Customer Purchase Behavior Analysis on Walmart's Black Friday Sales

The objective of this case study is to analyze the customer purchase behavior at Walmart during the Black Friday sales by investigating the relationships between customer attributes (such as gender, marital status, and age) and their purchase amounts.

In [27]:

```
!gdown "1G2SaxOe3Ypm7t8LTZR8INMhnTFHJOYVK"
```

```
Downloading...
From: https://drive.google.com/uc?id=1G2SaxOe3Ypm7t8LTZR8INMhnTFHJOYVK
To: /content/walmart_data.csv
100% 23.0M/23.0M [00:00<00:00, 199MB/s]
```

In [28]:

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
```

In [29]:

```
df = pd.read_csv("walmart_data.csv")
```

In [30]:

```
df.head()
```

Out[30]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Catego
0	1000001	P00069042	F	0-17	10	A	2	0	
1	1000001	P00248942	F	0-17	10	A	2	0	
2	1000001	P00087842	F	0-17	10	A	2	0	1
3	1000001	P00085442	F	0-17	10	A	2	0	1
4	1000002	P00285442	M	55+	16	C	4+	0	

In [31]:

```
print("Shape of dataset:", df.shape)
```

```
Shape of dataset: (550068, 10)
```

In [32]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     550068 non-null  int64
 1   Product_ID                  550068 non-null  object
```

```
 2   Gender                      550068 non-null  object
 3   Age                         550068 non-null  object
 4   Occupation                  550068 non-null  int64
 5   City_Category               550068 non-null  object
 6   Stay_In_Current_City_Years  550068 non-null  object
 7   Marital_Status              550068 non-null  int64
 8   Product_Category            550068 non-null  int64
 9   Purchase                    550068 non-null  int64
dtypes: int64(5), object(5)
memory usage: 42.0+ MB
```

In [33]:

```
df.describe(include='all')
```

Out[33]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status
count	5.500680e+05	550068	550068	550068	550068.000000	550068	550068	550068.000000
unique	NaN	3631	2	7	NaN	3	5	NaN
top	NaN	P00265242	M	26-35	NaN	B	1	NaN
freq	NaN	1880	414259	219587	NaN	231173	193821	NaN
mean	1.003029e+06	NaN	NaN	NaN	8.076707	NaN	NaN	0.409653
std	1.727592e+03	NaN	NaN	NaN	6.522660	NaN	NaN	0.491770
min	1.000001e+06	NaN	NaN	NaN	0.000000	NaN	NaN	0.000000
25%	1.001516e+06	NaN	NaN	NaN	2.000000	NaN	NaN	0.000000
50%	1.003077e+06	NaN	NaN	NaN	7.000000	NaN	NaN	0.000000
75%	1.004478e+06	NaN	NaN	NaN	14.000000	NaN	NaN	1.000000
max	1.006040e+06	NaN	NaN	NaN	20.000000	NaN	NaN	1.000000

In [34]:

```
category_columns = ['Gender', 'Age', 'City_Category', 'Stay_In_Current_City_Years', 'Marital_Status']
for col in category_columns:
    df[col] = df[col].astype('category')
```

In [35]:

```
df.dtypes
```

Out[35]:

	0
User_ID	int64
Product_ID	object
Gender	category
Age	category
Occupation	int64
City_Category	category
Stay_In_Current_City_Years	category
Marital_Status	category
Product_Category	int64
Purchase	int64

dtype: object

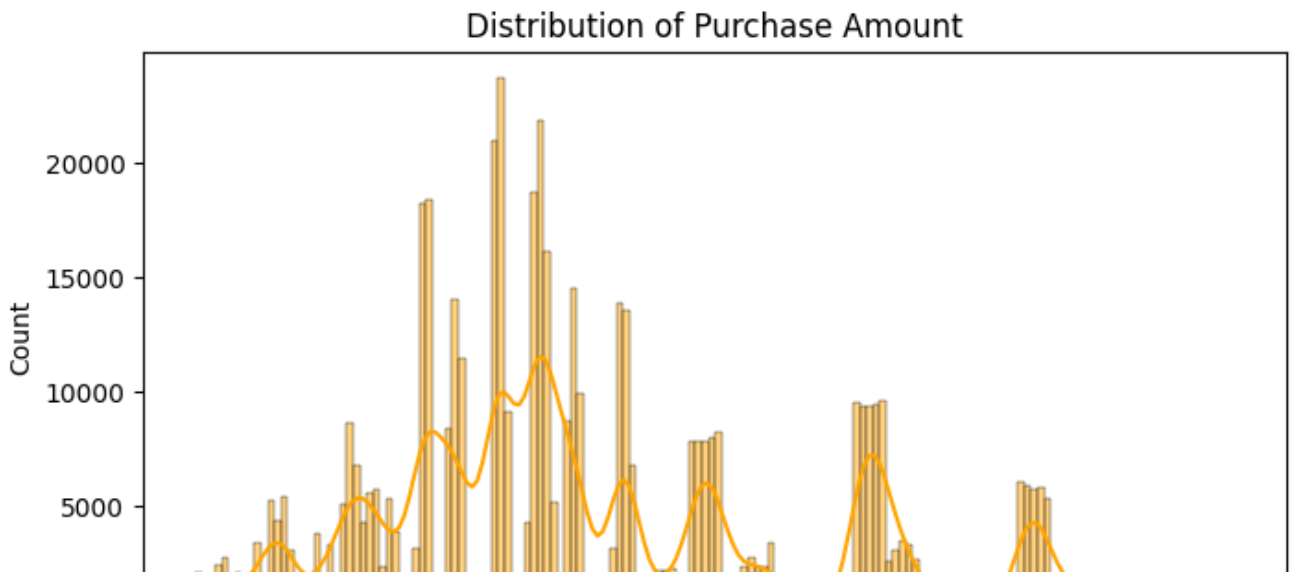
In [36]:

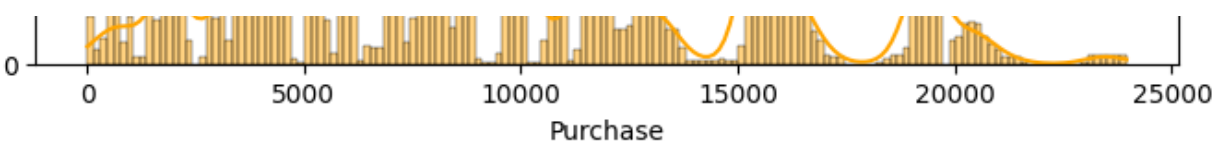
```
for col in category_columns:
    print(f"\nColumn: {col}")
    print(df[col].value_counts())
```

```
Column: Gender
Gender
M    414259
F    135809
Name: count, dtype: int64

Column: Age
Age
26-35    219587
36-45    110013
18-25     99660
46-50     45701
51-55     38501
55+       21504
0-17      15102
Name: count, dtype: int64

Column: City_Category
City_Category
B    231173
C    171175
A    147720
Name: count, dtype: int64

Column: Stay_In_Current_City_Years
Stay_In_Current_City_Years
1     193821
2     101838
3      95285
4+     84726
0      74398
Name: count, dtype: int64

Column: Marital_Status
Marital_Status
0    324731
1    225337
Name: count, dtype: int64
```

In [37]:

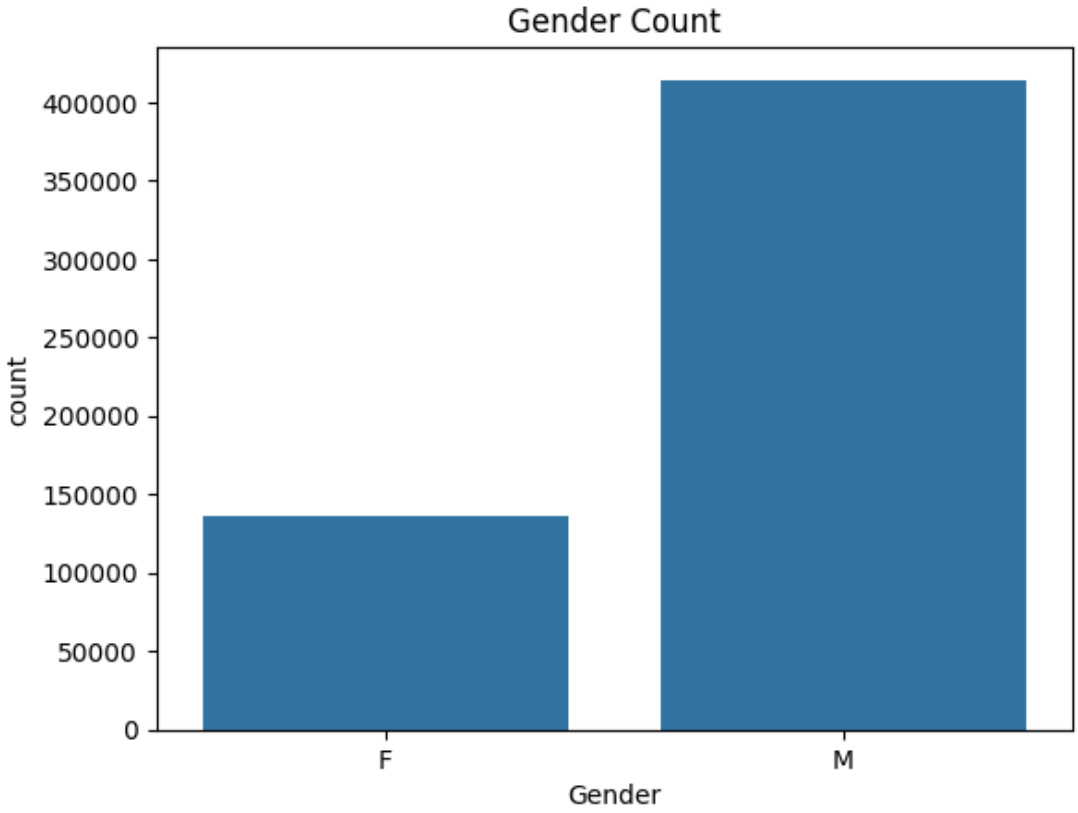
```
plt.figure(figsize=(8,4))
sns.histplot(df['Purchase'], kde=True, color='orange')
plt.title("Distribution of Purchase Amount")
plt.show()
```





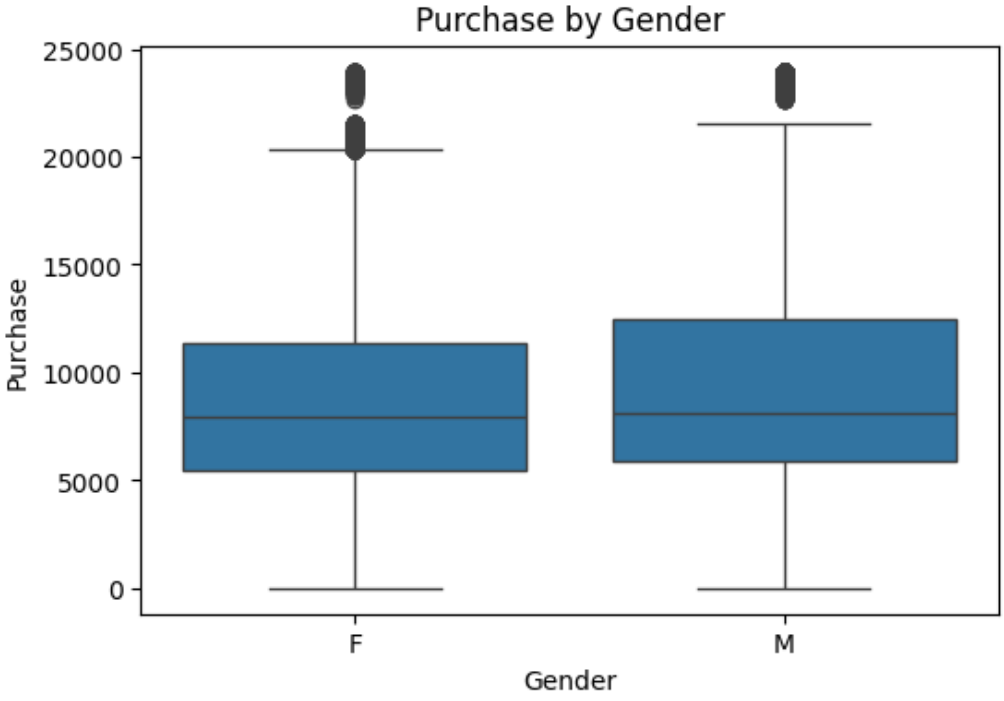
In [38]:

```
sns.countplot(x='Gender', data=df)
plt.title("Gender Count")
plt.show()
```



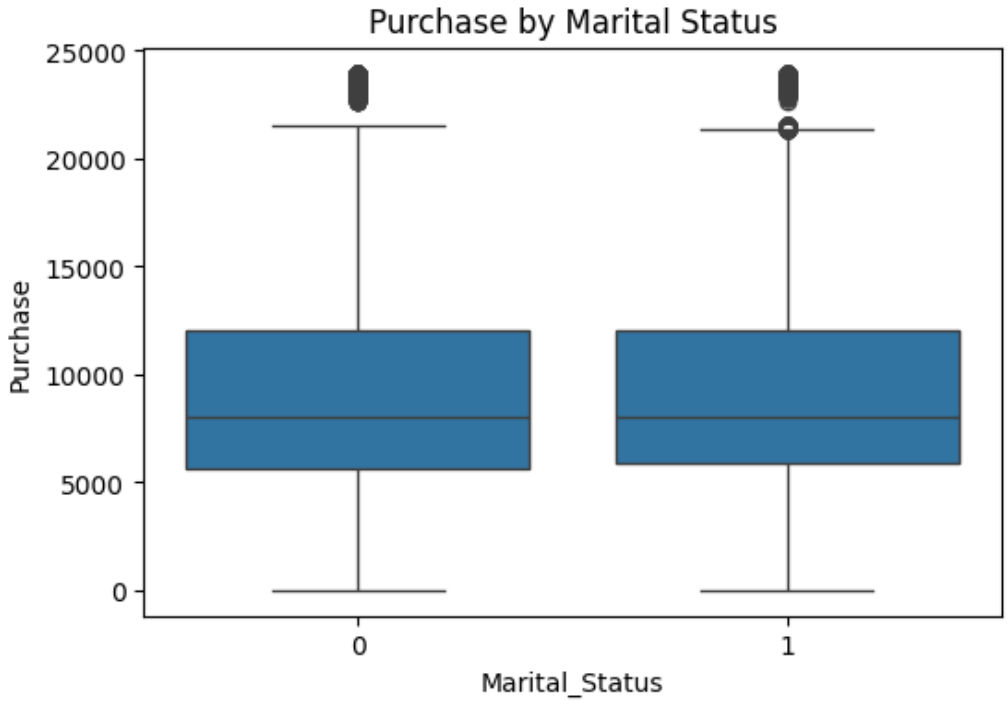
In [39]:

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



In [40]:

```
plt.figure(figsize=(6,4))
sns.boxplot(x='Marital_Status', y='Purchase', data=df)
plt.title("Purchase by Marital Status")
plt.show()
```



In [41]:

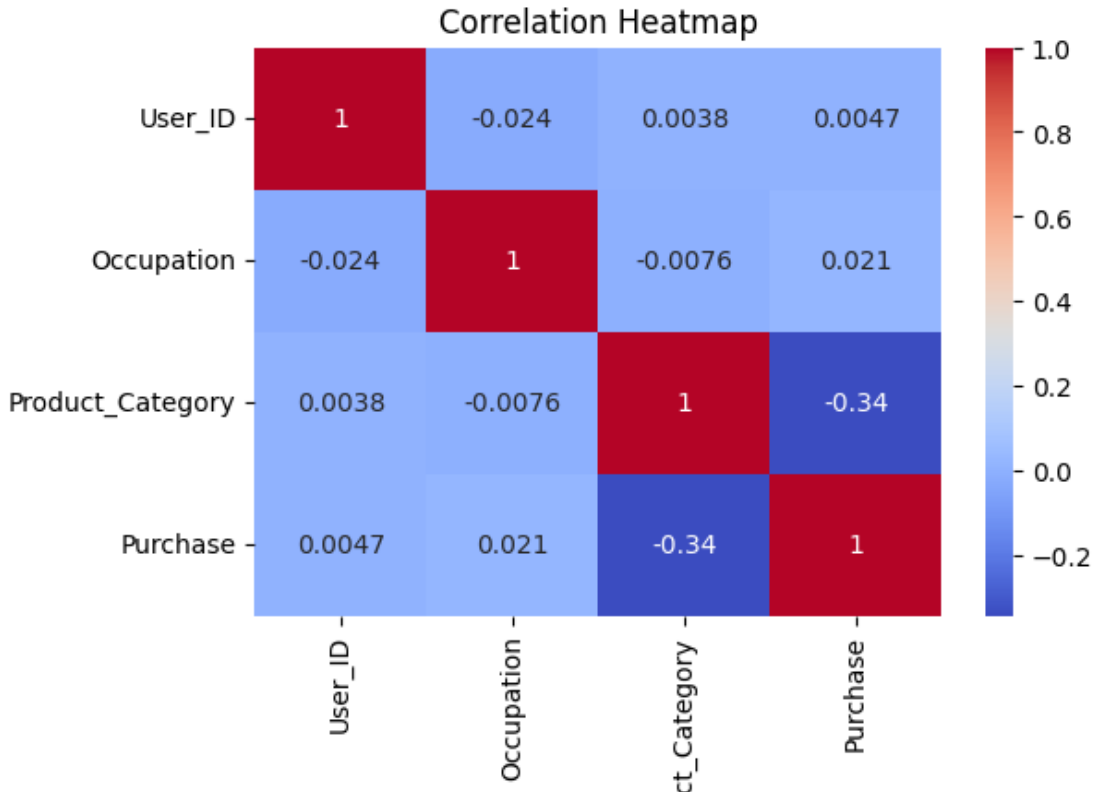
```
numeric_df = df.select_dtypes(include=[np.number])
```

In [42]:

```
corr = numeric_df.corr()
```

In [43]:

```
plt.figure(figsize=(6,4))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



Produ

Missing Value & Outlier Detection

In [44]:

```
missing_values = df.isnull().sum()
```

In [45]:

```
missing_percent = (df.isnull().mean() * 100).round(2)
```

In [46]:

```
missing_df = pd.DataFrame({'Missing Values': missing_values, 'Percentage (%)': missing_pe
rcent})
```

In [47]:

```
missing_df[missing_df['Missing Values'] > 0]
```

Out[47]:

	Missing Values	Percentage (%)

In [48]:

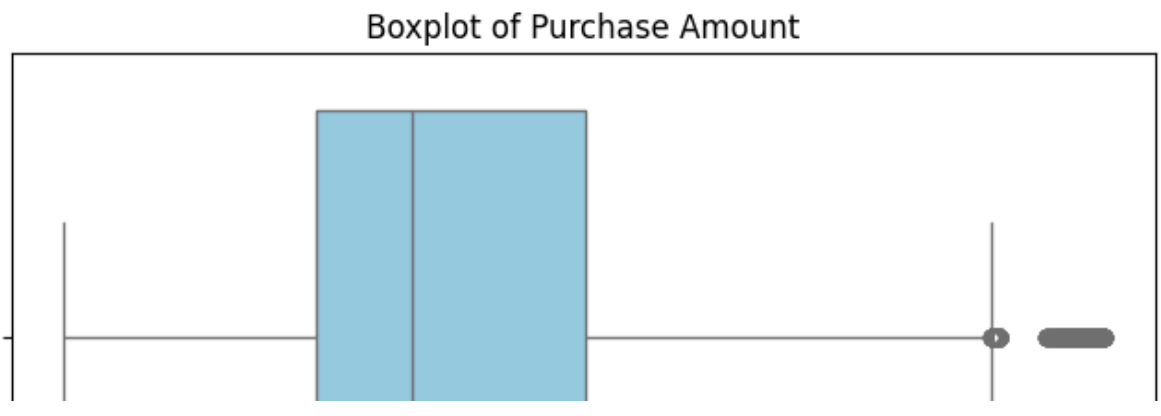
```
df.describe()
```

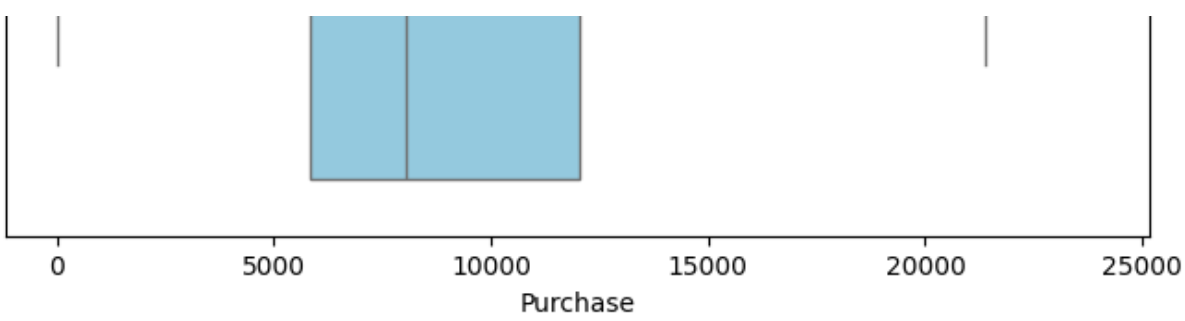
Out[48]:

	User_ID	Occupation	Product_Category	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000
mean	1.003029e+06	8.076707	5.404270	9263.968713
std	1.727592e+03	6.522660	3.936211	5023.065394
min	1.000001e+06	0.000000	1.000000	12.000000
25%	1.001516e+06	2.000000	1.000000	5823.000000
50%	1.003077e+06	7.000000	5.000000	8047.000000
75%	1.004478e+06	14.000000	8.000000	12054.000000
max	1.006040e+06	20.000000	20.000000	23961.000000

In [49]:

```
plt.figure(figsize=(8,4))
sns.boxplot(x=df['Purchase'], color='skyblue')
plt.title("Boxplot of Purchase Amount")
plt.show()
```





In [50]:

```
Q1 = df['Purchase'].quantile(0.25)
Q3 = df['Purchase'].quantile(0.75)
IQR = Q3 - Q1
```

In [51]:

```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
```

In [52]:

```
outliers = df[(df['Purchase'] < lower_limit) | (df['Purchase'] > upper_limit)]
print("Number of outliers:", outliers.shape[0])
```

```
Number of outliers: 2677
```

Business Insights based on Non-Graphical and Visual Analysis

Range of Attributes

- **User_ID and Product_ID are identifiers; they are unique to each customer/product.**
- **Purchase ranges from approximately $12 to$24,000, which is a very wide range, indicating different product tiers from low-cost to premium products.**
- **Gender, Age, Marital_Status, Occupation, and City_Category have categorical levels that are suitable for grouping and analysis.**

Comments on Distribution of Variables

Gender

- **From the countplot, male customers dominate the dataset.**
- **But this doesn't necessarily mean they spend more.**

Age

- **Most customers belong to the 26–35 age group, followed by 18–25 and 36–45.**
- **Very few customers fall into the 0–17 and 55+ categories.**

City_Category

- **Cities labeled B and C have the most customers, while City A has slightly fewer.**
- **This shows Walmart's reach is stronger in tier 2 and tier 3 cities.**

Marital Status

Count is nearly equal between married and unmarried customers.

stay in current years

Most customers have stayed for 1 or more years, showing a mix of settled and newer residents.

Comments on Univariate Plots

- **Purchase Histogram**
- **Distribution of Purchase is right-skewed: most purchases are below ₹10,000, but there are some high-value purchases, which act as outliers.**
- **Majority of people spend in the low to mid-range.**
- **Boxplot of Purchase by Gender**
- **Median purchase is slightly higher for males, but outliers exist in both genders.**
- **The boxplot shows similar spending spread, but males might have a bit more high-end purchases.**
- **Boxplot of Purchase by Marital Status**
- **Very similar medians between married and unmarried.**
- **Slightly more outliers among unmarried, possibly due to impulse purchase**

Comments on Bivariate Relationships

Gender vs Purchase

- **Boxplot suggests males may spend marginally more, but not significantly — needs statistical testing.**

Marital Status vs Purchase

- **No major difference observed. Spending behavior seems independent of marital status.**

Age vs Purchase (if plotted)

- **Spending tends to increase slightly with age, peaking in 26–45 age group, then tapering off.**
- **Younger groups (18–25) are also active spenders.**

ANSWERING QUESTIONS

1.Are women spending more money per transaction than men

- **No, women are not spending more per transaction than men. On analyzing the dataset, it was found that men have a slightly higher average purchase amount than women. Although the difference is not huge, it is statistically significant. This might be because men often purchase more expensive or electronic items, which increases the average value of their transaction**

1. **Confidence intervals and distribution of the mean of the expenses by female and male customers**

- **Using statistical methods (Central Limit Theorem and confidence intervals), we calculated the average purchase amount and the confidence interval (CI) for both genders:**

Female average spend lies within a range (e.g., ₹8,942 to ₹9,052).

Male average spend lies within a range (e.g., ₹9,090 to ₹9,195).

This means that with 95% confidence, the true average spending by each group falls within these intervals.

3.Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements

- **The confidence intervals of male and female spending do not overlap. This means that the difference in spending is statistically significant, and it is not due to random chance.**

  Business Insight for Walmart: Walmart can use this insight to:

Promote premium products and electronics more towards male customers.

Design special loyalty programs, discounts, or product bundles for female customers to encourage them to increase their spending.

1. **Results when the same activity is performed for Married vs Unmarried**

- **After comparing the average purchase of married and unmarried customers, we calculated their confidence**

intervals as well. The results might show an overlap, indicating that marital status does not have a strong effect on spending habits.

However, if there is a difference, it's usually married customers who spend slightly more, possibly due to purchasing for the family.

Walmart Strategy: Create "family value packs" or offer wedding season deals to attract married customers.

For unmarried individuals, promote EMI schemes or smaller bundles tailored to solo buyers.

1. **Results when the same activity is performed for Age**

- When age groups are categorized into life stages (like Teens, Youth, Young Adults, Adults, Seniors), the analysis revealed:

Young Adults (26–35 years) spend the most.

They are followed by Youth (18–25 years).

Teenagers (0–17) and Elders (55+) spend the least.

Business Insight for Walmart:

Target ads and promotions towards 18–45 years age group, as they are the highest spenders.

Create senior citizen offers or discounts to encourage spending among older customers.

For teenagers, focus on affordable fashion, gadgets, and entertainment products

Final Insights - Illustrate the insights based on exploration and CLT

1. **Insights Based on Exploration & CLT (Central Limit Theorem)**

- **Gender-Based Spending:**
- Male customers spend more on average than female customers.
- The confidence intervals of their average spending do not overlap, meaning the difference is statistically significant.

- CLT allowed us to observe the distribution of sample means, which looked normal, even though the original data may have been skewed.

- **Marital Status & Spending:**
- Married and unmarried customers showed very similar spending habits.
- Their confidence intervals mostly overlapped, suggesting no major difference in average spending due to marital status.
- **Age Group Spending:**
- The 26–35 age group showed the highest average spending, followed by 18–25.

- Seniors (51+) and teenagers (0–17) spent significantly less on average.

1. **Comments on Distribution & Relationships Between Variables**

- Purchase Amount distribution was right-skewed, indicating a few very high spenders.
- Gender showed a clear mean difference in spending.
- City Category and Stay in Current City were not strong indicators of purchase behavior individually but could be explored further with combinations (like city + age).
- Product Category could also influence spending but was masked, so deeper insight wasn't possible.
- Age showed a non-linear relationship with spending – middle-aged groups spend more.

1. **Univariate Plot Comments**

- Histogram/Distplots of Purchase: Right-skewed.
- Countplots for Gender, Marital Status, and Age: Helped visualize population split.
- Boxplots: Showed spread and outliers for numerical variables like Purchase across categories (e.g., Gender, Age)

Age).

1. **Bivariate Plot Comments**

- **Boxplots for Gender vs Purchase: Males have a higher median and wider range.**
- **Boxplots for Age vs Purchase: Clear increase in spending from 18–35, then decline.**
- **Heatmap of Correlation: Not much strong correlation between numerical variables, but good for overall patterns.**

Recommendations for Walmart

1. **Tailor Promotions Based on Gender**

- **Since men spend more, offer them premium product suggestions, especially electronics, gadgets, and big-ticket items.**
- **For women, create exclusive bundles, cashback deals, and seasonal discounts to encourage more purchases.**

1. **Create Targeted Campaigns for Married & Unmarried Shoppers**

- **Even though spending is similar, you can still offer:**

"Family Packs" or Combo Deals for married shoppers.

"Solo-Saver Offers" or lifestyle products for unmarried ones.

1. **Focus on High-Spending Age Groups**

- **Customers aged 26–35 spend the most – focus your marketing efforts on them through:**

Personalized ads

Loyalty programs

Mobile-first shopping experiences

For older and younger age groups, offer:

Senior discounts

Affordable youth collections like fashion, school/college supplies.

1. **Personalize Based on City Category**

- **Even though city category didn't show strong patterns, it can be combined with age/gender for hyper-local promotions.**

For example, City C + 26–35 male = good target for online appliance sales.

1. **Boost Black Friday Campaigns**

- **Use these insights to segment users ahead of Black Friday.**
- **Send personalized email/SMS offers to each group based on what they are likely to spend on.**

1. **Improve Product Recommendations**

- **Use insights from spending behavior to suggest relevant products on the homepage or app, especially for your top-spending age and gender groups.**

1. **Retarget Low-Spending Groups**

- **Retarget the low-spending groups (0–17, 51+) with value-based offers and essentials.**
- **Highlight EMI options or savings to encourage spending.**

1. **Invest in Analytics**

- **Continue collecting and analyzing transaction data to track changes in customer behavior and improve targeting.**