KLE Society's

KLE Technological University



**Exploratory Data Analysis**

**(22ECAC210)**

**Course Project Report on**

**“Superstore Data Analysis and Profit Prediction”**

*Submitted in partial fulfilment of the requirement for the award of*

**Degree of Bachelor of Engineering**

**in**

**Computer Science and Engineering**

*Submitted By*

|  |  |
| --- | --- |
| **Name** | **SRN** |
| **Debanshu Behera** | **01FE21BCI035** |
| **Vishesh Singh** | **01FE21BCI036** |
| **Manglam** | **01FE21BCI041** |
| **Devaj** | **01FE21BCI055** |
| **Gayatri Shenvi** | **01FE21BCI062** |

*Submitted To*

**SCHOOL OF COMPUTER SCIENCE & ENGINEERING,**

KLE TECHNOLOGICAL UNIVERSITY

HUBLI – 580 031 (India).

**Abstract**

The present data analysis project delves into a comprehensive exploration of a vast superstore dataset, aiming to uncover valuable insights into sales and profit trends. The dataset encompasses a wealth of attributes, such as sales figures, profit margins, order dates, ship dates, and more, enabling an in-depth investigation of the store's performance over a significant period.

The primary objective of this study is to extract meaningful patterns and correlations from the data to aid business decision-making processes. The analysis begins with data preprocessing, encompassing data cleaning, outlier detection, and handling missing values to ensure the reliability of subsequent findings.

Subsequently, exploratory data analysis techniques are applied to reveal an overview of the store's sales and profit trends over time. Visualizations such as line plots, bar charts, and heatmaps provide an intuitive understanding of patterns, seasonal variations, and potential factors influencing sales and profitability.

Furthermore, the project employs advanced statistical methods and machine learning algorithms to build predictive models for sales and profit forecasting. These models take into account various factors such as seasonal effects, promotional events, and economic indicators to provide accurate and reliable predictions.

The results of this data analysis endeavour have the potential to offer valuable insights into the store's performance, identify areas for optimization, and guide strategic decision-making. By leveraging the power of data, this project contributes to enhancing the store's efficiency, competitiveness, and overall success in the dynamic retail landscape.

**Table of Contents:**

**1. Introduction ----------------------------------------------------------------4**

1.1 Overview of superstore data analysis. --------------------------------------------4

1.2 Importance of EDA in data analysis. -----------------------------------------------4

1.3 Objectives of the course project. ---------------------------------------------------5

**2. Data Collection -------------------------------------------------------------6**

2.1 Dataset description. --------------------------------------------------------------------6

2.2 Data source. ------------------------------------------------------------------------------7

2.3 Data preprocessing steps. -------------------------------------------------------------7

**3. Data Exploration -----------------------------------------------------------8**

3.1 Descriptive statistics of the dataset. ------------------------------------------------8

3.2 Data visualization. -----------------------------------------------------------------------9

**4. Data Cleaning --------------------------------------------------------------14**

4.1 Steps taken to clean the dataset. --------------------------------------------------14

4.2 Data transformation and normalization. -----------------------------------------15

**5. Feature Selection ----------------------------------------------------------16**

5.1 Process of selecting relevant features. --------------------------------------------16

5.2 Rationale behind the feature selection process. --------------------------------16

**6. Data Analysis ----------------------------------------------------------------17**

6.1 Statistical analysis on the data and meaningful conclusion. ------------------17

**7. Insights and Findings ------------------------------------------------------21**

7.1 Key insights gained from the analysis. ----------------------------------------------21

7.2 Challenges encountered. ---------------------------------------------------------------21

**8. Recommendations ---------------------------------------------------------22**

8.1 Recommendations based on the findings. -----------------------------------------22

**9. Conclusion --------------------------------------------------------------------23**

9.1 The value of EDA in understanding the dataset. ----------------------------------24

**10. References ------------------------------------------------------------------ 25**

**CHAPTER 1: INTRODUCTION**

1.1 Overview of Superstore Data Analysis:

The superstore data analysis project aims to gain meaningful insights from a large dataset related to a retail superstore's sales and profit. The dataset contains several attributes, including sales, profit, order date, ship date, and more. By analysing this dataset, the project seeks to uncover patterns, trends, and factors influencing the store's performance to facilitate informed decision-making and drive business growth.

1.2 Importance of Exploratory Data Analysis (EDA) in Data Analysis:

* **Exploratory Data Analysis** (EDA) plays a pivotal role in the data analysis process. It involves the initial exploration and visualization of data to understand its underlying structure and characteristics. EDA serves several crucial purposes:
* **Data Understanding**: EDA helps analysts get a comprehensive grasp of the data, its distribution, and key features. It aids in identifying potential errors or outliers that might impact subsequent analyses.
* **Pattern Recognition**: Through visualizations and summary statistics, EDA reveals hidden patterns and relationships within the data, leading to valuable insights and hypotheses.
* **Feature Selection**: EDA assists in identifying relevant features and variables for further analysis, filtering out noise and redundant information.
* **Data Preprocessing**: EDA guides data cleaning and preparation steps, ensuring data integrity and improving the quality of the final analysis.

1.3 Objectives of the Course Project:

The primary objectives of the superstore data analysis project are as follows:

**Sales and Profit Trend Analysis**: Conduct a thorough exploration of sales and profit trends over time to identify patterns, seasonal variations, and potential growth opportunities.

**Identify Key Factors**: Determine the factors that significantly influence sales and profit. This may include factors like promotional events, regional influences, product categories, and more.

**Predictive Modelling**: Build predictive models using machine learning algorithms to forecast future sales and profit. These models will help the superstore plan inventory, manage resources, and optimize pricing strategies.

**Performance Evaluation**: Evaluate the store's performance by analysing key performance indicators (KPIs) derived from the dataset. This assessment will guide the store's strategic decision-making.

**Recommendations:** Provide actionable insights and recommendations based on the analysis, aiming to enhance the store's efficiency, profitability, and overall competitiveness in the market.

**CHAPTER 2: DATA COLLECTION**

2.1 Dataset Description:

The dataset used in the superstore project is a comprehensive collection of sales and profit data from a retail superstore. It contains a wide range of attributes that provide valuable insights into the store's performance, customer behaviour, and product trends. Below is a detailed description of the dataset attributes:

**Row ID**: A unique identifier for each row in the dataset.

**Order ID**: A unique identifier for each customer order, linking multiple items purchased by a single customer.

**Order Date**: The date when the customer placed the order for the products.

**Ship Date**: The date when the ordered products were shipped to the customer.

**Ship Mode**: The mode of shipping chosen by the customer (e.g., standard, express, etc.).

**Customer ID**: A unique identifier for each customer, helping to track individual customer behaviour.

**Customer Name**: The name of the customer who placed the order.

**Segment**: The segment to which the customer belongs, categorizing them as corporate, home office, or consumer.

**Country**: The country of residence of the customer.

**City**: The city of residence of the customer.

**State**: The state of residence of the customer.

**Postal Code**: The postal code of the customer's location, providing regional information.

**Region**: The region where the customer belongs (e.g., North, South, East, West).

**Product ID**: A unique identifier for each product in the store.

**Category**: The category to which the product belongs, such as office supplies, furniture, or technology.

**Sub-Category**: The specific sub-category of the product (e.g., chairs, phones, paper).

**Product Name**: The name of the product.

**Sales**: The total sales value of the product.

**Quantity**: The quantity of the product ordered.

**Discount**: The discount provided for the product.

**Profit**: The profit or loss incurred from the sale of the product.

2.2 DATA SOURCE:

The data collection process for the superstore dataset, sourced from Kaggle.com, was a meticulous and comprehensive endeavour aimed at capturing a vast array of information related to retail sales and profit.

Link: <https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>

2.3 DATA PREPROCESSING STEPS:

* Checking for NULL values.
* Checking for duplicate tuple and removing them.
* Splitting the order\_date and Ship\_date columns into year, month and date of week.
* To find difference between order date and ship date.
* Checking for outliers and removing them.

**CHAPTER 3: DATA EXPLORATION**

3.1 Descriptive statistics of the dataset:

Table 1

Table 1, shows the count, mean, standard deviation, minimum value and the values which are above 25%, 50% and 75% and also the maximum value of each column with datatype int64

3.2 Data visualization:  Fig. 01

From Fig. 01, we observe that Technology category gives the highest profit and Furniture category gives least profit.

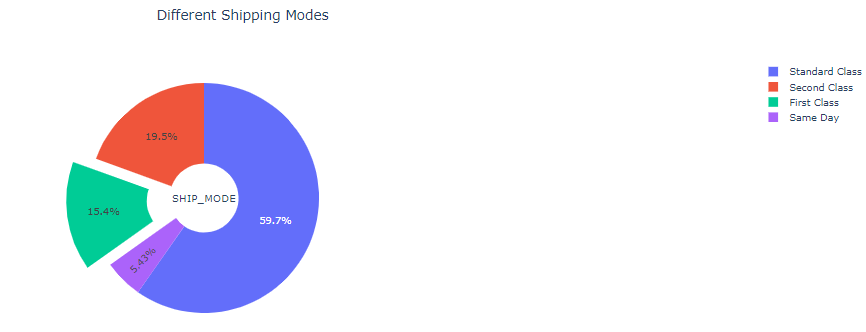


Fig. 02

From fig. 02, From this pie chart we come to know that majority of the products have been shipped through ‘Standard Class’ probably because it is the cheapest option.

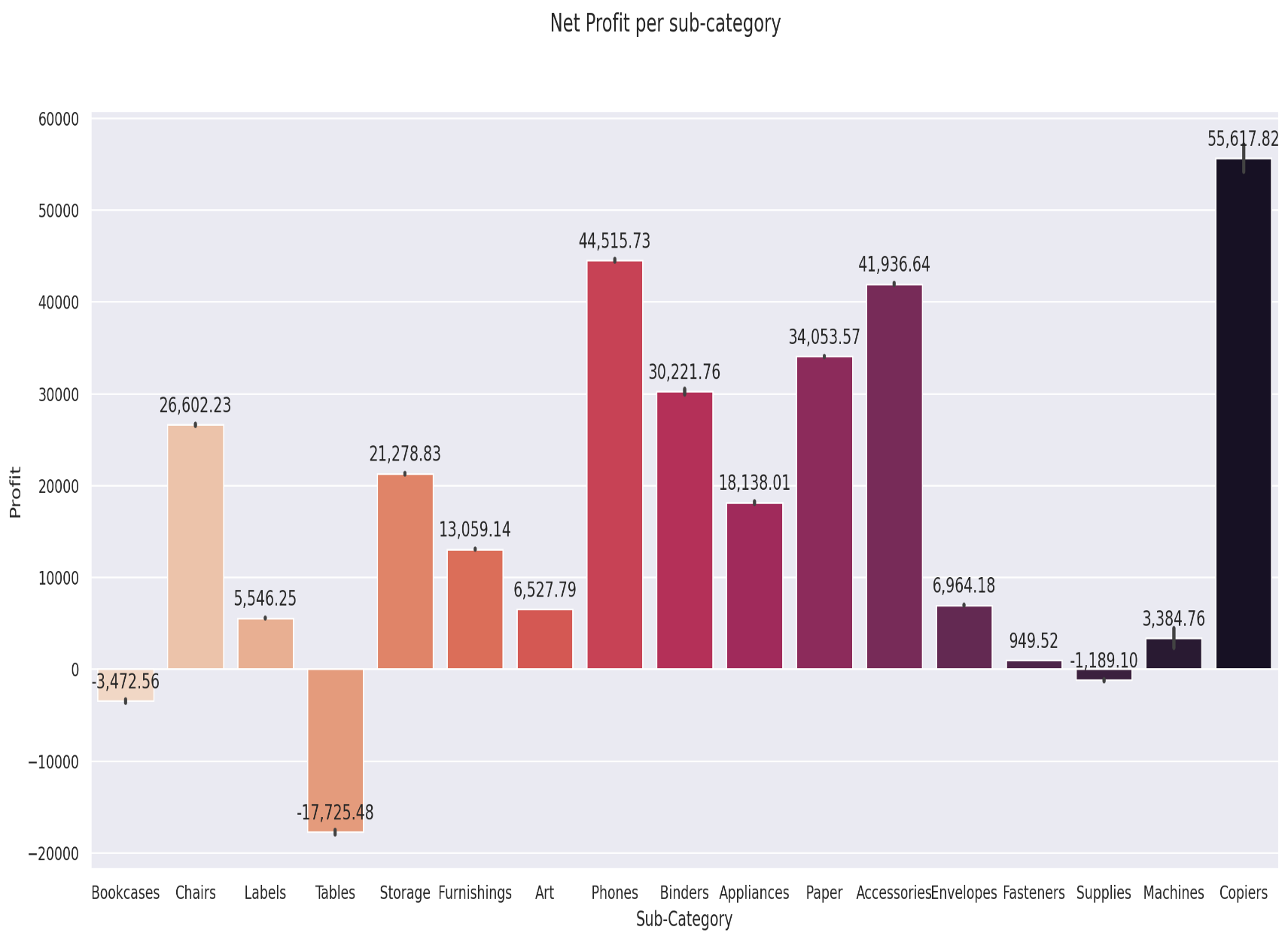


Fig. 03

From Fig. 03, From this graph we get to know that Copiers provide the most profit and Tables contribute to a huge loss. From the first graph we came to know that furniture provides the least profit. Now we know which furniture contributed towards that. So, the owner might need to avoid selling Tables.

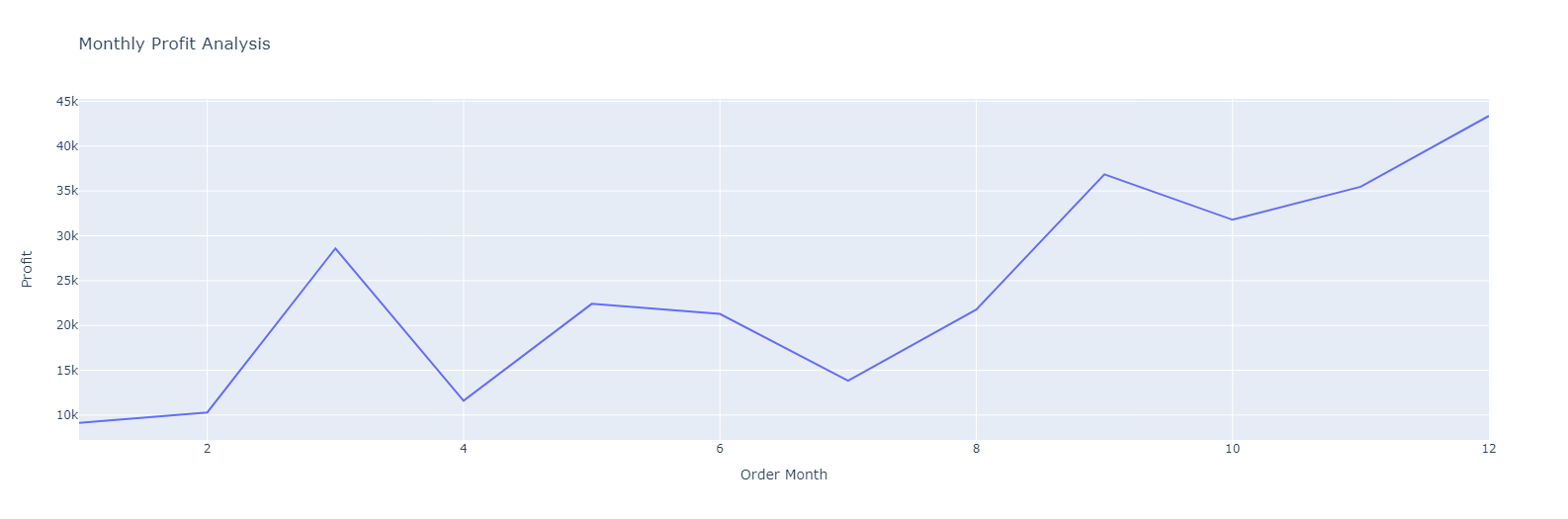


Fig. 04

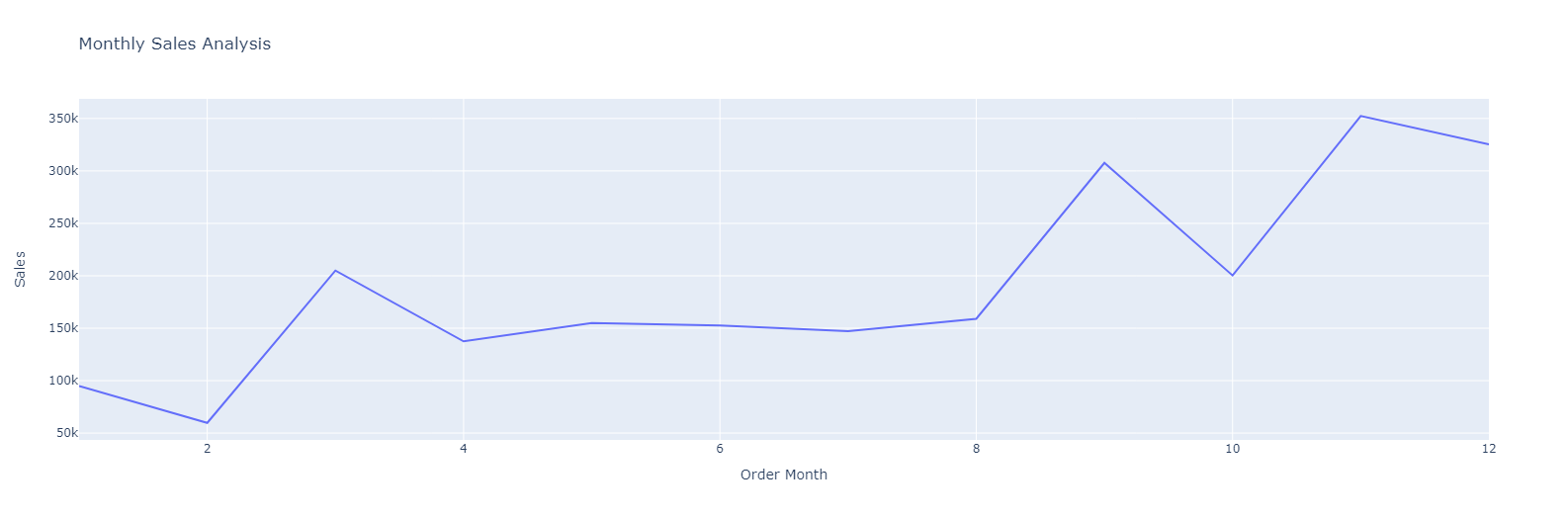


Fig. 05

* From Fig. 04 and Fig. 05 line plots we get to know that in the months of November and December the sales and profit were maximum probably because it is Christmas at that time during which many people shop due to sales and discount.
* Also, during the months of January and February the sales and profit are minimum probably because people have already shopped in November and December.
* This information might help the store owner to keep adequate supplies of products during November and December.



Fig. 06

Fig. 06 indicates the boxplot profit where there are many outliers.

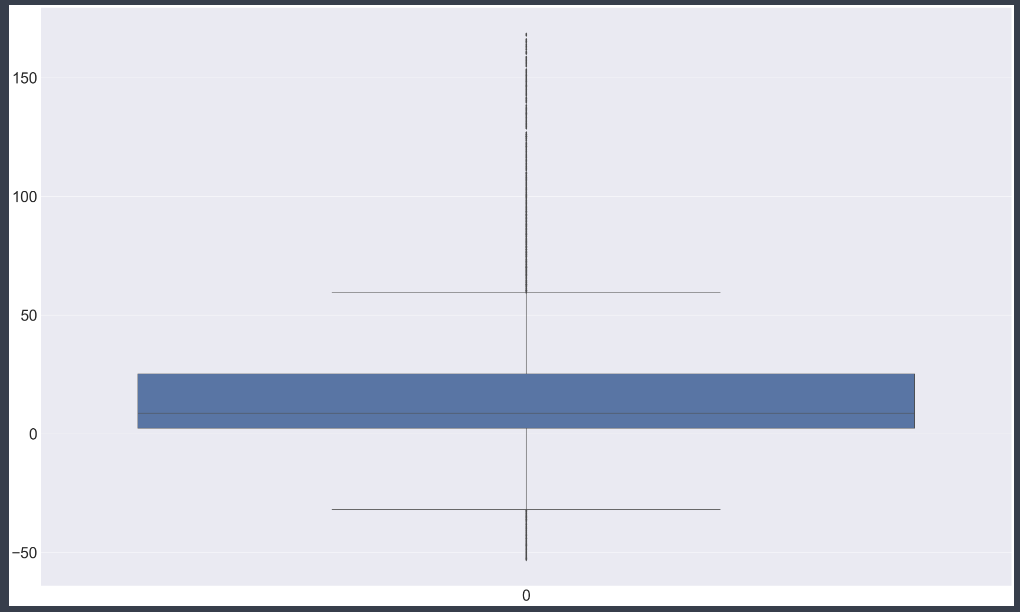


Fig. 07

Fig. 07 is the boxplot after the removal of the outliers.

**CHAPTER 4: DATA CLEANING**

4.1 Steps taken to clean the data:

* Handling Missing Values: Identify and handle any missing values in the dataset. Missing data can lead to inaccuracies and biased results. Depending on the attribute, missing values can be replaced with mean, median, mode, or dropped if the missing values are excessive.
* Dealing with Duplicates: Check for and remove any duplicate rows in the dataset. Duplicate entries can skew the analysis and affect the accuracy of the results.
* Outlier Detection and Treatment: Identify outliers, which are data points significantly different from the majority. Decide on whether to keep, remove, or transform outliers based on their impact on the analysis.
* Data Formatting: Standardize the format of data to ensure consistency. For example, dates should follow a consistent format, and numerical values should have the same number of decimal places.
* Feature Engineering: Create new relevant features from existing ones if needed. For example, extracting the month and year from the order date can be beneficial for analysing seasonal sales trends.
* Data Scaling and Normalization: While using machine learning algorithms, scale or normalize numerical data to bring them to a common scale, preventing any attribute from dominating the analysis.
* Final Data Review: After applying all the cleaning steps, perform a final review of the data to ensure that it is clean, consistent, and ready for analysis.

4.2 DATA TRANSFORMATION AND NORMALIZATION

* Feature Scaling: Feature scaling involves scaling individual features to a specific range to ensure they have similar magnitudes. It is essential when using algorithms sensitive to the scale of features, such as distance-based algorithms.
* Categorical Variable Encoding: Categorical variable "Segment" and "Category” need to be encoded into numerical representations for machine learning algorithms. Common techniques include one-hot encoding and label encoding.
* Date Feature Engineering: Extracting relevant information from date attributes like "Order Date" can be beneficial for time series analysis. For example, creating new features like "Month," "Quarter," or "Day of the Week" can capture seasonal patterns in sales data.
* Removing Irrelevant Features: Certain features have no significant impact on the analysis, they can be removed to reduce complexity and improve computational efficiency.

**CHAPTER 5: FEATURE SELECTION**

5.1 Process of selecting relevant features:

* Understanding the Business Problem: Gain a clear understanding of the business problem or objectives of the analysis. Identify the key questions that need to be answered, such as factors affecting sales growth, profitability drivers, or customer segmentation.
* Exploratory Data Analysis (EDA): Conduct EDA to explore the relationships between different features and the target variables (sales and profit). Use visualizations like scatter plots, bar charts, and heatmaps to identify potential correlations and patterns.
* Correlation Analysis: Calculate correlation coefficients between numerical features and the target variables. Features with high positive or negative correlations are more likely to be relevant in influencing sales and profit.

5.2 Rationale Behind the feature selection process:

* Simplification of the Model: By selecting only the most informative features, the complexity of the model is reduced. This helps to improve model interpretability and reduces the risk of overfitting, where the model performs well on the training data but poorly on unseen data.
* Enhanced Model Performance: Including irrelevant or redundant features can introduce noise and negatively impact model performance. Selecting relevant features improves the model's predictive accuracy by focusing on the most critical variables.
* Resource Efficiency: Working with a reduced set of relevant features can lead to faster computations and lower resource requirements during model training and evaluation.

**CHAPTER 6: DATA ANALYSIS**

6.1 Statistical analysis of the dataset and meaningful conclusions:

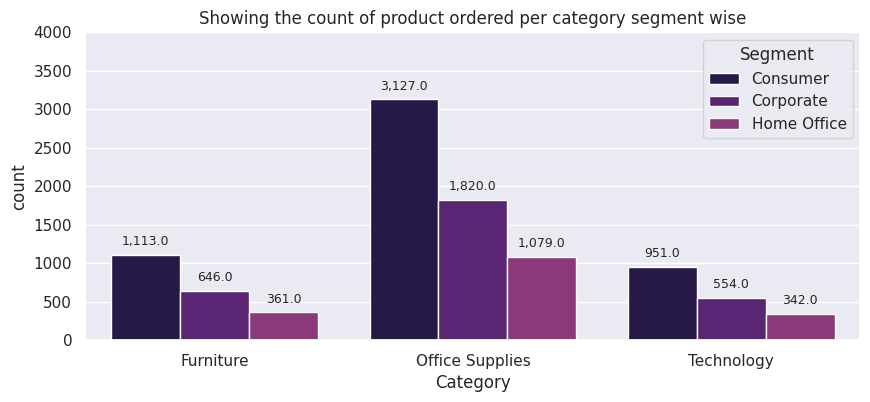


Fig. 08

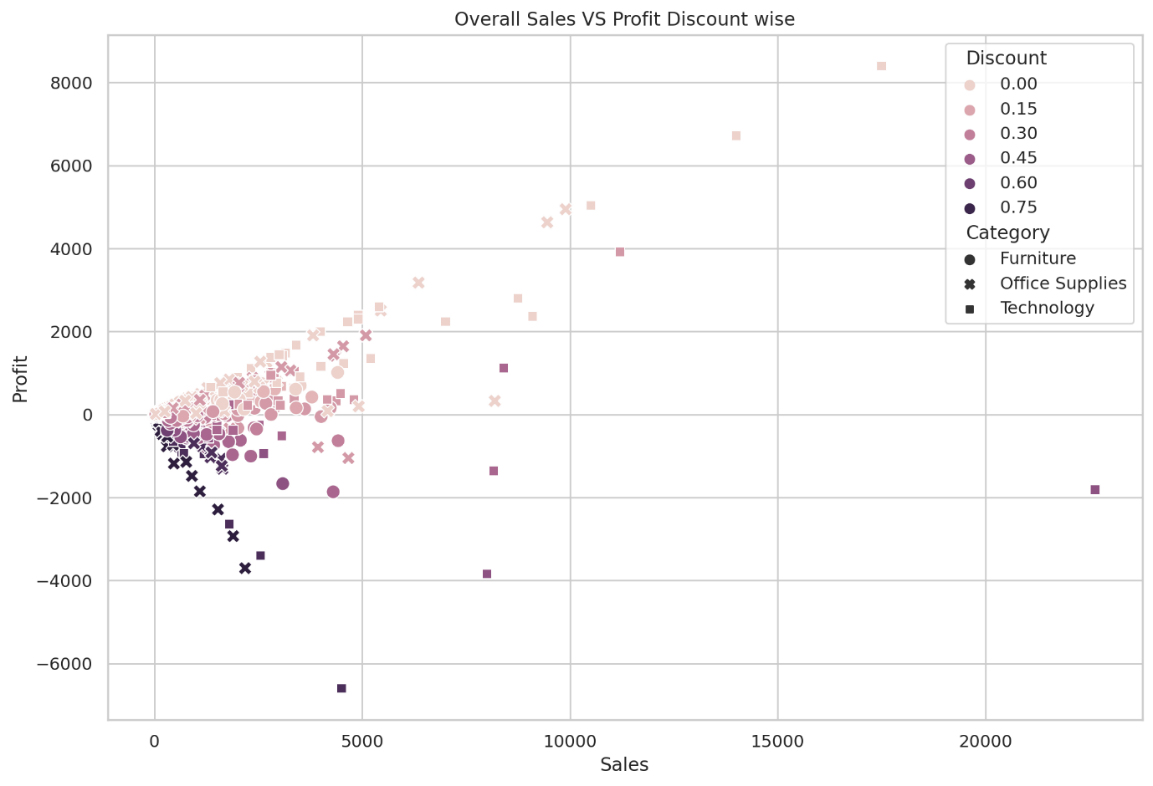


Fig. 09

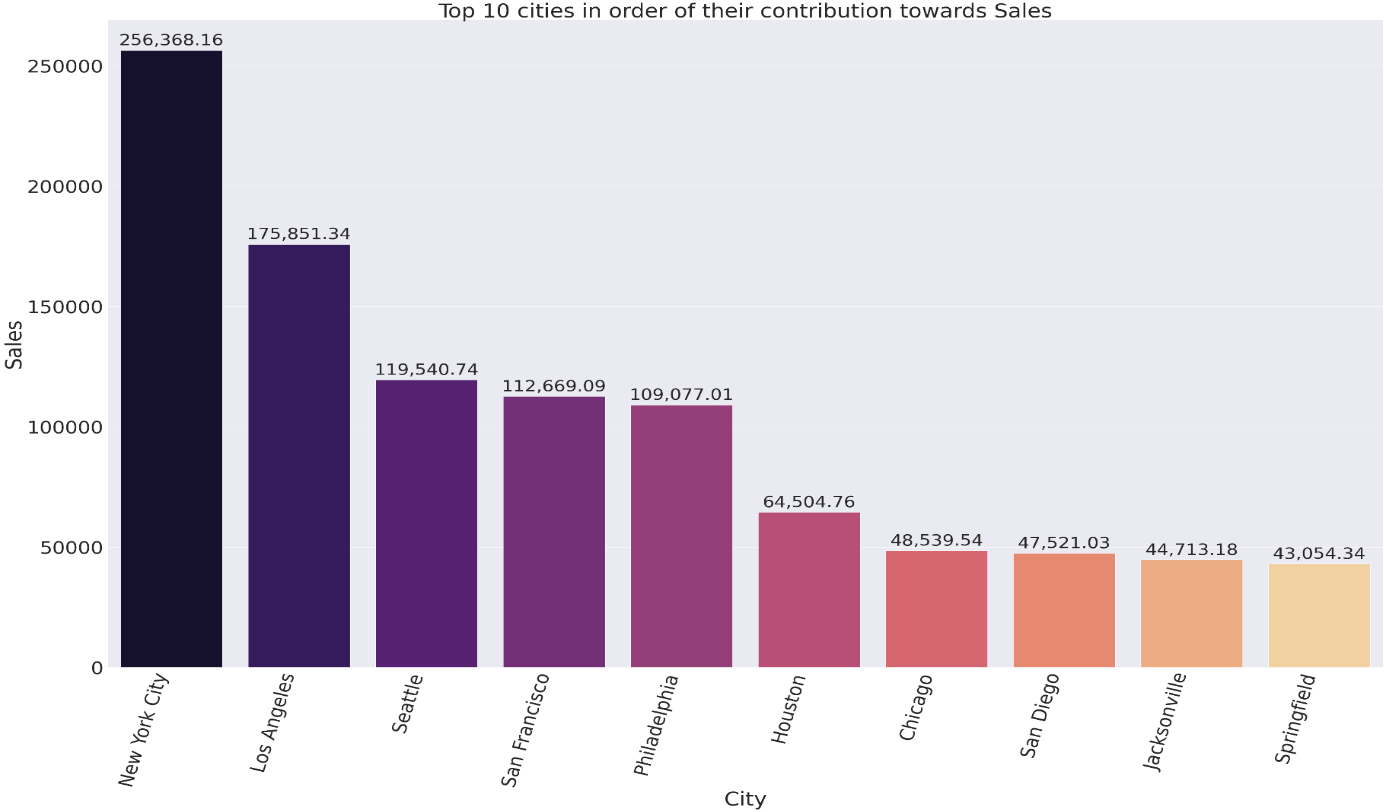


Fig. 10

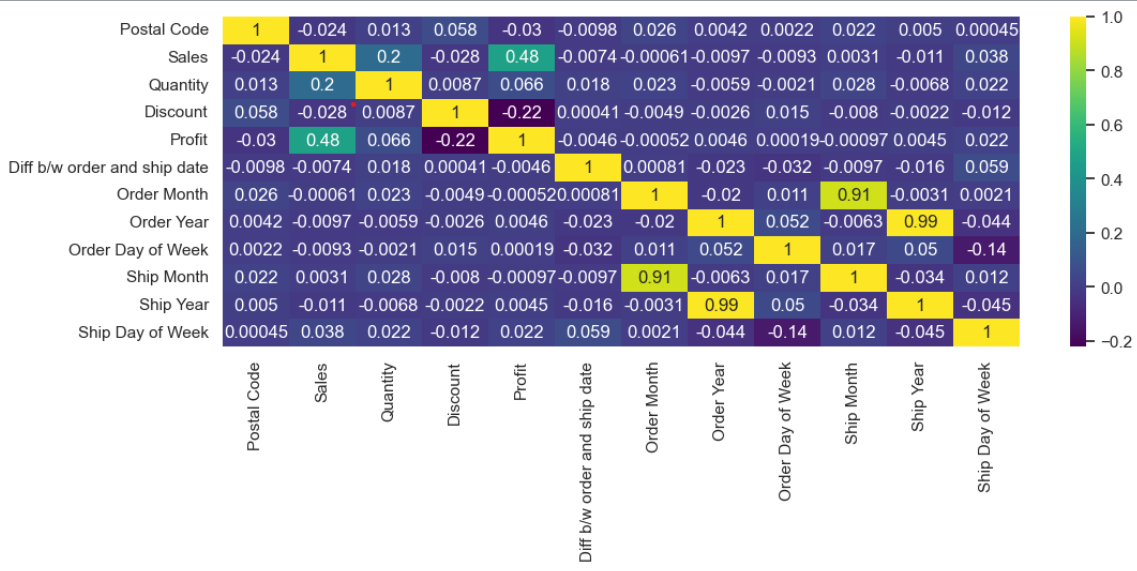


Fig. 11

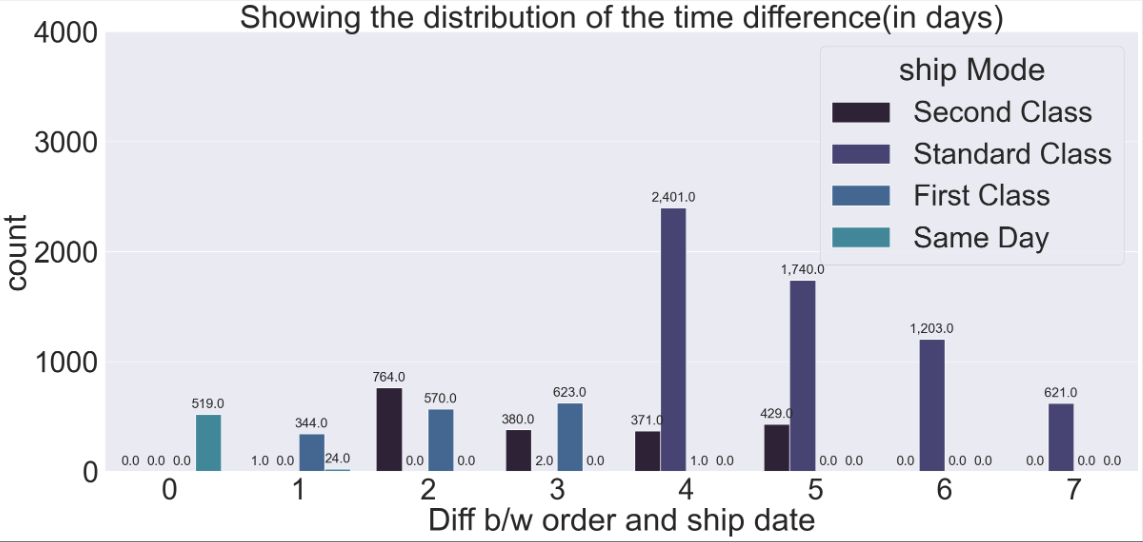


Fig. 12

Conclusions:

* From Fig. 08, we get to know that the consumer segment contributes most towards the number of products bought followed by corporate segment and home office segment in all three categories.
* From Fig. 09, This scatter plot provides us with great insights about the profit, sales, discount and category. For instance from this scatter plot we get to know that in technology category, discount offered is least and profit gained is highest.
* From Fig. 10, This plot gives us the top 10 cities in order of their contribution towards sales and New York City contributed the most towards the sales
* From Fig. 11, From the correlation heatmap we observe that Sales, Quantity and Discount are the features which impact Profit the most. So we select these three features.
* From Fig. 12, The count plot shows that standard class shipping mode took maximum number of days for shipment of the products.

**CHAPTER 7: INSIGHTS AND FINDINGS**

7.1 Key insights gained from the analysis:

* Top Selling Categories and Products: The analysis identified the top-performing product categories and specific products that contributed significantly to sales and profitability. This information can guide the superstore's marketing and product stocking strategies.
* Profitability by Region: The analysis showed variations in profitability across different regions. Understanding the reasons behind these differences can help the superstore allocate resources and tailor strategies for each region.
* Customer Segmentation: Customer segmentation analysis identified distinct customer groups based on their purchasing behaviour and preferences. This insight can be used for targeted marketing campaigns and personalized customer experiences.
* Shipping Mode Effectiveness: The analysis assessed the effectiveness of different shipping modes on customer satisfaction and profit. Understanding the most preferred shipping modes can optimize logistics and enhance customer experience.
* Predictive Sales and Profit Models: Built predictive models provided forecasts for future sales and profit based on historical data and relevant factors. These models can aid in business planning and decision-making.

7.2 Challenges encountered:

The dataset is not linear. So, we can’t apply linear regression model as it was giving large value of mean squared error. Hence, we used various classification algorithm to predict the profit.

**CHAPTER 8: RECOMMENDATIONS**

8.1 Recommendations based on the findings:

* Product Focus: Focus on the top-selling product categories and specific products identified in the analysis. Ensure adequate stock levels for these products to meet customer demand and optimize profit margins.
* Regional Optimization: Analyse the profitability of different regions and tailor marketing strategies and product offerings to cater to regional preferences and needs. Optimize shipping and logistics operations to reduce costs and improve customer satisfaction.
* Discount Optimization: Based on the impact of discounts on sales and profit, optimize the discount strategies. Experiment with different discount levels and timings to find the right balance that maximizes sales and profitability.
* Customer Segmentation: Leverage customer segmentation insights to create personalized marketing strategies for different customer groups. Offer targeted promotions and product recommendations to enhance customer loyalty and satisfaction.
* Efficient Shipping Modes: Identify the most preferred shipping modes to improve shipping efficiency and reduce costs. Offer incentives for customers to choose cost-effective shipping options.
* Predictive Models for Inventory Planning: Use the predictive sales and profit models to forecast future demand accurately. This can help in efficient inventory planning and prevent stockouts or overstock situations.
* Underperforming Product Analysis: Identify underperforming products or categories and conduct further analysis to understand the reasons behind their lower sales or profitability. Consider revising pricing or marketing strategies or discontinuing products with consistently poor performance.

**CHAPTER 9: CONCLUSION**

* Dataset Description: The superstore dataset contains information on sales, profit, order dates, ship dates, customer details, product categories, and more.
* Data Analysis Objective: The primary objective of the data analysis is to gain insights into sales and profit trends, identify influential factors, build predictive models, evaluate performance, and provide actionable recommendations to improve the superstore's efficiency and profitability.
* Data Preprocessing: The data preprocessing phase involves handling missing values, duplicates, outliers, and data formatting to ensure data integrity and reliability.
* Exploratory Data Analysis (EDA): EDA techniques are applied to explore relationships, patterns, and correlations in the data through visualizations and summary statistics.
* Feature Selection: Feature selection is performed to identify and retain the most relevant attributes that significantly impact sales and profit.
* Modelling Approach: Linear regression is not suitable due to the non-linear nature of the dataset, leading to the adoption of various classification algorithms for profit prediction.
* Key Insights Gained: The analysis provides valuable insights into seasonal sales trends, top-selling categories and products, regional profitability variations, the impact of discounts, customer segmentation, and predictive sales and profit models.

9.1 The value of EDA in understanding the dataset:

* Data Understanding: EDA provides a comprehensive view of the dataset's structure, size, and characteristics. It helps analysts get a sense of what the data represents, the number of features, and the range of values within each attribute.
* Data Quality Assessment: EDA helps identify data quality issues such as missing values, duplicates, outliers, and inconsistencies. Addressing these issues during EDA ensures the dataset's integrity and reliability for further analysis.
* Identifying Patterns and Relationships: Through visualizations and summary statistics, EDA reveals patterns, trends, and relationships within the data. It helps uncover correlations between variables and highlights potential cause-and-effect relationships.
* Feature Selection: EDA aids in identifying the most relevant features and attributes that significantly influence the target variable. By understanding which features are essential, analysts can focus on building more accurate and interpretable models.
* Insights Generation: EDA generates initial insights into the data without making assumptions or applying complex models. It serves as a starting point for formulating hypotheses and refining research questions.
* Outlier Detection: EDA facilitates the detection of outliers, which are data points significantly different from the majority. Outliers may indicate data entry errors or interesting phenomena in the dataset.
* Visualization of Distributions: EDA helps visualize the distributions of various variables, enabling analysts to understand data spread, skewness, and the presence of any clusters or groups.
* Data Preprocessing Guidance: EDA guides data preprocessing steps, such as imputation of missing values, handling outliers, and feature scaling, ensuring the data is well-prepared for subsequent analyses.

**CHAPTER 10: REFERENCES**

* <https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>
* <https://stackoverflow.com/questions/16686966/machine-learning-algorithms-which-algorithm-for-which-issue>
* <https://plotly.com/python/plotly-express/>
* <https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification>
* <https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/>