**Walmart Case Study – CLT and CI**

**Jypyter notebook pdf link:**

https://drive.google.com/file/d/1Bfo58uTu5FW7D1zYzaM\_AekgAxbERShj/view?usp=sharing

1. Import the dataset and do usual data analysis steps like checking the structure & characteristics of the dataset

**Basic info on the dataset**

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)

**Name of the columns:**

['User\_ID', 'Product\_ID', 'Gender', 'Age', 'Occupation', 'City\_Category',

'Stay\_In\_Current\_City\_Years', 'Marital\_Status', 'Product\_Category',

'Purchase']

**Dimension of the Data:**

Number of rows: 550,068

Number of columns: 10

**Number of Unique values:**

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

**Null values in the dataset:**

User\_ID : 0

Product\_ID : 0

Gender : 0

Age : 0

Occupation : 0

City\_Category : 0

Stay\_In\_Current\_City\_Years : 0

Marital\_Status : 0

Product\_Category : 0

Purchase : 0

**Unique values in the all the columns:**

| **count** | **unique** | **top** | **freq** |
| --- | --- | --- | --- |
| **User\_ID** | 550068 | 5891 | 1001680 | 1026 |
| **Product\_ID** | 550068 | 3631 | P00265242 | 1880 |
| **Gender** | 550068 | 2 | M | 414259 |
| **Age** | 550068 | 7 | 26-35 | 219587 |
| **City\_Category** | 550068 | 3 | B | 231173 |
| **Stay\_In\_Current\_City\_Years** | 550068 | 5 | 1 | 193821 |
| **Marital\_Status** | 550068 | 2 | 0 | 324731 |

**Unique user based on the Gender:**

Gender

F 1666

M 4225

In Percentage

F 28.280428

M 71.719572

**Purchase made based on the Gender:**

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |  |  |  |  |
| **F** | 135809.0 | 8734.565765 | 4767.233289 | 12.0 | 5433.0 | 7914.0 | 11400.0 | 23959.0 |
| **M** | 414259.0 | 9437.526040 | 5092.186210 | 12.0 | 5863.0 | 8098.0 | 12454.0 | 23961.0 |

**Percentage of User based on the age rage:**

0-17 3.700560

18-25 18.146325

26-35 34.849771

36-45 19.809879

46-50 9.013750

51-55 8.164997

55+ 6.314717

**Percentage of User Married/Unmarried:**

Marital\_Status

0 58.003735

1 41.996265

**Number of people from different cities:**

A 1045

B 1707

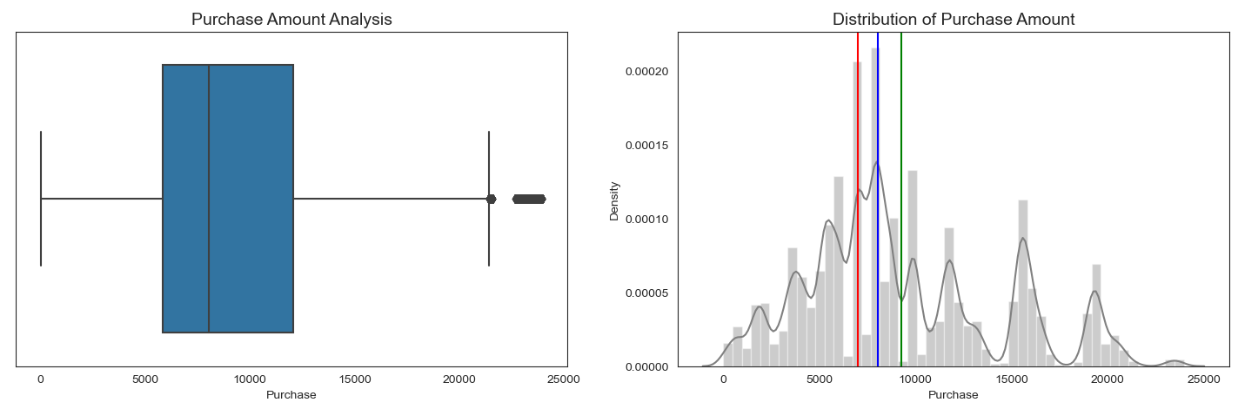
C 3139

2. Detect Null values and outliers

**Finding the outliers:**

The columns 'User\_ID','Product\_ID','Gender', 'Age','City\_Category','Marital\_Status' were converted to the Category type.

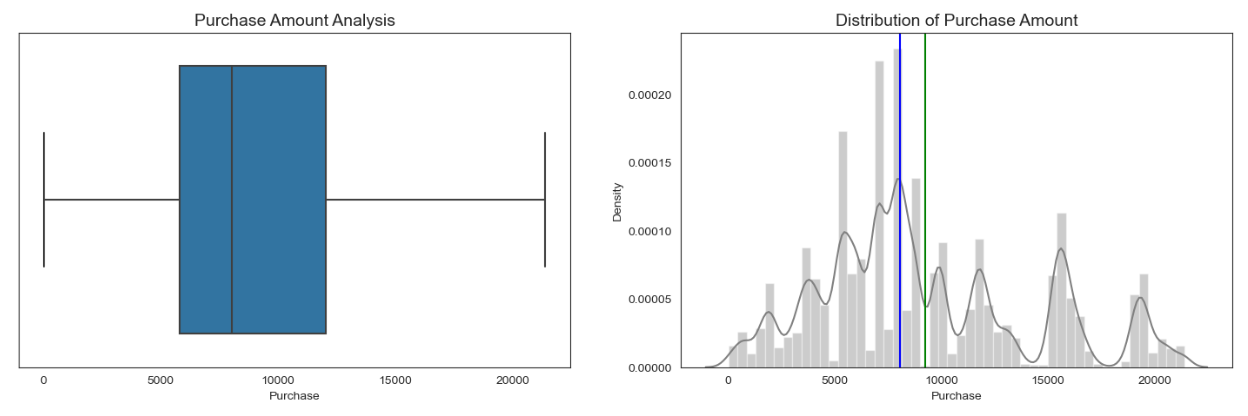
And the based on the graph the Purchase having the outliers

****

Mean of Purchase Amount = 9263.968712959126

Median of Purchase Amount = 8047.0

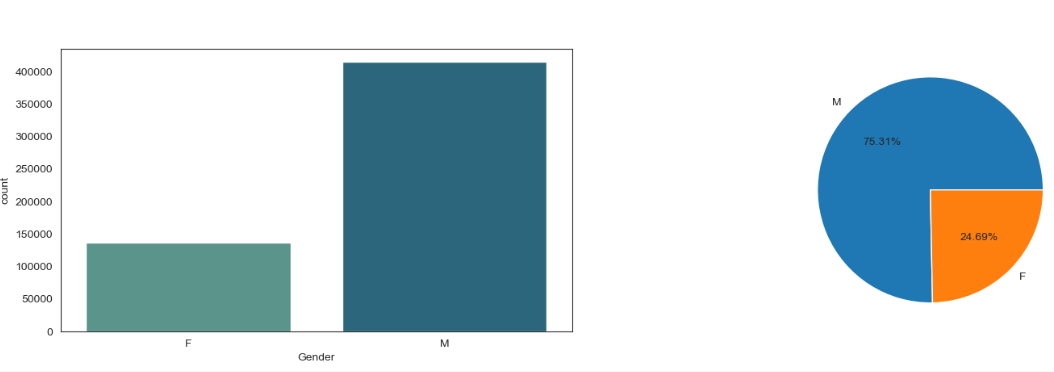
**After removing the outliers:**

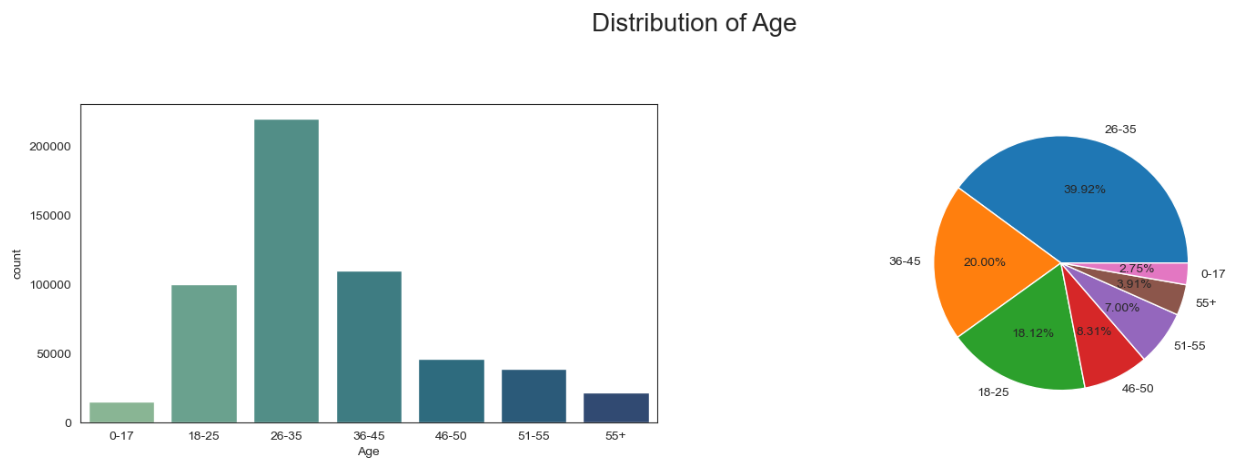
****

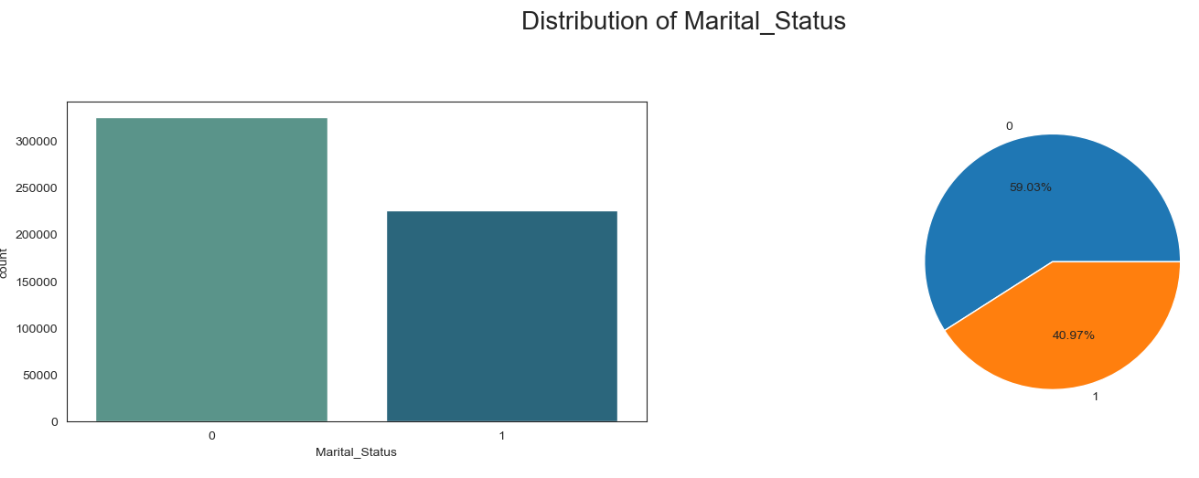
Mean of Purchase Amount = 9255.02429608703

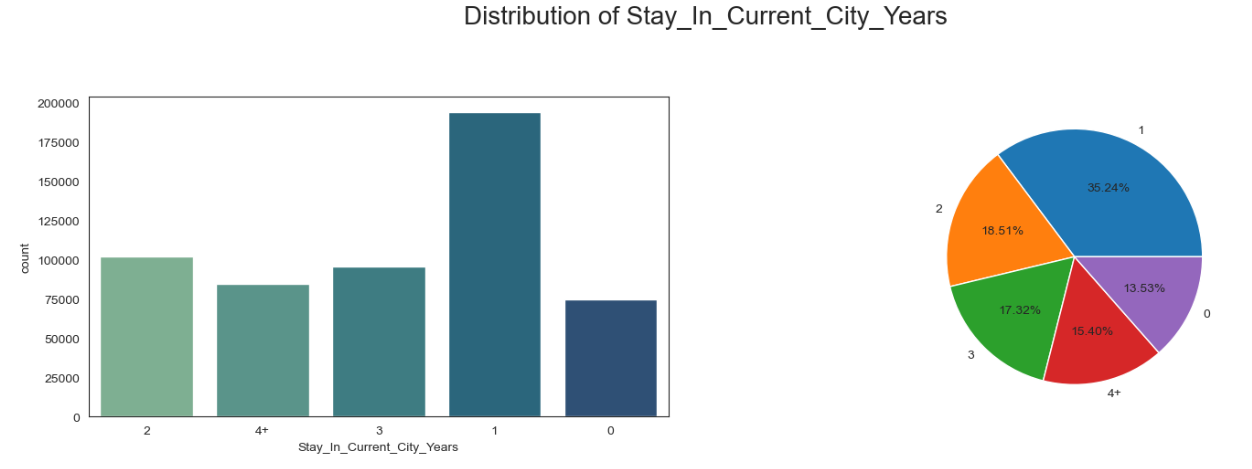
Median of Purchase Amount = 8047.0

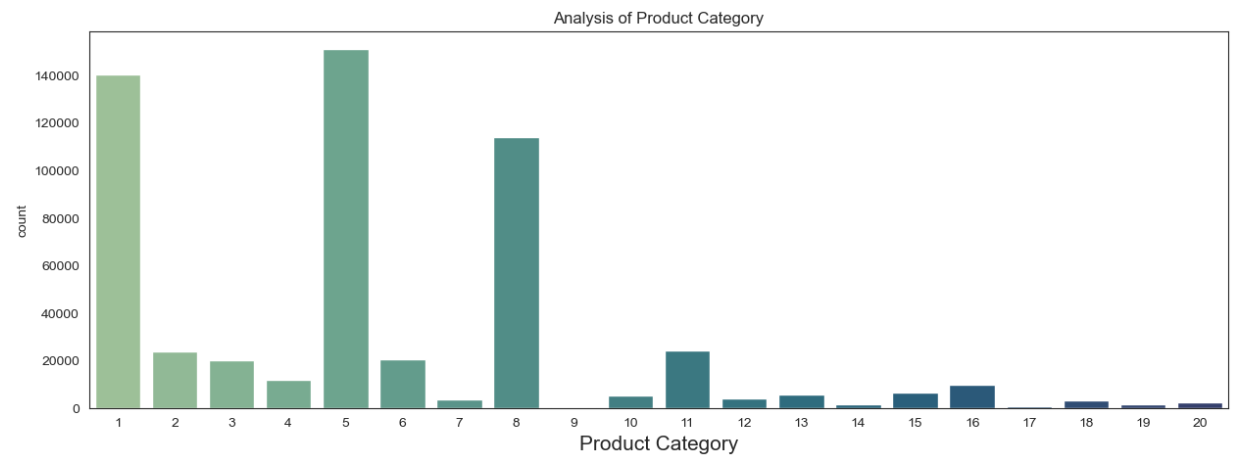
**Univariate Analysis:**

****

****

****

****

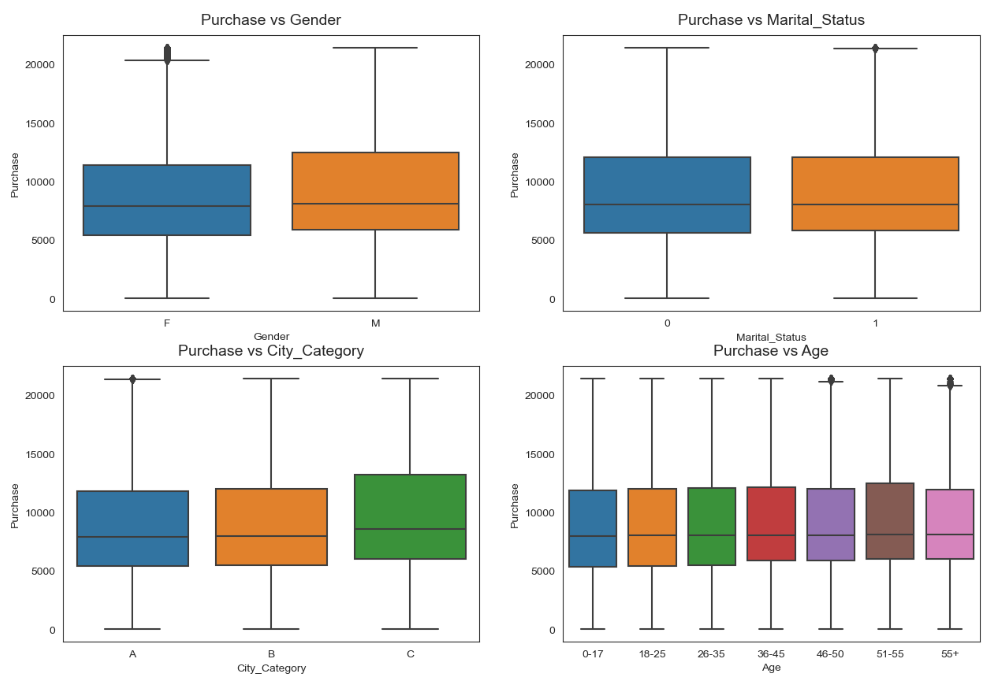
****

**Inference:**

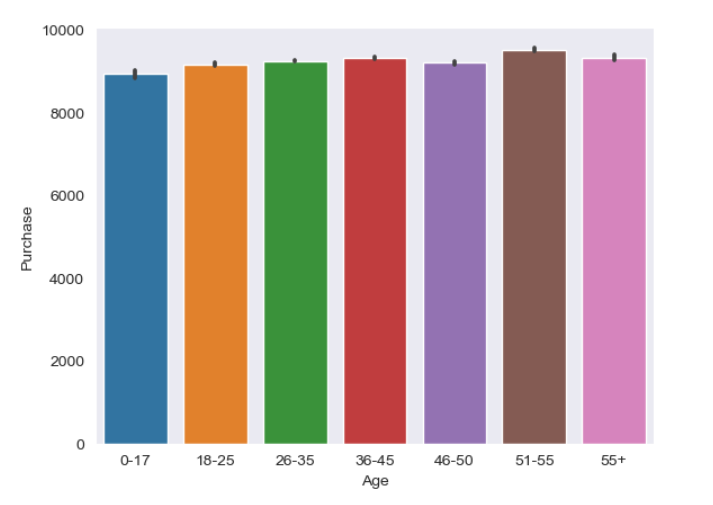
Inference based on the above univariate analysis are as follows:

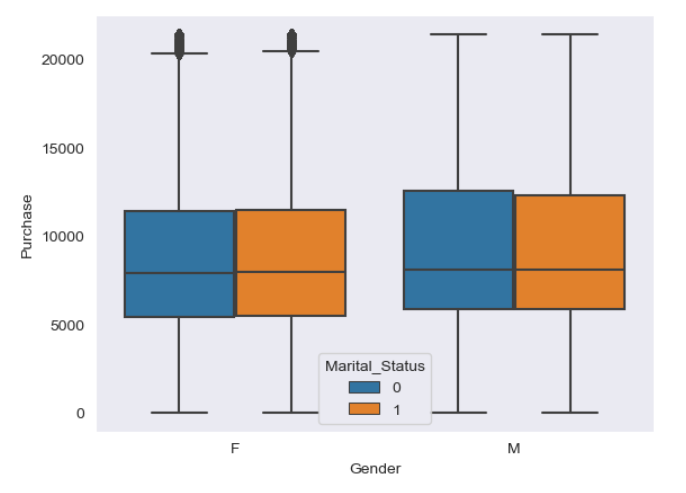
* Compared to female, Male buying the products more.
* The people in age group of 26-35 tends to do more purchase.
* Based on the graph Unmarried people purchasing more
* The people belong to city Category B purchasing more.
* The product category 5 is more selling.

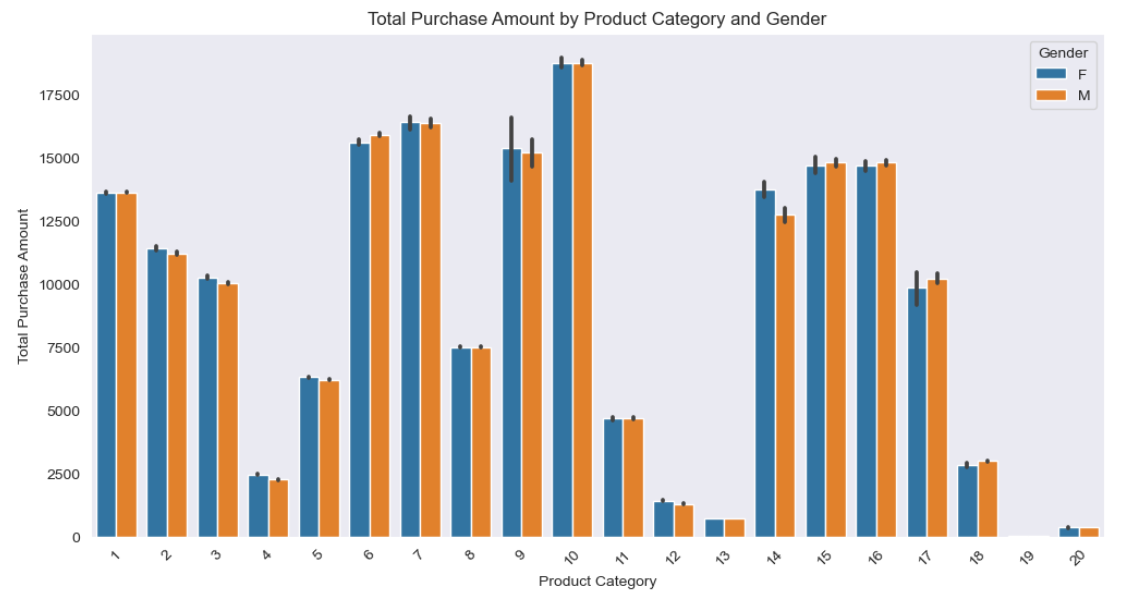
**Bi-variate Analysis:**

****

**Exploratory data analysis:**

****

****

****

**Inferences:**

**Based on the EDA:**

* The majority of our customers come from city category B but customers come from City category C spent more as mean is 9719.
* Males clearly purchase more than females. 75% of men and only 25% of women purchase products.
* Most customers are between the ages of 26 and 35.60% of purchases are made by people between the ages of 26 and 45.
* City Category B accounts for 42%, City Category C 31%, and City Category A represents 27% of all customer purchases.
* In City Category C, there are slightly more female customers.
* 58:42 is the ratio between married and unmarried customers

**Based on the Statistical Analysis (CLT and CI):**

* Using confidence interval 95%, the inference of the mean purchase value by gender
* As the sample size increases, the Male and female groups start to become distinct
* With increasing sample size, Standard error of the mean in the samples decreases.
* For Female (sample size 10000) range for mean purchase with confidence interval 95% is [8632.86, 8819.36]
* For Male range for mean purchase with confidence interval 95% is [9330.27, 9528.62]
* Overlappings are increasing with a confidence interval of 95%. Due to the increasing CI, we consider higher ranges within which the actual population might fall, so that both mean purchase are more likely to fall within the same range.
* Overlapping is evident for married vs un-married customer spend even when more samples are analyzed, which indicates that customers spend the same regardless of whether they are single or married.
* For Unmarried customer (sample size 100000) range for mean purchase with confidence interval 95% is [9225.25, 9289.03]
* For married customer range for mean purchase with confidence interval 95% is [9221.11, 9281.55]
* Spending by Age\_group 0-17 is low compared to other age groups.
* Customers in Age\_group 51-55 spend the most between 9381.9 and 9463.7

**Recommendations:**

* As per the analysis females spend less than males on average, management needs to focus on their specific needs differently.
* In order to attract more young shoppers, they can offer some discounts, video games subscription etc to lure more younger generation.
* There are more married customers than single. But as per analysis single customers mean purchase is higher than married one. Management needs to focus more on married customers as they are large in numbers but purchasing less.
* As product category 1 & 5 are popular among both male and female, in order to increase sales management need to give additional offers and introduce some strategry regarding cross-sale or upsell.
* Management need to find out the reason why proudct category 9 & 17 are less bought. If customers don't like this product that needs to be removed from inventory to reduce the cost.