Walmart Project:

Problem Statement:

Analyze the customer purchase behavior based on the criteria of Gender, age, marital status, occupation, and purchase amount of product.

Observations:

- Total number of data entries =550068
- No null entries in Dataset

```
wal.isnull().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
```

Attributes given in the data set:

User ID: User ID

Product_ID: Product ID

Gender: Sex of the User

Age: Age in bins

Occupation: Occupation (Masked)

City_Category: Category of the City(A,B,C)

Stay_In_Current_City_Years: Number of years stay in the current city

Marital_Status: 0 (Unmarried),1 (Married)

ProductCategory: Product Category (Masked)

Purchase: Purchase Amount

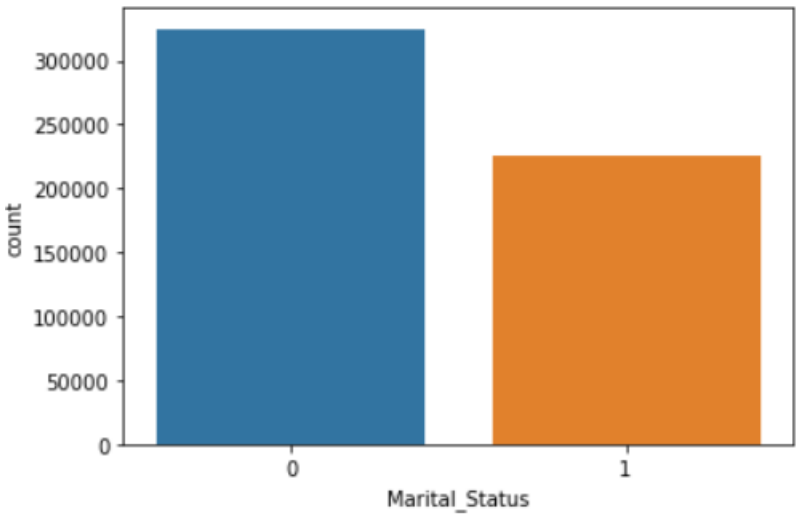
Numerical Attributes:

User_ID, Occupation, Marital_Status, Product_Category, Purchase

The descriptive analysis of numerical attributes is shown in the table below.

	User_ID	Occupation	Marital_Status	Product_Category	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000	550068.000000
mean	1.003029e+06	8.076707	0.409653	5.404270	9263.968713
std	1.727592e+03	6.522660	0.491770	3.936211	5023.065394
min	1.000001e+06	0.000000	0.000000	1.000000	12.000000
25%	1.001516e+06	2.000000	0.000000	1.000000	5823.000000
50%	1.003077e+06	7.000000	0.000000	5.000000	8047.000000
75%	1.004478e+06	14.000000	1.000000	8.000000	12054.000000
max	1.006040e+06	20.000000	1.000000	20.000000	23961.000000

- Marital Status can be categorized as Married(1) and Unmarried(0)
- 20 Product categories
- 20 Occupations



In Walmart customers Married people are fewer than unmarried 59.03% of unmarried customers.

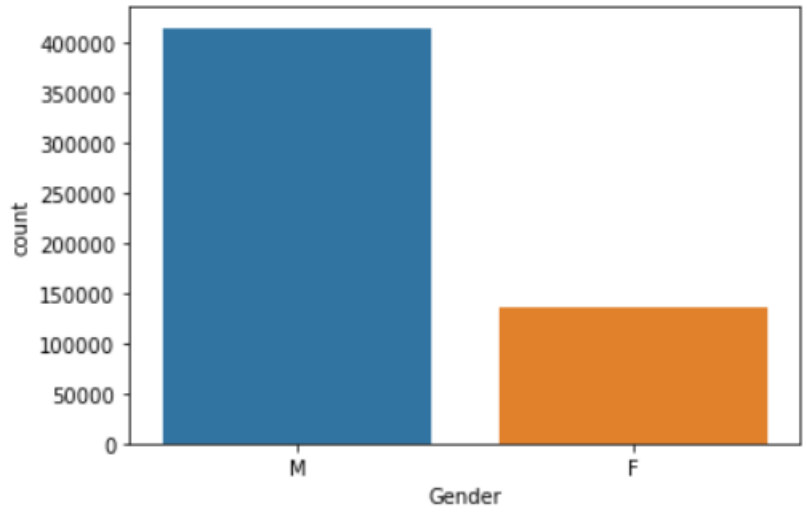
Categorical Attributes:

Categorial attributes are:
Gender, Age, , Stay_In_Current_City_Years, City_Category, Product_ID,

- **2 Gender categories: Male(M) and Female(F)**

```
wal["Gender"].value_counts()
```

```
M    414259
F    135809
Name: Gender, dtype: int64
```

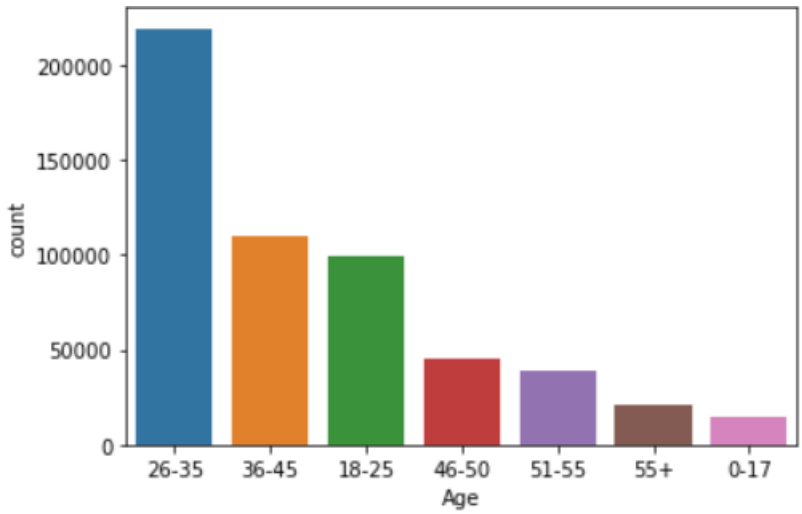


More male customers 75.31% Male

- **7 Age categories are present. 0-17, 18-25, 26-35, 36-45, 46-50 ,51-55 ,55+**

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

```
26-35    219587
36-45    110013
18-25     99660
46-50     45701
51-55     38501
55+       21504
0-17      15102
Name: Age, dtype: int64
```

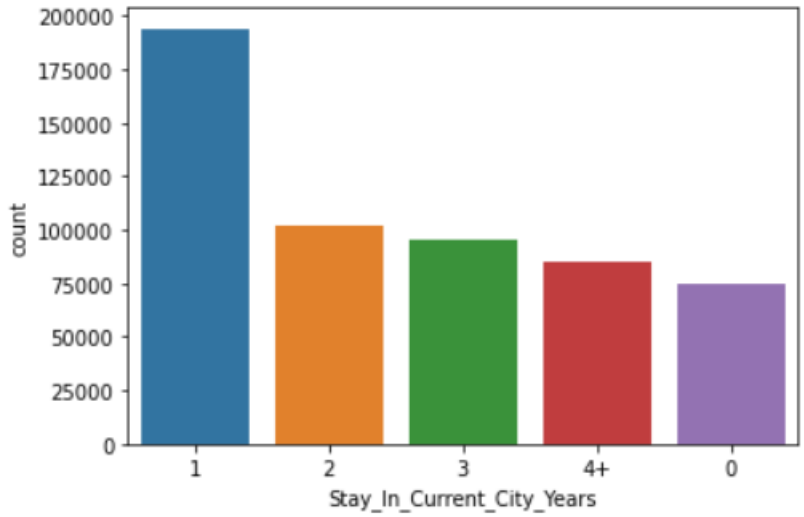


Highest number of people using Walmart are having age range 26 to 35 year .

- **5 categories are present. 0 years,1 year,2 years, 3 years, 4 years, 5years**

```
wal["Stay_In_Current_City_Years"].value_counts()
```

```
1     193821
2     101838
3      95285
4+     84726
0      74398
Name: Stay_In_Current_City_Years, dtype: int64
```

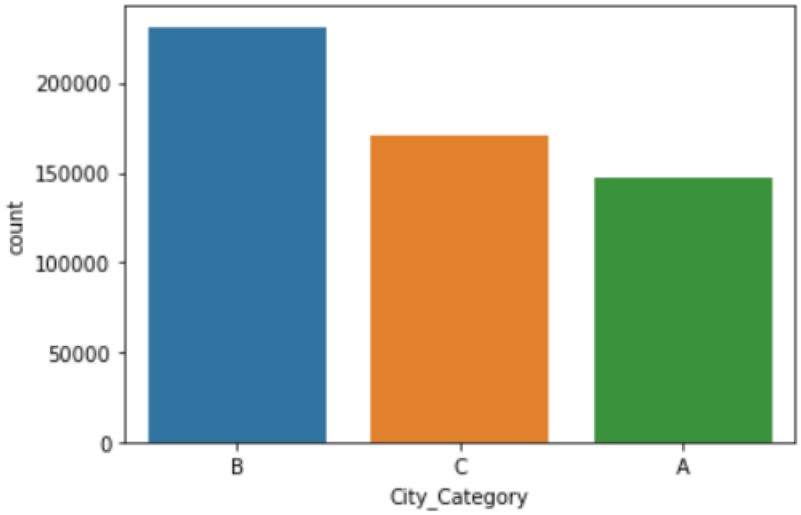


Customers who are using Walmart mostly 1 year of experience.

- **3 City categories A,B,C**

```
wal["City_Category"].value_counts()
```

```
B    231173
C    171175
A    147720
Name: City_Category, dtype: int64
```



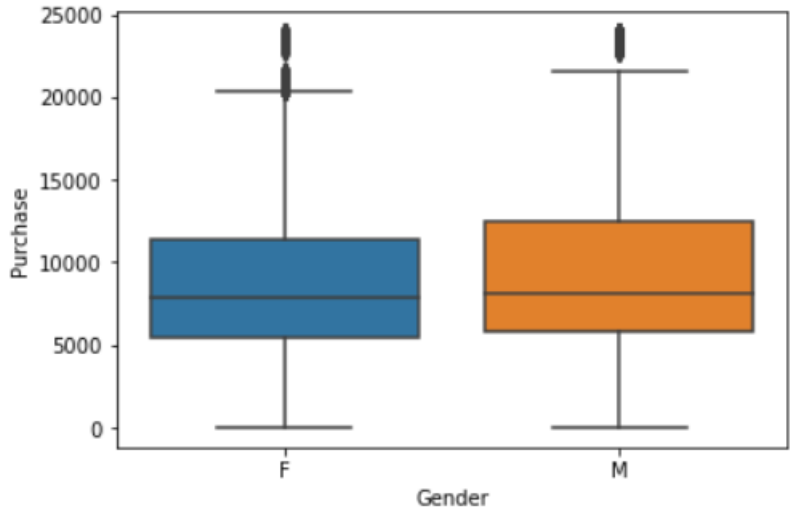
Most of the people are in **City Categories-B**

- 3631 Product IDs are present.

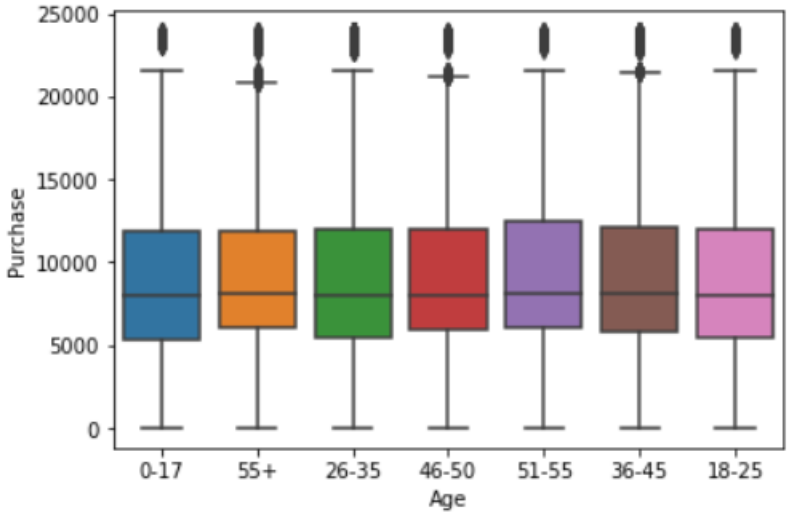
Bivariate analysis:

Purchase_Amount_Analysis

Purchase vs Gender:

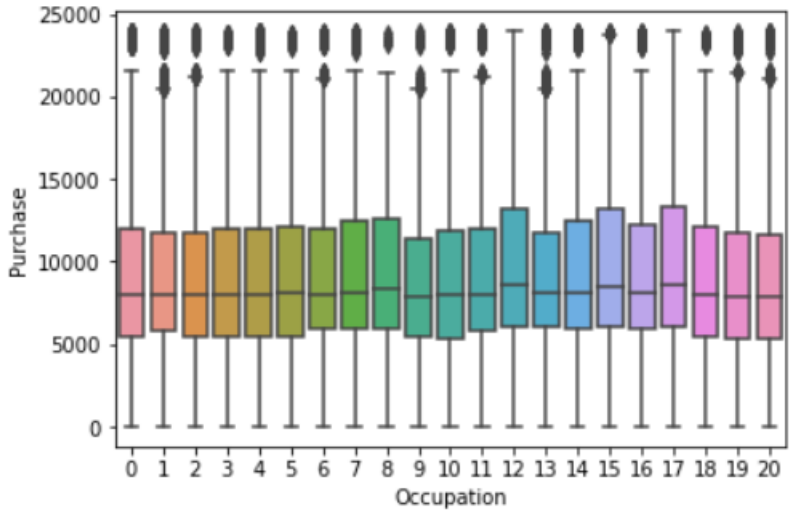


Purchase Vs Age-Range



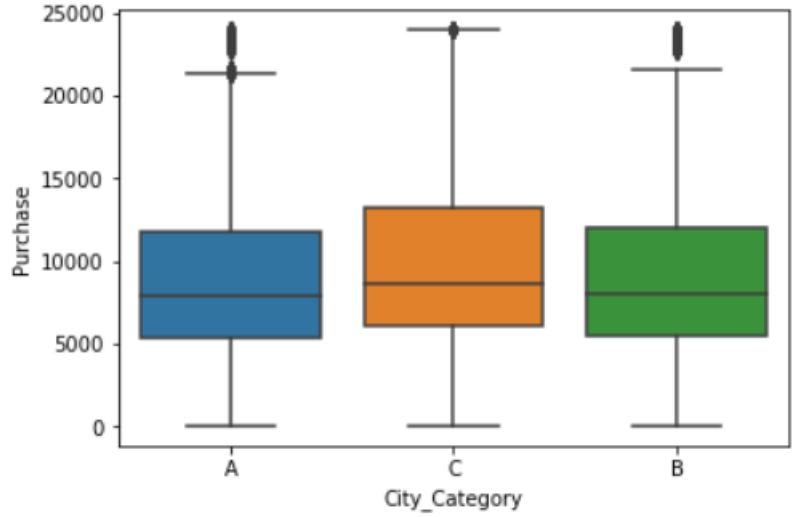
The average purchase is the same for all age ranges.

Purchase Vs Occupation



Purchase amount does not affect by occupation a lot

Purchase Vs City_Category



People of City_Category C has the highest purchase.

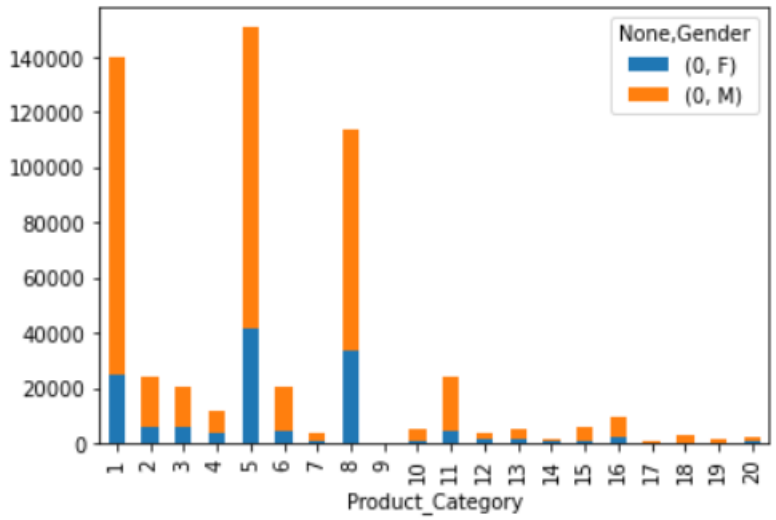
Product Category purchase analysis

Most selling product category is 5 .
% sale of each product category:

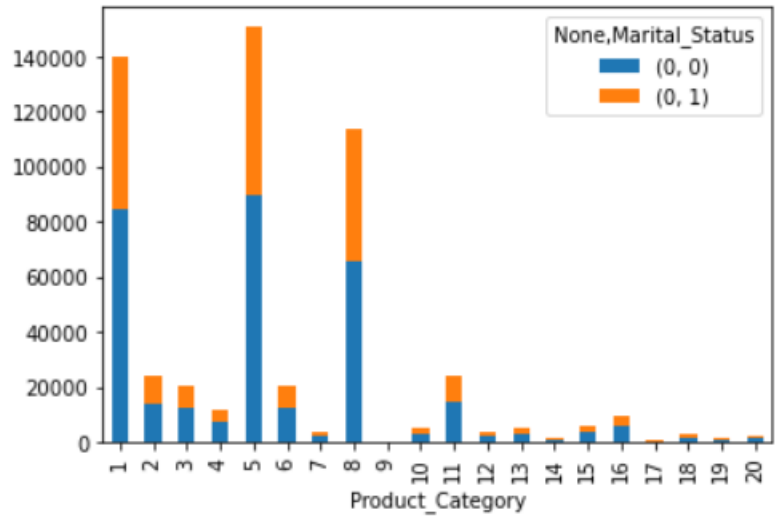
```
#% of most selling product category
wal["Product_Category"].value_counts(normalize=True)*100
```

```
5     27.438971
1     25.520118
8     20.711076
11     4.415272
2      4.338373
6      3.720631
3      3.674637
4      2.136645
16     1.786688
15     1.143495
13     1.008784
10     0.931703
12     0.717548
7      0.676462
18     0.568112
20     0.463579
19     0.291419
14     0.276875
17     0.105078
9      0.074536
Name: Product_Category, dtype: float64
```

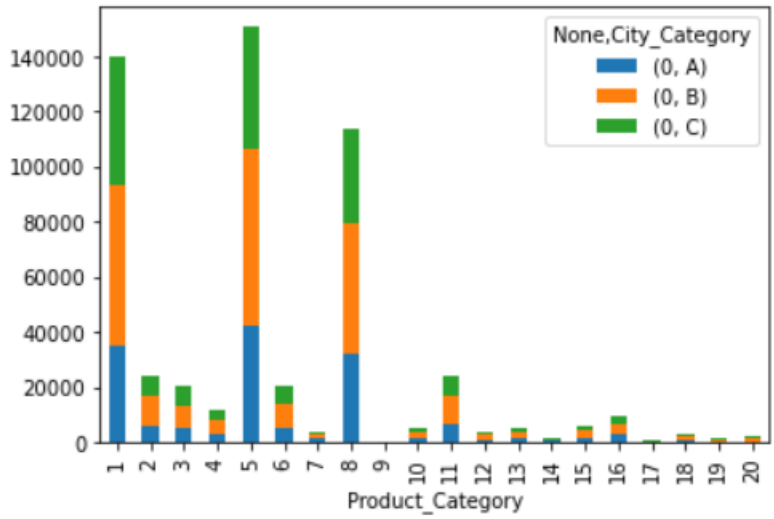
From above analysis it is clear that 73.7% of sale is from product categories1,5,8.



Mostly selling product categories are 1,5,8. And males & females both prefer these categories over others.



Most unmarried Walmart customers use categories 1,5,8.



From above plot it is clear that City_category B has most customers and again all city customers prefer product_ category1,5,8.

Joint Probabilities:

With respect to Product_Categories

Gender	F	M	All
Product_Category			
1	4.514169	21.005948	25.520118
2	1.028600	3.309773	4.338373
3	1.091865	2.582772	3.674637
4	0.661555	1.475090	2.136645
5	7.628330	19.810642	27.438971
6	0.828807	2.891824	3.720631
7	0.171433	0.505028	0.676462
8	6.100700	14.610375	20.711076
9	0.012726	0.061811	0.074536
10	0.211247	0.720456	0.931703
11	0.861530	3.553742	4.415272
12	0.278511	0.439037	0.717548
13	0.265785	0.742999	1.008784
14	0.113259	0.163616	0.276875
15	0.190158	0.953337	1.143495
16	0.436673	1.350015	1.786688
17	0.011271	0.093807	0.105078
18	0.069446	0.498666	0.568112
19	0.081990	0.209429	0.291419
20	0.131438	0.332141	0.463579
All	24.689493	75.310507	100.000000

- Probability of Males purchasing Product_Categories 1,5,8 are 21%,19%,14%.
- Females purchase category 5 most(7.6%) close to this it prefer Category 8 also (6.1%)

Age	0-17	18-25	26-35	36-45	46-50	51-55	55+	All
Product_Category								
1	0.651738	4.901576	10.589418	5.026288	1.904128	1.645069	0.801901	25.520118
2	0.146346	0.804991	1.623072	0.892981	0.382680	0.323778	0.164525	4.338373
3	0.218155	0.856258	1.392919	0.700641	0.250151	0.167979	0.088535	3.674637
4	0.137801	0.447763	0.762088	0.427947	0.179978	0.123257	0.057811	2.136645
5	0.787175	5.185177	11.175527	5.340612	2.176276	1.798505	0.975698	27.438971
6	0.072536	0.681552	1.542537	0.708821	0.294873	0.263604	0.156708	3.720631
7	0.009635	0.087444	0.300145	0.147073	0.059447	0.048358	0.024361	0.676462
8	0.410495	3.256143	8.045551	4.235113	1.937215	1.697972	1.128588	20.711076
9	0.002909	0.011453	0.027997	0.019452	0.005999	0.005272	0.001454	0.074536
10	0.020179	0.109623	0.324869	0.224518	0.094534	0.094352	0.063628	0.931703
11	0.134529	0.835715	1.795051	0.900434	0.382498	0.265058	0.101987	4.415272
12	0.022724	0.079808	0.199248	0.180705	0.094534	0.078718	0.061811	0.717548
13	0.020361	0.137438	0.381044	0.227245	0.100169	0.087807	0.054721	1.008784
14	0.007090	0.041813	0.102533	0.056720	0.027088	0.027997	0.013635	0.276875
15	0.029087	0.186159	0.431219	0.253605	0.109441	0.092352	0.041631	1.143495
16	0.041631	0.290510	0.748635	0.355411	0.159798	0.122167	0.068537	1.786688
17	0.001091	0.007454	0.023088	0.024542	0.017271	0.019452	0.012180	0.105078
18	0.004908	0.061629	0.189431	0.127621	0.063810	0.076900	0.043813	0.568112
19	0.010726	0.049994	0.102351	0.058175	0.027088	0.024361	0.018725	0.291419
20	0.016362	0.085262	0.163253	0.091989	0.041268	0.036359	0.029087	0.463579
All	2.745479	18.117760	39.919974	19.999891	8.308246	6.999316	3.909335	100.000000

- From the above analysis, it is clear that the age group of 26-35 purchases the most. (40%)
- And this age group prefers product category 1,5.
- Other than 26-35, the Age group 18-25 and the Age group 36-45 purchase 18% and 20% of products respectively.

Marital_Status	0	1	All
Product_Category			
1	15.339013	10.181105	25.520118
2	2.570228	1.768145	4.338373
3	2.246813	1.427823	3.674637
4	1.304748	0.831897	2.136645
5	16.299076	11.139895	27.438971
6	2.206818	1.513813	3.720631
7	0.370863	0.305599	0.676462
8	11.891439	8.819637	20.711076
9	0.044904	0.029633	0.074536
10	0.505028	0.426675	0.931703
11	2.666579	1.748693	4.415272
12	0.369772	0.347775	0.717548
13	0.574838	0.433946	1.008784
14	0.153799	0.123076	0.276875
15	0.658646	0.484849	1.143495
16	1.038599	0.748089	1.786688
17	0.054175	0.050903	0.105078
18	0.298327	0.269785	0.568112
19	0.171979	0.119440	0.291419
20	0.269058	0.194521	0.463579
All	59.034701	40.965299	100.000000

- From this analysis also it is clear that Product_category 1,5,8 are the most selling.
- And there is no significant purchase for any category is affected by their Marital Status.

City_Category	A	B	C	All
Product_Category				
1	6.377575	10.590145	8.552397	25.520118
2	1.116407	1.898674	1.323291	4.338373
3	0.898616	1.561080	1.214941	3.674637
4	0.554477	0.950064	0.632104	2.136645
5	7.673779	11.660013	8.105180	27.438971
6	1.001149	1.549990	1.169492	3.720631
7	0.222882	0.290691	0.162889	0.676462
8	5.850004	8.644931	6.216141	20.711076
9	0.019998	0.031632	0.022906	0.074536
10	0.242334	0.375045	0.314325	0.931703
11	1.200033	1.906128	1.309111	4.415272
12	0.193249	0.304508	0.219791	0.717548
13	0.293418	0.412858	0.302508	1.008784
14	0.087444	0.114895	0.074536	0.276875
15	0.312143	0.479577	0.351775	1.143495
16	0.517754	0.734091	0.534843	1.786688
17	0.021997	0.048539	0.034541	0.105078
18	0.136892	0.252514	0.178705	0.568112
19	0.049630	0.083990	0.157799	0.291419
20	0.085080	0.136892	0.241606	0.463579
All	26.854862	42.026259	31.118880	100.000000

- City_Category B has the most purchase (42%).
- No specific product_ category is preferred by any city.

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

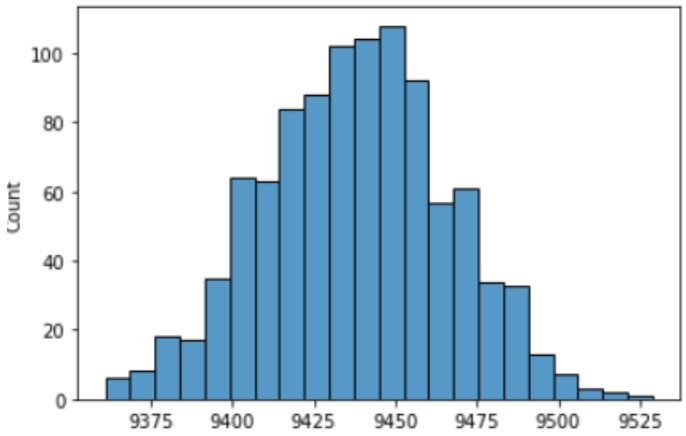


Fig: Histplot for males

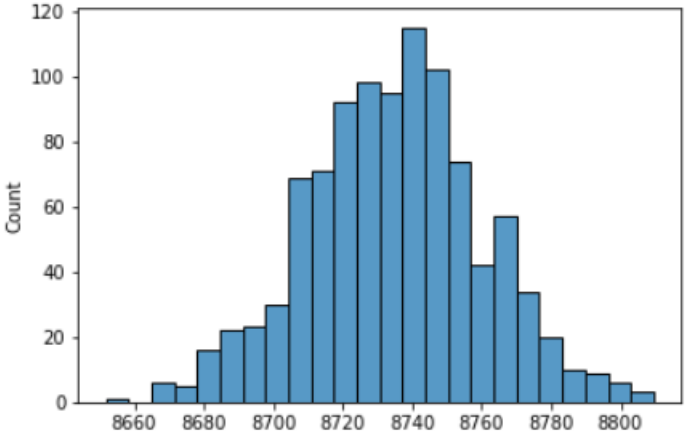


Fig: Histplot for females

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

For 90% confidence interval:

For Male:

```
from scipy.stats import norm
#for 90% confidence interval
z=norm.ppf(0.95)
x1=9437.52-z*std_err_male
x2=9437.52+z*std_err_male
x1,x2
```

```
(9389.161771649193, 9485.878228350808)
```

For Females:

```
n=30000
st_dev_female=np.std(wal_female["Purchase"])
std_err_female=st_dev_female/np.sqrt(n)
#for 90% confidence interval
z=norm.ppf(0.95)
x1=8734.80-z*std_err_female
x2=8734.80+z*std_err_female
x1,x2
```

```
(8689.527817089016, 8780.072182910983)
```

For 90% confidence interval range for male are spending more on Black Friday Sale.

Analysis for 95% Confidence interval:

```
# for 95% confidence interval
```

```
#for Males
n=30000
st_dev_male=np.std(wal_male["Purchase"])
std_err_male=st_dev_male/np.sqrt(n)
#for 90% confidence interval
z=norm.ppf(0.975)
x1=9437.52-z*std_err_male
x2=9437.52+z*std_err_male
x1,x2
```

```
(9379.897616846425, 9495.142383153576)
```

```
# for females
n=30000
st_dev_female=np.std(wal_female["Purchase"])
std_err_female=st_dev_female/np.sqrt(n)
#for 90% confidence interval
z=norm.ppf(0.975)
x1=8734.80-z*std_err_female
x2=8734.80+z*std_err_female
x1,x2
```

```
(8680.854866795966, 8788.745133204033)
```

With the above analysis, we can say that Males spent more than females.

Analysis for 99% Confidence interval:

```
# for 99% confidence interval
```

```
#for Males
n=30000
st_dev_male=np.std(wal_male["Purchase"])
std_err_male=st_dev_male/np.sqrt(n)
#for 90% confidence interval
z=norm.ppf(0.995)
x1=9437.52-z*std_err_male
x2=9437.52+z*std_err_male
x1,x2
```

```
(9361.791351801328, 9513.248648198673)
```

```
# for females
n=30000
st_dev_female=np.std(wal_female["Purchase"])
std_err_female=st_dev_female/np.sqrt(n)
#for 90% confidence interval
z=norm.ppf(0.975)
x1=8734.80-z*std_err_female
x2=8734.80+z*std_err_female
x1,x2
```

```
(8680.854866795966, 8788.745133204033)
```

With 99% confidence, we can say that Male spent more than female.

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

Analysis for 90% Confidence interval:

Married Vs unmarried:

```
np.mean(wal_mar["Purchase"].sample(30000))
sample_mean_mar=[(np.mean(wal_mar["Purchase"].sample(30000))) for i in range(1000)]
n=30000
st_dev_mar=np.std(wal_mar["Purchase"])
std_err_mar=st_dev_mar/np.sqrt(n)
#for 90% confidence interval
mean_mar=np.mean(wal_mar["Purchase"])
z=norm.ppf(0.95)
x1=mean_mar-z*std_err_mar
x2=mean_mar+z*std_err_mar
x1,x2
```

```
(9213.531378505573, 9308.817769659174)
```

```
np.mean(wal_unmar["Purchase"].sample(30000))
sample_mean_unmar=[(np.mean(wal_unmar["Purchase"].sample(30000))) for i in range(1000)]
n=30000
st_dev_unmar=np.std(wal_unmar["Purchase"])
std_err_unmar=st_dev_unmar/np.sqrt(n)
mean_unmar=np.mean(wal_unmar["Purchase"])
#for 90% confidence interval
z=norm.ppf(0.95)
x1=mean_unmar-z*std_err_unmar
x2=mean_unmar+z*std_err_unmar
x1,x2
```

```
(9218.165147449665, 9313.650090393348)
```

Fig: married data on top and unmarried on bottom

From analysis, it is clear that for a 90% confidence interval the purchase range of married customer is slightly less than unmarried.

Analysis for 95% Confidence interval:

```
np.mean(wal_mar["Purchase"].sample(30000))
sample_mean_mar=[(np.mean(wal_mar["Purchase"].sample(30000))) for i in range(1000)]
n=30000
st_dev_mar=np.std(wal_mar["Purchase"])
std_err_mar=st_dev_mar/np.sqrt(n)
#for 95% confidence interval
mean_mar=np.mean(wal_mar["Purchase"])
z=norm.ppf(0.975)
x1=mean_mar-z*std_err_mar
x2=mean_mar+z*std_err_mar
x1,x2
```

```
(9204.404205030149, 9317.944943134598)
```

```
np.mean(wal_unmar["Purchase"].sample(30000))
sample_mean_unmar=[(np.mean(wal_unmar["Purchase"].sample(30000))) for i in range(1000)]
n=30000
st_dev_unmar=np.std(wal_unmar["Purchase"])
std_err_unmar=st_dev_unmar/np.sqrt(n)
mean_unmar=np.mean(wal_unmar["Purchase"])
#for 95% confidence interval
z=norm.ppf(0.975)
x1=mean_unmar-z*std_err_unmar
x2=mean_unmar+z*std_err_unmar
x1,x2
```

```
(9209.018955344087, 9322.796282498926)
```

Fig: married data on top and unmarried on bottom

From analysis, it is clear that for a 95% confidence interval the purchase range of married customer is slightly less than unmarried

Analysis for 99% Confidence interval:

```
np.mean(wal_mar["Purchase"].sample(30000))
sample_mean_mar=[(np.mean(wal_mar["Purchase"].sample(30000))) for i in range(1000)]
n=30000
st_dev_mar=np.std(wal_mar["Purchase"])
std_err_mar=st_dev_mar/np.sqrt(n)
#for 99% confidence interval
mean_mar=np.mean(wal_mar["Purchase"])
z=norm.ppf(0.995)
x1=mean_mar-z*std_err_mar
x2=mean_mar+z*std_err_mar
x1,x2
```

```
(9186.565662218964, 9335.783485945783)
```

```
np.mean(wal_unmar["Purchase"].sample(30000))
sample_mean_unmar=[(np.mean(wal_unmar["Purchase"].sample(30000))) for i in range(1000)]
n=30000
st_dev_unmar=np.std(wal_unmar["Purchase"])
std_err_unmar=st_dev_unmar/np.sqrt(n)
mean_unmar=np.mean(wal_unmar["Purchase"])
#for 99% confidence interval
z=norm.ppf(0.995)
x1=mean_unmar-z*std_err_unmar
x2=mean_unmar+z*std_err_unmar
x1,x2
```

```
(9191.143241699116, 9340.671996143898)
```

Fig: married data on top and unmarried on bottom

From analysis, it is clear that for a 90% confidence interval the purchase range of married customer is slightly less than unmarried

5. Results when the same activity is performed for Age

Analysis for 90% Confidence interval:

Age range:
0-17 years: (8779.978974850774, 9086.950306039174)
18-25 years: (9018.47974249923, 9320.847470023347)
26-35 years: (9205.10793376461, 9300.273331975166)
36-50 years: (9247.78320900821, 9342.880276612863)
50+ years: (9415.598420431128, 9511.724935955841)

50+ year people purchase most with 90% interval.

Analysis for 95% Confidence interval:

0-17 years: (8750.575189506208, 9116.35409138374)
18-25 years:(9112.696222544599, 9226.630989977979)
26-35 years: (9195.99234980965, 9309.388915930125)
36-50 years: (9238.674170219734, 9351.989315401339)
50+ years: (9406.390774174824, 9520.932582212145)

50+ year people purchase most with 95% interval also.

Analysis for 99% Confidence interval:

0-17 years: (8693.107159004521, 9173.822121885427)
18-25 years: (9094.795773239142, 9244.531439283435)
26-35 years: (9178.176458058791, 9327.204807680984)
36-50 years: (9220.87107062549, 9369.792414995583)
50+ years: (9388.394951878296, 9538.928404508673)

50+ year people purchase most with 90% interval.

Recommendation:

1. Product Categories 1,5,8 are the most selling categories.
2. Males purchase more.
3. Unmarried purchase more.
4. 50+ year people purchase more.