WALMART -- BUSINESS CASE STUDY

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

We want 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. We 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 company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday. The dataset has the following features:
Dataset link: **Walmart_data.csv**

User_ID:	User ID
Product_ID:	Product ID
Gender:	Sex of User
Age:	Age in bins
Occupation:	Occupation(Masked)
City_Category:	Category of the City (A,B,C)
StayInCurrentCityYears:	Number of years stay in current city
Marital_Status:	Marital Status
ProductCategory:	Product Category (Masked)
Purchase:	Purchase Amount

PERFORMNIG USUAL DATA ANALYSIS

Various required libraries are imported and data is extracted in pandas data frame object with top 5 rows.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from scipy.stats import binom
```

```
df = pd.read_csv("Walmart.txt")
df.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Pu
0	1000001	P00069042	F	0-17	10	A	2	0.0	3.0	
1	1000001	P00248942	F	0-17	10	A	2	0.0	1.0	
2	1000001	P00087842	F	0-17	10	A	2	0.0	12.0	
3	1000001	P00085442	F	0-17	10	A	2	0.0	12.0	
4	1000002	P00285442	M	55+	16	C	4+	0.0	8.0	

Now to find the columns details below code is used .The code clearly depits the presence of Null values in data frame. Below are the columns with null values .

StayInCurrentCityYears:	Number of years stay in current city
Marital_Status:	Marital Status
ProductCategory:	Product Category (Masked)
Purchase:	Purchase Amount

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244594 entries, 0 to 244593
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     244594 non-null  int64
 1   Product_ID                  244594 non-null  object
 2   Gender                      244594 non-null  object
 3   Age                         244594 non-null  object
 4   Occupation                  244594 non-null  int64
 5   City_Category               244594 non-null  object
 6   Stay_In_Current_City_Years  244593 non-null  object
 7   Marital_Status              244593 non-null  float64
 8   Product_Category            244593 non-null  float64
 9   Purchase                    244593 non-null  float64
dtypes: float64(3), int64(2), object(5)
memory usage: 18.7+ MB

Double-click (or enter) to edit
```

Clearly there are 244594 rows and 10 columns

```
df.shape

(244594, 10)

to check null values
```

To check the total number of null values column wise

```
to check null values

[ ]  df.isna().sum()

User_ID                     0
Product_ID                  0
Gender                      0
Age                         0
Occupation                  0
City_Category               1
Stay_In_Current_City_Years  1
Marital_Status              1
Product_Category            1
Purchase                    1
dtype: int64
```

The code to fill the null values in purchase column with the average value of the column

```
0s
df["Purchase"].fillna(df["Purchase"].mean())
df.head(10)

Product_ID  Gender  Age  Occupation  City_Category  Stay_In_Current
```

To get a descriptive statistics summary of a given dataframe.

```
df.describe()
# there is a huge difference in 50% and mean it means there can be outliers
```

	User_ID	Occupation	Marital_Status	Product_Category	Purchase
count	4.327420e+05	432742.000000	432741.000000	432741.000000	432741.000000
mean	1.003004e+06	8.078458	0.409682	5.296840	9330.689847
std	1.738259e+03	6.523026	0.491775	3.748643	4978.923413
min	1.000001e+06	0.000000	0.000000	1.000000	185.000000
25%	1.001472e+06	2.000000	0.000000	1.000000	5866.000000
50%	1.003049e+06	7.000000	0.000000	5.000000	8061.000000
75%	1.004463e+06	14.000000	1.000000	8.000000	12071.000000
max	1.006040e+06	20.000000	1.000000	18.000000	23961.000000

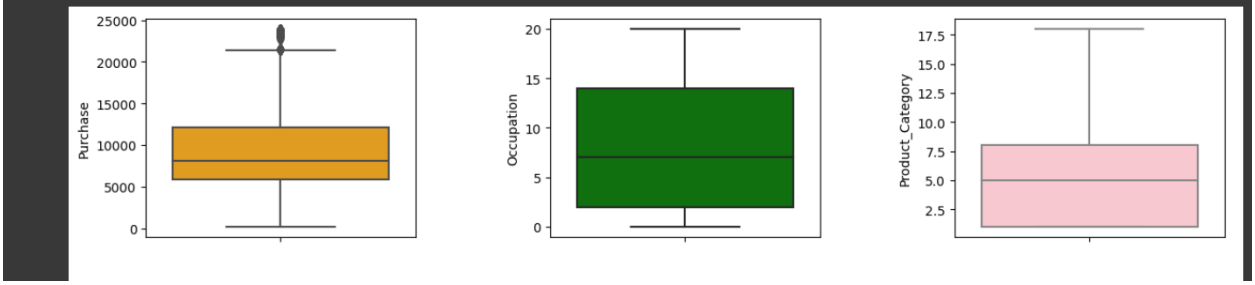
INFERENCE :

By the above image it is clear that there is presence of outliers in Purchase column , occupation as there is huge difference between the mean value and median (50%) value ..

To check outliers

To check outliers

```
[ ] fig,axis = plt.subplots(nrows = 2, ncols = 3, figsize =(15,7))
    sns.boxplot(data = df , y = "Purchase" ,ax = axis[0,0] , color = "orange")
    sns.boxplot(data = df , y = "Occupation" , ax = axis[0,1] , color = "green")
    sns.boxplot(data = df , y = "Product_Category" , ax = axis[0,2] , color = "pink")

    fig.subplots_adjust(hspace=0.3, wspace=0.5)
```



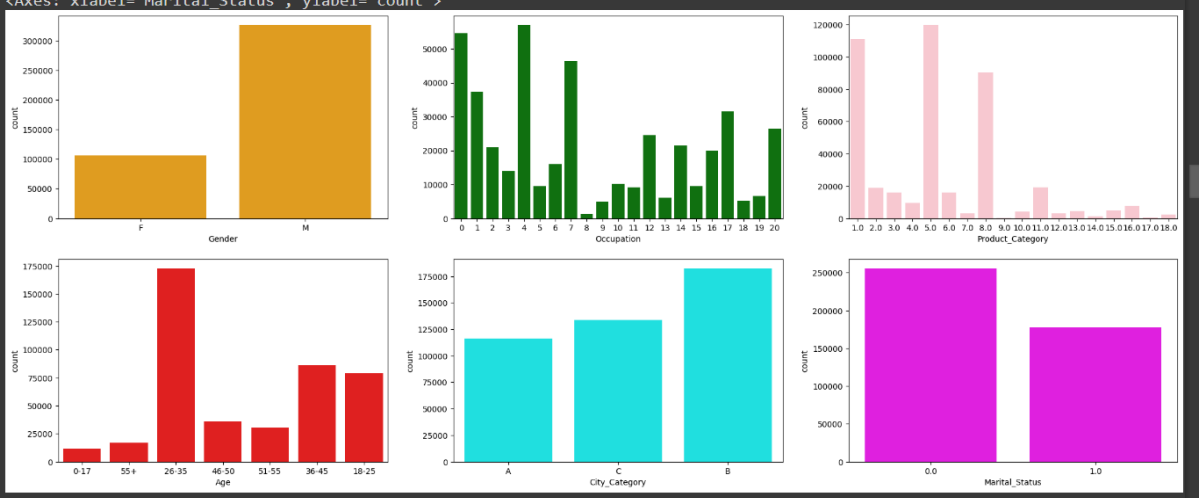
DATA EXPLORATION

UNIVARIATE ANALYSIS

Various categories are compared with the total count of purchase using the below code and got the desired output

```
fig,axis = plt.subplots(nrows = 2, ncols = 3, figsize =(25,10))
sns.countplot(data = df , x = "Gender" ,ax = axis[0,0] , color = "orange")
sns.countplot(data = df , x = "Occupation" , ax = axis[0,1] , color = "green")
sns.countplot(data = df , x = "Product_Category" , ax = axis[0,2] , color = "pink")
sns.countplot(data = df , x = "Age" , ax = axis[1,0] , color = "red")
sns.countplot(data = df , x = "City_Category" , ax = axis[1,1] , color = "aqua")
sns.countplot(data = df , x = "Marital_Status" , ax = axis[1,2] , color = "magenta")
```

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



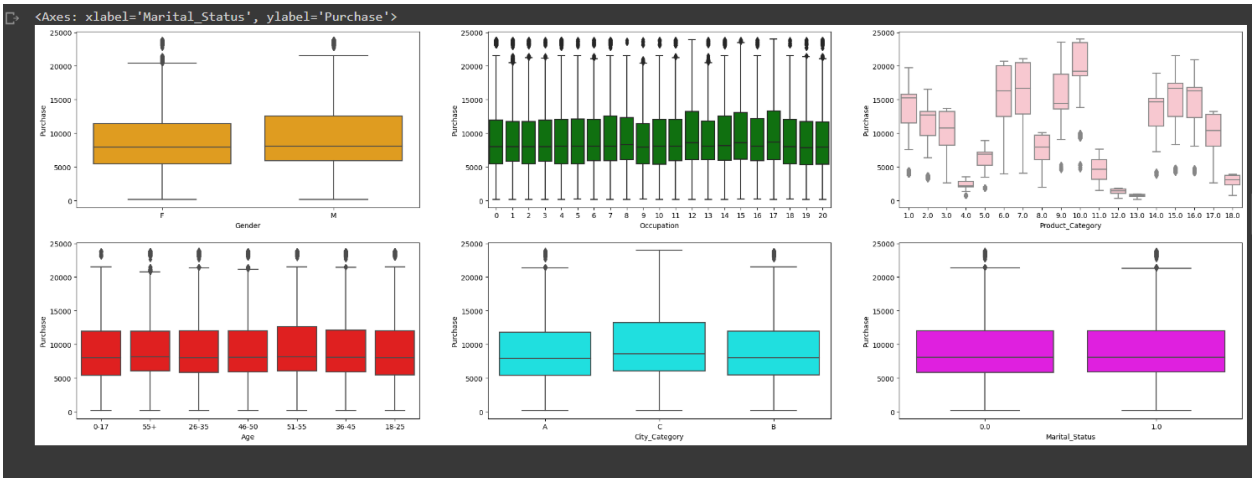
INFERENCE :

- First plot indicates that there are more male customers than females .
- Most customers belong to occupation 4.. Customers with occupation 8 are the least.
- The most sold product category is the product category 5.0 and category number 9.0 and 17.0 is least popular and sold.
- Customers with age group 26-35 are most frequent customers of Wal-Mart.. And teenagers are not fond of Walmart much.
- City B has the highest purchased items.
- There are more unmarried customers than married ones.

BIVARIATE ANALYSIS

To check the effect of categories on purchase .We can use boxplots for analysis using the below code

```
fig,axis = plt.subplots(nrows = 2, ncols = 3, figsize =(30,10))
sns.boxplot(data = df , x = "Gender" ,y = "Purchase",ax = axis[0,0] , color = "orange")
sns.boxplot(data = df , x = "Occupation" , y = "Purchase",ax = axis[0,1] , color = "green")
sns.boxplot(data = df , x = "Product_Category" , y = "Purchase",ax = axis[0,2] , color = "pink")
sns.boxplot(data = df , x = "Age" , y = "Purchase",ax = axis[1,0] , color = "red")
sns.boxplot(data = df , x = "City_Category" , y = "Purchase",ax = axis[1,1] , color = "aqua")
sns.boxplot(data = df , x = "Marital_Status" , y = "Purchase",ax = axis[1,2] , color = "magenta")
```



INFERENCE

- From first plot between gender and purchase . .There are few outliers for the females category .. and its clear than males spent more than females in walmart
- Second plot between occupation and purchase .there are several outlier in each point.However, though customers with occupation 17 are lesser as per univariate analysis. however , they have spent the maximum if outliers are ignored.
- Similary for product category vs purchase. Highest purchase was made by category 10 .However, it has few outliers .
- From purchase vs city , customers belong to city c has made highest transaction of nearly 24000.
- Both married and unmarried spent nearly equal amounts.

TO CALCULATE THE AVERAGE EXPENDITURE BY EACH OF THE GENDER

Tracking the amount spent per transaction of all the 50 million female customers, and all the 50 million male customers, calculate the average, and conclude the results.

```
df3 = df.loc[df["Gender"]=="F"]

result = df3["Purchase"].mean()

  #average of females
print("average purchase by females" , result)
```

```
average purchase by females 8808.37694283518
```

```
df4 = df.loc[df["Gender"]=="M"]
result1 = df4["Purchase"].mean()
print("average purchase by males" , result1)
# result1 = df4["Purchase"].mean()
#   # for males
# print("average purchase by males" , result1)
```

```
average purchase by males 9501.56640506399
```

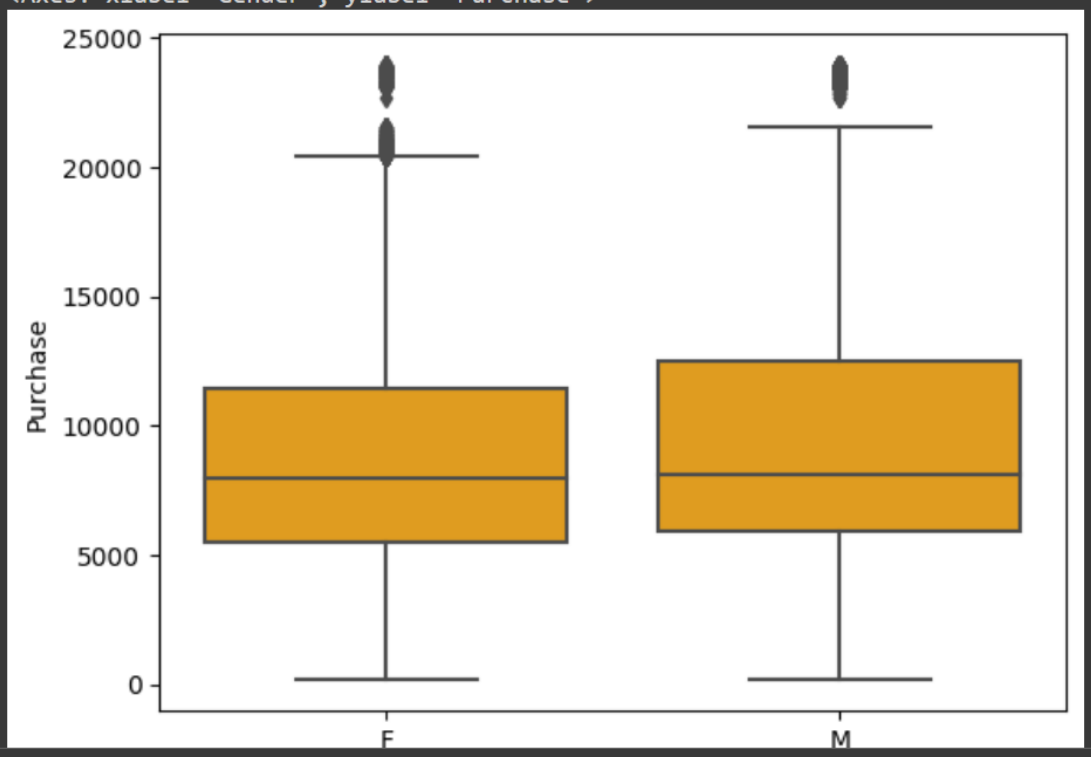
INFERENCE:

The average amount spent by female customers is 8808.37.. where as for males its 9501.56 . It clearly depicts that males spent more than females. Probably , males purchased expensive products .

More clearly from the below boxplot.. The highest amount spent by most of the female customers is 20000 with few outliers .. whereas as for males its slightly more.

```
sns.boxplot(data = df , x = "Gender" ,y = "Purchase" , color = "orange")
```

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



Allocating runtime

SAMPLE DISTRIBUTION ANALYSIS

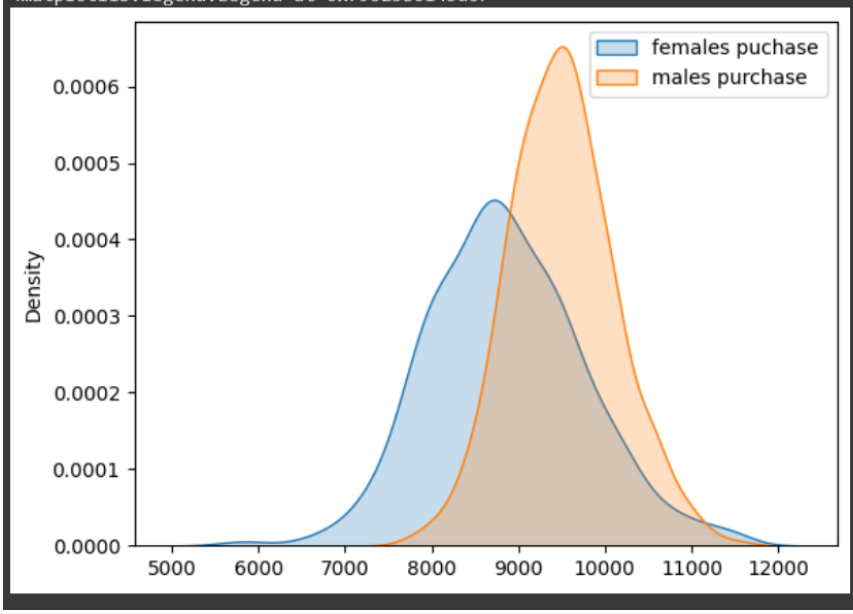
Using the sample average to find out an interval within which the population average will lie. Using the sample of female customers we will calculate the interval within which the average

spending of 50 million male and female customers may lie. We shall apply Central Limit Theorem to obtain sampling distribution as normal distribution and find the 95% confidence interval.

Using the below code taking 100 samples 500 times and obtain their mean and plotting the KDE plot respectively.

```
#  to find mean of random samples
l = []
for i in range(500):
    subset = df.sample(n=100)
    df1 = subset[subset["Gender"]=="F"]  #for female customers
    mu1 = df1["Purchase"].mean()
    l.append(mu1)
print("All average values of samples for female customers :",l)

l1 = []
for i in range(500):
    subset1 = df.sample(n=100)
    df2 = subset1[subset1["Gender"]=="M"]
    mu2 = df2["Purchase"].mean()
    l1.append(mu2)
print("All average values of samples for male customers :",l1)
sns.kdeplot(l, label = "females puchase",fill = True)
sns.kdeplot(l1, label = "males purchase", fill = True)
plt.legend()
```

```
All average values of samples for female customers : [8476.714285714286, 8732.52, 8882.772727272728, 7891.58333333
All average values of samples for male customers : [11034.154929577464, 11561.52564102564, 9827.825, 9915.95833333
<matplotlib.legend.Legend at 0x79523b0143d0>
```



INFERENCE:

- Above is the kde plot is the sampling distribution of sample means of both the genders .
- This works out only because of the Central limit theorem which says sampling distribution itself is Gaussian distributed.
- From the plot it can be estimated that the sample means for both genders follows a **normal distribution.**
- Also the sample mean from the plot is same as that of the population mean.
- Here SME(sample mean error) is population standard deviation /root(of no. sample)
- There is a strong overlapping in the plot. However the average amount spent by female customers is less than male customers.
- But there is a large population of females than males which can be visualised by the width of plots.
- Nearly 50% of sample female population has spent more than 8000.

TO FIND CONFIDENCE INTERVAL

For female customers

```
# 95% CONFIDENCE INTERVAL FOR PURCHASE MADE BY FEMALES
import scipy.stats
mu = df3['Purchase'].mean()  # as the population mean is same as the sampling distribution sample mean
sigma = df3['Purchase'].std()
SEM = scipy.stats.sem(df3['Purchase'])  # to calculate standard error
confidence = 0.95
ci = scipy.stats.norm.interval(confidence,
                               loc=mu,
                               scale=SEM)

  # for 95% ci
print("population mean :", mu)
print("population standard deviation :",sigma)
print("population standard mean error :", SEM)
print("95% confidence interval :", ci)
```

```
population mean : 8774.491872037916
population standard deviation : 4695.745295162423
population standard mean error : 22.85853044593786
95% confidence interval : (8729.689975624366, 8819.293768451465)
```

INFERENCE:

- Following the property of Gaussian distribution, 95% of values lie between [8729.68,8819.29].

- The interval [8729.68,8819.29].is 95% Confidence interval for the average spending of all the female transactions.

- In plain English, we can say that[8729.68,8819.29].covers 95% of the values of the average spending of all female customers.

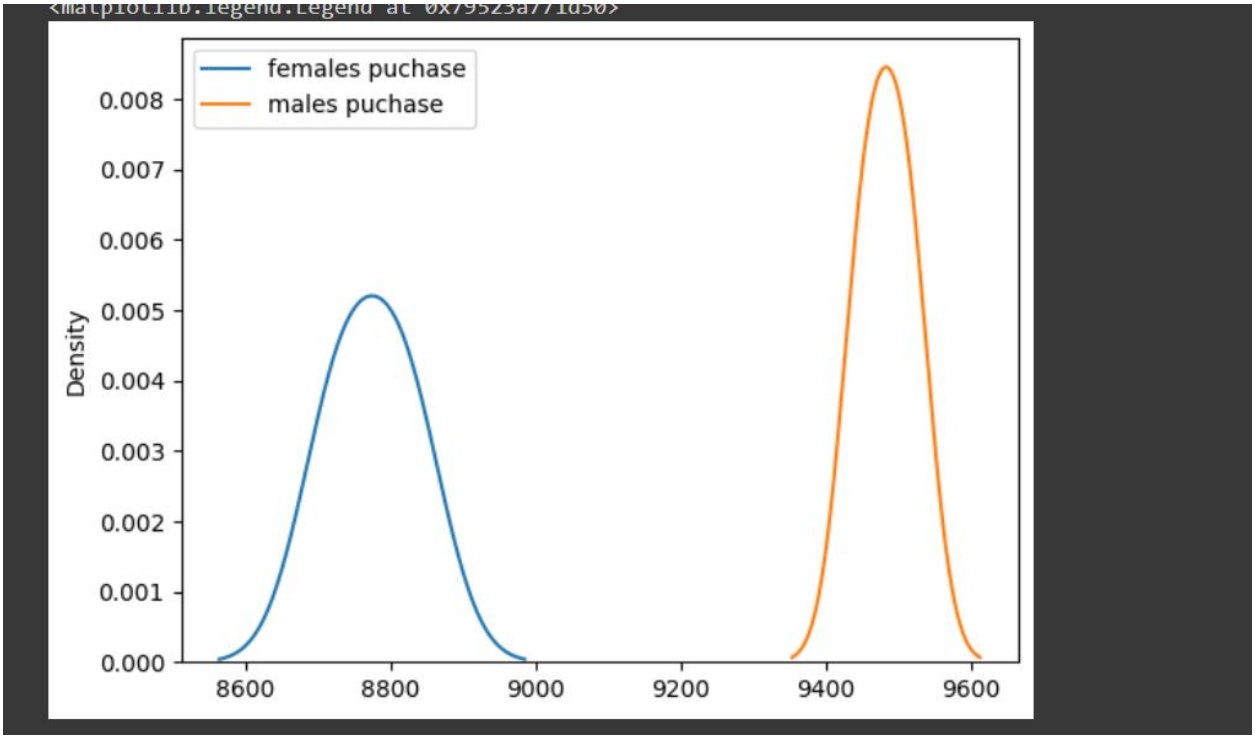
- Any value outside of this interval [8729.68,8819.29].occurs only 5% of the time. For example, the probability that the average spending of all female customers is 8200 is less than 0.05.

For male customers

```
# 95% CONFIDENCE INTERVAL FOR PURCHASE MADE BY MALES
import scipy.stats
mu2 = df4['Purchase'].mean()
sigma1 = df4['Purchase'].std()
sem1 = scipy.stats.sem(df4['Purchase'])  # to calculate standard error
confidence = 0.95
ci1 = scipy.stats.t.interval(confidence, len(df4['Purchase'])-1, loc=mu2, scale=sem1)

  # for 95% ci
print(mu2)
print(sigma1)
print(sem1)
print(ci1)

sns.kdeplot(ci ,label = "females puchase" )
sns.kdeplot(ci1 , label = "males puchase")
plt.legend()
```

```
9482.487524126223
5054.310624143144
14.071873570648895
(9454.906899968886, 9510.06814828356)
```

PLOTTING CONFIDENCE INTERVAL



INFERENCE:

for male customers the 95% CI IS 9454.90 AND 9510.06 .The population mean is 9501 which is within the interval. Same inference can be followed as of female customers .

From the above plots there is no overlapping in the confidence interval of both the genders. It implies statistical significance

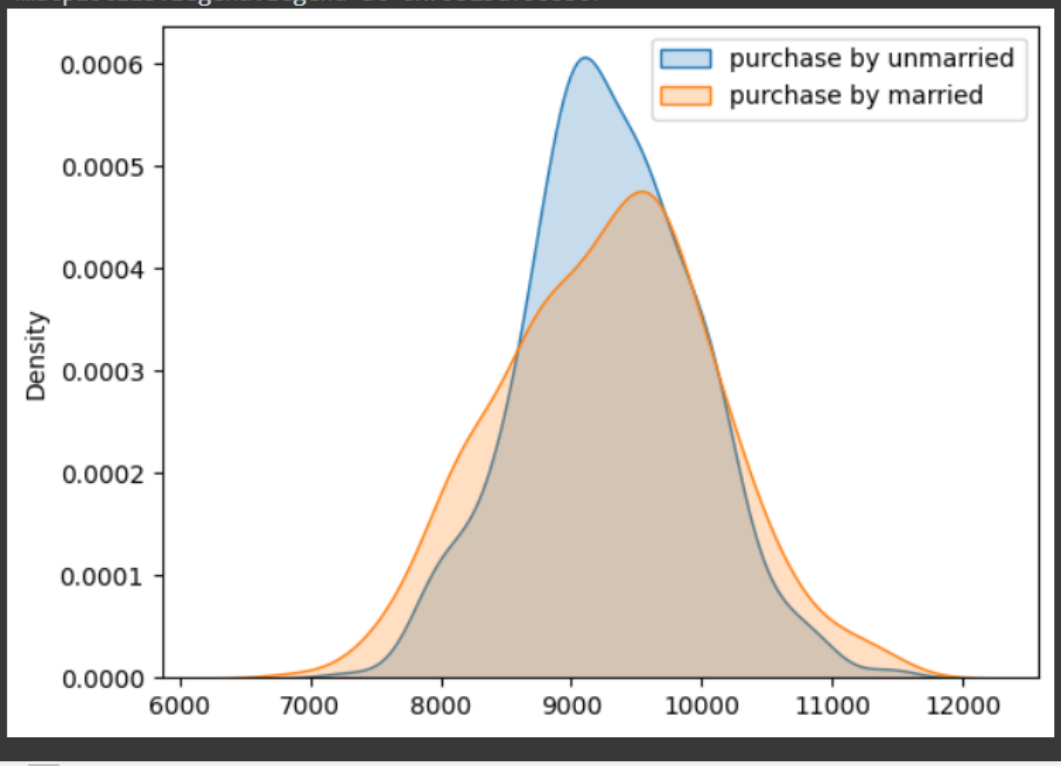
ANALYSIS DONE FOR MARRIED VS UNMARRIED CUSTOMERS

```
Average of purchase by  unmarried customers : 9296.875827420825
Average of purchase bymarried customers : 9323.91943721281
```

```
population mean of unmarried customers : 9296.875827420825

population standard deviation of unmarried customers : 4983.362638214957

population standard mean error of unmarried customers : 15.687093236435372

95% confidence intervalof unmarried customers: (9266.129320884753, 9327.622333956897)

population mean of married customers : 9323.91943721281

population standard deviation of married customers: 4969.505415604629

population standard mean error of married customers : 18.74377786952204

95% confidence interval of married customers : (9287.181675063455, 9360.657199362164)
```

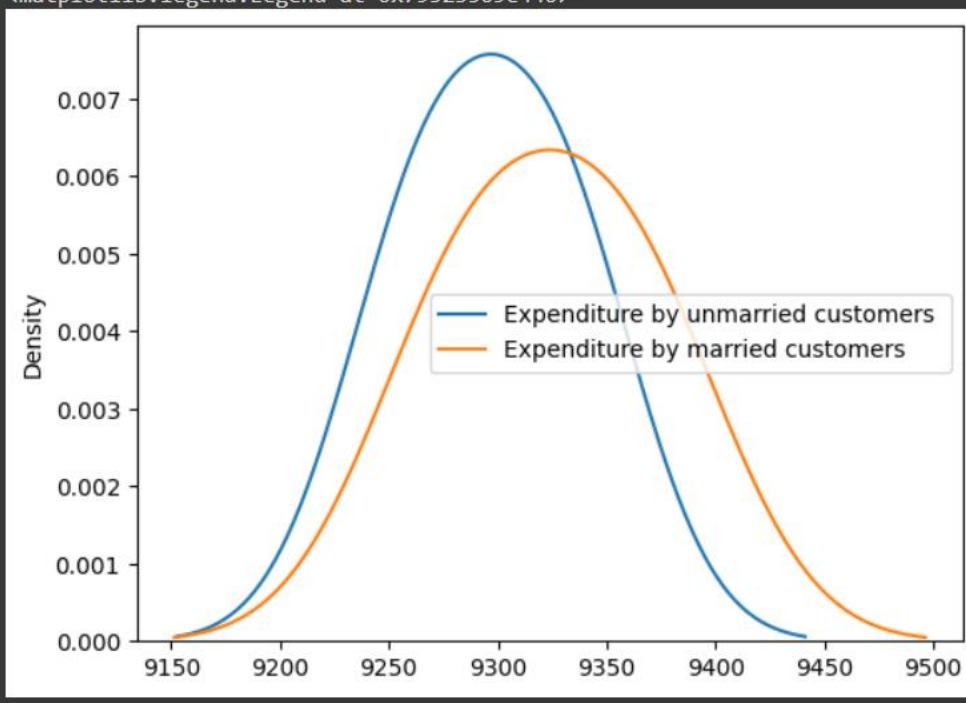
PLOTTING THE SAMPLING DISTRIBUTION MEAN FOR PURCHASE MADE BY MARRIED AND UNMARRIED CUSTOMERS

```
All average values of samples for unmarried customers : [8733.766666666666, 8790.553571428
All average values of samples for married customers : [8224.95744680851, 9256.457142857143
<matplotlib.legend.Legend at 0x79523af5bb50>
```



PLOTTING CONFIDENCE INTERVAL

```
<matplotlib.legend.Legend at 0x79523509e440>
```



INFERENCE

- For Both the categories the sampling distribution mean are normally distributed as per CLT.

- From the sampling distribution. The average purchase made by unmarried customers is more than married customers .
- Hence Confidence Interval is calculated and inferred that the population mean is with the intervals respectively.
- However both the CI are overlapping .. hence it is not statistically significant.

ANALYSIS DONE FOR CUSTOMERS WITH DIFFERENT AGE GROUPS

```
df[["Age","Purchase"]]
df1 = pd.DataFrame(df.groupby("Age")["Purchase"].mean().reset_index())  #average of each age group
df1
```

	Age	Purchase
0	0-17	9097.507300
1	18-25	9190.055329
2	26-35	9280.903252
3	36-45	9380.055199
4	46-50	9294.957209
5	51-55	9604.648294
6	55+	9419.929845

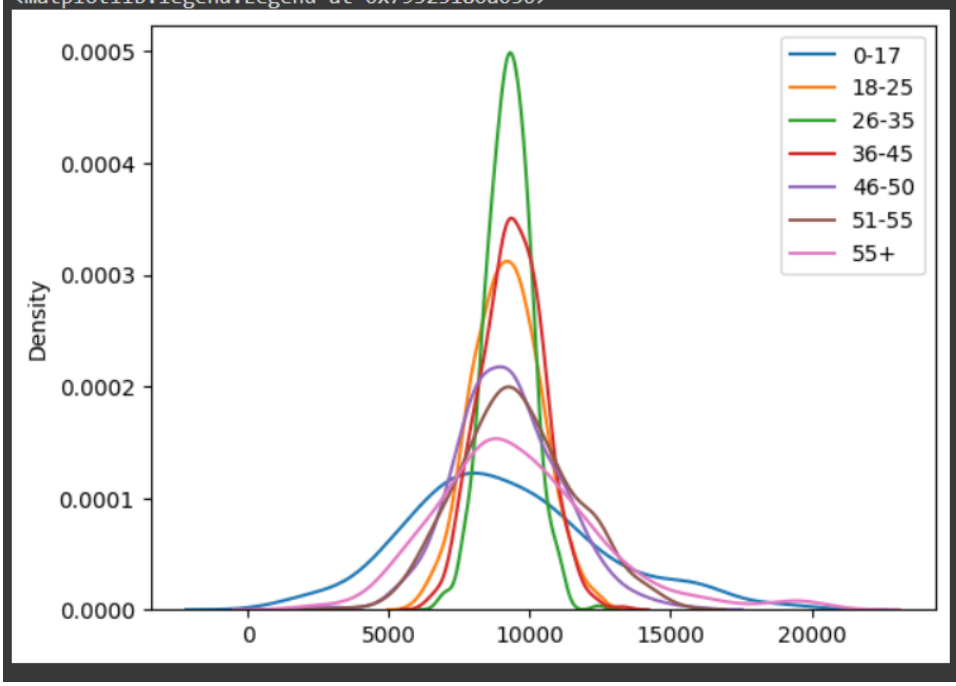
```
df["Age"].value_counts()
```

```
26-35    67929
36-45    33968
18-25    31629
46-50    14045
51-55    12047
55+       6728
0-17      4863
Name: Age, dtype: int64
```

find sampling distribution mean

PLOTTING THE SAMPLING DISTRIBUTION MEAN FOR DIFFERENT AGE GROUPS

```
for age group 36-45  : [8895.142857142857, 9446.777777777777, 7707.545454545455, 10605.380952380952, 1
for age group 46-50  : [9119.875, 11618.625, 8464.75, 8913.0, 9492.333333333334, 11000.666666666666, 7
for age group 51-55 : [10495.0, 4723.0, 9804.0, 7092.625, 10000.833333333334, 8064.75, 5479.75, 6701.5
for age group 55+  : [12629.333333333334, 6069.0, 10068.0, 11636.5, 12547.0, 12007.666666666666, 5061.
<matplotlib.legend.Legend at 0x79523186a050>
```



INFERENCE

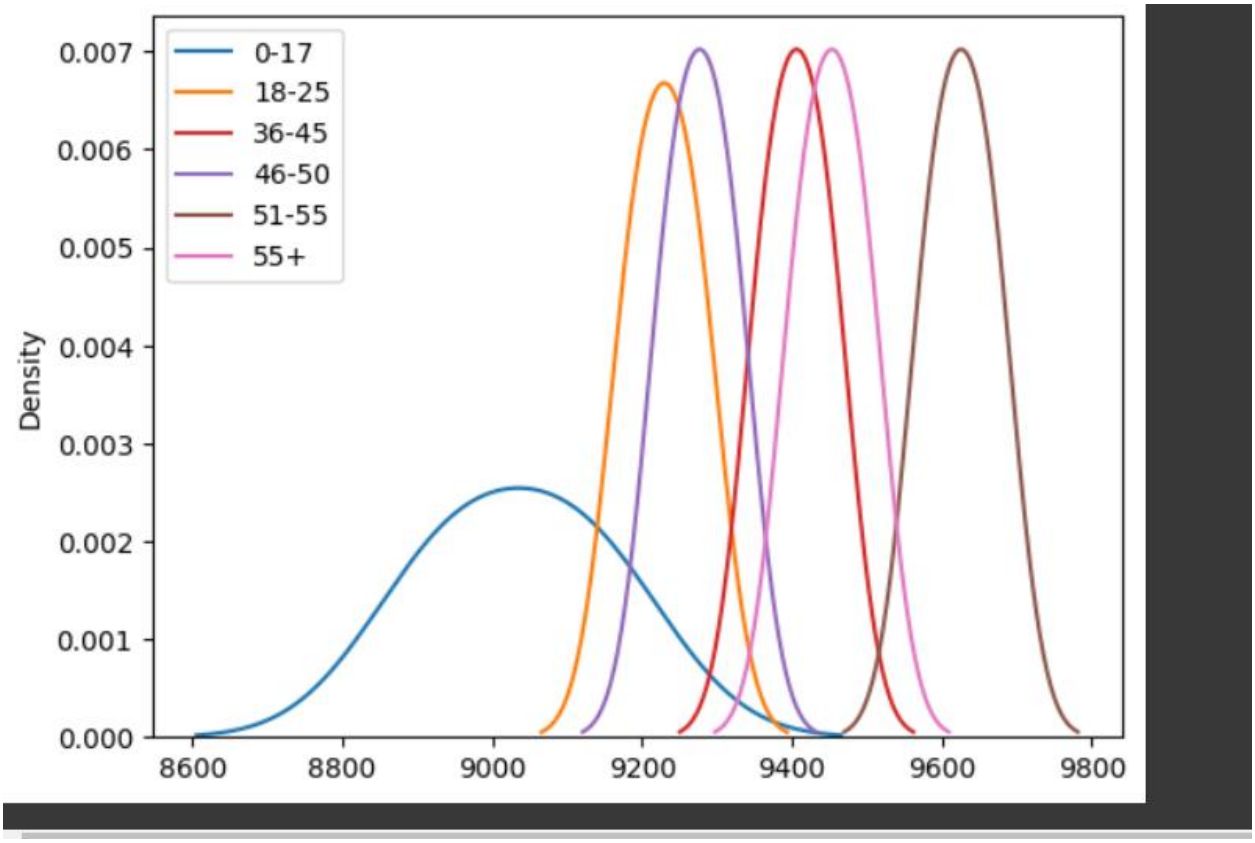
From the above plot of sample distribution as per clt. All the age groups are normally distributed.

Hence we can calculate the confidence interval for respective age groups.

Taking the population mean under consideration . the confidence interval for various age groups are :

```
confidence interval of 0-17 : (8943.721398830374, 9126.888767169625)
confidence interval of 18-25 : (9194.634958570758, 9264.405203429244)

confidence interval of 36-45 : (9372.878668309933, 9439.235067690066)

confidence interval of 46-50 : (9243.533124137693, 9309.890831862305)

confidence interval of 51-55 : (9592.24628615708, 9658.60440784292)

confidence interval of 55+ : (9419.988792505581, 9486.34903749442)
```

CONFIDENCE INTERVAL PLOTTED



INFERENCE:

Clearly all CI are overlapping with each other.

The difference between groups may not be statistically significant.

QUESTIONS AND ANSWERS

1. **Are women spending more money per transaction than men? Why or Why not?**

   Ans) By the above analysis the average expenditure made by women is less than men.

   ```
   average purchase by females 8732.438581163726
   ```

   ```
   average purchase by males 9411.924658265698
   ```

2. **Confidence intervals and distribution of the mean of the expenses by female and male customers**

```
FOR FEMALE CUSTOMERS:
population mean : 8732.438581163726
population standard deviation : 4644.747445457064
population standard mean error : 61.39743550716536
95% confidence interval : (8612.10181882656, 8852.77534350089)
```

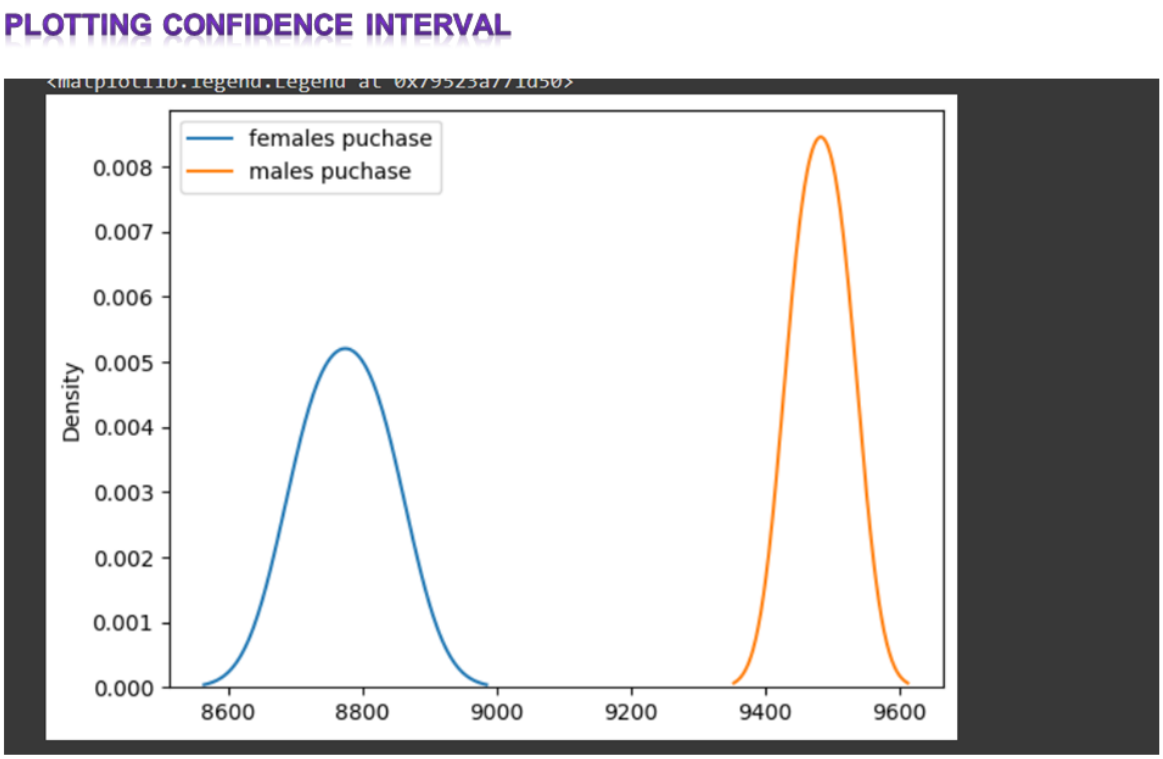
The 95% confidence interval: 8612.10 to 8852.77… indicating there are 95% of female customers who spend in average of the range.

3. FOR MALE CUSTOMERS

```
9482.487524126223
5054.310624143144
14.071873570648895
(9454.906899968886, 9510.06814828356)
```

The 95% confidence interval: 9454.90-9510.06… indicating there are 95% of male customers who spend in average of the range.

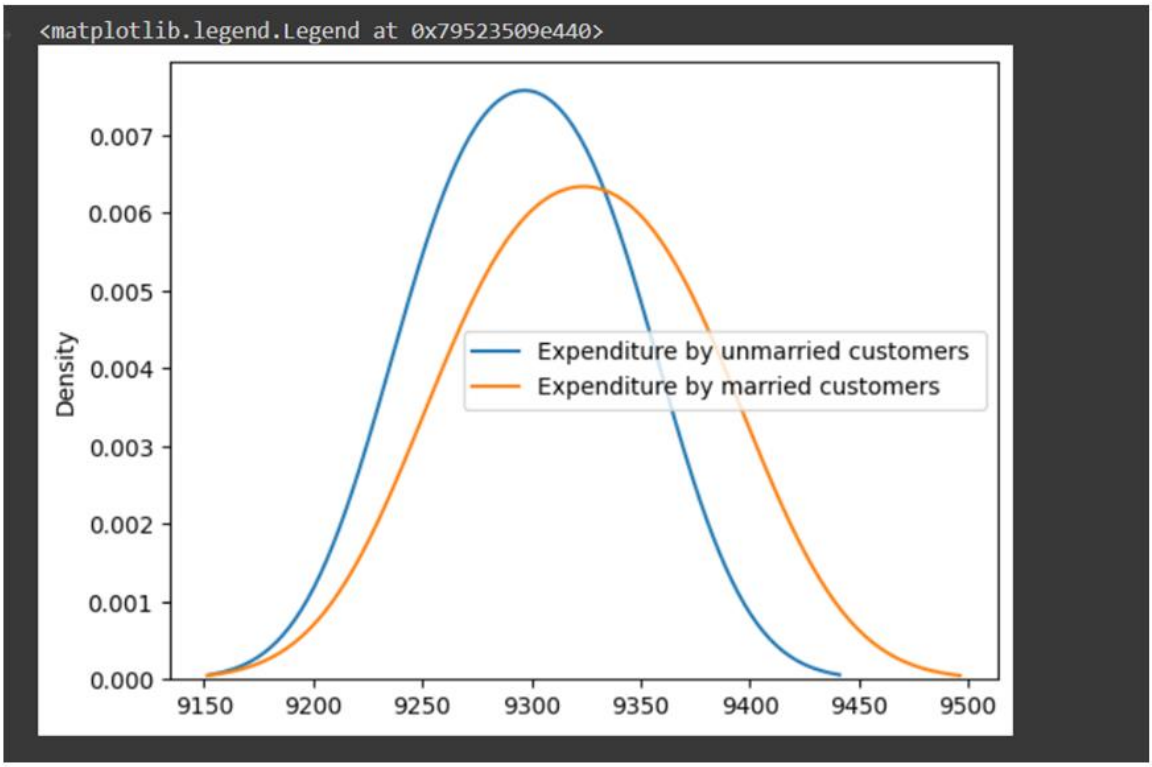
3. Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements?



Here confidence interval is not overlapping. Means the CI are statistically significant.

4. **Results when the same activity is performed for Married vs Unmarried**

PLOTTING CONFIDENCE INTERVAL



Ans:

Here both the intervals are overlapping . Hence , they are statistically insignificant.

For unmarried customers:

```
95% confidence intervalof unmarried customers: (9266.129320884753, 9327.622333956897)
```

```
Average of purchase by  unmarried customers : 9296.875827420825
Average of purchase bymarried customers : 9323.91943721281
```

```
  95% confidence interval of married customers : (9287.181675063455, 9360.657199362164)
```

SIMILARLY FOR AGE GROUPS

POPULATION MEAN

	Age	Purchase
0	0-17	9035.305083
1	18-25	9229.520081
2	26-35	9306.880289
3	36-45	9406.056868
4	46-50	9276.711978
5	51-55	9625.425347
6	55+	9453.168915

VARIOUS CONFIDENCE INTERVALS ARE CALCULATED ABOVE

BUSINESS INSIGHTS

- Based on the entire case study it can be inferred that women spend less than men in Walmart.
- People with age group 51-55 have spent the highest average amount. `confidence interval of 51-55 : (9592.24628615708, 9658.60440784292)`
- 95% of the customers with above age spent average amount of the above range.
- Teenagers tend to spend the least. As, they are minor and not economically stable.
- Most purchase was made in product category 5.
- Most customers are from age group 26-35 years. Also , the unmarried customers purchased more in Walmart.
- Hence , by considering the above observations , walmart can include the products related to female customers , may be more cosmetic brands , bag brands , apparels to attract more female customers.
- Also, as kids to accompany their moms, more toy brands, kids clothing's can too be introduced for little customers.
- Couple and family stuff too can be added such as home décor, electronics, utensils , bath ware to attract more married/ family customers.
- More seasonal offers can be introduced for better sales .
- Schemes Such as year-end sales, midyear sales , clear stock sales can be introduced to attract the customers for discounted price on the category which is least sold.

COLAB LINK

https://colab.research.google.com/drive/1ZL_s5HkDqmcY8QPWHIsIiFPnNhII13_Z?usp=sharing