**Project to**

**IBM NAAN MUTHALVAN**

**APPLIED DATA SCIENCE**

**Submitted by**

Selvin Raajkumar C (911721106315)

Vinoth Raajan K (911721106039)

Sriram N (911721106317)

Praveen Kumar (911721106310)

MOUNT ZION COLLEGE OF ENGINEERING AND TECHNOLOGY

NAAC Accredited Institution (with A+ Grade)

Pudukkottai-622507,Tamilnadu.

**Customer segmentation using data science**

**Project Title**: Customer segmentation

Phase:3 Preloading and processing of dataset Preloading and processing a dataset is a crucial step in preparing data for machine learning tasks. This process involves loading the dataset into memory, cleaning and organizing the data, handling missing values, feature engineering, and sometimes splitting the data into training and testing sets



**\*\*Problem Statement: Customer Segmentation using Data Science\*\***

**\*\*Background:\*\***

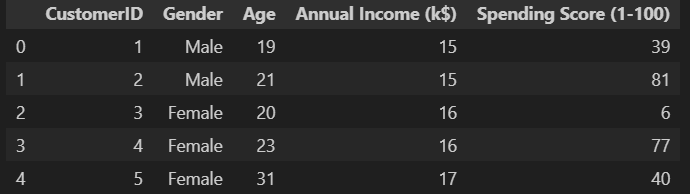
In today's competitive business landscape, understanding and effectively catering to the diverse needs of customers is crucial for sustainable growth. Customer segmentation is a powerful strategy that involves dividing a customer base into distinct groups based on similar characteristics, behaviors, or preferences. By doing so, businesses can tailor their marketing, product development, and customer service efforts to better meet the specific needs of each segment.

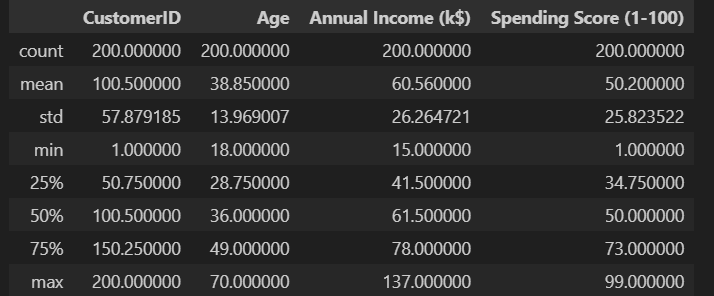
**\*\*Objective:\*\***

The goal of this project is to leverage data science techniques to perform customer segmentation for a given business. The identified segments should provide actionable insights that enable the company to enhance its marketing strategies, improve customer satisfaction, and drive overall business success.

**\*\*Dataset:\*\***

The dataset for this project will include relevant customer data such as:





1. \*\*Demographic Information:\*\* Age, gender, location, income, etc.

2. \*\*Transactional Data:\*\* Purchase history, frequency, recency, average transaction value, etc.

3. \*\*Behavioral Data:\*\* Online engagement, product/service usage patterns, etc.

**\*\*Key Tasks:\*\***

1. \*\*Data Exploration and Preprocessing:\*\*

- Explore and understand the structure of the dataset.

- Handle missing values, outliers, and perform necessary data cleaning.

- Conduct statistical and visual analyses to gain insights into the distribution of key variables.

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

n = 0

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

n +=1

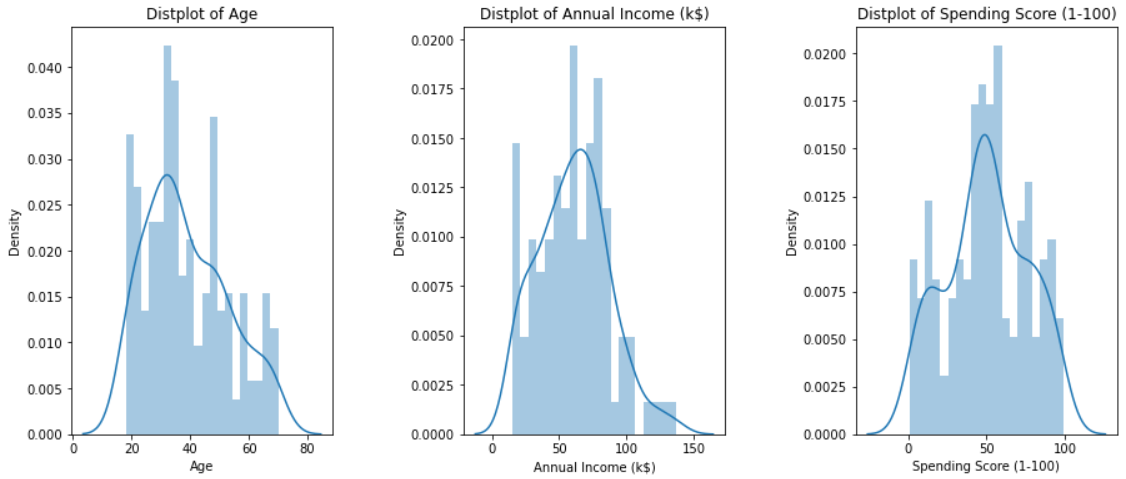
plt.subplot(1,3,n)

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

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

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

plt.show()



2. \*\*Feature Engineering:\*\*

- Create relevant features that can contribute to the segmentation process.

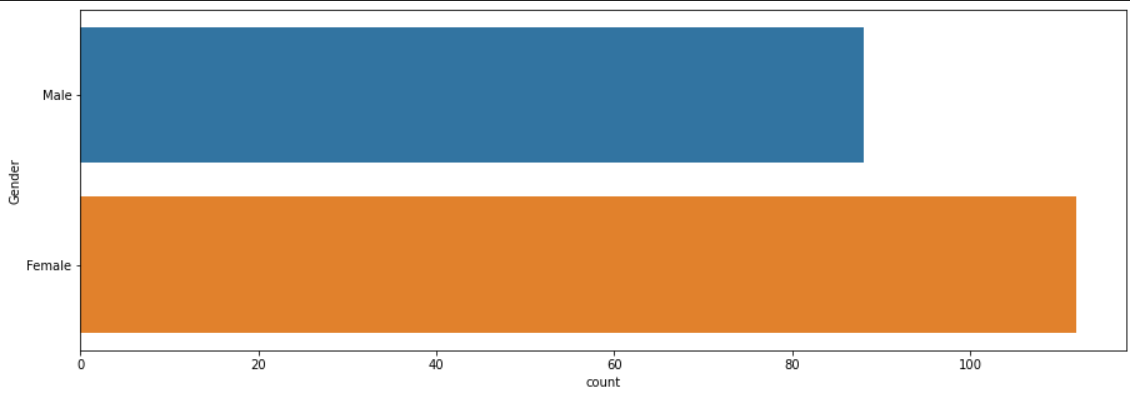
- Normalize or scale numerical features as needed.

- Encode categorical variables appropriately.

plt.figure(figsize=(15,5))

sns.countplot(y='Gender',data=df)

plt.show()



3. \*\*Customer Segmentation:\*\*

- Apply clustering algorithms (e.g., k-means, hierarchical clustering) to segment customers based on their characteristics.

- Experiment with different numbers of clusters and evaluate their effectiveness.

- Visualize the clusters to interpret and communicate the segmentation results effectively.

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")

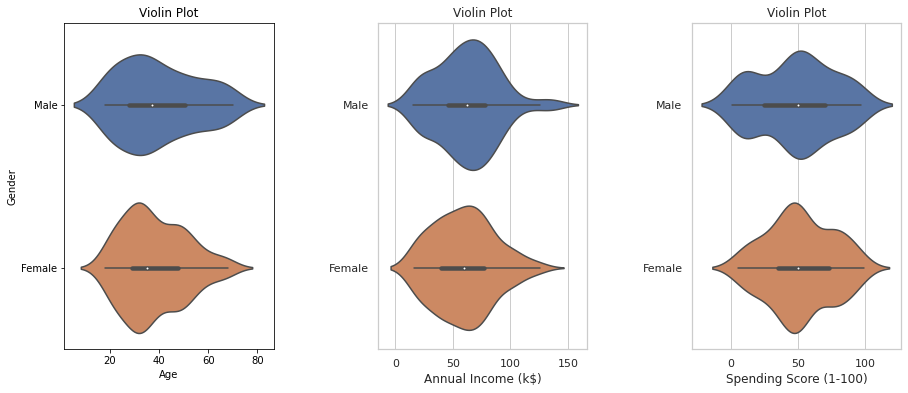
plt.subplots\_adjust(hspace=0.5,wspace=0.5)

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

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

plt.title('Violin Plot')

plt.show()



4. \*\*Segment Profiling:\*\*

- For each identified segment, create customer profiles detailing their common characteristics and behaviors.

- Analyze and interpret the distinguishing features of each segment

ss\_1\_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) & (df["Spending Score (1-100)"] <= 20)]

ss\_21\_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21) & (df["Spending Score (1-100)"] <= 40)]

ss\_41\_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41) & (df["Spending Score (1-100)"] <= 60)]

ss\_61\_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) & (df["Spending Score (1-100)"] <= 80)]

ss\_81\_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 81) & (df["Spending Score (1-100)"] <= 100)]

ssx= ["1-20","21-40","41-60","61-80","81-100"]

ssy=[len(ss\_1\_20.values),len(ss\_21\_40.values),len(ss\_41\_60.values),len(ss\_61\_80.values),len(ss\_81\_100.values)]

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

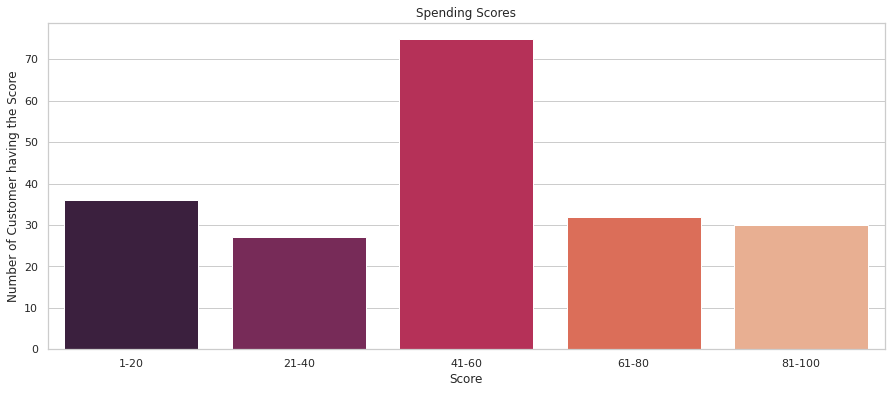
sns.barplot(x=ssx,y=ssy, palette="rocket")

plt.title("Spending Scores")

plt.xlabel("Score")

plt.ylabel("Number of Customer having the Score")

plt.show()

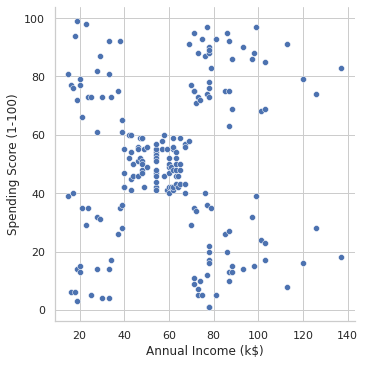


5. \*\*Validation and Model Evaluation:\*\*

- Validate the robustness of the segmentation through methods like cross-validation.

- Evaluate the performance of the clustering model using relevant metrics.

sns.relplot(x="Annual Income (k$)",y = "Spending Score (1-100)",data=df)



6. \*\*Business Recommendations:\*\*

- Provide actionable insights and recommendations for the business based on the identified customer segments.

ai\_0\_30 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 0) & (df["Annual Income (k$)"] <= 30)]

ai\_31\_60= df["Annual Income (k$)"][(df["Annual Income (k$)"] >=31)& (df["Annual Income (k$)"] <=60)]

ai\_61\_90= df["Annual Income (k$)"][(df["Annual Income (k$)"] >=61)& (df["Annual Income (k$)"] <=90)]

ai\_61\_90=df["Annual Income (k$)"][(df["Annual Income (k$)"] >=91)& (df["Annual Income (k$)"] <=120)]

ai\_121\_150 = df["Annual Income (k$)"][(df["Annual Income (k$)"]>=121) & (df["Annual Income (k$)"] <=150)]

aix = ["$ 0 - 30,000","$ 30,001 - 60,000","$ 60,001 - 90,000","$ 90,001 - 120,000","$ 120,001 - 150,000"]

aiy = [len(ai\_0\_30.values),len(ai\_31\_60.values),len(ai\_61\_90.values),len(ai\_61\_90.values),len(ai\_121\_150.values)]

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

sns.barplot(x=aix,y=aiy,palette="Spectral")

plt.title("Annual Incomes")

plt.xlabel("Income")

plt.ylabel("Numer of Customer")

plt.show()

- Suggest personalized marketing strategies, product/service improvements, or targeted promotions for each segment.

**\*\*Deliverables:\*\***

- Jupyter notebook or equivalent documenting the entire data science process.

- Visualizations and insights derived from the data.

- Detailed segment profiles and recommendations for business strategies.

**\*\*Success Criteria:\*\***

- Clearly defined and interpretable customer segments.

- Actionable insights that can be translated into business strategies.

- Effective communication of findings through visualizations and documentation.

- Demonstrated improvement in marketing or customer-related KPIs based on implemented recommendations.