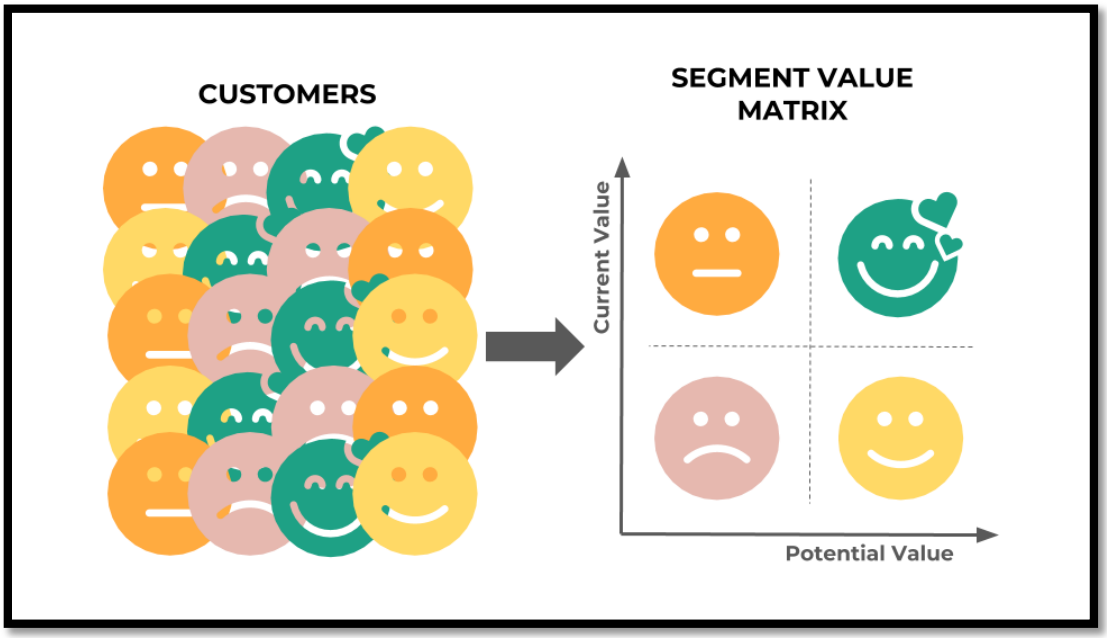
EXPOSYS DATA LABS INTERNSHIP

DATA SCIENCE PROJECT REPORT

Customer Segmentation Using KMeans.



BY:

Mamtha M V

BE,CSE(4th year)

CMRIT,Bangalore

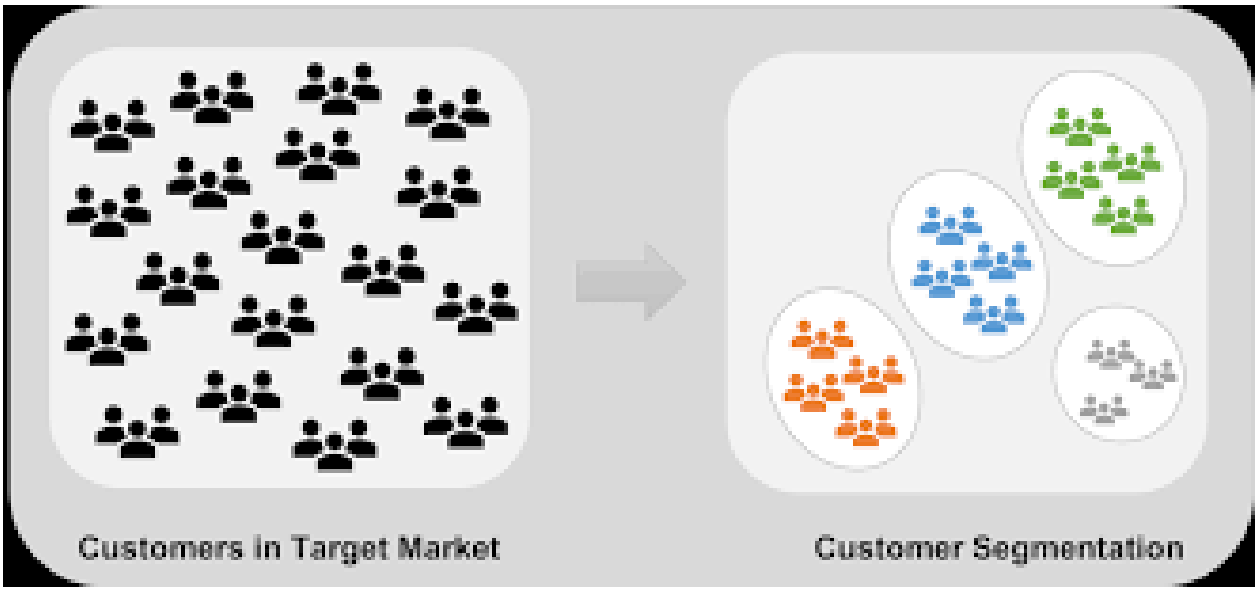
mamt17cs@cmrit.ac.in

ABSTRACT:

The process of grouping customers into sections of individuals who share common characteristics is called **Customer Segmentation**.

This segmentation enables marketers to create targeted marketing messages for a specific group of customers which increases the chances of the person buying a product. It allows them to create and use specific communication channels to communicate with different segments to attract them.

A simple example would be that the companies try to attract the younger generation through social media posts and older generation with maybe radio advertising. This helps the companies in establishing better customer relationships and their overall performance as an organisation.



Application of machine learning technique called k-means clustering can be used to solve this challenge. The algorithm can be used to group different potential customers on the basis of certain fairly similar characteristics. The outcome of the k-means algorithm are clusters that basically represent different customer classes on certain grounds.

TABLE OF CONTENTS:

INTRODUCTION:

In the domain of Machine Learning, Customer Segmentation is a popular problem that falls under the category of unsupervised learning. In this project, customer segmentation is implemented using KMeans clustering algorithm.

Consumer segmentation is the process of separating customer base into several groups of individuals that share a resemblance in various ways that are relevant to marketing, such as gender, age, preferences and diverse spending habits.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

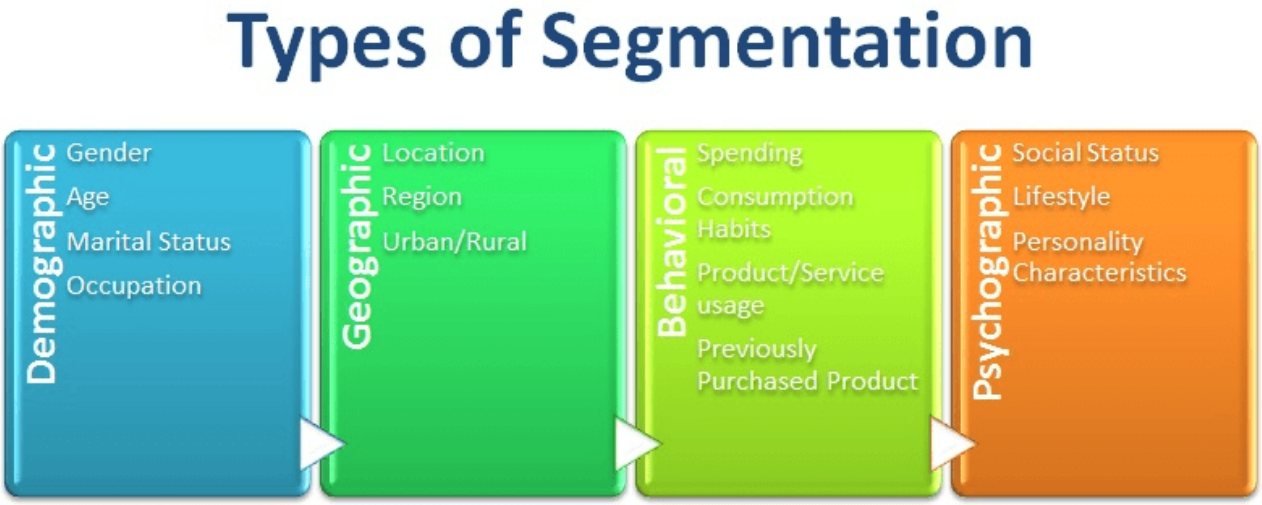
Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above data companies can then outperform the competition by developing uniquely appealing products and services.

A customer segmentation model allows for the effective allocation of marketing resources and the maximisation of cross and up-selling opportunities.

EXISTING METHODS:

The most common ways in which businesses segment their customer base are:

1. **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.
2. **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer's city, state, or even country of residence.
3. **Psychographics**, such as social class, lifestyle, and personality traits.
4. **Behavioral data**, such as spending and consumption habits, product/service usage, and desired benefits.



Apart from using KMeans clustering for performing Customer Segmentation. This can be readily performed using the below mentioned clustering algorithms as well:

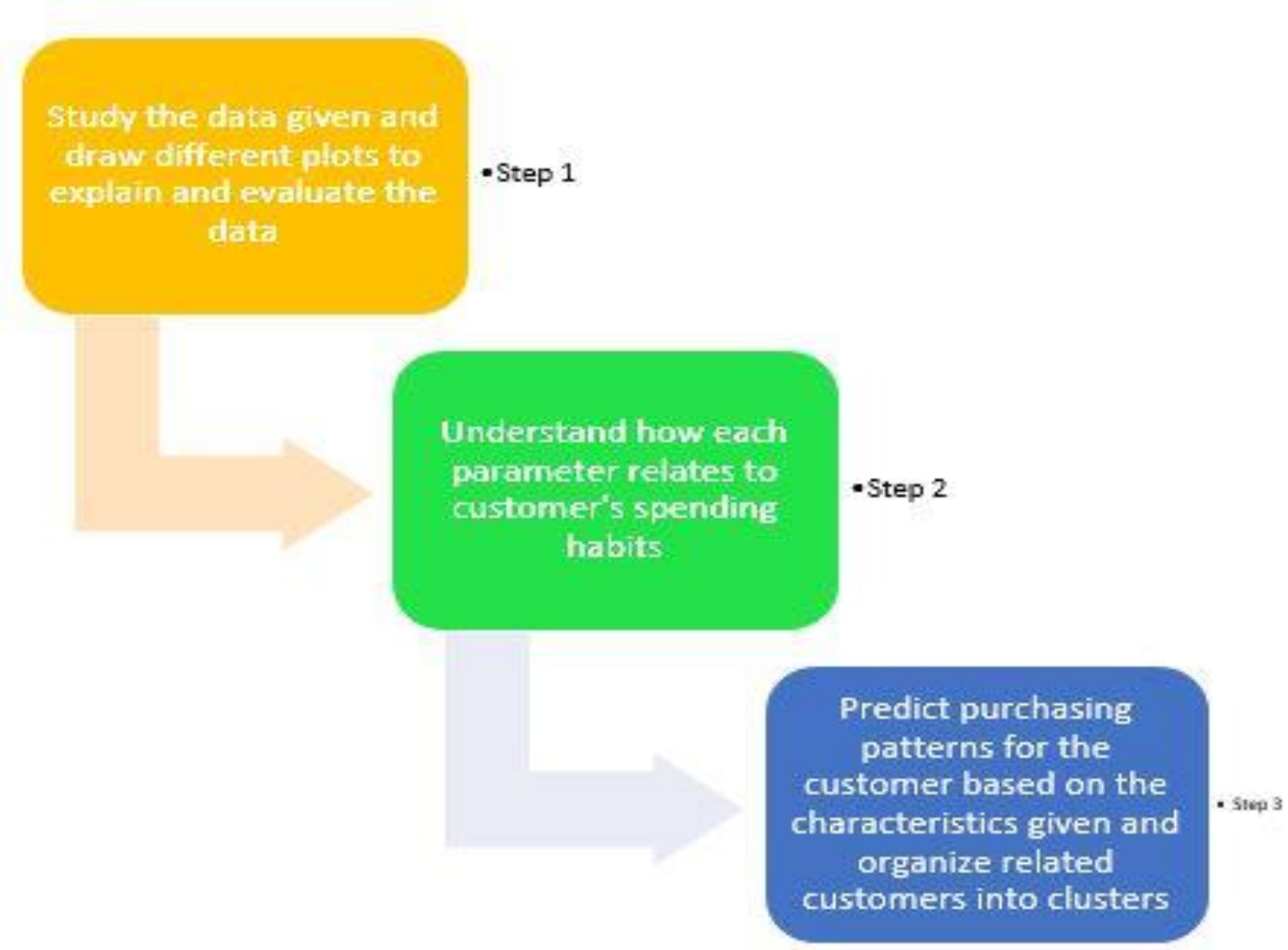
1) Hierarchical Clustering

2) DBSCAN(Density Based Spatial Clustering of Applications of Noise) Clustering

PROPOSED METHOD WITH ARCHITECTURE:

Customer Segmentation using Unsupervised Learning:-

K-Means clustering algorithm has been chosen for this task. Since it is simple and is apt for this task since it measures the distance between two observations to assign a cluster. This algorithm will help us in separating the general population with the help of the reduced features into a specified number of clusters and use this cluster information to understand the similarities in the general population and customer data. The number of clusters is selected can be decided with the help of an elbow plot.



ENVIRONMENT AND TOOLS :

- scikit-learn, version: 0.23.2
- seaborn, version 0.10.0
- numpy, version 1.18.1
- pandas, version 1.0.1
- matplotlib
- python , version 3
- anaconda
- jupyter notebook
- google colab

METHODOLOGY:

K Means Clustering Algorithm

1) Specify number of clusters K.
2) Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3) Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

Kmeans clustering where k=3

In order to determine k, we make use of Elbow plot. The value of k is equal to the elbow of the curve i.e, the point at which there is no longer a steep decrease in the sum of squared errors within a cluster.

IMPLEMENTATION:

The given dataset contains data of 200 customers of a mall.

The dataset includes the customerID, genre, age, annual income and spending score of each customer.

- Import appropriate libraries and dependencies needed.
- Read the Mall_customers.csv file using pandas.
- View the information after importing.

```
[1] import numpy as np
    import matplotlib.pyplot as plt
    import pandas as pd
    import seaborn as sns
    import os
```

```
/usr/local/lib/python3.6/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the functions i
  import pandas.util.testing as tm
```

```
#importing dataset of mall customers given
cusdata= pd.read_csv('Mall_Customers.csv')
cusdata.info()
```

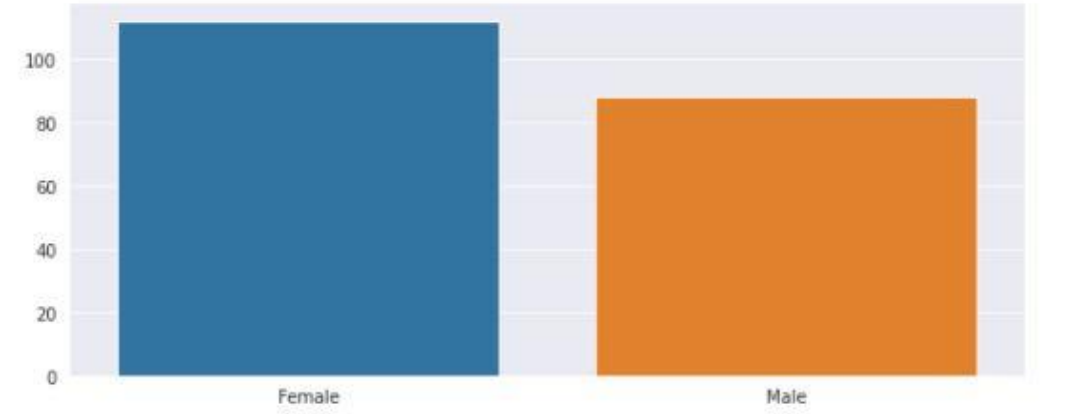
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    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
```

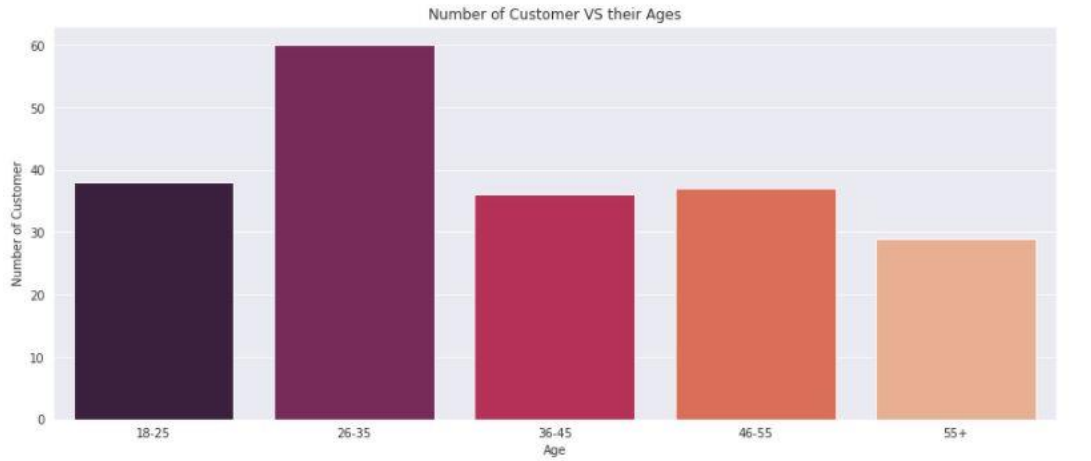
- View the first 5 entries to understand the data values.

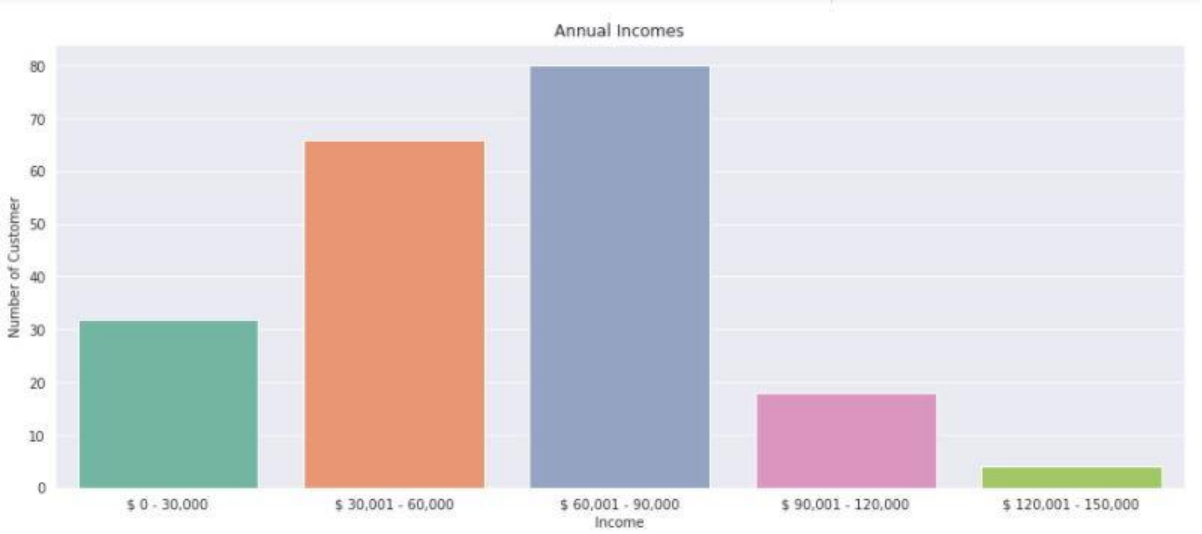
```
[3] cusdata.head()
```

	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

- Visualize the distribution of customers based on age, gender, annual income, spending score etc. effectively by using graphing and visualization libraries like matplotlib and seaborn.





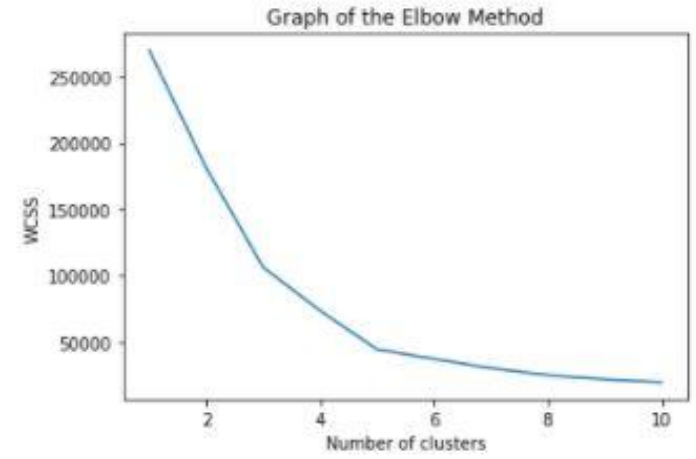


- Identify the features that are essential to build a KMeans clustering model namely Age , Gender, Annual Income and Spending Score.

- Find the k in KMeans using Elbow Plot Method.

```
for i in range(1,11):
  kmeans = KMeans(n_clusters=i, init ='k-means++',max_iter=300,n_init=10,random_state=0 )
  kmeans.fit(X)
  wcss.append(kmeans.inertia_)
```

```
plt.plot(range(1,11),wcss)
plt.title('Graph of the Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



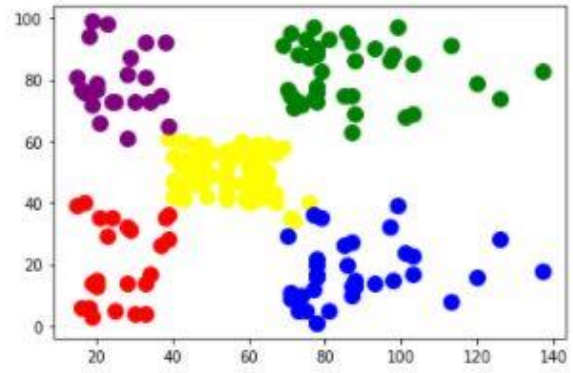
- From above plot we can determine that value of k=5 and build kmeans model for the same.

- Visualization of the clusters using scatter plots.

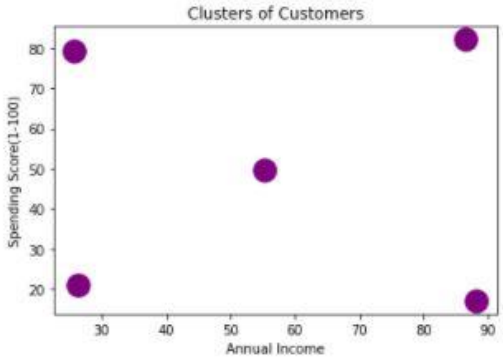
```
kmeans = KMeans(n_clusters=5, init ='k-means++',max_iter=300,n_init=10,random_state=0 )
y_kmeans = kmeans.fit_predict(X)
```

```
#visualizing the clusters obtained
plt.scatter(X[y_kmeans==0, 0], X[y_kmeans==0, 1], s=100, c='yellow', label ='Cluster 1')
plt.scatter(X[y_kmeans==1, 0], X[y_kmeans==1, 1], s=100, c='purple', label ='Cluster 2')
plt.scatter(X[y_kmeans==2, 0], X[y_kmeans==2, 1], s=100, c='green', label ='Cluster 3')
plt.scatter(X[y_kmeans==3, 0], X[y_kmeans==3, 1], s=100, c='red', label ='Cluster 4')
plt.scatter(X[y_kmeans==4, 0], X[y_kmeans==4, 1], s=100, c='blue', label ='Cluster 5')
```

```
<matplotlib.collections.PathCollection at 0x7fdc34c01cc0>
```




```
#plotting the centroid
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=300, c='purple', label ='Centroids')
plt.title('Clusters of Customers')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score(1-100)')
plt.show()
```



CONCLUSION:

The general population and customer population have been compared and segmented using an Unsupervised learning algorithm. We were able to determine which clusters have more customers and which are potential clusters to have probable customers.

The key inferences of customers in each cluster are:

1) **Cluster 1**: Earning moderate Annual Income and display moderate spending habits.

2) **Cluster 2**: Earning high Annual Income and display low spending habits.

3) **Cluster 3**: Earning low Annual Income and display low spending habits.

4) **Cluster 4**: Earning low Annual Income and display high spending habits.

5) **Cluster 5**: Earning high Annual Income and display high spending habits.

Thus this helps to get a better understanding of customers which in turn could be used to increase the revenue of the company.