Ex No: 2	Exploratory Data Analysis (EDA) Using Python

AIM:

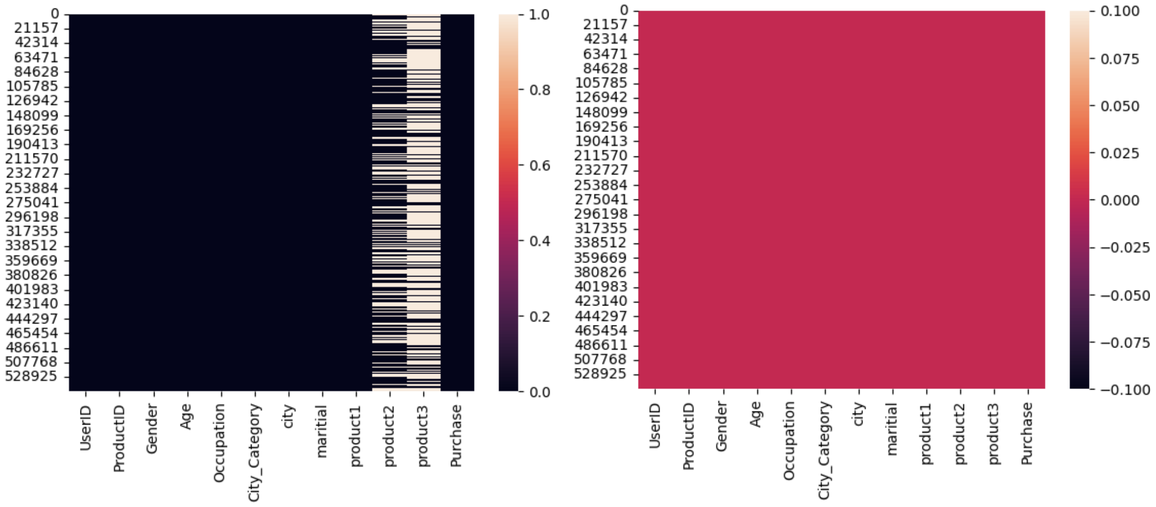
To perform Exploratory Data Analysis (EDA) on a dataset using python and generate insights.

PROCEDURE:

1. Import the required libraries and packages.
2. Download the sales dataset from Kaggle.
3. Load the dataset in Google Colab.
4. Get the summary and information about the dataset.
5. Count the data types of data available using the value_counts().
6. Get the shape of the dataset.
7. Preprocess the data
   - Rename the column names.
   - Handle null values.
   - Aggregate the ages into groups.
   - Change the marital status value in numerical to Boolean.
8. Analyse the data using various parameters in the dataset.
9. Create interactive and non – interactive graphs and bar charts accordingly.
10. Interpret the inferences obtained from the analysis and use it in decision-making.

EDA:

1) Heatmaps generated before and after handling null values:



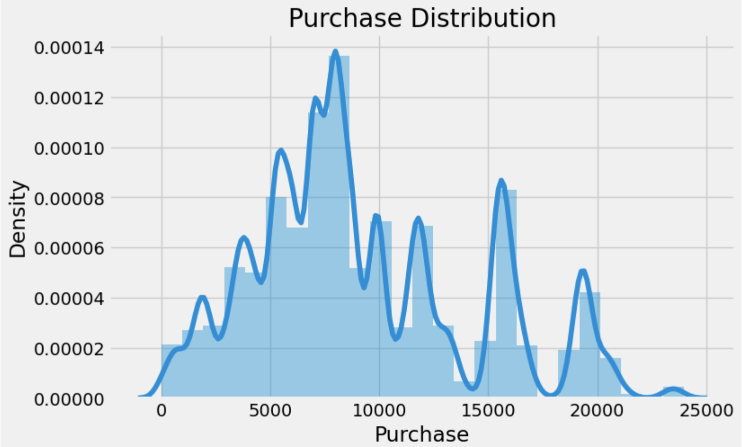
Inferences:

- Null values are present in the columns/features: product category 2 and product category 3.

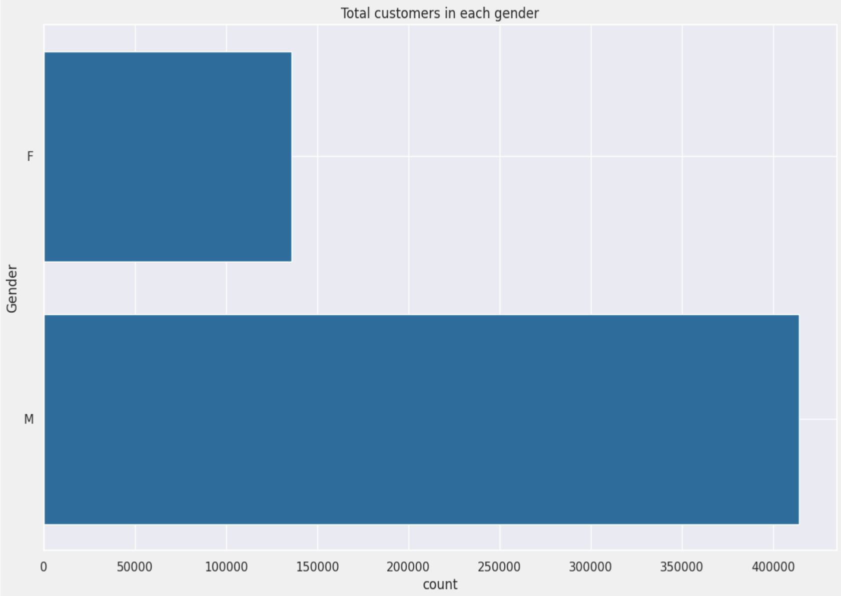
2) Purchase Distribution

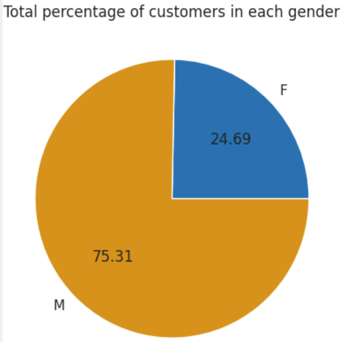
Inferences:

Density of purchase distribution is concentrated between 5000 and 10000 purchases.



3) Total customers in each gender

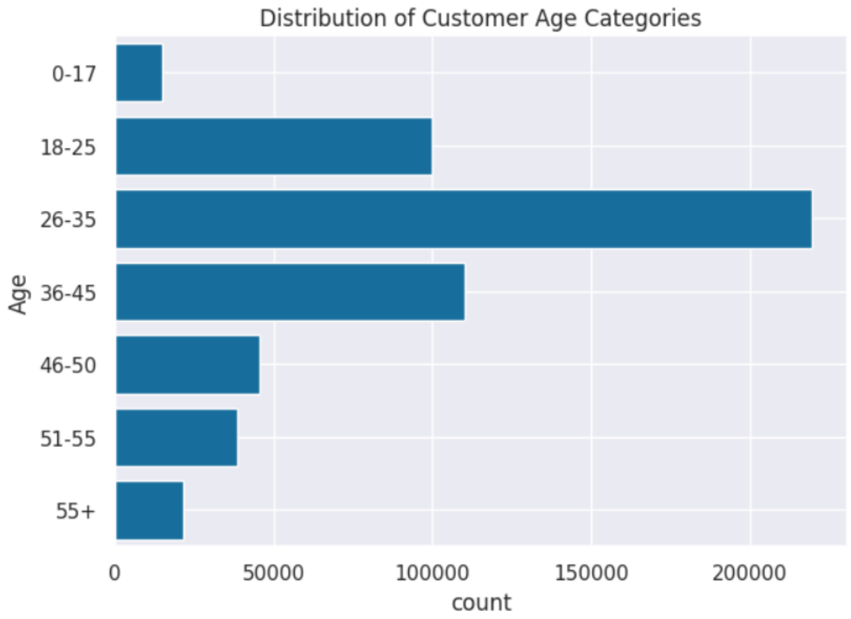




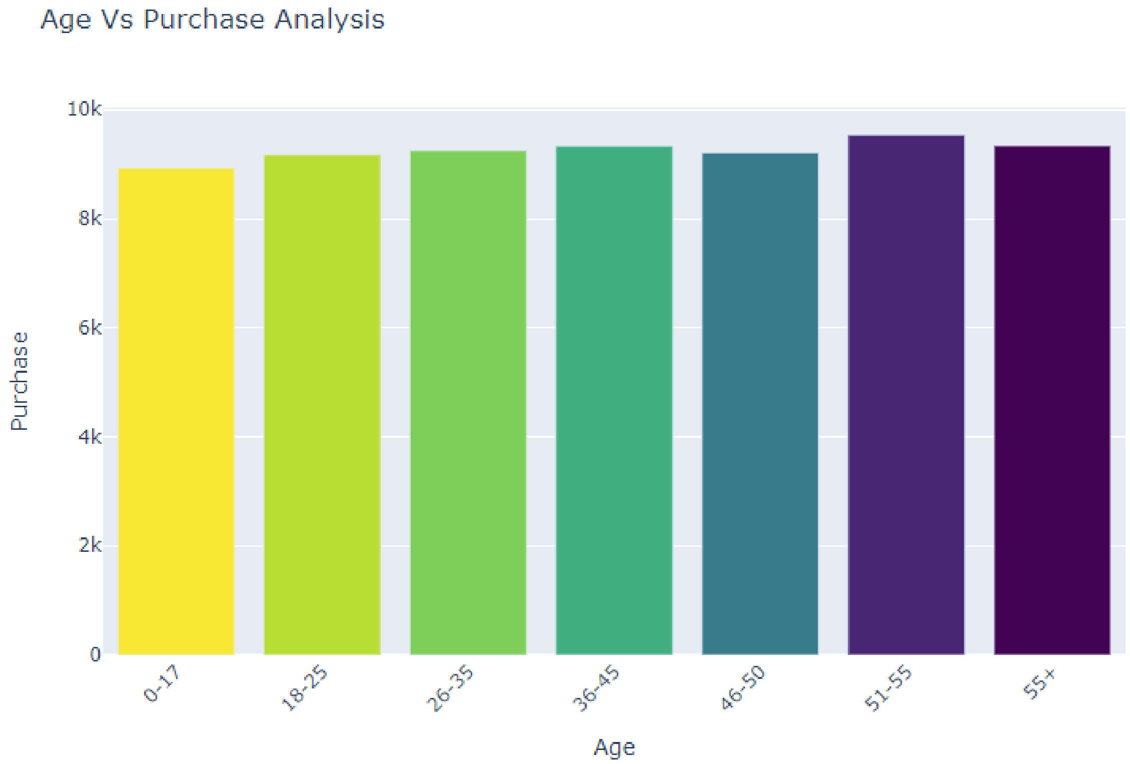
Inferences:
The number of male customers is way more than the females with a difference of around 27000 i.e., 50.62%.

4) Distribution of customer age categories

Inferences:
1. Largest Customer Group: The age group 26-35 contains the highest number of customers. This segment represents the largest share of the customer base.
2. Youth and Middle-Aged Customers: The 18-25 and 36-45 age groups also have significant customer representation, although slightly lower than the 26-35 group.
3. Decreasing Trend: As age increases beyond 45, the number of customers declines. The 55+ age category has the fewest customers.
4. Strategic Implications: Businesses targeting younger and middle-aged demographics should focus on the 18-45 age range, which constitutes the majority of their customer base.



5) Comparison of no of Purchases with respect to different age groups



Inferences:

1. Consistent Purchasing: Across different age groups, the number of purchases remains fairly consistent. Each age category makes between 8,000 to 10,000 purchases.
2. No Significant Variation: There isn’t a substantial difference in purchasing behavior from one age group to another based on this data. Whether young or middle-aged, customers exhibit similar buying patterns.

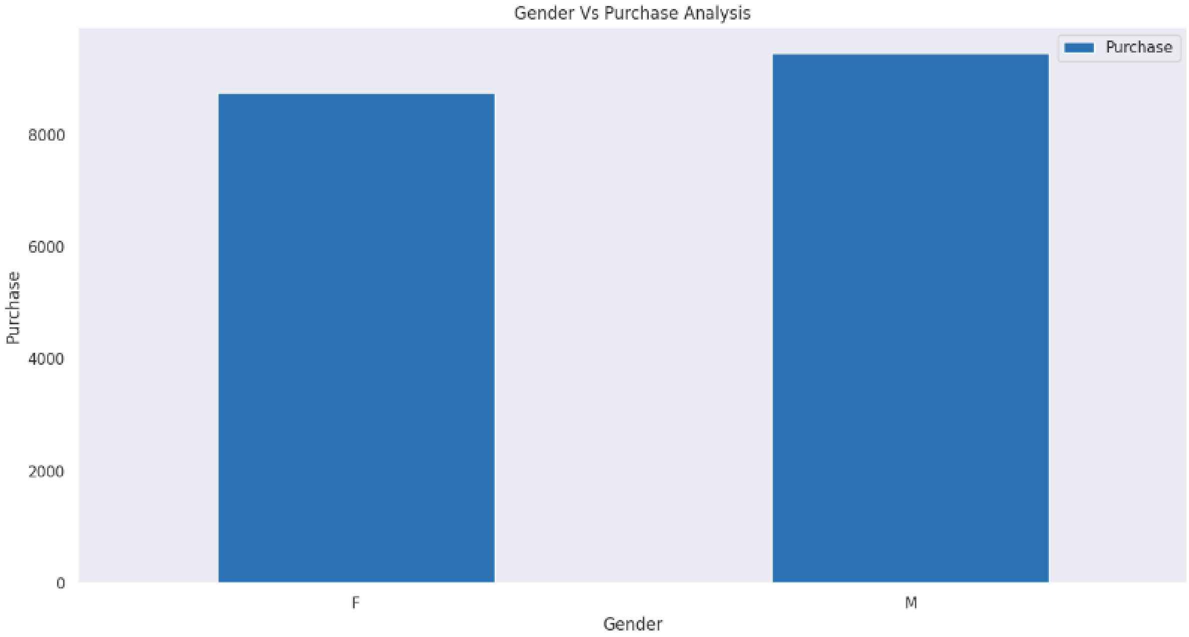
6) Gender Vs no of Purchases made – An analysis

Inferences:

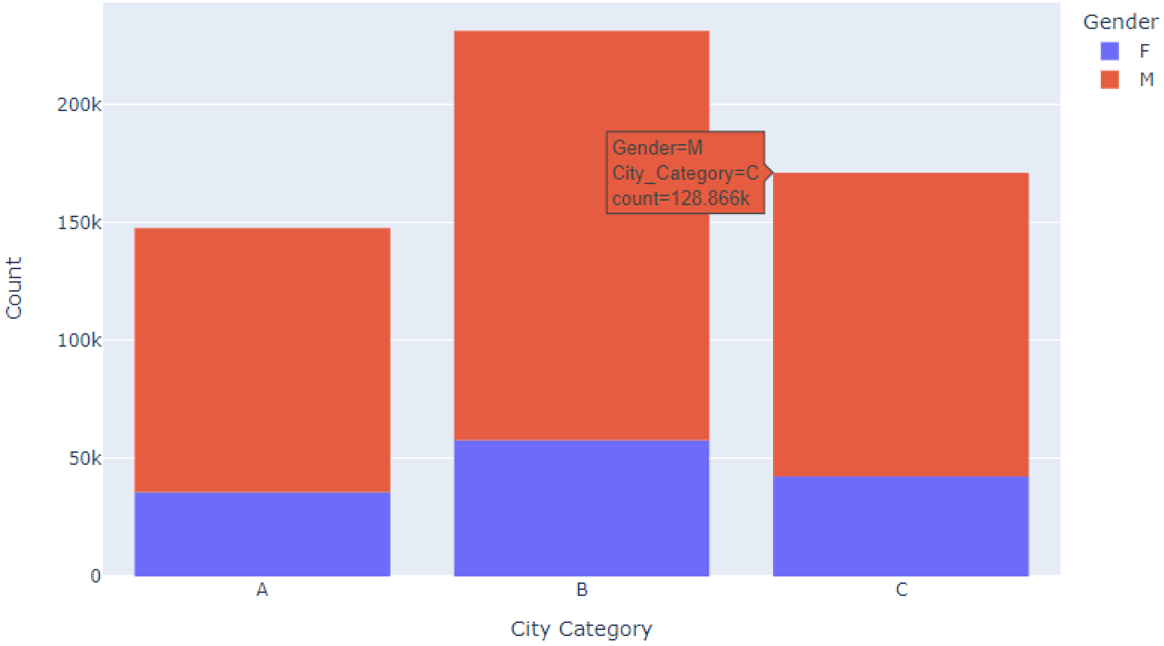
1. Balanced Purchasing: The number of purchases made by females (F) and males (M) is

nearly equal. Both genders exhibit similar buying patterns.

2. Similar Purchase Levels: Each gender group makes approximately 8,000 to 10,000 purchases, with males slightly edging out females.
3. Strategic Implications: Businesses can tailor marketing strategies without significant gender-specific variations, as both male and female customers contribute equally to the purchase volume.



7) City Category Distribution by Gender



Inferences:

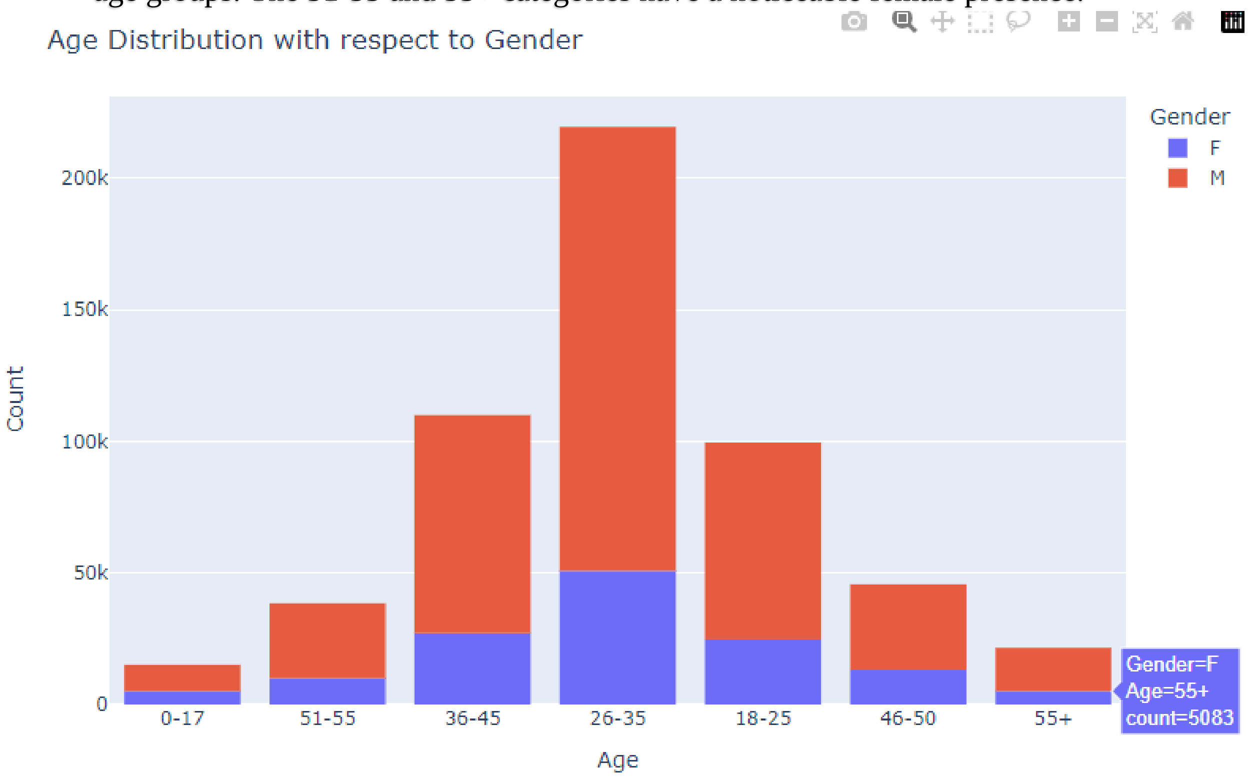
In each city, number of purchases made by a man is way more than that of a female (twice or more the no of purchases made by a female).

8) Distribution of customers with respect to Age and Gender

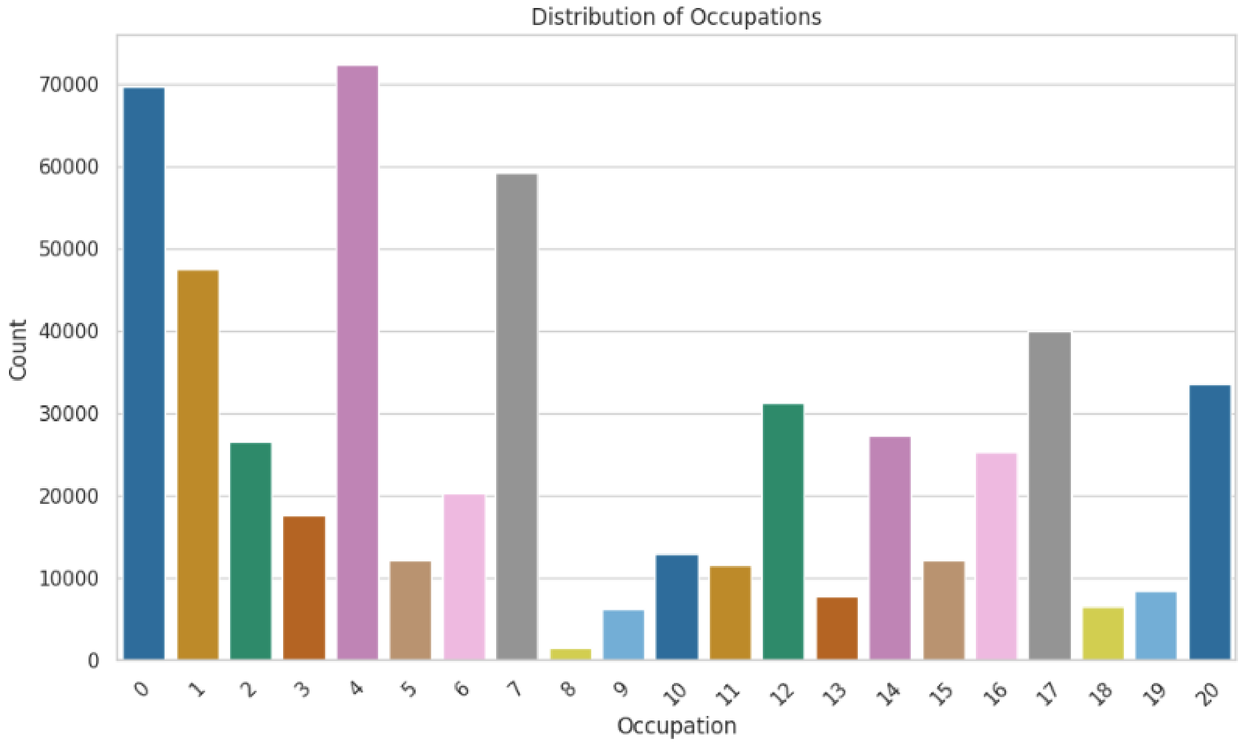
Inferences:

- Male Dominance in 26-35 Age Group: The most significant observation is the substantial number of males in the 26-35 age group. This category stands out with a notably higher count compared to other age segments.

- Even Female Distribution: In contrast, females are more evenly distributed across age groups. The 51-55 and 55+ categories have a noticeable female presence.



9) Distribution of customers based on Occupations



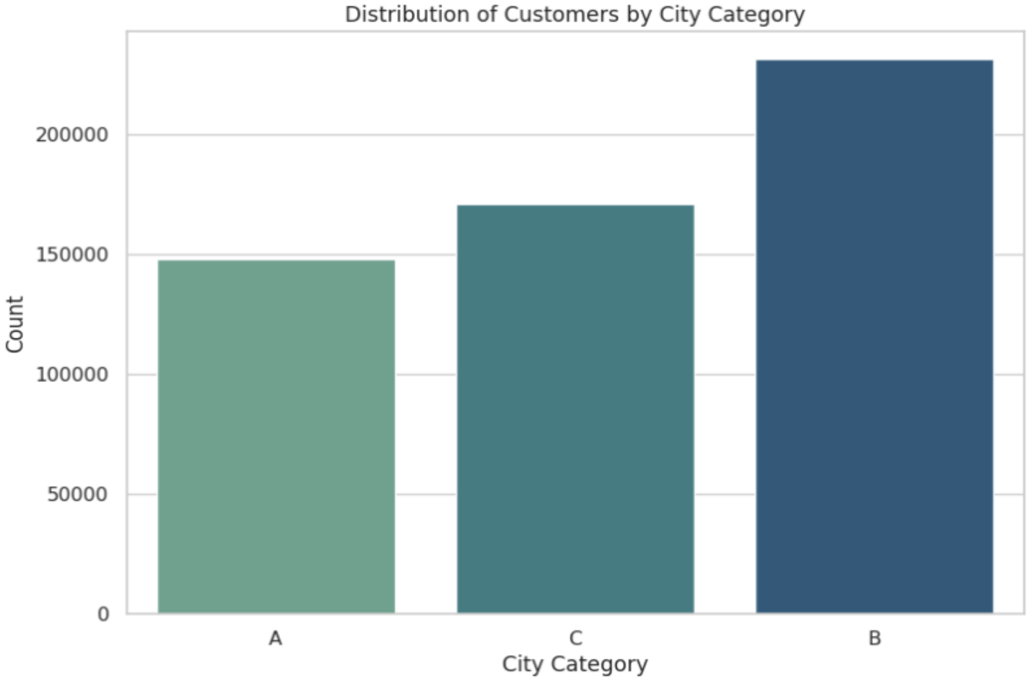
Inferences:

- The majority of the customers belong to occupation 4, 0 and 7.
- Least amount of customers belong to the occupation 8.

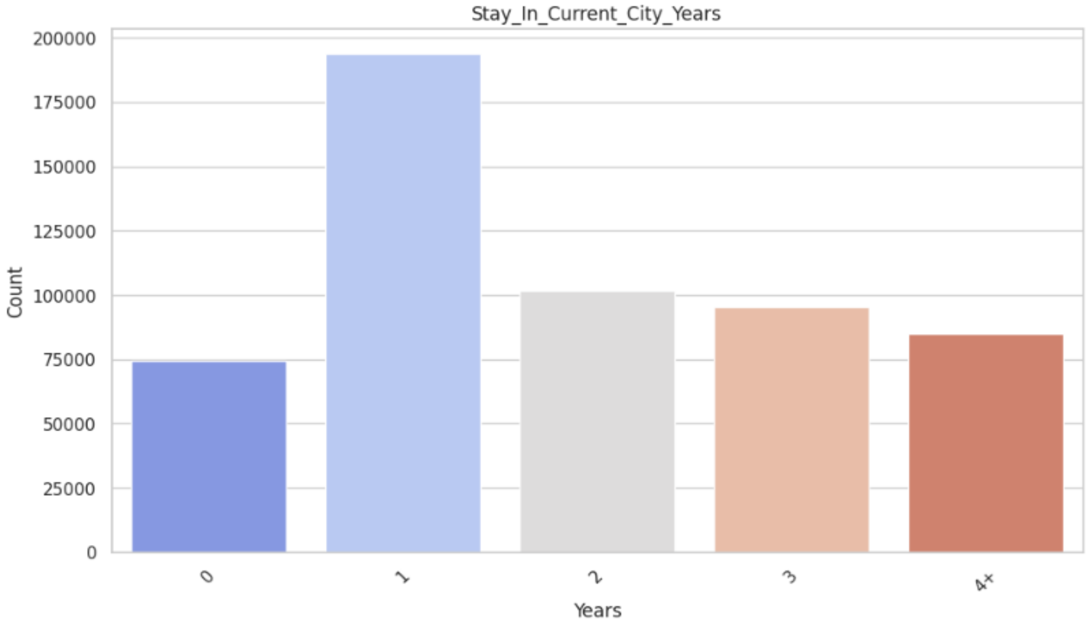
10)Distribution of Customers based on the type of city

Inferences:

Customers in city category B are more than A and C. Customers in category A and C vary slightly.



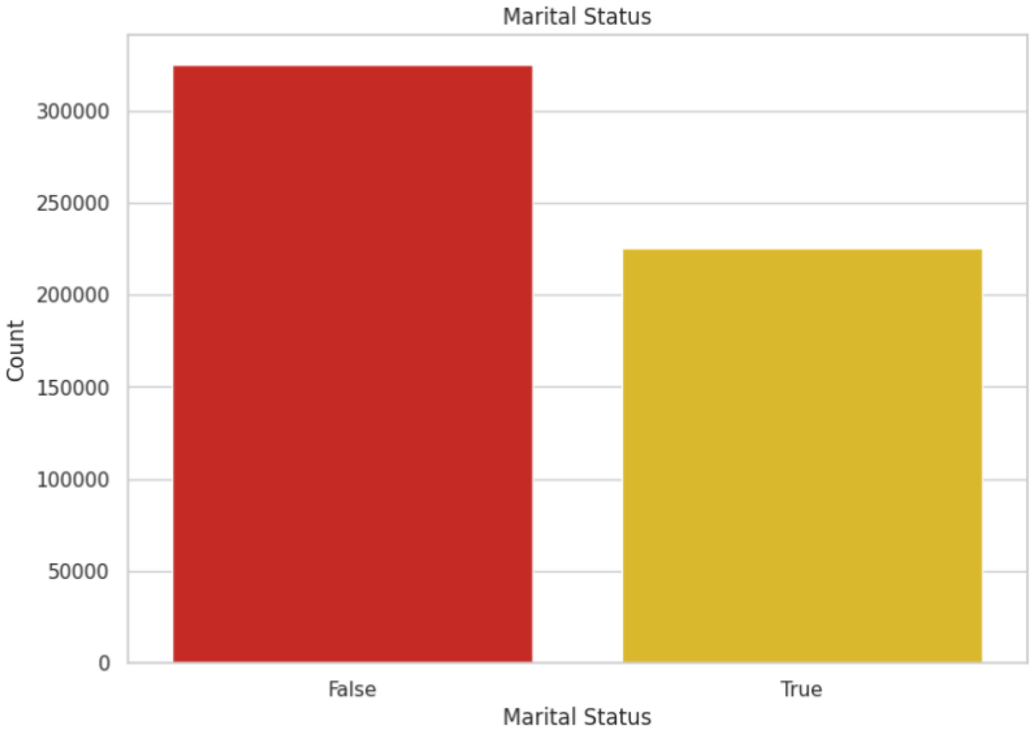
11) Categorizing the customers based on the duration of stay



Inferences:

1. Most Common Duration: The majority of people (almost 200,000) have stayed in their current city for 1 year. This is the tallest bar on the graph.
2. Second Most Common Durations: The next two categories are 2 years and 3 years, with similar counts of around 100,000 each. These durations are the second most common.
3. Less Common Durations: Fewer people have stayed for 0 years (indicating newcomers) or more than 4 years (long-term residents). The bars for these durations are shorter compared to the others.

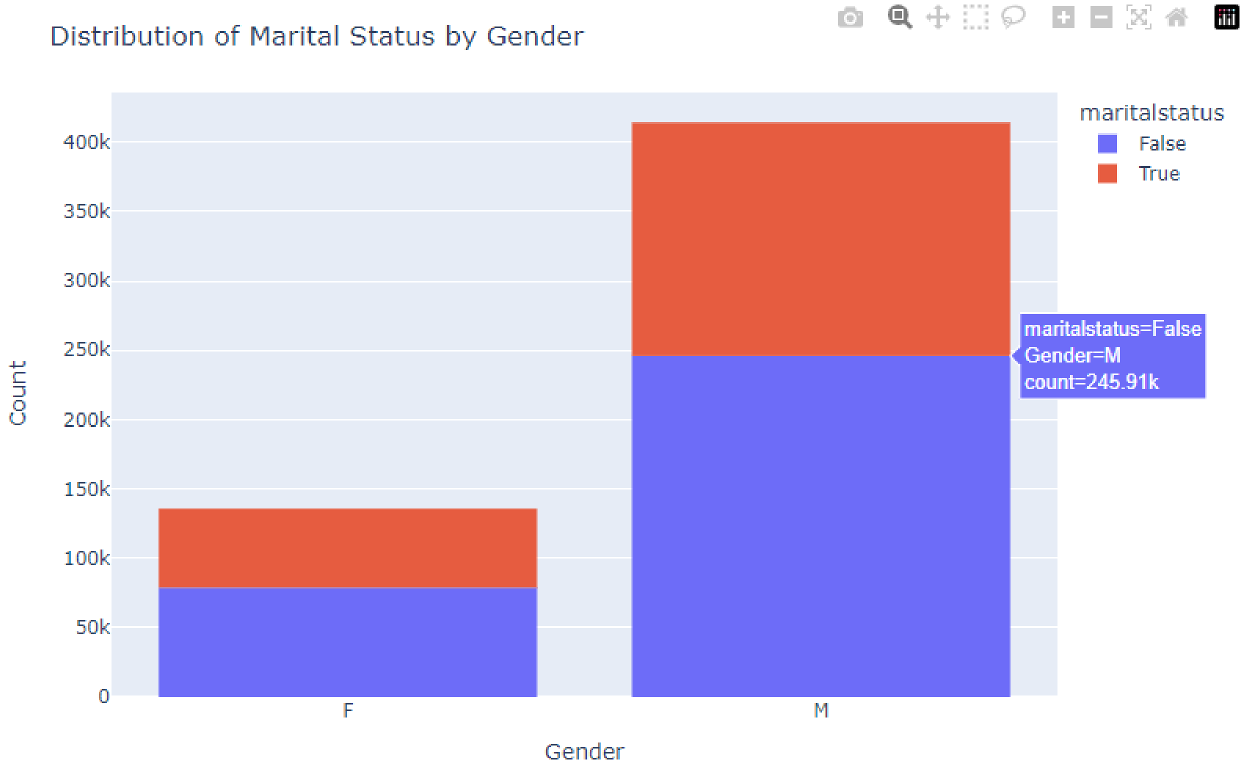
12)Marital Status of customers



Inferences:

Customers who are not married purchased more than the married customers.

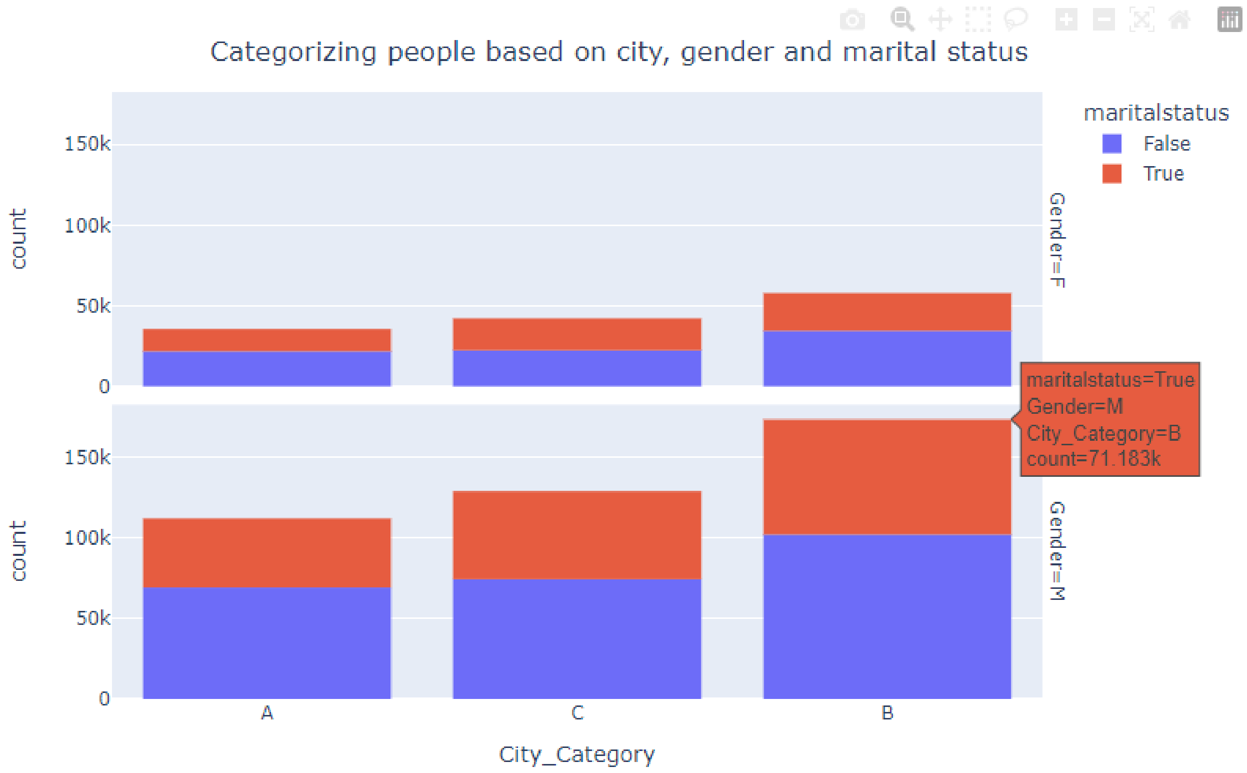
13)Distribution of customers based on their Gender and Marital Status



Inferences:

Based on Gender and Marital Status, more no of customers are in male category and count of non-married males is slightly more than that of married males and the same pattern of non-married females more than that of married females is observed.

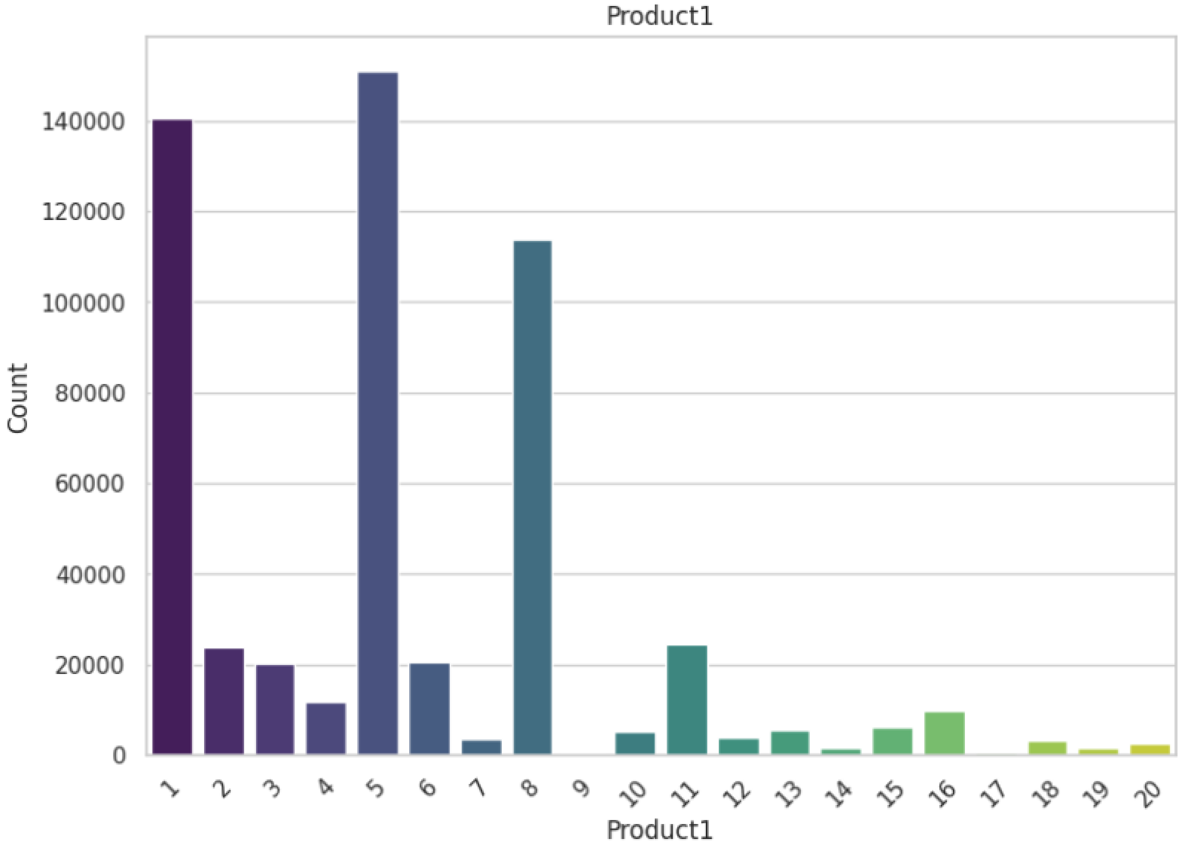
14)Categorizing customers based on city category, gender and marital status



Inferences:

The no of customers is the highest in City category B which predominantly has non - married male customers.

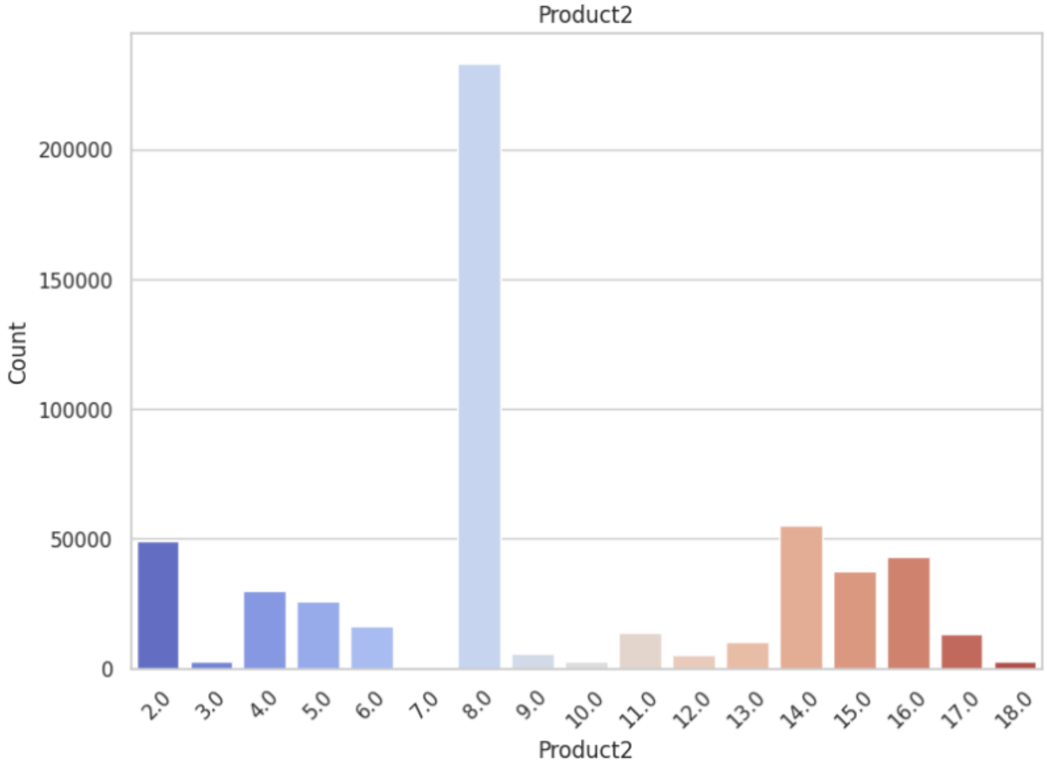
15)Distribution of products in Product1 category



Inferences:

Products 1,5 and 8 are mostly sold than other products in Product1 category.

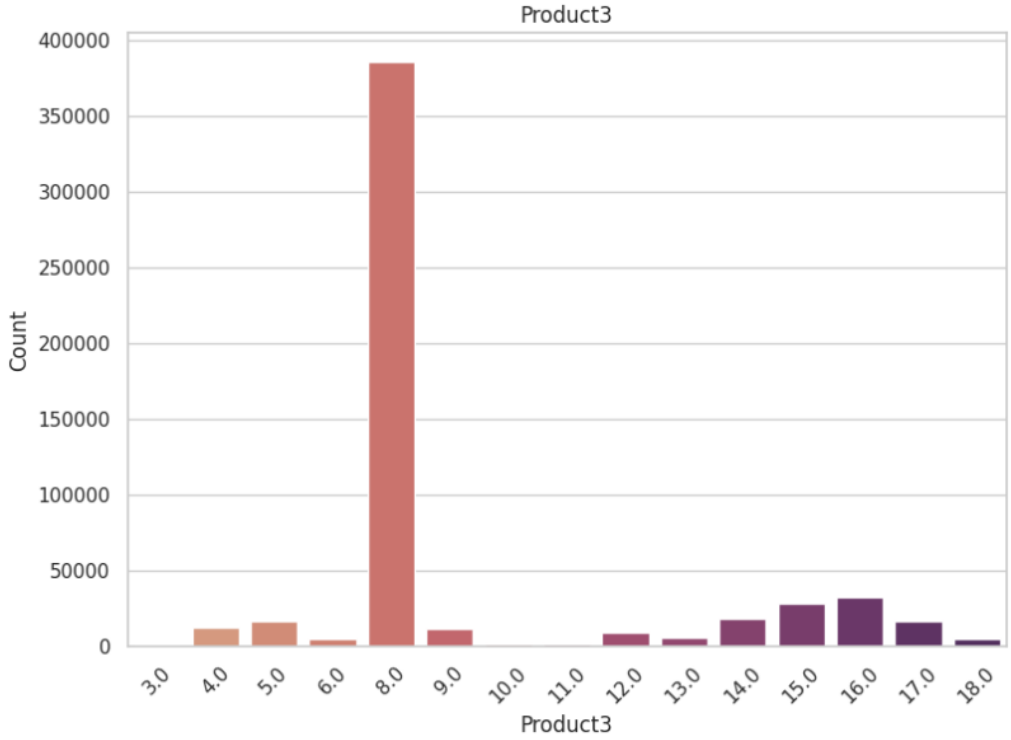
16)Distribution of products in Product2 category



Inferences:

Product 8 sold the most than the other products in Product2 category.

17)Distribution of products in Product3 category



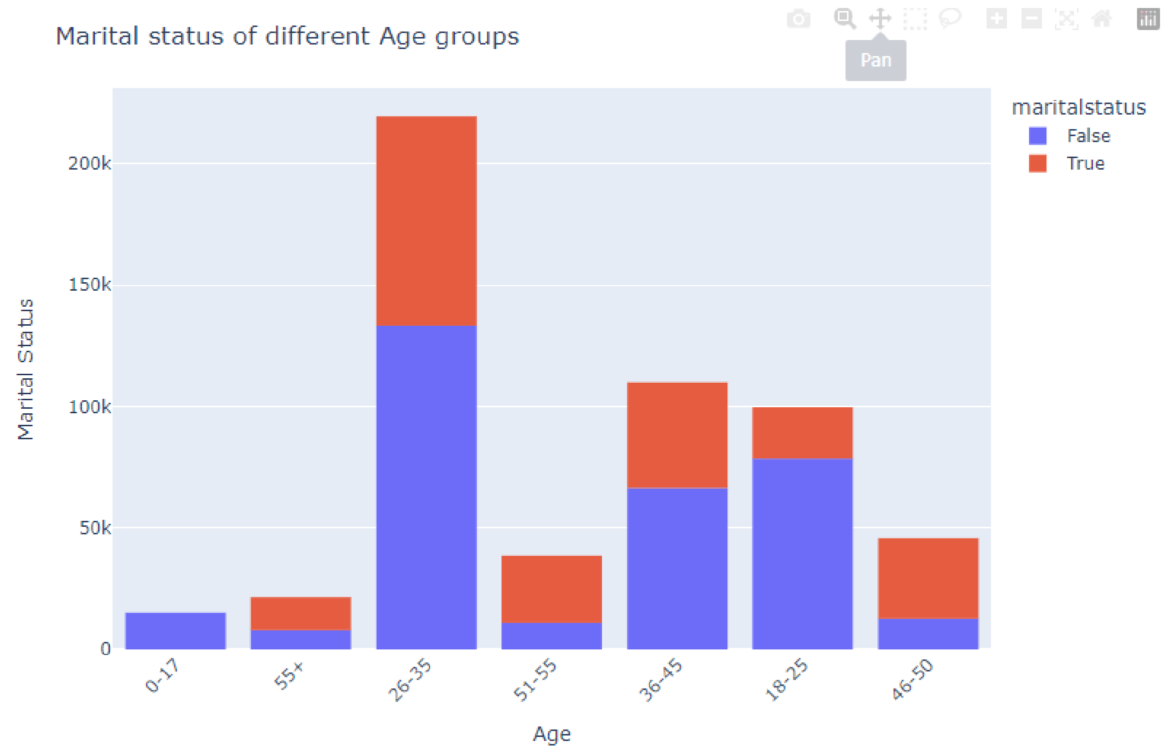
Inferences:

- Product 8 sold the most than the other products likeProduct2 category.
- Products 3, 10 and 11 have very few or zero sales.

18)Marital Status of customers across different age groups

Inferences:

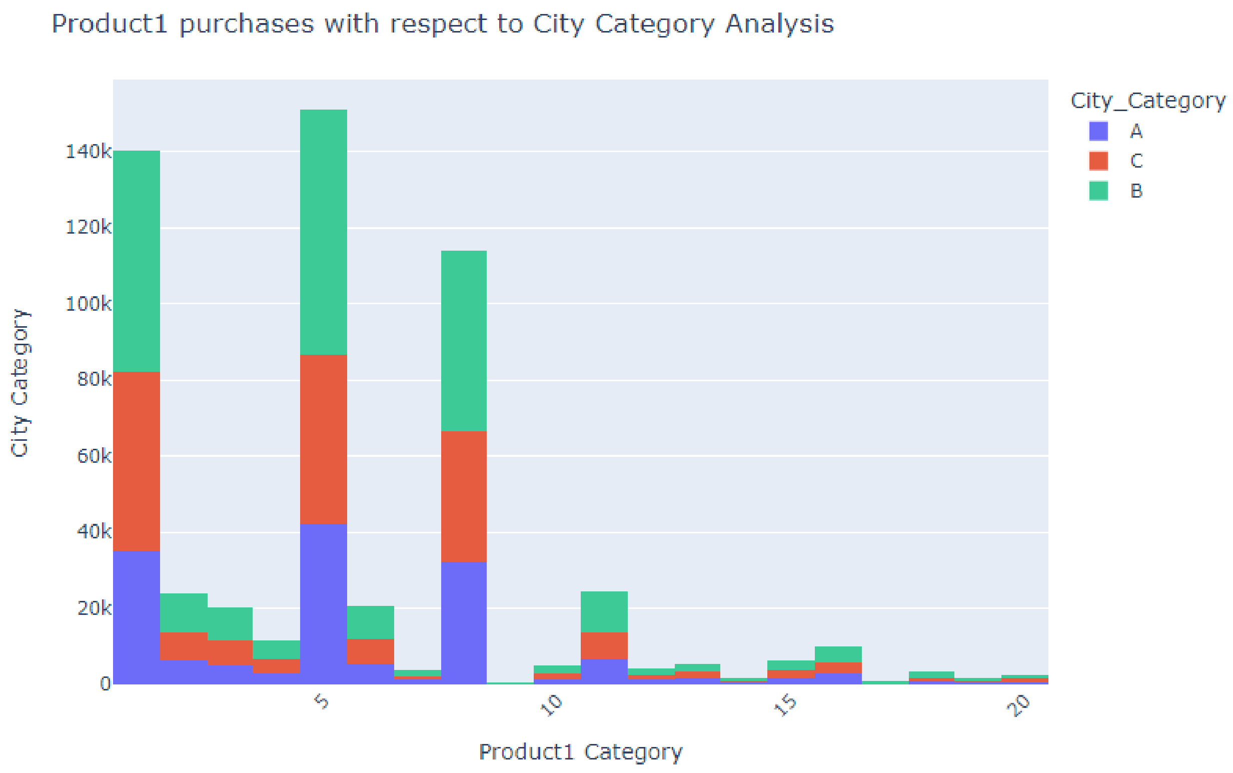
- **0-17**: Most individuals in this age group are **unmarried** (blue bar).
- **18-25**: Similar to the previous group, the majority are **unmarried**.
- **26-35**: The number of **married** individuals (red bar) starts to increase.
- **36-45**: The trend continues, with more married individuals.
- **46-50**: The count of married individuals remains high.
- **51-55**: Still a significant number of married individuals.
- 55+: The largest group of **married** individuals is in this age category.
- As age increases, the proportion of married individuals tends to rise.
- The graph reflects the common life trajectory where people are more likely to marry as they get older.

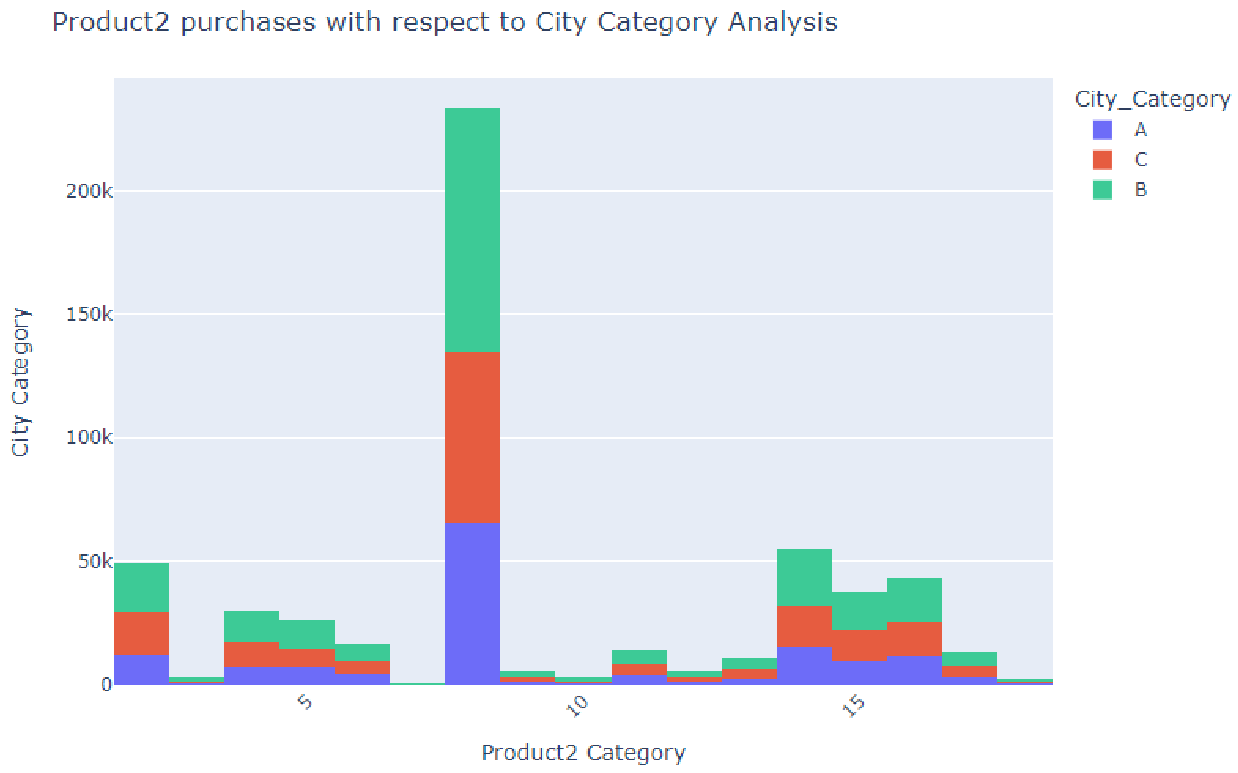


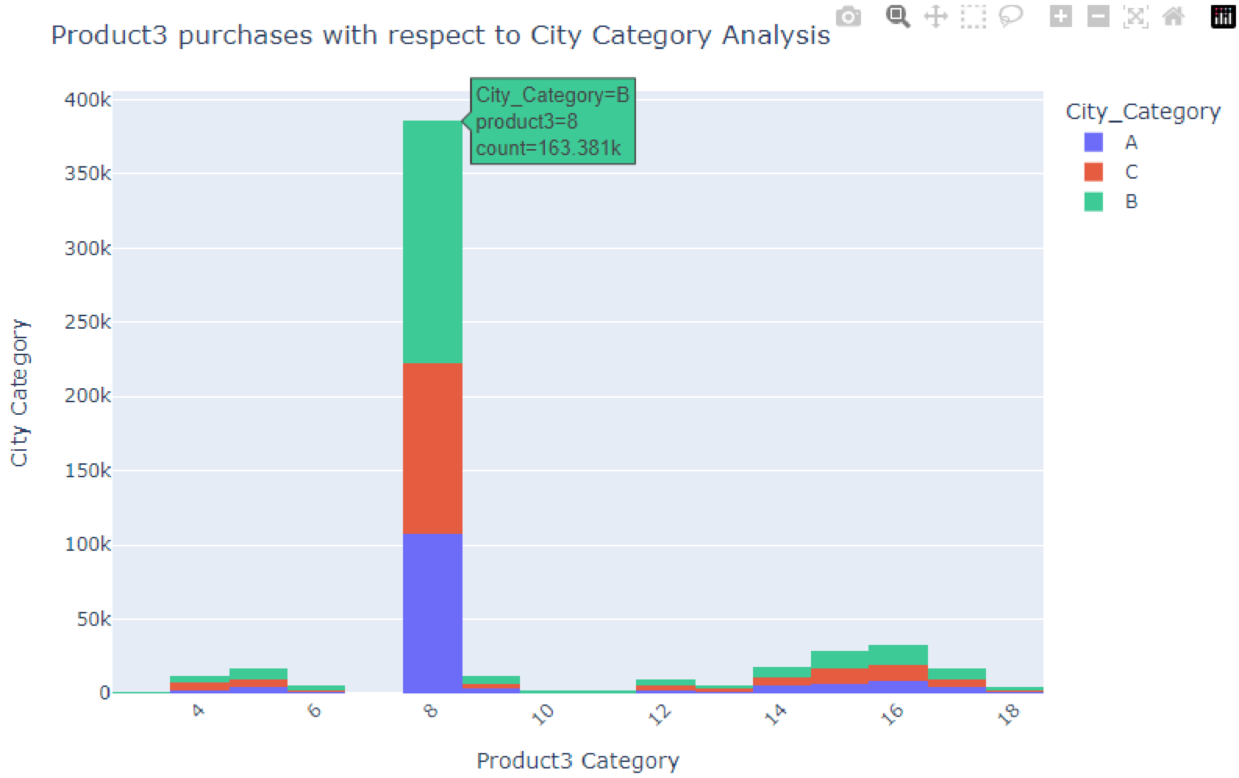
19)Product Purchases with respect to the type of cities

Inferences:

- Products 1,5,8 are the ones mostly sold in product 1 category and customers from city B purchase these products slightly higher than cities A and C.
- Product 8 is mostly sold in product 2 and product 3 categories.
- In all the product categories, customers from City B purchase slightly more than A and C.
- Least sold products in product 1 category are: 9 and 16.
- Least sold product in product 2 category is: 6.
- Least sold products in product 3 category are: 1,2,3,7,8 and 11.
- Total no of purchases in product 3 category is more than product 1 and 2 categories.


Product1 purchases with respect to City Category Analysis
City Category
0
20k
40k
60k
80k
100k
120k
140k
5
10
15
20
Product1 Category
City_Category
A
C
B


Product2 purchases with respect to City Category Analysis
City Category
0
50k
100k
150k
200k
5
10
15
Product2 Category
City_Category
A
C
B



CONCLUSION:

Exploratory Data Analysis (EDA) has been performed on sales dataset using python and insights has been obtained from the visualizations and documented successfully.