Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

Business Problem

The Management team at Walmart Inc. wants to analyze the customer purchase behavior (specifically, purchase amount) against the customer's gender and the various other factors to help the business make better decisions. They want to understand if the spending habits differ between male and female customers: Do women spend more on Black Friday than men? (Assume 50 million customers are male and 50 million are female).

Dataset

The company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday. The dataset has the following features:

Dataset link: Walmart_data.csv

User_ID: User ID

Product_ID: Product ID

Gender: Sex of User

Age: Age in bins

Occupation: Occupation(Masked)

City_Category: Category of the City (A,B,C)

StayInCurrentCityYears: Number of years stay in current city

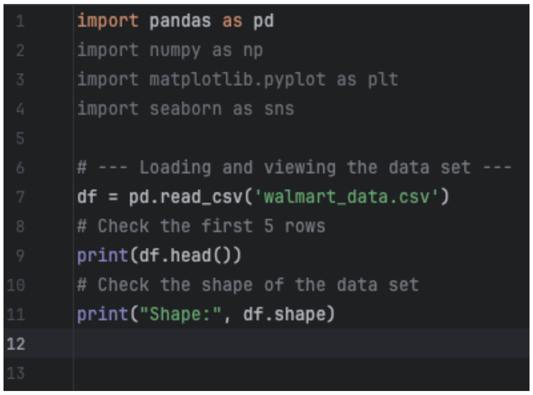
Marital_Status: Marital Status

ProductCategory: Product Category (Masked)

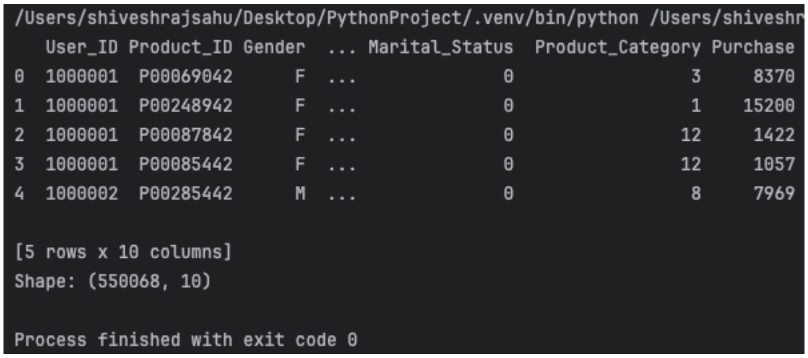
Purchase: Purchase Amount

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

SEGMENT 1: Import the libraries and load the Dataset



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

# --- Loading and viewing the data set ---
df = pd.read_csv('walmart_data.csv')
# Check the first 5 rows
print(df.head())
# Check the shape of the data set
print("Shape:", df.shape)


```



```
/Users/shiveshrajsahu/Desktop/PythonProject/.venv/bin/python /Users/shiveshr
   User_ID Product_ID Gender  ... Marital_Status  Product_Category Purchase
0  1000001  P00069042      F  ...              0                 3     8370
1  1000001  P00248942      F  ...              0                 1    15200
2  1000001  P00087842      F  ...              0                12     1422
3  1000001  P00085442      F  ...              0                12     1057
4  1000002  P00285442      M  ...              0                 8     7969

[5 rows x 10 columns]
Shape: (550068, 10)

Process finished with exit code 0
```

❖ **Import the Libraries:**

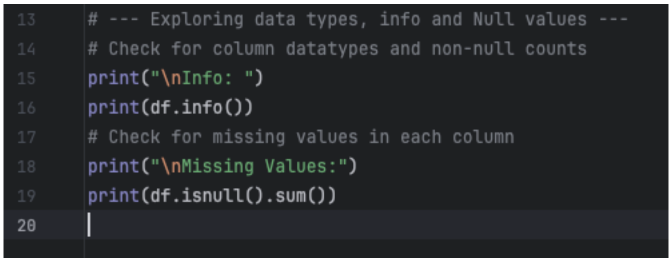
- *(import pandas as pd):* Pandas is the most popular library for data manipulation and analytics.
- *(import numpy as np):* Numpy stands for Numerical Python and is essential for working with arrays and numerical operations.
- *(import matplotlib.pyplot as plt):* Matplot is most widely used plotting library and pyplot module lets us create plots.
- *(import seaborn as sns):* Seaborn is a statistical data visualization library built on top of matplot, it makes it attractive and easy to read visualizations with less complex code.

❖ **Loading and viewing the data set:**

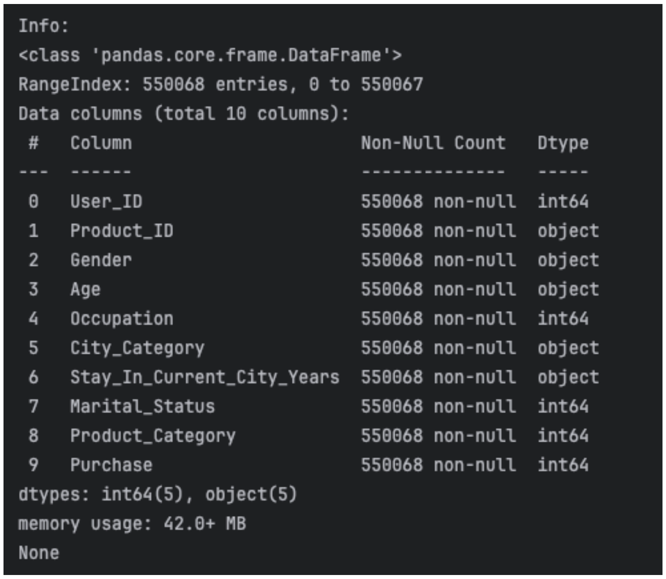
- *(df = pd.read_csv('walmart_data.csv')):* This loads the dataset, the data is loaded as a data frame.
- *(print(df.head())):* To check that the dataset is loaded properly and to see the first 5 rows of the loaded dataset.
- *(print("Shape:", df.shape)):* Confirms the size and shape of our dataset, for example, number of rows and columns.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

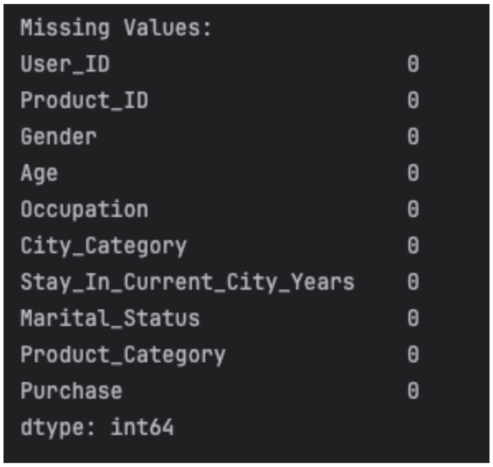
SEGMENT 2: Exploring Data Types, info and Null Values



```
# --- Exploring data types, info and Null values ---
# Check for column datatypes and non-null counts
print("\nInfo: ")
print(df.info())
# Check for missing values in each column
print("\nMissing Values:")
print(df.isnull().sum())

```



```
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
```



```
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
```

❖ Code Explanation:

- *(df,info());* Prints the number of rows, each column and number of non-null values.
- *(df.isnull().sum()):* Counts how many missing values are there in each column.

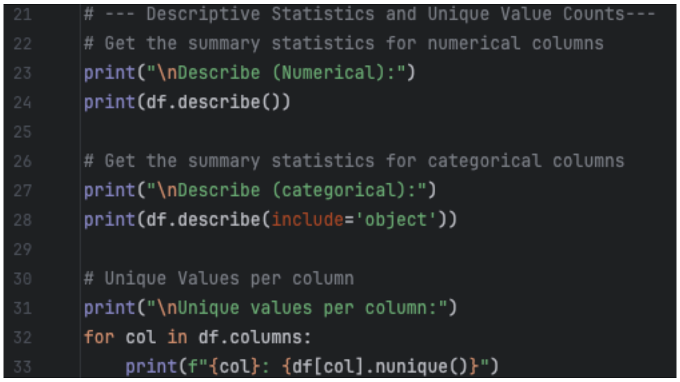
❖ Result Explanation:

- The *.info()* output shows that our dataset has 550,068 entries (rows) and 10 columns. Every column, including *User_Id*, *Gender*, *Age*, *Occupation*, *City_Category*, etc. has non-null values, meaning that there are no null values in any column.
- The *.isnull().sum()* output confirms this by printing 0 for every column, indicating that there are no null values.
- **Conclusion:**

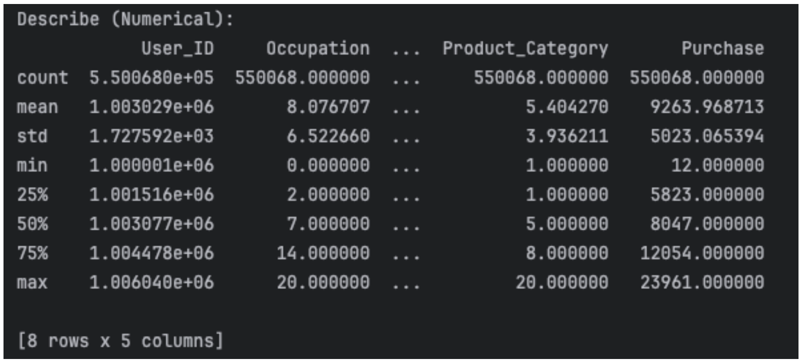
  The dataset is clean. There are no null or missing values to handle. This means we can proceed directly to future data analysis without any preprocessing for missing data.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

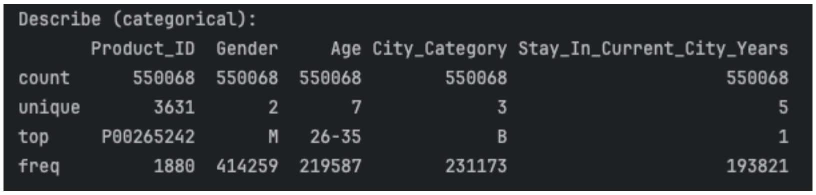
SEGMENT 3: Descriptive Statistics and Unique Value Count



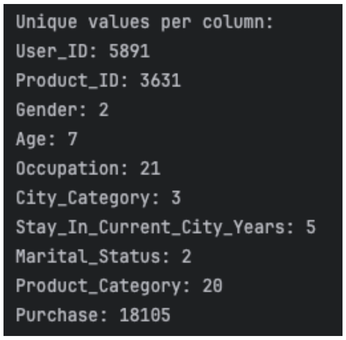
```
# --- Descriptive Statistics and Unique Value Counts---
# Get the summary statistics for numerical columns
print("\nDescribe (Numerical):")
print(df.describe())

# Get the summary statistics for categorical columns
print("\nDescribe (categorical):")
print(df.describe(include='object'))

# Unique Values per column
print("\nUnique values per column:")
for col in df.columns:
    print(f"{col}: {df[col].nunique()}")
```



```
Describe (Numerical):
            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

[8 rows x 5 columns]
```



```
Describe (categorical):
       Product_ID  Gender     Age City_Category Stay_In_Current_City_Years
count      550068  550068  550068        550068                     550068
unique       3631       2       7             3                          5
top     P00265242       M   26-35             B                          1
freq         1880  414259  219587        231173                     193821
```



```
Unique values per column:
User_ID: 5891
Product_ID: 3631
Gender: 2
Age: 7
Occupation: 21
City_Category: 3
Stay_In_Current_City_Years: 5
Marital_Status: 2
Product_Category: 20
Purchase: 18105
```

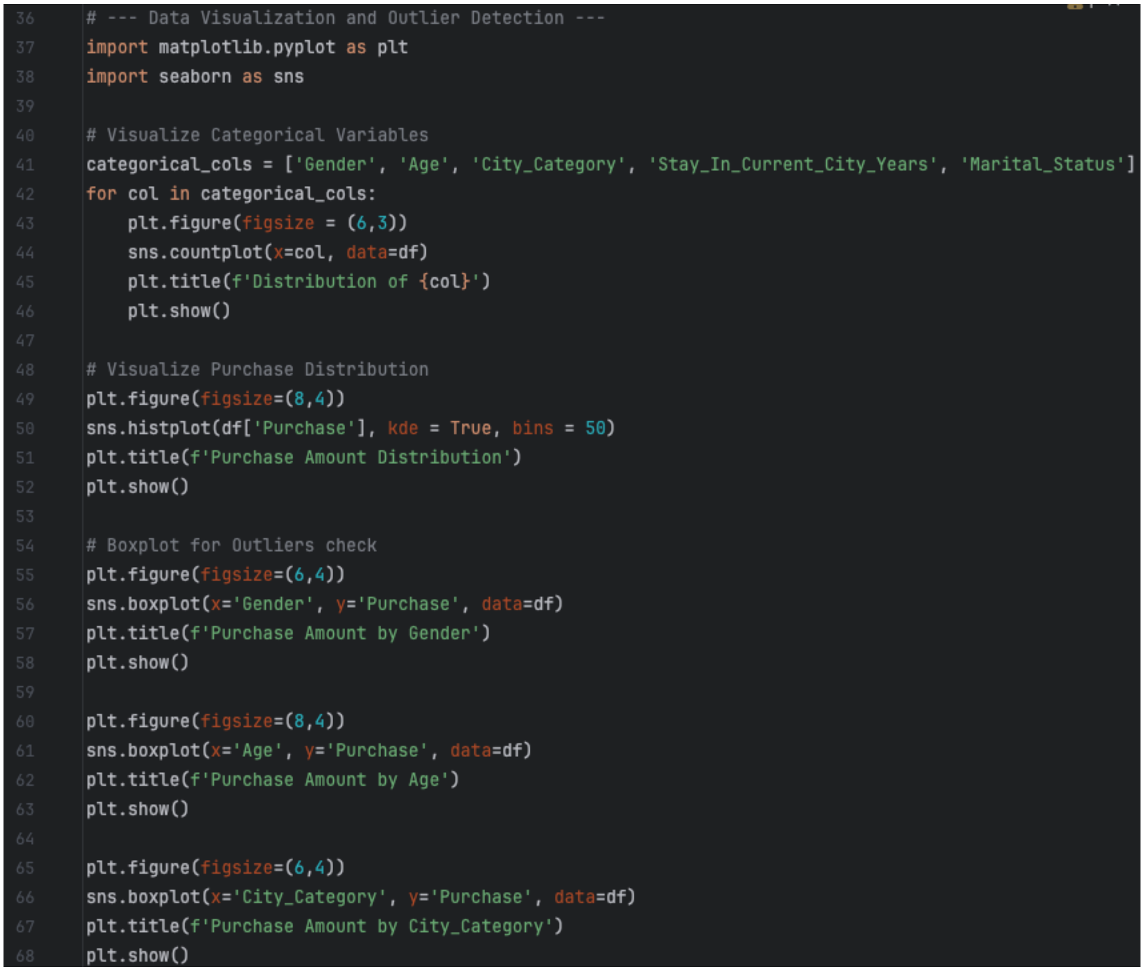
Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

- **Code Explanation:**
  - ***df.describe():***
    - Gives the summary stats for all numerical columns
      - Count: number of non-null values
      - Mean: average value
      - STD: standard deviation
      - Min: minimum value
      - 25% / 50% / 75%: quartiles for the spread of data
      - Max: maximum value
  - ***df.describe(include='object'):***
    - Gives the summary of all categorical and text columns:
      - Count: number of non-null values
      - Unique: number of unique values
      - Top: most common values or also known as mode
      - Freq: frequency of the most common value
  - ***Unique Value Count:***
    - The for loop goes through every column and prints:
      - The number of unique values in that column
      - Help us know which columns have a lot of variation
- **Result Explanation:**
  - ***The numeric summary by (df.describe()) reveal the following:***
      - The average purchase amount is Rs.9,264, what a standard deviation of Rs.5,023, indicating a wide spread of customer spendings.
      - The minimum purchase amount is Rs.12 and the maximum is Rs.23,961, showing that the purchases range from very small to relatively very large.
      - The quartiles tell us that: 25% of purchases are below Rs5,823
50% of purchases is Rs.8,047
75% of purchases are below Rs.12,054
  - ***For categorical columns(df.describe(include='object')):***
      - There are 2 genders (M,F) and 7 distinct groups.
      - The most common gender in the data is M (male), with 414,259 entries.
      - The largest age groups is 26-35, and the most common city category is B.
      - There are 3631 unique product ids and 20 different product categories.
  - ***The unique value counts per column confirm:***
      - Most columns have manageable number of unique values, except for User_id and Product_id, meaning to many users and products.

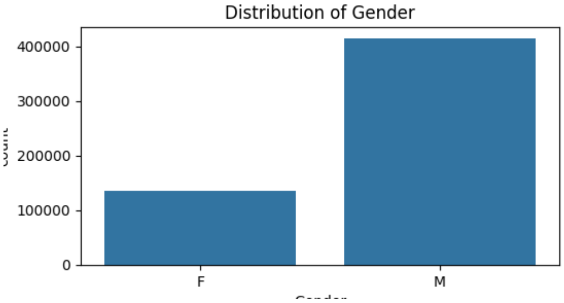
- Gender, Marital_Status, and City_Category have very few unique values, making them suitable for grouping and comparison.
- This initial exploration confirms that the dataset is diverse but well structured, with meaningful variations in both numerical and categorical features.
- The wide range in purchase amounts suggest the need for future analysis, like checking for outliers and grouping differences.
- Categorical columns like gender and age are well represented and ready for group wise analysis in the next upcoming steps.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

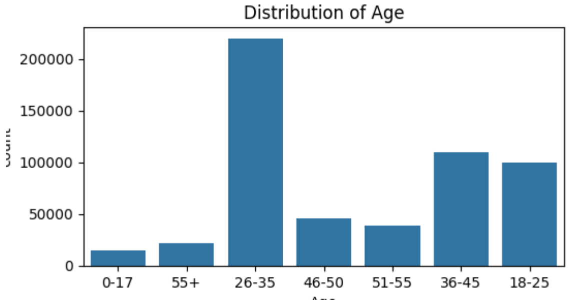
SEGMENT 4: Data Visualization and Outlier Detection



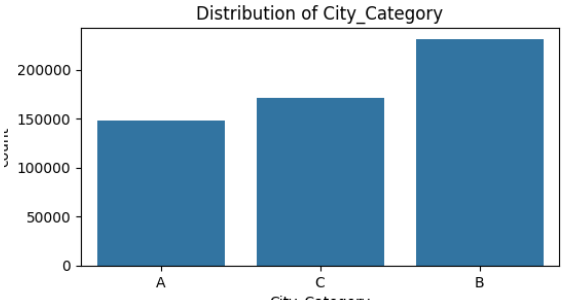
```python
# --- Data Visualization and Outlier Detection ---
import matplotlib.pyplot as plt
import seaborn as sns

# Visualize Categorical Variables
categorical_cols = ['Gender', 'Age', 'City_Category', 'Stay_In_Current_City_Years', 'Marital_Status']
for col in categorical_cols:
    plt.figure(figsize = (6,3))
    sns.countplot(x=col, data=df)
    plt.title(f'Distribution of {col}')
    plt.show()

# Visualize Purchase Distribution
plt.figure(figsize=(8,4))
sns.histplot(df['Purchase'], kde = True, bins = 50)
plt.title(f'Purchase Amount Distribution')
plt.show()

# Boxplot for Outliers check
plt.figure(figsize=(6,4))
sns.boxplot(x='Gender', y='Purchase', data=df)
plt.title(f'Purchase Amount by Gender')
plt.show()

plt.figure(figsize=(8,4))
sns.boxplot(x='Age', y='Purchase', data=df)
plt.title(f'Purchase Amount by Age')
plt.show()

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



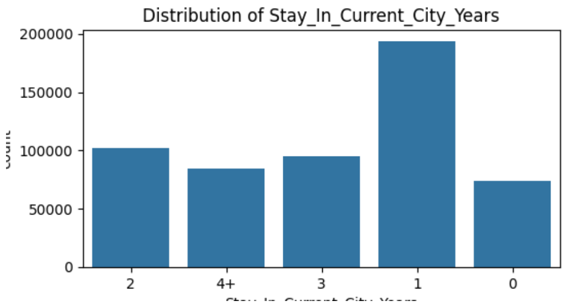
Distribution of Gender



Distribution of Age

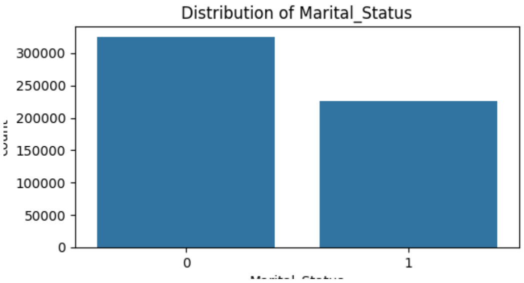


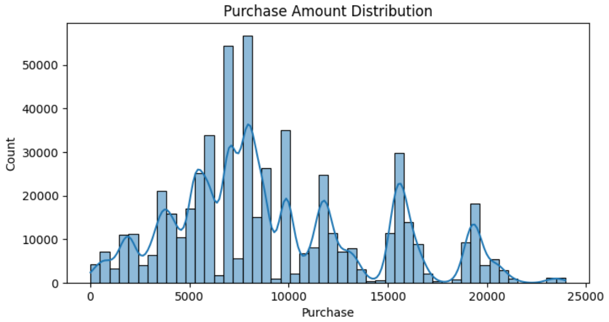
Distribution of City_Category

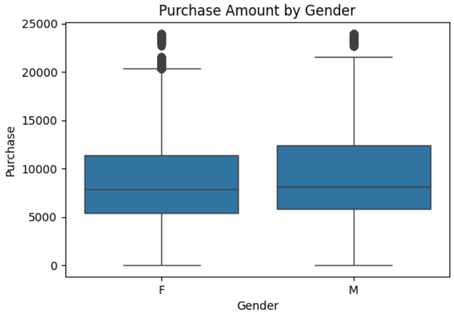


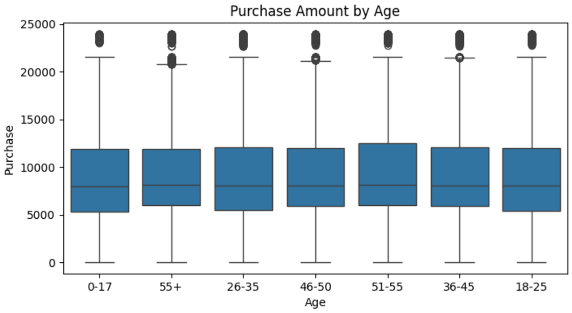
Distribution of Stay_In_Current_City_Years

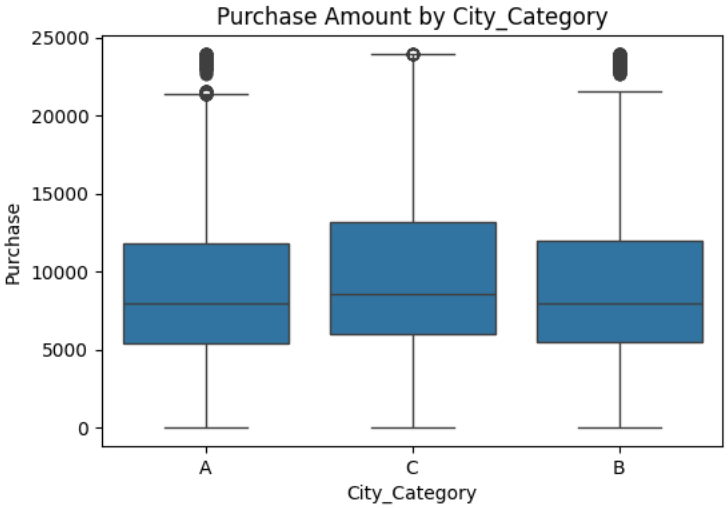
Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu











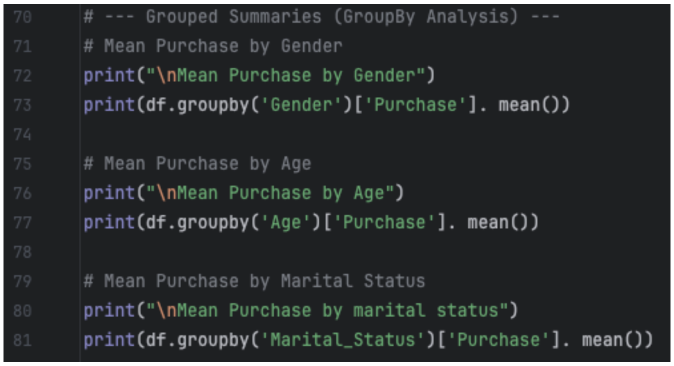
Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

Result Explanation for Visualizations & Outliers

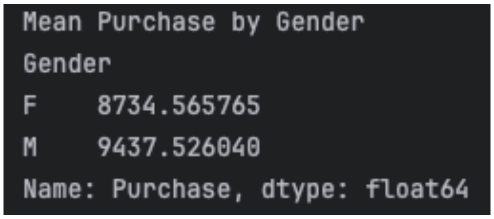
- **Result Explanation: Data Visualization & Outlier Detection**
  - ***Distribution Plots (Countplots):***
    - Gender: The dataset contains significantly more male than female customers.
    - Age: The largest group is 26–35 years, followed by 36–45 and 18–25. Groups like 0–17 and 55+ are much smaller.
    - City_Category: City B has the most customers, followed by C and then A.
    - Stay_In_Current_City_Years: Most customers have stayed in their current city for 1 year, with fewer in the “4+” or “0” years categories.
    - Marital_Status: More customers are marked as unmarried (0) than married (1).
  - ***Purchase Amount Distribution (Histogram):***
    - The purchase amount shows a right-skewed distribution: most purchases are between ₹5,000 and ₹12,000, but there are some purchases that go much higher (up to around ₹24,000).
    - There are visible “spikes” at regular intervals, which likely correspond to popular price points or product bundles.
  - ***Boxplots (Outlier and Group Differences):***
    - By Gender: Males generally spend more on average than females; both groups have similar ranges, but the median purchase is higher for males.
    - By Age: The 51–55 group seems to have the highest median purchase amount. Other age groups have similar spreads, but younger and oldest groups have lower counts.
    - By City Category: City C and City B customers spend more on average compared to City A, but all city groups show a wide range of purchases, with visible outliers (very high-value purchases).
    - Outliers: All boxplots show outliers (dots above the whiskers), which are expected in a large retail dataset—these represent customers who made unusually large purchases.
  - ***Summary:***
    - Walmart’s Black Friday customers are predominantly male and in the 26–35 age group.
    - Most spending falls within a moderate range, but there are frequent large purchases.Males, certain age groups, and residents of City C or B tend to spend more on average.
    - Outliers are common but are a natural part of large sales data.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

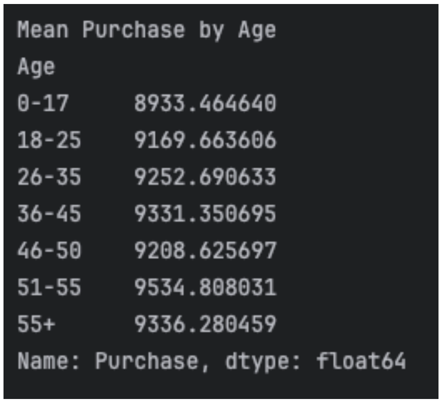
SEGMENT 5: Grouped Summaries (GroupBy Analysis)



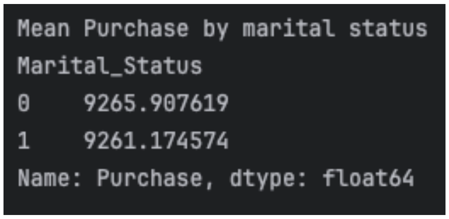
```python
# --- Grouped Summaries (GroupBy Analysis) ---
# Mean Purchase by Gender
print("\nMean Purchase by Gender")
print(df.groupby('Gender')['Purchase']. mean())

# Mean Purchase by Age
print("\nMean Purchase by Age")
print(df.groupby('Age')['Purchase']. mean())

# Mean Purchase by Marital Status
print("\nMean Purchase by marital status")
print(df.groupby('Marital_Status')['Purchase']. mean())
```



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



```
Mean Purchase by 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
Name: Purchase, dtype: float64
```

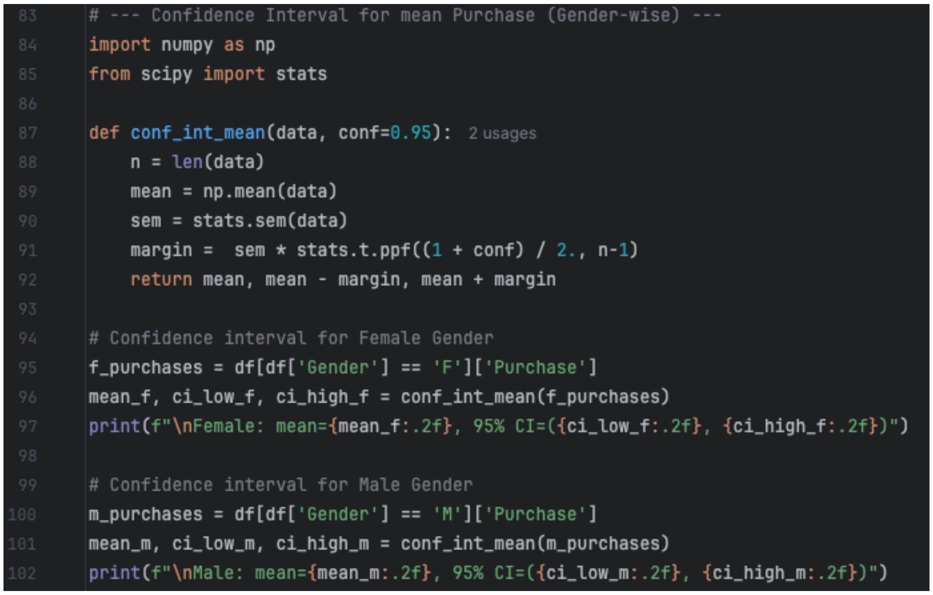


```
Mean Purchase by marital status
Marital_Status
0    9265.907619
1    9261.174574
Name: Purchase, dtype: float64
```

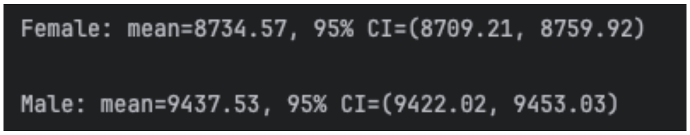
- **Result Explanation: Grouped Summaries (GroupBy Analysis)**
  - Gender: Male customers have a noticeably higher average purchase amount (Rs.9,437.53) than female customers (Rs.8,734.57). This suggests a significant gender-based difference in spending behavior.
  - Age: The highest average purchase is observed in the 51–55 age group (Rs.9,534.81), while the lowest is in the 0–17 group (Rs.8,933.46). Middle-aged groups (36–45, 46–50, 55+) also have higher average purchases, indicating that older customers tend to spend more.
  - Marital Status: There is almost no difference in average purchase amount between unmarried (Rs.9,265.91) and married (Rs.9,261.17) customers, suggesting marital status does not strongly influence purchase behavior in this dataset.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

SEGMENT 6: Confidence Interval for Mean Purchase (Gender-wise)



```
# --- Confidence Interval for mean Purchase (Gender-wise) ---
import numpy as np
from scipy import stats

def conf_int_mean(data, conf=0.95):  2 usages
    n = len(data)
    mean = np.mean(data)
    sem = stats.sem(data)
    margin =  sem * stats.t.ppf((1 + conf) / 2., n-1)
    return mean, mean - margin, mean + margin

# Confidence interval for Female Gender
f_purchases = df[df['Gender'] == 'F']['Purchase']
mean_f, ci_low_f, ci_high_f = conf_int_mean(f_purchases)
print(f"\nFemale: mean={mean_f:.2f}, 95% CI=({ci_low_f:.2f}, {ci_high_f:.2f})")

# Confidence interval for Male Gender
m_purchases = df[df['Gender'] == 'M']['Purchase']
mean_m, ci_low_m, ci_high_m = conf_int_mean(m_purchases)
print(f"\nMale: mean={mean_m:.2f}, 95% CI=({ci_low_m:.2f}, {ci_high_m:.2f})")
```



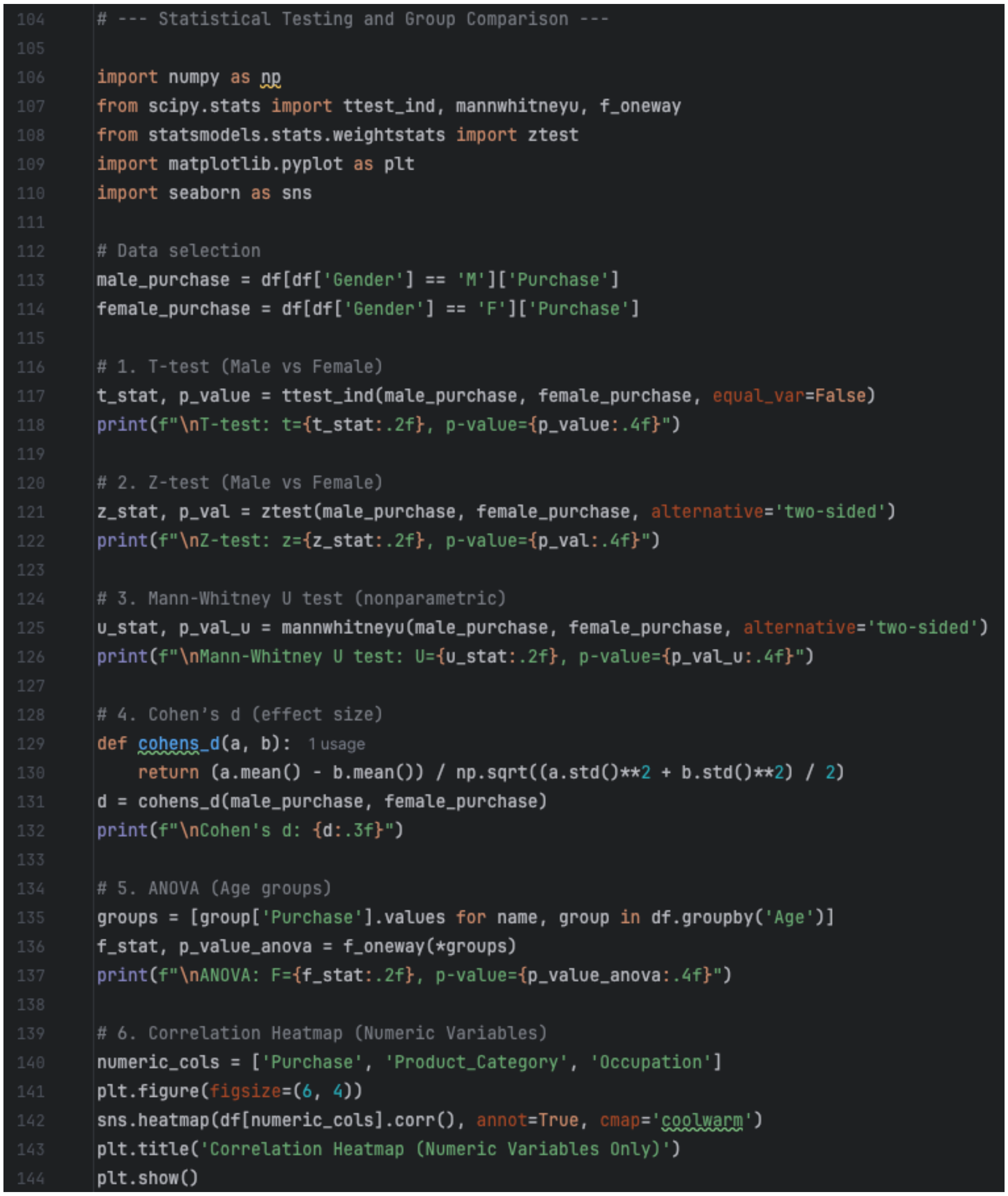
```
Female: mean=8734.57, 95% CI=(8709.21, 8759.92)

Male: mean=9437.53, 95% CI=(9422.02, 9453.03)
```

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

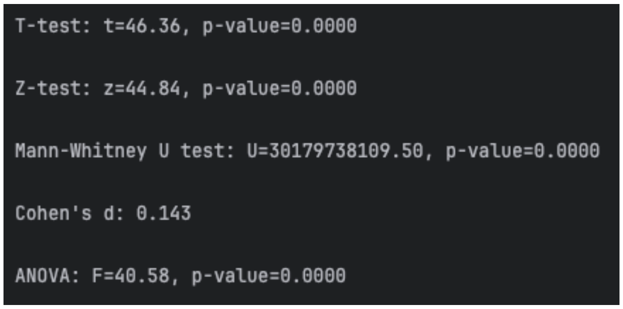
- **Code Explanation:**
  - ***Purpose of the code:***
    - We use a confidence interval as it gives a range that is likely to contain the "true mean" purchases of all customers, not just our sample, with a given confidence, where confidence is usually 95%
    - conf_int_mean() computes the mean, standard error of the mean (SEM), and then uses the "t-distribution" to find the margin error.
    - We use 95% confidence by default, but we can change it to 99%,etc.
- **Result Explanation:**
  - **Confidence Interval:**
    - The 95% confidence interval for female customers' mean purchase is Rs.8,709.21 to Rs.8,759.92, while for male customers it is Rs.9,422.02 to Rs.9,453.03.
    - These intervals do not overlap, confirming that male customers spend more on average than female customers, and the difference is statistically significant.
    - The narrow width of each interval (about +25 to -25) shows that the estimate is very precise, thanks to the large sample size.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

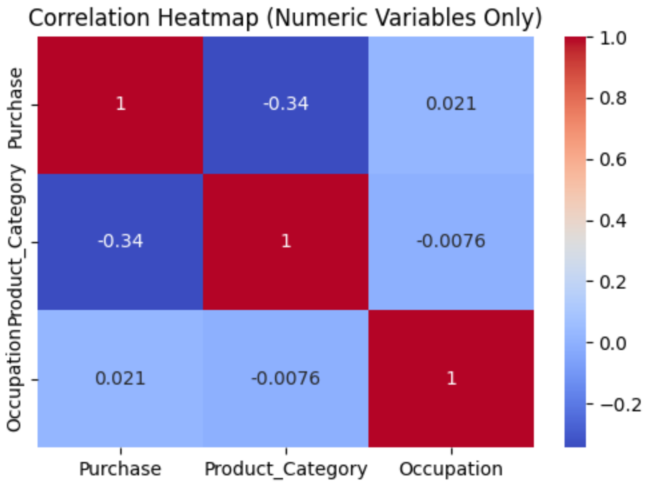
SEGMENT 7: Statistical Testing and Group Comparison



```python
# --- Statistical Testing and Group Comparison ---

import numpy as np
from scipy.stats import ttest_ind, mannwhitneyu, f_oneway
from statsmodels.stats.weightstats import ztest
import matplotlib.pyplot as plt
import seaborn as sns

# Data selection
male_purchase = df[df['Gender'] == 'M']['Purchase']
female_purchase = df[df['Gender'] == 'F']['Purchase']

# 1. T-test (Male vs Female)
t_stat, p_value = ttest_ind(male_purchase, female_purchase, equal_var=False)
print(f"\nT-test: t={t_stat:.2f}, p-value={p_value:.4f}")

# 2. Z-test (Male vs Female)
z_stat, p_val = ztest(male_purchase, female_purchase, alternative='two-sided')
print(f"\nZ-test: z={z_stat:.2f}, p-value={p_val:.4f}")

# 3. Mann-Whitney U test (nonparametric)
u_stat, p_val_u = mannwhitneyu(male_purchase, female_purchase, alternative='two-sided')
print(f"\nMann-Whitney U test: U={u_stat:.2f}, p-value={p_val_u:.4f}")

# 4. Cohen’s d (effect size)
def cohens_d(a, b):
    return (a.mean() - b.mean()) / np.sqrt((a.std()**2 + b.std()**2) / 2)
d = cohens_d(male_purchase, female_purchase)
print(f"\nCohen's d: {d:.3f}")

# 5. ANOVA (Age groups)
groups = [group['Purchase'].values for name, group in df.groupby('Age')]
f_stat, p_value_anova = f_oneway(*groups)
print(f"\nANOVA: F={f_stat:.2f}, p-value={p_value_anova:.4f}")

# 6. Correlation Heatmap (Numeric Variables)
numeric_cols = ['Purchase', 'Product_Category', 'Occupation']
plt.figure(figsize=(6, 4))
sns.heatmap(df[numeric_cols].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap (Numeric Variables Only)')
plt.show()
```



```
T-test: t=46.36, p-value=0.0000

Z-test: z=44.84, p-value=0.0000

Mann-Whitney U test: U=30179738109.50, p-value=0.0000

Cohen's d: 0.143

ANOVA: F=40.58, p-value=0.0000
```



❖ Explanation of Each Test:

- ***T-test*:**
  - Compares mean purchases between male and female customers to see if the difference is statistically significant.
- ***Z-test:***
  - Also compares means (useful for large samples), confirming the t-test results.
- ***Mann-Whitney U Test:***
  - A non-parametric test (does not assume normality), checks if there is a difference in purchase distributions between genders.
- ***Cohen's d:***
  - Measures the effect size (practical significance) of the difference in means.
    - 0.2 = small effect
    - 0.5 = medium
    - 0.8 = large
- ***ANOVA:***
  - Tests if there are significant differences in mean purchase across multiple age groups.
- ***Correlation Heatmap:***
  - Visualizes linear relationships between numeric features in the data.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

❖ Result Explanation:

- ***T-test:***
  - The t-test comparing male and female mean purchases yields t=46.36, p-value < 0.0001. This means the difference in average spending between genders is statistically significant.
- ***Z-test:***
  - The z-test returns z=44.84, p-value < 0.0001, confirming the statistical significance of the difference in means with a large sample size.
- ***Mann-Whitney U Test:***
  - U=30,179,738,109.50, p-value < 0.0001, showing the difference is significant even without assuming normal distribution.
- ***Cohen's d:***
  - Cohen's d = 0.143. This is a small effect size, meaning that while the difference is statistically significant, it is modest in practical terms.
- ***ANOVA (Age):***
  - F=40.58, p-value < 0.0001. There are significant differences in average purchases across different age groups.
- ***Correlation Heatmap:***
  - The heatmap shows very weak correlations between purchase amount and other numeric variables (product category, occupation), suggesting these features do not strongly influence spending.

Summary Table:

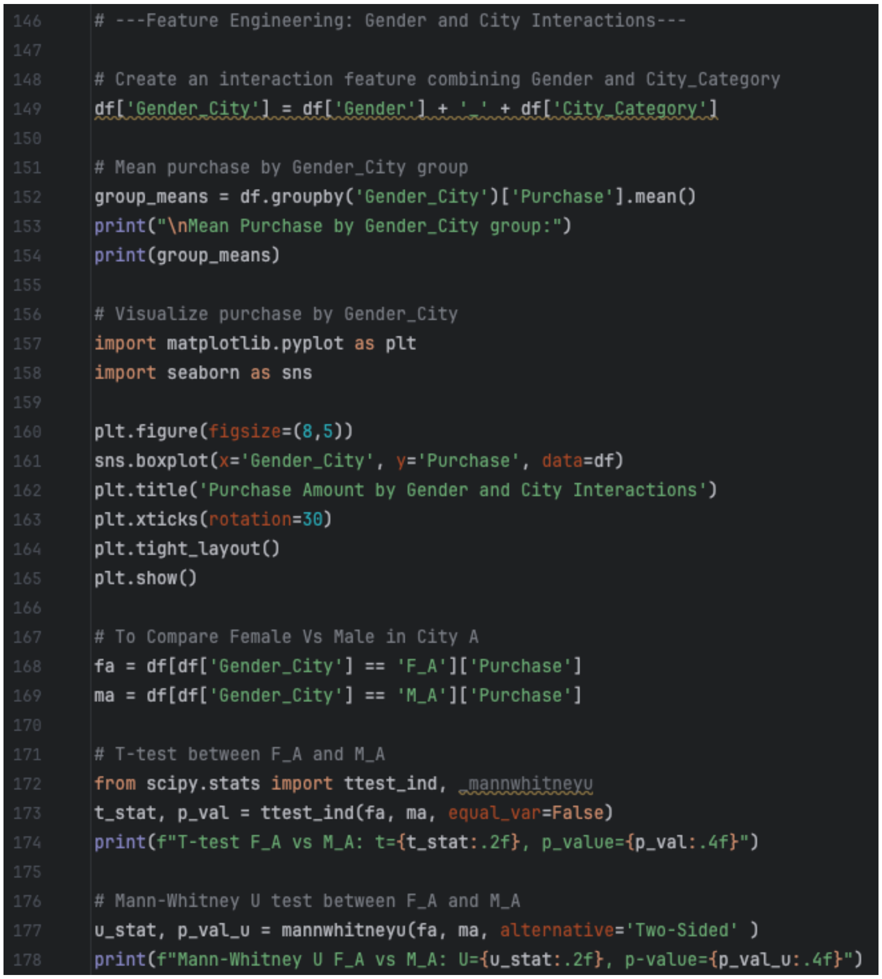
Test	Statistic/Value	p-value	Interpretation
t-test	t=46.36	<0.0001	Significant difference male vs female
z-test	Z=44.84	<0.0001	Significant, confirms the t-test
Mann-Whitney U test	U=30179738109.50	<0.0001	Significant, no normality assumption needed
Cohen's d	0.143	-	Small but real effect
ANOVA	F=40.58	<0.0001	Significant difference across the age groups

❖ Final Results and Conclusion:

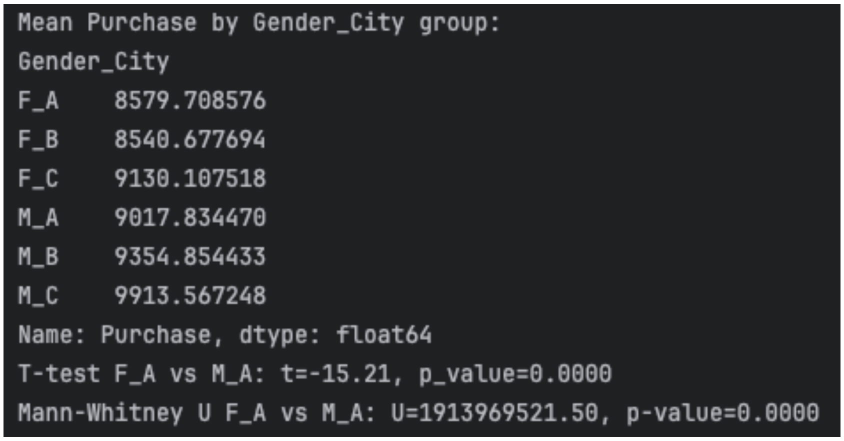
- These statistical tests together confirm that male customers spend significantly more than female customers at Walmart, although the effect size is small. Age also has a significant impact on purchase amount. Other numeric variables such as product category and occupation do not strongly predict spending behavior.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

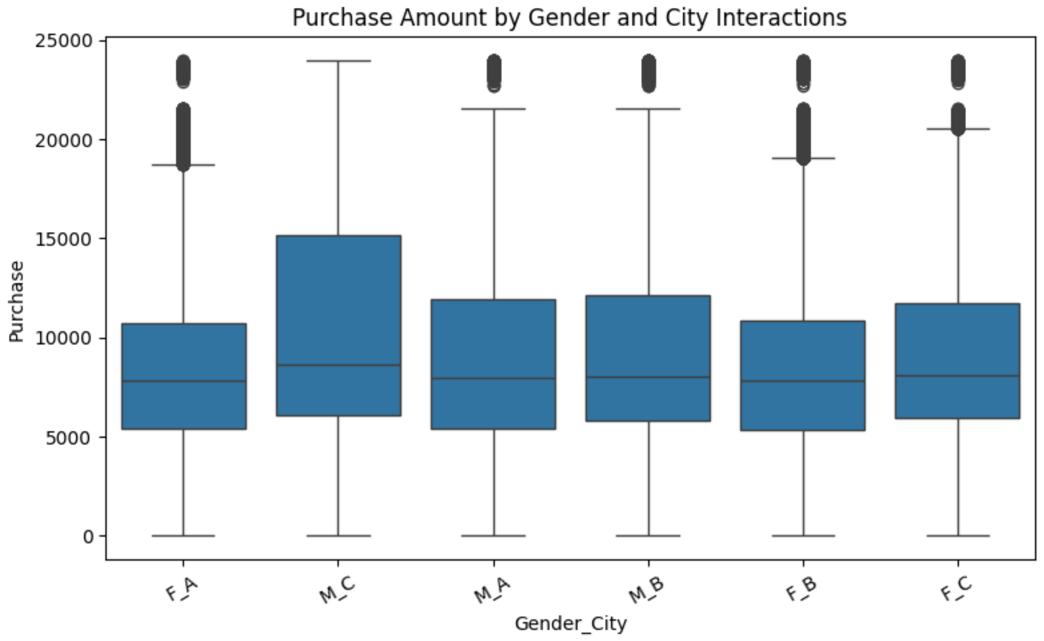
SEGMENT 8: Feature Engineering and Sub-Group Analysis



```python
# ---Feature Engineering: Gender and City Interactions---

# Create an interaction feature combining Gender and City_Category
df['Gender_City'] = df['Gender'] + '_' + df['City_Category']

# Mean purchase by Gender_City group
group_means = df.groupby('Gender_City')['Purchase'].mean()
print("\nMean Purchase by Gender_City group:")
print(group_means)

# Visualize purchase by Gender_City
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8,5))
sns.boxplot(x='Gender_City', y='Purchase', data=df)
plt.title('Purchase Amount by Gender and City Interactions')
plt.xticks(rotation=30)
plt.tight_layout()
plt.show()

# To Compare Female Vs Male in City A
fa = df[df['Gender_City'] == 'F_A']['Purchase']
ma = df[df['Gender_City'] == 'M_A']['Purchase']

# T-test between F_A and M_A
from scipy.stats import ttest_ind, mannwhitneyu
t_stat, p_val = ttest_ind(fa, ma, equal_var=False)
print(f"T-test F_A vs M_A: t={t_stat:.2f}, p_value={p_val:.4f}")

# Mann-Whitney U test between F_A and M_A
u_stat, p_val_u = mannwhitneyu(fa, ma, alternative='Two-Sided' )
print(f"Mann-Whitney U F_A vs M_A: U={u_stat:.2f}, p-value={p_val_u:.4f}")
```



```
Mean Purchase by Gender_City group:
Gender_City
F_A    8579.708576
F_B    8540.677694
F_C    9130.107518
M_A    9017.834470
M_B    9354.854433
M_C    9913.567248
Name: Purchase, dtype: float64
T-test F_A vs M_A: t=-15.21, p_value=0.0000
Mann-Whitney U F_A vs M_A: U=1913969521.50, p-value=0.0000
```



- **Code Explanation:**
  - ***Feature Engineering ("Gender_City"):***
    - By combining the Gender and City_Category columns, we can create a new feature, "Gender_City" to analyze mote granular sub-groups.
    - This allows us to compare, for example, purchases by females in city A ("F_A") vs the purchases made by males in city A ("M_A").
  - ***Group Means:***
    - Calculating mean purchases amounts for each sub-groups (F_A, F_B, F_C, M_A, M_B, M_C) highlights if certain gender-city combinations spend more.
  - ***Boxplot Visualizations:***
    - Boxplots help us to visualize the spread and outliers for each sub-groups, making patterns and differences easy to spot.
  - ***Statistical Comparison:***
    - T-test and Mann-Whitney U test between F_A and M_A test whether the difference in mean purchases is statistically significant, both under normal and non-normal assumptions.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

❖ Result Explanation:

Gender_City	Mean Purchases
F_A	8579.708576
F_B	8540.677694
F_C	9130.107518
M_A	9017.834470
M_B	9354.854433
M_C	9913.567248

➢ *Interpretation:*

- Males spend more than Females across all the city categories.
- Purchases in city C are the highest for both the genders, suggesting possible city-based effects.

❖ Statistical Tests:

➢ *T-test (F_A vs M_A):*

- T-stat = -15.21, p-value = 0.0000
- The negative t-statistic indicates that males spend more than females in city A; p-value < 0.05 means that this is statistically significant.

➢ *Mann-Whitney U (F_A vs M_A):*

- U = 1913969521.50, p-value=0.0000
- The difference is robust even if the purchase distributions are not normal.

➢ *Boxplot Interpretation:*

- Boxplots visually confirms that males in each city category tend to have higher and more varied purchase amounts than females.
- Some outliers are present, but the group differences remain clear.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

SEGMENT 9: Summary, Insights and Final Business Recommendations

- **Key Statistical Insights:**
  - ***Gender:***
    - Male customers consistently spend more on purchases than female customers across all city categories.
    - The difference is statistically significant (t-test, z-test, Mann-Whitney U), but the practical effect size is small (Cohen's d ≈ 0.14).
  - ***Age:***
    - Middle-aged and older customers (36–55+) tend to spend more on average than younger ones.
    - ANOVA confirms significant differences in mean purchase amounts across age groups.
  - ***City Category:***
    - Customers in City C spend the most, followed by B and A.
    - Both gender and city have interaction effects; City C males are the highest spenders.
  - ***Marital Status:***
    - Minimal impact on spending behavior; married and unmarried customers have nearly identical average purchase amounts.
  - ***Other Numeric Factors:***
    - Product category and occupation have weak correlations with purchase amount.

- **Visualization & Outlier Insights**
    - Boxplots revealed outliers and skewed distributions, especially in purchase amounts for higher-spending groups.
    - Heatmaps showed weak correlations between numeric features (occupation, product category) and purchase.

- **Limitations:**
    - No time variable: Data is not time-series; unable to track changing trends over time.
    - No product details: Only product categories, no information about brands or product price.
    - No additional demographic variables: For example, income, education, etc.

- Possible data bias: More males than females in the dataset may influence findings.

Business Recommendations:

- ***Targeted Marketing:***
  - Prioritize marketing campaigns for male customers, especially in City C, as they represent the highest spenders.
  - Consider loyalty programs for middle-aged and older customers who spend more on average.
- ***Female Customer Engagement:***
  - Since females spend less on average, identify barriers (through surveys or focus groups) and create offers or promotions tailored for women to increase their spend.
- ***City-Specific Promotions:***
  - Launch premium products or exclusive deals in City C to leverage higher spending behavior.
  - For City A and B, focus on customer retention and upselling strategies.
- ***Personalization:***
  - Use age and city information for personalized email offers, in-app notifications, or ad targeting.
- **Investigate Outliers:**
  - Analyze high-value transactions for potential fraud or to identify VIP customers for further engagement.

Deeper Analysis:

- Explore predictive modeling (regression/classification) to forecast future purchases or segment customers.
- Add more demographic/behavioral features if available (income, device used, online/offline, etc.).
- Consider time-based or cohort analyses to track how behavior changes.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

SEGMENT 10: Conclusion

This comprehensive analysis of Walmart's purchase data reveals clear and actionable patterns that can be leveraged to inform both immediate business strategy and longer-term planning. By examining customer demographics, city-specific behavior, and purchasing patterns, several key findings emerge that provide a blueprint for data-driven decision-making.

First, the data demonstrates that male customers consistently outspend female customers across all city categories, with the difference being not only statistically significant but also persistent across various statistical tests. This trend is most pronounced in City C, where both men and women tend to spend more, but men stand out as the highest spenders. While the overall effect size of this gender difference is modest, its consistency suggests an opportunity for tailored engagement: by targeting male customers—particularly in high-spending urban centers like City C—Walmart can amplify its revenue from its best-performing segments.

However, the story does not end with existing high spenders. The analysis also points to untapped potential among female customers and younger age groups. While these groups currently spend less on average, their large numbers present a substantial growth opportunity. By investigating possible barriers to higher spending—such as product selection, marketing tone, store experience, or promotional strategies—Walmart can develop initiatives aimed at increasing engagement and basket size among these segments. For example, personalized offers, loyalty rewards, or community engagement campaigns could be tested to drive incremental sales.

Age is another important factor: middle-aged and older customers (36–55+) tend to spend more per transaction, and ANOVA analysis confirms significant differences in purchase behavior by age. This suggests that marketing efforts can be further refined based on life stage, possibly with differentiated product bundles or targeted messaging. Interestingly, marital status appears to have little impact on spending, which means resources can be better allocated toward more influential factors like gender, age, and location. The correlation analysis also shows that numeric variables such as occupation and product category have relatively weak relationships with purchase amount, indicating that customer demographics and city-based factors are more critical drivers of spending.

Business Case: Walmart - Confidence Interval and CLT
By: Shivesh Raj Sahu

SEGMENT 11: Strategic Outlook and Recommendations

❖ **The findings from this study offer immediate guidance for Walmart's business strategy:**

- Enhance male-focused and city-specific promotions to capitalize on the highest-spending segments.
- Innovate and invest in engaging female and younger shoppers through targeted campaigns, improved store layouts, or digital experiences that better resonate with their needs.
- Leverage data-driven personalization by using demographic and location data for tailored offers and communications.
- Continue monitoring and analyzing outliers and high-value transactions to identify new customer segments or prevent fraud.

❖ **Forward Path**

While this analysis provides robust initial insights, it also highlights the importance of continuous learning. Walmart can further improve its understanding by incorporating additional variables—such as income, shopping frequency, or product-level preferences and employing predictive analytics to forecast trends. Time-based analyses and segmentation can reveal how customer behavior evolves, allowing for proactive strategy adjustments.

In conclusion, this project not only uncovers actionable patterns but also establishes a foundation for Walmart's ongoing journey toward deeper customer understanding and smarter business decisions.