**TEAM MEMBER**

732521104030 : NAVEEN S

**PHASE 2** Submission Document

**Problem Definition:**

The problem is to implement data science techniques to segment customers based on their behavior, preferences, and demographic attributes. The goal is to enable businesses to personalize marketing strategies and enhance customer satisfaction. This project involves data collection, data preprocessing, feature engineering, clustering algorithms, visualization, and interpretation of results.

**Dataset Link:**[**https://www.kaggle.com/datasets/akram24/mall-customers**](https://www.kaggle.com/datasets/akram24/mall-customers)



**Design Thinking:**

Data collection:

Collect customer data, including attributes like purchase history, demographic information, and interaction behavior.

***Purchase history*** : Purchase history refers to **a record of a customer's past transactions with a business or retailer**. It typically includes details such as the items purchased, dates, quantities, and prices.

***Demographic information*** : Demographic information examples include: **age, race, ethnicity, gender, marital status, income, education, and employment**. You can easily and effectively collect these types of information with survey questions.

**Interaction behavior**: The quality factors influence the quality aspects through the interaction performance aspects. They include **the characteristics of the user, system, and context of use that have an impact on perceived quality**. User factors. All characteristics of the user that influence her interaction behavior and quality judgment.

**Data preprocessing**:

 Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.

**Feature Engineering**:

Create additional features that capture customer behavior and preferences, such as total spending, frequency of purchases, etc.

***Total spending***: The sum of the price paid for one or more products or services multiplied by the amount of each item purchased.

**Frequency of purchases**: Purchase frequency describes the number of times that your customers make a purchase from you within a specified period of time. This information is crucial in helping you to understand your customer retention rate, your customers' buying behaviors, and even the degree to which they're satisfied.

***Clustering Algorithms*:** Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

***K-Means* :** K-means is **an unsupervised learning method for clustering data points**. The algorithm iteratively divides data points into K clusters by minimizing the variance in each cluster.

***hierarchical clustering*:** Hierarchical clustering is **a popular method for grouping objects**. It creates groups so that objects within a group are similar to each other and different from objects in other groups. Clusters are visually represented in a hierarchical tree called a dendrogram.

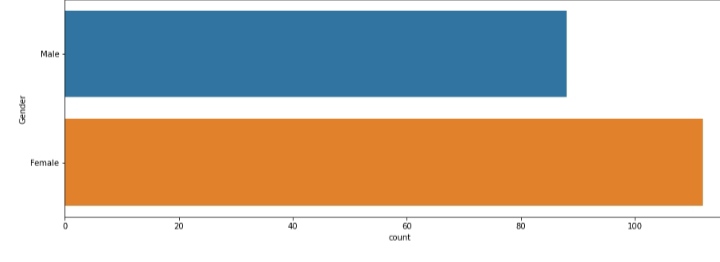
**Visualization:**

Visualize the customer segments using techniques like scatter plots, bar charts, and heatmaps.

***plots*** : A scatter plot is **a diagram where each value in the data set is represented by a dot**. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis: x = [5,7,8,7,2,17,2,9,4,11,12,9,6]

***Bar charts*** : The bar chart **displays data using a number of bars, each representing a particular category**. The height of each bar is proportional to a specific aggregation (for example the sum of the values in the category it represents). The categories could be something like an age group or a geographical location.

**Barchart:**



***Heatmaps*** : A heatmap is **a graphical representation of data that uses a system of color coding to represent different values**. Heatmaps are used in various forms of analytics but are most commonly used to show user behavior on specific web pages or webpage templates.

**Interpretation:**

Analyze and interpret the characteristics of each customer segment to derive actionable insights for marketing strategies.

Program:

import numpy as np linear algebra

import pandas as pd

import os

for dirname,. filenames in os.walk(' C:\Users\Students\Documents\New folder\Mall Customers '):

for

filename in filenames:

print(os.path.join(dirname, filename))

import numpy as np

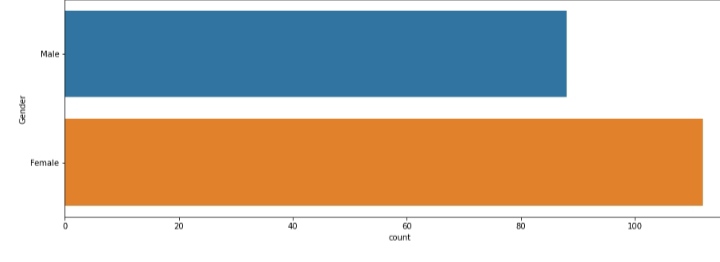
import pandas as pd import matplotlib.pyplot as plt

import seaborn as sns

df pd.read\_csv('/kaggle/input/mall-customers/Mall Customers.csv')

df.rename(columns=('Genre: 'Gender' }, inplace=True)

df.head()

**output:** 

plt.figure(1,figsize=(15,6))

n = 0

for x **in** ['Age','Annual Income (k$)','Spending Score (1-100)']:

n +=1

plt.subplot(1,3,n)

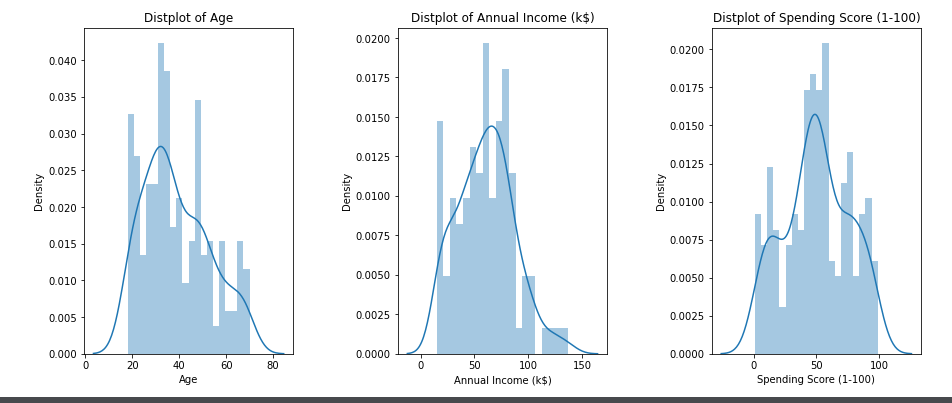
plt.subplots\_adjust(hspace=0.5,wspace=0.5)

sns.distplot(df[x],bins=20)

plt.title('Distplot of **{}**'.format(x))

plt.show()

**output :**

****

**plt.figure(1,figsize=(15,6))**

**n = 0**

**for** **cols in ['Age','Annual Income (k$)','Spending Score (1-100)']:**

**n +=1**

**plt.subplot(1,3,n)**

**sns.set(style="whitegrid")**

**plt.subplots\_adjust(hspace=0.5,wspace=0.5)**

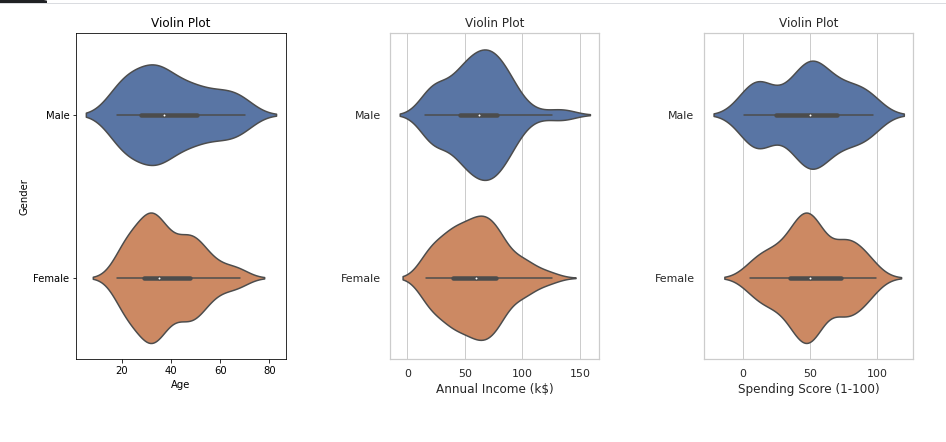
**sns.violinplot(x = cols,y = 'Gender',data=df)**

**plt.ylabel('Gender' if n== 1 else '')**

**plt.title('Violin Plot')**

**plt.show()**

**OUTPUT:**

****