**EXPOSYS DATA LABS**

Data Science - Customer Segmentation project

**A Report by -**

Name: Masabathula V S Raghavendra Rao

Internship stream chosen: Data Science/AI/ML

Topic: Customer Segmentation

Branch: CSE

College: National Institute of Technology- Andhra Pradesh.

email: iamraghava.22@gmail.com

Phone: 9390255829

**1.Abstract:**

The main intention of this project is to divide the customers into various segments based on their common characteristics, so that the business can organize and market each group effectively and with ease.

**2.Tabel of contents:**

1. Introduction
   1. What is customer segmentation.
   2. Why customer segmentation

3.3 Types of customer segmentation

1. Existing Method
2. Proposed method with Architecture

5.1 What is the basic procedure involved to form clusters?

1. Methodology

6.1 KMeans Clustering Methodology

6.1.1 Advantages of KMeans clustering

6.1.2 Disadvantages of KMeans clustering

6.2 Hierarchical Clustering Methodology

6.2.1 Advantages of Hierarchical clustering

6.2.2 Disadvantages of Hierarchical clustering

6.3 DBSCAN Clustering Methodology

6.3.1 Advantages of DBSCAN clustering

6.3.2 Disadvantages of DBSCAN clustering

1. Implementation
   1. KMeans Clustering Algorithm, code
   2. Hierarchical Clustering Algorithm, code
   3. DBSCAN Clustering Algorithm, code
   4. Silhouette Cluster Validation technique
   5. Finding the BEST model
2. Conclusion
   1. Visualizing the Clustered data for different combinations of features and analyzing them.
   2. Finally Ranking the customers based on their spending scores
   3. Treating separate clusters effectively

**3.Introduction:**

**3.1 What is customer segmentation?**

Customer segmentation is the process of separating your customers into groups based on the certain traits (e.g., personality, interests, habits) and factors (e.g., demographics, industry, income) they share, so that each group can be treated effectively.

Segmentation offers a simple way of organizing and managing your company’s relationships with your customers.

**3.2 Why customer segmentation?**

There are a number of other reasons why customer segmentation is so important. Here are some of the things this process can help your business accomplish:

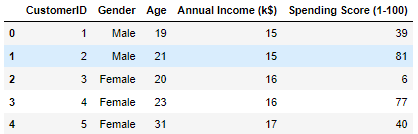
* Learning about your customers on a deeper level so you can tailor your content to their unique needs and challenges.
* Creating targeted campaigns and ads to resonate with and convert segments of customers.
* Improving your customer service and customer support efforts by understanding and preparing for challenges different groups are likely to experience.
* Increasing customer loyalty with customized content and interactions.
* Understanding who your most valuable customers are and why.
* Communicating with segments of customers via preferred channel or platform.
* Meeting specific groups of customers where they are.
* Identifying new opportunities for products, support, and service efficiently.
* We can efficiently target a group of customers at a time, instead of looking at each customer individually.

**3.3 Types of customer segmentations:**

* Demographic Customer Segmentation (By Age, Gender, Income, occupation etc.)
* Geographic Customer Segmentation (By Location, Language, Transportation, etc...)
* Behavioral Customer Segmentation (By Lifecycle stage, Website Activity, etc...)
* Psychographic Segmentation.
* Social Media Segmentation, and more ....

In our data set, we have the following columns:

CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100).



So, we will be segmenting customers into groups based on These four important features: Gender, Age, Annual Income (k$), Spending Score (1-100).

**4.Existing Method:**

We have seen what are the important features in our dataset.

So, our job now is to cluster the 200 customers into optimal number of clusters so that we could differentiate them easily.



The very basic segmentation is using demographics like age, Gender, Income etc. Taking age as ultimate basis, we can segment customers into various segments. Like for example we can group the customers whose age is less than 18 into one cluster, 18-30 into one group, 30-45 into another, above 45 into another.

And now we can deal with each group differently based on their interests and habits.

Also, the clustering is can be done based on the spending scores given to us, and rank them accordingly.

For instance, for customers having spending scores > 75 % will be treated separately compared to the customers having spending scores < 50%.

Basically, for some datasets scores will not be given. So, we have to calculate these **RFM scores** ourselves from the data given, and then rank them based on these scores.

But, How RFM scores are calculated? Let’s understand briefly. RFM stands for Recency, Frequency, Modelling.

We first calculate these 3 for each customer:

- **Recency:** no of days after the customer’s recent purchase

- **Frequency:** No of Times the customer has purchased.

- **Monetary**: Total money spent by the customer.

Lower the Recency better the RFM score. Higher the Frequency and monetary better the RFM score. Now we divide customers into n segments corresponding to each of Recency, frequency, monetary, i.e., Basing recency we assign each customer a group value under column 'R’. Similarly basing Frequency, Monetary we do the same creating the columns 'F', 'M'.

Finally, RFM score = R+F+M values for each customer. Now based on the RFM score, we divide customers into N groups. Therefore, we have successfully divided into N segments of customers.

Fortunately, these scores are given to us.

**5.Proposed method with Architecture:**

Previously we divided customers into segments based on just one feature, such as age, spending score. But now we will extend this a bit. What if we group them into clusters considering two or three or four features at a time?

But our dataset has Gender, Income, Spending Score too. So let us take advantage of these and get much more organized clusters. Now, we can not only divide customers into segments based on 1 feature, but we can divide them more efficiently using 2 or more features too.

We can plot them as a 2-D graph or 3-D graph and observe the clusters that can be formed.

We can plot between many features like:

Age VS Income, Age VS spending Score, Age VS Gender, Income VS Spending Score, Income VS Gender, Gender VS Spending Score!

So, we label these clusters and treat each cluster separately.

So, what is the procedure we have to follow? Machine **learning** provides us various kinds of approaches to do this. However, the intention and basic steps are same, the technical concept is different, which actually makes them advantageous and disadvantageous sometimes. So, it’s our job determine the best model out of all.

**5.1 What is the basic procedure involved to form clusters?**

The generalized method is as follows:

1. Perform the Feature Engineering to the available data so that the Machine Learning algorithm understands it, betters the accuracy, and free the errors from the data.

2. Create a suitable Machine Learning Model with suitable parameters. (Suitable parameters for KMeans and hierarchical clustering are found using elbow method, and for DBSCAN, the parameters are hyper tuned using silhouette validation technique.

3. Fit the n-Dimensional data to the model. (n-Dimensional data has n features)

4. Predict the labels, and add append these values to the data frame.

5. Visualize the data with different clusters formed.

6. Now we can treat each cluster differently and efficiently based on their habits.

**6.Methodology:**

Here we will see how the clustering is actually carried out.

Now, we have come across the ideas that on what basis we can divide the customers into groups using machine learning. But how this is technically done?

Let’s go through different clustering algorithms.

1. K-Means Clustering Algorithm

2. Hierarchical Clustering

3. DBSCAN Clustering

and then we use silhouette validation technique to validate the clustering models.

**6.1 KMeans Clustering Methodology:**

K-Means algorithm performs division of clusters into clusters which are similar in characteristics, and those datapoints which are dissimilar belong to another cluster.

Procedure:

1. The Elbow Method: Find the value of k, the optimal number of clusters for the data we have. We do this using the elbow method. We do this by calculating an important value known as wcss (within cluster sum of squares), meaning the sum of k SSEs of each cluster among 'k' clusters. Where SSE for a cluster is the sum of squares of distances from each point in a cluster with its centroid.

So, we calculate WCSS values for each k value. Let those be WCSS(k).

Next, we plot the graph between WCSS(K) and k. The k value at which you observe an elbow kind of decrease shape in the graph is our optimum number of clusters. So finally, we got optimal 'K' value.

2. Now coming to the main procedure, plot the graph, and assign 'K' random centroids on the plane. Now calculate the distances between the points and find out which centroid is closer to each point, and assign this point that corresponding cluster. So now we have k clusters.

Calculate new centroids for each cluster based on the points we have in each cluster.

Repeat the same process iteratively until the position of the new centroids do not change. So, this completes out KMeans clustering. The clusters to which the point is associated are the final cluster labels.

3. So now we have each point assigned to each cluster. We retrieve those labels, and treat the customers based on the cluster labels.

4. We can even visualize the clusters with a 2-D or 3-D plot, and view the characteristics of each cluster plotted within those set of variables.

**6.1.1 Advantages of KMeans clustering:**

* Relatively simple to implement.
* Scales to large data sets.
* Guarantees convergence.
* Can start well with the positions of centroids.
* Easily adapts to new examples.

**6.1.2 Disadvantages of KMeans clustering:**

* Choosing k value manually.
* Being dependent on initial values of centroids.
* Clustering outliers.

**6.2 Hierarchical Clustering Methodology:**

Hierarchical clustering does same as KMeans clustering, but in a different methodology. It basically follows two kinds of approaches. They are Aggloromative clustering and divisive clustering.

Aggloromative clustering is a bottom-up approach. Divisive clustering is a bottom-up approach.

Procedure:

1. First we find out the optimal clusters k using the same elbow method.

2. Now we use Aggloromative clustering here. In Aggloromative clustering, we initially consider each data point as an individual cluster.

3. We find the nearest clusters and group them as a single cluster.

4. We iterate the step 3 until we remail with k number of clusters finally.

5. So we have formed k clusters using Hierarchical clustering.

6. Similarly we can perform the clustering using Divisive clustering.

7. Initially consider all the data points as a same cluster, and keep dividing the dissimilar clusters iteratively until we remain with k clusters.

8. So now we have each point assigned to each cluster. We retrieve those labels, and treat the customers based on the cluster labels.

9. We can even visualize the clusters with a 2-D or 3-D plot, and view the characteristics of each cluster plotted within those set of variables.

**6.2.1 Advantages of Hierarchical clustering:**

* easy to implement.
* Dendrogram visualization helps to view big picture of clusters

**6.2.2 Disadvantages of Hierarchical clustering:**

* Does not handle big datasets efficiently.
* Dendrogram is commonly misinterpreted.

**6.3 DBSCAN Clustering Methodology:**

DBSCAN stands for Density Based Spatial Clustering of Applications with Noise.

In DBSCAN, we cluster the data points based on the density of the points. DBSCAN can really perform very well sometimes and even better than KMeans and Hierarchical clustering.

Procedure:

1. Here in DBSCAN, we don’t need the optimal clusters parameter, it finds k value itself. But there are certain parameters we need to provide.

2. The parameters to be provided are:

* Epsilon (the radius of the circle the algorithm forms)
* Min points (the minimum number of points to be inside of the circle with radius equal to epsilon).

3. So based on these parameters, we categorize each point in the dataset as:

* core point: a point is considered as core point if no. of points in its circle is at least equal to Min points
* Border points: does not satisfy the core point property but has at least 1 core point in its circle.
* Noise points: Neither core points, nor Border points.

4. The noise points are eliminated by the DBSCAN algorithm and are not considered under any cluster. Hence DBSCAN handles outliers very efficiently, bus has own drawbacks due to this.

5. So now we have each point assigned to each cluster. We retrieve those labels, and treat the customers based on the cluster labels.

6. We can even visualize the clusters with a 2-D or 3-D plot, and view the characteristics of each cluster plotted within those set of variables.

**6.3.1 Advantages of DBSCAN clustering:**

* Resistant to noise and handles outliers efficiently
* Can handle clusters for various sizes and of different shapes

**6.3.2 Disadvantages of DBSCAN clustering:**

* Cannot cluster data with large differences in densities well.
* Choosing a proper value for min\_points and eps is somewhat difficult.

**7.Implementation:**

**7.1 KMeans Clustering Algorithm, code:**

Assume that the optimal number of clusters are calculated using elbow method.

1. Pick k random points as cluster centroids. centroids = [c1, c2, c3,.ck]

2. Assign each point xi to nearest cluster by calculate its distance to each cluster centroid.

3. Find new centroids by taking the average of all the points in each cluster separately.

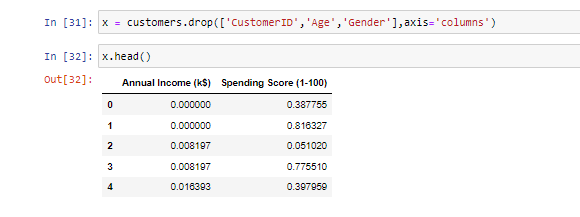
4. Repeat steps 2,3 until none of cluster positions change.

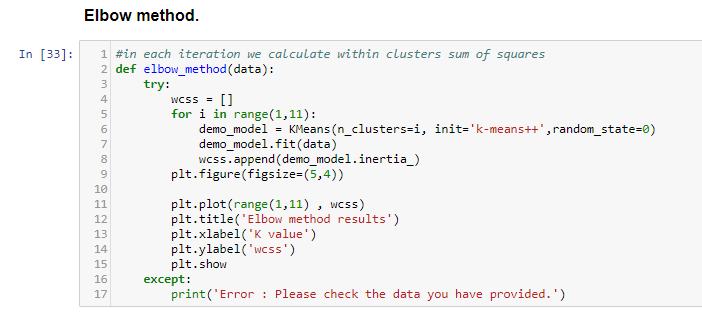
5. Get all the labels of each customer, which denotes the customer's cluster.

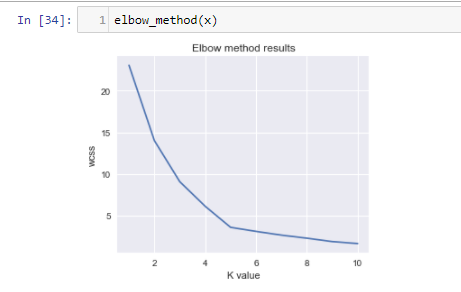
6. Now save these labels in the data frame and use them to treat the customers effectively

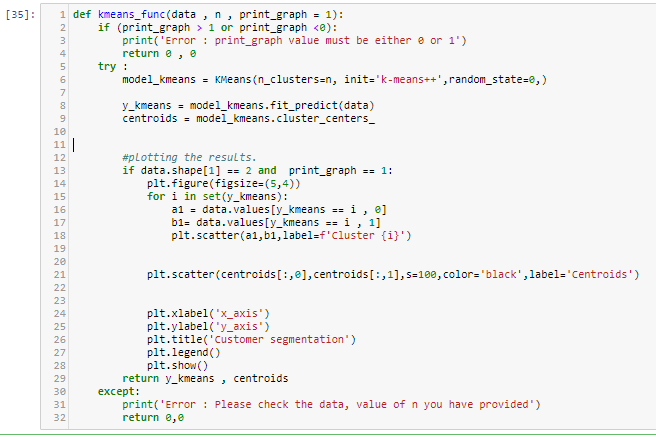
Now let’s go to the actual code and group the customers based on the relationship between:

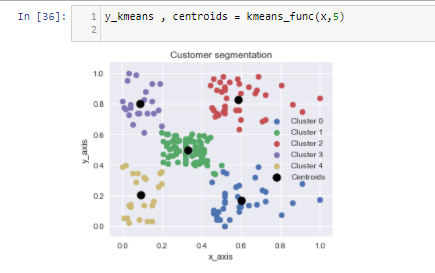
‘Annual Income’ and spending score

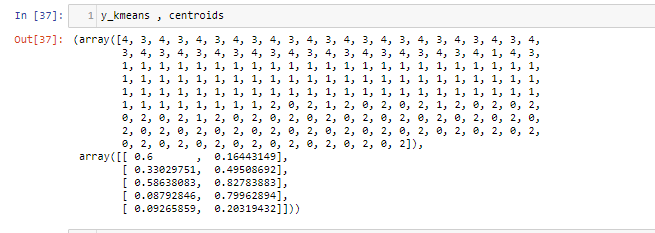












**7.2 Hierarchical Clustering Algorithm, code:**

Algorithm for Aggloromative clustering:

Assume that the optimal number of clusters are calculated using elbow method.

1. We have n data points. So, we have n clusters initially.

2. Calculate the centroids ci of each cluster.

3. Find the nearest clusters and make them as a single cluster.

4. Calculate the new centroids for the new clusters formed.

5. Repeat the steps 3,4 iteratively until we remain with k clusters.

6. Get all the labels of each customer, which denotes the customer's cluster.

7. Now save these labels in the database and use them to treat the customers effectively

Algorithm for Divisive clustering:

Assume that the optimal number of clusters are calculated using elbow method.

1. Initially consider all the data points as a single cluster.

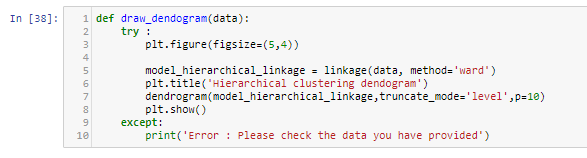
2. Divide this cluster into 2 clusters, for which the cluster split has largest cluster sum of squares.

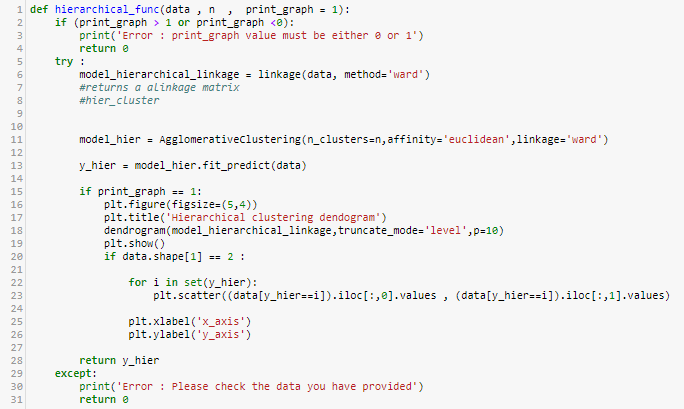
3. Follow step 2 iteratively until we are left with k clusters.

4. Get all the labels of each customer, which denotes the customer's cluster.

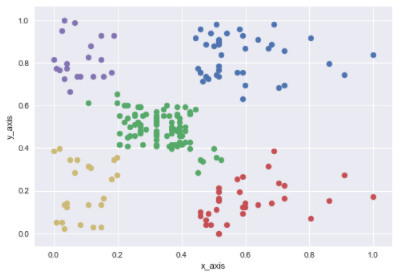
5. Now save these labels in the data frame and use them to treat the customers effectively

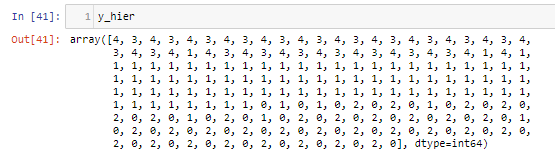
Drawing Dendrogram:











**7.3 DBSCAN Clustering Algorithm, code:**

1. Select a point from the dataset.

2. Provide the parameters - epsilon, min\_samples

3. Match all the reachable points and check whether a point falls under core point category, border point category, or noise point category.

4. Classify each point into one of these three categories until there are no points are reachable.

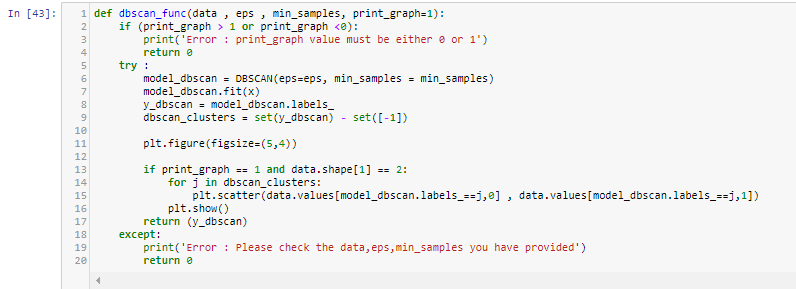
5. So this is the end of calculating 1 cluster.

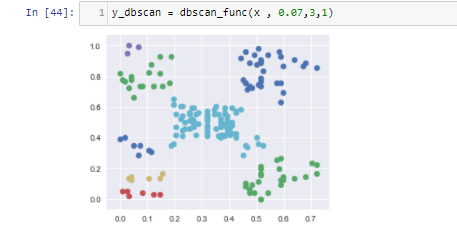
6. Now pick up any one point from the data which is not clustered.

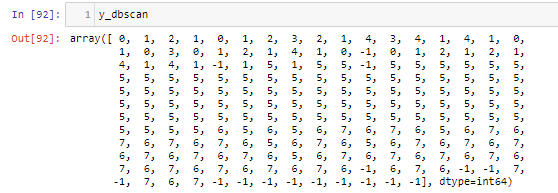
7. Repeat the steps 3,4,5 until there are no points left those are not clustered in the dataset.

8. Get all the labels of each customer, which denotes the customer's cluster.

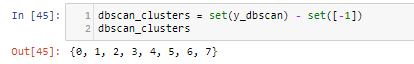
9. Now save these labels in the data frame and use them to treat the customers effectively







The customers with labels -1, are outliers and doesn’t fall in any cluster.



**7.4 Silhouette Cluster Validation technique:**

It is used to find/Check/Validate whether the chosen number of clusters for a specific model is correct and optimal or not.

Silhouette score (or value) for a data point i, is given in terms of a(i) and b(i), where ai is the mean distance from point i to all other points in the same cluster. And bi is the mean distance from point i to all other points in the nearest cluster.

silhouette value for a point i = s(i) = (b(i) - a(i)) / max{a(i), b(i)}, if |Ci| >=1

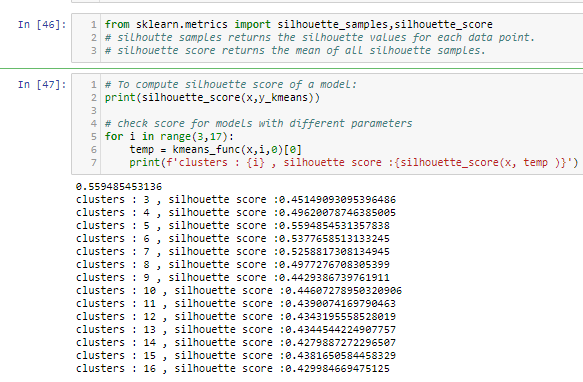
(Ci is the cluster in which i is present)

silhouette value ranges between -1 to +1.

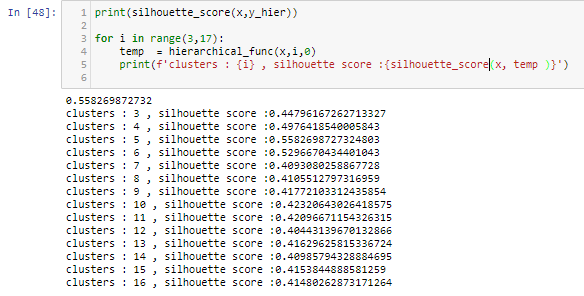
Final silhouette value of a model = avg(s(i)), i =1 to n

More the silhouette value, clustering is done properly.

Calculating silhouette score for KMeans for different values of clusters:

****

Calculating silhouette score for Hierarchical for different values of clusters:



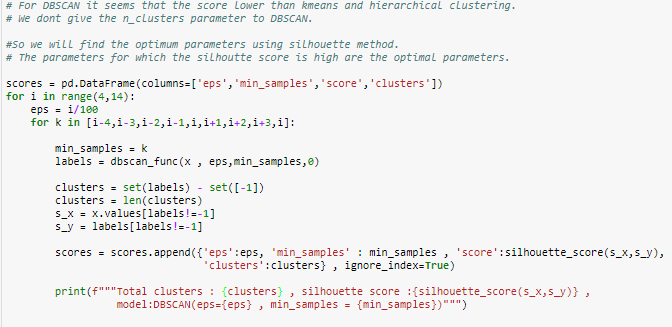
Calculating silhouette score for DBSCAN model:

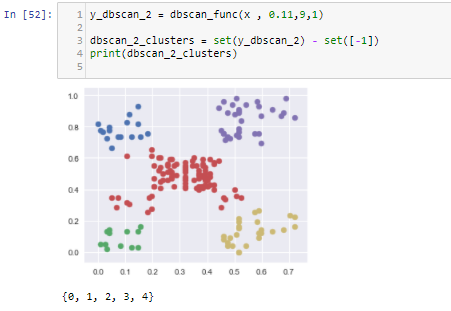


This score is lesser than the other models, so lets find the optimum values by hyper parameter tuning using Silhouette validation technique.

Hyper parameter tuning for DBSCAN:

Initially the silhouette score of clusters formed by DBSCAN algorithm might be low because of the unknown optimal parameters. We can hyper tune these parameters using silhouette validation technique.





**7.5 Finding the BEST model:**

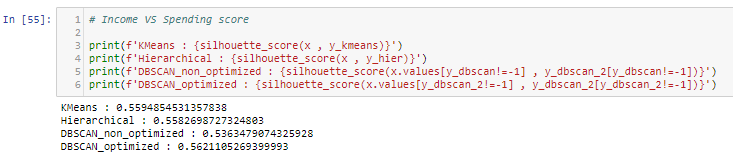
We have clustered all the data into k clusters. So, each data point has its own cluster label value. These label values are predicted by three algorithms now, KMeans, Hierarchical, DBSCAN. How to find the best model among them.?

Simple find the model with highest silhouette score. A good idea but there is a trap here. Even if we have highest silhouette score for DBSCAN algorithm, we should not choose this because we know that DBSCAN handles outliers by eliminating them. But we don’t want to lose any one of the customer data points here. So, we eliminate DBSCAN.

Now among KMeans and Hierarchical models, for all combinations of input data points, the silhouette score is more for the KMeans Model.

So, the best model for this customer segmentation dataset is KMeans Clustering Model.

The labels outputted by this model is considered ultimately.



**8. Conclusion:**

We have found out the best model, which has high silhouette score, and using this model, we predict the cluster labels, save these labels in the data frame and then visualize the data.

Till now I have only discussed that I have calculated the cluster labels for the data: Annual VS Income.

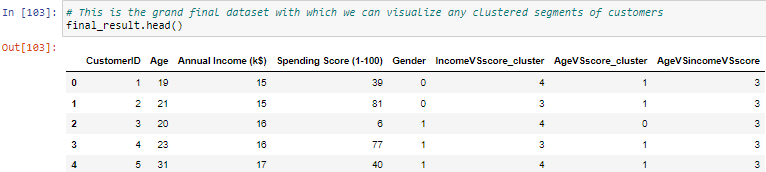
But I also have done clustering based on 2 other **meaningful** combinations.

They are: Age VS score, and Age VS Income VS score.

So finally, we get three columns of cluster labels.

The reason that we don’t have any traces of Gender in clustering because, Gender itself is a classification of two different clusters. And even if we consider Gender in the input data combination, we get the label values same as the Gender values. So, this doesn’t make sense.

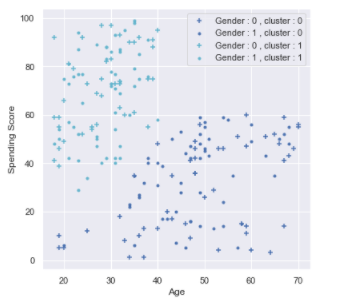
The final data frame we have now is:



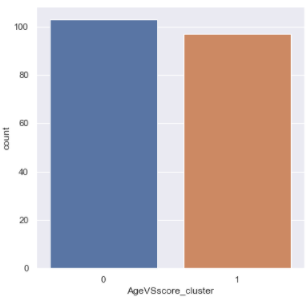
Now we got clusters of the customers corresponding to various combinations of features. Now using these cluster labels, we visualize the data and hence we can market each group effectively.

**8.1 Visualizing the Clustered data for different combinations of features and analyzing them:**

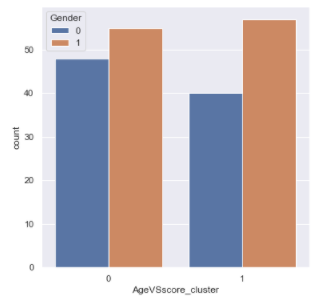
1. Visualizing the clustered data for the plot: Age **VS** Spending score:

****

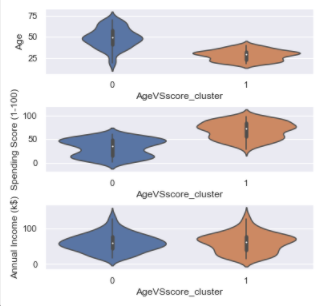
This is scatter plot, with two different clusters, along with Gender as hue. Gender = 0 means that customer is a male, and if Gender = 1 means, customer is a female.



This is a count plot for this cluster column. It says that, above 100 customers belong to cluster 0, and below 100 customers belong to cluster 1.

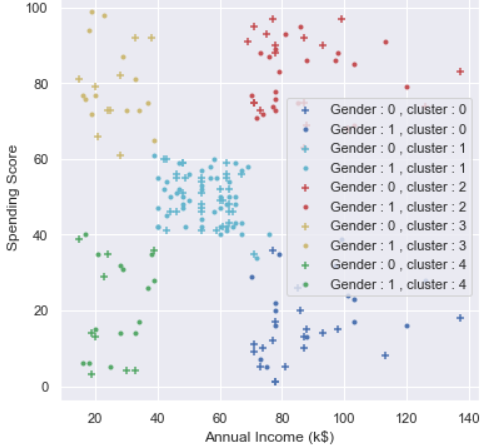
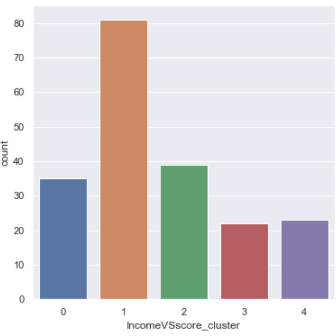


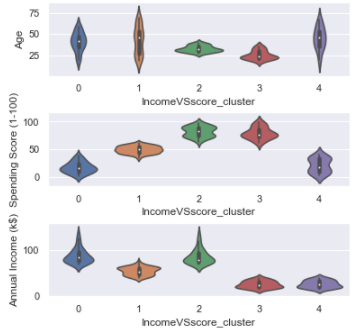
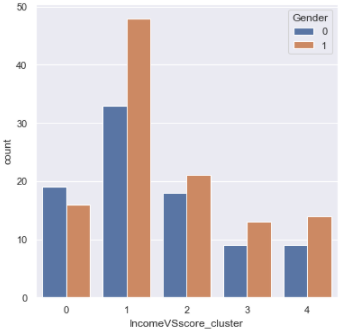
Now This count plot provides little more information about each individual cluster. It says how many are male and female in both the clusters separately.



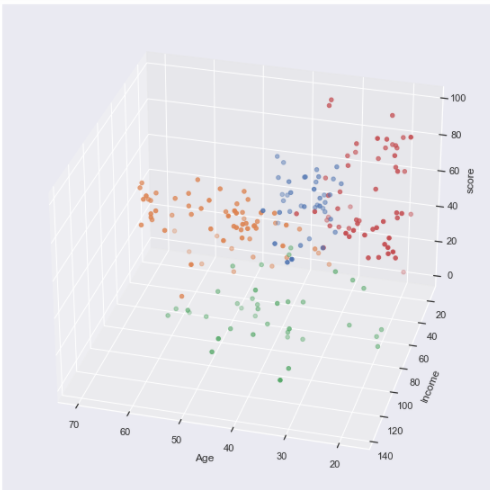
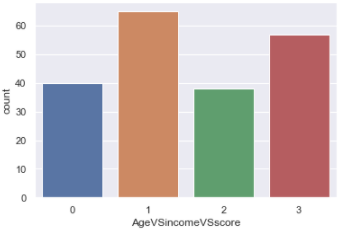
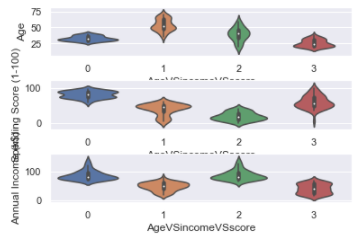
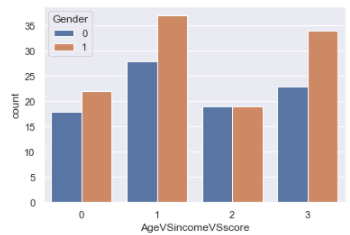
This violin plot provides the information about the distribution of different columns in each cluster.

1. Visualizing the clustered data for the plot: Annual Income **VS** Spending score:

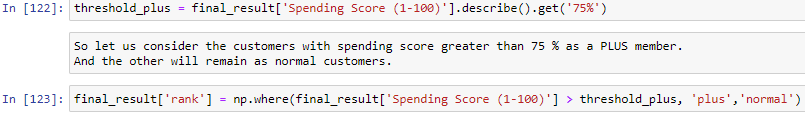
** **



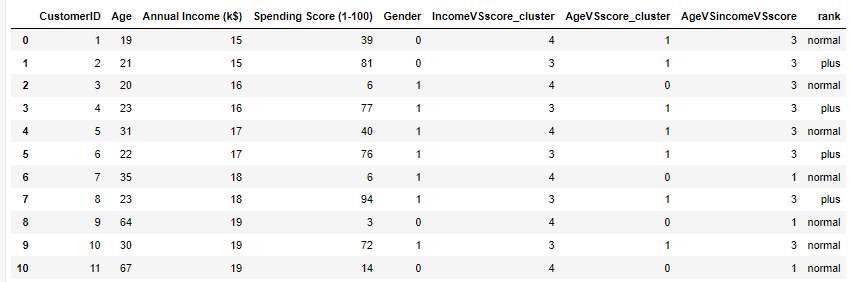
1. Visualizing the clustered data for the plot: Age **VS** Annual Income **VS** Spending score:

**** **** ****

**8.2 Finally Ranking the customers based on their spending scores:**



Customers with assigned rank:

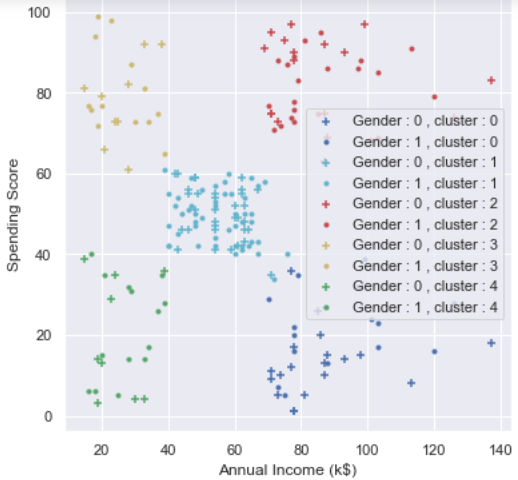


**8.3 Treating separate clusters effectively:**

Now we have also visualized the data, so let’s take one of those cluster columns, and find out what should we do in order to perform an effective marketing.

Let’s take example of cluster: Income VS score.

The scatter plot for this is:



Let’s do a small analysis which is useful and meaningful.

Let’s say, our Aim is to increase the spending score. We have to look at the clusters which has less spending scores. Clusters 0,4 has very less spending score.

* Treating cluster 0 (violet colored cluster): The customers of these cluster have reasonable annual income, so there is no problem of financial shortage here. We can make them spend more money on our products buy sending more notifications on our products, and making more attractive deals.
* Treating cluster 4 (green colored cluster): On the other hand, the customers of this cluster has less annual income. So, they can’t spend more money at a time. The solution to increase the spending scores of this category is to offer more discounts and offers than usual.

So, this is a small analysis I have done to increase the spending scores of the customers in order to improve the business and marketing strategy.

Therefore, I want to conclude that, with the help of customer segmentation, we can market each group effectively and with ease, providing and satisfying the customers with appropriate needs.

This is the report of my project on **Data Science - Customer Segmentation**.

Thankyou **Exposys Data Labs** for this opportunity.

**A Report by:** Masabathula V S Raghavendra Rao