

Super Store Sales Analysis

Submitted for

DATA VISULIZATION AND DASHBOARD

Submitted by:

E23MCAG0078

SHUBHANGI KUNDU

Submitted to-

Dr. Nitin Arvind Shelke

July-Dec 2023

**SCHOOL OF COMPUTER SCIENCE AND
ENGINEERING**



INDEX

Sr No.	Content	Page No.
1	Introduction and Related Work	4
2	Software Used	6
3	Methodology	7
4	Experimental Results	10
5	Conclusion	22
6	References	22

LIST OF FIGURES

Figure No.	Title	Page No.
1	Flowchart of Methodology	8
2	Dashboard for overall region	11
3	Dashboard for west region	11
4	Dashboard for south region	12
5	Dashboard for east region	12
6	Dashboard for central region	13
7	Sales and profit for each sub-category	13
8	Number of orders for each region	14
9	Profit and sales by category	14
10	Number of orders for each category	15
11	Sales and Profit time analysis	15
12	Time analysis of sales	16
13	Categorical profit analysis	16
14	Sales with Positive Profit and Negative Profit	17
15	Number of orders for each Sub Category	17
16	Profit and sales by Region	18

1. ABSTRACT

The "Superstore Sales Data Analysis" uses Python for data preprocessing and Power BI for comprehensive analysis. This project examines sales dynamics using a dataset that contains transactional data, customer, product, and geographic information.

Python is used methodically to clean and transform data, resulting in the creation of a complex data model. Then, using charts and graphs, Power BI's visualization tools create interactive dashboards and reports that visually represent key metrics like sales revenue, profit, and margin.

The Superstore Sales Visualization project, powered by Python and Power BI, provides actionable insights for companies looking to enhance their sales processes and make wise decisions based on a deep understanding of their sales dynamics.

2. INTRODUCTION AND RELATED WORK

Introduction:

This project utilizes the potent features of Power BI for in-depth analysis and seamlessly integrates Python for meticulous data preparation, all in an age where making decisions based on data is essential. The main objective of the project is to use data from a big dataset that contains transactional information, customer profiles, product specifications, and geographical variances to grasp the complexity of Superstore's sales environment. The use of Python ensures a solid foundation for preparing data, setting the stage for a thorough analysis that goes beyond the apparent.

With the deft use of Power BI, the project tells the tale of Superstore's sales journey visually. Power Query enables comprehensive data transformation and cleaning prior to creating an intricate data model.

The end product is a set of interactive dashboards and reports with charts, tables, and graphs that clearly illustrate important parameters like sales revenue, profit, and margin.

To put it simply, the Superstore Sales Visualization and Analysis initiative is a lighthouse for companies looking for practical insights. Through the project's seamless integration of Power BI's visualization skills with Python's data

pretreatment expertise, businesses are better equipped to make educated decisions, fine-tune their sales tactics, and confidently traverse the challenging landscape of contemporary retail.

Related Work:

A thorough understanding of superstore sales analysis and data visualization can be obtained from the literature review, which synthesizes a multitude of sources. Microsoft's "Analyze and visualize Superstore data in Power BI" reference is one particularly valuable tool that highlights the potential of Power BI for in-depth data analysis and visualization. Microsoft's expertise emphasizes the significance of leveraging state-of-the-art technologies to glean valuable insights from intricate datasets. The tutorial, which offers helpful advice on how to use Power BI for efficient data visualization in the context of retail analytics, fits in perfectly with the project's goals because it concentrates on Superstore data.

The significance of sales forecasting for manufacturing, logistics, and retail is emphasized in the Shivankar et al. study. When applied to the Global Superstore sales dataset from Kaggle, their machine learning model performs better than existing algorithms, which is particularly helpful for companies like Big Mart. This project provides informative data that aligns with the increasing importance of predictive analytics in the strategic decision-making process of retail companies. Our understanding of the dynamics of superstore sales as a whole is enhanced by the emphasis on applying machine learning techniques to generate accurate sales predictions.

The Exploratory Data Analysis (EDA) of the Sample Superstore dataset post from the Data Science Blogathon offers helpful hints for visualizing sales data from superstores. The author's use of Pandas, Matplotlib, and Python demonstrates these tools' versatility for manipulating and visualizing data. Targeted advertising is one of the practical solutions for improvement that the EDA offers. It also identifies potential loss areas and areas where discounts could have a negative impact on profits. Visualizations of regional sales, shipping methods, and subcategories offer a comprehensive foundation for effective data visualization in the context of superstore sales.

Critical insights into sales trends and patterns are explored in the Kaggle project on Superstore sales analysis using Python, which contributes to the literature even more. The research examines subcategories, shipping methods, and regional sales dynamics using libraries like Seaborn, Matplotlib, and Pandas. Actionable steps for increasing profitability are suggested, and the analysis highlights the importance of discount strategies. This in-depth investigation is a useful resource that provides applicable methods for improving data visualization strategies related to superstore sales analytics. Altogether, these sources strengthen the body of literature by highlighting industry best practices, cutting-edge visualization methods, and the

incorporation of potent instruments for well-informed decision-making concerning superstore sales analysis.

3. SOFTWARE USED

Google Colab: Collaboratory, also referred to as Google Colab, is a free cloud-based platform that enables group Python program development and execution. It supports real-time code sharing and commenting and is based on the Jupyter Notebook environment. Given that Colab provides customers with powerful hardware accelerators such as GPUs and TPUs, it is a popular choice for machine learning tasks. Users may easily import datasets, work with Google Drive using a browser-based interface, and use a variety of libraries for machine learning and data analysis.

Power BI Desktop: Power BI Desktop is a business analytics tool developed by Microsoft that enables users to build dynamic dashboards and reports. Users are able to transform unstructured data into meaningful knowledge by establishing connections with a range of data sources, including databases, web services, and spreadsheets. Drag-and-drop data visualization features, a wide range of customizable tables, graphs, and charts, and an easy-to-use user interface characterize Power BI Desktop. Users can apply complex data transformations, create connections across different data sets, and share their discoveries with others. Power BI Desktop makes it simple to connect to other Power BI services as part of the broader Power BI suite.

4. METHODOLOGY

In the recommended approach, we analyze a wide range of items from many stores that are spread out over the globe. The market, product categories, country, state, and city perspectives are all taken into consideration while analyzing product sales.

An explanation of the information We leverage the US Superstore sales dataset in our methodology. It consists of tables with the following kinds of columns:

- 'Row ID',
- 'Order ID',
- 'Order Date',
- 'Ship Date',
- 'Ship Mode',
- 'Customer ID',
- 'Customer Name',
- 'Segment',
- 'Country',
- 'City',
- 'State',
- 'Region',
- 'Product ID',
- 'Category',
- 'Sub-Category',
- 'Product Name',
- 'Sales',
- 'Quantity',
- 'Profit',
- 'Returns',
- 'Payment Mode'

The sales analysis method for the Superstore dataset consists of several parts. The following are these steps:

Data collection: The Superstore dataset was obtained through reliable sources like Kaggle.

Preprocessing data in Python: Use Python for comprehensive data preprocessing that ensures data quality and integrity. To ensure that data is compatible with later analytic algorithms, handle null values, eliminate anomalies, and prepare the data.

Metrics selection: The crucial metrics for sales analysis, such as profit and sales, are identified and selected.

Data visualization: The selected metrics are shown using graphs, tables, and charts. Power BI offers a wide range of visualization options, such as bar charts, maps, and more.

Analysis and interpretation: The visuals are analyzed and interpreted in order to provide additional insight about Superstore's sales success. identifying patterns, trends and areas of strength and weakness.

Insights and Recommendations: Derive actionable insights from the analysis, highlighting areas of strength and opportunities for improvement. Provide strategic recommendations based on the identified patterns, enabling informed decision-making for optimizing Superstore's sales performance.

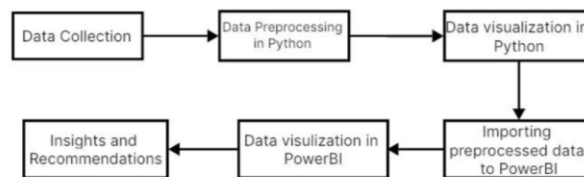


Fig 1: Flowchart Of Methodology

In fig 1, the methodology of sales analysis on Superstore dataset involves several steps, which include:

Data collection: The Superstore dataset is collected from a reliable source, such as Kaggle.

Data Preprocessing with python: Utilize Python for rigorous data preprocessing, ensuring data quality and integrity. Handle missing values, clean outliers, and format data for compatibility with subsequent analysis tools.

Metrics selection: The key metrics for sales analysis, such as sales revenue, profit, etc are identified and selected.

Data visualization: The selected metrics are visualized using charts, tables, and graphs. Power BI provides a variety of visualization options, including bar charts, line charts, and maps.

Analysis and interpretation: The visualizations are analysed and interpreted to gain insights into the sales performance of Superstore. This includes identifying

trends, patterns, and areas of strength and weakness.

LIBRARIES USED IN PYTHON:

- **Pandas-** Pandas is a Python library for data manipulation and analysis. It offers data structures like DataFrame for efficient data handling, cleaning, and analysis. Pandas is widely used for working with structured data.
- **Plotly Express-** Plotly Express is a high-level data visualization library built on Plotly. It enables the creation of interactive and expressive visualizations with a concise syntax. Plotly Express supports various chart types and is known for its ease of use.
- **Matplotlib.pyplot-** Matplotlib.pyplot is a popular 2D plotting library in Python. It provides a MATLAB-like interface for creating static, animated, and interactive visualizations. Matplotlib is highly customizable and is often used in conjunction with other libraries for data visualization.
- **Seaborn-** Seaborn is a statistical data visualization library based on Matplotlib. It simplifies the process of creating informative and attractive statistical graphics. Seaborn comes with built-in themes and color palettes to enhance the visual appeal of plots and is particularly useful for working with complex datasets.
- **CSV (Comma-Separated Values) Library-** CSV is a file format commonly used for tabular data storage. In Python, the built-in csv module provides functionality for reading from and writing to CSV files. It simplifies the process of handling comma-separated data, allowing easy manipulation and interaction with tabular datasets in Python programs.

5. EXPERIMENTAL RESULTS

5.1 Data visualization & Forecasting:

In the realm of data analysis, both Python and Power BI emerge as indispensable tools for effective visualization and forecasting. Python, with its rich libraries like Matplotlib and Seaborn, offers robust capabilities for crafting expressive visuals and leveraging forecasting techniques using libraries such as Statsmodels. On the other hand, Power BI provides a user-friendly interface for seamless data visualization and forecasting, enabling users to create compelling narratives and predict future trends effortlessly.

5.1.1 Data visualization and dashboard creation using PowerBi:

In this comprehensive data visualization and analysis project, four key performance indicators (KPIs) have been identified—namely, orders, average delivery days, sales, and profit. To provide a holistic overview of the dataset, a map has been employed to visualize overall sales

and profit across the country. A slicer enhances user interactivity by allowing categorization of data based on 4 major regions i.e., Central, East, West, and South.

For a nuanced understanding of sales dynamics, three donut charts have been strategically implemented. The first chart reveals the preferred payment mode among various options, offering insights into customer payment preferences. The second chart focuses on segment-wise contributions to overall sales, shedding light on which customer segments are instrumental in driving revenue. The third donut chart identifies the major customer base by region, offering a regional breakdown of customer distribution.

To delve into product-related insights, three clustered charts have been deployed. The first chart illustrates the sales distribution across product categories, providing clarity on which categories contribute most significantly. The second chart narrows down to sub-categories, highlighting specific product types that dominate sales. The third chart focuses on shipping dynamics, elucidating the predominant ship mode chosen by customers for their orders.

To capture temporal trends, two stacked area charts have been applied, offering year-over-year (YoY) insights. The first chart delves into monthly sales performance over time, providing a dynamic visualization of sales trends. The second chart does the same for monthly profit, offering a comprehensive view of profit trends over the specified time period.

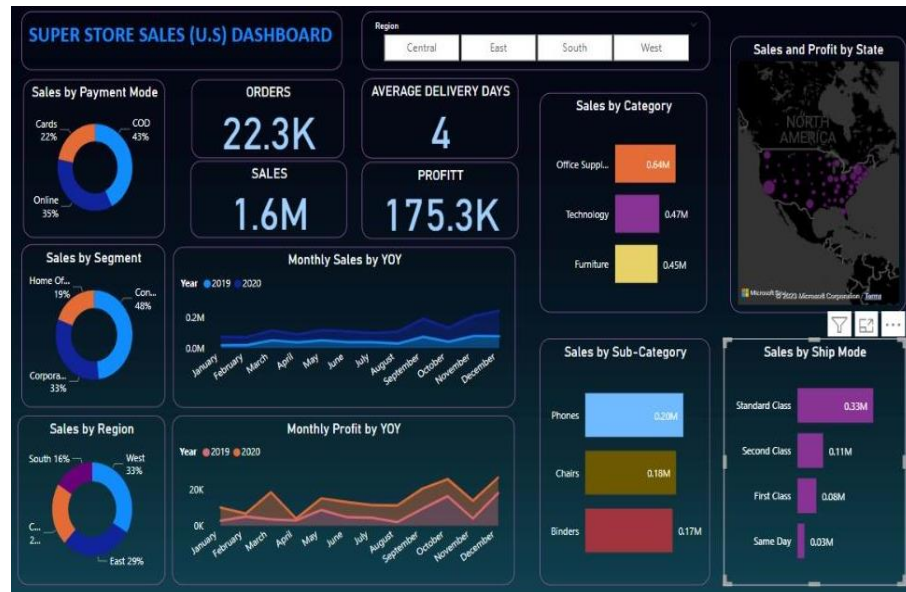


Fig -2 - Dashboard for Overall Region



Fig – 3 - Dashboard for West Region



Fig – 4 - Dashboard for South Region



Fig – 5 -Dashboard for East Region



Fig – 6 -Dashboard for Central Region

5.1.2 Data visualization using Python:

Python serves as a powerful and versatile tool for data visualization in the context of sales analysis. Leveraging libraries such as Matplotlib, Seaborn, and Plotly, businesses can create a wide range of visualizations, including line plots, bar charts, and other interactive visuals, to represent sales trends and patterns. The integration of Pandas facilitates efficient data wrangling, ensuring that sales data is well-organized and prepared for visualization. Jupyter Notebooks provide an interactive environment, enabling the creation of narrative-driven analyses.

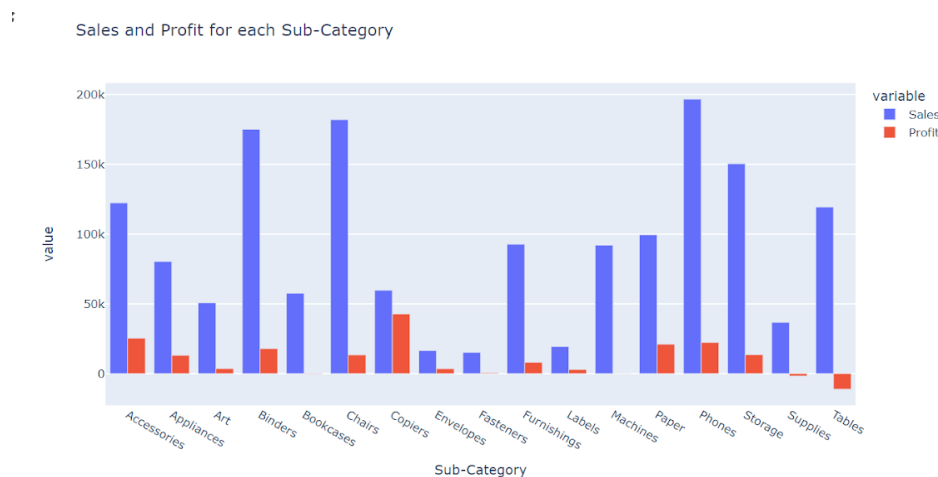


Fig – 7 - Sales and Profit for each sub-category

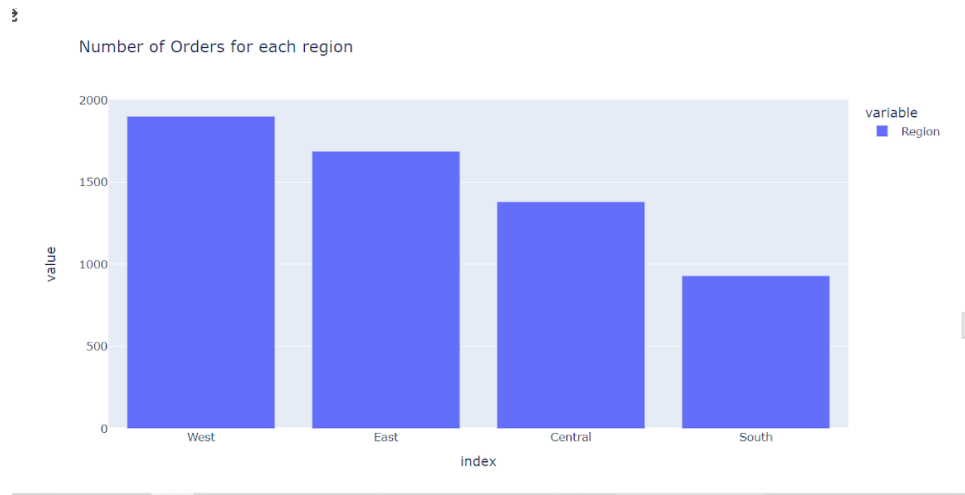


Fig – 8 -Number of orders for each region

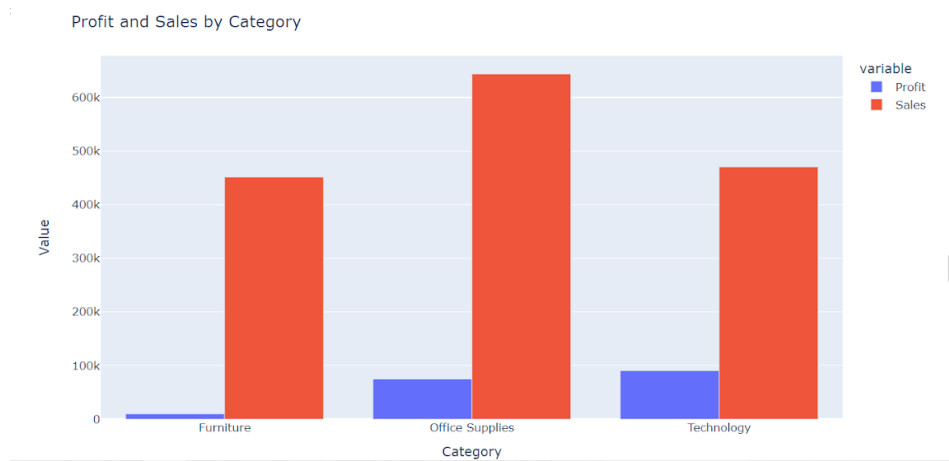


Fig – 9 - Profit and sales by category

Number of Orders for each Category

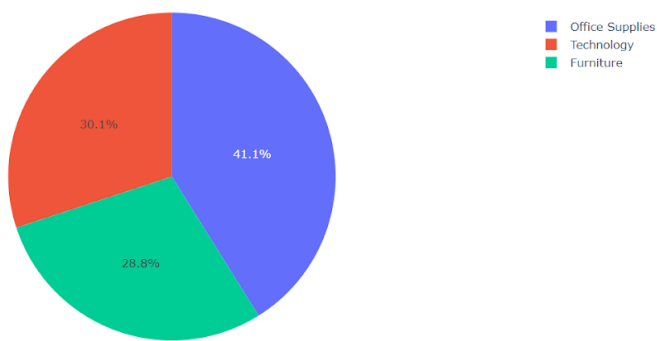


Fig – 10 -Number of Orders for each Category

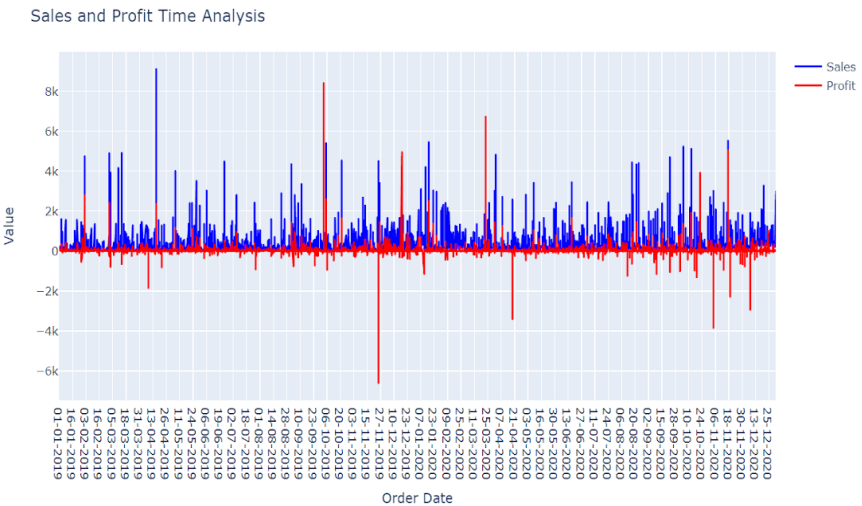


Fig – 11 -Sales and Profit Time Analysis

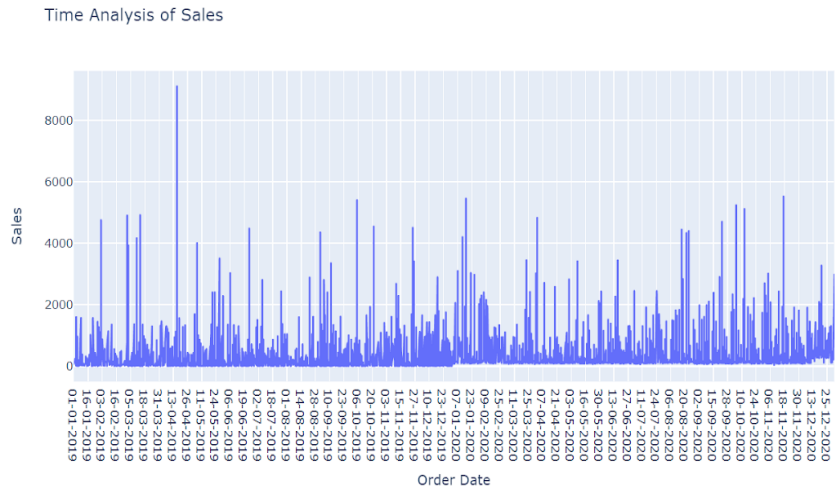


Fig – 12 -Time Analysis of Sales

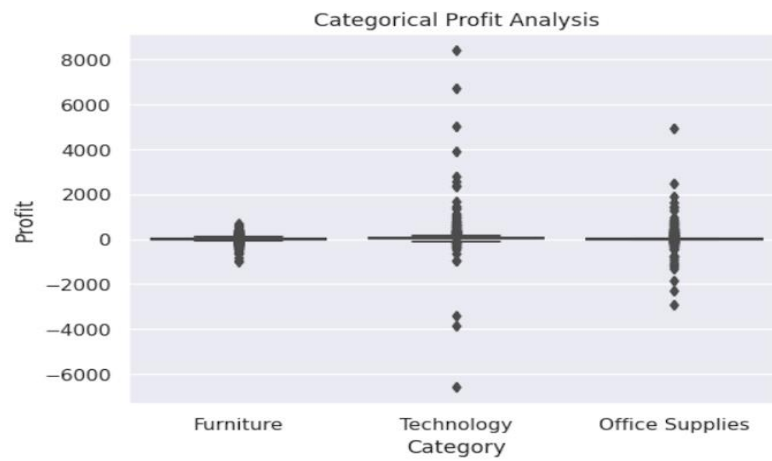


Fig – 13 -Categorical Profit Analysis

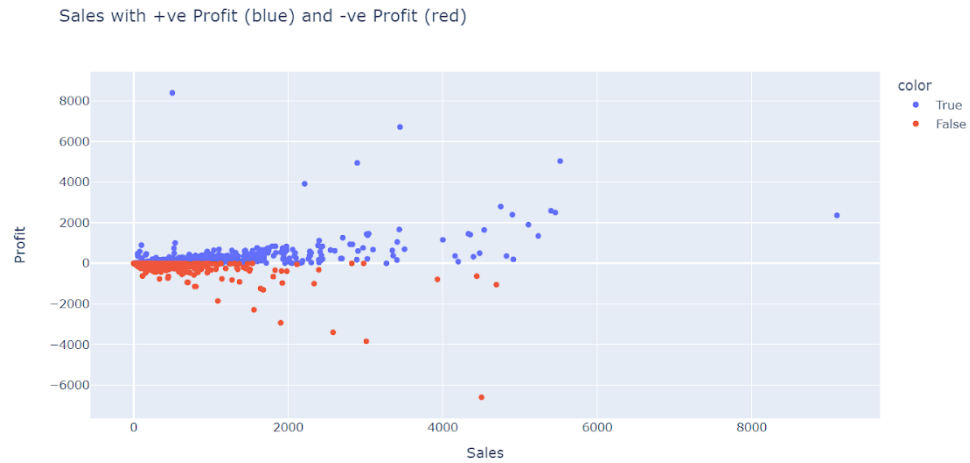


Fig – 14 -Sales with Positive Profit and Negative Profit

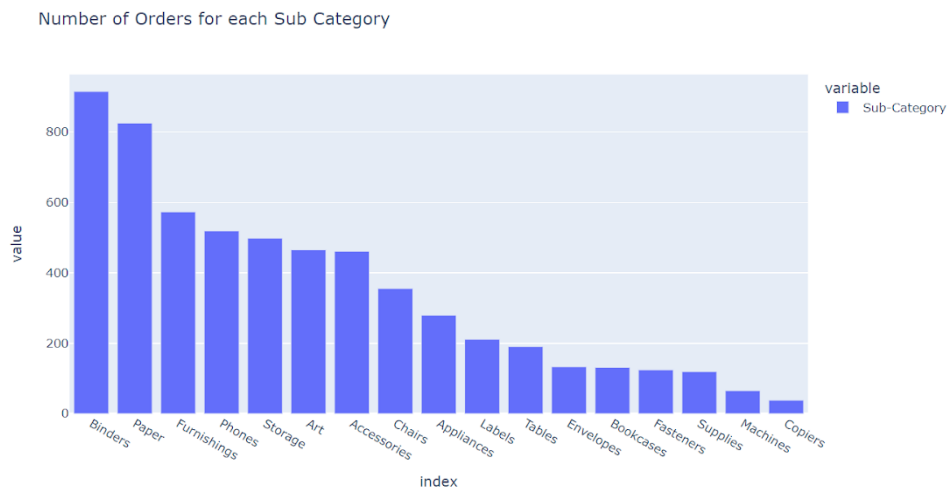


Fig – 15 -Number of Orders for each sub category



Fig – 16 -Profit and sales by Region

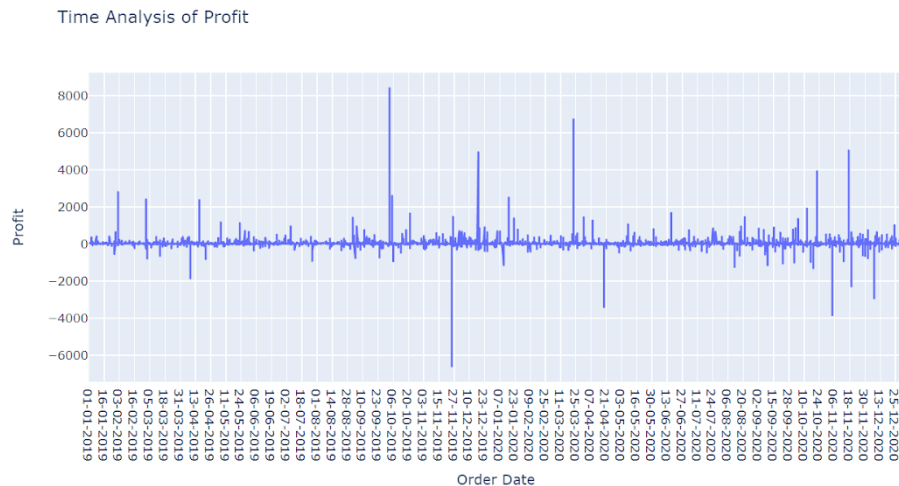


Fig – 17 -Time Analysis of Profit

5.1.3 Forecasting using Power BI:

In the given figure(fig 18), a line chart has been employed in Power BI, visualizing the relationship between sales and order dates. This line chart serves as the foundation for forecasting future sales over the next 20 days. The forecasting feature in Power BI is activated with a confidence level set at 95%, indicating the level of certainty in the predicted sales values. This confidence level parameter is flexible and can be adjusted based on the specific requirements and confidence thresholds desired by the user.

The forecasting algorithm, embedded in the line chart, employs sophisticated methods such as exponential smoothing or ARIMA to project future sales trends based on historical patterns. The confidence interval, set at 95%, establishes a range within which the actual sales values are likely to fall.

To enhance user interactivity and focus on specific periods of interest, a zoom slider has been incorporated. This feature allows users to dynamically adjust the time range displayed on the line chart. With the zoom slider, stakeholders can narrow down their focus to a particular time frame, facilitating a more detailed examination of the forecasted sales trends. The blue lines on the chart represent historical data, offering a contextual understanding of past sales performance. In contrast, the red lines symbolize the forecast, indicating the predicted trajectory of sales over the specified period.

This combination of a line chart, forecasting functionality, and a zoom slider in Power BI provides a comprehensive and dynamic platform for analyzing historical sales data, predicting future trends, and allowing users to interactively explore and understand the forecasted sales patterns with a high level of confidence.

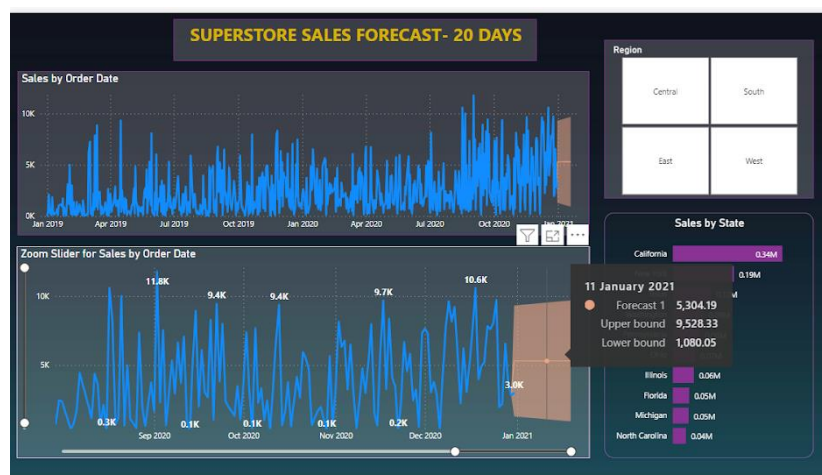


Fig – 18 - Forecasting Sales by Order Date

5.2 INSIGHTS:

5.2.1 From PowerBI

1. Preferred Payment Method:

Cash on Delivery (COD) emerges as the most favored payment method across the country, surpassing other payment options.

2. Consumer Dominance:

The consumer segment stands out as the primary contributor to overall sales, demonstrating its significant impact on revenue.

3. Major Customer Concentration:

The West region of the country commands the largest customer base, showcasing a substantial market presence.

4. Best-Selling Category:

"Office Supplies" emerges as the top-selling category nationwide, indicating a consistent demand for office-related products.

5. Top-Performing Sub-Category:

Within product sub-categories, "Phone" emerges as the leading choice among consumers, boasting the highest sales figures across all regions.

6. Preferred Shipping Mode:

"Standard Class" proves to be the most preferred shipping mode, highlighting the prominence of standard delivery options among customers.

7. Order, Sales, and Profit Overview:

Nationally, a total of 22.3k orders have been placed, resulting in a commendable overall sales figure of 1.6m and a profitable outcome of 175.3k.

8. Peak Sales Months:

The months of November and December in both 2019 and 2020 witness the highest sales activity, indicating a seasonal trend in consumer behavior.

9. Peak Profit Months:

Maximum profit is achieved during October and December in both 2019 and 2020, with an additional peak in March 2020.

10. Regional Profit Dynamics:

While the overall country reports no loss in profit, specific regions reveal isolated instances of loss. Notably, the East region in November 2019 (*Ref Fig - 5*), South region in October 2020 (*Ref Fig - 4*), and West region in April 2020 (*Ref Fig - 3*) experienced temporary setbacks.

5.2.2 From Python

1. Max Sales:

- a. 18/04/19 and 17/11/20 recorded the highest sales.

2. Max Loss and Profit:

- a. Max loss occurred on 25/11/19.
 - b. Max profit was achieved on 16/04/19 for both years.
- 3. Product Categories:
 - a. Office supplies dominated sales across all categories.
 - b. Technology category led in terms of maximum profit.
- 4. Regional Impact:
 - a. The West region emerged as the top contributor for both sales and profits.
- 5. Top Sub-Category Insights:
 - a. Binders witnessed the maximum number of orders among all sub-categories.
 - b. Although phones had the highest sales, copiers generated the maximum profit among all sub-categories.

6. CONCLUSION

The analysis reveals a clear pattern in consumer behavior and regional dynamics. Cash on Delivery (COD) remains the preferred payment method, emphasizing the importance of flexible payment options. The dominance of the consumer segment underscores its pivotal role in driving overall sales and revenue. The West region stands out as a major market hub, requiring strategic focus and tailored approaches to capitalize on its substantial customer base.

The popularity of "Office Supplies" highlights a consistent demand for workplace-related products, offering a prime opportunity for sustained sales growth. Simultaneously, the success of the "Phone" sub-category indicates a strong consumer preference for tech-related items. While the overall country reports no loss in profit, isolated setbacks in specific regions underline the importance of region-specific strategies to mitigate potential challenges.

To optimize overall performance, businesses should leverage the preference for standard shipping and strategically plan marketing and inventory management around the peak sales and profit months. By understanding these insights and recommendation, businesses can refine their strategies, enhance customer experiences, and capitalize on key market trends for sustained success.

7. REFERENCES

1. Microsoft. (n.d.). Analyse and visualize Superstore data in Power BI:
<https://community.fabric.microsoft.com/t5/Data-Stories-Gallery/Superstore-Data-Visualization-with-Power-BI/td-p/2977837>
 2. SALES ANALYSIS ON SUPERSTORE DATASET Abdul Razzaque*1, Anupam Nimawat*2, Burhanuddin Khamgaonwala*3, Danish Sheikh*4, Lokesh Yadav
https://www.irjmets.com/uploadedfiles/paper//issue_4_april_2023/36572/final/fin_irjmets1682186035.pdf
 3. Global Superstores Sales Prediction and Data Visualization Using Power BISHruti Shivankar, Shardul Mehetar, Neha Darade ,Saachi Bhimanpalli, Dnyanada Dafale
<https://journal.ijresm.com/index.php/ijresm/article/view/2670/2647>
 4. <https://www.analyticsvidhya.com/blog/2022/03/eda-on-superstore-dataset-using-python/>
 5. <https://www.kaggle.com/code/kamilatul/superstore-sales-analysis-in-python>
 6. <https://www.kaggle.com/code/arloblanco/superstore-sales-forecasting-in-python/notebook>
-
