Walmart Case Study

Business Problem Statement

Walmart wants to analyze customer purchase behavior during Black Friday to understand how purchase amount varies across customer demographics, specifically:

- Gender
- Marital Status
- Age

The primary business question is:
Do women spend more money per transaction than men during Black Friday?

Basic Metrics

Metric	Description	Output
Total Records	Number of transactions in the dataset	**550,068**
Number of Columns	Total features available	**10**
Unique Customers	Number of distinct users	**5,891**
Unique Products	Number of distinct products	**3,631**
Gender Distribution	Transactions by gender	**Male – ~75%,** **Female – ~25%**
Age Groups	Number of age bins	**7 age groups**
Marital Status Split	Distribution by marital status	**Unmarried – ~59%,** **Married – ~41%**
City Categories	Types of cities	**3 (A, B, C)**
Product Categories	Distinct product categories	**20**
Purchase Range	Minimum to maximum purchase amount	**₹12 – ₹23,961**
Average Purchase	Mean transaction value	**~₹9,264**
Missing Values	Percentage of null values	**0% (No missing data)**
Duplicate Records	Redundant rows	**0 duplicates**
Data Types	Numerical vs Categorical columns	**5 Numerical, 5 Categorical**
Data Skewness	Shape of purchase distribution	**Right-skewed**

Comments on Attributes

Attribute	Description	Comment / Use in Analysis
User_ID	Unique customer identifier	Used to identify unique customers
Product_ID	Unique product identifier	Helps analyze product-level purchasing patterns
Gender	Gender of the customer (M/F)	Key attribute for comparing male vs female spending behaviour
Age	Age group of customer (binned)	Used to study spending behaviour across different age groups
Occupation	Occupation code	Can indicate profession-based purchasing trends
City_Category	City classification (A, B, C)	Used to compare urban vs semi-urban spending behavior
Stay_In_Current_City_Years	Years spent in current city	Indicates customer stability and loyalty potential
Marital_Status	Marital status (0 = Unmarried, 1 = Married)	Used to compare spending habits of married vs unmarried customers
Product_Category	Product category	Helps identify high-revenue and popular product categories
Purchase	Transaction purchase amount	Target variable; central to all spending, CI, and CLT analysis

Missing Value Analysis

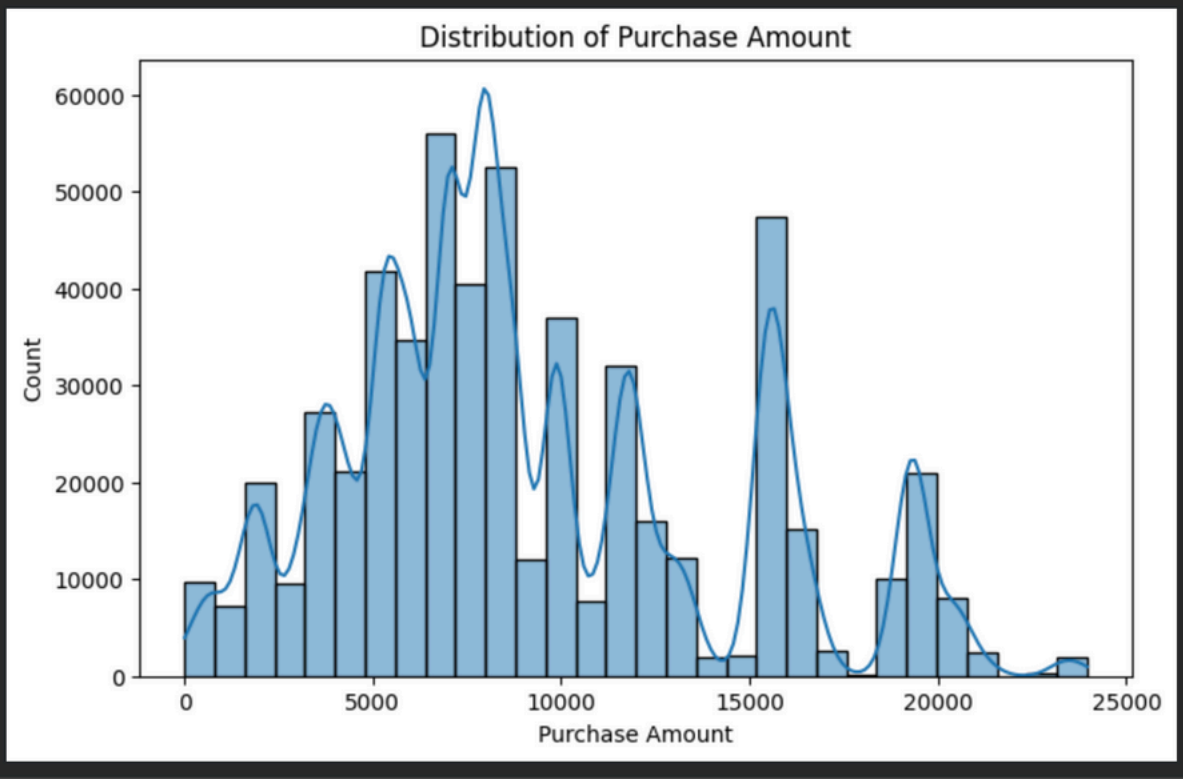
No missing values were found in the dataset. This ensures that the analysis is not biased due to incomplete data.

Univariate Analysis

1. Purchase (Continuous Variable)

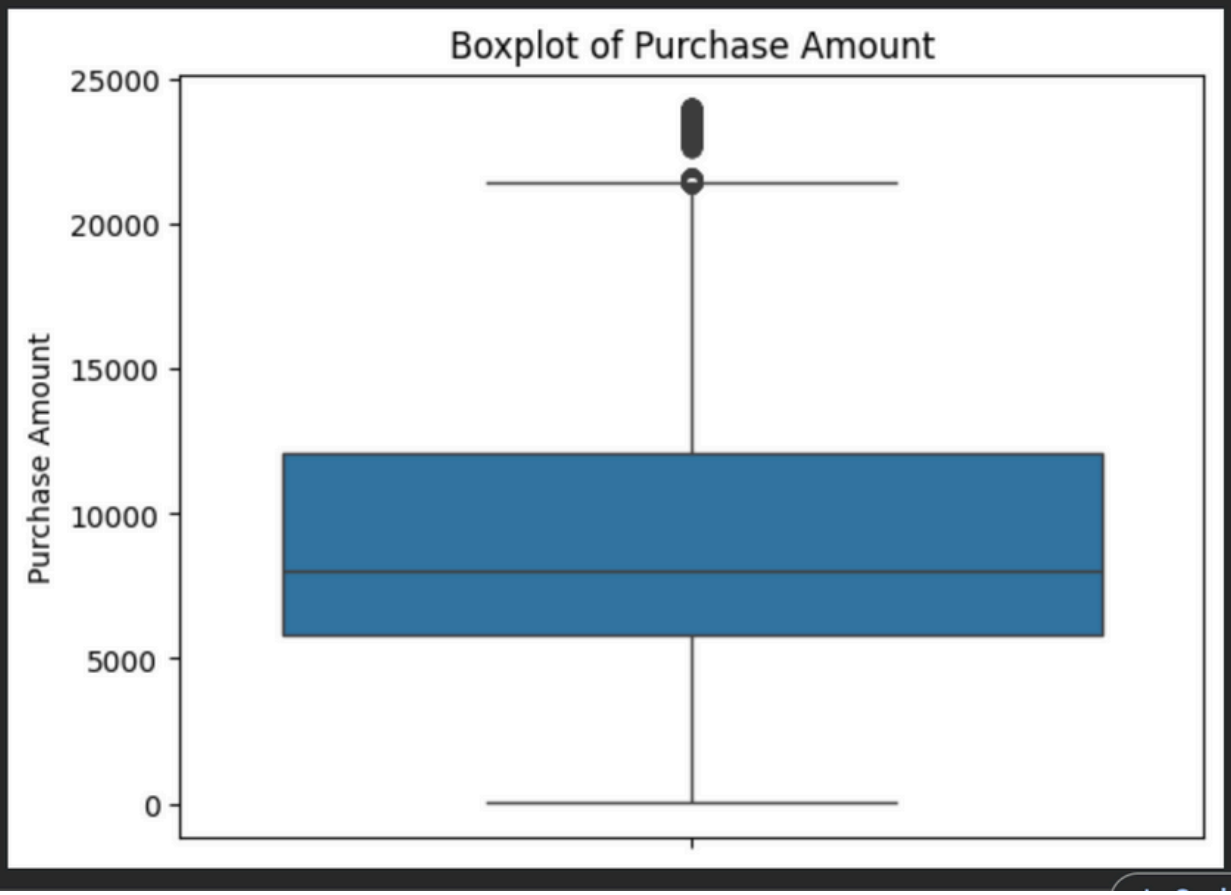
HISTPLOT

```
plt.figure(figsize=(8,5))
sns.histplot(data['Purchase'], kde=True, bins=30)
plt.xlabel('Purchase Amount')
plt.title('Distribution of Purchase Amount')
plt.show()
```



BOXPLOT

```
sns.boxplot(y=data['Purchase'])
plt.title('Boxplot of Purchase Amount')
plt.ylabel('Purchase Amount')
plt.show()
```



Insight from above plots:

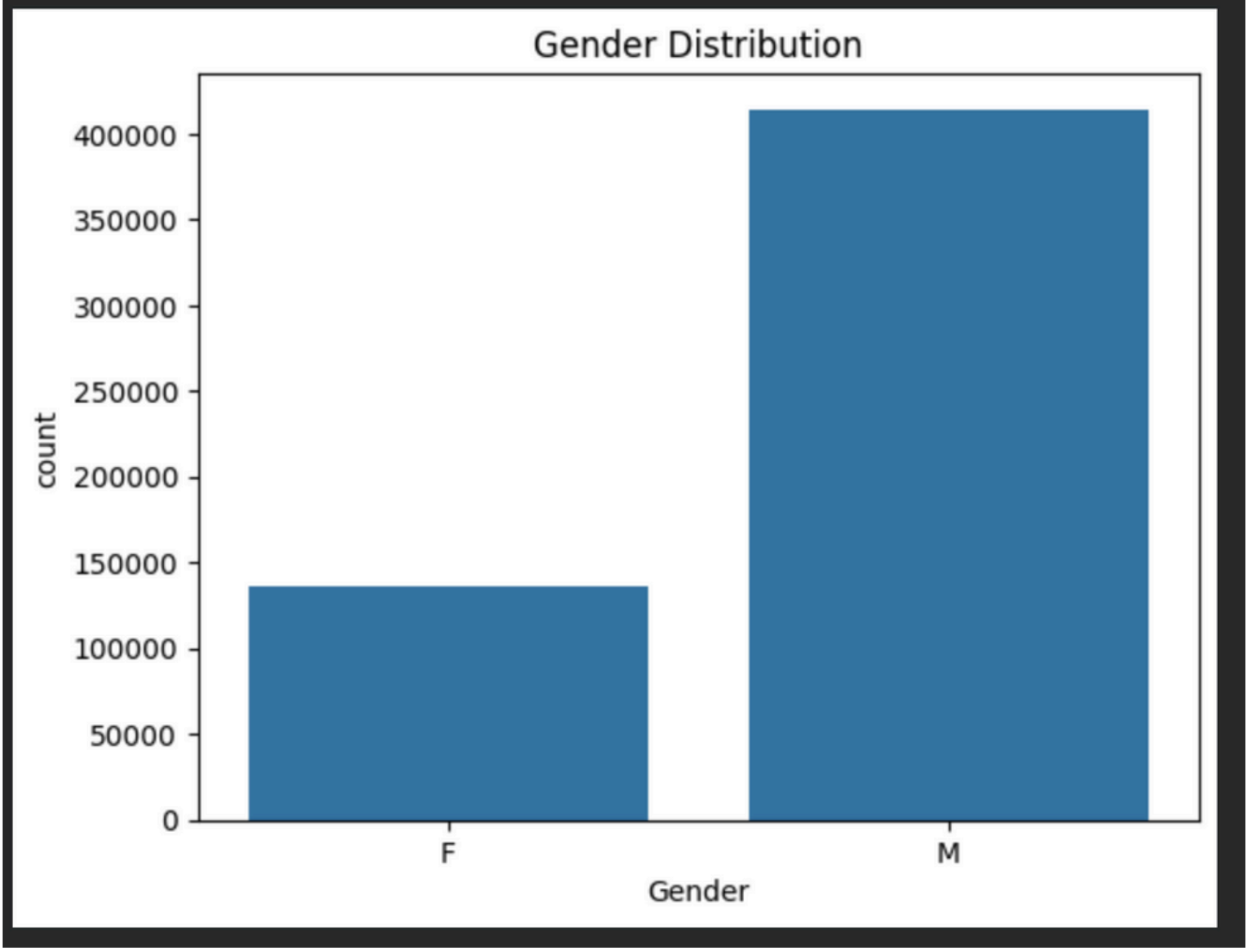
The boxplot of purchase amount shows the presence of several high-value outliers on the upper side, indicating a right-skewed distribution. The median is closer to the lower quartile, and the right whisker is longer than the left. This indicates that while most customers make moderate purchases, a small proportion contributes significantly higher transaction values.

Although the histogram does not appear smoothly right-skewed due to discrete purchase values and multiple price points, the boxplot confirms that the distribution is positively skewed.

2. Gender (Categorical Variable)

COUNTPLOT

```
sns.countplot(x='Gender', data=data)
plt.title('Gender Distribution')
plt.show()
```

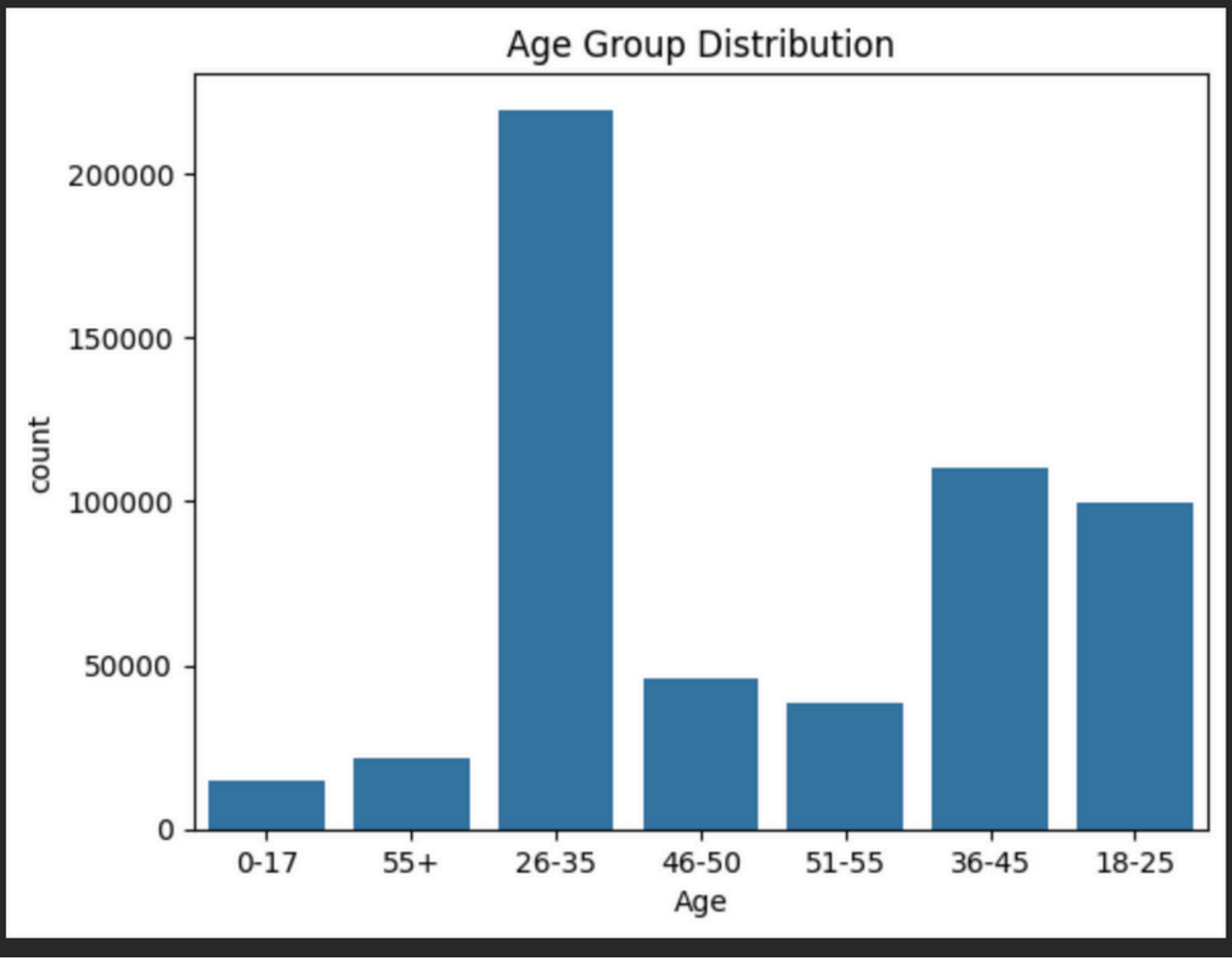


Male transactions are more frequent than female transactions.

3. Age (Categorical Variable)

COUNTPLOT

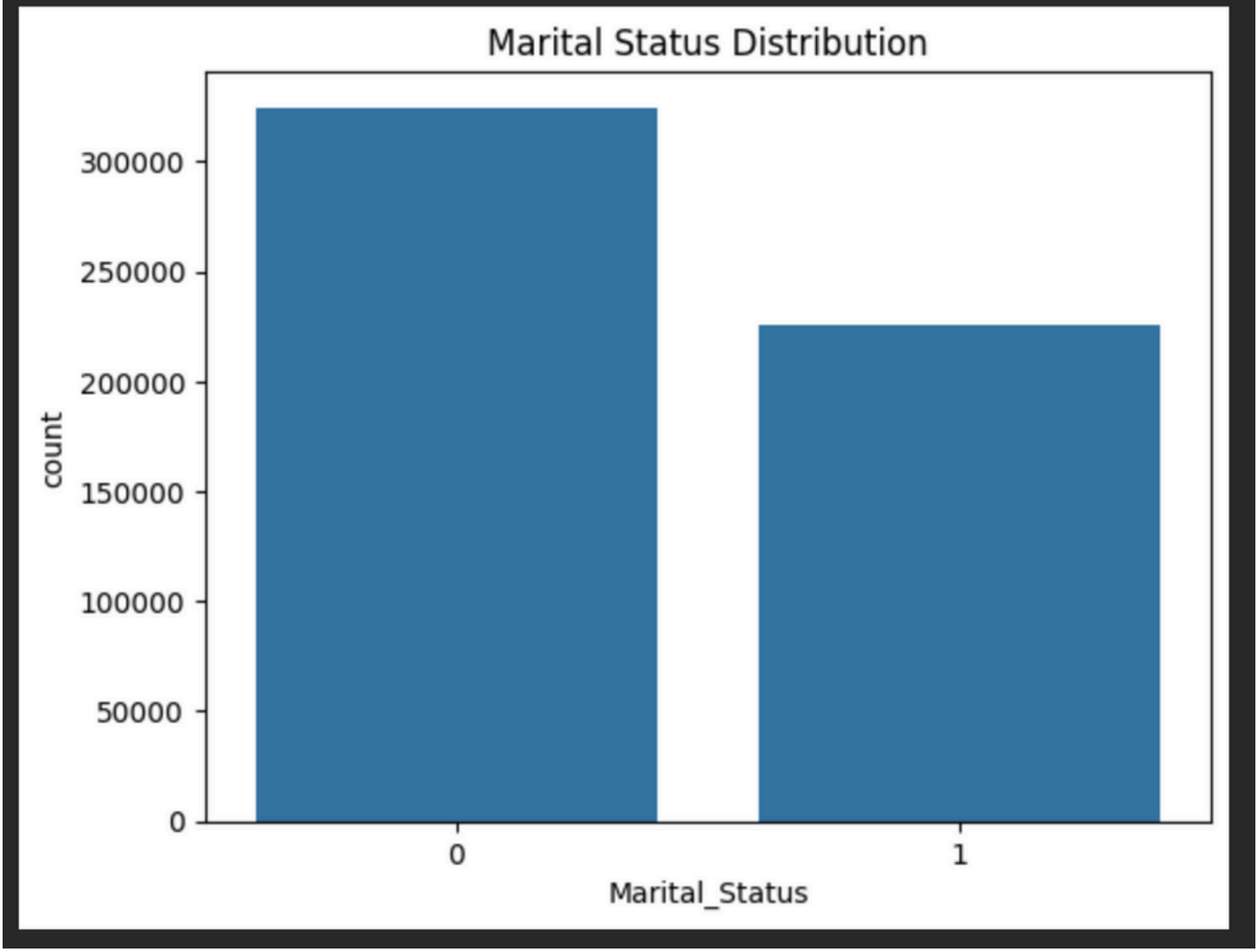
```
sns.countplot(x='Age', data=data)
plt.title('Age Group Distribution')
plt.show()
```



Age group 26–35 appears most dominant.

4. Marital Status (Categorical Variable)

```
sns.countplot(x='Marital_Status', data=data)
plt.title('Marital Status Distribution')
plt.show()
```



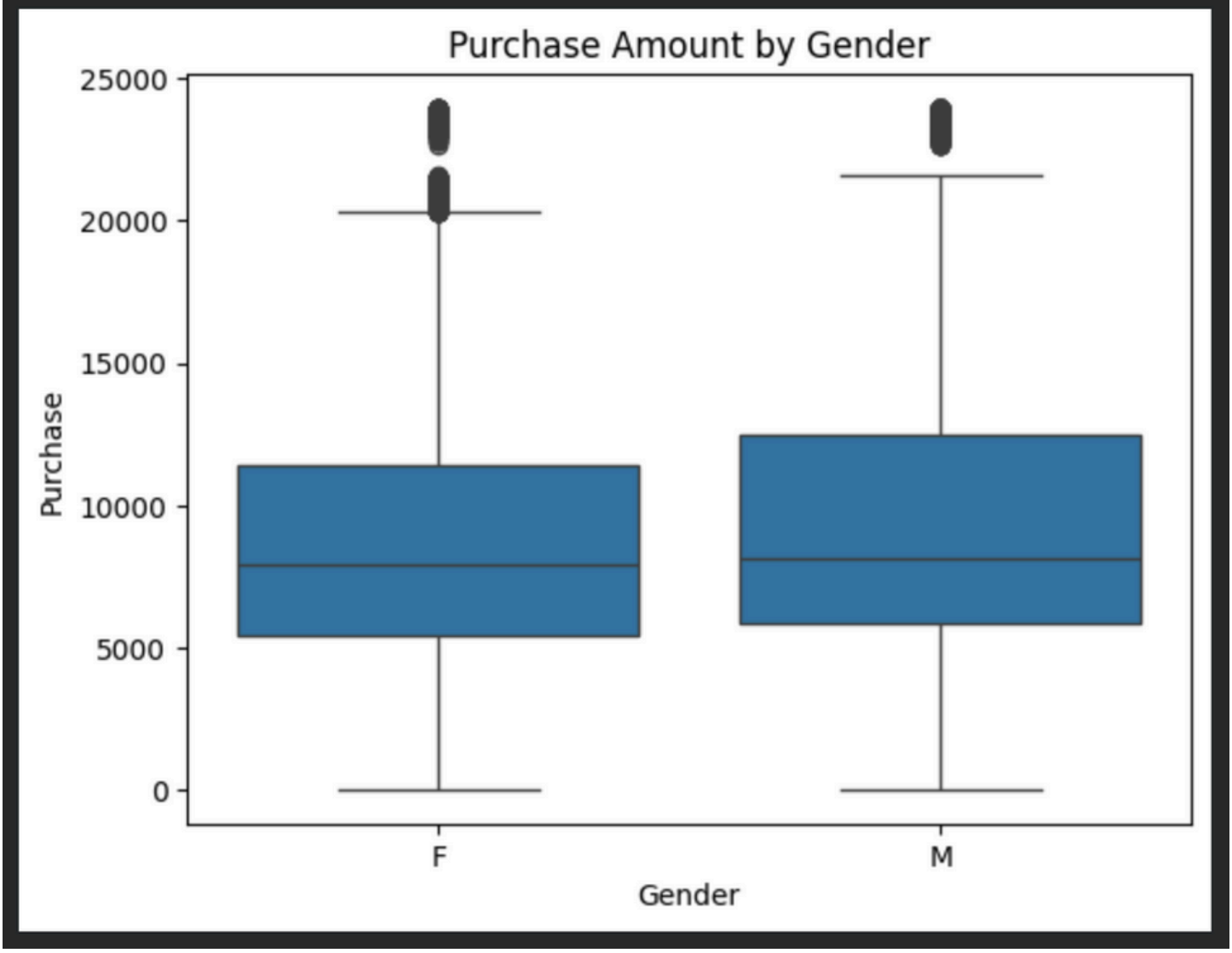
Transactions made by unmarried are more than married.

Bivariate Analysis

1. Gender vs Purchase

BOXPLOT

```
sns.boxplot(x='Gender', y='Purchase', data=data)
plt.title('Purchase Amount by Gender')
plt.show()
```



Male customers show:

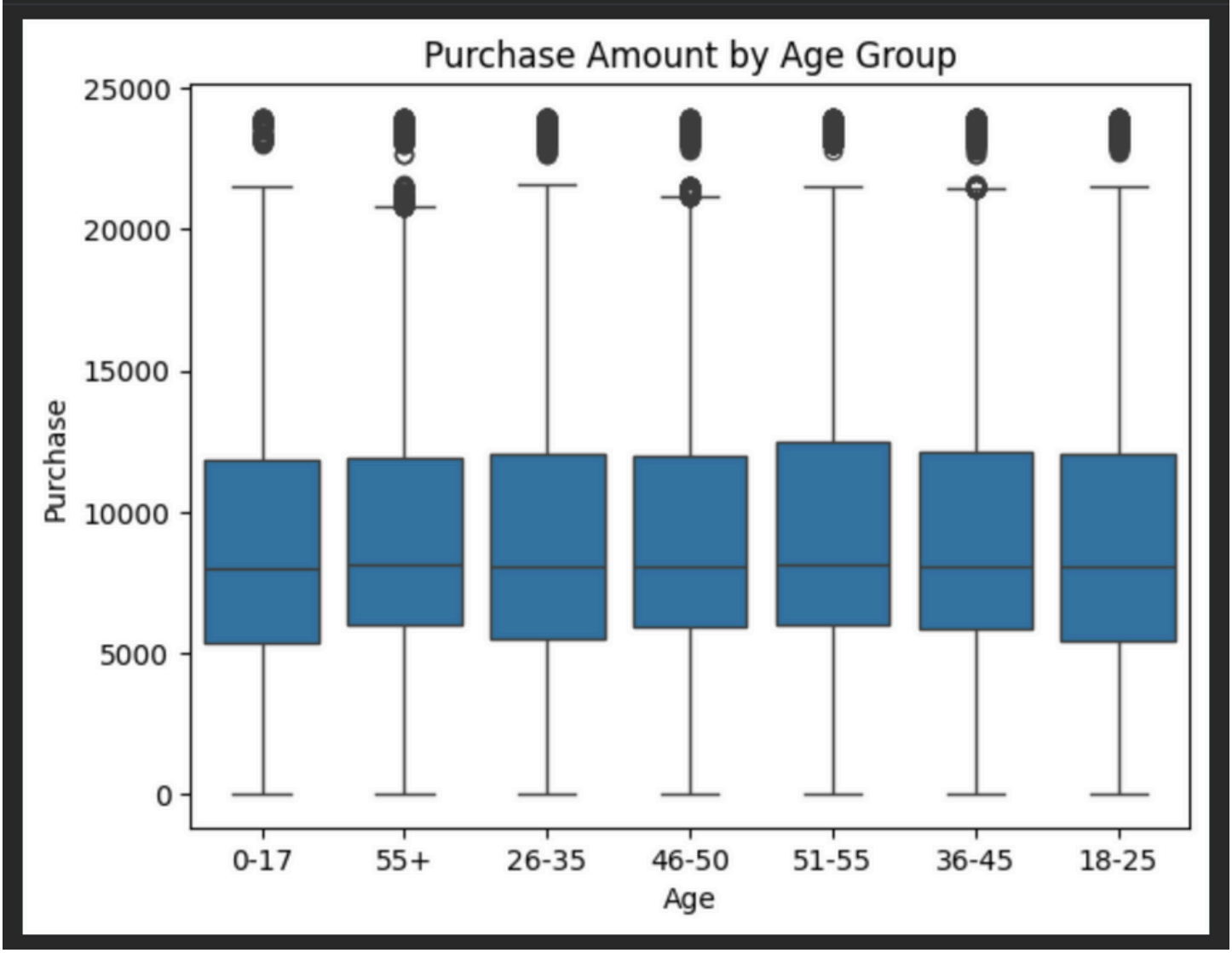
- Higher median purchase
- Wider interquartile range

This indicate male customers exhibit higher variability and higher spending behavior.

2. Age vs Purchase

BOXPLOT

```
sns.boxplot(x='Age', y='Purchase', data=data)
plt.title('Purchase Amount by Age Group')
plt.show()
```



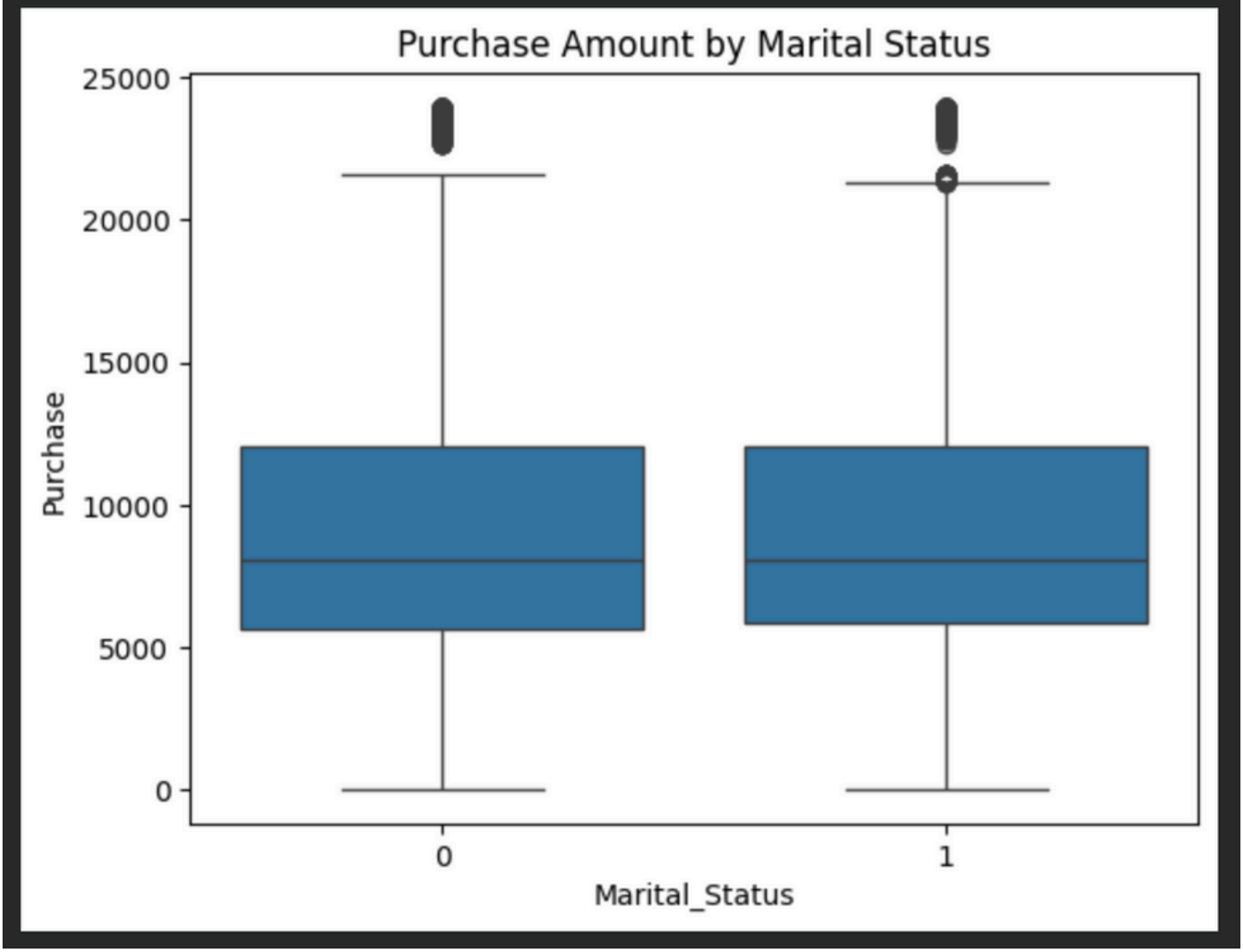
26–35 age group shows higher median and wider spread. Younger and older groups show lower purchase values.

Spending peaks during prime working age.

3. Marital Status vs Purchase

BOXPLOT

```
sns.boxplot(x='Marital_Status', y='Purchase', data=data)
plt.title('Purchase Amount by Marital Status')
plt.show()
```



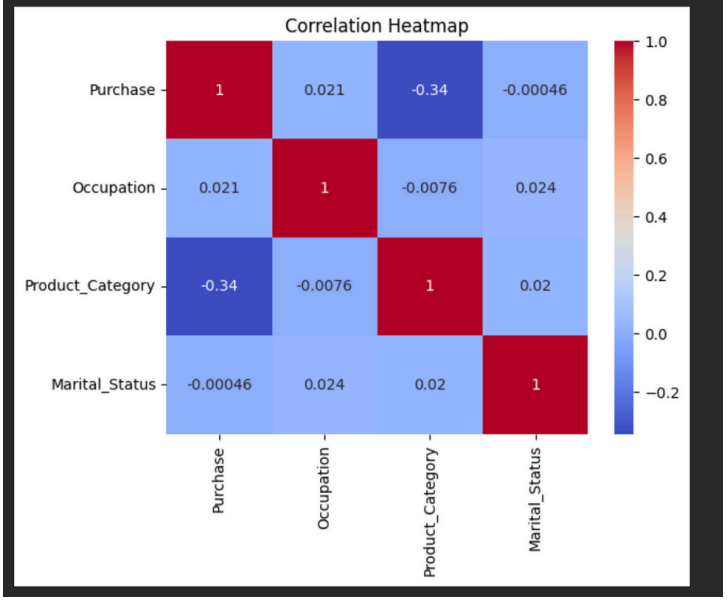
Married and unmarried customers show similar medians. Unmarried customers have slightly higher variability.

Marital status does not strongly influence purchase amount.

Correlation Analysis

HEATMAP

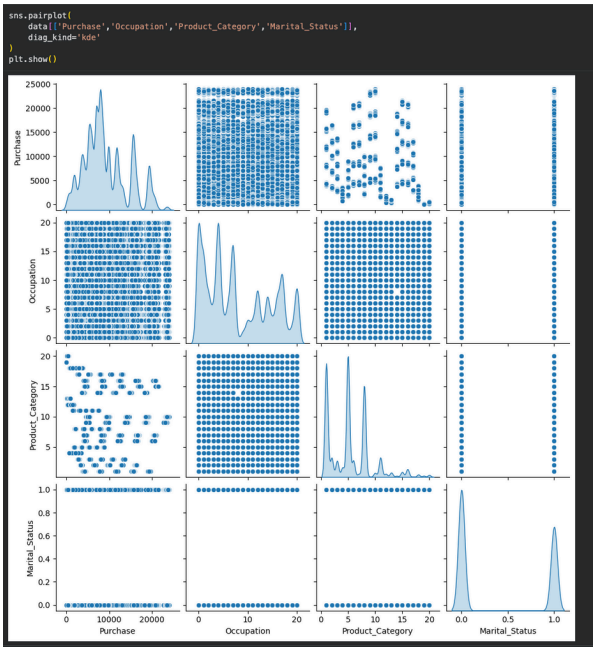
```
sns.heatmap(
    data[['Purchase','Occupation','Product_Category','Marital_Status']].corr(),
    annot=True,
    cmap='coolwarm'
)
plt.title('Correlation Heatmap')
plt.show()
```



The heatmap shows that the purchase amount has very weak correlation with occupation, product category, and marital status. No strong linear relationship is observed between purchase and any numerical variable. This indicates that customer spending is not strongly influenced by a single numeric factor, but rather by categorical and behavioral attributes.

PAIRPLOT

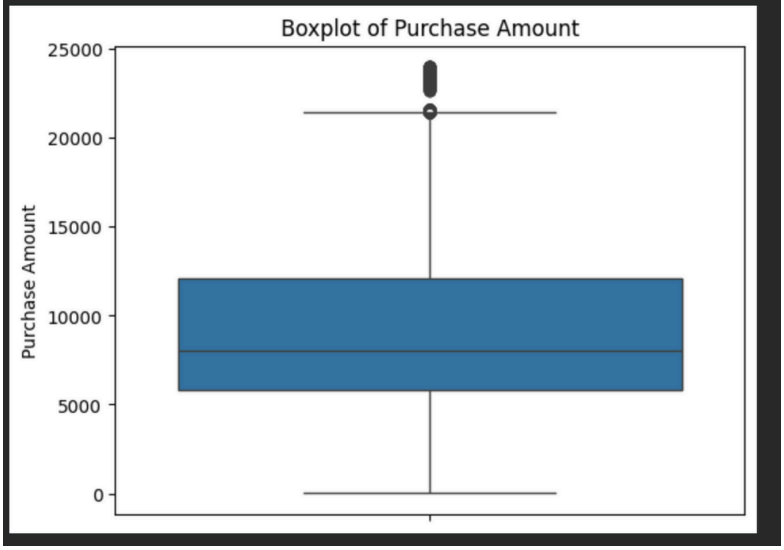
```
sns.pairplot(
    data[['Purchase','Occupation','Product_Category','Marital_Status']],
    diag_kind='kde'
)
plt.show()
```



The pairplot shows no clear linear or non-linear relationship between purchase amount and other numerical variables. Scatterplots are widely dispersed, indicating high variability in purchase behavior. This further confirms that customer spending cannot be explained by numeric attributes alone.

Outlier Detection

```
sns.boxplot(y=data['Purchase'])
plt.title('Boxplot of Purchase Amount')
plt.ylabel('Purchase Amount')
plt.show()
```



Boxplot indicates the presence of outliers in purchase amounts. The mean is higher than the median, confirming positive skewness. These outliers represent high-value purchases and are retained since they are valid business transactions.

What is CLT?

For a sufficiently large sample size, the sampling distribution of the sample mean approaches a normal distribution, regardless of the shape of the original population distribution, provided the population has a finite mean and variance.

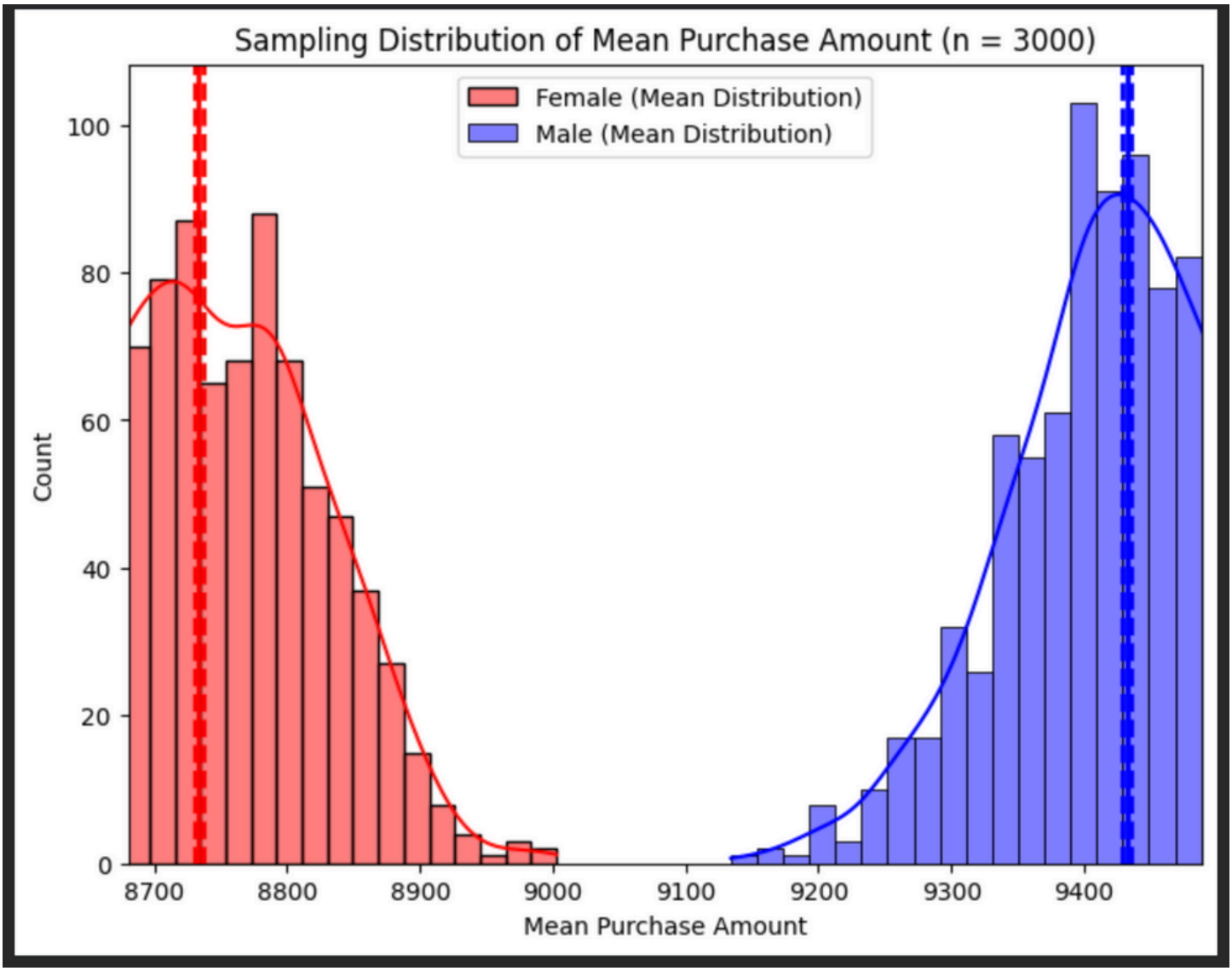
Derived Conclusion from CLT

- Increasing sample size improves the precision of the estimated mean, resulting in a narrower confidence interval. Because SE=SD/Sqrt(n).
- The confidence intervals for different sample sizes overlap and the overlap decreases as the sample size increases.
- All samples are drawn from the same underlying population.
- The true mean remains the same.
- Smaller samples → higher variability → wider CI
- Overlapping confidence intervals indicate consistent estimation of the same population mean, while reduced overlap reflects greater estimation precision at larger sample sizes.
- As sample size increases, the sampling distribution of the mean becomes more normal and more concentrated.
- Larger sample sizes produce tighter, smoother, and more normally distributed sampling distributions of the mean.

Gender - Based Spending Analysis

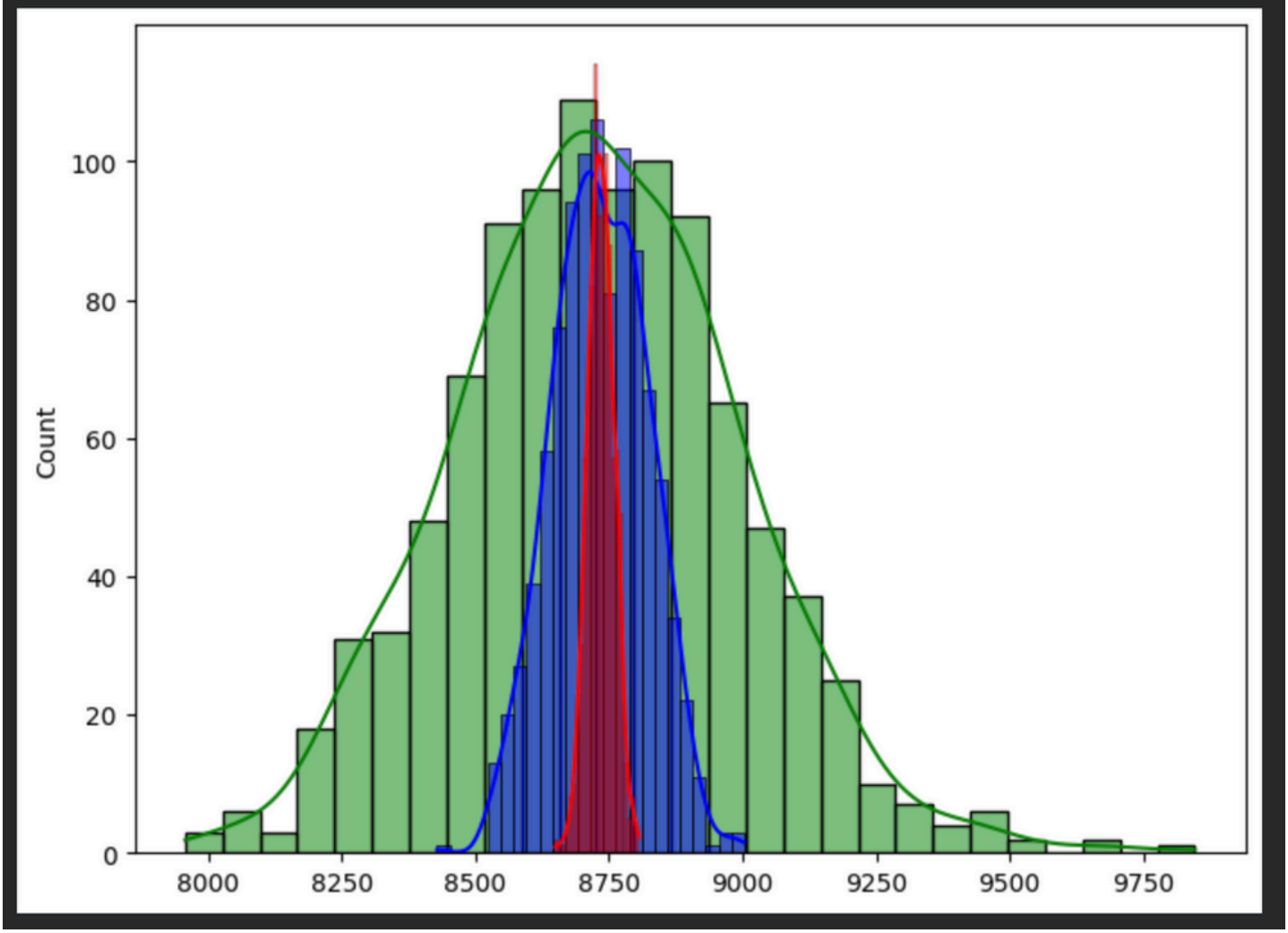
Sample Size	GENDER	MEAN	CI	CI width
300	Female	8727.985963333333	8697.754177163119, 8758.217749503476	60.463572340286 19
300	Male	9432.041933333334	9399.920040074676, 9464.163826591992	64.243786517316 04
3000	Female	8733.713777333334	8730.583578250847, 8736.843976415821	6.260398164973594
3000	Male	9432.288065333332	9429.091733124465, 9435.484397542199	6.392664417733613
30000	Female	8735.146586666666	8734.870060280 55, 8735.423113052782	0.553052772233 0589
30000	Male	9437.752103533334	9437.442965590126, 9438.061241476542	0.618275886416086

- **Average spend:** The mean purchase amount for **male customers is higher** than that for female customers.
- **Confidence intervals:** The 95% confidence intervals for the average spend show that the **male interval lies entirely above the female interval**. Since these intervals **do not overlap**, the difference is **statistically significant**.



- **Variability:** Male spending shows **greater variability** (wider CIs and boxplots with more extreme values), indicating more high-ticket purchases.

```
fig, ax = plt.subplots(figsize=(8, 6))
sns.histplot(female_sample_300,kde=True,ax=ax,color='g',label='300')
sns.histplot(female_sample_3000,kde=True,ax=ax,color='b',label='3000')
sns.histplot(female_sample_30000,kde=True,ax=ax,color='r',label='30000')
plt.show()
```



- **Robustness:** As sample size increases, confidence intervals narrow but the **male–female separation remains**, strengthening the conclusion.

Business Advice

The confidence intervals for average male and female spending do not overlap, indicating a statistically significant difference in per-transaction purchase amounts. Male customers consistently show higher average spending and greater variability. This provides strong evidence that male customers constitute a higher-value segment during Black Friday. Walmart can leverage this insight by designing targeted promotions, premium bundles, and personalized campaigns for male customers, while continuing to engage female customers through value-driven and loyalty-focused strategies.

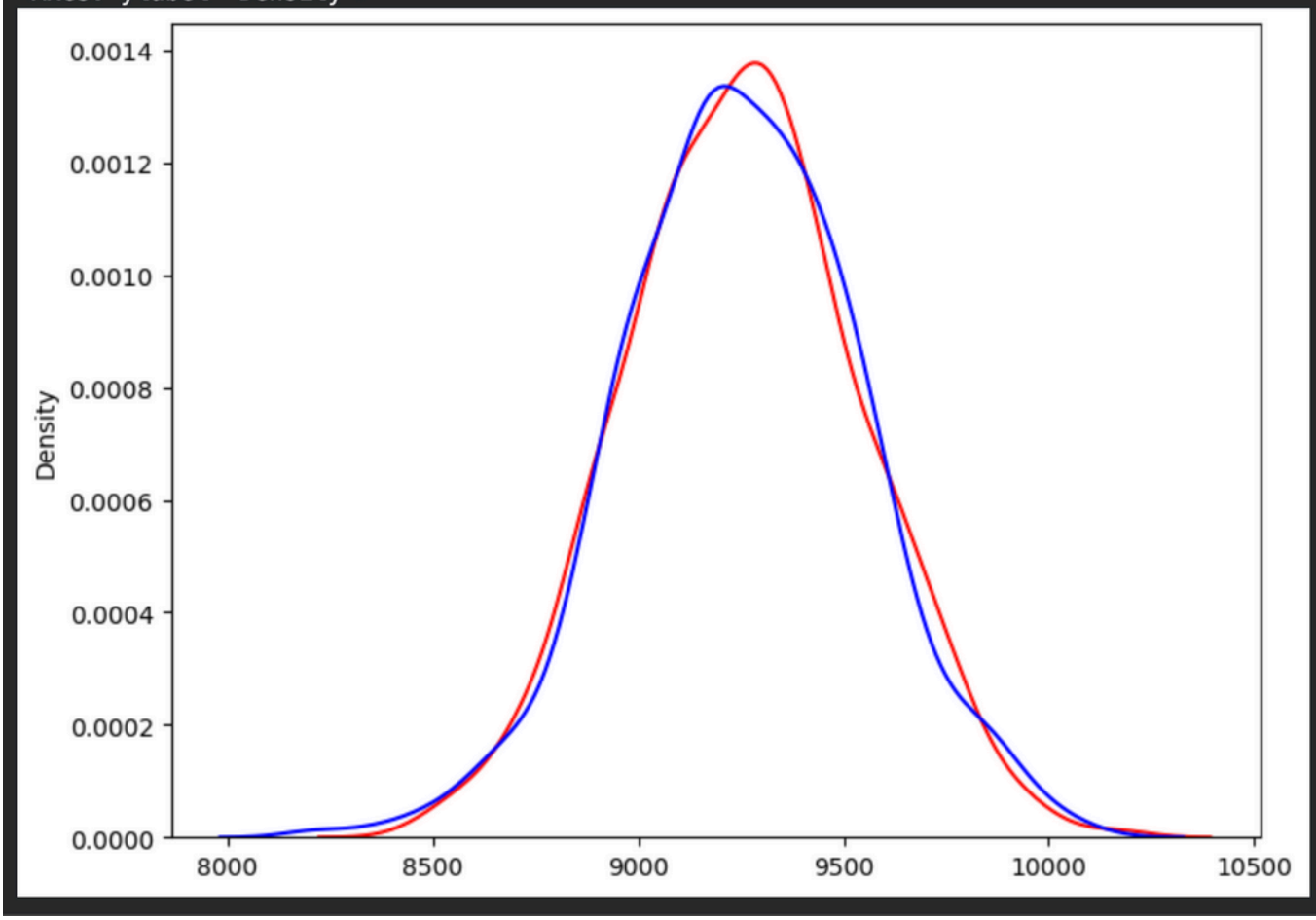
Marital Status - Based Spending Analysis

Sample Size	GENDER	MEAN	CI	CI width
300	Married	9256.3827	9224.173685895 31), np.float64(9288. 59171410469	64.418028209380 28
300	Unmarried	9252.901976666 666	9219.949703480 04), np.float64(9285. 854249853292	65.904546373250 92
3000	Married	9255.499629333 333	9252.244092053 486), np.float64(9258. 75516661318	6.511074559693 952
3000	Unmarried	9265.255932666 667	9262.002929244 634), np.float64(9268. 5089360887	6.506006844065 2685
30000	Married	9260.568656666 666	9260.260216281 074), np.float64(9260. 877097052258	0.616880771183 7871
30000	Unmarried	9266.024416233 333	9265.707096731 461), np.float64(9266. 341735735205	0.634639003743 3594

- **Average spend:** The mean purchase amount for **married and unmarried customers is very similar**, with only a marginal difference observed between the two groups.
- **Confidence intervals:** The 95% confidence intervals for the average spending of married and unmarried customers **overlap significantly**. This overlap indicates that the observed difference in mean purchase amounts is **not statistically significant** at the population level.

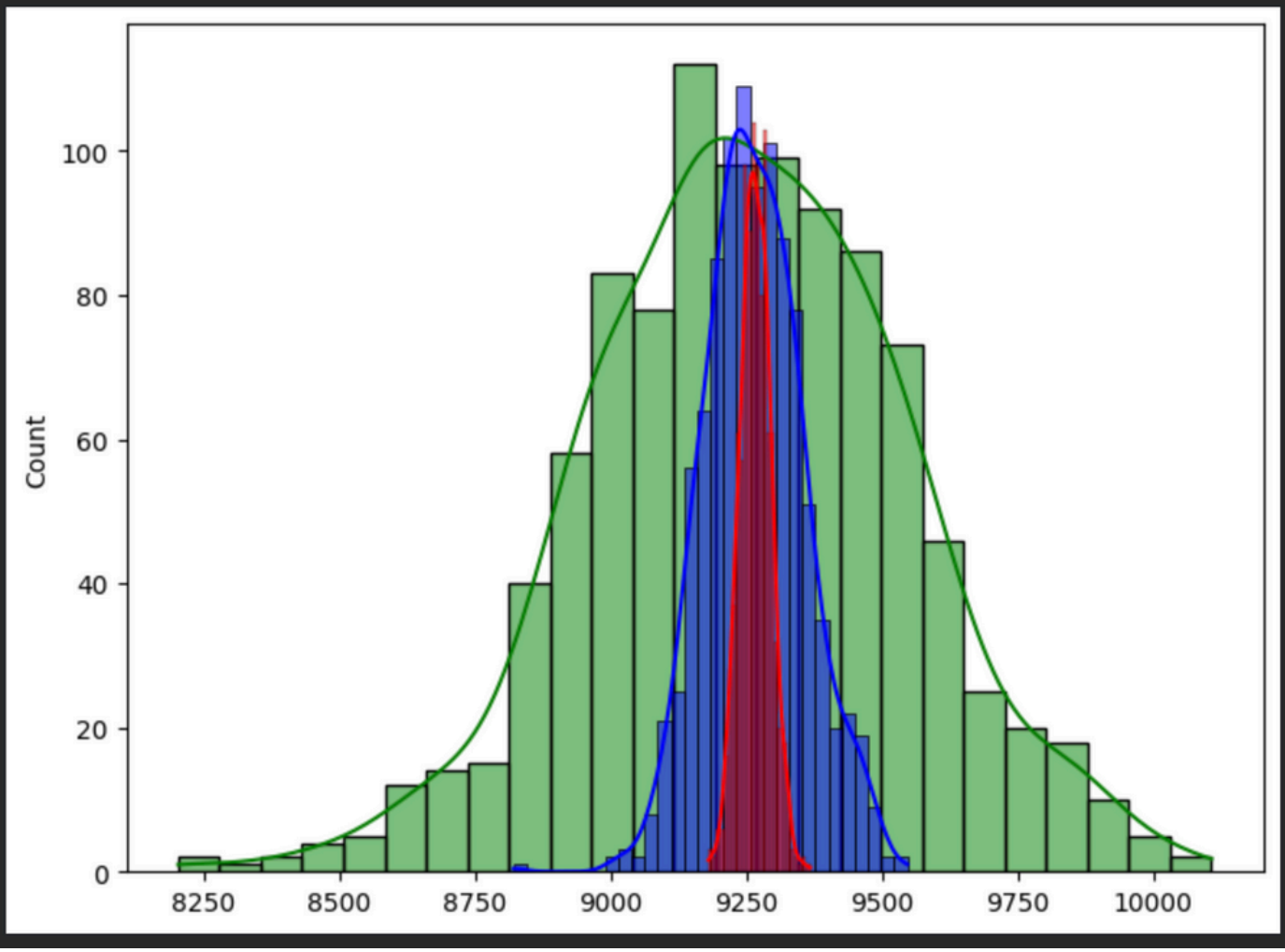
```
fig, ax = plt.subplots(figsize=(8, 6))
sns.kdeplot(married_sample_300,ax=ax,color='r',label='300')
sns.kdeplot(unmarried_sample_300,ax=ax,color='b',label='300')
```

```
<Axes: ylabel='Density'>
```



- **Variability:** Both married and unmarried customers exhibit **comparable variability** in spending behaviour. The width of the confidence intervals and the spread observed in boxplots are similar, suggesting that neither group consistently makes more extreme high-value purchases than the other.

```
fig, ax = plt.subplots(figsize=(8, 6))
sns.histplot(unmarried_sample_300,kde=True,ax=ax,color='g',label='300')
sns.histplot(unmarried_sample_3000,kde=True,ax=ax,color='b',label='3000')
sns.histplot(unmarried_sample_30000,kde=True,ax=ax,color='r',label='30000')
plt.show()
```



- **Robustness:** As the sample size increases, the confidence intervals for both groups become narrower, reflecting improved precision of the estimated means. However, the **overlap between the intervals persists across different sample sizes**, reinforcing the conclusion that marital status alone does not strongly differentiate spending behavior during Black Friday.

Business Advice

The overlapping confidence intervals for married and unmarried customers indicate that **marital status is not a strong predictor of per-transaction spending** during Black Friday. Since both groups demonstrate similar average spending and variability, Walmart should avoid designing promotions solely based on marital status. Instead, Walmart can focus on other more influential factors such as **age group, product category preferences, and customer purchase history**. Campaigns that emphasize lifestyle needs, family-oriented bundles, or individual convenience can be applied broadly rather than segmented strictly by marital status.

Age Group - Based Spending Analysis

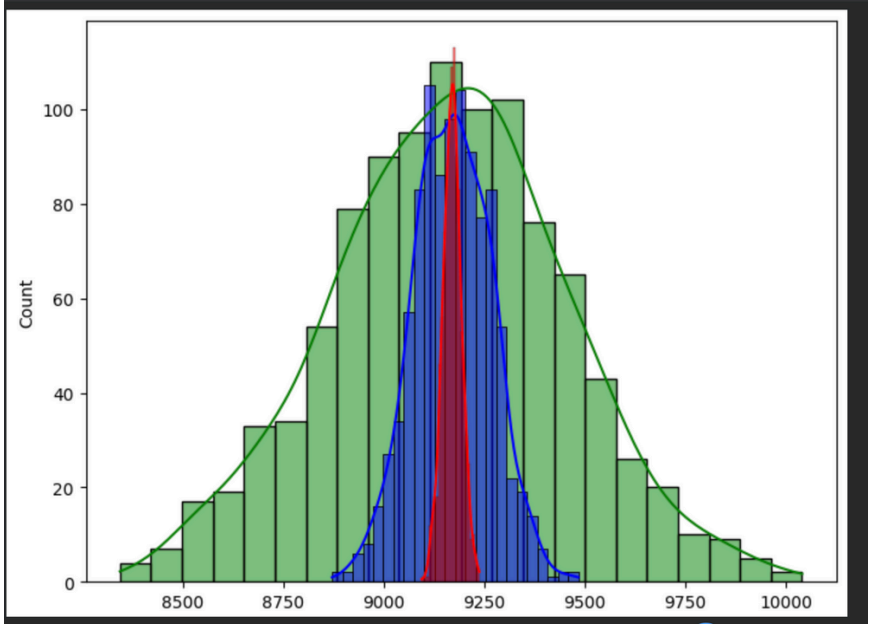
Sample Size	Age Group	MEAN	CI	CI width
300	0-17	8925.080933333333	8892.58511402882 7, 8957.576752637839	64.99163860901172
300	18-25	9160.742353333333	9127.754394829729 3, 9193.730311836945	65.97591700722478
300	26-35	9253.571026666667	9220.947759596787 , 9286.194293736553	65.24653413977285
300	36-45	9337.70335	9305.475266660105, 9369.931433339894	64.45616667978902
300	46-50	9216.430106666665	9183.286657477216, 9249.573555856114	66.28689837889 76
300	51-55	9518.914406666667	9484.378829170199, 9553.449984163135	69.07115499293 65
300	55+	9329.944826666668	9297.370263917861, 9362.519389415475	65.14912549761357
3000	0-17	8931.769551	8928.760049726674, 8934.779052273325	6.019002546650881
3000	18-25	9171.088726333333	9167.699141120193,	6.7791704262817 8

			9174.478311546474	
3000	26-35	9253.49091	9250.05918226019, 9256.922637739812	6.863455479622644
3000	36-45	9331.669187000001	9328.485532687571, 9334.852841312431	6.367308624860016
3000	46-50	9207.099226	9203.867620600799, 9210.330831399202	6.463210798403452
3000	51-55	9532.825469999998	9529.596867434 72, 9536.054072565275	6.457205130554939
3000	55+	9338.378199666666	9335.345715996 06, 9341.410683337272	6.0649673412117 41
30000	18-25	9170.1038453	9169.844616658189, 9170.363073941811	0.5184572836224106
30000	26-35	9252.380493699999	9252.084139608689, 9252.676847791308	0.5927081826193898
30000	36-45	9332.145438166666	9331.876890422 86, 9332.413985910473	0.5370954876125325
30000	46-50	9208.9351192	9208.743669811027,	0.3828987779452291

			9209.126568588972	
30000	51-55	9534.6617216	9534.509077370405, 9534.814365829594	0.3052884591888869

- **Average spend:**The mean purchase amount varies across age groups. Customers in the **26–35, 36–45, and 51–55** age brackets consistently exhibit **higher average spending per transaction** compared to younger age groups such as **0–17 and 18–25**. Among all groups, the **51–55 age group shows the highest mean purchase amount**, followed by customers aged **36–45**.
- **Confidence intervals:**The 95% confidence intervals computed for each age group reveal **partial overlap across adjacent age bands**, while **clear separation is observed between younger and older age groups**, particularly when using larger sample sizes (n = 3000). This indicates that although spending differences between neighboring age groups are modest, **older age groups tend to have a higher central tendency in purchase amounts**.
- **Variability:**At smaller sample sizes (n = 300), confidence intervals are wider across all age groups, reflecting higher sampling variability. As sample size increases to n = 3000, the confidence intervals narrow substantially and become more stable across age groups. The **51–55 age group consistently shows a slightly wider confidence interval**, indicating marginally higher variability and the presence of higher-value purchases within this segment.

```
fig, ax = plt.subplots(figsize=(8, 6))
sns.histplot(age1_sample_300,kde=True,ax=ax,color='g',label='300')
sns.histplot(age1_sample_3000,kde=True,ax=ax,color='b',label='3000')
sns.histplot(age1_sample_30000,kde=True,ax=ax,color='r',label='30000')
plt.show()
```



- **Robustness:**With increasing sample size, the confidence intervals across all age groups become narrower, confirming improved precision of the estimated mean purchase amounts. Importantly, the **relative ordering of age groups remains consistent** across different sample sizes, reinforcing the robustness of the observed age-based spending patterns.

Business Advice

The age-based confidence interval analysis indicates that **customer spending behaviour varies meaningfully across life stages**. Customers aged **26–45 and 51–55** represent the most valuable segments in terms of average transaction value, while younger age groups tend to spend less per transaction. Walmart can leverage this insight by tailoring Black Friday promotions based on life-stage needs—such as premium electronics, home improvement products, and family-oriented bundles for higher-spending age groups—while offering budget-friendly deals and entry-level product discounts to younger customers. This age-specific targeting can help maximise revenue while maintaining broad customer engagement.

Age has a stronger influence on spending behaviour than marital status, with middle-aged and older customers consistently demonstrating higher average purchase values and more stable spending patterns.

CONCLUSION

Comments on Distribution of Variables & Relationships

- **Purchase Amount** shows a positively skewed distribution across all segments, driven by high-value purchases.
- **Gender vs Purchase** reveals a noticeable shift in central tendency, with male customers spending more on average.
- **Age vs Purchase** demonstrates a clear life-stage pattern, where spending increases during prime working and earning years.
- **Marital Status vs Purchase** shows overlapping distributions, indicating minimal influence on spending.
- Numerical variables exhibit **weak correlation** with purchase amount, suggesting that customer behavior is influenced more by demographic and categorical factors than by individual numeric attributes.

Comments on Univariate and Bivariate Plots

Univariate Plots

- **Histogram (Purchase):** Shows concentration of transactions at lower values with a long right tail.
- **Boxplot (Purchase):** Highlights valid high-value outliers and confirms positive skewness.
- **Countplots (Gender, Age, Marital Status):** Indicate balanced and meaningful category distributions suitable for comparative analysis.

Bivariate Plots

- **Gender vs Purchase (Boxplot):** Males show higher median and wider spread, indicating greater spending variability.

- **Age vs Purchase (Boxplot):** Spending peaks in the 26–45 and 51–55 age groups.
- **Marital Status vs Purchase (Boxplot):** Medians are similar with overlapping ranges, showing limited impact.
- **Heatmap & Pairplot:** Reveal no strong linear relationships among numerical variables, reinforcing the importance of categorical segmentation.

Generalizing Insights to the Population

Although the dataset contains over **550,000 transactions**, it is still treated as a sample. Confidence interval and CLT-based analysis confirm that:

- Sample means are accurate estimates of population means.
- Larger samples reduce uncertainty and improve precision.
- Observed demographic patterns are **not due to random variation**, but reflect true population behavior.

This allows Walmart to make confident, data-driven decisions at a population level rather than relying solely on sample observations.

Recommendations (Business-Focused)

1. Focus Black Friday strategies on **high-value customer segments**, especially males and customers aged 26–45 and 51–55.
2. Use **age-based segmentation** rather than marital status for targeted promotions.
3. Avoid one-size-fits-all discounting; instead, design offers based on spending capacity and life stage.
4. Use confidence-interval-based metrics to evaluate campaign success and reduce decision risk.

Actionable Items for Walmart

- Increase inventory and marketing spend for premium and high-demand product categories.
- Run targeted digital campaigns for customers in prime spending age groups.
- Design premium bundles and upsell offers aimed at high-spending segments.
- Offer budget-friendly deals to younger customers to encourage volume growth.
- Continuously monitor customer spending patterns using large samples for reliable insights.