Customer Segmentation using Data Science

TEAM MEMBER: Sakthibala A

NM ID : aut22leai03

Introduction:

- It is the process of grouping customers according to how and why they are buy products.
- 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.
- The main goal for the customer segmentation using data science is to divide the customer base into distinct groups based on similar characteristics.
- This segment will helpful for many Business purpose.

Project Phase 3:

Building the project by loading and preprocessing the dataset(Mall Dataset).

Dataset:

Dataset link :(https://www.kaggle.com/datasets/akram24/mall-customers)

CustomerI	Genre	Age	Annual Inc	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29
20	Female	35	23	98
21	Male	35	24	35
22	Male	25	24	73
23	Female	46	25	5
24	Male	31	25	73
25	Female	54	28	14
26	Male	29	28	82
27	Female	45	28	32
28	Male	35	28	61

Data Loading :

Data loading is the process of copying and loading data or data sets from a source file, folder or application to a database or similar application.

Program :

Data Loading

```
import pandas as pd
df = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
data.head(5)
```

Output :

	CustomerID	Genre	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

Data Preprocessing :

Data preprocessing is a important step in data analysis and machine learning. It involves cleaning, integrating, and transforming raw data into a format suitable for analysis or model training. Data preprocessing is essential because real-world data is often incomplete and inconsistent. By preparing the data properly, we can improve the quality and reliability of our analysis or machine learning models.

Techniques :

1) Data Cleaning
2) Data Transformation
3) Data Reduction

Data Cleaning :

- Data cleaning is also known as data cleansing or data scrubbing. It is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset. It is a crucial step in data preprocessing, necessary to ensure that the data used for analysis or machine learning is accurate, reliable, and free from noise.
- **Handling Missing Data:** Identifying and dealing with missing values in the dataset. This can involve filling in missing values, removing rows or columns with too many missing values
- **Removing Duplicates:** Identifying and removing duplicate records or observations from the dataset.

Program :

Correcting “Gender” Column name

```
import pandas as pd
df = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
df=df.rename(columns={'Genre':'Gender'})
```

df.head()

Checking the null values

```
import pandas as pd
df = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
df.isnull().sum()
```

Output :

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

```
CustomerID                0
Genre                     0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

Data Transformation :

Data transformation is the context of data analysis, refers to the process of converting data from one format, structure, or representation into another. This transformation is performed to make the data suitable for analysis, reporting, visualization, or modeling.

Program & Output :

Describing the dataset

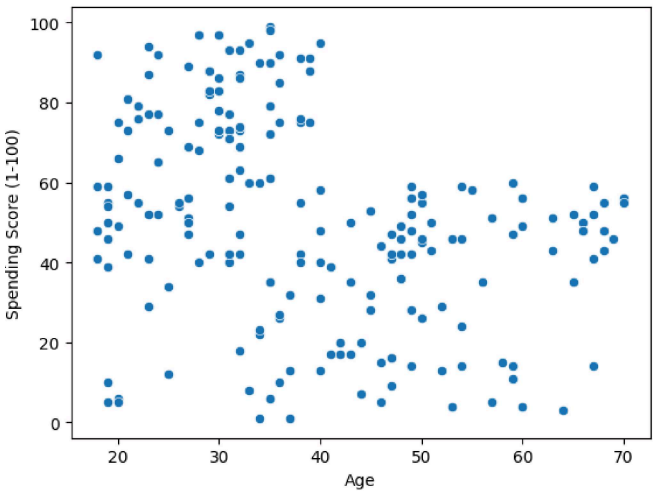
```
import pandas as pd
df = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
df.describe()
```

	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

Scatter Plot

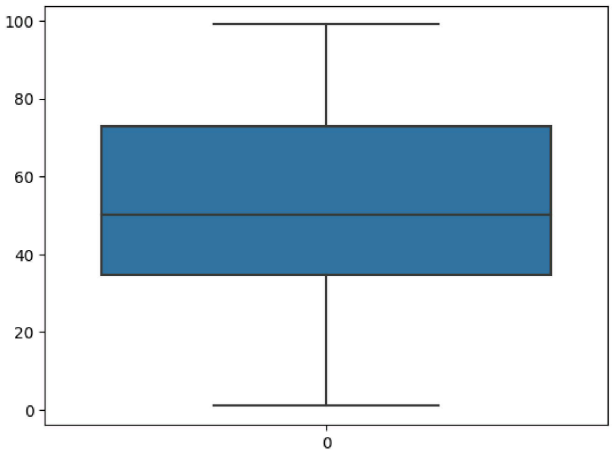
```
import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
sns.scatterplot(x='Age',y='Spending Score (1-100)',data=df)
```



Box Plot

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
sns.boxplot(df['Spending Score (1-100)'])
```



Data Reduction :

Data reduction is a process used in data analysis and data mining to decrease the volume but produce the same or similar analytical results.It is also called as “Dimensionality Reduction”.

Program & Output :

Reducing the Column

```
import pandas as pd
mall_data = pd.read_csv("C:\\Users\\Student\\Downloads\\archive (1)\\Mall_Customers.csv")
selected_features = ["CustomerID", "Age", "Genre"]
reduced_mall_data = mall_data[selected_features]
reduced_mall_data.head()
```

	CustomerID	Age	Genre
0	1	19	Male
1	2	21	Male
2	3	20	Female
3	4	23	Female
4	5	31	Female

conclusion :

Data Loading and Data Preprocessing provides more ability to the Mall_customers dataset.It reduce the unnecessary rows or columns,null values and noisy data.The model cannot provide correct accuracy with the null values.The model should be perform on well dataset.These steps are increasing the accuracy of the model.