IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Customer Segmentation using K-Means Clustering

Mrs. J. Sirisha[1], V. Lakshmi Prathyusha[2], P. Naga Anupriya[3], M. Suma Sri[4], P. Naga Hema[5]
Assistant Professor, Department of Information Technology[1]
B.Tech Students, Department of Information Technology[2,3,4,5]
Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: *Good revenue has to be generated to run a company. Company needs data to segregate their customers to forecast their product sales or profits they wish to get. Proper decisions need to be taken by evaluating the data from their database. Paper work of grouping customers is a huge task and it is unsure about the accuracy of the results. After the introduction of Machine Learning, techniques in it are being used widely in various areas to uplift the accuracy of the outcomes and to find a better way of evaluation in every area. Here, customer segmentation using K-Means clustering helps in grouping the customers with the similar traits that helps to business in a better way. The elbow method applied here helps in finding the optimal number of clusters to be visualized.*

Keywords: Numpy, Pandas, Sklearn, Matplotlib, Seaborn, Clustering, Elbow Method, K-Means Algorithm.

I. INTRODUCTION

Every business requires the data of the customers to perform some operations where the data can be grouped or ungrouped. Data mining is the technique that extracts the data from the database in human readable format but the dataset couldn't provide any benefits.

Customer segmentation is used to group the customers based on their demographics, age, expenditure, and annual income. The groups that are made are known as clusters. By clustering of customers using annual income and amount spent, one can know which group is contributing more to the successful run of the business and which group of customer needs marketing of the products for making them contribute to the business profits and its stability. Likewise, the comparison can also be done between age and count of the customers, gender and count of customers.

The main goal is to group the customers into optimal number of clusters so that each cluster has the similar trails. Elbow method is used to find the optimal number of clusters to group the customers and K-Means clustering algorithm is used to cluster the customers and visualization [2-6].

II. PROPOSED SYSTEM

Paper work, which had used much before the introduction of Machine Learning, hadn't given the best results but takes much time to evaluate the customer's data and this can be replaced by the machine learning technique, which isa powerful innovation to predict the outcome through its defined algorithms.

Grouping of customers can be beneficiary to the business for running successfully. This can be done by an unsupervised Machine Learning algorithm known as "K-Means Clustering" algorithm. This comes under unsupervised which contains the unlabelled data.

So, we used the K-Means clustering algorithm which is considered as the unsupervised learning technique that uses only unlabelled data as input. Number of clusters can be known through the elbow method and the process of clustering groups the customers there by we will visualize the grouped data appropriately by using the same algorithm [2-6].

III. TOOLS REQUIRED

3.1 Google Colaboratory

Colaboratory, or "Colab" for short allows the users to write and execute the python code through the browser that runs entirely in the cloud. Unlike other python notebooks, colab does not require any setups to run the python code. It is the free cloud resource that can be viewed by anybody which is hosted by Google.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

IV. PYTHON LIBRARIES REQUIRED

The following are the libraries of python that are used in this project:

1. **Numpy:** "numpy" is to operate the arrays and matrices.
2. **Pandas:**"pandas" are used to deal with the ML tasks and data analytics and also to load the data from the csv file to the data frame of panda.
3. **Matplotlib:** "matplotlib" is used to have static, animated and interactive visualizations for our data.
4. **Seaborn:** "seaborn" is a data visualization library of python which is based on the matplotlib and that gives the interface at high-level.

V. RELATED WORK

5.1 Dataset

The data is taken in the form of csv file which contains the data in the form of rows and columns. In this project we used the dataset https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis/data that has the columns named customer ID, age, gender, annual income and spending score.

Here, we majorly focused on annual income and spending in major. The spending score is in the range of 1-100 where the score tells the level of amount customer spent [1].

5.2 Stepwise Procedure for Clustering

Step 1: Upload the required dataset that contains the columns named customer ID, age, gender, annual income and spending score [1].

Step 2: Data Collection and Analysis:

- Importing the required libraries.
- Loading the dataset into pandas dataframe.
- Finding the shape of the dataset taken i.e., no. of rows and columns.
- Getting the dataset information for every column mentioned to check whether there are any null values present in the dataset.

Step 3: Data Cleaning is done to fill up the cells of null values with the appropriate values in the dataset.

Step 4: Deleting all the columns that are not considered to the project.

Step 5: Elbow Method is choosing the number of clusters i.e., optimal no. of clusters to group the customers using the formula of *Within Clusters Sum of Squares (WCSS)*.

Step 6: Training the K-Means clustering model which means assigning the values to each customer in the dataset according to the similar traits and which group they belong.

Step 7: Visualizing the clusters of customers through graph with considered columns taken on x-axis and y-axis.

5.3 Visualization

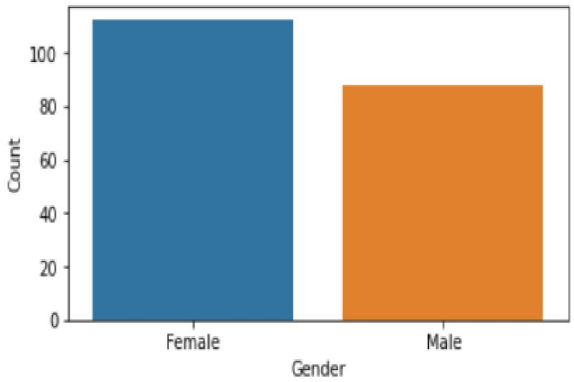

Figure 5.3.1: Gender plot analysis.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

The above bar graph shows that there are more females customers than the male customers i.e., female customers are more than 100 whereas male customers are nearly 80.

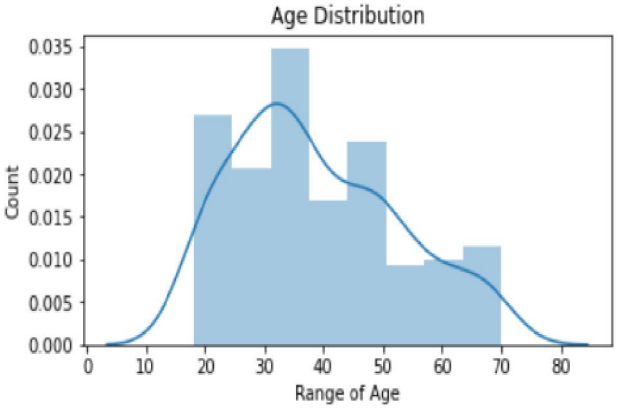

Figure 5.3.2: Age plot.

The above age plot gives the information that the customers age is nearly from 20 to 70 in the dataset. The customers between 30-40 are more in number and later 20 - 30 likewise.

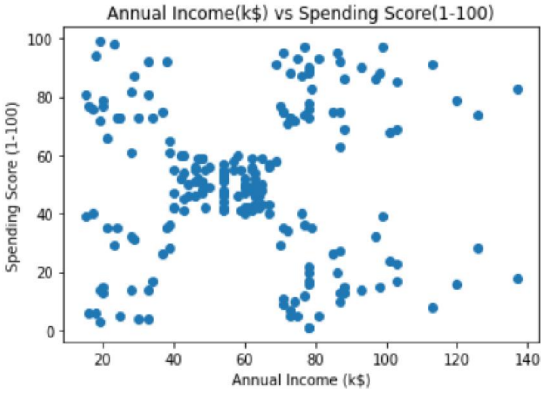

Figure 5.3.3: Annual Income vs Spending Score.

The above scatterplot takes the x-axis as Annual Income and y-axis as Spending Score and this shows that the plot varies from low annual income with low expenditure to high annual income with high expenditure. The above has not shown any division of clusters as everything in the same colour blue. We need to group the customers into clusters so that we need number of clusters which can get from plotting the elbow graph.

VI. EXPERIMENTAL RESULTS

```
[ ] # first 5 rows in the dataframe
    customer_data.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

Figure 6.1: First five rows of dataset.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

The above figure gives an overview of the column names of the dataset that is taken for the evaluation of customers to be grouped into different clusters. The column Customer ID contains unique IDs, annual income in k$ and spending score in the range of 1-100. The score tells the level of spending of each customer [1].

```
[ ]  # finding the number of rows and columns
     customer_data.shape

     (200, 5)
```

Figure 6.2: Shape of the dataset.

It gives the number of rows and columns present in the dataset. Here there are 200 rows and 5 columns which can be said as a small dataset taken from some mart to visualize the customers into clusters [1].

```
# getting some informations about 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
```

Figure 6.3: Overview of the dataset.

The above gives the information of the columns of the dataset i.e., whether the column contains any null values and its count, the data type of each column and memory used [1].

```
# finding wcss value for different number of clusters

wcss = []

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

  wcss.append(kmeans.inertia_)
```

```
# plot an elbow graph

sns.set() #this gives the basic theme of the graph.
plt.plot(range(1,11), wcss)
plt.title('The Elbow Point Graph')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```

Figure 6.4: (a) Finding WCSS value and no. of clusters using elbow graph.

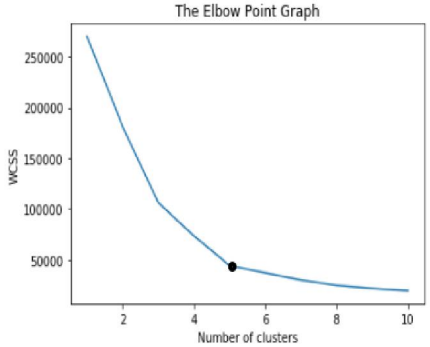

Figure 6.5: (b) Elbow Graph.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

The above graph is the Elbow Point graph which is based on the observation that the increasing 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 have finer groups of data objects that are more similar to each other.The plot determines the approximate number of clusters required in our model according to the range of the k we take say 1-10.

The x-axis is taken as number of clusters and the y-axis is taken as WCSS which is the sum of squared distance between each point and the centroid in a cluster.

$$WCSS = \sum_{P_i \text{ in Cluster 1}} distance(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} distance(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} distance(P_i, C_3)^2$$

Where C1,C2, C3 are the centroids of the respective clusters and Pi are the data points in the clusters.

The optimal number of clusters can be found through the bend of the elbow graph. There are two bends in the above graph but here we take k=5 since there is any other bend after k=5 and hence it is the optimal value of k [4].

```
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)
```

```
[4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
 3 4 3 4 3 4 1 4 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 2 0 2 1 2 0 2 0 2 1 2 0 2 0 2 0 2 0 2 1 2 0 2 0 2
 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0
 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2]
```

Figure 6.6: Training the K-Means clustering model.

The dataset contains the various customer IDs and they need to be clustered so that according to above each customer is assigned a number through which one can cluster the customer to the respective cluster. Here, the elbow method gives the optimal number to divide the clusters is k=5. Hence, the numbers assigned are from 0 to 4.

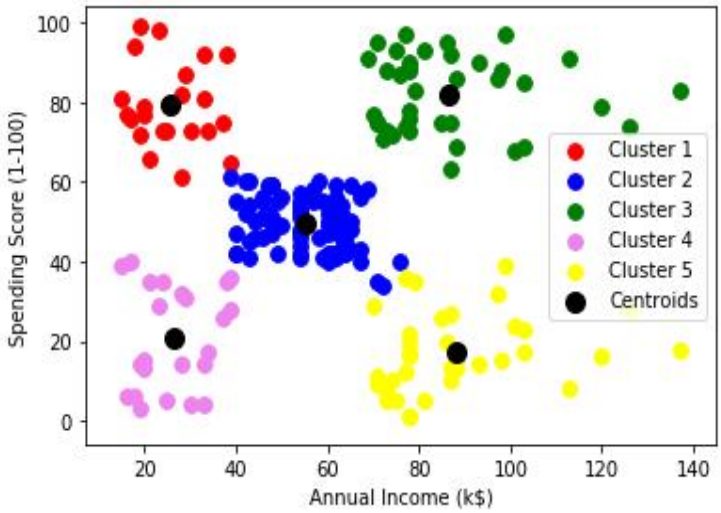

Figure 6.7: Visualization of the clusters.

The clusters have visualized based on the annual income and spending score of them mentioned in the dataset taken. We have given a separate colour to each of the five clusters as shown above. The black spot in the each cluster is the centroid of the each cluster. The colour points are the data points that are plotted by taking x-axis and y-axis as shown.

- The cluster 1 is about the customers with low annual income and high spending score.
- The cluster 2 is about the customers with average annual income and average spending score.
- The cluster 3 is about the customers with high annual income and high spending score.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

- The cluster 4 is about the customers with low annual income and low spending score.
- The cluster 5 is about the customers with high annual income but with less spending score [5].

VII. SCOPE OF FUTURE USE

This can be further extended by taking an extra column products and by considering three columns (spending score, annual income and products) one can get the data of the products that are being purchased by the customers so that the evaluation of the most selling products can be more easy and accurate. So, the owners can also get an idea on which products the marketing should be done to increase the sales and decreasing the churning of the customers.

VIII. CONCLUSION

Therefore, we have visualized that the customers with high income and high spending score are more beneficial since they increases the business and the customers with high income and low spending score should be taken into consideration for the feedback for increase or maintaining stability in the business. So, advertising can be done to meet the business requirements to these type of customers.

REFERENCES

[1]. Dataset - https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis/data
[2]. https://www.geeksforgeeks.org/supervised-unsupervised-learning/
[3]. https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a
[4]. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
[5]. https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3
[6]. https://towardsdatascience.com/clustering-algorithm-for-customer-segmentation-e2d79e28cbc3