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 :

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category
count	550068	550068	550068	550068	550068	550068	550068	550068	550068
unique	5891	3631	2	7	21	3	5	2	20
top	1001680	P00265242	Male	26-35	4	B	1	Singe	5
freq	1026	1880	414259	219587	72308	231173	193821	324731	150933

from data set :

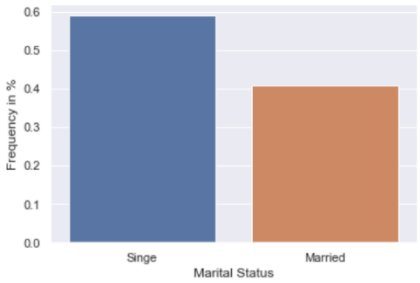
	count	mean	std	min	25%	50%	75%	max
Purchase	550068.0	9263.968713	5023.065394	12.0	5823.0	8047.0	12054.0	23961.0

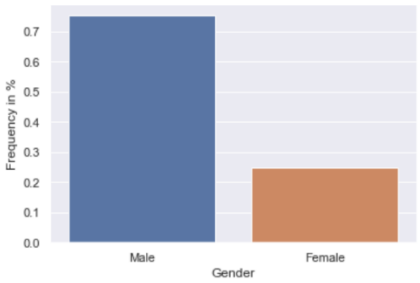
There is 1200 of difference in Median Customer Purchase and Mean. (outlier detection is in Confidence Interval Analysis below)

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

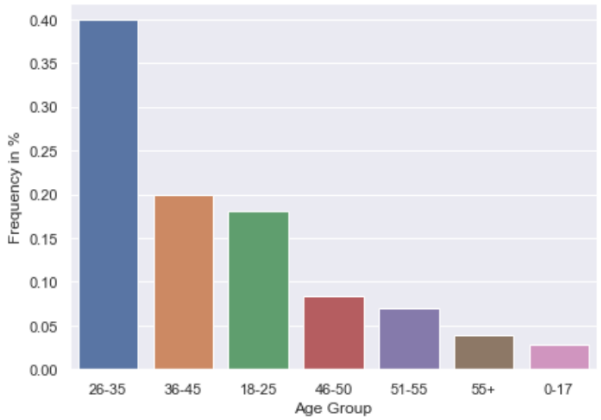
24.68 % customers are female

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:

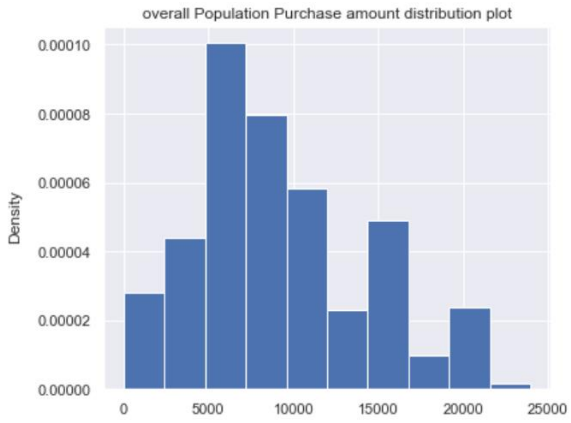
	count	mean	std	min	25%	50%	75%	max
Purchase	414259.0	9437.52604	5092.18621	12.0	5863.0	8098.0	12454.0	23961.0

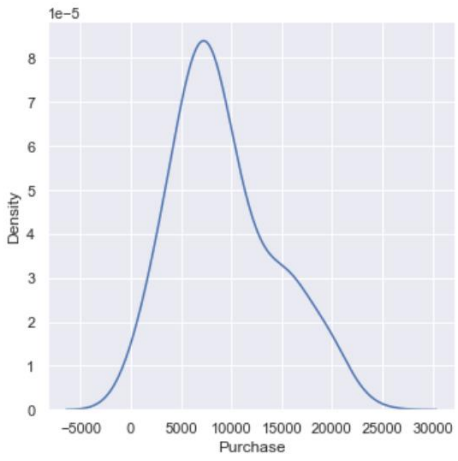
Purchase Statistics for Female Customers:

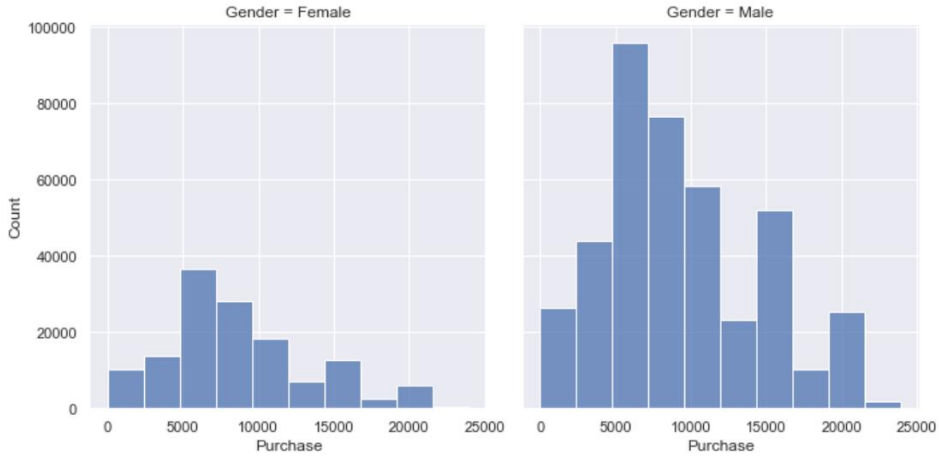
	count	mean	std	min	25%	50%	75%	max
Purchase	135809.0	8734.565765	4767.233289	12.0	5433.0	7914.0	11400.0	23959.0

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.)

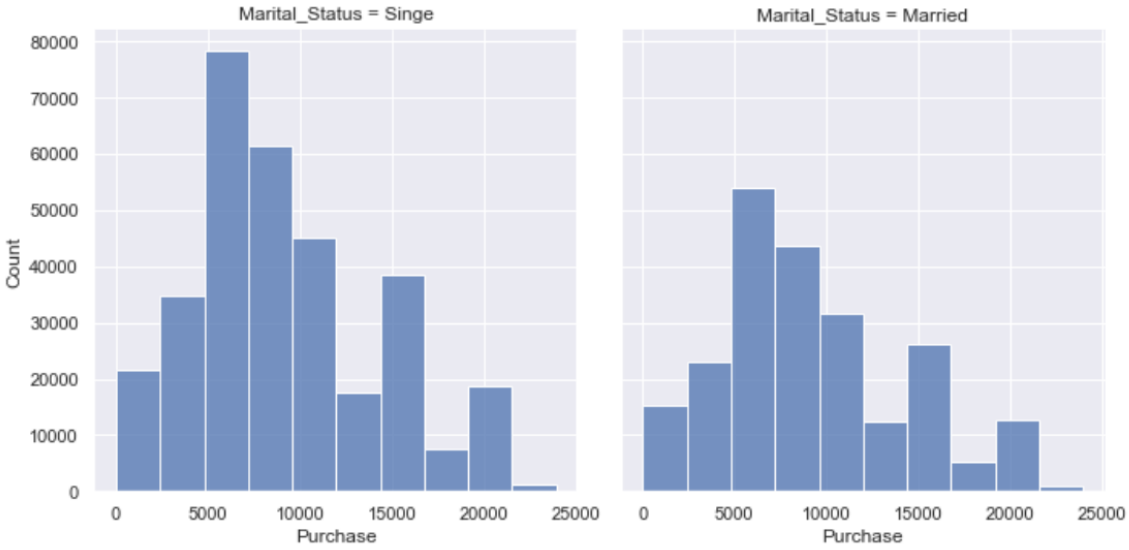
For Male and Female Customer Purchase Amount's Distribution



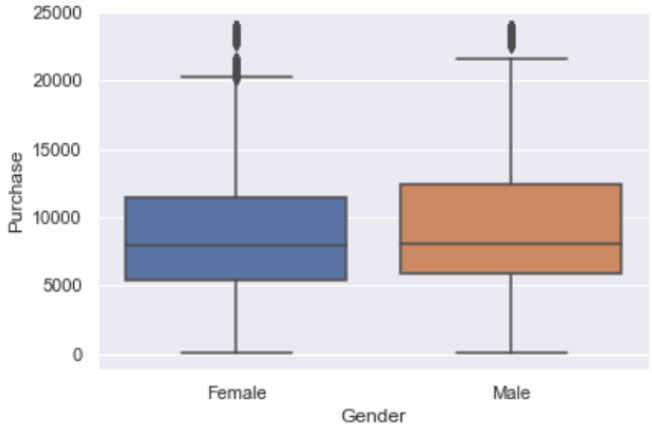




For Single and Married Customer Purchase Amount's Distribution



0.0049% outlier data detected from given data.



after removing outliers:

note : Confidence Interval and some analysis parameter can be slightly different than Jupyter notebook due to Random Sampling. though all the figures are double checked, if they are not having a large gap.

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

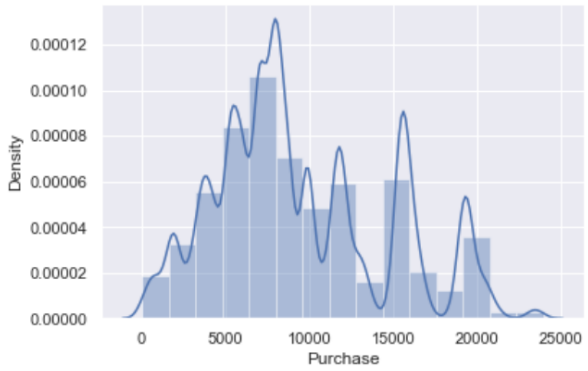
Sample Size : 10,000

Trials : 500

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

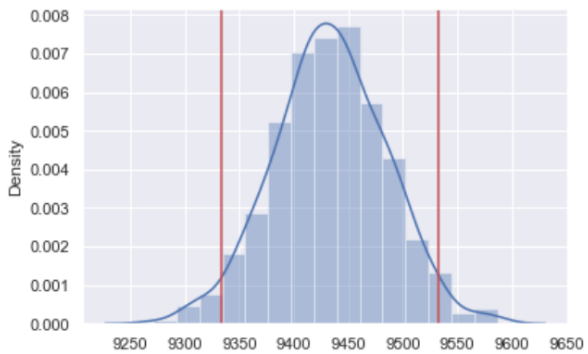
Confidence Interval For Male (Purchase)

Data Distribution before Sampling/Bootstrap:



Data Distribution After Sampling/Bootstraping:

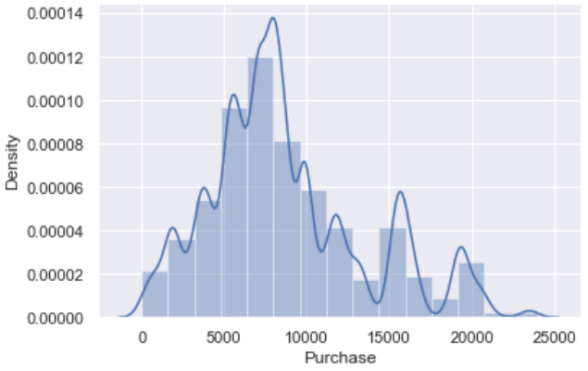
```
t: 1.9602012636213575
sample mean : 9433.369057200001
sample standard deviation : 5092.180063635943
sample size:  10000
standard error : 50.92180063635943
Margin of Error : 99.8169779532666
```



Confidence Interval : (9333.552079246734, 9533.186035153269)

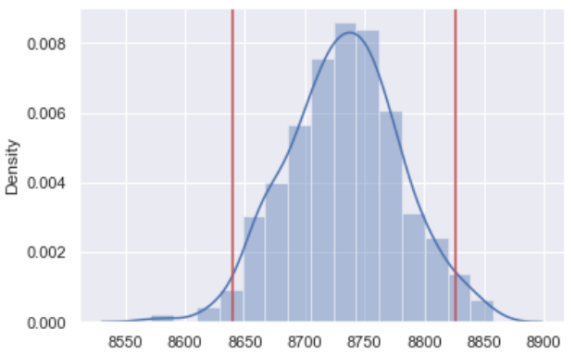
Confidence Interval For Female (Purchase)

Data Distribution before Sampling/Bootstrap:



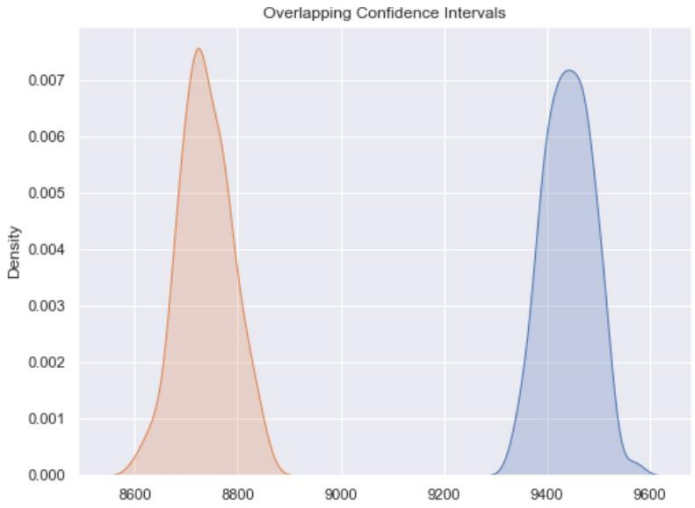
Data Distribution After Sampling/Bootstraping:

```
t: 1.9602012636213575
sample mean : 8733.0957912
sample standard deviation : 4767.215738016988
sample size:  10000
standard error : 47.672157380169885
Margin of Error : 93.44702313616523
```



Confidence Interval : (8639.648768063835, 8826.542814336164)

Confidence Interval Comparison for Male and Female Purchase Amount Distribution .



(with 95% confidence and sample size of 10000 with 500 repetition)

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 .

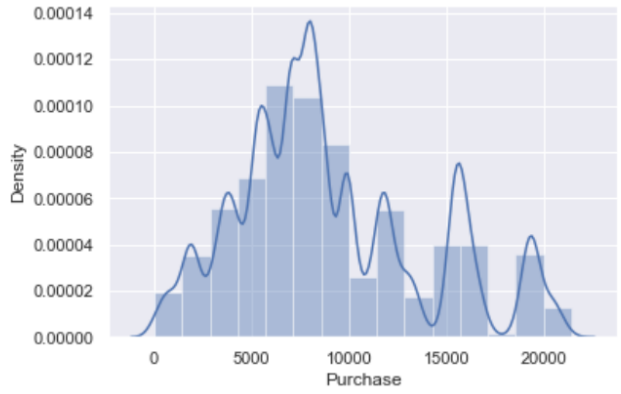
Average Male Spending Amount from all 100 million customers lies in Range of 9333 to 9533 as per Bootstrapping Method .

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 :

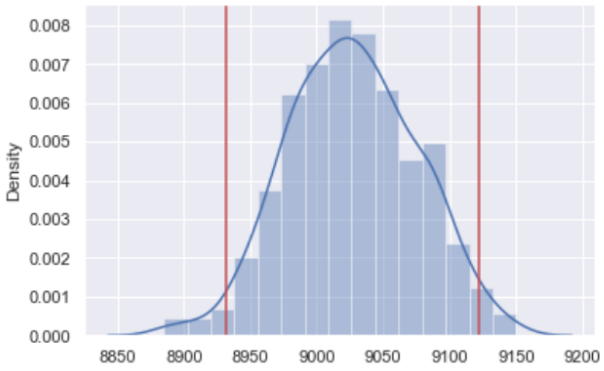
because of gender bias , data resampling was needed. after resampling , using CLT and Bootstrap method ,

Data Distribution before Sampling/Bootstrap:



Data Distribution After Sampling/Bootstraping:

```
sample mean : 9026.832013799998
sample standard deviation : 4862.0213666725895
sample size:  10000
standard error : 48.620213666725896
Margin of Error : 95.3054042670565
```



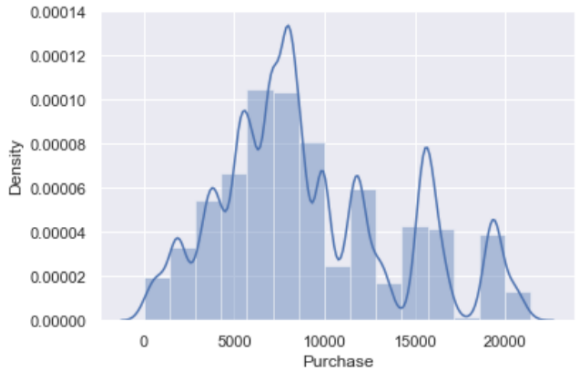

```
Confidence Interval :  (8931.52660953294, 9122.137418067055)
```

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 with 500 repetition)

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

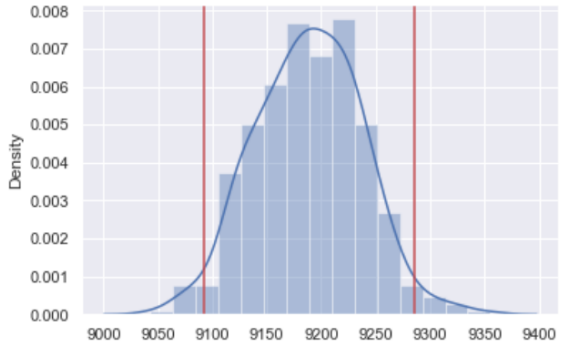
Confidence Interval for Married

Data Distribution before Sampling/Bootstrap:



Data Distribution After Sampling/Bootstraping:

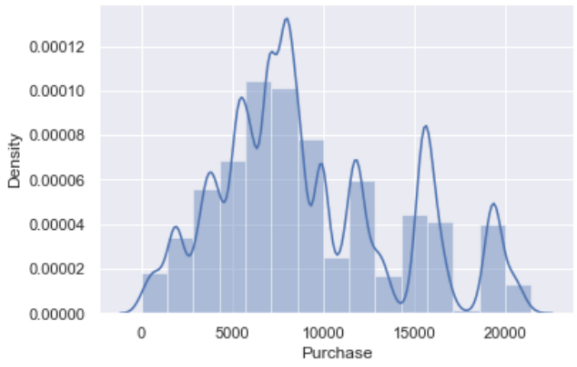
t: 1.9602012636213575
sample mean : 9188.576745999999
sample standard deviation : 4925.194245385293
sample size: 10000
standard error : 49.25194245385293
Margin of Error : 96.54371983384888



Confidence Interval : (9092.03302616615, 9285.120465833848)

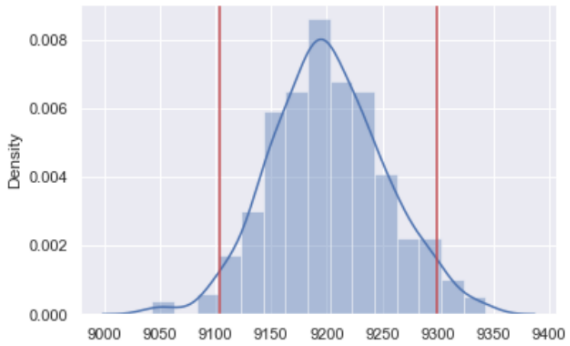
Confidence Interval for Single People Purchase Data

Data Distribution before Sampling/Bootstrap:



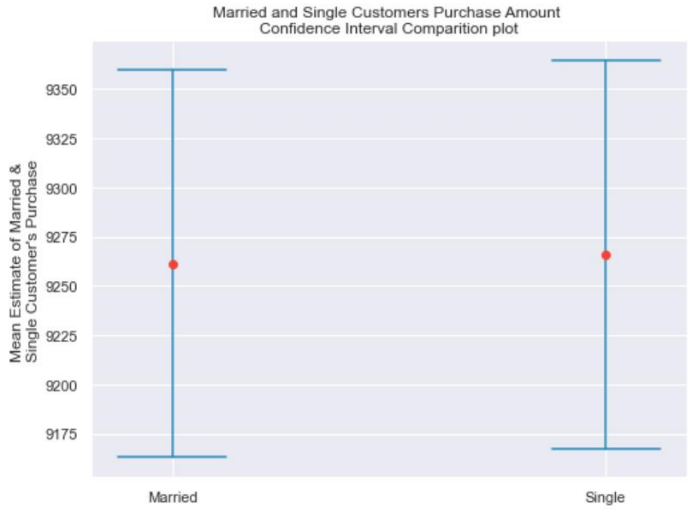
Data Distribution After Sampling/Bootstraping:

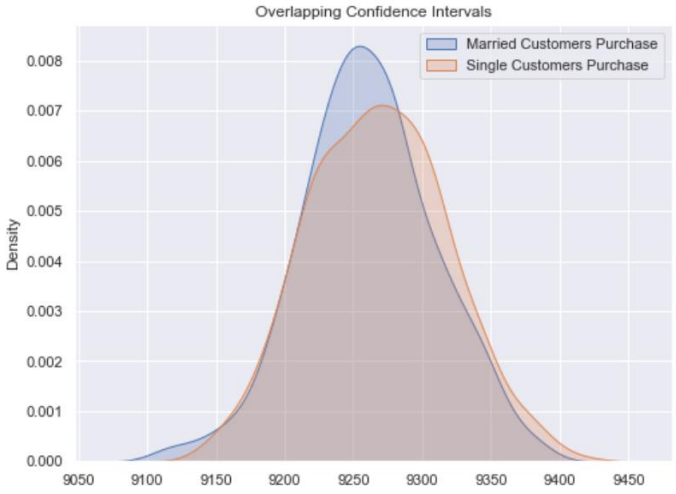
t: 1.9602012636213575
sample mean : 9201.501541200001
sample standard deviation : 4948.319743238383
sample size: 10000
standard error : 49.48319743238383
Margin of Error : 96.9970261349839



Confidence Interval : (9104.504515065017, 9298.498567334986)

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





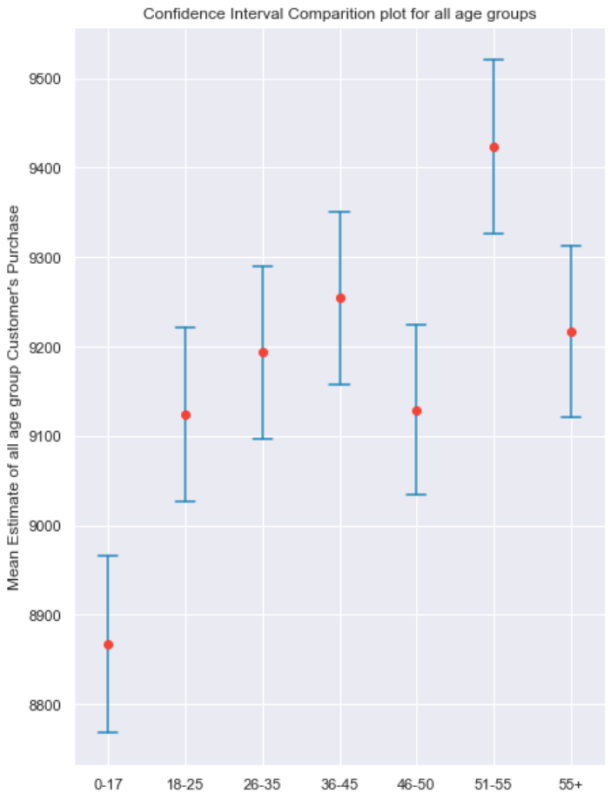
As per confidence Interval comparison for both **Single and Married Customer's average purchase data** , that **there is not much difference between their average spending amounts**. Married and Single Customer's spending amounts distribution are almost lies with same distribution. (with 95% confidence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

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

(with 95% confidence and sample size of 10000 with 500 repetition)

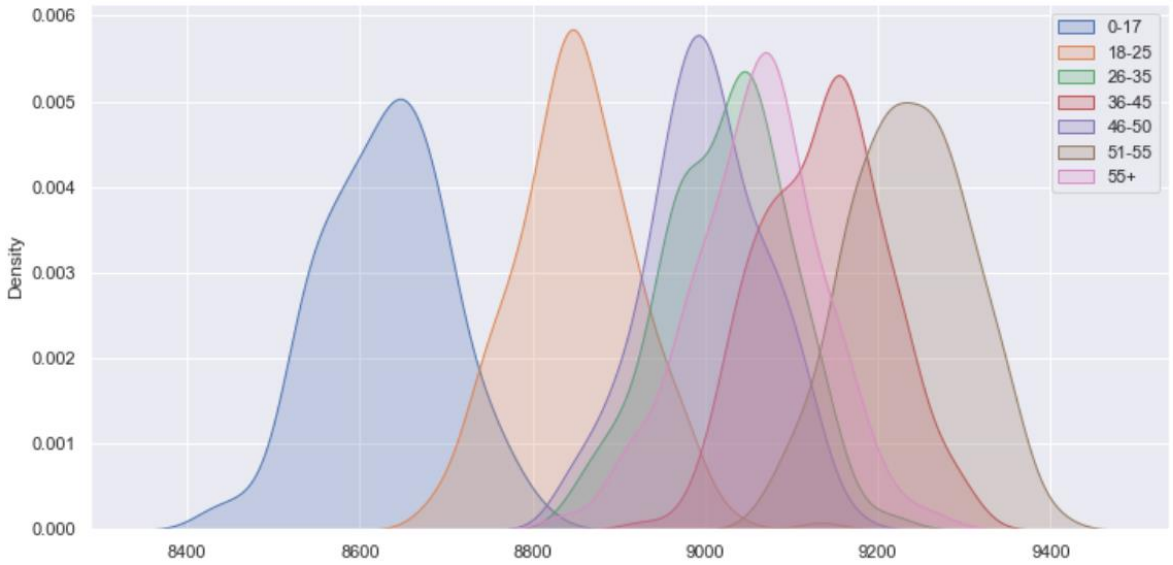
Age group Value counts in %.

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




```
Age Group :  0-17
Confidence Interval :  (8768.851166551302, 8966.042926051146)
Sample Mean : 8867.447046301224  and  Margin of Error : 98.59587974992306

Age Group :  18-25
Confidence Interval :  (9026.437113190344, 9221.62634947098)
Sample Mean : 9124.031731330662  and  Margin of Error : 97.59461814031708

Age Group :  26-35
Confidence Interval :  (9096.686954194787, 9290.25289333175)
Sample Mean : 9193.469923763269  and  Margin of Error : 96.78296956848119

Age Group :  36-45
Confidence Interval :  (9157.608946498, 9350.795480925703)
Sample Mean : 9254.202213711851  and  Margin of Error : 96.59326721385233

Age Group :  46-50
Confidence Interval :  (9033.575019923457, 9224.395139840639)
Sample Mean : 9128.985079882048  and  Margin of Error : 95.410059958591

Age Group :  51-55
Confidence Interval :  (9326.02157031302, 9520.221837819787)
Sample Mean : 9423.121704066403  and  Margin of Error : 97.1001337533831

Age Group :  55+
Confidence Interval :  (9121.3547892136, 9311.945651645607)
Sample Mean : 9216.650220429603  and  Margin of Error : 95.2954312160032
```

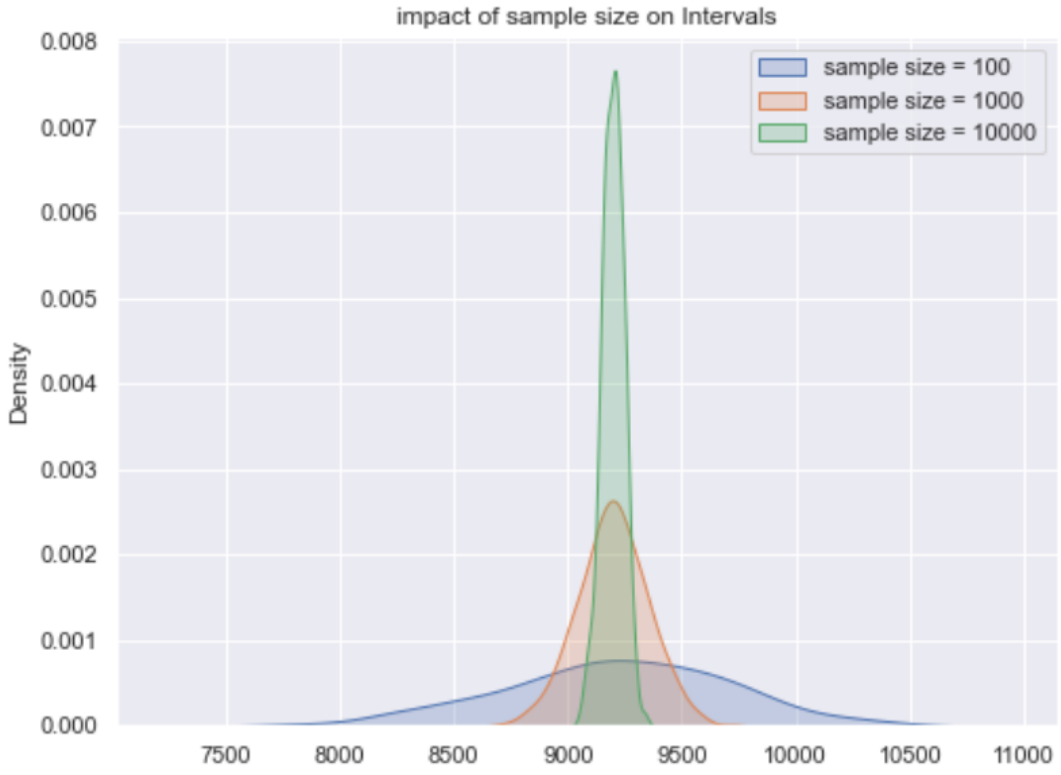
Customers from age 26-35 are 40% of all customers. and their Average Spending amount is near to overall customers average spending amount.



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.

as per distribution , Age 26- 35 has the highest frequency of customers.

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

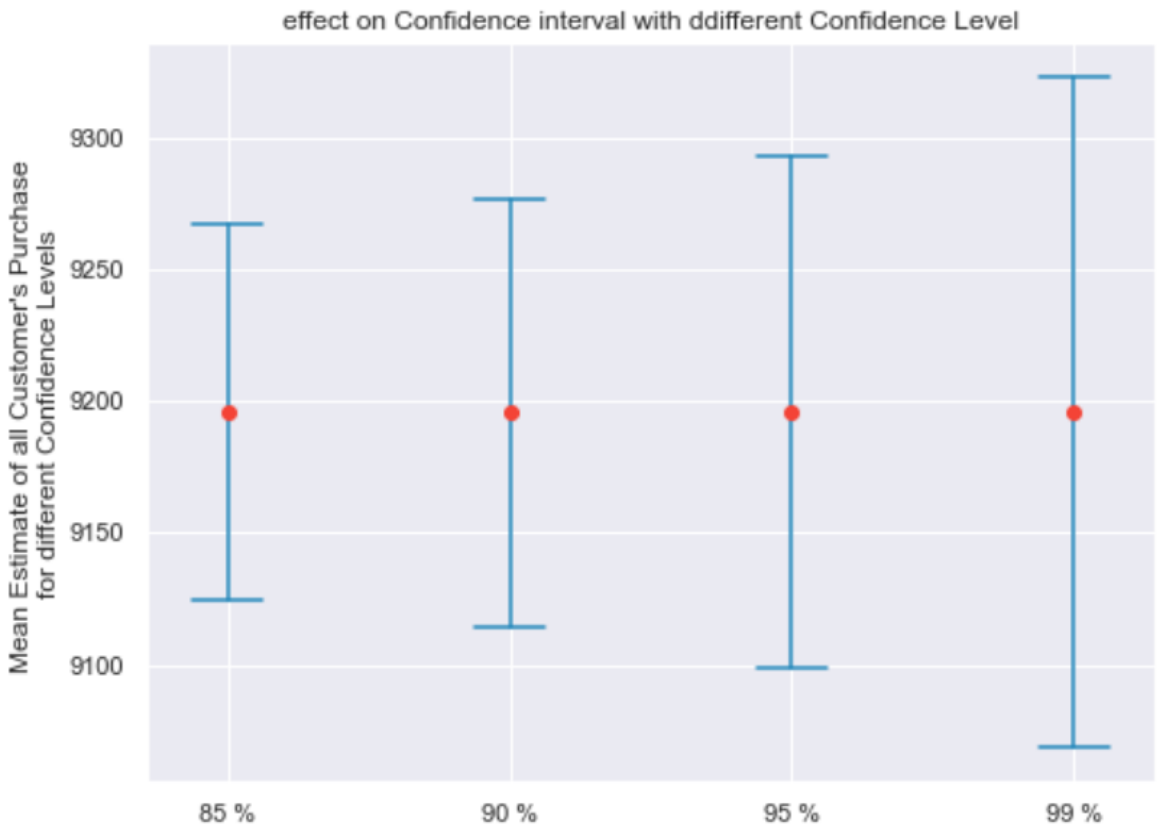


As per calculations(in Jupyter notebook) and above distribution plot, as we increase the sample size, standard error decreases , means that the average spending amount gets closers and closer to the actual mean spending amount of the all customer average spending amount.

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

sample size = 10,000

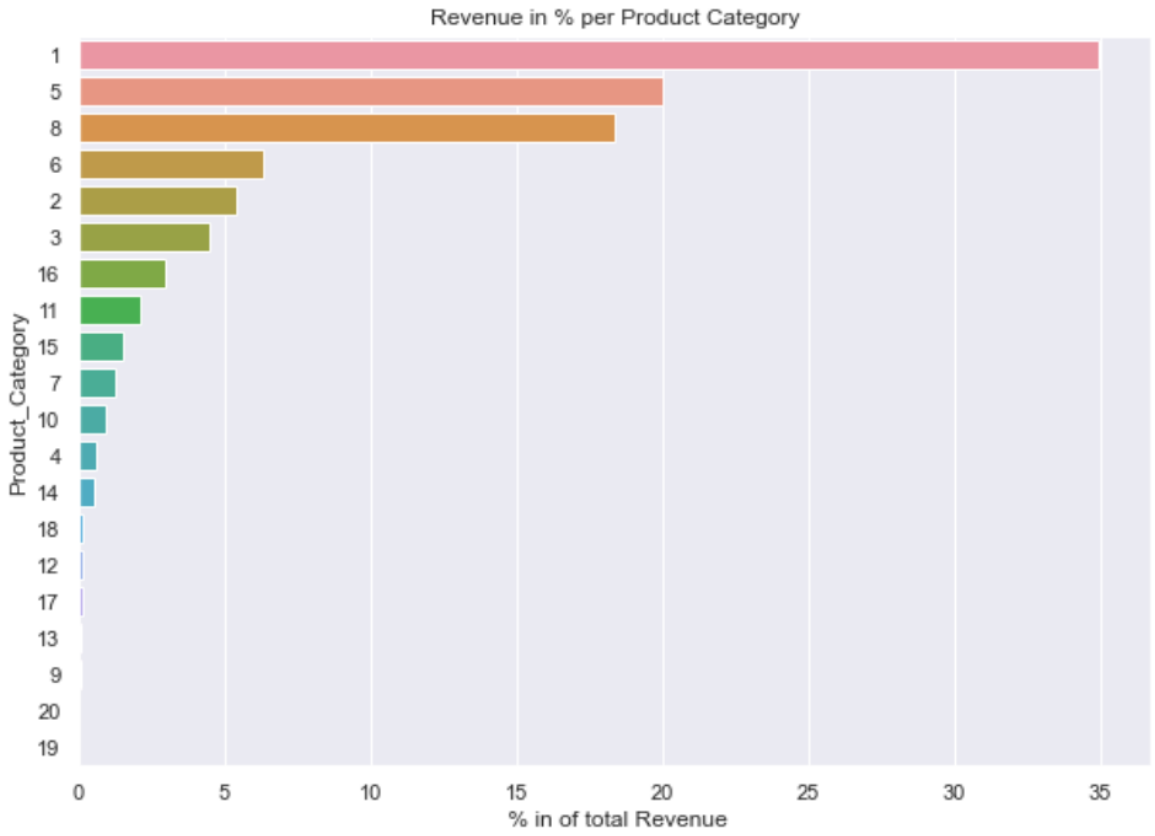
500 repetitions



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:

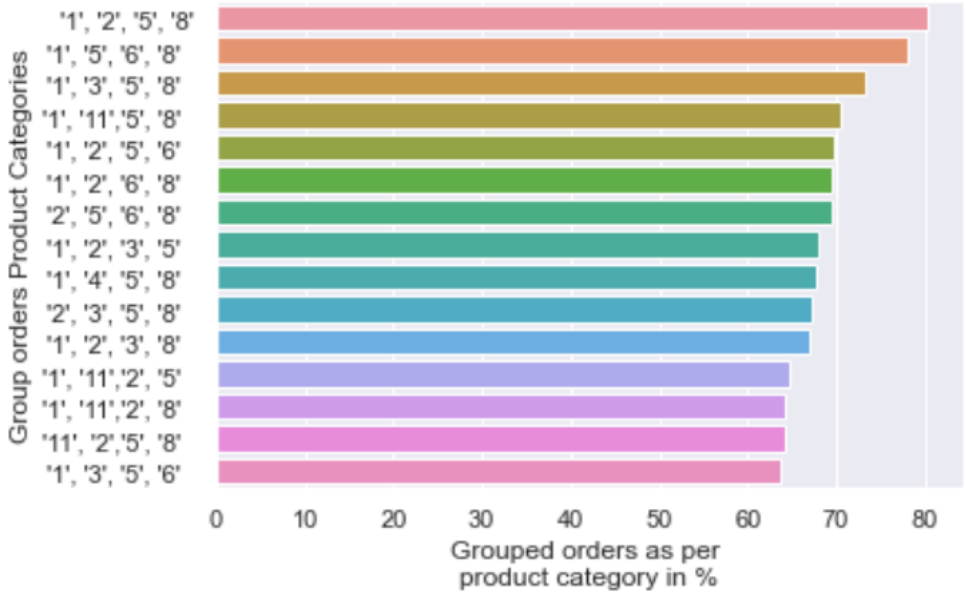
Top 3 Highest Revenue generating Product Categories :

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 :

Product Categories Sold Together	Number of orders
'1', '2', '5', '8'	217471
'1', '5', '6', '8'	211573
'1', '3', '5', '8'	198821
'1', '11','5', '8'	190851
'1', '2', '5', '6'	189101
'1', '2', '6', '8'	188376
'2', '5', '6', '8'	188322
'1', '2', '3', '5'	184308
'1', '4', '5', '8'	183249
'2', '3', '5', '8'	182173
'1', '2', '3', '8'	181778
'1', '11','2', '5'	175369
'1', '11','2', '8'	174108
'11', '2','5', '8'	173908
'1', '3', '5', '6'	172663

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



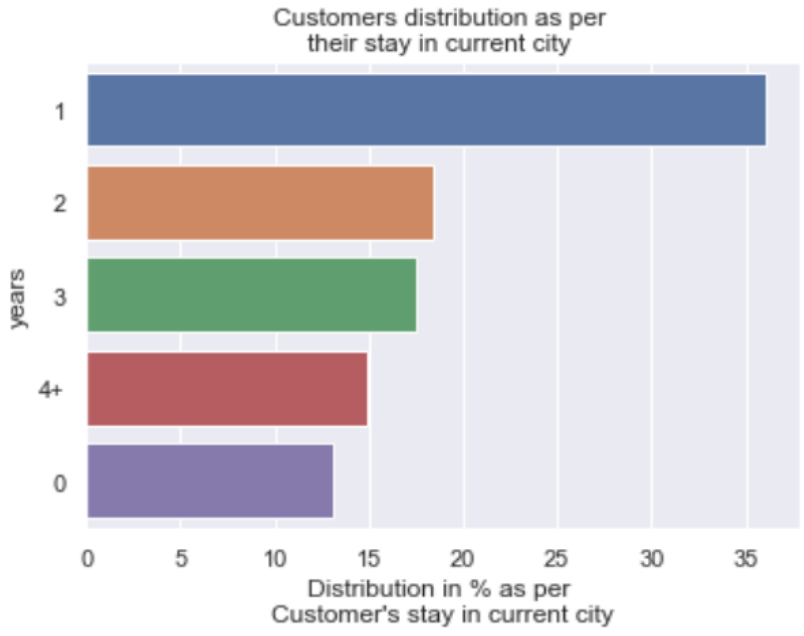
Above product categories group are most commonly sold together.

City Category :

Customers per City Category :

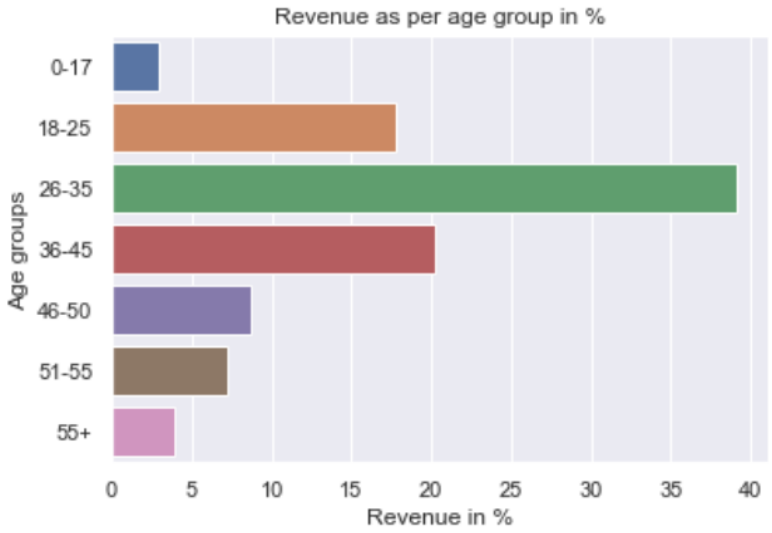
City Category B has highest Customers Base compared to C and A category.

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

Recommendations:

Observations :

- Majority of customers are between age 18-45 years .
- 75% male and 25% are female customers as per given sample data
- from the given data:

    75.31 % customers are male

24.68 % customers are female from the problem statement : company has 50 million customers are male and 50 million are female overall. Given Sample is a gender biased .

- (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 .
- All 100 million customer's average spending amount lies between 8931 to 9122. and sample mean is 9026.
- As per confidence Interval comparison for both Single and Married Customer's average purchase data
- There is not much difference between their average spending amounts. Married and Single Customer's spending amounts distribution are almost lies with same distribution.
- 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.
- 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.

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

Recommendations :

- City Category B has the highest customer base compared to C and A . Since City Category A and C customers, have the lesser spending average amount that city category B customers, more infrastructure and marketing strategies can be focused on City category A.
- There is not much significant difference between Married and Single Category Customers, no changes needs to be taken in that area.
- And there is a huge gap and difference between Male and Female spending average amounts and intervals, We can introduce special offers for particularly women like Women's day offer , or mother special or something like that.
- Age group 0-25 has the lowest spending compared to other age groups. Since most of the 0-25 age customers would be students , more products related students / teenage / kids recommended to introduce and university/student discount can help increase the revenue from this age group.