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

Aditya Kumar Jha, adityajha0129@gmail.com

Under the Guidance of

Abdul Aziz Md

Master Trainer, Edunet Foundations

ACKNOWLEDGEMENT

I am profoundly grateful for the guidance, encouragement, and steadfast support that have shaped this journey. This achievement wouldn't have been possible without the contributions of several incredible individuals and organizations, to whom I owe my deepest appreciation.

First and foremost, my sincerest thanks go to my exceptional mentor, Abdul Aziz Md, whose insightful guidance and unwavering belief in my abilities have been instrumental at every stage of this project. Their expertise, patience, and ability to provide both innovative solutions and constructive feedback have been invaluable. Through every challenge, they have offered clarity and direction, pushing me to go beyond my limits and strive for excellence. Their mentorship has been nothing short of transformative, and I am truly honored to have had their support.

I am also immensely grateful to TechSaksham for creating such a dynamic platform that fosters innovation and exploration in the realm of artificial intelligence. The experience, resources, and mentorship provided through this initiative have not only deepened my technical expertise but have also played a crucial role in my professional growth. This internship was a remarkable opportunity to bridge theory with real-world application, and I sincerely appreciate TechSaksham's vision in empowering aspiring minds like mine to turn ideas into impact.

A special note of thanks to Pavan Sir, whose support behind the scenes made all the difference. From keeping us informed about the program schedule to ensuring we never missed out on any important updates, his efforts brought much-needed structure and clarity to the process. His diligence and commitment in facilitating smooth communication were invaluable, and I deeply appreciate the role he played in making this experience seamless and enriching.

ABSTRACT

This report presents the development of a data-driven approach to identifying shopping trends using advanced data analysis techniques. Retail businesses collect vast amounts of transactional and behavioral data from multiple channels, yet many struggle to extract meaningful insights that drive decision-making. Ineffective trend analysis can lead to lost revenue, inventory mismanagement, and suboptimal marketing strategies.

To address these challenges, this project proposes a comprehensive solution involving data collection, preprocessing, exploratory analysis, automated reporting, and actionable recommendations. By leveraging statistical and machine learning techniques, the system identifies emerging shopping patterns, seasonal fluctuations, and customer preferences with high accuracy. The integration of interactive visualizations and automated reporting enhances accessibility for decision-makers, enabling timely and informed actions.

The project also highlights challenges related to data quality, scalability, and real-time trend detection. Experimental analysis demonstrates the system's effectiveness in uncovering valuable insights for retailers, with future enhancements focusing on real-time data streaming, predictive analytics, and personalized recommendations. This approach aims to empower businesses with a data-driven framework for optimizing inventory, refining marketing strategies, and staying competitive in an evolving retail landscape.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem Statement:

Retail businesses generate vast amounts of shopping data from multiple channels, including in-store transactions, e-commerce platforms, and customer interactions. However, the challenge lies in effectively analyzing this data to uncover emerging trends, understand customer preferences, and identify seasonal buying patterns. Without proper trend analysis, businesses may face financial losses due to poor inventory management, ineffective marketing strategies, and an inability to stay competitive in a rapidly evolving retail landscape. This project seeks to develop a data-driven approach that enables businesses to make informed decisions and optimize their operations.

Significance of the Study:
Efficiency: Automates trend detection, reducing manual effort and boosting productivity.
Accuracy: Minimizes errors in trend identification and reporting.
Scalability: Adapts to businesses of all sizes.
Competitive Advantage: Helps businesses anticipate market trends and refine strategies.
Data-Driven Decisions: Provides actionable insights for marketing, inventory, and customer engagement.

1.2 Motivation:

In an era where data is a valuable asset, businesses that fail to leverage analytical tools risk falling behind. The ability to identify shopping trends in real time can significantly impact sales forecasting, personalized marketing, and inventory planning. This project is motivated by the need to bridge the gap between data collection and actionable insights, enabling retailers to make strategic decisions with confidence. Potential applications include optimizing product recommendations in e-commerce, improving demand forecasting in supply chains, and enhancing customer segmentation strategies. The project's impact extends to boosting operational efficiency, maximizing revenue potential, and fostering data-driven decision-making across multiple industries.

1.3 Objective:

With this project we aim to develop a robust system for collecting and integrating retail shopping data from various sources.

To do this we have to understand the specific objectives of this project. These include but are not limited to:

1. Implement preprocessing and cleaning techniques to enhance data quality.
2. Conduct exploratory data analysis (EDA) to identify trends, correlations, and customer behaviors.
3. Automate the generation of reports and visualizations for easy interpretation by stakeholders.
4. Provide actionable recommendations based on identified shopping patterns to improve decision-making.

1.4 Scope of the Project:

This project focuses on analyzing shopping trends using retail data to enhance business decision-making. It encompasses data collection, preprocessing, exploratory analysis, visualization, and reporting to identify consumer behavior patterns and market trends. The analysis is designed for small to medium-sized retail businesses looking to leverage data for improved operations. While the project will provide valuable insights, its initial implementation does not include real-time analytics, predictive modeling, or advanced AI-driven forecasting, which can be explored in future enhancements.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain:

The retail industry has increasingly adopted data analytics to gain insights into consumer behavior, optimize operations, and enhance decision-making processes. The integration of data analytics enables retailers to analyze vast amounts of information from various sources, including in-store transactions, e-commerce platforms, and customer interactions. This approach facilitates the identification of shopping trends, customer preferences, and seasonal buying patterns.

A study titled "Predictive Analysis of Big Data in Retail Industry" highlights the significance of big data analytics in retail. The research emphasizes that retailers can leverage analytics to gain a unified view of their customers and operations across different channels, thereby making strategic decisions that contribute to the industry's growth. The study focuses on the Singapore retail sector, employing a quantitative research method involving 500 participants. The findings indicate that social media analytics is predominantly utilized within Singapore's retail industry, underscoring the importance of analyzing customer interactions on social platforms to inform business strategies.

The COVID-19 pandemic has also influenced consumer shopping behaviors, leading to a surge in online shopping. An article titled "Analysis and modeling of changes in online shopping behavior due to Covid-19 pandemic: A Florida case study" examines this shift. The study analyzes data from the first quarter of 2022, noting a decrease in new COVID-19 cases and the lifting of restrictions, which allowed consumers to return to physical stores. However, the research highlights that the pandemic has led to lasting changes in shopping behaviors, with a significant portion of consumers continuing to prefer online shopping due to its convenience and safety.

Furthermore, the role of big data in retail is explored in the study "Analyzing the role of big data and its effects on the retail industry." The research discusses how the expansion of social media, technology, and online shopping accelerates the development of the retail industry. Retail organizations are increasingly recognizing the value of big data analysis in making informed decisions. The study emphasizes that big data analytics enables retailers to understand customer preferences, optimize pricing strategies, and enhance inventory management.

2.2 Existing models, Techniques, or Methodologies:

Existing methods for shopping trend analysis in the retail industry encompass a range of statistical and machine learning approaches:

- **Time-Series Analysis:** Traditional statistical techniques, such as time-series analysis, are employed to forecast future sales and demand by analyzing historical data. These methods help in identifying seasonal patterns and trends over time.
- **Machine Learning Models:** Advanced machine learning models are utilized for clustering and classification of customer behaviors. These models analyze purchasing patterns to segment customers into distinct groups, enabling personalized marketing strategies.
- **Data Mining Approaches:** Data mining techniques are applied to discover hidden patterns within large datasets. In the retail context, this involves analyzing transaction data to identify associations between products, commonly known as market basket analysis.
- **A comparative study titled "Retail Demand Forecasting:** A Comparative Study for Multivariate Time Series" explores various regression and machine learning models to predict retail demand accurately. The study enriches time series data with macroeconomic variables, such as the Consumer Price Index (CPI) and unemployment rates, to enhance the predictive accuracy of the models.

2.3 Limitations in existing Systems:

Despite the advancements in data analytics within the retail industry, several limitations persist:

- **Computational Resources**: Many existing models require extensive computational resources, making them less accessible for small to medium-sized enterprises.
- **Real-Time Adaptability**: Some methods lack the capability to adapt in real-time, limiting their effectiveness in dynamic retail environments where consumer behaviors can change rapidly.
- **Privacy Concerns**: The collection and analysis of customer data raise privacy issues. Ensuring compliance with data protection regulations and maintaining customer trust are significant challenges.
- **Capturing Rapidly Changing Behaviors**: Traditional models may not effectively capture rapidly changing consumer behaviors, especially in the context of unforeseen events such as pandemics or economic shifts.

Addressing these limitations requires the development of scalable, adaptable, and privacy-conscious analytical models that can operate efficiently across various retail contexts.

1. **Scalability**

   Utilize cloud-based infrastructure (AWS, Google Cloud) and distributed databases (Hadoop, MongoDB) for handling large retail datasets.

   Optimize data storage and retrieval to ensure smooth processing of high-volume transactions.

2. **Adaptability**

   Implement real-time analytics using streaming platforms (Kafka, Flink) to detect emerging shopping trends dynamically.

   Integrate multi-source data from POS systems, e-commerce platforms, and customer feedback for holistic analysis.

   Employ AI-powered predictive analytics for demand forecasting and personalized recommendations.

3. **Privacy-Conscious Analytics**

   Ensure compliance with GDPR, CCPA through data anonymization, encryption, and access control.

   Leverage federated learning to analyze customer data without compromising privacy.

   Use transparent data collection methods to enhance consumer trust.

4. **Operational Efficiency**

   Automate decision-making for inventory optimization, pricing adjustments, and targeted marketing strategies.

   Minimize computational costs with efficient algorithms, making analytics accessible to small and medium-sized retailers.

   Develop interactive dashboards for easy interpretation of insights by business stakeholders.

   By integrating these advancements, businesses can enhance customer insights, improve decision-making, and stay competitive in a data-driven retail environment.

CHAPTER 3

Proposed Methodology

3.1 System Design

The proposed methodology follows a structured data-driven approach to identify shopping trends. It consists of several key stages:

1. **Data Collection**:

   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**:

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

4.1 Snap Shots of Result:

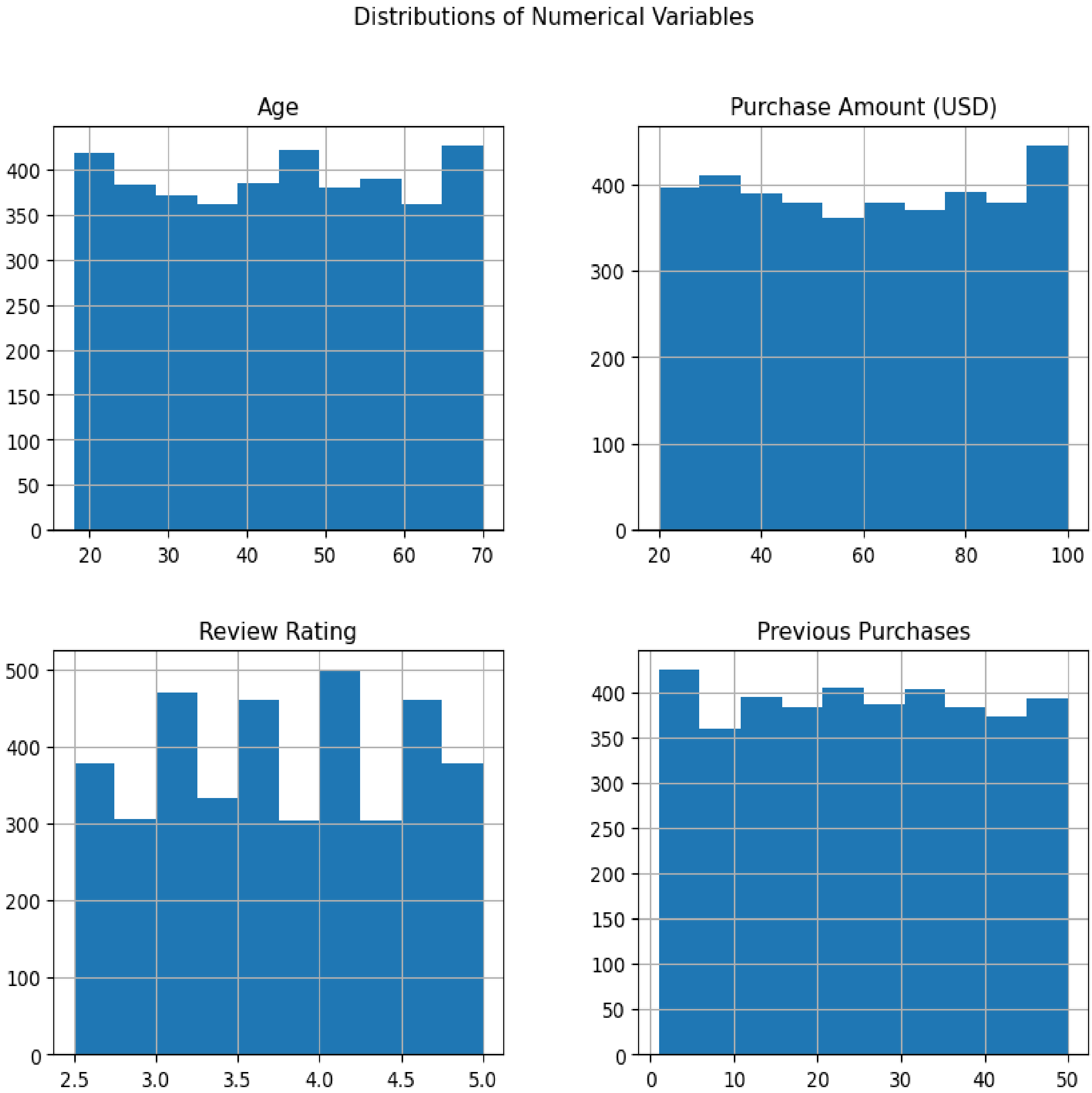
```
Summary Statistics:
         Customer ID          Age Gender Item Purchased  Category  \
count    3900.000000  3900.000000   3900           3900      3900
unique           NaN          NaN      2             25         4
top              NaN          NaN   Male         Blouse  Clothing
freq             NaN          NaN   2652            171      1737
mean     1950.500000    44.068462    NaN            NaN       NaN
std      1125.977353    15.207589    NaN            NaN       NaN
min         1.000000    18.000000    NaN            NaN       NaN
25%       975.750000    31.000000    NaN            NaN       NaN
50%      1950.500000    44.000000    NaN            NaN       NaN
75%      2925.250000    57.000000    NaN            NaN       NaN
max      3900.000000    70.000000    NaN            NaN       NaN

         Purchase Amount (USD) Location  Size  Color  Season  Review Rating  \
count              3900.000000     3900  3900   3900    3900    3900.000000
unique                     NaN       50     4     25       4            NaN
top                        NaN  Montana     M  Olive  Spring            NaN
freq                       NaN       96  1755    177     999            NaN
mean                 59.764359      NaN   NaN    NaN     NaN       3.749949
std                  23.685392      NaN   NaN    NaN     NaN       0.716223
min                  20.000000      NaN   NaN    NaN     NaN       2.500000
25%                  39.000000      NaN   NaN    NaN     NaN       3.100000
```

```
50%                60.000000      NaN   NaN    NaN     NaN
3.700000
75%                81.000000      NaN   NaN    NaN     NaN
4.400000
max               100.000000      NaN   NaN    NaN     NaN
5.000000

       Subscription Status  Shipping Type Discount Applied Promo Code
Used  \
count                 3900           3900             3900
3900
unique                   2              6                2
2
top                     No  Free Shipping               No
No
freq                  2847            675             2223
2223
mean                   NaN            NaN              NaN
NaN
std                    NaN            NaN              NaN
NaN
min                    NaN            NaN              NaN
NaN
25%                    NaN            NaN              NaN
NaN
50%                    NaN            NaN              NaN
NaN
75%                    NaN            NaN              NaN
NaN
max                    NaN            NaN              NaN
NaN

        Previous Purchases Payment Method Frequency of Purchases
count          3900.000000           3900                   3900
unique                 NaN              6                      7
top                    NaN         PayPal         Every 3 Months
freq                   NaN            677                    584
mean             25.351538            NaN                    NaN
std              14.447125            NaN                    NaN
min               1.000000            NaN                    NaN
25%              13.000000            NaN                    NaN
50%              25.000000            NaN                    NaN
75%              38.000000            NaN                    NaN
max              50.000000            NaN                    NaN

Data Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
```

```
 9   Season                   3900 non-null   object
 10  Review Rating            3900 non-null   float64
 11  Subscription Status      3900 non-null   object
 12  Shipping Type            3900 non-null   object
 13  Discount Applied         3900 non-null   object
 14  Promo Code Used          3900 non-null   object
 15  Previous Purchases       3900 non-null   int64
 16  Payment Method           3900 non-null   object
 17  Frequency of Purchases   3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB

First few rows of the data:
    Customer ID  Age Gender Item Purchased  Category  Purchase Amount
(USD)  \
0            1   55   Male         Blouse  Clothing
53
1            2   19   Male        Sweater  Clothing
64
2            3   50   Male          Jeans  Clothing
73
3            4   21   Male        Sandals  Footwear
90
4            5   45   Male         Blouse  Clothing
49

        Location Size      Color  Season  Review Rating Subscription
Status  \
0       Kentucky    L       Gray  Winter            3.1
Yes
1          Maine    L     Maroon  Winter            3.1
Yes
2  Massachusetts    S     Maroon  Spring            3.1
Yes
3   Rhode Island    M     Maroon  Spring            3.5
Yes
4         Oregon    M  Turquoise  Spring            2.7
Yes

     Shipping Type Discount Applied Promo Code Used  Previous
Purchases  \
0        Express              Yes             Yes
14
1        Express              Yes             Yes
2
2  Free Shipping              Yes             Yes
23
3   Next Day Air              Yes             Yes
49
4  Free Shipping              Yes             Yes
31

   Payment Method Frequency of Purchases
0           Venmo            Fortnightly
1            Cash            Fortnightly
2     Credit Card                 Weekly
3          PayPal                 Weekly
4          PayPal               Annually

Missing Values:
```

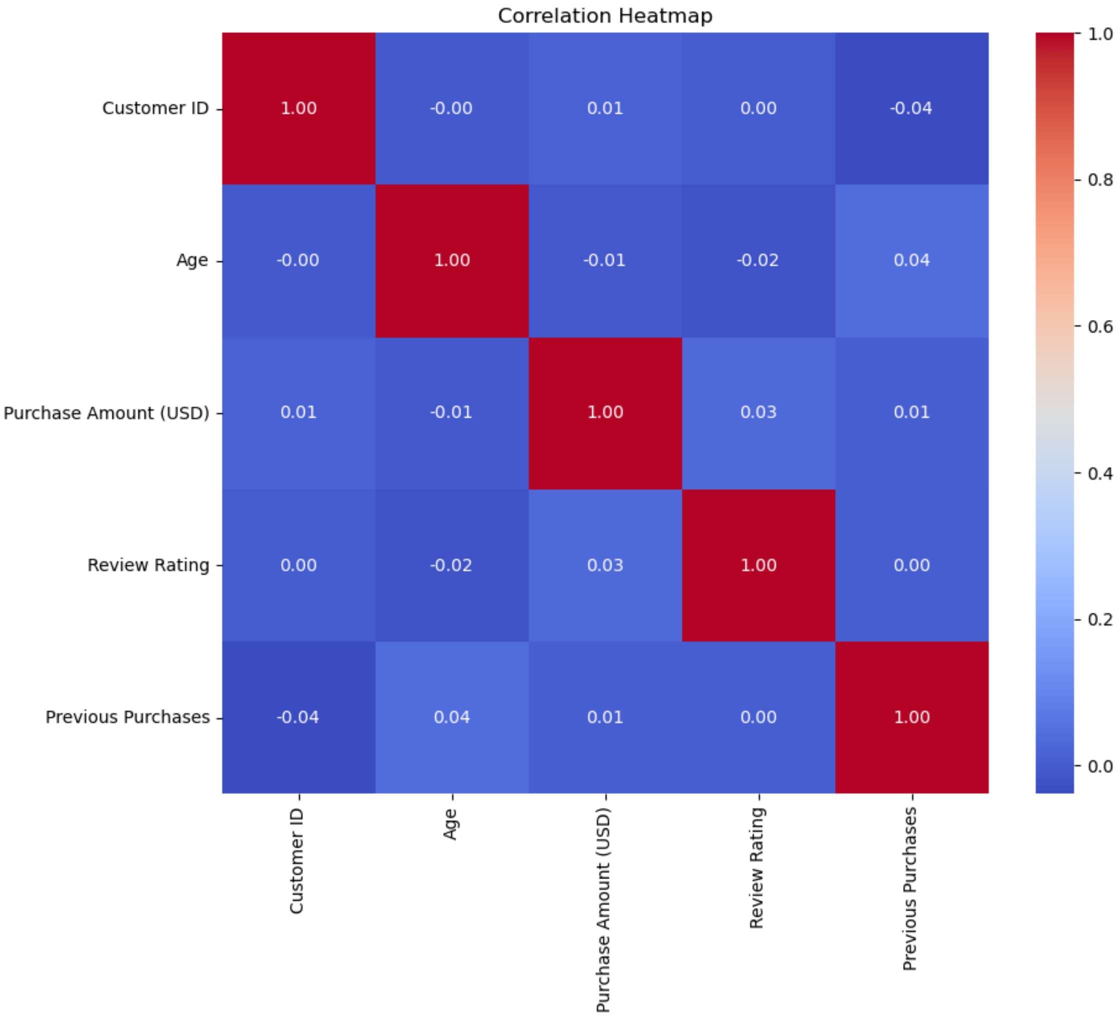
```
 Customer ID                0
Age                        0
Gender                     0
Item Purchased             0
Category                   0
Purchase Amount (USD)      0
Location                   0
Size                       0
Color                      0
Season                     0
Review Rating              0
Subscription Status        0
Shipping Type              0
Discount Applied           0
Promo Code Used            0
Previous Purchases         0
Payment Method             0
Frequency of Purchases     0
dtype: int64

Number of duplicate rows: 0

Value counts for Gender:
 Gender
Male      2652
Female    1248
Name: count, dtype: int64

Value counts for Category:
 Category
Clothing       1737
Accessories    1240
Footwear        599
Outerwear       324
Name: count, dtype: int64

Value counts for Season:
 Season
Spring    999
Fall      975
Winter    971
Summer    955
Name: count, dtype: int64

Value counts for Subscription Status:
 Subscription Status
No     2847
Yes    1053
Name: count, dtype: int64

Value counts for Shipping Type:
 Shipping Type
Free Shipping     675
Standard          654
Store Pickup      650
Next Day Air      648
Express           646
2-Day Shipping    627
Name: count, dtype: int64

```

```
Value counts for Discount Applied:
 Discount Applied
No      2223
Yes     1677
Name: count, dtype: int64
```

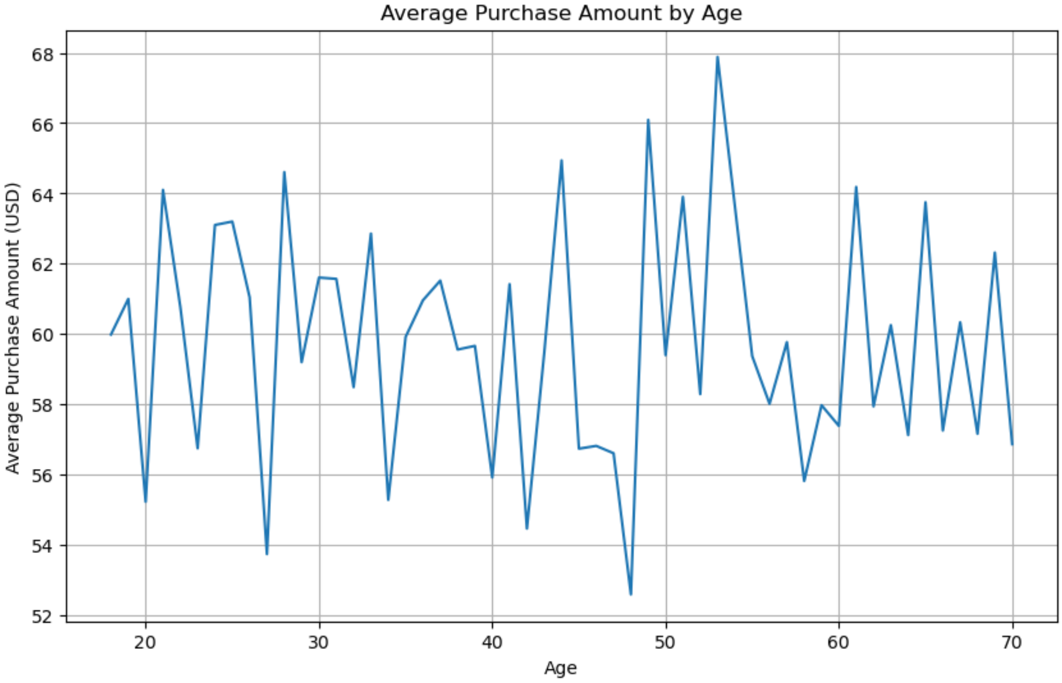



```
Correlation Matrix:
                         Customer ID       Age  Purchase Amount
(USD)  \
Customer ID                 1.000000 -0.004079
0.011048
Age                        -0.004079  1.000000                -
0.010424
Purchase Amount (USD)       0.011048 -0.010424
1.000000
Review Rating               0.001343 -0.021949
0.030776
Previous Purchases         -0.039159  0.040445
0.008063

                       Review Rating  Previous Purchases
Customer ID                 0.001343           -0.039159
```

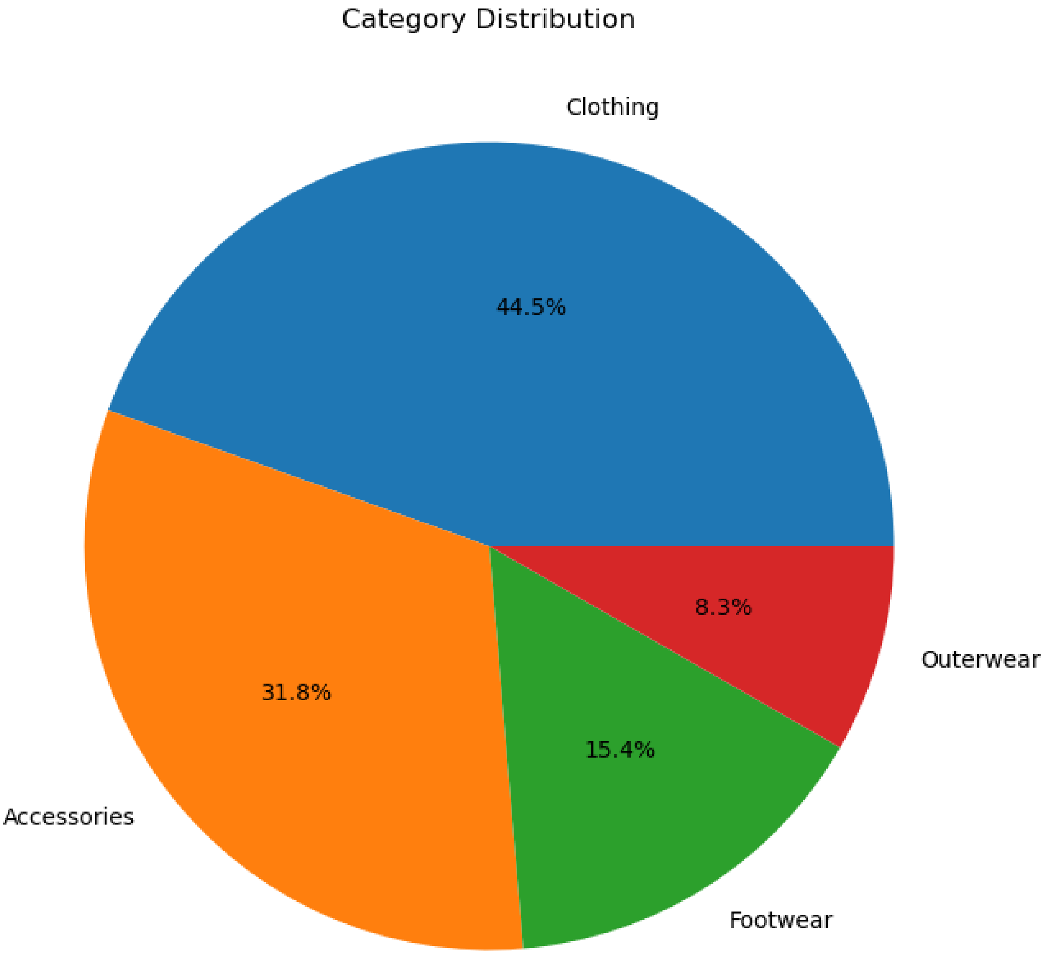
```
192      Age                      -0.021949         0.040445
193      Purchase Amount (USD)     0.030776         0.008063
194      Review Rating             1.000000         0.004229
195      Previous Purchases        0.004229         1.000000
```



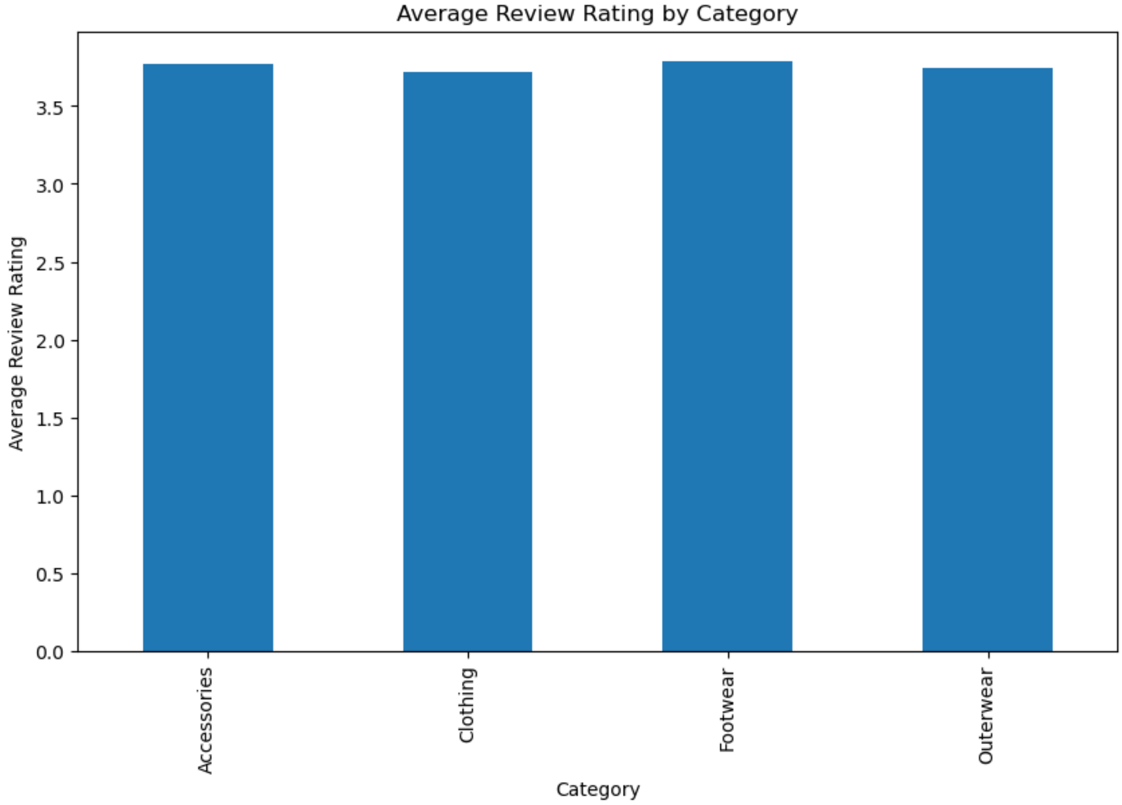
196

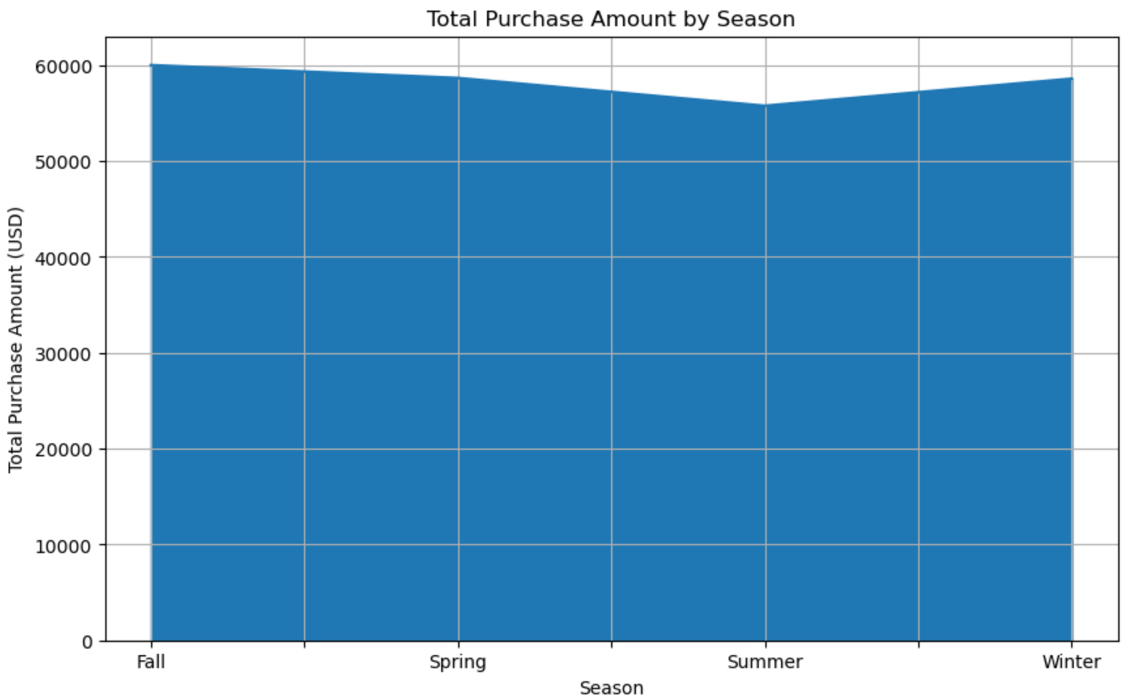

Average Purchase Amount by Age
Average Purchase Amount (USD)
Age
68
66
64
62
60
58
56
54
52
20
30
40
50
60
70

197


Category Distribution
Clothing
44.5%
Accessories
31.8%
Footwear
15.4%
Outerwear
8.3%

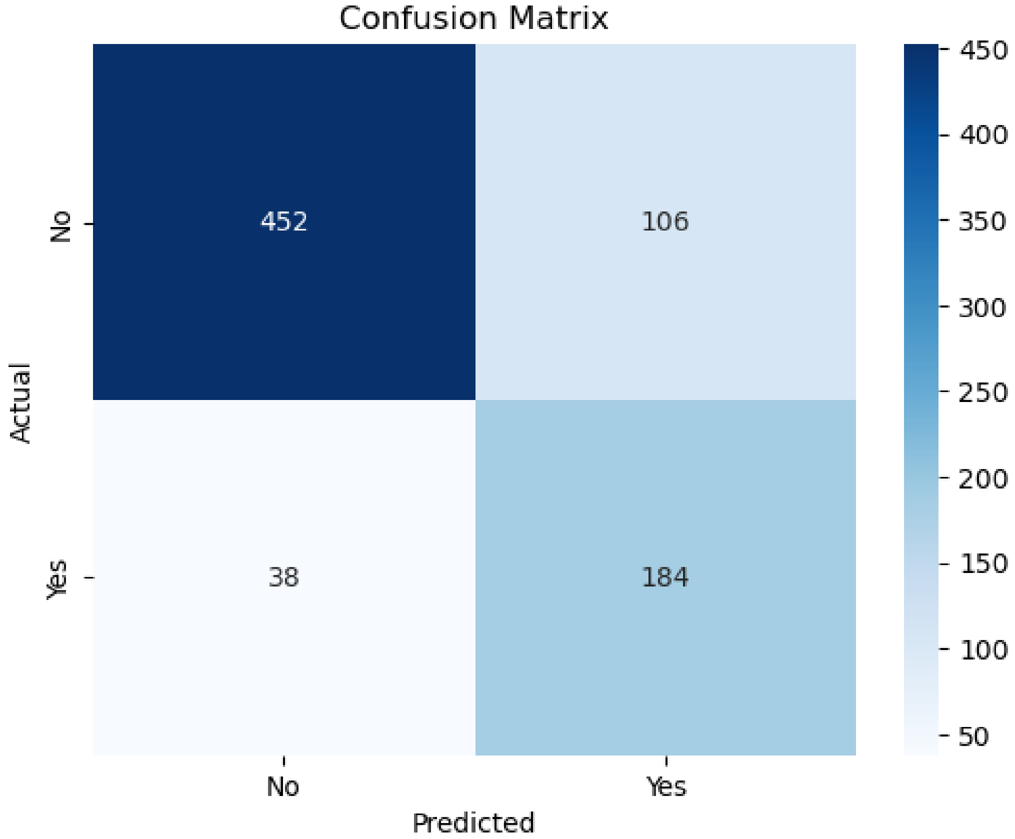
198



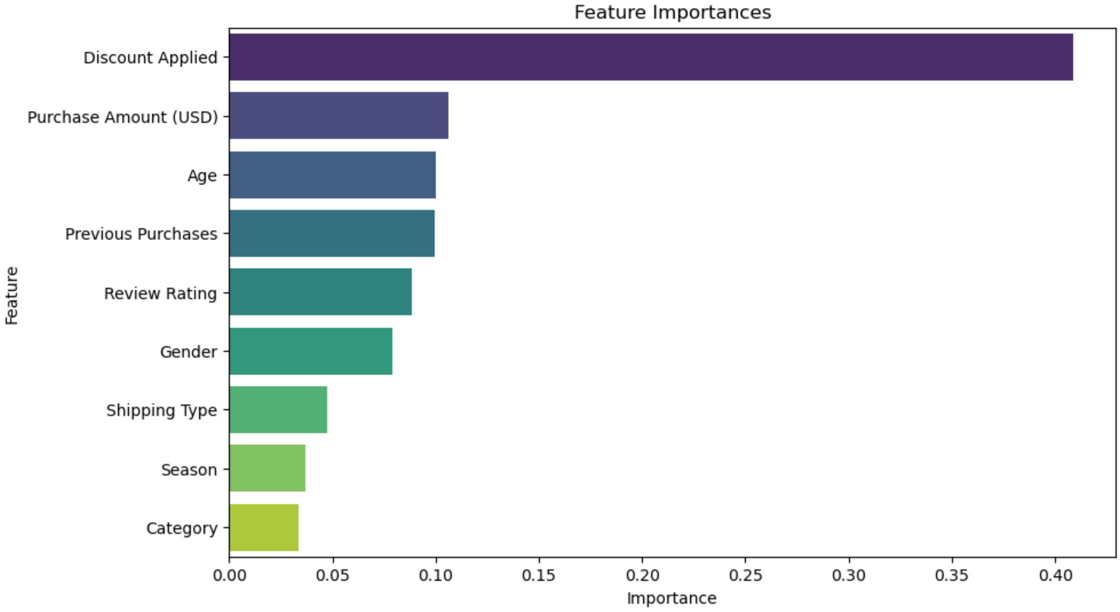



```
Cleaned data saved to: shopping_trends_cleaned.csv

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.81      0.86       558
           1       0.63      0.83      0.72       222

```

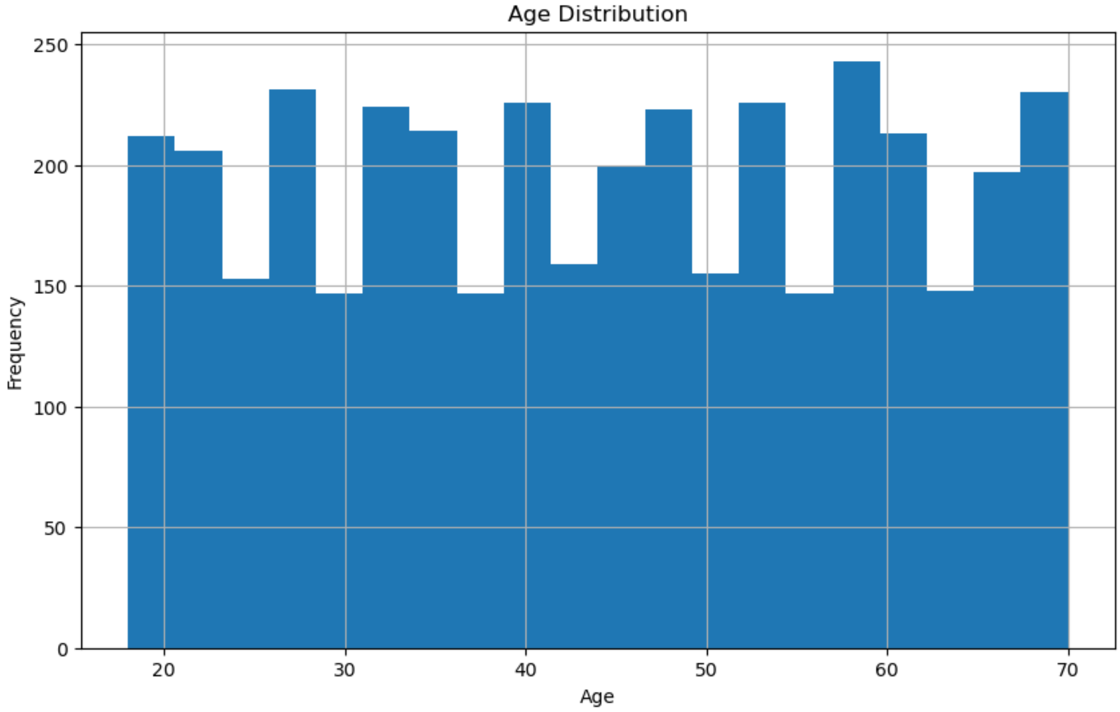
```
      accuracy                           0.82       780
     macro avg       0.78      0.82      0.79       780
  weighted avg       0.84      0.82      0.82       780

Accuracy Score: 0.8153846153846154
```



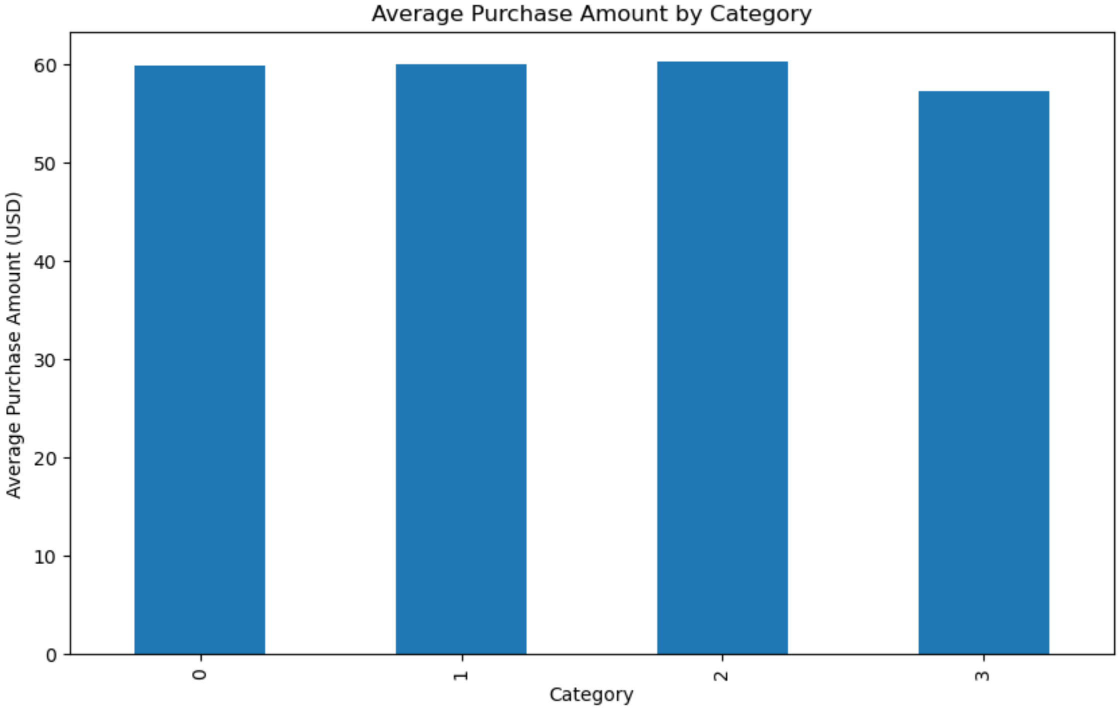

```
Feature Importances:
                  Feature  Importance
8        Discount Applied    0.408965
1  Purchase Amount (USD)    0.106326
0                     Age    0.100002
3     Previous Purchases    0.099473
2          Review Rating    0.088333
4                  Gender    0.079090
7          Shipping Type    0.047457
6                  Season    0.036807
5                Category    0.033547
C:\Users\dattu\AppData\Local\Temp\ipykernel_14220\373244521.py:143: FutureWarning:

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.

  sns.barplot(data=importance_df, x='Importance', y='Feature', palette='viridis')
```

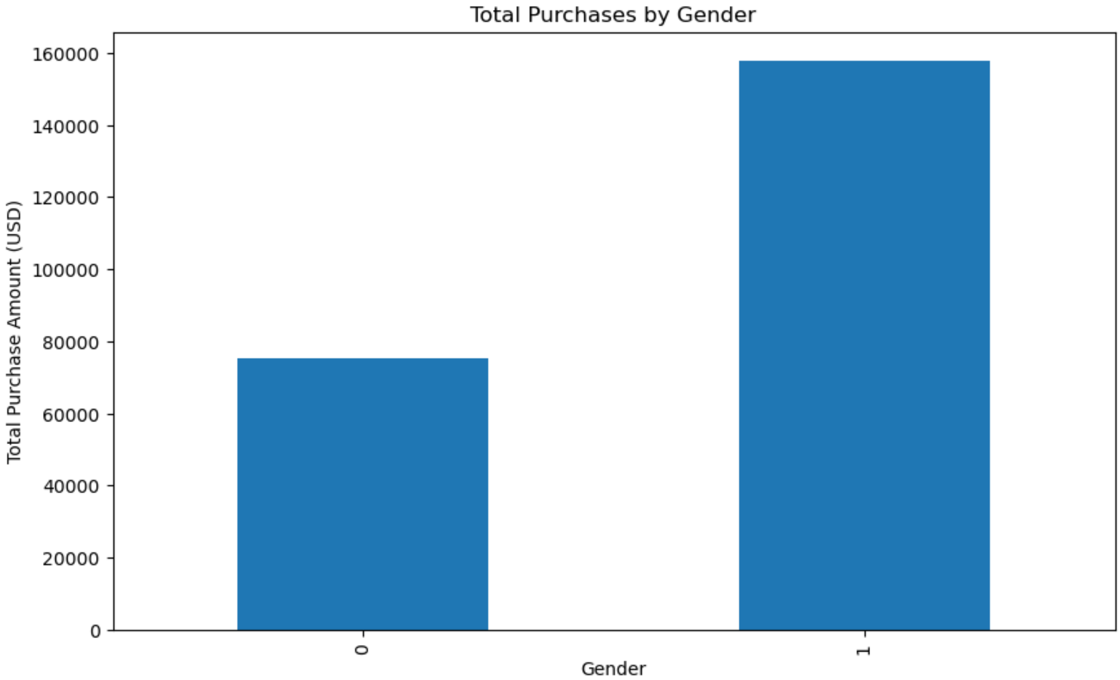

Feature Importances
Discount Applied
Purchase Amount (USD)
Age
Previous Purchases
Review Rating
Gender
Shipping Type
Season
Category
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance

232

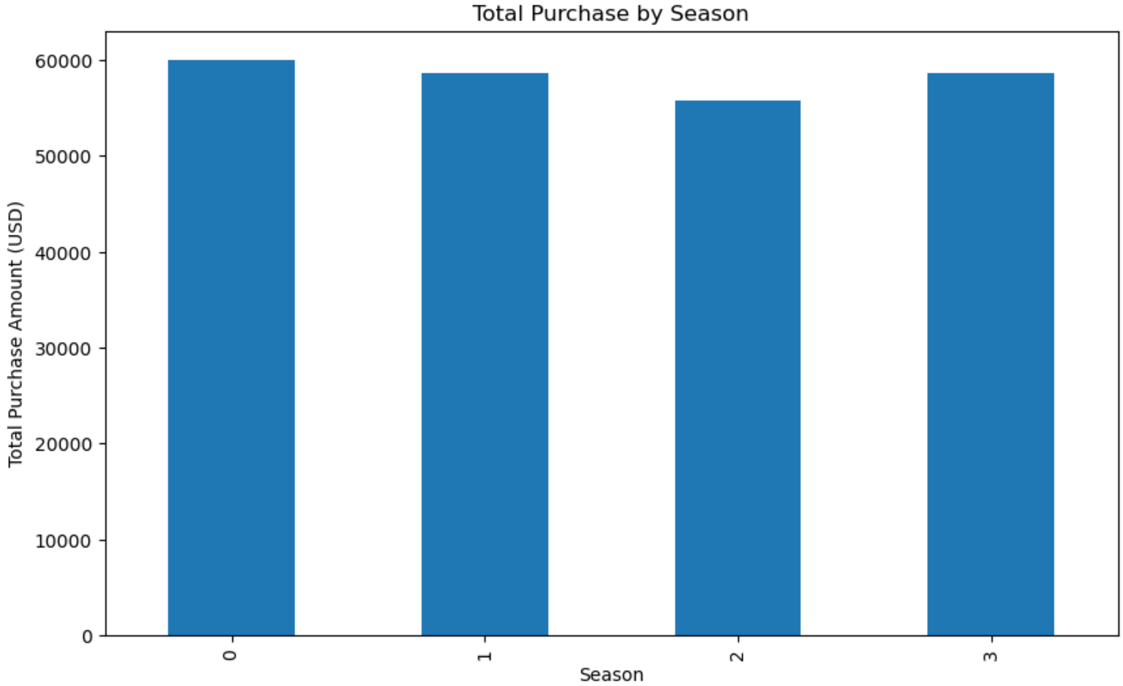

Age Distribution
250
200
150
100
50
0
Frequency
20
30
40
50
60
70
Age

233

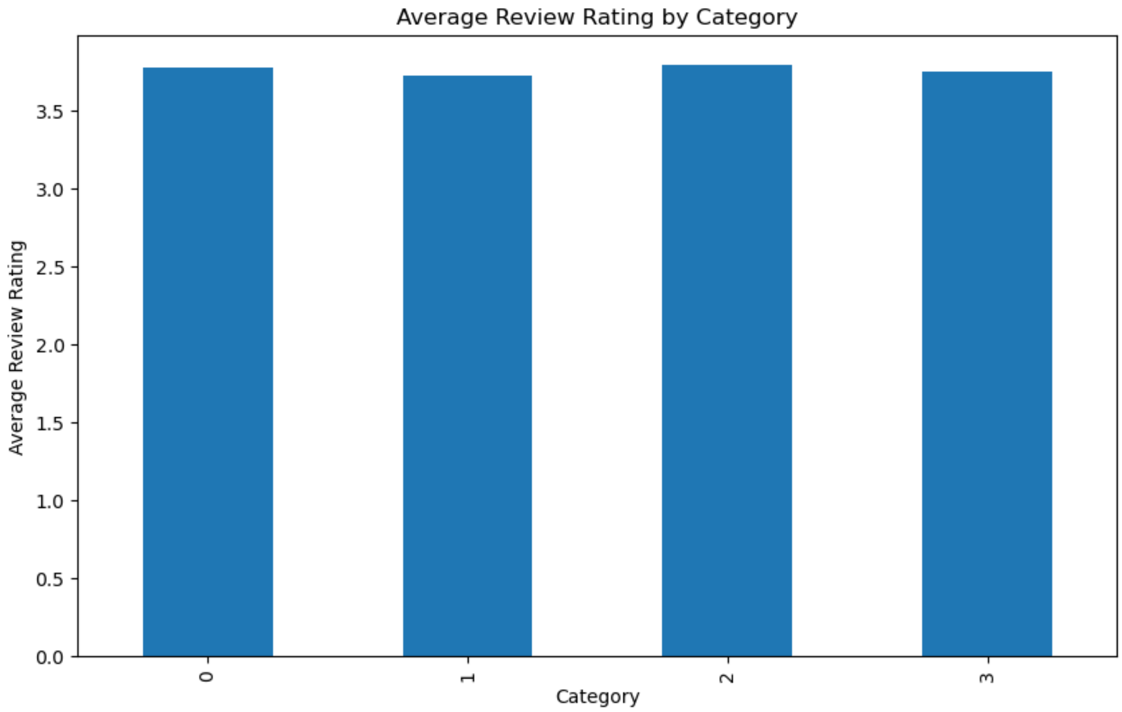





```
Most Common Items in Each Category:
 Category
0    Jewelry
1     Blouse
2    Sandals
3     Jacket
Name: Item Purchased, dtype: object
```


Total Purchase by Season
Total Purchase Amount (USD)
60000
50000
40000
30000
20000
10000
0
0
1
2
3
Season

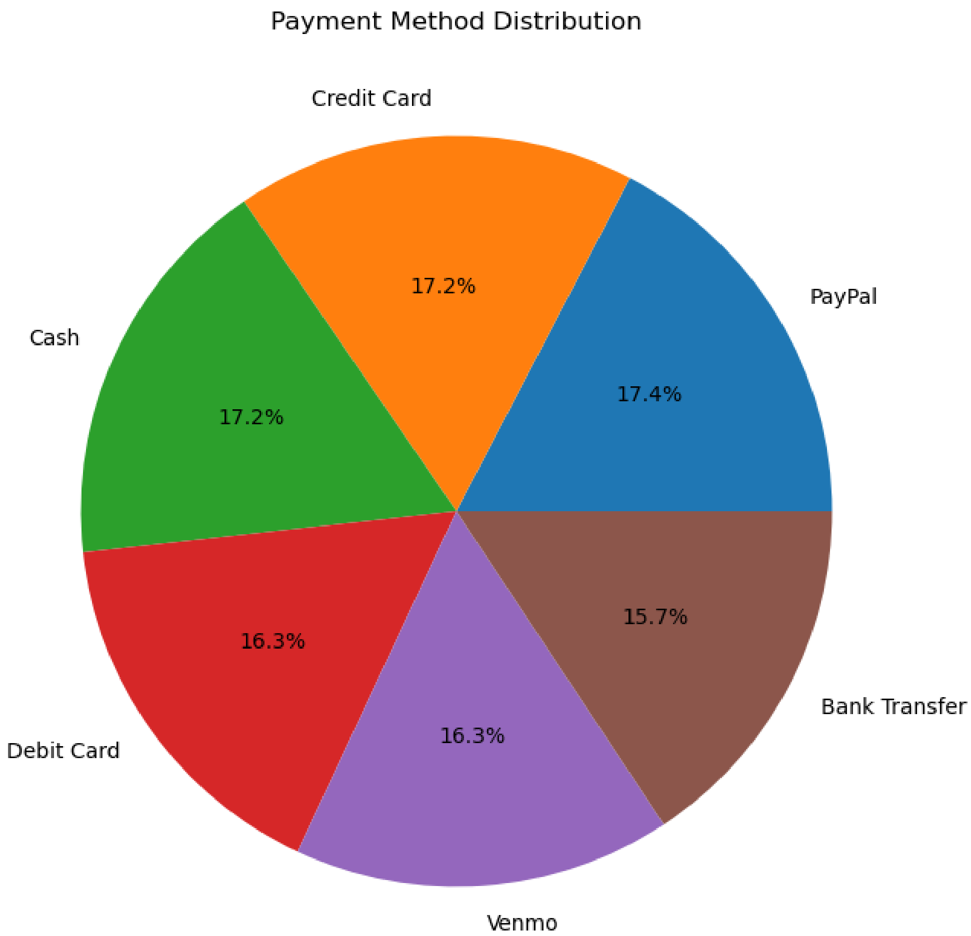
244


Average Review Rating by Category
Average Review Rating
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
Category

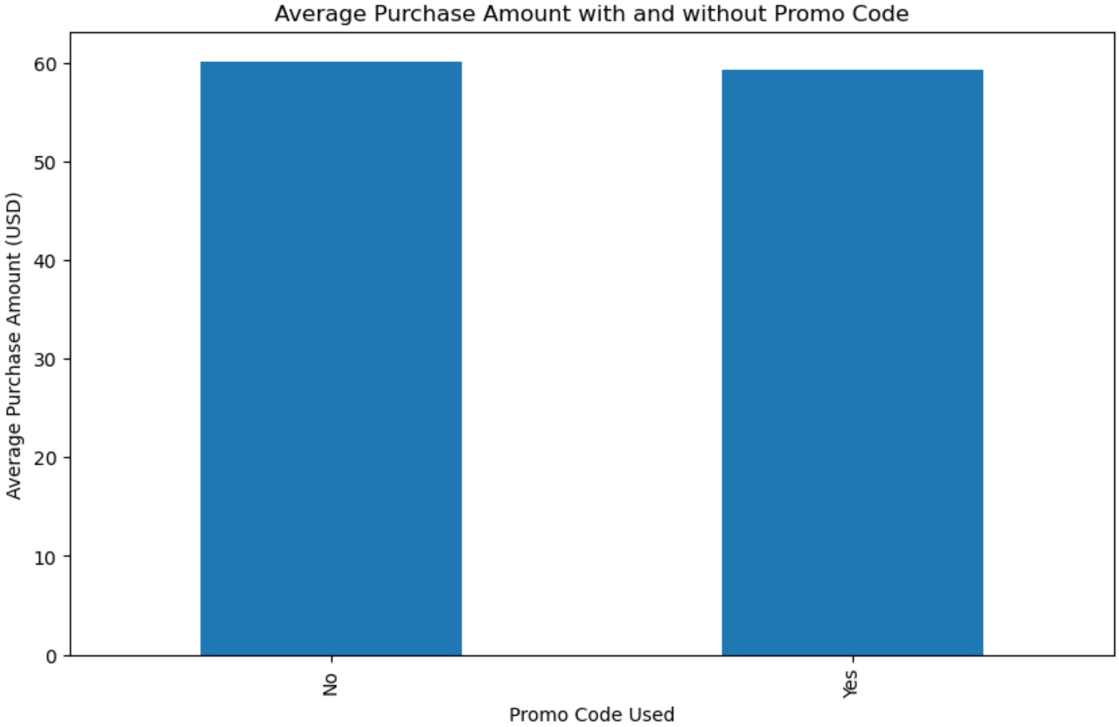
245


Total Purchase by Subscription Status
Total Purchase Amount (USD)
160000
140000
120000
100000
80000
60000
40000
20000
0
No
Yes
Subscription Status

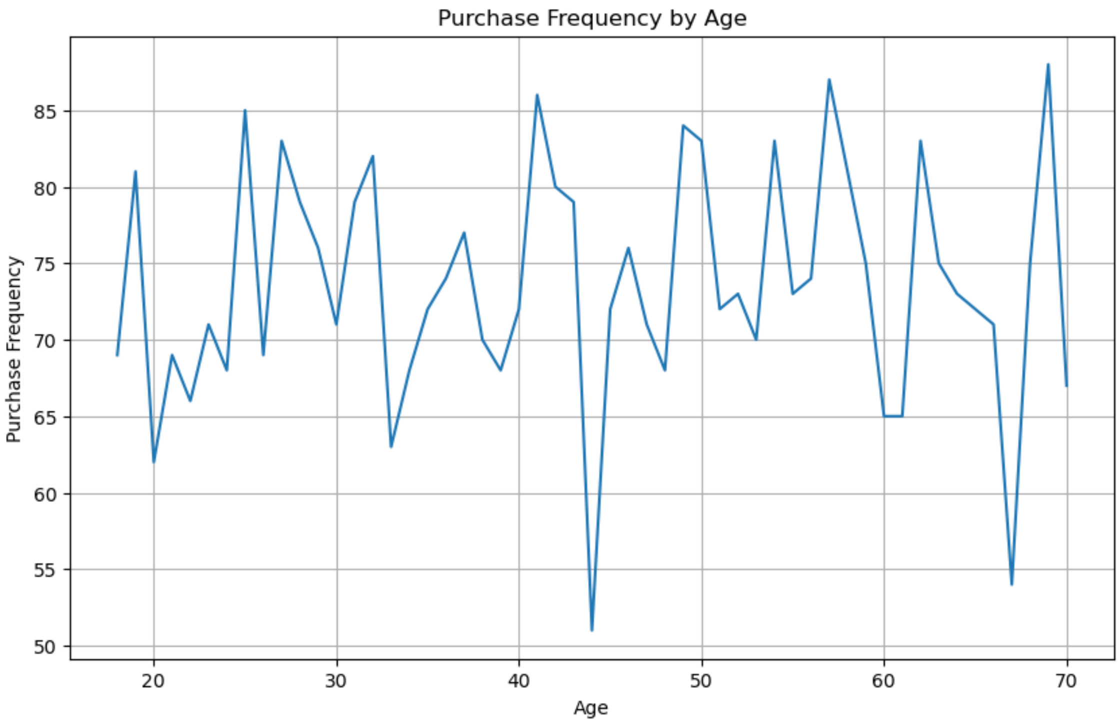
246


Payment Method Distribution
Credit Card
17.2%
PayPal
17.4%
Cash
17.2%
16.3%
15.7%
Bank Transfer
Debit Card
16.3%
Venmo

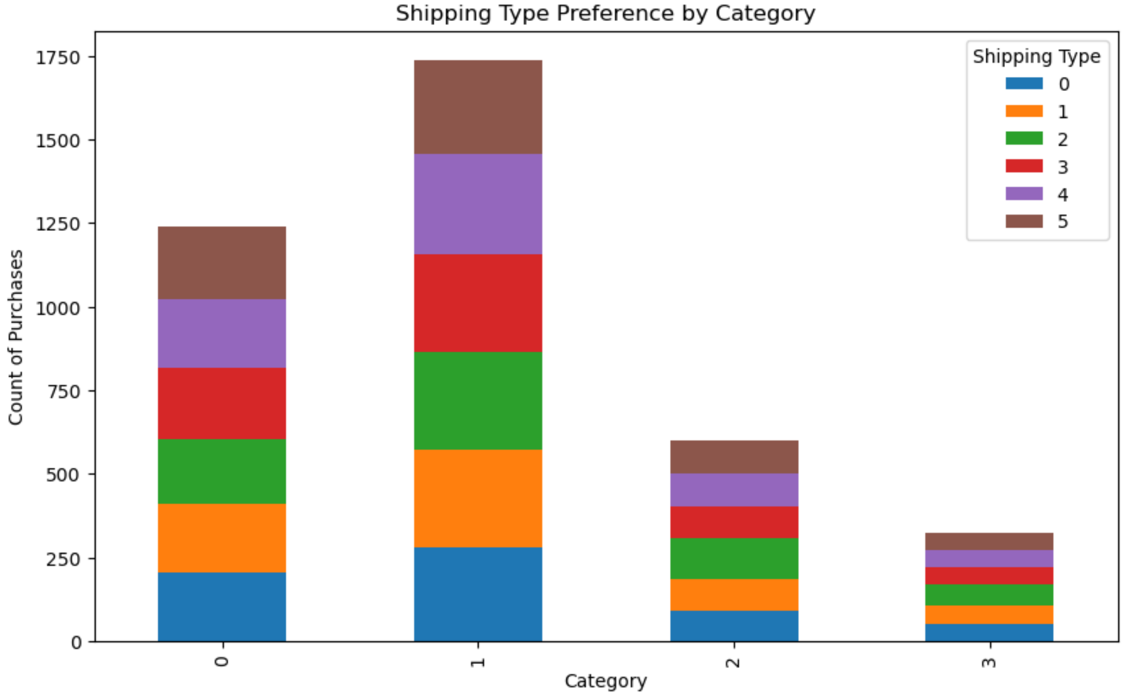
247


Average Purchase Amount with and without Promo Code
Average Purchase Amount (USD)
0
10
20
30
40
50
60
No
Yes
Promo Code Used

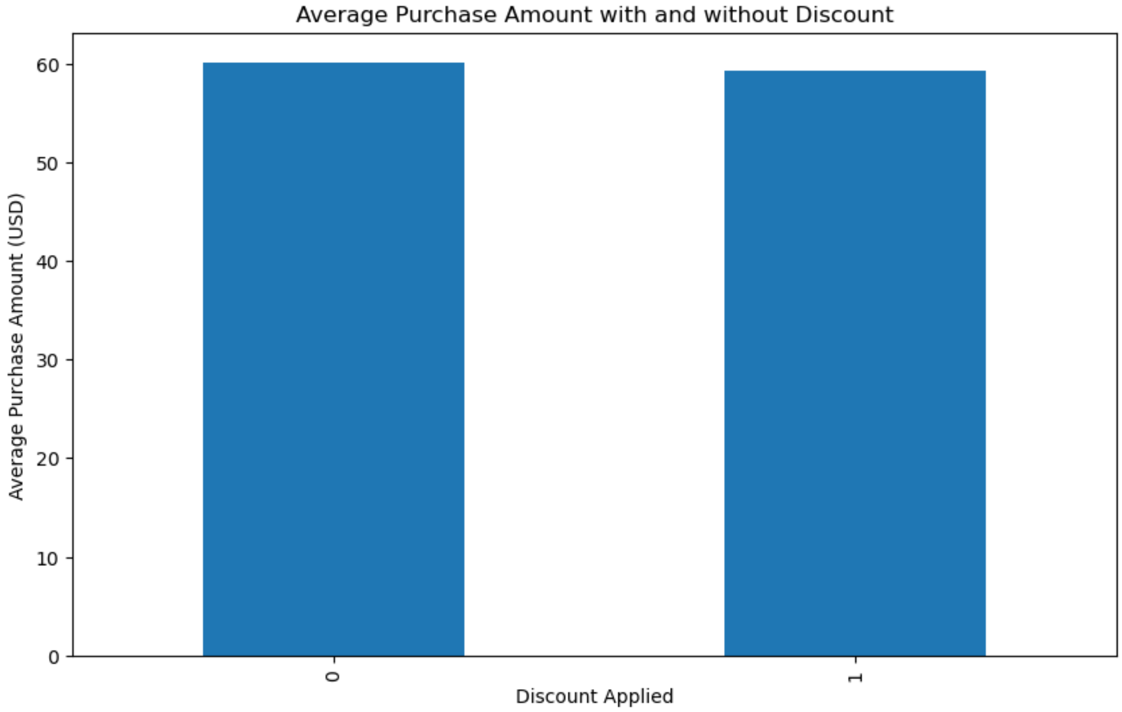
248


Purchase Frequency by Age
Purchase Frequency
50
55
60
65
70
75
80
85
20
30
40
50
60
70
Age

249

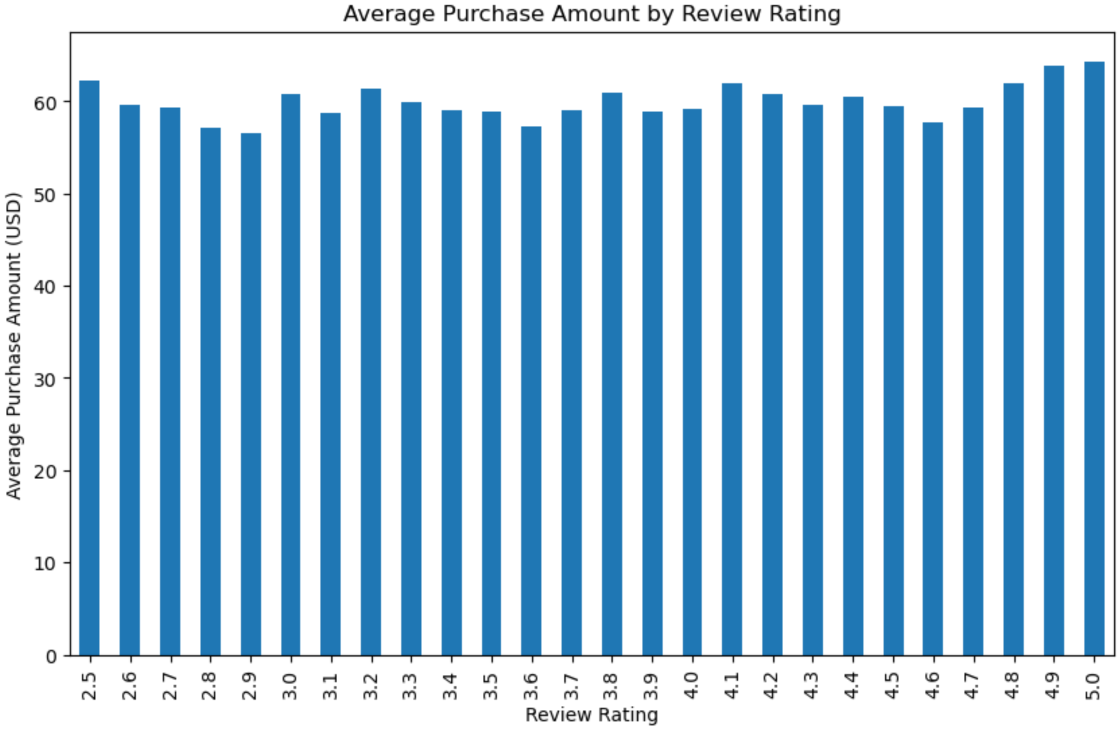


250

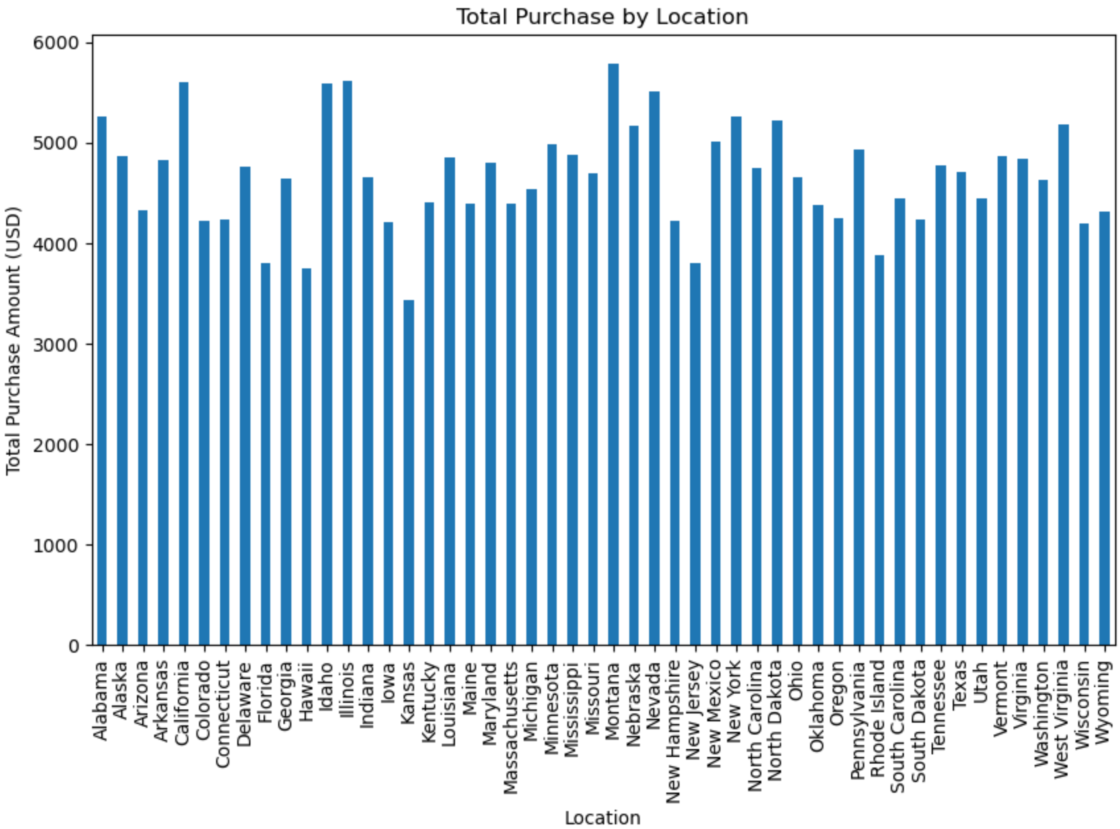


251

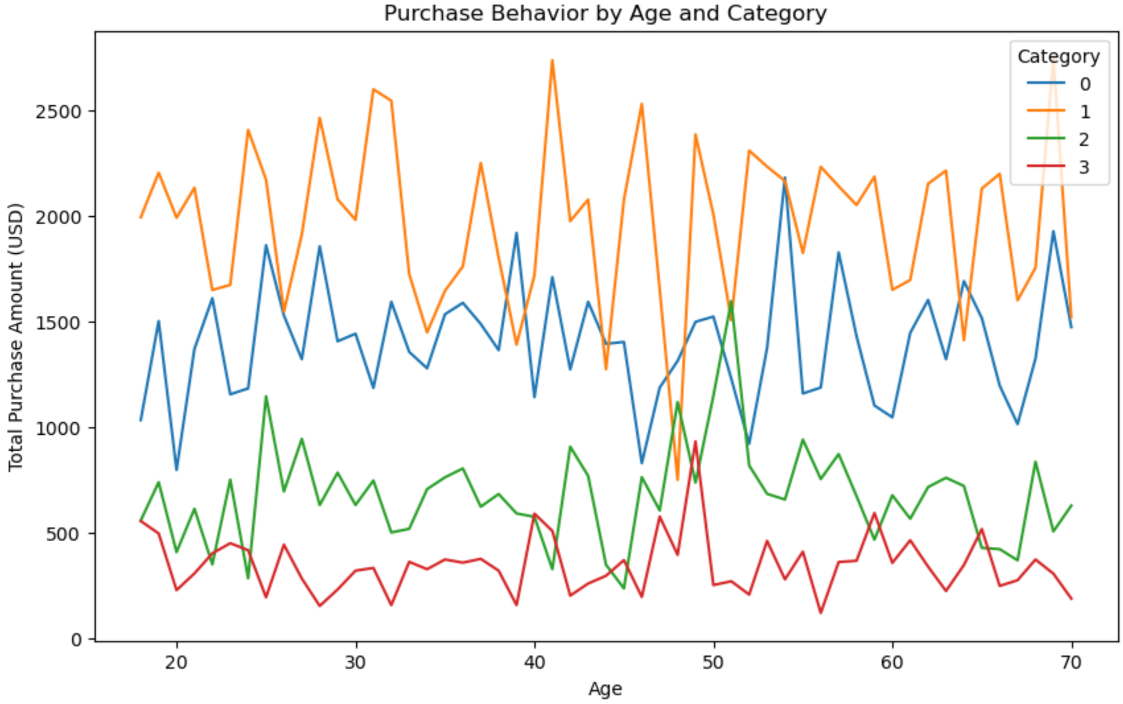
252
253 Average Number of Previous Purchases: 25.35153846153846


Average Purchase Amount by Review Rating
Average Purchase Amount (USD)
0
10
20
30
40
50
60
2.5
2.6
2.7
2.8
2.9
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5.0
Review Rating

254


Total Purchase by Location
Total Purchase Amount (USD)
0
1000
2000
3000
4000
5000
6000
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Location

255



256

257

258 In [ ]:

4.2 GitHub Link for Code:

https://github.com/AdityaJha2901/Identifying-Shopping-Trends-Using-Data-Analysis

CHAPTER 5

Discussion and Conclusion

5.1 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.

5.2 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.