Task-01

Create a bar chart or histogram to visualize the distribution of a categorical or continuous variable, such as the distribution of ages or genders in a population.

Dataset:- https://www.kaggle.com/datasets/fahmidachowdhury/customer-segmentation-data-for-marketing-analysis/data

This dataset contains simulated customer data that can be used for segmentation analysis. It includes demographic and behavioral information about customers, which can help in identifying distinct segments within the customer base. This can be particularly useful for targeted marketing strategies, improving customer satisfaction, and increasing sales.

```
#Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
#Read the dataset
df = pd.read_csv('/content/customer_segmentation_data.csv')
df
```

	id	age	gender	income	spending_score	membership_years	purchase_frequency
0	1	38	Female	99342	90	3	24
1	2	21	Female	78852	60	2	42
2	3	60	Female	126573	30	2	28
3	4	40	Other	47099	74	9	5
4	5	65	Female	140621	21	3	25
...	...	...	...	...	...	...	...
995	996	57	Male	112170	57	6	1
996	997	23	Other	65337	76	10	23
997	998	23	Male	113097	40	5	42
998	999	22	Female	113695	63	7	44
999	1000	36	Female	90420	7	2	31

1000 rows × 9 columns

```
df.head()
```

	id	age	gender	income	spending_score	membership_years	purchase_frequency	pre
0	1	38	Female	99342	90	3	24	
1	2	21	Female	78852	60	2	42	
2	3	60	Female	126573	30	2	28	
3	4	40	Other	47099	74	9	5	
4	5	65	Female	140621	21	3	25	

```
df.shape #Total no. of rows & cols
```

```
(1000, 9)
```

```
df.info() #summary of df
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   id                    1000 non-null   int64  
 1   age                   1000 non-null   int64  
 2   gender                1000 non-null   object 
 3   income                1000 non-null   int64  
 4   spending_score        1000 non-null   int64  
 5   membership_years      1000 non-null   int64  
 6   purchase_frequency    1000 non-null   int64  
 7   preferred_category    1000 non-null   object 
 8   last_purchase_amount  1000 non-null   float64
dtypes: float64(1), int64(6), object(2)
```

```
memory usage: 70.4+ KB
```

```
df.isnull().sum() #check for null values
```

```
id                    0
age                   0
gender                0
income                0
spending_score        0
membership_years      0
purchase_frequency    0
preferred_category    0
last_purchase_amount  0
dtype: int64
```

There are no missing values in the dataset, no duplicate values as well.

```
df.duplicated().sum() #check for duplicate values
```

```
0
```

Categorical Variables dataset are:-

1.gender: Represents categories (e.g., Male, Female).

2.preferred_category: Represents categories of products (e.g., Electronics, Clothing, Home & Garden, Sports, Groceries).

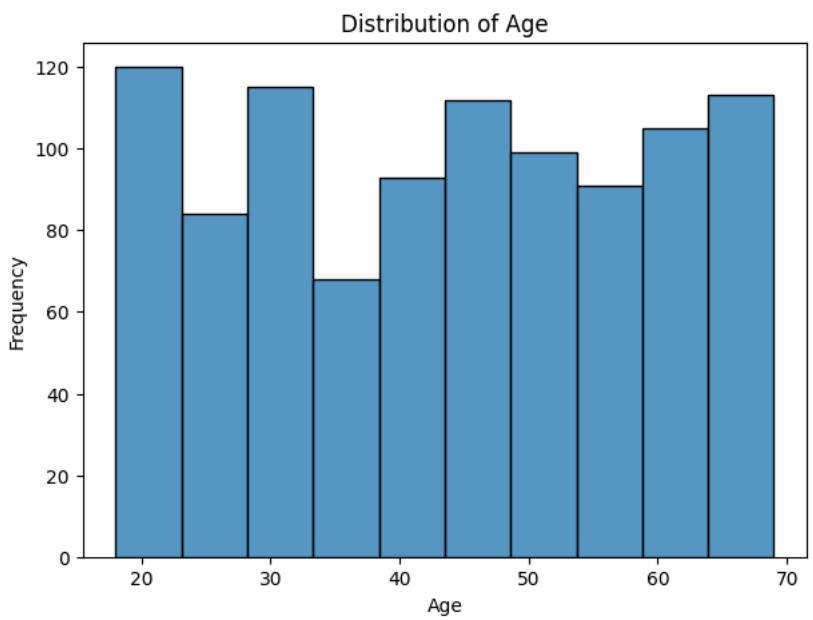
Continuous Variables

1.income: Represents a continuous measure of income.

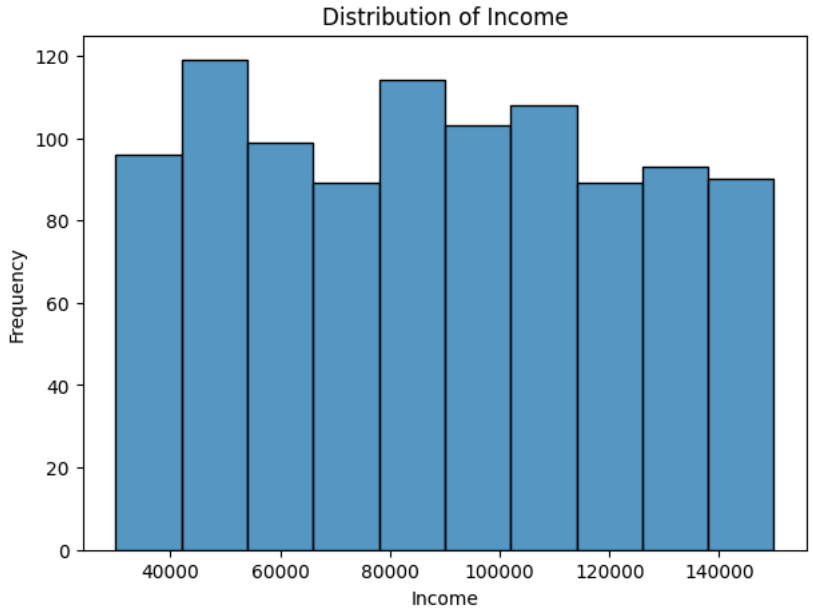
2.spending_score: Represents a continuous score based on spending behavior.

3.last_purchase_amount: Represents a continuous measure of the amount spent in the last purchase.

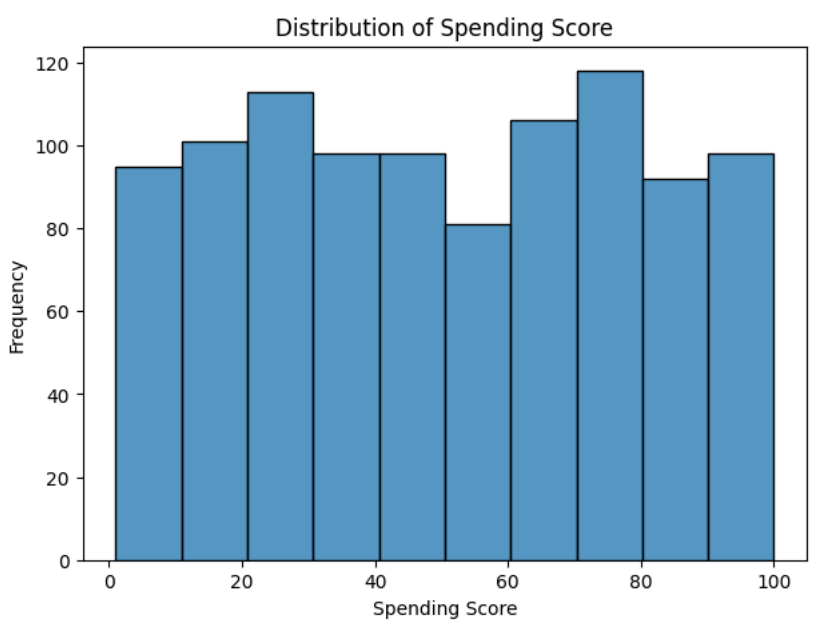
```
import seaborn as sns       #Histogram
plt.figure(figsize=(7, 5))
sns.histplot(df['age'], bins=10)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



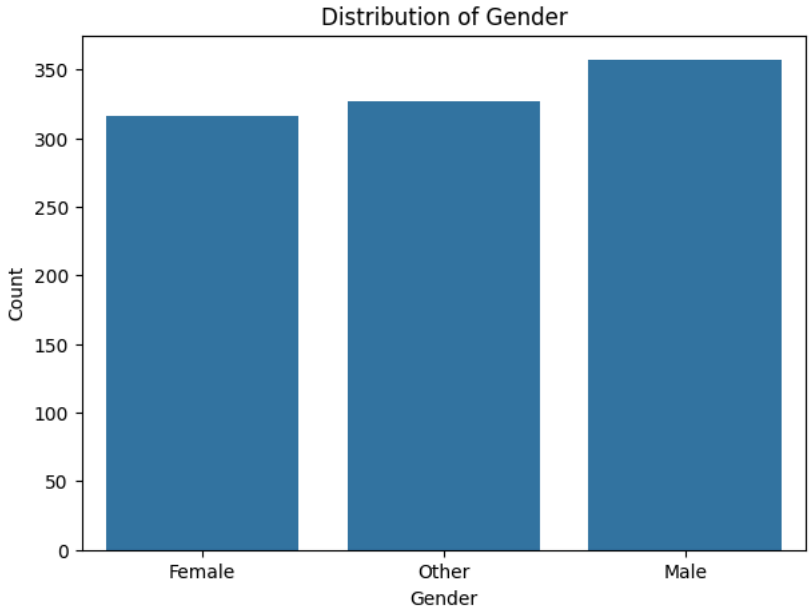

```
plt.figure(figsize=(7, 5))     #Hist
sns.histplot(df['income'], bins=10)
plt.title('Distribution of Income')
plt.xlabel('Income')
plt.ylabel('Frequency')
plt.show()
```



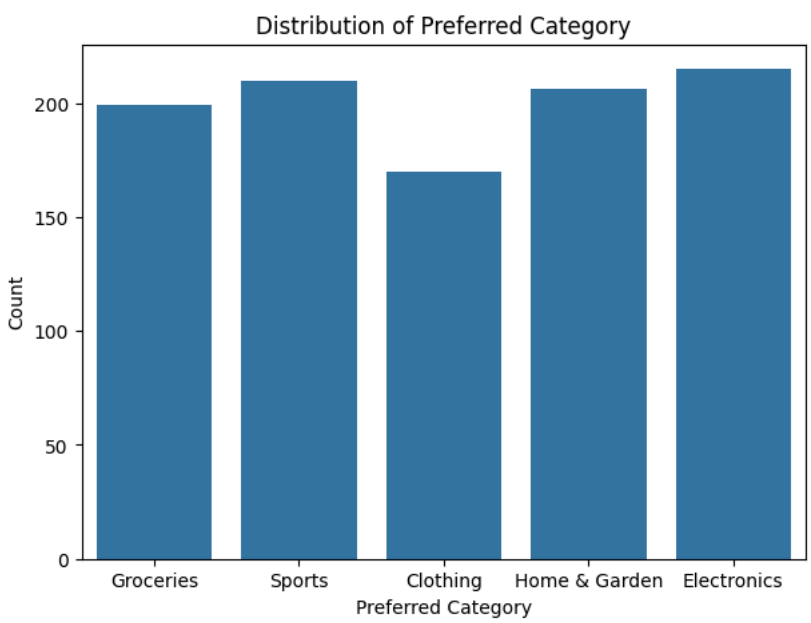

```
plt.figure(figsize=(7, 5))     #Hist
sns.histplot(df['spending_score'], bins=10)
plt.title('Distribution of Spending Score')
plt.xlabel('Spending Score')
plt.ylabel('Frequency')
plt.show()
```



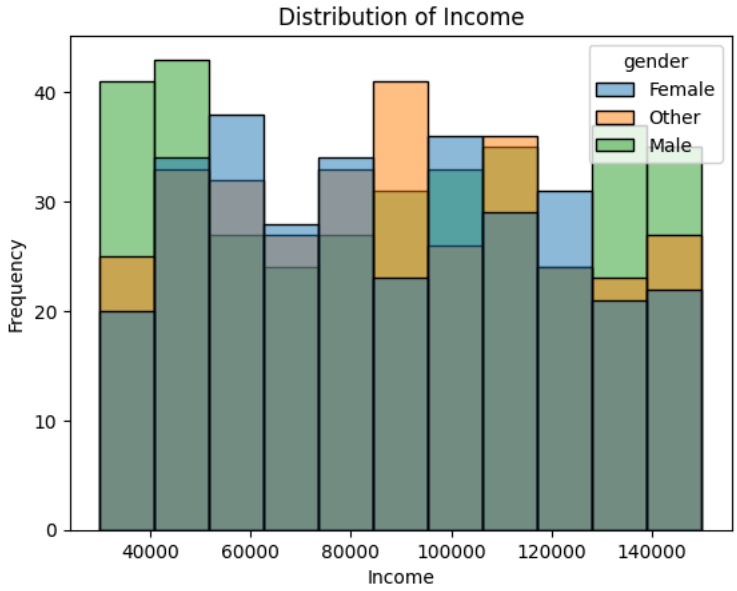

```
#Bar Chart
plt.figure(figsize=(7, 5))
sns.countplot(x='gender', data=df)
plt.title('Distribution of Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```




```
#Bar Chart
plt.figure(figsize=(7, 5))
sns.countplot(x='preferred_category', data=df)
plt.title('Distribution of Preferred Category')
plt.xlabel('Preferred Category')
plt.ylabel('Count')
plt.show()
```




```
sns.histplot(x='income', data=df, hue='gender')
plt.title('Distribution of Income')
plt.xlabel('Income')
plt.ylabel('Frequency')
plt.show()
```




```
sns.histplot(x='last_purchase_amount', data=df, hue='preferred_category')
plt.title('Distribution of last_purchase_amount')
plt.xlabel('last_purchase_amount')
plt.ylabel('Frequency')
plt.show()
```