# A PROJECT REPORT ON “CUSTOMER SEGMENTATION USING CLUSTERING”

EXPOSYS
DATA LABS

DEVELOPED AND SUBMITTED BY
GORANTLA SAI SATHVIK

# ABSTRACT

- Now-a-days,Customer Segmentation has become a very popular method for dividing customers in order to retain them and as well as making profits based on the segmentation.
- In the following study, customers have been classified on the basis of their behavioral characteristics such as spending score and annual income.
- For this classification, a machine algorithm named as k-means clustering algorithm is used and based on the behavioral characteristic’s, customers are classified.
- Formed clusters help the company to target individual customer and advertise the content to them through marketing campaigns and social media sites in which they’re usually active in.

# TABLE OF CONTENTS

# INTRODUCTION

- Clustering is a technique of grouping the similar items,which is widely used in unsupervised classification.
- K-means is one of the best clustering strategies among all partitioning based clustering strategies which is widely used in clustering due to its uniform effect on the clusters produced with relatively uniform size,though the input data has different size of clusters.
- It is well known as market segmentation, customer segmentation is the division of potential customers in a given market into discrete groups.
- That division is based on customers having similar needs and buying charactersitics.

# EXISTING METHODS

- CLUSTERING

- Clustering is one of the most common methods used in exploring data to obtain a clear understanding of the data structure.It can be characterized as the task of finding the subgroups in the complete dataset.
- Similar data is clustered in the same subgroup.A cluster refers to a collection of aggregated data points due to some similarities.
- Clustering is used in Market Basket Analysis used to segment the customers based on their behaviours.

# K-MEANS CLUSTERING ALGORITHM

K-Means clustering is the most common and simplest Machine learning algorithm and it follows an iterative Approach which attempts to partition the dataset into “K” Different number of predefined and non overlapping Subgroups where each data point belongs to only one Subgroup.

# PROPOSED METHOD

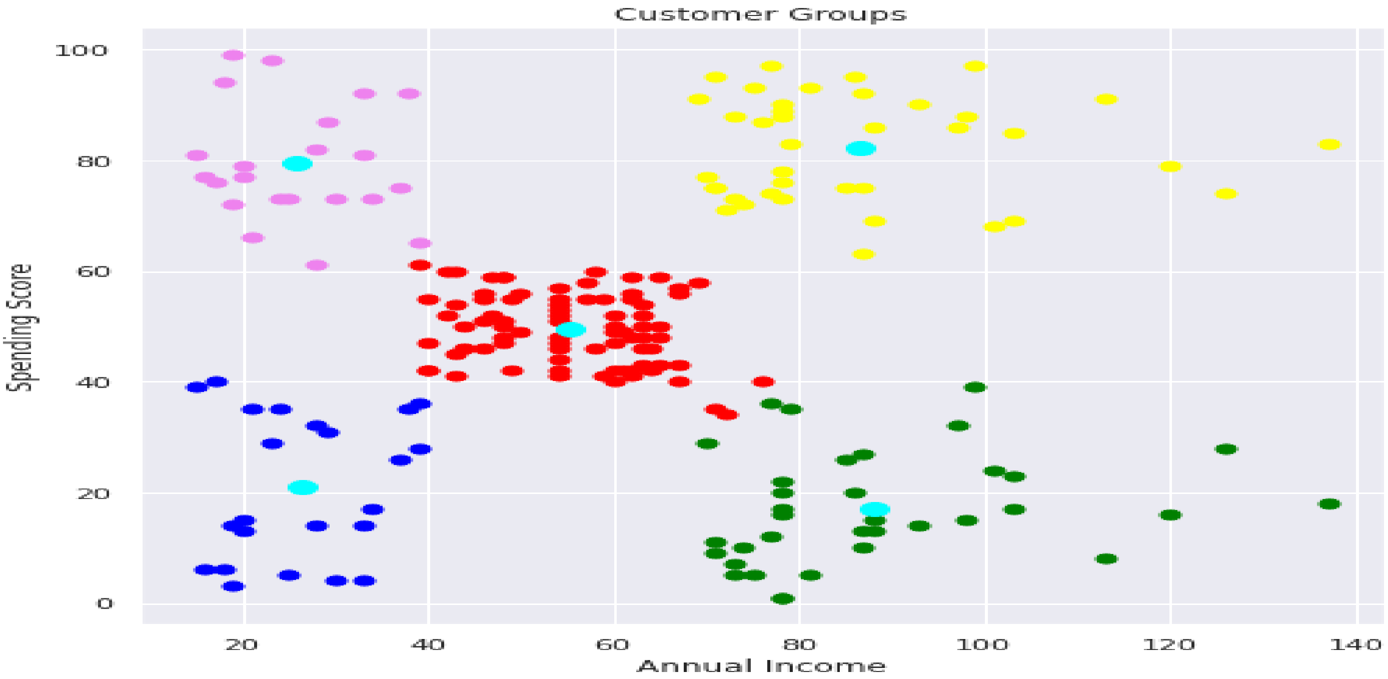
- It is a web application for segmentation of mall customers which is useful for the shopkeepers to market the product based on the planned strategy. The cluster which is generated by the application is stored in the image format.
- Login module:
- The marketing team logins the application using the username and  password and each time the marketing team should register with the system in order to access the details.
- Stimulus: Marketing team enters the username and password.
- Response: Navigates to the Mall Customer Segmentation Registration page.
- Register Module:
- The web application provides registration feature for the marketing team. After successfully logged into the system, each time the marketing team should register with the system in order to access the details.
- Stimulus: Marketing team enters the registration details.
- Response: Navigates to the Upload Dataset Module.

- MCS - Upload Dataset Module:
- The Marketing team is provided with MCS dataset feature in which the dataset will be imported and it gets navigated to the K-Means execution module.
- In this module MCS dataset will be imported and after the display imported successfully message it gets navigated to the K-Means Algorithm execution module.
- Stimulus: Marketing team imports the dataset.
- Response: The system provides the message dataset imported successfully.
- MCS – K-Means Algorithm Module:
- The K-Means algorithm will be executed that provides five different clusters which indicate the customers based on spending score and annual income.
- Stimulus: The marketing team presses the K-Means execution button.
- Response: It navigates to the visualization page

- Visualization Module:
- The visualization module provides the results based on the following clusters. The results are generated in the form of a graph and stored as image which is retrieved for the marketing team.

Stimulus: The marketing team presses the K-Means execution button

Response: The results are generated in the form of a graph and stored as image with clusters K=5.



# METHODOLOGY

## DATA VISUALIZATION

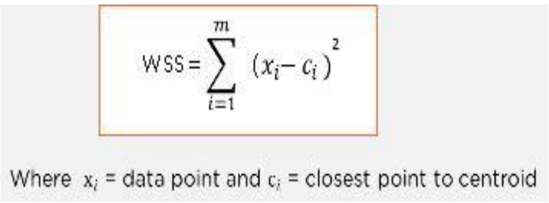
- Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.
- The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets
- Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.
- It provides an increased understanding of the next steps to be taken to improve the organization.

# • K-MEANS CLUSTERING

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset
- Here, we divide a data space into K clusters and assign a mean value to each. The data points are placed in the clusters closest to the mean value of that cluster. There are several distance metrics available that can be used to calculate the distance.

# • THE ELBOW METHOD

- The Elbow method is the best way to find the number of clusters. The elbow method constitutes running  K-Means clustering on the dataset.
- Next, we use within-sum-of-squares as a measure to find the optimum number of clusters that can be formed for a given data set. Within the sum of squares (WSS) is defined as the sum of the squared distance between each member of the cluster and its centroid.

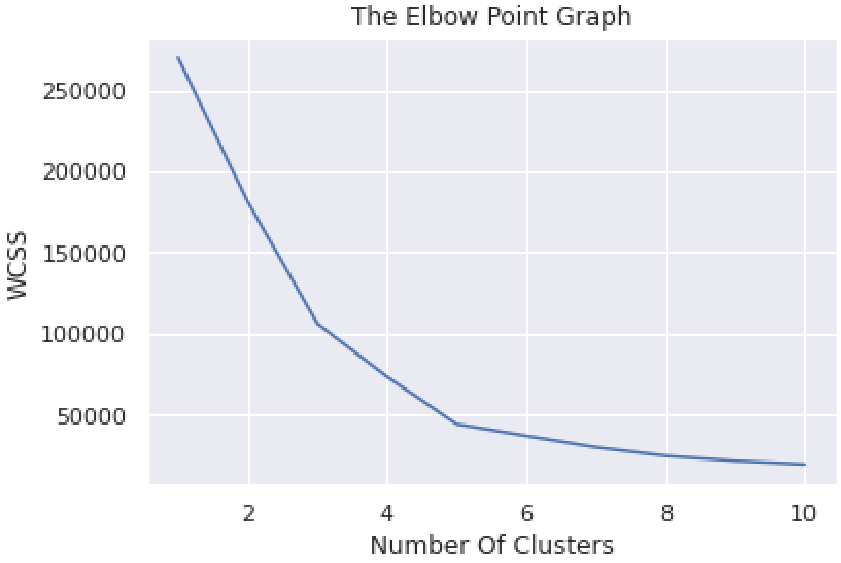


$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Where $x_i$ = data point and $c_i$ = closest point to centroid

- The WSS is measured for each value of K. The value of K, which has the least amount of WSS, is taken as the optimum value.

Now, we draw a curve between WSS and the number of clusters



- Here, WCSS is on the y-axis and number of clusters on the x-axis.
- You can see that there is a very gradual change in the value of WSS as the K value increases from 5
- So, you can take the elbow point value as the optimal value of K. It should be either five, six, or at most seven. But, beyond that, increasing the number of clusters does not dramatically change the value in WSS, it gets stabilized.

# IMPLEMENTATION

The implementation of code for Customer Segmentation using K-Means Clustering is as follows:

1. Importing The Dependencies

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- from sklearn.cluster import KMeans

2.Data Analysis And Collection

Loading the data from csv file to a Pandas DataFrame

customer_data = pd.read_csv('/content/Mall_Customers.csv’)

- first 5 rows in the dataframe
- customer_data.head()

The first 5 rows of the dataset are obtained

```
customer_data.shape
```

`(200,5)`–Indicating that there are 200 rows and 5 columns in the dataset.

## Checking for missing values

`customer_data.isnull().sum()`-There are no missing values in the dataset and it is of int64 type.

### Choosing the Annual Income Column & Spending Score column

```
X= customer_data.iloc[:,[3,4]].values
```

print(X)- Prints all the Annual Income and Spending Score column's values.

### Choosing the number of clusters

WCSS- Within Clusters Sum Of Squares

```
Finding wcss value for different number of clusters
```

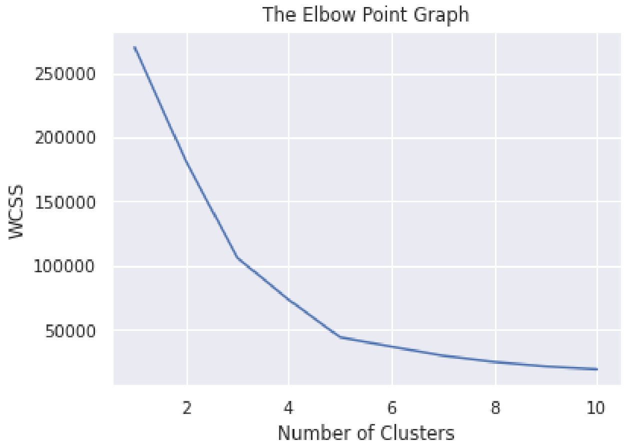
```
wcss = []

for i in range(1,11):
  kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
  kmeans.fit(X)

  wcss.append(kmeans.inertia_)
```

- Plot an elbow graph

```
sns.set()
plt.plot(range(1,11), wcss)
plt.title('The Elbow Point Graph')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```



Optimum Number of Clusters = 5

## Training the k-Means Clustering Model

```
kmeans = KMeans(n_clusters=5, init='k-means++', random_state=0)

# return a label for each data point based on their cluster
Y = kmeans.fit_predict(X)

print(Y)
```

- 5 Clusters - 0, 1, 2, 3, 4

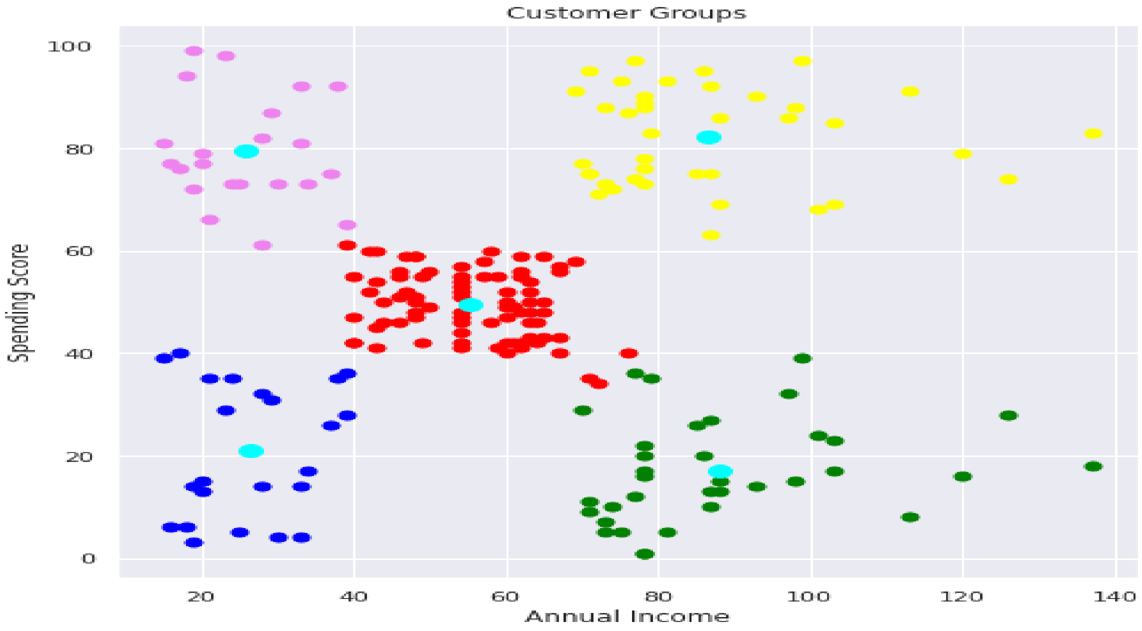
- Plotting all the clusters and their centroids

```
plt.figure(figsize=(8,8))
plt.scatter(x[Y==0,0], x[Y==0,1], s=50, c='green', label='Cluster 1')
plt.scatter(x[Y==1,0], x[Y==1,1], s=50, c='red', label='Cluster 2')
plt.scatter(x[Y==2,0], x[Y==2,1], s=50, c='yellow', label='Cluster 3')
plt.scatter(x[Y==3,0], x[Y==3,1], s=50, c='violet', label='Cluster 4')
plt.scatter(x[Y==4,0], x[Y==4,1], s=50, c='blue', label='Cluster 5')
```

# Plotting the centroids

```
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s=100, c='cyan', label='centroids')
```

```
plt.title('Customer Groups')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.show()
```



# CONCLUSIONS

- K-Means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset.
- The goal of K-Means is to group data points into distinct non-overlapping subgroups.
- One of the major application of K-Means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

# VERSIONS

- Python: 3.8.16.final.0
- pandas: 1.3.5
- numpy: 1.21.6
- seaborn: 0.11.2
- matplotlib: 3.2.2
- sklearn: 1.0.2