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.

The Management team at Walmart Inc. wants to analyze the customer purchase behavior (specifically, purchase amount) against the customer's gender and the various other factors to help the business make better decisions. They want to understand if the spending habits differ between male and female customers: Do women spend more on Black Friday than men? (Assume 50 million customers are male and 50 million are female).

The given data set has 180 rows and 9 columns. The following are ten columns: User_id, Product_id, Age, Gender, Occupation, City Category, Stay_in_current _city_year ,Marital_Status, Product_Category, Purchase .Each columns provides which different inputs of the customers like Product column is to show type of product chosen. Age column specifies the different ages of the customers. Gender of the customers is shown in Gender column. Stay_in_current _city_year columns shows the years stayed in the particular city. Marital status column shows whether is partnered or single. Product category columns how frequently the product is purchased. Purchase column the purchase pattern of the customers. City Category columns different cities of the customers.

The data type of the given file is:

#	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	object
5	City_Category	550068 non-null	object
6	Stay_In_Current_City_Years	550068 non-null	object
7	Marital_Status	550068 non-null	object
8	Product_Category	550068 non-null	object
9	Purchase	550068 non-null	int64

Table 1: Data type of the file

From Table 1, we can see that columns which are object are object type, including product, Gend er,Age,Occupation,City_Category,Stay_In_current_City_Years ,Product_Category and Martial s tatus. The User_id, Pruchase are in integer type. The file has zero null values.

1. Initial analysis data

There are no missing values in the data. There are 3631 unique product IDs in the dataset. P00265242 is the most sold Product ID. 7 unique age groups are present and most of the

purchase belongs to age 26-35 group. There are 3 unique cities categories with category B being the highest.5 unique values for Stay_in_current_citi_years with 1 being the highest. The difference between mean and median seems to be significant for purchase that suggests outliers in the data. Minimum & Maximum purchase is 12 and 23961 suggests the purchasing behavior is quite spread over a significant range of values. Mean is 9264 and 75% of purchase is of less than or equal to 12054. It suggest most of the purchase is not more than 12k. Out of 550068 data points, 414259's gender is Male and rest are the female. Male purchase count is much higher than female. Standard deviation for purchase have significant value which suggests data is more spread out for this attribute. Few numerical data is in object type so converted into integer type.

Observation post modifying the categorical variable's data type:

1. There are 5891 unique users, and user_id 1001680 being with the highest count.
2. The customers belongs to 21 distinct occupation for the purchases being made with Occupation 4 being the highest.
3. Marital status unmarried contribute more in terms of the count for the purchase.
4. There are 20 unique product categories with 5 being the highest.

Observations by doing basic data exploration using contingency tab

1. Age 26-35 do 40% of the purchase and 78% purchase are done by the customers aged between the age 18-45 (40%: 26-35, 18%: 18-25, 20%: 36-45).

2. 75% of the purchase count are done by Male and 25% by Female

3. 60% Single, 40% Married contributes to the purchase count.

4. 35% Staying in the city from 1 year, 18% from 2 years, 17% from 3 years

5. There are 20 product categories in total.

6. There are 20 different types of occupations in the city.

We can observe from the data that 35% of the users are aged 26-35. 73% of users are aged b etween (18-45). Using contingency tad for data exploration we observed 40% of the purcha se are done by users aged 26-35. And, we have 35% of users aged between (26-35) and the y are contributing 40% of total purchase count. So, we can infer users aged (26-35) are mor e frequent customers.

We have 72% male users and 28% female users. Combining with previous observations we can see 72% of male users contributing to 75% of the purchase count and 28% of female us ers are contributing to 25% of the purchase count.

We have 58% of the single users and 42% of married users. With observation, single users contributes more as 58% of the single contributes to the 60% of the purchase count.

53% of the users belong to city category C whereas 29% to category B and 18% belong to c ategory A. Combining from the previous observation category B purchase count is 42% and Category C purchase count is 31%. We can clearly see category B are more actively purcha sing in spite of the fact they are only 28% of the total users. On the other hand, we have 53 % of category C users but they only contribute 31% of the total purchase count.

We have seen earlier that city category B and A constitutes less percentage of total populati on, but they contribute more towards purchase count. We can see from above results large p ercentage of customers aged 26-35 for B (40%) and A (50%) which can be the reason for th ese city categories to be more actively purchasing.

We can observe male (72% of the population) contributes to more than 76% of the total pur chase amount whereas female (28% of the population) contributes 23% of the total purchase amount.

Single users are contributing 59% towards the total purchase amount in comparison to 41% married users. From the Occupation data we observe 0, 4, 7 has contributed more towards t otal purchase amount. The data from the product category says 1, 8, 5 are among the highe st yielding product categories and 19, 20, 13 are among the lowest in terms of their contribut ion to total amount.

1.1 **Using univariate analysis for quantitative variables**

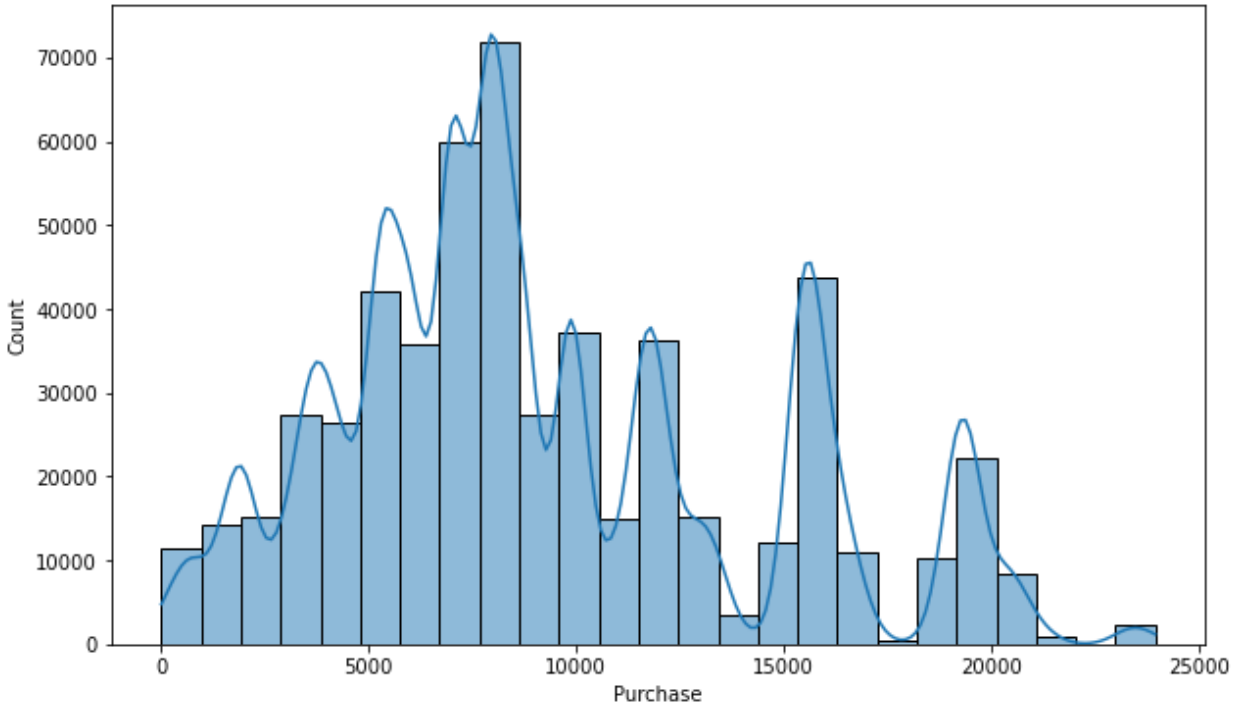

Figure 1: Count plot for Purchase

We can observe the value between 5000 and 10000 have higher count. From the initial observation we have already seen the mean and median is 9263 and 8047 respectively. Also, we can see there are outliers in the data.

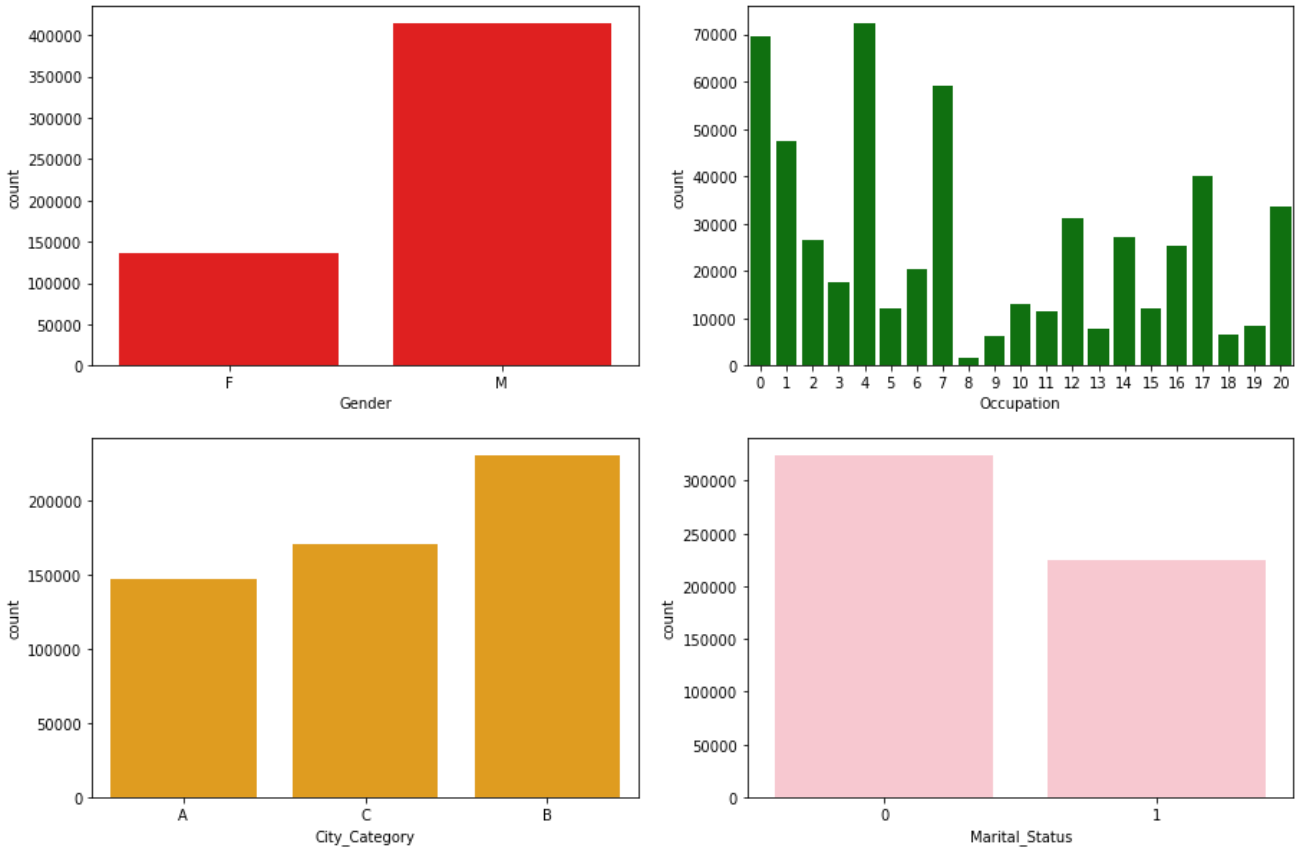

Figure 2: count plot for quantitative data

We can clearly observe from the graphs of figure 2 the purchases done by males are m uch higher than females.

We have 21 occupations categories. Occupation category 4, 0, and 7 are with higher nu mber of purchases and category 8 with the lowest number of purchase.

The purchases are highest from City category B.

Single customer purchases are higher than married users.

There are 20 product categories with product category 1, 5 and 8 having higher purcha sing frequency.

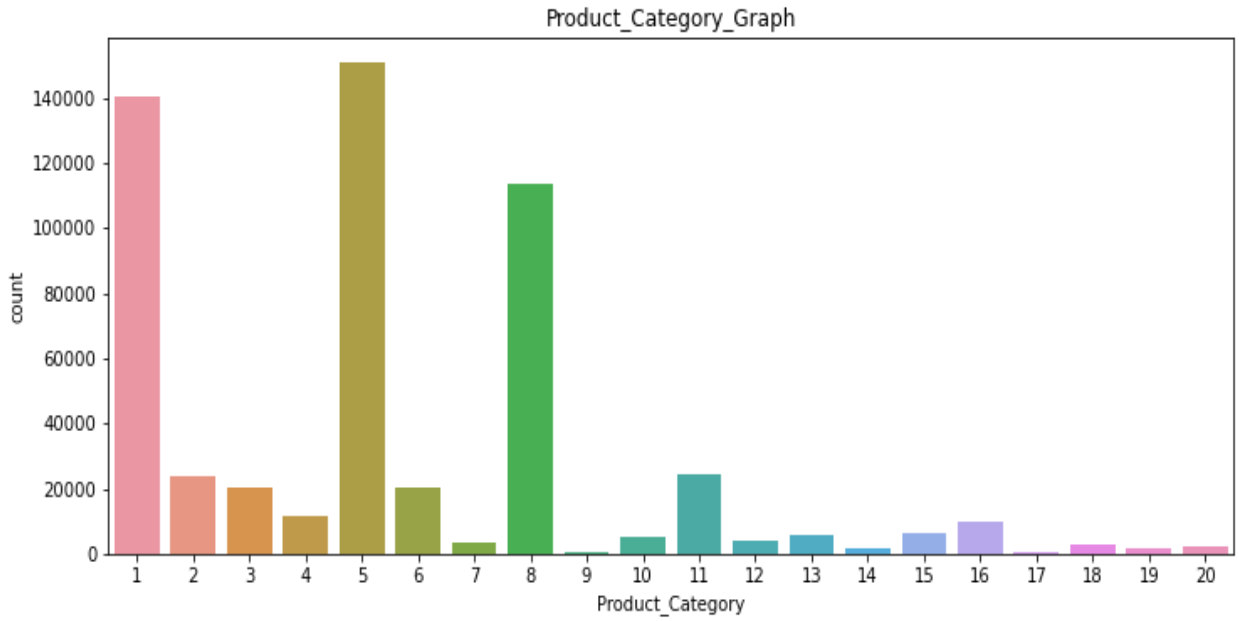

Figure 3: count plot of Product_Category

2. Analyzing purchase behavior according to the data

2.1 Gender V/S Purchase

From the Figure 4 and 5 his plot, we can clearly see spending behavior is very m uch similar in nature for both males and females as the maximum purchase count are between the purchase value range of 5000-10000 for both. But, the purchase count are more in case of males.

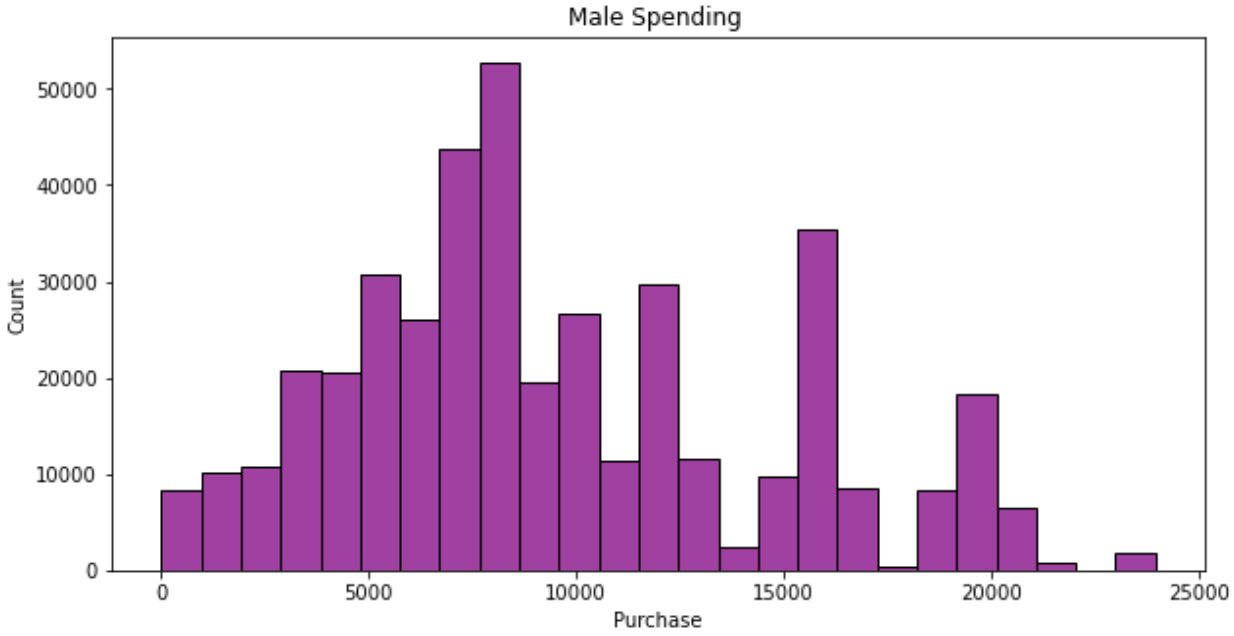

Figure 4: Male Spending

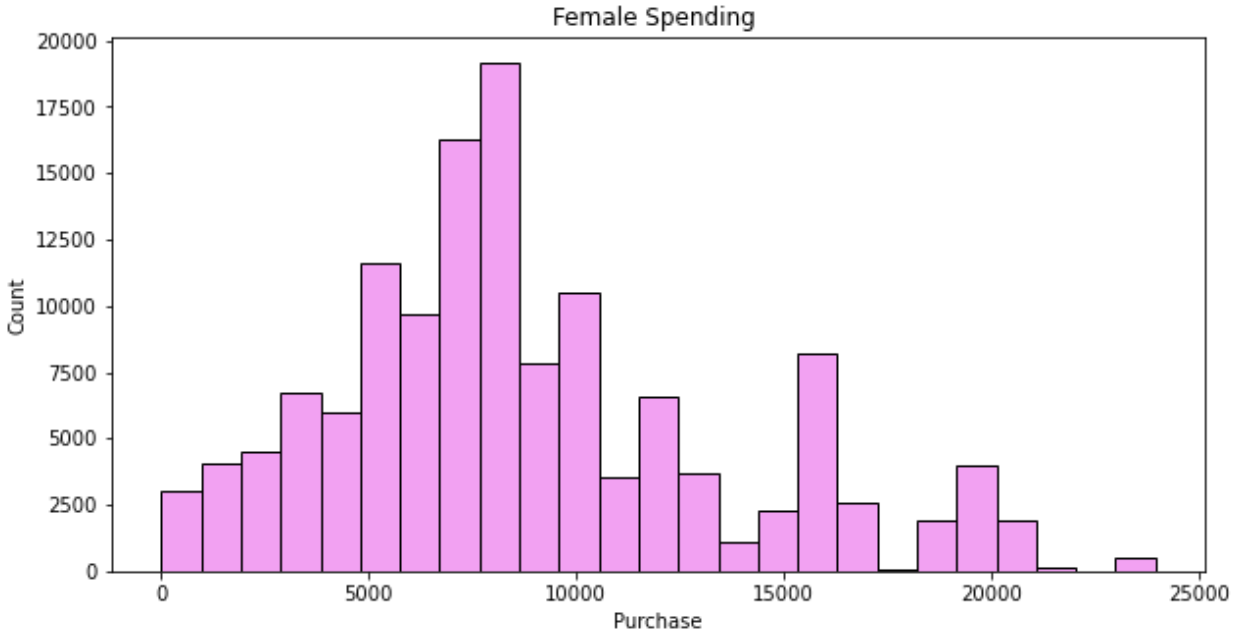

Figure 5: Female spending

2.2 Occupation V/S Purchase

Figure 6: Boxplot of occupation and purchase

Among different occupation as well, we see similar purchasing behaviour in terms o f the purchase values.

2.3 City_Category V/S Purchase

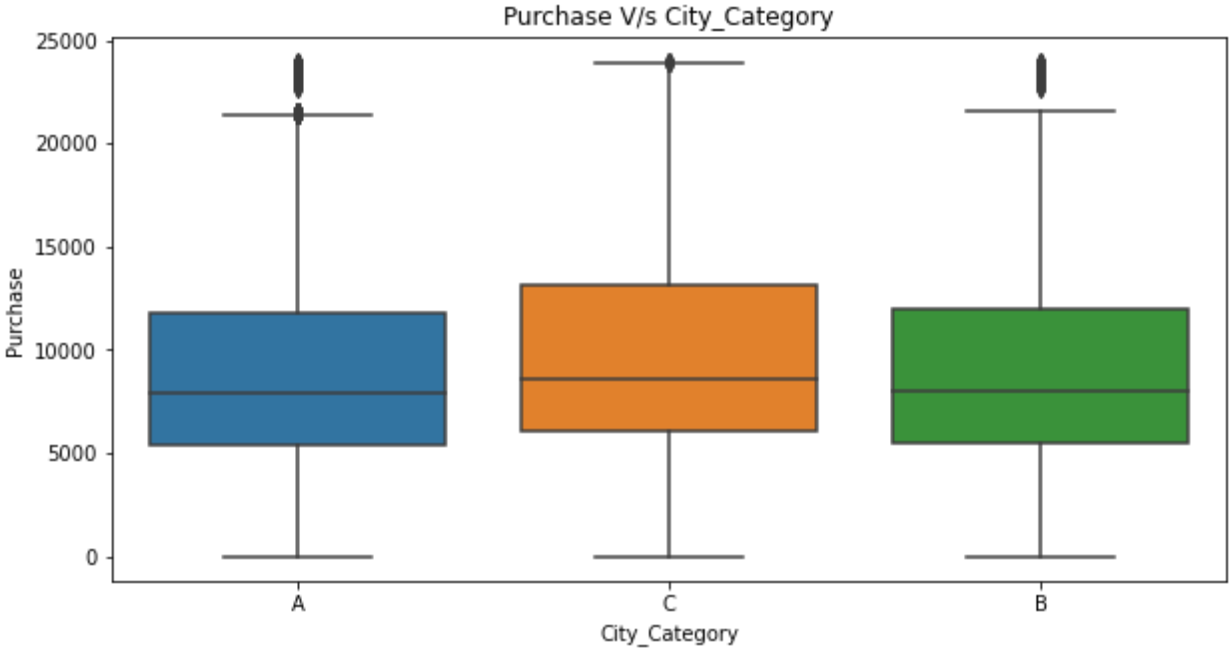

Figure 7: Boxplot of City_Category and purchase

The purchase patterns of the all the cities is almost similar by observe the median and th e spending also is mostly in between 5k-10k. They are some outliners which are negotia ble.

2.4 City_Category V/S Purchase

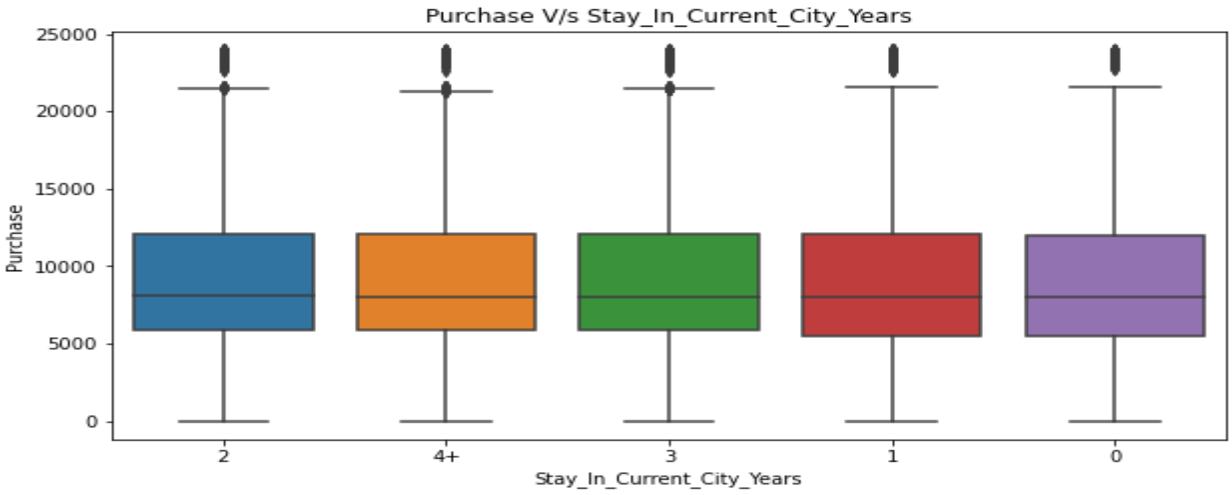

Figure 8: Boxplot of stay_in_current_city_years and purchase

The purchase patterns is almost similar by observe the median and the spending also is mostly in between 5k-10k. They are some outliners which are negotiable.

3 Analyzing the data by performing Multivariate analysis

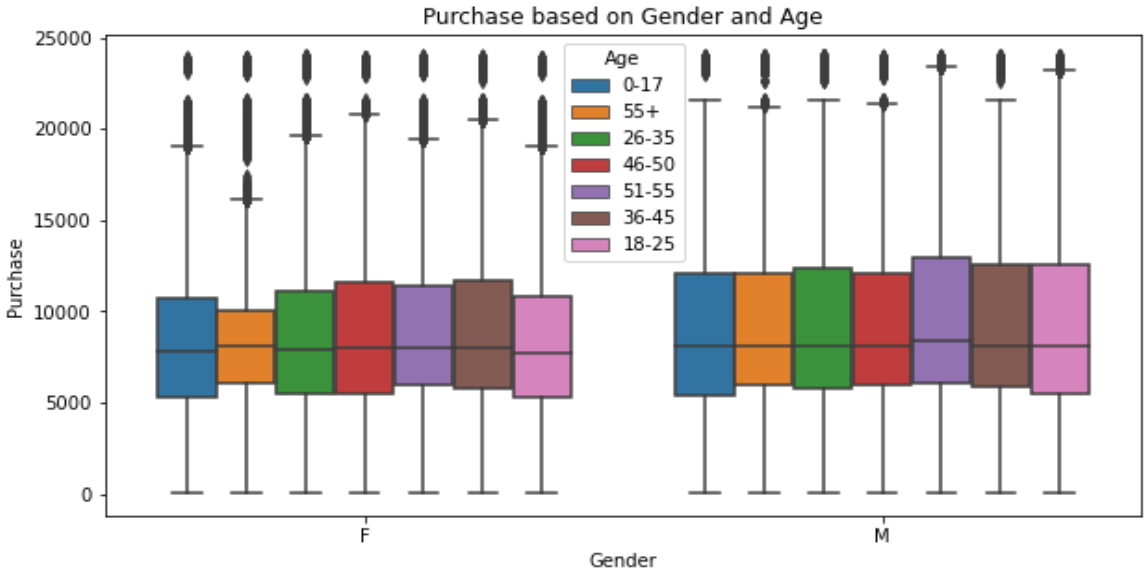

Figure 9

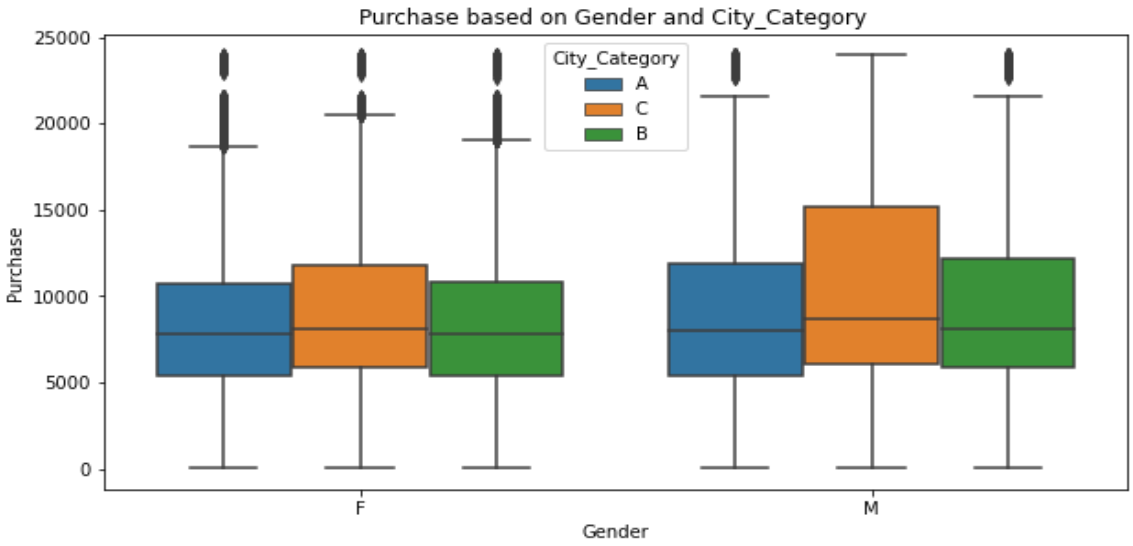

Figure 10

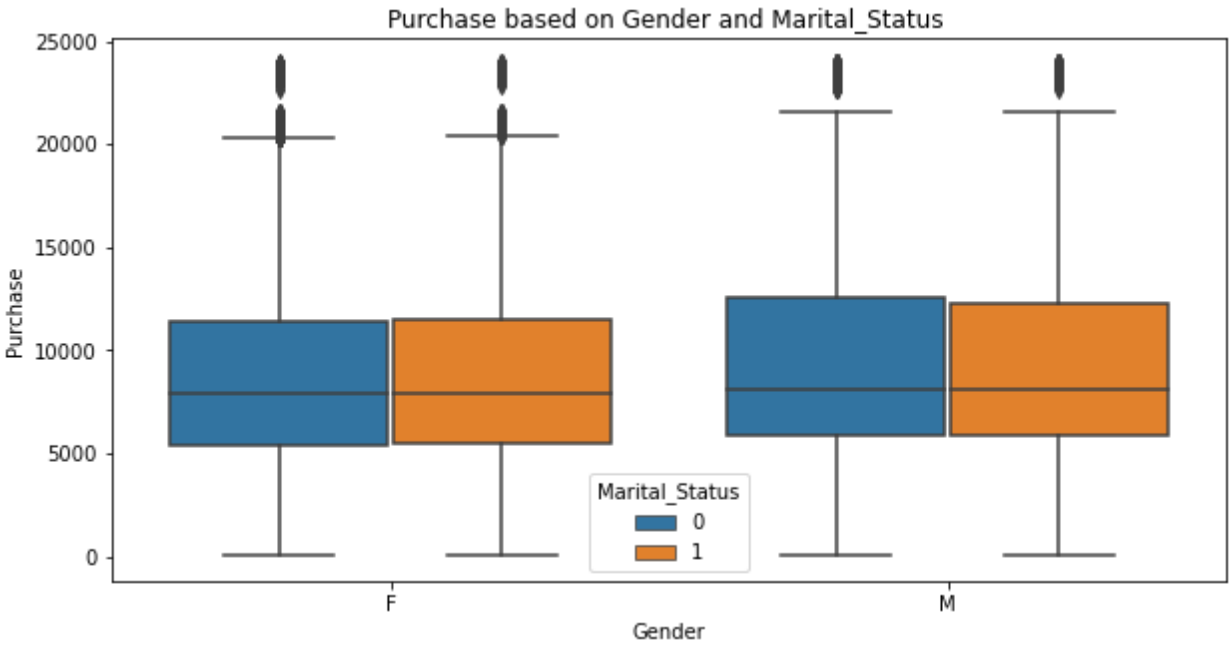

Figure 11

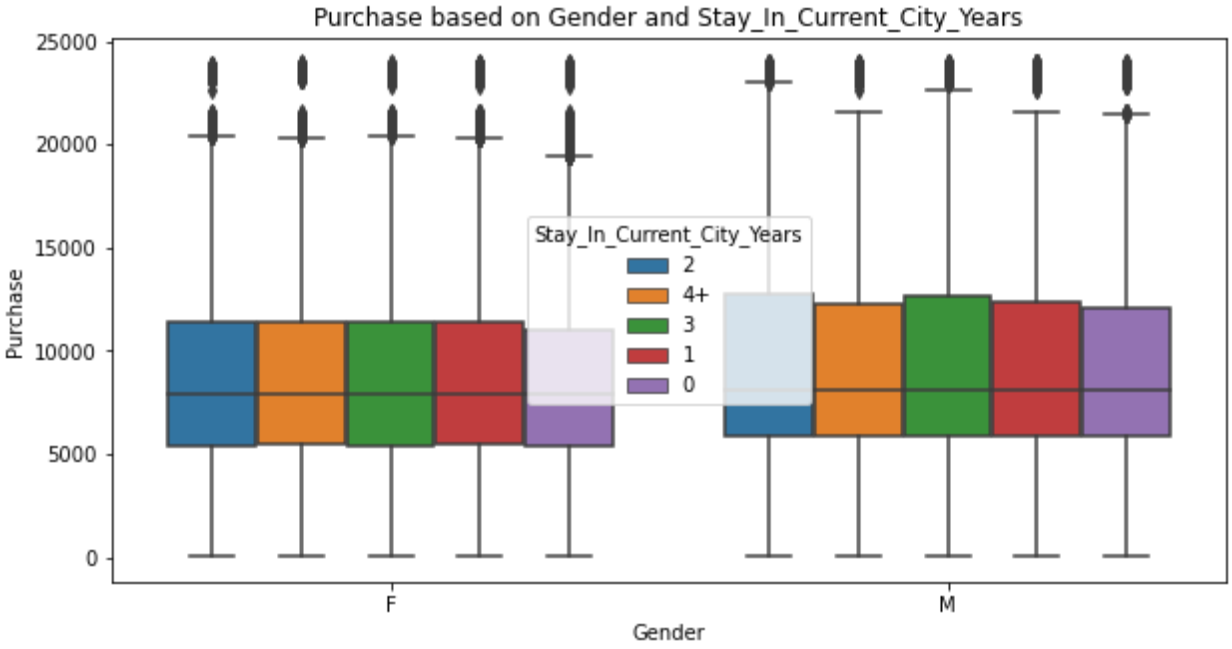

Figure 12

The purchasing pattern is very much similar for males and females even among differe nt age groups.

The purchasing behavior of males and females basis different cities categories is also si milar in nature. Still, males from city category B tends to purchase costlier products in comparison to females.

Males and females spending behavior remains similar even when take into account thei r marital status.

Purchase values are similar for males and females basis Stay_in_current_city_years. Al though, Males buy slightly high value product..

3.1 Average money spend by per customer of male and female

There are total 4225 males and 1666 female. Average amount for the males is 925344 for the entire population whereas it's much lesser for females (712024).

Total amount spend by males is around 4 billion whereas for females it's 1.2 billion.

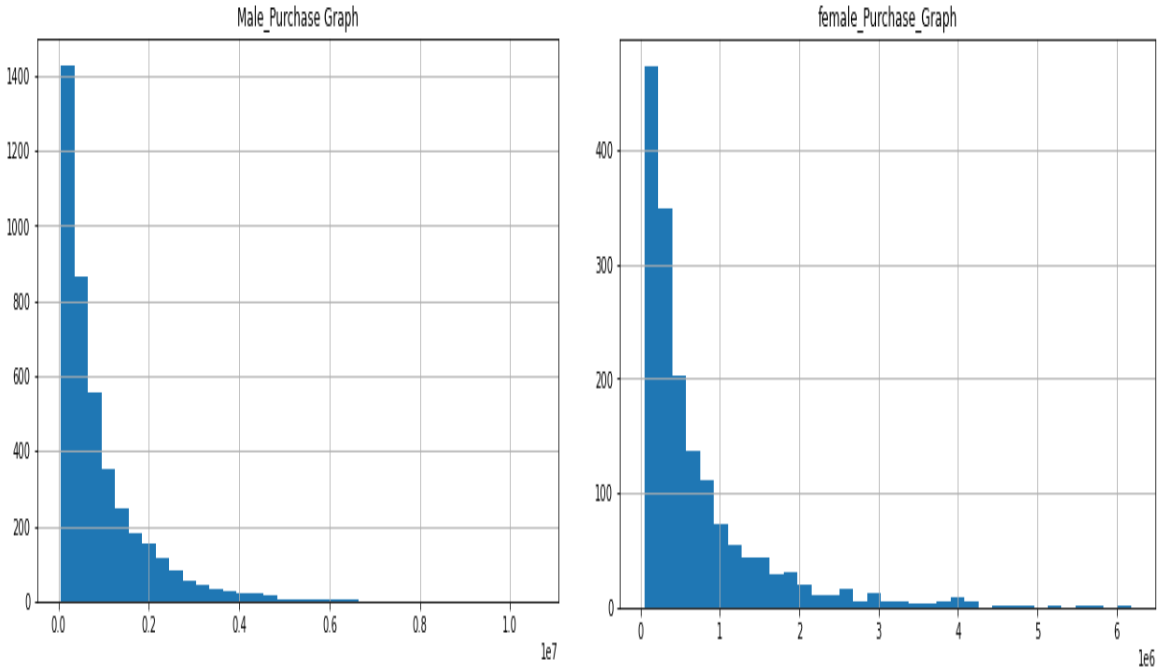

Figure 13: Male and Female purchase graph

5. Sampling the data by using CTL

5.1 For Gender and Purchase

From figure 14 we can observe the both the means of the male and female gender

are very closer and the outliners are negotiable . So from this figure we can't conclu ded the purchase patterns. So now we going to perform CLT to analyse the data mor e precisely

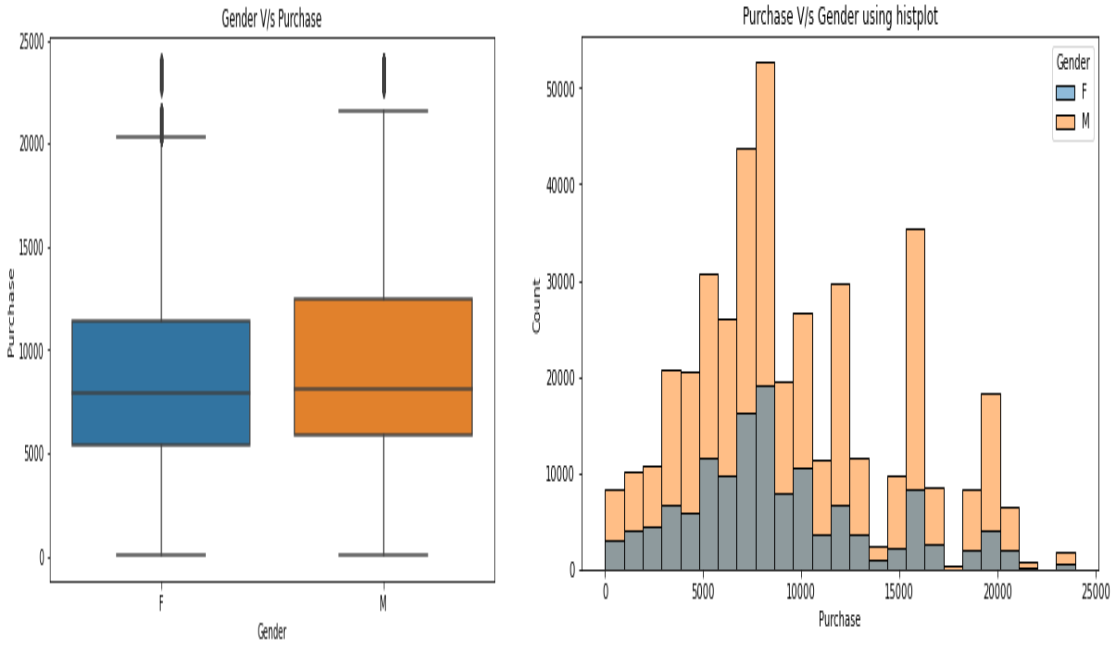

Figure 14. Observing the purchase pattern of the Gender using his plot and boxplot

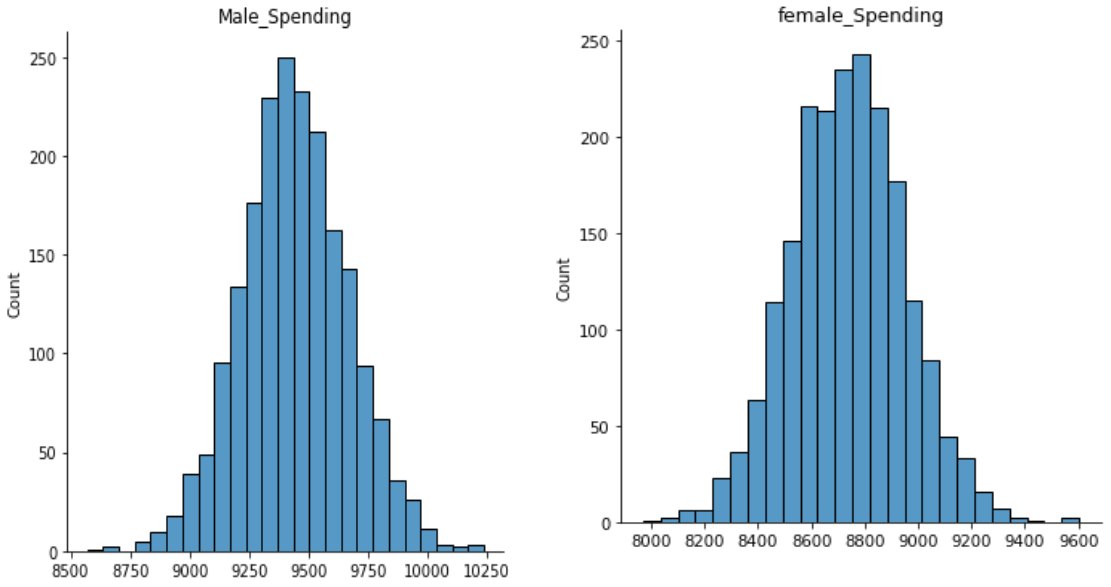

Figure 15: Male and Female spending using CLT

The sample of 500 and iterations of 2000 are taken the results been observed from the f igure 15 is both male and females are almost similar purchase pattern and show normal distribution. Also, we can see the mean of the sample means are closer to the populatio n mean as per central limit theorem.

5.2 Marital Status and Purchase

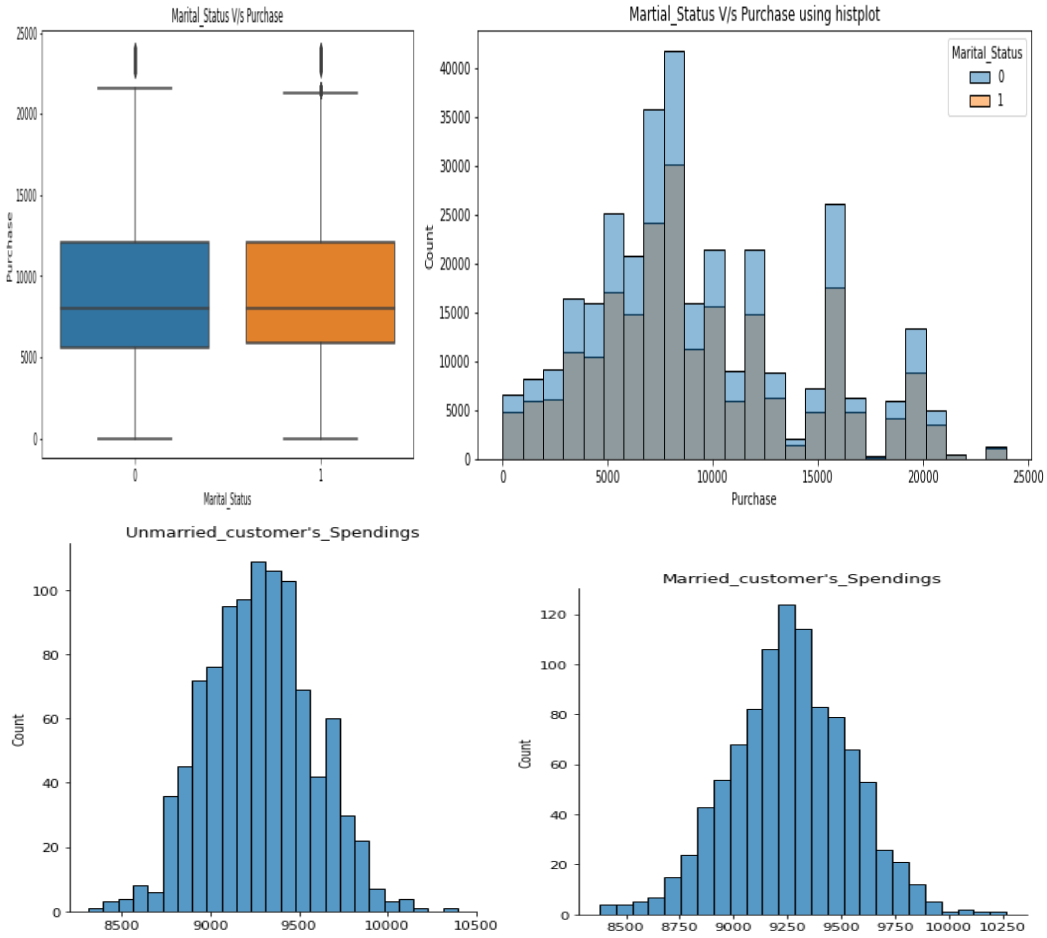

Figure 16: Unmarried and married spending using CLT

From figure 15 we can observe in the both boxplot and his plot the means of the sing le and married customer are very closer and the outliners are negotiable. So from tho se plots we can't concluded the purchase patterns. So now we going to perform CLT to analyze the data more precisely

The sample of 300 and iterations of 1000 are taken the results been observed from th e figure 16 is both unmarried and married are almost similar purchase pattern and sh ow normal distribution. Also, we can see the mean of the sample means are closer to the population mean as per central limit theorem.

5.3. Age and Purchase

From Figure 17 we can observe the means sample seems to be normally distributed for all age groups. Also, we can see the mean of the sample means are closer to the population mean as per central limit theorem.

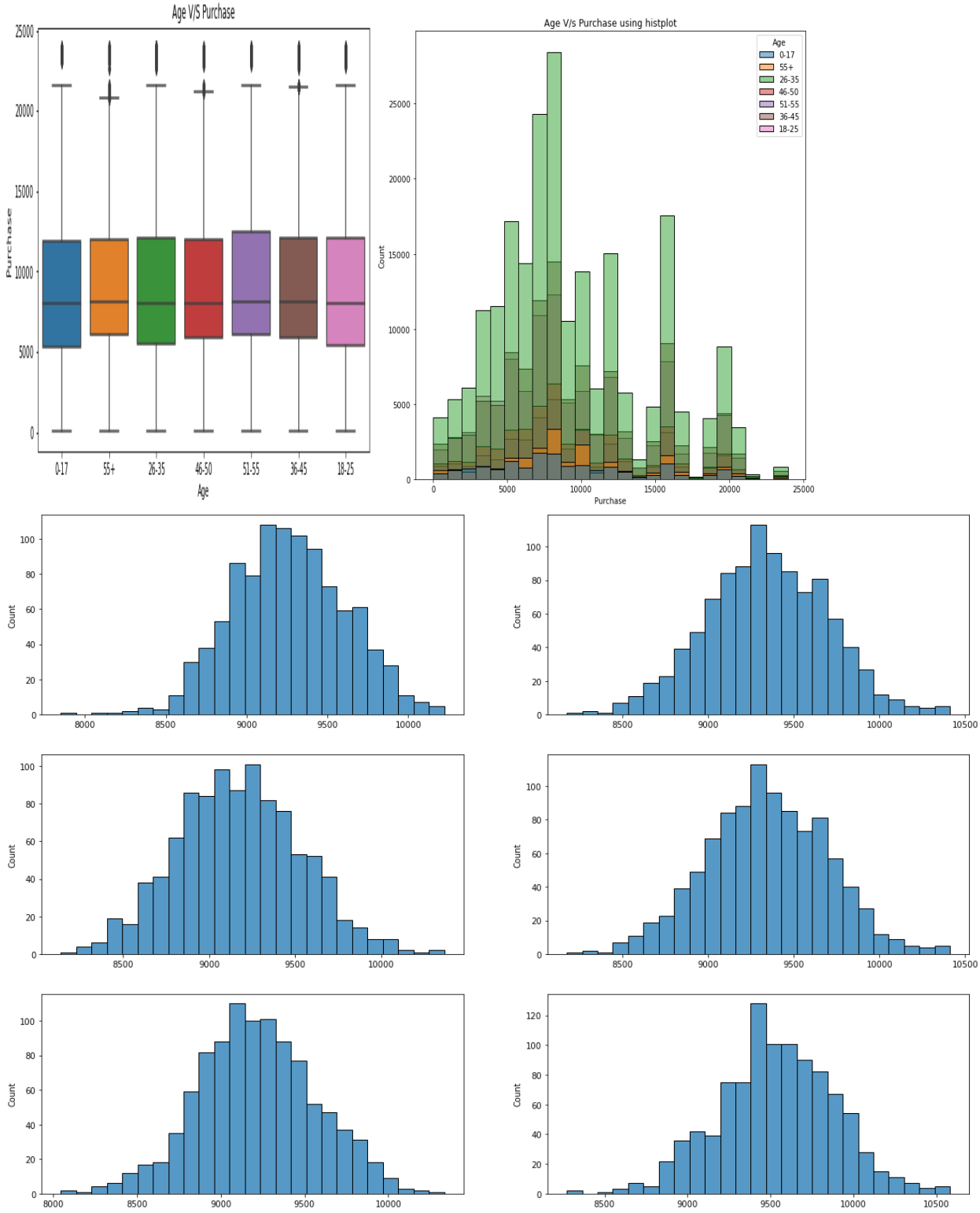

Figure 17: **Observing different age groups spending spending using CLT**

6. Sampling the data using Confidence Interval

6.1. Gender V/S Purchase

Now using the Confidence interval at 95%, by using z score method we can say that:

Average amount spend by male customers lie in the range 8991.606 – 9882.60

Average amount spend by female customers lie in range 8318.825 – 9147.63

The purchase ranges of male and female are almost overlapping.

6.2. Marital Status V/S Purchase

Now using the Confidence interval at 90%, by using z score method we can say that:

Average amount spend by unmarried customers lie in the range 8898.50 – 9623.94

Average amount spend by married customers lie in range 8790.23 – 9748.80

The purchase ranges of unmarried and married are almost same.

6.3 Age v/s Purchase

Now using the Confidence interval at 90%, by using z score method we can say that:
Average amount spend by age group [0-17] lie in the range 8336.43- 9495.89
Average amount spend by age group [18-25] lie in the range 8585.55 - 9759.16
Average amount spend by age group [26-35] lie in the range 8683.86 9838.04
Average amount spend by age group [36-45] lie in the range 8757.90 9923.14
Average amount spend by age group [46-50] lie in the range 8634.01- 9795.17
Average amount spend by age group [51-55] lie in the range 8959.60 - 10112.89
Average amount spend by age group 55 + lie in the range 8746.52- 9916.89

7. Business Insights

- CI's of male and female do not overlap and upper limits of female purchase CI are le sser than lower limits of male purchase CI. This proves that men usually spend more than women (NOTE: as per data 77% contributions are from men and only 23% pur chases are from women).
- The reason for less purchase by women could have several factors
    - Males might be doing the purchase for females.
    - Salary can be a factor in less purchase.
    - If the female based products quality/quantity needs to be improved for wome n purchasing.
- Confidence intervals of average male and female spending are not overlapping but t hey are very slow. The male CI limits are higher than female's. This trend can be ch anged via introducing female centric marketing strategies by Walmart so that more f emale customers are attracted to increase female purchases to achieve comparable st atistics close to 50%.

- The results when the CI is performed on married and unmarried the are almost overl apping. But the when do univariate and bivariate analysis it shows single's spend mo re.
- The age groups results after performing CI are the limits of age groups (26-35) are much higher than the remaining groups. When we have done various analysis like u nivariate and bivariate and same result came that age group of (26-35) are the more s penders.

8. Recommendations

- Men spent more money than women, company can focus on retaining the male customers and getting more male customers.
- Product Category - 1, 5, 8 have highest purchasing frequency. It means these ar e the products in these categories are in more demand. Company can focus on s elling more of these products.
- Unmarried customers spend more money than married customers, so company should focus on acquisition of Unmarried customers.
- Customers in the age 26-35 spend more money than the others, so company sho uld focus on acquisition of customers who are in the age 26-35.
- We have more customers aged 26-35 in the city category B and A, company ca n focus more on these customers for these cities to increase the business.
- Male customers living in City Category C spend more money than other male c ustomers living in B or C, Selling more products in the City Category C will he lp the company increase the revenue.
- Some of the Product category like 19, 20 and 13 have very less purchase. Comp any can think of dropping it.
- The top 10 users who have purchased more the company should give more offe rs and discounts so that they can be retained and can be helpful for companies b usiness.
- The occupation which are contributing more company can think of offering cre dit cards or other benefits to those customers by liaising with some financial pa rtners to increase the sales.
- The top products should be given focus in order to maintain the quality in order to further increase the sales of those products.
- People who are staying in city for a year have contributed to 35% of the total p urchase amount. Company can focus on such customer base who are neither to o old nor too new residents in the city.
- We have highest frequency of purchase order between 5k and 10k, company ca n focus more on these mid-range products to increase the sales.