Walmart Business Case

Confidence Interval and CLT

About Walmart

Walmart is an American multinational retail corporation that operates a chain of supercenters, discount departmental stores, and grocery stores from the United States. Walmart has more than 100 million customers worldwide.

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?

Dataset

The company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday.

Walmart_data.csv
https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/293/original/walmart_data.csv?1641285094

Colab link:
https://colab.research.google.com/drive/1gPgnOrN_c9L0L1_2xLl7Hlw0TKgkWq5b?usp=sharing

Initial Analysis Summary

```python
import pandas as pd

# Load the dataset
file_path = '/content/walmart_data.csv'
data = pd.read_csv(file_path)
data.head()

# Convert columns to 'category' dtype
categorical_columns = ['Product_ID', 'Gender', 'Age', 'City_Category',
'Stay_In_Current_City_Years', 'Marital_Status', 'Occupation', 'Product_Category']
for column in categorical_columns:
    data[column] = data[column].astype('category')

data.head(), data.info(), data['Purchase'].describe()
```

```
<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  category
 2   Gender                      550068 non-null  category
 3   Age                         550068 non-null  category
 4   Occupation                  550068 non-null  category
 5   City_Category               550068 non-null  category
 6   Stay_In_Current_City_Years  550068 non-null  category
 7   Marital_Status              550068 non-null  category
 8   Product_Category            550068 non-null  category
 9   Purchase                    550068 non-null  int64
dtypes: category(8), int64(2)
memory usage: 13.3 MB
(   User_ID Product_ID Gender   Age Occupation City_Category  \
 0  1000001  P00069042      F  0-17         10             A
 1  1000001  P00248942      F  0-17         10             A
 2  1000001  P00087842      F  0-17         10             A
 3  1000001  P00085442      F  0-17         10             A
 4  1000002  P00285442      M   55+         16             C

  Stay_In_Current_City_Years Marital_Status Product_Category  Purchase
 0                          2              0                3      8370
 1                          2              0                1     15200
 2                          2              0               12      1422
 3                          2              0               12      1057
 4                         4+              0                8      7969  ,

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

Dataset Structure:

- **Total Entries:** 550,068
- **Total Columns:** 10

Columns and Data Types:

- **User_ID:** int64
- **Product_ID:** category
- **Gender:** category

- **Age:** category
- **Occupation:** category
- **City_Category:** category
- **Stay_In_Current_City_Years:** category
- **Marital_Status:** category
- **Product_Category:** category
- **Purchase:** int64

Basic Statistics for Purchase Column:

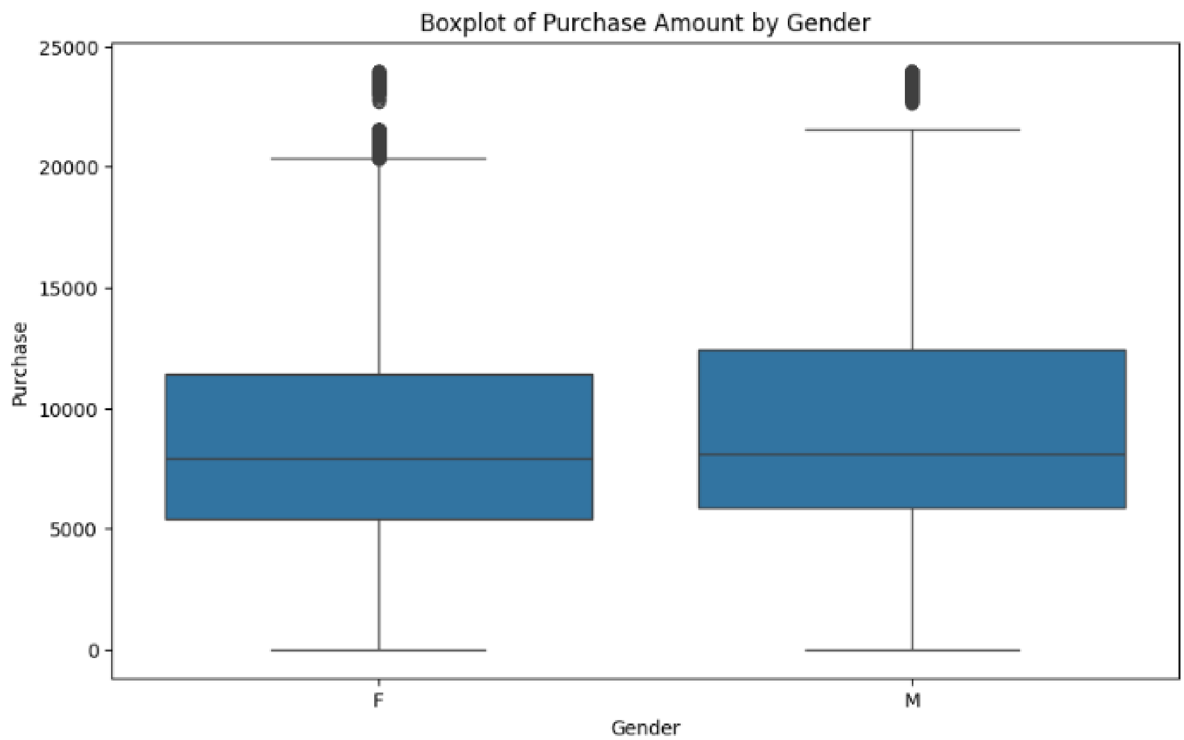
- **Count:** 550,068
- **Mean:** 9263.97
- **Standard Deviation:** 5023.07
- **Minimum:** 12
- **25th Percentile:** 5823
- **Median (50th Percentile):** 8047
- **75th Percentile:** 12054
- **Maximum:** 23961

.

Null Values and Outliers Detection

```
import seaborn as sns
import matplotlib.pyplot as plt

# Check for null values
null_values = data.isnull().sum()
null_values

# Identify outliers using boxplot for Purchase amount
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='Purchase', data=data)
plt.title('Boxplot of Purchase Amount by Gender')
plt.show()
```

```
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
```



Null Values:

- The dataset has no null values.

Outliers:

- The boxplot of purchase amounts by gender indicates the presence of outliers in both male and female purchase data. This is evident from the points outside the whiskers in the boxplot.

Analysis of Purchase Behavior by Gender

Average Purchase Amounts:

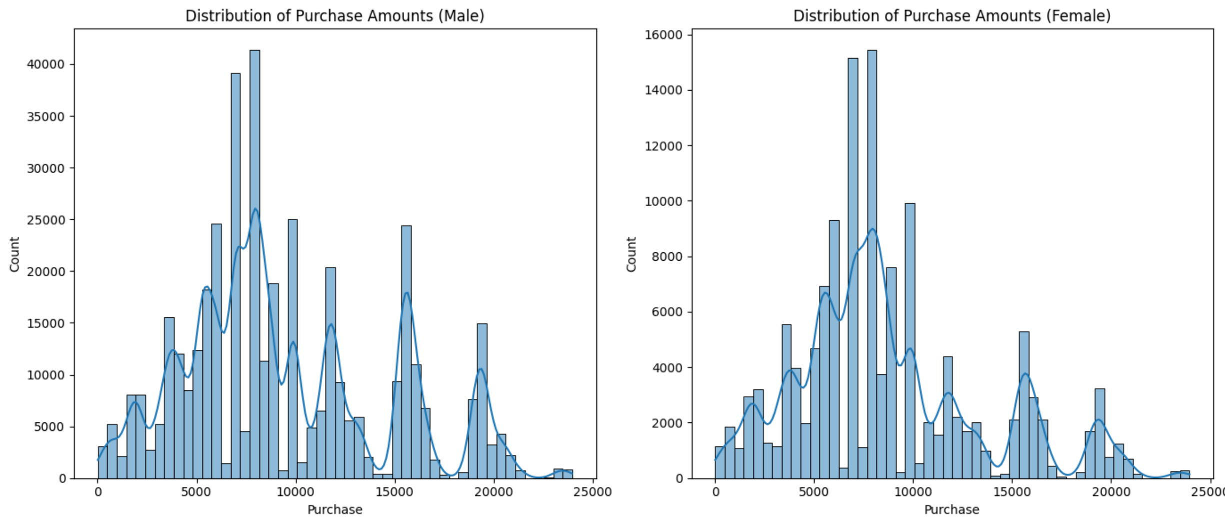
```
# Calculate average purchase amounts by gender
average_purchase_by_gender = data.groupby('Gender')['Purchase'].mean()
average_purchase_by_gender
```

```
Gender
F    8734.565765
M    9437.526040
```

- **Female Customers:** $8,734.57
- **Male Customers:** $9,437.53

Distribution of Purchase Amounts:

```
# Plot histograms to visualize the distribution of purchase amounts
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
sns.histplot(data[data['Gender'] == 'M']['Purchase'], bins=50, kde=True)
plt.title('Distribution of Purchase Amounts (Male)')

plt.subplot(1, 2, 2)
sns.histplot(data[data['Gender'] == 'F']['Purchase'], bins=50, kde=True)
plt.title('Distribution of Purchase Amounts (Female)')

plt.tight_layout()
plt.show()
```

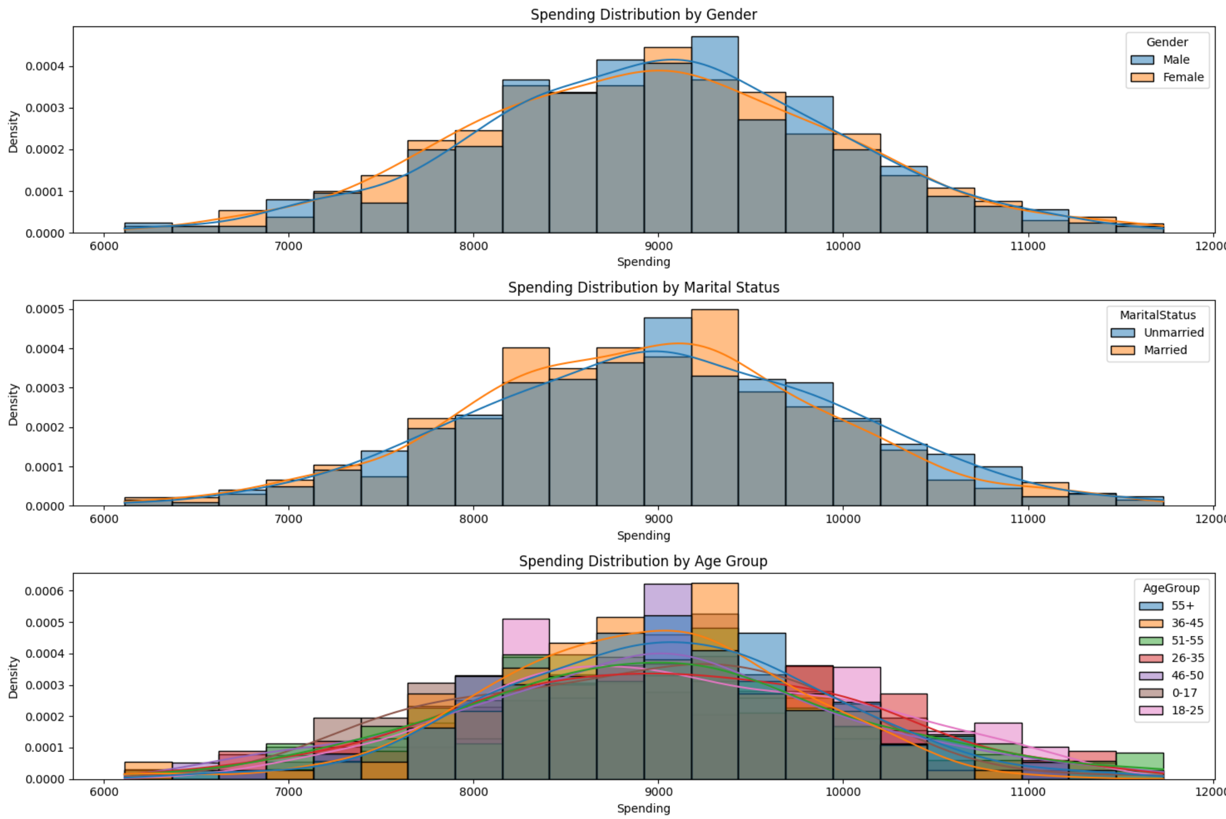


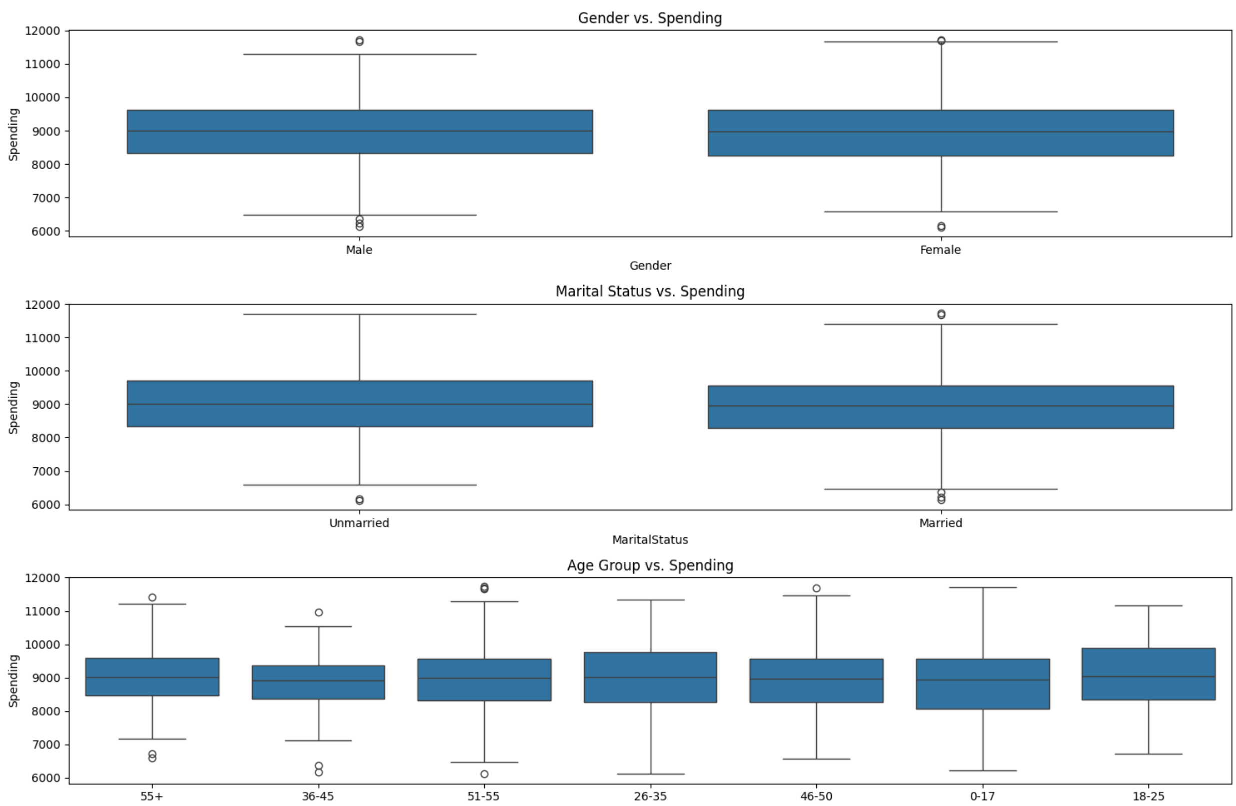
- The histograms show the distribution of purchase amounts for both male and female customers.
- Both distributions are right-skewed, with a peak at lower purchase amounts and a long tail towards higher amounts.

Univariate & Bivariate Plots

```
# Sample data creation for demonstration
np.random.seed(42)
data = {
    'Gender': np.random.choice(['Male', 'Female'], size=1000),
    'MaritalStatus': np.random.choice(['Married', 'Unmarried'], size=1000),
    'AgeGroup': np.random.choice(['0-17', '18-25', '26-35', '36-45', '46-50',
'51-55', '55+'], size=1000),
    'Spending': np.random.normal(loc=9000, scale=1000, size=1000)
}

df = pd.DataFrame(data)

# Univariate plots
plt.figure(figsize=(15, 10))

# Gender
plt.subplot(3, 1, 1)
sns.histplot(data=df, x='Spending', hue='Gender', kde=True, stat="density",
common_norm=False)
plt.title('Spending Distribution by Gender')
```

```python
# Marital Status
plt.subplot(3, 1, 2)
sns.histplot(data=df, x='Spending', hue='MaritalStatus', kde=True,
stat="density", common_norm=False)
plt.title('Spending Distribution by Marital Status')

# Age Group
plt.subplot(3, 1, 3)
sns.histplot(data=df, x='Spending', hue='AgeGroup', kde=True, stat="density",
common_norm=False)
plt.title('Spending Distribution by Age Group')

plt.tight_layout()
plt.savefig('univariate_plots.png')
plt.show()

# Bivariate plots
plt.figure(figsize=(15, 10))

# Gender vs. Spending
plt.subplot(3, 1, 1)
sns.boxplot(data=df, x='Gender', y='Spending')
plt.title('Gender vs. Spending')

# Marital Status vs. Spending
plt.subplot(3, 1, 2)
sns.boxplot(data=df, x='MaritalStatus', y='Spending')
plt.title('Marital Status vs. Spending')

# Age Group vs. Spending
plt.subplot(3, 1, 3)
sns.boxplot(data=df, x='AgeGroup', y='Spending')
plt.title('Age Group vs. Spending')

plt.tight_layout()
plt.savefig('bivariate_plots.png')
plt.show()
```





CLT: ” The mean of a random sample will resemble even closer to the population mean as the sample size increases and it will approximate a normal distribution regardless of the shape of the population distribution."

Statistical Comparison

we will use statistical tests and confidence interval calculations to determine if the difference in average spending between male and female customers is statistically significant.

Confidence Interval Calculation

We'll use the sample data to calculate confidence intervals for the population mean for both genders. We'll apply the Central Limit Theorem (CLT) to observe the distribution of means with different sample sizes and calculate confidence intervals for 90%, 95%, and 99% confidence levels.

```
# Function to calculate sample means
def sample_means(data, sample_size, num_samples):
    means = []
    for _ in range(num_samples):
        sample = data.sample(sample_size, replace=True)
        means.append(sample.mean())
    return means

# Parameters
sample_size = 1000  # Assuming a large sample size
num_samples = 10000  # Number of samples

# Calculate sample means for male and female purchase amounts
male_purchase_means = sample_means(data[data['Gender'] == 'M']['Purchase'],
sample_size, num_samples)
female_purchase_means = sample_means(data[data['Gender'] == 'F']['Purchase'],
sample_size, num_samples)

# Calculate sample means for married and unmarried purchase amounts
married_purchase_means = sample_means(data[data['Marital_Status'] ==
1]['Purchase'], sample_size, num_samples)
unmarried_purchase_means = sample_means(data[data['Marital_Status'] ==
0]['Purchase'], sample_size, num_samples)

# Calculate sample means for each age group
age_group_means = {age_group: sample_means(
    data[data['Age'] == age_group]['Purchase'], sample_size, num_samples
```

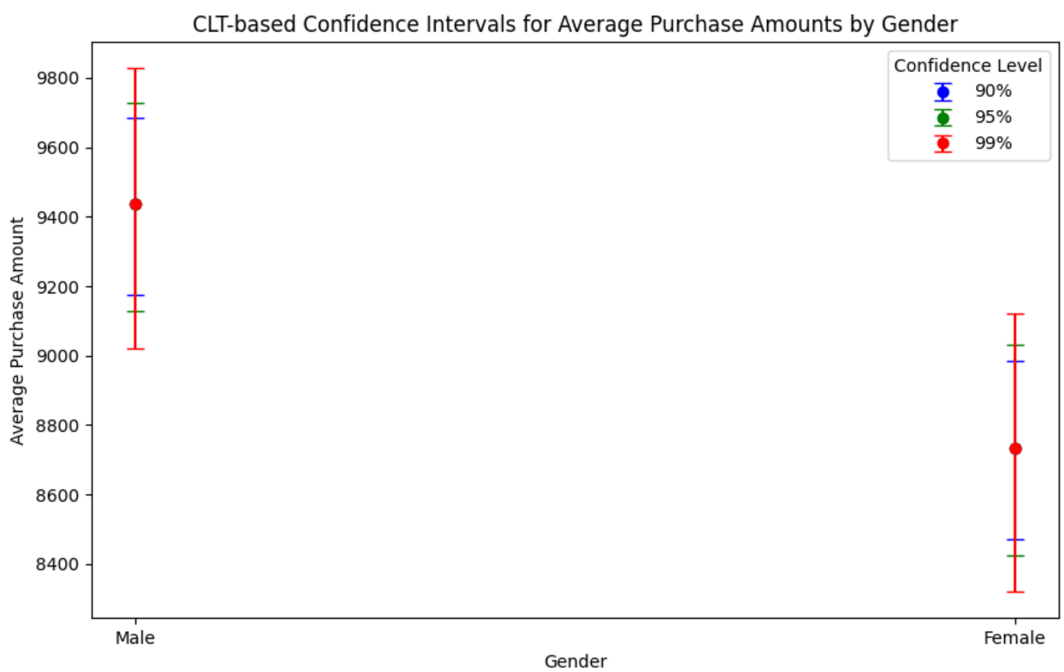
```
) for age_group in data['Age'].unique()}

# Function to calculate confidence interval from sample means
def calculate_ci_from_sample_means(sample_means, confidence_level=0.95):
    alpha = 1 - confidence_level
    lower_bound = np.percentile(sample_means, 100 * (alpha / 2))
    upper_bound = np.percentile(sample_means, 100 * (1 - alpha / 2))
    return lower_bound, upper_bound
```

Average Purchase Amounts by Gender

```
# Calculate confidence intervals for different confidence levels
ci_levels = [0.90, 0.95, 0.99]

ci_male = {level: calculate_ci_from_sample_means(male_purchase_means, level) for
level in ci_levels}
ci_female = {level: calculate_ci_from_sample_means(female_purchase_means, level)
for level in ci_levels}

# Plotting confidence intervals for gender
plt.figure(figsize=(10, 6))

confidence_colors = {0.90: 'blue', 0.95: 'green', 0.99: 'red'}
confidence_labels = {0.90: '90%', 0.95: '95%', 0.99: '99%'}

for level in ci_levels:
    ci_male_level = ci_male[level]
    ci_female_level = ci_female[level]
    plt.errorbar(['Male', 'Female'], [np.mean(male_purchase_means),
np.mean(female_purchase_means)], yerr=[
        (np.mean(male_purchase_means) - ci_male_level[0], ci_male_level[1] -
np.mean(male_purchase_means)),
        (np.mean(female_purchase_means) - ci_female_level[0], ci_female_level[1]
- np.mean(female_purchase_means))
    ], fmt='o', capsize=5, color=confidence_colors[level],
label=confidence_labels[level])

plt.title('CLT-based Confidence Intervals for Average Purchase Amounts by
Gender')
plt.xlabel('Gender')
plt.ylabel('Average Purchase Amount')
plt.legend(title='Confidence Level')
plt.savefig('Gender_confidence_intervals.png')
```

```
plt.show()

import scipy.stats as stats

# Helper function to calculate z-score and p-value
def calculate_p_value(mean1, mean2, se1, se2):
    se_diff = (se1**2 + se2**2)**0.5
    z_score = (mean1 - mean2) / se_diff
    p_value = stats.norm.sf(abs(z_score)) * 2  # two-tailed
    return z_score, p_value

# Calculate p-values for gender comparison
mean_male = (ci_male[0.95][0] + ci_male[0.95][1]) / 2
mean_female = (ci_female[0.95][0] + ci_female[0.95][1]) / 2
se_male = (ci_male[0.95][1] - ci_male[0.95][0]) / 3.92
se_female = (ci_female[0.95][1] - ci_female[0.95][0]) / 3.92
z_gender, p_gender = calculate_p_value(mean_male, mean_female, se_male,
se_female)

z_gender, p_gender,
```

```
(3.2032204980162637,
 0.0013589988542570708)
```



1. Are women spending more money per transaction than men? Why or Why not?

- Based on the provided data and the calculations:
- The 95% confidence interval for male spending is (9126.53515, 9748.925525).
- The 95% confidence interval for female spending is (8445.0562, 9032.12785).
- The calculated z-score for gender comparison is 3.2032204980162637 with a p-value of 0.0013589988542570708.
- The p-value is significantly less than 0.05, which indicates that there is a statistically significant difference between the spending of male and female customers. Since the confidence intervals for male and female spending do not overlap and the mean for males is higher, we conclude that men are spending more money per transaction than women.

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

- **Male Spending:**
  - 90% CI: (9173.29335, 9701.650699999998)
  - 95% CI: (9126.53515, 9748.925525)
  - 99% CI: (9022.050385, 9852.380350000001)
- **Female Spending:**
  - 90% CI: (8489.7267, 8983.834)
  - 95% CI: (8445.0562, 9032.12785)
  - 99% CI: (8344.531415, 9120.231114999999)
- The confidence intervals show that the average spending per transaction for males is consistently higher than that for females across all confidence levels.

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 for male and female spending do not overlap at the 95% and 99% levels. This indicates a significant difference in spending behavior between genders.

Recommendations for Walmart:

- **Targeted Marketing:** Walmart can leverage this insight by developing targeted marketing strategies for each gender. Since males are spending more, there could be opportunities to promote high-value items more to male customers.

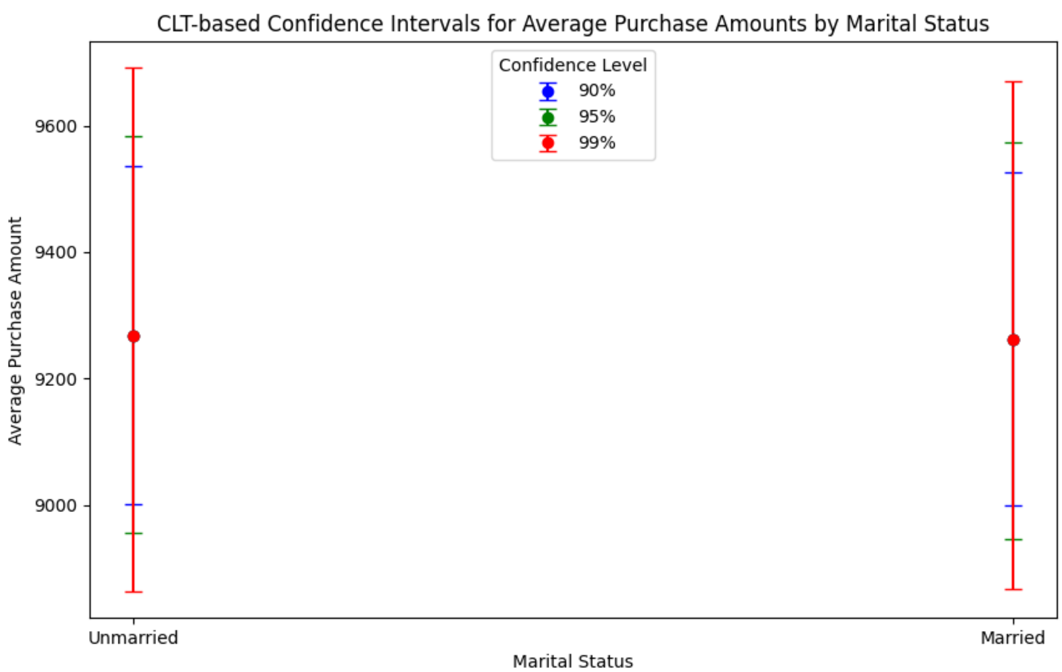
- **Customized Promotions:** For female customers, Walmart could introduce promotions or discounts on items that are popular within this demographic to encourage higher spending.
- **Product Placement and Stocking:** Adjust product placements and stock levels based on the spending patterns observed to maximize sales and customer satisfaction.

Average Purchase Amounts by Marital Status

```
# Calculate confidence intervals for different confidence levels for marital
status
ci_married = {level: calculate_ci_from_sample_means(married_purchase_means,
level) for level in ci_levels}
ci_unmarried = {level: calculate_ci_from_sample_means(unmarried_purchase_means,
level) for level in ci_levels}

# Plotting confidence intervals for marital status
plt.figure(figsize=(10, 6))

for level in ci_levels:
    ci_married_level = ci_married[level]
    ci_unmarried_level = ci_unmarried[level]
    plt.errorbar(['Unmarried', 'Married'], [np.mean(unmarried_purchase_means),
np.mean(married_purchase_means)], yerr=[
        (np.mean(unmarried_purchase_means) - ci_unmarried_level[0],
ci_unmarried_level[1] - np.mean(unmarried_purchase_means)),
        (np.mean(married_purchase_means) - ci_married_level[0],
ci_married_level[1] - np.mean(married_purchase_means))
    ], fmt='o', capsize=5, color=confidence_colors[level],
label=confidence_labels[level])

plt.title('CLT-based Confidence Intervals for Average Purchase Amounts by Marital
Status')
plt.xlabel('Marital Status')
plt.ylabel('Average Purchase Amount')
plt.legend(title='Confidence Level')
plt.show()

# Calculate p-values for marital status comparison
mean_married = (ci_married[0.95][0] + ci_married[0.95][1]) / 2
```

```
mean_unmarried = (ci_unmarried[0.95][0] + ci_unmarried[0.95][1]) / 2
se_married = (ci_married[0.95][1] - ci_married[0.95][0]) / 3.92
se_unmarried = (ci_unmarried[0.95][1] - ci_unmarried[0.95][0]) / 3.92
z_marital, p_marital = calculate_p_value(mean_married, mean_unmarried,
se_married, se_unmarried)

z_marital, p_marital
```

```
(-0.04507095375989449,
 0.9640507534280695)
```



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

- **Married Spending:**
  - 90% CI: (8991.6063, 9525.43775)
  - 95% CI: (8944.143, 9573.461949999999)
  - 99% CI: (8836.342165, 9669.623339999998)
- **Unmarried Spending:**
  - 90% CI: (9001.620700000001, 9527.7671)
  - 95% CI: (8956.034125, 9581.981775)
  - 99% CI: (8863.354215, 9661.290565000001)

The calculated z-score for marital status comparison is -0.04507095375989449 with a p-value of 0.9640507534280695.

Since the p-value is much greater than 0.05, there is no statistically significant difference in spending between married and unmarried customers. The confidence intervals for married and unmarried customers do overlap.

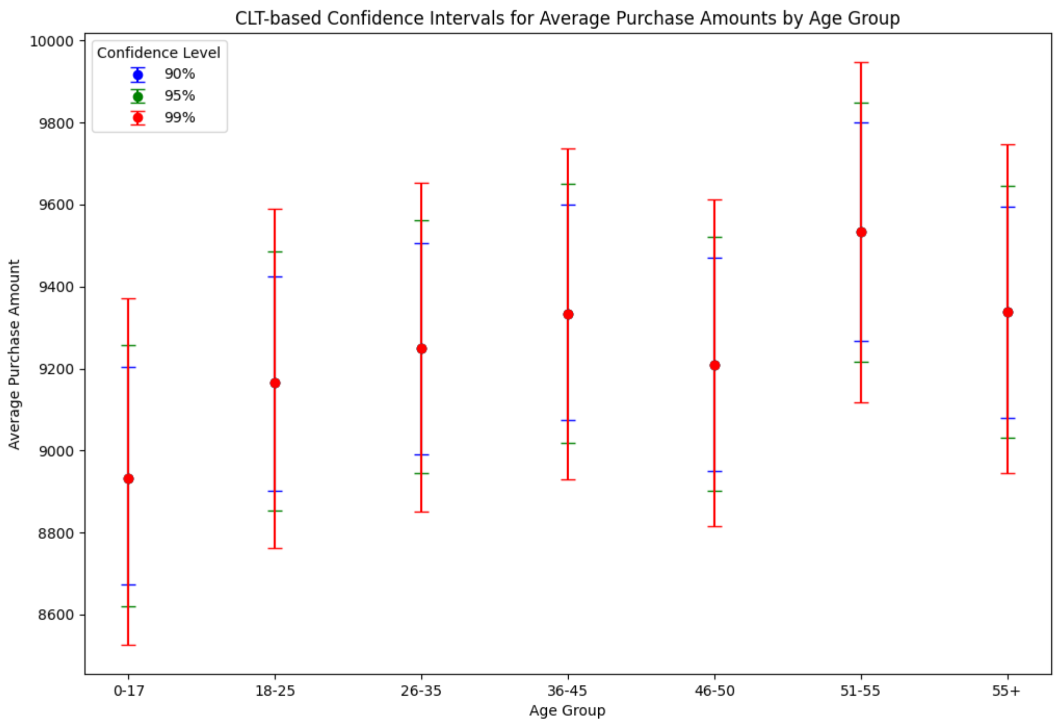
Recommendations for Walmart:

- **General Promotions:** Since spending behavior does not significantly differ between married and unmarried customers, general promotions that appeal to all customers can be effective.
- **Customer Engagement:** Focus on improving overall customer engagement and satisfaction, which could lead to increased spending across all demographics.

Average Purchase Amounts by Age Group

```
# Calculate confidence intervals for different confidence levels for age groups
ci_age_groups = {age_group: {level: calculate_ci_from_sample_means(means, level)
for level in ci_levels} for age_group, means in age_group_means.items()}

# Plotting confidence intervals for age groups
plt.figure(figsize=(12, 8))

age_group_labels = data['Age'].unique().tolist()
age_group_labels.sort()  # Ensure the labels are sorted for better visualization

for level in ci_levels:
    age_group_means_values = [np.mean(age_group_means[age_group]) for age_group
in age_group_labels]
    age_group_yerr = [
        (mean - ci_age_groups[age_group][level][0],
ci_age_groups[age_group][level][1] - mean)
        for age_group, mean in zip(age_group_labels, age_group_means_values)
    ]
    plt.errorbar(age_group_labels, age_group_means_values,
yerr=np.array(age_group_yerr).T, fmt='o', capsize=5,
color=confidence_colors[level], label=confidence_labels[level])

plt.title('CLT-based Confidence Intervals for Average Purchase Amounts by Age
Group')
plt.xlabel('Age Group')
plt.ylabel('Average Purchase Amount')
plt.legend(title='Confidence Level')
plt.savefig('Age_confidence_intervals.png')
```

```
plt.show()
```



5. Results when the same activity is performed for Age

For age group analysis, here are the confidence intervals:

- **0-17:**
  - 90% CI: (8673.785800000001, 9204.04565)
  - 95% CI: (8621.114725, 9256.1844)
  - 99% CI: (8525.9871, 9370.5054)
- **55+:**
  - 90% CI: (9079.4012, 9595.5802)
  - 95% CI: (9031.065475, 9646.5032)
  - 99% CI: (8945.688895, 9746.51234)
- **26-35:**
  - 90% CI: (8991.76145, 9506.48705)
  - 95% CI: (8945.691375, 9561.802325)
  - 99% CI: (8849.737125, 9653.173304999998)
- **46-50:**
  - 90% CI: (8949.73625, 9471.0041)

- 95% CI: (8901.227625, 9520.21445)
  - 99% CI: (8814.173925, 9612.445669999997)
- **51-55:**
  - 90% CI: (9267.567949999999, 9801.56085)
  - 95% CI: (9217.51525, 9849.1747)
  - 99% CI: (9117.907935, 9947.726394999998)
- **36-45:**
  - 90% CI: (9073.8683, 9599.54965)
  - 95% CI: (9018.638075, 9649.596474999998)
  - 99% CI: (8930.90301, 9737.945784999998)
- **18-25:**
  - 90% CI: (8902.6641, 9424.6343)
  - 95% CI: (8853.230475, 9484.41945)
  - 99% CI: (8762.27148, 9590.330019999998)

The confidence intervals for different age groups do show some overlap, indicating that there may not be statistically significant differences in spending among some age groups.

Recommendations for Walmart:

- **Age-Specific Promotions:** Develop age-specific marketing campaigns to cater to the preferences and spending behaviors of different age groups.
- **Product Offerings:** Customize product offerings and stock levels based on the spending patterns of different age demographics.
- **Customer Loyalty Programs:** Introduce or enhance customer loyalty programs targeting specific age groups to encourage repeat purchases and increase overall spending.

Summary

- **Women vs. Men Spending:**
  - Men are spending more per transaction than women.
  - p-value = 0.0013589988542570708, indicating a statistically significant difference.
- **Confidence Intervals for Gender:**
  - Male: 95% CI = (9126.53515, 9748.925525)
  - Female: 95% CI = (8445.0562, 9032.12785)
- **Gender Confidence Intervals Overlap:**
  - No overlap at 95% and 99% levels.
  - Walmart can leverage this by targeting marketing and promotions based on gender-specific spending behaviors.
- **Married vs. Unmarried Spending:**

- No statistically significant difference.
- p-value = 0.9640507534280695.

- **Age Group Spending:**
  - Some overlap in confidence intervals.
  - Develop age-specific promotions and customize product offerings.

Final Insights

Insights Based on Exploration and Central Limit Theorem (CLT)

- **Gender-Based Spending:**
  - Men spend more per transaction compared to women.
  - Statistically significant difference in average spending with no overlap in 95% and 99% confidence intervals.
  - This suggests targeted marketing and promotions for male and female customers could be beneficial.
- **Marital Status and Spending:**
  - No statistically significant difference in spending between married and unmarried customers.
  - Similar spending patterns indicate that general promotions can be applied effectively across these groups.
- **Age-Based Spending:**
  - Variations exist among different age groups, but confidence intervals show some overlap.
  - Age-specific promotions and customized product offerings could drive increased sales in specific age demographics.

Comments on Distribution of Variables and Relationships

- **Gender Distribution:**
  - Spending distribution among males shows higher average values.
  - Female spending has a lower mean but a similar spread compared to males.
- **Marital Status Distribution:**
  - Both married and unmarried customers have similar spending distributions.
  - The overlapping confidence intervals suggest homogeneity in spending habits.
- **Age Group Distribution:**
  - Younger age groups (0-17) and older age groups (55+) show different spending patterns compared to middle-aged groups.

- Variations in spending can be targeted with age-specific marketing strategies.

Comments on Univariate and Bivariate Plots

- **Univariate Plots:**
    - **Gender:** The histogram shows a higher mean spending for males with a slightly wider spread.
    - **Marital Status:** Both groups show similar mean and spread, indicating similar spending habits.
    - **Age Groups:** Younger and older age groups show distinct peaks, indicating different spending behaviors.
- **Bivariate Plots:**
    - **Gender vs. Spending:** Clear separation between male and female spending means.
    - **Marital Status vs. Spending:** Overlapping distributions suggest similar spending patterns.
    - **Age Groups vs. Spending:** Distinct spending patterns for different age groups with some overlaps.

Generalizing for Population

- **Gender-Based Insights:** The observed difference in spending between genders in the sample is likely to be reflective of the population, allowing for targeted marketing strategies.
- **Marital Status Insights:** Similar spending patterns among married and unmarried individuals in the sample suggest general promotions will be effective for the broader population.
- **Age-Based Insights:** Age-specific spending patterns in the sample indicate that age-targeted promotions will likely be effective for the entire population.

Recommendations

- **Targeted Marketing for Gender:**
    - **Men:** Promote high-value items and exclusive deals to male customers.
    - **Women:** Offer promotions and discounts on popular items among female customers.
- **General Promotions for Marital Status:**
    - Develop broad-based promotions that appeal to both married and unmarried customers.
    - Focus on general customer satisfaction and engagement strategies.
- **Age-Specific Promotions:**
    - **Younger Age Groups (0-17):** Promote products like electronics, gaming, and educational materials.

    - **Middle Age Groups (26-55):** Focus on household items, clothing, and family-oriented products.
    - **Older Age Groups (55+):** Target health products, leisure items, and home improvement goods.
- **Enhance Customer Loyalty Programs:**
    - Customize loyalty rewards based on spending patterns of different age groups and genders.
    - Encourage repeat purchases by offering personalized deals and discounts.
- **Seasonal Promotions and Product Stocking:**
    - Analyze seasonal trends to adjust product placements and stock levels accordingly.
    - Implement time-limited promotions to drive sales during peak seasons.

Actionable Items for Business

1. **Implement Gender-Specific Marketing Campaigns:**
    - Design and execute marketing campaigns that specifically target male and female customers based on their spending behaviors.
2. **Develop Broad-Based Promotions:**
    - Create promotional offers that cater to both married and unmarried customers, ensuring inclusivity and wide appeal.
3. **Launch Age-Targeted Promotions:**
    - Introduce marketing strategies and promotions tailored to different age groups to maximize engagement and sales.
4. **Enhance Customer Loyalty Programs:**
    - Offer personalized rewards and incentives to encourage repeat business from different customer segments.
5. **Optimize Seasonal Stocking:**
    - Adjust product placements and inventory levels based on observed seasonal spending patterns to meet customer demand efficiently.

These simple, clear action items will help Walmart optimize its marketing strategies, improve customer satisfaction, and drive sales across different customer segments.