absolute difference between the predicted and actual values. It is calculated as the average of the absolute differences between the predicted and actual values. MAE treats all errors equally, is less sensitive to outliers and is not affected by the scale of the dependent variable.
4. The MAE is greater than all of them because it does not give more weight to larger errors, unlike MSE and RMSE. Therefore, it may be more appropriate to use MAE when the model should be evaluated on the magnitude of the errors rather than their squared values.

Hypothesis testing using EDA method

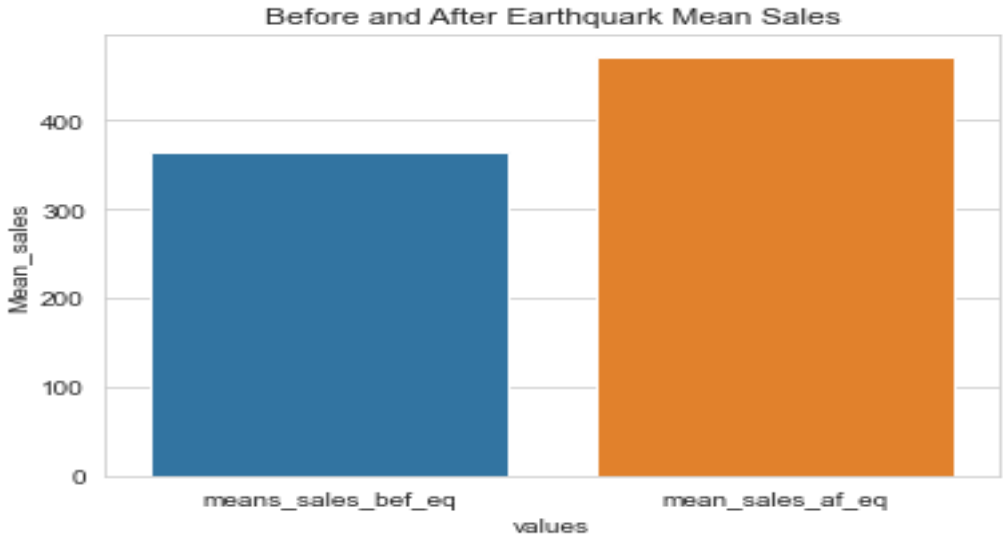
NULL: Earth quark has significant impact on sales

ALTERNATIVE: Earth quark has no significant impact on sales

As we are made aware that earth quark of magnitude 7.8 stroked Ecuador on the 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.

So we have assumed that this will have significant impact on sales as people won't go shopping as usual. It worth noting that this same situation can also lead to great increase as people will need different items to replace the damaged ones. To conclude on this, we need to test this using our dataset

The basic idea is to compare the sales before and after the earth quark and see if there is a significant difference between the two periods



Building the Sales Forecasting Model

The problem that was tackled is a regression problem. Three models were used to solve the problem: linear regression, decision tree, and random forest. The evaluation metrics used for comparing these models are the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Root Mean Squared Logarithmic Error (RMSLE).

The linear regression model produced an MSE of 0.76, an RMSE of 0.87, and an RMSLE of 0.29. The decision tree model produced an MSE of 0.16, an RMSE of 0.40, and an RMSLE of 0.10. Finally, the random forest model produced an MSE of 0.08, an RMSE of 0.28, and an RMSLE of 0.08.

A lower value of MSE, RMSE, and RMSLE indicates better performance of the model. Based on the evaluation metrics, the random forest model performed the best with an MSE of 0.08, an RMSE of 0.28, and an RMSLE of 0.08. Therefore, the random forest model is a good choice to report as it provided the best performance among the three models.

Conclusion

In this article, we explored the Favorita Grocery Sales Forecasting dataset. We analyzed the various attributes of the dataset and visualized the sales patterns over time. We then used various time-series forecasting techniques to build a robust sales forecasting model. With this model, we can predict the future sales of different products in different stores and make informed decisions about inventory management, pricing, and promotions. The dataset provides a great opportunity for further analysis and prediction, and we encourage you to explore it further.