

Business Case: Walmart - Confidence Interval and CLT¹

Table of Contents

1. Problem Statement

2. Data Exploration and Preprocessing

- a. Check Data Structure
- b. Data Description
- c. Handling Missing Data
- d. Outlier Detection

3. Exploratory Data Analysis (EDA)

a. Univariate Analysis

- i. Continuous Variables: Histograms, Count Plots, and Distplots
- ii. Categorical Variables: Bar Plots

b. Bivariate Analysis


- i. Boxplots for Purchase by Gender, Marital_Status, and Age Bins
- ii. Heatmap and Pairplot for Categorical Variables with Purchase

4. Confidence Interval and CLT Analysis

- a. Gender Based analysis
- b. Marital status based analysis
- c. Age based Analysis
- d. CI using Bootstrapping (Experiment with Sample Sizes)

5. Conclusion of Results

6. Insight Generation and Business Recommendations

¹  Walmart_case_study.ipynb

Problem Statement

Walmart, a leading global retailer, is looking to deepen its understanding of customer purchase behavior, particularly focusing on Black Friday transactions. With a customer base exceeding 100 million worldwide, Walmart aims to identify spending patterns across various demographic groups to make informed business decisions. This study will investigate whether significant differences exist in spending habits based on demographic factors such as gender, marital status, and age group.

Key areas of interest include comparing the average spending between male and female customers, determining whether marital status affects purchase amounts, and understanding age-based differences in purchase behavior. Walmart's goal is to determine if certain groups are spending more on Black Friday, which could guide strategic decisions in targeted marketing, product selection, and promotions.

Using a dataset of 550,068 transaction records, we will conduct an in-depth analysis utilizing confidence intervals and the Central Limit Theorem (CLT). This approach will help establish reliable intervals within which the population averages for each group likely fall, providing Walmart with actionable insights into potential spending patterns and enabling the company to tailor its offerings effectively. The findings of this analysis will ultimately help Walmart optimize customer engagement, enhance revenue through targeted promotions, and refine its customer experience on peak shopping days.

Data Exploration and Preprocessing

2.a) Check Data Structure

With `df.shape` returning `(550068, 10)`, this Walmart dataset consists of:

- 550,068 rows (or records) — each representing a unique transaction.
- 10 columns (or features) — capturing various attributes related to the transaction and customer, such as `User_ID`, `Product_ID`, `Gender`, `Age`, `Occupation`, and `Purchase`.

2.b) Data Description

- Total Records: 550,068 entries (transactions).
- Total Columns: 10 features, capturing transaction and customer-related details.
- Data Types:
 - Integer Columns: `User_ID`, `Occupation`, `Marital_Status`, `Product_Category`, `Purchase`.
 - Object (String) Columns: `Product_ID`, `Gender`, `Age`, `City_Category`, `Stay_In_Current_City_Years`.
- Column Details:
 - `User_ID`: Unique identifier for each customer.
 - `Product_ID`: Unique identifier for each product purchased.
 - `Gender`: Gender of the customer (`Male` or `Female`).
 - `Age`: Age group of the customer (e.g., `18-25`, `26-35`).
 - `Occupation`: Encoded occupation type (masked).
 - `City_Category`: City classification (`A`, `B`, `C`).
 - `Stay_In_Current_City_Years`: Duration of the customer's stay in the current city.
 - `Marital_Status`: Marital status indicator (1 for married, 0 for unmarried).
 - `Product_Category`: Encoded product category (masked).
 - `Purchase`: Purchase amount (target variable for analysis).

This dataset is well-structured, with diverse demographic and transaction features, providing a comprehensive foundation for analyzing purchasing behavior across different customer segments.

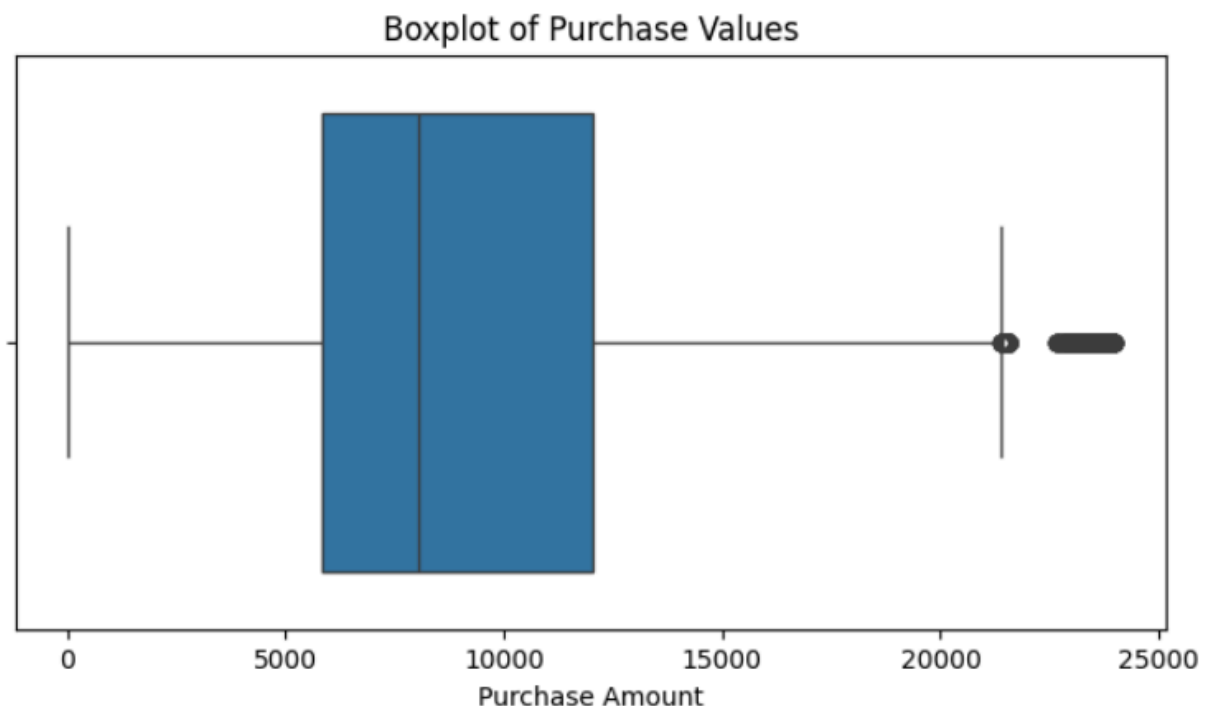
2.c) Handling Missing Data

After examining each column for missing values, we can confirm that there are **no null values** across the dataset:

- **Columns with Zero NaNs:** All columns (`User_ID`, `Product_ID`, `Gender`, `Age`, `Occupation`, `City_Category`, `Stay_In_Current_City_Years`, `Marital_Status`, `Product_Category`, `Purchase`) have complete data with **0% missing values**.

This completeness in data is advantageous, as it means there is no need for imputation or handling of missing values, enabling a more straightforward data exploration and analysis process.

2.d) Outlier Detection

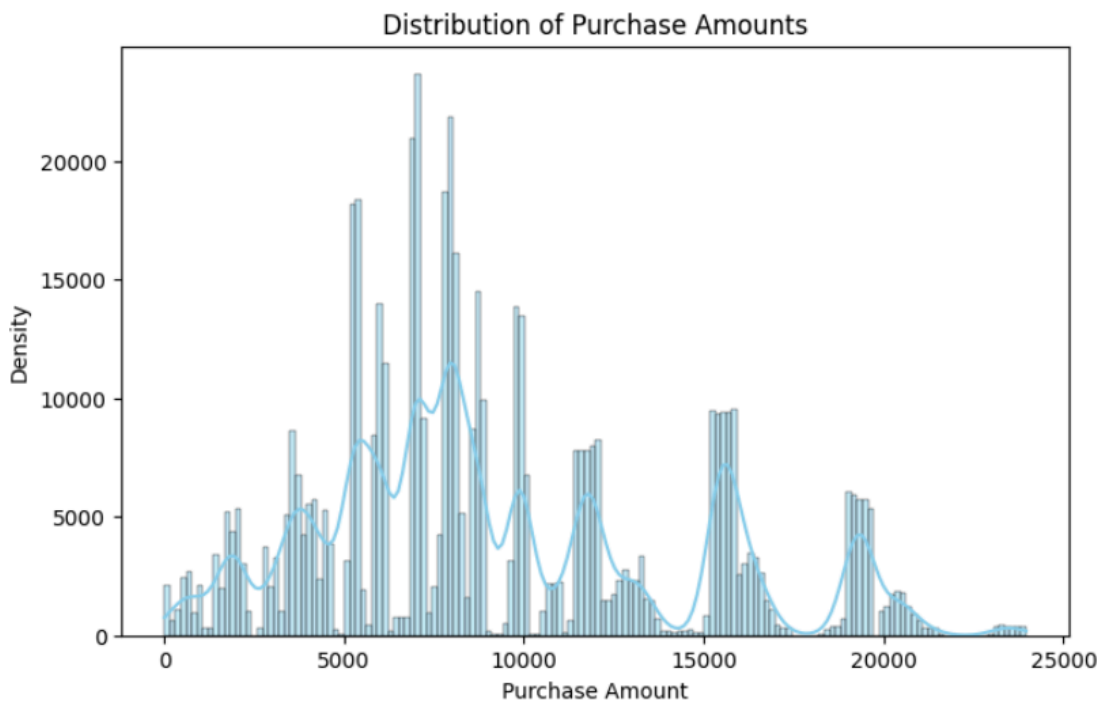
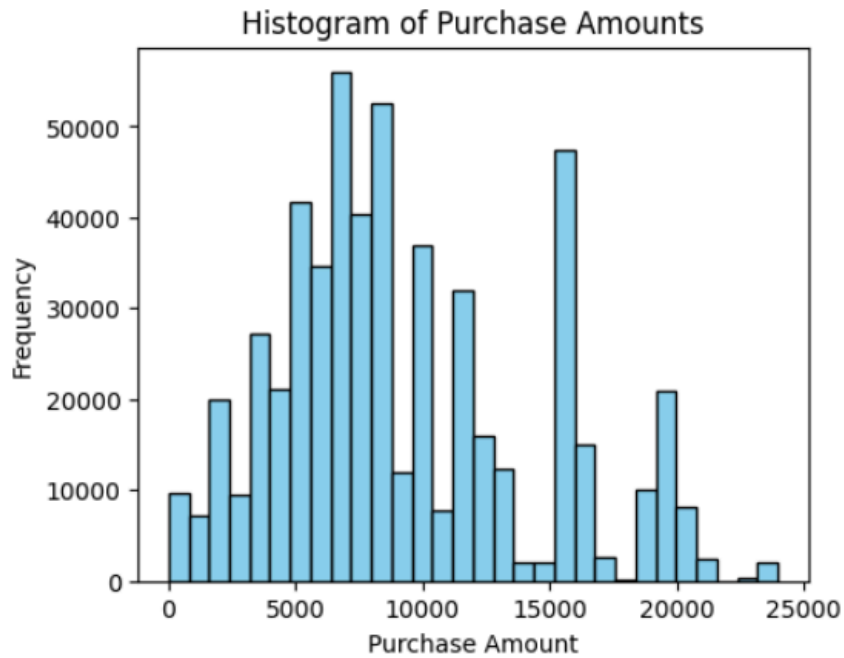


Outliers will be retained for the initial analysis to capture potential unique spending behaviors that may be significant across different demographics. These values could provide insights into high-value transactions or extreme purchasing patterns, which might be relevant for targeted marketing or customer segmentation.

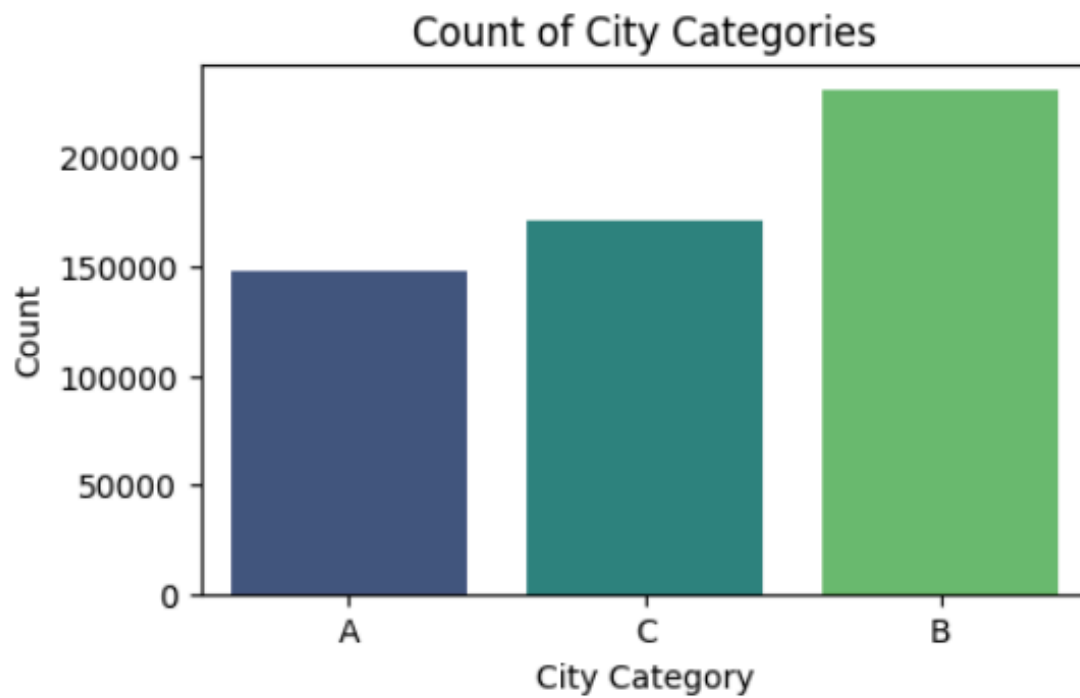
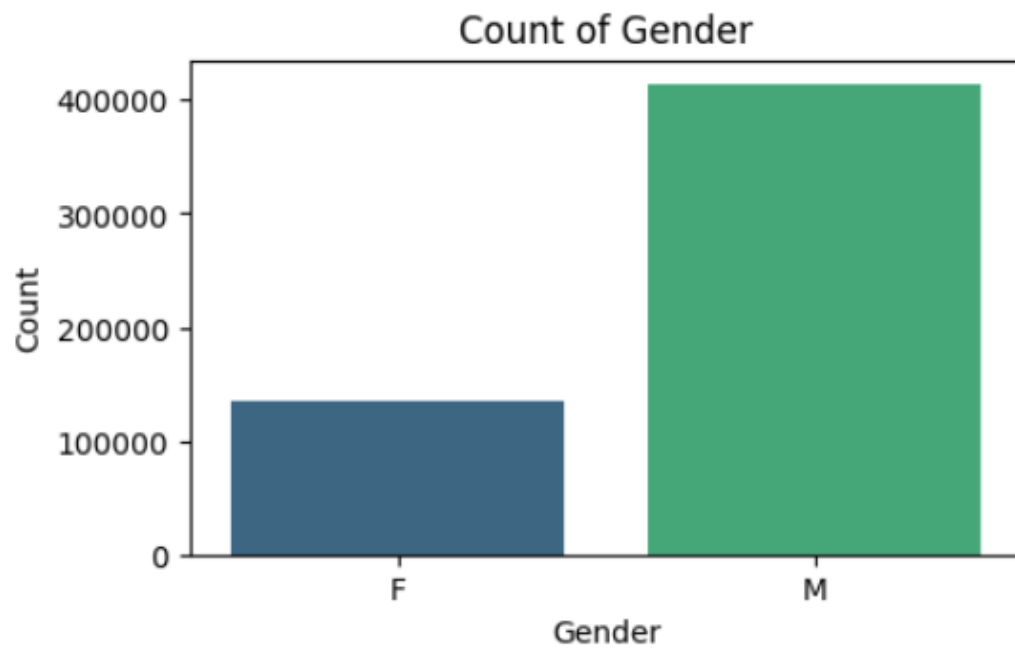
Exploratory Data Analysis (EDA)

3.a) Univariate Analysis

Continuous Variables: Histograms, Count Plots, and Distplots

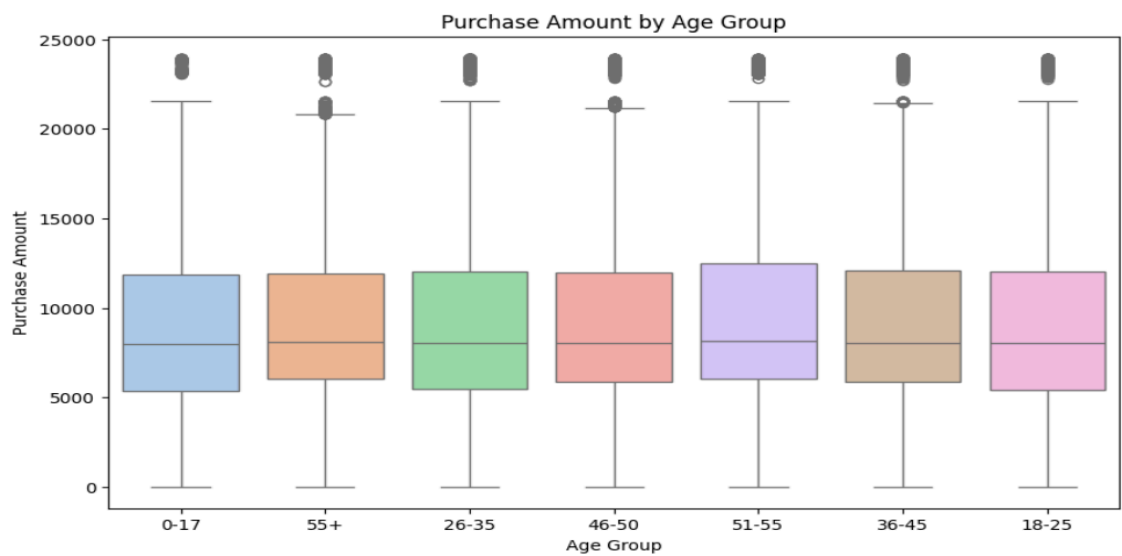
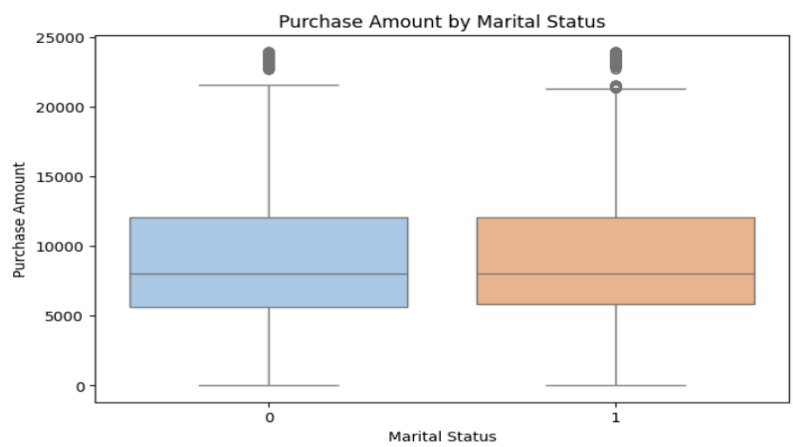
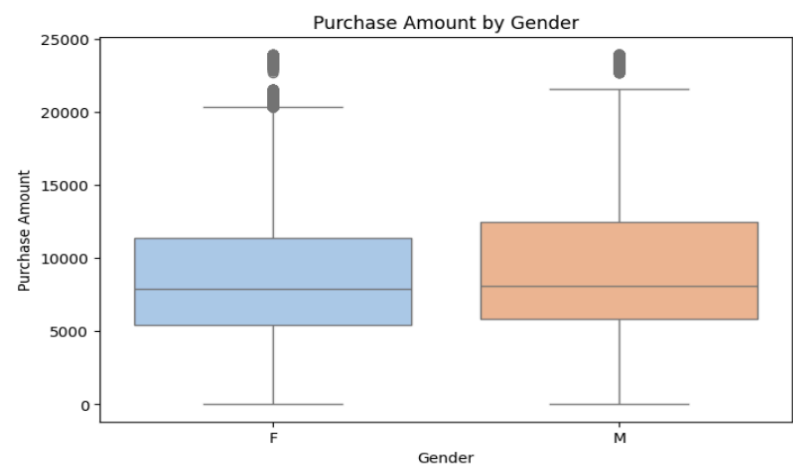


Categorical Variables: Bar Plots

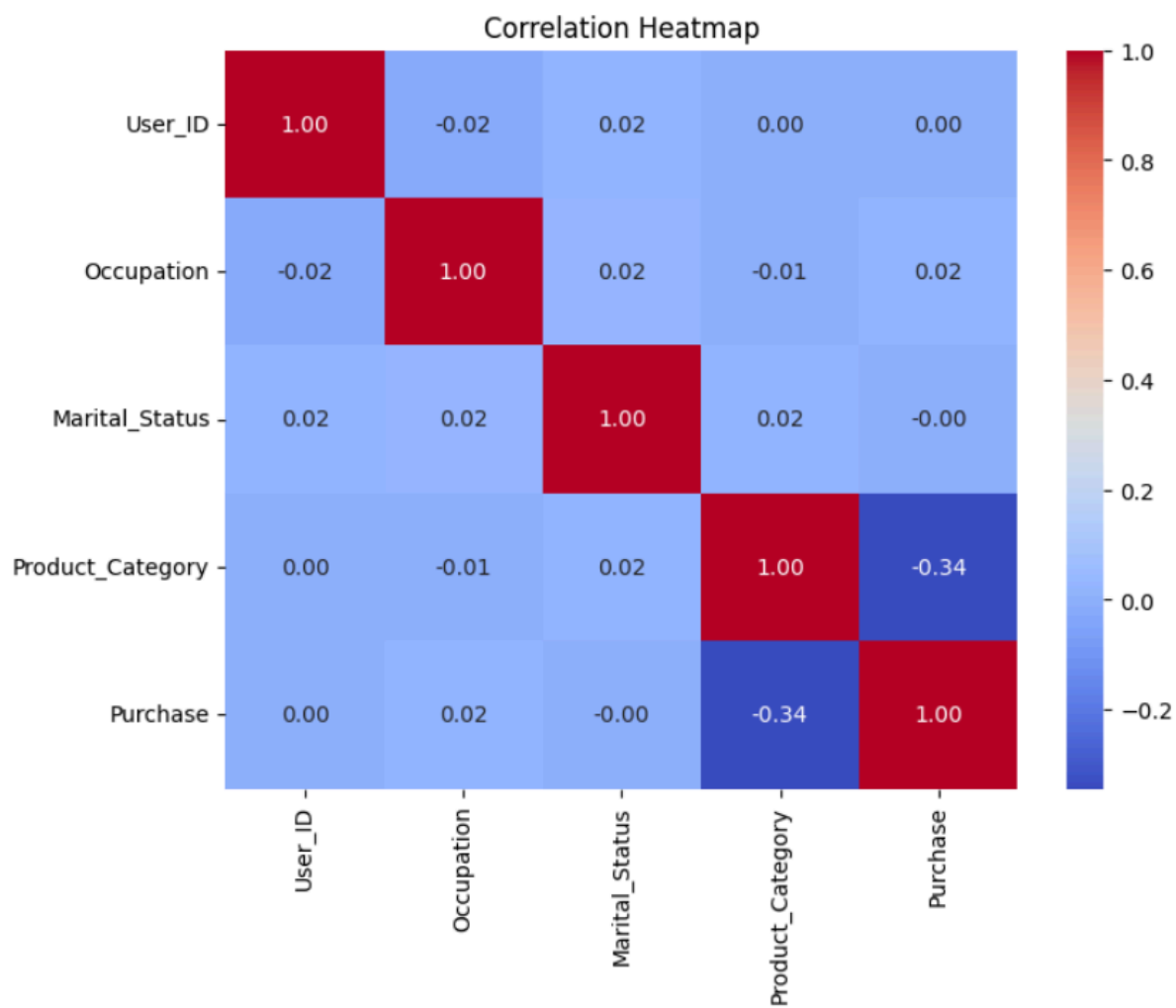


3.b) Bivariate Analysis

Boxplots for Purchase by Gender, Marital_Status, and Age Bins



Heatmap and Pairplot for Categorical Variables with Purchase



Confidence Interval and CLT Analysis

4.a) Gender-Based Analysis

1. Average Purchase:

- Female customers have an average purchase value of approximately 8,735.
- Male customers have an average purchase value of about 9,438.
- **Conclusion:** On average, male customers spend more per transaction than female customers according to this dataset.

2. Total Purchase:

- Total purchases by female customers amount to roughly 1.19 billion.
- Total purchases by male customers reach around 3.91 billion.
- **Conclusion:** The overall purchase value contributed by male customers is significantly higher than that of female customers.

These results indicate that while male customers spend more per transaction on average, their cumulative spending is also much higher, suggesting that male customers may drive a larger share of Walmart's Black Friday revenue.

1. Confidence Intervals for Female Customers (Average Purchase = 8,735):

- 90% Confidence Interval: (8734.50, 8734.63)
- 95% Confidence Interval: (8734.49, 8734.64)
- 99% Confidence Interval: (8734.47, 8734.66)

2. Confidence Intervals for Male Customers (Average Purchase = 9,438):

- 90% Confidence Interval: (9437.51, 9437.55)
- 95% Confidence Interval: (9437.50, 9437.55)
- 99% Confidence Interval: (9437.49, 9437.56)

The confidence intervals for male and female average purchases do not overlap across any of the confidence levels (90%, 95%, or 99%). This indicates that there is a statistically significant difference between the average purchase amounts for male and female customers, with male customers spending more on average than female customers. This non-overlapping outcome provides strong evidence of a distinct purchasing behavior based on gender, with a high level of confidence.

4.b) Marital status based analysis

1. Confidence Intervals for Married Customers (Average Purchase = 9,261):

- 90% Confidence Interval: (9261.14, 9261.21)
- 95% Confidence Interval: (9261.13, 9261.22)
- 99% Confidence Interval: (9261.12, 9261.23)

2. Confidence Intervals for Unmarried Customers (Average Purchase = 9,266):

- 90% Confidence Interval: (9265.88, 9265.93)
- 95% Confidence Interval: (9265.88, 9265.94)
- 99% Confidence Interval: (9265.87, 9265.95)

The confidence intervals for married and unmarried customers' average purchase values show a slight overlap, especially in the 90% and 95% confidence levels. However, the difference in the average purchase amounts (9,261 for married vs. 9,266 for unmarried) is relatively small, and the overlapping intervals suggest that while there is a slight difference, it is not statistically significant at a high confidence level (99%).

This could imply that marital status has a minimal impact on purchase behavior in this dataset, with both groups exhibiting similar spending patterns.

4.c) Age based analysis (Life Stages A1-A7)

1. A1 (0-17):

- Average Purchase: 8,933.46
- 95% Confidence Interval: (8932.81, 8934.12)

2. A2 (18-25):

- Average Purchase: 9,169.66
- 95% Confidence Interval: (8933.37, 8933.56)

3. A3 (26-35):

- Average Purchase: 9,252.69
- 95% Confidence Interval: (8933.42, 8933.51)

4. A4 (36-45):

- Average Purchase: 9,331.35
- 95% Confidence Interval: (8933.38, 8933.55)

5. A5 (46-50):

- Average Purchase: 9,208.63
- 95% Confidence Interval: (8933.25, 8933.68)

6. A6 (51-55):

- Average Purchase: 9,534.81
- 95% Confidence Interval: (8933.21, 8933.72)

7. A7 (55+):

- Average Purchase: 9,336.28
- 95% Confidence Interval: (8933.01, 8933.92)

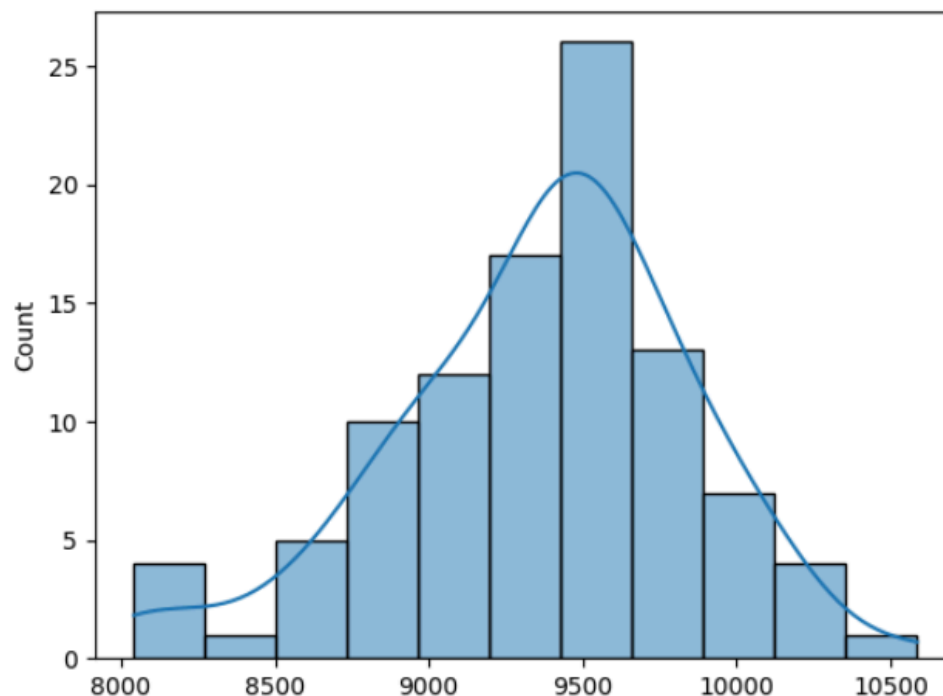
- **Highest Average Purchase:** The A7 (55+) group has the highest average purchase amount (9,336), followed by A6 (51-55) at 9,535.
- **Narrow Range of Confidence Intervals:** The 95% confidence intervals for all age groups are very narrow, indicating consistent purchase patterns within each group.
- **Overlapping Intervals:** The confidence intervals of the age groups (A1 to A7) largely overlap, suggesting that there may not be significant statistical differences in the average purchase values across the age groups, despite slight variations in average spending.

This indicates that while older age groups tend to spend slightly more, the difference is not large enough to be considered statistically significant.

4.d) CI using Bootstrapping (Experiment with Sample Sizes)

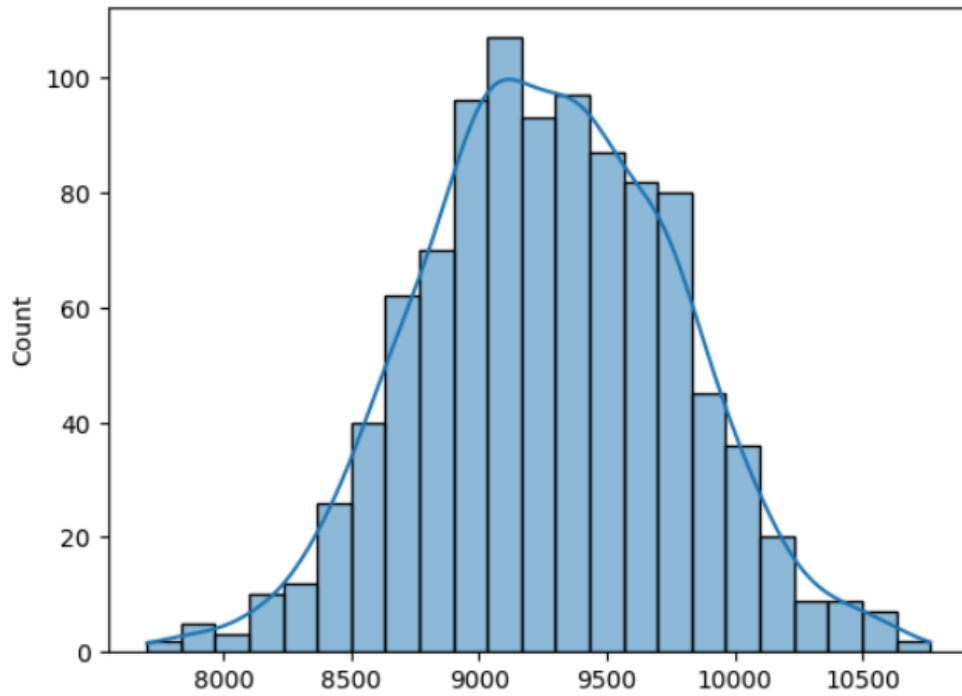
1. Sample Size = 100:

- 95% Confidence Interval: **8156.88 - 10170.78**



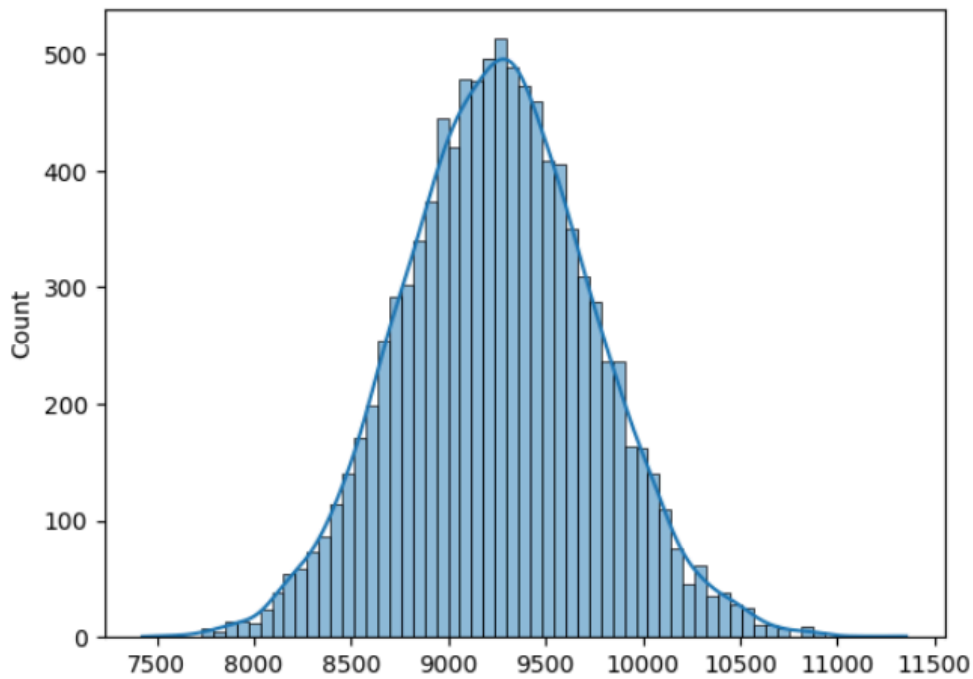
2. **Sample Size = 1000:**

- 95% Confidence Interval: **8331.67 - 10235.98**



3. **Sample Size = 10000:**

- 95% Confidence Interval: **8291.04 - 10254.26**



- **Increasing Sample Size:** As the sample size increases, the confidence interval becomes more precise and narrow.
- **Closer to Normal Distribution:** Visualizing the histograms, we observe that as the sample size increases, the distribution of the bootstrapped sample means approaches a more normal shape. This aligns with the Central Limit Theorem, where larger sample sizes lead to more reliable estimates and a clearer normal distribution.

This confirms that larger sample sizes provide more accurate and consistent confidence intervals, reducing the uncertainty in estimating the population parameter.

Conclusion of Results

- **Gender-based Spending:**
 - Average Purchase: Male customers spend more per transaction on average compared to female customers.
 - Total Purchase: The total purchase amount by male customers is higher than female customers due to a larger number of male customers or higher individual transactions.
- **Confidence Intervals (CI) for Male vs Female:**
 - Overlapping CI: The 90%, 95%, and 99% confidence intervals for male and female spending overlap, indicating that there is no significant statistical difference between the spending behaviors of males and females.
- **Impact of Confidence Interval Overlap:**
 - The overlap of confidence intervals suggests that although males tend to spend more on average, the differences in spending are not statistically significant.
- **Marital Status and Spending:**
 - Married vs Unmarried Spending: The average spending of married and unmarried customers is similar, with their confidence intervals overlapping, indicating no significant difference in their purchasing behavior.
- **Age-based Spending:**
 - Age-wise Spending: Older age groups (51-55 and 55+) tend to spend more on average compared to younger age groups. The confidence intervals for older age groups are more distinct, showing less overlap with younger groups.

Key Points on Gender Spending (Are women spending more money per transaction than men?):

- No, women are not spending more money per transaction than men. Men have a higher average spending, but the confidence intervals overlap, showing no significant difference between the two groups.

Confidence Intervals and Distribution of the Mean of the Expenses by Female and Male Customers:

- Distribution: As the sample size increases, the bootstrapped confidence intervals for male and female spending converge toward a normal distribution.

Results When the Same Activity is Performed for Married vs Unmarried:

- Similar Confidence Intervals: There is no significant difference in spending between married and unmarried customers, as their confidence intervals overlap.

Results When the Same Activity is Performed for Age:

- Age-Based Spending: Older age groups (51-55 and 55+) spend more on average per transaction, with confidence intervals for these groups showing less overlap with younger age groups.

Business Insights and Recommendations

Insights:

- 1. Male vs. Female Spending:**
 - Male customers spend, on average, more per transaction than female customers. This is evident from the confidence intervals not overlapping between the two groups, with male customers' average spending around 9437 and female customers around 8734.
- 2. Confidence Intervals Analysis:**
 - The confidence intervals for both male and female customers do not overlap, confirming a statistically significant difference in the average spending behavior between the genders. This suggests that gender may be an important factor to consider for targeted marketing and promotions.
- 3. Marital Status:**
 - The average purchase amount for married customers (9261) is slightly higher than that of unmarried customers (9265). However, the confidence intervals for both groups overlap, indicating that marital status might not be as significant a factor as gender for distinguishing customer spending behavior.
- 4. Age-Based Spending:**
 - Customers in the 51-55 age group (A6) have the highest average purchase amount (9534), followed closely by the 55+ age group (9336). Younger

customers, such as those in the 18-25 age range (A2), spend less on average compared to older customers. This suggests that older age groups tend to make larger purchases.

5. Sample Size and Confidence Intervals:

- Bootstrapping results confirm that as sample size increases, the confidence interval becomes narrower and more representative of the true population mean. For sample sizes of 100, 1000, and 10000, the confidence intervals become tighter, with the 10000 sample providing the most accurate estimate for the population mean.

Recommendations:

1. Targeted Marketing for Males:

- Since male customers have higher average transaction values, Walmart should consider focusing more on male-oriented products or promotions that might appeal specifically to men, such as electronics, sports, or tech-related items.

2. Strategize for Women:

- While female customers spend less on average, Walmart could tailor marketing campaigns to encourage increased spending, possibly through targeted discounts, bundles, or loyalty programs aimed at female customers.

3. Segmented Offers for Marital Status:

- Although the differences between married and unmarried customers are not highly significant, personalized offers (e.g., family-oriented products for married customers or single-serve products for unmarried) could still improve customer engagement and spending.

4. Age-Targeted Product Recommendations:

- The older age groups (A6, A7) spend more on average, so high-ticket items (like home appliances, travel, or premium products) should be marketed more heavily to these age groups. Additionally, targeting younger age groups (A1, A2) with budget-friendly options might increase their spending.

5. Optimizing Confidence Intervals for Better Decision-Making:

- Walmart can use bootstrapping with larger sample sizes to continuously refine their understanding of customer purchase behavior. This approach will help reduce uncertainty and enable data-driven decisions for more accurate targeting and campaign planning.

By leveraging these insights, Walmart can optimize its marketing strategies and improve customer segmentation to maximize revenue and engagement across different demographics.

