**Business Case: Walmart - Confidence Interval and CLT**

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

**Business Problem**

Analyze the customer purchase behavior (specifically, purchase amount) against the customer’s gender and the various other factors.

How the spending habits differ between male and female customers : Do women spend more on Black Friday than men? 

**Dataset**

The company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday.

Data contains 5,50,068 rows of data in 10 columns as explained below.

***Analyzing Basic Metrics and Non-Graphical Analysis :***

(550068, 10)

**Columns Contains :**

Column                                            | number of non-null values   |  Datatype   |    About Column

0 User\_ID 550068 non-null int64 Unique User Id of Customer

1 Product\_ID 550068 non-null object Product ID

2 Gender 550068 non-null object Sex of Customers

3 Age 550068 non-null object Age group

4 Occupation 550068 non-null int64 Occupation (Masked)

5 City\_Category 550068 non-null object Category of the City (A,B,C)

6 Stay\_In\_Current\_City\_Years 550068 non-null object Number of years stay in Current City

7 Marital\_Status 550068 non-null int64 Marital Status

8 Product\_Category 550068 non-null int64 Product Category

9 Purchase 550068 non-null int64 Purchase Amount

No Null Values Found.

**Number of Unique Values Per Column.**

* 5891 unique customers
* 3631 unique products
* 7 different Age groups
* 3 different City Categories
* stay in current city from 0 to 5 years
* Gender , Marital status
* 20 different Product Category

**Categorial Data:**

User\_ID

Product\_ID

Gender # changed values F and M to Female and Male.

Age

Occupation

City\_Category

Stay\_In\_Current\_City\_Years

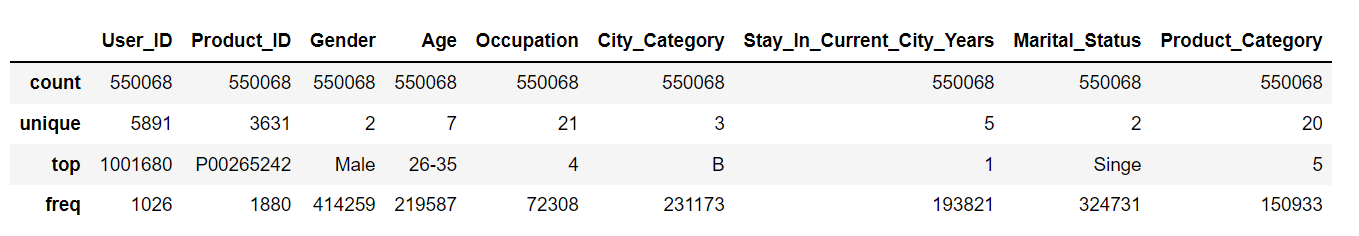
Marital\_Status # changed values 0 and 1 to single and married

Product\_Category

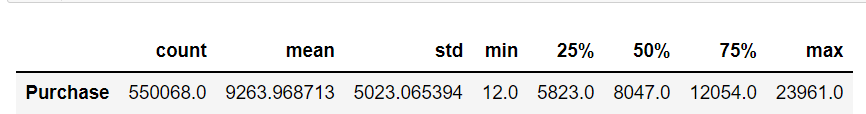
**Numerical Data:**

 Purchase Amount

**Statistical Summery :**

****

from data set :

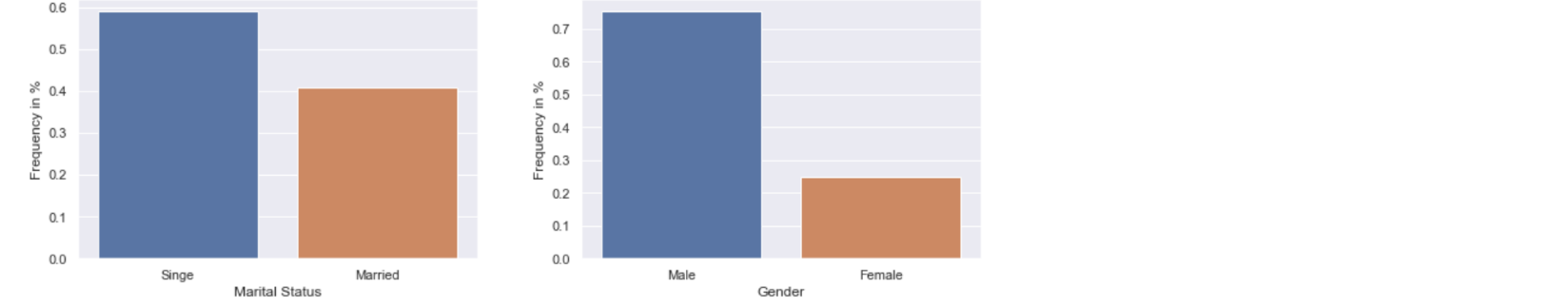
****

There is 1200 of difference in Median Customer Purchase and Mean. (outlier detection is in Confidence Interval Analysis below)

***Visual Analysis :***

Distributions with in categories  from given dataset :

Singe : 59.03 % Married : 40.96 % Male : 75.31 % Female : 24.68 %

****

59%(majority) customers are single and 41% are married.

**from the given data:**

75.31 % customers are male

24.68 % customers are female

**from the prproblem statement :**

company has 50 million customers are male and 50 million are female overall.

Given Sample has a gender bias.

Age Group Frequency **AGE**

26-35 39.91 %

36-45 19.99 %

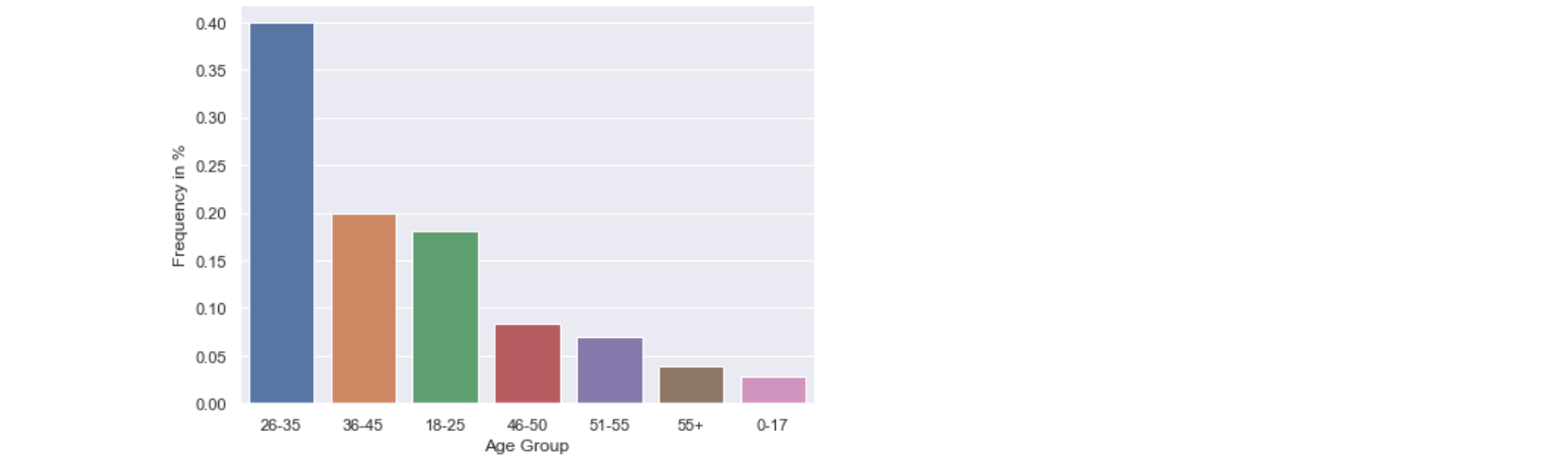
18-25 18.11 %

46-50 8.30 %

51-55 6.99 %

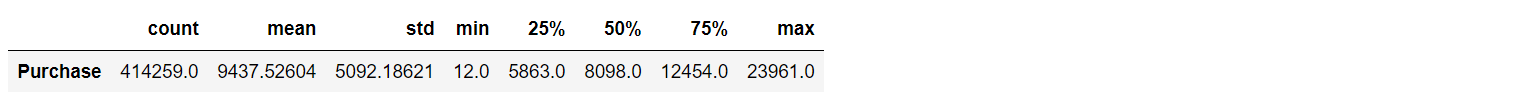
55+ 3.90 %

0-17 2.74 %

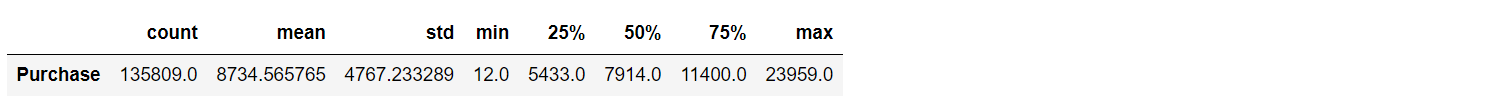
****

Majority customers are from age 18 to 45 years.

**Purchase Statistics for Male Customers:**

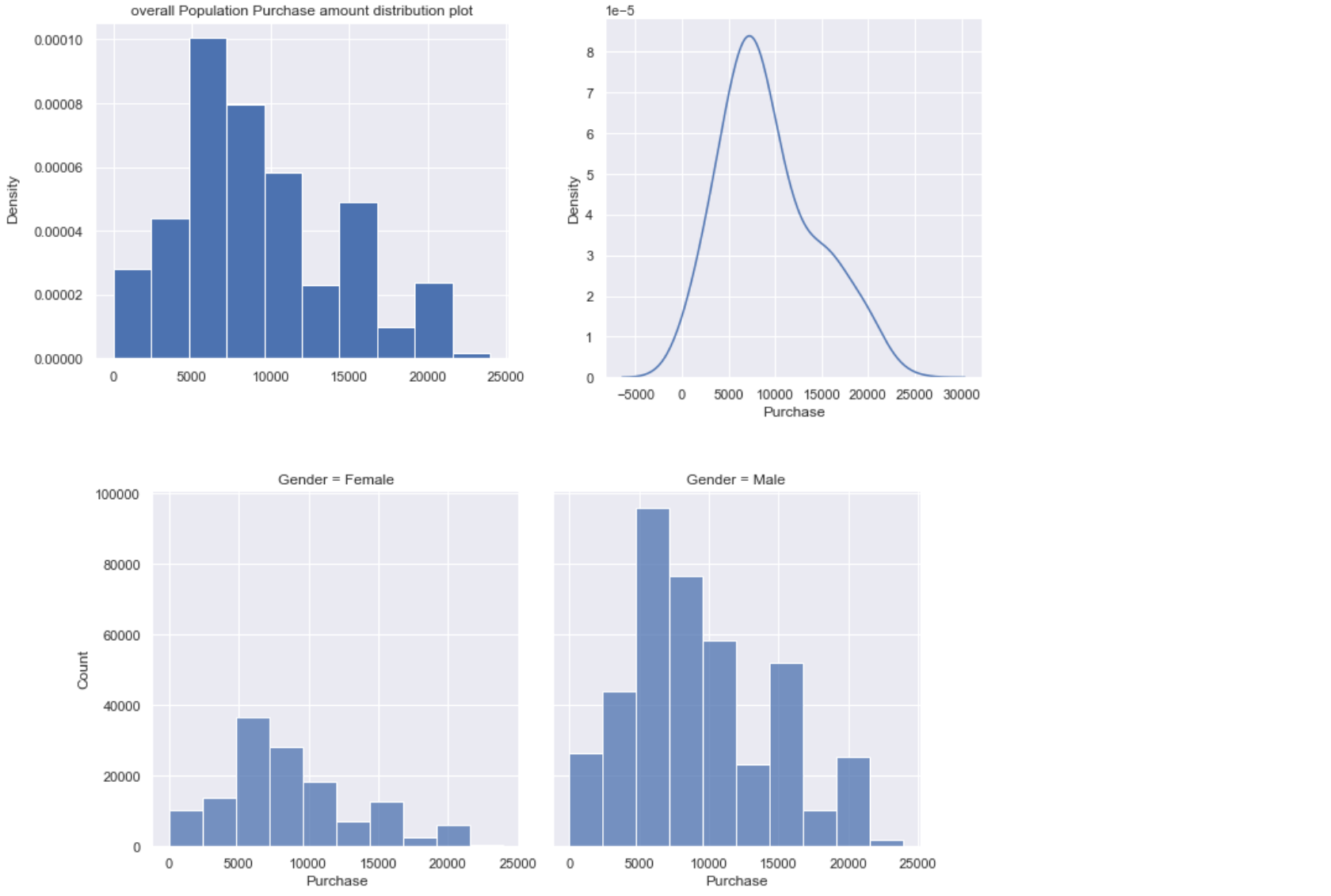
****

**Purchase Statistics for Female Customers:**

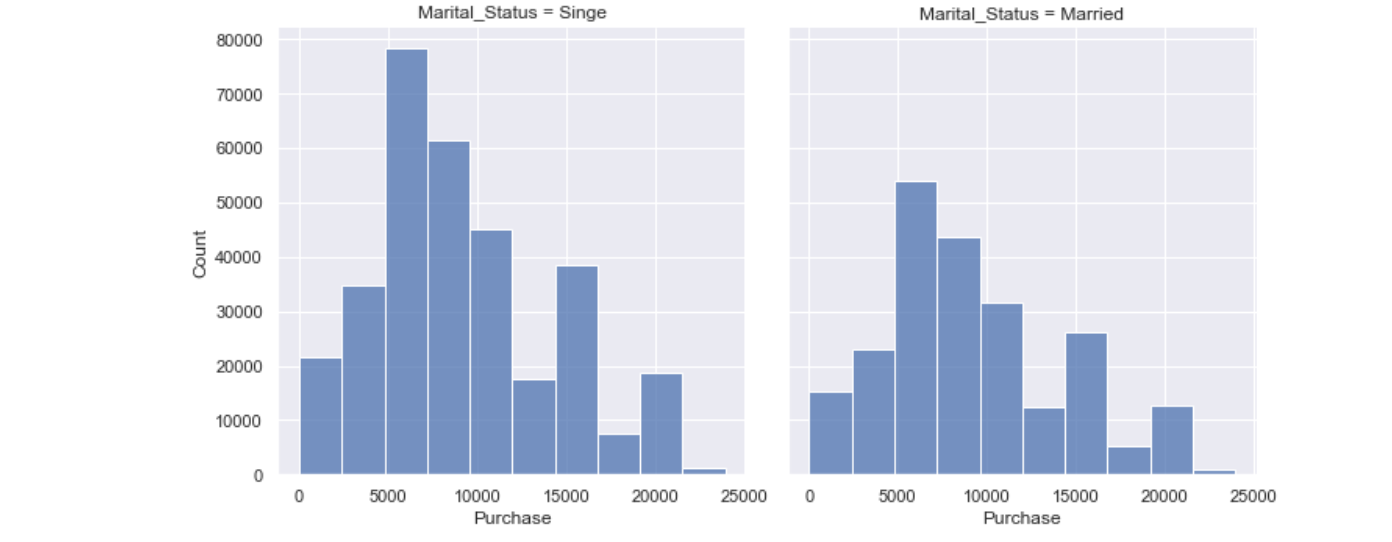
****

**Distribution Plot for Purchase Amount   :  for overall sample , and below two graphs  for Male and Female.  (check if data is distributed normally for further analysis.)**

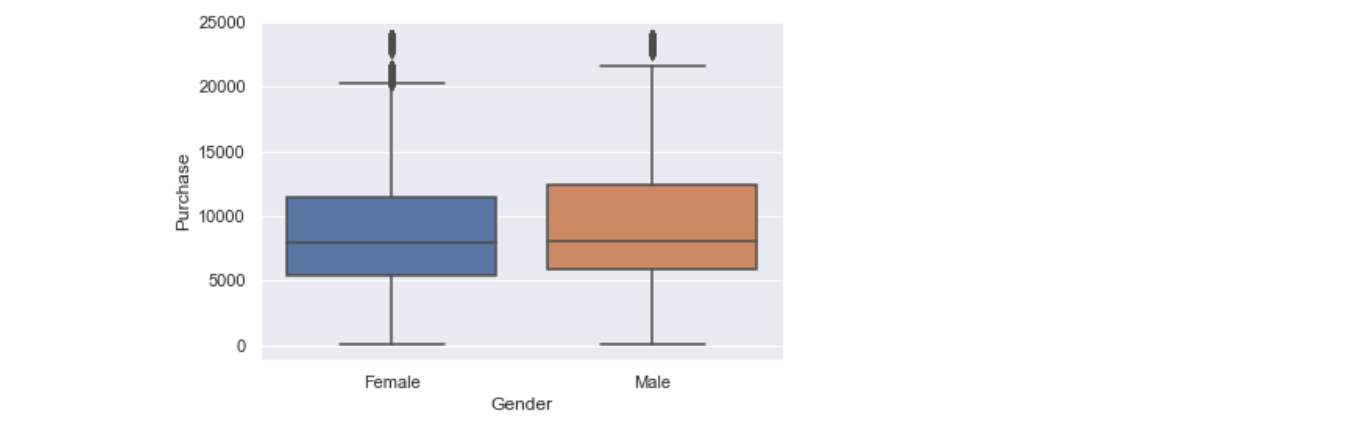
For Male and Female Customer Purchase Amount's Distribution

****

For Single and Married Customer Purchase Amount's Distribution

****

0.0049% outlier data detected from given data.

****

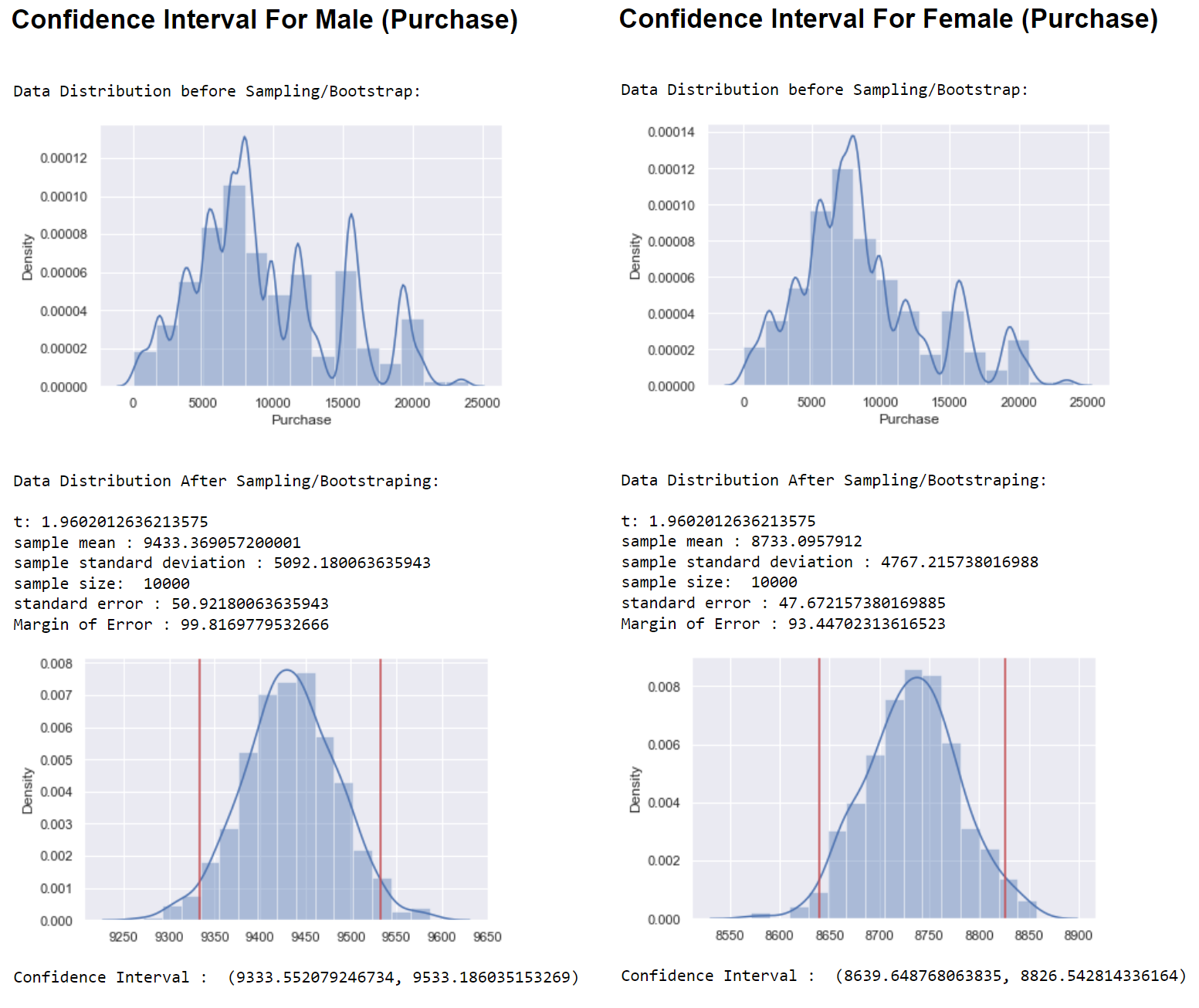
after removing outliers:

**Purchase data in given dataset is not perfectly normally distributed . so we can handle the data using Central Limit Theorem and Bootstrapping Method.**

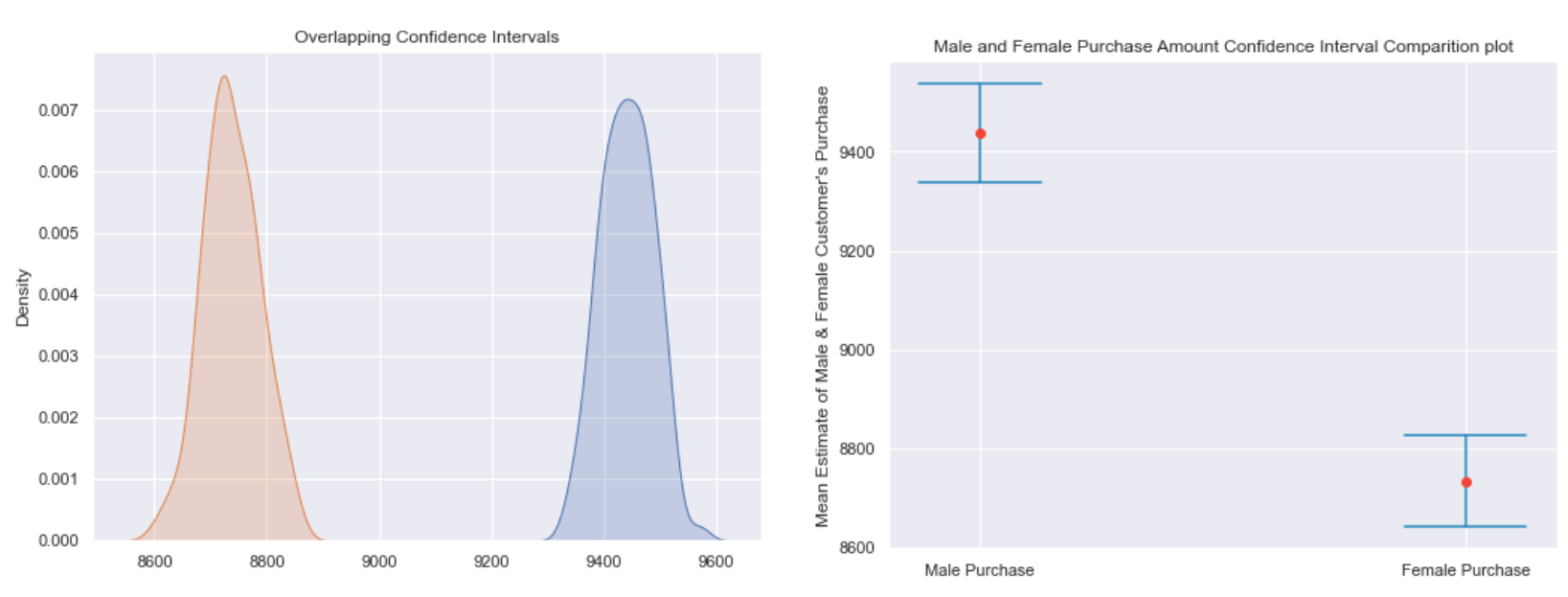
**Sample Size : 10,000**

**Trials : 500**

***Confidence Interval for Male and Female Customer's Purchase Amount:***

****

Confidence Interval Comparison for Male and Female Purchase Amount Distribution .

****

(with 95% confidence and sample size of 10000 , 500 trials. .)

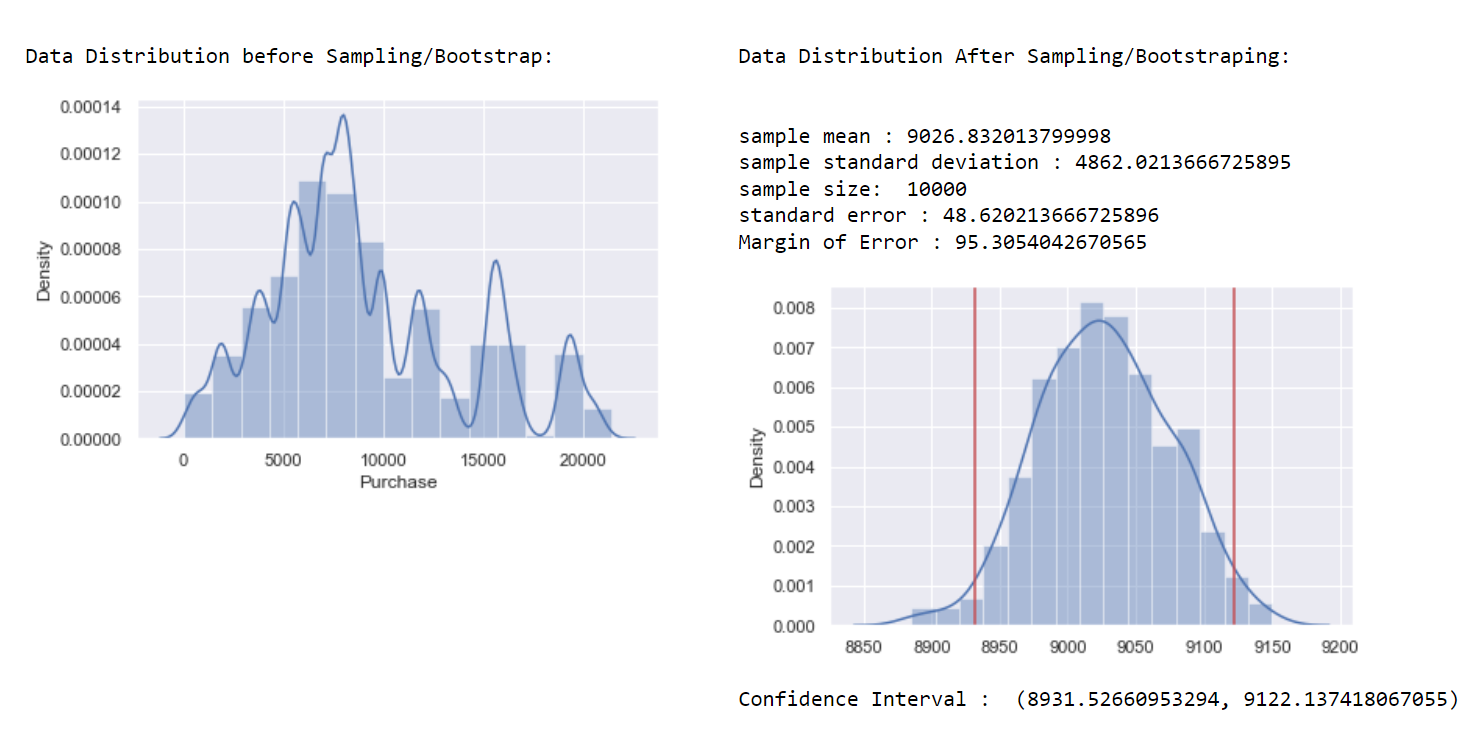
As per confidence Interval comparison for both female purchase and male purchase data , its clear  that there's no over lapping , and hence there's a good amount of difference between Male and Female Spending amounts .   
**Male Customers are more likely to spend more amount than female customers .**

Average Male Spending Amount from all 100 million customers lies in Range of 9333 to 9533 as per Bootstrapping Method .

Average Female Spending Amount from all 100 million customers lies in Range of 8639 to 8826 as per Bootstrapping Method .

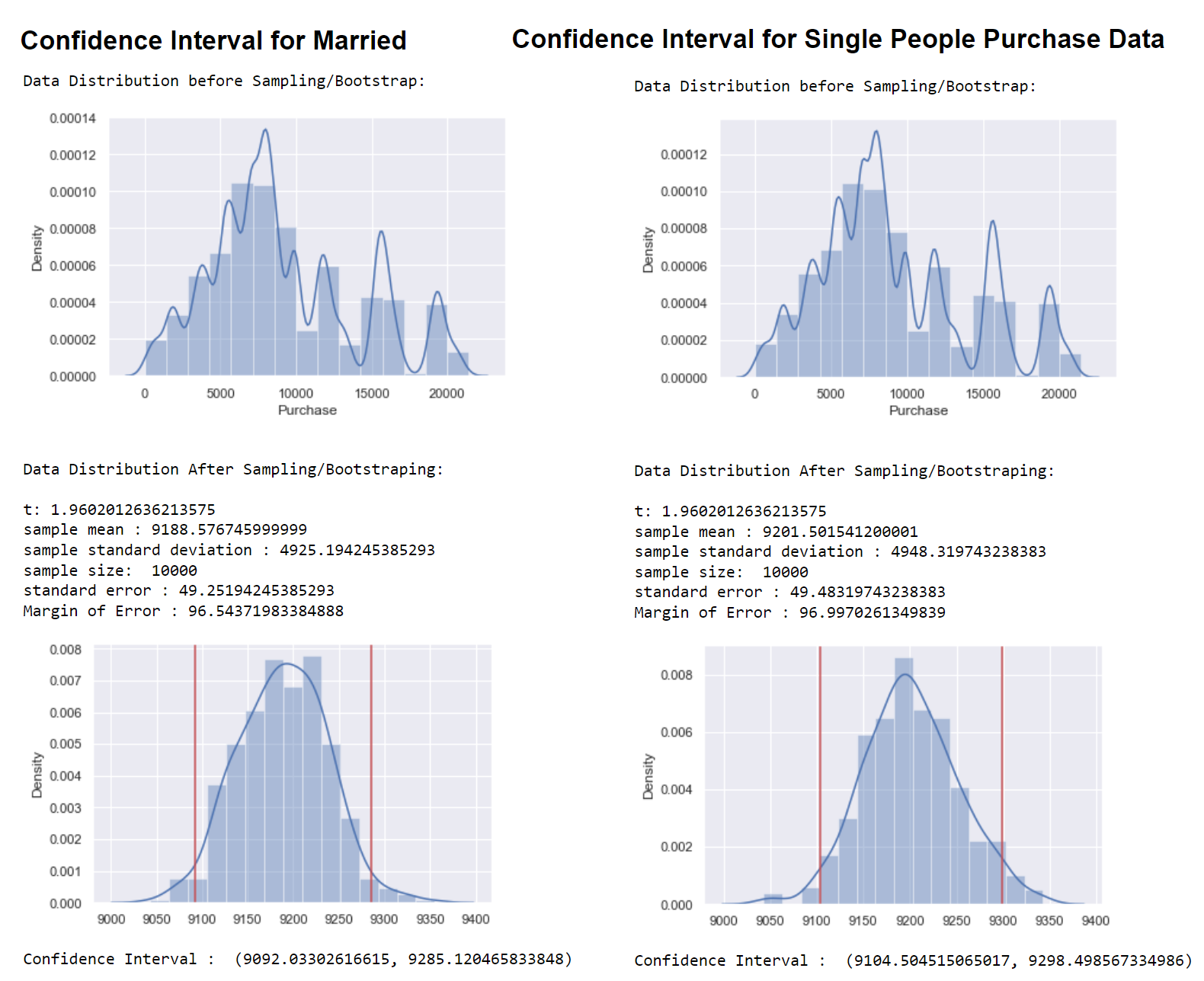
***Estimation and Confidence Interval for all customers (unknow population data) average spending/ purchase amount :***

because of gender bias , data resampling was needed.  after resampling , using CLT and Bootstrap method ,

****

All 100 million customer's average spending amount lies between 8931 to 9122  and sample mean is 9026. (with 95% confidence and sample size of 10000 , 500 trials. .)

***Confidence Interval for Single and Married Customer's Purchase Amount:***

****

Confidence Interval Comparison for Married and Single Customer's Purchase Amount Distribution .

****

As per confidence Interval comparison for both **Single and Married Customer's average purchase data** , that**there is not much difference between their average spending amounts**. Married and Single Customer's spending amounts distribution are almost lies with same distribution.  (with 95% confidence and sample size of 10000 , 500 trials. .)

***Confidence Interval for for different age group Customer's Purchase Amount:***

(with 95% confidence and sample size of 10000 , 500 trials. .)

Age group  Value counts in %.

26-35 39.919974

36-45 19.999891

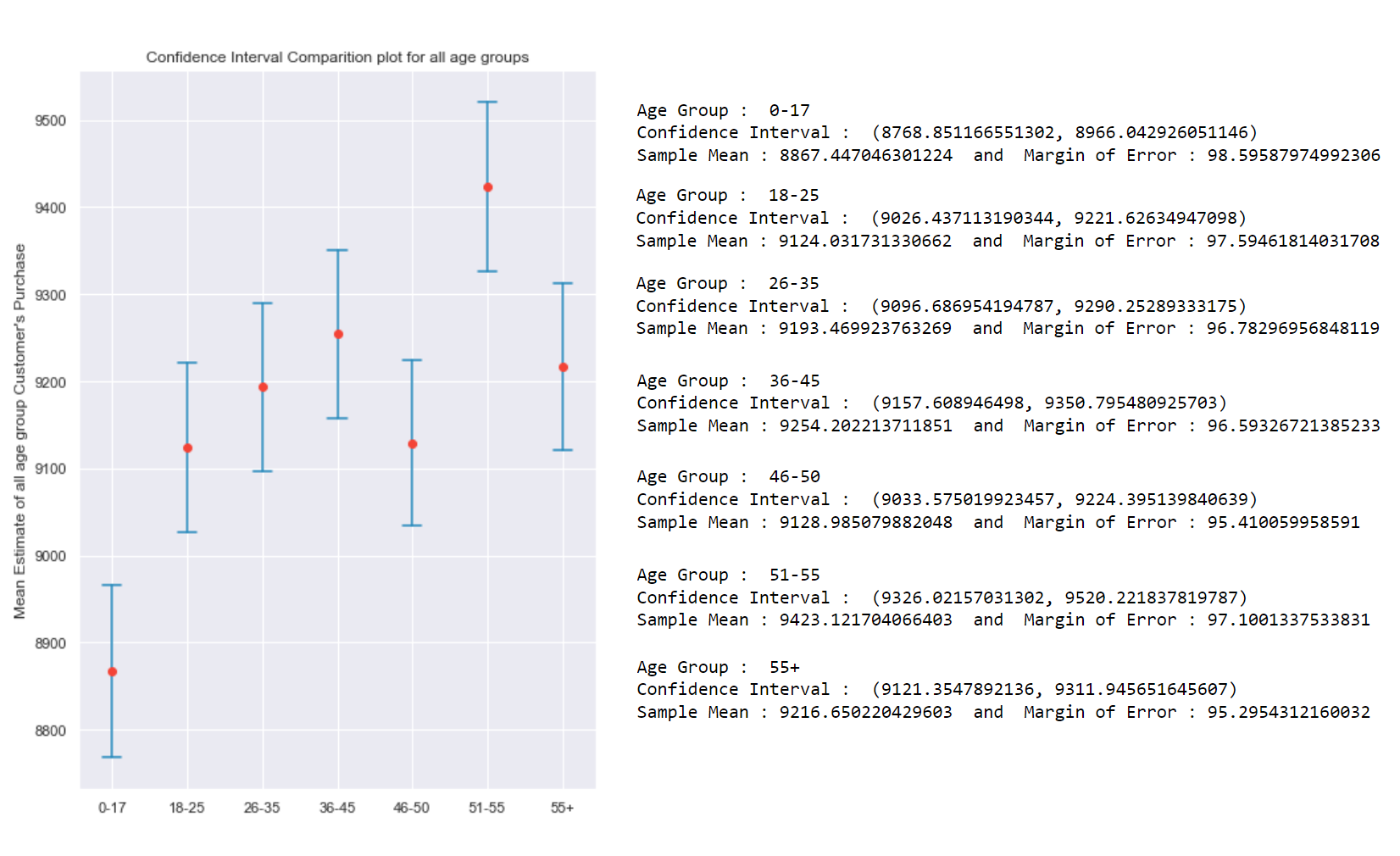
18-25 18.117760

46-50 8.308246

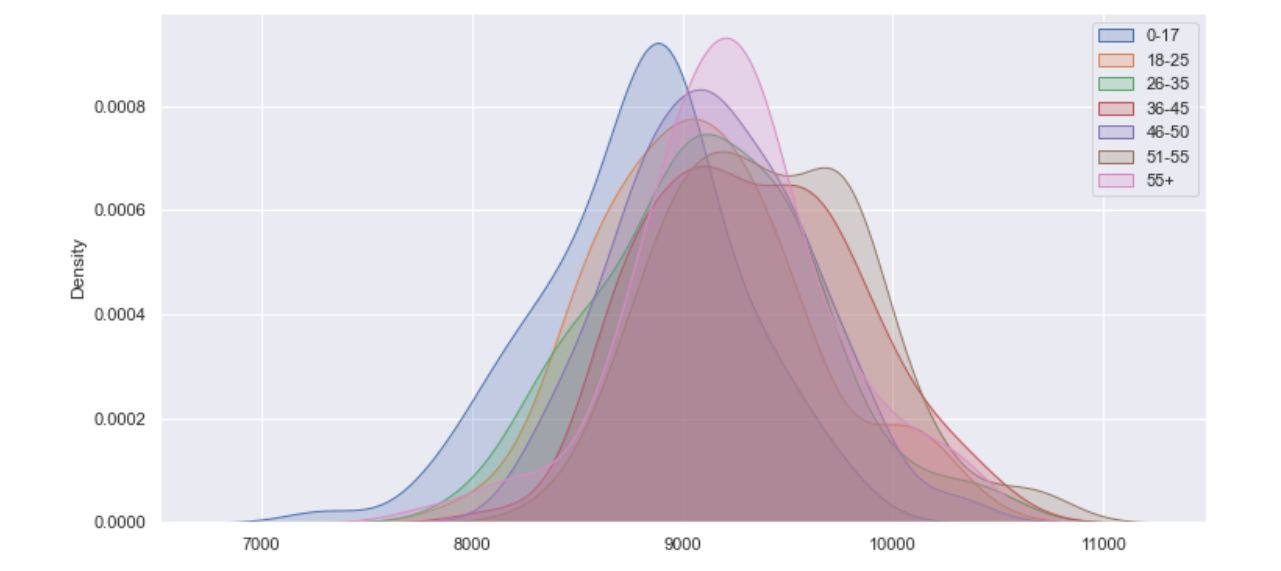
51-55 6.999316

55+ 3.909335

0-17 2.745479

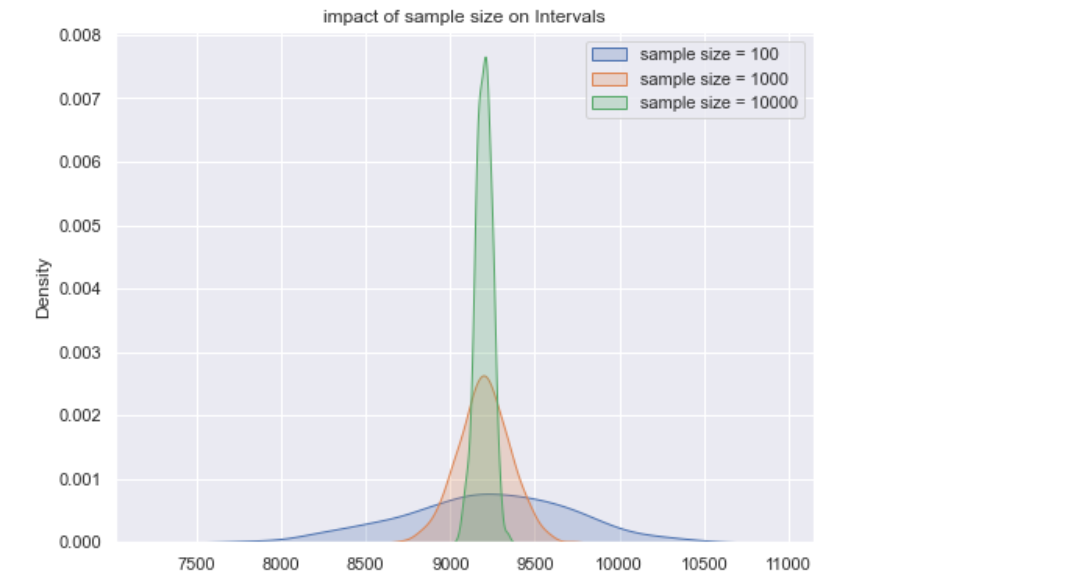
****

Customers from age 26-35 are 40% of all customers. and their Average Spending amount is near to overall customers average spending amount.

****

Age group 51-55 customers are more likely to spend more amount than all other groups. and customers under 17 age are the least spending average amount.

**Impact on Confidence Interval for All Customer Purchase data Average, according to different different Sample Sizes : Confidence level  95% .**

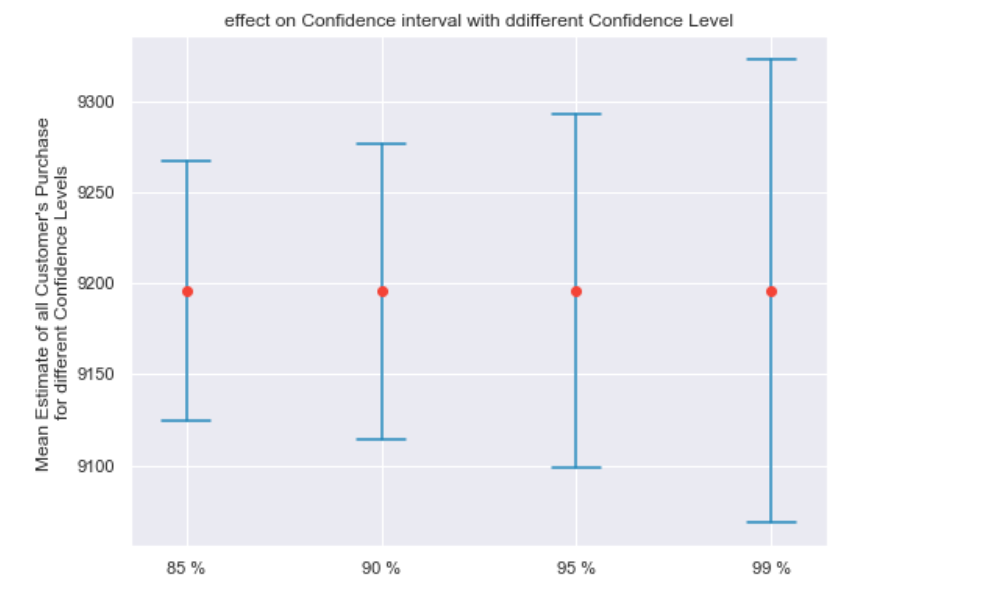
****

As per calculations(in Jupyter notebook) and above distribution plot, as we increase the sample size, standard error decreases , means that the average spending amount gets closers and closer to the actual mean spending amount of the all customer average spending amount.

**Effect on Confidence Interval for All Customer Purchase data Average, according to different different Confidence Level :**

**sample size = 10,000**

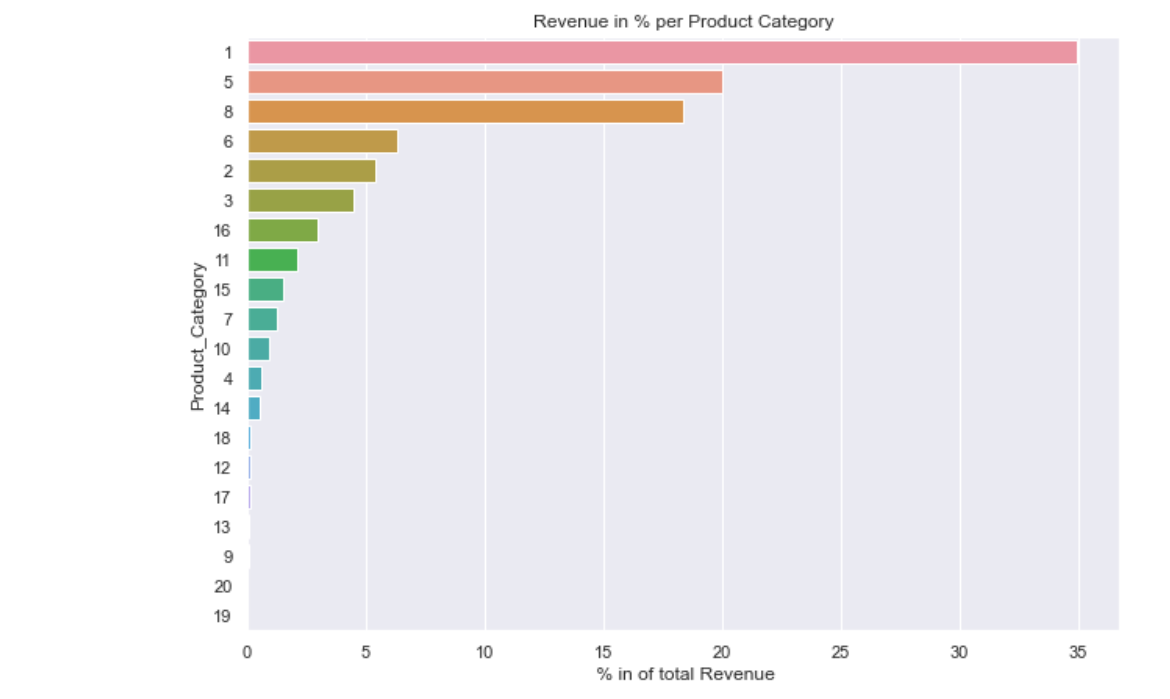
**500 trials**

****

If we require more confidence for estimate of overall customer purchase data , as per above plot , the interval of purchase average amount gets wider.

***Product Category :***

Revenue Generated (in %) per Product Category :

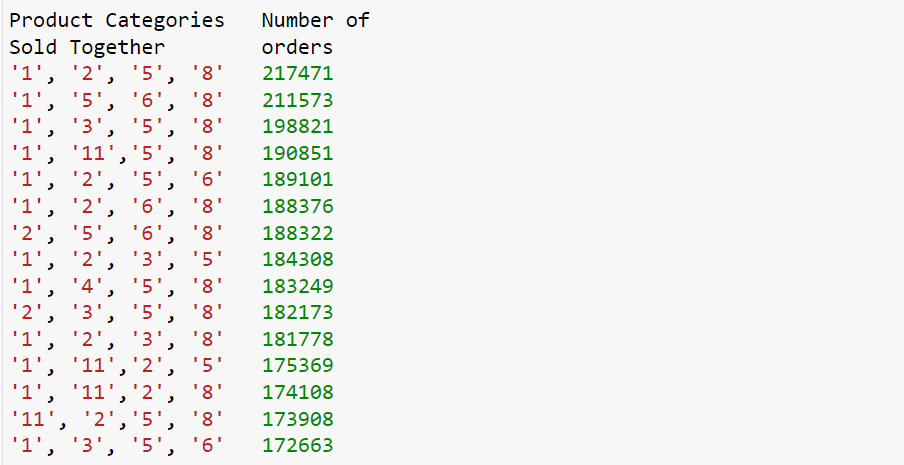
****

Bar Plot of Revenue per Product Category in Percentages:

Top 3 Highest Revenue generating Product Categories :

Category **1** is the highest revenue generating Product category ( 35 % )  
Category **5** is the highest revenue generating Product category ( 20 % )  
Category **8** is the highest revenue generating Product category ( 18 % )

**Most Common Product Categories which are sold together :**

****

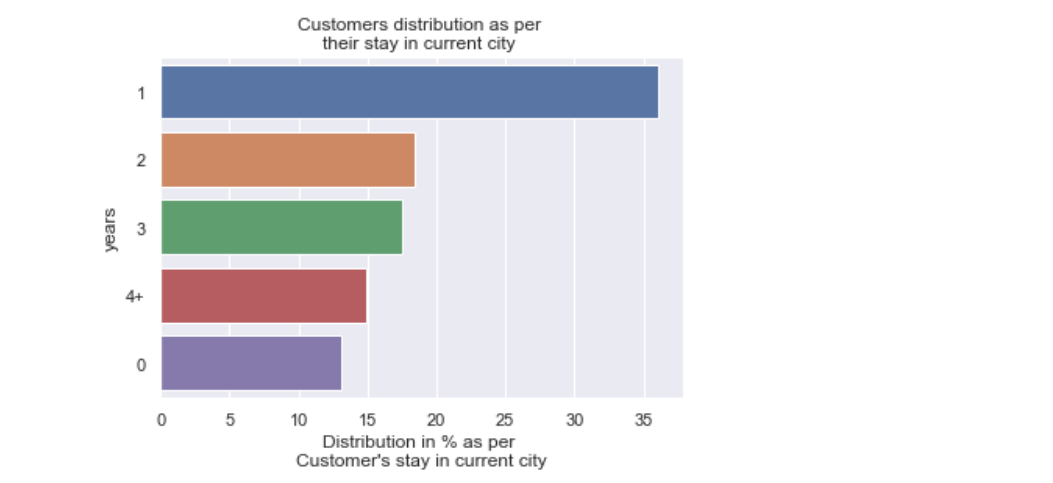
***City Category :***

Customers per City Category  :

****

City Category B has highest Customers Base compared to C and A category.

***Customers their stay in current city (in years) :***

****

***Revenue Generated as per different Age Groups:***

Age Group   Revenue(%)

0-17 2.953258

18-25 17.741691

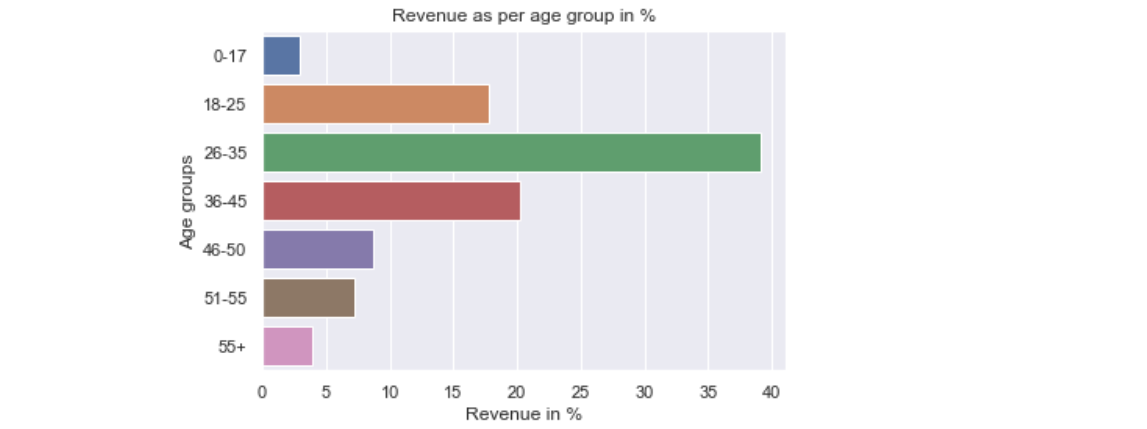
26-35 39.086842

36-45 20.289386

46-50 8.754860

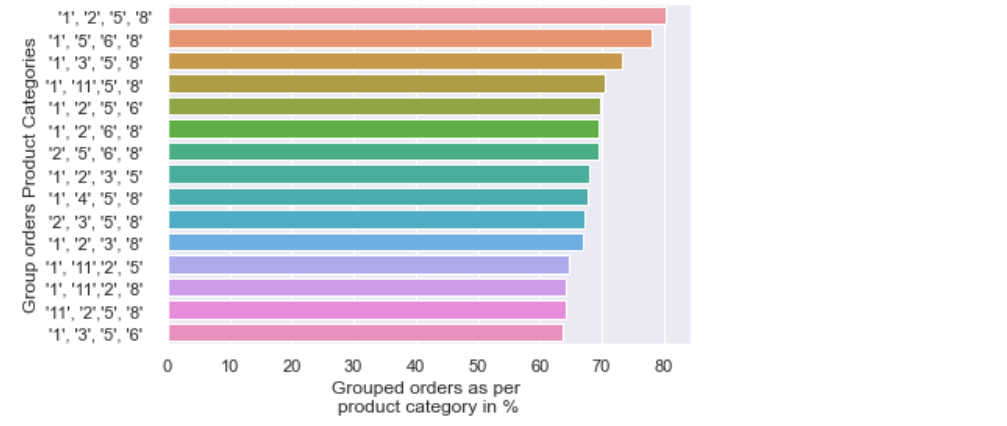
51-55 7.294806

55+ 3.879157

****

Most of the revenue is generated from Age 18 to 45

***Most Common (Top 15) Product Categories which are sold Together :***

****

Above product categories group are most commonly sold together.