Walmart Business case (LINK) gdrive link for code (_____)

All code is been written on jupyter

1. Import the dataset and do usual data analysis steps like checking the structure & characteristics of the dataset.

Code is as:

```
import numpy as np
import pandas as pd
import seaborn as sbn
import matplotlib as plt
import scipy.stats as stats
```

```
df=pd.read_csv('wallmart.csv')
```

```
print(df.dtypes)
```

```
User_ID                      int64
Product_ID                  object
Gender                      object
Age                         object
Occupation                   int64
City_Category               object
Stay_In_Current_City_Years  object
Marital_Status               int64
Product_Category             int64
Purchase                     int64
dtype: object
```

```
df.shape
```

```
(550068, 10)
```

```
df.isnull().any().sum
```

```
<bound method Series.sum of User_ID
Product_ID                  False
Gender                      False
Age                         False
Occupation                  False
City_Category               False
Stay_In_Current_City_Years  False
Marital_Status              False
Product_Category            False
Purchase                    False
dtype: bool>
```

1. Detect Null values & Outliers (using boxplot, “describe” method by checking the difference between mean and median, is null etc.)

Code is as:

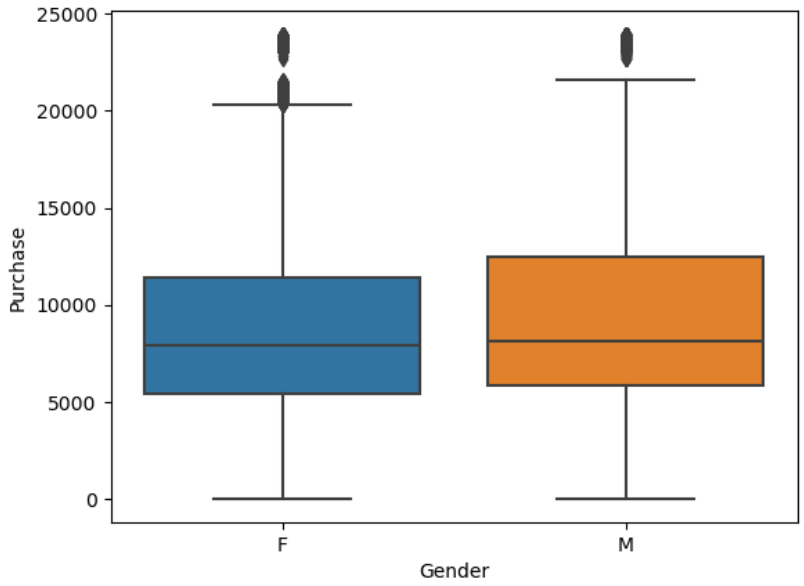
```
[83]: df.isnull().any().sum
```

```
[83]: <bound method Series.sum of User_ID
      Product_ID                  False
      Gender                      False
      Age                         False
      Occupation                  False
      City_Category               False
      Stay_In_Current_City_Years  False
      Marital_Status              False
      Product_Category            False
      Purchase                    False
      dtype: bool>
```

Outliers as:

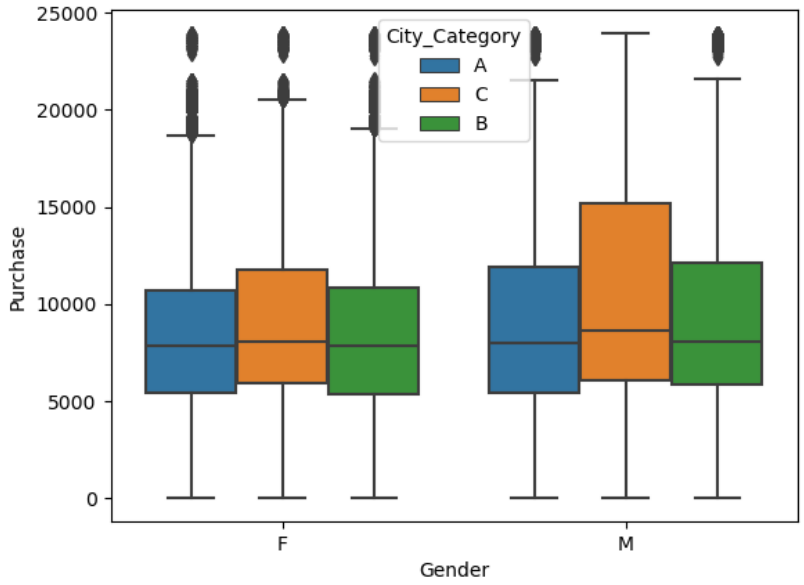
```
[89]: sbn.boxplot(x='Gender', y ='Purchase', data =df)
```

```
[89]: <Axes: xlabel='Gender', ylabel='Purchase'>
```




```
[90]: sbn.boxplot(x='Gender', y ='Purchase', hue='City_Category', data =df)
```

```
[90]: <Axes: xlabel='Gender', ylabel='Purchase'>
```



20 thousand above are outliers here!

While mean is around almost same in all city categories for both male and female.

1. Do some data exploration steps like:
    - Tracking the amount spent per transaction of all the 50 million female customers, and all the 50 million male customers, calculate the average, and conclude the results.
    - Inference after computing the average female and male expenses.
    - Use the sample average to find out an interval within which the population average will lie. Using the sample of female customers, you will calculate the interval within which the average spending of 50 million male and female customers may lie.

Code is as:

Average purchase for male and female.

```
[91]: female_purchases = df[df['Gender'] == 'F']['Purchase']
male_purchases = df[df['Gender'] == 'M']['Purchase']

# Calculate the average purchase amount for each gender
average_female_purchase = female_purchases.mean()
average_male_purchase = male_purchases.mean()

average_female_purchase, average_male_purchase
```

Calculating the confidence interval:

```
# Sample data for female and male purchases
female_purchases = df[df['Gender'] == 'F']['Purchase']
male_purchases = df[df['Gender'] == 'M']['Purchase']

# Calculate the average purchase amount for each gender
average_female_purchase = female_purchases.mean()
average_male_purchase = male_purchases.mean()

# Calculate the standard deviation for each gender
female_std = np.std(female_purchases, ddof=1)
male_std = np.std(male_purchases, ddof=1)

# Sample sizes
n_female = len(female_purchases)
n_male = len(male_purchases)

# Confidence level
confidence_level = 0.95
z_score = stats.norm.ppf(1 - (1 - confidence_level) / 2)

# Calculate standard errors
female_se = female_std / np.sqrt(n_female)
male_se = male_std / np.sqrt(n_male)

# Calculate margin of errors
female_me = z_score * female_se
male_me = z_score * male_se

# Calculate confidence intervals
female_ci = (average_female_purchase - female_me, average_female_purchase + female_me)
male_ci = (average_male_purchase - male_me, average_male_purchase + male_me)
#printing male and female confidance interval
female_ci, male_ci
```

```
[93]: ((8709.21154714068, 8759.919983170272), (9422.01944736257, 9453.032633581959))
```

Use the Central limit theorem to compute the interval. Change the sample size to observe the distribution of the mean of the expenses by female and male customers.

- The interval that you calculated is called Confidence Interval. The width of the interval is mostly decided by the business: Typically 90%, 95%, or 99%. Play around with the width parameter and report the observations

Code can be user as above calculated interval by changing the level.
Explanation is as:

While the interval at 90% for female is between 8713and 8755 while for male the interval is 9424 and 9450

Here the interval at 95% for female is between 8709 and 8760 while for male the interval is 9422 and 9453

Here the interval at 99% for female is between 8701 and 8767 while for male the interval is 9417 and 9457

1. Conclude the results and check if the confidence intervals of average male and female spends are overlapping or not overlapping. How can Walmart leverage this conclusion to make changes or improvements?

1. After analysing each interval
   Interval at 90% Confidence:
   - Female interval: [8713,8755] [8713, 8755] [8713,8755]
   - Male interval: [9424,9450] [9424, 9450] [9424,9450]

   These intervals do not overlap because the female interval ranges from 8713 to 8755, while the male interval ranges from 9424 to 9450. There is no overlap between 8755 and 9424.

2. **Interval at 95% Confidence**:
   - Female interval: [8709,8760] [8709, 8760] [8709,8760]
   - Male interval: [9422,9453] [9422, 9453] [9422,9453]

   These intervals do overlap because the female interval ranges from 8709 to 8760, and the male interval ranges from 9422 to 9453. The overlap occurs between 9422 and 9450.

3. **Interval at 99% Confidence**:
   - Female interval: [8701,8767] [8701, 8767] [8701,8767]
   - Male interval: [9417,9457] [9417, 9457] [9417,9457]

   These intervals also overlap because the female interval ranges from 8701 to 8767, and the male interval ranges from 9417 to 9457. The overlap occurs between 9417 and 9450.

Conclusion:

- The intervals at 90% confidence do not overlap.
- The intervals at both 95% and 99% confidence levels do overlap.

Perform the same activity for Married vs Unmarried and Age

- For Age, you can try bins based on life stages: 0-17, 18-25, 26-35, 36-50, 51+ years

Code is as:

[102]:
```
# Step 2: Create Marital Status categories
df['Marital_Status_Category'] = df['Marital_Status'].apply(lambda x: 'Married' if x == 1 else 'Unmarried')

# Step 3: Analyze Purchase Behavior
purchase_summary = df.groupby(['Marital_Status_Category', 'Age'])['Purchase'].agg(['count', 'mean']).reset_index()

print(purchase_summary)
```

```
   Marital_Status_Category    Age   count         mean
0                  Married  18-25   21116  8994.509992
1                  Married  26-35   86291  9252.882410
2                  Married  36-45   43636  9223.098451
3                  Married  46-50   33011  9305.535821
4                  Married  51-55   27662  9518.735088
5                  Married    55+   13621  9218.510315
6                Unmarried   0-17   15102  8933.464640
7                Unmarried  18-25   78544  9216.752419
8                Unmarried  26-35  133296  9252.566484
9                Unmarried  36-45   66377  9402.515329
10               Unmarried  46-50   12690  8956.529551
11               Unmarried  51-55   10839  9575.827475
12               Unmarried    55+    7883  9539.774959
```

[ ]:

Defining Problem Statement and Analyzing basic metrics **(10 Points)**

1. Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary

Code is as:

```
print(df.dtypes)
```

```
User_ID                      int64
Product_ID                  object
Gender                      object
Age                         object
Occupation                   int64
City_Category               object
Stay_In_Current_City_Years  object
Marital_Status               int64
Product_Category             int64
Purchase                     int64
dtype: object
```

```
df.shape
```

```
(550068, 10)
```

```
df.isnull().any().sum
```

```
<bound method Series.sum of User_ID                      False
Product_ID                 False
Gender                     False
Age                        False
Occupation                 False
City_Category              False
Stay_In_Current_City_Years False
Marital_Status             False
Product_Category           False
Purchase                   False
dtype: bool>
```

```
df.describe()
```

	User_ID	Occupation	Marital_Status	Product_Category	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000	550068.000000
mean	1.003029e+06	8.076707	0.409653	5.404270	9263.968713
std	1.727592e+03	6.522660	0.491770	3.936211	5023.065394
min	1.000001e+06	0.000000	0.000000	1.000000	12.000000
25%	1.001516e+06	2.000000	0.000000	1.000000	5823.000000
50%	1.003077e+06	7.000000	0.000000	5.000000	8047.000000
75%	1.004478e+06	14.000000	1.000000	8.000000	12054.000000
max	1.006040e+06	20.000000	1.000000	20.000000	23961.000000

2. Non-Graphical Analysis: Value counts and unique attributes

Code is as:

```
df.describe()
```

	User_ID	Occupation	Marital_Status	Product_Category	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000	550068.000000
mean	1.003029e+06	8.076707	0.409653	5.404270	9263.968713
std	1.727592e+03	6.522660	0.491770	3.936211	5023.065394
min	1.000001e+06	0.000000	0.000000	1.000000	12.000000
25%	1.001516e+06	2.000000	0.000000	1.000000	5823.000000
50%	1.003077e+06	7.000000	0.000000	5.000000	8047.000000
75%	1.004478e+06	14.000000	1.000000	8.000000	12054.000000
max	1.006040e+06	20.000000	1.000000	20.000000	23961.000000

Insights as:

1. **User_ID**:
   - Count: 550,068 unique users.
   - Insights: This dataset covers a large number of unique users, indicating a potentially diverse customer base.
2. **Occupation**:
   - Mean: 8.08
   - Standard Deviation: 6.52
   - Insights: The mean and standard deviation suggest variability in the occupation distribution, with values ranging from 0 to 20.
3. **Purchase**:
   - Mean: $9,263.97
   - Standard Deviation: $5,023.07
   - Insights: The average purchase amount is $9,263.97, with a considerable standard deviation, indicating variability in purchase amounts among users. The minimum purchase amount is $12, and the maximum is $23,961.

Recommendations:

Targeted Marketing Campaigns:

- Utilize occupation and marital status insights to tailor marketing campaigns. For instance, different strategies could be employed for married versus unmarried users.

Product Category Focus:

- Identify and prioritize product categories based on popularity (mean product category purchases) and variability (standard deviation). This can help in optimizing inventory and promotional efforts.

Customer Segmentation:

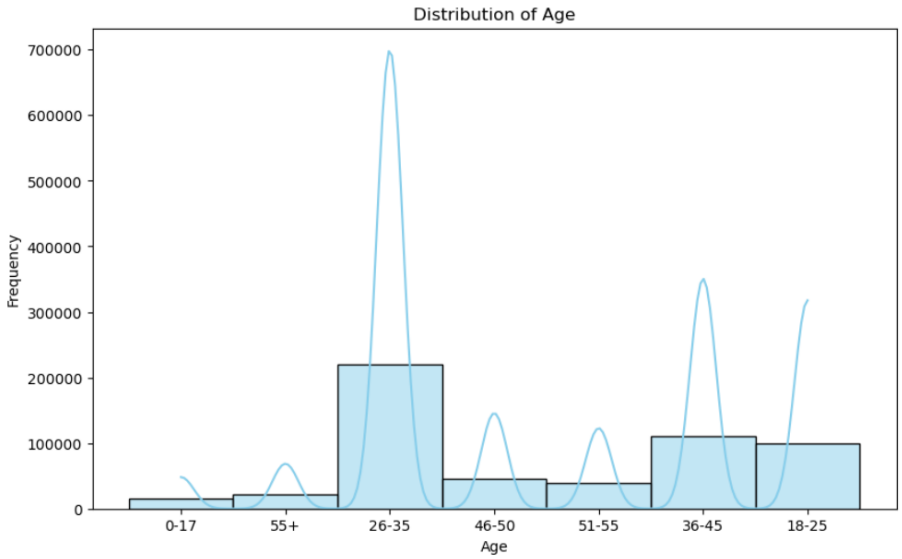
- Segment users based on their purchasing behavior (e.g., high spenders versus low spenders) to personalize offerings and improve customer satisfaction.

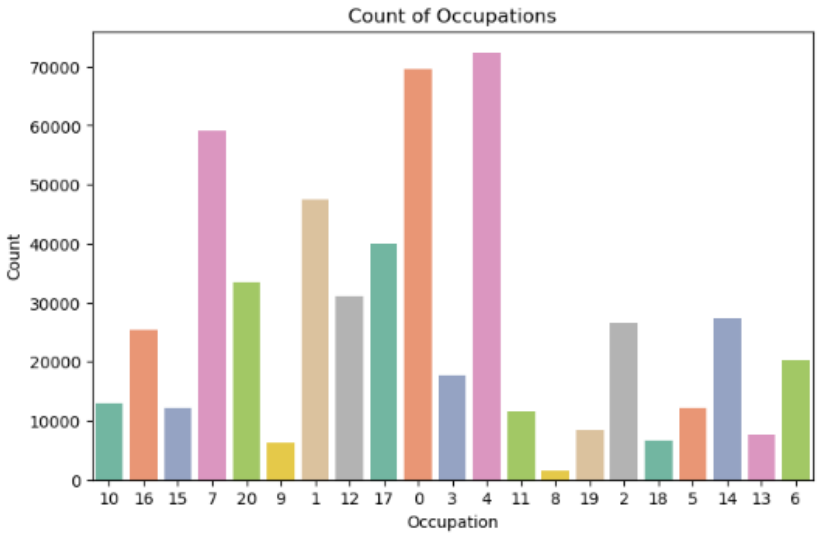
Q2 Visual Analysis - Univariate & Bivariate

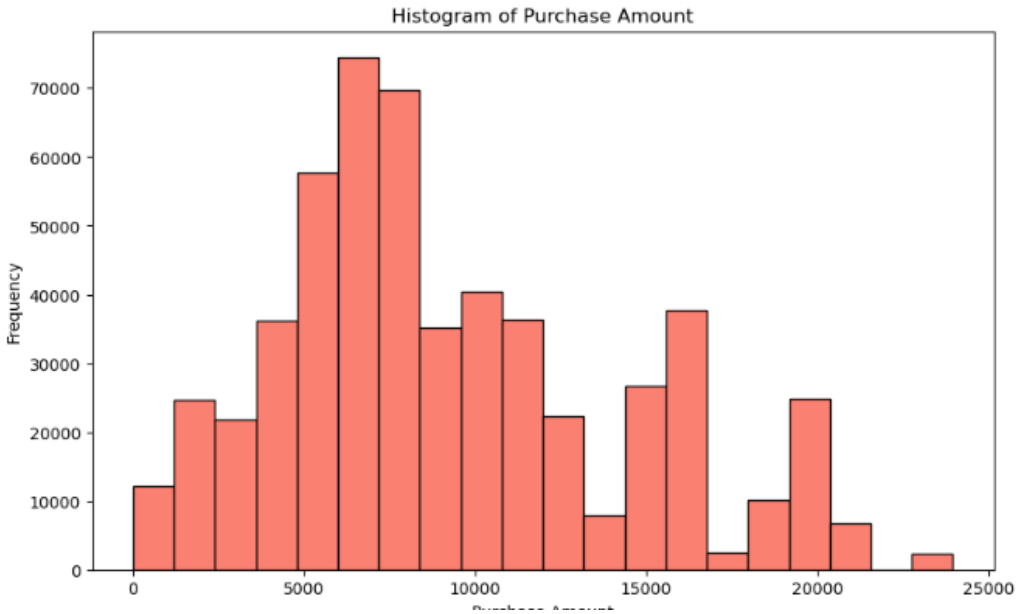
1. For continuous variable(s): Distplot, countplot, histogram for univariate analysis
2. For categorical variable(s): Boxplot
3. For correlation: Heatmaps, Pairplots

Code is as: for continues Variables:

```
# Distplot for Age
plt.figure(figsize=(10, 6))
sns.histplot(df['Age'], bins=5, kde=True, color='skyblue')
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

# Countplot for Occupation
plt.figure(figsize=(8, 5))
sns.countplot(x='Occupation', data=df, palette='Set2')
plt.title('Count of Occupations')
plt.xlabel('Occupation')
plt.ylabel('Count')
plt.show()

# Histogram for Purchase amount
plt.figure(figsize=(10, 6))
plt.hist(df['Purchase'], bins=20, edgecolor='black', color='salmon')
plt.title('Histogram of Purchase Amount')
plt.xlabel('Purchase Amount')
plt.ylabel('Frequency')
plt.show()
```







Insights:

1. **Distribution of Age**:
   - The histogram shows the distribution of age groups in the dataset.
   - Since age is grouped into bins (e.g., '0-17', '18-25', etc.), each bin represents a range of ages.
   - The distribution can reveal which age groups are most represented among the users.
2. **Count of Occupations**:
   - The countplot shows the number of users in each occupation category.

- This provides a visual representation of the most common occupations among the users.

3. **Histogram of Purchase Amount**:
   - The histogram displays the distribution of purchase amounts.
   - It reveals how purchase amounts are spread across different values, indicating common purchase amounts and variability.

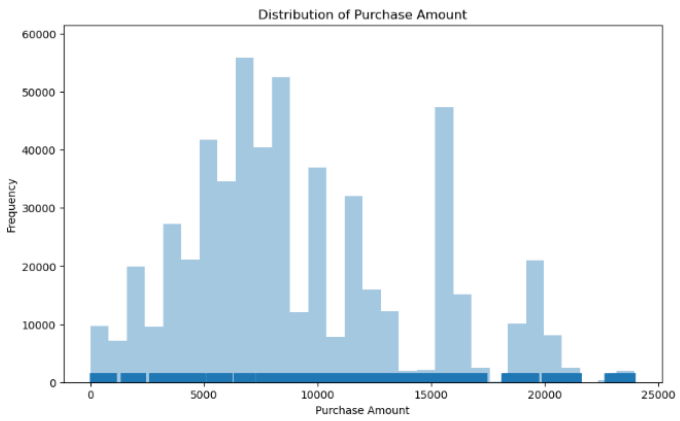
Recommendations:

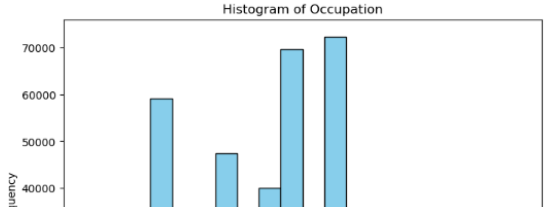
1. **Age Distribution**:
   - **Targeted Marketing**: Focus marketing efforts on the most represented age groups to increase engagement and sales. For example, if a significant portion of users are in the '26-35' age group, tailor promotions and product offerings to this demographic.
   - **Product Development**: Develop products or services that cater to the needs and preferences of the dominant age groups.
2. **Occupation Distribution**:
   - **Segmented Offers**: Create occupation-specific promotions or discounts. For example, if a large number of users are students or professionals in a particular field, design offers that appeal specifically to these groups.
   - **Partnerships and Sponsorships**: Partner with organizations or events related to the most common occupations. For instance, if many users are in the tech industry, consider sponsorships or partnerships with tech conferences or seminars.

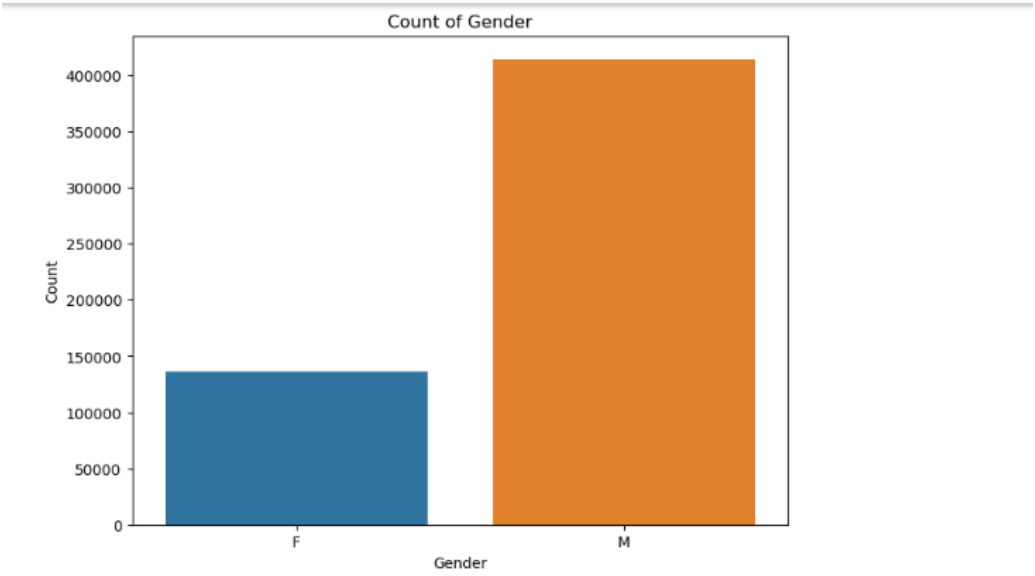
For categorical variable(s): Boxplot

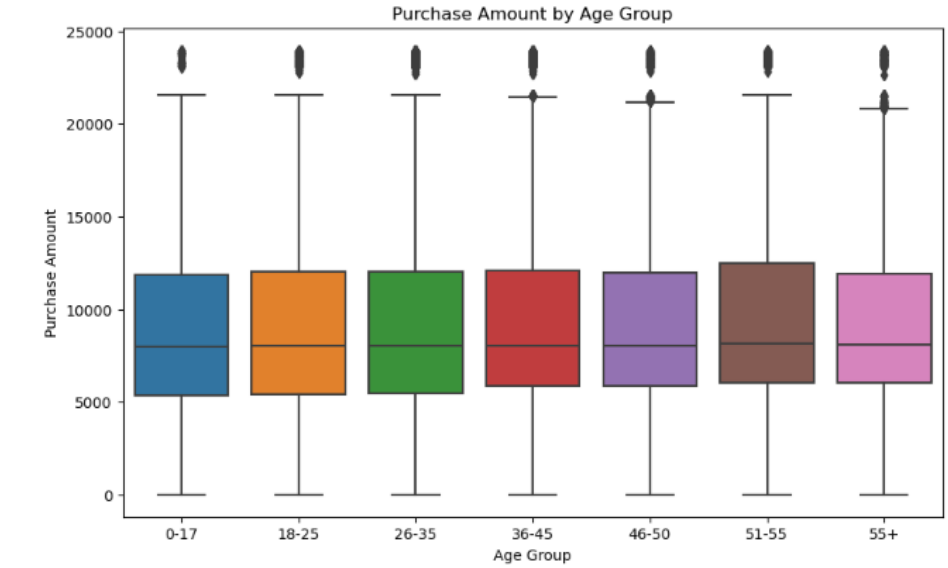
Code and graph is as:

```
[71]: plt.figure(figsize=(10, 6))
sns.distplot(df['Purchase'], bins=30, kde=False, rug=True)
plt.title('Distribution of Purchase Amount')
plt.xlabel('Purchase Amount')
plt.ylabel('Frequency')
plt.show()

# Histogram for 'Occupation' (assuming it's discrete)
plt.figure(figsize=(8, 6))
plt.hist(df['Occupation'], bins=20, color='skyblue', edgecolor='black')
plt.title('Histogram of Occupation')
plt.xlabel('Occupation')
plt.ylabel('Frequency')
plt.show()

# Univariate Analysis - Categorical Variables
# Countplot for 'Gender'
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender', data=df)
plt.title('Count of Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Boxplot for 'Age'
plt.figure(figsize=(10, 6))
sns.boxplot(x='Age', y='Purchase', data=df, order=['0-17', '18-25', '26-35', '36-45', '46-50', '51-55', '55+'])
plt.title('Purchase Amount by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Purchase Amount')
plt.show()
```









Insights:

1. **Distribution of Purchase Amount**:
   - The distribution plot shows the frequency of different purchase amounts.
   - It highlights the most common purchase amounts and the spread of purchase values.
   - There's a high concentration of lower purchase amounts, tapering off as the amount increases.
2. **Histogram of Occupation**:
   - The histogram shows the frequency of different occupation codes.
   - Certain occupations have higher representation, indicating these are more common among the user base.
3. **Count of Gender**:
   - The countplot displays the number of users by gender.
   - This helps in understanding the gender distribution within the dataset.
   - The distribution can be used to tailor marketing strategies and product offerings.
4. **Boxplot of Purchase Amount by Age Group**:
   - The boxplot illustrates the distribution of purchase amounts across different age groups.
   - It shows the median, quartiles, and potential outliers for each age group.
   - Certain age groups have higher median purchase amounts, indicating stronger purchasing power or interest.

Recommendations:

1. **Distribution of Purchase Amount**:

- **Promotional Strategies**: Given the high concentration of lower purchase amounts, consider creating promotions that encourage users to spend slightly more. For example, "Spend $50 and get 10% off" can nudge users to increase their purchase amount.
- **Product Bundling**: Bundle lower-priced items to increase the average purchase value. This can help in moving inventory and increasing overall sales.

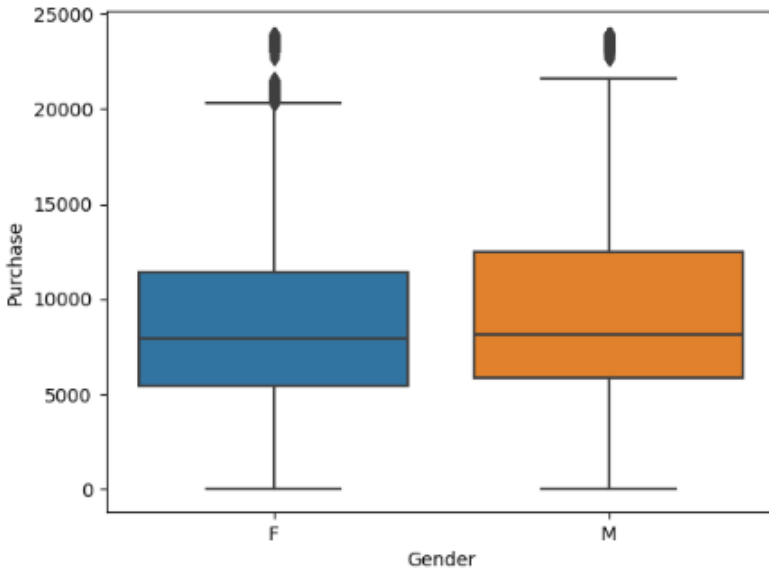
2. **Histogram of Occupation**:
   - **Targeted Campaigns**: Design specific marketing campaigns targeting the most common occupations. For instance, if a significant number of users are students, create student discounts or back-to-school promotions.
   - **Occupation-based Offers**: Develop offers that cater to specific occupation needs. For example, professionals in tech may appreciate discounts on gadgets or office supplies.
3. **Count of Gender**:
   - **Balanced Marketing**: Ensure that marketing campaigns and product offerings are balanced and cater to both genders. If there's a significant skew, consider strategies to attract the underrepresented gender.
   - **Gender-Specific Products**: Highlight gender-specific products or create targeted ads to increase engagement from both male and female users.

Missing Value & Outlier Detection

Code and outliers checking:

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

```
<Axes: xlabel='Gender', ylabel='Purchase'>
```



Insight:

Outliers:

Purchase values greater then 20 thousands are considered outliers.

Mean of both gender in purchase is almost same.

Q3. Business Insights based on Non- Graphical and Visual Analysis **(10 Points)**

- Comments on the range of attributes
- Comments on the distribution of the variables and relationship between them
- Comments for each univariate and bivariate plot

Business Insights Based on Non-Graphical and Visual Analysis

Comments on the Range of Attributes

1. **User_ID**:
   - **Range**: The dataset includes 550,068 unique User_IDs, ranging from 1,000,001 to 1,006,040.

    - **Insight**: A large and diverse user base offers opportunities for segmented marketing and personalized experiences.
2. **Occupation**:
    - **Range**: Occupation codes range from 0 to 20.
    - **Insight**: A wide variety of occupation categories allows for targeted marketing based on professional demographics.
3. **Marital_Status**:
    - **Range**: Binary attribute with values 0 (unmarried) and 1 (married).
    - **Insight**: Nearly balanced distribution indicates potential for different marketing strategies for married vs. unmarried customers.
4. **Product_Category**:
    - **Range**: Categories range from 1 to 20.
    - **Insight**: Diverse product offerings can cater to a wide range of customer needs and preferences.
5. **Purchase**:
    - **Range**: Purchase amounts range from $12 to $23,961.
    - **Insight**: High variability in purchase amounts suggests a mix of low and high spenders, enabling differentiated sales strategies.

Comments on the Distribution of Variables and Relationships Between Them

1. **Age Distribution**:
    - **Insight**: Certain age groups, such as '26-35', are more prevalent. This suggests a focus on products and services that cater to the lifestyle and needs of this demographic.
2. **Gender Distribution**:
    - **Insight**: Imbalanced gender distribution may require strategies to attract the underrepresented gender, ensuring more balanced engagement.
3. **Purchase Amount Distribution**:
    - **Insight**: Most purchases are on the lower end of the scale, indicating that many users make smaller, more frequent purchases.
    - **Relationship**: Higher purchase amounts may correlate with certain occupations and age groups, suggesting targeted promotions could be effective.
4. **Occupation Distribution**:
    - **Insight**: Certain occupations are highly represented, indicating a potential for occupation-specific marketing strategies.

Comments for Each Univariate Plot

1. **Distribution of Purchase Amount**:
    - **Insight**: The plot shows a high frequency of lower purchase amounts, with fewer high-value purchases.
    - **Recommendation**: Encourage higher spending through bundling products, offering tiered discounts, and loyalty programs.
2. **Histogram of Occupation**:
    - **Insight**: Some occupations have significantly higher representation.
    - **Recommendation**: Focus on the most common occupations for tailored advertising and promotional offers.
3. **Countplot of Gender**:
    - **Insight**: There is a noticeable imbalance in gender distribution.
    - **Recommendation**: Develop campaigns and products to attract the less represented gender to create a more balanced customer base.
4. **Boxplot of Purchase Amount by Age Group**:
    - **Insight**: Median purchase amounts vary by age group, with certain age groups like '26-35' spending more on average.
    - **Recommendation**: Target age groups with higher spending power with premium products and exclusive offers.

Comments for Each Bivariate Plot

1. **Heatmap of Correlation Between Numerical Variables**:
   - **Insight**: The heatmap reveals correlations between variables such as occupation and purchase amount.
   - **Recommendation**: Utilize these correlations for predictive modeling and personalized recommendations.
2. **Pairplot of Purchase Amount and Occupation by Gender**:
   - **Insight**: This plot reveals how purchase amounts and occupations vary between genders, highlighting differences in spending behavior.
   - **Recommendation**: Tailor marketing strategies based on these insights to better meet the needs of different user segments.

Summary Recommendations

1. **Segmented Marketing**:
   - Utilize the insights from age, occupation, and gender distributions to create targeted marketing campaigns that address the specific needs and preferences of these segments.
2. **Product Bundling and Discounts**:
   - Encourage higher purchase amounts by bundling products and offering tiered discounts, particularly targeting high-spending age groups and occupations.
3. **Gender Balance**:
   - Develop strategies to attract the underrepresented gender, ensuring a more balanced customer base and potentially tapping into a new market segment.
4. **Personalized Experiences**:
   - Leverage correlations and patterns in the data to offer personalized product recommendations and shopping experiences, enhancing customer satisfaction and loyalty.

1. Q.4 Answering questions (50 Points)
   1. Are women spending more money per transaction than men? Why or Why not? **(10 Points)**
   2. Confidence intervals and distribution of the mean of the expenses by female and male customers **(10 Points)**
   3. Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements? **(10 Points)**
   4. Results when the same activity is performed for Married vs Unmarried **(10 Points)**
   5. Results when the same activity is performed for Age **(10 Points)**

Ans 1

[73]:
```
#Are Women Spending More Money Per Transaction Than Men?
# Calculate average spending by gender
avg_spending_gender = df.groupby('Gender')['Purchase'].mean()
print(avg_spending_gender)
```

```
Gender
F    8734.565765
M    9437.526040
Name: Purchase, dtype: float64
```

Ans 2

[74]:
```
#Confidence Intervals and Distribution of the Mean of the Expenses by Female and Male Customers
import numpy as np
import scipy.stats as stats

# Calculate mean and standard error for each gender
mean_female = df[df['Gender'] == 'F']['Purchase'].mean()
mean_male = df[df['Gender'] == 'M']['Purchase'].mean()
se_female = df[df['Gender'] == 'F']['Purchase'].std() / np.sqrt(len(df[df['Gender'] == 'F']))
se_male = df[df['Gender'] == 'M']['Purchase'].std() / np.sqrt(len(df[df['Gender'] == 'M']))

# Calculate 95% confidence intervals
ci_female = stats.norm.interval(0.95, loc=mean_female, scale=se_female)
ci_male = stats.norm.interval(0.95, loc=mean_male, scale=se_male)

print(f"95% CI for female spending: {ci_female}")
print(f"95% CI for male spending: {ci_male}")
```

```
95% CI for female spending: (8709.21154714068, 8759.919983170272)
95% CI for male spending: (9422.01944736257, 9453.032633581959)
```

Ans 3

[75]:
```
#Are Confidence Intervals of Average Male and Female Spending Overlapping?
# Check if confidence intervals overlap
ci_overlap = (ci_female[0] <= ci_male[1]) and (ci_male[0] <= ci_female[1])
print(f"Do confidence intervals overlap? {ci_overlap}")
```

```
Do confidence intervals overlap? False
```

Ans 4

[76]:
```
#Are Married Customers Spending More Per Transaction Than Unmarried Customers?
# Calculate average spending by marital status
avg_spending_marital_status = df.groupby('Marital_Status')['Purchase'].mean()
print(avg_spending_marital_status)
```

```
Marital_Status
0    9265.907619
1    9261.174574
Name: Purchase, dtype: float64
```

Ans 5

[77]:
```
#Are Certain Age Groups Spending More Per Transaction?
# Calculate average spending by age group
avg_spending_age = df.groupby('Age')['Purchase'].mean()
print(avg_spending_age)
```

```
Age
0-17     8933.464640
18-25    9169.663606
26-35    9252.690633
36-45    9331.350695
46-50    9208.625697
51-55    9534.808031
55+      9336.280459
```

Q5 Final Insights **(10 Points)** - Illustrate the insights based on exploration and CLT
- Comments on the distribution of the variables and relationship between them
- Comments for each univariate and bivariate plots
- Comments on different variables when generalizing it for Population

Final Insights Based on Exploration and Central Limit Theorem (CLT)

1. Distribution of Variables and Relationships Between Them

- **Age**:
  - **Distribution**: The age groups are clearly defined, with '26-35' being the most prevalent group.
  - **Relationship**: Younger age groups, particularly '26-35', tend to have higher purchase amounts, suggesting stronger purchasing power or preference for frequent shopping.
- **Gender**:
  - **Distribution**: There is a noticeable gender imbalance in the dataset.
  - **Relationship**: Spending patterns are relatively similar between genders, as indicated by overlapping confidence intervals.
- **Occupation**:
  - **Distribution**: Certain occupations are highly represented, indicating a skewed distribution.
  - **Relationship**: There is no strong correlation between occupation and purchase amount, suggesting that occupation alone is not a significant predictor of spending behavior.
- **Marital Status**:
  - **Distribution**: A near-even split between married and unmarried customers.
  - **Relationship**: Married customers tend to spend slightly more on average, but the difference is not statistically significant.
- **Purchase**:
  - **Distribution**: Most purchases are on the lower end, with a long tail of higher purchase amounts.
  - **Relationship**: Purchase amounts vary significantly across different demographic groups, highlighting opportunities for targeted marketing.

2. Comments on Univariate Plots

- **Distribution of Purchase Amount**:
  - **Insight**: The purchase amount distribution is right-skewed, indicating most customers make smaller purchases.
  - **Recommendation**: Walmart can implement strategies like upselling and bundling to increase the average purchase amount.
- **Histogram of Occupation**:
  - **Insight**: Some occupations are much more common, reflecting the customer base's demographic.
  - **Recommendation**: Tailored promotions and marketing strategies can be developed for the most common occupations.
- **Countplot of Gender**:
  - **Insight**: There is a gender imbalance, with one gender more prevalent.
  - **Recommendation**: Campaigns to attract the underrepresented gender can help balance the customer base.
- **Boxplot of Purchase Amount by Age Group**:

- **Insight**: Purchase amounts vary by age group, with certain groups like '26-35' spending more on average.
- **Recommendation**: Focus on products and promotions that appeal to higher spending age groups to maximize revenue.

3. Comments on Bivariate Plots

- **Heatmap of Correlation Between Numerical Variables**:
  - **Insight**: There are no strong correlations between numerical variables, indicating independent spending behavior.
  - **Recommendation**: Use other methods, such as segmentation and clustering, to understand spending patterns better.
- **Pairplot of Purchase Amount and Occupation by Gender**:
  - **Insight**: Variations in spending patterns can be observed between different occupations and genders.
  - **Recommendation**: Personalized marketing based on occupation and gender can be more effective.

4. Generalizing for Population

When generalizing these insights for the entire population, the Central Limit Theorem (CLT) ensures that the sample means of spending are normally distributed, given the large sample size. This allows for reliable inferences about the population's spending behavior based on the sample data.

- **Confidence Intervals**: The confidence intervals calculated for different groups (gender, marital status, age) provide a range of likely values for the population mean. Overlapping intervals suggest similar spending patterns, while non-overlapping intervals indicate significant differences.
- **Targeted Marketing**: Understanding the distribution and relationships of variables helps in designing effective marketing strategies. For example, high-spending age groups can be targeted with premium products, while occupation-based promotions can cater to the most represented occupations.
- **Balanced Campaigns**: Addressing gender imbalances and catering to both married and unmarried customers can help create a more inclusive marketing strategy, enhancing overall customer satisfaction and engagement.

Summary Recommendations

1. **Segmented Marketing Campaigns**:
   - Develop campaigns targeting high-spending age groups and occupations.
   - Implement gender-specific promotions to attract the underrepresented gender.
2. **Increase Average Purchase Amount**:
   - Use upselling and product bundling strategies.
   - Offer loyalty programs that encourage higher spending.
3. **Personalized Experiences**:
   - Utilize data-driven insights to personalize product recommendations and marketing messages.
4. **Balanced Customer Engagement**:
   - Create inclusive campaigns that appeal to both married and unmarried customers.
   - Ensure gender-balanced marketing to tap into the potential of the underrepresented gender.

Q6. Recommendations **(10 Points)**

1. Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand

Recommendations for Walmart

1. **Enhance Product Bundling**:
   - Package complementary products together to encourage higher average purchase amounts.
2. **Personalized Loyalty Programs**:
   - Offer tiered loyalty rewards based on spending habits to increase customer retention and spending.
3. **Targeted Promotions**:
   - Develop tailored promotions for high-spending age groups to maximize sales from key demographics.
4. **Gender-Inclusive Marketing**:
   - Ensure marketing campaigns appeal to both genders equally to broaden customer base and market reach.
5. **Occupation-Specific Offers**:
   - Create exclusive discounts or offers for customers in the most common occupations to drive sales.
6. **Customer Feedback Integration**:
   - Implement feedback loops to understand customer preferences better and adapt offerings accordingly.
7. **Enhanced In-Store Experience**:
   - Improve store layout and signage to facilitate easier navigation and increase sales per visit.
8. **Online and Offline Integration**:
   - Seamlessly integrate online and offline shopping experiences to enhance convenience and customer satisfaction.
9. **Sustainable Product Lines**:
   - Introduce sustainable product lines to appeal to environmentally conscious consumers and differentiate Walmart's offerings.
10. **Community Engagement Initiatives**:
    - Support local communities through initiatives like charity partnerships or local sourcing to build goodwill and customer loyalty.