**Data Analysis of Shopping Trends**

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

**Shashank Dattu**

**dattushashanka@gamil.com**

Under the Guidance of

**Jay Rathod**

**ACKNOWLEDGEMENT**

#### I would like to take this opportunity to express my heartfelt gratitude to my mentor, Jay Rathod and P Raja sir, for their unwavering guidance and support throughout the course of this project on data analysis of shopping trends. Their deep expertise, constructive feedback, and encouragement have been instrumental in shaping the direction of this study and enhancing the quality of my work.

#### From the initial brainstorming sessions to the final stages of analysis, my mentors has provided invaluable insights that not only improved my technical skills but also enriched my understanding of the complexities involved in analyzing real-world shopping trends. Their ability to simplify challenging concepts and their constant availability for discussions have made this learning journey both inspiring and fulfilling.

#### This project has been a significant milestone in my academic and personal growth, and I attribute much of its success to Mr. Jay Rathod and Mr. P Raja's mentorship. Their patience, motivation, and dedication to fostering a collaborative and supportive environment encouraged me to explore innovative approaches and think critically about data-driven decision-making.

#### I would also like to extend my gratitude to Microsoft and AICTE for their support, whether it be through resources, encouragement, or constructive suggestions.

#### Finally, I am deeply thankful for the opportunity to immerse myself in the fascinating domain of data analysis, which has been both challenging and rewarding. This project has been a stepping stone toward achieving my aspirations in data analysis, and I am profoundly grateful to all those who contributed to its success.

#### Sincerely, Shashank Dattu

#### **ABSTRACT**

This project focuses on analyzing shopping trends using a dataset comprising customer demographics, purchase behaviors, and transactional details. The objective was to identify key factors influencing shopping behavior, predict subscription status, and derive actionable insights to enhance decision-making in retail strategies. The study began with data collection and preprocessing, including cleaning, handling missing values, and encoding categorical variables. Exploratory Data Analysis (EDA) was conducted to understand customer demographics, purchase patterns, and seasonal trends. Visualizations such as histograms, heatmaps, and bar charts were employed to analyze variables like age distribution, purchase frequency, and category preferences. Machine learning techniques were applied to predict customers' subscription status using a Random Forest Classifier. Features included demographic variables (e.g., age, gender) and transactional data (e.g., purchase amount, previous purchases). The model achieved a notable accuracy score, with feature importance highlighting key contributors such as "Purchase Amount" and "Previous Purchases."

Key findings include:

1.Younger customers showed higher purchase frequencies but lower average purchase amounts.

2. Seasonal trends indicated a significant rise in sales during festive periods.

3. Discount applications and promo codes significantly influenced purchase amounts.

4. Gender and category preferences varied, with notable differences in purchase behavior.

In conclusion, this analysis provides valuable insights into customer behavior and purchasing trends. Retailers can use these findings to optimize marketing strategies, personalize customer experiences, and improve subscription engagement. The integration of EDA with machine learning demonstrates a robust approach to deriving data-driven insights in the retail domain.

**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**

**Chapter 4.**  **Implementation and Results**

**Chapter 5. Discussion and Conclusion**

**References**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Caption** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table. No.** | **Table Caption** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CHAPTER 1**

**Introduction**

* 1. **Problem Statement:**

In the competitive retail landscape, understanding customer shopping trends is crucial for improving customer engagement, optimizing marketing strategies, and increasing revenue. Retail businesses often face challenges in identifying key factors that influence purchasing decisions, predicting customer behaviors such as subscription status, and tailoring strategies to meet customer demands. The availability of large datasets containing customer demographics, purchasing patterns, and transactional details provides an opportunity to address these challenges. However, effectively analyzing such datasets to extract actionable insights requires robust data processing, exploratory analysis, and predictive modeling techniques. This project aims to address these challenges by conducting a comprehensive analysis of shopping trends. The primary focus is to identify patterns in customer demographics, purchasing behavior, and seasonal trends, while leveraging machine learning to predict customer subscription status. The insights derived will help retailers make data-driven decisions to enhance customer satisfaction and optimize business performance.

* 1. **Motivation:**

The motivation behind this project stems from the growing importance of data-driven decision- making in the retail industry. With the rapid growth of e-commerce and brick-and-mortar stores, retailers face an overwhelming amount of customer data. However, transforming this data into actionable insights remains a significant challenge. Understanding customer shopping trends, preferences, and behaviors is essential to maintaining a competitive edge, improving customer experiences, and driving revenue growth.

Several key factors inspired this project:

I. Improving Customer Understanding: Retailers need to segment customers, identify their

preferences, and understand what influences their purchasing decisions. This knowledge allows

businesses to offer personalized experiences, increasing customer loyalty.

2. Optimizing Marketing and Sales Strategies: By uncovering trends such as the impact of

discounts, promotional codes, and seasonal variations, businesses can design targeted

campaigns that are more effective in reaching their audience.

3. Leveraging Data for Predictive Insights: Predictive modeling, such as forecasting subscription

likelihood, can empower businesses to retain customers and focus on those likely to engage

further with their services.

4. Bridging the Gap Between Data and Decision-Making: With the availability of sophisticated

tools and techniques for data analysis, there is a growing need to bridge the gap between raw

data and actionable business strategies.

This project seeks to contribute to the field by exploring how data analysis and machine learning can uncover patterns in shopping trends and provide insights that can be applied across retail operations. It demonstrates the value of harnessing data science to address real-world business challenges.

* 1. **Objective:**

The objective of this project is to analyze shopping trends using customer demographic and

transactional data to uncover key patterns, factors, and insights that influence purchasing behavior. Additionally, the project aims to leverage machine learning to predict customer subscription status, providing actionable insights that can enhance decision-making, optimize marketing strategies, and improve overall business performance in the retail industry.

Specific goals include:

1. Data Cleaning and Exploration: Process and clean the dataset to handle missing values,

duplicates, and inconsistencies while conducting exploratory data analysis (EDA) to identify trends and patterns.

2. Customer Behavior Analysis: Investigate the impact of factors such as age, gender, category preferences, seasonal trends, and discounts on customer purchases.

3. Visualization of Trends: Use data visualization techniques to represent key insights, such as category distributions, purchase behaviors, and review ratings.

4. Predictive Modeling: Build a machine learning model (Random Forest Classifier) to predict subscription status based on customer demographics, purchasing history, and other relevant features.

5. Feature Importance Analysis: Identify the most significant factors contributing to customer subscription decisions and overall purchase trends.

6. Actionable Insights: Provide recommendations for optimizing retail strategies, including

targeted marketing, inventory management, and customer retention initiatives.

By achieving these objectives, the project aims to demonstrate the value of data-driven approaches in enhancing retail operations and creating personalized, customer-centric strategies.

* 1. **Scope of the Project:**

1. **Data Analysis and Trend Identification**:

* Conduct in-depth analysis of customer demographic data, purchasing patterns, and seasonal trends to identify key behaviours and preferences.
* Visualize trends using charts, graphs, and heatmaps for better understanding and communication of findings.

2. **Predictive Modelling**:

* Develop a machine learning model (Random Forest Classifier) to predict customer subscription status based on demographic, behavioural, and transactional features.
* Analyz feature importance to identify key drivers influencing subscription decisions.

3. **Retail Strategy Insights**:

* Provide actionable insights to optimize marketing campaigns, inventory management, and personalized customer experiences.
* Identify customer segments that require targeted strategies, such as discount-focused customers or high-value subscribers.

4. **Exploratory Questions**:

* Address key retail questions, such as the influence of discounts, most purchased categories, and seasonal spending patterns.

**Limitations of the Project:**

1. **Dataset Quality**:
   * The accuracy of the findings depends on the quality and completeness of the dataset. Missing values, duplicates, or biases in the data may impact the analysis.
2. **Feature Scope**:
   * Some potentially influential variables (e.g., customer income, specific marketing campaigns, geographic locations) may not be available in the dataset, limiting the scope of insights.
3. **Generalizability**:
   * The results and insights are specific to the dataset used. The findings may not generalize to other retail settings or industries without additional validation.
4. **Machine Learning Model**:
   * The performance of the predictive model depends on the quality and quantity of data. Limited data points, imbalanced classes, or unaccounted features may reduce model accuracy.
5. **Real-Time Application**:
   * The project focuses on historical data analysis and prediction. Real-time applications, such as live recommendation systems, are outside the scope.
6. **Causal Relationships**:
   * While correlations and patterns are identified, the project does not establish causal relationships between variables (e.g., discounts and increased purchases).
7. **Behavioural Factors**:
   * The analysis does not consider external factors such as economic conditions, competitor actions, or customer sentiment, which may influence shopping trends.

By recognizing these limitations, the project aims to provide valuable insights within the defined scope while acknowledging areas for future improvement and research.

**CHAPTER 2**

**Literature Survey**

**2.1 Review of Relevant Literature or Previous Work in the Domain**

**The analysis of shopping trends and customer behaviour has been extensively studied in retail analytics. The following are notable contributions in this domain:**

1. **Customer Segmentation:** Previous studies have focused on segmenting customers based on demographics, purchasing patterns, and behaviour using clustering algorithms (e.g., K-means, hierarchical clustering). These segments help businesses understand different customer groups and target them effectively.
2. **Predictive Modelling in Retail:** Research has utilized machine learning algorithms such as decision trees, random forests, and logistic regression to predict customer churn, subscription status, and purchase intent.
3. **Impact of Discounts and Promotions:** Studies have shown the importance of pricing strategies, discounts, and loyalty programs in influencing customer purchases. Behavioural economics also highlights how customers react to promotions and seasonal discounts.
4. **Exploratory Data Analysis (EDA):** EDA techniques have been widely used to understand trends such as purchase frequency, preferred product categories, and seasonal patterns. Visualization tools like heatmaps and histograms are commonly employed for trend analysis.
5. **Feature Engineering for Retail Analytics:** Research often emphasizes the importance of feature engineering to improve the predictive power of models. Features such as "time since last purchase" or "discount sensitivity" have proven to be valuable in enhancing insights.

**2.2 Existing Models, Techniques, or Methodologies**

1. **Recommendation Systems:** Collaborative filtering and content-based filtering techniques are commonly used to suggest products based on customer preferences and past purchases.
2. **Machine Learning Models:**

Supervised learning models like Random Forest, Support Vector Machines, and Neural Networks are often used for classification tasks like subscription prediction.Unsupervised learning models, such as K-means, are used for customer segmentation.

1. **Correlation and Trend Analysis:** Correlation matrices and visualization techniques like bar charts, pie charts, and line plots are widely used to identify relationships between variables such as age, gender, and purchase amount.
2. **Sentiment Analysis:** Text analysis models are employed to analyze customer reviews and extract sentiments, helping businesses understand customer satisfaction.

**2.3 Gaps or Limitations in Existing Solutions**

**While significant work has been done in the domain of retail analytics, there are notable gaps and limitations:**

1. **Limited Data Integration:** Many studies rely on siloed datasets that focus on specific aspects (e.g., only transactional or demographic data), limiting a holistic understanding of customer behaviour.
2. **Generic Models:** Existing predictive models often use generalized features that may not capture the nuances of specific retail environments, such as seasonal influences or category-specific trends.
3. **Lack of Interpretability:** Advanced machine learning models often act as black boxes, making it difficult to explain the factors driving predictions and trends.
4. Dynamic Behaviour: Customer behaviour changes over time, influenced by external factors like economic trends or competitors’ actions. Existing models often fail to capture these dynamic behaviours.
5. **Underutilization of Visualization:** Although EDA is widely used, many studies lack a focus on actionable insights derived from visualizations, which limits their usability for business decision-making.

**How This Project Addresses the Gaps**

1. **Comprehensive Analysis:** This project integrates both customer demographic and transactional data to provide a holistic understanding of shopping trends.
2. **Custom Predictive Model:** A Random Forest Classifier tailored to predict subscription status based on both numerical and categorical features will be developed, addressing retail-specific challenges.
3. **Feature Importance and Interpretability:** The project will emphasize feature importance analysis, providing actionable insights into the key drivers of subscription and purchasing behaviours.
4. **Dynamic Insights:** Seasonal and category-specific trends will be highlighted, allowing businesses to adapt their strategies dynamically.
5. **Visualization for Actionable Insights:** Advanced visualizations (e.g., heatmaps, bar charts, line graphs) will be used to convey trends clearly, enabling better decision-making for business stakeholders.

By addressing these gaps, the project aims to make significant contributions to the domain of shopping trend analysis and customer behaviour prediction.

**CHAPTER 3**

**Proposed Methodology**

**3.1 System Design**

**1. Data Collection (CSV Dataset)**

**Objective:  
To load and examine the dataset for further analysis.**

**Tasks:**

**Load the CSV file: Read the dataset using pandas.read\_csv().**

**Initial Inspection:**

**View the first few records (data.head()) to understand the dataset structure.**

**Check data types and null values (data.info()).**

**Outcome:  
The dataset is successfully loaded and inspected to understand its structure, columns, and data types, which will guide further analysis and cleaning.**

**2. Data Preprocessing**

**Objective:  
To clean and prepare the data for analysis by handling missing values, duplicates, and encoding categorical variables.**

**Tasks:**

* **Handling Missing Values:**
  + **Check for missing values using data.isnull().sum() and decide how to handle them (e.g., imputation or removal).**
* **Removing Duplicates:**
  + **Check for and remove any duplicate rows using data.drop\_duplicates().**
* **Encoding Categorical Variables:**
  + **Convert categorical features (e.g., 'Gender', 'Category', 'Season', etc.) to numeric using LabelEncoder or pd.get\_dummies().**
* **Scaling Numerical Features:**
  + **Standardize numerical features (e.g., 'Age', 'Purchase Amount', etc.) using StandardScaler to ensure they are on the same scale.**

**Outcome:  
A cleaned and prepared dataset, data\_cleaned, ready for exploration, analysis, and modelling.**

**3. Exploratory Data Analysis (EDA)**

**Objective:  
To explore the dataset visually and numerically, identifying key trends, patterns, and relationships.**

**Tasks:**

* **Summary Statistics:**
  + **Generate descriptive statistics (data.describe()), including mean, median, standard deviation, etc., to understand distributions.**
* **Distribution of Variables:**
  + **Visualize the distribution of numerical variables (e.g., age, purchase amount) using histograms (data\_cleaned.hist()).**
  + **Explore the distribution of categorical variables (e.g., gender, category) using value counts and pie charts.**
* **Correlation Analysis:**
  + **Calculate correlations between numerical variables and visualize them in a heatmap to uncover relationships (sns.heatmap()).**

**Outcome:  
Key insights into the distributions, correlations, and patterns in the dataset. Helps to identify relationships for feature engineering.**

**4. Feature Engineering**

**Objective:  
To create or modify features to improve model performance and prepare the data for machine learning.**

**Tasks:**

* **Feature Selection:**
  + **Identify and select relevant features for the model (e.g., 'Age', 'Purchase Amount', 'Review Rating', etc.).**
* **Categorical Encoding:**
  + **Use LabelEncoder or one-hot encoding to convert categorical columns (e.g., 'Gender', 'Season') into numerical format.**
* **Create Interaction Features (Optional):**
  + **Create new features based on interactions between existing features (e.g., age \* purchase amount, or discount applied with shipping type).**

**Outcome:  
A transformed dataset, where features are now suitable for machine learning models. The categorical features are encoded, and any additional interaction features are included.**

**5. Visualization and Trend Analysis**

**Objective:  
To generate visualizations that reveal patterns and trends in the data, aiding both exploration and insight generation.**

**Tasks:**

* **Line Charts:**
  + **Visualize relationships such as average purchase amount by age or purchase frequency by age using line charts.**
* **Bar Charts:**
  + **Show comparisons such as average purchase amount by category or total purchases by gender.**
* **Pie Charts:**
  + **Display distributions of categorical data, such as purchase category distribution or subscription status distribution.**
* **Area Graphs:**
  + **Analyz seasonal trends in total purchase amounts.**
* **Trend Analysis:**
  + **Investigate temporal patterns, such as purchases during different seasons, and the impact of promotions or discounts.**

**Outcome:  
Visual insights that show trends in purchases by different customer segments (age, gender, category, etc.), allowing for a better understanding of customer behavior.**

**6. Machine Learning Model (Random Forest)**

**Objective:  
To predict customer behaviour, specifically subscription status, using a machine learning model.**

**Tasks:**

* **Target and Features Selection:**
  + **Identify the target variable (Subscription Status) and the features (numerical and categorical).**
* **Model Training:**
  + **Split the dataset into training and testing sets (train\_test\_split()).**
  + **Train a Random Forest model using RandomForestClassifier().**
* **Model Evaluation:**
  + **Evaluate the model’s performance using metrics like accuracy (accuracy\_score()), classification report (classification\_report()), and confusion matrix (confusion\_matrix()).**
  + **Visualize the confusion matrix with seaborn.heatmap() to understand the true vs. predicted classifications.**

**Outcome:  
A trained Random Forest model that can predict subscription status based on customer data, along with performance evaluation metrics that measure its predictive accuracy.**

**7. Results and Insights**

**Objective:  
To summarize the findings and provide actionable business insights from the analysis and the machine learning model.**

**Tasks:**

* **Key Insights from EDA:**
  + **What are the most common categories purchased, and how do different demographics (age, gender) influence purchase behaviour?**
  + **What are the spending patterns for customers with different subscription statuses?**
  + **How do seasonal changes impact purchases?**
* **Key Insights from Machine Learning:**
  + **Which features (e.g., age, review rating) are most important in predicting subscription status based on feature importance?**
  + **Model performance insights such as accuracy, precision, recall, and confusion matrix.**
* **Business Recommendations:**
  + **Targeted marketing strategies based on customer segments (e.g., customers in certain age groups or with a high likelihood of subscribing).**
  + **Insights into seasonal sales patterns and promotional effectiveness.**

**Outcome:  
A comprehensive set of actionable insights that can help drive marketing, sales, and customer engagement strategies based on the data trends and predictions.**

**3.2 Requirement Specification**

**3.2.1 Hardware Requirements**

* **Processor: Intel Core i5 or equivalent (minimum)**
* **RAM: 8 GB (minimum), 16 GB (recommended for larger datasets)**
* **Storage: 10 GB free space for datasets, libraries, and outputs**
* **Graphics Card: Optional but useful for high-performance visualizations**

**3.2.2 Software Requirements**

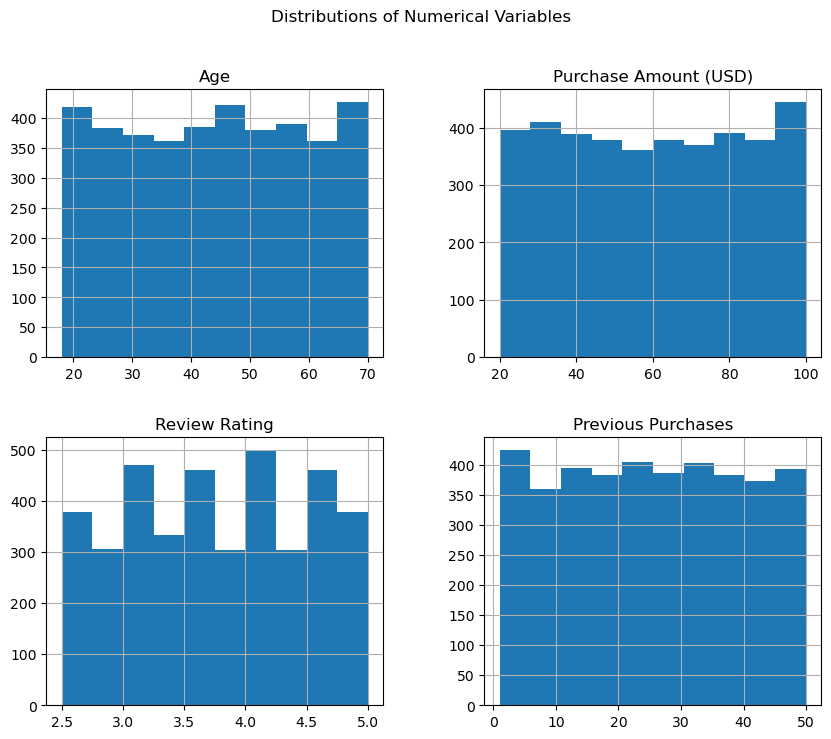
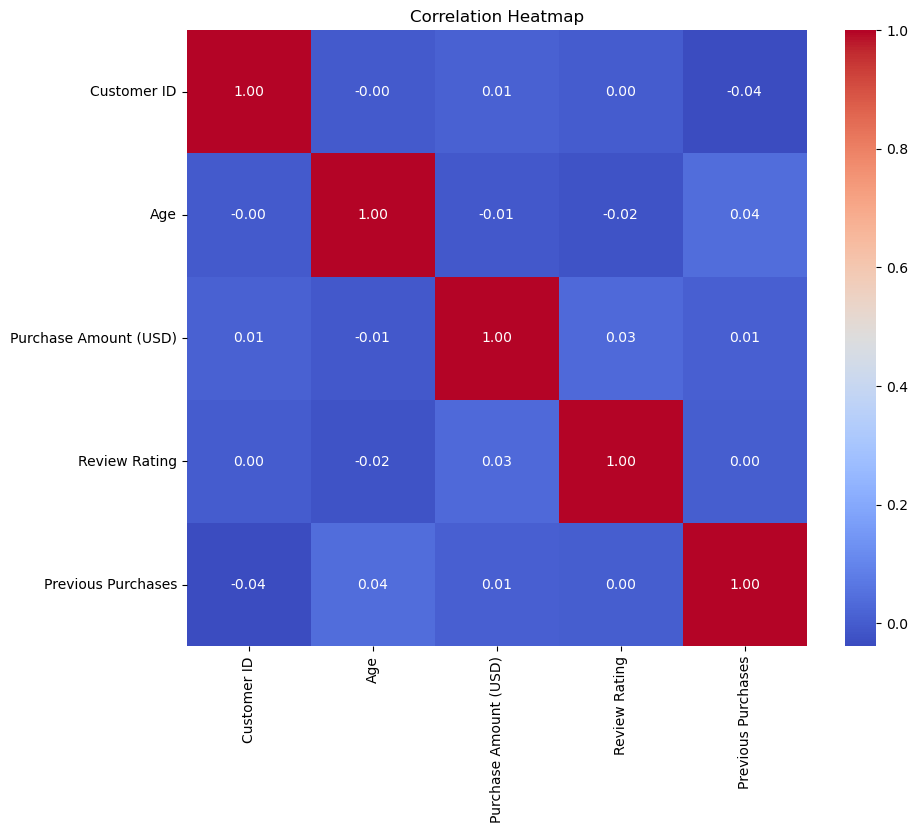
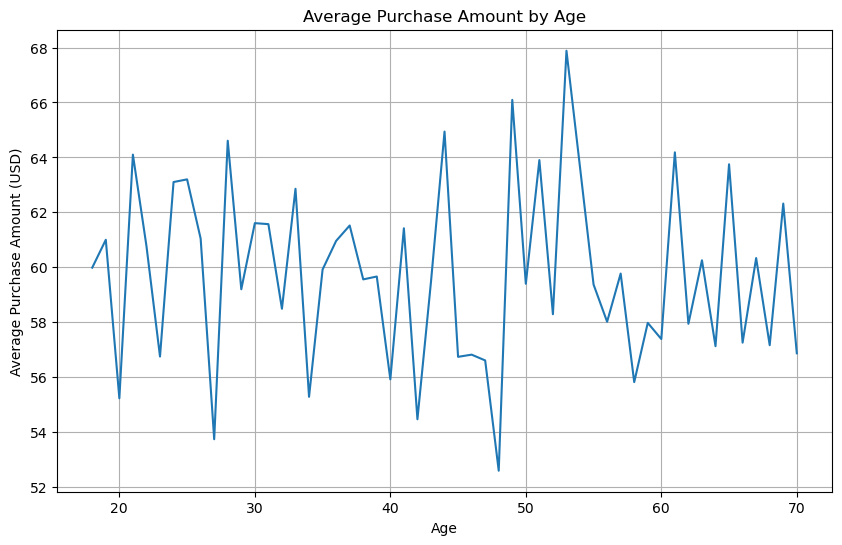
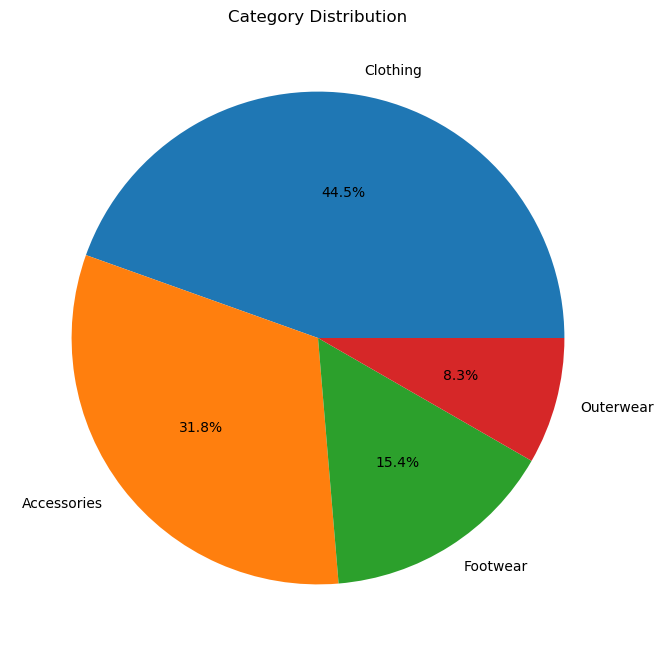
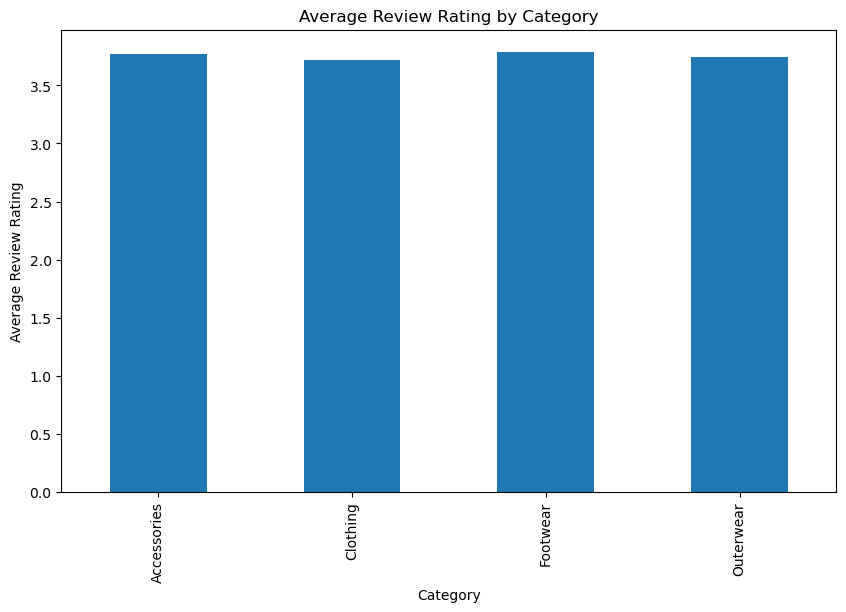
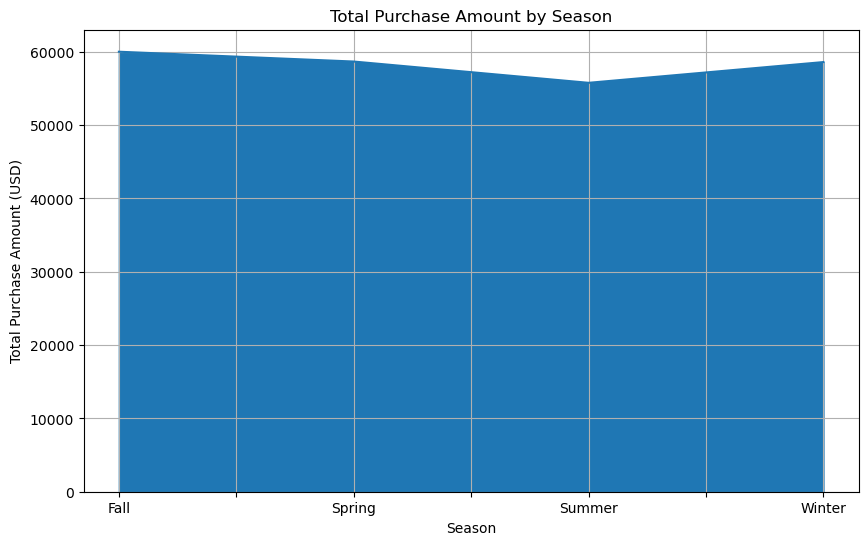
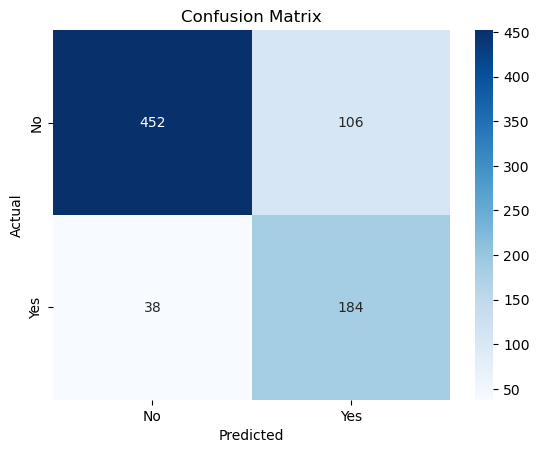
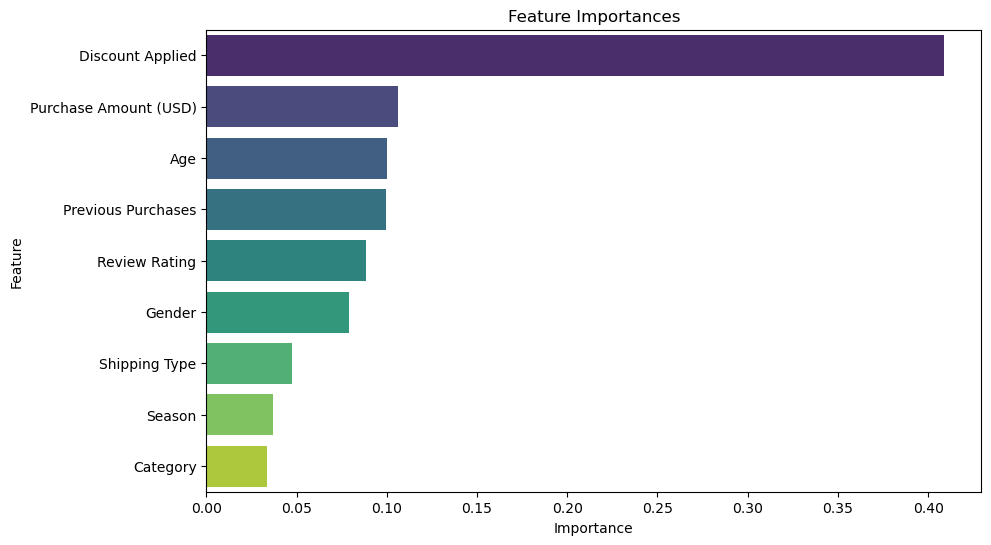
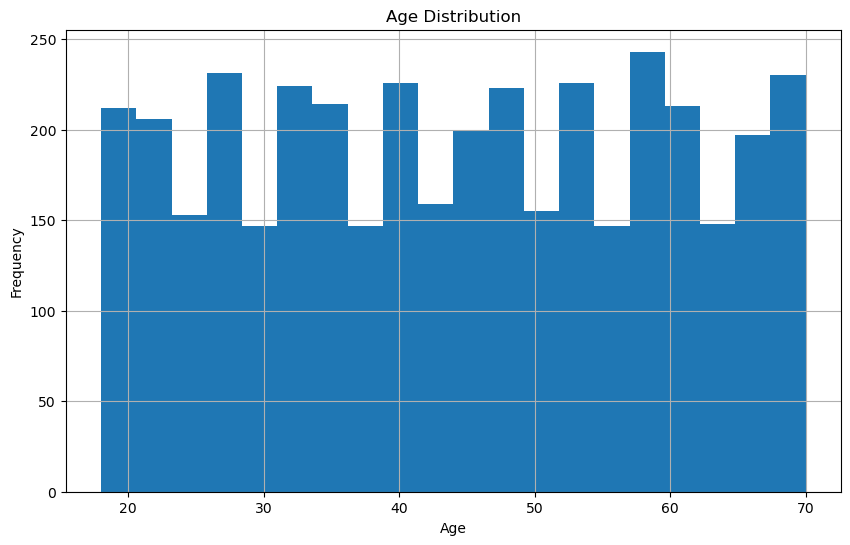
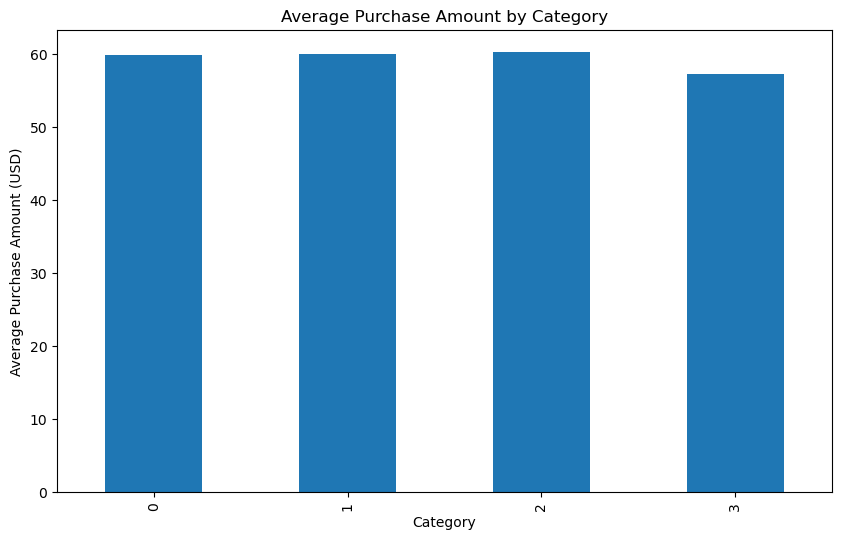
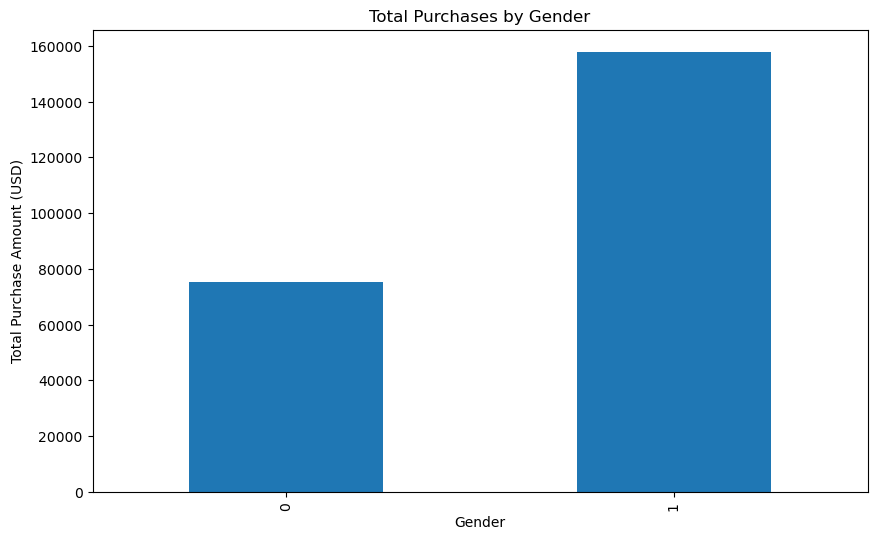
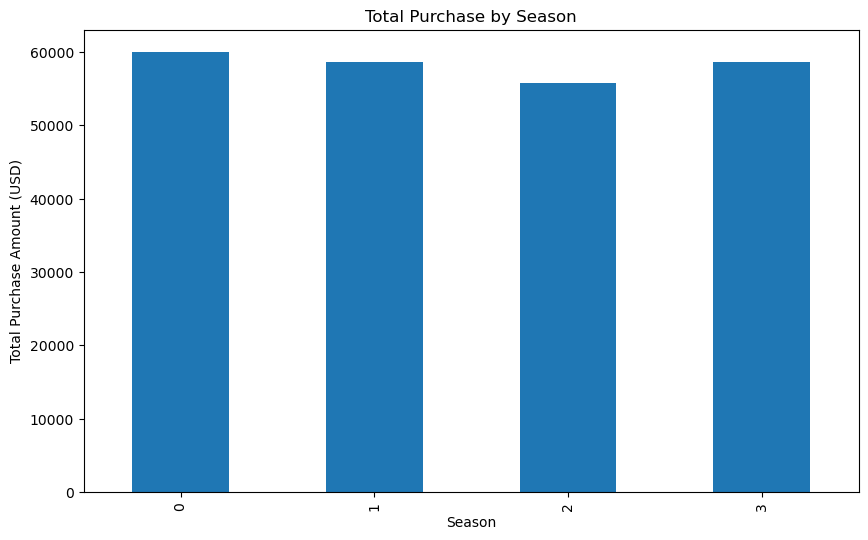
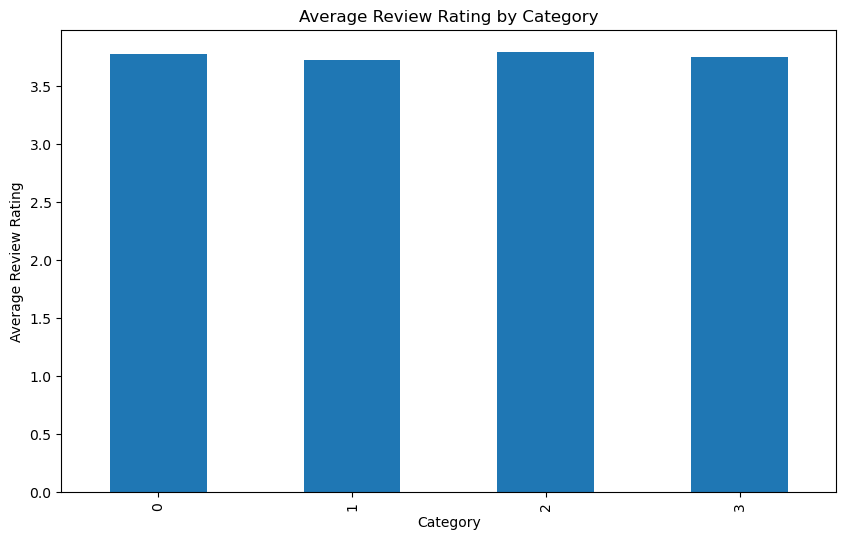
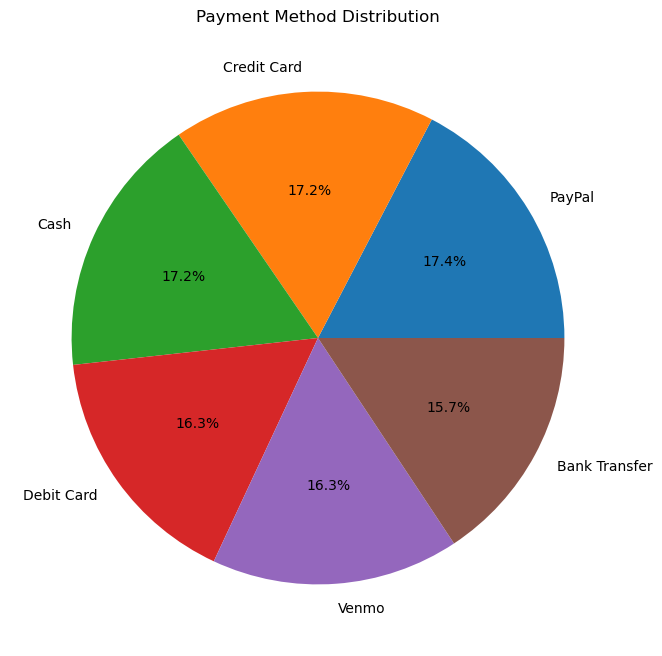
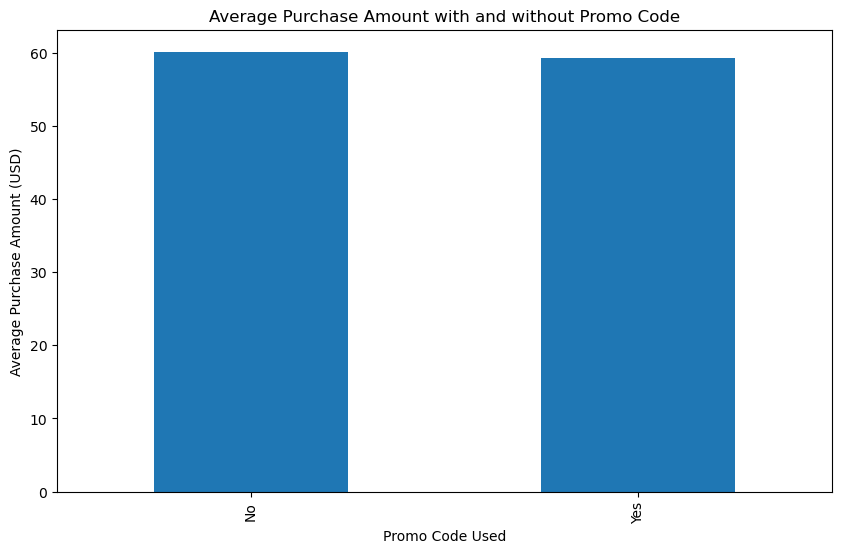
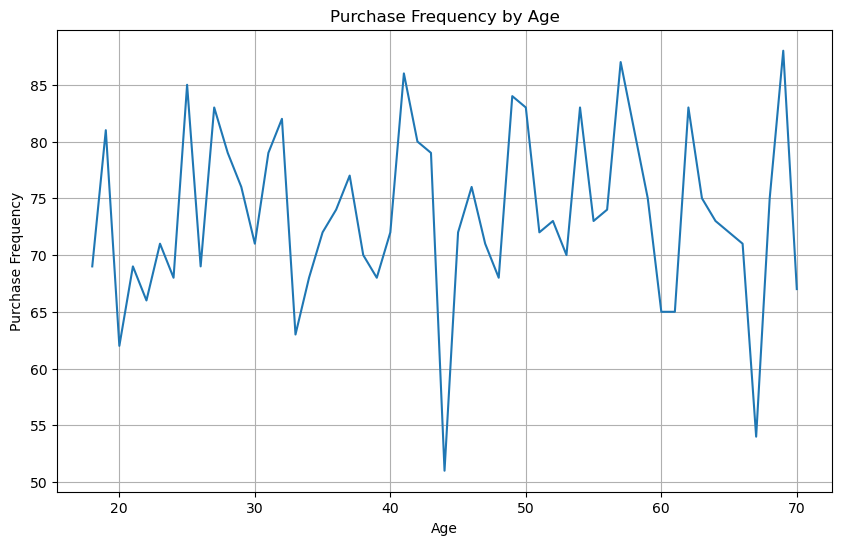
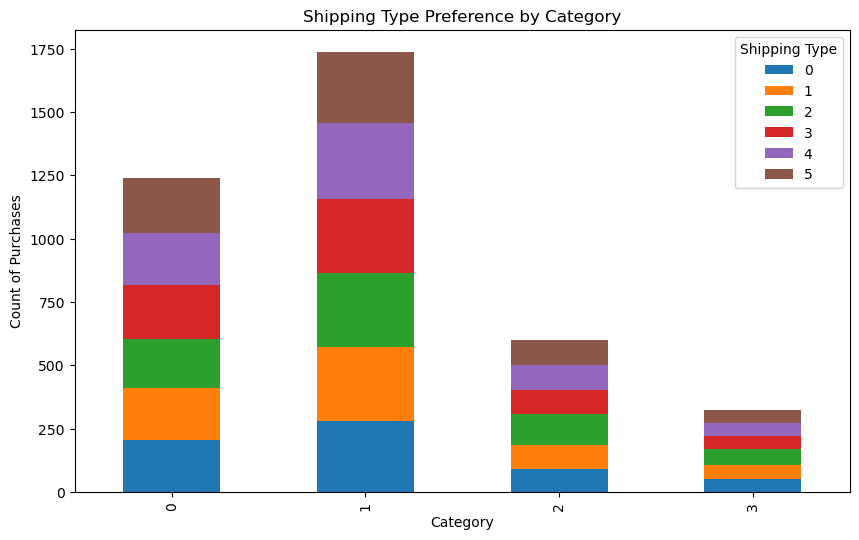
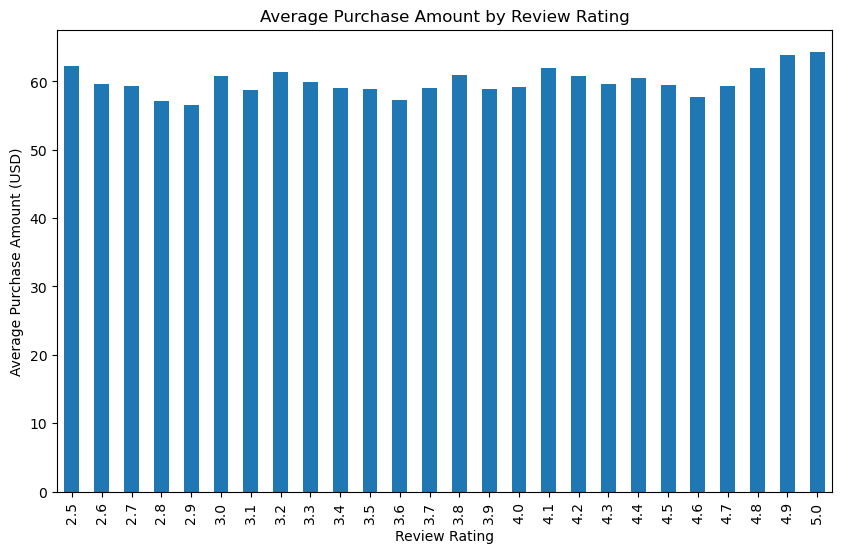
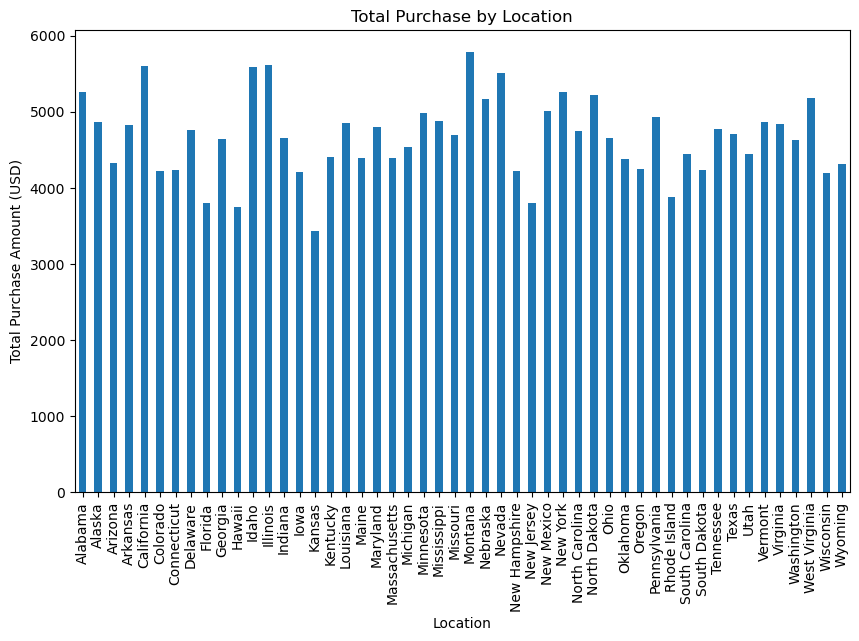
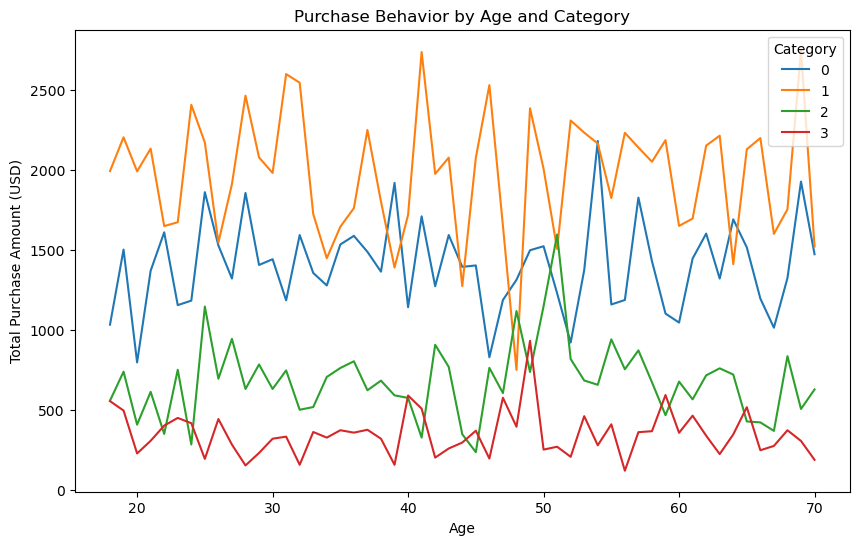
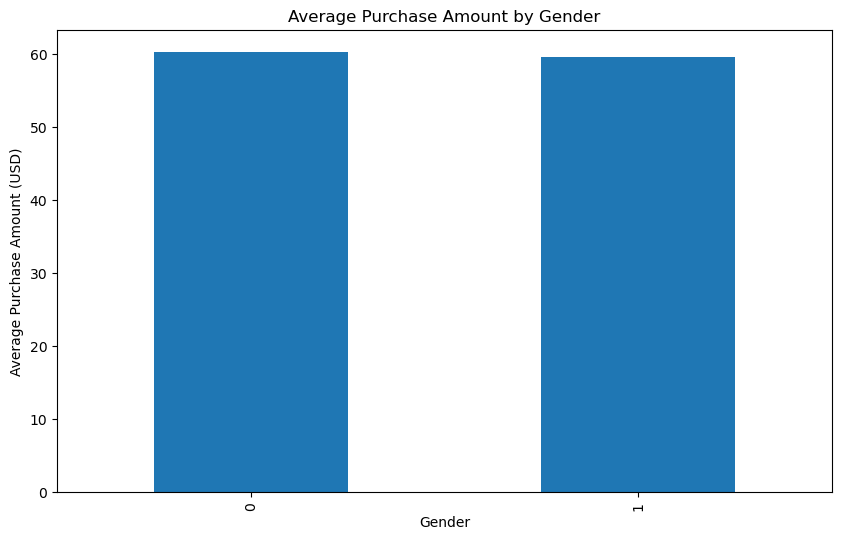
* **Operating System: Windows 10, macOS, or Linux**
* **Programming Language: Python 3.8 or above**
* **Python Libraries:**
  + **Data Manipulation and Analysis: Pandas, NumPy**
  + **Data Visualization: Matplotlib, Seaborn**
  + **Machine Learning: Scikit-learn**
  + **Other Tools: LabelEncoder, StandardScaler**
* **Development Environment:**
  + **Jupyter- Notebook, Google Collab, or any Python IDE (e.g., PyCharm, VSCode)**
* **Dataset: CSV file containing customer demographic and transactional details**

**By following this methodology, the project ensures a systematic approach to understanding shopping trends and developing predictive capabilities.**

**CHAPTER 4**

**Implementation and Result**

* 1. **Snap Shots of Result:**

1. Summary Statistics:
2. Customer ID Age Gender Item Purchased Category \
3. count 3900.000000 3900.000000 3900 3900 3900
4. unique NaN NaN 2 25 4
5. top NaN NaN Male Blouse Clothing
6. freq NaN NaN 2652 171 1737
7. mean 1950.500000 44.068462 NaN NaN NaN
8. std 1125.977353 15.207589 NaN NaN NaN
9. min 1.000000 18.000000 NaN NaN NaN
10. 25% 975.750000 31.000000 NaN NaN NaN
11. 50% 1950.500000 44.000000 NaN NaN NaN
12. 75% 2925.250000 57.000000 NaN NaN NaN
13. max 3900.000000 70.000000 NaN NaN NaN
14. Purchase Amount (USD) Location Size Color Season Review Rating \
15. count 3900.000000 3900 3900 3900 3900 3900.000000
16. unique NaN 50 4 25 4 NaN
17. top NaN Montana M Olive Spring NaN
18. freq NaN 96 1755 177 999 NaN
19. mean 59.764359 NaN NaN NaN NaN 3.749949
20. std 23.685392 NaN NaN NaN NaN 0.716223
21. min 20.000000 NaN NaN NaN NaN 2.500000
22. 25% 39.000000 NaN NaN NaN NaN 3.100000
23. 50% 60.000000 NaN NaN NaN NaN 3.700000
24. 75% 81.000000 NaN NaN NaN NaN 4.400000
25. max 100.000000 NaN NaN NaN NaN 5.000000
26. Subscription Status Shipping Type Discount Applied Promo Code Used \
27. count 3900 3900 3900 3900
28. unique 2 6 2 2
29. top No Free Shipping No No
30. freq 2847 675 2223 2223
31. mean NaN NaN NaN NaN
32. std NaN NaN NaN NaN
33. min NaN NaN NaN NaN
34. 25% NaN NaN NaN NaN
35. 50% NaN NaN NaN NaN
36. 75% NaN NaN NaN NaN
37. max NaN NaN NaN NaN
38. Previous Purchases Payment Method Frequency of Purchases
39. count 3900.000000 3900 3900
40. unique NaN 6 7
41. top NaN PayPal Every 3 Months
42. freq NaN 677 584
43. mean 25.351538 NaN NaN
44. std 14.447125 NaN NaN
45. min 1.000000 NaN NaN
46. 25% 13.000000 NaN NaN
47. 50% 25.000000 NaN NaN
48. 75% 38.000000 NaN NaN
49. max 50.000000 NaN NaN
50. Data Information:
51. <class 'pandas.core.frame.DataFrame'>
52. RangeIndex: 3900 entries, 0 to 3899
53. Data columns (total 18 columns):
54. # Column Non-Null Count Dtype
55. --- ------ -------------- -----
56. 0 Customer ID 3900 non-null int64
57. 1 Age 3900 non-null int64
58. 2 Gender 3900 non-null object
59. 3 Item Purchased 3900 non-null object
60. 4 Category 3900 non-null object
61. 5 Purchase Amount (USD) 3900 non-null int64
62. 6 Location 3900 non-null object
63. 7 Size 3900 non-null object
64. 8 Color 3900 non-null object
65. 9 Season 3900 non-null object
66. 10 Review Rating 3900 non-null float64
67. 11 Subscription Status 3900 non-null object
68. 12 Shipping Type 3900 non-null object
69. 13 Discount Applied 3900 non-null object
70. 14 Promo Code Used 3900 non-null object
71. 15 Previous Purchases 3900 non-null int64
72. 16 Payment Method 3900 non-null object
73. 17 Frequency of Purchases 3900 non-null object
74. dtypes: float64(1), int64(4), object(13)
75. memory usage: 548.6+ KB
76. First few rows of the data:
77. Customer ID Age Gender Item Purchased Category Purchase Amount (USD) \
78. 0 1 55 Male Blouse Clothing 53
79. 1 2 19 Male Sweater Clothing 64
80. 2 3 50 Male Jeans Clothing 73
81. 3 4 21 Male Sandals Footwear 90
82. 4 5 45 Male Blouse Clothing 49
83. Location Size Color Season Review Rating Subscription Status \
84. 0 Kentucky L Gray Winter 3.1 Yes
85. 1 Maine L Maroon Winter 3.1 Yes
86. 2 Massachusetts S Maroon Spring 3.1 Yes
87. 3 Rhode Island M Maroon Spring 3.5 Yes
88. 4 Oregon M Turquoise Spring 2.7 Yes
89. Shipping Type Discount Applied Promo Code Used Previous Purchases \
90. 0 Express Yes Yes 14
91. 1 Express Yes Yes 2
92. 2 Free Shipping Yes Yes 23
93. 3 Next Day Air Yes Yes 49
94. 4 Free Shipping Yes Yes 31
95. Payment Method Frequency of Purchases
96. 0 Venmo Fortnightly
97. 1 Cash Fortnightly
98. 2 Credit Card Weekly
99. 3 PayPal Weekly
100. 4 PayPal Annually
101. Missing Values:
102. Customer ID 0
103. Age 0
104. Gender 0
105. Item Purchased 0
106. Category 0
107. Purchase Amount (USD) 0
108. Location 0
109. Size 0
110. Color 0
111. Season 0
112. Review Rating 0
113. Subscription Status 0
114. Shipping Type 0
115. Discount Applied 0
116. Promo Code Used 0
117. Previous Purchases 0
118. Payment Method 0
119. Frequency of Purchases 0
120. dtype: int64
121. Number of duplicate rows: 0
122. Value counts for Gender:
123. Gender
124. Male 2652
125. Female 1248
126. Name: count, dtype: int64
127. Value counts for Category:
128. Category
129. Clothing 1737
130. Accessories 1240
131. Footwear 599
132. Outerwear 324
133. Name: count, dtype: int64
134. Value counts for Season:
135. Season
136. Spring 999
137. Fall 975
138. Winter 971
139. Summer 955
140. Name: count, dtype: int64
141. Value counts for Subscription Status:
142. Subscription Status
143. No 2847
144. Yes 1053
145. Name: count, dtype: int64
146. Value counts for Shipping Type:
147. Shipping Type
148. Free Shipping 675
149. Standard 654
150. Store Pickup 650
151. Next Day Air 648
152. Express 646
153. 2-Day Shipping 627
154. Name: count, dtype: int64
155. Value counts for Discount Applied:
156. Discount Applied
157. No 2223
158. Yes 1677
159. Name: count, dtype: int64
160. 
161. Correlation Matrix:
162. Customer ID Age Purchase Amount (USD) \
163. Customer ID 1.000000 -0.004079 0.011048
164. Age -0.004079 1.000000 -0.010424
165. Purchase Amount (USD) 0.011048 -0.010424 1.000000
166. Review Rating 0.001343 -0.021949 0.030776
167. Previous Purchases -0.039159 0.040445 0.008063
168. Review Rating Previous Purchases
169. Customer ID 0.001343 -0.039159
170. Age -0.021949 0.040445
171. Purchase Amount (USD) 0.030776 0.008063
172. Review Rating 1.000000 0.004229
173. Previous Purchases 0.004229 1.000000
174. 
175. 
176. 
177. 
178. 
179. Cleaned data saved to: shopping\_trends\_cleaned.csv
180. Classification Report:
181. precision recall f1-score support
182. 0 0.92 0.81 0.86 558
183. 1 0.63 0.83 0.72 222
184. accuracy 0.82 780
185. macro avg 0.78 0.82 0.79 780
186. weighted avg 0.84 0.82 0.82 780
187. Accuracy Score: 0.8153846153846154
188. 
189. Feature Importances:
190. Feature Importance
191. 8 Discount Applied 0.408965
192. 1 Purchase Amount (USD) 0.106326
193. 0 Age 0.100002
194. 3 Previous Purchases 0.099473
195. 2 Review Rating 0.088333
196. 4 Gender 0.079090
197. 7 Shipping Type 0.047457
198. 6 Season 0.036807
199. 5 Category 0.033547
200. C:\Users\dattu\AppData\Local\Temp\ipykernel\_14220\373244521.py:143: FutureWarning:
201. Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.
202. sns.barplot(data=importance\_df, x='Importance', y='Feature', palette='viridis')
203. 
204. 
205. 
206. 
207. Most Common Items in Each Category:
208. Category
209. 0 Jewelry
210. 1 Blouse
211. 2 Sandals
212. 3 Jacket
213. Name: Item Purchased, dtype: object
214. 
215. 
216. 
217. 
218. 
219. 
220. 
221. 
222. Average Number of Previous Purchases: 25.35153846153846
223. 
224. 
225. 
226. 
227. In [ ]:
228. 1. **GitHub Link for Code:**

**https://github.com/Shankadude/Data\_Analysis\_on\_shopping\_trends**

**CHAPTER 5**

**Discussion and Conclusion**

* 1. **Future Work:**

My project is well-structured, and it covers a wide range of analysis techniques, including exploratory data analysis (EDA), visualizations, feature engineering, and machine learning. Here are some suggestions for improvement and addressing potential issues in future work:

**1. Data Cleaning**

Handling Missing Values: In my data cleaning section, I check for missing values but only print them. Consider actually addressing these missing values by either imputing them (mean/median for numerical columns, mode for categorical ones) or dropping rows/columns that are too incomplete.

Outlier Detection: Consider handling outliers in numerical columns, as they might distort statistical analyses and model performance.

Standardization of Column Names: Ensure that column names are standardized (e.g., no leading/trailing spaces, consistent capitalization) to avoid potential errors during data preprocessing.

**2. Data Exploration**

Additional Visualizations: While I do cover several visualizations, it might help to include scatter plots, box plots, or pair plots to reveal relationships between different features, particularly for numerical variables.

Time Series Analysis: If my dataset includes a timestamp, I could consider analyzing trends over time, which can help uncover seasonality and temporal patterns.

3. Feature Engineering

Interaction Features: I can try creating interaction features between certain variables (e.g., Age x Category or Purchase Amount x Discount Applied) to capture more complex relationships.

Categorical Variable Handling: Instead of Label Encoding for categorical variables, I could consider One-Hot Encoding for non-ordinal categories (e.g., Shipping Type, Category). This would help prevent the model from mistakenly assuming a hierarchy between categories.

Feature Selection: I could perform feature selection techniques (like Recursive Feature Elimination or Feature Importance) to reduce noise in my model.

4. Model Improvement

Model Hyperparameter Tuning: My Random Forest model could benefit from hyperparameter tuning. I can use GridSearchCV or RandomizedSearchCV to find the optimal hyperparameters (like n\_estimators, max\_depth, etc.) to improve the model's performance.

Cross-Validation: Instead of using a single train-test split, I could consider applying cross-validation (e.g., k-fold cross-validation) to get more reliable estimates of model performance.

Model Comparison: I could try comparing the performance of Random Forest Classifier with other classification algorithms, such as Logistic Regression, SVM, or XGBoost. It’s always good to benchmark multiple models.

Handling Imbalanced Classes: If my target variable (Subscription Status) has imbalanced classes, I could consider using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weights to handle this imbalance.

5. Evaluation Metrics

ROC Curve: Besides accuracy and confusion matrix, I could consider plotting the ROC curve and calculating the AUC (Area Under the Curve) for more insight into model performance, especially for imbalanced datasets.

Precision-Recall Curve: For imbalanced datasets, the Precision-Recall Curve might give more information than the ROC curve.

6. Scalability and Performance

Data Size: If my dataset is large, I could consider optimizing performance by using techniques like batch processing or parallelizing some operations. Alternatively, if I plan to scale the model in production, I could consider using cloud-based ML platforms like AWS, GCP, or Azure.

Efficient Model Deployment: Once I finalize the model, I could explore how to deploy it for real-time predictions or batch processing.

7. Additional Questions for Future Analysis

Segmentation Analysis: I could explore customer segmentation based on purchase behaviour using clustering techniques (e.g., K-means or DBSCAN).

Customer Lifetime Value (CLV): Another useful analysis could be to estimate the Customer Lifetime Value based on purchase history, which could guide marketing strategies.

Sentiment Analysis: If I have textual data like product reviews, performing sentiment analysis can add valuable insights into customer opinions.

8. Documenting Assumptions and Insights

I should document the assumptions I made during the analysis, such as why I chose specific features or why I handled missing values in a certain way.

Summarize key insights from my analysis, such as trends or patterns I discovered and their potential business implications.

9. Future Work

Model Interpretability: If I intend to deploy this model, I could consider model interpretability techniques like SHAP values to explain the model's decisions, especially when dealing with black-box models like Random Forest.

Continuous Data Pipeline: I could consider implementing a pipeline for continuous data collection and model retraining, especially if I want to predict subscription statuses based on new customer data over time.

**Conclusion:**

The overall impact and contribution of this project lies in its comprehensive approach to analyzing shopping trends and predicting customer behaviours. By applying data analysis techniques such as exploratory data analysis (EDA), feature engineering, and machine learning, this project provides valuable insights into various aspects of customer behaviour, including purchase patterns, preferences, and demographic influences.

**Key Contributions:**

1. In-depth Data Exploration: Through thorough examination of the dataset, including statistical summaries, distribution analysis, and correlation matrices, the project uncovers key relationships between variables like age, purchase amount, and category preferences.
2. Visualization of Trends: The project effectively communicates insights through a variety of visualizations such as histograms, pie charts, and heatmaps, helping stakeholders understand customer behaviour patterns in an intuitive manner.
3. Predictive Model: By building and evaluating a Random Forest Classifier to predict subscription status, the project demonstrates the potential to forecast customer actions, which can guide marketing, sales, and customer retention strategies.
4. Actionable Business Insights: The analysis answers specific business questions, such as identifying the most purchased items by category, segmenting customers based on gender or age, and evaluating the impact of discounts and promo codes on purchase behaviour. These insights can directly inform business decisions like product stocking, pricing, and targeted promotions.
5. Scalability and Future Potential: The project sets a foundation for scaling and deploying predictive models in real-world applications, offering businesses the ability to continually update and improve predictions as new data comes in.

**Overall Impact:**

The project provides a strong analytical foundation for understanding customer purchasing behaviour, making it highly valuable for businesses seeking to optimize sales strategies, enhance customer engagement, and predict future trends. Additionally, it opens doors for further research, such as customer segmentation, sentiment analysis, and the development of a continuous data pipeline for real-time predictions. By offering both descriptive and predictive analysis, the project contributes to data-driven decision-making and a deeper understanding of consumer behaviour.

**REFERENCES**

[1]. **Wes McKinney**, "Python for Data Analysis," O'Reilly Media, 2nd Edition, 2017.  
[2]. **DataCamp**, "Exploratory Data Analysis using Python," DataCamp, Available at: https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python, 2020.  
[3]. **Real Python**, "Data Preprocessing with Pandas," Real Python, Available at: https://realpython.com/python-data-cleaning-numpy-pandas/, 2020.  
[4]. **Kaggle**, "Exploratory Data Analysis on Kaggle," Kaggle, Available at: https://www.kaggle.com/learn/eda, 2020.  
[5]. **DataCamp**, "Feature Engineering for Machine Learning," DataCamp, Available at: https://www.datacamp.com/community/tutorials/feature-engineering-python, 2020.  
[6]. **Towards Data Science**, "Feature Engineering 101: A Guide for Data Scientists," Towards Data Science, Available at: https://towardsdatascience.com/feature-engineering-for-machine-learning-d7d72b7b5c11, 2019.  
[7]. **Scikit-learn**, "Scikit-learn Documentation," Scikit-learn, Available at: https://scikit-learn.org/stable/documentation.html, 2020.  
[8]. **GitHub**, "Random Forest Classifier Project on GitHub," GitHub, Available at: <https://github.com/justmarkham/DAT8/blob/master/notebooks/03_feature_selection_and_random_forest.ipynb>, 2019.  
[9]. **Towards Data Science**, "A Comprehensive Guide to Hyperparameter Tuning," Towards Data Science, Available at: https://towardsdatascience.com/hyperparameter-tuning-with-scikit-learn-68fa2ff7595a, 2019.  
[10]. **Machine Learning Mastery**, "How to Evaluate Machine Learning Models," Machine Learning Mastery, Available at: https://machinelearningmastery.com/compare-machine-learning-algorithms-python-using-resampling/, 2020.  
[11]. **Kunal Sood**, "Data Visualization with Python and Matplotlib," O'Reilly Media, 2021.  
[12]. **GitHub**, "Matplotlib and Seaborn Projects on GitHub," GitHub, Available at: <https://github.com/rougier/matplotlib-tutorial>, 2019.  
[13]. **Kaggle**, "Titanic: Machine Learning from Disaster," Kaggle, Available at: https://www.kaggle.com/c/titanic, 2020.  
[14]. **GitHub**, "Customer Segmentation using K-means Clustering," GitHub, Available at: <https://github.com/nikbearbrown/Customer-Segmentation>, 2020.  
[15]. **Towards Data Science**, "Customer Segmentation with Python," Towards Data Science, Available at: https://www.datacamp.com/community/tutorials/customer-segmentation-python, 2019.  
[16]. **Towards Data Science**, "A Comprehensive Intuitive Guide to XGBoost," Towards Data Science, Available at: https://towardsdatascience.com/a-comprehensive-intuitive-guide-to-xgboost-92f6b16bace9, 2020.  
[17]. **Coursera**, "Deep Learning Specialization by Andrew Ng," Coursera, Available at: <https://www.coursera.org/specializations/deep-learning>, 2021.  
[18]. **Towards Data Science**, "How to Deploy a Machine Learning Model," Towards Data Science, Available at: https://towardsdatascience.com/deploying-your-machine-learning-model-as-an-api-37ec7e773041, 2019.  
[19]. **Coursera**, "Deploying Machine Learning Models on AWS," Coursera, Available at: <https://www.coursera.org/learn/machine-learning-with-aws>, 2020.