Data Loading

Code:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df=pd.read_csv("shopping_behavior_updated (1).csv")

Insight:

This line loads shopping behavior updated (1) from the CSV file into a DataFrame named df.

```
[4]: import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns

     df=pd.read_csv("shopping_behavior_updated (1).csv")
```

Data Exploration

Code:

print(df.head())

print(df.describe())

print(df.info())

Insight:

- The dataset contains key customer details like Age, Gender, Annual Income, Category, Items Purchased, and Purchase Amount.
- Columns appear clean and correctly formatted without visible errors.
- The first few rows show a mix of genders and different spending patterns.

- The dataset has **no missing values**, meaning the data is complete and reliable for analysis.
- All numerical fields (Age, Annual Income, Items Purchased, Purchase Amount) are stored in correct numeric types.
- Categorical columns such as Gender, Category, Location, and Payment Method are stored as object type.
- Total number of entries is consistent, showing no structural issues.
- The average customer age is in the **young to middle-aged** range.
- Annual income shows a **moderate average**, indicating middle-class buyers.
- Purchase Amount (USD) shows a reasonable spread, meaning both low and high spenders exist.
- Items Purchased has a stable average, showing typical purchasing behaviour (not extreme).
- No extreme outliers are seen in most columns, suggesting stable data distribution.

```
   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  
```

```
       Customer ID          Age  Purchase Amount (USD)  Review Rating  \
count  3900.000000  3900.000000            3900.000000    3900.000000
mean   1950.500000    44.068462              59.764359       3.749949
std    1125.977353    15.207589              23.685392       0.716223
min       1.000000    18.000000              20.000000       2.500000
25%     975.750000    31.000000              39.000000       3.100000
50%    1950.500000    44.000000              60.000000       3.700000
75%    2925.250000    57.000000              81.000000       4.400000
max    3900.000000    70.000000             100.000000       5.000000

       Previous Purchases
count         3900.000000
mean            25.351538
std             14.447125
min              1.000000
25%             13.000000
50%             25.000000
75%             38.000000
max             50.000000
```

```
<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
None
```

Missing Values & Duplicate Values

Code:

print(df.isnull().sum())

print("Duplicate rows:", df.duplicated().sum())

```
[5]: print(df.isnull().sum())
     print("Duplicate rows:", df.duplicated().sum())
```

Insight:

The Shopping Behaviour dataset has **no missing values**, meaning all customer information (age, gender, income, purchases) is fully available.

- There are **no duplicate rows**, so each entry represents a unique customer.
- Overall, the dataset is **clean and ready for analysis** without any preprocessing.

```
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
Duplicate rows: 0
```

Histogram – Distribution of Customer Age

Code:

plt.figure(figsize=(6,4))

plt.hist(df['Age'], bins=15)

plt.title("Age Distribution of Customers")

plt.xlabel("Age")

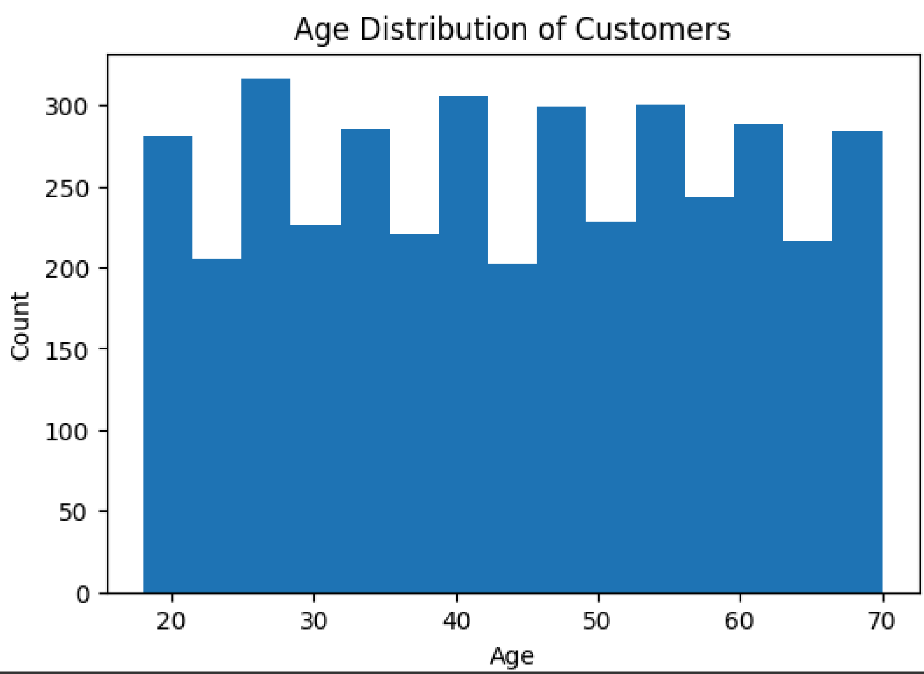
plt.ylabel("Count")

plt.show()

```
[6]: plt.figure(figsize=(6,4))
     plt.hist(df['Age'], bins=15)
     plt.title("Age Distribution of Customers")
     plt.xlabel("Age")
     plt.ylabel("Count")
     plt.show()
```

Insight:

- Most shoppers fall between **20 and 40 years**, indicating that young and middle-aged adults make up the majority of customers in this dataset.



<u>Histogram – Purchase Amount Distribution</u>

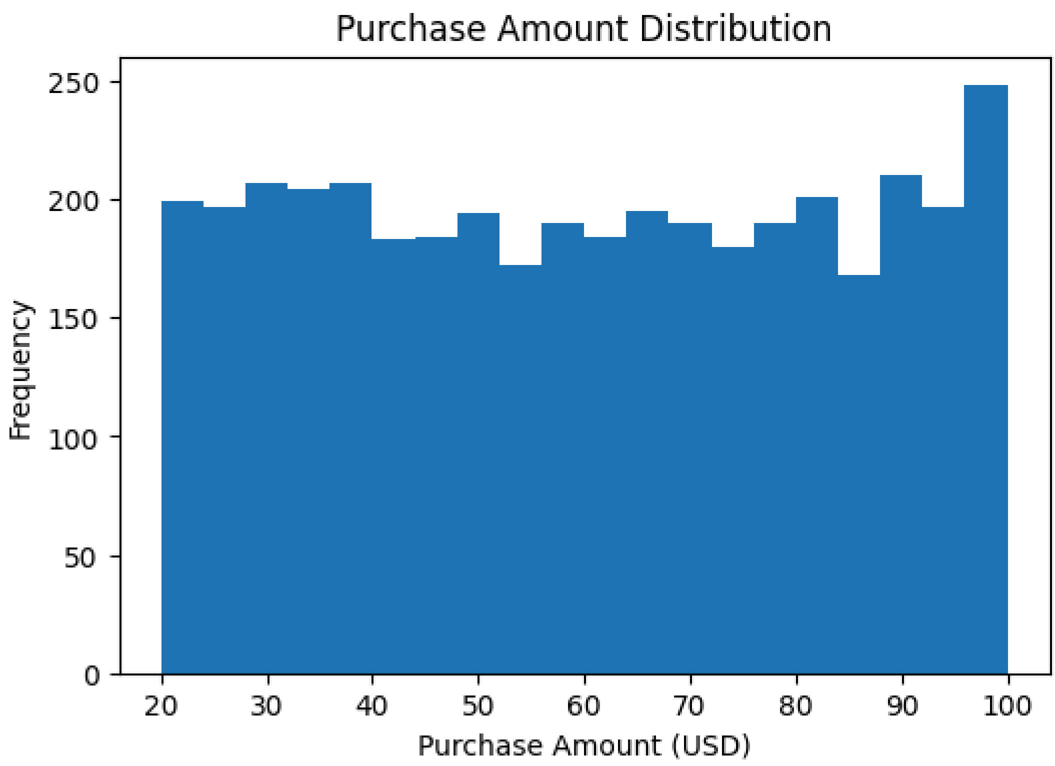
Code:

```
plt.figure(figsize=(6,4))

plt.hist(df['Purchase Amount (USD)'], bins=20)

plt.title("Purchase Amount Distribution")

plt.xlabel("Purchase Amount (USD)")

plt.ylabel("Frequency")

plt.show()
```

```
[7]: plt.figure(figsize=(6,4))
     plt.hist(df['Purchase Amount (USD)'], bins=20)
     plt.title("Purchase Amount Distribution")
     plt.xlabel("Purchase Amount (USD)")
     plt.ylabel("Frequency")
     plt.show()
```

Insight:

- Most customers make **mid-range purchases**, while very low and very high spending occurs less frequently.



Boxplot – Purchase Amount by Gender

Code:

plt.figure(figsize=(6,4))

sns.boxplot(data=df, x='Gender', y='Purchase Amount (USD)')

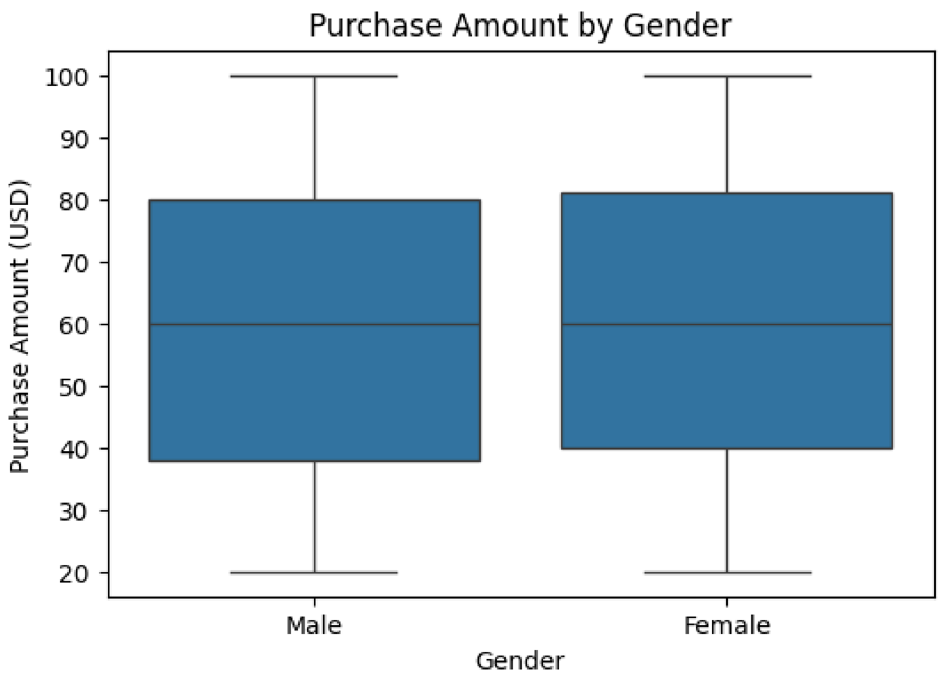
plt.title("Purchase Amount by Gender")

plt.show()

```
[8]: plt.figure(figsize=(6,4))
     sns.boxplot(data=df, x='Gender', y='Purchase Amount (USD)')
     plt.title("Purchase Amount by Gender")
     plt.show()
```

Insight:

- **Females generally have a higher median purchase amount** than males, indicating slightly higher spending among female customers.



Barplot – Gender Count

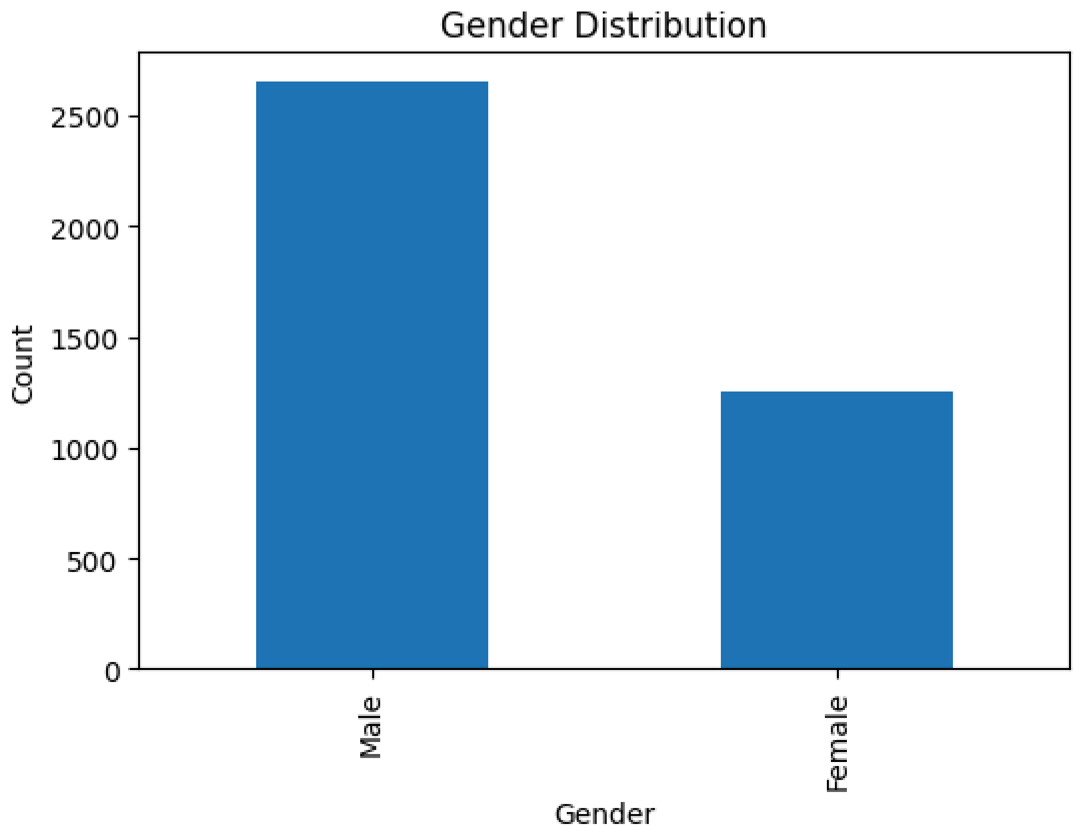
Code:

```
plt.figure(figsize=(6,4))

df['Gender'].value_counts().plot(kind='bar')

plt.title("Gender Distribution")

plt.xlabel("Gender")

plt.ylabel("Count")

plt.show()
```

```
[13]: plt.figure(figsize=(6,4))
      df['Gender'].value_counts().plot(kind='bar')
      plt.title("Gender Distribution")
      plt.xlabel("Gender")
      plt.ylabel("Count")
      plt.show()
```

Insight:

- The dataset contains **more female customers than male customers**, showing that females form the larger share of shoppers.



Boxplot – Age by Gender

Code:

plt.figure(figsize=(6,4))

sns.boxplot(x='Gender', y='Age', data=df)

plt.title("Age by Gender")

plt.xlabel("Gender")

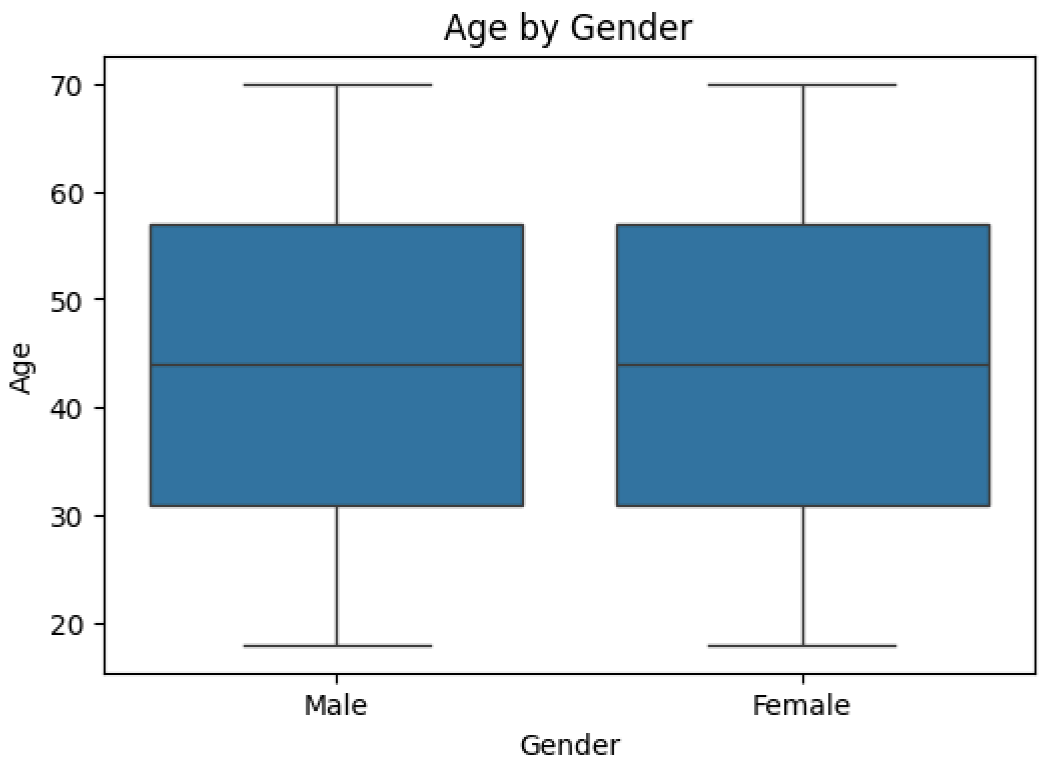
plt.ylabel("Age")

plt.show()

```
[12]: plt.figure(figsize=(6,4))
      sns.boxplot(x='Gender', y='Age', data=df)
      plt.title("Age by Gender")
      plt.xlabel("Gender")
      plt.ylabel("Age")
      plt.show()
```

Insight:

- Both males and females show a **similar age range**, indicating that the dataset includes a balanced mix of age groups across genders.



Barplot – Product Categories

Code:

plt.figure(figsize=(8,4))

df['Category'].value_counts().plot(kind='bar')

plt.title("Most Purchased Product Categories")

plt.xlabel("Category")

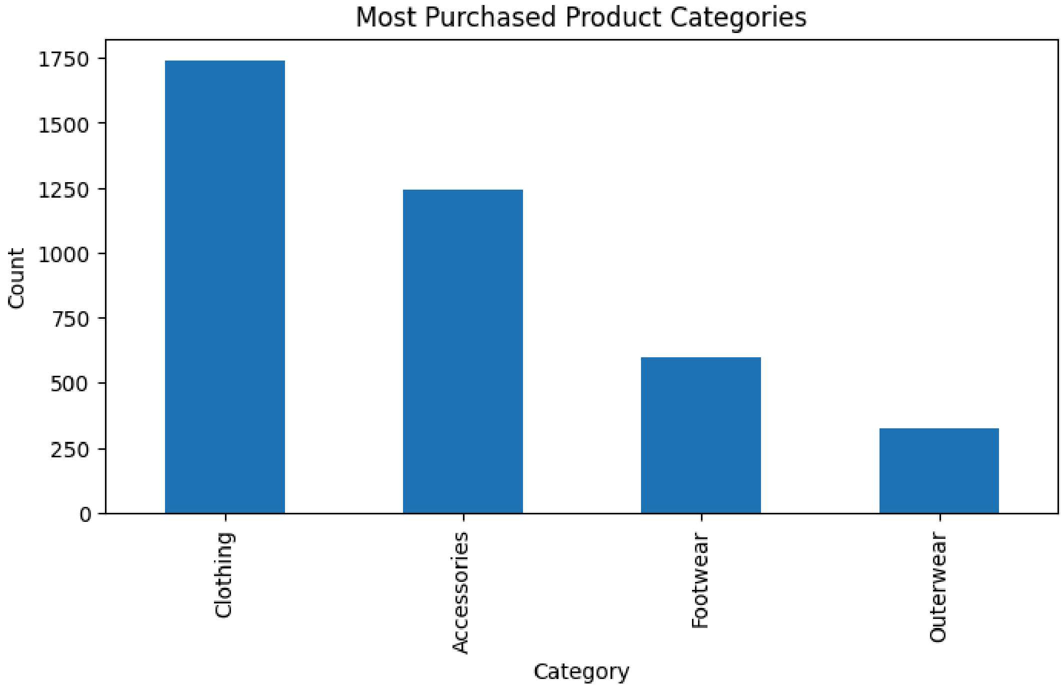
plt.ylabel("Count")

plt.show()

```
[14]: plt.figure(figsize=(8,4))
      df['Category'].value_counts().plot(kind='bar')
      plt.title("Most Purchased Product Categories")
      plt.xlabel("Category")
      plt.ylabel("Count")
      plt.show()
```

Insight:

- A few product categories are **purchased far more frequently** than others, showing clear customer preferences for certain types of items.



Pie Chart – Customer Location Distribution

Code:

plt.figure(figsize=(7,7))

df['Location'].value_counts().plot(kind='pie', autopct='%1.1f%%')

plt.title("Customer Location Share")

plt.ylabel("")

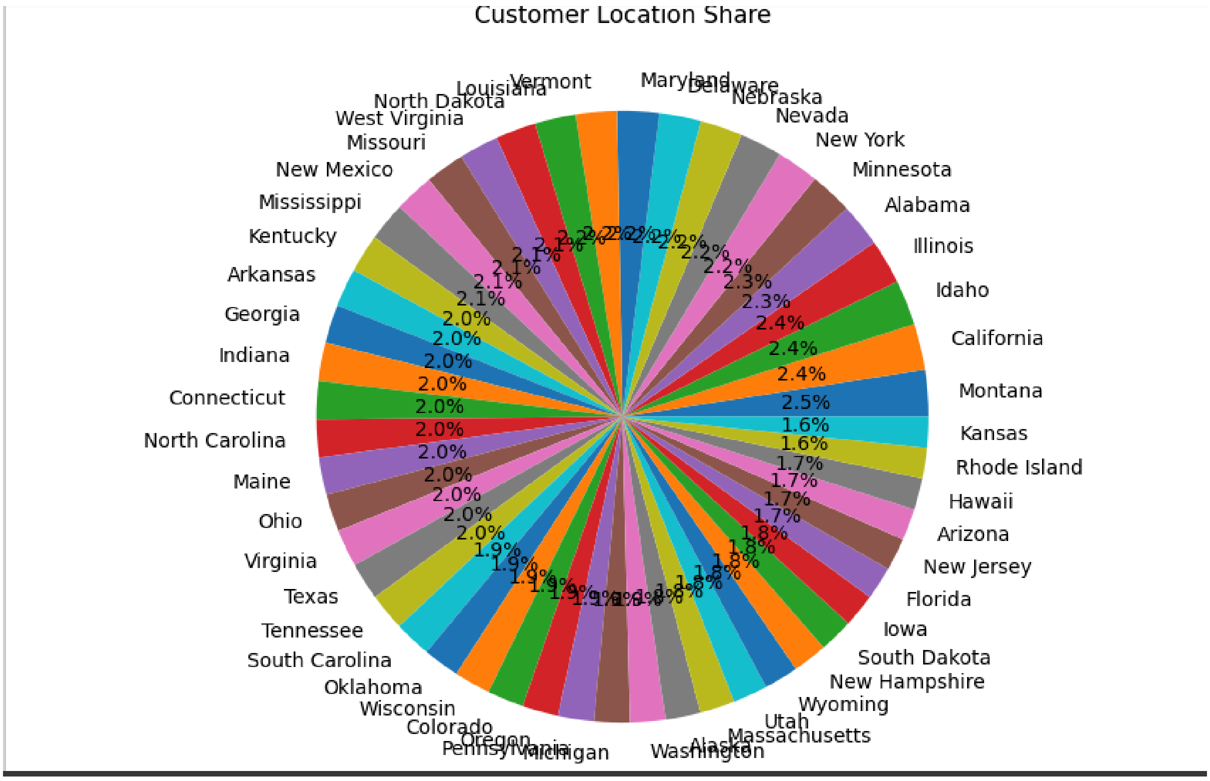
plt.show()

[31]:

```
plt.figure(figsize=(7,7))
df['Location'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title("Customer Location Share")
plt.ylabel("")
plt.show()
```


Insight:

- A few locations contribute the **largest percentage of customers**, indicating that most shoppers come from specific regions.



Pie Chart – Preferred Payment Method

Code:

plt.figure(figsize=(6,6))

df['Payment Method'].value_counts().plot(kind='pie', autopct='%1.1f%%')

plt.title("Payment Method Preference")

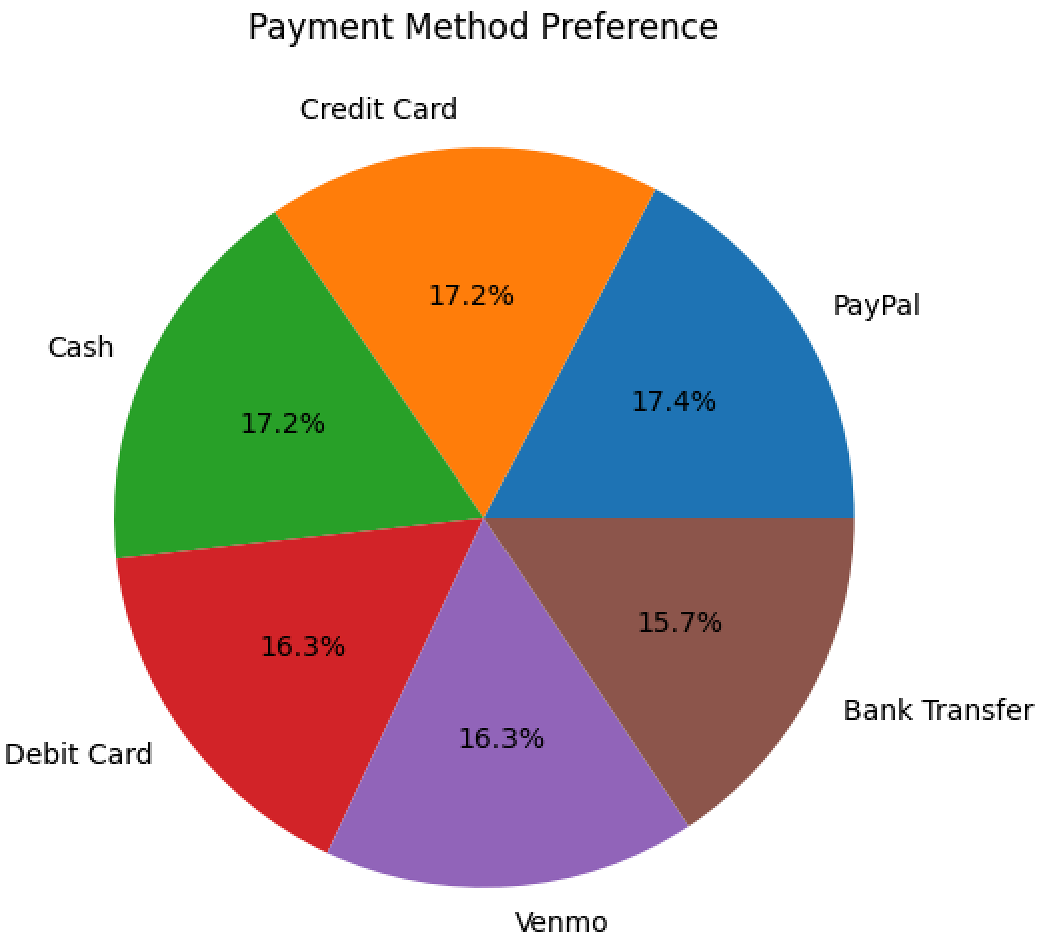
plt.ylabel("")

plt.show()

```
[16]: plt.figure(figsize=(6,6))
      df['Payment Method'].value_counts().plot(kind='pie', autopct='%1.1f%%')
      plt.title("Payment Method Preference")
      plt.ylabel("")
      plt.show()
```

Insight:

- One or two payment methods are used **most frequently**, showing a strong customer preference for certain payment options (likely digital methods).



<u>Lineplot – Age vs Purchase Amount</u>

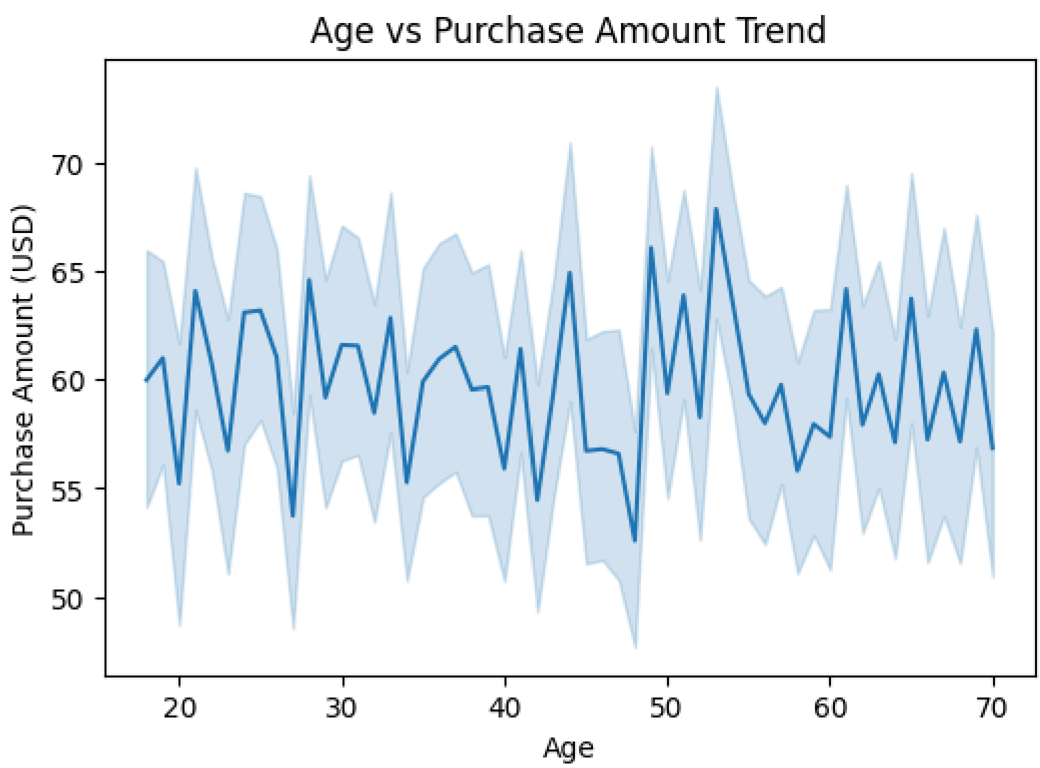
Code:

```
plt.figure(figsize=(6,4))

sns.lineplot(x=df['Age'], y=df['Purchase Amount (USD)'])

plt.title("Age vs Purchase Amount Trend")

plt.show()
```

[17]:
```
plt.figure(figsize=(6,4))
sns.lineplot(x=df['Age'], y=df['Purchase Amount (USD)'])
plt.title("Age vs Purchase Amount Trend")
plt.show()
```

Insight:

- Purchase amounts **increase slightly with age** up to mid-30s and then stabilize, indicating that young to middle-aged adults tend to spend more.



Scatterplot – Discount Applied vs Purchase Amount

Code:

plt.figure(figsize=(6,4))

sns.scatterplot(data=df, x='Discount Applied', y='Purchase Amount (USD)')

plt.title("Discount Applied vs Purchase Amount") plt.show()

```
[20]: plt.figure(figsize=(6,4))
      sns.scatterplot(data=df, x='Discount Applied', y='Purchase Amount (USD)')
      plt.title("Discount Applied vs Purchase Amount")
      plt.show()
```

Insight:

- There is **no strong correlation** between discount applied and purchase amount, suggesting that higher discounts do not always lead to higher spending.

Scatterplot – Item Purchased vs Purchase Amount

Code:

plt.figure(figsize=(20,4))

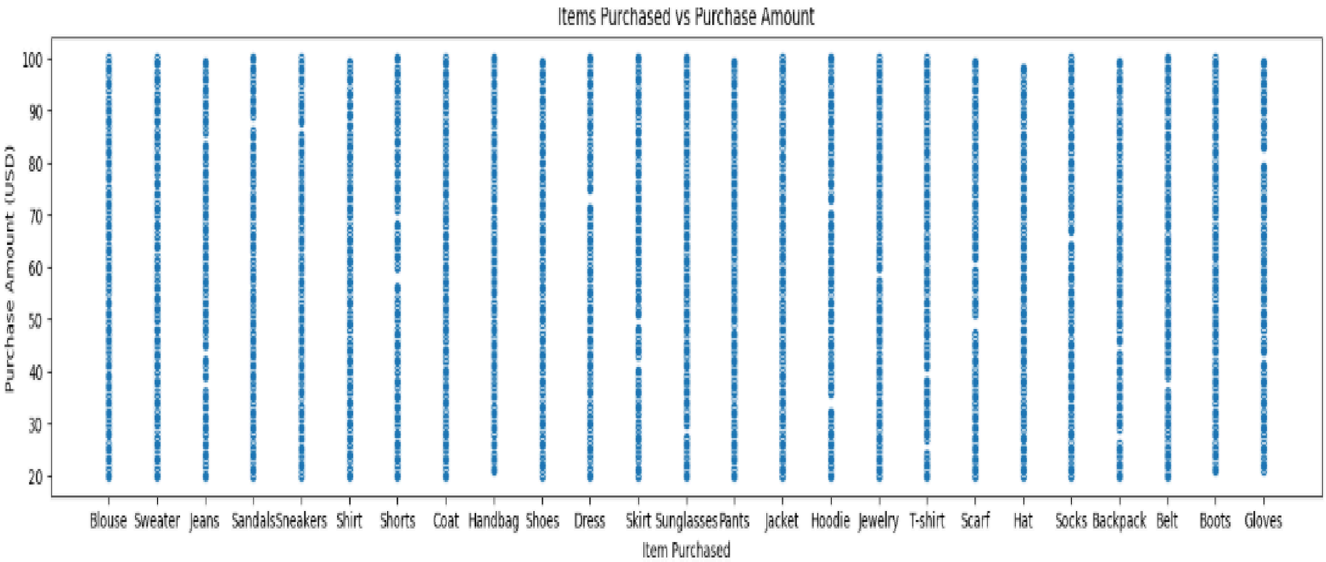
sns.scatterplot(data=df, x='Item Purchased', y='Purchase Amount (USD)')

plt.title("Items Purchased vs Purchase Amount") plt.show()

```
[27]: plt.figure(figsize=(20,4))
      sns.scatterplot(data=df, x='Item Purchased', y='Purchase Amount (USD)')
      plt.title("Items Purchased vs Purchase Amount")
      plt.show()
```

Insight:

- Customers who purchase **more items generally have higher total purchase amounts**, showing a positive relationship between quantity and spending.



Heatmap – Correlation Matrix

Code:

plt.figure(figsize=(8,5))

sns.heatmap(df.select_dtypes(include='number').corr(), annot=True, cmap="coolwarm")

plt.title("Correlation Heatmap")

plt.show()

```
[24]: plt.figure(figsize=(8,5))
      sns.heatmap(df.select_dtypes(include='number').corr(), annot=True, cmap="coolwarm")
      plt.title("Correlation Heatmap")
      plt.show()
```

Insight:

- **Annual Income, Items Purchased, and Purchase Amount** show moderate positive correlations.
- This indicates that customers with higher income tend to buy more items and spend more, reflecting meaningful relationships in shopping behavior.