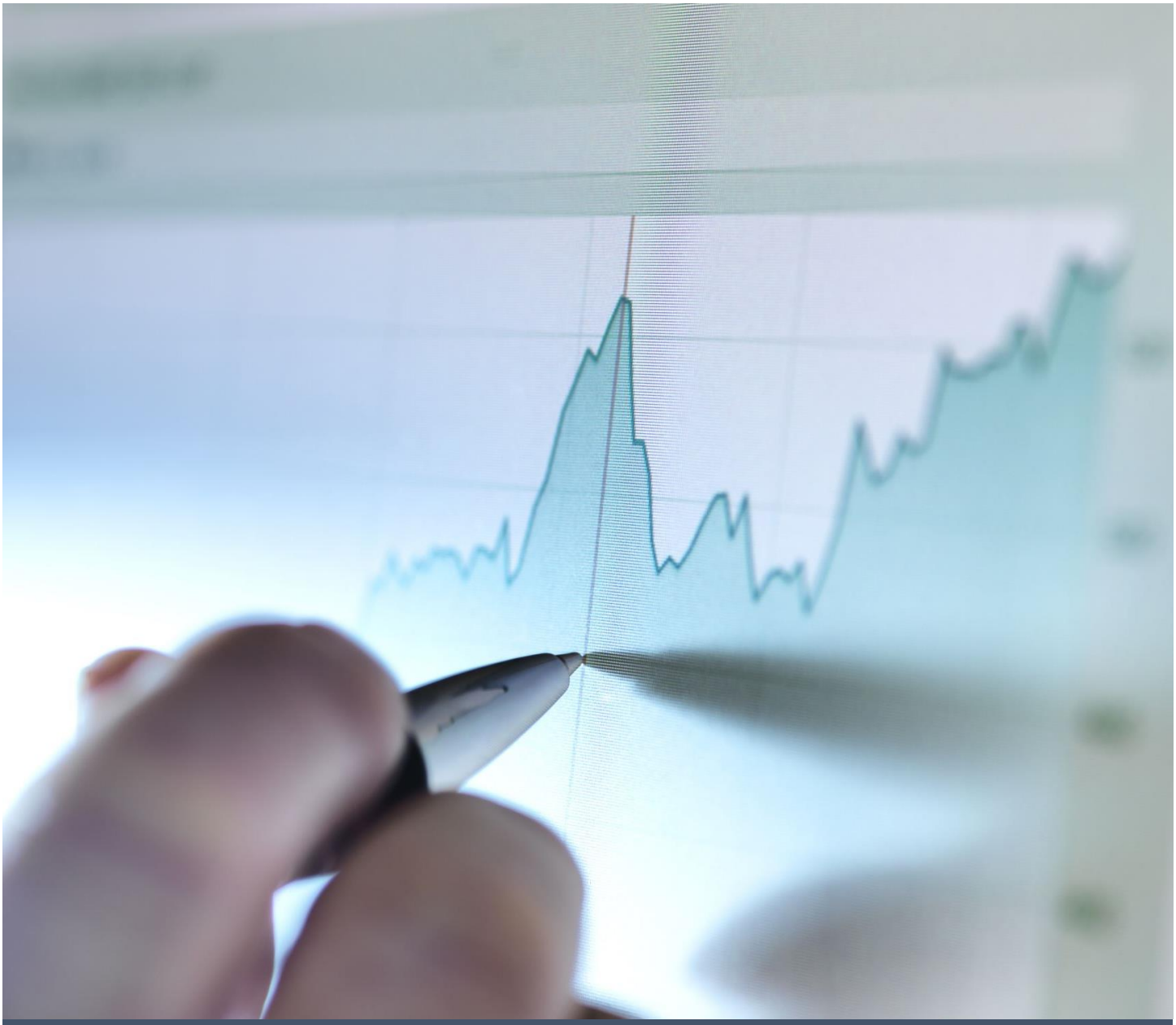


Business Case: Walmart - Confidence Interval and CLT

Nikhil K A

SCALER DSML

Business Case: Walmart - Confidence Interval and CLT

The Management team at Walmart Inc. wants to analyse the customer purchase behaviour (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).

1. Defining Problem Statement and Analysing basic metrics.

a) Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary

Walmart Inc. is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores in the United States, headquartered in Bentonville, Arkansas. The business case envisages analysing of the given Walmart dataset and using different methods like visual and non-visual analysis and statistical analysis and formulate insights which will help Walmart in decision making. The business case leverages Python's robust data analytics and visualization capabilities to extract valuable insights from the data set, purchasing patterns, and product performance metrics. By harnessing Python libraries such as Pandas, NumPy, SciPy, Matplotlib, and Seaborn, the case aims to gain a comprehensive understanding of user preferences, market trends, and market dynamics. Through data-driven analysis and visualization techniques, the case tries to optimize content recommendations to cater to diverse customer segments. This data-centric approach empowers Walmart to make informed decisions, drive customer retention and growth, and maintain a leading position in the ever-evolving retail industry landscape.

The data set have the following columns:

1	User_ID	:	User ID
2	Product_ID	:	Product_ID
3	Gender	:	Sex of User
4	Age	:	Age in bins
5	Occupation	:	Occupation (Masked)
6	City_Category	:	Category of the City (A, B, C)
7	StayInCurrentCityYears	:	Number of years stay in current city
8	Marital_Status	:	Marital Status
9	ProductCategory	:	Product Category (Masked)

10 Purchase : Purchase Amount

The data set is downloaded as ‘walmart_data.csv’ and saved as dataframe named ‘df’.

```
!gdown https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/293/original/walmart_data.csv?1641285094
```

```
Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/293/original/walmart_data.csv?1641285094
To: /content/walmart_data.csv?1641285094
100% 23.0M/23.0M [00:00<00:00, 151MB/s]
```

```
[2] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[8] df=pd.read_csv('walmart_data.csv?1641285094')
```

The data sample is observed by df.head()

```
[11] df.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	7969

The data is divided into 10 columns and there are 550068 rows in the dataset.

Shape of the dataframe : df.shape showed

```
[13] df.shape
```

```
(550068, 10)
```

The basic information about dataframe. df.info()

```
[15] 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
```

Data type of 5 of the 10 columns are object type, other 5 columns being int64.

Detecting missing values by isna().

```
[9] df.isna().sum()
```

```
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
dtype: int64
```

It shows there are no null values or missing values in the dataset.

b) Non-Graphical Analysis: Value counts and unique attributes.

The dataset consists of data of purchases by 5891 unique customers

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

```
5891
```

The dataset shows there are 5891 unique products.

```
[6] df['Product_ID'].nunique()
```

```
3631
```

Since the data contains entries of purchases there are multiple entries of same customer. Drop duplicate method is used to create a new dataframe with unique entries of users.

The data contains 4225 male customers and 1666 female customers.

```
[12] df1=df.drop_duplicates(subset='User_ID')
```

```
[14] df1['Gender'].value_counts()
```

```
Gender
M    4225
F    1666
Name: count, dtype: int64
```

The data shows that among the 5891 customers, the distribution among different age bins.

```
[16] df1['Age'].value_counts()
```

```
Age
26-35    2053
36-45    1167
18-25    1069
46-50     531
51-55     481
55+       372
0-17      218
Name: count, dtype: int64
```

Among the customers 3139 are from city category C, 1707 from B category and remaining 1045 from category A city.

```
[18] df1['City_Category'].value_counts()
```

```
City_Category
C    3139
B    1707
A    1045
Name: count, dtype: int64
```

The data shows that among the customers, 2474 are partnered and 3417 are single.

```
[19] df1['Marital_Status'].value_counts()
```

```
Marital_Status
0    3417
1    2474
Name: count, dtype: int64
```

c) Visual Analysis - Univariate & Bivariate

i. For continuous variable(s): Distplot, countplot, histogram for univariate analysis

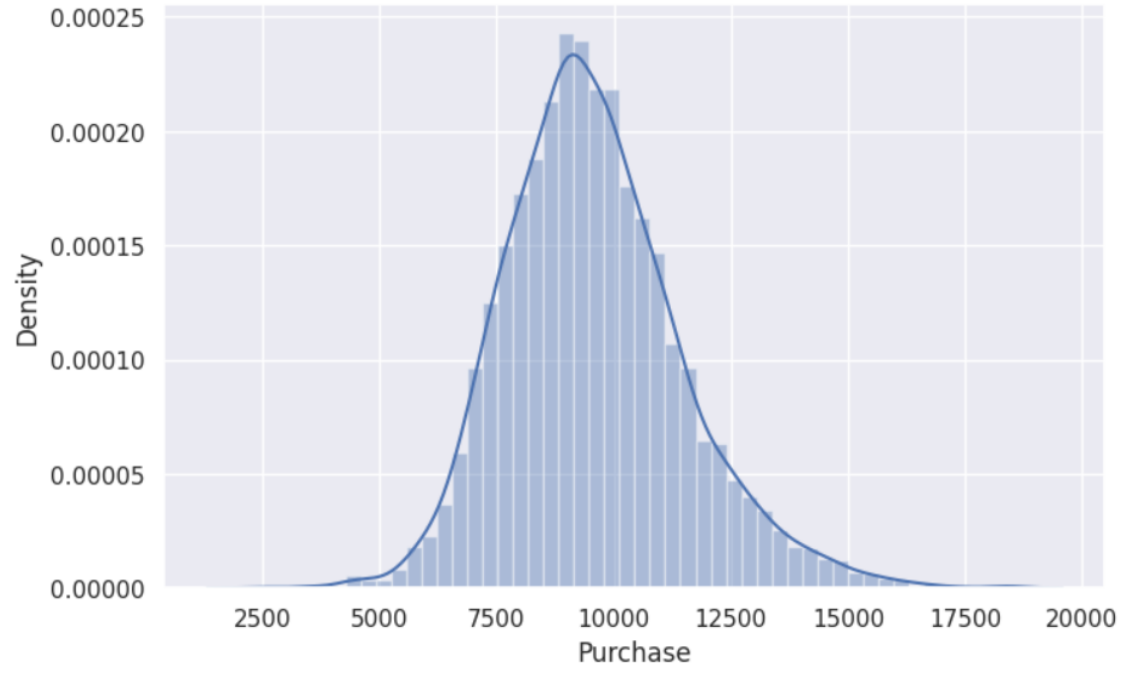
The purchase amount is repeated for different purchases of a single user. The data is grouped by user name and mean of the purchase amount is taken for analysis.

```
[5] df_mean=df.groupby(by=["User_ID","Gender","Age","Occupation","City_Category","Marital_Status"])[['Purchase']].mean()
```

Distribution of customers purchases mean amount are depicted using distplot.

```
[6] sns.set(rc={"figure.figsize":(8,5)})
    sns.distplot(df_mean['Purchase'])
```

```
sns.distplot(df_mean['Purchase'])
<Axes: xlabel='Purchase', ylabel='Density'>
```



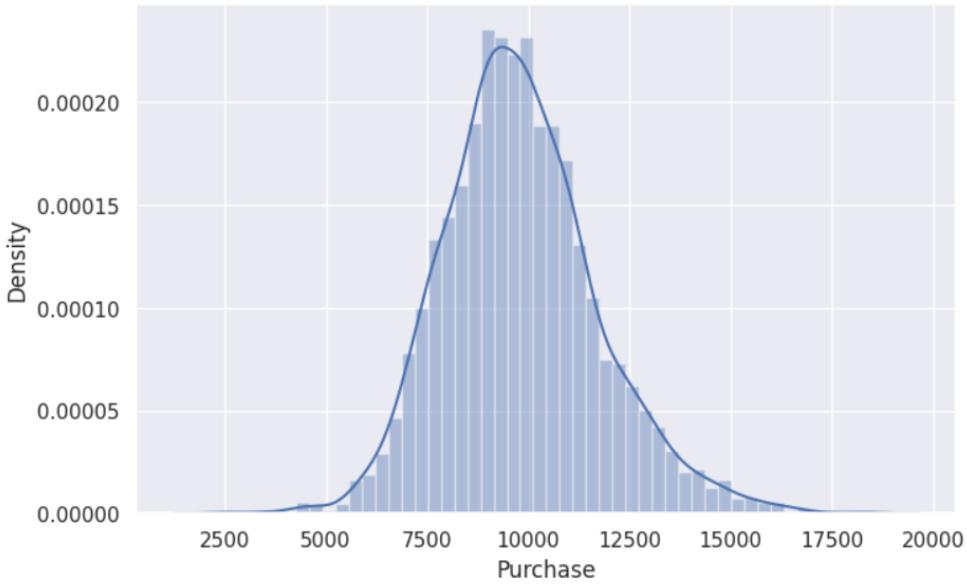
The number of customers peaks at about a purchase amount of 10000 and then gradually decreases.

Similarly, the distribution of male customers and female customers are analysed separately.

```
[17] df_men=df[df['Gender']=='M']
     df_women=df[df['Gender']=='F']
```

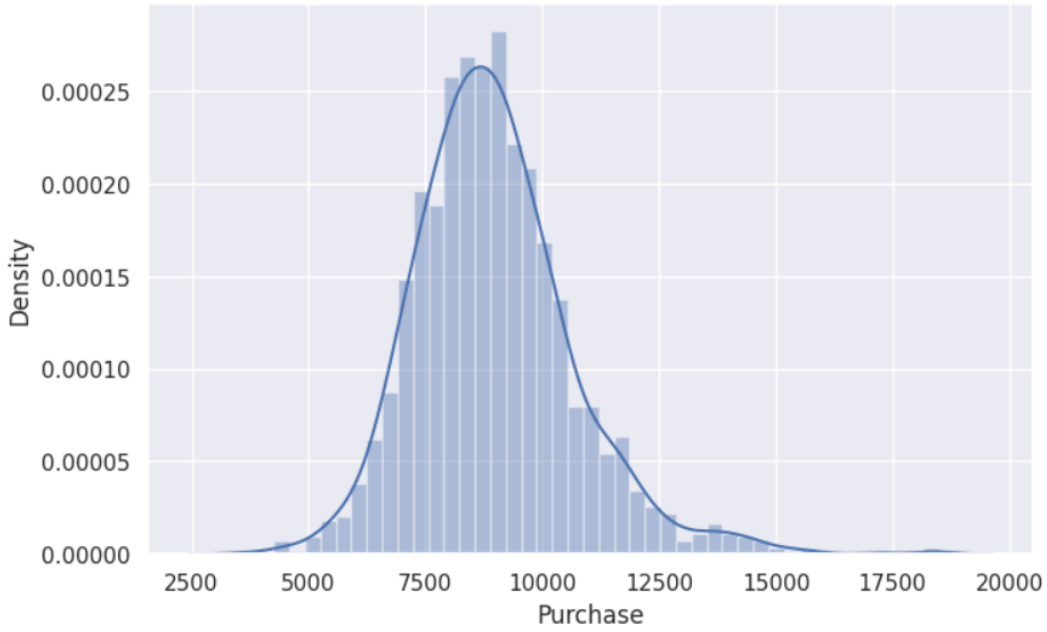
```
[18] df_men_mean=df_men.groupby(by=["User_ID","Gender","Age","Occupation","City_Category","Marital_Status"])[['Purchase']].mean()
     sns.set(rc={"figure.figsize":(8,5)})
     sns.distplot(df_men_mean['Purchase'])
```

```
sns.distplot(df_men_mean['Purchase'])
<Axes: xlabel='Purchase', ylabel='Density'>
```




```
[19] df_women_mean=df_women.groupby(by=["User_ID","Gender","Age","Occupation","City_Category","Marital_Status"])[['Purchase']].mean()
     sns.set(rc={"figure.figsize":(8,5)})
     sns.distplot(df_women_mean['Purchase'])
```

```
sns.distplot(df_women_mean['Purchase'])
<Axes: xlabel='Purchase', ylabel='Density'>
```



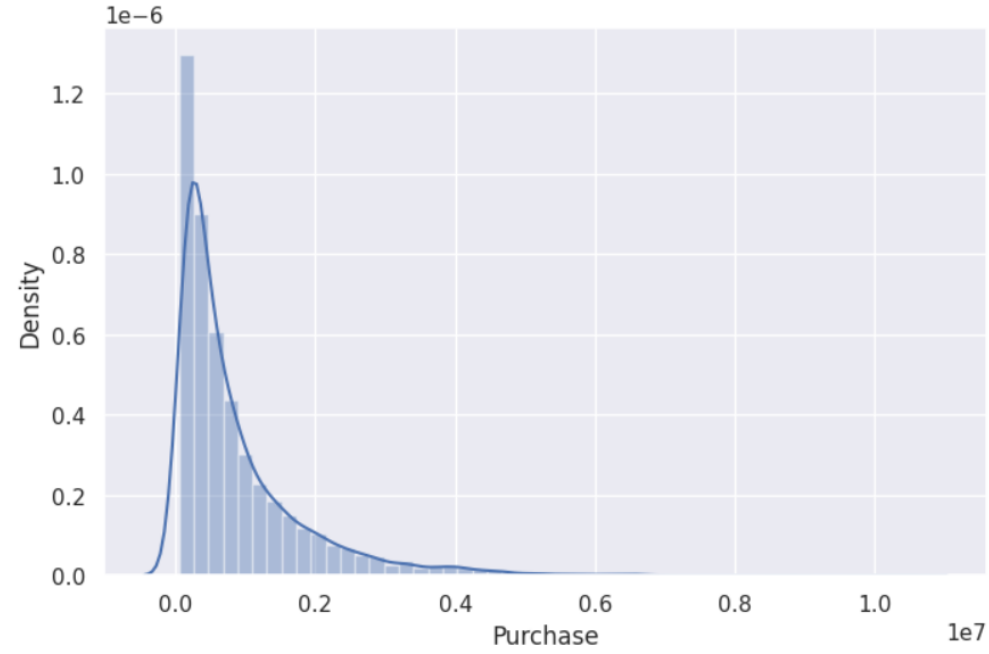
The peak of number of customers are about 10000$ purchase amount in mens category. In case of women, the peak of number of customers are about 87500$ and shows a narrow peak.

The sum of purchases by each user is found out and the variation of number of purchases along the sum is analysed.

```
df_sum=df.groupby(by=["User_ID","Gender","Age","Occupation","City_Category","Marital_Status"])[['Purchase']].sum()
```

```
[27] sns.set(rc={"figure.figsize":(8,5)})
     sns.distplot(df_sum['Purchase'])
```

```
sns.distplot(df_sum['Purchase'])
<Axes: xlabel='Purchase', ylabel='Density'>
```

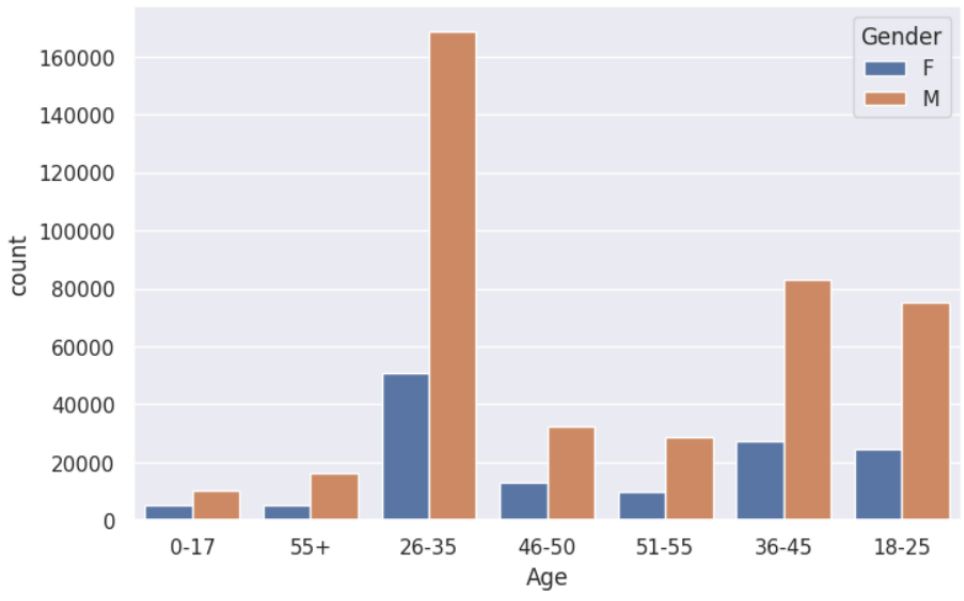


The distribution of number of purchases shown an exponential reduction with the increase in amount spent in total.

Distribution of count of customer purchases of products across age and gender.

```
[22] sns.set(rc={"figure.figsize":(8,5)})
     sns.countplot(data=df,hue='Gender',x='Age')
```

```
<Axes: xlabel='Age', ylabel='count'>
```

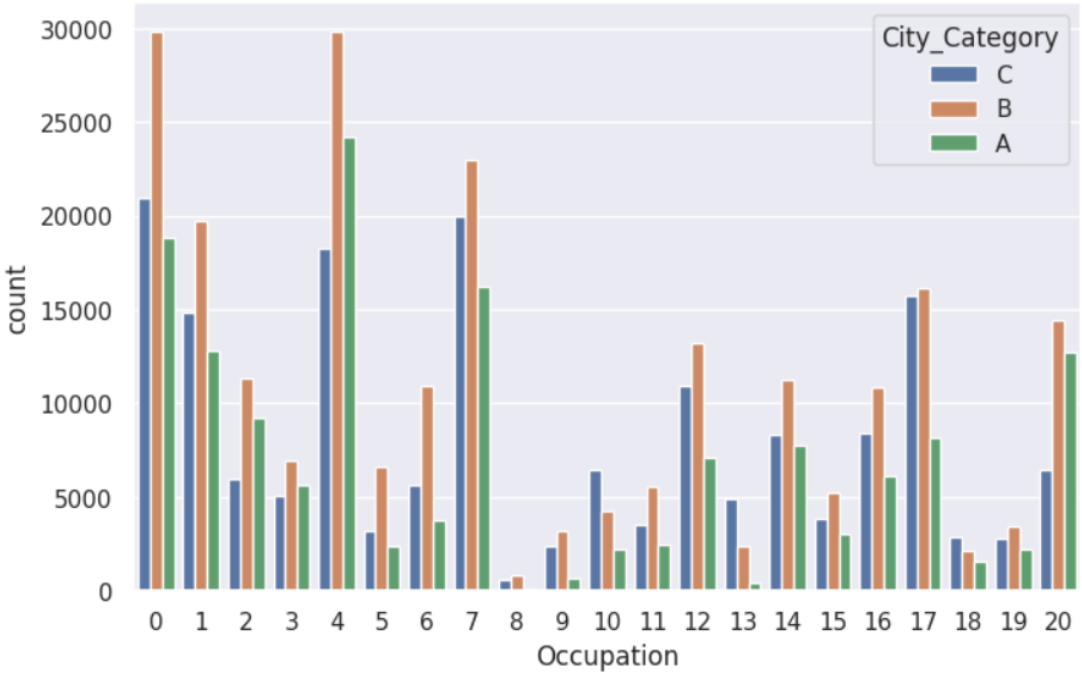


Here, the largest number of customer purchases are from the age group 26-35. And among all age groups men are predominant in the number of purchases.

When analysed across occupation and city category, the data shown the following trend:

```
[21] sns.set(rc={"figure.figsize":(8,5)})
     sns.countplot(data=df,hue='City_Category',x='Occupation')
```

```
<Axes: xlabel='Occupation', ylabel='count'>
```

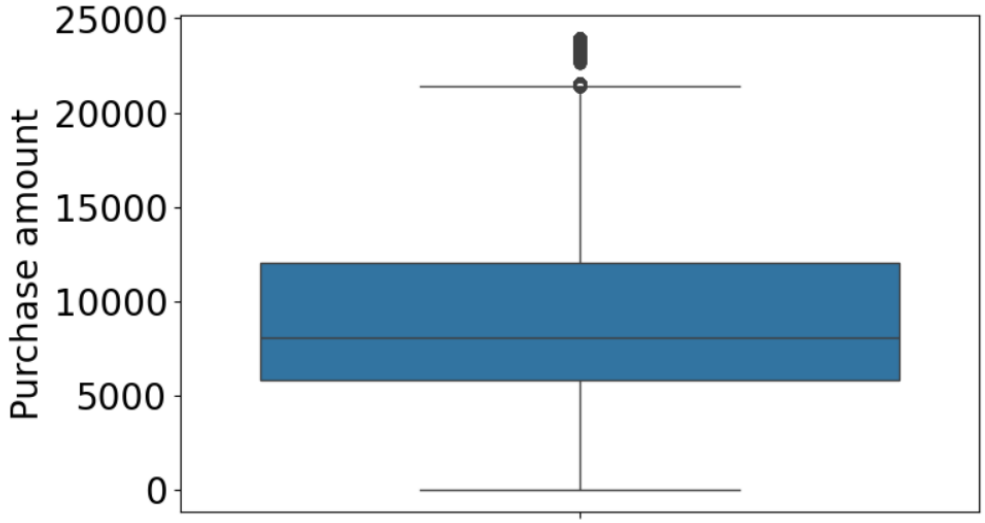


The leading number of purchases are from occupation 0, 4 and 7 and among all occupations except a few, B category city contributes to major number of purchases.

ii. For categorical variable(s): Boxplot

The data of customer purchases is plotted as a box plot.

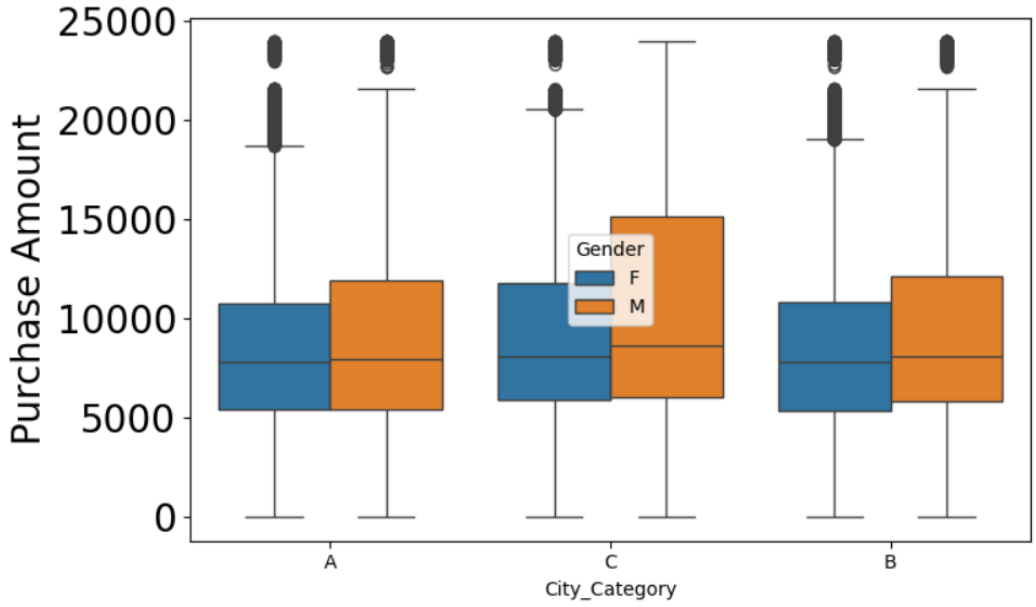
```
[6]  plt.figure(figsize=(8,5))
     sns.boxplot(data=df,y='Purchase')
     plt.yticks(fontsize=20)
     plt.ylabel('Purchase amount', fontsize=20)
```



Here we can observe that there are some outliers.

The customer purchases for different city categories over gender:

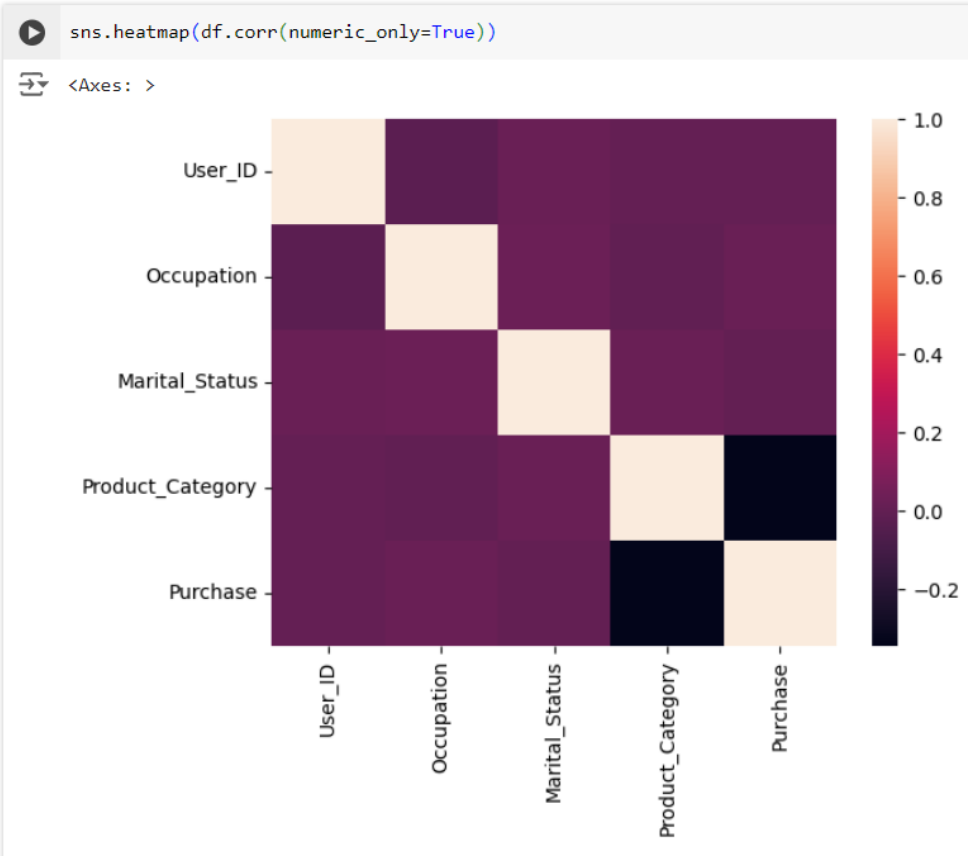
```
[8]  plt.figure(figsize=(8,5))
     sns.boxplot(data=df,x='City_Category',hue='Gender',y='Purchase')
     plt.yticks(fontsize=20)
     plt.ylabel('Purchase Amount', fontsize=20)
```



All the city categories show outliers. Even though we have seen above that city category B is more in number of purchases, city category C shows higher median and range.

iii) For correlation: Heatmaps, Pairplots

The correlation between different numerical data is depicted in the heat map.



The heatmap shows there are no direct correlation between the numeric parameters.

2) Missing Value & Outlier detection

Detecting missing values by isna().

```
[ ] df.isna().sum()
```

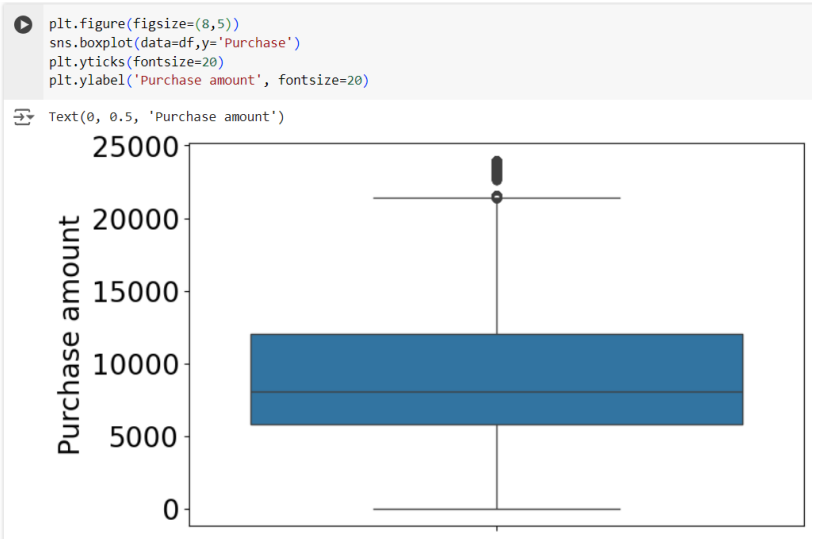
```
User_ID                     0
Product_ID                  0
Gender                      0
Age                         0
Occupation                  0
City_Category               0
Stay_In_Current_City_Years  0
Marital_Status              0
Product_Category            0
Purchase                    0
dtype: int64
```

It shows there are no null values or missing values in the dataset.

In the box plot drawn using the purchase amount customers, outliers have been spotted.

```
plt.figure(figsize=(8,5))
sns.boxplot(data=df,y='Purchase')
plt.yticks(fontsize=20)
plt.ylabel('Purchase amount', fontsize=20)
```

```
Text(0, 0.5, 'Purchase amount')
```



3) Business Insights based on Non-Graphical and Visual Analysis

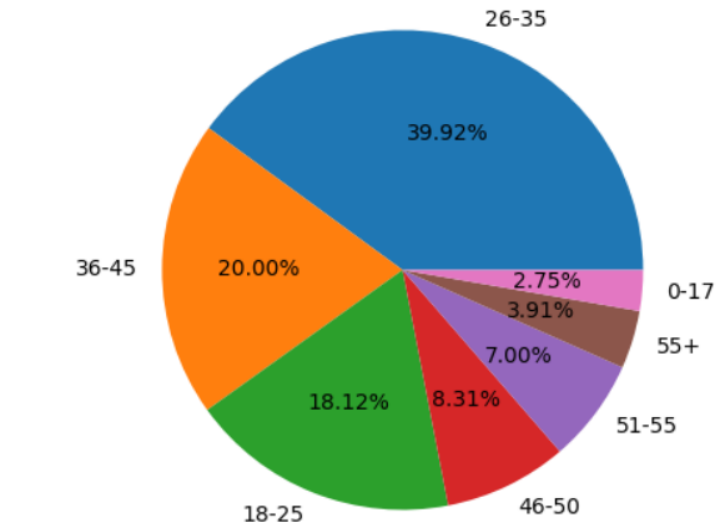
i. Comments on the range of attributes

- The data contains 550 thousand entries corresponding to different customer purchases of variety of products.
- There are 10 different attributes corresponding to each customer purchase.
- There are data of 5891 customers who have purchased from a set of 3631 different products.
- Among the customers, 4225 are men and 1666 are women.
- The customers consist of ages 18-45 majorly.
- The customers are from cities categorised into three: A category, B category and C category.
- Even though more customers are from city C, the greatest number of purchases are from city B.
- The products in the data set are categorised into 20 different categories. In both men and women categories 1,5 and 8 are popular products.
- The occupation of customers is categorised into 20. Among men, 4, 0 and 7 are common. And among women, 0,1 and 4 are common.

ii. Comments on the distribution of the variables and relationship between them

- The distribution of customer purchases over different age bins:

```
counts=df['Age'].value_counts()
plt.figure(figsize = (5,5))
plt.pie(counts,
        labels=counts.index,
        startangle=0,
        autopct = '%.2f%%')
plt.show()
```

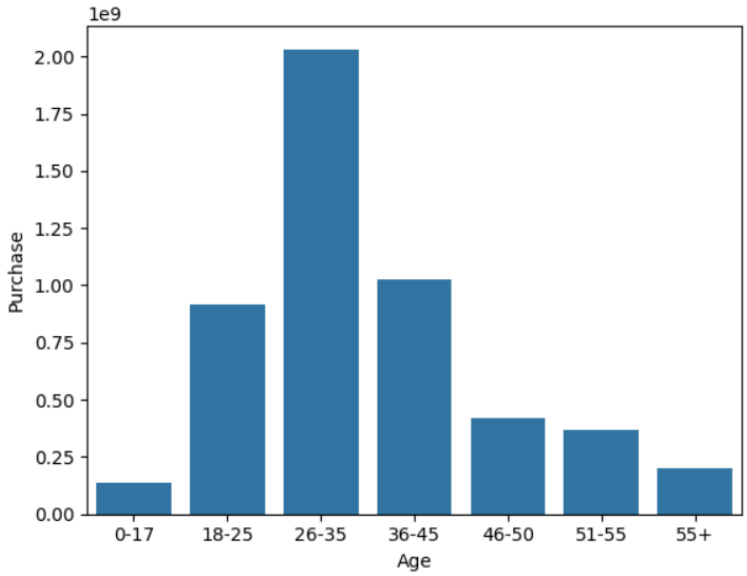


The major part of customer purchases is from 26-35 age bin constituting 40% of total number of purchases. Age groups of 36-45 and 18-25 following with nearly 20% of number of purchases.

- When considering the sum of purchases, the distribution is as follows:

```
[16] sns.barplot(data=df.groupby(by='Age')[['Purchase']].sum().reset_index(),x='Age',y='Purchase')
```

```
<Axes: xlabel='Age', ylabel='Purchase'>
```

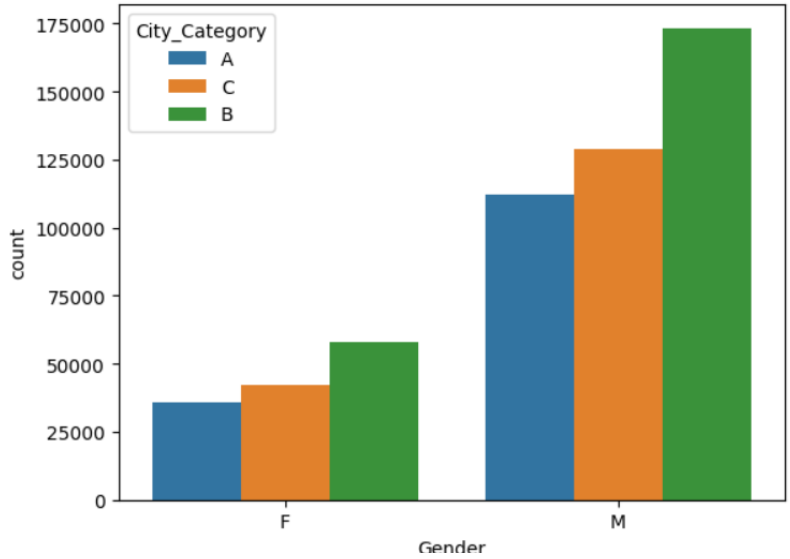


Here also, the major contributors are 26-35 age bin, followed by 36 to 45 and 18-25.

- The distribution of customer purchases across gender from different categories of cities:

```
[10] sns.countplot(data=df,x='Gender',hue='City_Category')
```

```
<Axes: xlabel='Gender', ylabel='count'>
```



The major part of customer purchases is from men. Among different categories of cities, B-category city shows higher purchase percentage.

iii. Comments for each univariate and bivariate plot

- The number of customers peaks at about a purchase amount of 10000 and then gradually decreases. The peak of number of customers are about 10000$ purchase amount in men's category. In case of women, the peak of number of customers are about 87500$ and shows a narrow peak.

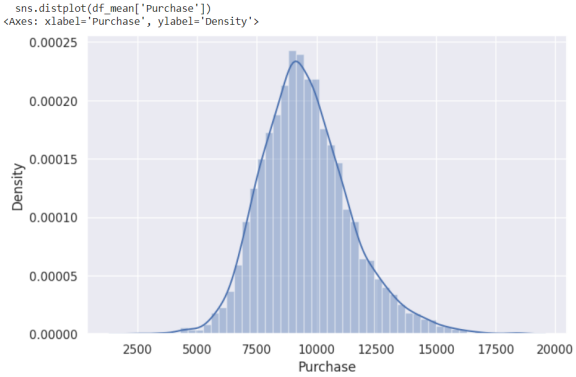


Figure 1Distribution of purchases of all customers

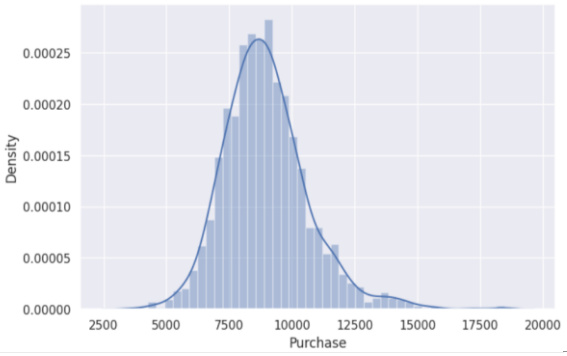


Figure 2Distribution of purchases of male customers

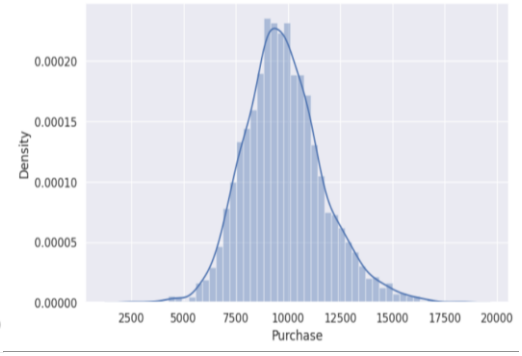
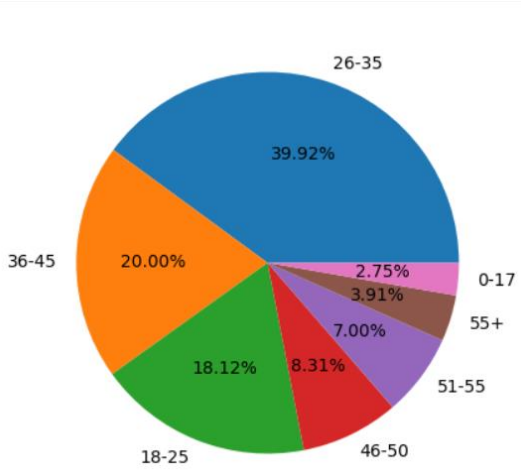


Figure 3Distribution of purchases of female customers

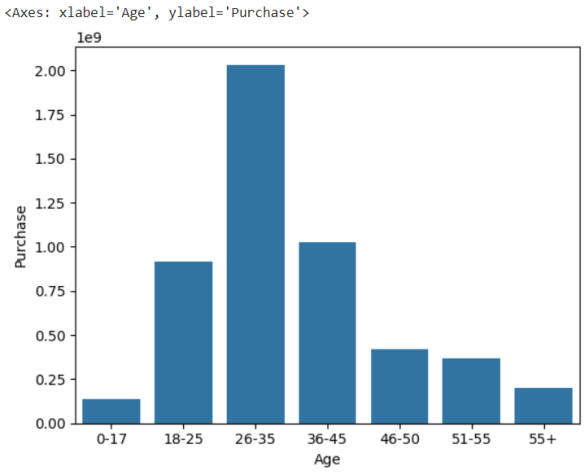
- The sum of purchases by each user is found out and the variation of number of purchases along the sum is analysed. The distribution of number of purchases shown an exponential reduction with the increase in amount spent in total. This shows an inverse relation between the amount of purchase and number of purchases. It also indicates that both the large amount and small amount purchases are important for the firm since they both contribute to the revenue equally importantly.

- The distribution of customer purchases over different age bins:



The major part of customer purchases is from 26-35 age bin constituting 40% of total number of purchases. Age groups of 36-45 and 18-25 following with nearly 20% of number of purchases.

- In case of the sum of purchases contributed among different age levels also,



Here also, the major contributors are 26-35 age bin, followed by 36 to 45 and 18-25.

4) Answering questions

a) Are women spending more money per transaction than men? Why or Why not?

When analysed, the distribution of purchase amounts of men and women shown the following parameters.

```
[13] df_men['Purchase'].describe()
```

```
count    414259.00000
mean       9437.52604
std        5092.18621
min          12.00000
25%        5863.00000
50%        8098.00000
75%       12454.00000
max       23961.00000
Name: Purchase, dtype: float64
```

```
[14] df_women['Purchase'].describe()
```

```
count    135809.000000
mean       8734.565765
std        4767.233289
min          12.000000
25%        5433.000000
50%        7914.000000
75%       11400.000000
max       23959.000000
Name: Purchase, dtype: float64
```

From the mean value of purchase amounts of men and women we can assume that men are spending more than women.

The hist plot of purchase amounts of men and women are observed to be:

```
plt.figure(figsize = (5,5))
sns.histplot(df_men['Purchase'])
```

```
<Axes: xlabel='Purchase', ylabel='Count'>
```

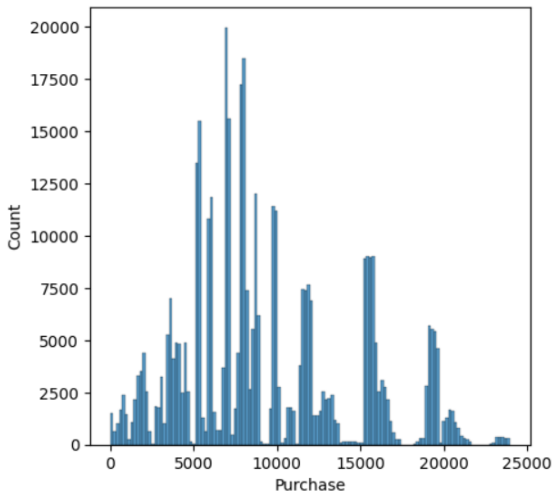


Figure 4 Histplot of purchase amount of male customers

```
plt.figure(figsize = (5,5))
sns.histplot(df_women['Purchase'])
```

```
<Axes: xlabel='Purchase', ylabel='Count'>
```

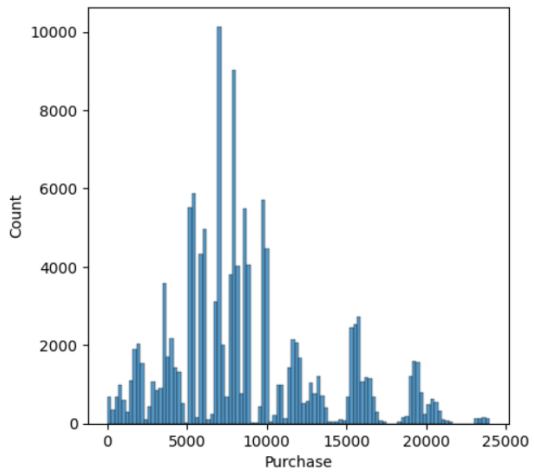


Figure 5 Histplot of purchase amount of female customers

The distribution of purchases by male and female customers are not found to be a standard distribution. Central Limit Theorem (CLT) is used to approximate it to normal distribution.

Central Limit Theorem

The central limit states that "the mean of a random sample will resemble even closer to the population mean as the sample size increases and it will approximate a normal distribution regardless of the shape of the population distribution"

- Means, if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed.

- The CLT tends to work well when the sample size (n) is sufficiently large, typically considered as $n \geq 30$. However, it is not a strict rule but a rough guideline.

At first, 10000 number of samples of 5 is taken from the data of male customers.

```
[18] sample_mean5=[np.mean(df_men['Purchase'].sample(5)) for i in range(10000)]
```

Here, the mean value of means can be found out to be:

```
[20] np.mean(sample_mean5)

9425.91734
```

Where the mean of total values from previous calculation is 9437.5.

For higher accuracy, the sample size is increased to 20.

```
[21] sample_mean20=[np.mean(df_men['Purchase'].sample(20)) for i in range(10000)]
```

Here the mean value of means is:

```
[22] np.mean(sample_mean20)
     9444.2978
```

The sample size is increased to 30.

```
[23] sample_mean30=[np.mean(df_men['Purchase'].sample(30)) for i in range(10000)]
```

Here the obtained mean value of means is:

```
[24] np.mean(sample_mean30)
     9431.307203333334
```

The sample size is increased to 35

```
[26] sample_mean35=[np.mean(df_men['Purchase'].sample(35)) for i in range(10000)]
```

Here the obtained mean value of means is found to be closer to the actual mean value i.e., 9437.5

```
[28] np.mean(sample_mean35)
     9435.16338857143
```

Hence the sample size is fixed at 35.

The mean values of means of samples for male and female customers:

b) Confidence intervals and distribution of the mean of the expenses by female and male customers.

Distribution of sample means of expenses by male and female customers:

```
plt.figure(figsize = (5,5))
sns.histplot(sample_men,kde=True)
```

```
<Axes: ylabel='Count'>
```

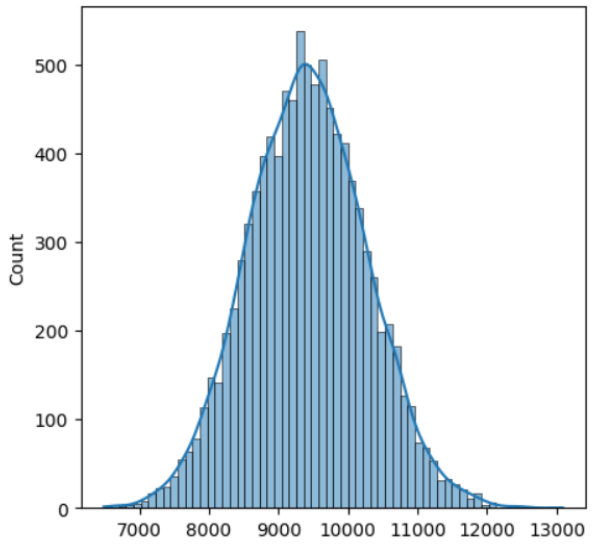


Figure 6 Distribution of sample means of purchases by male customers

```
plt.figure(figsize = (5,5))
sns.histplot(sample_women,kde=True)
```

```
<Axes: ylabel='Count'>
```

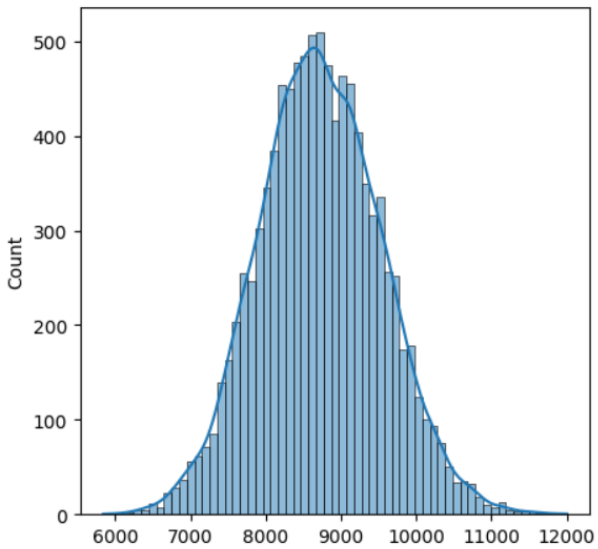


Figure 7 Distribution of sample means of purchases by female customers

Confidence intervals

Confidence intervals are calculated using norm.interval() method from the library scipy.stats.

```
[38] from scipy.stats import norm

[39] norm.interval(confidence=.90,loc=np.mean(sample_men),scale=np.std(sample_men))

     (8019.6149308967915, 10816.276029103208)
```

The confidence intervals for different confidence levels (90%,95%,99%) are as follows.

Confidence Level	Confidence interval for	
	Male customers	Female customers
90%	(8019.6149308967915, 10816.276029103208)	(7396.765704212212, 10057.709861502073)
95%	(7751.731860856329, 11084.159099143671)	(7141.882518502584, 10312.593047211702)
99%	(7228.169596352978, 11607.721363647022)	(6643.72779049456, 10810.747775219725)

c) Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements?

The confidence interval of average male spending is found to be overlapping with the average spending of female customers. For example, in 90% confidence level, male average spending shows a confidence interval of (8019.6, 10816.27). While, for the same confidence level, female average spending confidence interval is (7396.76, 10057.71). Here the confidence interval of female spending is inclusive to that of male spending.

This insight can be used to improve the income to Walmart by

- Introducing more products whose price lies in the confidence interval.
- Discounts and other methods can be used to bring the prices of products into this range.
- Unisex products can be promoted more in this price range.
- Since this price range is more popular from the confidence interval, combos and such offers can be formulated to provide product ranges in this price range.

d) Results when the same activity is performed for Married vs Unmarried

The dataframe is divided into single and married entries.

```
[6]  df_single=df[df['Marital_Status']==0]
     df_marr=df[df['Marital_Status']==1]
```

Single Customers:

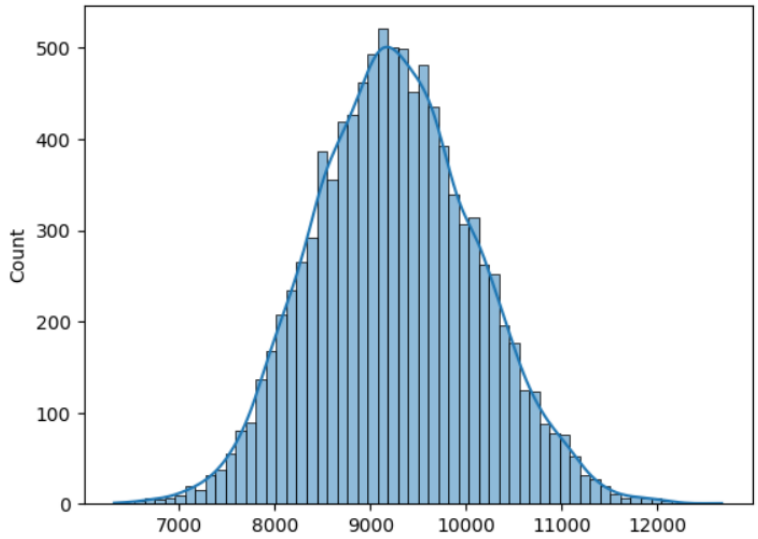
```
[8]  sample_single=[np.mean(df_single['Purchase'].sample(35)) for i in range(10000)]
```

```
[9]  np.mean(sample_single)
```

```
9258.264448571428
```

```
[14] sns.histplot(sample_single,kde=True)
```

```
<Axes: ylabel='Count'>
```



Married Customers:

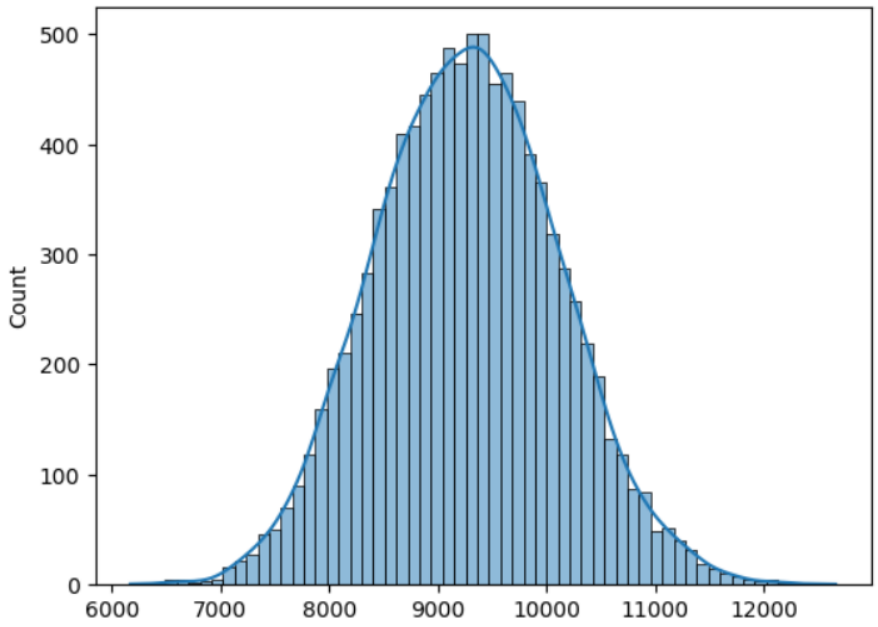
```
[11] sample_marr=[np.mean(df_marr['Purchase'].sample(35)) for i in range(10000)]
```

```
[12] np.mean(sample_marr)
```

```
9269.881322857142
```

```
[15] sns.histplot(sample_marr,kde=True)
```

```
<Axes: ylabel='Count'>
```



The confidence intervals for different confidence levels (90%,95%,99%) are as follows.

Confidence Level	Confidence interval for	
	Single customers	Married customers
90%	(7867.14 , 10649.39)	(7882.43 , 10657.33)
95%	(7600.63 , 10915.90)	(7616.63 , 10923.13)
99%	(7079.77 , 11436.76)	(7097.14 , 11442.62)

Here, the confidence intervals of amount of purchase per purchase of single customers are found to be overlapping with that of married customers. The confidence interval obtained for married customers average spending is found to be completely overlapping with that of single customers.

e) Results when the same activity is performed for Age

The dataframe is divided based on the age bins

```
[23] df_0=df[df['Age']=='0-17']
     df_18=df[df['Age']=='18-25']
     df_26=df[df['Age']=='26-35']
     df_36=df[df['Age']=='36-45']
     df_46=df[df['Age']=='46-50']
     df_51=df[df['Age']=='51-55']
     df_55=df[df['Age']=='55+']
```

Samples are taken for each dataframe separately.

```
[33] sample_0=[np.mean(df_0['Purchase'].sample(35)) for i in range(10000)]
     sample_18=[np.mean(df_18['Purchase'].sample(35)) for i in range(10000)]
     sample_26=[np.mean(df_26['Purchase'].sample(35)) for i in range(10000)]
     sample_36=[np.mean(df_36['Purchase'].sample(35)) for i in range(10000)]
     sample_46=[np.mean(df_46['Purchase'].sample(35)) for i in range(10000)]
     sample_51=[np.mean(df_51['Purchase'].sample(35)) for i in range(10000)]
     sample_55=[np.mean(df_55['Purchase'].sample(35)) for i in range(10000)]
```

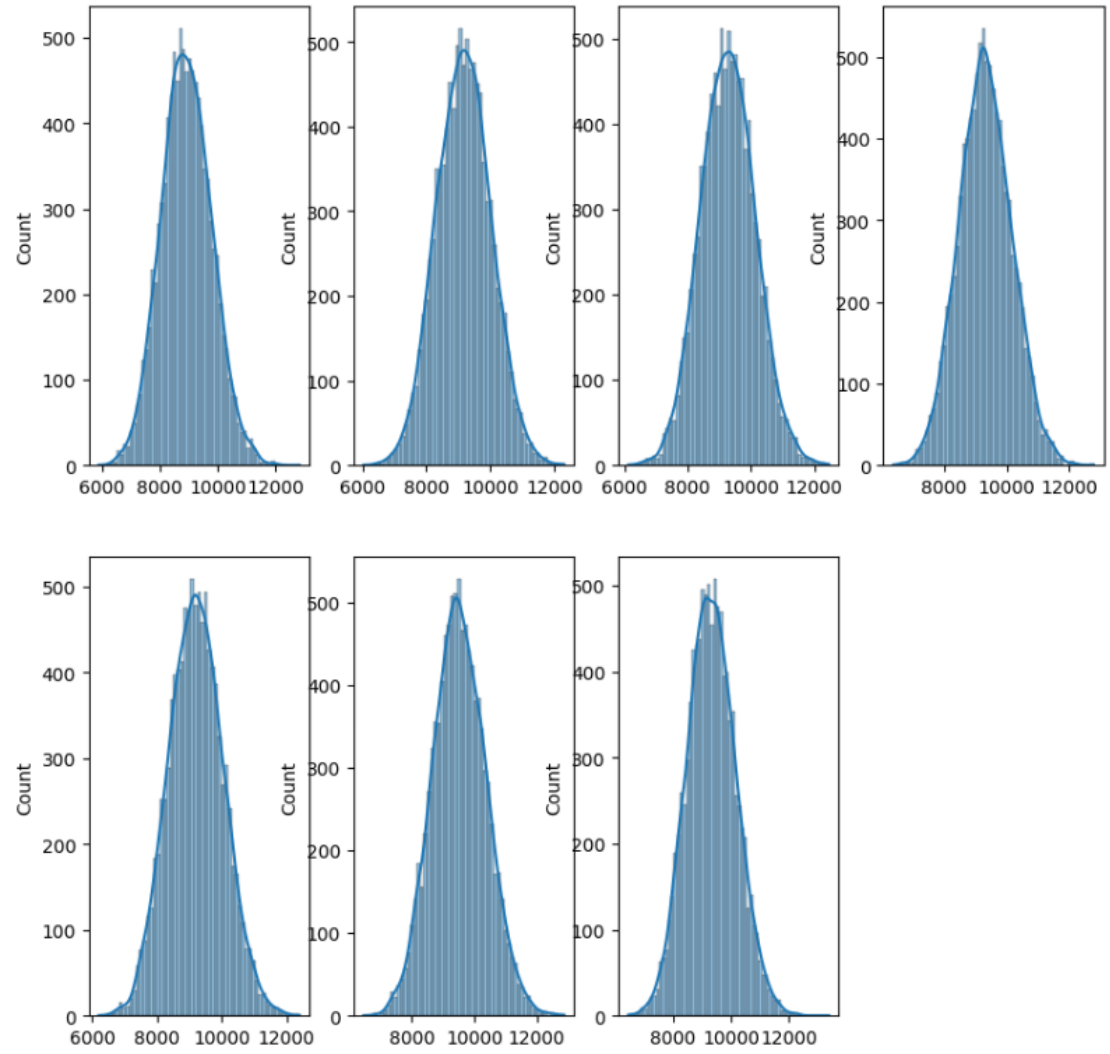
Mean values are calculated:

```
[34] np.mean(sample_0)
8929.845417142858

[35] np.mean(sample_18)
9171.32786

[36] np.mean(sample_26)
9271.744154285714

[37] np.mean(sample_36)
9320.196185714287

[38] np.mean(sample_46)
9203.224065714287

[39] np.mean(sample_51)
9526.989634285714

[40] np.mean(sample_55)
9330.59685142857
```

The distribution of sample means are plotted as histplot:

```
plt.figure(figsize = (10,10))
plt.subplot(2,4,1)
sns.histplot(sample_0,kde=True)
plt.subplot(2,4,2)
sns.histplot(sample_18,kde=True)
plt.subplot(2,4,3)
sns.histplot(sample_26,kde=True)
plt.subplot(2,4,4)
sns.histplot(sample_36,kde=True)
plt.subplot(2,4,5)
sns.histplot(sample_46,kde=True)
plt.subplot(2,4,6)
sns.histplot(sample_51,kde=True)
plt.subplot(2,4,7)
sns.histplot(sample_55,kde=True)
```

```
<Axes: ylabel='Count'>
```



Confidence intervals:

Age bin	Confidence interval for		
	90%	95%	99%
0-17	(7505.07 , 10354.62)	(7232.12 , 10627.56)	(6698.66 , 11161.03)
18-25	(7782.82 , 10559.83)	(7516.82 , 10825.83)	(6996.94 , 11345.72)
26-35	(7858.34 , 10685.14)	(7587.57 , 10955.91)	(7058.37 ,11485.12)
36-45	(7937.66 , 10702.72)	(7672.81 , 10967.58)	(7155.16 , 11485.23)
46-50	(7819.18 , 10587.26)	(7554.04 , 10852.4)	(7035.83 , 11370.62)
51-55	(8123.62 , 10930.35)	(7854.77 , 11199.2)	(7329.32 , 11724.65)
55+	(7944.74 , 10716.45)	(7679.24 , 10981.95)	(7160.35 , 11500.84)

Here, when analysing the confidence interval of customer's average spending across age bins, the age group 0-17 is found to be having a lower mean and confidence interval i.e., probability of lower amount of average spending. The age groups 18 to 50 and 55+ have a comparable confidence interval with each other. Among them 55+ and 36-45 being the higher end. The age bin 51-55 shows a higher confidence interval.

Even though in the volume analysis of purchases the age bins of 26-45 showed higher amount spent in total and number of purchases, the average spending seems to be higher on 51-55 and 55+ age bins.

5) Final Insights

- The number of customers peaks at about a purchase amount of 10000 and then gradually decreases. The peak of number of customers are about 10000$ purchase amount in men's category. In case of women, the peak of number of customers are about 87500$ and shows a narrow peak. The confidence interval of average male spending is found to be overlapping with the average spending of female customers. The confidence interval of female spending is inclusive to that of male spending.

- The sum of purchases by each user is found out and the variation of number of purchases along the sum is analysed. The distribution of number of purchases shown an exponential reduction with the increase in amount spent in total. This shows an inverse relation between the amount of purchase and number of purchases. It also indicates that both the large amount and small amount purchases are important for the firm since they both contribute to the revenue equally importantly.

- The confidence intervals of amount of purchase per purchase of single customers are found to be overlapping with that of married customers. The confidence interval obtained for married customers average spending is found to be completely overlapping with that of single customers.

- The major part of customer purchases is from 26-35 age bin. Age groups of 36-45 and 18-25 following. These metrices can be used to focus the target customers in these age ranges.

- The age group 0-17 is found to be having a lower mean and confidence interval i.e., probability of lower amount of average spending. The age groups 18 to 50 and 55+ have a comparable confidence interval with each other. Among them 55+ and 36-45 being the higher end. The age bin 51-55 shows a higher confidence interval.

- Even though in the volume analysis of purchases the age bins of 26-45 showed higher amount spent in total and number of purchases, the average spending seems to be higher on 51-55 and 55+ age bins.

6) Recommendations.

- The distribution of spending in volume and rate was found to be more in the age groups of 18-45 with more concentration on the age group 26-35. This refers to the section of people who are more inclined to using the service of Walmart in the future also. This can be used to concentrate the target customers in this age range and formulate further marketing techniques based on that. Promotional activities for this section of customers can be thus very focused and much cost-effective.

- The confidence interval of average male spending is found to be overlapping with the average spending of female customers. The confidence interval of female spending is completely inclusive to that of male spending. Walmart can use these inputs to improve the income by introducing more products whose price lies in the confidence interval, discounts and other methods can be used to bring the prices of products into this range, unisex products can be promoted more in this price range, combos and such offers can be formulated to provide product ranges in this price range.

- The confidence intervals of average spending of single customers are found to be overlapping with that of married customers. The confidence interval obtained for married customers average spending is found to be completely overlapping with that of single customers. Here also Walmart can make use of this insights in formulating promotional methods. The comfortable price range of both single and married customers are coinciding.

- Another interesting trend is in the case of average spending of customers over different age bins. the age group 0-17 is found to be having probability of lower amount of average spending. The age groups 18 to 50 and 55+ have a comparable confidence interval with each other. The age bin 51-55 shows a higher confidence interval. Even though in the volume analysis of purchases the age bins of 26-45 showed higher amount spent in total and number of purchases, the average spending seems to be higher on 51-55 and 55+ age bins. This may point to the fact that customers on the lower age bins may tend to use the service of Walmart for everyday consumptions and small purchases where the higher age bin customers will make use of it for higher and larger purchases.

  Both the customer segments shall be separately treated in terms of marketing activities and formulation of promotional formulae.