Colab notebook link:
https://colab.research.google.com/drive/1t3bGHjXkVk1flmixVbUOoaEiUfYiJht3?usp=sharing

Insights:

Analyzing the dataset:

There are total 550068 rows and 10 columns in the dataset.

Columns are: ['User_ID', 'Product_ID', 'Gender', 'Age', 'Occupation', 'City_Category', 'Stay_In_Current_City_Years', 'Marital_Status', 'Product_Category', 'Purchase'].

Data type of columns: 5 integer type columns, 5 object type columns given in the dataset. But we can observe that other than 'Purchase' column, all of the rest columns are categorical type. So, it's better to change the data types of all such columns into category.

Total no. of unique 'User_ID' is: 5891.

Total no. of unique product is: 3631.
And highest selling product is: P00265242 with a maximum of 1880 units sold.

There are two types of gender: Male and Female. And frequency of buying products by male is greater than female.

Total no. of unique age group is: 7.
Age-groups are: ['0-17', '18-25', '26-35', '36-45', '46-50', '51-55', '55+'].
Highest transactions are done by age-group: '26-35'.

Total no. of unique occupation is: 21 and it is categorized by '0' to '20'.

Total three types of city category: 'A', 'B' and 'C'. And highest transactions have done by city category 'B'.

Customer staying in a same city are categorized by 5 unique category: '0','1','2',3, and '4+'.
And maximum transactions are accounted by Customers with '1' year of stay in current city.

Marital statuses are categorized by '0' and '1'. '0' for unmarried and '1' for married.
And maximum transactions were done by Unmarried Customers.

Total no. of unique product category is: 20, which is categorized by '0' to '20'.

The minimum purchased amount is $12 and maximum purchased amount is $23961.
The median purchase amount of $8047 is notably lower than the mean purchase amount of $9264, indicating that there are outliers.

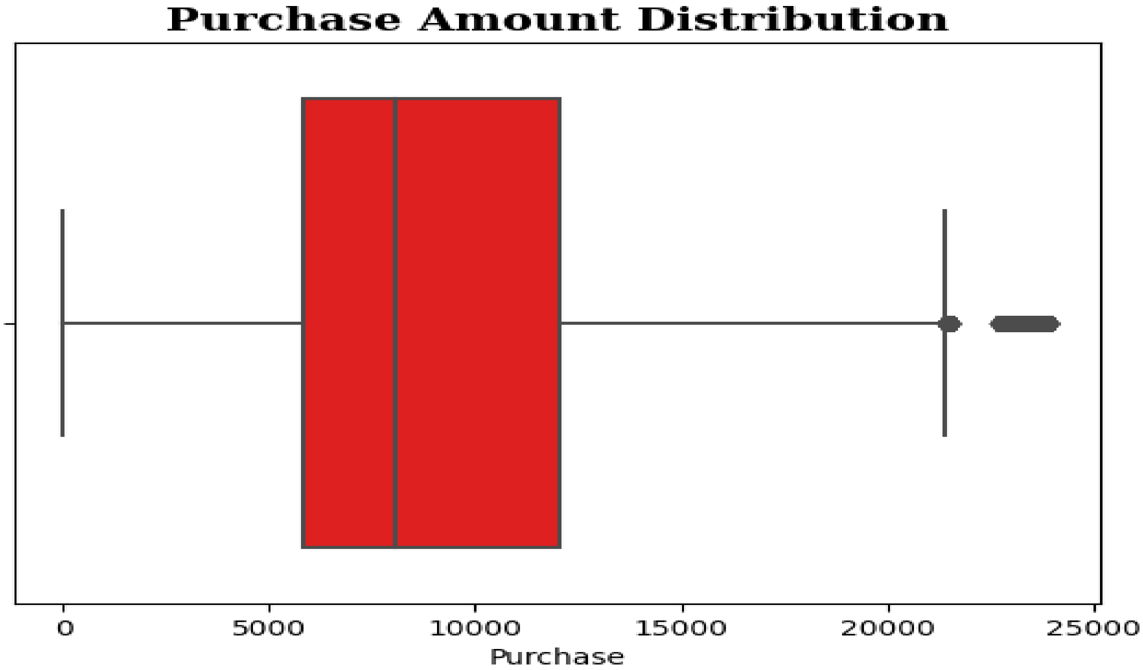
Missing Values & Duplicates:

There are no missing values in the dataset.

There are no duplicates in the dataset.

Detect Outliers:

Data suggests that the majority of customers spent between 5,823 USD and 12,054 USD, with the median purchase amount being 8,047 USD.
And mean of all the purchases is 9,264 USD (approx.).
Difference between mean and median is 1,217 (approx). It means 'Purchase' have more outliers.
We also can detect the outliers by box plot chart.

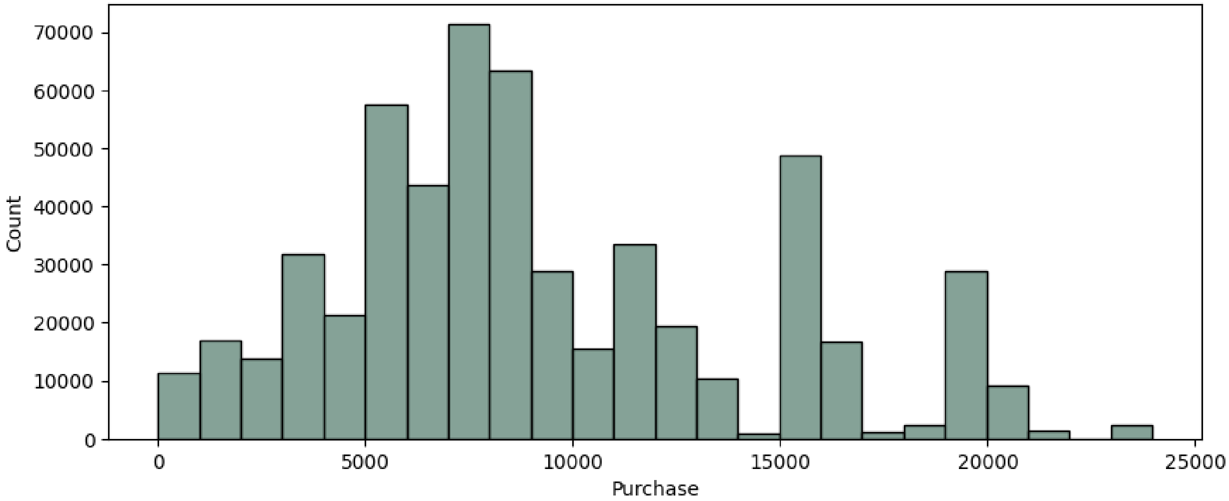


Replacing the values of Marital_Status column:

Replacing the values of marital status into ‘married’ and ‘unmarried’ for better understanding.

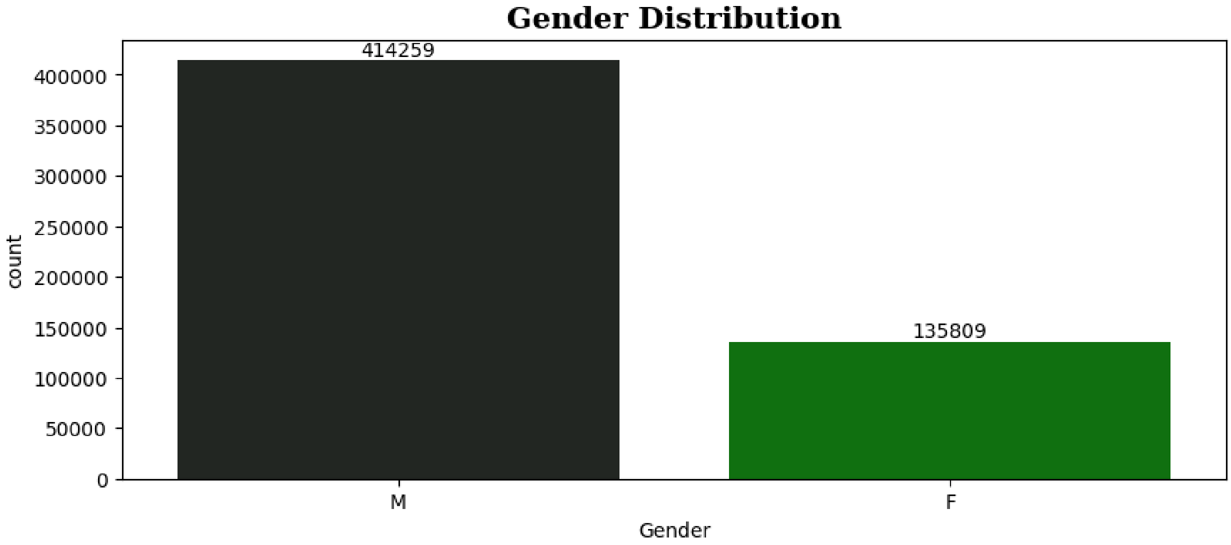
Univariate Analysis:

Continuous Variable



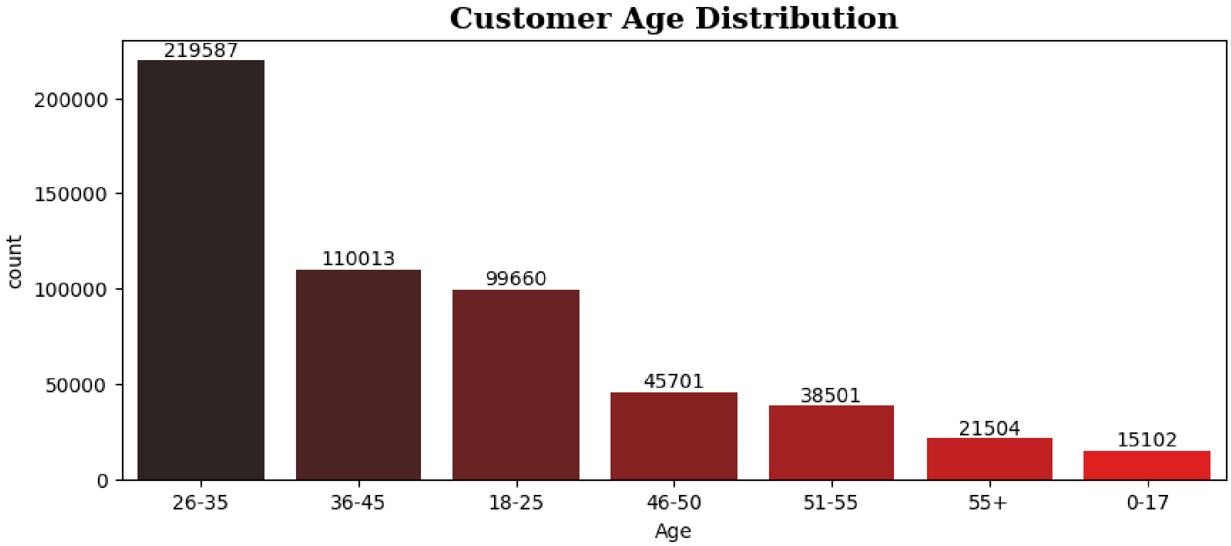
From the above chart, we can consider that the majority of customer spent between $5000 USD and $10000 USD (approx).

Categorical Variables



Gender Vs Purchase:

We can clearly observe that, there is a huge difference in purchase behavior between male and female during Black Friday event. Most of the users are male.

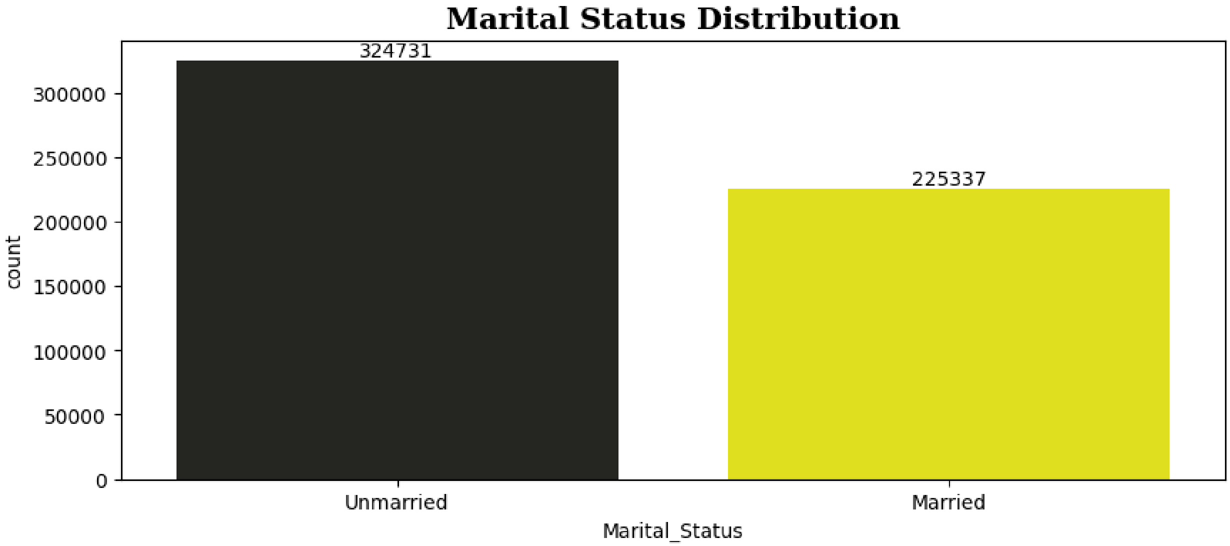


Age Vs Purchase:

The age group of 26-35 dominating the purchase chart with highest percentage. It indicates that the young and middle-aged adults are the most active and interested in shopping for deals and discounts.

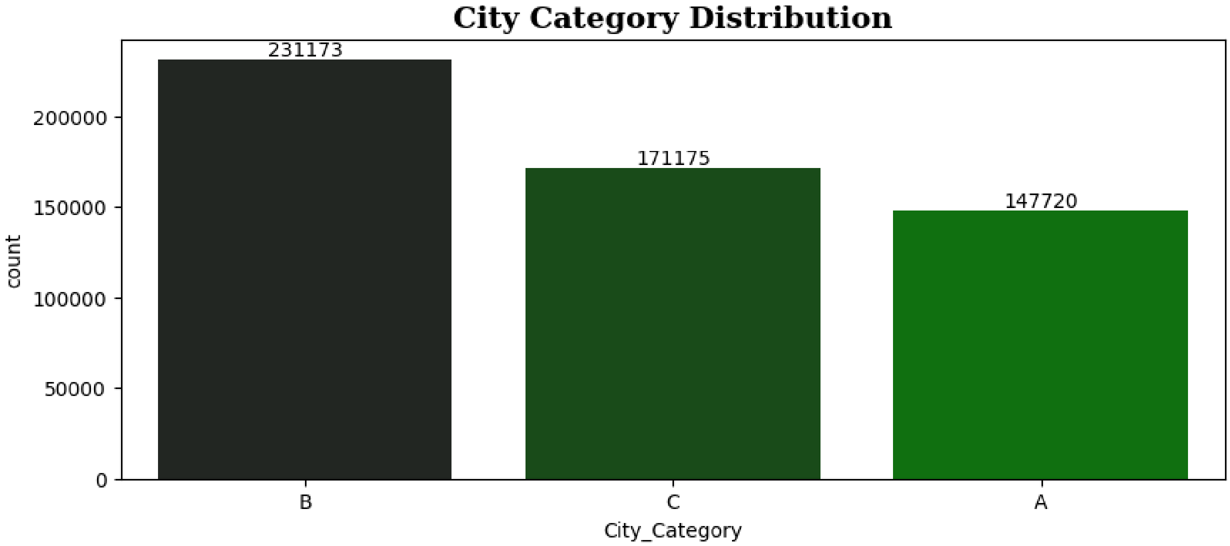
The second and third positions are captured by 36-45 and 18-25 age groups respectively.

It indicates that, the young and adult aged groups were active and interested in shopping during Black Friday event.



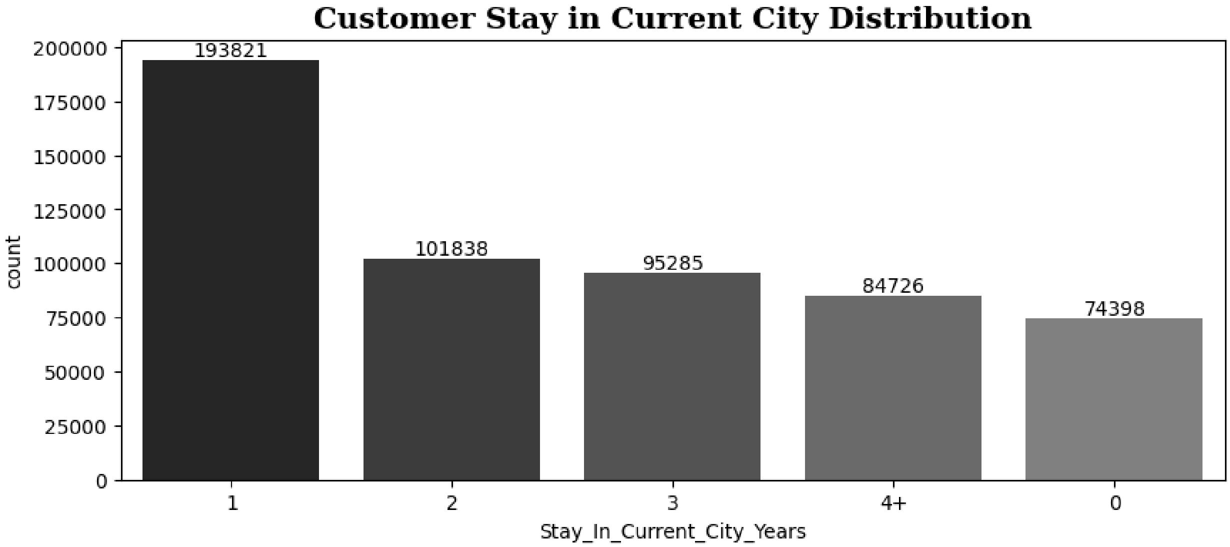
Marital Status Vs Purchase:

Unmarried customers purchasing more with respect to married customers during Black Friday event.



City Category Vs Purchase:

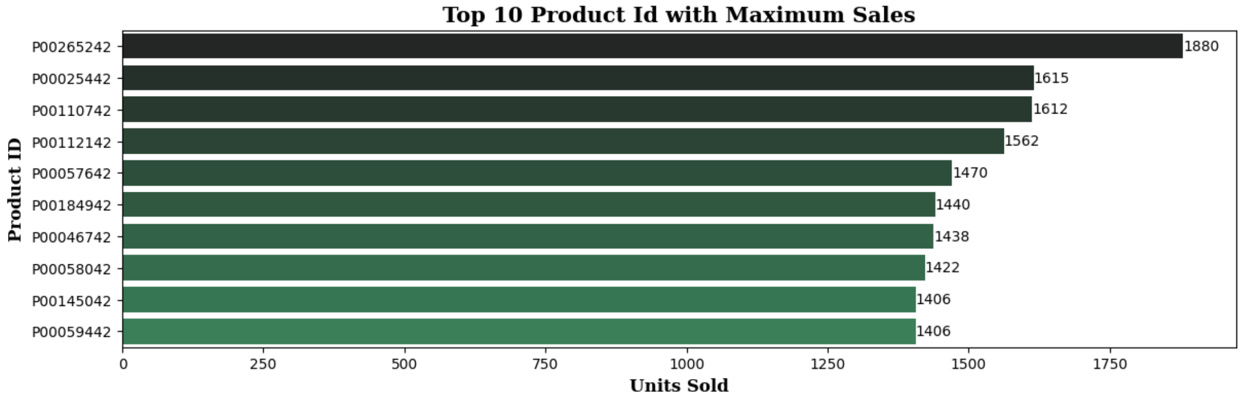
We can observe that city B is the most number of transactions followed by city C and city A respectively.



Current City Vs Purchase:

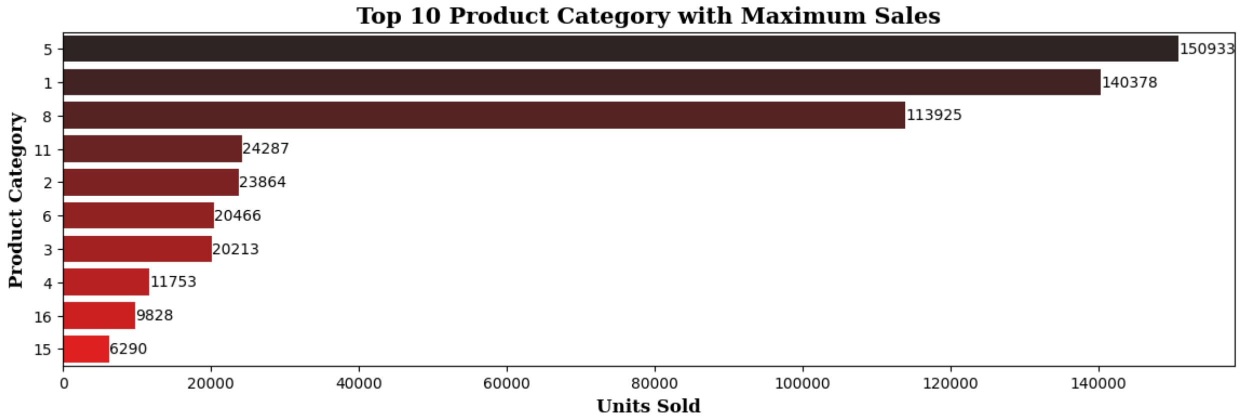
We can observe that most of the customers stayed in current city for 1 year. And if we add it to less than 1 year customers then it's a huge number. It indicates that, the Walmart need to showing strong interest to newcomers who may be looking for affordable and convenient shopping options.

The percentage of customers decreases as the stay in the current city increases, but they are the loyal customers for Walmart's in respective to purchasing. It means that Walmart may also get benefit by targeting long-term residents with loyalty programs and promotions.



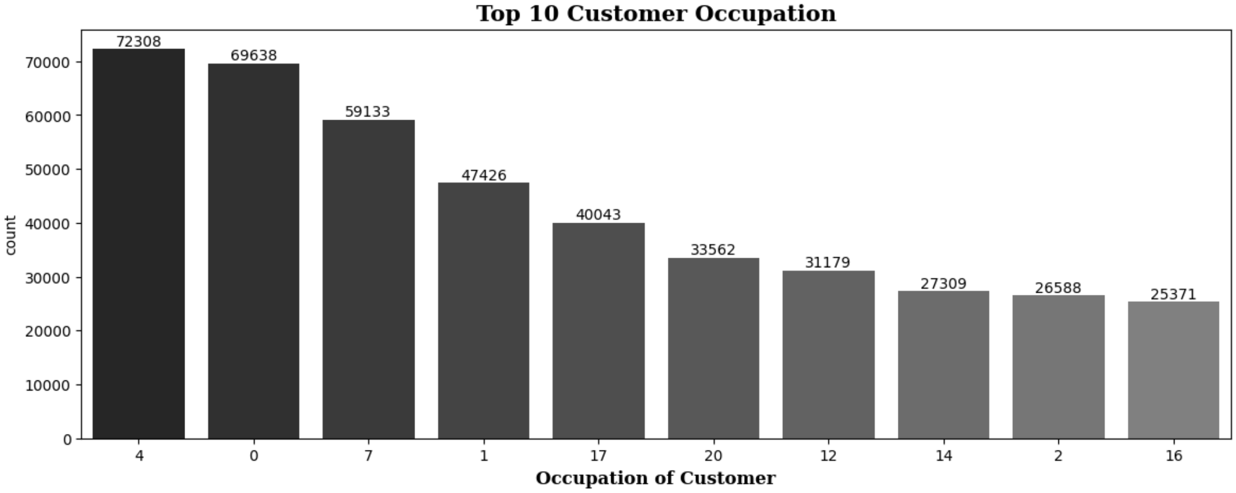
Top 10 Products Sold:

Top selling product during Black Friday event is P00265242 with 1880 sales, which is followed by P00025442 and P00110742 with 1615 and 1612 respectively. There is not much more difference in Top 10 products list, which suggesting that Walmart offers a variety of products that many different customers like to buy.



Top 10 Product Categories:

Top three product categories are 5, 1 and 8 with combined sales is nearly 75 %(approx.) of the total sales. It means that these three products are most preferable products among customers during Black Friday event.



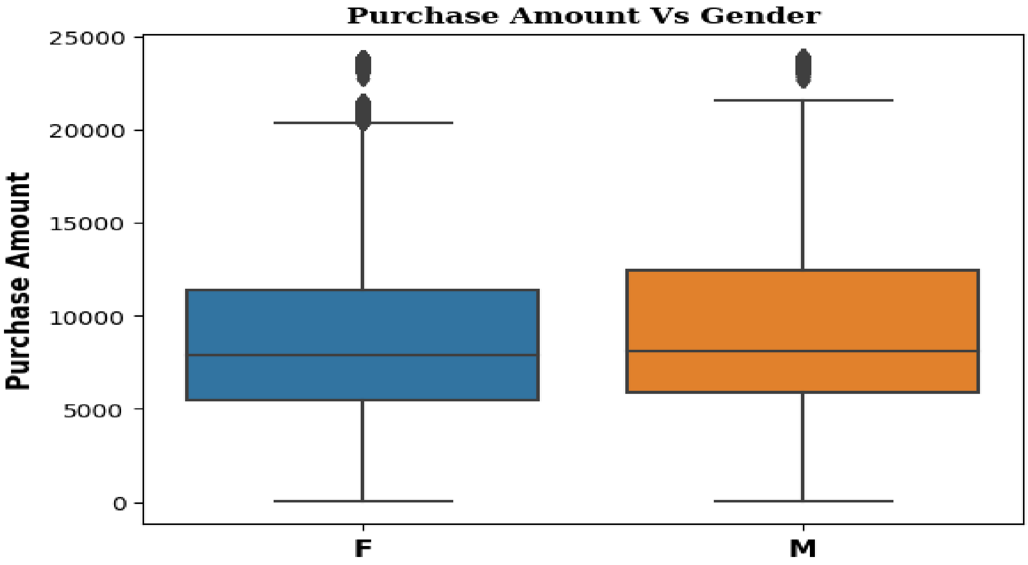
Customer Occupation Vs Purchase:

The above chart shows that the Top 10 customer occupations whose are more likely to purchase from Walmart. Customers with Occupation category 4, 0 and 7 are top three occupations in this chart.
It means that these occupations have a high demand for Walmart products or services.

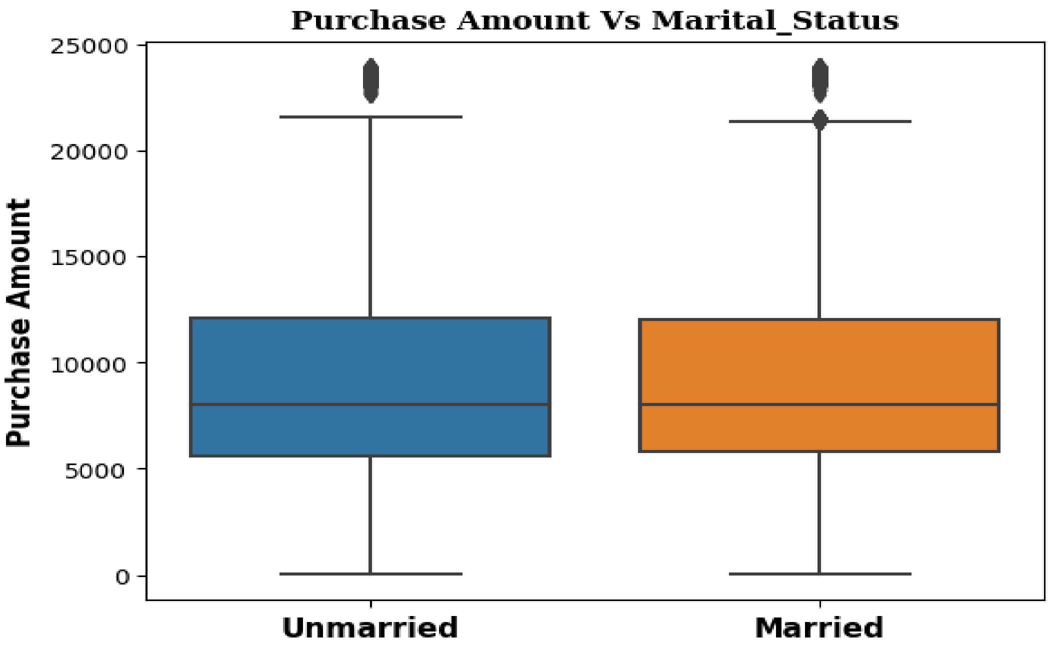
Bivariate Analysis:

Exploring purchase pattern by box plots.

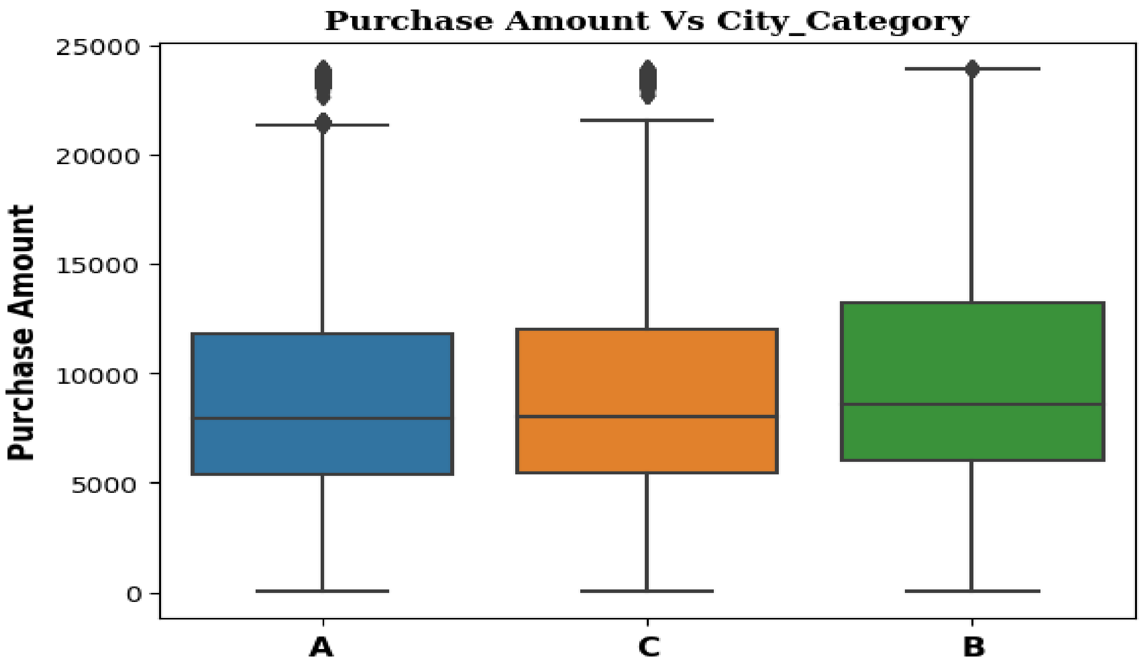
Purchase Amount Vs Gender



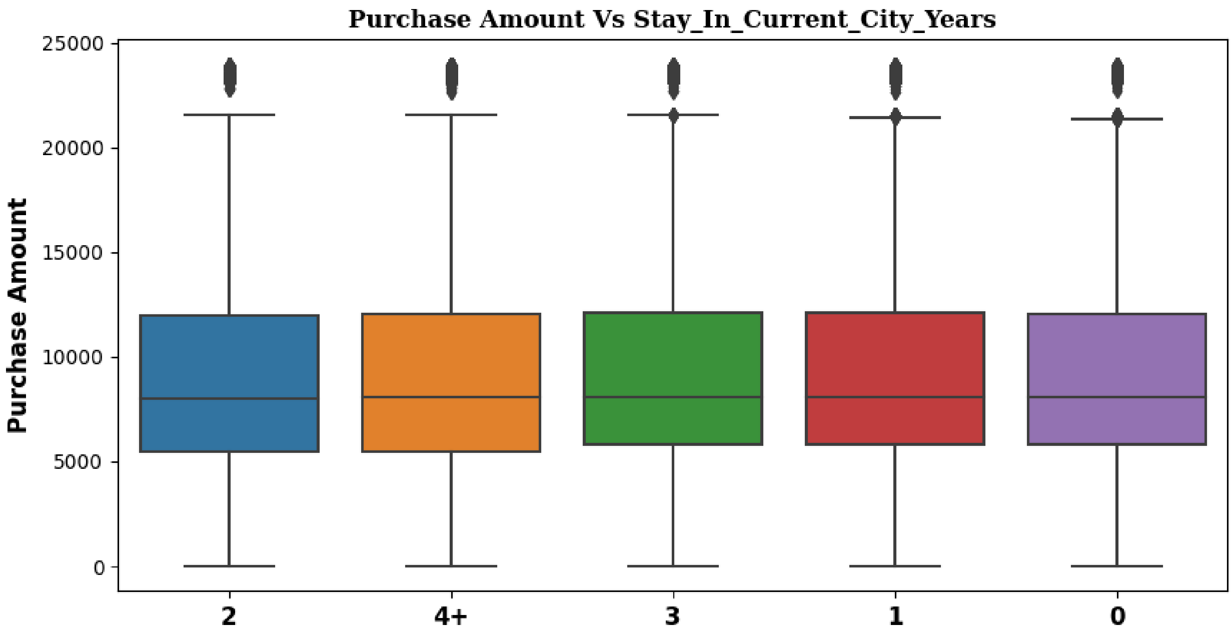
Purchase Amount Vs Marital Status



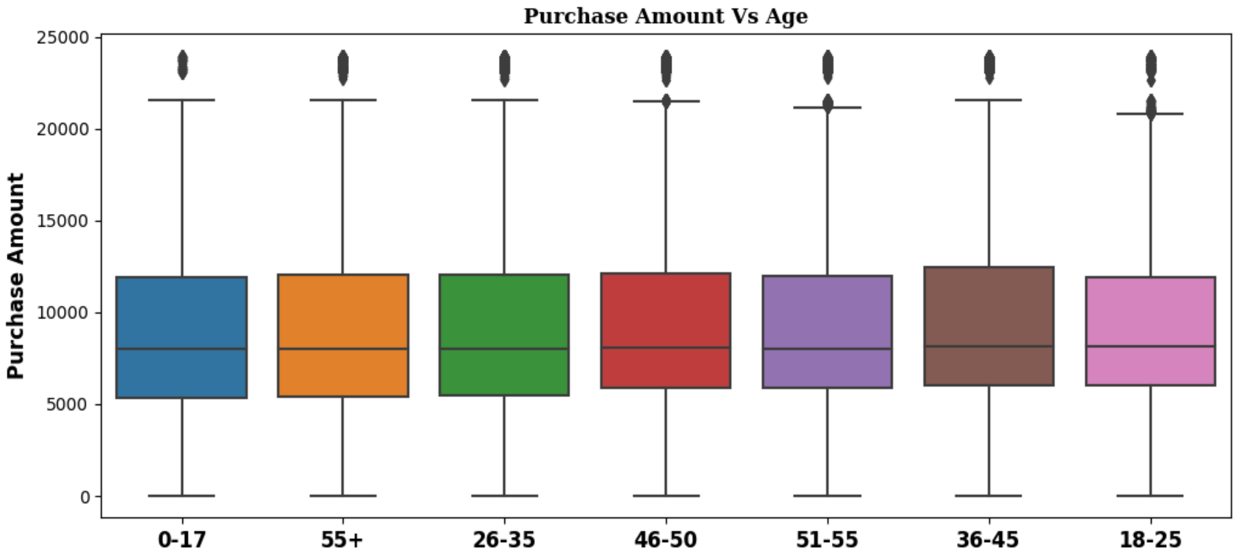
Purchase Amount Vs City Category



Purchase Amount Vs Stay in Current City Years



Purchase Amount Vs Age

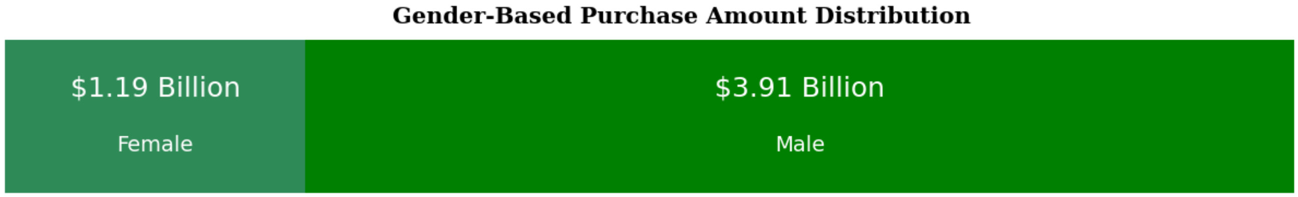


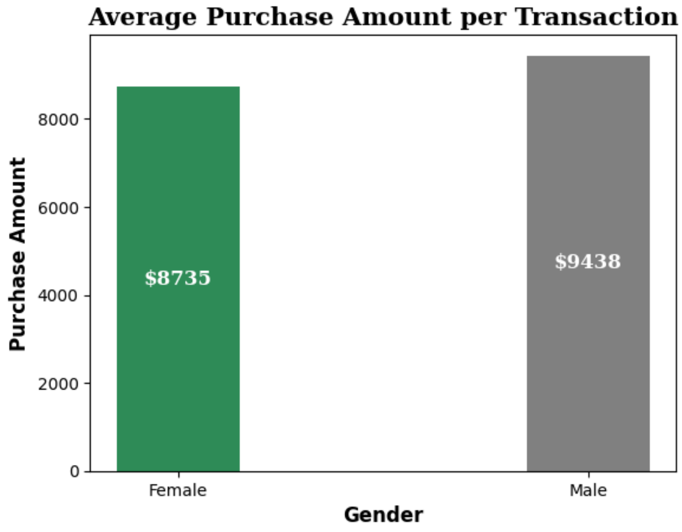
We can observe from the above box plot charts, that there is not much more variation in purchase amount. The purchase amount relatively stable regardless of the variable under consideration.

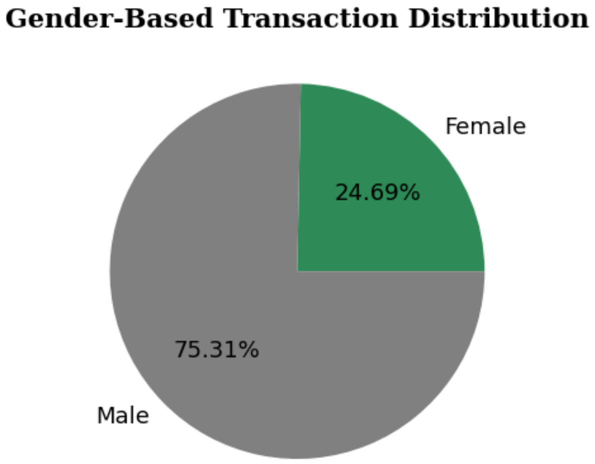
Gender VS Purchase Amount:

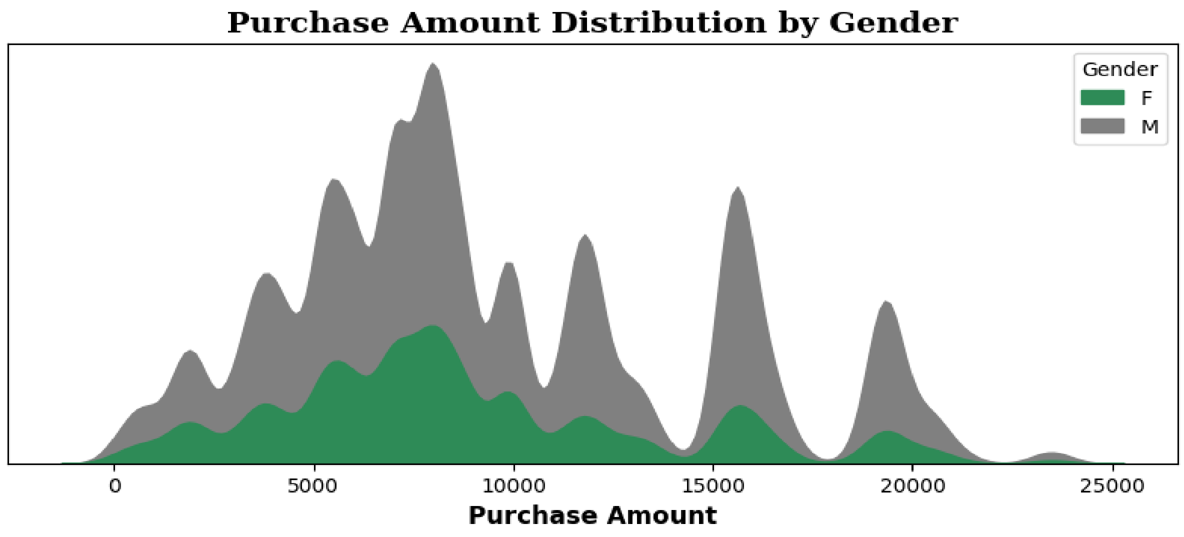
Data Visualization:
Sample Checking Gender-wise

For better understanding taking the 50 thousand sample of data and aggregate the data gender wise.
Compare the sample data with actual data, we can find out that the amount of money spent by female is less than male. Also percentage of sum is same for both the cases and gender-wise per person purchase (mean) is nearly same.


Gender-Based Purchase Amount Distribution


Average Purchase Amount per Transaction


Gender-Based Transaction Distribution


Purchase Amount Distribution by Gender

Total Purchase and Transactions Comparison

The total purchase amount and number of transactions by male customers was three times more than female customers. It's indicating that male customers have made a significant impact on the Black Friday sales.

Average Transaction Value

The average purchase amount per transaction was slightly higher for male customers than female customers.

Distribution of Purchase Amount

The purchase amount distribution for both the genders is not normally distributed.

Question: Are women spending more money per transaction than men?

Answer: If we compare average purchase amount per transaction gender-wise, then we can say that the women spending slightly less money per transaction than men. The reason behind this, the number of transaction by men is nearly three times of the number of transactions by women.

Confidence Interval and CLT:

CLT(Central Limit Theorem)

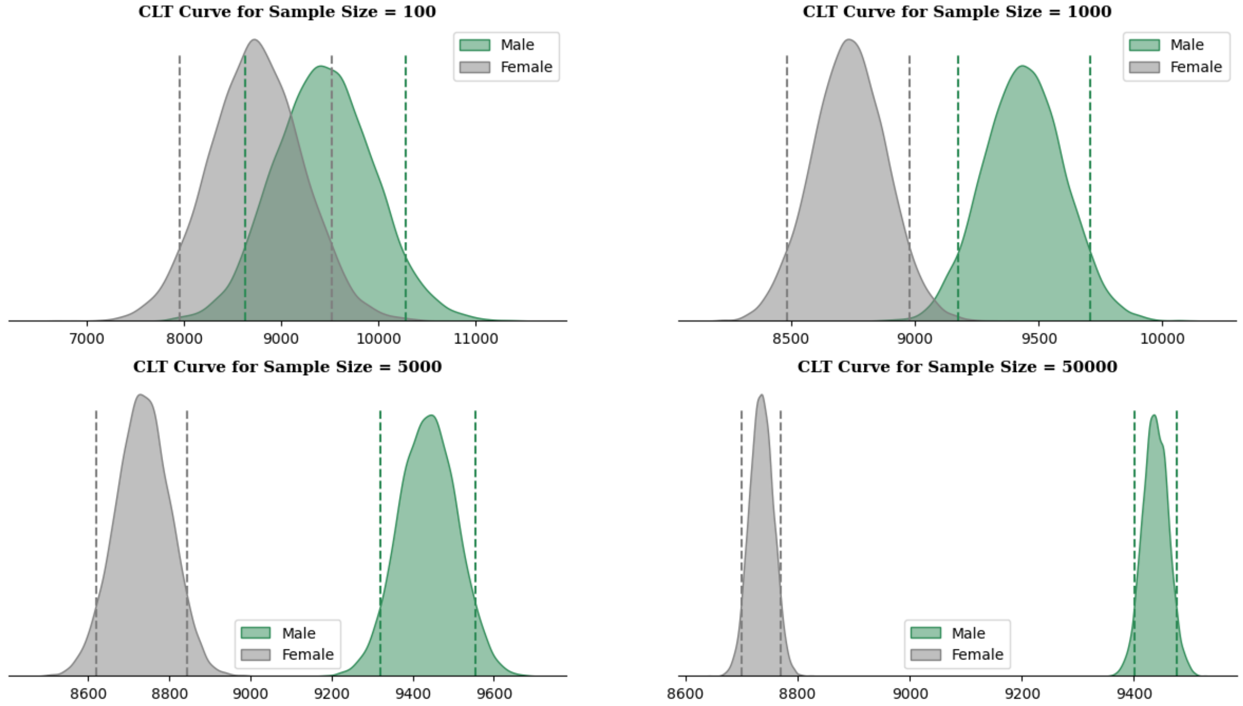
The purchase amount distribution is not normal. So, we need to use central limit theorem. It states that the sampling distribution of a sample mean is approximately normal if the sample size is large enough, even if the population distribution is not normal.

We are using different sample size for building CLT curve.
Sample sizes are: [500, 1000, 5000, 50000].

Confidence Interval

After building CLT curve of different samples, we will create a confidence interval predicting population mean at 99%, 95% and 90% Confidence level.

90% Confidence Interval

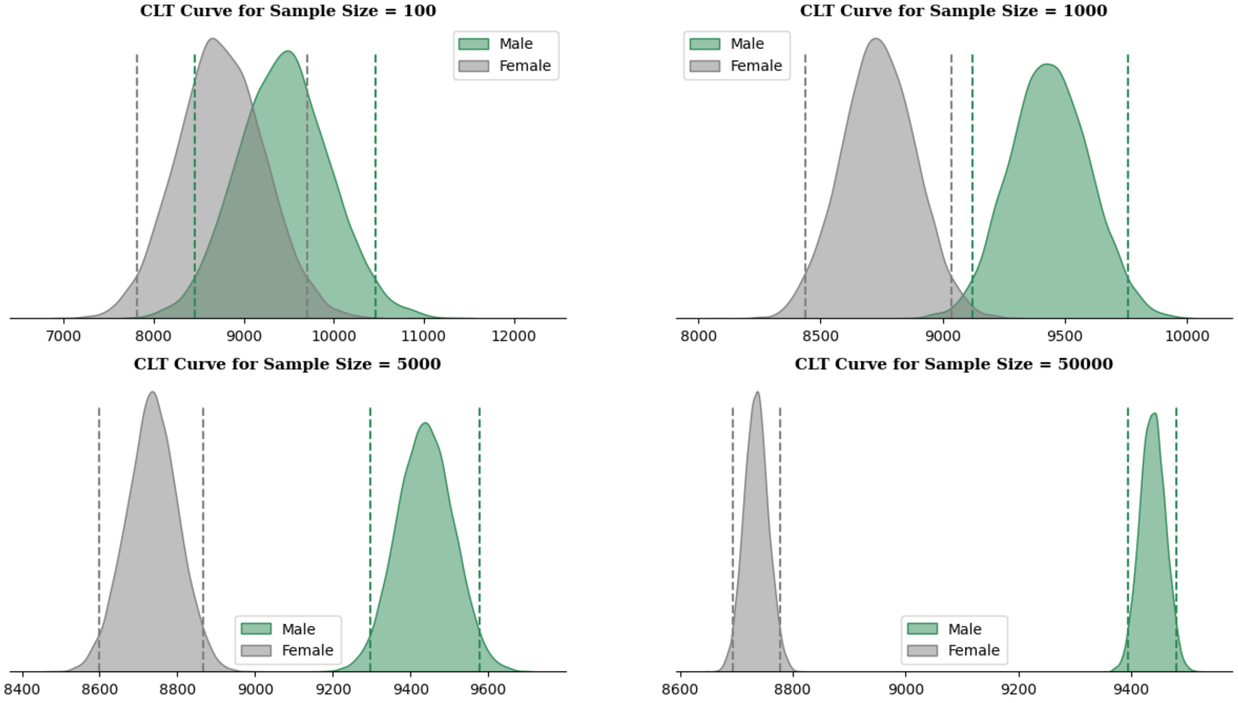

CLT Curve for Sample Size = 100

CLT Curve for Sample Size = 1000

CLT Curve for Sample Size = 5000

CLT Curve for Sample Size = 50000

95% Confidence Interval

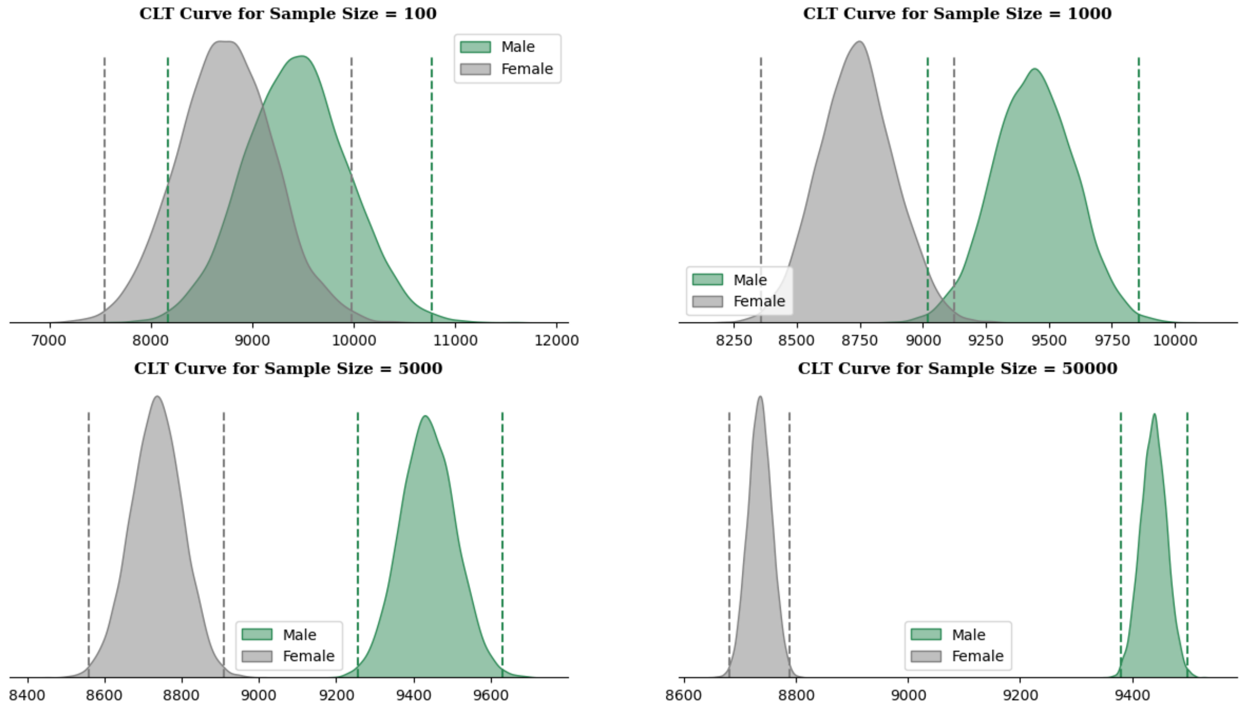

CLT Curve for Sample Size = 100

CLT Curve for Sample Size = 1000

CLT Curve for Sample Size = 5000

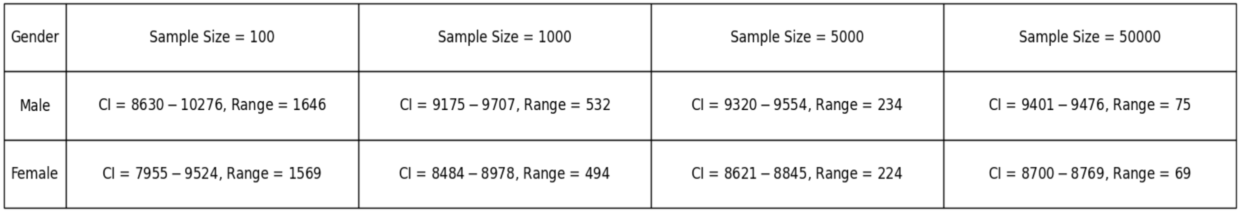
CLT Curve for Sample Size = 50000

99% Confidence Interval



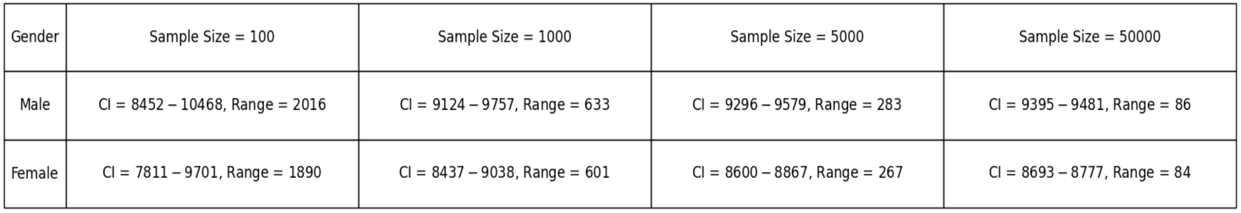
Question: Are confidence intervals of average male and female spending overlapping?

90% Confidence Interval Summary



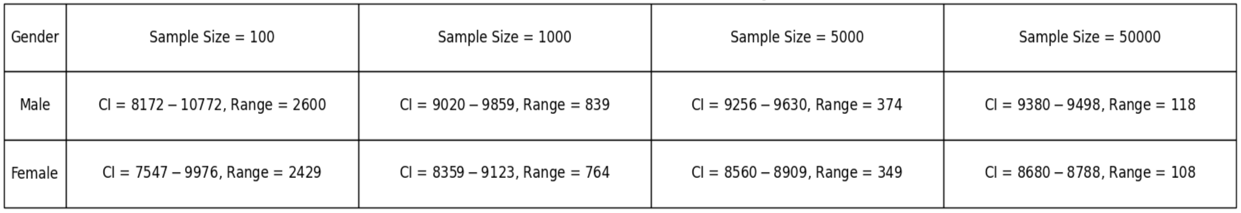
Gender	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
Male	CI = 8630 – 10276, Range = 1646	CI = 9175 – 9707, Range = 532	CI = 9320 – 9554, Range = 234	CI = 9401 – 9476, Range = 75
Female	CI = 7955 – 9524, Range = 1569	CI = 8484 – 8978, Range = 494	CI = 8621 – 8845, Range = 224	CI = 8700 – 8769, Range = 69

95% Confidence Interval Summary



Gender	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
Male	CI = 8452 – 10468, Range = 2016	CI = 9124 – 9757, Range = 633	CI = 9296 – 9579, Range = 283	CI = 9395 – 9481, Range = 86
Female	CI = 7811 – 9701, Range = 1890	CI = 8437 – 9038, Range = 601	CI = 8600 – 8867, Range = 267	CI = 8693 – 8777, Range = 84

99% Confidence Interval Summary



Gender	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
Male	CI = 8172 – 10772, Range = 2600	CI = 9020 – 9859, Range = 839	CI = 9256 – 9630, Range = 374	CI = 9380 – 9498, Range = 118
Female	CI = 7547 – 9976, Range = 2429	CI = 8359 – 9123, Range = 764	CI = 8560 – 8909, Range = 349	CI = 8680 – 8788, Range = 108

From the above analysis, we can see that except for the Sample Size of 100, the confidence interval does not overlap as the sample size increases. It means that as the sample size increases, the confidence intervals become narrower and more precise

For 90% confidence interval

For Female (sample size 50000) range for mean purchase with confidence interval 90% is [8700, 8769]
For Male (sample size 50000) range for mean purchase with confidence interval 90% is [9401, 9476]

For 95% confidence interval

For Female (sample size 50000) range for mean purchase with confidence interval 95% is [8600, 8867]
For Male (sample size 50000) range for mean purchase with confidence interval 95% is [9296, 9579]

For 99% confidence interval

For Female (sample size 50000) range for mean purchase with confidence interval 99% is [8560, 8909]
For Male (sample size 50000) range for mean purchase with confidence interval 99% is [9256, 9630]

Question: How can Walmart leverage this conclusion to make changes or improvements?

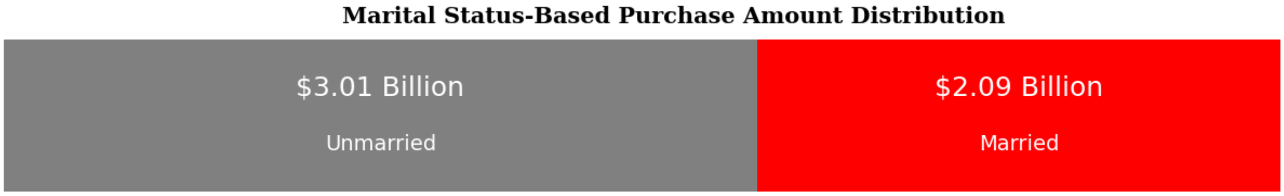
Answer: Walmart can take the following steps to leverage this conclusion:

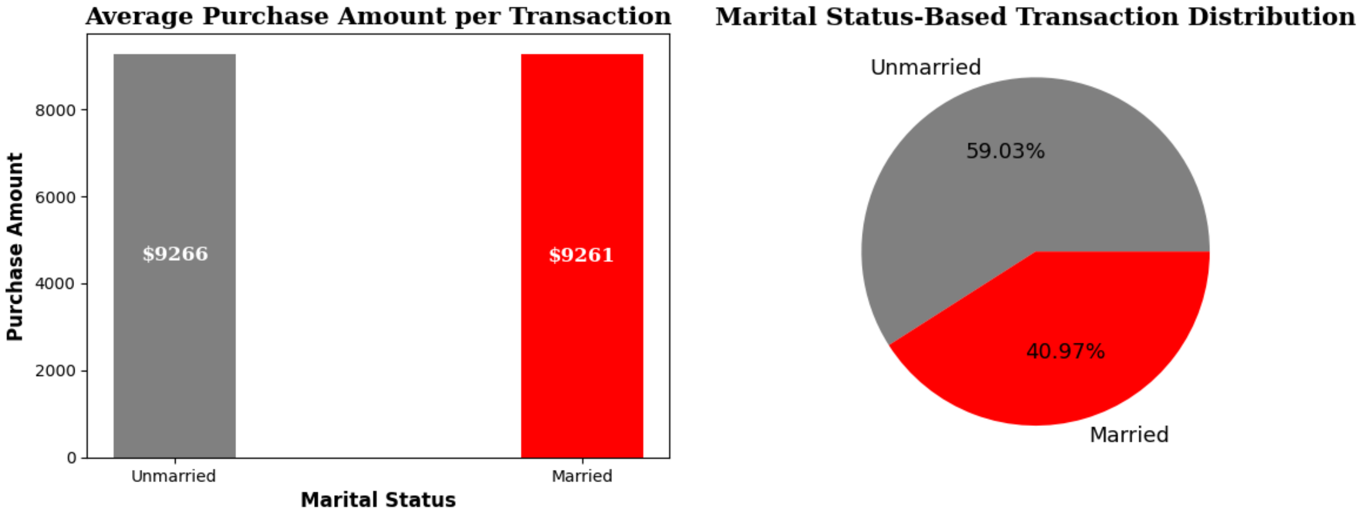
- Based on gender spending behavior of customers, Walmart can create targeted marketing campaigns, loyalty programs and bundles of products. It can increase the purchase amount for each customer and help to maximize the revenue for Walmart's.

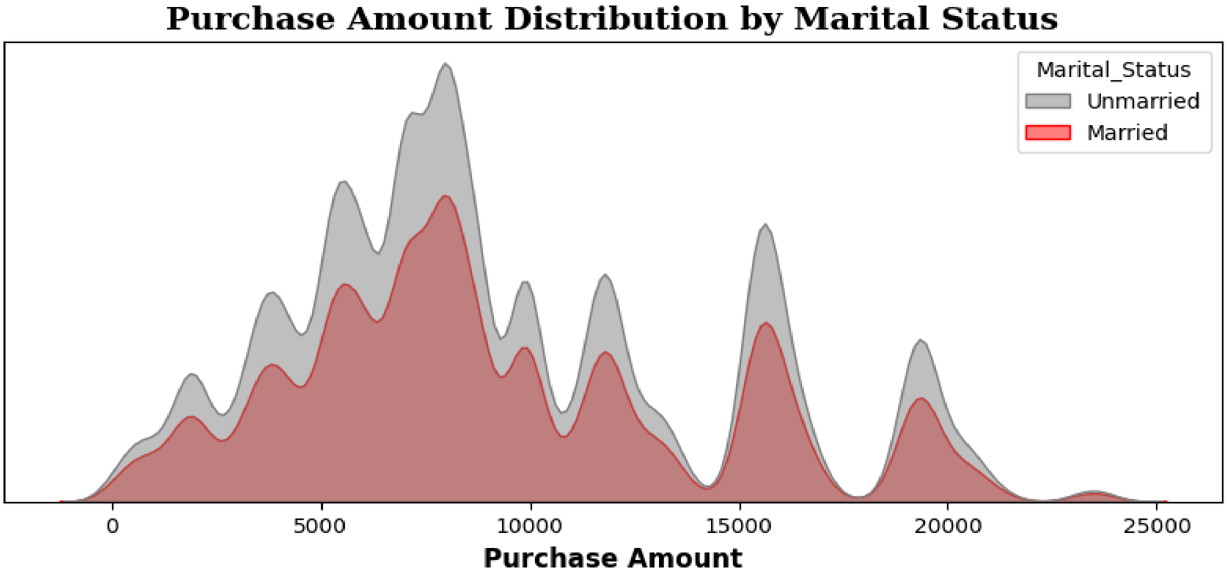
- Also based on the average spending per transaction by gender, Walmart can start discount or price adjusting strategies to incentivize customers for higher spending.

Marital Status VS Purchase Amount:

Data Visualization:







Total Purchase and Transactions Comparison

The total purchase amount and number of transactions by unmarried customers was nearly 1.5 times of married customers. It's indicating that unmarried customers made a significant impact on the Black Friday sales.

Average Transaction Value

The average purchase amount per transaction was almost similar for married and unmarried customers, $9266 for unmarried customers and $9261 for married customers.

Distribution of Purchase Amount

The purchase amount distribution for both married and unmarried customers is not normally distributed.

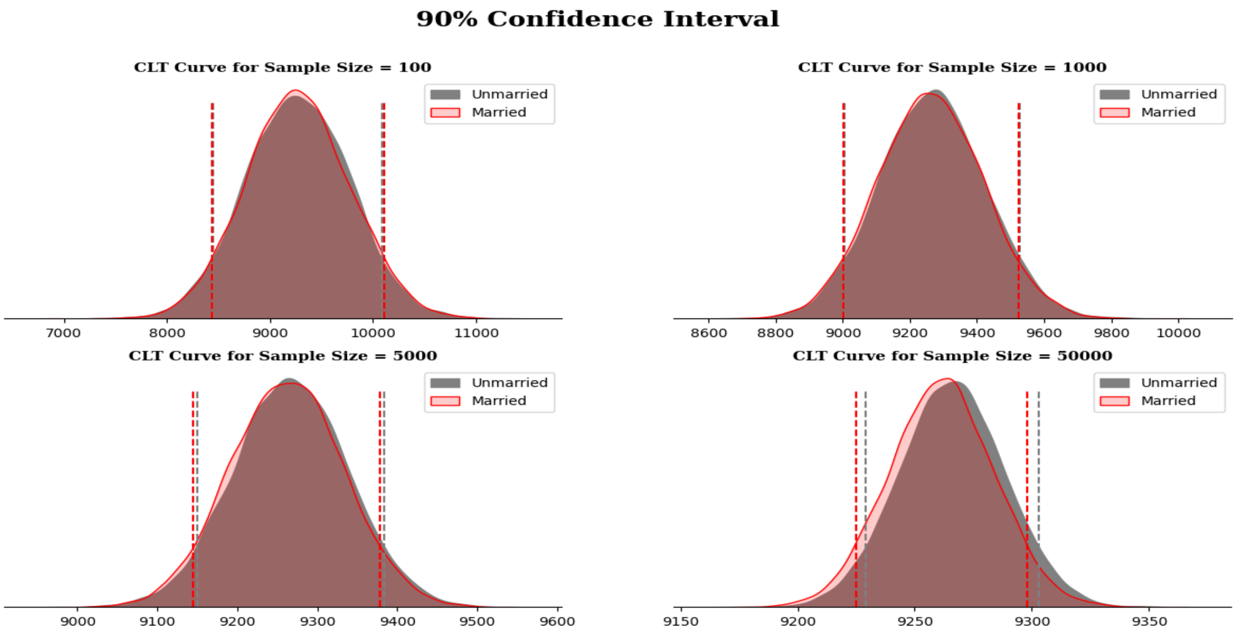
Confidence Interval and CLT:

CLT(Central Limit Theorem)

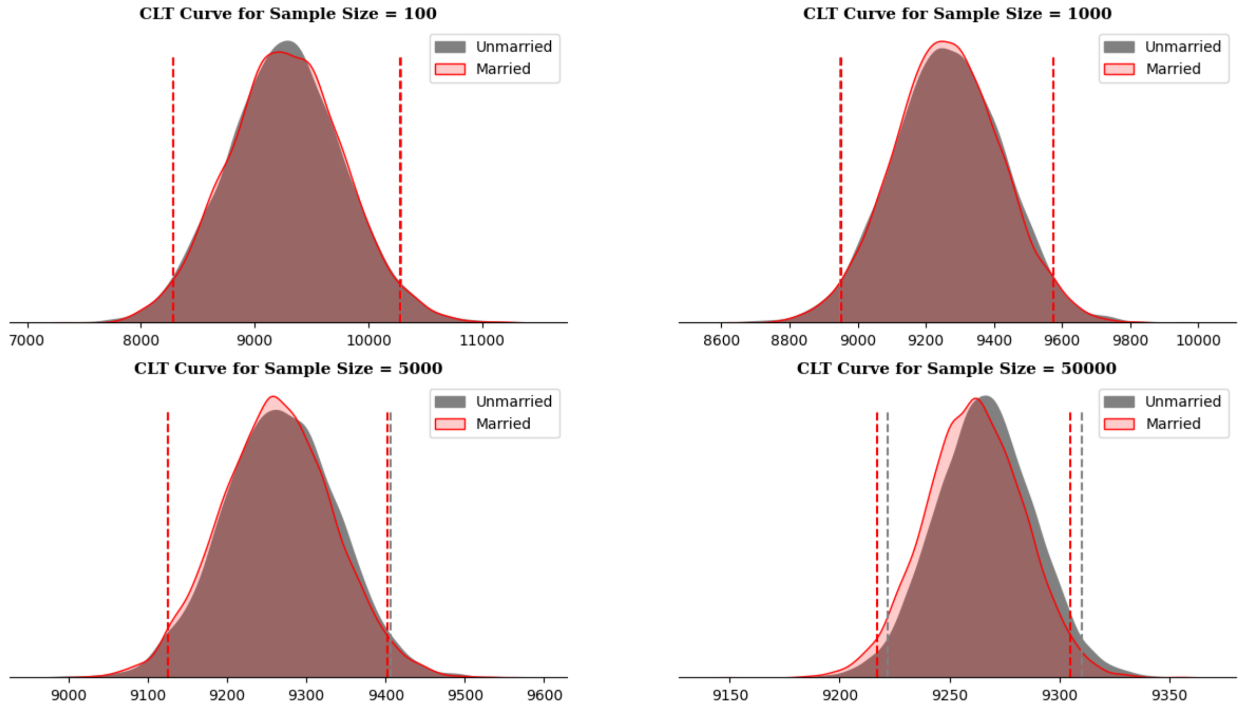
The purchase amount distribution is not normal.
We are using sample size [500, 1000, 5000, 50000] for building CLT curve.

Confidence Interval

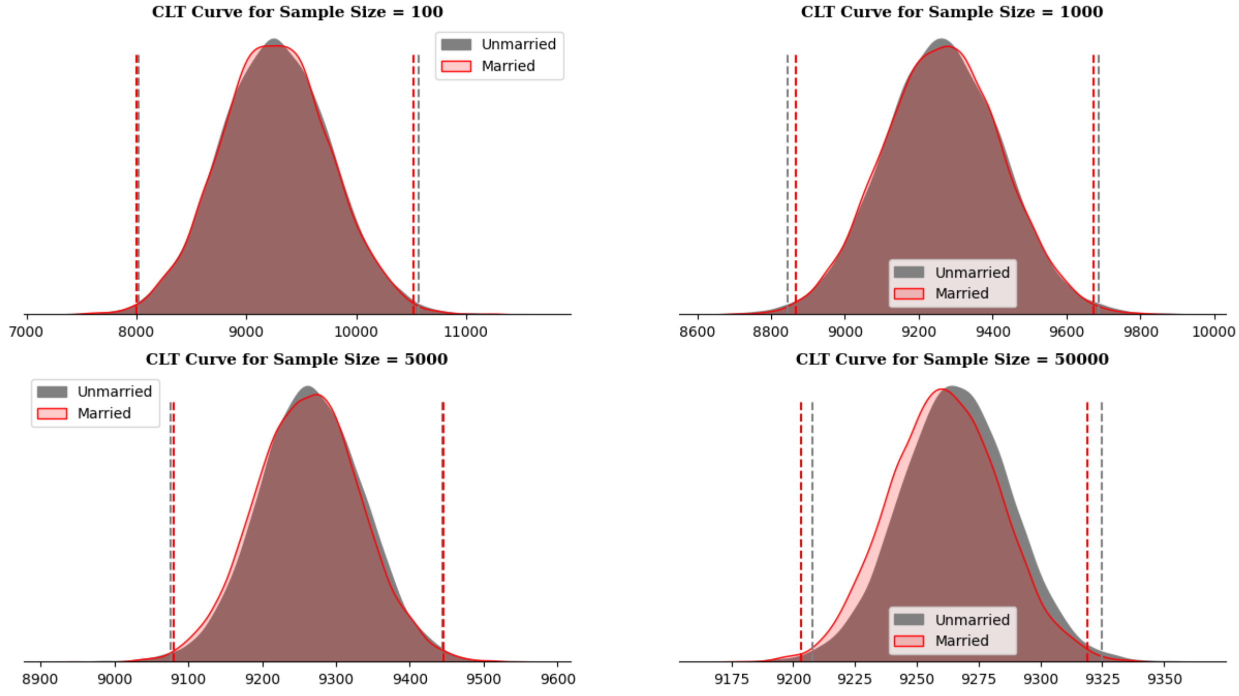
After building CLT curve of different samples, we will create a confidence interval predicting population mean at 99%, 95% and 90% Confidence level.



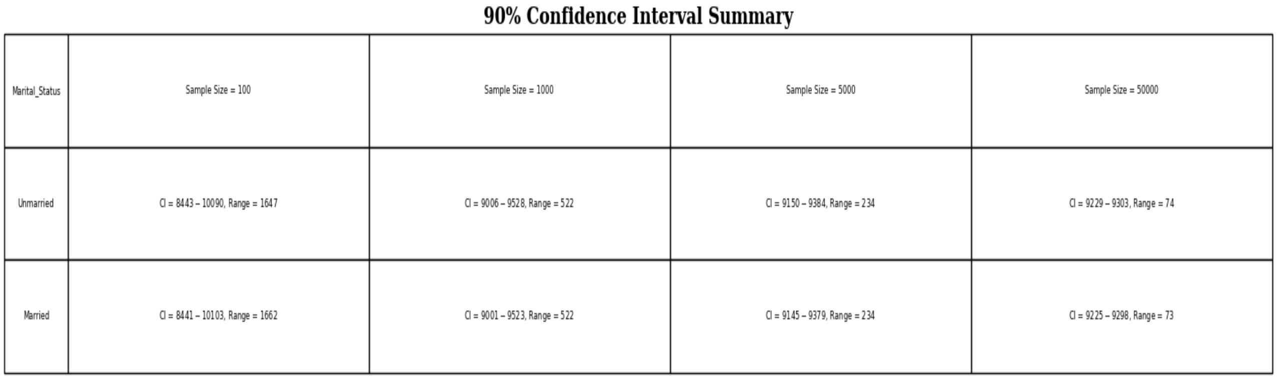
95% Confidence Interval



99% Confidence Interval

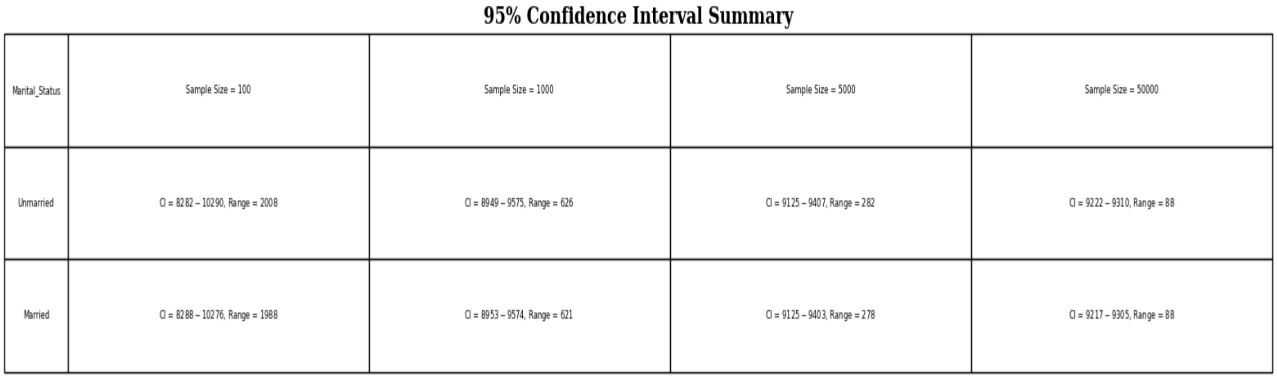


Question: Are confidence intervals of average married and unmarried customer spending overlapping?



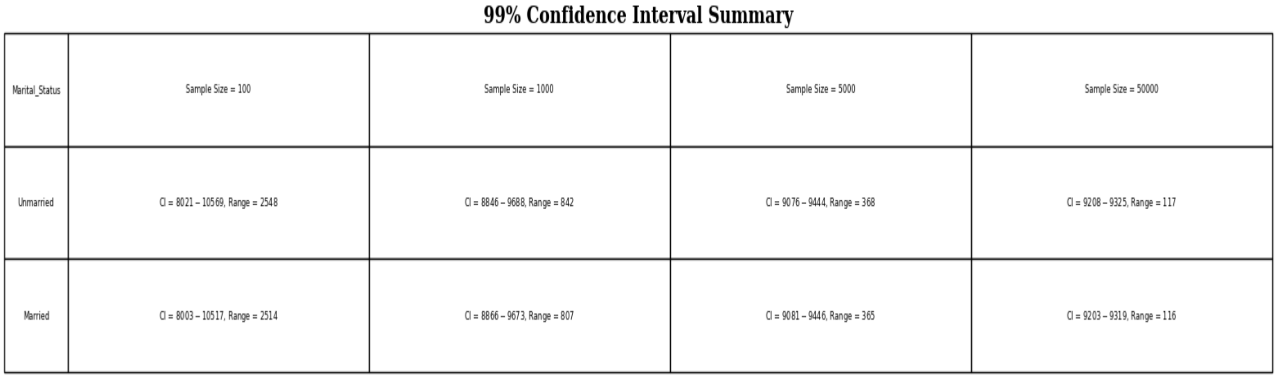
90% Confidence Interval Summary

Marital_Status	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
Unmarried	CI = 8443 – 10090, Range = 1647	CI = 9006 – 9528, Range = 522	CI = 9150 – 9384, Range = 234	CI = 9229 – 9303, Range = 74
Married	CI = 8441 – 10103, Range = 1662	CI = 9001 – 9523, Range = 522	CI = 9145 – 9379, Range = 234	CI = 9225 – 9298, Range = 73



95% Confidence Interval Summary

Marital_Status	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
Unmarried	CI = 8282 – 10290, Range = 2008	CI = 8949 – 9575, Range = 626	CI = 9125 – 9407, Range = 282	CI = 9222 – 9310, Range = 88
Married	CI = 8288 – 10276, Range = 1988	CI = 8953 – 9574, Range = 621	CI = 9125 – 9403, Range = 278	CI = 9217 – 9305, Range = 88



99% Confidence Interval Summary

Marital_Status	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
Unmarried	CI = 8021 – 10569, Range = 2548	CI = 8846 – 9688, Range = 842	CI = 9076 – 9444, Range = 368	CI = 9208 – 9325, Range = 117
Married	CI = 8003 – 10517, Range = 2514	CI = 8866 – 9673, Range = 807	CI = 9081 – 9446, Range = 365	CI = 9203 – 9319, Range = 116

From the above analysis, we can see that the confidence interval overlap for all sample sizes. This means that there is no statistically significant difference between the average spending per transaction for married and unmarried customers within the given samples.

The overlapping confidence intervals of average spending for married and unmarried customers indicate that both married and unmarried customers spend a similar amount per transaction.

For 90% confidence interval

For unmarried (sample size 50000) range for mean purchase with confidence interval 90% is [9229, 9303].
For married (sample size 50000) range for mean purchase with confidence interval 90% is [9225, 9298]

For 95% confidence interval

For unmarried (sample size 50000) range for mean purchase with confidence interval 95% is [9222, 9310]
For married (sample size 50000) range for mean purchase with confidence interval 95% is [9217, 9305]

For 99% confidence interval

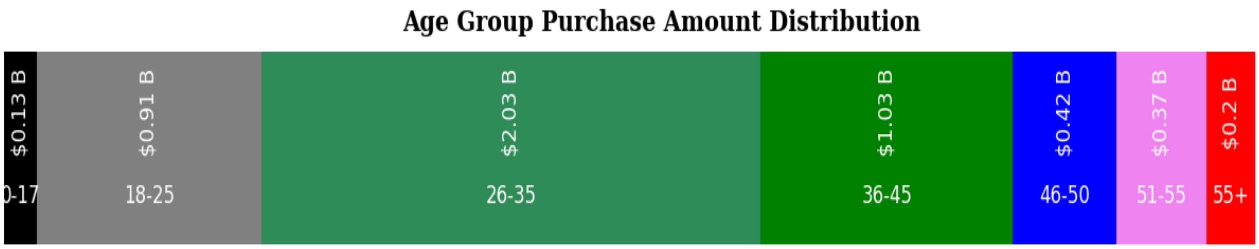
For unmarried (sample size 50000) range for mean purchase with confidence interval 99% is [9208, 9325]
For married (sample size 50000) range for mean purchase with confidence interval 99% is [9203, 9319]

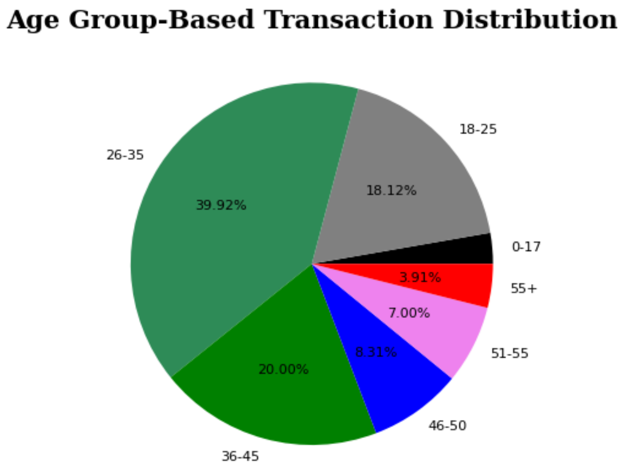
Question: How can Walmart leverage this conclusion to make changes or improvements?

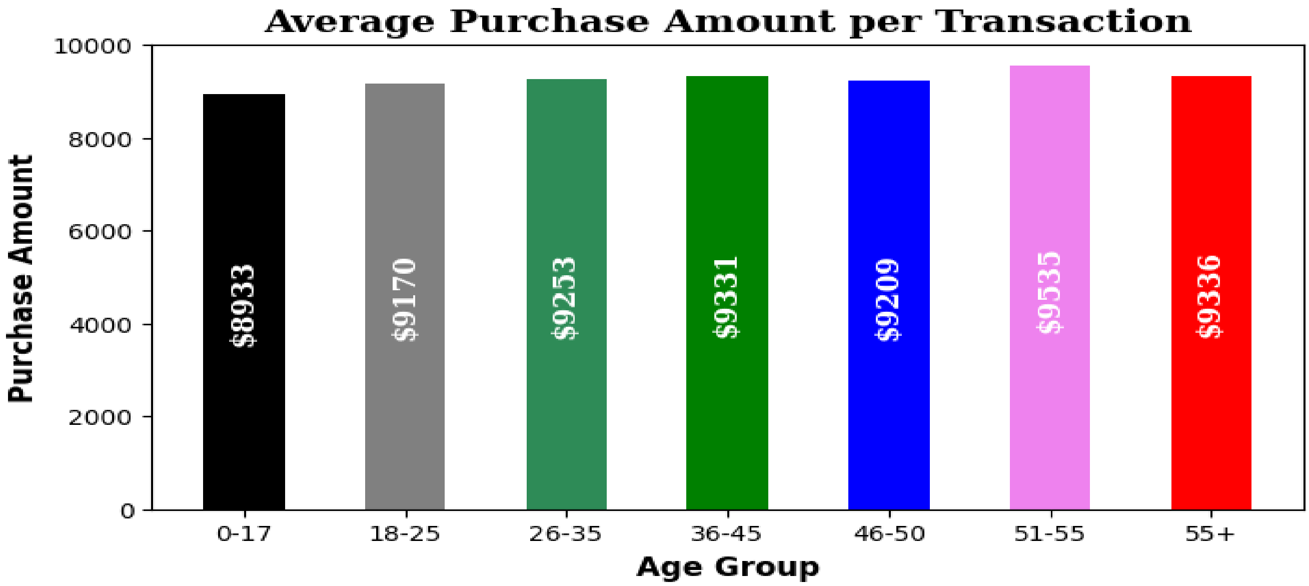
Answer: We can analyze from the above data that both married and unmarried customers are spending same amount per transaction.
So, no need to allocate marketing resources for specifically targeting one group over the other. Instead, they can focus on broader marketing strategies that appeal to both groups.

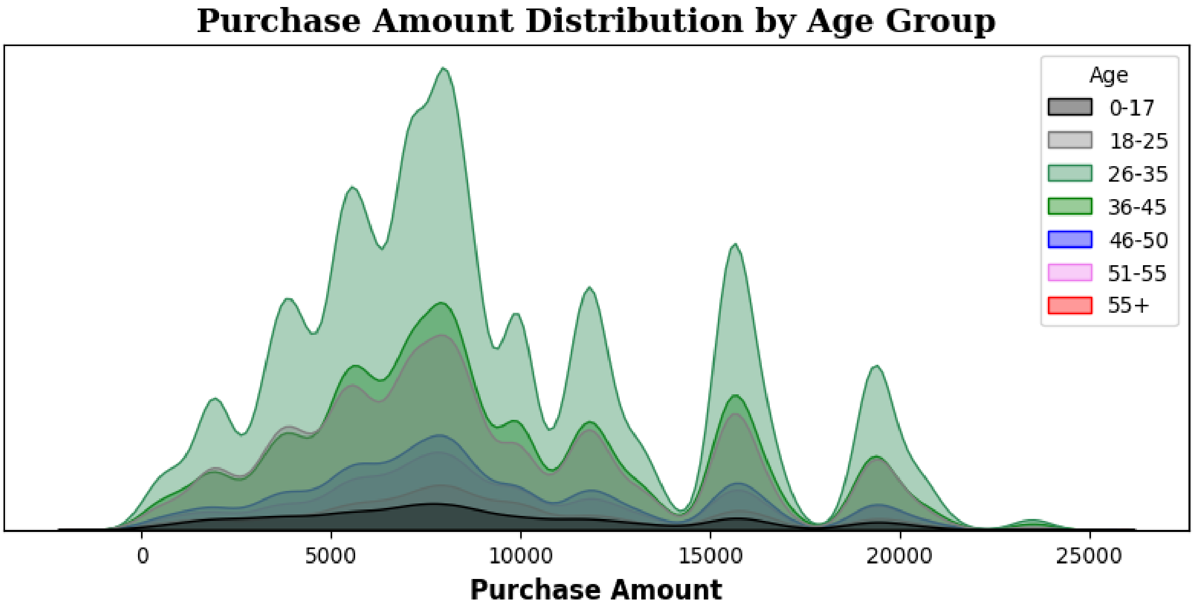
Customer Age VS Purchase Amount:

Data Visualization:




Age Group-Based Transaction Distribution
26-35
39.92%
18-25
18.12%
0-17
55+
3.91%
7.00%
51-55
8.31%
20.00%
46-50
36-45


Average Purchase Amount per Transaction
Purchase Amount
10000
8000
6000
4000
2000
0
$8933
$9170
$9253
$9331
$9209
$9535
$9336
0-17
18-25
26-35
36-45
46-50
51-55
55+
Age Group


Purchase Amount Distribution by Age Group
Age
0-17
18-25
26-35
36-45
46-50
51-55
55+
0
5000
10000
15000
20000
25000
Purchase Amount

Total Purchase and Transactions Comparison

Age group between 26 - 45 accounts to almost 60% of the total sales suggesting that Walmart's Black Friday sales are most popular among these age groups. The age group 0-17 has the lowest sales percentage (2.6%), which is expected as they may not have as much purchasing power. Understanding their preferences and providing special offers could be beneficial, especially considering the potential for building customer loyalty as they age.

Average Transaction Value

The average purchase amount per transaction was almost similar for among all the age groups. The 51-55 age groups have a relatively low sales percentage but they have the highest per purchase spending at 9535.
So, Walmart could consider strategies to attract and retain this high-spending demographic.

Distribution of Purchase Amount

The purchase amount distribution for all age groups is not normally distributed.

Confidence Interval and CLT:

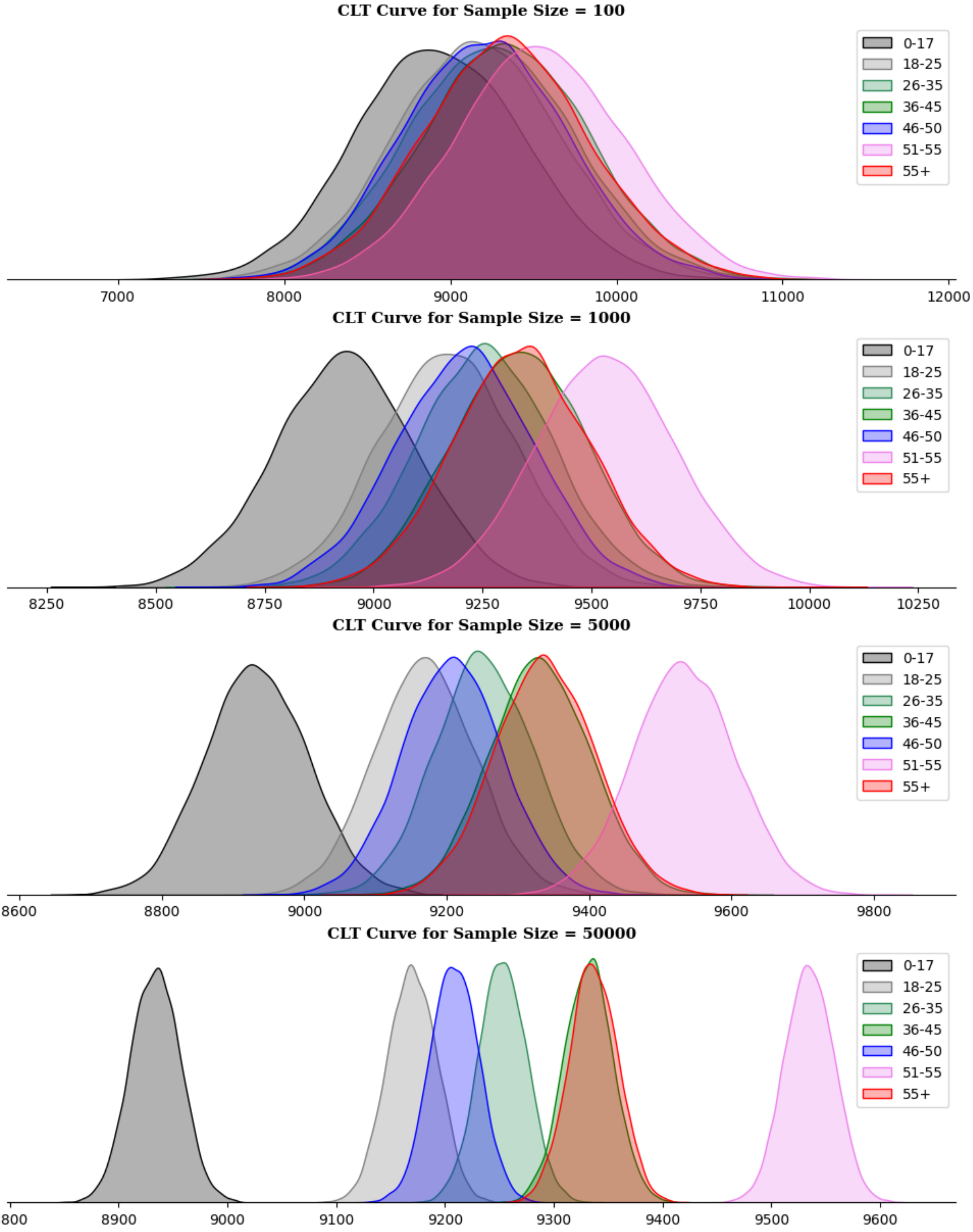
CLT(Central Limit Theorem)

The purchase amount distribution is not normal.
We are using sample size [500, 1000, 5000, 50000] for building CLT curve.

Confidence Interval

After building CLT curve of different samples, we will create a confidence interval predicting population mean at 99%, 95% and 90% Confidence level.

90% Confidence Interval



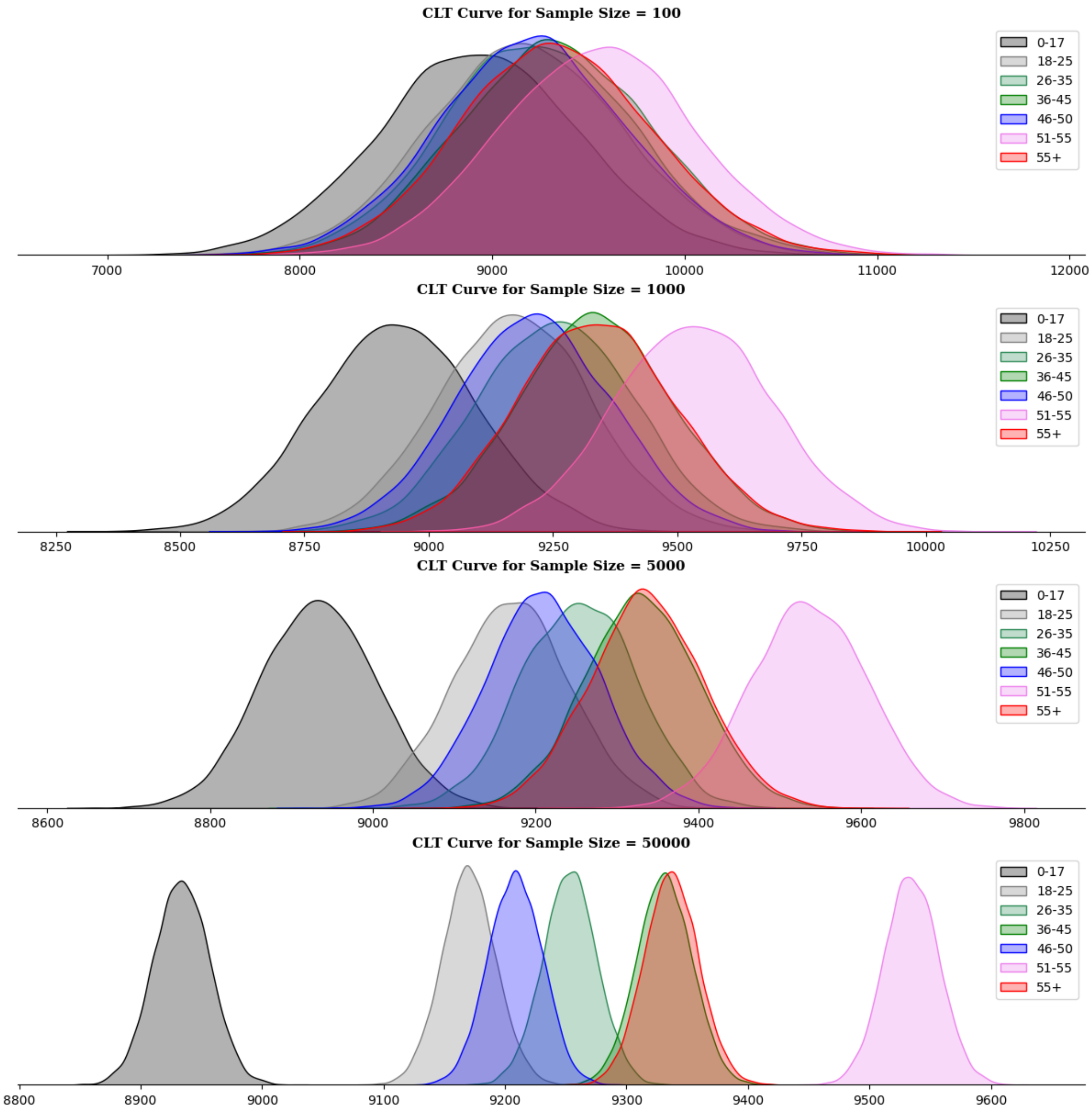
CLT Curve for Sample Size = 100

CLT Curve for Sample Size = 1000

CLT Curve for Sample Size = 5000

CLT Curve for Sample Size = 50000

95% Confidence Interval

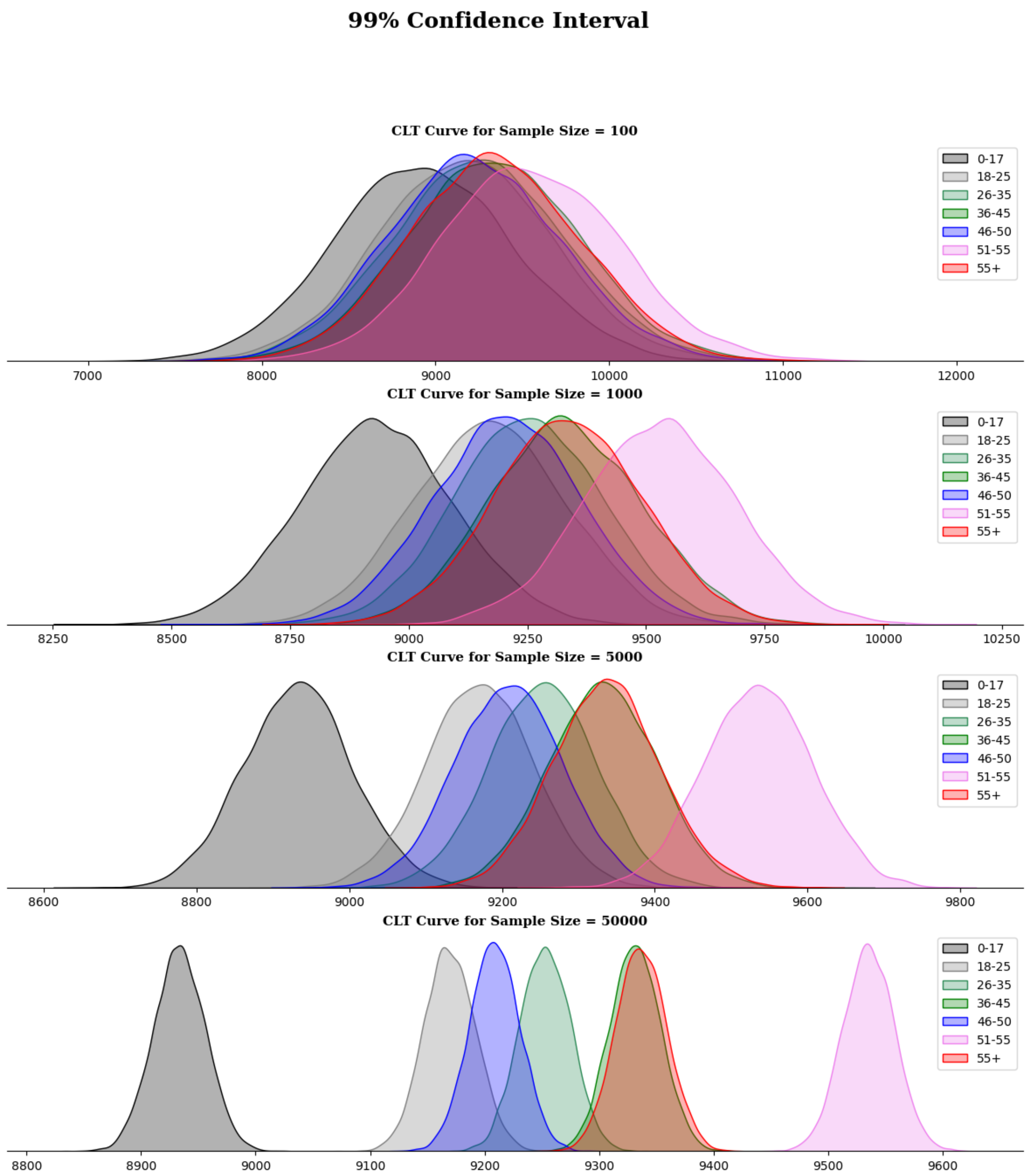


CLT Curve for Sample Size = 100

CLT Curve for Sample Size = 1000

CLT Curve for Sample Size = 5000

CLT Curve for Sample Size = 50000



99% Confidence Interval

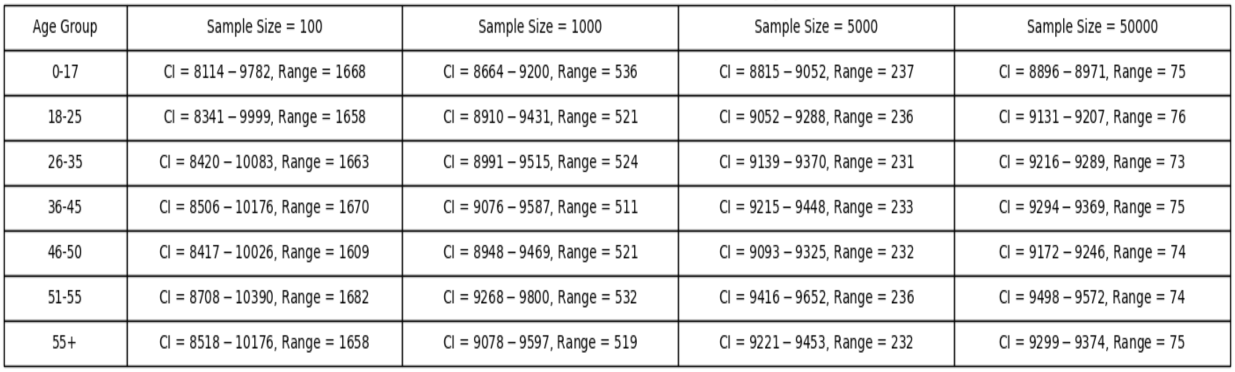
CLT Curve for Sample Size = 100

CLT Curve for Sample Size = 1000

CLT Curve for Sample Size = 5000

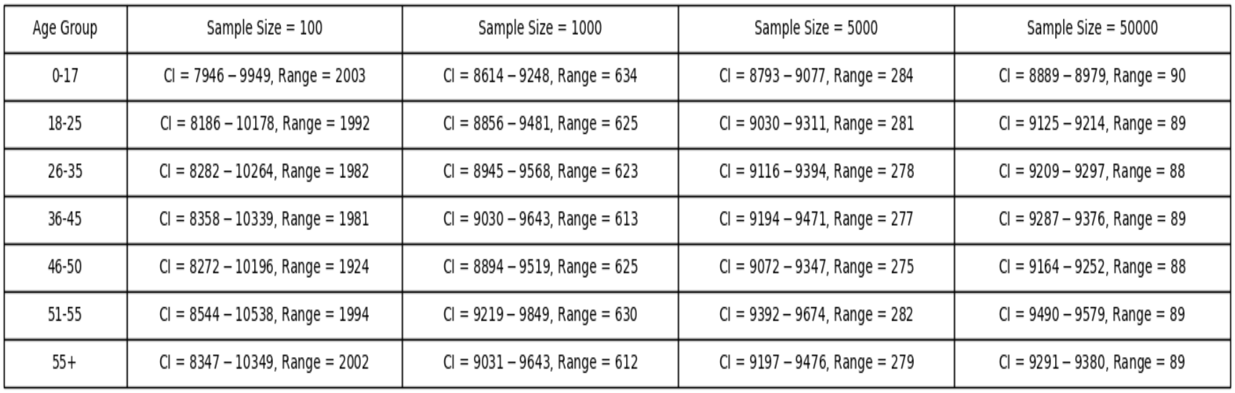
CLT Curve for Sample Size = 50000

Question: Are confidence intervals of customer's age-group spending overlapping?



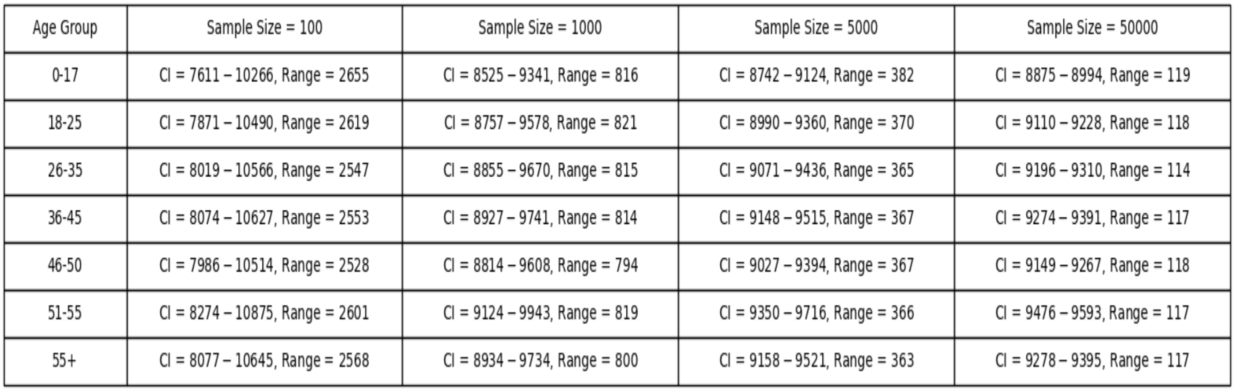
90% Confidence Interval Summary

Age Group	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
0-17	CI = 8114 – 9782, Range = 1668	CI = 8664 – 9200, Range = 536	CI = 8815 – 9052, Range = 237	CI = 8896 – 8971, Range = 75
18-25	CI = 8341 – 9999, Range = 1658	CI = 8910 – 9431, Range = 521	CI = 9052 – 9288, Range = 236	CI = 9131 – 9207, Range = 76
26-35	CI = 8420 – 10083, Range = 1663	CI = 8991 – 9515, Range = 524	CI = 9139 – 9370, Range = 231	CI = 9216 – 9289, Range = 73
36-45	CI = 8506 – 10176, Range = 1670	CI = 9076 – 9587, Range = 511	CI = 9215 – 9448, Range = 233	CI = 9294 – 9369, Range = 75
46-50	CI = 8417 – 10026, Range = 1609	CI = 8948 – 9469, Range = 521	CI = 9093 – 9325, Range = 232	CI = 9172 – 9246, Range = 74
51-55	CI = 8708 – 10390, Range = 1682	CI = 9268 – 9800, Range = 532	CI = 9416 – 9652, Range = 236	CI = 9498 – 9572, Range = 74
55+	CI = 8518 – 10176, Range = 1658	CI = 9078 – 9597, Range = 519	CI = 9221 – 9453, Range = 232	CI = 9299 – 9374, Range = 75



95% Confidence Interval Summary

Age Group	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
0-17	CI = 7946 – 9949, Range = 2003	CI = 8614 – 9248, Range = 634	CI = 8793 – 9077, Range = 284	CI = 8889 – 8979, Range = 90
18-25	CI = 8186 – 10178, Range = 1992	CI = 8856 – 9481, Range = 625	CI = 9030 – 9311, Range = 281	CI = 9125 – 9214, Range = 89
26-35	CI = 8282 – 10264, Range = 1982	CI = 8945 – 9568, Range = 623	CI = 9116 – 9394, Range = 278	CI = 9209 – 9297, Range = 88
36-45	CI = 8358 – 10339, Range = 1981	CI = 9030 – 9643, Range = 613	CI = 9194 – 9471, Range = 277	CI = 9287 – 9376, Range = 89
46-50	CI = 8272 – 10196, Range = 1924	CI = 8894 – 9519, Range = 625	CI = 9072 – 9347, Range = 275	CI = 9164 – 9252, Range = 88
51-55	CI = 8544 – 10538, Range = 1994	CI = 9219 – 9849, Range = 630	CI = 9392 – 9674, Range = 282	CI = 9490 – 9579, Range = 89
55+	CI = 8347 – 10349, Range = 2002	CI = 9031 – 9643, Range = 612	CI = 9197 – 9476, Range = 279	CI = 9291 – 9380, Range = 89



99% Confidence Interval Summary

Age Group	Sample Size = 100	Sample Size = 1000	Sample Size = 5000	Sample Size = 50000
0-17	CI = 7611 – 10266, Range = 2655	CI = 8525 – 9341, Range = 816	CI = 8742 – 9124, Range = 382	CI = 8875 – 8994, Range = 119
18-25	CI = 7871 – 10490, Range = 2619	CI = 8757 – 9578, Range = 821	CI = 8990 – 9360, Range = 370	CI = 9110 – 9228, Range = 118
26-35	CI = 8019 – 10566, Range = 2547	CI = 8855 – 9670, Range = 815	CI = 9071 – 9436, Range = 365	CI = 9196 – 9310, Range = 114
36-45	CI = 8074 – 10627, Range = 2553	CI = 8927 – 9741, Range = 814	CI = 9148 – 9515, Range = 367	CI = 9274 – 9391, Range = 117
46-50	CI = 7986 – 10514, Range = 2528	CI = 8814 – 9608, Range = 794	CI = 9027 – 9394, Range = 367	CI = 9149 – 9267, Range = 118
51-55	CI = 8274 – 10875, Range = 2601	CI = 9124 – 9943, Range = 819	CI = 9350 – 9716, Range = 366	CI = 9476 – 9593, Range = 117
55+	CI = 8077 – 10645, Range = 2568	CI = 8934 – 9734, Range = 800	CI = 9158 – 9521, Range = 363	CI = 9278 – 9395, Range = 117

From the above analysis, we can see that the confidence interval overlap for some of the age groups. So, for better understanding we can club the average spending into some groups:

- 18 - 25, 26 - 35, 46 - 50 : Customers in these age groups have overlapping confidence intervals indicating similar buying characteristics.

- 36 - 45, 55+ : Customers in these age groups have overlapping confidence intervals indicating and similar spending patterns

For 90% confidence interval

For 0-17 (sample size 50000) is [8896, 8971].
For 18-25 (sample size 50000) is [9131, 9207].
For 26-35 (sample size 50000) is [9216, 9289].
For 36-45 (sample size 50000) is [9294, 9369].
For 46-50 (sample size 50000) is [9172, 9246].
For 51-55 (sample size 50000) is [9498, 9572].
For 55+ (sample size 50000) is [9299, 9374].

For 95% confidence interval

For 0-17 (sample size 50000) is [8889, 8979].
For 18-25 (sample size 50000) is [9125, 9214].
For 26-35 (sample size 50000) is [9209, 9297].
For 36-45 (sample size 50000) is [9287, 9376].
For 46-50 (sample size 50000) is [9164, 9252].
For 51-55 (sample size 50000) is [9490, 9579].
For 55+ (sample size 50000) is [9291, 9380].

For 99% confidence interval

For 0-17 (sample size 50000) is [8875, 8994].
For 18-25 (sample size 50000) is [9110, 9228].
For 26-35 (sample size 50000) is [9196, 9310].
For 36-45 (sample size 50000) is [9274, 9391].
For 46-50 (sample size 50000) is [9149, 9267].
For 51-55 (sample size 50000) is [9476, 9593].
For 55+ (sample size 50000) is [9278, 9395].

Question: How can Walmart leverage this conclusion to make changes or improvements?

Answer: Walmart can take the following steps to leverage this conclusion:

- Customers in the 0 - 17 age group have the lowest spending per transaction, Walmart can try to increase their spending per transaction by offering them more attractive discounts, coupons, or rewards programs.

- Customers in the 18 - 25, 26 - 35, and 46 - 50 age groups exhibit similar buying characteristics; Walmart can optimize its product selection to cater to the preferences of these age groups. Also, Walmart can use this information to adjust their pricing strategies for different age groups.
- Customers in the 51 - 55 age group have the highest spending per transaction, Walmart can explore opportunities to enhance the shopping experience for this demographic. This might involve offering premium services, personalized recommendations, or loyalty programs that cater to the preferences and spending habits of this age group.

Colab notebook link:
https://colab.research.google.com/drive/1t3bGHjXkVk1flmixVbUOoaEiUfYiJht3?usp=sharing