Walmart

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.

Business Problem

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.

Observations:

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

```
[89] data= pd.read_csv("https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/293/original/walmart_data.csv?1641285094")
     data.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

```
[90] data.shape
```

```
(550068, 10)
```

```
[91] data.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
```

We can observe that the columns User_ID, Occupation, Marital_Status , Product_Category are of integer data type, which we can convert them into object type.

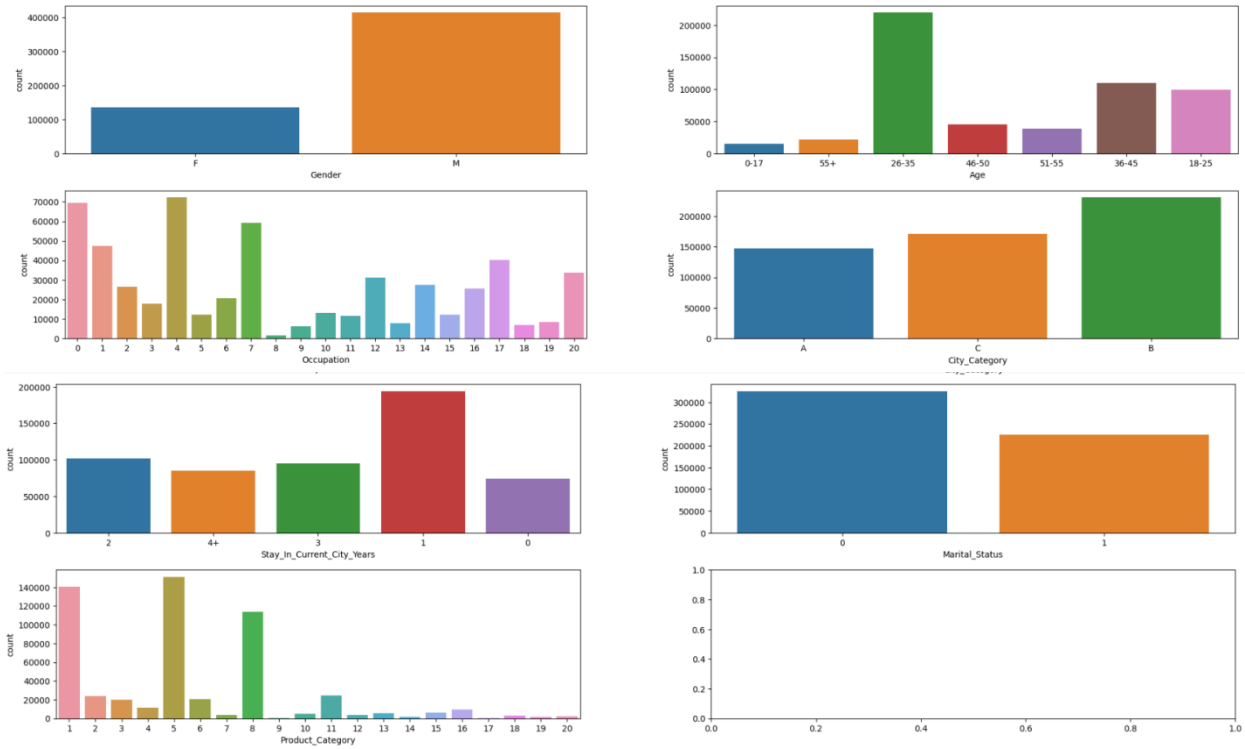
```
data.info()
```

```
<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  object
 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
dtypes: int64(1), object(9)
memory usage: 42.0+ MB
```

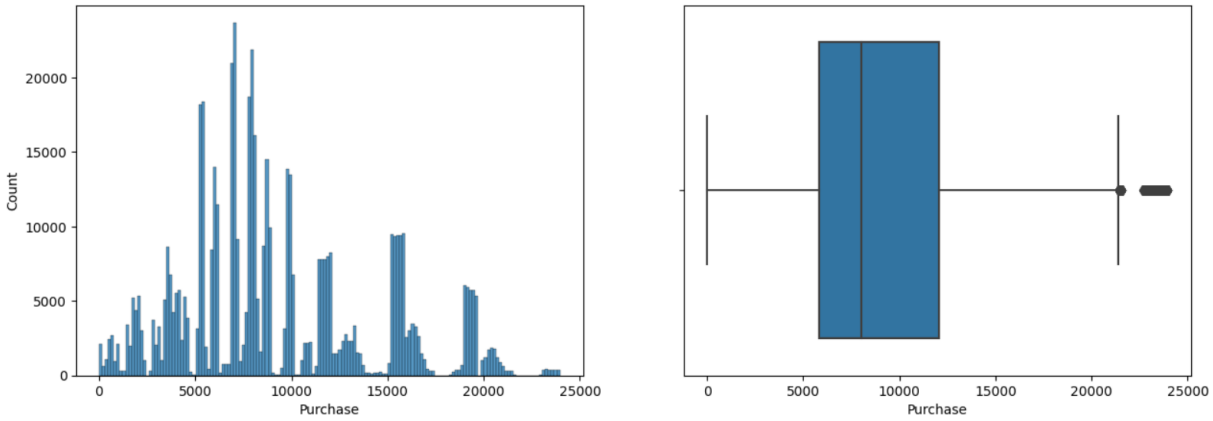
```
[153] data.describe(include= "all")
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
count	550068.0	550068	550068	550068	550068.0	550068	550068	550068.0	550068.0	550068.000000
unique	5891.0	3631	2	7	21.0	3	5	2.0	20.0	NaN
top	1001680.0	P00265242	M	26-35	4.0	B	1	0.0	5.0	NaN
freq	1026.0	1880	414259	219587	72308.0	231173	193821	324731.0	150933.0	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	9263.968713
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5023.065394
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	12.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5823.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8047.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	12054.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	23961.000000

- There are no null values in the dataset.
- There are total 5891 unique user id available in the dataset with user id 1001680 having the highest number of count.
- The product id has 3631 unique values and the from the dataset we can get that the product id P00265242 is the most purchased product.
- Most of the customers are male.
- There are 7 age groups and most of the customers belong to the age group 26 to 35.
- The customers are from 3 cities and most of the customers belong to city B.
- There are 5 categories in the number of years stay in current city and most of the customers have mentioned that they are staying in their current city for about an year.
- The customers who have purchased belongs to 21 distinct occupation with Occupation 4 being the highest.

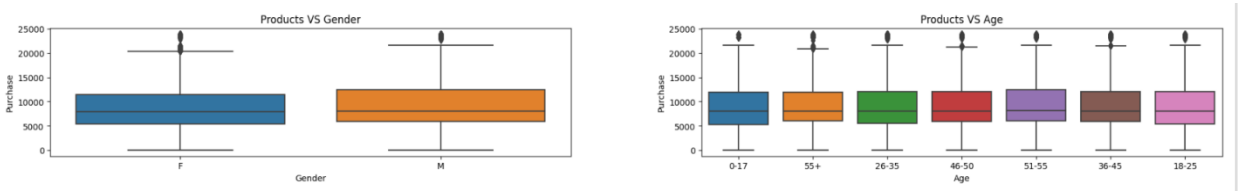


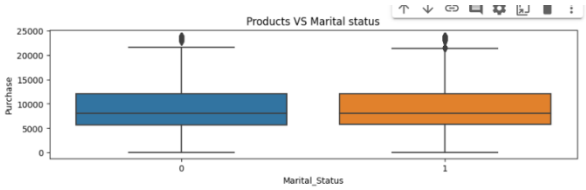
<Axes: xlabel='Purchase'>

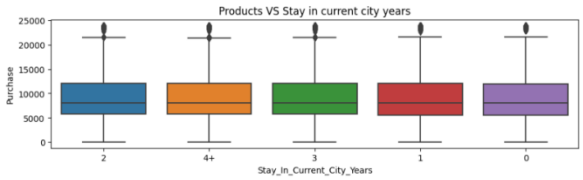


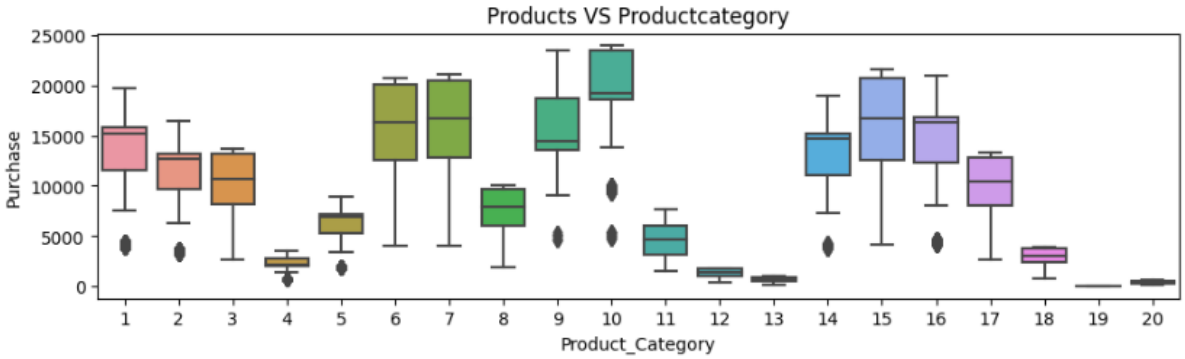
From the above graphs we can observe that

- Around 76% of purchases are made by male customers and 24% purchases are made by female customers.
- Around 40% of the customers who have made purchases are in ages between 26 to 35 years and around 20% of purchases are made by customers of age between 36 to 40 years.
- Around 12% of the customers who have made purchases belong to the occupation 4 and 0.
- Around 42% of the customers belong to city B, 31% belong to city C and 26% belong to city A.
- Around 59% of customers are unmarried and 41% of customers are married.
- Around 35% of the customers have stated that they are staying from 1 year in their current city.
- There are totally 20 product categories and the product categories 5, 1, 8 are most purchased products.
- From the histogram and boxplot we can observe that most of the purchase amount belong to the range 5000 to 10000, with an average amount of 9264. Also we can observe that there are outliers in the data.





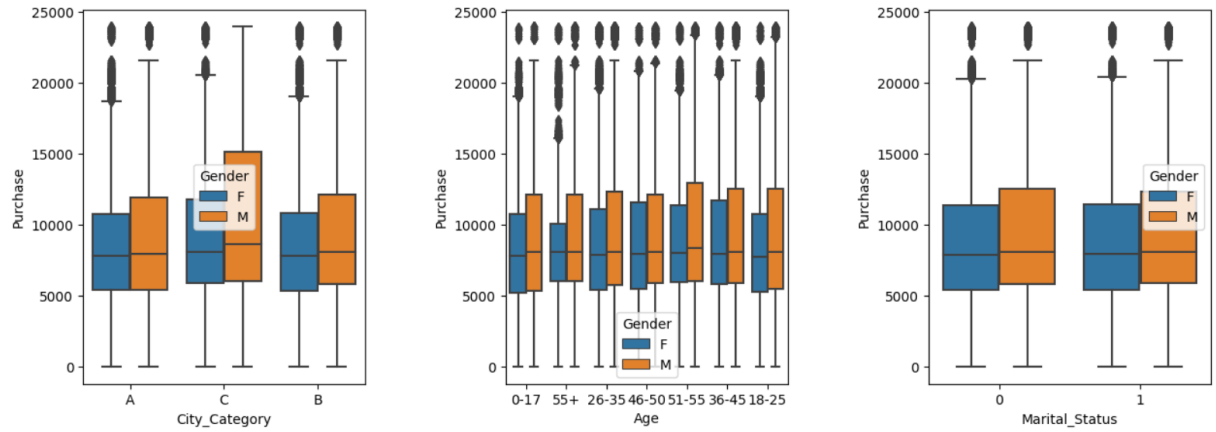




From the above Bi-variant graphs plotted against the purchase amount, we can observe:

- The spending behavior for males and females are. however the average purchase amount of male are little higher than females.
- Among different age categories, we observe similar purchasing patterns. For all age groups, most of the purchase amount fall between 5k to 12k.
- Among different occupation as well, we see similar purchasing behavior in terms of the purchase amount.
- Similarly for City category, stay in current city years, marital status - we see the users spends mostly in the range of 5k to 12k.
- We can observe variations in the purchase amount among different product categories.

<Axes: xlabel='City_Category', ylabel='Purchase'>



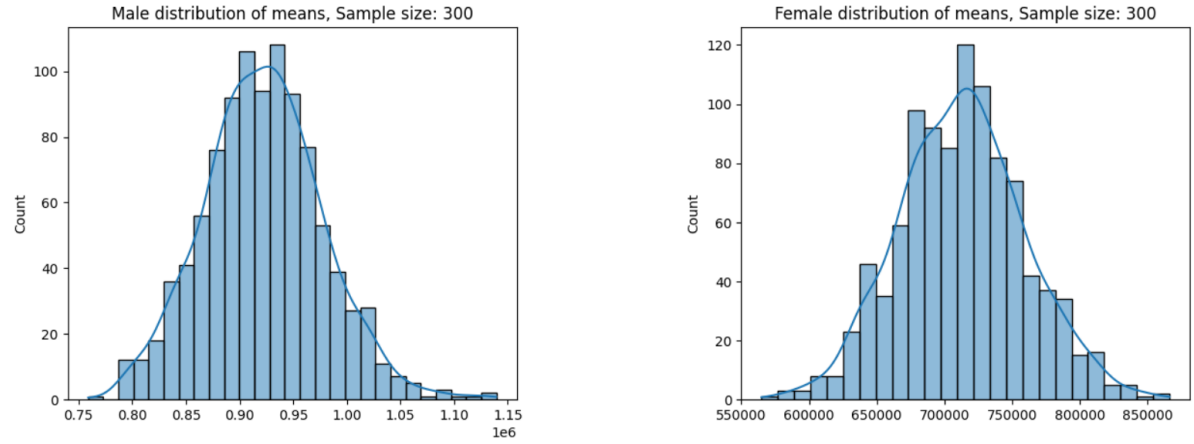
From the above graphs we can observe the following:

- The spending habits of male and female customers from all the cities have somewhat have similar pattern and males are spending little more when compared to female customers.
- The spending habits of male and female customers of all age group also have the same behavior and here also male customer are spending little more when compared to female customers.
- The spending habits of married and unmarried male and female customers have same pattern and male customer are spending little more when compared to female customers

Confidence Interval of average expense:

- By Gender:

Text(0.5, 1.0, 'Female distribution of means, Sample size: 300')



The average male population expense is found to be approximately 922198.49 and we can conclude that

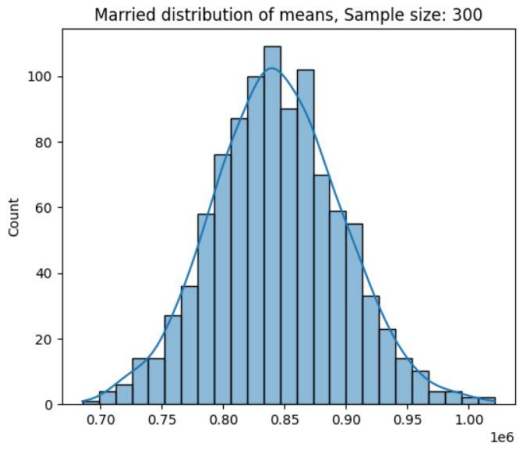
- At 90% confidence the true population average falls within the range 916954.44 to 927442.54.
- At 95% confidence the true population average falls within the range 915950.26 to 928446.72.
- At 99% confidence the true population average falls within the range 914015.22 to 930381.76.

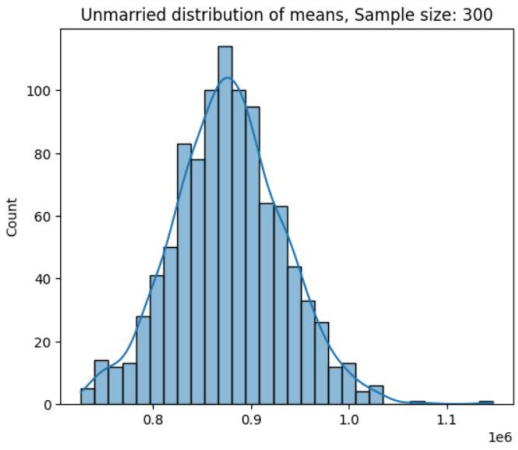
The average female population expense is found to be approximately 712928.06 and we can conclude that

- At 90% confidence the true population average falls within the range 708524.12 to 717332.0.
- At 95% confidence the true population average falls within the range 707680.81 to 718175.31.
- At 99% confidence the true population average falls within the range 706055.77 to 719800.35.

From the above we can get that the average male and female expenditure are not overlapping and thus we can conclude that the average male customer expenditure is more than that of female customer.

By Marital status:





The average married population expense is found to be approximately 844023.15 and we can conclude that

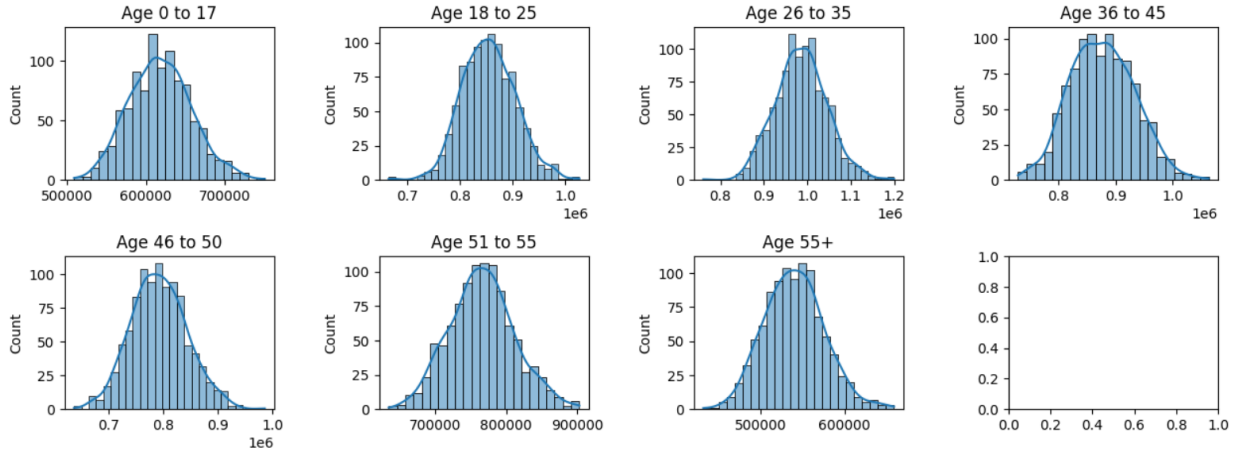
- At 90% confidence the true population average falls within the range 839047.82 to 848998.48.
- At 95% confidence the true population average falls within the range 838095.1, 849951.2.
- At 99% confidence the true population average falls within the range836259.22, 851787.08.

The average unmarried population expense is found to be approximately 876984.19 and we can conclude that

- At 90% confidence the true population average falls within the range 871653.8 to 882314.58.
- At 95% confidence the true population average falls within the range 870633.09, 883335.29.
- At 99% confidence the true population average falls within the range 868666.19 to 885302.19.

From the above we can see that the average expenditure of married and unmarried customers are overlapping and we can conclude that there is no much difference in the expenditure of married and unmarried customers.

By Age:



The average population expense of customer of age 0 to 17 is found to be approximately 618411.27 and we can conclude that

- At 90% confidence the true population average falls within the range 614676.3, 622146.24.
- At 95% confidence the true population average falls within the range 613961.09, 622861.45.
- At 99% confidence the true population average falls within the range 612582.9, 624239.64.

The average population expense of customers of age between 18 to 25 years is found to be approximately 855635.1 and we can conclude that

- At 90% confidence the true population average falls within the range 850891.23, 860378.97.
- At 95% confidence the true population average falls within the range 849982.83, 861287.37

- At 99% confidence the true population average falls within the range 848232.36, 863037.84

The average population expense of customer of age between 26 to 35 years is found to be approximately 985691.68 and we can conclude that

- At 90% confidence the true population average falls within the range 980102.63, 991280.73
- At 95% confidence the true population average falls within the range 979032.38, 992350.98.
- At 99% confidence the true population average falls within the range 976970.04, 994413.32.

The average population expense of customer of age between 36 to 45 years is found to be approximately 876618.44 and we can conclude that

- At 90% confidence the true population average falls within the range 871304.04, 881932.84.
- At 95% confidence the true population average falls within the range 870286.39, 882950.49
- At 99% confidence the true population average falls within the range 868325.39, 884911.49

The average population expense of customer of age between 46 to 50 years is found to be approximately 792067.12 and we can conclude that

- At 90% confidence the true population average falls within the range 787230.28, 796903.96.
- At 95% confidence the true population average falls within the range 786304.08, 797830.16
- At 99% confidence the true population average falls within the range 784519.3, 799614.94

The average population expense of customer of age between 51 to 55 years is found to be approximately 766022.87 and we can conclude that

- At 90% confidence the true population average falls within the range 761675.99, 770369.75
- At 95% confidence the true population average falls within the range 760843.61, 771202.13
- At 99% confidence the true population average falls within the range 759239.63, 772806.11

The average population expense of customer of age above 55 years is found to be approximately 539988.38 and we can conclude that

- At 90% confidence the true population average falls within the range 536597.79, 543378.97
- At 95% confidence the true population average falls within the range 535948.53, 544028.23
- At 99% confidence the true population average falls within the range 534697.41, 545279.35

From the above we can get that the average expenditure of the customer of age 26-30 is the highest among other group ages and the customer of age above 55 and age between 0 to 17 have lowest average expenditure. The averages does not overlap.

Recommendations:

- From data set we can get that the male customers are spending more when compared to females, so the company should focus promoting gender specific promotions to retain the male customers and attract female customers.
- Also unmarried customer contribute more when compared to married customers, so the company should focus on retaining them and also should take measures to attract more married customers.
- Most of the customers are in the ages between 18 to 45 and should focus on attracting more customers from these age groups.
- Most of the customers are staying only from an year in the city, so the company should focus on the customers who are leaving more than an year which increases the repeat purchase.
- The products 5, 1, 8 are most purchased products and the company should focus on selling more of these and also the company should focus on selling the products 2, 3, 6, 11 that have potential to increase purchase.
- Along with city B the company should focus on other cities to increase the sale.
- The company can introduce loyalty schemes and other schemes providing special offers and discounts which can attract more customers.
- The company can provide different offers on products to different category of customers to increase the sale.