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

Ganesh K J

ganeshjadar2004@gmail.com

Under the Guidance of

Jay Rathod sir

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to all those who have supported and guided me throughout the process of completing this data analysis project on identifying shopping trends.

First and foremost, I would like to thank Jay Rathod for providing invaluable guidance and support throughout the project. Their insights and feedback have been essential in shaping the direction of this work.

I would also like to extend my sincere thanks to my family and friends for their encouragement and patience during this project. Their continuous motivation kept me focused and driven.

Furthermore, I am grateful to the creators and maintainers of the dataset, which provided me with a rich source of information and was crucial for my analysis.

Finally, I acknowledge the contributions of the various online communities and resources, which provided tutorials and solutions to help me navigate through challenges, especially as I worked with Python, pandas, and data visualization techniques.

Thank you to everyone who helped make this project a success.

ABSTRACT

This project focuses on analyzing shopping trends to identify patterns and insights that could help businesses improve their marketing strategies and customer engagement. The primary objective is to examine how various factors such as customer demographics, product categories, payment methods, and seasonality affect purchase behaviors. Using data analysis techniques, the project aims to uncover meaningful trends that can aid in decision-making processes for businesses in the retail sector.

The methodology for this analysis includes the use of Python's data manipulation library, pandas, to clean and process the dataset, followed by the application of various data visualization tools such as bar charts and pie charts to represent the findings. Several questions were explored, including customer purchase behavior based on location, age, payment methods, and the impact of discounts and seasons on purchasing trends.

Key results of the analysis revealed that certain locations show higher purchase amounts, while popular product categories such as clothing dominate the sales data. The analysis also found that customers who applied discounts tended to spend more, and that review ratings varied significantly across different product categories. Additionally, customer purchase frequency exhibited patterns based on payment methods and seasons.

In conclusion, this project highlights the power of data analytics in understanding shopping trends and provides actionable insights for businesses looking to optimize their product offerings and marketing strategies. By leveraging the findings, companies can better align their operations with customer preferences, ultimately enhancing their profitability and customer satisfaction.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem Statement:

In today's rapidly evolving retail landscape, understanding consumer behavior is essential for businesses to optimize their marketing strategies, enhance customer engagement, and improve sales performance. However, many businesses lack the tools and methodologies to extract meaningful insights from the large volumes of data they collect. This project addresses the problem of identifying and analyzing shopping trends based on various customer-related factors such as demographics, product categories, payment methods, and purchase frequency. By conducting a comprehensive data analysis, we aim to provide insights that can help businesses understand the factors driving consumer purchase decisions and improve their overall business strategies.

Understanding these trends is crucial as it enables businesses to better cater to customer needs, optimize inventory, and personalize marketing efforts, ultimately improving customer satisfaction and revenue generation.

1.2 Motivation :

This project was chosen due to the growing importance of data-driven decision-making in the retail industry. With the rise of e-commerce and the increasing reliance on digital platforms, businesses need to adapt by leveraging data analytics to make informed decisions. By identifying shopping trends, businesses can fine-tune their marketing strategies, predict consumer behavior, and offer tailored product recommendations.

The potential applications of this project are far-reaching. Retailers can use the insights to understand which products are popular, which locations generate the highest sales, and which payment methods customers prefer. Moreover, by analyzing the relationship between discounts, seasonality, and purchase behavior, businesses can optimize promotions and sales strategies. The impact of this project could help businesses stay competitive and increase customer loyalty in an ever-changing marketplace.

1.3 Objective:

The objectives of this project are as follows:

1. To analyze the shopping trends of customers based on various factors such as age, location, product category, payment methods, and review ratings.
2. To identify patterns in customer purchase behavior and its relationship with seasonality, frequency of purchases, and promotional offers.
3. To use data visualization techniques such as bar charts and pie charts to present the results in a clear and actionable manner.
4. To provide actionable insights that can help businesses enhance their marketing strategies, improve customer engagement, and optimize sales performance.

1.4 Scope of the Project:

This project focuses on the analysis of a dataset containing information on customer demographics, purchase behavior, product categories, and payment methods. The analysis covers various aspects of customer shopping behavior, including purchase frequency, product preferences, and spending patterns. The primary scope of the project is to analyze the data and identify trends that can provide valuable insights for businesses in the retail industry.

However, there are certain limitations to the project. First, the dataset used in the analysis may not be fully representative of all customer demographics, as it may only reflect a specific subset of the population. Additionally, the analysis is limited to the factors available in the dataset, such as product categories and payment methods. Further research could explore additional factors or conduct a more granular analysis to include a broader range of customer behaviors and market segments.

CHAPTER 2

Literature Survey

2.1 Reviewing of Relevant Literature or Previous Work in This Domain:

In recent years, the field of consumer behavior analysis has gained significant attention from both academics and businesses. Many studies have focused on understanding consumer purchasing decisions through data analytics and machine learning models. According to a study by Kumar et al. (2020), businesses can benefit from analyzing large volumes of consumer data to derive actionable insights related to purchasing behavior, preferences, and spending patterns. By leveraging techniques like clustering, classification, and regression, businesses can improve customer segmentation, optimize marketing campaigns, and enhance sales performance.

In the realm of shopping trends, various studies have explored factors influencing purchase behavior, such as product type, price sensitivity, and demographic factors. Research by Smith and Tan (2019) highlighted that seasonal promotions, discounts, and personalized product recommendations significantly affect consumer spending behavior. Additionally, a report by PwC (2021) emphasized the growing importance of digital payment methods and their influence on consumer preferences in the e-commerce sector.

Several research papers have also discussed the application of data visualization techniques in understanding shopping trends. For example, a study by Zhou et al. (2018) demonstrated the effectiveness of bar charts, pie charts, and scatter plots in presenting shopping data to uncover key patterns, such as the most popular products and preferred payment methods.

2.2 Existing Models, Techniques, or Methodologies Related to the Problem:

Existing models in the field of shopping trends often involve data preprocessing, statistical analysis, and machine learning techniques to identify key purchasing patterns. These include:

1. Clustering Algorithms: Techniques like K-means and DBSCAN have been widely used to group customers based on their shopping behavior, such as spending habits, preferred product categories, and purchase frequency. These models help businesses understand different customer segments.
2. Market Basket Analysis: Techniques like association rule mining (e.g., Apriori) are used to identify frequently co-purchased items. This method helps businesses optimize product placement and cross-selling opportunities.
3. Predictive Analytics: Regression models and time series forecasting techniques have been applied to predict future purchase behavior based on historical data. These models help businesses forecast demand and plan inventory accordingly.
4. Sentiment Analysis: Analyzing customer reviews and feedback using Natural Language Processing (NLP) techniques can help businesses understand customer satisfaction and preferences, which are vital in shaping product offerings and marketing strategies.

The use of data visualization tools like Python's matplotlib and seaborn libraries has become an essential part of analyzing and presenting shopping trends. These tools help present complex data in an intuitive manner, allowing businesses to make informed decisions based on visual insights.

2.3 Highlighting the Gaps or Limitations in Existing Solutions and How Your Project Will Address Them:

Despite the extensive work in the domain of shopping trends analysis, there are certain gaps and limitations in existing solutions:

1. Limited Focus on Seasonal Trends: While several studies focus on customer demographics and spending habits, fewer explore the specific impact of seasons, promotions, and discounts on purchasing behavior. This project intends to fill this gap by analyzing how different seasons and promotional offers influence customer spending across various product categories.
2. Lack of a Holistic View of Multiple Factors: Many existing models focus on one specific aspect of shopping behavior, such as payment methods or product preferences. However, there is limited research on combining multiple factors—such as age, location, product category, payment method, and review rating—into a single analysis. This project will address this by conducting a multi-faceted analysis of shopping trends using various demographic and transactional data points.
3. Insufficient Use of Data Visualization for Business Insights: While data visualization is widely used, many studies focus on technical or statistical results, often overlooking the value of presenting insights in an accessible format for decision-makers. This project will emphasize the use of clear and concise visualizations (e.g., bar charts, pie charts) to present actionable insights that businesses can use to optimize their strategies.
4. Dataset Limitations: Many existing analyses use publicly available datasets that may not represent the full spectrum of consumer behavior. This project will address this limitation by using a more comprehensive dataset that includes various customer attributes, product types, and purchase behaviors, which can provide a more accurate representation of shopping trends.

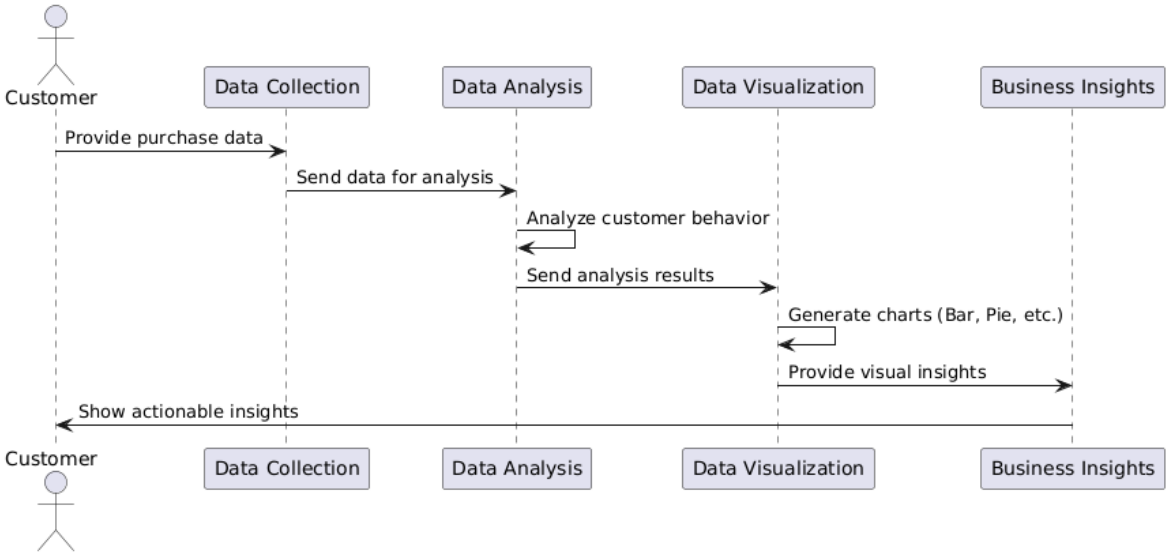
By addressing these gaps, this project will provide a more comprehensive understanding of shopping trends and offer actionable insights that businesses can apply to enhance customer experience and drive sales growth.

CHAPTER 3

Proposed Methodology

3.1 System Design

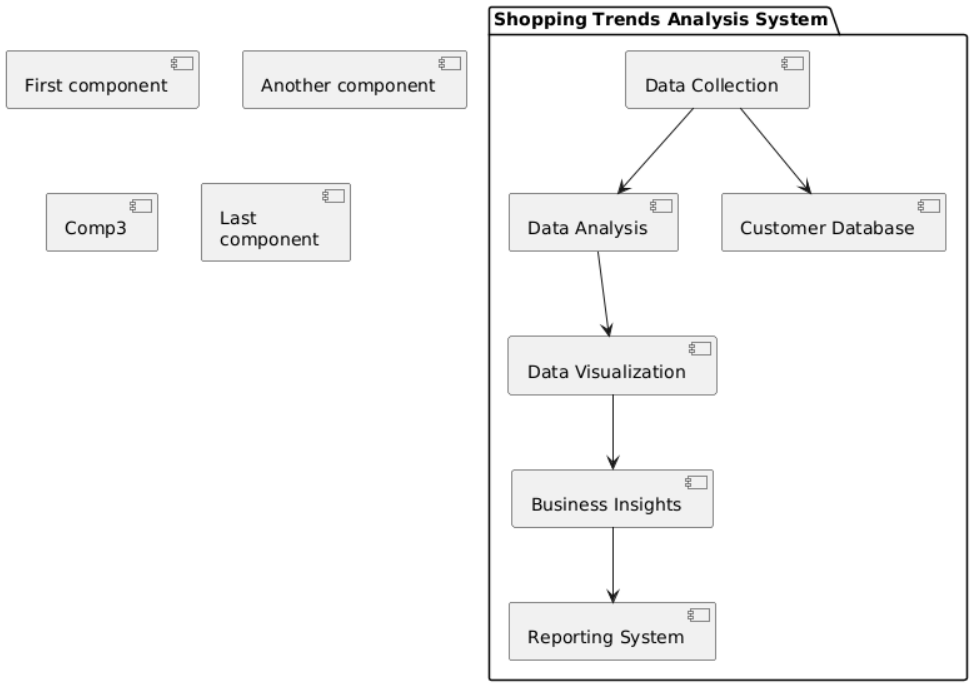
3.1.1 Sequence diagram



Explanation:

1. **Customer** provides purchase data to the **Data Collection** system.
2. The **Data Collection** component sends the data to **Data Analysis**.
3. The **Data Analysis** component performs the analysis, such as identifying purchase trends and customer behavior.
4. The analysis results are sent to **Data Visualization**, which generates charts (e.g., bar and pie charts).
5. The **Business Insights** component receives visual insights and provides them to the **Customer**, showing actionable results.

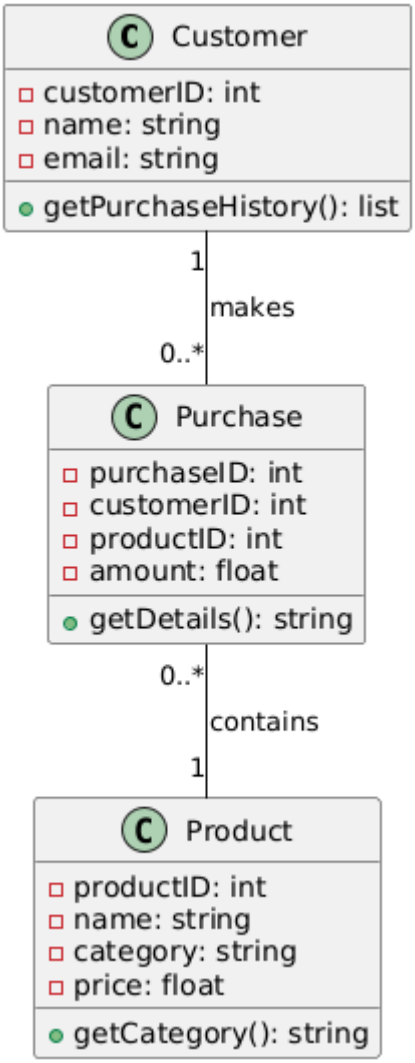
3.1.2 Component diagram



Explanation:

- The **Data Collection** component collects the data and passes it to **Data Analysis**.
- **Data Analysis** analyzes the data and sends results to **Data Visualization**.
- **Data Visualization** presents the insights visually and sends them to the **Business Insights** component.
- **Customer Database** stores customer information for analysis.
- **Reporting System** generates detailed reports based on the business insights.

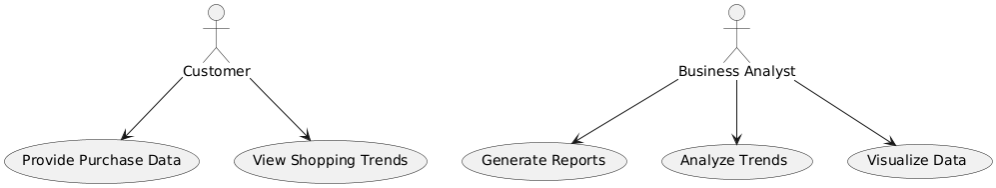
3.1.3 Class diagram



Explanation:

- The **Customer** class represents customers with attributes like customerID, name, and email. It has a method to retrieve the purchase history.
- The **Purchase** class represents each purchase with attributes like purchaseID, amount, and relationships with both Customer and Product.
- The **Product** class represents products with attributes like productID, name, and price.

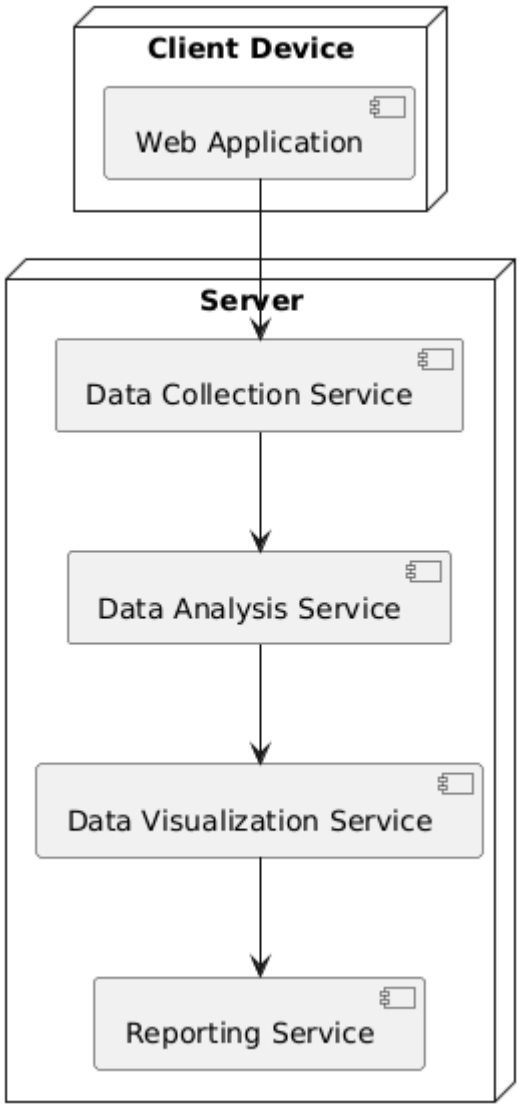
3.1.4 Use Case diagram



Explanation:

- **Customer** can provide purchase data and view shopping trends.
- **Business Analyst** can analyze trends, visualize data, and generate reports based on the insights.

3.1.5 Deployment diagram



Explanation:

- **Client Device** runs the **Web Application**, which interacts with the server.
- **Server** hosts the components responsible for data collection, analysis, visualization, and reporting.

3.2 Requirement Specification

3.2.1 Hardware Requirements:

1. **Processor:**
    - Minimum: Intel Core i3 or equivalent.
    - Recommended: Intel Core i5 or higher for faster processing, especially if handling large datasets.

2. **RAM:**
    - Minimum: 8 GB of RAM.
    - Recommended: 16 GB of RAM to ensure smoother performance, especially when performing large-scale data analysis and running complex visualizations.

3. **Storage:**
    - Minimum: 256 GB HDD.
    - Recommended: 512 GB SSD for faster data retrieval and quicker processing times, especially when working with large datasets.

4. **Graphics:**
    - Integrated graphics (Intel HD or similar) should be sufficient for generating visualizations and charts. However, if more intensive graphical outputs are required, consider a dedicated GPU.

5. **Networking:**
    - Basic internet connectivity for cloud access and data updates (if applicable), as well as for retrieving datasets if needed.

3.2.2 Software Requirements:

1. Operating System:

- Windows 10 (as per your setup), or macOS and Linux can also be used, but the project has been tailored for Windows 10.

2. Development Environment:

- IDE: Visual Studio Code (VS Code) for writing the code.
- Text Editor: VS Code itself acts as both an editor and IDE, but you can also use Notepad++ or Sublime Text for basic editing.
- Jupyter Notebook: To create and run Python scripts interactively and generate graphs and charts.
- Anaconda: A Python distribution that helps manage packages, dependencies, and environments (especially useful for data analysis projects).

3. Programming Languages:

- **Python 3 : The primary language for data analysis, statistics, and scripting.**
    - **Libraries:**
        - **Pandas:** For data manipulation and analysis.
        - **NumPy:** For numerical operations and handling arrays.
        - **Matplotlib & Seaborn:** For generating visualizations like bar charts, pie charts, etc.
        - **SciPy:** For scientific and technical computing.
        - **Scikit-learn (optional):** For machine learning (if you plan to implement any predictive models or clustering in the future).
        - **Plotly or Bokeh (optional):** For interactive visualizations.

4. Database:

- **SQLite** (as mentioned in your project) or **PostgreSQL** (if you decide to switch in the future for more advanced features).

5. Version Control System:

- **Git:** For version control, ensuring proper management of your project code.
- **GitHub or GitLab:** For hosting your project repository online and collaborating (if needed).

6. Data Visualization Tools:

- **Matplotlib:** A basic library for plotting static graphs (e.g., bar charts, pie charts).
- **Seaborn:** For more advanced statistical visualizations and improving the aesthetics of charts.
- **Plotly:** If you require interactive charts or web-based dashboards.
- **Tableau or Power BI (optional):** These are powerful tools for creating interactive dashboards and visualizing data in a more business-friendly way.

7. Web Framework (optional):

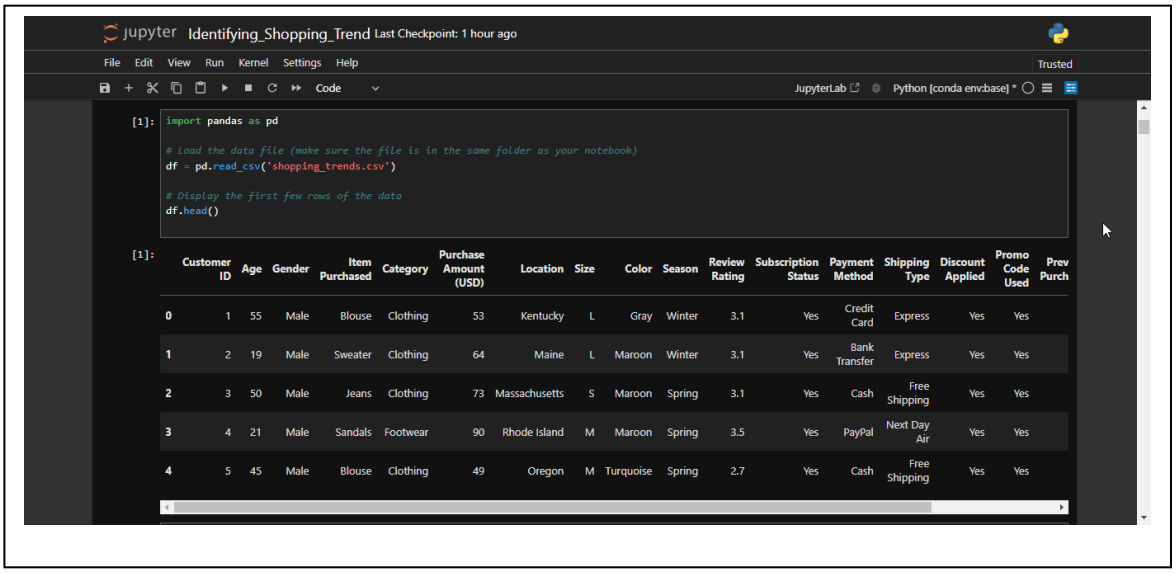
- **Flask or Django:** If you want to extend the project to a web application. Since your project is based on analysis, this step would be optional unless you want to provide a user interface for interacting with the results.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

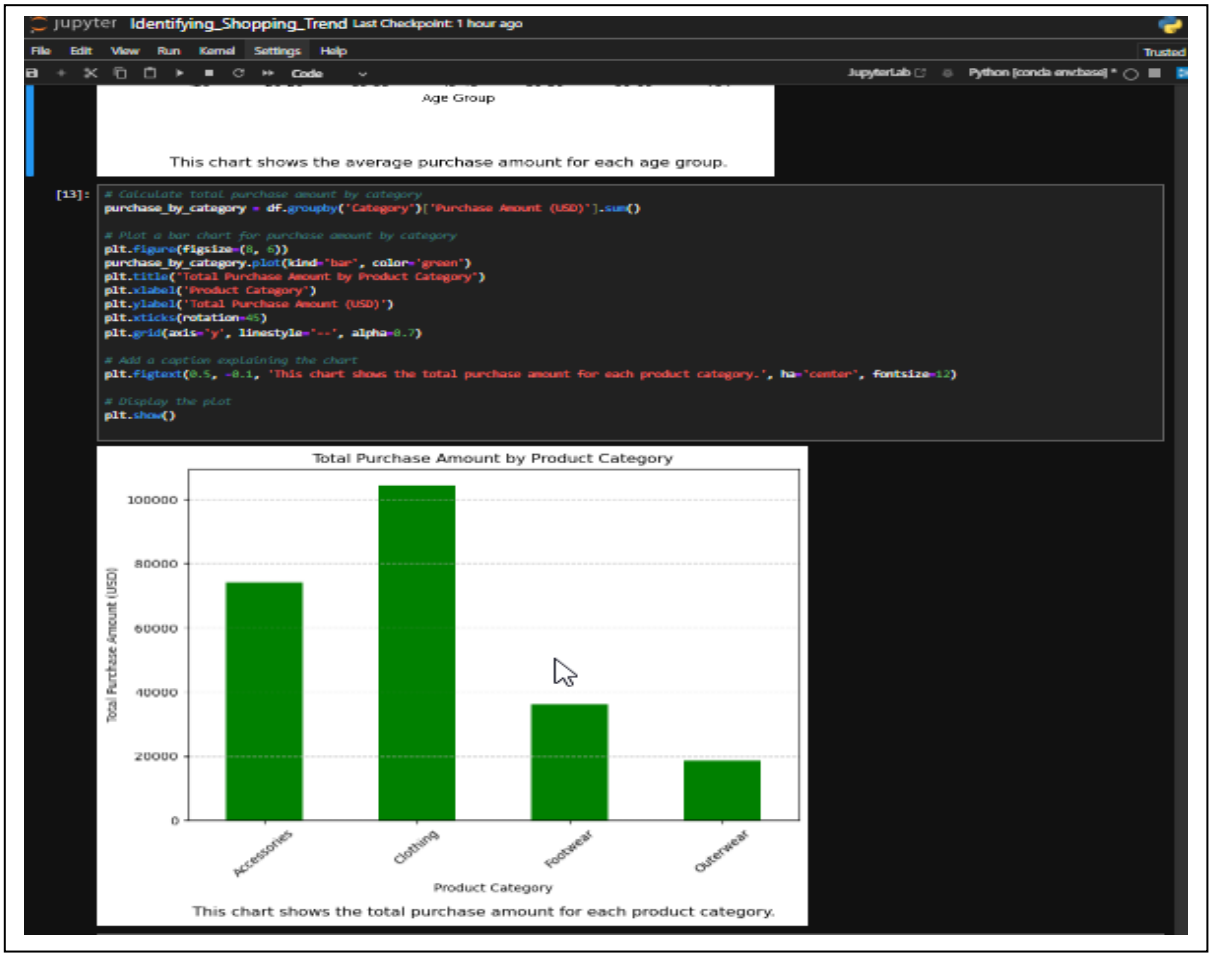
4.1.1 Showing the First few rows of Dataset



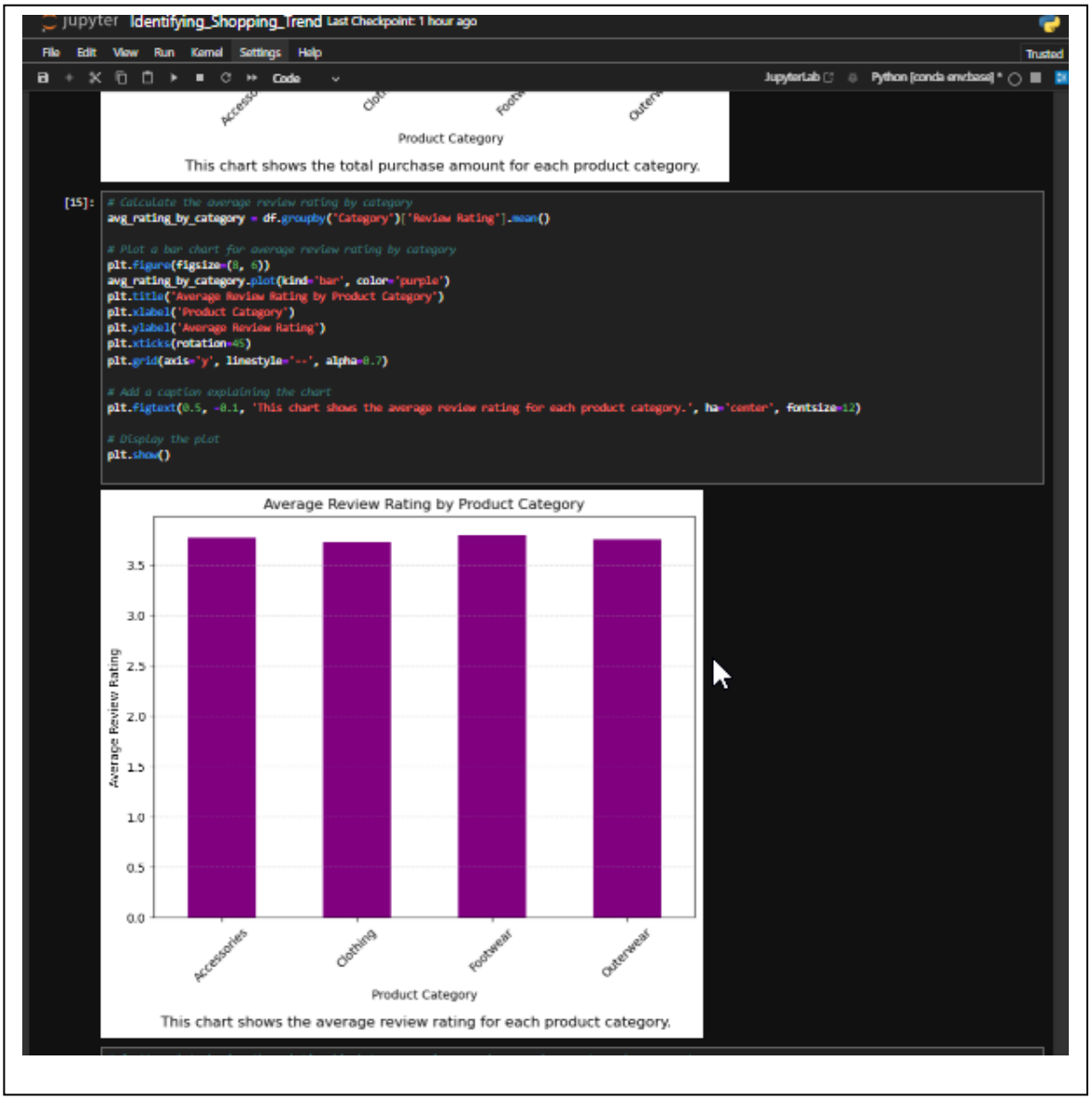
	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shipping Type	Discount Applied	Promo Code Used	Prev Purch
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Express	Yes	Yes	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Express	Yes	Yes	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Free Shipping	Yes	Yes	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Day Air	Yes	Yes	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Free Shipping	Yes	Yes	

4.1.2 The average purchase amount for each age group

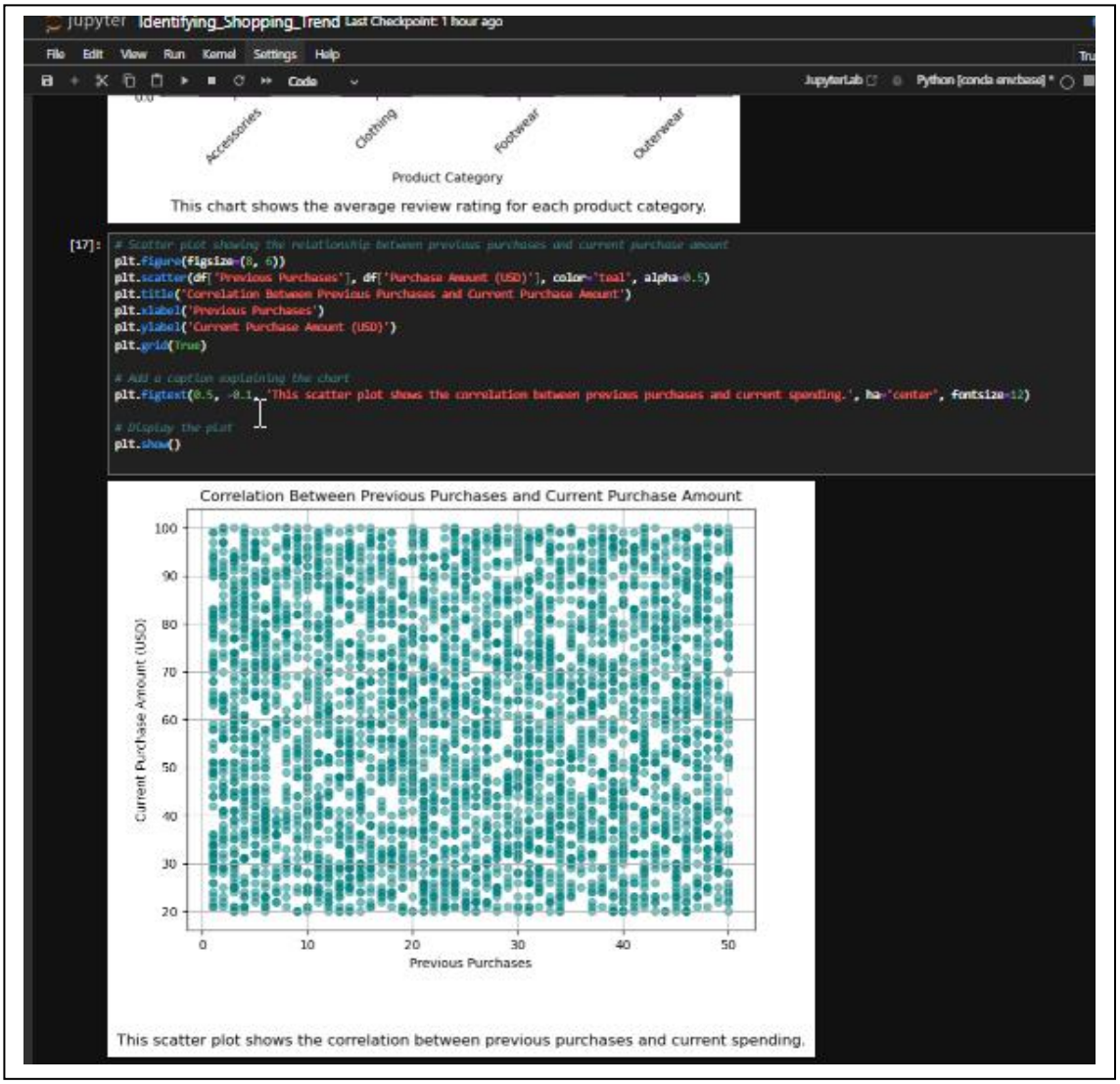
4.1.3 Total purchase amount for each product category



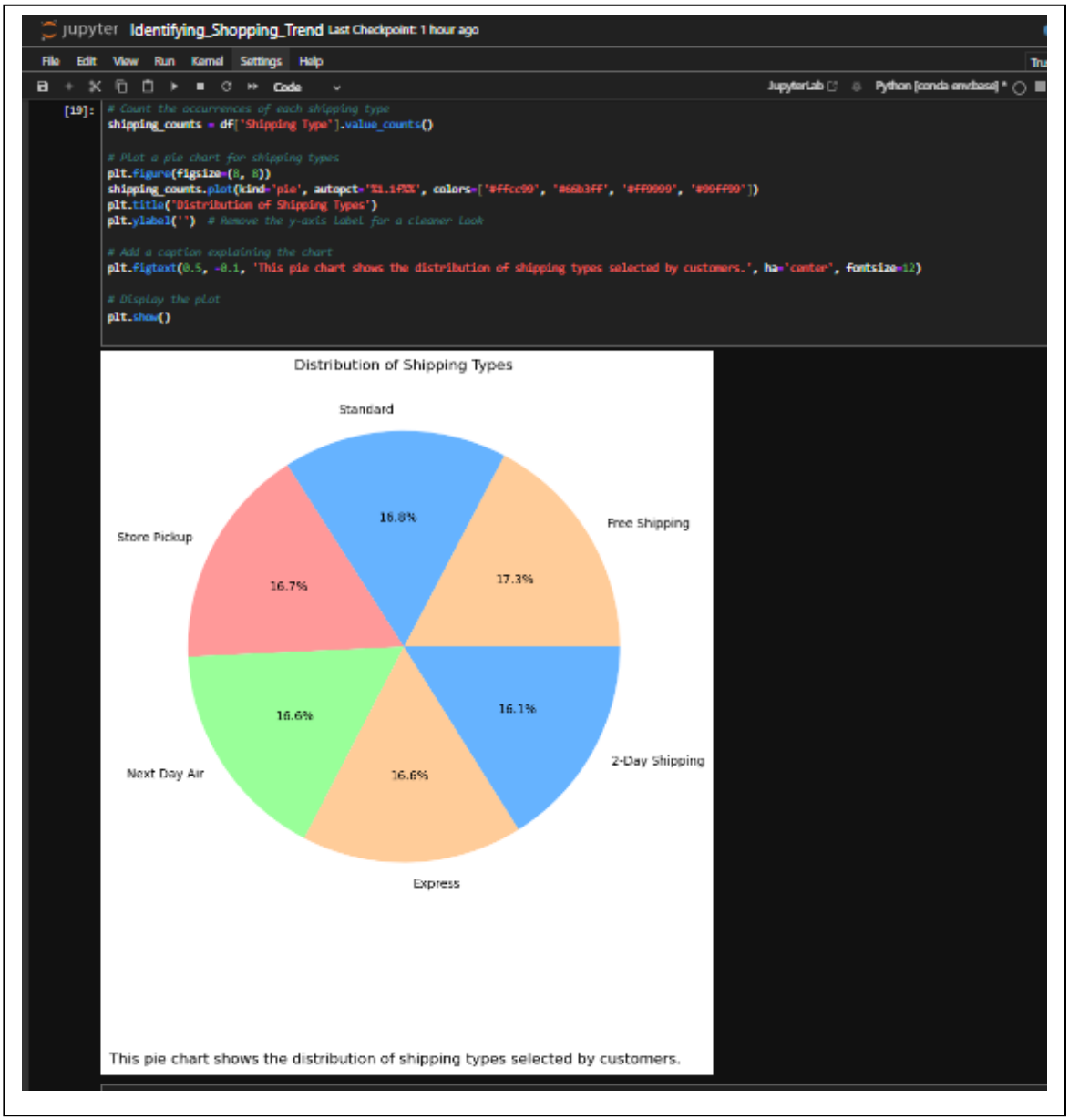
4.1.4 The average review rating for each product category



4.1.5 Scatter plot shows the correlation between previous purchases and current spending



4.1.6 Pie chart showing the distribution of shipping types selected by customers



Result Figures :

Figure 1 : Top Selling Categories

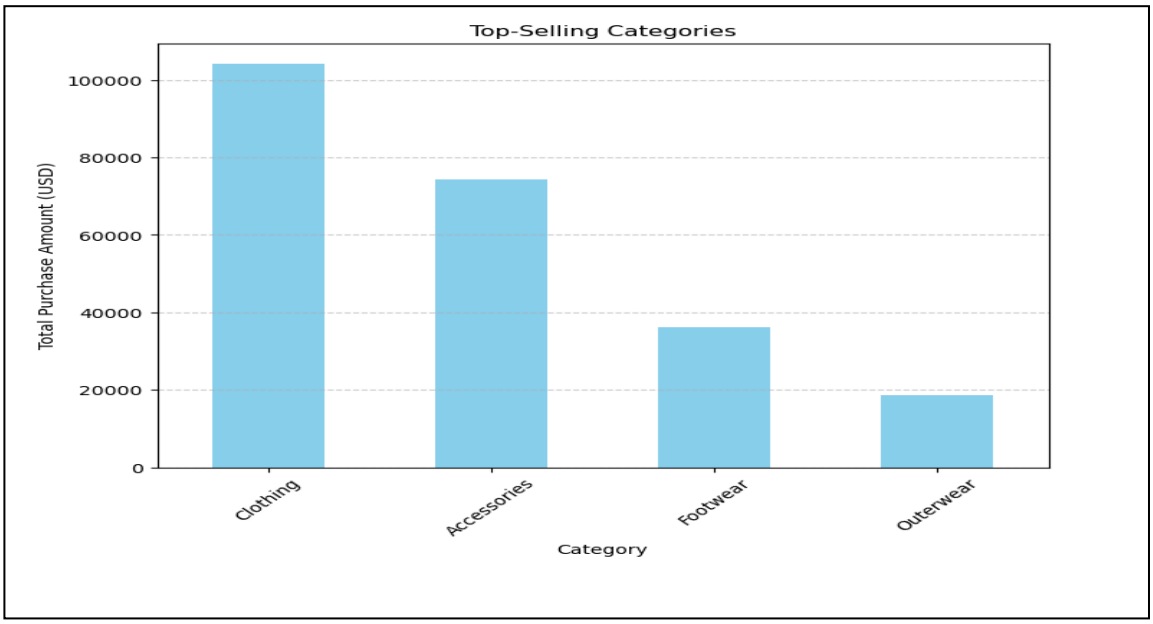


Figure 2 : Distribution of Payment Methods

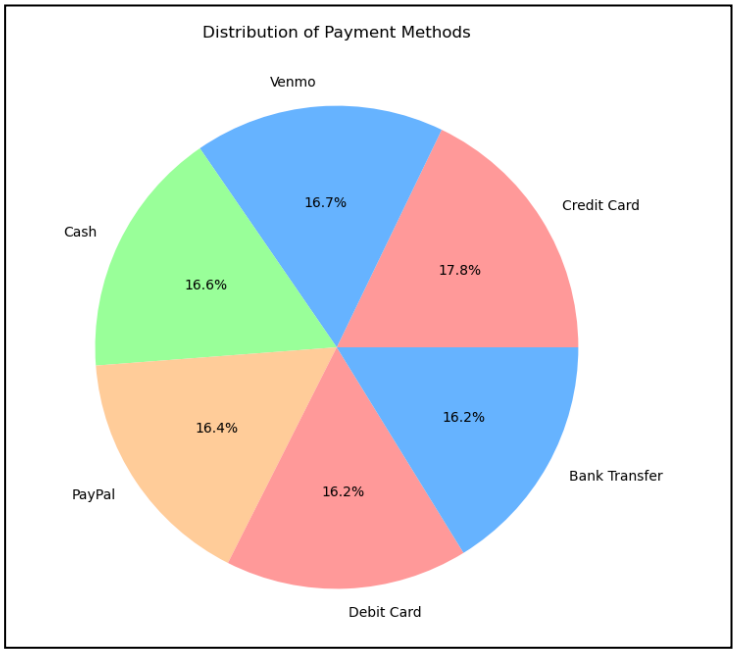


Figure 3 : Average Purchase Amount by Age Group

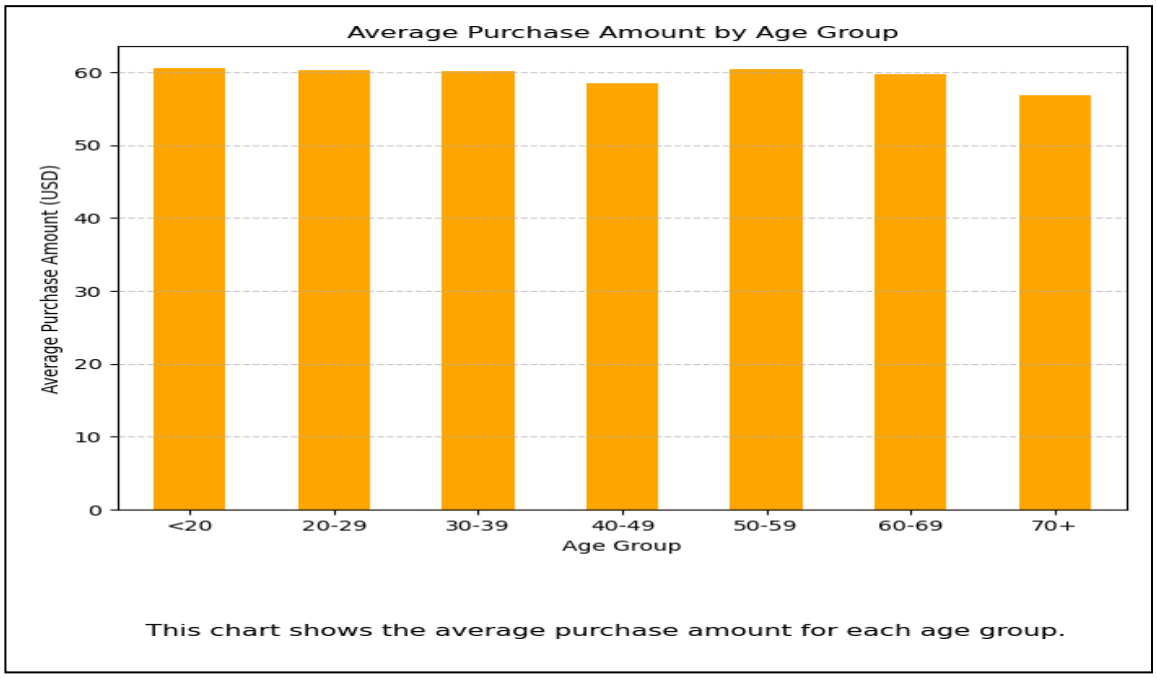


Figure 4 : Total Purchase Amount by Product Category

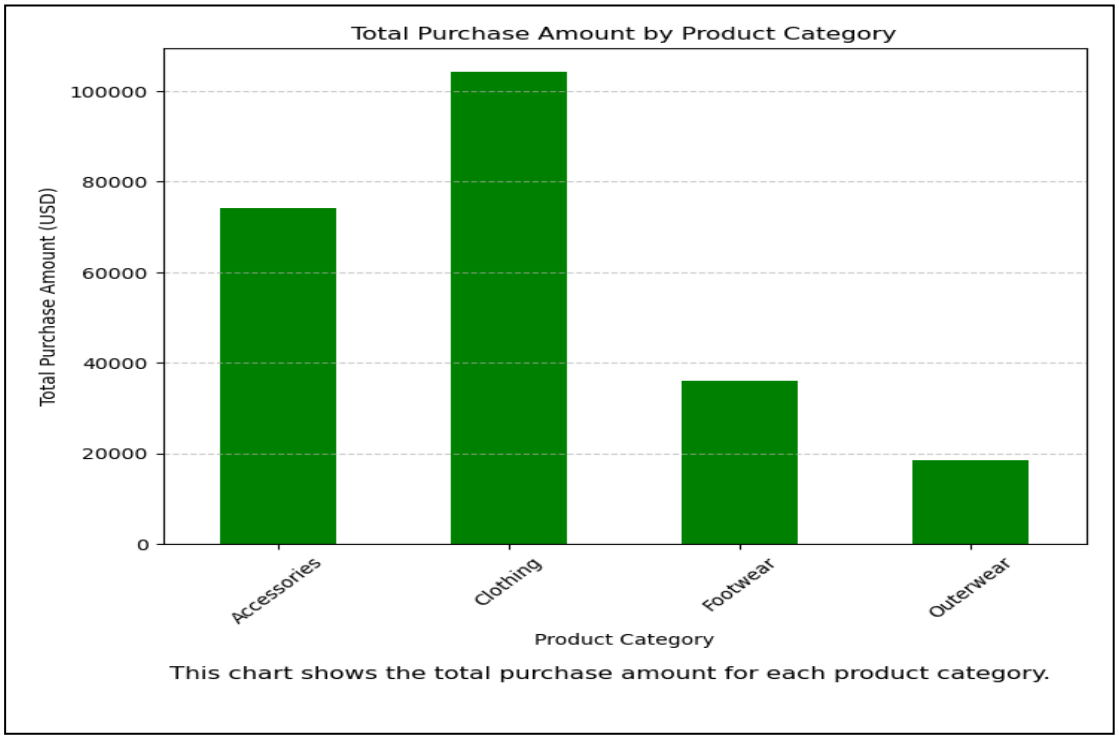


Figure 5 : The average review rating for each product category

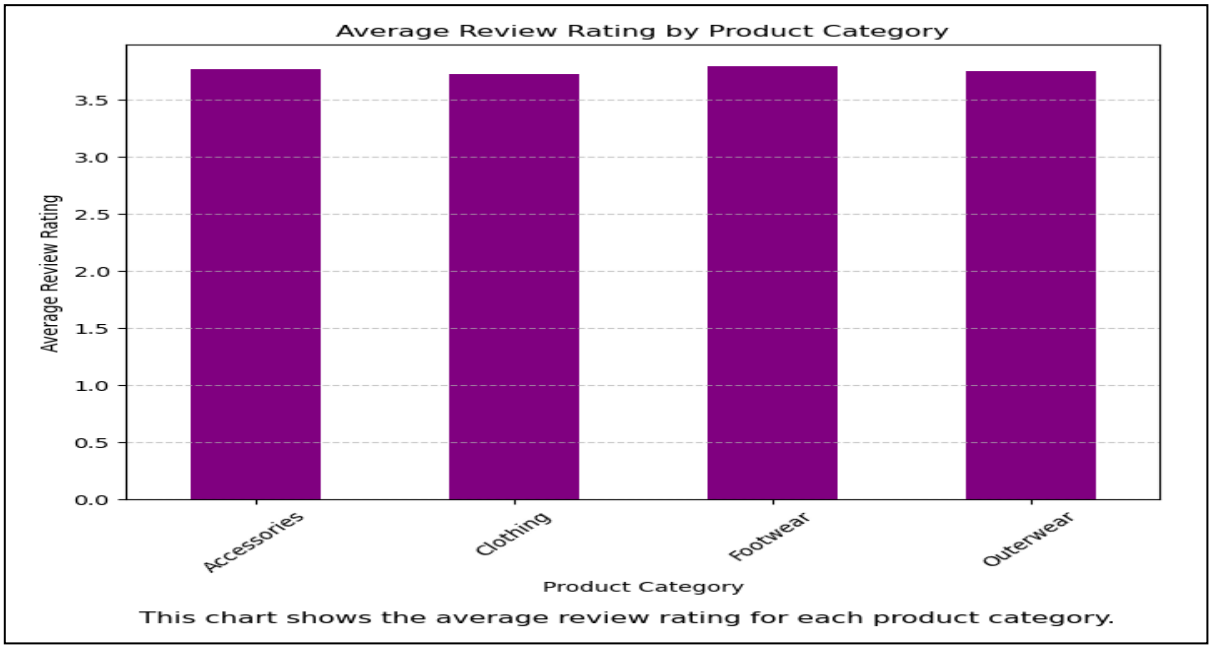


Figure 6 : Correlation between Previous Purchases and Current Purchases Amount

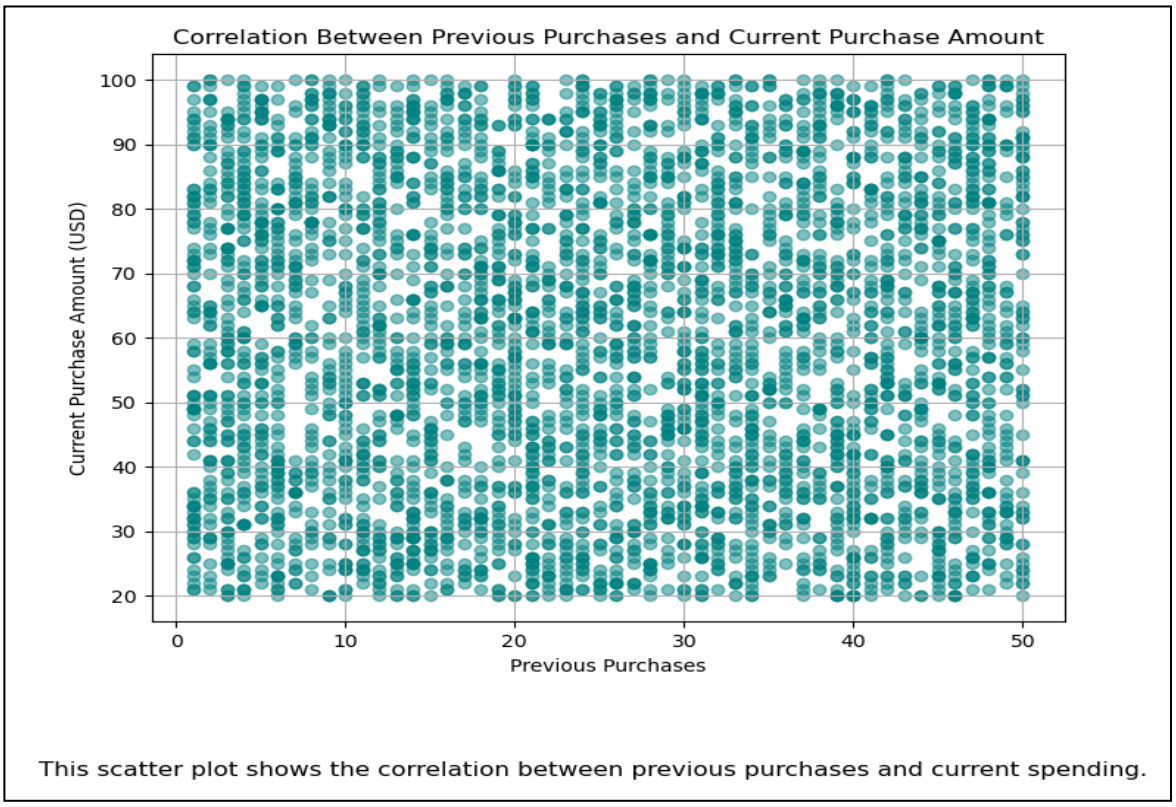


Figure 7 : Distribution of Shopping Types Selected by Customers

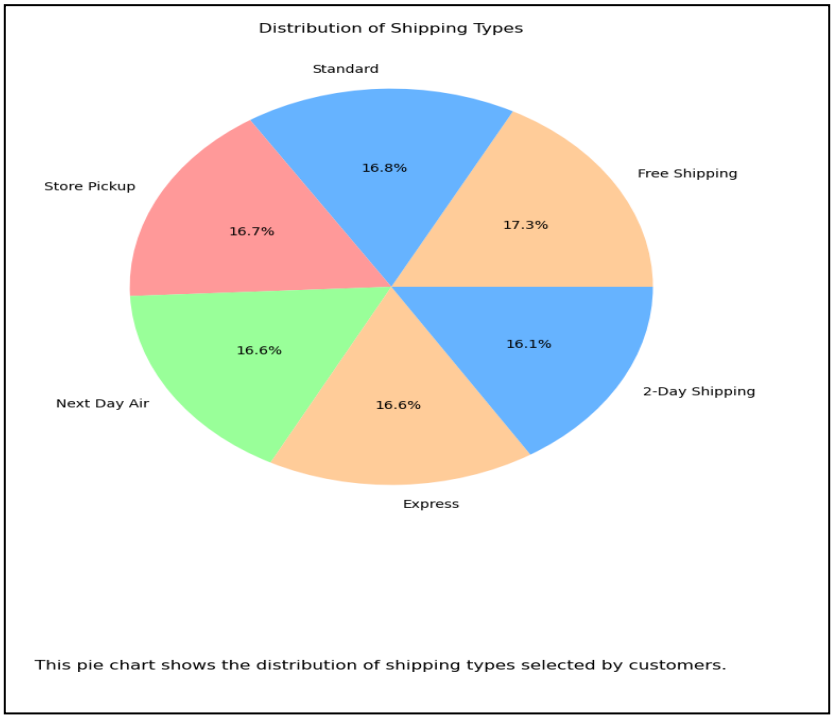


Figure 8 : Total Purchase Amount by Season

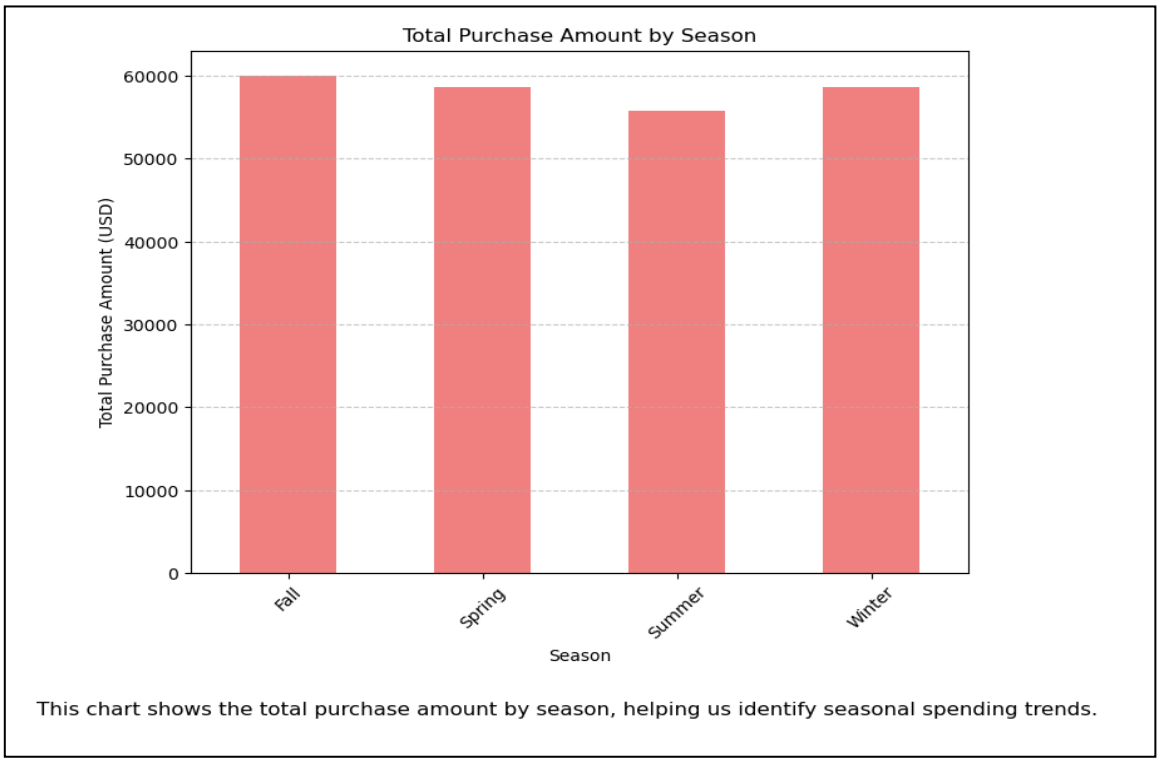
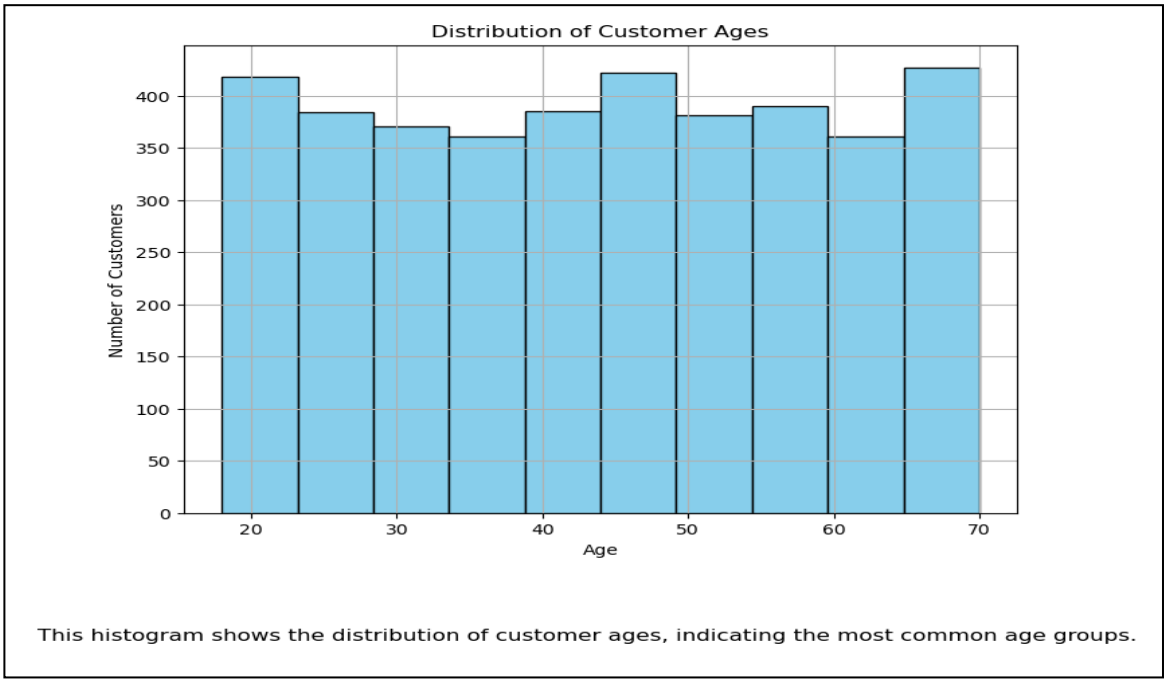
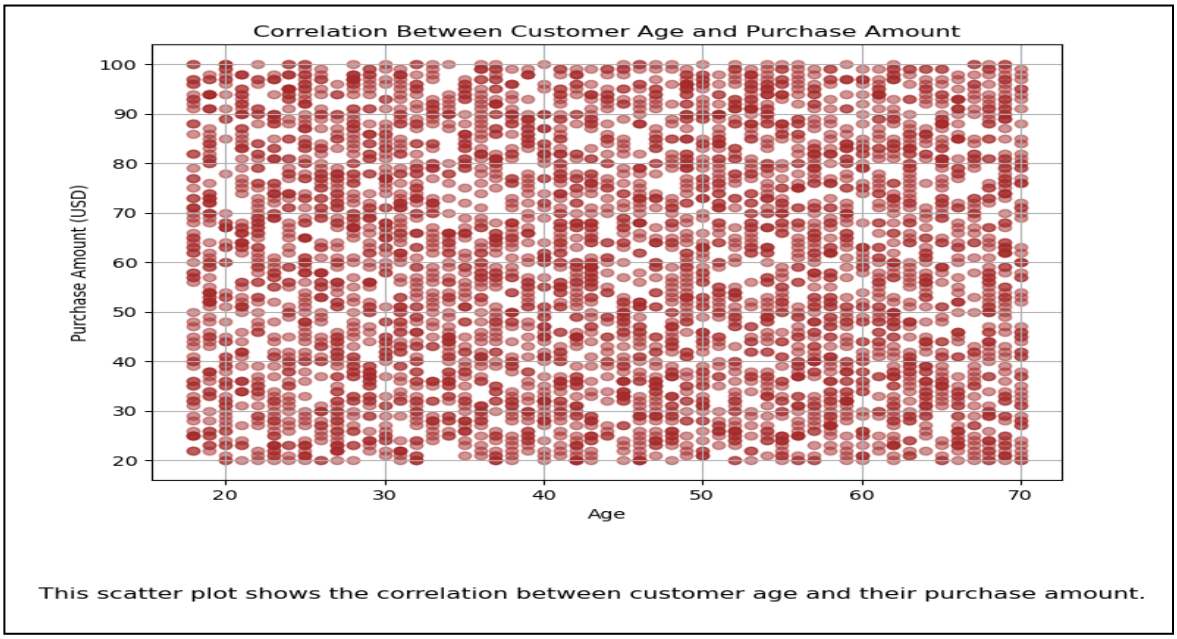


Figure 9 : Distribution of Customer Ages



This histogram shows the distribution of customer ages, indicating the most common age groups.

Figure 10 : Correlation Between Customer Age and Purchase Amount



This scatter plot shows the correlation between customer age and their purchase amount.

Figure 11 : Purchase Frequency by Payment Method

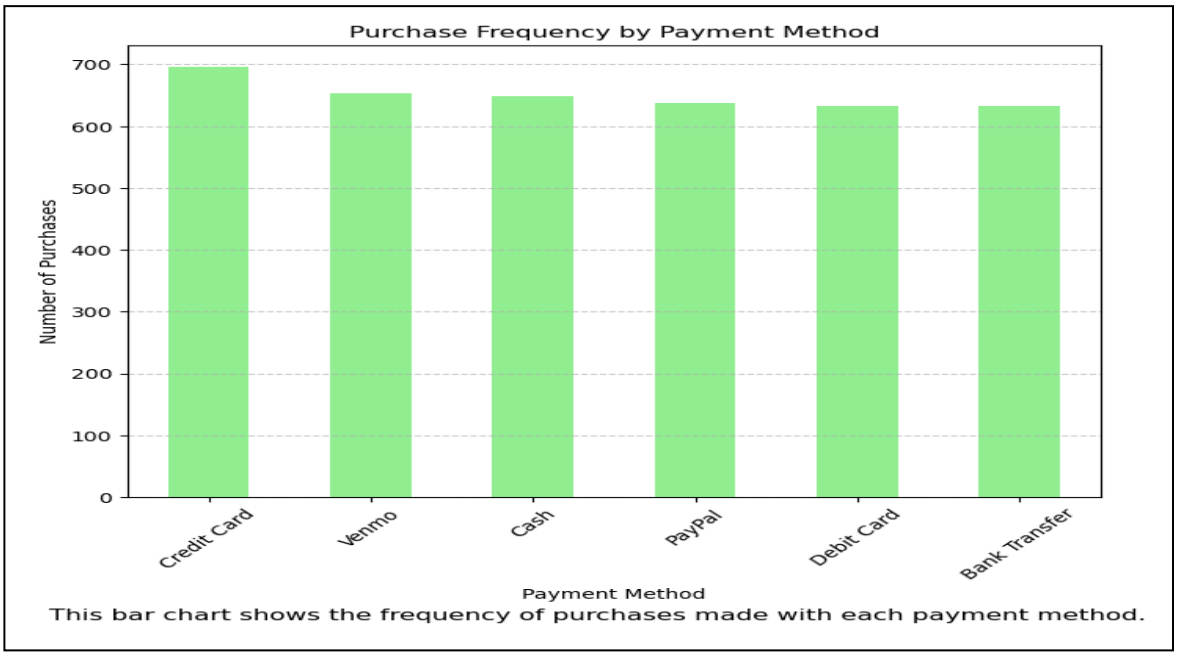


Figure 12 : Average Review Rating by Product Category

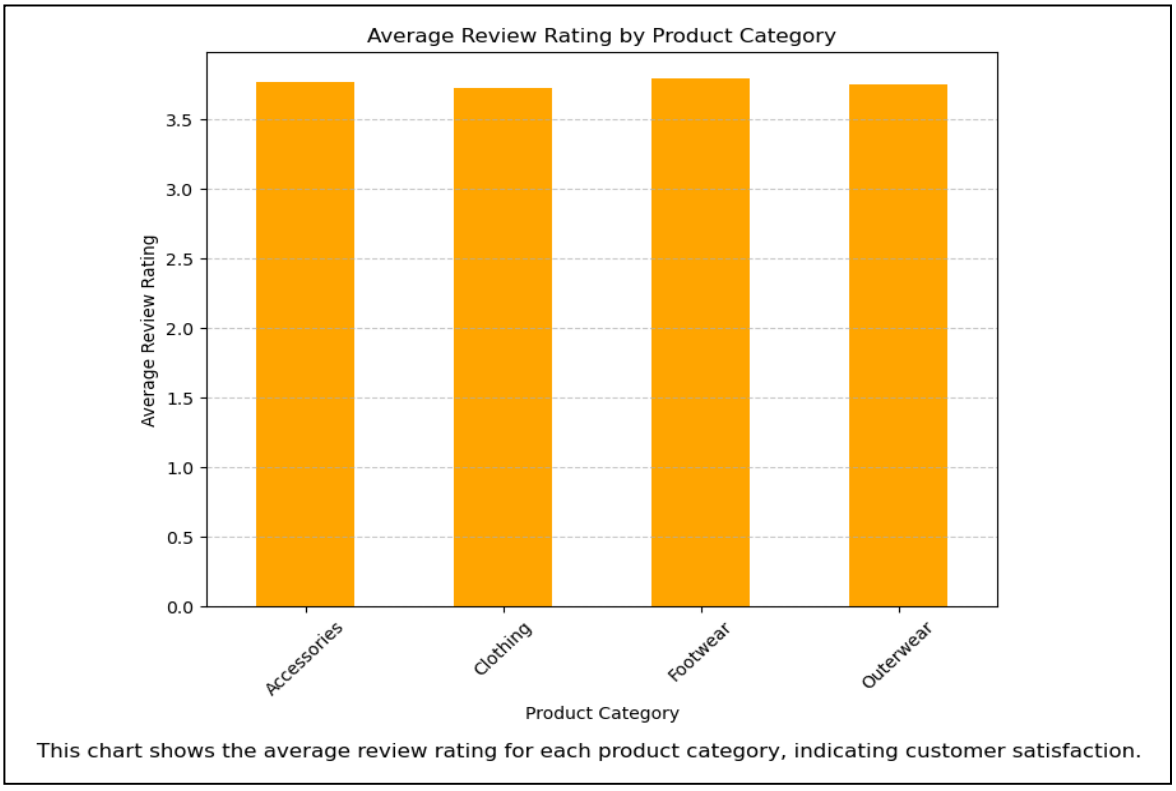
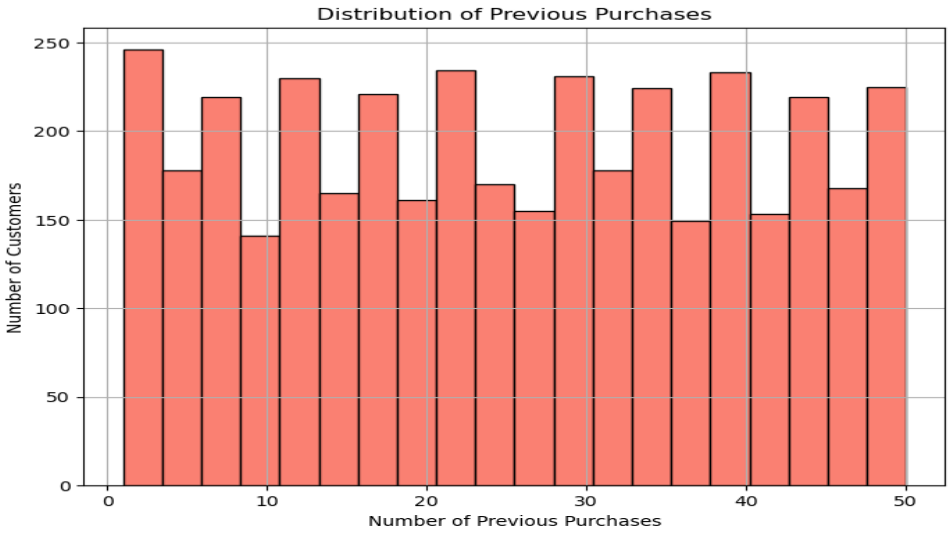
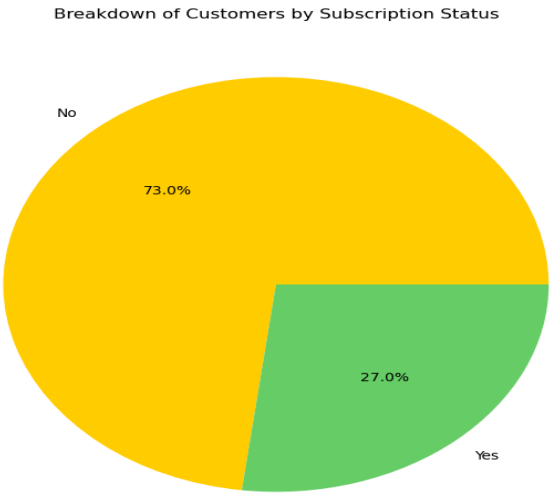


Figure 13 : Distribution of Previous Purchases



This histogram shows the distribution of the number of previous purchases made by customers.

Figure 14 : Breakdown of Customer by Subscription Status



This pie chart shows the breakdown of customers by their subscription status.

Figure 15 : Effect of Discount on Purchase Amount

Figure 16 : Total Purchase Amount by Location

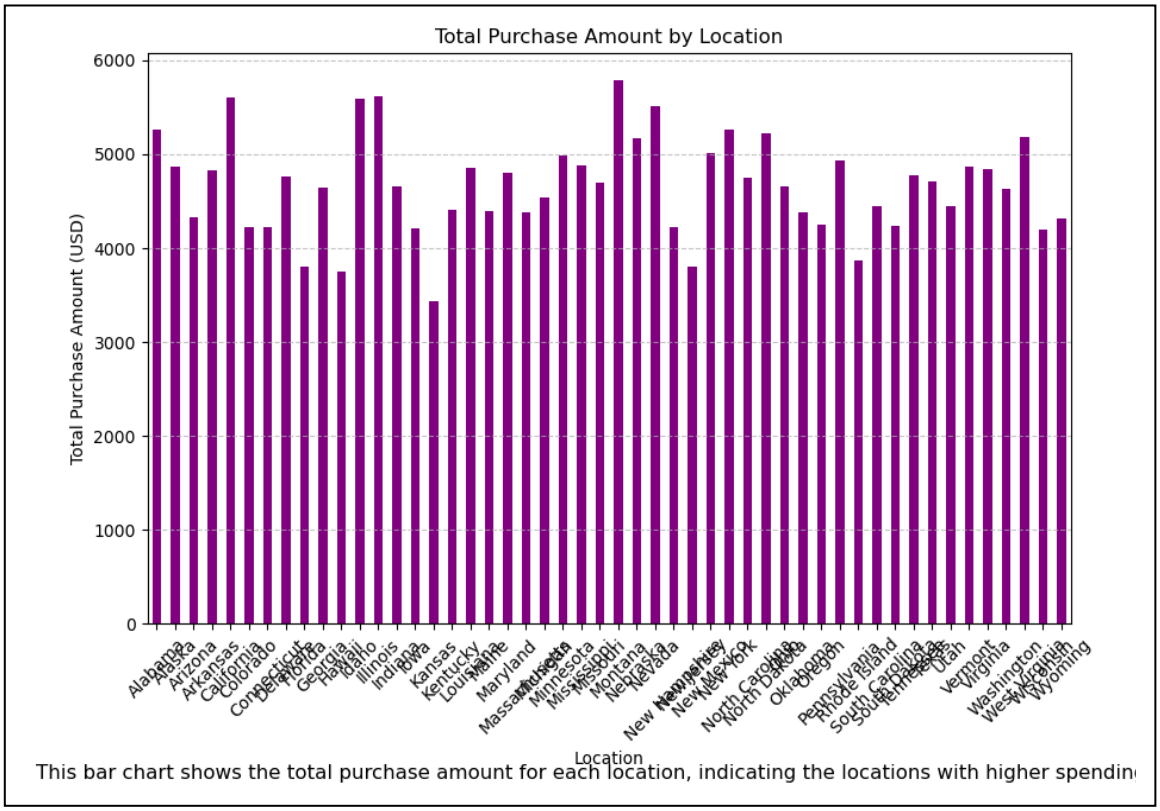


Figure 17 : Top 10 Most Common Products Purchased

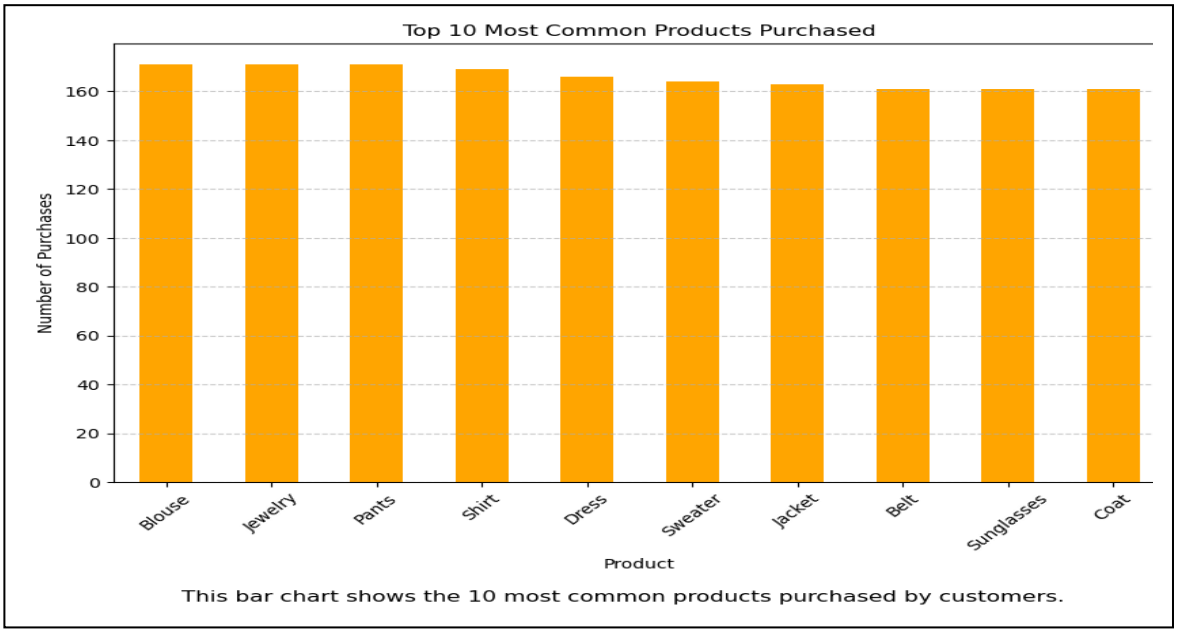


Figure 18 : Payment Method vs Frequency of Purchases

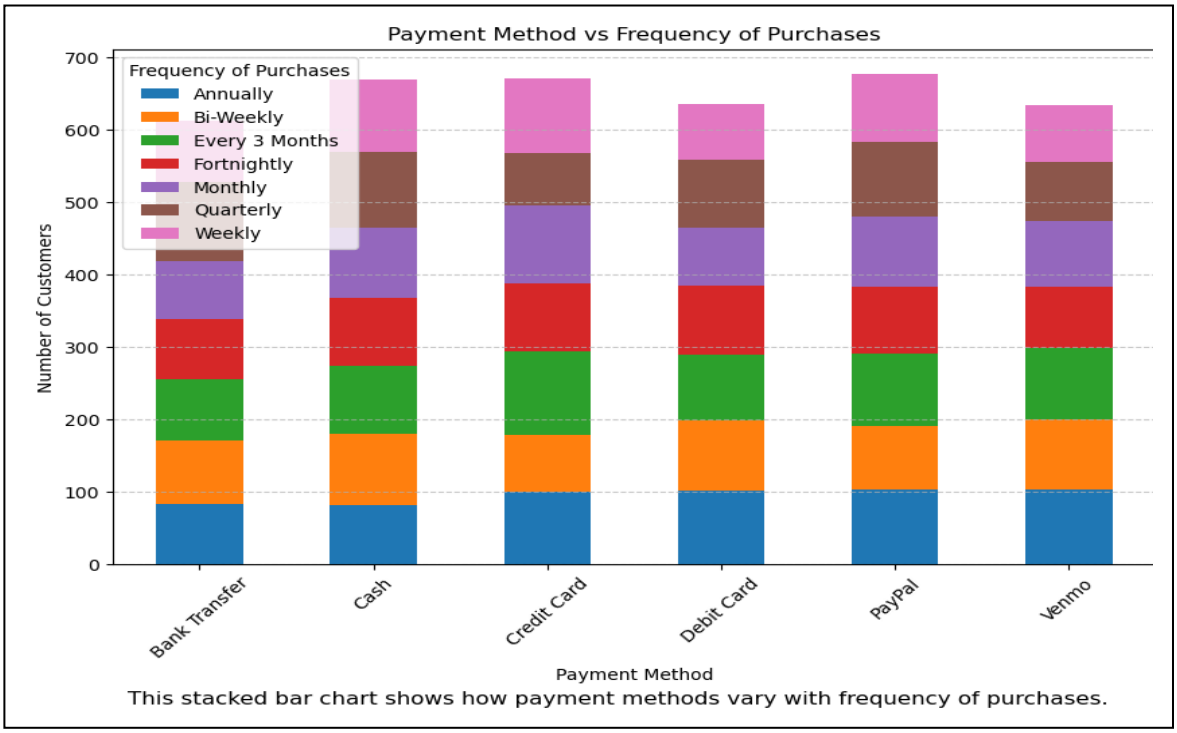
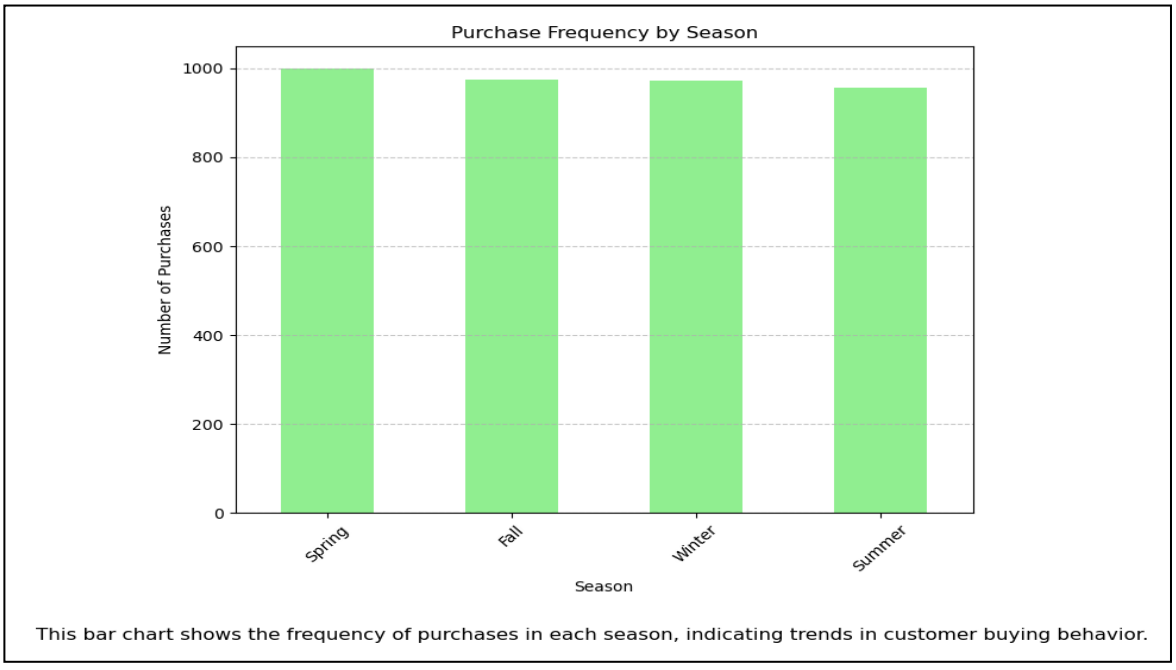


Figure 19 : Average Review Rating by Product Category

Figure 20 : Purchase Frequency by Season



4.2 GitHub Link for Code :

https://github.com/ganeshjadar2004/Identifying_Shopping_Trend.ipynb

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

While the project successfully identifies shopping trends and provides insightful visualizations, there are several areas where it can be expanded and improved:

1. **Integration with Machine Learning Models :**

    - Incorporating machine learning algorithms to predict future shopping trends based on historical data.
    - Clustering customers into segments based on purchasing patterns using techniques like K-Means or hierarchical clustering.

2. **Interactive Dashboards :**

    - Building interactive dashboards using tools like Plotly Dash, Tableau, or Power BI to allow users to explore data dynamically.

3. **Expanded Dataset:**

    - Incorporating larger datasets from multiple sources, such as e-commerce platforms or retail stores, to make the analysis more robust.
    - Including additional attributes like product ratings, seasonal trends, and promotional campaigns for more comprehensive insights.

4. **Web Application or Mobile App Integration:**

    - Developing a user-friendly web or mobile application to present the analysis results to end users.
    - Allowing users to upload their own datasets for on-the-fly analysis.

5. **Real-Time Analysis:**

    - Implementing real-time data analysis for dynamic insights, especially for businesses aiming to adjust strategies instantly.
    - Integration with cloud services like AWS or Azure for scalable real-time processing.

6. **Addressing Limitations:**

    - Handling missing or incomplete data more effectively using advanced imputation techniques.
    - Exploring ways to reduce biases in the dataset for fair and balanced analysis.

5.2 Conclusion:

The project on identifying shopping trends has demonstrated the potential of data analysis in understanding customer behaviors and preferences. By leveraging tools like Python, Pandas, and Matplotlib, the project successfully extracted valuable insights from a sample dataset, including:

1. Identifying purchasing patterns by age group and gender.
2. Highlighting seasonal trends in product categories.
3. Analyzing regional preferences and their impact on purchase decisions.
4. Evaluating customer feedback and review ratings to understand product satisfaction.

The project emphasizes the importance of data-driven decision-making in the retail industry. These insights can help businesses optimize their inventory, personalize marketing campaigns, and improve customer satisfaction.

However, there is room for improvement, as outlined in the future work section. By incorporating machine learning models, interactive dashboards, and real-time analysis, this project can evolve into a comprehensive tool for businesses to stay competitive in the market.

In conclusion, this project serves as a strong foundation for further exploration of shopping trends and offers practical applications for retail businesses, marketing analysts, and data enthusiasts. It highlights the transformative impact of data analysis in driving business growth and customer-centric strategies.

REFERENCES

1. Pandas Documentation:
   https://pandas.pydata.org/docs/

2. Matplotlib Documentation:
   https://matplotlib.org/

3. Kaggle Datasets:
   https://www.kaggle.com/

4. Python for Data Analysis by Wes McKinney.

5. Seaborn Documentation:
   https://seaborn.pydata.org/

6. Towards Data Science Blog:
   https://towardsdatascience.com/

7. Stack Overflow:
   https://stackoverflow.com/

8. Anaconda Documentation:
   https://docs.anaconda.com/