Customer Segmentation using k-means and DBSCAN clustering algorithm

Machine Learning

Mini Project Report

Subject Code - CSL701

Semester VII

Department of Computer Engineering

Academic Year 2023-2024

(In accordance with University of Mumbai syllabus)

Course: BE - Computer Engineering

Name: Sakshi Dond

Content

Chapter 1
Introduction

Arthur Samuel, a pioneer in the field of artificial intelligence and computer gaming, coined the term “Machine Learning”. He defined machine learning as – a “Field of study that gives computers the capability to learn without being explicitly programmed”. Machine Learning is a branch of artificial intelligence that develops algorithms by learning the hidden patterns of the datasets used to make predictions on new similar type data, without being explicitly programmed for each task.

Machine learning history starts in 1943 with the first mathematical model of neural networks presented in the scientific paper "A logical calculus of the ideas immanent in nervous activity" by Walter Pitts and Warren McCulloch.Then, in 1949, the book The Organization of Behavior by Donald Hebb is published. The book had theories on how behavior relates to neural networks and brain activity and would go on to become one of the monumental pillars of machine learning development.In 1950 Alan Turing created the Turing Test to determine if a computer has real intelligence. To pass the test, a computer must be able to fool a human into believing it is also human.

Will be using Customer Segmentation in the retail industry, a Mall, to segment customers into various groups and target potential. The industry can then work towards attractive ideas to sell products and services inclined towards these specific customers

Segmenting customers goes beyond putting people into categories. When you segment customers, you learn about them deeply and use that info to create content for each segment's unique needs and challenges.Segmenting can improve your customer service and support efforts and help internal teams prepare for challenges different groups are likely to experience.It also allows you to communicate with segments of customers through preferred channels or platforms, and help you find new opportunities for products, support, and service efficiently.

Chapter 2
Literature Review

In the paper[1], presented by Patel Monil,Pyla. Srinivas Dileep, Dr. M. Seshashayee, Customer segmentation is done by using K-means clustering algorithm.By using clustering techniques, customers with similar means, end and behavior are grouped together into homogeneous clusters.Certain parameters are considered while segmenting the customer. The clustering parameters can broadly be classified as geographic, demographic, psychographic and behavioural

In the paper[2] represented by Pyla. Srinivas Dileep, Dr. M. Seshashayee,Mall Customer segmentation is done by using K-means clustering algorithm.They used a clustering approach called K-means clustering, in particular. Kmeans clustering is one of the most popular clustering methods, and it's frequently the first thing practitioners try when they're working on a clustering problem.

In the paper[3] presented by Musthofa Galih Pradana, Hoang Thi Ha,they have divided the customer into 5 clusters based on the relationship between annual income and their spending score, and it has been concluded that customers who have high-income levels & have a high spending score are also very appropriate targets for implementing market strategies.

Chapter 3
Methodology

Clustering:

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

Various clustering algorithms available:

1. K-Means algorithm:
2. Mean-shift algorithm
3. DBSCAN Algorithm
4. Agglomerative Hierarchical algorithm
5. Affinity Propagation

K-means clustering algorithm:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

The working of the K-Means algorithm is explained in the below steps:

- Step-1: Select the number K to decide the number of clusters.
- Step-2: Select random K points or centroids. (It can be other from the input dataset).
- Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
- Step-4: Calculate the variance and place a new centroid of each cluster.
- Step-5: Repeat the third steps, which means re-assign each datapoint to the new closest centroid of each cluster.
- Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
- Step-7: The model is ready.

Density-Based Spatial Clustering Of Applications With Noise (DBSCAN):

Clusters are dense regions in the data space, separated by regions of the lower density of points. The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Parameters Required For DBSCAN Algorithm

- eps: It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to ‘eps’ then they are considered neighbors. If the eps value is chosen too small then a large part of the data will be considered as an outlier. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the k-distance graph.
- MinPts: Minimum number of neighbors (data points) within eps radius. The larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. The minimum value of MinPts must be chosen at least 3.

In this algorithm, we have 3 types of data points.

Core Point: A point is a core point if it has more than MinPts points within eps.

Border Point: A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.

Noise or outlier: A point which is not a core point or border point.

Steps Used In DBSCAN Algorithm

- Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.
- For each core point if it is not already assigned to a cluster, create a new cluster.
- Find recursively all its density-connected points and assign them to the same cluster as the core point.
- Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

Chapter 4
Exploratory Data Analysis

Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

EDA makes it simple to comprehend the structure of a dataset, making data modelling easier. The primary goal of EDA is to make data 'clean' implying that it should be devoid of redundancies. It aids in identifying incorrect data points so that they may be readily removed and the data cleaned. Furthermore, it aids us in comprehending the relationship between the variables, providing us with a broader view of the data and allowing us to expand on it by leveraging the relationship between the variables. It also aids in the evaluation of the dataset's statistical measurements.

Outliers or abnormal occurrences in a dataset can have an impact on the accuracy of machine learning models. The dataset might also contain some missing or duplicate values. EDA may be used to eliminate or resolve all of the dataset's undesirable qualities.

Data preprocessing and cleansing are critical components of EDA. Understanding the variables and the structure of the dataset is the initial stage in this process. The data must then be cleaned. The dataset may contain redundancy such as irregular data, missing values or outliers that may cause the model to overfit or underfit during training. Removing or resolving these redundancies is known as data cleaning. The last part is analysing the relationship between the variables.

Importance of using EDA for analyzing data sets is:

- Helps identify errors in data sets.
- Gives a better understanding of the data set.
- Helps detect outliers or anomalous events.
- Helps understand data set variables and the relationship among them.

Libraries and Functions

- Pandas library :

  pandas is an open-source library built on top of numpy providing high performance, easy to use data structures and data analysis tools for python. It allows for fast analysis and data cleaning and preparation.

- Numpy library:

  NumPy is a library for Python that adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

- Matplotlib library :

  Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

- Seaborn library :

  Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Chapter 5

Implementation

```
#importing libraries

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as pl
from mpl_toolkits.mplot3d import Axes3D
```

```
# loading the datasets
df = pd.read_csv('/kaggle/input/customer-segmentation-tutorial-in-python/Mall_Customers.csv')
df.head()
```

	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

```
df.info()
```

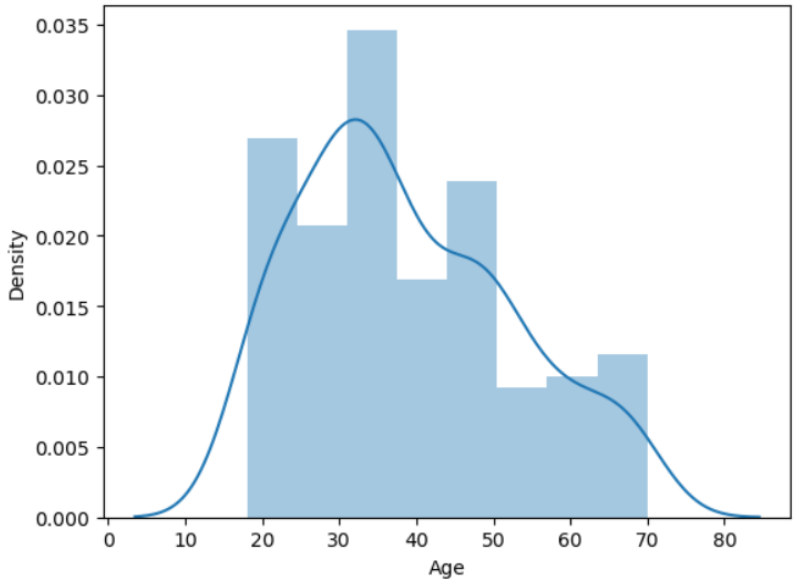
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

```
# getting stastical details about data
df.describe()
```

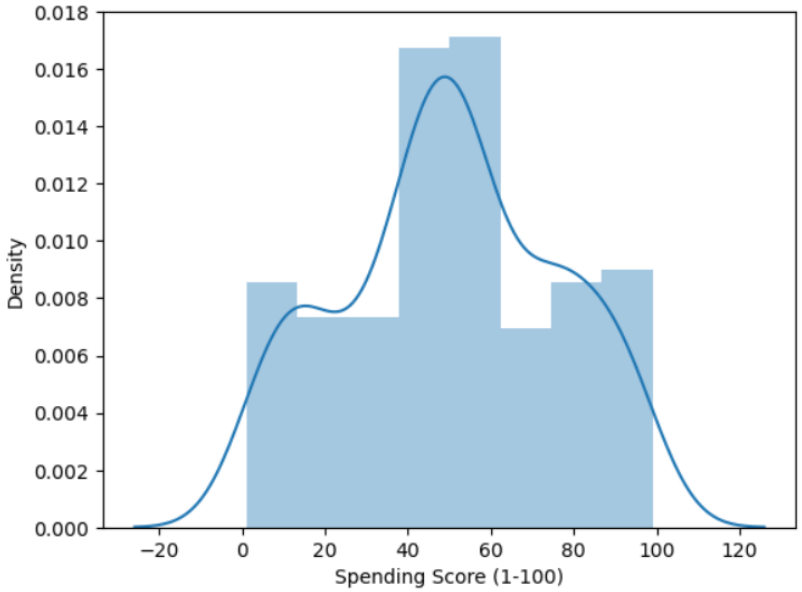
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000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

EDA:

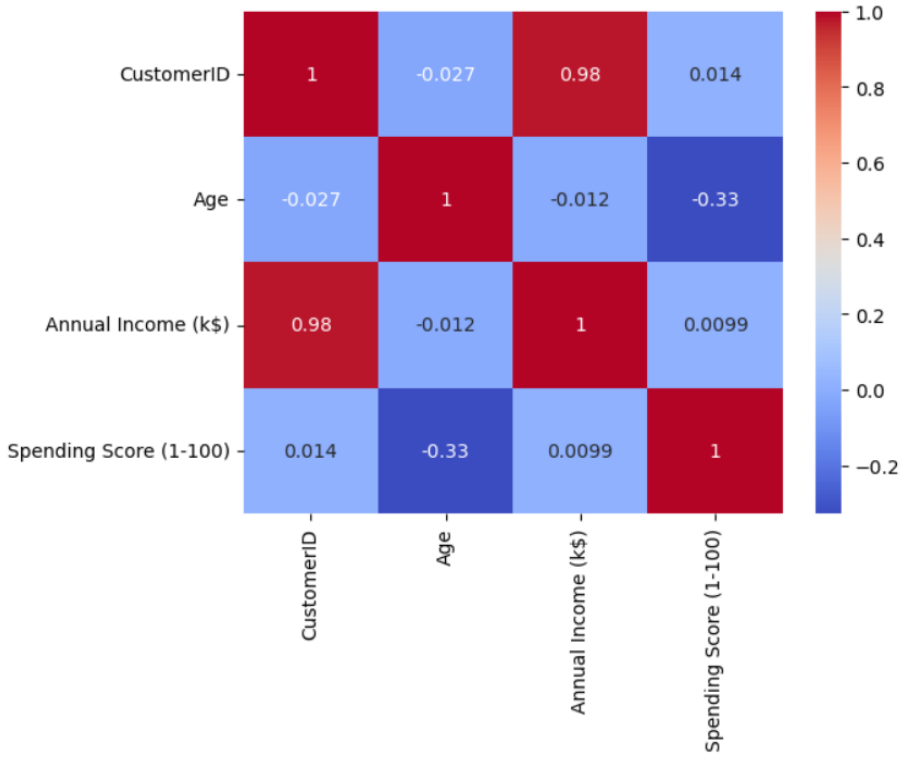
```
sns.distplot(df['Age'])
```




```
sns.distplot(df['Spending Score (1-100)'])
```




```
# finding correlation of all columns in the dataset
corr = df.corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
```




```
# Dropping the least correlated columns
df=df.drop(['CustomerID','Gender','Age'],axis=1)
```

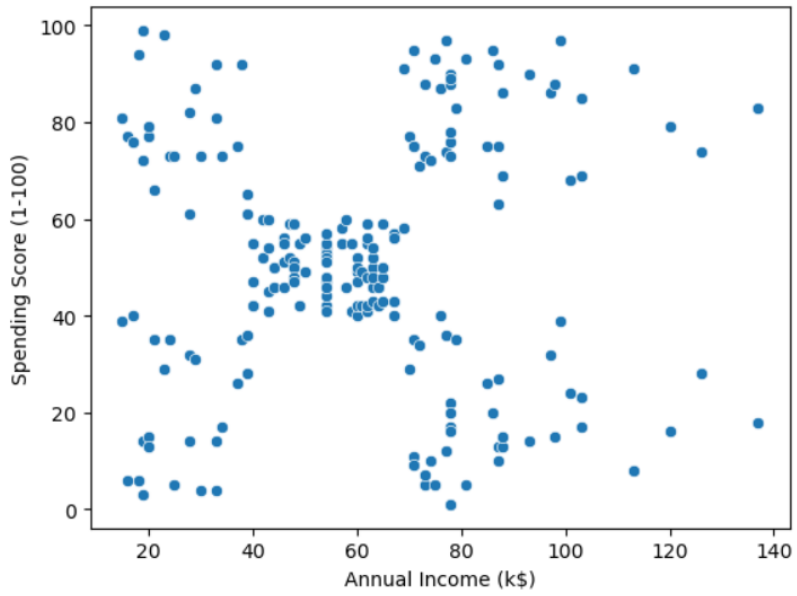
+ Code + Markdown

```
df.head()
```

	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

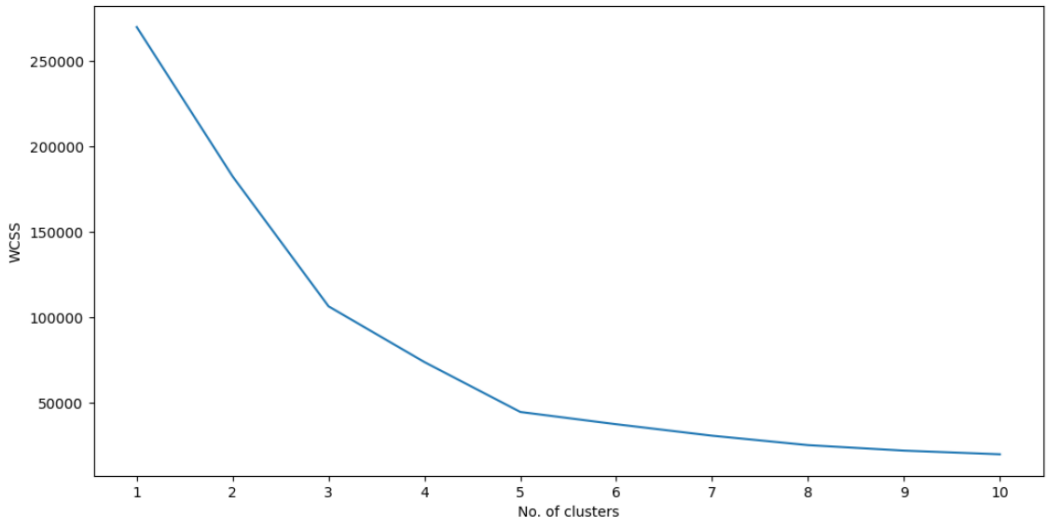
```
# Creating Scatter plots
sns.scatterplot(x=df['Annual Income (k$)'], y=df['Spending Score (1-100)'])
```

```
<Axes: xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```




```
# importing kmeans and finding elbow point
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(df)
    wcss.append(kmeans.inertia_)
```

```
import matplotlib.pyplot as plt
plt.figure(figsize=(12,6))
plt.plot(range(1,11), wcss)
plt.xlabel('No. of clusters')
plt.ylabel('WCSS')
plt.xticks(np.arange(1,11,1))
plt.show()
```



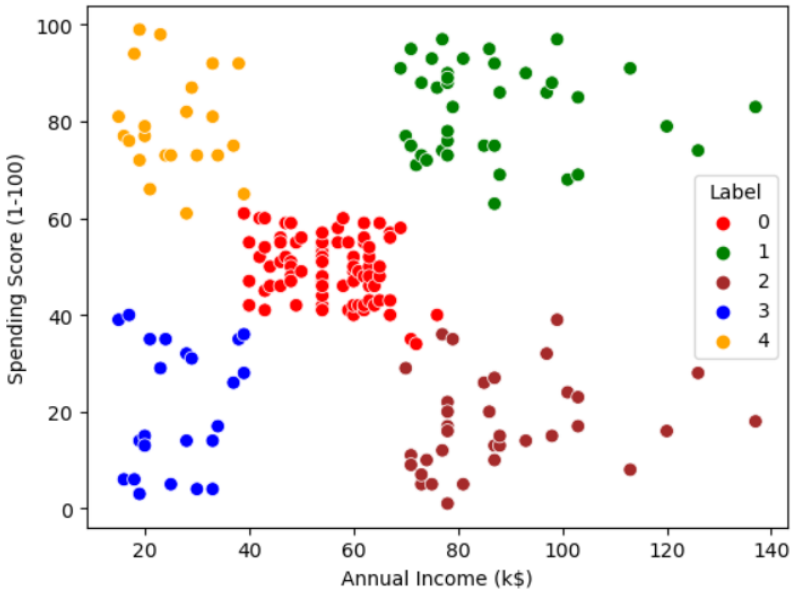
```
# making clusters and labelling them
km = KMeans(n_clusters=5)
km.fit(df)
y = km.predict(df)
df['Label'] = y
df.head()
```

```
/opt/conda/lib/python3.10/site-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will cha
nge from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```

	Annual Income (k$)	Spending Score (1-100)	Label
0	15	39	3
1	15	81	4
2	16	6	3
3	16	77	4
4	17	40	3

```
sns.scatterplot(x='Annual Income (k$)', y='Spending Score (1-100)', data=df, hue='Label', s=50, palette=[
```

```
<Axes: xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



```
from sklearn.cluster import DBSCAN
```

```
X=df[['Annual Income (k$)','Spending Score (1-100)']]
X.head()
```

	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

```
# Converting Dataframe into a numpy array
X=np.array(X)
```

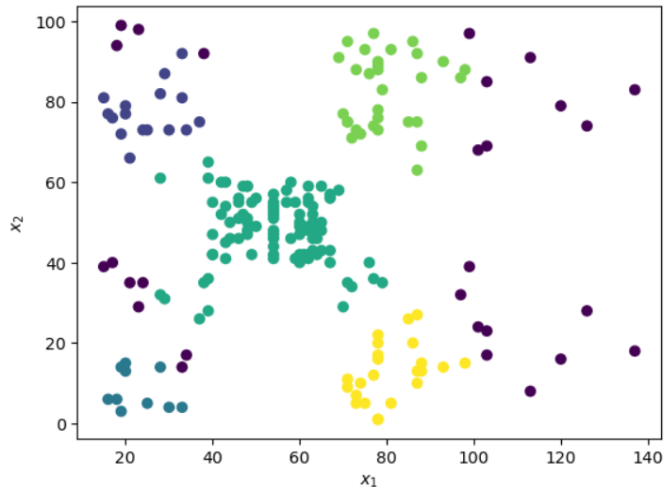
```
dbscan_cluster = DBSCAN(eps=12, min_samples=10)
dbscan_cluster.fit(X)

# Visualizing DBSCAN
plt.scatter(X[:, 0],
X[:, 1],
c=dbscan_cluster.labels_,
label=y)
plt.xlabel("$x_1$")
plt.ylabel("$x_2$")

# Number of Clusters
labels=dbscan_cluster.labels_
N_clus=len(set(labels))-(1 if -1 in labels else 0)
print('Estimated no. of clusters: %d' % N_clus)

# Identify Noise
n_noise = list(dbscan_cluster.labels_).count(-1)
print('Estimated no. of noise points: %d' % n_noise)
```

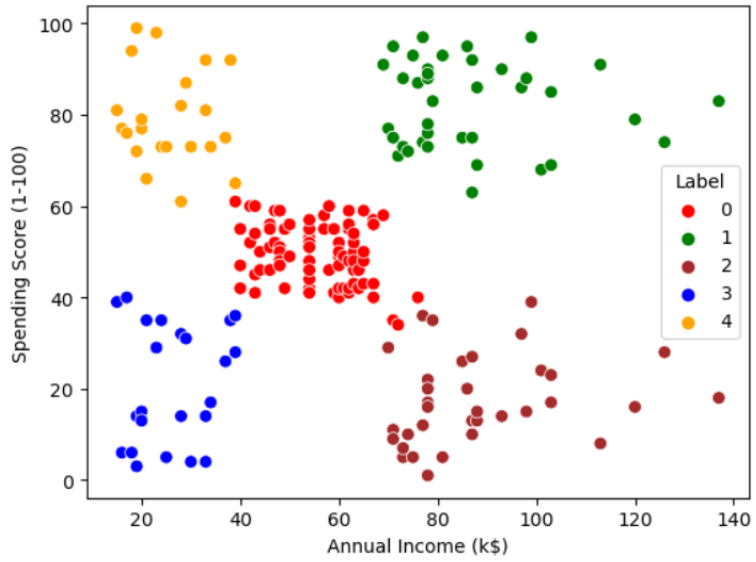
```
Estimated no. of clusters: 5
Estimated no. of noise points: 28
```



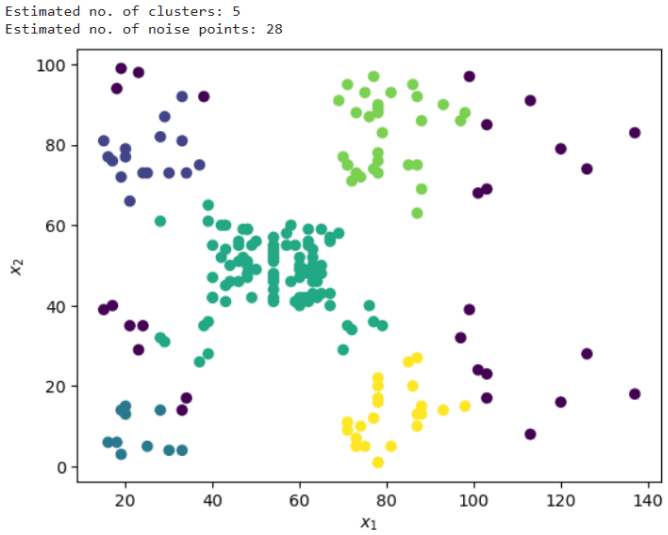
Chapter 6

Results & Analysis

Output (k-means): It can be seen from the below scatter plot that there are 5 group of customers based on their Annual income and spending score.



Output (DBSCAN): It can be seen from the below scatter plot that our DBSCAN clustering model has estimated 5 clusters and the estimated number of noise points are 28.



Chapter 7
Conclusion

In this project, segments of customers are created using the k-means and DBSCAN clustering algorithm. Visualization of the data set has been done for a better understanding about all the elements and its relation between the data. We used a clustering approach called K-means clustering and DBSCAN clustering in particular. Kmeans clustering is one of the most popular clustering methods, and it's frequently the first thing practitioners try when they're working on a clustering problem. K- means are used to divide data points into discrete, non overlapping groupings. One of the most common uses of K-means clustering is client segmentation in order to gain a better understanding of them, which can then be used to boost the company's income.It also helps them in maintaining customer relationships and customer retention by executing different marketing strategies.

References

[1] Patel Monil , Patel Darshan , Rana Jecky, Chauhan Vimarsh , Prof. B. R. Bhatt, "Customer Segmentation using Machine Learning," International Journal for Research in Applied Science & Engineering Technology (IJRASET) ,Volume 8, June 2020, http://doi.org/10.22214/ijraset.2020.6344

[2] Pyla. Srinivas Dileep, Dr. M. Seshashayee,"CUSTOMER SEGMENTATION USING MACHINE LEARNING",International Research Journal of Modernization in Engineering Technology and Science,Volume:04,May-2022

[3] Musthofa Galih Pradana, Hoang Thi Ha,"MAXIMIZING STRATEGY IMPROVEMENT IN MALL CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING",Journal of Applied Data Sciences,Vol 2,2021, https://doi.org/10.47738/jads.v2i1.18