Walmart Business case study

Introduction:

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

Summary statistics of the dataset:

Jupiter Notebook Link: https://drive.google.com/file/d/1UlNt4l_cvXIQH0Ego1-wnQuGvTM1Y7-g/view?usp=sharing

Shape: (550068, 10) -> 5,50,068 unique contents, each with 10 columns providing various details about it.

Columns:

- User_ID: User ID
- Product_ID: Product ID
- Gender: Sex of User
- Age: Age in bins
- Occupation: Occupation
- City_Category: Category of the City (A,B,C)
- StayInCurrentCityYears: Number of years stay in current city
- Marital_Status: Marital Status
- ProductCategory: Product Category
- Purchase: Purchase Amount

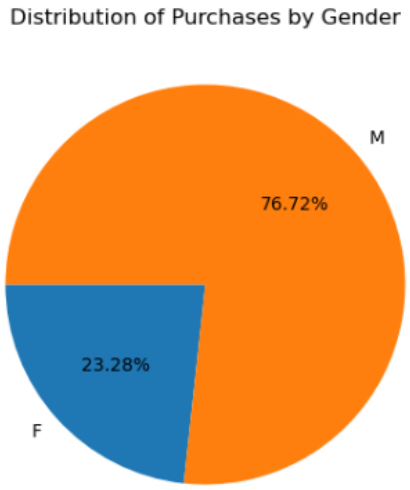
Datatype and No. of Nulls in each column:

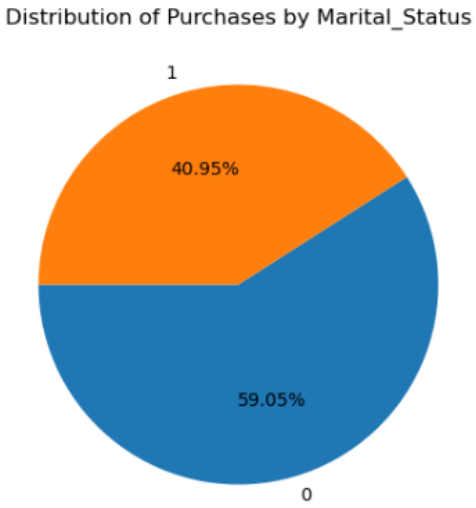
#	Column	Non Null Count	Dtype	New Dtype
0	User_ID	550068	int64	int64
1	Product_ID	550068	object	object
2	Gender	550068	object	category
3	Age	550068	object	category
4	Occupation	550068	int64	category
5	City_Category	550068	object	category
6	Stay_In_Current_City_Years	550068	object	category
7	Marital_Status	550068	int64	category
8	Product_Category	550068	int64	category
9	Purchase	550068	int64	int64

No. of Unique Values in each 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

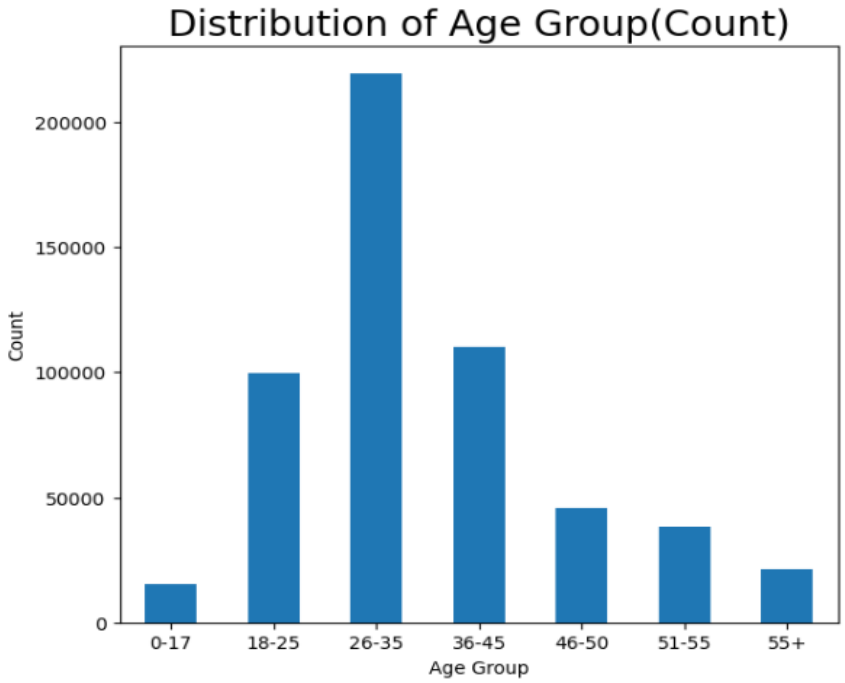
Univariate Analysis:





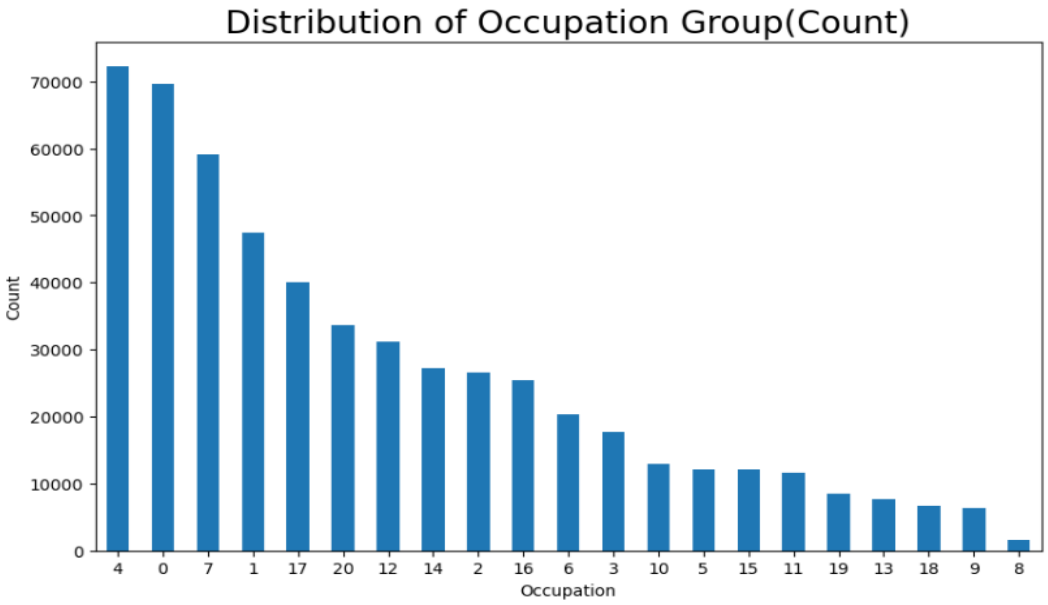
Insight:

- Based on the above pie chart, we can infer that the contribution of males is higher, at 76.72%, compared to females at 23.28%.
- Similarly, the purchase value for unmarried individuals is higher, at 59.05%, compared to married individuals at 40.95%.



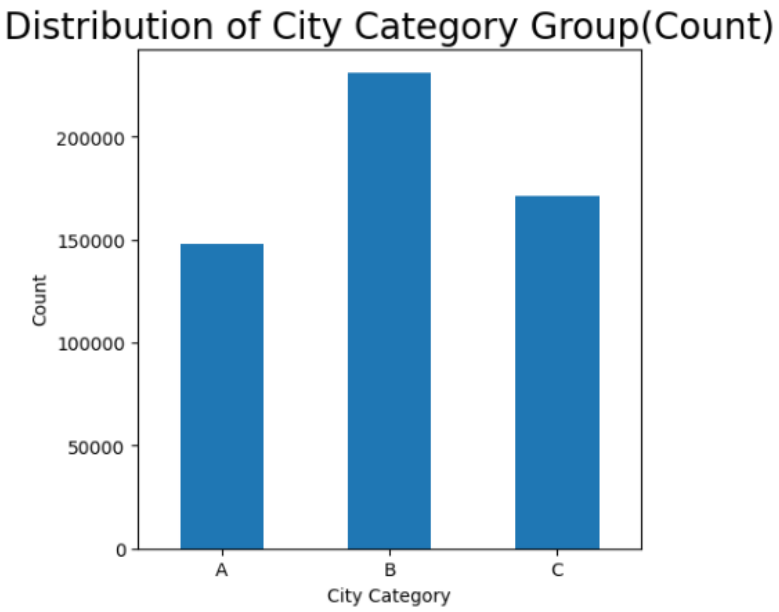
Insight:

- The age group 26-35 has the highest count of transactions, indicating that this age group is the most active in making purchases at Walmart during Black Friday.
- The age groups 18-25 and 36-45 also show significant transaction counts, suggesting these age groups are also active shoppers, but not as much as the 26-35 age group.
- The age groups 46-50, 51-55, and 55+ have lower transaction counts compared to younger age groups, indicating that older customers are less active in making purchases during Black Friday.
- The age group 0-17 has the lowest transaction count, which is expected as this group consists of minors who may not have significant purchasing power.



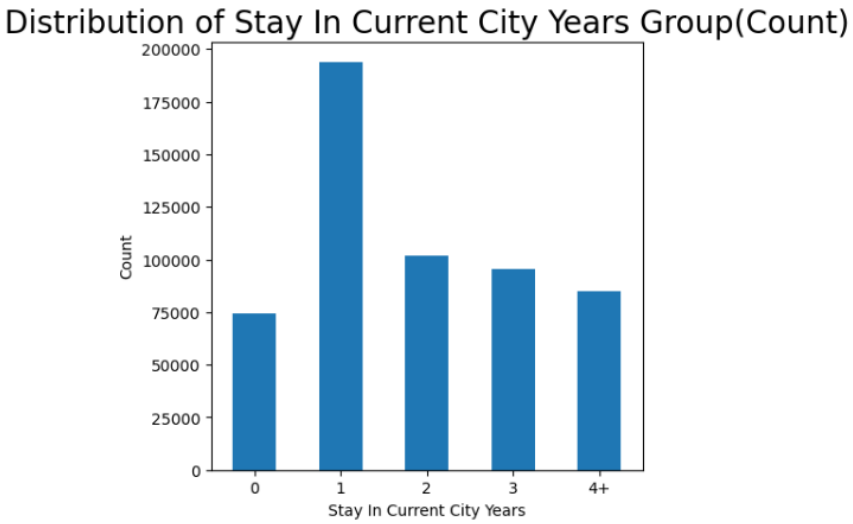
Insight:

- Highest Transaction Count in Occupation Group 4**:** The occupation group labelled '4' has the highest transaction count, indicating that individuals in this occupation are the most active in making purchases at Walmart during Black Friday.
- High Transaction Count in Occupation Groups 0 and 7**:** Occupation groups '0' and '7' also have significant transaction counts, suggesting that individuals in these occupations are also frequent shoppers, though not as much as those in occupation group '4'.
- Moderate Transaction Count in Occupation Groups 1, 17, 20, 12, 14, 2, and 16**:** These occupation groups have moderate transaction counts, indicating a decent level of shopping activity among individuals in these occupations.
- Lower Transaction Count in Remaining Occupation Groups**:** The remaining occupation groups have lower transaction counts, with the lowest counts observed in occupation groups '9' and '8'.



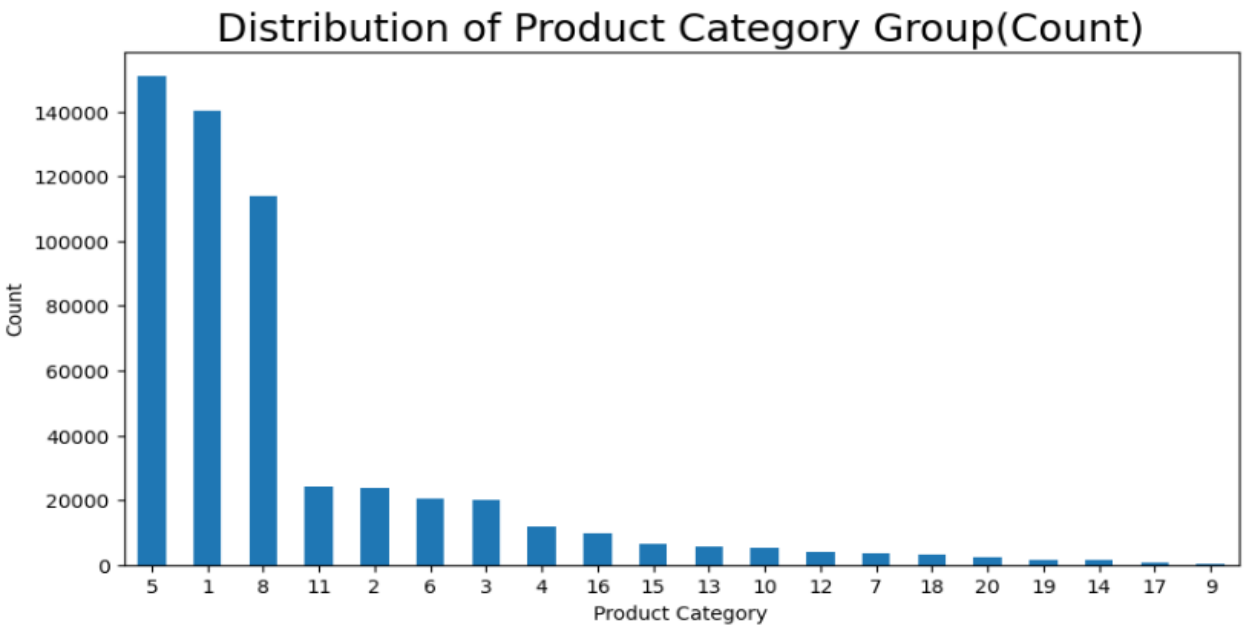
Insight:

- City B has the highest number of transactions, indicating that it is the most active market among the three city categories. This suggests that a significant portion of Walmart's customer base resides in City B, making it a crucial area for focused marketing efforts and resource allocation.
- While City A and City C have fewer transactions compared to City B, they still represent substantial markets with around 150,000 and 160,000 transactions respectively. These numbers indicate a strong customer presence and consistent shopping behaviour in these cities.



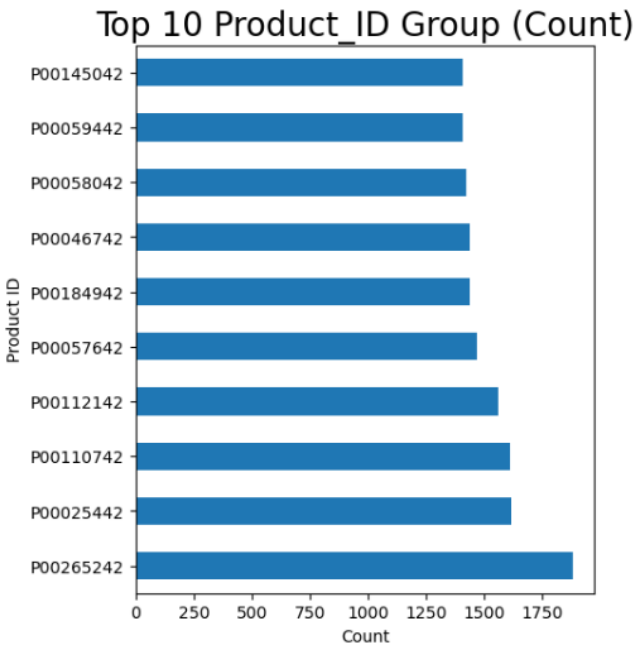
Insights:

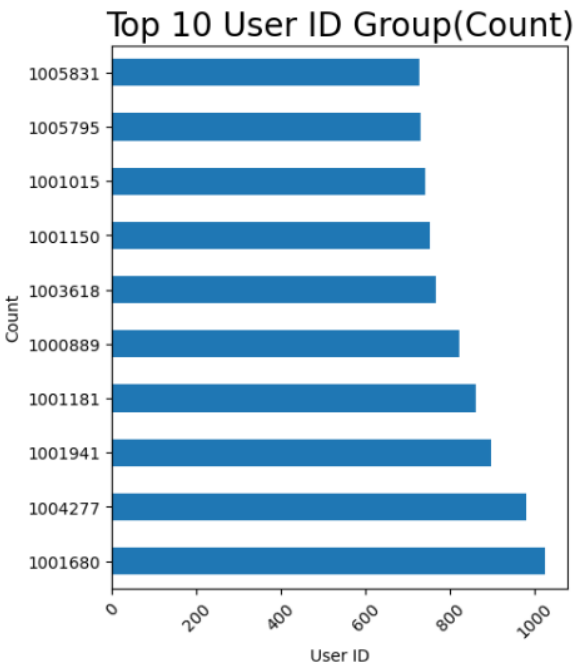
- The largest group of customers has been living in their current city for just 1 year. This indicates a high level of recent mobility among Walmart's customers.
- There is a noticeable decline in the number of customers as the duration of stay increases. This trend suggests that fewer customers remain in the same city for extended periods, with the numbers dropping significantly for those who have been in their city for 2 years, 3 years, and 4+ years.
- A substantial number of customers are new residents who have just moved to their current city (0 years). This could indicate a trend of people moving to new cities and immediately engaging in shopping activities, such as those available at Walmart.



Insights:

- From the above bar graph, we can see that product categories labelled 5, 1, and 8 are the fastest-moving products during Black Friday sales.
- The moderately popular product categories include 11, 2, 6, and 3. However, there is a significant difference in the number of transactions between the high-level and moderate-level product categories.
- Product category 9 is the least popular product during Black Friday sales.



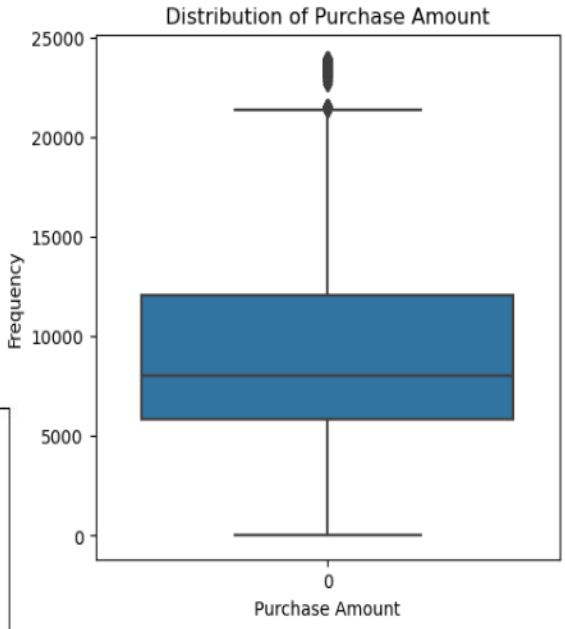


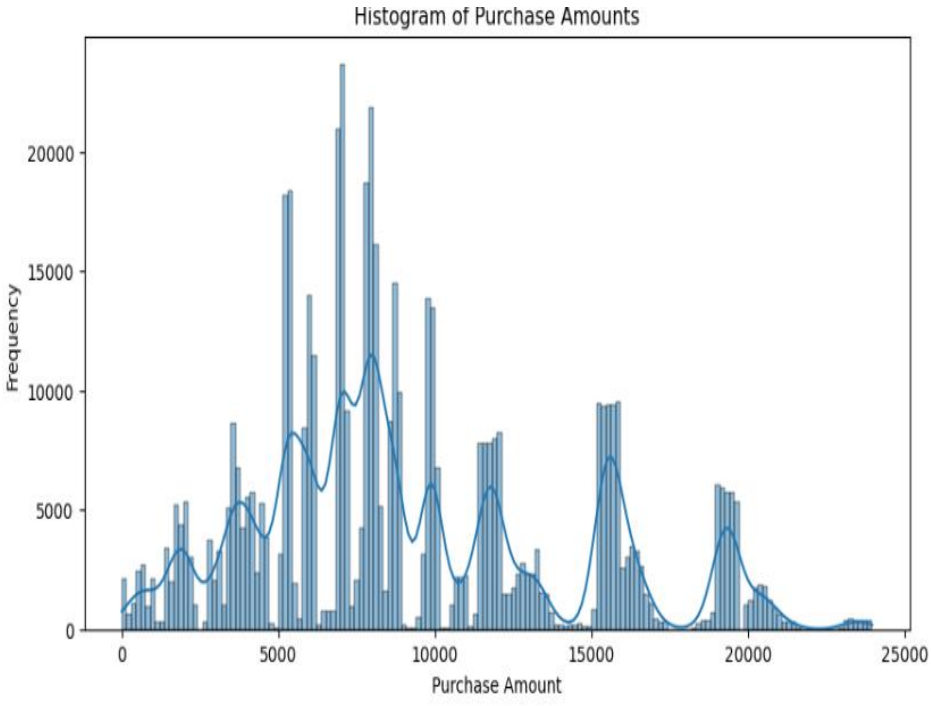
Insights:

- The above chart represents the top 10 Product IDs. The top 10 Product IDs range from 1,406 to 1,880. Product ID P00265242 is the fastest-moving product with 1,880 purchases.
- Similarly, the top 10 User IDs range from 727 to 1,026. User ID 1001680 is the customer who made the highest number of purchases, exceeding 1,000 transactions during Black Friday.

Purchasing Behaviour Analysis:

count	550068
mean	9263.97
std	5023.06
min	12.00
25%	5823.00
50%	8047.00
75%	12054.00
max	23961.00





- The minimum purchase amount observed is $12.00, while the maximum is $23,961.00, indicating a wide range of spending behaviours among customers.
- The mean purchase amount ($9,263.97) falls close to the median ($8,047.00), suggesting a roughly symmetric distribution of purchase amounts, with no significant skewness.
- The standard deviation of $5,023.06 indicates a considerable spread or variability in purchase amounts around the mean. This suggests that there might be significant differences in spending habits among customers.

Percentile Analysis:

- 25% of purchases are below $5,823.00.
- 50% of purchases are below $8,047.00 (median).
- 75% of purchases are below $12,054.00.
- Min is $12.00
- IQR = 75% - 25% (12,054.00 – 5,823) i.e.$6,231.00
- Max is $21400.50
- Outliers: 0.49%

How does gender affect the amount spent?

Confidence interval for average spending of Male customers: (9422.02, 9453.03)
Confidence interval for average spending of Female customers: (8709.21, 8759.92)

Sample Size: 300
Male CI: (9290.76, 10418.49)
Female CI: (8158.76, 9223.01)

Sample Size: 3000
Male CI: (9161.88, 9519.48)
Female CI: (8638.11, 8980.87)

Sample Size: 30000
Male CI: (9380.64, 9495.47)
Female CI: (8700.49, 8808.40)

i. Wider CI for Female Customers:

- Yes, the confidence interval for female customers tends to be wider compared to males, indicating higher variability in spending behaviour among females.

ii. Effect of Sample Size:

- As sample size increases, the width of the confidence interval decreases, indicating more precise estimates.
- Larger sample sizes lead to narrower confidence intervals.

iii. **Overlap of Confidence Intervals**:

- Confidence intervals for different sample sizes overlap for both genders, indicating no statistically significant differences in average spending.

iv. **Effect of Sample Size on Distribution Shape**:

- With increasing sample size, the distributions of the means become narrower and more symmetric around the population mean.

How does Marital Status affect the amount spent?

Confidence interval for average spending of Male customers: (9248.62, 9283.20)
Confidence interval for average spending of Female customers: (9240.46, 9281.89)

Sample Size: 300
Married CI: (8443.99, 9578.56)
Unmarried CI: (8578.08, 9675.96)

Sample Size: 3000
Married CI: (8991.66, 9354.23)
Unmarried CI: (9224.96, 9589.20)

Sample Size: 30000
Married CI: (9189.65, 9303.39)
Unmarried CI: (9229.41, 9343.60)

i. Wider CI for Unmarried Customers:

- Yes, the confidence interval tends to be slightly wider for unmarried customers compared to married customers, indicating slightly higher variability in spending behaviours among unmarried customers.

ii. Effect of Sample Size:

- As sample size increases, the width of the confidence interval decreases, indicating more precise estimates.
- Larger sample sizes lead to narrower confidence intervals.

iii. Overlap of Confidence Intervals:

- Confidence intervals for different sample sizes overlap for all categories, indicating no statistically significant differences in average spending.

iv. Effect of Sample Size on Distribution Shape:

- With increasing sample size, the distributions of the means become narrower and more symmetric around the population mean for all categories.

How does Marital Status affect the amount spent?

Age Group	Sample Size	Confidence Interval (Lower, Upper)	Confidence Interval (Lower, Upper)
0-17	300 3000 30000	(8400.22, 9477.12) (8812.31, 9180.28) (8859.10, 8974.70)	(8851.95, 9014.98)
18-25	300 3000 30000	(8480.17, 9659.91) (9125.76, 9491.88) (9096.36, 9211.66)	(9138.41, 9200.92)
26-35	300 3000 30000	(8537.61, 9663.56) (9031.34, 9390.61) (9250.56, 9264.11)	(9231.73, 9273.65)
36-45	300 3000 30000	(8487.87, 9555.64) (9161.02, 9520.42) (9261.33, 9374.50)	(9301.67, 9361.03)
46-50	300	(8995.58, 10172.21)	(9163.09, 9254.17)

	3000 30000	(9092.74, 9451.44) (9129.64, 9241.98)	
51-55	300 3000 30000	(9261.96, 10400.10) (9220.97, 9578.41) (9478.12, 9593.40)	(9483.99, 9585.62)
55+	300 3000 30000	(8533.13, 9617.47) (9141.83, 9502.41) (9280.45, 9394.58)	(9269.30, 9403.26)

i. Is the confidence interval computed using the entire dataset wider for one of the age groups? Why is this the case?

Yes, the confidence interval computed using the entire dataset is wider for some age groups.

The reason for the wider intervals is due to the larger variability in the spending patterns within these age groups. This larger variability can be attributed to diverse income levels, spending habits, and financial responsibilities among individuals within these groups.

ii. How is the width of the confidence interval affected by the sample size?

The width of the confidence interval decreases as the sample size increases. This is evident in the following examples:

Age 0-17:

- Sample (300): (8400.22, 9477.12) -> Width: 1076.90
- Sample (3000): (8812.31, 9180.28) -> Width: 367.97
- Sample (30000): (8859.10, 8974.70) -> Width: 115.60

Age 18-25:

- Sample (300): (8480.17, 9659.91) -> Width: 1179.74
- Sample (3000): (9125.76, 9491.88) -> Width: 366.12
- Sample (30000): (9096.36, 9211.66) -> Width: 115.30

As sample size increases, the margin of error decreases, leading to narrower confidence intervals. This is because larger sample sizes provide more information, reducing the standard error of the mean.

iii. Do the confidence intervals for different sample sizes overlap?

Yes, the confidence intervals for different sample sizes do overlap, though the extent of overlap decreases as the sample size increases and the intervals become narrower. For example:

Age 26-35:

- Sample (300): (8537.61, 9663.56)
- Sample (3000): (9031.34, 9390.61)
- Sample (30000): (9250.56, 9264.11)
- Population CI: (9231.73, 9273.65)

There is overlap between the intervals for the different sample sizes, indicating that despite the variation in sample sizes, the estimates are consistent.

- Some sample confidence intervals are narrower than the population confidence intervals. This is particularly noticeable for larger sample sizes (e.g., 30,000) and the age group 0-17, 26-35, 46-50, and 55+.

iv. How does the sample size affect the shape of the distributions of the means?

As the sample size increases, the shape of the distribution of the sample means becomes more normal, and the distribution's spread decreases. This results in:

More Normal Distribution: Due to the Central Limit Theorem, as sample size increases, the distribution of the sample means tends to approximate a normal distribution, regardless of the population distribution.

Reduced Spread: With larger sample sizes, the variability (standard deviation) of the sample means decreases, resulting in narrower confidence intervals.

For example, in the age group 51-55:

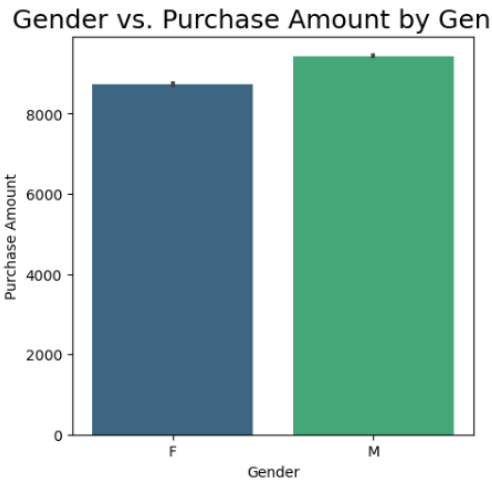
Sample 300: (9261.96, 10400.10) -> Width: 1138.14

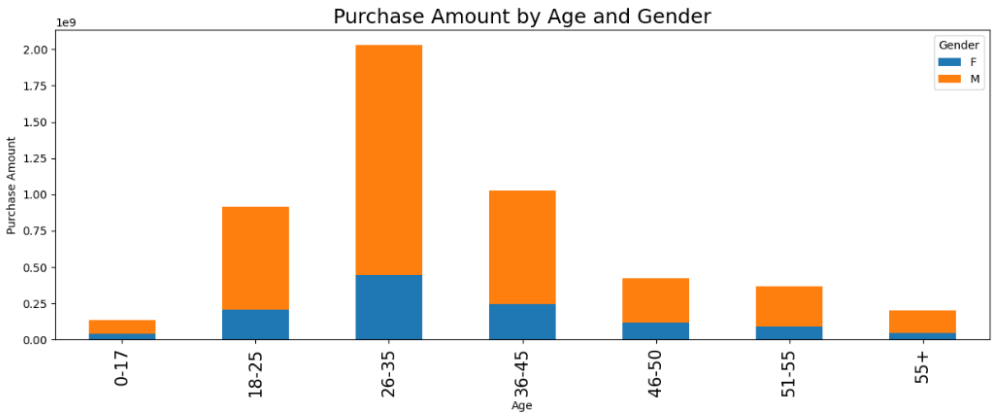
Sample 3000: (9220.97, 9578.41) -> Width: 357.44

Sample 30000: (9478.12, 9593.40) -> Width: 115.28

This decreasing width with increasing sample size illustrates a more precise and reliable estimate of the population mean, aligning with the properties described by the Central Limit Theorem.

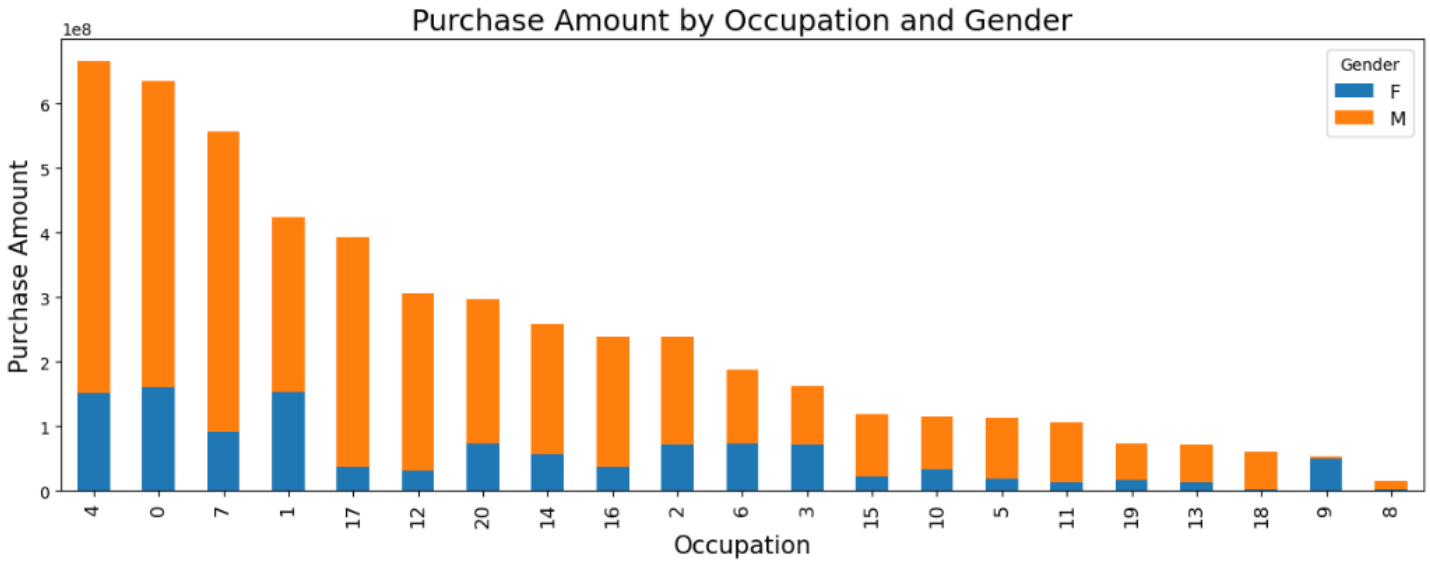
Bivariate and Multivariate Analysis:





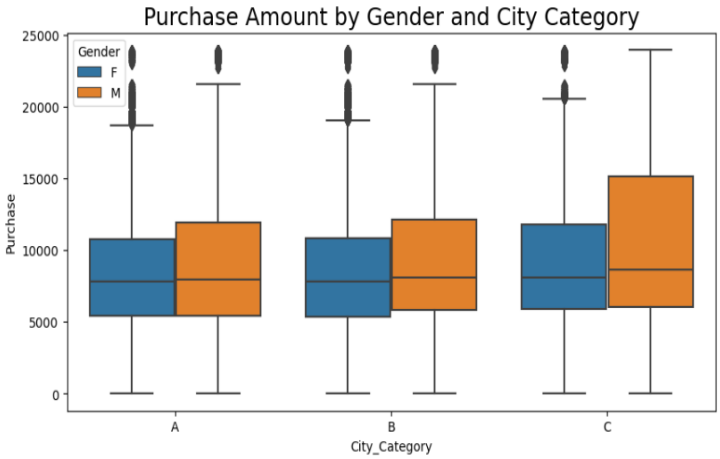
Insights:

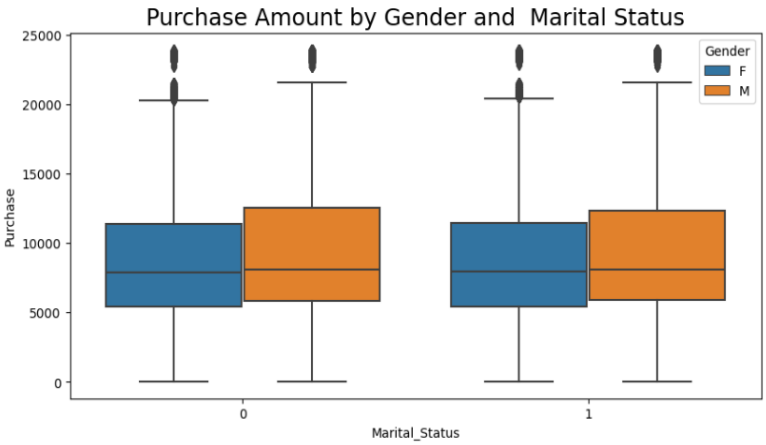
- From the above bar chart, we can see that the total purchase amount made by males is higher than that made by females.
- Similarly, in every age group, the purchase amount by females is very much less than that by males.



Insight:

- The contribution of males is generally higher compared to females across most occupations, except for occupation 9, which stands out despite being one of the occupations with the lowest overall contribution.





Insights:

- The purchase ratio between City A and B shows a similar pattern, with almost equal proportions of both genders. However, residents from City C exhibit a higher purchase ratio. Despite this, the median purchase value across all cities and genders appears to be nearly identical, with females slightly lower than males.
- From the second graph, it appears that there is no significant difference in purchase values based on marital status and gender.

Recommendation:

- Design promotions and marketing campaigns that resonate with female shoppers. Highlight products that are popular among women or offer special discounts on items typically preferred by female customers.
- Create a welcoming and comfortable shopping environment within stores, with dedicated sections featuring products relevant to female shoppers. Consider adding amenities such as seating areas or beauty stations.
- Create loyalty programs with rewards and incentives tailored to both male and female shoppers' preferences. Offer rewards such as exclusive access to new tech releases, gaming events, or sports tickets for male and skincare, cosmetics, and wellness items for female customers.
- Establish channels for collecting feedback from both male and female customers. Actively listen to their suggestions and preferences to improve product offerings and shopping experiences.