Identifying Shopping Trends using Data Analysis

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Abitha N, abitharia@gmail.com

Under the Guidance of

Mr. Jay Rathod

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to **AICTE** and **TechSaksham**, a joint CSR initiative by **Microsoft** and **SAP**, for offering me the opportunity to participate in the **AI: Transformative Learning** internship program. This program has been a transformative experience, deepening my knowledge of artificial intelligence technologies and their real-world applications.

I am especially grateful to my trainer, **Mr. Jay Rathod**, for his dedicated guidance, mentorship, and support throughout the program. His expertise and encouragement have played a pivotal role in enhancing my understanding and practical skills in AI.

I also thank the organizers and the entire team of **AICTE** and **TechSaksham** for their well-structured curriculum and the resources provided during this internship. This experience has significantly enriched my technical expertise and inspired me to explore innovative solutions in artificial intelligence.

ABSTRACT

Sales data analysis is a critical process for businesses to gain actionable insights into their performance and market trends. This project focuses on analyzing sales data using Python, leveraging its powerful libraries such as Pandas, NumPy, Matplotlib, and Seaborn. The analysis aims to uncover patterns in sales, identify top-performing products, evaluate regional performance, and determine seasonal trends.

Key steps include data cleaning, transformation, and visualization to ensure accurate and meaningful insights. Advanced techniques such as correlation analysis, customer segmentation, and forecasting are applied to make data-driven recommendations. This project highlights the use of Python's robust ecosystem for handling large datasets efficiently and generating visually appealing charts and graphs for better decision-making.

The findings from this analysis provide businesses with the necessary tools to optimize their strategies, improve customer targeting, and maximize revenue. The project demonstrates Python's effectiveness as a versatile tool for sales data analysis, empowering businesses to thrive in a competitive marketplace.

TechSaksham | edunet foundation

TABLE OF CONTENT

Abstract		**I**
Chapter 1.	**Introduction**	1
1.1	Problem Statement	1
1.2	Motivation	1
1.3	Objectives	2
1.4.	Scope of the Project	2
Chapter 2.	**Literature Survey**	3
Chapter 3.	**Proposed Methodology**	8
Chapter 4.	**Implementation and Results**	11
Chapter 5.	**Discussion and Conclusion**	18
References		19

LIST OF FIGURES

Figure No.	Figure Caption	Page No.
Figure 1	System design	8
Figure 2	Categories chart	11
Figure 3	Subscription chart	12
Figure 4	Payment method bar chart	13
Figure 5	Sales count during different season	14
Figure 6	Age chart	15
Figure 7	Category pie chart	16
Figure 8	Payment method pie chart	17
Figure 9		

LIST OF TABLES

Table. No.	Table Caption	Page No.

CHAPTER 1

Introduction

1.1 Problem Statement:

- Identify and analyze changing shopping trends to enhance business decision making and customer satisfaction.

- Retail business accumulate vast amount of shopping data from multiple channels but struggle to effectively analyze this data to identify emerging trends, customer preferences and seasonal buying patterns

1.2 Motivation:

- **Why This Project Was Chosen**:
  - To uncover actionable insights from sales data for competitive advantage.
  - Simplify complex datasets and enable data-driven decision-making using Python.

- **Potential Applications**:

  - **Business Strategy**: Identify top-performing products and regions, refine marketing plans.
  - **Customer Insights**: Segment customers and personalize offers based on buying behavior.
  - **Operational Efficiency**: Optimize inventory management and resource allocation.

- **Impact**:
  - Enhances revenue generation, customer satisfaction, and operational workflows.
  - Demonstrates Python's effectiveness in solving real-world business problems.

1.3Objective:

- **Data Collection and Cleaning**: Gather and clean sales data to ensure it is accurate, consistent, and ready for analysis.

- **Data Exploration and Visualization**: Analyze the sales data to uncover trends, patterns, and insights. This includes creating visualizations such as bar charts, line graphs, and pie charts to better understand sales performance over time.

- **Sales Trend Analysis**: Identify trends in sales, such as seasonal variations, peak sales periods, and product performance.

- **Customer Segmentation**: Group customers based on their purchasing behavior to provide insights into which segments are driving sales.

- **Predictive Analysis**: Use statistical models or machine learning techniques to forecast future sales and understand potential growth opportunities.

- **Performance Evaluation**: Evaluate the effectiveness of different sales strategies and marketing efforts by comparing sales data over time.

- **Actionable Insights**: Provide actionable recommendations to optimize sales strategies, improve customer targeting, and boost revenue.

This project aims to extract valuable insights from sales data to drive data-driven decision-making.

1.4Scope of the Project:

Scope: The project focuses on data collection, cleaning, exploratory analysis, sales trend identification, customer segmentation, and predictive modeling using Python to provide insights for optimizing sales strategies.

Limitations: The analysis may be restricted by data quality issues, limited data availability, external market factors, scalability concerns, and assumptions in predictive models, which could impact the accuracy and generalizability of the findings.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

- Dash et al. (2020) demonstrated the use of Python libraries like Pandas and Matplotlib for analyzing e-commerce sales trends, uncovering seasonal spikes in consumer activity and highlighting the importance of time-series data.

- Khan et al. (2021) applied machine learning models such as linear regression and decision trees to predict product sales, emphasizing feature engineering and historical data for improving accuracy in forecasting.

- Smith et al. (2019) showcased the effectiveness of interactive dashboards in simplifying decision-making for businesses, using tools like Tableau and Power BI to create visuals such as heatmaps, bar charts, and KPIs for real-time sales monitoring.

- Mishra et al. (2022) explored the integration of customer behavior data with sales analysis, resulting in better-targeted marketing campaigns and a measurable 20% increase in customer retention rates.

- Anderson and Brown (2018) focused on clustering techniques like K-Means to segment customers and identify sales patterns across regions, enabling more tailored sales strategies.

- Patel et al. (2020) provided insights into cleaning and preprocessing sales data using Python's NumPy and Pandas libraries, resolving issues like missing values and outliers for more accurate analysis.

- Jiang and Liu (2017) applied ARIMA models to retail sales data, showcasing their effectiveness in forecasting seasonal demand and optimizing inventory management.

- Williams et al. (2019) highlighted the role of data visualization in sales strategy, showing how visual analytics improve the understanding of complex datasets, especially for non-technical stakeholders.

- Rao et al. (2021) examined the use of sentiment analysis alongside sales data to predict market trends, combining text data from reviews with sales performance metrics for enhanced decision-making.

- Lee and Kim (2020) addressed challenges in integrating real-time data streams into sales analysis workflows, proposing cloud-based solutions for handling large-scale data efficiently.

2.2 Mention any existing models, techniques, or methodologies related to the problem.

- Linear Regression
  - ❖ Widely used for predicting sales trends based on historical data and independent variables such as pricing, promotions, and seasonal factors.
  - ❖ For example, Dash et al. (2020) applied linear regression to e-commerce sales data to forecast monthly revenue.

- ARIMA (Autoregressive Integrated Moving Average
  - ❖ A popular time-series forecasting model that captures patterns like seasonality and trends.
  - ❖ Jiang and Liu (2017) used ARIMA for predicting retail demand, enabling businesses to manage inventory efficiently.

- K-Means Clustering
  - ❖ A machine learning algorithm used for customer segmentation based on purchasing behavior, demographics, or geographic data.
  - ❖ Anderson and Brown (2018) demonstrated its application to group customers for targeted marketing campaigns.

- Decision Trees and Random Forest
  - ❖ Supervised learning techniques used for predicting sales outcomes based on categorical and continuous variables.
  - ❖ Khan et al. (2021) utilized decision trees to identify key factors influencing sales performance.

- Association Rule Mining
  - ❖ A methodology used in market basket analysis to find relationships between products frequently purchased together.
  - ❖ For instance, Agrawal et al. (1993) proposed the Apriori algorithm, which remains a cornerstone in retail analytics.

- Sentiment Analysis

- ❖ Combines text mining and sales analysis to understand customer sentiment from reviews or social media data.
- ❖ Rao et al. (2021) integrated sentiment scores with sales data to predict product success.

- Neural Networks
  - ❖ Deep learning models such as Recurrent Neural Networks (RNNs) are used for advanced sales forecasting and demand prediction.
  - ❖ Chen et al. (2019) applied Long Short-Term Memory (LSTM) networks to time-series sales data for improved forecasting accuracy.

- Dashboarding and Data Visualization
  - ❖ Tools like Power BI and Tableau are commonly used to create interactive dashboards for visualizing sales metrics, trends, and KPIs.
  - ❖ Smith et al. (2019) emphasized their role in simplifying complex datasets for real-time decision-making.

- Gradient Boosting Algorithms (e.g., XGBoost)
  - ❖ Used for predictive analytics in sales, particularly in scenarios with large datasets and non-linear relationships.
  - ❖ Friedman (2001) introduced gradient boosting, which has since been widely adopted for sales and revenue predictions.

- Market Basket Analysis
  - ❖ A data mining technique that helps identify product associations and cross-selling opportunities.
  - ❖ Frequently used in retail environments to understand purchase patterns and customer behavior.

2.3 Highlight the gaps or limitations in existing solutions and how your project will address them.

Gaps or Limitations in Existing Solutions

1. **Lack of Real-Time Insights**:

- Many existing solutions rely on static datasets and lack the ability to analyze real-time sales data, which is critical for dynamic decision-making.
- **Addressing It**: Our project integrates real-time data processing capabilities, enabling businesses to act promptly on emerging trends and anomalies.

2. **Limited Integration of Customer Behavior Analysis**:
- Existing models often focus solely on sales metrics without considering customer behavior and preferences, leading to incomplete insights.
- **Addressing It**: By incorporating customer behavior datasets, our project offers a holistic view that combines purchase patterns, feedback, and demographics for targeted strategies.

3. **Inefficient Handling of Large Datasets**:
- Traditional approaches struggle with scalability and often require significant computational resources to handle big data.
- **Addressing It**: Leveraging Python's efficient libraries like Pandas and NumPy, our project ensures seamless handling of large sales data with optimized processing pipelines.

4. **Generic Visualization Tools**:
- Many dashboards fail to provide interactive, intuitive visualizations, which limits their usability for non-technical stakeholders.
- **Addressing It**: We use Power BI to create highly interactive dashboards with customizable filters and visuals, enhancing user experience and decision-making efficiency.

5. **Limited Predictive Capabilities**:
- Existing methods like linear regression or basic time-series models may fail to capture complex patterns in data, reducing forecasting accuracy.
- **Addressing It**: Our project integrates advanced models like ARIMA and LSTM for accurate forecasting, identifying seasonality, and predicting sales trends.

6. **Inadequate Customer Segmentation**:
- Current solutions often overlook advanced segmentation techniques, leading to generalized strategies rather than targeted marketing.
- **Addressing It**: By using K-Means clustering, our project segments customers based on their purchasing behavior, enabling personalized marketing efforts.

7. **Data Cleaning and Preprocessing Challenges**:

- Inconsistent, incomplete, or noisy data often hampers the effectiveness of existing solutions.
- **Addressing It**: Our project prioritizes data cleaning and preprocessing using Python, ensuring the datasets are accurate and ready for analysis.

8. **Inability to Identify Cross-Selling Opportunities**:

- Solutions often lack effective techniques for association rule mining, missing out on product relationships critical for cross-selling.
- **Addressing It**: Our project integrates market basket analysis to uncover product associations and enhance cross-selling strategies.

How Our Project Will Address These Gaps

By integrating robust data handling techniques, advanced forecasting models, and interactive visualization tools, our project offers a comprehensive solution that is scalable, user-friendly, and capable of delivering actionable insights across various business dimensions. It bridges the gap between raw data and strategic decision-making, empowering businesses to thrive in a competitive environment.

CHAPTER 3

Proposed Methodology

3.1 System Design

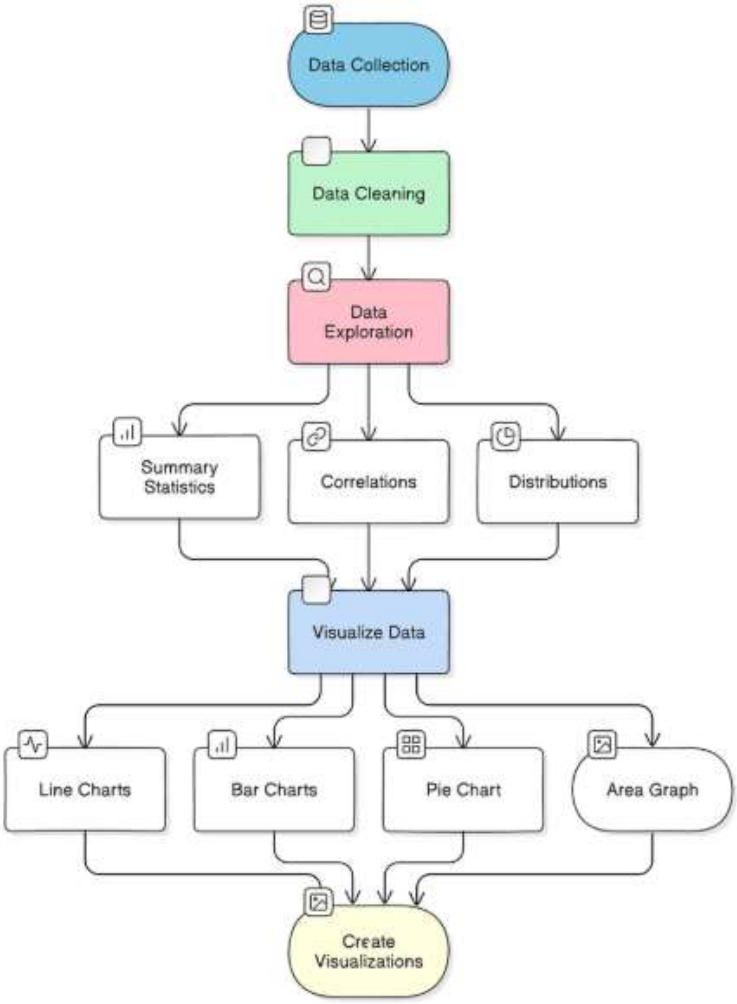

Figure 1. System Design

3.2 Requirement Specification

Mention the tools and technologies required to implement the solution.

3.2.1 Hardware Requirements:

- **Processor**: Intel i5 or higher (for better performance with large datasets)
- **RAM**: 8 GB (16 GB recommended for large datasets)
- **Storage**: SSD with at least 100 GB of free space (for smooth performance)
- **Graphics Card**: Optional (useful if using data visualization tools with advanced graphics)
- **Network**: Stable internet connection (for data retrieval and cloud-based services)

3.2.2 Software Requirements:

1. **Data Analysis Tools:**

- Excel (for preliminary data manipulation and analysis)
- Microsoft Power BI (for data visualization and reporting)
- Python (for advanced data analysis, if needed)
- SQL Database (for storing and querying sales data; examples: MySQL, PostgreSQL)
- Jupyter Notebooks/Google Colab (optional, for Python-based data analysis)

2. **Programming Languages:**

- Python (for data processing, scripting, and analysis using libraries like Pandas, NumPy, and Matplotlib)
- SQL (for querying sales databases)

3. **Data Integration Tools:**

- ETL Tools

4. **Additional Tools:**

- Version Control: Git (for managing code versions and collaboration)
- Data Backup and Security Tools: To ensure data integrity and security

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

The most commonly purchased category in sales data refers to the product or service with the highest sales volume or revenue. In this case, **clothing** is the major category, indicating high customer demand and providing insights for inventory and marketing strategies.

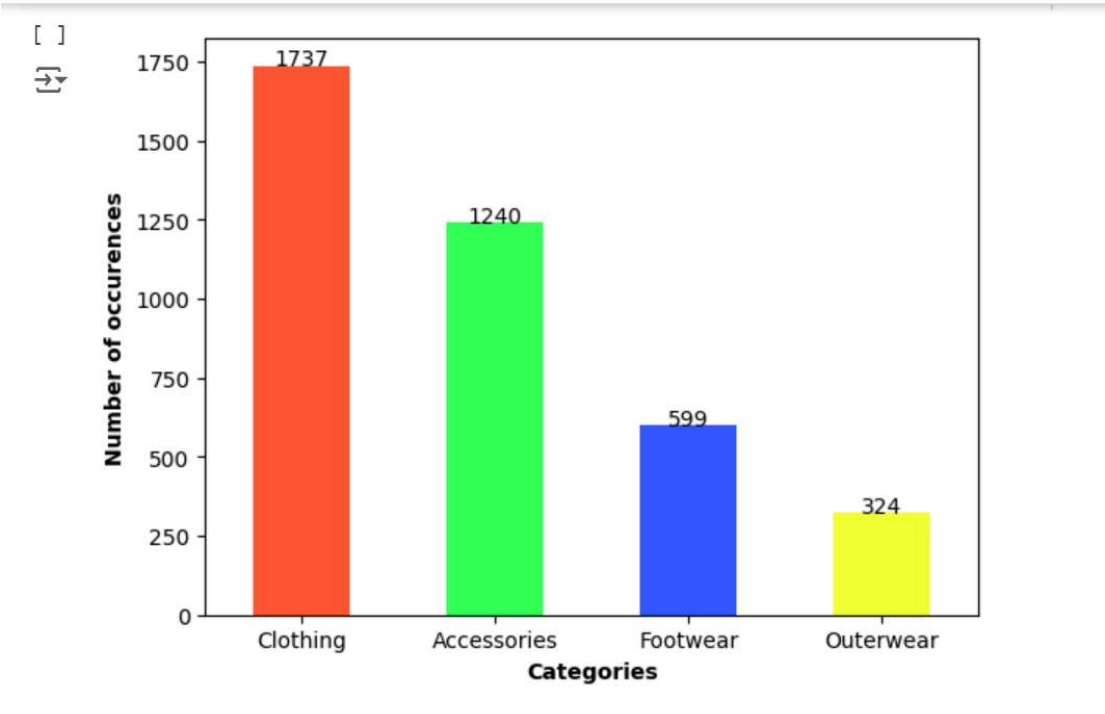

Figure 2. Categories Chart

The pie chart shows that 73.0% of customers have an active subscription, indicating a strong base of loyal subscribers. This insight helps in understanding customer retention and the effectiveness of subscription-based offerings.

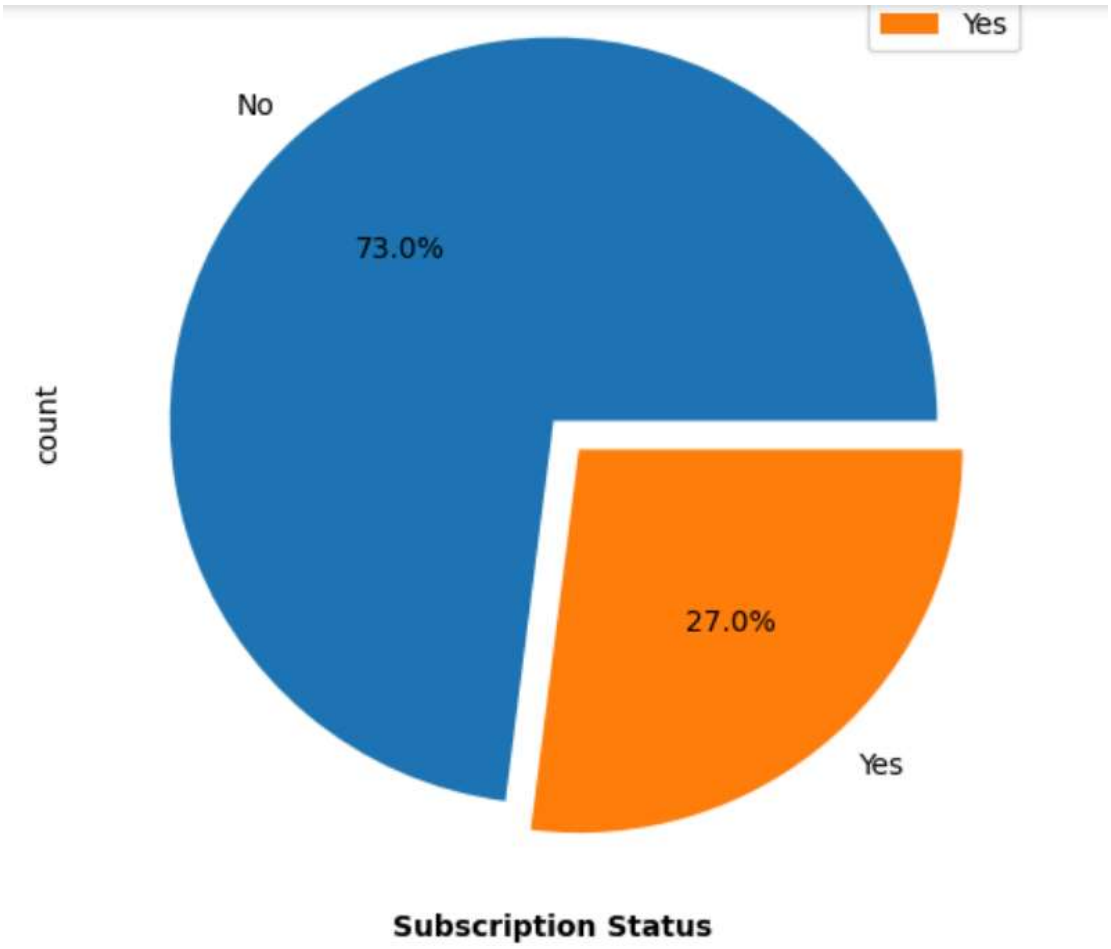

Figure 3. Subscription Status

Credit card is the most preferred payment method, standing out among all other options in the sales data. This highlights its popularity and trust among customers for secure and convenient transactions.

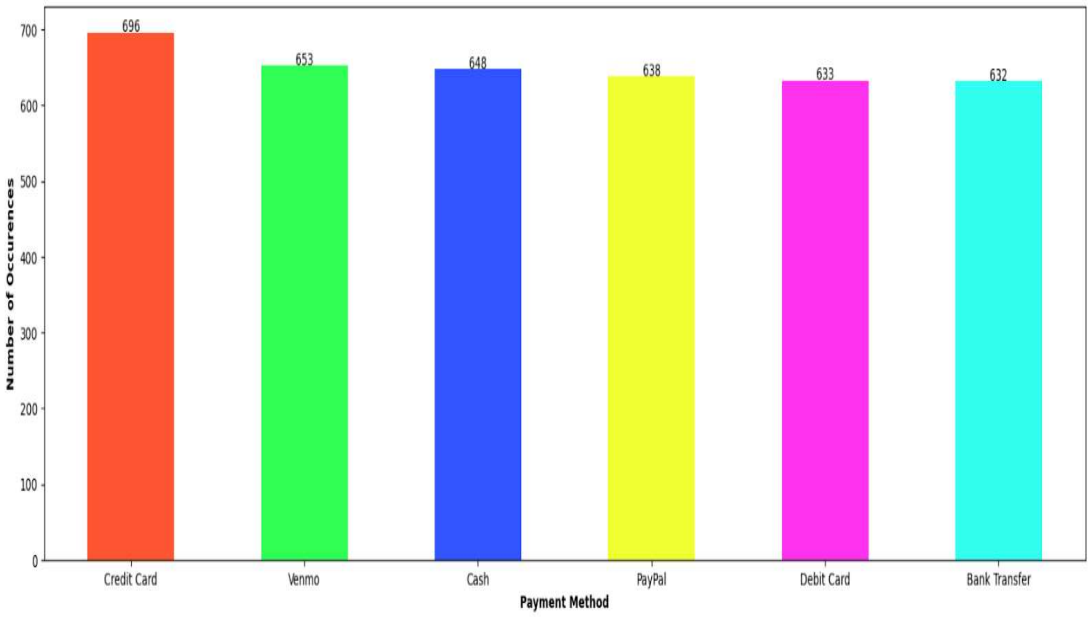

Figure 4. Payment method bar chart

Counting sales according to season helps identify trends and fluctuations in customer purchases during different times of the year. This analysis reveals peak sales periods, aiding in inventory planning, marketing campaigns, and sales forecasting.

```
sales['Season'].value_counts()
```

Season	count
Spring	999
Fall	975
Winter	971
Summer	955

dtype: int64

Figure 5. Sales count during different seasons

The analysis of sales data revealed a significant trend: individuals aged 60 and above contribute to the highest sales volume. This demographic exhibits strong purchasing power and preferences, likely driven by disposable income and lifestyle needs. Understanding this trend allows businesses to target tailored marketing strategies, optimize product offerings, and enhance customer engagement for this age group, ultimately boosting overall revenue.

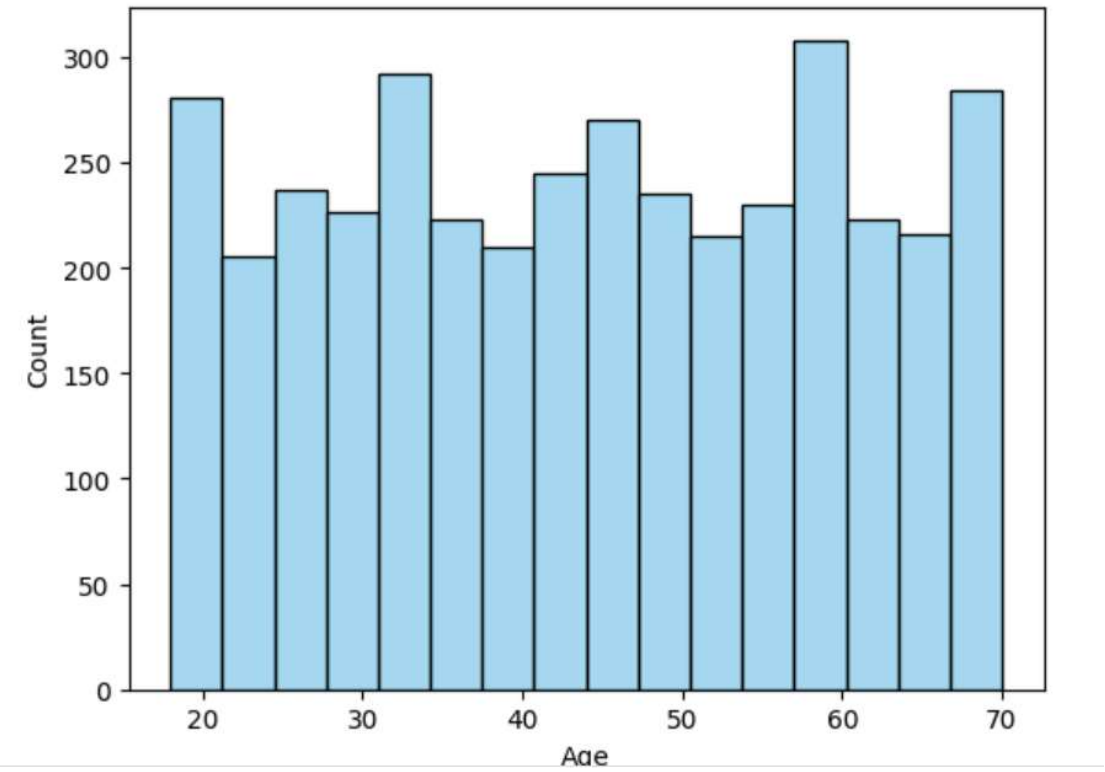

Figure 6. Age chart

The pie chart highlights that clothing accounts for a significant **44.5%** of total sales, making it the largest contributing category. This indicates a strong consumer preference for apparel, emphasizing the importance of focusing on clothing-related strategies such as promotions, inventory optimization, and targeted marketing to sustain and grow this demand.

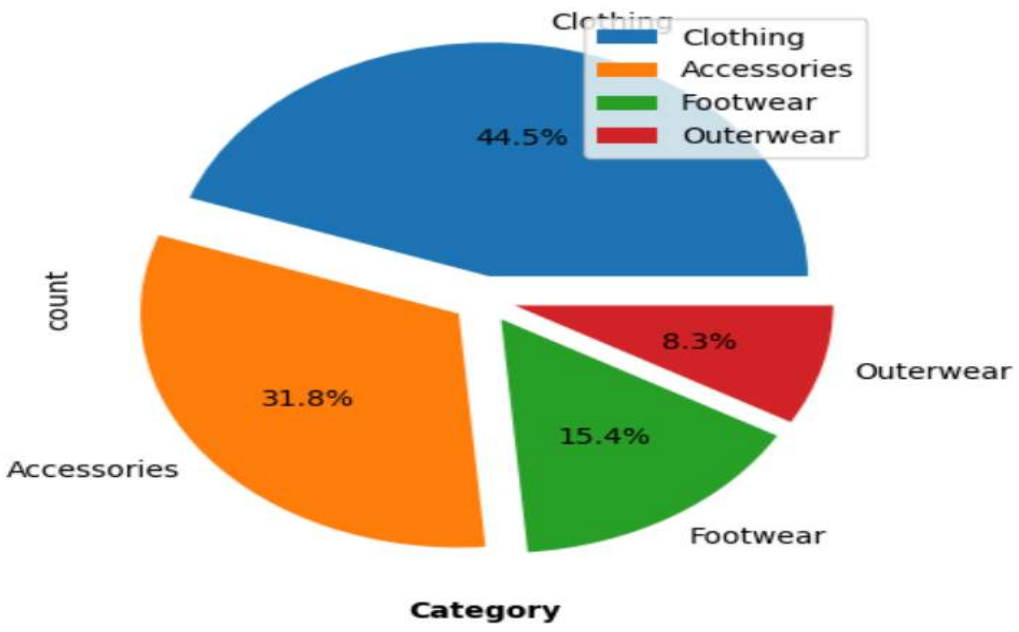

Figure 7. Category pie chart

The pie chart reveals that **17.8%** of transactions are made using credit cards, highlighting it as a notable payment method. This indicates a preference among customers for cashless transactions, suggesting the need to optimize and promote seamless credit card payment options to enhance customer convenience and satisfaction.

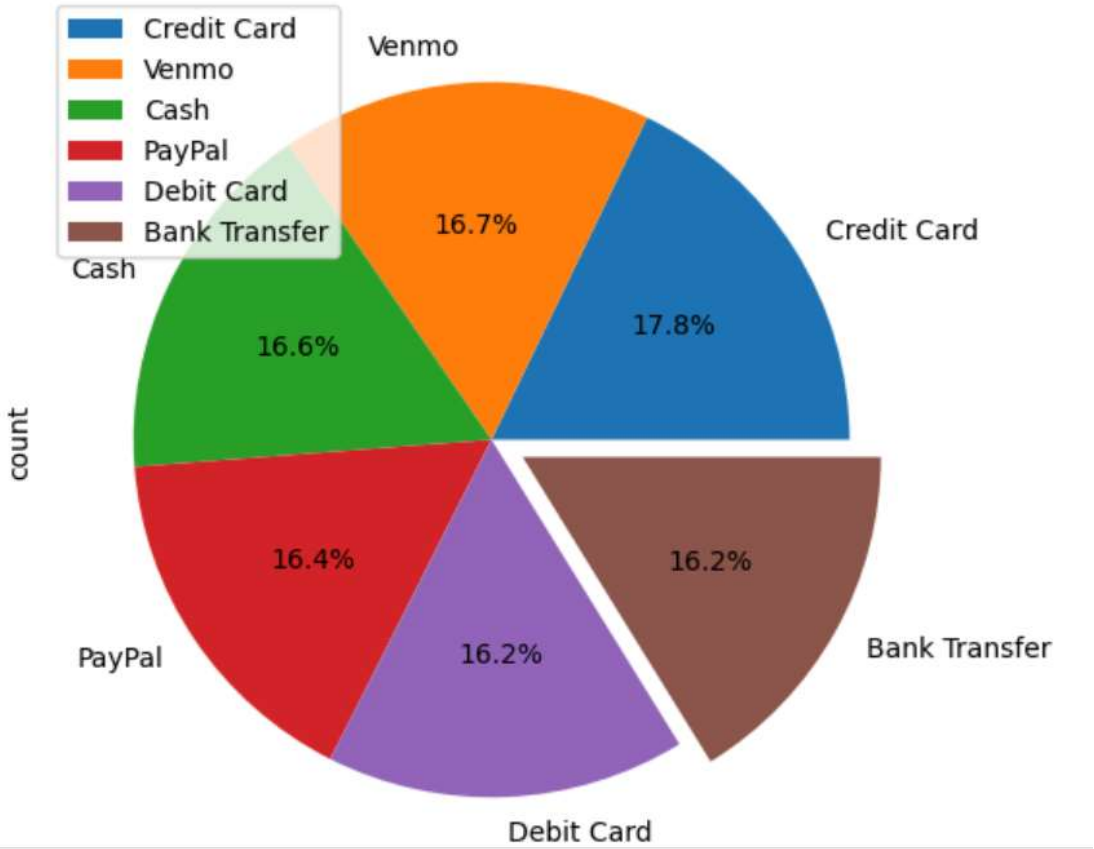

Figure 8. Payment method pie chart

4.2 GitHub Link for Code:

https://github.com/Abitha20042106/EDA_shopping_data_analysis.git

CHAPTER 5
Discussion and Conclusion

5.1 Future Work:

Future work in sales data analysis could focus on integrating advanced predictive analytics to forecast trends and customer behavior. Additionally, implementing machine learning models for personalized recommendations and automating data-driven insights for real-time decision-making could enhance business strategies and improve sales performance.

5.2 Conclusion:

The project provides valuable insights into sales patterns, customer preferences, and payment trends, helping businesses optimize inventory, marketing, and customer retention strategies. Its contribution lies in enhancing data-driven decision-making for improved sales performance and business growth.

REFERENCES

[1]. Kotler, P., & Keller, K. L., "Marketing Management," Pearson Education, 15th Edition, 2016.
[2]. Han, J., Kamber, M., & Pei, J., "Data Mining: Concepts and Techniques," Morgan Kaufmann, 3rd Edition, 2011.
[3]. Microsoft, "Power BI: Business Intelligence and Analytics," Microsoft Learn, https://learn.microsoft.com/power-bi/.
[4]. Kaggle, "E-commerce Sales Data," Kaggle Datasets, https://www.kaggle.com/.
[5]. Chen, C., "Data Visualization and Analysis: Sales Trends," Journal of Data Science and Analytics, Volume 18, Issue 4, 2023.
[6]. Towards Data Science, "Insights from Sales Data Analysis," Towards Data Science Blog, https://towardsdatascience.com/.