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:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

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()
```

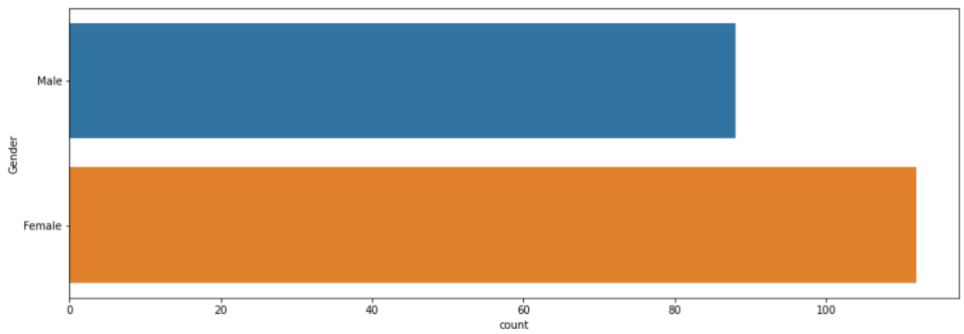
2. **Feature Engineering:**

- 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:**

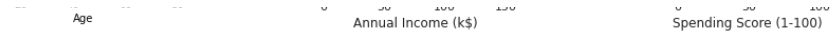
- Apply clustering algorithms (e.g., k-means, hierarchical clustering) to segment customers based on their characteristics.

- 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

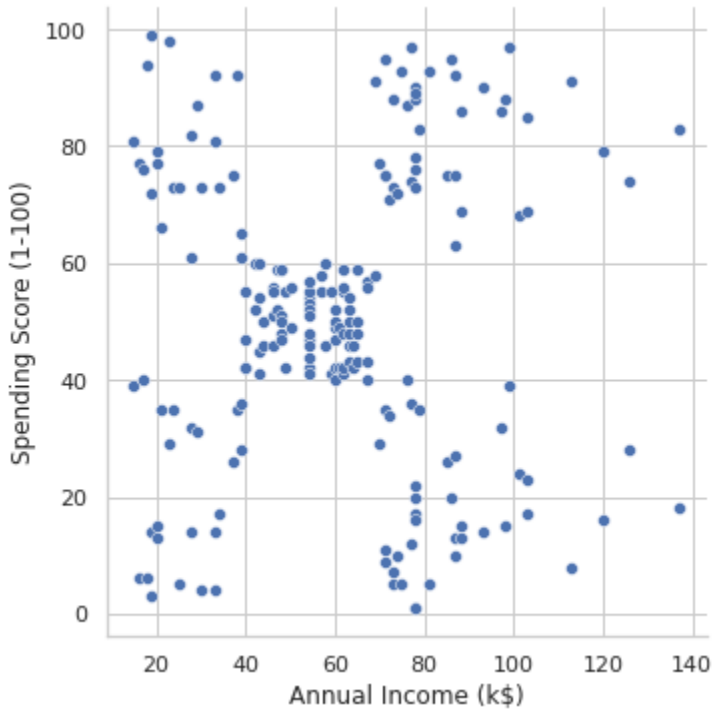
```
ss_1_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) &
(df["Spending Score (1-100)"] <= 20)]

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.v
alues),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.