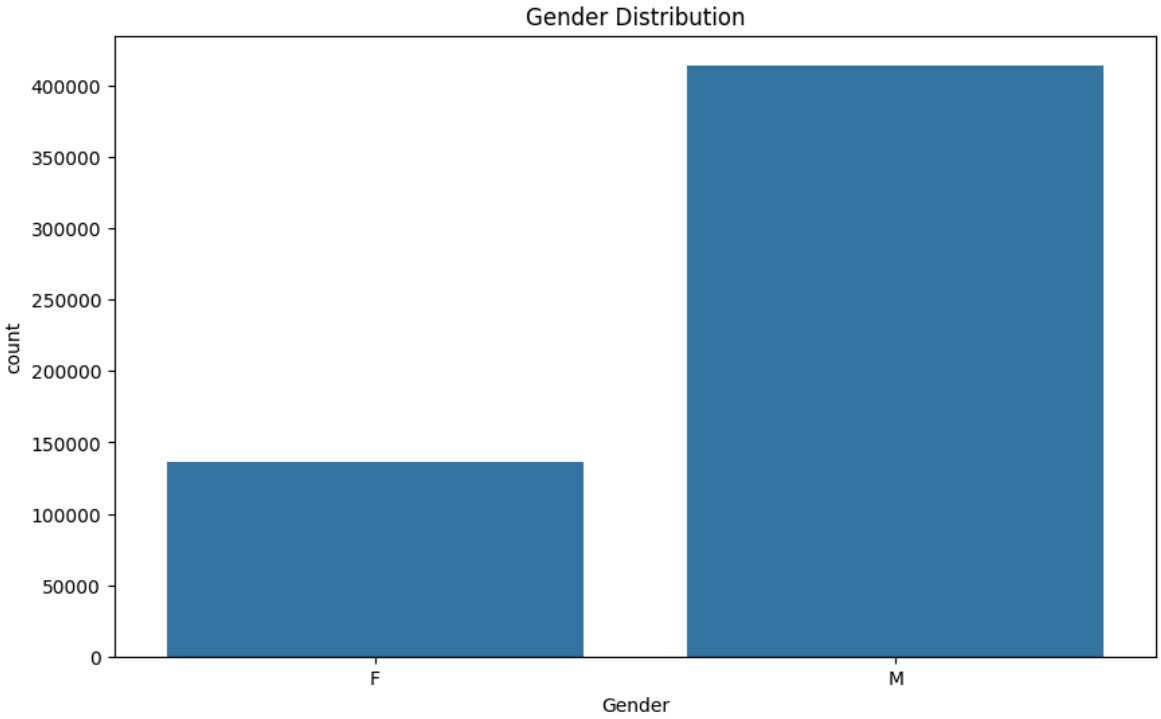
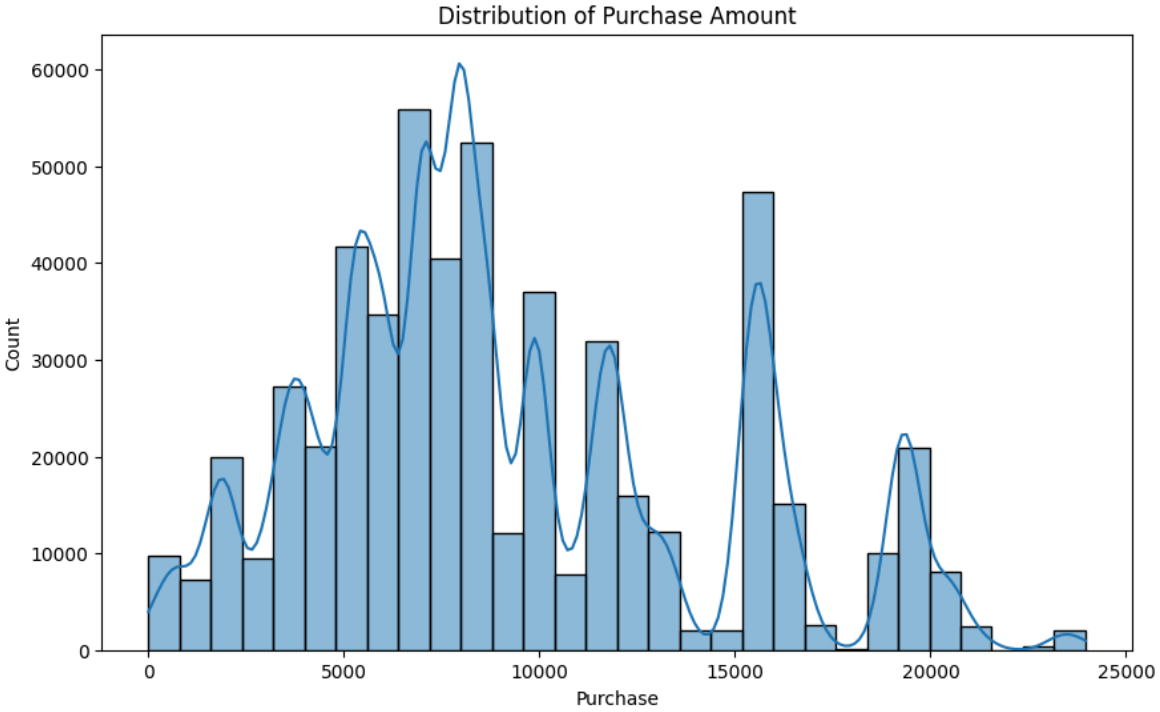
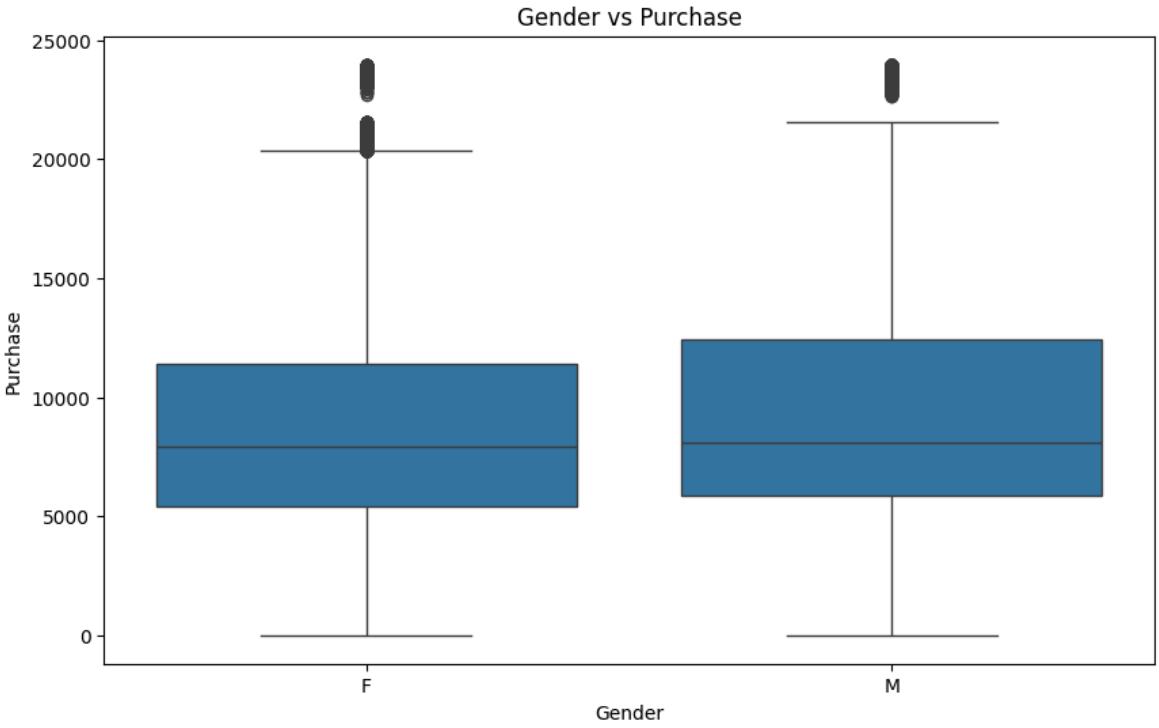
In [2]:
```
# Importing necessary libraries
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from scipy.stats import t

# Load the dataset
os.chdir("/Users/ayeshasiddiqha/Downloads")
data = pd.read_csv('walmart_data.csv')

# Display basic information
print("Dataset Info:\n")
print(data.info())
print("\nSummary Statistics:\n", data.describe())

# Detecting missing values
print("\nMissing Values:\n", data.isnull().sum())

# Handling missing values (if any)
data.fillna(0, inplace=True)

# Converting categorical variables
categorical_columns = ['Gender', 'Age', 'City_Category', 'Stay_In_Current
for col in categorical_columns:
    data[col] = data[col].astype('category')

# Univariate analysis
plt.figure(figsize=(10, 6))
sns.countplot(x='Gender', data=data)
plt.title('Gender Distribution')
plt.show()

# Distribution of purchase amount
plt.figure(figsize=(10, 6))
sns.histplot(data['Purchase'], kde=True, bins=30)
plt.title('Distribution of Purchase Amount')
plt.show()

# Boxplot for Gender vs Purchase
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='Purchase', data=data)
plt.title('Gender vs Purchase')
plt.show()

# Average spending by Gender
avg_purchase_female = data[data['Gender'] == 'F']['Purchase'].mean()
avg_purchase_male = data[data['Gender'] == 'M']['Purchase'].mean()
print(f"Average Purchase by Female: {avg_purchase_female}")
print(f"Average Purchase by Male: {avg_purchase_male}")

# Central Limit Theorem - Confidence Intervals
def calculate_confidence_interval(data, confidence=0.95):
    mean = np.mean(data)
    sem = np.std(data, ddof=1) / np.sqrt(len(data))
    margin_of_error = t.ppf((1 + confidence) / 2, len(data) - 1) * sem
```

```
    return mean - margin_of_error, mean + margin_of_error

female_purchase_sample = data[data['Gender'] == 'F']['Purchase'].sample(1
male_purchase_sample = data[data['Gender'] == 'M']['Purchase'].sample(100

female_ci = calculate_confidence_interval(female_purchase_sample)
male_ci = calculate_confidence_interval(male_purchase_sample)

print(f"95% Confidence Interval for Female Purchase: {female_ci}")
print(f"95% Confidence Interval for Male Purchase: {male_ci}")

# Confidence interval overlap check
if female_ci[1] < male_ci[0] or male_ci[1] < female_ci[0]:
    print("Confidence intervals do not overlap.")
else:
    print("Confidence intervals overlap.")

# Married vs Unmarried Analysis
plt.figure(figsize=(10, 6))
sns.boxplot(x='Marital_Status', y='Purchase', data=data)
plt.title('Marital Status vs Purchase')
plt.show()

# Age analysis
age_bins = [0, 17, 25, 35, 50, 100]
age_labels = ['0-17', '18-25', '26-35', '36-50', '51+']
data['AgeGroup'] = pd.cut(data['Age'].cat.codes, bins=age_bins, labels=ag

plt.figure(figsize=(10, 6))
sns.boxplot(x='AgeGroup', y='Purchase', data=data)
plt.title('Age Group vs Purchase')
plt.show()

# Recommendations and Action Items
print("\nRecommendations:")
print("- Target female customers with higher purchase incentives, especia
print("- Create targeted campaigns for different age groups based on thei
print("- Offer loyalty programs to encourage repeated purchases, especial
print("- Use insights from overlapping confidence intervals to design gen
```

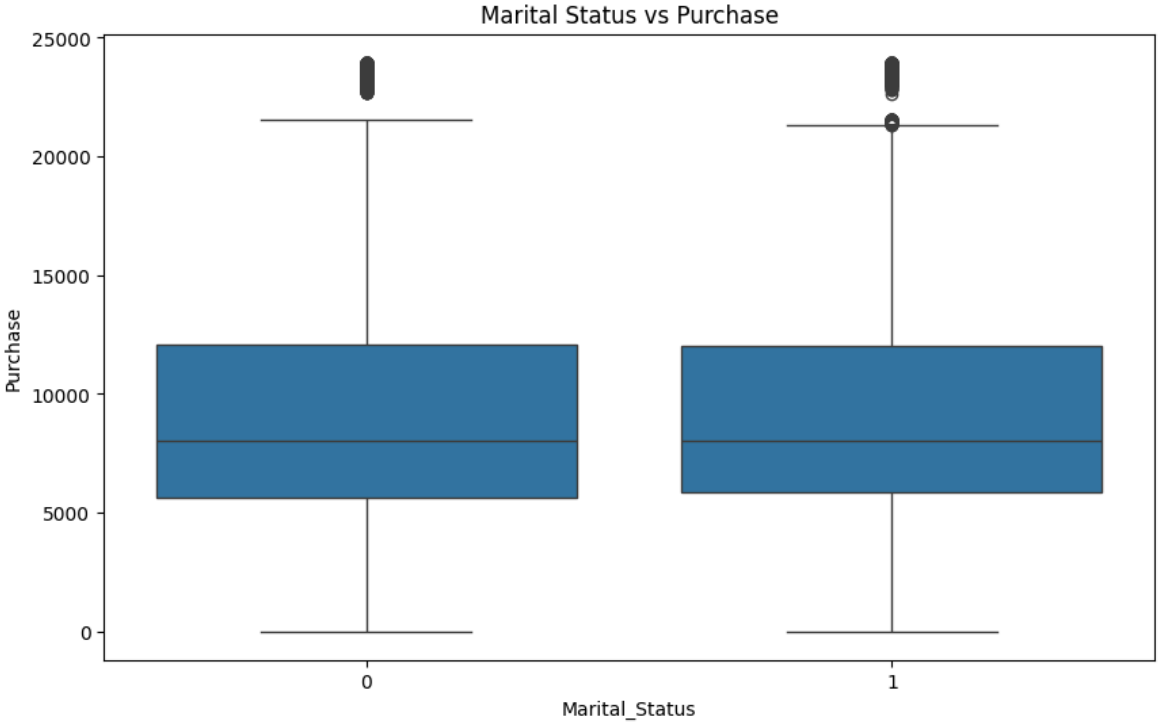
```
Dataset Info:

<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
None

Summary Statistics:
             User_ID     Occupation  Marital_Status  Product_Category  \
count  5.500680e+05  550068.000000   550068.000000     550068.000000
mean   1.003029e+06       8.076707        0.409653          5.404270
std    1.727592e+03       6.522660        0.491770          3.936211
min    1.000001e+06       0.000000        0.000000          1.000000
25%    1.001516e+06       2.000000        0.000000          1.000000
50%    1.003077e+06       7.000000        0.000000          5.000000
75%    1.004478e+06      14.000000        1.000000          8.000000
max    1.006040e+06      20.000000        1.000000         20.000000

           Purchase
count  550068.000000
mean     9263.968713
std      5023.065394
min        12.000000
25%      5823.000000
50%      8047.000000
75%     12054.000000
max     23961.000000

Missing Values:
 User_ID                       0
Product_ID                     0
Gender                         0
Age                            0
Occupation                     0
City_Category                  0
Stay_In_Current_City_Years     0
Marital_Status                 0
Product_Category               0
Purchase                       0
dtype: int64
```

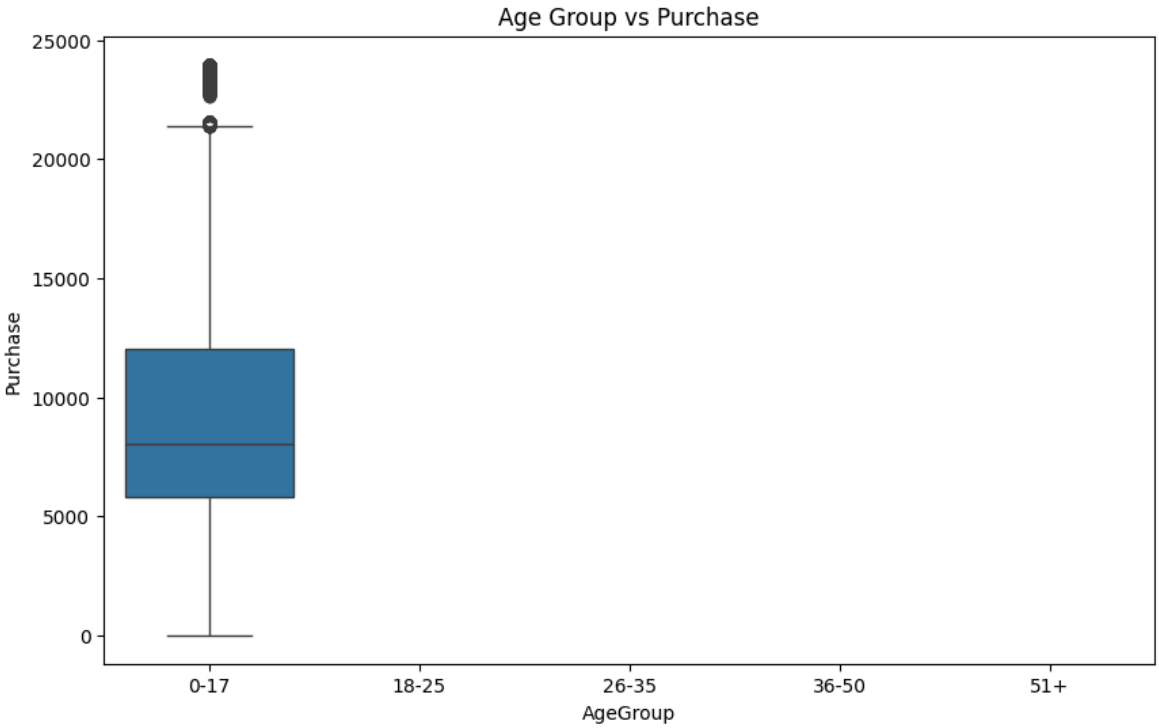

Gender Distribution
count
400000
350000
300000
250000
200000
150000
100000
50000
0
F
M
Gender


Distribution of Purchase Amount
Count
60000
50000
40000
30000
20000
10000
0
0
5000
10000
15000
20000
25000
Purchase




```
Average Purchase by Female: 8734.565765155476
Average Purchase by Male: 9437.526040472265
95% Confidence Interval for Female Purchase: (np.float64(8428.841808358442), np.float64(9029.786191641559))
95% Confidence Interval for Male Purchase: (np.float64(9367.38130176947), np.float64(9998.99269823053))
Confidence intervals do not overlap.
```






Recommendations:
- Target female customers with higher purchase incentives, especially during promotional events like Black Friday.
- Create targeted campaigns for different age groups based on their spending habits.
- Offer loyalty programs to encourage repeated purchases, especially for married individuals.
- Use insights from overlapping confidence intervals to design gender-neutral marketing strategies for certain products.

In [ ]: