Jiteshwar Anjale
DSC680, Weeks 5-8
21-October-2023

# *Project 2: Milestone 3*

## Topic:

Effective use of data science methodologies in supermarkets industry so that it can efficiently predict Relation of customers with Supermarket, Payment methods used in supermarket, quantity of goods sold with more accuracy.

## Business Problem:

A supermarket is self-service shop offering a wide variety of food, beverages and household products, organized into sections. It is larger and has a wider selection than earlier grocery stores, but is smaller and more limited in the range of merchandise than a hypermarket or big-box market.

The supermarkets deals with multiple customers and it is very important to understand customers behaviors so that the supermarkets can stock the most sold items or goods prior to its unavailability. Unfortunately, there is no reliable way to predict the customer's behavior. This is a challenge to the supermarket and its partners.

## Background/History

Supermarkets are retail stores that sell a variety of food and household items to consumers. Supermarkets have become an integral part of modern-day life, offering convenience and a one-stop-shop for all kinds of products. With increasing competition in the supermarket industry, it has become imperative for supermarkets to understand consumer behavior and purchasing patterns to remain competitive.

The purpose of data analytics in supermarkets is to analyze consumer data and predict future sales trends, which can help supermarkets optimize their operations and make data-driven decisions. Data analytics can help supermarkets identify which products are selling well and which products need improvement, track inventory levels, and understand consumer behavior and preferences.

Currently, supermarket run-centers, Favorita keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Favorita stores located in Ecuador, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Linear Regression, ARIMA, Random Forest, XGBoost and LSTM (Long Short-Term Memory) techniques for forecasting the sales of a business such as Big-Mart, and it was discovered that the model outperforms existing models.

# Data Explanation

For this phase of the project, we will use traditional stock parameters. This dataset will have.

## Datasets

To establish an initial proof of concept for supermarket sales data, I will use the historical data for items sold in a supermarket Favorita stores located in Ecuador. I have collected around 5 different data sets from Kaggle. 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 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. 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.

### 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.

## Data Preparation

Store, holiday, oil and train Datasets will be augmented with a column for label. These datasets are merged based on the 'store_nbr'.

```
train_df = train_df.merge(stores_df, on ='store_nbr')
train_df = train_df.merge(oil_df, on ='date', how='left')
holiday_event_df = holiday_event_df.rename(columns={'type': 'holiday_type'})
train_df = train_df.merge(holiday_event_df, on='date', how='left')
```

```
train_df.head(3)
```

| | id | date | store_nbr | family | sales | onpromotion | city | state | type | cluster | dcoilwtico | holiday_type | locale | locale_name | description | tra |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 2013-01-01 | 1 | AUTOMOTIVE | 0.0 | 0 | Quito | Pichincha | D | 13 | NaN | Holiday | National | Ecuador | Primer dia del ano | |
| **1** | 1 | 2013-01-01 | 1 | BABY CARE | 0.0 | 0 | Quito | Pichincha | D | 13 | NaN | Holiday | National | Ecuador | Primer dia del ano | |
| **2** | 2 | 2013-01-01 | 1 | BEAUTY | 0.0 | 0 | Quito | Pichincha | D | 13 | NaN | Holiday | National | Ecuador | Primer dia del ano | |

## Data Analysis

### Does the type of stores affect the store sales?

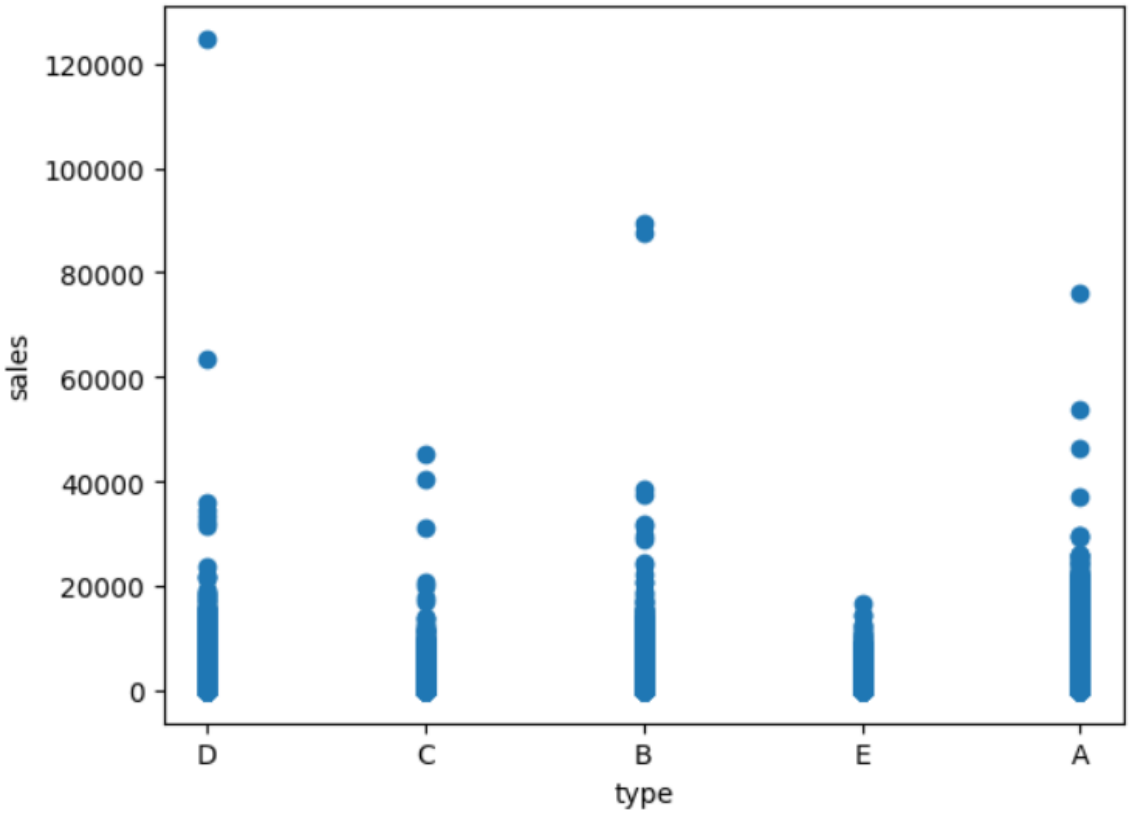
To answer the first question 'Does the type of stores affect the store sales?' , i will use ANOVA test. ANOVA (Analysis of Variance) is a statistical test used to determine whether there are significant differences between the means of two or more groups. It compares the variation between the groups (due to the different categories or factors) to the variation within the groups.

H0 (>0.05)= The type of stores does not affect store sales. There is no significant difference in store sales between different types of stores.

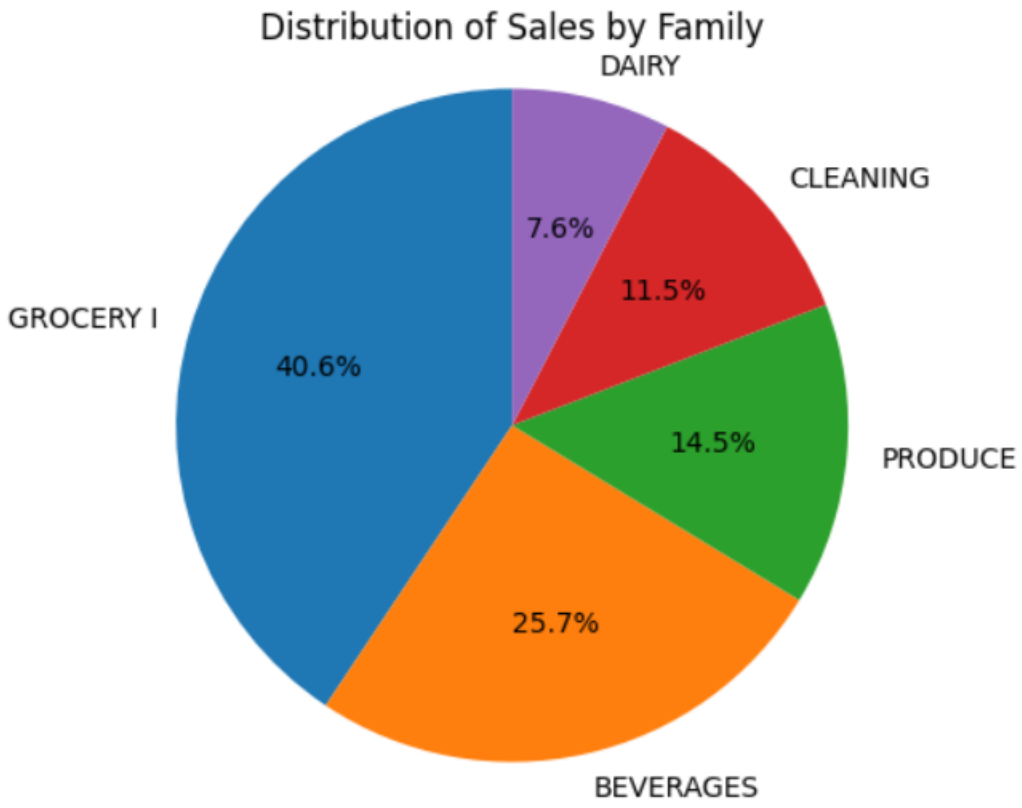
H1 (<0.05)= The type of stores does affect store sales. There is a significant difference in store sales between different types of stores.

```
F-Statistic: 17766.023730362205
p-value: 0.0
```

Based on the F-statistics and p-value above, we reject null hypothesis and accept alternative hypothesis. Hence, the type of stores does affect the store sales. There is a significant difference in store sales between different type.



## Which product family is having the highest sales?




Based on the pie chart above, the GROCERY I is having the highest sales, and Baverages comes second highest.

## Does promotion able to improve the sales?

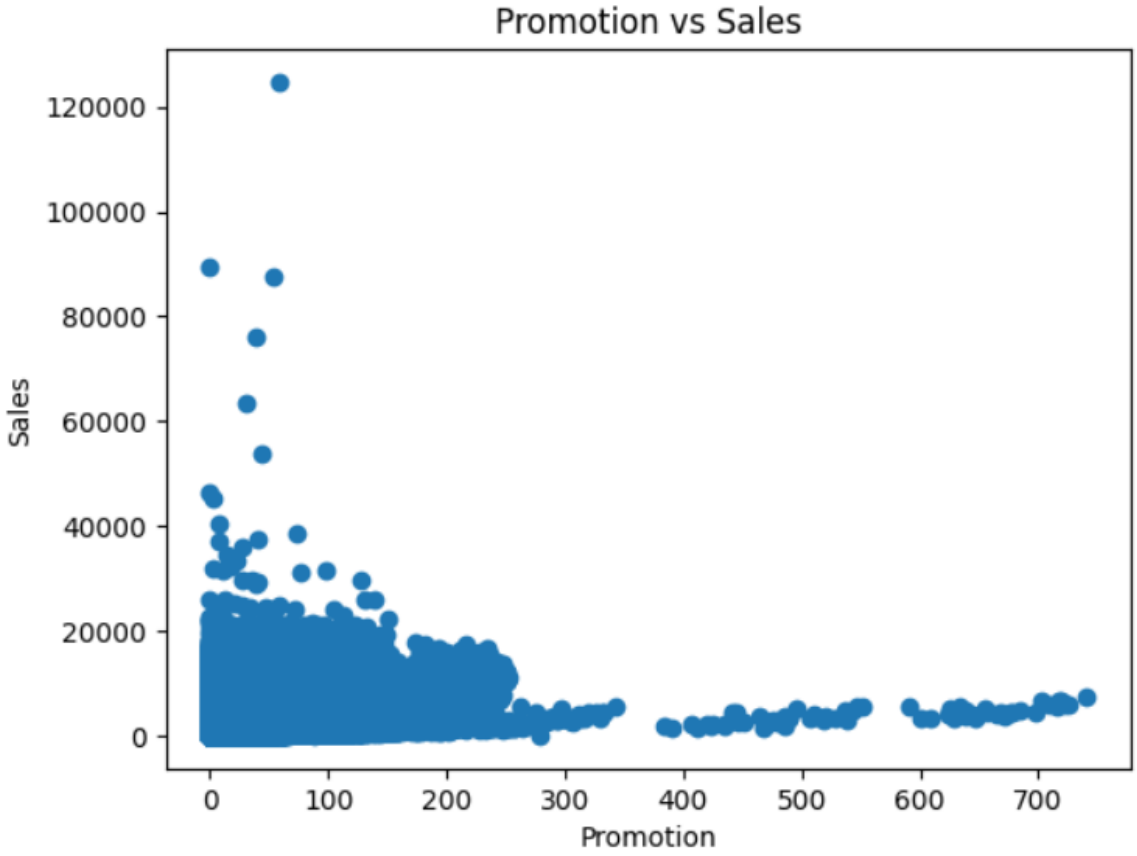
To answer the 3rd question "Does promotion able to improve the sales?" I will use Pearson correlation test to determine the relationship between the two variables, as both of the variables are numericals. The Pearson correlation coefficient measures the linear relationship between two continuous variables and ranges from -1 to +1.

H0 (>0.05)= The promotion does not affect store sales.

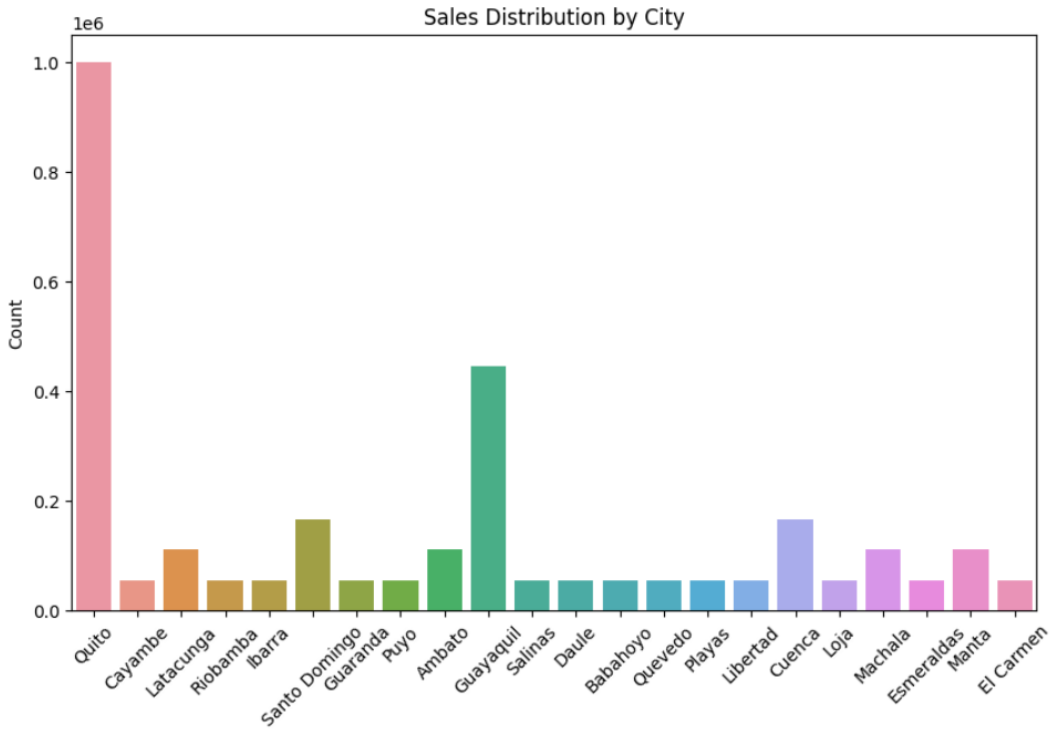
H1 (<0.05)= The promotion does affect store sales.

```
Pearson correlation coefficient: 0.42792320481209284
p-value: 0.0
```

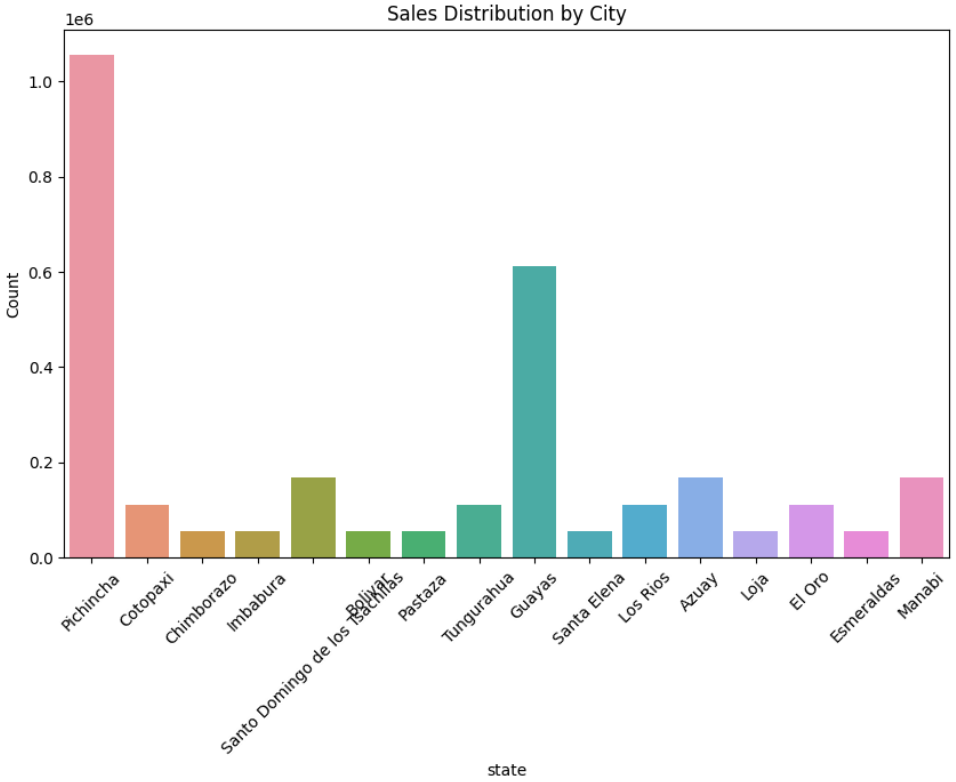
Based on the Pearson correlation coefficient of 0.4279 and the p-value of 0.0, we can reject the null hypothesis (H0) and conclude that there is a significant relationship between promotion and store sales. Therefore, the promotion does affect store sales.


Promotion vs Sales
120000
100000
80000
60000
40000
20000
0
Sales
0
100
200
300
400
500
600
700
Promotion

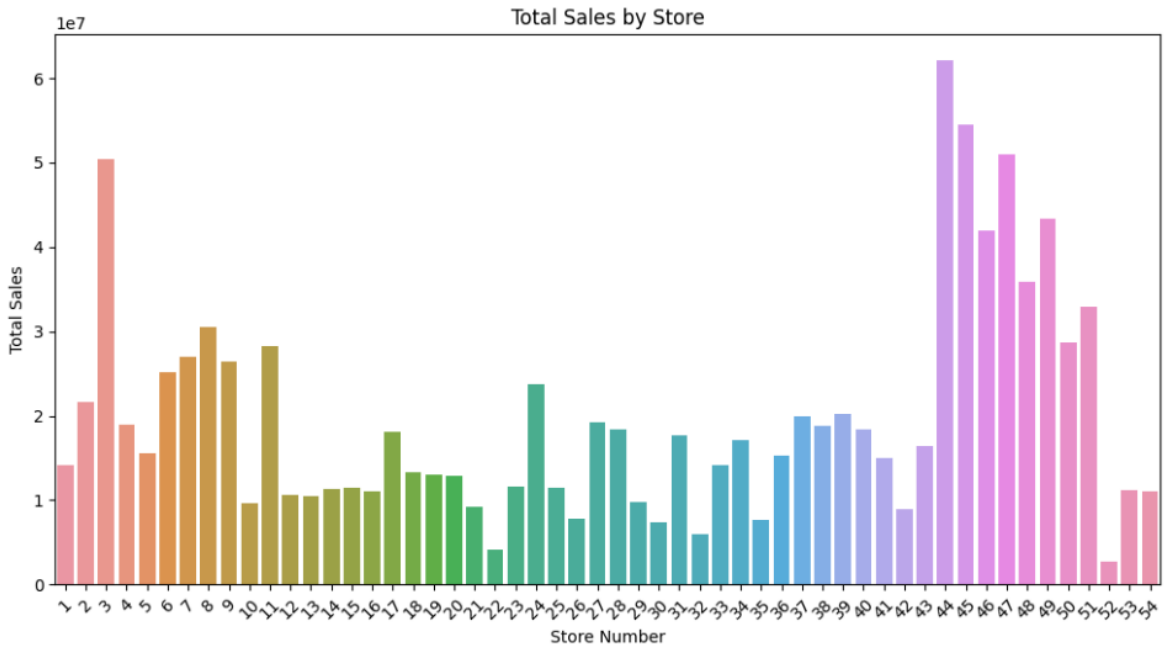
## Which city has the most most number of customers?




## Which state has the most number of customers?



Which of the stores has the highest sales.




## Prediction Methods

To predict supermarket sales using big data analytics, I am thinking to evaluate some of below machine learning algorithms that are commonly used in time-series forecasting. These algorithms were selected based on their ability to handle large volumes of data, capture complex patterns, and generate accurate predictions. The following is a brief overview of the machine learning algorithms used in this study:

1. Linear Regression: This is a basic machine learning algorithm that models the relationship between a dependent variable and one or more independent variables. In this study, linear regression was used to model the relationship between sales revenue and weather variables such as temperature, precipitation, and humidity.

2. ARIMA (Autoregressive Integrated Moving Average): This is a popular time-series forecasting algorithm that models the trend, seasonality, and noise in the data. ARIMA is widely used for sales forecasting, and was used in this study to model the sales patterns observed during the two defined seasons.

3. Random Forest: This is an ensemble learning algorithm that builds a multitude of decision trees to generate predictions. Random Forest is often used in classification tasks, but can also be used for regression tasks such as sales forecasting.

4. XGBoost (Extreme Gradient Boosting): This is a tree-based ensemble learning algorithm that uses gradient boosting to iteratively improve the performance of weak models. XGBoost is highly efficient and accurate and has been used in a variety of applications including sales forecasting.

5. LSTM (Long Short-Term Memory): This is a type of neural network that is capable of capturing long-term dependencies in time-series data. LSTM is widely used in natural language processing and speech recognition but can also be used for sales forecasting.

## Conclusion

We showed that the sales at a store are dependent on multiple factors like product category, promotions, store type, city , state, month. We have calculated autocorrelation of sales on training dataset and then performed stationarity tests such as Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test . Then using the ARIMA machine learning algorithm, we can successfully predict sales at different store locations.

Mean Absolute Error (MAE) , Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) helped me to evaluate the performance of the model.

The ARIMA(2,1,1) model seems to give a directionally correct forecast. And the actual observed values lie within the 95% confidence band. That seems fine.

But each of the predicted forecasts is consistently below the actuals. That means, by adding a small constant to our forecast, the accuracy will certainly improve. So, there is scope for improvement.

## Assumptions

The model is evaluated based on open source data available on the internet. The analysis is limited to Favorita stores located at Ecuador. The model performance can improve by using more feature available in production data. The machine learning aspect of this project is mainly the responsibility of the developers and data scientists.

## Challenges/Issues

As of now it is very early to document all the challenges and issues that I could face during the course of the project. But things like data quality of the datasets used in the project, algorithm or approach to use for the analysis and prediction, accuracy of the algorithm can be used as primary challenges in this project. I am confident that I rely on the previous course notes and professor's guidance to mitigate all these challenges.

## Implementation Plan

To implement this project, we will perform the following main tasks:

1. Gather and prepare dataset.
2. Prepare Design Document.
3. Code, and test
4. Present
5. Deploy

## Ethical Considerations

There are some traditional ethical considerations in a typical service provider relationship.

Bias and fairness: Ensuring that the algorithms and models used in the project do not perpetuate or amplify biases and discrimination against any group.

Data privacy: Ensuring that the data collected is obtained in a legal and ethical manner and that personal information is not misused or disclosed without consent.

Transparency: Ensuring that the data sources, analysis methods, and findings are transparent and easily understandable to all stakeholders.

Security: Ensuring that the data is secure and protected against unauthorized access or theft.

Informed consent: Ensuring that individuals whose data is being used in the project are fully informed about the project's purpose, risks, and benefits and have given their informed consent to participate.

Impact on society: Ensuring that the project's results do not have a negative impact on society or vulnerable populations.

## Questions and Answers

Q: What is technical analysis?

Q: What is machine learning?

Q: How do graphics help?

Q. Explain RNN algorithm.

Q. Explain Autoregressive Integrated Moving Average Model (ARIMA) model algorithm.

Q. Can the use of ML adversely affect sales forecast of supermarket store?

Q. Does this type of stores affect the store sales?

Q. Which family has the highest sales?

Q. Does promotion able to improve the sales?

Q. Which city has the greatest number of customers?

Q. Which state has the greatest number of customers?

Q. Which of the stores has the highest sales.

Q. Which month has the most sales, and least sales.

# References:


Kaggle Datasets https://www.kaggle.com/datasets/yapwh1208/supermarket-sales-data/data?select=annex3.csv

Applied Machine Learning for Supermarket Sales Prediction

https://www.researchgate.net/publication/338681895_Applied_Machine_Learning_for_Supermarket_Sales_Prediction

Predicting Sales https://towardsdatascience.com/predicting-sales-611cb5a252de

Time Series Forecasting of the monthly sales with LSTM and BiLSTM
https://denisechendd.github.io/Time-Series-Forecasting-of-the-monthly-sales-with-LSTM-and-BiLSTM/

Exploratory data analysis using supermarket sales data in Python

https://towardsdatascience.com/exploratory-data-analysis-using-spermarket-sales-data-in-python-e99d329a07fc

How to Interpret ARIMA Results https://analyzingalpha.com/interpret-arima-results

ARIMA Model – Complete Guide to Time Series Forecasting in Python
https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/