

Walmart - Confidence Interval and CLT - Case Study

Business Problem

Analyzing the customer purchasing behaviour based on gender to make better decision :

*Analyzing customer behaviour based on Male and Female on Sale (Black Friday Sale)season.

*Analyzing and comparing the spending habits of thier customers .

Basic Matrix :

*Uploading the dataset.

*Observations on shape of data

*Data types of all the attributes

```
import pandas as pd
```

```
df = pd.read_csv('walmart_data.csv')
df
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	P
0	1000001	P00069042	F	0-17	10	A	2	0		3
1	1000001	P00248942	F	0-17	10	A	2	0		1
2	1000001	P00087842	F	0-17	10	A	2	0		12
3	1000001	P00085442	F	0-17	10	A	2	0		12
4	1000002	P00285442	M	55+	16	C	4+	0		8
...
550063	1006033	P00372445	M	51-55	13	B	1	1		20
550064	1006035	P00375436	F	26-35	1	C	3	0		20

Observation on shape of Data

```
df.shape
```

```
(550068, 10)
```

By observing there are 550068 rows in the data and 10 columns in the data set

Data types of all attributes :

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   User_ID                              550068 non-null  int64
 1   Product_ID                           550068 non-null  object
 2   Gender                               550068 non-null  object
 3   Age                                   550068 non-null  object
 4   Occupation                           550068 non-null  int64
 5   City_Category                        550068 non-null  object
 6   Stay_In_Current_City_Years          550068 non-null  object
 7   Marital_Status                      550068 non-null  int64
 8   Product_Category                    550068 non-null  int64
 9   Purchase                            550068 non-null  int64
dtypes: int64(5), object(5)
```

memory usage: 42.0+ MB

Checking the Data types of all the columns

by observing we get to know there are 7 Categorical Columns and 3 numeric

Non- Graphical Analysis

OBSERVATION ON PURCHASING AMOUNT SPENT BY CUSTOMERS


```
df['Purchase'].describe() #Purchasing amount statistics
```



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

dtype: float64

```
df.describe()
```



	User_ID	Occupation	Marital_Status	Product_Category	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000	550068.000000
mean	1.003029e+06	8.076707	0.409653	5.404270	9263.968713
std	1.727592e+03	6.522660	0.491770	3.936211	5023.065394
min	1.000001e+06	0.000000	0.000000	1.000000	12.000000
25%	1.001516e+06	2.000000	0.000000	1.000000	5823.000000
50%	1.003077e+06	7.000000	0.000000	5.000000	8047.000000
75%	1.004478e+06	14.000000	1.000000	8.000000	12054.000000
max	1.006040e+06	20.000000	1.000000	20.000000	23961.000000

Identifying Unique values for each columns

```
unique_columns = df.nunique()  
unique_columns
```



	0
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

dtype: int64

EDA - Exploratory Data Analysis

```
df['User_ID'].nunique()
```

```
↗ 5891
```

```
per_person_purchase_sum = df.groupby(['User_ID'])['Purchase'].sum()
apoo = per_person_purchase_sum.reset_index()
minimum_purchase = apoo.min()
maximum_purchase = apoo.max()
print('The Maximum Sales is by',maximum_purchase)
print('The Minimum Sales is by', minimum_purchase)
```

```
↗ The Maximum Sales is by User_ID      1006040
Purchase      10536909
dtype: int64
The Minimum Sales is by User_ID      1000001
Purchase      46681
dtype: int64
```

```
#Tracking the amount spent per transaction of all the 50 million female customers,
# and all the 50 million male customers, calculate the average, and conclude the results.
```

```
average_amt = df.groupby(['Gender'])[['Purchase']].mean()
average_amt
```

```
↗
```

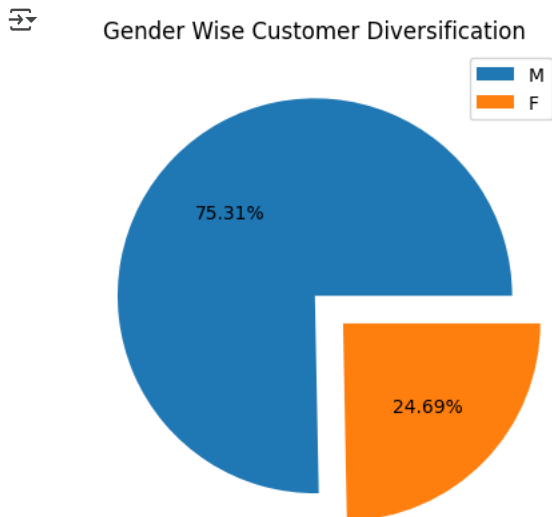
	Purchase
Gender	
F	8734.565765
M	9437.526040

Next steps:

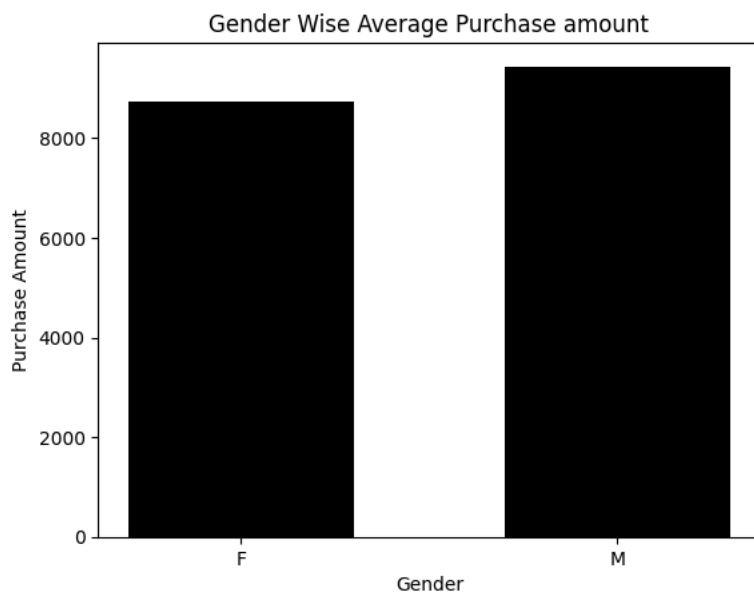
[Generate code with average_amt](#)[View recommended plots](#)[New interactive sheet](#)

Visual Analysis

```
import matplotlib.pyplot as plt
gender_data = df['Gender'].value_counts()
plt.pie(gender_data,autopct = '%.2f%%', explode= [0,0.2])
plt.title('Gender Wise Customer Diversification')
plt.legend(gender_data.index)
plt.show()
```

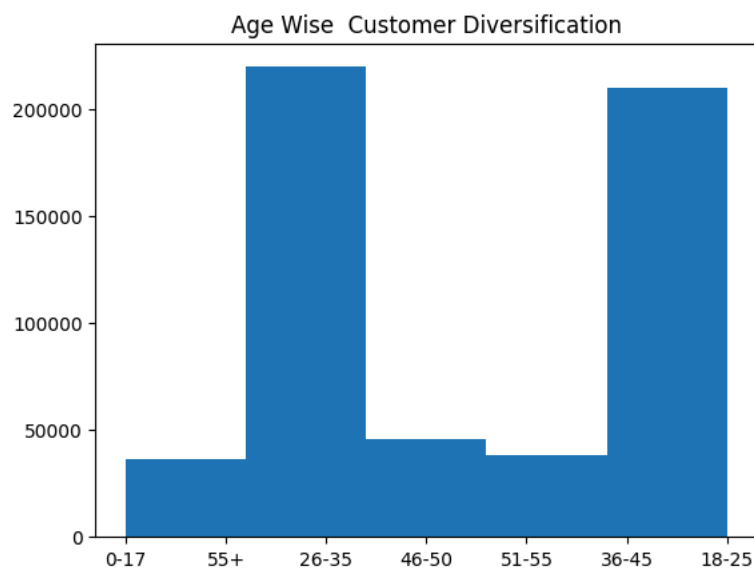


```
plt.bar(average_amt.index , average_amt['Purchase'] , color = 'black', width = 0.6)
plt.title('Gender Wise Average Purchase amount')
plt.xlabel('Gender')
plt.ylabel('Purchase Amount')
plt.show()
```



```
age_data = df['Age']
```

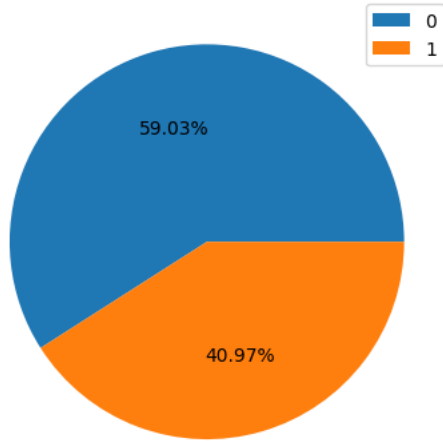
```
plt.hist(age_data , bins =5)  
plt.title('Age Wise Customer Diversification')  
plt.show()
```



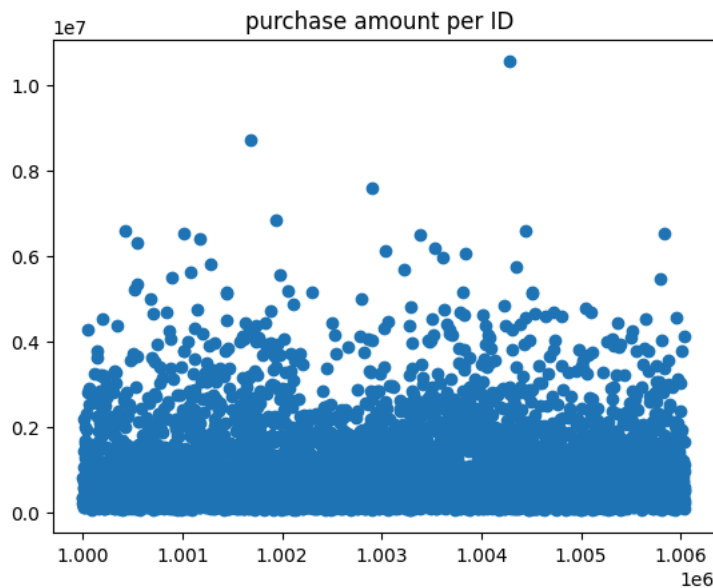
```
marital_status = df['Marital_Status'].value_counts()  
marital_status.reset_index()  
plt.pie(marital_status, autopct = '%.2f%%')  
plt.legend(marital_status.index)  
plt.title('Marital Status diversification')  
plt.show()
```



Marital Status diversification



```
per_person_purchase_sum = df.groupby(['User_ID'])['Purchase'].sum()
apoo = per_person_purchase_sum.reset_index()
apoorv_unique = apoo['User_ID']
apoorv_sum = apoo['Purchase']
apoorv_unique
plt.scatter(apoorv_unique , apoorv_sum )
plt.title('purchase amount per ID ')
plt.show()
```



✓ Missing Value & Outlier Detection

CHECKING OUT NULL VALUES

```
df.isnull().value_counts()
```



User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
False	False	False	False	False	False	False	False	False	False

```
dtype: int64
```

DROP ALL NULL VALUES TO MAKE DATA CLEAN

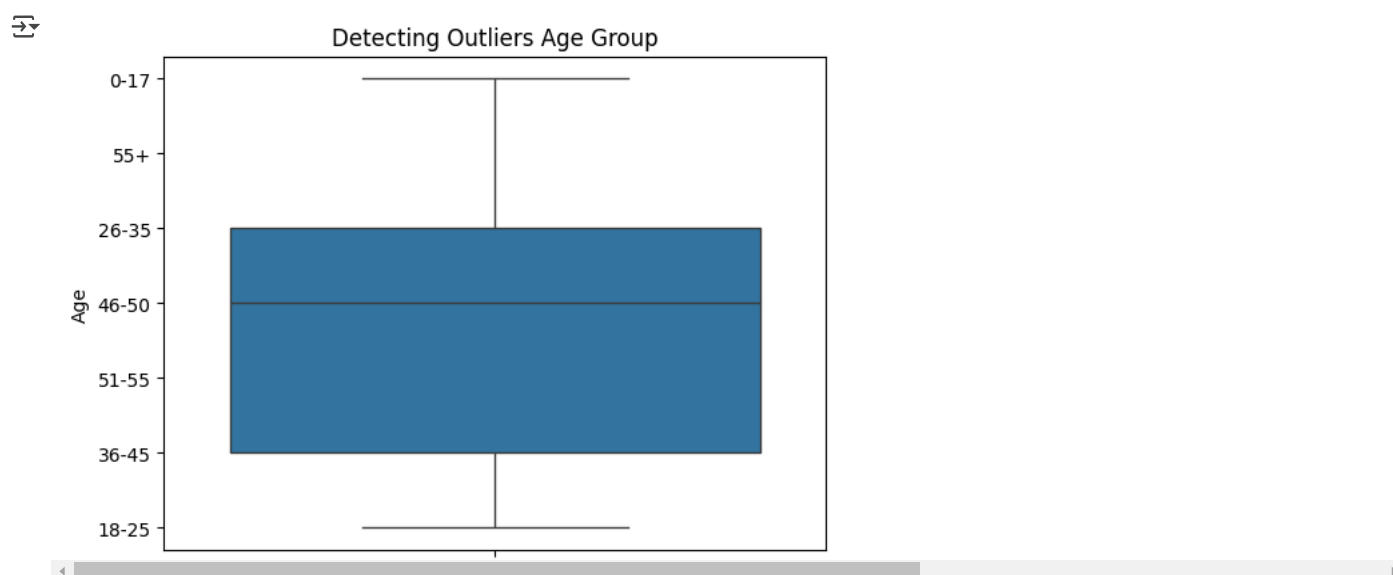
```
droping_null = df.dropna()
droping_null
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	P
0	1000001	P00069042	F	0-17	10	A	2	0	3	
1	1000001	P00248942	F	0-17	10	A	2	0	1	
2	1000001	P00087842	F	0-17	10	A	2	0	12	
3	1000001	P00085442	F	0-17	10	A	2	0	12	
4	1000002	P00285442	M	55+	16	C	4+	0	8	
...
550063	1006033	P00372445	M	51-55	13	B	1	1	20	
550064	1006035	P00375436	F	26-35	1	C	3	0	20	

✓ *OUTLIER DETECTIONS *

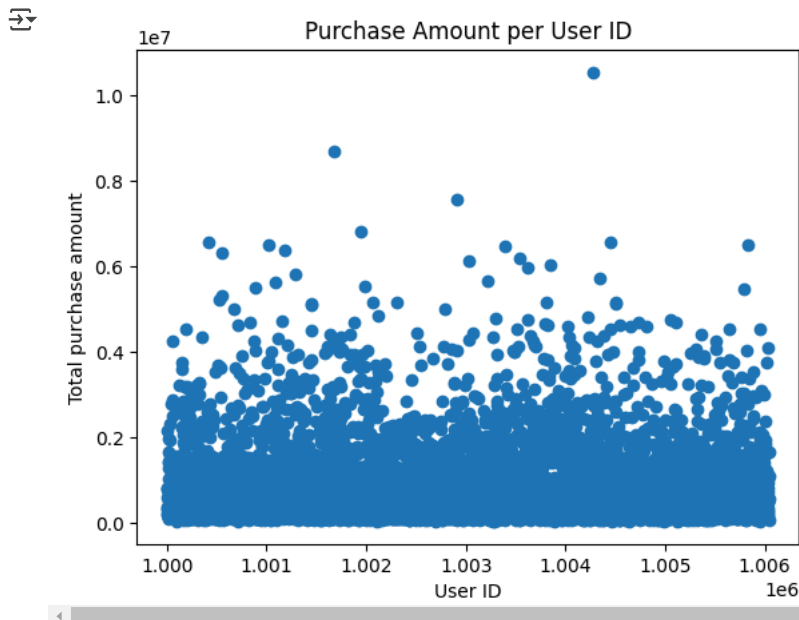
```
import matplotlib.pyplot as plt
import seaborn as sns
age_data = df['Age']
purchase_data = df['Purchase']
sns.boxplot(age_data)
```

```
plt.title('Detecting Outliers Age Group')
plt.show()
```



By This Box Plot we can conclude that - the Median Age range for the customers is 46-50 years

```
total_ammount_per_id = df.groupby(['User_ID'])['Purchase'].sum()
total_ammount_per_id1 = total_ammount_per_id.reset_index()
total_ammount_per_id12 = total_ammount_per_id1['User_ID']
total_ammount_per_id_sum = total_ammount_per_id1['Purchase']
plt.scatter(total_ammount_per_id12 , total_ammount_per_id_sum)
plt.title('Purchase Amount per User ID')
plt.xlabel('User ID')
plt.ylabel('Total purchase amount')
plt.show()
```



Business Insights -

There are follow key points

- There are more Males customer then Females. so they must focus on the population of females
- There are more Single customer then Married once, The perenatge contribution of Single are 59.04% where married were contributing 40.96% , which indicates that they need to attract more married customers
- The Median Age of the customers is between 46-50 according to the boxplot we create , showing up that the teenage group and old age group is less focused.
- There are 5891 Unique User IDs only out of which User ID - 1006040 has the max amount sales off rupee 10536909 and minimum sales is of 1000001 of rupee is 46681
- The mean of purchase amount by gender - females -8734.565765 where male has - 9437.58, which shows us that Females have more spending habit then males , perhaps males have three times more population then females.

Some Questions and Answeres -

What is the mean amount spend by per male and female ? Who has more Spending habit bbetween Male and Female ?

answere - We Calculate the mean amount spend by Male and Female below , which indicates that females has spent 6.43 per women , But Males only spend 2.24 rupee per men , It Makes an Insight that the spending habit is more in the females then males

#total average amount divide gender wise

```
female_unique = df[df['Gender']== 'F'].value_counts()
total_females = len(female_unique)
qq= female_unique.reset_index()
female_purchase_mean =qq['Purchase'].mean()
per_female_amount = (female_purchase_mean / total_females) * 100
per_female_amount
```

6.431507311853762

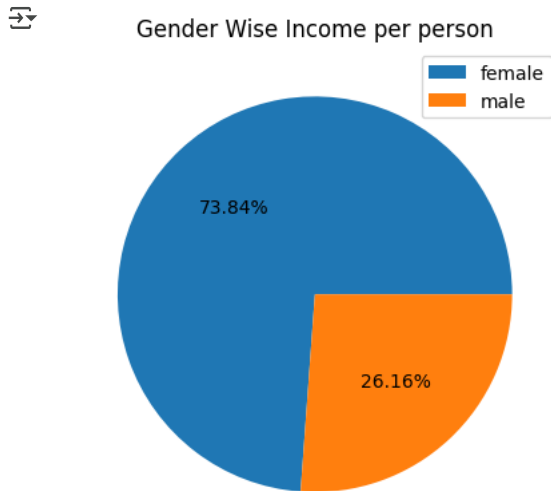
```
male_unique = df[df['Gender']== 'M'].value_counts()
total_males = len(male_unique)
pp= male_unique.reset_index()
male_purchase_mean =pp['Purchase'].mean()
per_male_amount = (male_purchase_mean / total_males) * 100
per_male_amount
```

```
2.2781704297244634
```

```
gender_wise = (per_female_amount , per_male_amount)
gender_wise
```

```
(6.431507311853762, 2.2781704297244634)
```

```
plt.pie(gender_wise, autopct= '%.2f%%')
plt.title('Gender Wise Income per person')
plt.legend(['female', 'male'])
plt.show()
```



Question - What is the ratio of spending habit between Males and Females?

The Ratio of spending habit of Females vs Males (**2.8 : 1**)

This indicates that Females has highest spending habit then Males , where the population of Males is 3x then females

Question- Use the sample average to find out an interval within which the population average will lie ?

The Graph indicates that the purchasing amount of customers will lie mostly in the range of 6000 - 12000 , where the 50% population is lying.

```
# Use the Central limit theorem to compute the interval.
# Change the sample size to observe the distribution of the mean of the expenses by female and male customers.
# The interval that you calculated is called Confidence Interval.

# The width of the interval is mostly decided by the business: Typically 90%, 95%, or 99%.
# Play around with the width parameter and report the observations.
```

```
mens_mean = df[df['Gender'] == 'M']
mens_mean_purchase = mens_mean['Purchase'].mean()
standard_devmen = mens_mean['Purchase'].std()
standard_devmen
```

```
5092.186209777949
```

```
female_mean = df[df['Gender'] == 'F']
females_mean_purchase = female_mean['Purchase'].mean()
standard_devfemale = female_mean['Purchase'].std()
standard_devfemale
```

```
4767.233289291444
```

```
length_female = len(female_mean)
length_men = len(mens_mean)
length_men , length_female
```

```
(414259, 135809)
```


Questions - Confidence intervals and distribution of the mean of the expenses by female and male customers

ans- We calculate the Confidence interval at different level of confidence (i.e, 95% ,90% ,99%) please check below


```

# Men CLT
# 95% Confidence Level:
# z=1.96
Confidence_interval_men = (mens_mean_purchase) * (standard_devmen) /length_men
Confidence_interval_mens_upper = (Confidence_interval_men) + 1.96
Confidence_interval_mens_lower = (Confidence_interval_men) - 1.96
CLT_95 = Confidence_interval_mens_lower ,Confidence_interval_mens_upper
print('This means we are 95% confident that the true population mean lies between' , CLT_95)
# 90% Confidence Level:
# z ~ 1.645
Confidence_interval_men = (mens_mean_purchase) * (standard_devmen) /length_men
Confidence_interval_mens_upper = (Confidence_interval_men) + 1.645
Confidence_interval_mens_lower = (Confidence_interval_men) - 1.645
CLT_90 = Confidence_interval_mens_lower ,Confidence_interval_mens_upper
print('This means we are 90% confident that the true population mean lies between' , CLT_90)
# 99% Confidence Level:
# z ~ 2.576
Confidence_interval_men = (mens_mean_purchase) * (standard_devmen) /length_men
Confidence_interval_mens_upper = (Confidence_interval_men) + 2.576
Confidence_interval_mens_lower = (Confidence_interval_men) - 2.576
CLT_99 = Confidence_interval_mens_lower ,Confidence_interval_mens_upper
print('This means we are 99% confident that the true population mean lies between' ,CLT_99)


```

 This means we are 95% confident that the true population mean lies between (114.0486804576682, 117.96868045766819)
 This means we are 90% confident that the true population mean lies between (114.3636804576682, 117.65368045766819)
 This means we are 99% confident that the true population mean lies between (113.4326804576682, 118.58468045766818)

```

# Female CLT
# 95% Confidence Level:
# z=1.96
Confidence_interval_female = (females_mean_purchase) * (standard_devfemale) /length_female
Confidence_interval_females_upper = (Confidence_interval_female) + 1.96
Confidence_interval_females_lower = (Confidence_interval_female) - 1.96
CLT_95 = Confidence_interval_females_lower ,Confidence_interval_females_upper
print('This means we are 95% confident that the true population mean lies between' , CLT_95)
# 90% Confidence Level:
# z ~ 1.645
Confidence_interval_female = (females_mean_purchase) * (standard_devfemale) /length_female
Confidence_interval_females_upper = (Confidence_interval_female) + 1.645
Confidence_interval_females_lower = (Confidence_interval_female) - 1.645
CLT_90 = Confidence_interval_females_lower ,Confidence_interval_females_upper
print('This means we are 90% confident that the true population mean lies between' , CLT_90)
# 99% Confidence Level:
# z ~ 2.576
Confidence_interval_female = (females_mean_purchase) * (standard_devfemale) /length_female
Confidence_interval_females_upper = (Confidence_interval_female) + 2.576
Confidence_interval_females_lower = (Confidence_interval_female) - 2.576
CLT_99 = Confidence_interval_females_lower ,Confidence_interval_females_upper
print('This means we are 99% confident that the true population mean lies between' , CLT_99)

```

 This means we are 95% confident that the true population mean lies between (304.64495757390586, 308.5649575739058)
 This means we are 90% confident that the true population mean lies between (304.95995757390585, 308.2499575739058)
 This means we are 99% confident that the true population mean lies between (304.0289575739058, 309.18095757390586)

✓ **Question - Confidence intervals and distribution of the mean of the expenses by married and singles **

Answers -

*The CLT for married at 95% confidence level lies between 204.23 to 208.15

*The CLT for Single Marital Status at 95% confidence level lies between 145.41 to 141.49

```
married1 = df[df['Marital_Status'] == 1]
married_len1 = len(married1)
married_groupby = married1.groupby(['Marital_Status'])['Purchase'].mean()
married12 = married_groupby.reset_index()
married12_mean = married12['Purchase']
married12_mean
```

```
↗
Purchase
0    9261.174574

dtype: float64
```

```
married0 = df[df['Marital_Status'] == 0]
married_len0 = len(married0)
married_groupby = married0.groupby(['Marital_Status'])['Purchase'].mean()
married123 = married_groupby.reset_index()
married123_mean = married123['Purchase']
married123_mean
```

```
↗
Purchase
0    9265.907619

dtype: float64
```

```
married_std0 = married0['Purchase'].std()
married_std0
```

```
↗ 5027.347858674457
```

```
married_std1 = married1['Purchase'].std()
married_std1
```

```
↗ 5016.89737779313
```

```
married_clt0 = (married123_mean) * (married_std0) / (married_len0)
clt_married0upper = married_clt0 + 1.96
clt_married0low = married_clt0 - 1.96
clt_married0upper , clt_married0low
```

```
↗ (0    145.410859
   Name: Purchase, dtype: float64,
   0    141.490859
   Name: Purchase, dtype: float64)
```

```
married_clt1 = (married12_mean) * (married_std1) / (married_len1)
clt_married1upper = married_clt1 + 1.96
clt_married1low = married_clt1 - 1.96
clt_married1upper , clt_married1low
```

```
↗ (0    208.150561
   Name: Purchase, dtype: float64,
   0    204.230561
   Name: Purchase, dtype: float64)
```

Answers -

*The CLT for married at 95% confidence level lies between 204.23 to 208.15

*The CLT for Single Marital Status at 95% confidence level lies between 145.41 to 141.49

✓ Question - Confidence intervals and distribution of the mean of the expenses by Age group

Double-click (or enter) to edit

```
age_group = df.groupby(['Age'])['Purchase'].sum()
age_group1 = age_group.reset_index()
age_mean = age_group1['Purchase'].mean()
age_mean
```

```
727973248.8571428
```

```
df['Age'].unique()
```

```
array(['0-17', '55+', '26-35', '46-50', '51-55', '36-45', '18-25'],
      dtype=object)
```

```
age_std = age_group1['Purchase'].std()
age_std
```

```
668059612.6201671
```

```
len_age = len(age_group1)
```

```
CLT_age_by_expense = (age_mean) * (age_std) / len_age
CLT_age_by_expense1 = CLT_age_by_expense + 1.96
CLT_age_by_expense2 = CLT_age_by_expense - 1.96
CLT_age_by_expense2 ,CLT_age_by_expense1
```

```
(6.9475646661335336e+16, 6.9475646661335336e+16)
```

The CLT Interval is overlapping here for both upper and lower limit

✓ Final Insights :

The Following key points regarding CLT intervals are :

1- When we calculate the Confidence interval at different level of confidence for Males Expense (i.e, 95% ,90% ,99%)

- This means we are 95% confident that the true population mean lies between (114.0486804576682, 117.96868045766819)
- This means we are 90% confident that the true population mean lies between (114.3636804576682, 117.65368045766819)
- This means we are 99% confident that the true population mean lies between (113.4326804576682, 118.58468045766818)

2- When we calculate the Confidence interval at different level of confidence for Females Expense (i.e, 95% ,90% ,99%)

- This means we are 95% confident that the true population mean lies between (304.64495757390586, 308.5649575739058)
- This means we are 90% confident that the true population mean lies between (304.95995757390585, 308.2499575739058)
- This means we are 99% confident that the true population mean lies between (304.0289575739058, 309.18095757390586)

3- Confidence intervals and distribution of the mean of the expenses by married and singles

- *The CLT for married at 95% confidence level lies between 204.23 to

208.15

- The CLT for Single Marital Status at 95% confidence level lies between 145.41 to 141.49

4- Confidence intervals and distribution of the mean of the expenses by Age group

- The lower and the upper limit is overlaped here
- Same lower and upper limit - (6.9475646661335336e+16, 6.9475646661335336e+16)

✓ Recommendations

- There are more Males customer then Females. so they must focus on the population of females
- There are more Single customer then Married once, The percenatge contribution of Single are 59.04% where married were contributing 40.96% , which indicates that they need to attract more married customers
- There are 5891 Unique User IDs only out of which User ID - 1006040 has the max amount sales off rupee 10536909 and minimum sales is of 1000001 of rupee is 46681

- The mean of purchase amount by gender - females -8734.565765 where male has - 9437.58, which shows us that Females have more spending habit then males , perhaps males have three times more population then females.
 - They should attract more Male customer so that they spend more as the population is 3x but the spending habit is 3x less then females
 - Sample Size Considerations: CLT suggests that as the sample size increases, the sample mean will be more normally distributed, and confidence intervals will become narrower.
 - If the company is currently working with a small sample size, you could suggest increasing the sample size to improve precision and narrow the confidence interval further. This will improve the estimate of the true population mean and make it more reliable for decision-making.
 - Tailor confidence interval levels based on the company's decision-making needs—use the 90% interval for more precision and quicker decisions, the 95% interval as a balanced approach, and the 99% interval for maximum caution.
 - The confidence intervals can help the company estimate male expenses more accurately and forecast budget allocations with different levels of certainty
-

- There are more Males customer then Females. so they must focus on the population of females
- There are more Single customer then Married once, The percenatge contribution of Single are 59.04% where married were contributing 40.96% , which indicates that they need to attract more married customers
- There are 5891 Unique User IDs only out of which User ID - 1006040 has the max amount sales off rupee 10536909 and minimum sales is of 1000001 of rupee is 46681
- The mean of purchase amount by gender - females -8734.565765 where male has - 9437.58, which shows us that Females have more spending habit then males , perhaps males have three times more population then females.
- They should attract more Male customer so that they spend more as the population is 3x but the spending habit is 3x less then females
- Sample Size Considerations: CLT suggests that as the sample size increases, the sample mean will be more normally distributed, and confidence intervals will become narrower.