Project Report

on

Data Analysis and Sales Forecasting of Walmart Retail Data using PySpark and Machine Learning.

Submitted in partial fulfilment for the award of
Post Graduate Diploma in Big Data Analytics (PG-DBDA)
From Know IT(Pune)

Guided by:

Mrs. Trupti Joshi & Mr. Prasad Deshmukh

Submitted by:

Kiran Thorat (240343025022)

Manu Singh (240343025025)

Payal Jadhav (240343025034)

Shubham Mule (240343025056)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Kiran Thorat (240343025022)

Manu Singh (240343025025)

Payal Jadhav (240343025032)

Shubham Mule (240343025056)

Have successfully completed their project on

Data analysis and Sales Forecasting of Walmart Retail Data using PySpark and Machine Learning.

Under the guidance of

Mrs. Trupti Joshi

Mr. Prasad Deshmukh

ACKNOWLEDGEMENT

Our project "Data Analysis and Sales Forecasting of Walmart Retail Data using PySpark and Machine learning" was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention Mrs Trupti Joshi and Mr. Prasad Deshmukh for their valuable guidance while working on this project. Their guidance and support helped us overcome various obstacles during the project work.

We are highly grateful to Mr. Vaibhav Inamdar Manager (Know IT), CDAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through CDAC ACTS, Pune.

Our most heartfelt thanks go to Mr. Prasad Deshmukh (Course Coordinator, PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility, and extra Lab hours to complete the project and throughout the course up to the last day here in CDAC Know-IT, Pune

TABLE OF CONTENTS

ABSTRACT

This project focuses on the historical data analysis, modelling, and forecasting of Walmart's extensive retail data using PySpark, machine learning, and Tableau. By exploring Walmart's detailed historical sales data, we aim to uncover patterns and trends that can drive better business decisions. Retailers rely on such data to enhance operational efficiency, refine inventory management, and optimize sales strategies. The methodology includes data preprocessing and feature engineering, followed by advanced machine learning techniques to develop robust predictive models. PySpark is used for processing large datasets efficiently, while machine learning algorithms provide accurate sales forecasts. Tableau creates intuitive dashboards to present these insights in a user-friendly manner, making complex data accessible and actionable. This approach will support Walmart in making informed decisions about inventory management, demand forecasting, and strategic planning.

INTRODUCTION

Accurate predictions and insights derived from historical retail data are essential for making informed business decisions. This project centers on the analysis of Walmart's historical sales data to develop machine learning models capable of forecasting future sales trends. These predictions will support key business operations such as inventory management and pricing strategies. Leveraging PySpark's distributed computing capabilities allows us to efficiently process Walmart's extensive dataset, while Tableau is utilized to create intuitive visualizations, making the data insights easily accessible and actionable for stakeholders.

Dataset Collection and Features

1-Data Sources:

the dataset used in this project was obtained from GOOGLE

2-Data Structure:

Each row in the dataset represents a single retail transaction, with columns describing attributes related to customers, products, sales, shipping, and more.

3-Dataset Size:

The dataset contains a substantial number of records, offering insights into various aspects of retail sales, customer behavior, and product performance.

Features / Attributes:

1. **City:** The city where the transaction took place.
2. **Customer Age:** The age of the customer involved in the transaction.
3. **Customer Name:** The name of the customer who made the purchase.
4. **Customer Segment:** The segment to which the customer belongs (e.g., Corporate, Small Business, Consumer).
5. **Discount:** The discount applied to the transaction, represented as a percentage.
6. **Order Date:** The date when the order was placed.
7. **Order ID:** A unique identifier for the order.
8. **Order Priority:** The priority level of the order (e.g., High, Medium, Low).

9. **Order Quantity:** The quantity of items ordered.
10. **Product Base Margin:** The profit margin of the product, represented as a percentage.
11. **Product Category:** The category to which the product belongs (e.g., Furniture, Technology).
12. **Product Container:** The type of container used for the product (e.g., Jumbo Drum, Small Box).
13. **Product Name:** The specific name or model of the product.
14. **Product Sub-Category:** The sub-category of the product (e.g., Chairs & Chairmats, Tables).
15. **Profit:** The profit made from the transaction, usually calculated as sales minus costs.
16. **Region:** The geographic region where the transaction occurred (e.g., Central, West, East).
17. **Sales:** The sales amount for the transaction.
18. **Ship Date:** The date when the product was shipped to the customer.
19. **Ship Mode:** The mode of shipment used (e.g., Delivery Truck, Regular Air, Express Air).
20. **Shipping Cost:** The cost associated with shipping the product.
21. **State:** The state where the transaction took place.
22. **Unit Price:** The price per unit of the product sold.
23. **Zip Code:** The postal code where the transaction occurred.

SYSTEM REQUIREMENTS

Hardware Requirements

1. **Computer:** A computer with sufficient processing power and memory to run data processing and analysis tasks. A modern multicore processor and at least 8 GB of RAM are recommended.

2. **Storage:** Adequate storage space to store the generated dataset and any additional datasets if required. An SSD (Solid State Drive) is recommended for faster data access.

3. **Internet Connection:** A stable internet connection for downloading and installing software packages and libraries, as well as for any online resources needed during the project.

Software Requirements

1. **Operating System:** Windows 10 or higher
2. **Python**: The project heavily relies on Python for data generation, analysis, and machine learning. Ensure Python is installed on your system.
3. **Python Libraries**: Install the following Python libraries and dependencies using package managers like pip:
4. **Matplotlib and Seaborn**: data visualization
5. **PySpark:** For preprocessing and ML.

6. **Apache Spark:** For Preprocessing of dataset and ML.
7. **Integrated Development Environment (IDE):** Choose a Python friendly IDE, such as Jupyter Notebook, or your preferred text editor.

4.Visualization Software

Tableau: If you plan to visualize and analyze data with Tableau, install Tableau public.

FUNCTIONAL REQUIREMENTS

1.Python:

1. Python is a general purpose and high-level programming language.
2. It is use for developing desktop GUI applications, websites and web applications.
3. Python allows to focus on core functionality of the application by taking care of common programming tasks.
4. Python is derived from many other languages, including ABC, Modula3, C, C++, Algol68, Small Talk, and Unix shell and other scripting languages.

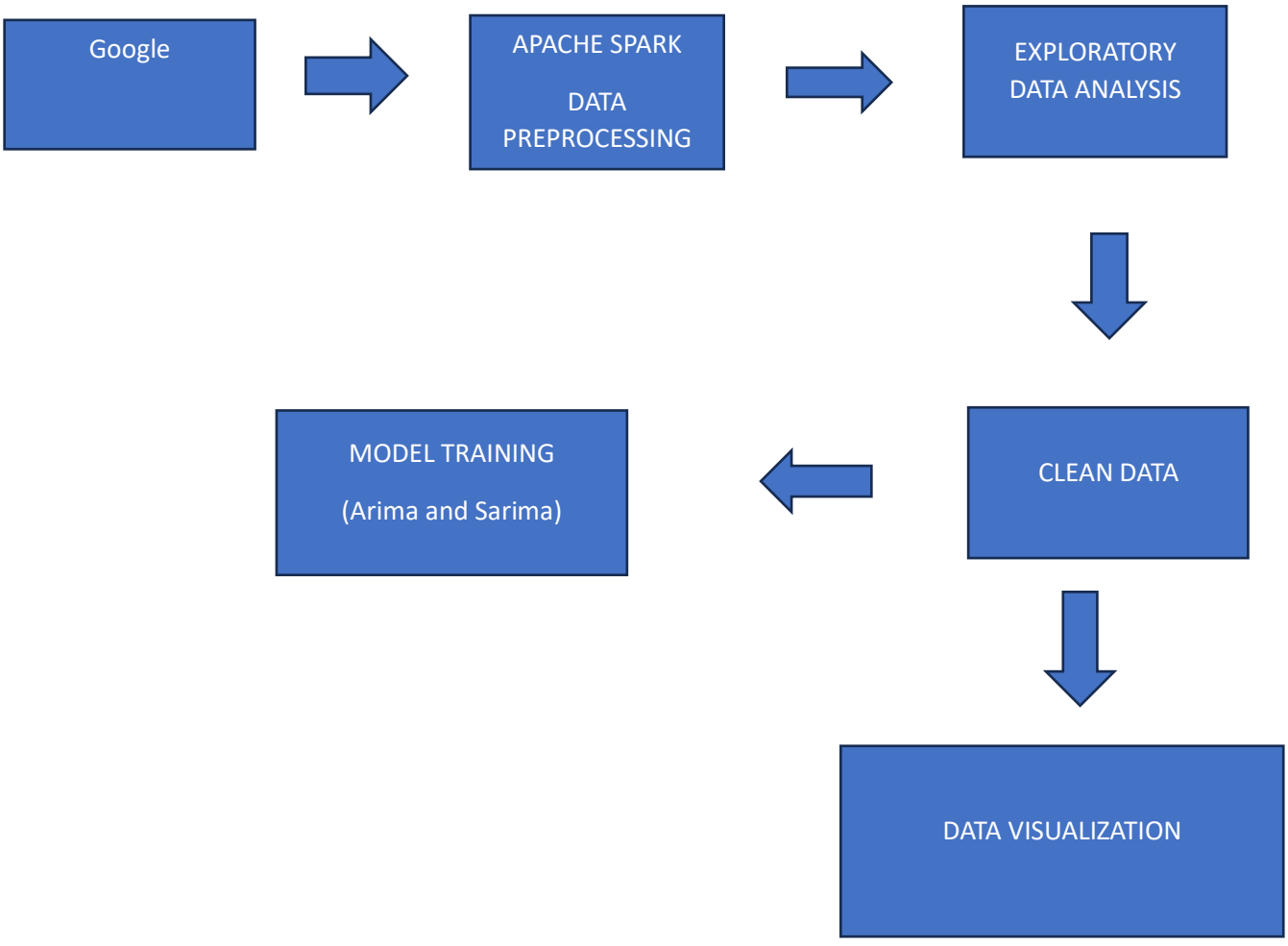
2.APACHE SPARK:

1. What is Spark: Apache Spark is an opensource distributed computing system designed for processing large volumes of data.
2. Key Features: Spark provides a number of key features that make it well-suited for processing big data, including in memory processing, support for various data sources and formats, fault tolerance, and scalability.
3. Spark also provides a range of APIs, including SQL, streaming, machine learning, and graph processing, making it a versatile platform for a wide range of use cases.

3.Tableau:

1. Data visualization is the graphical representation of information and data.
2. It helps create interactive elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
3. It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.
4. Tableau is widely used for Business Intelligence but is not limited to it.
5. All of this is made possible with gestures as simple as drag and drop.

ARCHITECTURE



PRE - PROCESSING

1.Data Loading: The dataset is loaded into a PySpark DataFrame from a CSV file. This DataFrame will be used for subsequent preprocessing steps.

2.Column Removal: Certain columns that are not relevant to the sales prediction task or contain mostly null values are dropped from the DataFrame.

3.Duplicate Removal: Duplicate records in the DataFrame are removed to ensure data integrity and avoid bias in the analysis.

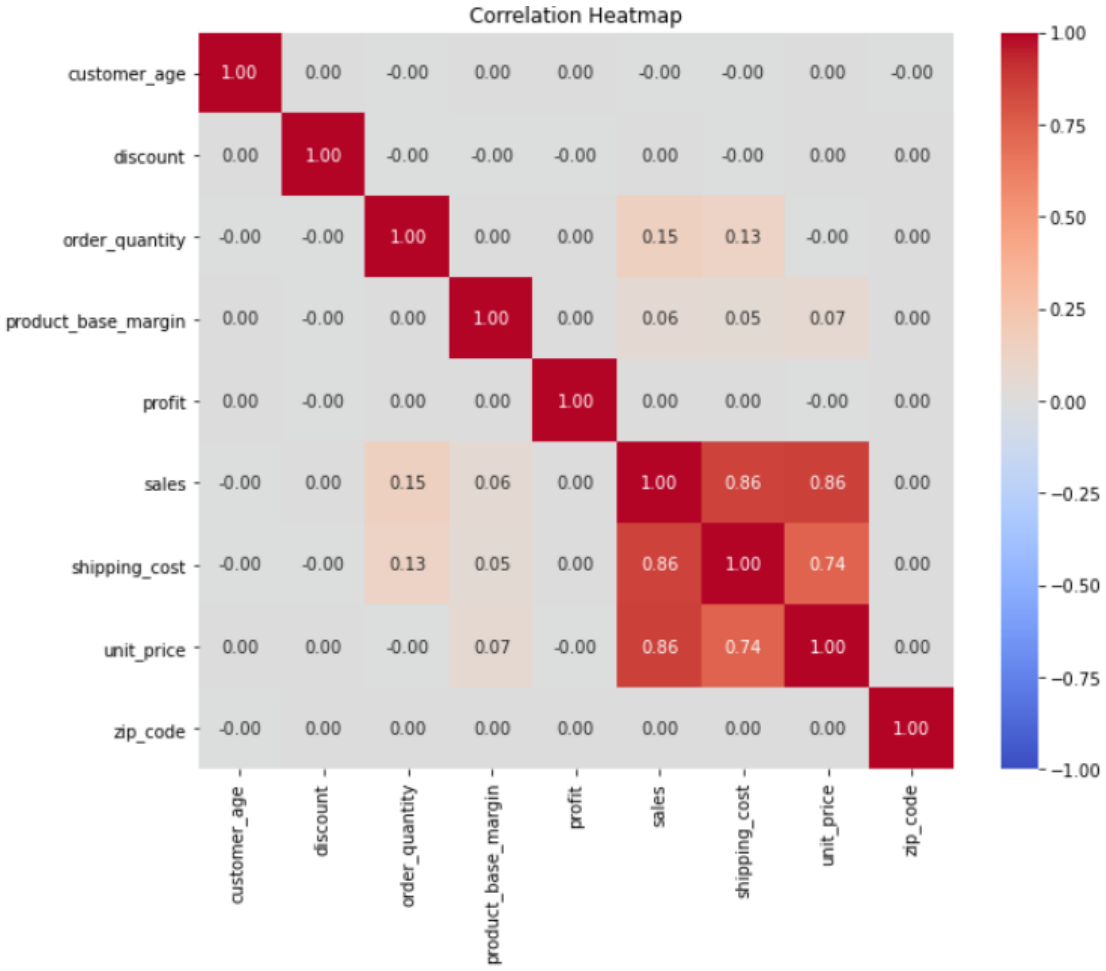
4.Handling Missing Values: Null values in specific columns are dropped.

5.Outlier handling:-outliers were found based on IQR and replaced by respective upper bound and lower bound

6.Feature Engineering: A new feature 'month' is created based on the 'order date' column to represent the monthly sales at the time of prediction.

7.Final Cleanup: Unnecessary columns are dropped from the DataFrame, leaving only the relevant features for the prediction task

EDA


Correlation Heatmap
customer_age
discount
order_quantity
product_base_margin
profit
sales
shipping_cost
unit_price
zip_code
1.00 0.00 -0.00 0.00 0.00 -0.00 -0.00 0.00 -0.00
0.00 1.00 -0.00 -0.00 -0.00 0.00 -0.00 0.00 0.00
-0.00 -0.00 1.00 0.00 0.00 0.15 0.13 -0.00 0.00
0.00 -0.00 0.00 1.00 0.00 0.06 0.05 0.07 0.00
0.00 -0.00 0.00 0.00 1.00 0.00 0.00 -0.00 0.00
-0.00 0.00 0.15 0.06 0.00 1.00 0.86 0.86 0.00
-0.00 -0.00 0.13 0.05 0.00 0.86 1.00 0.74 0.00
0.00 0.00 -0.00 0.07 -0.00 0.86 0.74 1.00 0.00
-0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

MACHINE LEARNING ALGORITHM

ARIMA Model

The ARIMA (AutoRegressive Integrated Moving Average) model was utilized in our project to analyze and forecast time series data. ARIMA is a popular statistical method for modeling time series data that exhibits patterns over time.

Purpose and Application: In our project, ARIMA was chosen to forecast [sales]. The model's ability to handle both trend and noise made it suitable for our analysis, allowing us to forecast future values based on historical data.

Data Preparation: Before applying the ARIMA model, the data underwent several preprocessing steps. This included handling missing values, transforming the data to achieve stationarity through differencing, and splitting the data into training and test sets to evaluate the model's performance.

Model Selection: The ARIMA model parameters (p, d, q) were selected based on the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The chosen parameters were [specific values for p, d, q], which were determined to best capture the underlying patterns in the time series data.

Model Evaluation: The performance of the ARIMA model was assessed using various metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics provided insights into the accuracy of the model's forecasts and its ability to generalize to new data.

Results and Interpretation: The ARIMA model generated forecasts that [forecasted future values, identified significant trends]. The results were visualized using [describe visualization tools, e.g., line charts], which helped in interpreting the forecasted trends and patterns.

Challenges and Limitations: Some challenges faced during the application of the ARIMA model included [dealing with non-stationarity in the data, selecting appropriate parameters]. Additionally, the model's limitations, such as its inability to handle seasonality without adjustments, were noted.

Future Work: To improve the forecasting accuracy, future work could explore alternative models, such as SARIMA (Seasonal ARIMA), or incorporate additional features and data sources. Further analysis may also involve experimenting with different parameter settings and validation techniques.

Conclusion: The ARIMA model provided valuable insights into [specific area, e.g., future sales trends, demand forecasting]. Its application demonstrated the model's effectiveness in capturing and forecasting time series patterns, contributing significantly to the project's objectives.

MODEL PERFORMANCE

For Arima model

```
[433]: # Evaluation parameter
mae = mean_absolute_error(test, forecast)
mse = mean_squared_error(test, forecast)
rmse = np.sqrt(mse)

print(f'Mean Absolute Error (MAE): {mae}')
print(f'Mean Squared Error (MSE): {mse}')
print(f'Root Mean Squared Error (RMSE): {rmse}')
```

```
Mean Absolute Error (MAE): 232958.08282243312
Mean Squared Error (MSE): 69676752399.09314
Root Mean Squared Error (RMSE): 263963.5436932402
```

For Sarima model

```
[433]: # Evaluation parameter
mae = mean_absolute_error(test, forecast)
mse = mean_squared_error(test, forecast)
rmse = np.sqrt(mse)

print(f'Mean Absolute Error (MAE): {mae}')
print(f'Mean Squared Error (MSE): {mse}')
print(f'Root Mean Squared Error (RMSE): {rmse}')
```

```
Mean Absolute Error (MAE): 232958.08282243312
Mean Squared Error (MSE): 69676752399.09314
Root Mean Squared Error (RMSE): 263963.5436932402
```

For prophet Model

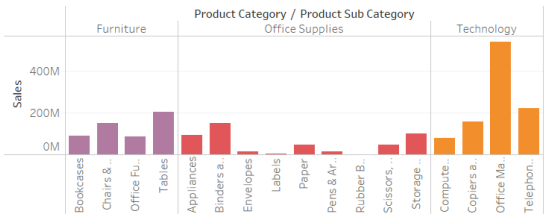
	horizon	mse	rmse	mae	mape	mdape	smape	coverage
0	37 days	2.284778e+06	1511.548068	1237.210749	9.445529	1.785533	1.122055	0.835184
1	38 days	2.285730e+06	1511.863196	1237.396101	9.479430	1.786709	1.122518	0.834992
2	39 days	2.284619e+06	1511.495641	1236.912257	9.453175	1.794311	1.122623	0.835197
3	40 days	2.289053e+06	1512.961657	1238.057588	9.505826	1.797847	1.123045	0.834901
4	41 days	2.295054e+06	1514.943509	1239.367707	9.530154	1.785564	1.122508	0.834406

TABLEAU DASHBOARD

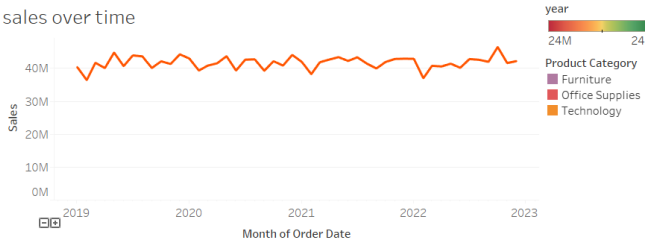
DATA ANALYSIS OF WALMART DATA

SALES PERFORMANCE | PROFIT PERFORMANCE | REGIONAL PERFORMANCE | TOP 10 PRODUCTS BY

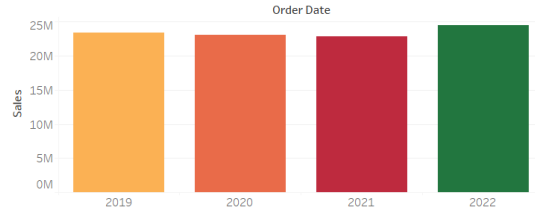
sales by individual product



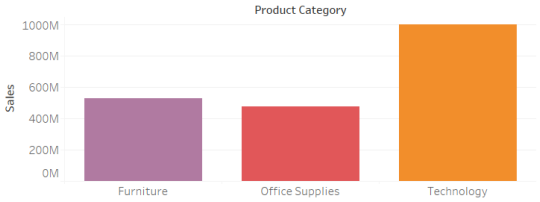
sales over time



sales comparison bar chart



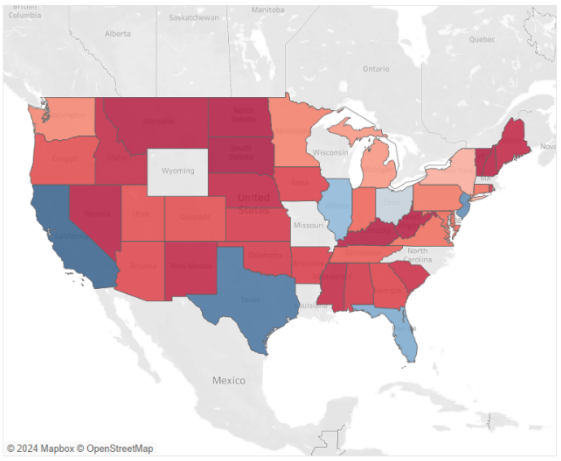
sales by category bar chart



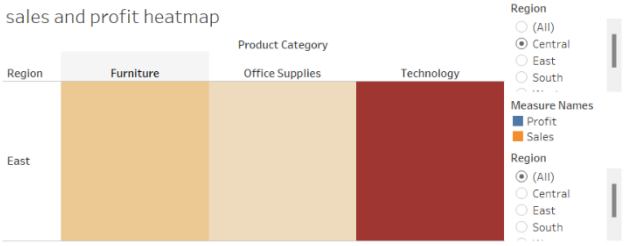
DATA ANALYSIS OF WALMART DATA

SALES PERFORMANCE | PROFIT PERFORMANCE | REGIONAL PERFORMANCE | TOP 10 PRODUCTS BY

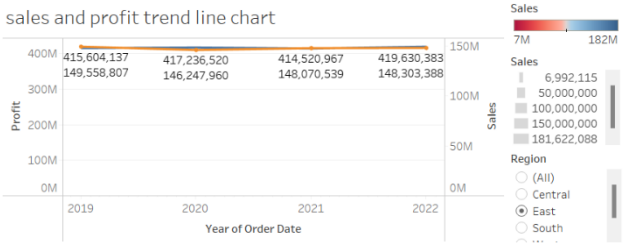
sales by region

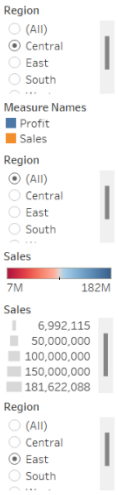


sales and profit heatmap

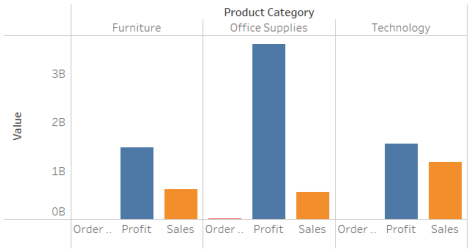


sales and profit trend line chart

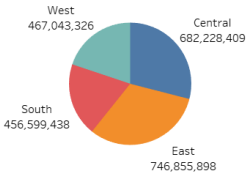




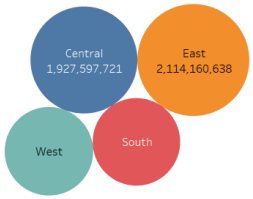
Product Performance

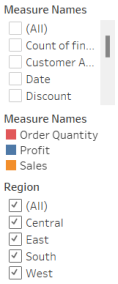


sales by region

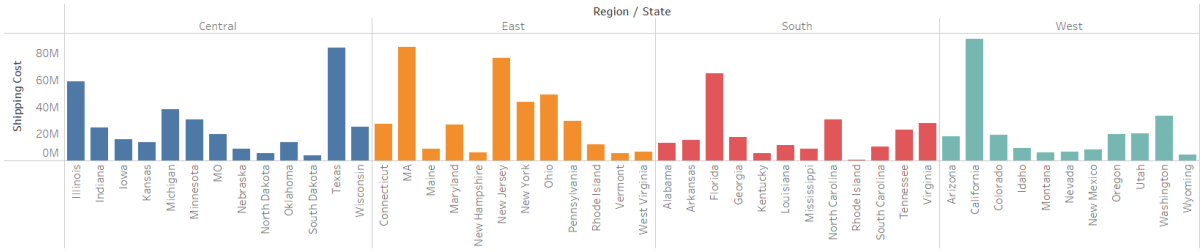


Profit by Region

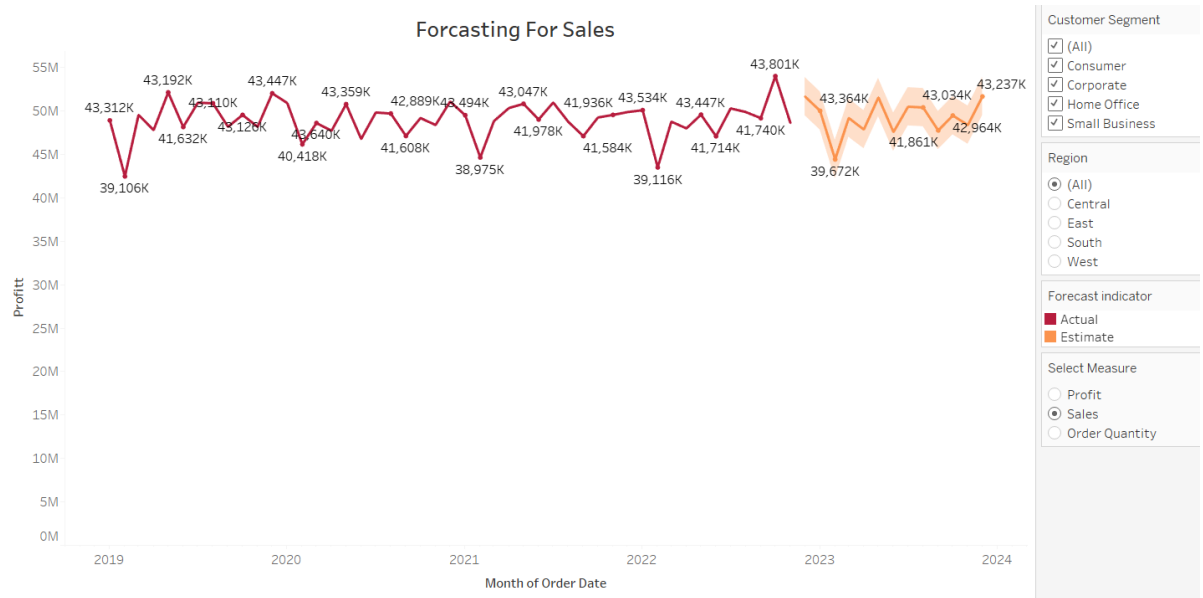




Shipping Costs Analysis



Forcasting For Sales



CONCLUSION

1. The analysis of Walmart sales data provided critical insights into sales trends, seasonality, and the impact of various factors like holidays, promotions, and economic indicators.

2. The predictive models developed during the analysis were effective in forecasting sales performance, identifying key drivers of sales fluctuations.

3. The project successfully demonstrated the importance of data-driven decision-making in retail, particularly in inventory management, supply chain optimization, and promotional planning.

4. The findings emphasized the significance of continuously monitoring sales data to adapt to changing market conditions and customer preferences.

FUTURE SCOPE

1. Incorporating External Data: Integrating additional datasets, such as weather patterns, economic indicators, and competitor pricing, can enhance the accuracy of sales forecasts.

2. Real-Time Analytics: Developing a real-time analytics dashboard could enable Walmart to react quickly to sales trends and operational challenges, optimizing inventory and pricing strategies on the fly.

3. Advanced Machine Learning Models: Exploring more sophisticated machine learning techniques like deep learning or ensemble methods could further improve the precision of sales predictions.

4. Customer Behavior Analysis: Conducting a more in-depth analysis of customer purchasing behavior and preferences could lead to personalized marketing strategies, increasing customer loyalty and sales.

5. Analysis: Analyzing sales data on a regional basis to tailor strategies for different markets, ensuring more effective promotions and inventory distribution.

References:-

1. 1.Apache spark(pyspark package — PySpark 2.4.5 documentation (apache.org))
2. Python. [https://www.python.org/]
3. scikit-learn. (https://scikit-learn.org/stable/index.html)
4. https://matplotlib.org/stable/api/pyplot_summary.html
5. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.quantile.html
6. https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html
7. https://prophet.readthedocs.io/en/latest/api.html
8. https://www.statsmodels.org/dev/generated/statsmodels.graphics.tsaplots.plot_acf.html
9. https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html
10. 10.https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html.