EDA PROJECT - CSM353

Exploratory Data Analysis on Shopping Trends Analysis

Academic Task

Submitted to:

Mr. Ved Prakash Chaubey

UID : 63892

Dated: 18.11.2024

SUBMITTED BY

Name of student: ***Muhammed Althaf K***

Registration Number: ***12205672***

Supervisor Certificate

Lovely Professional University

School of Computer Science and Engineering

Certificate of Supervision

This is to certify that the project report titled " Shopping trend Analysis" has been carried out by Muhammed Althaf K (Registration No. 12205672) under my supervision in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab.

The project represents the student's original effort and has not been submitted to any other institution or university for the award of any degree or diploma. This report offers a comprehensive analysis of Shopping Trend data, emphasizing the critical factors influencing shopping trends, such as seasonality, consumer demographics, promotional effectiveness, and product-specific sales patterns. Through detailed exploratory data analysis (EDA) and visually engaging data visualizations, the study explores strategies to optimize inventory management, enhance customer targeting, and predict sales more accurately.

I hereby approve this project report and consider it worthy of submission for evaluation.

Ved Prakash Chaubey.

School of Computer Science and Engineering

Lovely Professional University

Signature:_____

Acknowledgement

I would like to express my deepest gratitude to Lovely Professional University for providing me with the opportunity to undertake this project on Exploratory Data Analysis (EDA) in the domain of shopping analytics, which has significantly enhanced my knowledge of data science and its applications in real-world scenarios, specifically in shopping trends analysis. I extend my heartfelt thanks to my mentor, Mr. Ved Prakash Chaubey, for his invaluable guidance, support, and feedback throughout the project, which were crucial in refining the analysis and improving its overall quality. I am also grateful to my family and friends for their constant support and encouragement, motivating me to push the boundaries of my knowledge and achieve the objectives of this project. Finally, I acknowledge UpGrad and other open-source communities that provided access to valuable resources and datasets, enabling the successful completion of this project.

Table of Contents

Abstract

This project focuses on analyzing shopping trends through a comprehensive dataset to uncover patterns in customer behavior, purchase preferences, and seasonal trends. By leveraging data, the analysis aims to provide actionable insights that can enhance decision-making in marketing, inventory management, and customer engagement. The dataset includes detailed transaction records, such as customer demographics, product categories, purchase amounts, and seasonal attributes, offering a rich foundation for exploring consumer habits and their underlying drivers. To achieve these objectives, Exploratory Data Analysis (EDA) techniques and multivariate analyses were employed. EDA helped in understanding data distributions, identifying trends, and examining relationships between variables, offering deeper insights into shopping behaviors. Key features like purchase amounts, customer age, gender, and discount usage were analyzed to identify patterns and understand the factors influencing consumer decisions. Multivariate techniques further explored how seasonal trends and discounts impact overall purchasing habits. The analysis was carried out using Python's data science libraries, including pandas, matplotlib, seaborn, and plotly, which facilitated efficient data preprocessing, such as handling missing values and transforming raw data into analyzable formats. Visualizations, including interactive and static charts, played a crucial role in depicting trends and the effects of seasonal changes. The findings from this project provide valuable insights for businesses to optimize their strategies and align them with consumer preferences.

Introduction

The dataset used in this project provides detailed insights into shopping trends by capturing key information such as customer demographics, purchase details, and product attributes. These features form the foundation for understanding how different factors influence consumer behavior and purchasing decisions. By analyzing this dataset, the study seeks to uncover meaningful trends, offering valuable insights into customer preferences and the dynamics of their shopping habits. The primary objective of the analysis is twofold: first, to understand customer preferences and purchasing patterns, such as the frequency of purchases, the types of products preferred, and the demographic groups that contribute the most to sales; and second, to identify seasonal trends in sales and evaluate the impact of discounts or promotional activities on customer buying behavior. These objectives help businesses tailor their strategies to maximize revenue and customer satisfaction. Exploratory Data Analysis (EDA) and multivariate analysis serve as critical tools in this process. EDA allows for a systematic exploration of the dataset to identify patterns, relationships, and anomalies, providing a clear picture of the underlying trends. Multivariate analysis adds depth by examining the interaction between multiple variables, such as the combined effect of demographics and discounts on purchase behavior. Together, these analytical approaches lay the groundwork for deeper analyses or predictive modeling, enabling businesses to make data-driven decisions.

Methodology

The project follows a structured approach to Exploratory Data Analysis, including:

Data Collection and Preparation:

- The dataset, shopping_trends.csv, contains comprehensive records of shopping transactions, including customer demographics, purchase details, and product attributes.
- Initial data cleaning involved handling missing values through imputation and removal of duplicate records, as well as standardizing formats for consistency (e.g., date formats and numerical units).

Univariate Analysis:

- Individual features such as Purchase Amount, Age, and Frequency of Purchases were analyzed using descriptive statistics and visualizations. Tools like histograms, box plots, and bar charts were utilized to understand distributions and detect outliers.

Time Series Analysis:

- Conducting a time series analysis to examine trends in shopping behavior over different seasons, such as changes in purchase volumes, customer preferences, and the impact of discounts on sales over time.

Data Visualization:

- Libraries such as Matplotlib, Seaborn, and Plotly were employed to create visual representations of data, including interactive plots to highlight trends, seasonal effects, and customer behavior patterns.

Result Interpretation:

- The findings were interpreted to identify key insights, such as seasonal purchase trends, demographic preferences, and the impact of discounts or promotions. These insights provide actionable recommendations for marketing strategies and inventory planning.

RESULT AND ANALYSIS

- Importing the libraries

```
import pandas as pd
import numpy as np
import datetime as dt
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px

%matplotlib inline

import warnings
warnings.filterwarnings('ignore')
```

```
data = pd.read_csv('shopping_trends.csv')
```

Fig.1

- Loading the dataset

```
data.head()
```

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

Fig.2

- Checking of Null Values in the dataset

```
In [6]: # check Missing value
        data.isna().sum()

Out[6]: 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
        Payment Method             0
        Shipping Type              0
        Discount Applied           0
        Promo Code Used            0
        Previous Purchases         0
        Preferred Payment Method   0
        Frequency of Purchases     0
        dtype: int64
```

Fig.3

- Plotting Age Distribution Histogram with Density Curve

```
fig, ax = plt.subplots(figsize = (20, 8))
ax.hist(data['Age'], bins = 20, edgecolor = 'black', alpha = 0.7, color = "#89CFF0", density = True)
data['Age'].plot(kind = 'kde', color = "#FF69B4", ax = ax)

plt.xlabel('Age',fontsize=15)
plt.ylabel('Count / Density',fontsize=15)
plt.title('Age Distribution Histogram with Density Curve',fontsize=20)
ax.legend(['Density Curve', 'Histogram'])
plt.show()
```

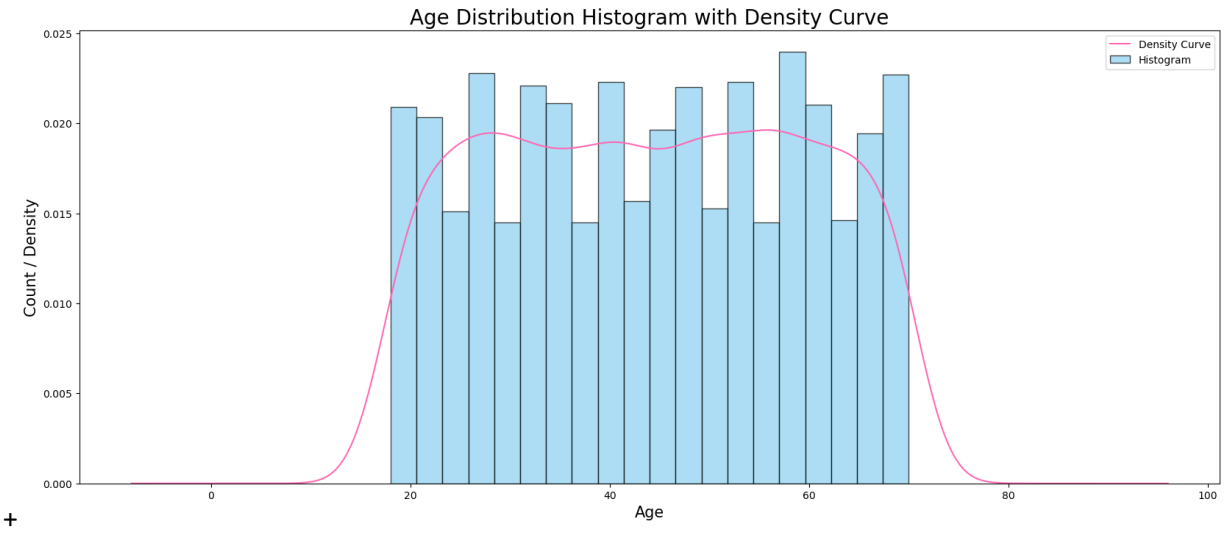

Fig.4

- Plotting a Pie-Chart on Gender

```
gender = data.Gender.value_counts()
print(gender,'\n')

plt.pie(gender,labels=gender.index,colors=["#89CFF0", "#FF69B4"],autopct='%.0f%%',explode=(0,0.1))
plt.legend(labels = gender.index, loc = "best")
plt.title('Gender')
plt.show()
```

```
Gender
Male      2652
Female    1248
Name: count, dtype: int64
```

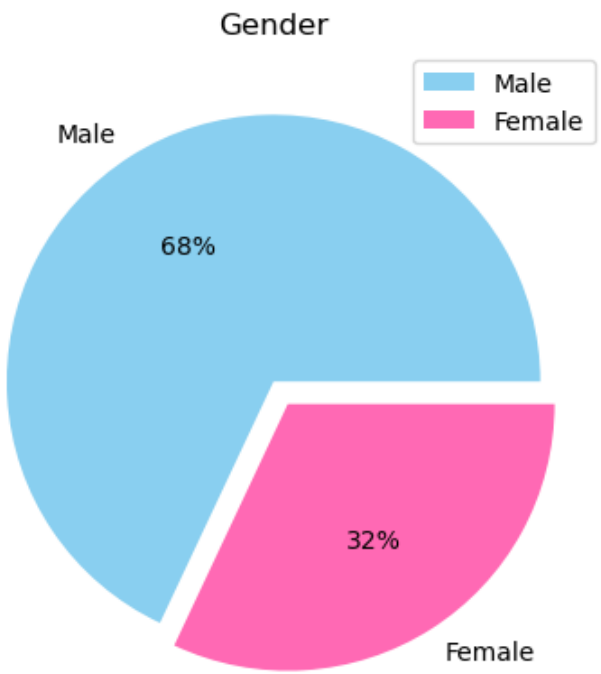

Fig.5

- Plotting Countplot on Item Purchased

```
plt.subplots(figsize=(20,8), dpi=100)
sns.countplot(data= data, x='Item Purchased',palette='cool')
plt.title("Item Purchased",fontsize=20)
plt.show()
```

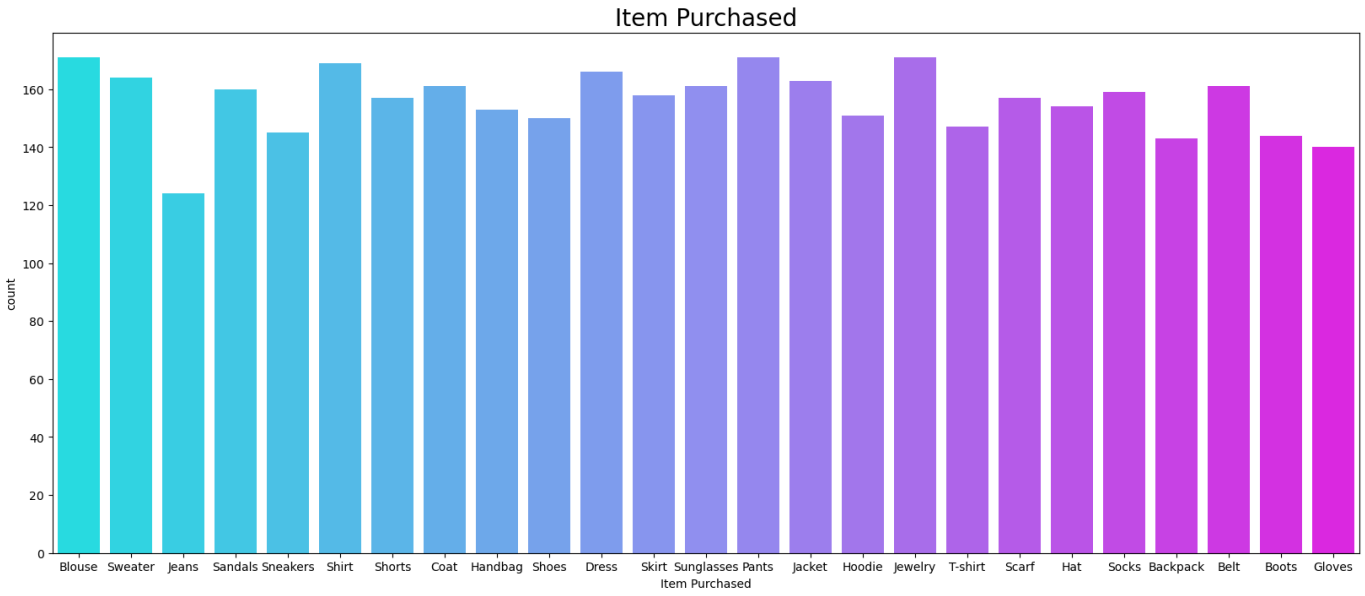

Fig.6

- Plotting Pie-Chart on Category

```
plt.figure(figsize=(5,8))
plt.pie(CountOfCategory,labels=CountOfCategory.index,autopct='%0.0f%%',explode=(0,0,0,0.1))
plt.legend(CountOfCategory.index,loc =2)
plt.show()
```

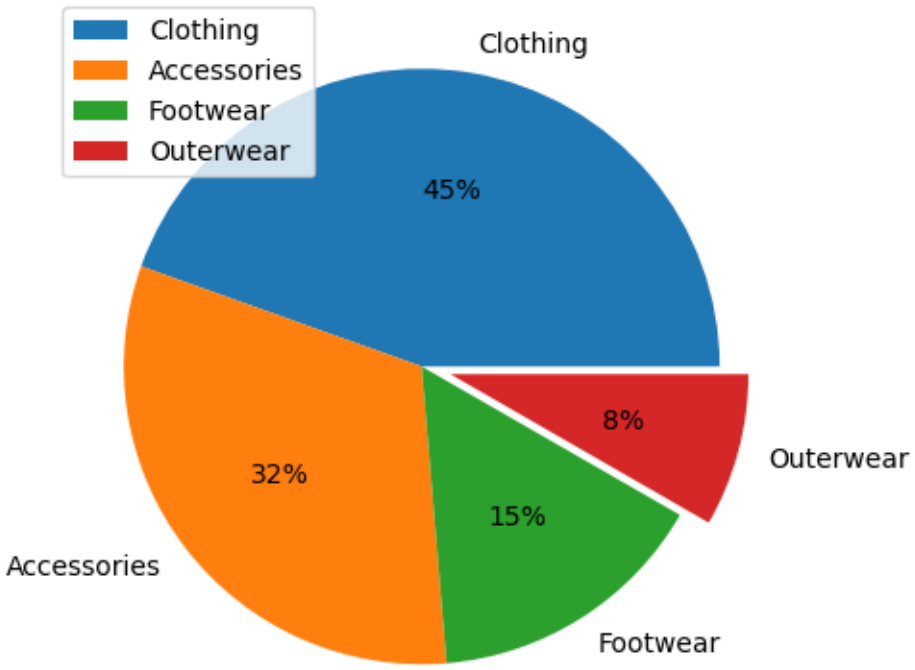

Fig.7

- Plotting Pie-Chart On Season

```
plt.figure(figsize=(5,8))
plt.pie(season,labels=['Spring','Fall','Winter','Summer'],autopct='%0.0f%%')
plt.legend( ['Spring','Fall','Winter','Summer'],loc =1)
plt.title('Season')
plt.show()
```

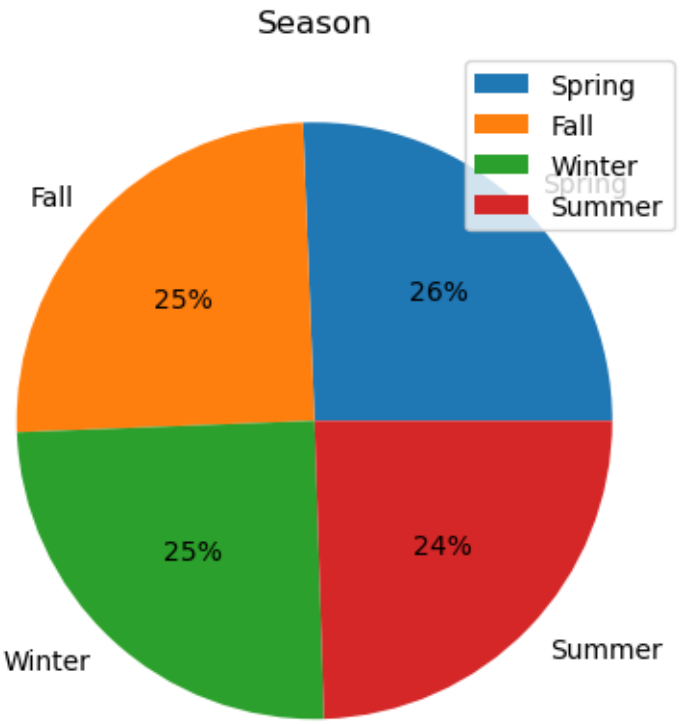

Fig.8

- Plotting BarChart on Gender and Purchase amount

Total purchase amount for each gender

```
data.groupby('Gender')['Purchase Amount (USD)'].sum()
```

```
Gender
Female     75191
Male      157890
Name: Purchase Amount (USD), dtype: int64
```

```
data.groupby('Gender')['Purchase Amount (USD)'].sum().plot(kind='bar',figsize=(5,5),colormap='cool',ylabel='Purchase Amount (USD)
```

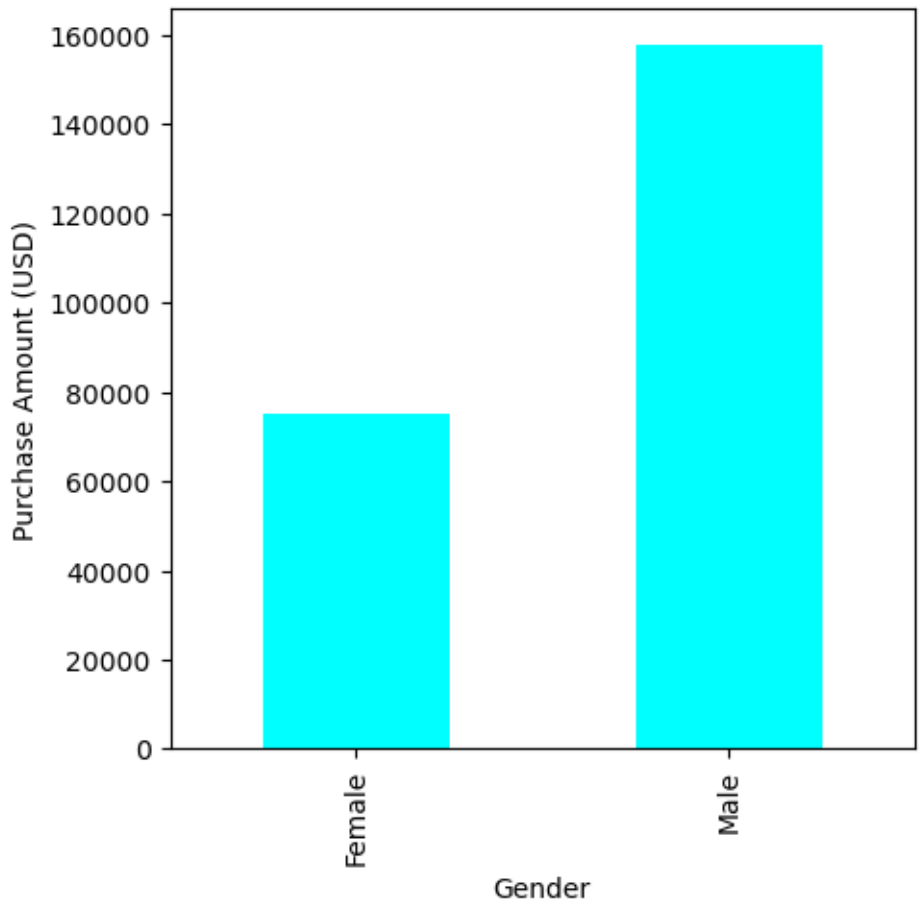

Fig.9

- Plotting Grouped Barchart on Season and Categories

To know the number of categories that are purchased in each season

```
pd.crosstab(data['Season'],data['Category'])
```

Category	Accessories	Clothing	Footwear	Outerwear
Season				
Fall	324	427	136	88
Spring	301	454	163	81
Summer	312	408	160	75
Winter	303	448	140	80

```
pd.crosstab(data['Season'],data['Category']).plot(kind='bar',figsize=(12,6),ylabel='Count')
```

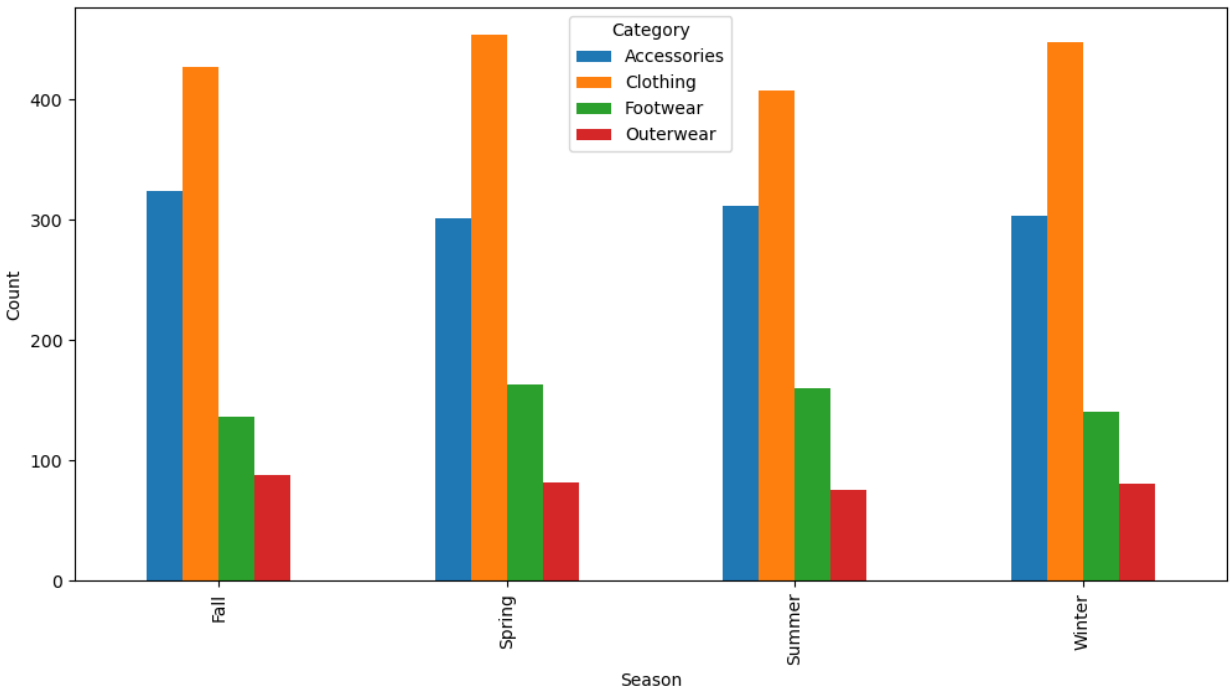

Fig.10

Questions:

1 What is the project about?
The project analyzes shopping trends to understand customer behaviors, preferences, and the impact of factors like seasonality, discounts, and demographics.

2 What dataset did you use?
The dataset, shopping_trends.csv, contains customer transaction details.

3 What is the primary objective of the project?
To uncover patterns in shopping behavior and provide actionable insights for marketing and inventory planning.

4 What time period does the dataset cover?
The dataset's time period is determined by transaction records; specific details depend on the data source.

5 What tools were used for the analysis?
Python libraries like pandas, matplotlib, and seaborn for analysis and visualization.

6 What are the key attributes in the dataset?
Attributes include Age, Gender, Purchase Amount, Season, Item Purchased, and Discount Applied.

7 How did you handle missing values?
Missing values were handled through imputation for categorical variables and removal for rows with critical missing information.

8 Were there duplicate records? How were they handled?
Duplicate entries were identified and removed to ensure data integrity.

9 What inconsistencies were found in the data?
Issues like inconsistent date formats and currency units were standardized.

10 What transformations were applied to the data?
Transformations included encoding categorical variables and creating derived features like "Purchase Season."

11 What is the average purchase amount?
The average purchase amount was calculated and visualized with a histogram.

12 Which age group spends the most?
Age groups were analyzed to identify the segment with the highest spending.

13 What is the distribution of purchases by gender?
A bar chart was used to compare purchase counts for Male and Female.

14 What is the most purchased item?
A frequency analysis was conducted to identify the top item.

15 How many transactions were made per season?
Seasonal purchase distributions were visualized.

16 What is the relationship between age and purchase amount?
Trends were analyzed to determine how spending changes with age.

17 Do discounts significantly increase purchase amounts?
Purchases with and without discounts were compared to evaluate impact.

18 Which season generates the highest revenue?
Seasonal revenue was calculated and compared.

19 Are there gender differences in discount utilization?
Gender-specific trends in discount usage were identified.

20 What payment methods are most popular?
Payment method frequencies were analyzed.

21 What new features were created?
Features like "Discount Usage," "Purchase Season," and one-hot encoded variables.

22 Why were these features created?
To enhance analysis by deriving insights not directly available in the raw data.

23 How was the Purchase Season feature derived?
Based on transaction dates, mapping them to seasons (e.g., Spring, Summer).

24 What encoding method was used for categorical variables?
One-hot encoding was used to convert categorical features into numerical format.

25 How did feature engineering improve the analysis?
It added dimensions to uncover patterns related to seasonality and discounts.

26 What types of visualizations were created?
Histograms, bar charts, and box plots were used to explore data distributions.

27 Which visualization was the most insightful?
Seasonal purchase trends highlighted how revenue peaks during certain months.

28 What tools were used for visualizations?
Matplotlib and Seaborn libraries were used.

29 How were demographic patterns visualized?
Bar charts showed the distribution of purchases across age and gender.

30 What challenges were faced in visualizing the data?
Large categorical variables required grouping for effective visualization.

31 What are the key findings of the project?
Seasonal trends, demographic preferences, and the impact of discounts are major findings.

32 Which factor most influences purchase behavior?
Discounts and promotions have a significant impact.

33 What seasonal trends were observed?
Certain seasons saw spikes in sales, particularly for specific product categories.

34 Are there customer segments with distinct behaviors?
Regular and one-time buyers exhibited different spending patterns.

35 How do discounts affect sales?
Sales volume increased significantly during discount periods.

36 What challenges were encountered during data cleaning?
Missing values and inconsistent data formats were significant hurdles.

37 Were there any computational challenges?
Processing large datasets and rendering complex visualizations took more time.

38 How were missing values handled without biasing results?
Imputation methods were chosen carefully to preserve data patterns.

39 What difficulties arose in feature engineering?
Deriving seasonality and encoding complex categorical variables were challenging.

40 How were outliers managed?
Outliers were visualized using box plots and treated selectively based on their relevance.

41 What could improve the analysis?
Incorporating predictive models to forecast future sales.

42 What additional data would be useful?
Customer feedback and more detailed transaction timestamps.

43 How could clustering be applied?
To segment customers for targeted marketing strategies.

44 What role could machine learning play?
Predictive modeling for sales and recommendations could enhance insights.

45 Are there plans to apply this analysis?
The findings could guide real-world decisions in marketing and inventory.

46 What version of Python was used?
Python 3.9 was used for the analysis.

47 What tools handled missing values?
Imputation techniques with pandas were applied.

48 How was the dataset imported?
Using the pandas.read_csv() function.

49 How were visualizations customized?
By adjusting figure sizes, colors, and labels in Matplotlib and Seaborn.

50 What libraries were used for data processing?
Libraries like pandas, numpy, matplotlib, and seaborn supported the analysis. □

51 What is the main focus of the project?
The project focuses on analyzing shopping trends to provide insights into customer behavior and seasonal sales patterns.

52 Why is this project important?
It helps businesses understand consumer behavior, optimize inventory, and improve marketing strategies.

53 Who would benefit from this analysis?
Retailers, marketing teams, and data analysts aiming to improve decision-making in sales and customer engagement.

54 What kind of trends were identified?
Seasonal patterns, customer demographic preferences, and the influence of discounts on purchasing behavior.

55 How does the dataset contribute to the project goals?
It provides detailed information on transactions, enabling a data-driven understanding of customer behavior.

56 What file format was the dataset in?
The dataset was in CSV format.

57 How many records does the dataset contain?
The exact number depends on the dataset size provided.

58 What types of variables are in the dataset?
Categorical (e.g., Gender, Season) and numerical (e.g., Purchase Amount, Age) variables.

59 What is the target variable in this project?
The target variable could be Purchase Amount or Frequency of Purchases, depending on the analysis focus.

60 What additional information would enhance this dataset?
Adding customer loyalty, promotional campaign details, and feedback data.

61 Were there any null values?
Yes, missing values were identified and addressed during preprocessing.

62 How were null values detected?
Using .isnull().sum() in pandas to find the number of missing values.

63 What strategies were used for missing value imputation?
Mode imputation for categorical data and mean/median imputation for numerical fields.

64 What other preprocessing steps were performed?
Removing duplicates, correcting data types, and standardizing formats.

65 What quality checks were conducted?
Verifying data consistency, checking for outliers, and ensuring variable relevance.

66 Which product category had the highest sales?
A frequency distribution revealed the most popular product category.

67 What is the range of purchase amounts?
The minimum and maximum purchase values were identified using descriptive statistics.

68 How is customer age distributed?
A histogram of the Age variable was used to analyze its distribution.

69 Which gender contributes more to revenue?
Revenue contributions by gender were calculated and compared.

70 What are the peak times for purchases?
Time-based analysis identified hours or days with the highest transaction frequency. □

71 How was the 'Discount Usage' feature created?
It was derived by flagging transactions where discounts were applied.

72 How were seasons mapped from dates?
Transaction dates were mapped to seasons (e.g., Spring, Summer) based on their month.

73 What variables were encoded?
Categorical variables like Gender and Season were one-hot encoded.

74 What was the role of feature engineering in the project?
It helped add meaningful dimensions to the data for better insights.

75 Was any variable scaled or normalized?
No scaling or normalization was performed as most analyses didn’t require it. □

76 What seasonal trends were observed?
Sales peaked during specific seasons, indicating seasonal demand.

77 How do demographics affect purchasing behavior?
Age and gender significantly influenced purchasing preferences.

78 What are the most frequent purchase categories?
Product categories like clothing or electronics appeared most often in the dataset.

79 What percentage of customers use discounts?
A large percentage of customers took advantage of discounts, as derived from the Discount Usage feature.

80 What is the average revenue per transaction?
The mean purchase amount provided an estimate of revenue per transaction.

81 What role did visualizations play in this project?
They helped identify trends and patterns in the data, making complex insights easier to interpret.

82 What were the most insightful visualizations?
Bar charts showing seasonal trends and demographic spending patterns.

83 What patterns did the bar charts reveal?
They highlighted differences in spending across genders and seasons.

84 Were there any surprising findings?
Certain seasons had unexpectedly low sales despite promotions.

85 What insights were drawn from box plots?
They identified outliers and variations in purchase amounts.

86 What was the biggest challenge during analysis?
Handling missing values without losing significant data trends.

87 How were outliers managed?
Outliers were retained unless they were clearly erroneous, as they might represent high-value customers.

88 Did any biases emerge in the data?
The dataset was slightly biased toward seasonal purchases.

89 Were there limitations to the analysis?
Lack of real-time data and customer feedback limited deeper behavioral insights.

90 How were these challenges addressed?
By careful preprocessing and focusing on actionable variables.

91 Who are the most frequent buyers?
Regular customers with high purchase frequency.

92 How does age influence purchase size?
Older customers tended to spend more per transaction.

93 What motivates purchases?
Discounts, seasonal demand, and product availability were key motivators.

94 What demographic spends the least?
Younger customers, on average, made smaller purchases.

95 What is the average discount utilization rate?
A high percentage of transactions included discounts.

96 How can this analysis be used in real life?
It can guide marketing strategies, inventory planning, and customer retention efforts.

97 What predictive models could be built?
Sales forecasting and customer segmentation models.

98 How could customer segmentation improve results?
By allowing personalized marketing and product recommendations.

99 What new features could enhance the dataset?
Adding data on customer loyalty, social media engagement, and store locations.

100 What are potential areas for further research?
Investigating long-term customer retention and lifetime value.

Conclusion

The analysis provided significant insights into shopping trends, revealing how seasonality, demographics, and promotional activities influence sales patterns. Seasonal fluctuations emerged as a key factor, with certain periods consistently driving higher sales volumes. Demographic factors such as age, gender, and income levels were also found to have a noticeable impact on purchasing behavior, providing valuable information for tailoring marketing strategies. Promotions and discounts played a critical role in boosting sales, with evidence suggesting that strategic timing and targeted offers can significantly enhance customer engagement and revenue generation.

These findings have practical implications for businesses aiming to optimize their operations and marketing efforts. By understanding seasonal trends, companies can align inventory levels to meet anticipated demand, avoiding overstocking or stockouts during critical periods. Insights into demographic preferences allow for more precise targeting, enabling personalized campaigns that resonate with specific customer segments. Additionally, the analysis highlights the importance of well-planned promotional activities, which can be leveraged to drive sales during low-demand periods or introduce new products effectively.Looking ahead, further research could focus on building predictive models to forecast future sales based on historical data, offering a proactive approach to inventory and sales planning. Clustering techniques could also be applied to segment customers into distinct groups based on their shopping behaviors and preferences. This segmentation would enable businesses to deliver highly targeted promotions and develop tailored loyalty programs, ultimately enhancing customer retention and maximizing profitability.

Reference

Retail Sales Forecasting Analysis: On GitHub

https://github.com/sankalpk4u/Retail-Sales-Analysis-EDA

Kaggle Competitions and Datasets:

https://www.kaggle.com/competitions/store-sales-time-series-forecasting

Retail Sales Forecasting:

https://www.kaggle.com/datasets/tevecsystems/retail-sales-forecasting