E-COMMERCE SALES ANALYSIS

PROJECT REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY
(Information Technology)

Submitted By:

Natasha

Submitted To:

ABC
(Assistant Professor)

Department of Information Technology,
Guru Nanak Dev Engineering College,
Ludhiana-141006

DECLARATION

I hereby certify that the work which is being presented in this report with the project entitled "E-COMMERCE SALES ANALYSIS" by Natasha, University Roll No. in partial fulfilment of requirements for the award of degree of B.Tech.(Information Technology) submitted in the Department of Information Technology ata Guru Nanak Dev Engineering College, Ludhiana under I.K. Gujral Punjab Technical University in an authentic record of my work carried out under the supervision of ABC(Assistant Professor), Department of Information Technology of GNDEC, Ludhiana. The matter presented has not been submitted by me to any other University/Institute for the award of B.Tech. Degree.

Natasha
URN:

The External Viva-Voce Examination of the student has been held on.

Signature of Internal Examiner Signature of External Examiner

ABSTRACT

This report presents an in-depth analysis of Superstore USA's sales data, aiming to uncover key trends and insights to inform strategic decisions. The dataset includes various attributes related to sales transactions such as order details, customer demographics, product categories, and financial metrics.

The analysis process begins with data cleaning and preprocessing to ensure data accuracy. This is followed by Exploratory Data Analysis (EDA), which involves examining the distribution of sales and profit, regional sales performance, and product category performance. Correlation analysis is also conducted to identify relationships between numerical features like sales, profit, discount, and unit price.

Key findings reveal that most sales transactions are within a specific range, with a few high-value outliers. Profit distribution shows that while the majority of transactions are profitable, some result in losses. Regional analysis highlights significant differences in sales performance across regions. Product category analysis identifies both top-performing and underperforming product lines. Correlation analysis provides insights into factors driving sales and profitability.

The report concludes with several recommendations: focusing on high-performing regions, enhancing sales in low-performing areas, optimizing the product mix, improving profit margins, and leveraging correlation insights for strategic planning. These recommendations aim to improve overall sales performance and profitability, providing a foundation for informed business decisions and future growth strategies.

ACKNOWLEDGEMENT

We are highly grateful to the Principal, Guru Nanak Dev Engineering College (GNDEC), Ludhiana, for providing this opportunity to
carry out the minor project work at E-COMMERCE SALES ANALYSIS

The constant guidance and encouragement received from H.O.D., IT Department, GNDEC Ludhiana has been of great help in carrying out the project work and is acknowledged with reverential thanks.

WE would like to express a deep sense of gratitude and thanks profusely to Prof. ABC, without her wise counsel and able guidance, it would
have been impossible to complete the project in this manner

We express gratitude to other faculty members of Information Technology Department of GNDEC for their intellectual support throughout the course of this work.

Finally, We are indebted to all whosoever have contributed in this report work.

NATASHA

CONTENTS

1 INTRODUCTION

This report delves into an extensive analysis of sales data from Superstore USA, a leading retail chain. The primary objective of this analysis is to uncover critical patterns, trends, and insights that can inform strategic business decisions aimed at boosting sales and profitability.

The dataset used for this analysis encompasses a wide range of attributes related to sales transactions. These attributes include order-specific details like Order ID, Order Priority, and Order Date, as well as customer-related information such as Customer ID, Customer Name, and Customer Segment. Additionally, the dataset includes product-related details like Product Category, Product Sub-Category, and Product Name, along with financial metrics such as Sales, Profit, Discount, and Shipping Cost. This comprehensive dataset provides a rich foundation for in-depth analysis.

The analysis process begins with data cleaning and preprocessing to ensure the data is accurate and reliable. This step involves handling missing values, correcting data types, and removing duplicates. Once the data is clean, Exploratory Data Analysis (EDA) is conducted to summarize the main characteristics of the data and visualize key trends.

The EDA focuses on several critical areas:

- **Sales and Profit Distribution**: Understanding the spread and central tendencies of sales and profit values to identify common transaction ranges and outliers.

- **Regional Sales Performance**: Analyzing sales data across different regions to identify high and low-performing areas, which can inform regional marketing and resource allocation strategies.
- **Product Category Performance**: Evaluating sales and profit data by product category and sub-category to identify top-performing and underperforming product lines, aiding in inventory management and product development strategies.
- **Correlation Analysis**: Examining the relationships between numerical features such as sales, profit, discount, and unit price to understand how these variables interact and influence each other.

Key findings from this analysis provide valuable insights into sales performance, profit margins, and regional and product category trends. Based on these insights, the report offers several recommendations to enhance sales and profitability, such as focusing on high-performing regions, improving sales strategies in low-performing areas, optimizing the product mix, and leveraging correlation insights for strategic planning.

This comprehensive analysis aims to equip Superstore USA with actionable insights and data-driven recommendations that can guide future business strategies and operational improvements, ultimately contributing to the store's growth and success.

2 Data Description

The Superstore USA sales dataset provides a comprehensive overview of the company's sales transactions. Each attribute in the dataset captures specific details about the orders, customers, products, and financial metrics. Here's a detailed description of each attribute:

- **Order ID:** A unique identifier for each order placed. This helps in tracking and distinguishing individual orders.
- **Order Priority**: Indicates the urgency of the order. Possible values include High, Medium, Low, and Critical. This attribute helps in understanding the distribution of orders based on their urgency.
- **Discount:** The discount percentage applied to the order. This helps in analyzing the impact of discounts on sales and profit.
- **Unit Price**: The price per unit of the product sold. This attribute is essential for calculating total sales and understanding pricing strategies.
- **Shipping Cost**: The cost incurred to ship the product. This helps in analyzing the overall expenses related to shipping and its impact on profitability.
- **Customer ID**: A unique identifier for each customer. This is crucial for tracking customer-specific transactions and analyzing customer behavior.

- **Customer Name**: The name of the customer who placed the order. This attribute, coupled with Customer ID, helps in detailed customer analysis.
- **Ship Mode**: The mode of shipping chosen for the order, such as Standard, Express, or First-Class. This helps in analyzing the preferences for different shipping modes.
- **Customer Segment**: The segment to which the customer belongs, such as Consumer, Corporate, or Home Office. This helps in segmenting the market and understanding sales trends within each segment.
- **Product Category**: The category of the product, such as Technology, Office Supplies, or Furniture. This attribute is essential for analyzing sales and profitability across different product categories.
- **Product Sub-Category**: The sub-category of the product within the main category. This helps in detailed analysis within each product category.
- **Product Container**: The type of container used for the product, such as Box, Wrap, or Pack. This helps in understanding packaging preferences and costs.
- **Product Name**: The name of the product sold. This is useful for detailed product-level analysis.
- **Product Base Margin**: The base margin percentage for the product. This helps in analyzing profit margins and pricing strategies.

- **Region**: The region where the order was placed, such as East, West, Central, or South. This is crucial for geographical analysis of sales.
- **State or Province**: The state or province of the order's shipping address. This helps in more granular geographical analysis.
- **City**: The city of the order's shipping address. This further helps in detailed geographical analysis.
- **Postal Code**: The postal code of the shipping address. This is useful for very detailed geographical analysis and for mapping sales data.
- **Order Date**: The date when the order was placed. This is essential for time-series analysis and understanding sales trends over time.
- **Ship Date**: The date when the order was shipped. This helps in analyzing the shipping times and delays.
- **Profit**: The profit made from the order. This is a key financial metric for analyzing the profitability of sales transactions.
- **Quantity Ordered**: The quantity of the product ordered. This helps in understanding the volume of sales and inventory requirements.
- **Sales**: The total sales amount for the order. This is the primary metric for analyzing revenue.

3 Data Cleaning and Preprocessing

Data cleaning and preprocessing are crucial steps in any data analysis project. For the Superstore USA sales dataset, these steps ensure that the data is accurate, consistent, and ready for analysis. Here's a detailed description of the data cleaning and preprocessing process:

1. Handling Missing Values

- **Identification**: First, identify any missing values in the dataset. Missing values can skew the analysis results if not handled properly.
- **Treatment**: Depending on the nature and proportion of missing values, different strategies can be employed:
    - **Deletion**: If the number of missing values is small and they appear randomly, those rows or columns can be deleted.
    - **Imputation**: For columns with significant missing data, missing values can be imputed using mean, median, or mode for numerical columns, or the most frequent value for categorical columns.

2. Removing Duplicates

- **Detection**: Identify duplicate rows in the dataset which can inflate the results.
- **Removal**: Remove any duplicate rows to ensure each transaction is unique.

3. Correcting Data Types

- **Review**: Review the data types of each column to ensure they are appropriate for analysis.

- **Conversion**: Convert columns to their appropriate data types (e.g., dates to `datetime`, categorical variables to `category`, and numerical values to `int` or `float`).

4. Handling Outliers

- **Detection**: Use statistical methods or visualization techniques like box plots to detect outliers in numerical columns (e.g., Sales, Profit).
- **Treatment**: Depending on the analysis requirements, outliers can be:
    - **Removed**: If they are errors or irrelevant.
    - **Transformed**: Using techniques like log transformation.
    - **Capped/Floored**: To limit their impact on the analysis.

5. Standardizing and Normalizing Data

- **Standardization**: Adjust numerical data to have a mean of zero and a standard deviation of one, especially useful when comparing different scales.
- **Normalization**: Scale numerical data to a range of [0, 1] to handle varying scales and ensure fair comparison.

6. Encoding Categorical Variables

- **Label Encoding**: Convert categorical values into numerical labels.
- **One-Hot Encoding**: Create binary columns for each category to avoid ordinal relationships in categorical data.

7. Creating New Features

- **Date Features**: Extract additional features from date columns, such as month, quarter, and day of the week, to capture temporal patterns.
- **Interaction Features**: Create new features by combining existing features (e.g., Sales per Unit, Profit Margin).

8. Ensuring Consistency

- **Data Consistency**: Check for consistency in data entries (e.g., standardized naming conventions for categories, uniform units for numerical data).

9. Handling Data Imbalance

- **Balancing**: If the dataset has an imbalance (e.g., more records for certain regions or product categories), techniques like sampling or weighted analysis can be used to ensure balanced representation.

Data cleaning and preprocessing are essential for ensuring the reliability and accuracy of the analysis. By addressing missing values, removing duplicates, correcting data types, handling outliers, standardizing and normalizing data, encoding categorical variables, creating new features, ensuring data consistency, and handling data imbalance, we prepare the dataset for thorough analysis. These steps help in deriving meaningful insights and making informed business decisions based on the cleaned and processed data.

4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in understanding the underlying patterns and relationships within a dataset. For the Superstore USA dataset, EDA involves analyzing various aspects of the sales data to gain insights and inform strategic decisions. Below are the detailed steps and analyses for each specified area of focus:

4.1 Sales Distribution

Objective: To understand the distribution of sales values and identify any patterns or anomalies.

Approach:

- **Histogram**: Plot a histogram to visualize the distribution of sales.
- **Descriptive Statistics**: Calculate mean, median, mode, standard deviation, and percentiles to summarize the sales data.

Insights:

- Identify common sales ranges and the frequency of different sales amounts.
- Detect any outliers or unusually high sales values that may need further investigation.

4.2 Profit Distribution

Objective: To analyze the distribution of profit values and assess the profitability of transactions.

Approach:

- **Histogram**: Plot a histogram to visualize the distribution of profit.
- **Descriptive Statistics**: Calculate mean, median, mode, standard deviation, and percentiles to summarize the profit data.

Insights:

- Understand the range of profit margins and identify any transactions that resulted in losses.
- Detect patterns in profitability that could inform pricing and discount strategies.

4.3 Sales by Region

Objective: To evaluate the sales performance across different geographical regions.

Approach:

- **Bar Plot**: Create a bar plot to compare total sales across regions.
- **Descriptive Analysis**: Summarize total sales, average sales, and number of transactions for each region.

Insights:

- Identify high-performing and low-performing regions.
- Determine regional sales trends and potential areas for targeted marketing efforts.

4.4 Sales by Category and Sub-Category

Objective: To analyze sales performance across different product categories and sub-categories.

Approach:

- **Bar Plot**: Create bar plots to compare total sales across product categories and sub-categories.
- **Descriptive Analysis**: Summarize total sales, average sales, and number of transactions for each category and sub-category.

Insights:

- Identify best-selling and least-selling product categories and sub-categories.
- Determine product categories that contribute most to overall sales and profitability.

4.5 Correlation Analysis

Objective: To understand the relationships between numerical features in the dataset.

Approach:

- **Correlation Matrix**: Calculate the correlation matrix to quantify the relationships between numerical variables like sales, profit, discount, and unit price.
- **Heatmap**: Visualize the correlation matrix using a heatmap.

5. CONCLUSION

The analysis of the Superstore USA sales dataset has provided several valuable insights and actionable recommendations that can significantly inform strategic business decisions and operational improvements. Here are the key conclusions drawn from the project:

1. Sales and Profit Distribution

- The sales distribution reveals that most transactions fall within a specific range, with a few high-value outliers. This indicates that while there are standard sales amounts, there are occasional large orders that can significantly impact total sales.
- Profit distribution analysis shows that the majority of transactions are profitable, but there are some instances of negative profits. These loss-making transactions need further investigation to understand the underlying causes, such as excessive discounts or high shipping costs.

2. Regional Sales Performance

- The analysis of sales by region highlights significant disparities in sales performance across different regions. Certain regions, such as the West and East, show higher sales volumes, indicating strong market presence and customer base.
- Regions with lower sales, such as the South and Central, present opportunities for targeted marketing efforts and promotional activities to boost sales.

3. Product Category and Sub-Category Analysis

- The evaluation of sales by product category and sub-category reveals that some categories, like Technology and Office Supplies, are top performers contributing significantly to overall sales.
- Underperforming categories and sub-categories need attention for potential improvement in marketing strategies, product placement, and inventory management. Identifying and addressing issues in these areas can help optimize the product mix and enhance profitability.

4. Correlation Analysis

- The correlation analysis uncovers important relationships between numerical features. For instance, a strong correlation between sales and profit suggests that higher sales generally lead to higher profits.
- Understanding the impact of discounts on sales and profit helps in optimizing pricing strategies. The negative correlation between discount and profit indicates that while discounts can drive sales, they also reduce profit margins.

6 RECOMMENDATIONS

Based on the insights gained from the analysis, the following recommendations are proposed to enhance sales performance and profitability for Superstore USA:

1. **Focus on High-Performing Regions**: Continue to invest in and expand operations in high-performing regions. Implement targeted marketing campaigns to further strengthen market presence in these areas.
2. **Improve Sales in Low-Performing Regions**: Develop targeted promotional activities and customer engagement strategies for regions with lower sales. Consider region-specific product offerings and pricing strategies to attract more customers.
3. **Optimize Product Mix**: Focus on promoting high-performing product categories and sub-categories. Analyze customer preferences and sales trends to adjust the product mix and ensure a balanced inventory.
4. **Enhance Profit Margins**: Investigate loss-making transactions and implement measures to reduce costs, such as optimizing shipping methods and negotiating better supplier deals. Adjust discount strategies to find a balance between driving sales and maintaining healthy profit margins.

5. **Leverage Correlation Insights**: Use the findings from correlation analysis to inform pricing and discount strategies. Ensure that discounts are strategically applied to maximize sales without significantly compromising profit margins.
6. **Continuous Monitoring and Analysis**: Establish a system for continuous monitoring and analysis of sales data. Regularly review performance metrics and adjust strategies as needed to respond to changing market conditions and customer preferences.

Final Thoughts

The comprehensive analysis of the Superstore USA sales data provides a solid foundation for making informed business decisions. By understanding the key drivers of sales and profitability, the company can implement targeted strategies to enhance performance, optimize operations, and achieve sustainable growth. The insights gained from this project will serve as a valuable resource for future strategic planning and decision-making.

7. REFERENCES

- **Books and Academic Papers**

  - Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
  - Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.
  - Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

- **Online Courses and Tutorials**

  - Coursera. (2021). *Applied Data Science with Python Specialization*. University of Michigan.
  - Kaggle. (2021). *Data Cleaning*. Available at: Kaggle Data Cleaning Course.

- **Software and Libraries Documentation**

  - Python Software Foundation. (2021). *Python Documentation*. Available at: Python Docs.
  - pandas Development Team. (2021). *pandas Documentation*. Available at: pandas Docs.
  - seaborn Development Team. (2021). *seaborn Documentation*. Available at: seaborn Docs.
  - Matplotlib Development Team. (2021). *Matplotlib Documentation*. Available at: Matplotlib Docs.

- **Research Articles and Case Studies**

  - Manyika, J., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
  - Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press.