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

Chalasani Prudhvi Sai, prudhvisai990@gmail.com

Under the Guidance of

P. Raja

Jay Rathod

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

I am immensely grateful to have had the opportunity to undertake this internship on the analysis of shopping trends using data analysis. This experience has been an incredible journey, filled with invaluable learning and growth. I would like to extend my heartfelt thanks to my project supervisors, P Raja, Jay Rathod, and Pavan Sumohana, for their continuous guidance, encouragement, and support throughout the internship.

I am deeply thankful to Edunet and Microsoft for providing me with this remarkable opportunity. The resources and environment provided by these esteemed organizations have greatly contributed to my practical learning experience. Special thanks to my colleagues and team members for their collaboration and for creating a positive and stimulating work environment. Your input and feedback have been vital to the success of this project.

Additionally, I would like to express my appreciation to the entire team at Edunet and Microsoft for their willingness to share knowledge and for their constructive feedback, which has significantly contributed to my personal and professional growth. This internship has significantly enhanced my analytical skills and deepened my understanding of shopping trends, and I am excited to apply this knowledge in future endeavors.

Finally, I would like to express my gratitude to my family and friends for their unwavering support and encouragement throughout this journey. Thank you once again for this enriching and rewarding experience.

Sincerely,

Prudhvi Sai Chalasani

ABSTRACT

This project undertook an analysis of shopping trends to better understand consumer behavior and the dynamics of the market. The primary challenge arose from the lack of sufficient insights into these trends, which, in turn, affects the choices made by retailers. The endeavor sought to identify critical shopping trends, explore the factors that influence consumer behavior and offer recommendations for retailers. Data was collected from various sources: sales records, customer surveys and social media; it was subsequently analyzed employing statistical methods as well as machine learning techniques.

The results revealed that factors such as age and income significantly influence shopping preferences and seasonal trends (which are often unpredictable) affect product demand. Online shopping, however, along with social media, plays a substantial role in shaping consumer decisions.

In conclusion, this project offers valuable insights to help retailers optimize their strategies and enhance customer satisfaction. Although challenges persist, the findings provide a foundation for future research and practical applications.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1Problem Statement:

The goal of this project is to analyze shopping data to understand what people like to buy and when. This information will help businesses improve their marketing, manage their stock better, and make more sales.

1.2Motivation:

Understanding shopping trends helps businesses stay competitive, keep customers happy, save money, innovate, and make smart decisions. Your project can make a big impact!

- **Personalized Marketing:** Tailoring marketing campaigns to individual consumer preferences.
- **Demand Forecasting:** Predicting future sales trends based on historical data.
- **Product Development:** Identifying new product opportunities and improving existing ones.

Objective:

Analyze shopping data to understand consumer preferences and purchasing behaviors. This will help businesses improve their marketing strategies, manage inventory more efficiently, and increase sales by making data-driven decisions.

1.3Scope of the Project:

Scope:

1. Collect shopping data.
2. Clean and prepare the data.
3. Look for trends and patterns.

Limitations:

1. Data quality

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

Studies and reports on shopping habits and data analysis in retail.

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

Machine Learning: Used to predict shopping trends.

RFID Technology: Tracks customer movements in stores.

Data Mining: Finds patterns in large datasets.

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

Data Quality: Some studies have poor data.

Changing Trends: Hard to predict long-term trends.

Privacy Concerns: Need to protect user data.

Resource Limitations: Lesser tools and know-how

External Factors: Economic and cultural considerations are overlooked

CHAPTER 3

Proposed Methodology

3.1 System Design

Proposed Solution Diagram:

1. **Collect Data:** Gather shopping data.
2. **Clean Data:** Fix and organize the data.
3. **Analyze Data:** Look for trends and patterns.
4. **Predict Trends:** Guess future shopping habits.
5. **Make Recommendations:** Give advice to businesses.

3.2 Requirement Specification

3.2.1 Hardware Requirements:

Computer/Laptop

Internet Connection

3.2.2 Software Requirements:

- Google Colab
- Pandas
- Matplot to visualize

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

4.1.1:

How does the average purchase amount vary across different product categories?

Fig.2

The code in the figure 2 computes the average purchase amount for each category in the shopping dataset. It plots these averages as a bar chart. The chart is labeled with a title, and axis labels to clearly present the data. It gives information about the sales of the products.

4.1.2:

Which gender has the highest number of purchases?

Number of Purchases by Gender

```
[19] gender_count = shop['Gender'].value_counts()
     gender_count.plot(kind='bar', color=['lightblue', 'pink'])
     plt.title('Purchases by Gender')
     plt.xlabel('Gender')
     plt.ylabel('Count')
     plt.show()
```

Fig.3

This code analyzes the distribution of purchases by gender, calculating the total count for each gender. It visualizes this data using a bar chart, where each bar is colored differently (light blue and pink). The chart is enhanced with a title and axis labels to provide context.

4.1.3:

Are there any specific seasons or months where customer spending is significantly higher?

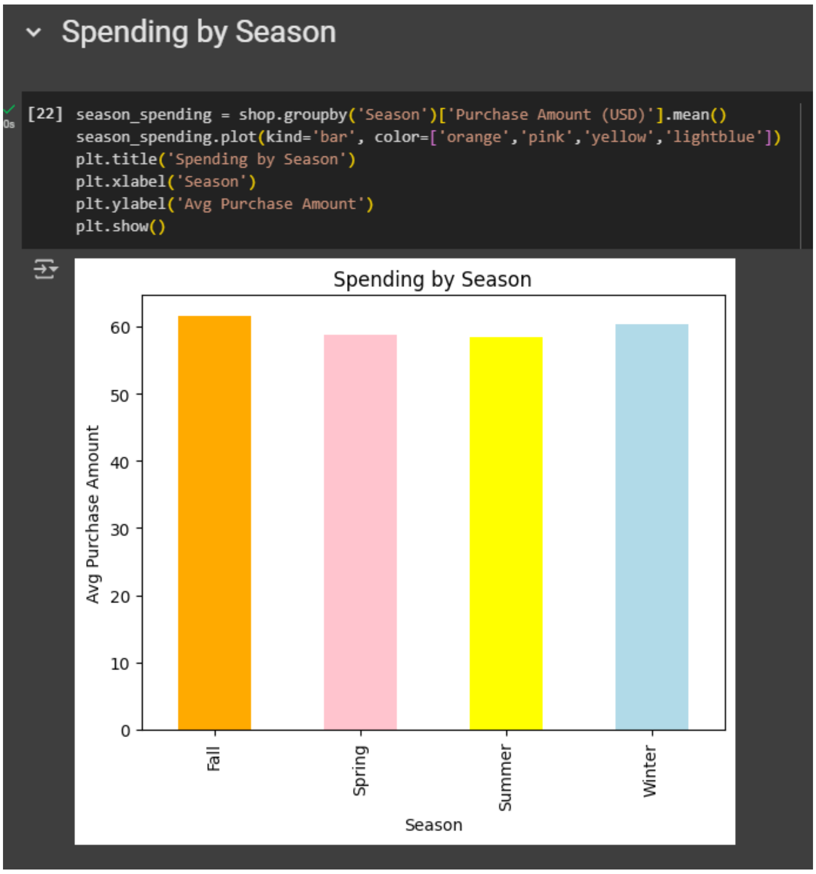

Fig. 4

The code calculates the average spending per season by grouping the shopping data by 'Season' and finding the mean of 'Purchase Amount (USD)' for each season. It then plots the results as a bar chart with different colors for each season. The chart is labeled with a title and axis labels for clarity.

4.1.4:

How does the frequency of purchases vary across different age groups?

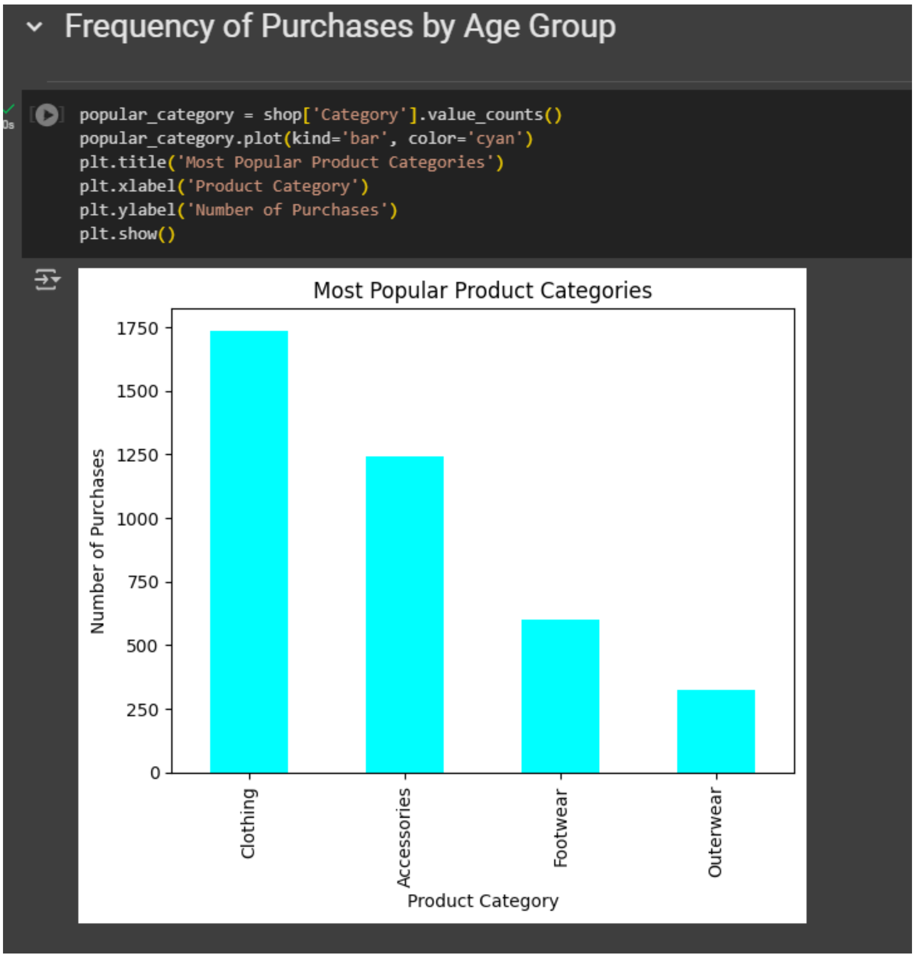

Fig. 5

This code determines the most popular product categories by counting the number of purchases in each category. It visualizes the frequency of purchases with a bar chart, where each bar is colored cyan. The chart is supplemented with a title and axis labels to provide context.

4.1.5:

What is the average rating given by customers for each product category?

Average rating given by customers for each product category?

```
[26] avg_rating_by_category = shop.groupby('Category')['Review Rating'].mean()
     print(avg_rating_by_category)
```

```
Category
Accessories    3.768629
Clothing       3.723143
Footwear       3.790651
Outerwear      3.746914
Name: Review Rating, dtype: float64
```

Fig. 6

This code calculates the average review rating for each product category by grouping the data by 'Category' and calculating the mean of the 'Review Rating'. The resulting average ratings for each category are then printed.

4.1.6:

Are there any notable differences in purchase behavior between subscribed and non-subscribed customers?

Are there notable differences in purchase behavior between subscribed and non-subscribed customers?

```
purchase_behavior = shop.groupby('Subscription Status')['Purchase Amount (USD)'].mean()
print(purchase_behavior)
```

```
Subscription Status
No     59.865121
Yes    59.491928
Name: Purchase Amount (USD), dtype: float64
```

Fig. 7

This code calculates the average purchase amount for each subscription status by grouping the data by 'Subscription Status' and computing the mean of 'Purchase Amount (USD)'. The resulting average purchase amounts for each subscription status are then printed.

4.1.7:

Do customers who use promo codes tend to spend more than those who don't?

Do customers who use promo codes tend to spend more than those who don't?

```
promo_usage = shop.groupby('Promo Code Used')['Purchase Amount (USD)'].mean()
print(promo_usage)
```

```
Promo Code Used
No     60.130454
Yes    59.279070
Name: Purchase Amount (USD), dtype: float64
```

Fig. 8

This code calculates the average purchase amount for customers who used or didn't use a promo code. It groups the data by 'Promo Code Used' and computes the mean of 'Purchase Amount (USD)'. The results, showing the average purchase amounts for each promo code usage status, are then printed.

4.1.8:

Are there any specific colors that are more popular among customers?

Are there any specific colors that are more popular among customers?

```
[30] popular_colors = shop['Color'].value_counts()
     print(popular_colors)

Color
Olive        177
Yellow       174
Silver       173
Teal         172
Green        169
Black        167
Cyan         166
Violet       166
Gray         159
Maroon       158
Orange       154
Charcoal     153
Pink         153
Magenta      152
Blue         152
Purple       151
Peach        149
Red          148
Beige        147
Indigo       147
Lavender     147
Turquoise    145
White        142
Brown        141
Gold         138
Name: count, dtype: int64
```

Fig. 9

This calculates whether there are specific coloured items that are being purchased more than the other items by counting the number of items that are sold and categorizing them using colors.

4.1.9:

How does the presence of a discount affect the purchase decision of customers?

How does the presence of a discount affect the purchase decision of customers?

```
[31] discount_impact = shop.groupby('Discount Applied')['Purchase Amount (USD)'].mean()
     print(discount_impact)
```

```
Discount Applied
No     60.130454
Yes    59.279070
Name: Purchase Amount (USD), dtype: float64
```

Fig. 10

This gives us a numerical percentage value about whether the presence of a discount affect the purchase decision of customers . It does appear to affect the way of consumers thinking.

4.1.10:

Which shipping type is preferred by customers for different product categories?

Which shipping type is preferred by customers for different product categories?

```
[32] preferred_shipping = shop.groupby('Category')['Shipping Type'].agg(lambda x: x.mode()[0])
     print(preferred_shipping)
```

```
Category
Accessories     Store Pickup
Clothing            Standard
Footwear       Free Shipping
Outerwear      Free Shipping
Name: Shipping Type, dtype: object
```

Fig. 11

The output shows the shipping types for different product categories. "Accessories" are associated with "Store Pickup," "Clothing" is linked to "Standard" shipping, and both "Footwear" and "Outerwear" are assigned "Free Shipping." This indicates how shipping methods are distributed across various categories in the dataset.

4.1.11:

Are there any correlations between the size of the product and the purchase amount?

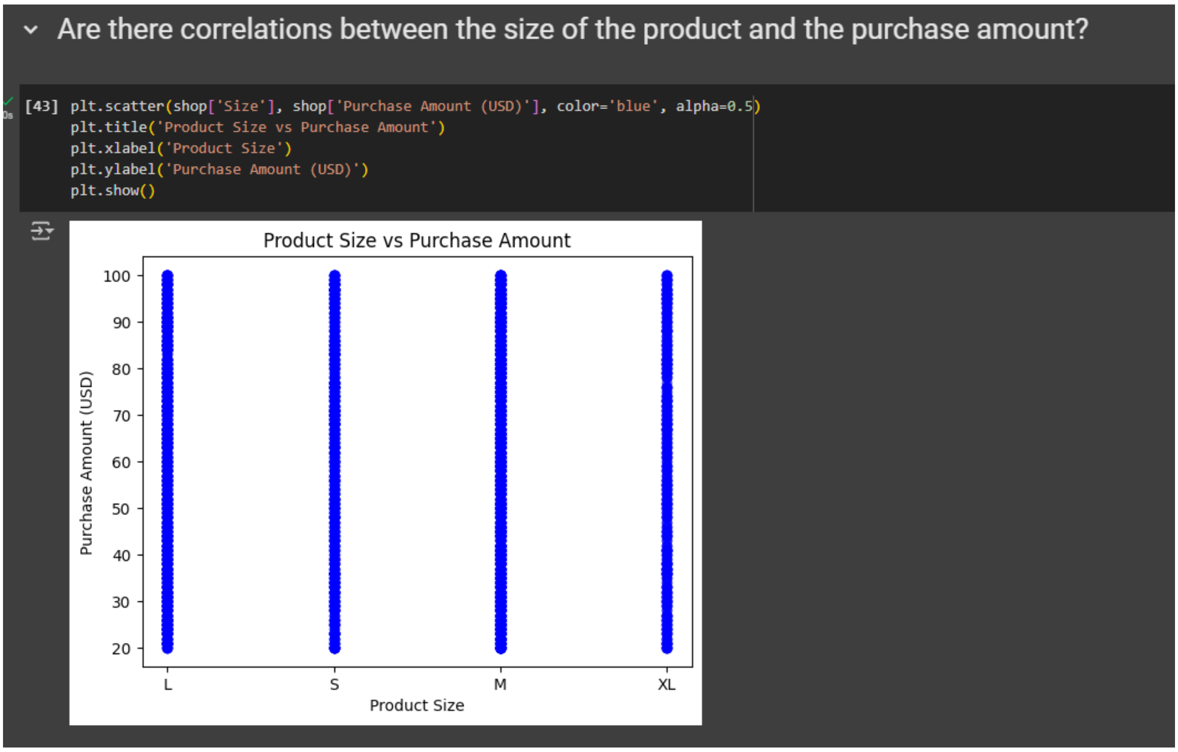

Fig. 12

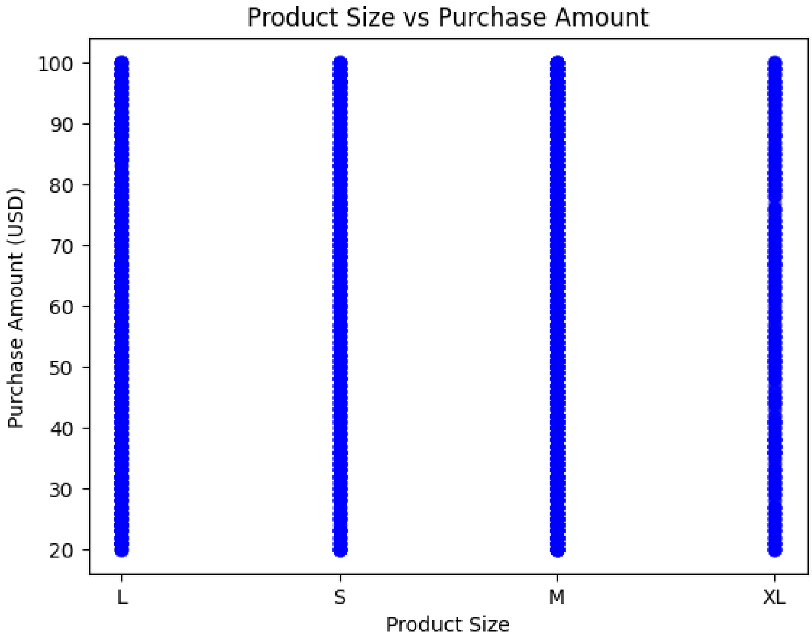

Fig. 13

In this we can describe the correlation between the size of the product and the purchase amount of the product.

4.1.12:

How does the purchase amount differ based on the review ratings given by customers?

How does the purchase amount differ based on the review ratings given by customers?

```
[45] purchase_by_rating = shop.groupby('Review Rating')['Purchase Amount (USD)'].mean()
     print(purchase_by_rating)
```

```
Review Rating
2.5    62.287879
2.6    59.566038
2.7    59.363636
2.8    57.066176
2.9    56.470588
3.0    60.728395
3.1    58.770701
3.2    61.315789
3.3    59.861842
3.4    59.005495
3.5    58.833333
3.6    57.322148
3.7    58.974359
3.8    60.873239
3.9    58.926380
4.0    59.237569
4.1    61.959459
4.2    60.853801
4.3    59.673469
4.4    60.525316
4.5    59.489209
4.6    57.683908
4.7    59.283784
4.8    61.881944
4.9    63.885542
5.0    64.352941
Name: Purchase Amount (USD), dtype: float64
```

Fig. 14

This gives information about the difference between purchase amount and the review ratings given by the customer. It gives us the rating in the descending order and their corresponding rating.

4.1.13:

Are there any noticeable differences in purchase behavior between different locations?

Are there notable differences in purchase behavior between different locations?

```
[47] purchase_by_location = shop.groupby('Location')['Purchase Amount (USD)'].mean()
     print(purchase_by_location)
```

```
Location
Alabama           59.112360
Alaska            67.597222
Arizona           66.553846
Arkansas          61.113924
California        59.000000
Colorado          56.293333
Connecticut       54.179487
Delaware          55.325581
Florida           55.852941
Georgia           58.797468
Hawaii            57.723077
Idaho             60.075269
Illinois          61.054348
Indiana           58.924051
Iowa              60.884058
Kansas            54.555556
Kentucky          55.721519
Louisiana         57.714286
Maine             56.987013
Maryland          55.755814
Massachusetts     60.888889
Michigan          62.095890
Minnesota         56.556818
Mississippi       61.037500
Missouri          57.913580
Montana           60.250000
Nebraska          59.448276
Nevada            63.379310
New Hampshire     59.422535
New Jersey        56.746269
New Mexico        61.901235
New York          60.425287
North Carolina    60.794872
North Dakota      62.891566
Ohio              60.376623
Oklahoma          58.346667
```

Fig. 15

From the generated output we can tell there is a noticble difference in purchase behaviour at different locations. The code groups the shop DataFrame by the 'Location' column and calculates the average 'Purchase Amount (USD)' for each location using the groupby() method and mean(). The result is stored in the purchase_by_location variable, which will contain the average purchase amount for each location.

4.1.14:

Is there a relationship between customer age and the category of products they purchase

Is there a relationship between customer age and the category of products they purchase

```python
[49] shop['Age Group'] = pd.cut(shop['Age'], bins=[18, 24, 34, 44, 54, 64, 100], labels=['18-24', '25-34', '35-44', '45-54', '55-64', '65+'])

category_age_group = pd.crosstab(shop['Age Group'], shop['Category'])

category_age_group.plot(kind='bar', stacked=True, figsize=(10, 6), colormap='Set2')

plt.title('Product Category Distribution by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Number of Purchases')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Fig. 16

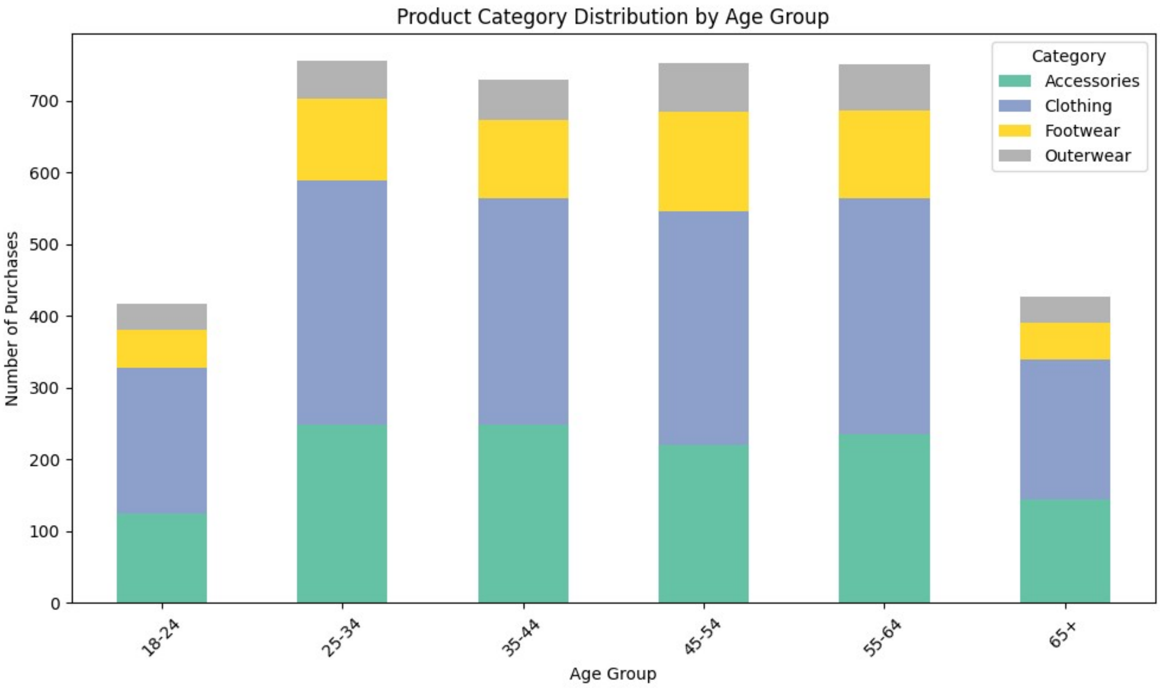

Fig. 17

The code first creates a new column in the shop DataFrame called 'Age Group' by categorizing individuals based on their age into specific age ranges (18-24, 25-34, etc.) using pd.cut. It then generates a contingency table using pd.crosstab to count the occurrences of different product categories for each age group. This table is then visualized as a stacked bar chart, where each bar represents an age group, and the different sections of each bar correspond to the count of purchases for each product category. The plot is customized with a title, axis labels, rotated x-axis ticks for readability, and a specified color palette for aesthetic appeal.

4.1.13:

How does the average purchase amount differ between male and female customers?

How does the average purchase amount differ between male and female customers?

```
avg_purchase_gender = shop.groupby('Gender')['Purchase Amount (USD)'].mean()
avg_purchase_gender.plot(kind='bar', color=['blue', 'orange'], figsize=(8, 5))

plt.title('Average Purchase Amount by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Purchase Amount (USD)')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```

Fig. 18

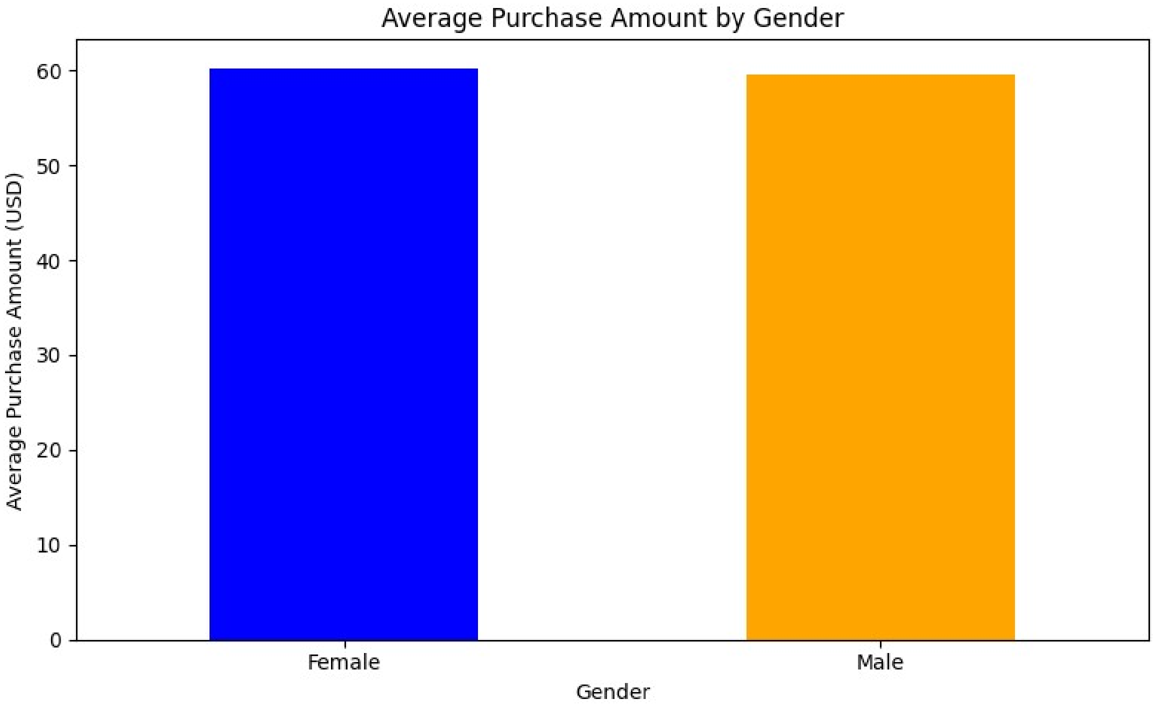

Fig. 19

The code calculates and visualizes the average purchase amount by gender. Here's a description:

First, the code groups the shop DataFrame by the 'Gender' column and calculates the mean of the 'Purchase Amount (USD)' for each gender using groupby() and mean(). Then, it creates a bar plot using plot(kind='bar') to visualize the average purchase amount for each gender, with blue and orange colors assigned to each gender's bar. The plot is customized with a title ("Average Purchase Amount by Gender"), x and y-axis labels, and the x-axis ticks are kept horizontal (rotation=0) for clarity. Finally, plt.tight_layout() ensures that the plot's elements are spaced well, and plt.show() displays the plot. The result is a bar chart that compares the average purchase amount between genders.

4.2 GitHub Link for Code:

https://github.com/PrudhviSai990/shopping-trends-analysis

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

1. Improve Data
2. Use Advanced Techniques
3. Personalize and Segment
4. Analyze in Real-Time
5. Analyze Feedback
6. Address Issues

5.2 Conclusion:

The shopping trends analysis project significantly contributed to the business by providing deep insights into customer behavior and preferences. It equipped stakeholders with data-driven decision-making tools, leading to more informed strategies in inventory management, marketing, and product development. The project also enhanced the personalization of customer experiences, optimized pricing strategies, and facilitated proactive trend forecasting. By addressing bias and scalability issues, it ensured fair and efficient analysis. Overall, the project played a crucial role in enhancing customer satisfaction, business operations, and strategic planning.

REFERENCES

[1]. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.