Customer Segmentation Using **K-Means** **Clustering**

Sravanthi Tulabandula

1. Abstract

Effective decisions are mandatory for any company to generate good revenue. In these days competition is huge and all companies are moving forward with their own different strategies. We should use data and take a proper decision. Every person is

different from one another and we don't know what he/she buys or what their likes are. But, with the help of machine learning technique one can sort out the data and can find the target group by applying several algorithms to the dataset. Without this, It will be very difficult and no better techniques are available to find the group of people with similar character and interests in a large dataset. Here, The customer segmentation using K-Means clustering helps to group the data with same attributes which exactly helps to business the best. We are going to use elbow method to find the number of clusters and at last we visualize the data.

2. Keywords

Clustering, Elbow Method, K-Means Algorithm, Customer Segmentation, Visualization.

3. Introduction

3.1 Introduction

Nowadays the competition is vast and lot of technologies came into account for effective growth and revenue generation. For every business the most important component is data. With the help of grouped or ungrouped data, we can perform some operations to find customer interests.

Data mining helpful to extract data from the database in a human readable format. But, we may not known the actual beneficiaries in the whole dataset. Customer Segmentation is useful to divide the large data from dataset into several groups based on their age, demographics, spent, income, gender, etc. These groups are also known as clusters. By this, we can get to know that, which product got huge number of sales and which age group are purchasing etc. And, we can supply that product much for better revenue generation.

Initially we are going to take the old data. As we know that old is gold so, by using the old data we are going to apply K-means clustering algorithm and we have to find the number of clusters first. So, at lastly, we have to visualize the data. One can easily find the potential group of data while observing that visualization. The goal of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm. The elbow method determines the optimal
clusters.

3.2 Problem Statement

Customer Segmentation is the best application of unsupervised learning. Using clustering, identify segments of customers in the dataset to target the potential user base. They divide customers into various groups according to common characteristics like gender, age, interest, and spending habits so they can market to each group effectively. Use K-Means Clustering and also visualize the gender and age distributions. Then analyze their annual income and spending scores. As it describes about how we can divide the customers based on their similar characteristics according to their needs by using k-means clustering which is a classification of unsupervised machine learning.

4. Existing System

The existing method is storing customer data through paperwork and computer software (digital data) is increasing day by day. At end of the day they will analyse their data as how many things are sold or actual customer count etc. By analyzing the collected data they got to know who is beneficial to their business and increase their sales. It requires more time and more paperwork. Also, it is not much effective solution to find the desired customers data.

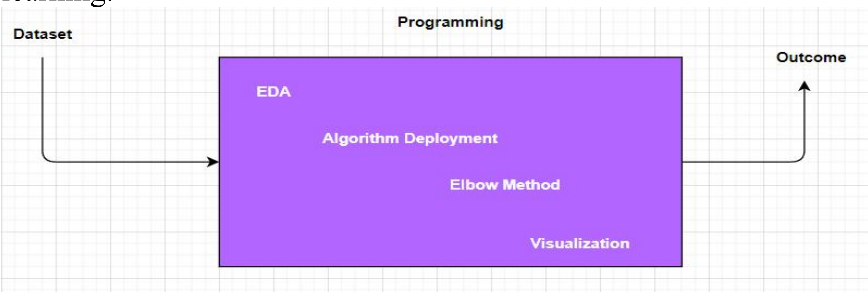
5. Proposed System

5.1 Proposed Method

To overcome the traditional method i.e paper work and computerized digital data this new method will play vital role. As we collect a vast data day by day which requires more paperwork and time to do. As new technologies were emerging in today's world. Machine Learning which is powerful innovation which is used to predict the final outcome which has many algorithms. So for our problem statement we will use K-Means Clustering which groups the data into different clusters based on their similar characteristics. And then we will visualize the data.

5.2 System Architecture

Initially we will see the dataset and then we will perform exploratory data analysis which deals with the missing data, duplicates values and null values. And then we will deploy our algorithm k-means clustering which is unsupervised learning in machine learning.



As in order to find the no of clusters we use elbow method where distance will be calculate through randomly chosen centers and repeat it until there is no change in cluster centers. Thereafter we will analyse the data through data visualization. Finally we will get the outcome.

5.3 Algorithm

5.3.1 K-Means Clustering

- K Means algorithm in an iterative algorithm that tries to partition the dataset into K predefined distinct non overlapping sub groups which are called as cluster.
- Here K is the total no of clusters.
- Every point belongs to only one cluster.
- Clusters cannot overlap.

5.3.2 Steps of Algorithm

- Arbitrarily choose k objects from D as the initial cluster centers.
- Repeat.
- Assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.

- Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
- Until no change.

6. Methodology

1. First of all we will import all the necessary libraries or modules (pandas, numpy, sea born).
2. Then we will read dataset and analyse whether it contains any null values, missing values and duplicate values. So we will fix them by dropping or fixing the value with their means, medians etc which is technically named as Data Preprocessing.
3. We will deploy our model algorithm K-Means Clustering, which divides the data into group of clusters based on similar characteristics. To find no.of clusters we will use elbow method.

4. Finally, we will visualize our data using matplotlib, which concludes the customers divided into groups who are similar to each other on their group.

7. Implementation And Analysis

7.1 Overview of a Dataset

This is a mall customer segmentation data which contains 5 columns and 200 rows.

	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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

7.2 Exploratory Data Analysis

It deals with the data preprocessing, whether it contains any missing values or null values. There after we will see the information and description of the dataset.

7.2.1 Information of the dataset

#customer_data.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
```

7.2.2 Description of the data

#df.describe()

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

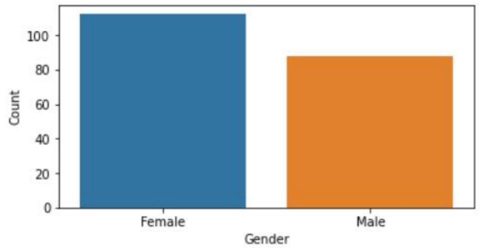
It describes about the count which counts the no of rows in it, mean of the columns, standard deviations, maximum and minimum and percentiles etc.

7.3 Gender plot Analysis

Here it overview the gender analysis

```
genders=customer_data['Gender'].value_counts()
plt.figure(figsize=(6,3))
sns.distplot(X=genders.index, y=genders.values)
plt.xlabel('Gender')
plt.ylabel('Count')
```

So we label the x-axis as Gender and y-axis as Count and we plot it by using bar plot.

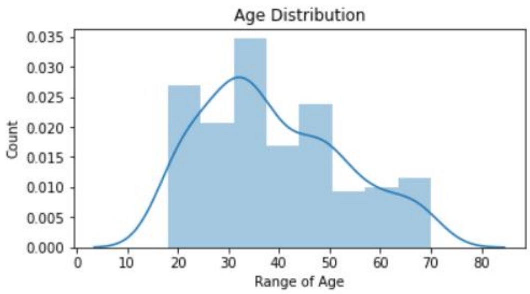


From the plot we will con clued that the there are more female customers than the male customers i.e female customers are more than 100 whereas male customers are nearly 80.

7.4 Age plot

We will use dist plot for the distribution of age of the customers.
So we label X-axis as range of age and y-axis as count.

```
plt.figure(figsize=(16,5))
sns.distplot(customer_data['Age'])
plt.title('Age Distribution')
plt.xlabel('Range of Age')
plt.ylabel('Count')
```



From the plot, it varies the age from nearly 20 to 70. it is evident that the age of the customers between 30 - 40 are more, then after 20-30 etc.

7.5 Annual Income vs Spending Score

As we will use scatter-plot and labeled x-axis as Annual Income(k$) and y-axis as Spending Score(1-100)

```
plt.figure(1 , figsize = (15 , 6))
for gender in ['Male' , 'Female']:
    plt.scatter(x = 'Annual Income (k$)' , y = 'Spending Score (1-
100)' ,
                data = customer_data[customer_data['Gender'] ==
gender] ,
                s = 200 , alpha = 0.5 , label = gender)
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.title('Annual Income (k$) vsSpending Score (1-100)')
plt.legend()
plt.show()
```

From the plot we observed that it varies from low annual income with low expenditure or spending money to high annual income with high expenditure.

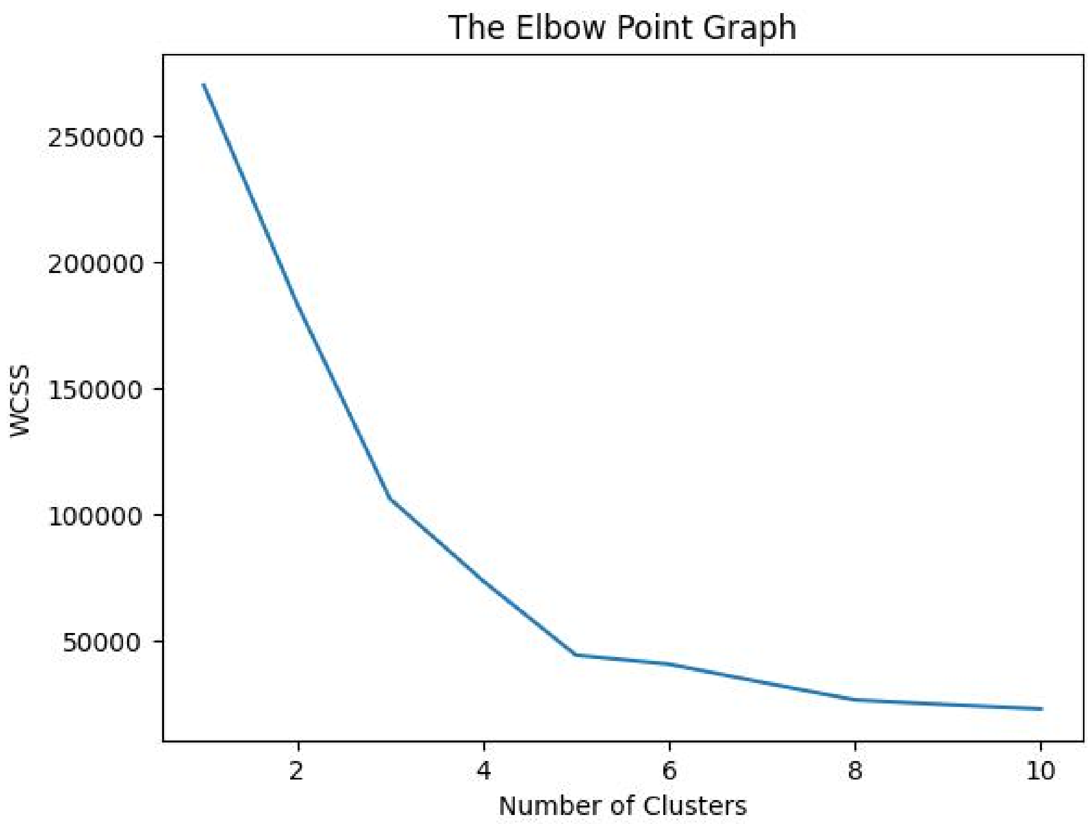
7.6 Elbow Method

The elbow method is based on the observation that increasing the number of clusters can help to reduce
the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture
finer groups of data objects that are more similar to each other.
To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k. This is
done by ranging k from 1 to 10 clusters. Then we calculate the total intra-cluster sum of square. Then,
we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the
approximate number of clusters required in our model. The optimum clusters can be found from the graph
where there is a bend in the graph.
First we will consider the data X which as only two columns they are annual income and spending score.
X=customer_data[['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

```
#finding wcss for value 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()
```



So from the graph we observed that the at 5 there is bend and it can be considered as k which is no of clusters.
Therefore, k=5 i.e no of clusters are equal to 5.

7.7 Fitting the Algorithm

```
kmeans = K Means(n_clusters=5, init='k-means++', random_state=0)
Y = kmeans.fit_predict(X)
customer_data['cluster']=Y
customer_data.head()
```

As here we initialized the k means as kmeans with 5 clusters and we will fit it. There after we will predict the data and store it in y. And then we will add new column named as Cluster and data as y.

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

So from the figure we observed that each customer is labeled with cluster which is based on their characteristics.

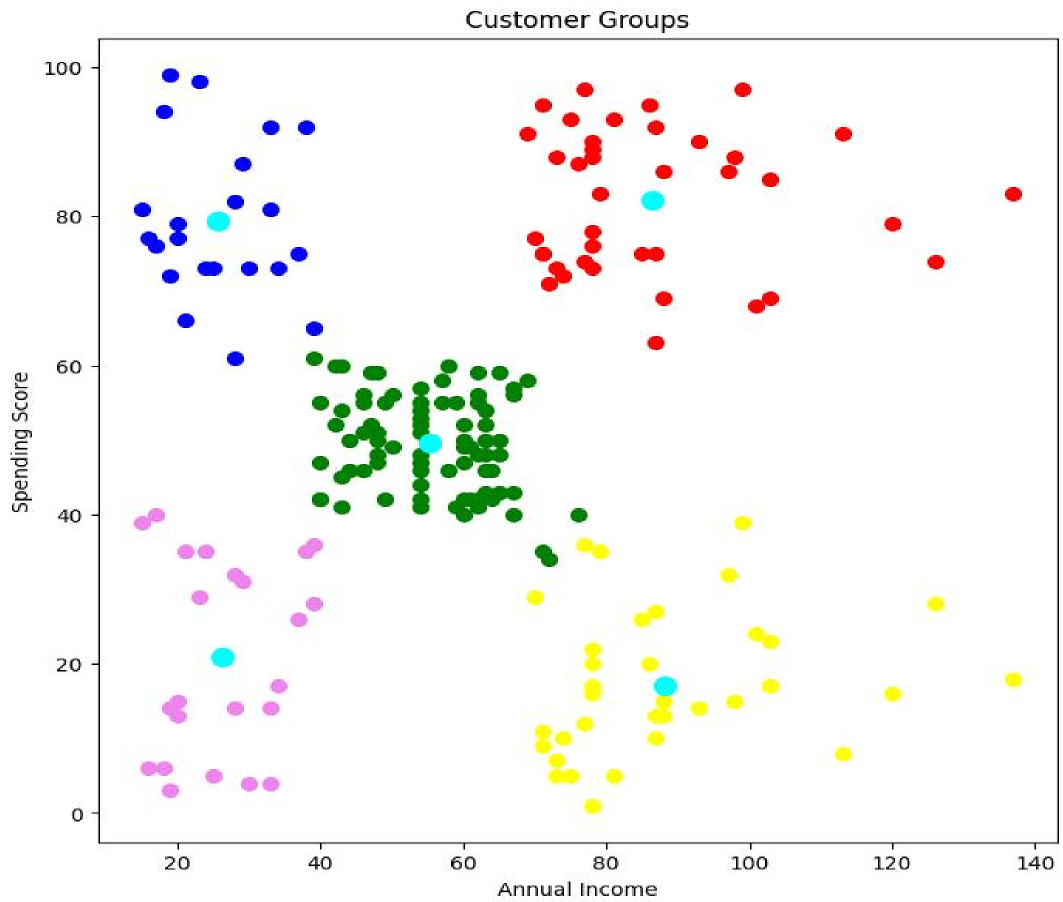
7.8 Visualization the clusters

Visualizing the clusters based on Annual Income and Spending Score of the customers. As here we plot a graph named as Clusters of Customers to visualize the data in terms of groups or cluster.

```
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')

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

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



Cluster 0 (Center):

This cluster is located in the center of the plot, representing customers with **average annual income** and **average spending score**. These customers likely have moderate financial habits, neither spending excessively nor being particularly frugal.

Cluster 1 (Top Right):

This cluster is positioned in the top-right corner, indicating customers with the **highest annual income** and the **highest spending score**. These are high-income individuals who tend to spend significantly, possibly representing luxury shoppers or those with strong purchasing power.

Cluster 2 (Top Left):

Cluster 2 is in the top-left corner, representing customers with the **lowest annual income** but the **highest spending score**. These customers, despite having lower incomes, tend to spend a lot. This group may include people who spend beyond their means or those who prioritize spending on certain items.

Cluster 3 (Bottom Right):

This cluster is located in the bottom-right corner, indicating customers with **high annual income** but a **low spending score**. These are high earners who prefer to save money or spend conservatively. They likely have financial stability but choose to limit their spending.

Cluster 4 (Bottom Left):

This cluster is in the bottom-left corner, representing customers with the **lowest annual income** and the **lowest spending score**. These individuals have limited income and tend to spend very little, likely being more cautious about their purchases or financially constrained.

8. Conclusion

So we concluded that the ,

- The Highest income , high spending can be target these type of customers as they earn more money and spend as much as they want.
- Highest income, low spending can be target these type of customers by asking feedback and advertising the product in a better way.
- Average income, Average spending may or may not be beneficial to the mall owners of this type of customers.
- Low income, High spending can be target these type of customers by providing them with low-cost EMI's etc.
- Low income, Low spending don't target these type of customers because they earn a bit and spend some amount of money.

So high income, high spending are the most beneficial ones to the mall owners which increases the owner's business. (Cluster 1)

9. References

[1] Cooil, B., Aksoy, L. & Keiningham, T. L. (2008), 'Approaches to customer segmentation', Journal of Relationship Marketing 6(3-4), 9–39.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.Silverman, and A. Y. Wu, "An efficient K-means clustering
algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence,
vol. 24, pp. 881-892, 2002.

[4] Bhatnagar, Amit,Ghose, S. (2004), 'A latent class segmentation analysis of e-shoppers', Journal of Business Research 57, 758–767.

[5] Marcus, C. (1998), 'A practical yet meaningful approach to customer segmentation approach to customer segmentation', Journal of Consumer Marketing15, 494–504.