**Demand Forecasting on Retail (web e-tailer chain) products:**

**Executive Summary:**

* Demand forecasting for a web e-tailer chain selling retail products is a critical process that helps the organization optimize its inventory levels, manage its supply chain, and maximize its sales and revenue. To perform demand forecasting, I analyzed a dataset that includes various columns such as Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, and Profit.
* Accurate demand forecasting is crucial for ensuring that the company has the right products in the right quantities at the right time, which is essential for meeting customer needs and expectations.
* To improve the accuracy of demand forecasting, I recommend leveraging advanced technologies. These technologies can help automate the forecasting process, analyze large amounts of data quickly, and provide real-time insights into customer behavior, sales trends, and market conditions.
* In addition, I recommend implementing a continuous improvement process to refine the demand forecasting model based on ongoing feedback and results. This process can help the organization stay agile and responsive to changes in the market and customer demand, ultimately leading to improved sales, higher customer satisfaction, and increased revenue.

# Project motivation/background:

The background for this project is based on the challenges faced by web e-tailer chains selling retail products, particularly in the context of the growing e-commerce industry. With an increasing number of competitors and changing customer behavior and preferences, it is essential for these organizations to have an accurate understanding of customer demand and sales trends to stay competitive and meet customer needs effectively.

Therefore, this project aims to provide the organization with a data-driven approach to demand forecasting that incorporates both quantitative and qualitative methods, leverages advanced technologies and implements a continuous improvement process to refine the forecasting model over time.

# Key Questions:

## Do the product Analysis

## Find the product with highest demand in each segment by our analysis.

## Find the category of the product with highest demand in each region

## Find the category of the product which is being sold with highest discounts

## Find the most Used Ship Mode?

## Find the Impact of Discount on Profit%

## Which Sub-Category should be focused?

## Which state should be chosen for expansion & which products should be focused?[¶](http://localhost:8888/notebooks/Downloads/Project.ipynb#Which-state-should-be-chosen-for-expansion-&-which-products-should-be-focused?)

## Do the Sales/Profit Analysis

## Find the TOP 5 customers with most number of orders and revenue

## Find the category of the product that earns highest profit

## Find the city with highest sales and Profit

## Product Categories and Sub-Categories that made loss

## Represent all the insights in 2D.

## Estimate the demand of different type of products based on the sales history using predictive analysis.

* Forecasting which model is better basis the accuracy

# Data Source:

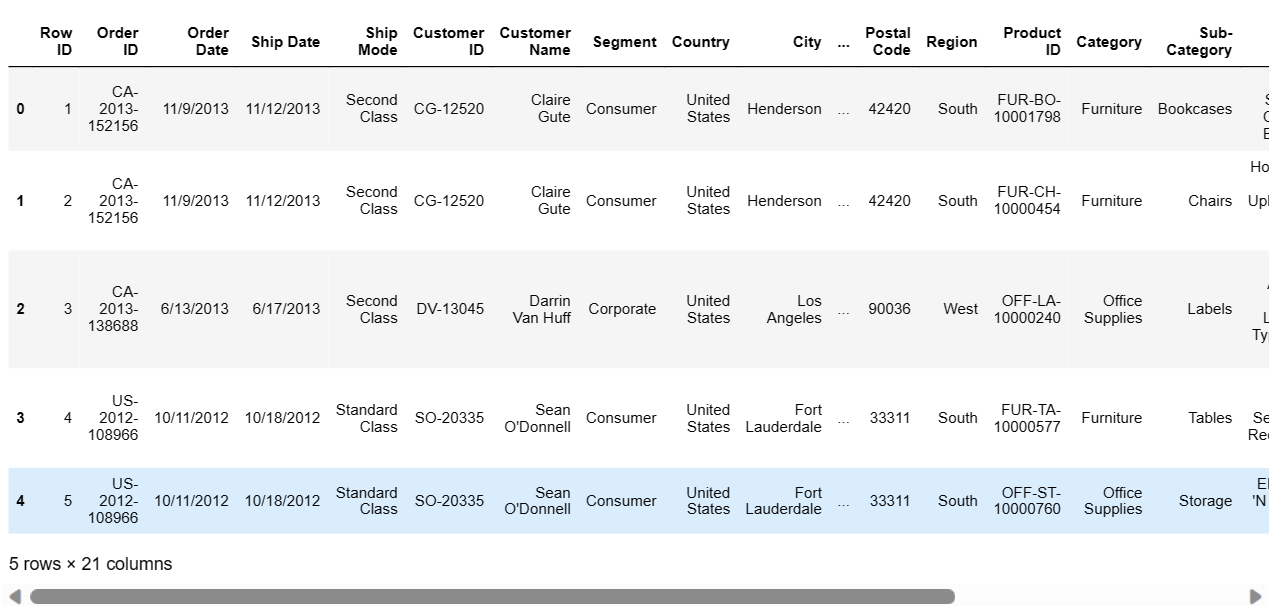
* It is a Real time data of a Retail (web e-tailer chain). We have got this data from online sources.

# Data Description:

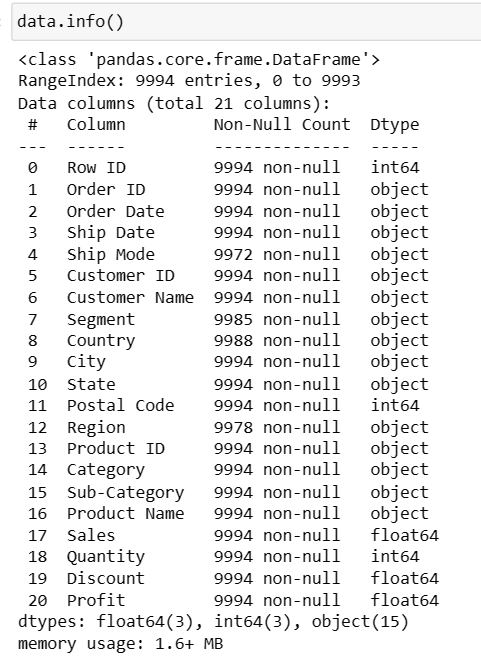
* This is a web e-tailer data set of a startup in E-commerce Industry.
* The data set have the volume of 9984 rows and 21 columns.
* The 21 columns include Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, Profit.
* Row ID: A unique identifier assigned to each row of data.
* Order ID: A unique identifier assigned to each order made by a customer.
* Order Date: The date on which the customer placed the order.
* Ship Date: The date on which the order was shipped to the customer.
* Ship Mode: The shipping method used to deliver the order.
* Customer ID: A unique identifier assigned to each customer.
* Customer Name: The name of the customer who placed the order.
* Segment: The market segment to which the customer belongs (e.g., consumer, corporate, or home office).
* Country: The country where the order was placed.
* City: The city where the order was placed.
* State: The state or province where the order was placed.
* Postal Code: The postal code where the order was placed.
* Region: The geographic region where the order was placed.
* Product ID: A unique identifier assigned to each product.
* Category: The broad category to which the product belongs (e.g., furniture, office supplies, or technology).
* Sub-Category: The sub-category to which the product belongs (e.g., chairs, desks, or phones).
* Product Name: The name of the product.
* Sales: The total sales revenue generated by the order.
* Quantity: The quantity of the product ordered.
* Discount: The discount applied to the order.
* Profit: The profit earned from the order.

# Data Transformation/Exploratory Data Analysis:

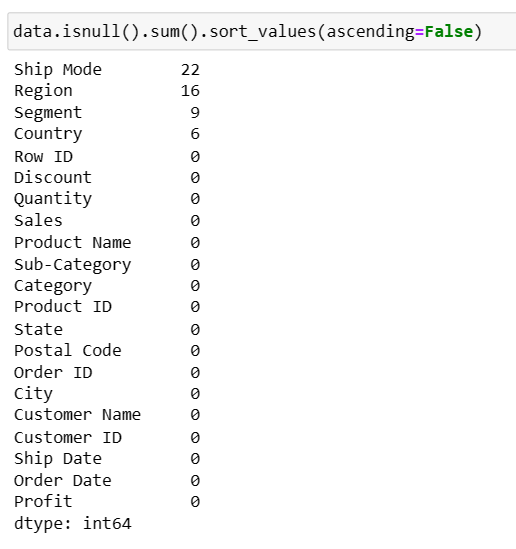
* Data transformation involves preparing data for analysis by cleaning, preprocessing, aggregating, and transforming it into a more useful format.
* Exploratory data analysis is the process of exploring and summarizing the characteristics of the data to gain insights into patterns, relationships, and trends that can inform further analysis. This involves using statistical and visual methods to examine the data and identify any outliers, missing values, or other anomalies that may impact the analysis.
* We have done the EDA in the following steps:
  1. **Data Loading**: In this part we have loaded the data in to our system.



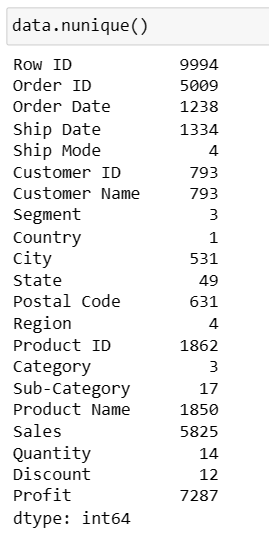
* 1. **Data Info:**



* 1. **Data Cleaning**
* Finding the number of null values that are present in the data

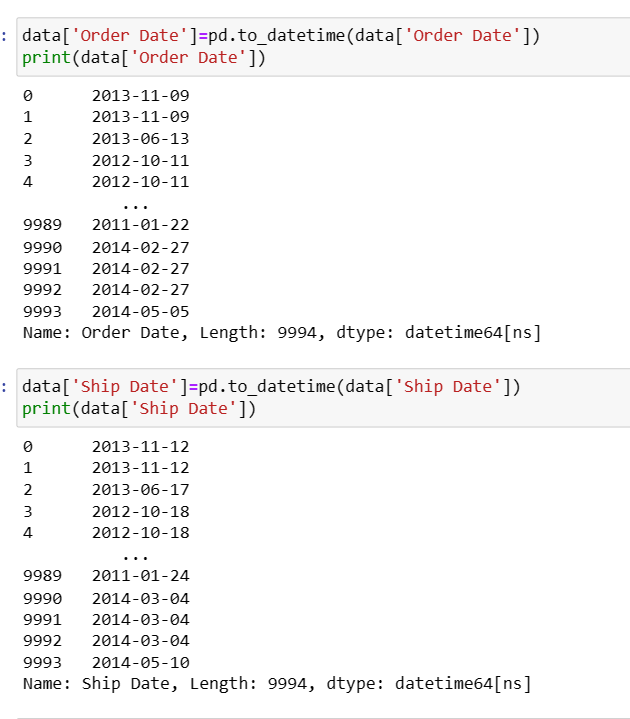


* We have found the number of unique values for each column.



These are the count of unique values present in the dataset.

* Correcting data types: Ensuring that the data types of each variable or column are correct and consistent with their intended use.



We have ensured that the data is consistent and standardized, such as by converting all text to a common case, or ensuring consistent date and time formats.

* The cleaning of data is done
  1. **Exploratory Data Analysis**:

### Univariate analysis, bi variate analysis and calculation of some important values is also done

### In this section the patterns of how the particular variable is affected is shown in the graphs

### We have used the different graphs like histograms, crosstabs, count plots …

### Univariate and bivariate analysis for few factors is shown below:

### Category:

### 

### Sub-Category:

### 

### Ship Modes:

### 

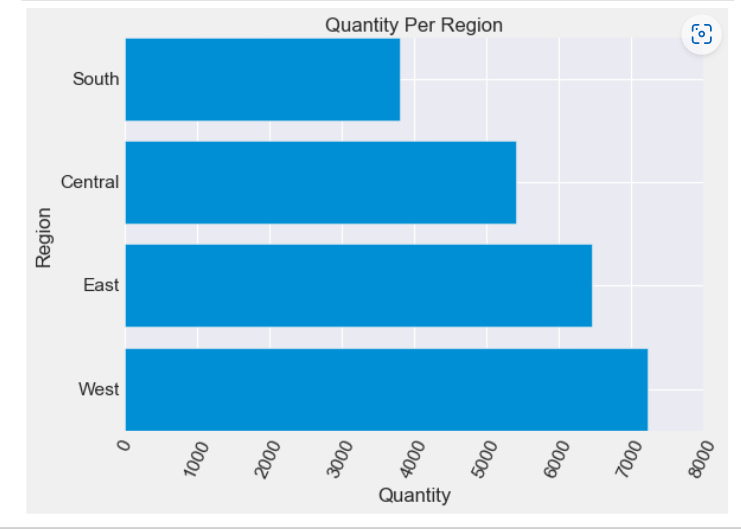
## **Product with highest demand in each segment:**

## 

## **Category of the product with highest demand in each region:**

## 

### 



## **Impact of Discount on Profit%:**

## 

## **Which Sub-Category should be focused?**

## 

## 

## **Which state should be chosen for expansion & which products should be focused?**

## 

## 

# Sales/Profit Analysis:

## 

## 

## 

# The TOP 5 customers with the greatest number of orders and amount spent:

## 

## 

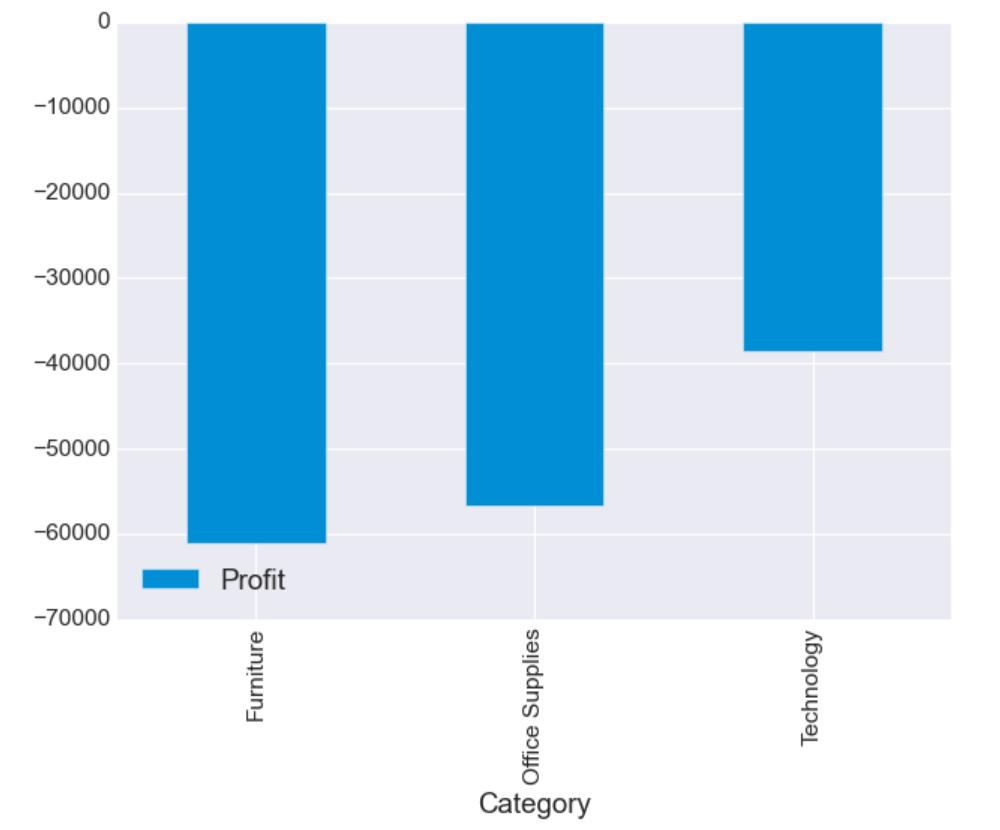
## **Most valuable customer:**

## 

## **The segment which yields highest profit:**

## 

## **Product Categories that made loss:**



* Furniture category products had made the highest loss accounting to -60936

# Models and Analysis:

As the data contains various metrics related to sales like orders, categories, sub categories, sales etc. and the quantity is a continuous metric, In this scenario we are taking the quantity as the target variable and predict it using various features. This is clearly a regression problem as the target variable is continuous. We are using numerous features and those features are correlated to the target variable quantity. So we are using two most used regression models namely XGBoost and Random Forest

# Random Forest Regression Model:

With the aid of several decision trees and a method known as Bootstrap and Aggregation, also ref erred to as bagging, Random Forest is an ensemble methodology capable of handling both regression and classification tasks. This method's fundamental principle is to integrate several decision trees to get the final result rather than depending solely on one decision tree.

Random Forest is utilized for tasks like classification and regression. To increase accuracy and lessen overfitting of the model, this ensemble learning technique integrates several decision trees. Each tree in a Random Forest model is trained using a unique subset of the training data and a randomly selected subset of the features. Combining all of the trees' projections yields the final conclusion. The advantages of Random Forest are its high accuracy, resilience, and capacity for handling highly dimensional data with numerous attributes. It is extensively employed in many different industries, including banking, healthcare, and image analysis.

In our scenario, the mapping between the predicted and actual quantities are matching with each other with a MAPE of 43 which means the accuracy is 1-MAPE i.e 57%

**XGBoost Model:**

Extreme Gradient Boosting (XGBoost), a potent machine learning algorithm, is a member of the ensemble learning methodology family. This tree-based model combines numerous weak learners into a single strong learner using boosting. XGBoost is a well-liked option for a variety of applications, including data mining, natural language processing, and image recognition, because to its speed and scalability. Additionally, a variety of hyperparameters are provided, which can be adjusted to enhance model performance.

Being able to manage missing data and outliers, which can frequently be an issue for other machine learning algorithms, is one of XGBoost's primary advantages. It is also quite adaptable and may be used for both classification and regression problems.

In our scenario, the mapping between the predicted and actual quantities are matching with each other with a MAPE of 0.41 which means the accuracy is 1-MAPE i.e. ~60%.

As we see the accuracies with both the models, they look to be similar varying around 55 to 60% as we reiterate the training again and again. This is considerably low as the dataset is not continuous and has a lot of missing data. The accuracies can further be improved by performing hyper parameter tuning, where we tune the parameters of the model using grid search CV where it attempts to reduce the cost function as much as it can in order to give the best model.

# Findings and Managerial Implications:

* Technology products have more competitors, therefore proper advertisement and marketing strategy to be implemented in the regions rather states where its incurring losses, for example: Otherwise technology seems to a cash cow in the product category bringing good revenue and thus good profits.
* Office Supplies is definitely a dark horse category to be looked out for as they are regular used products
* Furniture needs to check upon its discount rates i.e., not only boosting up sales but also keeping Profit in mind and thus need to work upon Pricing Strategy so as to get an edge over the competitive market.
* Discount can be pertained when the customer is more regular or the product is more popular yielding profit, beside generating Sales.
* It is obvious that Wyoming, is the state with Sales. But, Vermont is producing highest Profit.
* Looking at the Profit Percentage table we see states like Columbia, Minnesota coming up, although sales is low.
* Hence the expansion plan should be based on these states maybe by changing different strategies as applicable for the different states keeping customer satisfaction in mind.
* From the Past Trends of the sales data, it was observed that a minimum of 15% increase in sales is to be met in coming year to have a smooth marketing growth

# Conclusions on Insights Observation and prediction analysis:

* For the Product Line I observed the following insights:
* Office Supplies is the most ordered.

1. Progressive Profit and Sales Growth.

2. Moderate sales and profit value.

* Technology is the least ordered.

1. Increasing Sales and profit over the years.

* Furniture has made huge losses despite having good sales.
* Customer Segment:
* Consumer yield the highest profit, quantity sold is also most.
* Office supplies are the products with highest demand in each segment
* Most valuable customers are from technology and office supplies
* Standard as shipment is chosen over any other modes by all segments.
* Geographical:
* West – High quantities, High profit.
* South- least sales but, good profit.
* East shows several fluctuations, Central incurred losses.
* States with highest profit percentage and loss percentages are in East.
* New York city is the best performing city

1. Corporate

2. Consumer

3. Home Office

* I also did the predictive analysis with 2 models,
* For XGBOOST: In this scenario, the mapping between the predicted and actual quantities are matching with each other with a MAPE of 0.32 which means the accuracy is 1-MAPE i.e 68%
* For Random Forest: In this scenario, the mapping between the predicted and actual quantities are matching with each other with a MAPE of 0.49 which means the accuracy is 1-MAPE i.e 51%
* This shows that the random forest model is performing with a MAPE of 0.49 and the accuracy is 51%
* We can see a slight difference in accuracies obtained from XGBoost and Random Forest models, whereas XGBoost performs better with an accuracy of 68%.
* These models can further be improved by tuning the hyper parameters using grid search CV with helps in finding the optimal parameters and getting the best accuracies.[¶](http://localhost:8888/notebooks/Downloads/Project.ipynb)