SegmentWise: An Unsupervised Approach to Customer Behavior Clustering via K-Means

Vijayakumar Polepally
Department of Computer Science & Engineering
Kakatiya Institute of Technology & Science
Warangal, Telangana
vijay.cse@kitsw.ac.in

Jagannadha Rao D.B.
Department of Computer Science & Engineering
Malla Reddy University
Hyderabad, Telangana
jagandb@gmail.com

***Abstract*— When someone wants to setup a company there are a lot of things, they should consider to get good profits/revenue because to setup a big company they need a lot of resources and a lot of budget to put together and one small mistake can give huge losses. So, we should use better strategies so that the risk factor becomes low. Decision making is a very important thing before setting up a company and people who are working in the company should use better techniques to reduce the risk to the company. their decision will make the final result a better result or give worst results. One of the important techniques is to make use of technology and one of the best techniques we could use to get better yield for a company is machine learning where we use several past datasets and train models and get important data from it so that it will be very useful to the company's growth.**

Keywords—*Customer segmentation, Clustering, K-Means Algorithm*

I. INTRODUCTION

1.1 Overview

Competition now a days is very huge and literally every company is using their own strategy or techniques to get good revenue for their companies. For every company competition, the very important role is played by the customer as they are the ones that are buying the product in the first place and so, if we collect the data of all the customers like age, income and spending score etc, we could get some useful information from the data which is useful for business as well as for getting good revenue for the company. Data mining is very useful in this case because what we do is we process the old data that is fetched from several customers and that data is mined and we get some useful information from the data like which age people are buying the product or which gender is most buying the product and who are spending a lot of money on certain product etc.

The main theme of this project is to try to determine the audience for the product and by using the clustering algorithms on the customer base, we need to sell the product to the targeted audience such that we get high revenue and high profits. This is very helpful for all businesses as well as all the people out there who wants to set up a company with less risks.

1.2 Motivation

Firstly, we need to collect the data from the customers this data is considered as the OLD data, this is collected when people are purchasing the product so simultaneously, we collect the data from them and after we collect the data, we need to get the whole data together and visualize it.

The group of people who have more income are most eligible to buy a product so, the income of the customer is also very important for the data. These are the main important things we need to consider for now for customer segmentation into clusters and i think that this data will be very useful for the companies out there.

1.3 Scope

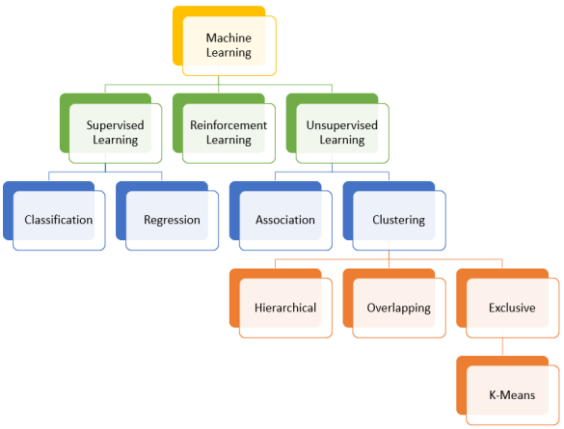
The main goal for any company is to increase their profits and if there is good competition between several companies who produce the same product then, they need to think of certain strategy which makes the product unique and so that to achieve better profits, we need a good technique and customer data segmentation is a good marketing strategy as it helps in giving the information we want. we fetch the data from the old data we discussed above and then we try to segment the data using K-Means algorithm and in that K-Means algorithm, we use elbow method which is very helpful to give the number of clusters that we need to do from a set of data. So, after we segment the data, we can clearly and safely use algorithms at visualize the data and increase company's profit as well. So, in this project we make use of this k means algorithm to divide the data into clusters and after that we can use the clustered data in any way we want.

1.4 Machine Learning

Machine learning is an important topic everyone needs to learn and implement in everyone's business to improve their standards of their company. We can use this machine learning algorithms everywhere like recommendation systems, data segmentations etc. we are using this technique in our project as well as we are segmenting the customer data that we got from the old data.

Machine learning is of two types, one is supervised learning and others is unsupervised learning. the main difference between these both learnings is that supervised learning has a label for the segmented data while in unsupervised learning, the data which is segmented is not labeled, it is just segmented into parts that's it.

Mostly data analysts use supervised learning as they deal with problems in day to day business life. the unsupervised learning is unlike supervised learning is based on a model (mathematical model) and it is useful in marketing business, supermarkets etc. here in our project we are going to use unsupervised learning where we segment the whole data into clusters which have no label at all. So, to generate good profits, and to increase customer revenue, we going to use this kind of learning.

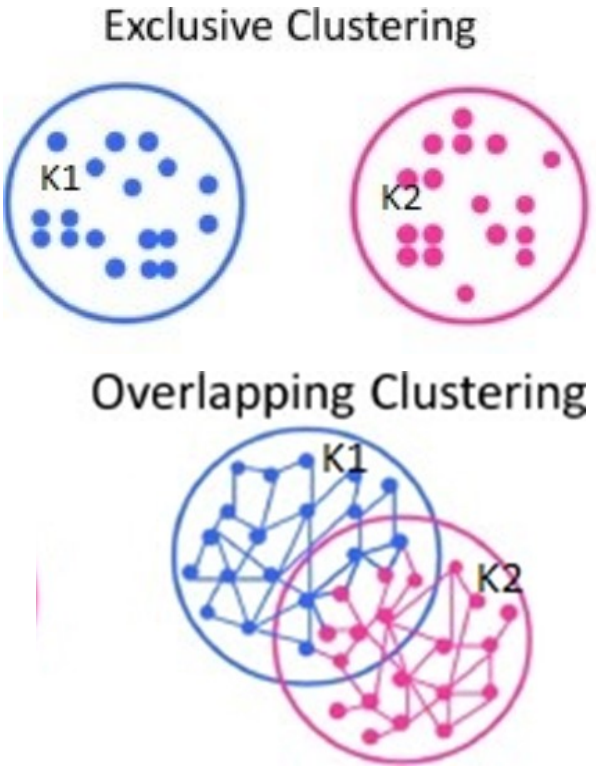


The above figure is the tree diagram of the machine learning and we can clearly see the K Means at the bottom of the tree which means the K means is a derivative of the Machine learning program. There are three types of learnings, supervised, unsupervised, and reinforcement learning. Clearly, the K means falls under the unsupervised learning as there are no labels while clustering the data in k means.

1.5 Clustering

Another important topic we need to understand is Clustering topic. The main reason that we convert the useless data into clusters is because we make some use of it. Many companies use this Cluster based classification to divide the useless data into clusters. First, we need to plot a graph from the useless data and then we need to apply the elbow method to find the number of optimal clusters from the given data and after we get the optimal number of clusters, we need to divide the whole data into those number of clusters, and each and every cluster has its own centroid.

There are two types of strategies in clustering, they are factor analysis and the other is cluster analysis, both have their pros and cons but we mainly use cluster analysis as factor analysis has difficult approach. This clustering method is a basic strategy in machine learning as in machine learning we make use of whole lot of data and without making them into clusters, we cannot advance anywhere.



II. PROBLEM STATEMENT

Now, let us discuss about the problem statement of our project. The main problem is to get good revenue to the company by making scattered data into useful data. By clustering or segmenting the customer data, we can make useless data into useful data. Companies now a days are struggling a lot because they don't have good strategies to get good profits out of the products.

There is also a lot of competition between the companies as different companies using different techniques and the best technique wins. We are proposing to use the K-Means Clustering where we segment the customer data and along gender, age, spending score and annual income. After clustering we visualize the clusters and get the important information, we want for the sake of companies to get good revenue or profits.

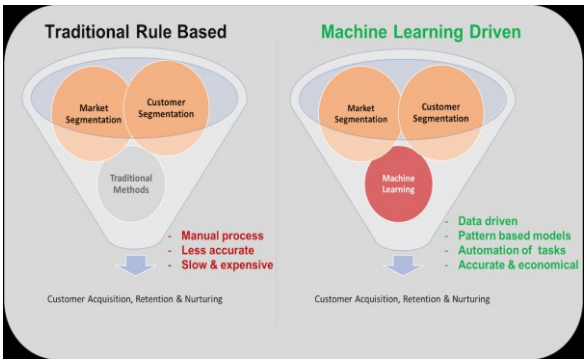
By this technique, we can know the group of people who are buying the product and who are not buying it and with this information we can avoid the problem of getting low revenues for the companies by analyzing the dataset.

III. PROPOSED METHOD

In order to overcome the above-mentioned problem, we use customer data segmentation using K-Means method to convert the useless data into some useful ones and by doing that we convert the data into some sort of clusters. After we make them into clusters, we are going to visualize the clusters so that it will lead to good information that we can make use of it to get better revenues for the company.

The k-means clustering is a famous machine learning technique coming from ages used by many

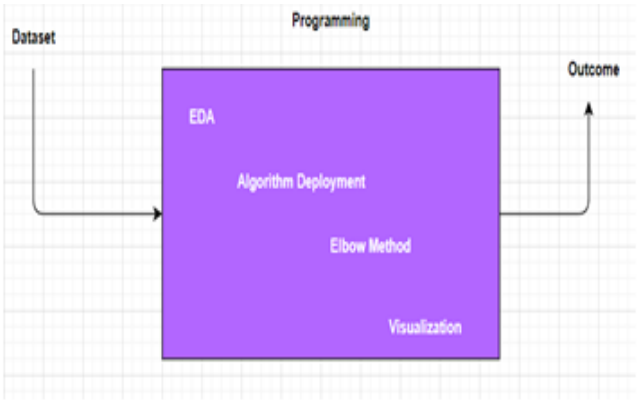
companies to get more profits. The data set we collect from the people contain their age, spending score, annual income etc., which are powerful data very useful for us. The main thing is it is useful to increase company's profits by this algorithm and that's why we are using it.



IV. SYSTEM ARCHITECTURE

The first thing we need to do is to clean the data from the dataset. What we do is we remove the duplicate values, and null values and such from our dataset so that we will have less stuff to work on and also useful stuff too.

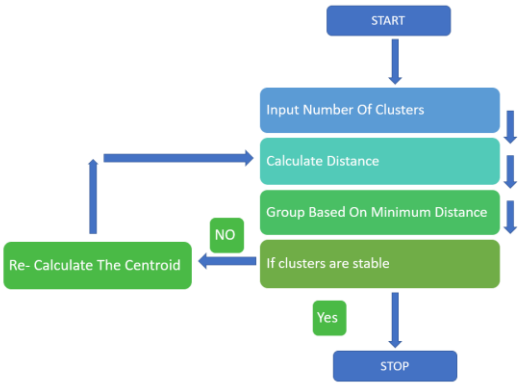
To find the no of clusters, we make use of elbow method where we use a formula, distance that is calculated from the data centers. we will repeat this method until we get same number of clusters and that is what we want and after that we proceed to make clusters.



V. ALGORITHM

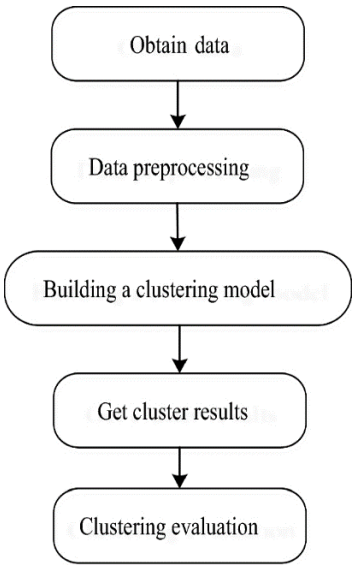
K-Means Clustering

- K-Means algorithm is used to divide the data into k number of clusters using elbow method. This is how we classify the whole data.
- K is the total no. of clusters we get from elbow method.
- We divide the data such that one point belongs to only one cluster.
- The clusters cannot be one on top of another cluster.
- This is the process flow of the algorithm:



STEPS OF ALGORITHM

- We select K random points from the data we have and find the centroid and save that point in a certain cluster.
- Repeat this process until we get same clusters.
- We have to find the average or mean values from the random data and after that we need to send the point to appropriate cluster where it is nearer to the average value.
- After we divide them into clusters, we find the new average values and continue the same process again and again until we get same clusters.
- And when we find no change in previous and present cluster values, those are the final clusters.



VI. METHODOLOGY

The first step is to import all the important needed libraries like pandas, NumPy and seaborn which is used to make plots etc. Then we need to clean the data like we need to remove all the null values and also the duplicate values etc. After dropping and cleaning the data, the remaining data is the data which is useful for applying k means algorithm. This process is known as data preprocessing.

After the preprocessing we use k means clustering to cluster the whole data into clusters, we use elbow method for finding

the number of clusters(k), in each cluster, the data is somehow similar to each other so that we can easily visualize the data.

Finally, with the help of seaborn package, we will plot the graphs for every data available for us and make it convenient for us to use and visualize.

VII. IMPLEMENTATION AND ANALYSIS

7.1 Overview of the Dataset

Below, we can see the data from the csv file which is also known as dataset. We can see the first five rows and last five rows.

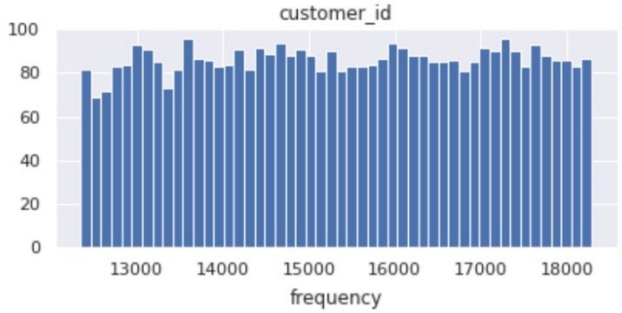
	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

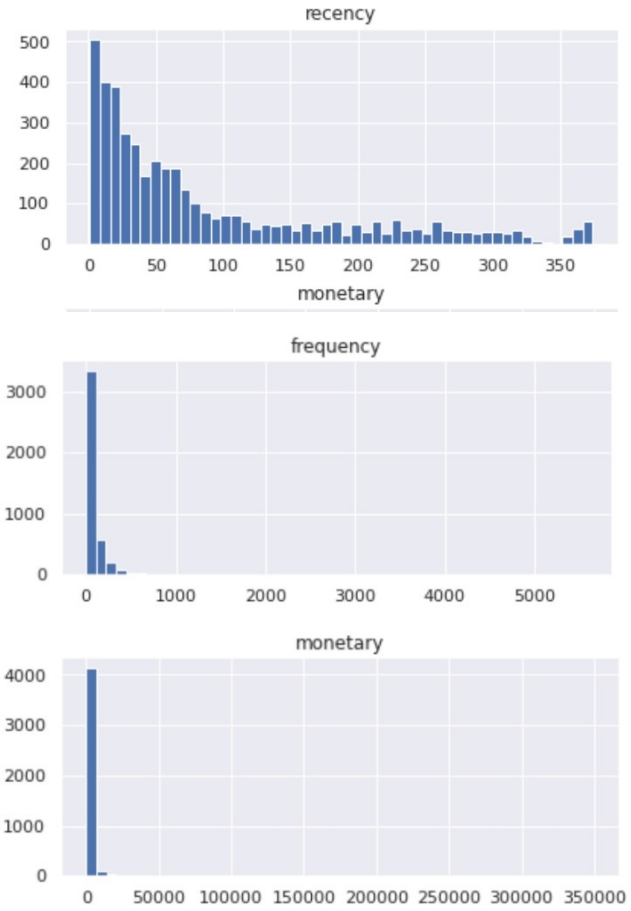
200 rows × 5 columns

The above data is the old and primary information we need to make any algorithm to work.

7.2 Exploratory Data Analysis

We then clean the data which is also known as data preprocessing where we clear the null values, duplicate values etc., to make the data more elegant and useful. After preprocessing, the number of columns might decrease as we are removing the redundant data from the dataset.





7.3 Information of the dataset

```
<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
```

The above figure shows the information of the dataset where we can clearly see the customerID, the gender, age, annual income and spending score of the Customers which are essentially the column names.

7.4 Description of the data

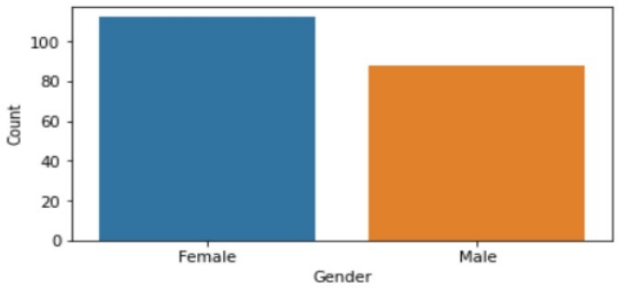
	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823522
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000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

here the code describe the count to find the no.of rows, columns and the average values, min values and max values etc.

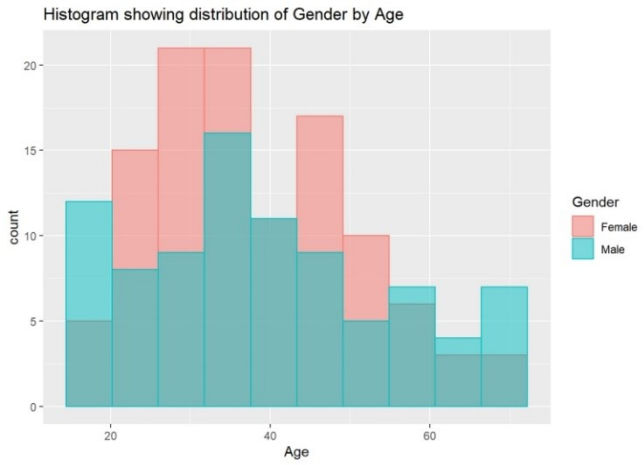
7.5 Gender plot Analysis

```
#Gender Distribution
genders=df.Gender.value_counts()
plt.figure(figsize=(6,3))
sns.barplot(x=genders.index,y=genders.values)
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show
```

In the below figure , there is a plot with count of the customers who are buying the product versus the gender.



If you clearly look at the above graph you can see that the ratio of female who are buying a certain product is a lot more than the male.



The above figure is an histogram. We can clearly see that we have mixed the

Age plot as well as the gender plot against each other.

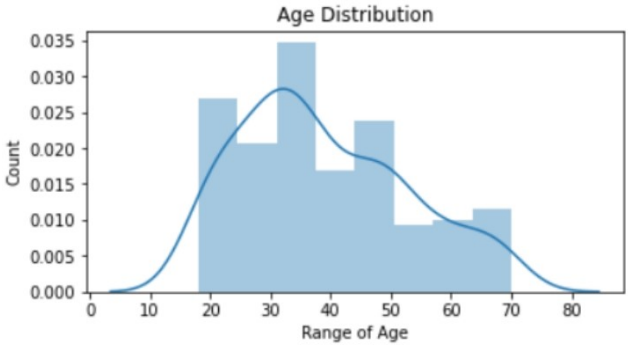
This histogram is like two in one plot which gives twice as information in just one plot. So this is we can do best while doing analysis.

7.6 Age plot

Below is the code for plot age versus count

```
plt.figure(figsize=(6,3))
sns.distplot(df['Age'])
plt.title('Age Distribution')
plt.xlabel('Range of Age')
plt.ylabel('Count')
plt.show()
```

We took age on X axis and count on Y axis



From the above plot, clearly, we can see that ages in the range 30 to 40 have more count and as age increases, the count also decreasing exponentially.

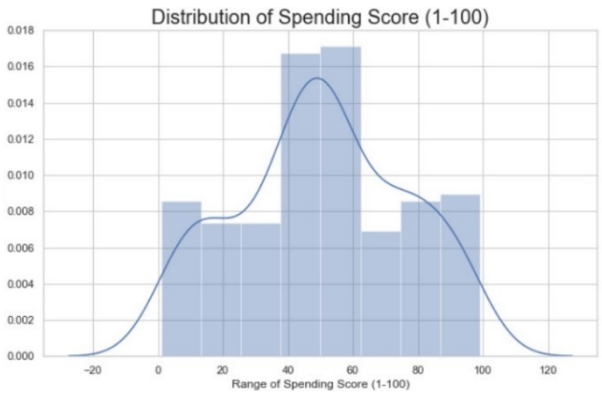
7.7 Annual Income vs. Spending Score

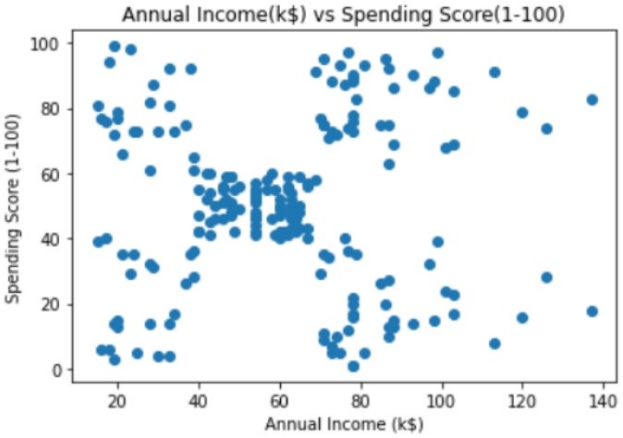
Next, we drew the plot between annual income of a customer versus the spending score of that particular customer.

```
plt.scatter(df['Annual Income (k$)'],df['Spending Score (1-100)'])
plt.title('Annual Income(k$) vs Spending Score(1-100)')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.show()
```

In the plot that we drew between annual income and spending score,

The annual income is on X axis ranges from zero to 140 and the spending score is on Y axis while this ranges from zero to 100.





From the above plot, we can clearly see that the density is more when annual income of the customer and spending score of that customer are average and the other points are too scattered than normal.

7.8 Elbow Method

The elbow method is a special technique to find the number of clusters that we are supposed to divide the data into. There can be any number of clusters we can make but we need to make only the optimal number of clusters. so to find out the optimal number of clusters, we use this method where we use elbow method formula and k ranging from 1 to 10 and plot a graph against WCSS score.

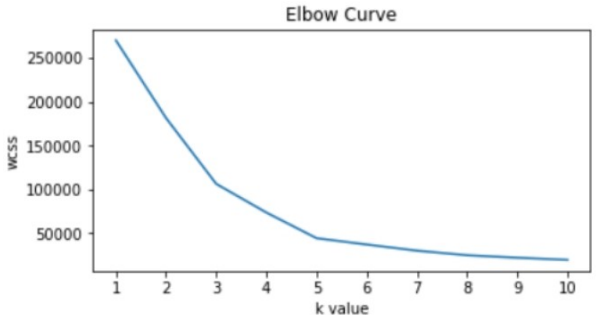
when we find a sharp steep in the plot we made, that particular k value is the optimal number of clusters to build our model. In our project, we got sharp steep at k= 5 so, we took total number of optimal clusters as 5.The below figure shows the first five columns of annual income and spending score.

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

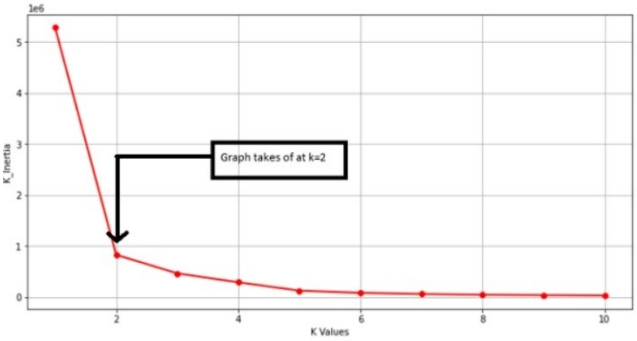
X.head() // for getting the first five rows

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

Code(raw):
Wcss= []
for i in range (1, 11):
Km=KMeans (n_clusters=i)
km.fit(X)
wcss.append (km.inertia_)
plt.figure (fig size= (6,3))
plt.plot (range (1,11),wcss)
plt.title ('Elbow Curve')
plt.xlabel ('k value')
plt.xticks (np.arange (1, 11, 1))
plt.ylabel ('wcss')
plt.show ()



As we discussed above that we plot an elbow curve against wcss and k value, we can see clearly that a sharp steep at k value of 5, so the number of optimal number of clusters is 5 and as we found the k value, we can move on and apply the k means algorithm and divide the data into clusters.



This is the Euclidean distance formula between two points.

$$k = \sqrt{(x_n - x_c)^2 + (y_n - y_c)^2}$$

This is the midpoint formula and to find centroid

$$k_{new} = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right)$$

$$k_{new} = \left(\frac{147 + 138}{2}, \frac{330 + 309}{2}\right) = 142.5\,, 319.5$$

The Data:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	label
0	1	Male	19	15	39	5
1	2	Male	21	15	81	3
2	3	Female	20	16	6	4
3	4	Female	23	16	77	3
4	5	Female	31	17	40	5

From the above data table, the last column which is nothing but label indicates the cluster number the row belongs to.

So, each row has one label which means that row belongs to that particular cluster and after we get labels for all the rows, by plotting them, we can get clear clusters.

7.9 Fitting the Algorithm

```
km=KMeans(n_clusters=5)
km.fit(X)
y=km.predict(X)
df['Cluster']=y
df.head()
```

We already found that the optimal number of clusters for the customer data segmentation project is 5 so we now move on and apply the algorithm. We need to divide the data into 5 optimal clusters where there are similarities between the data in each cluster. The below picture shows the head (first five rows) of the whole dataset we are using right now and the cluster number as the final column.

	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

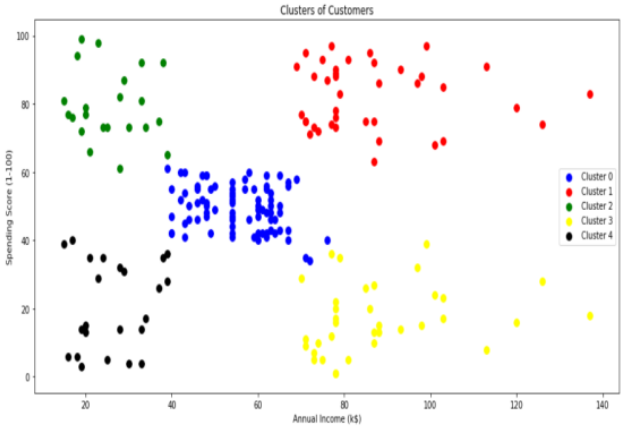
In the above figure, we can clearly see the cluster number of each and every row or data present in the dataset. This way, we divide the whole data and store them in the 5 different clusters.

7.10 Visualization of the clusters

Cluster visualization is an important thing and here we visualize the clusters we obtained. Here what we do is that we plot a graph against age, spending score and also the annual income and point the dots in the graph accordingly. We can clearly see the clusters are formed.

```
plt.figure (fig size= (15,7))
plt.scatter (df ["Annual Income (k$)"][df.Cluster == 0], df["Spending Score (1-100)"][df.Cluster == 0], c='blue', s=60,label='Cluster 0')
plt.scatter (df ["Annual Income (k$)"][df.Cluster == 1], df["Spending Score (1-100)"][df.Cluster == 1], c='red', s=60,label="Cluster 1")
plt.scatter (df ["Annual Income (k$)"][df.Cluster == 2], df["Spending Score (1-100)"][df.Cluster == 2], c='green', s=60,label='Cluster 2')
plt.scatter (df ["Annual Income (k$)"][df.Cluster == 3], df["Spending Score (1-100)"][df.Cluster == 3], c='yellow', s=60,label='Cluster 3')
plt.scatter (df ["Annual Income (k$)"][df.Cluster == 4], df["Spending Score (1-100)"][df.Cluster == 4], c='black', s=60,label='Cluster 4')
plt.title ('Clusters of Customers')
plt.legend ()
plt.xlabel ('Annual Income (k$)')
plt.ylabel ('Spending Score (1-100)')
plt.show ()
```

VIII. RESULTS



So, as part of the visualization of the clusters, we can clearly see from the above plot that:

- The first Cluster which is in blue colour has average spending score and also average annual income which is much denser than any other clusters, so this is more important data.
- The second cluster which is in green colour has high spending score but low annual income which is a little bit scattered but also a useful information.
- The third cluster which is in black colour has low spending score and also low annual income which is too scattered might be not useful cluster.
- The fourth cluster which is in red colour has high spending score as well as high annual income which is also scattered but people with higher income and higher spending score are so common as they are directly proportional.
- The fifth cluster which is in yellow colour has low spending score but high annual income which is a lot scattered because , people who have high annual income tend to spend a lot more.

IX. CONCLUSION

- We can clearly target the group of people who has high spending score as well as high annual income as they are the people who mostly be useful for the revenue of our company.
- For people who have high annual income but low spending score, we can actually use advertisements and recommendations so that they tend to buy the products and fall into the high income, high spending category.
- Average income, average spending score is not much beneficial but from the cluster analysis, we can see that most people fall in this range, so we can keep an eye on them.
- Low income and low spending score are not much beneficial for us as some people are poor and they cannot buy the products but we can make discounts for them so that they can buy some of our products.

- We can also target the people who spends more but less income as clearly, we can see that they tend to spend more, so we can take advantage of that.

X. REFERENCES

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

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

[3] Neil, D. Akshay, Kelvin, P.C (2006), 'customer segmentation applications', Journal of customer segmentation.

[4] Ms, Celine. (1995), 'uses and needs of customer segmentation and approach to customer segmentation',

[5] Sebastian, B., Arin, sean, T. L. (2007), 'customer segmentation, how to work with it', Journal of k means and data segmentation.

[6] Brian, K. (1995), 'segmentation of data and its uses', Marketing genius.