Name: Rahul Vijayvargiya
Student: 245784

Task:
Heart failure clinical records Link for dataset:

Dataset Link

The collection contains data related to cases of heart failure. Models learned from this dataset were used to assess patient survival and aid treatment selection. The collection is adapted for classification and clustering tasks – for this assignment it will be used for clustering.

Sol.

Hello, We have a dataset of clinical Heart Disease and people who got affected with respect their age, gender and cause of death and survive from the disease,

Here we are going to use clustering, unsupervised machine learning problem, with respect K-Means and DBSCAN

K-Means:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.

DBSCAN:

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers). The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

Is DBSCAN better than KMeans?

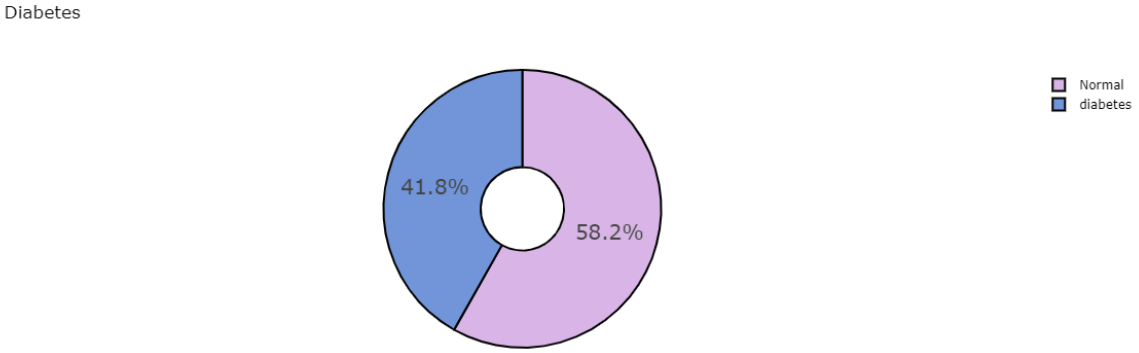
K-Means	**DBSCAN**
K-means has difficulty with non-globular clusters and clusters of multiple sizes.	DBSCAN is used to handle clusters of multiple sizes and structures and is not powerfully influenced by noise or outliers.

Here we go with our report and data preprocessing part:

```
Index(['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes',
       'ejection_fraction', 'high_blood_pressure', 'platelets',
       'serum_creatinine', 'serum_sodium', 'sex', 'smoking', 'time',
       'DEATH_EVENT'], dtype='object')
```

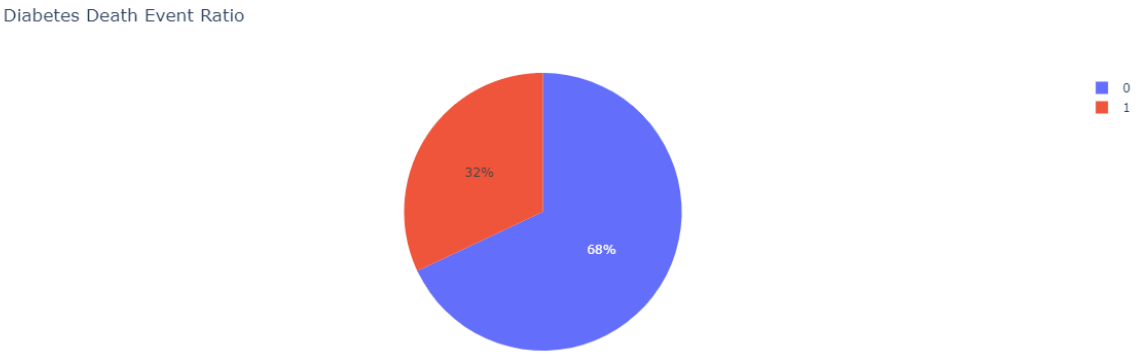
As you can see above the columns name of our dataset tells us about age, the disease and sex and death, he or she survived or died from diease

1.



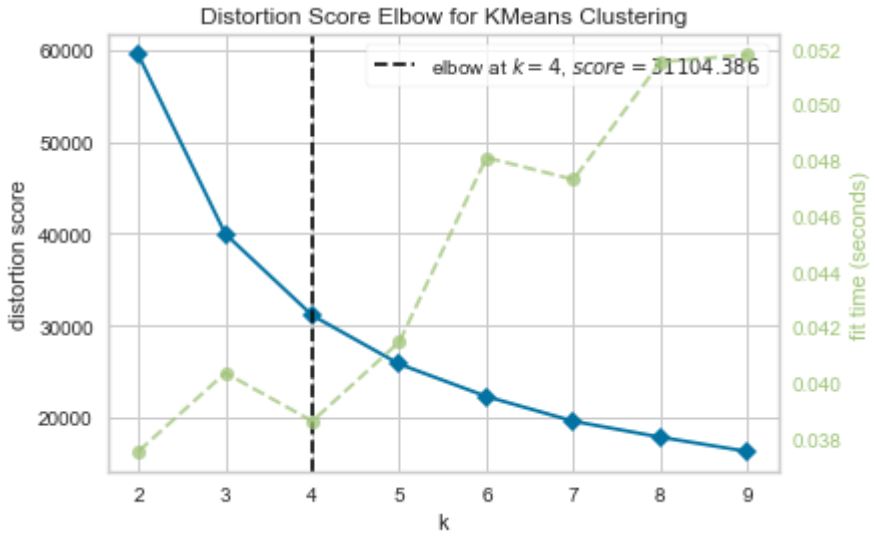
Above you see total no. of people are affected from diabetes or normal person in the dataset

2. plot



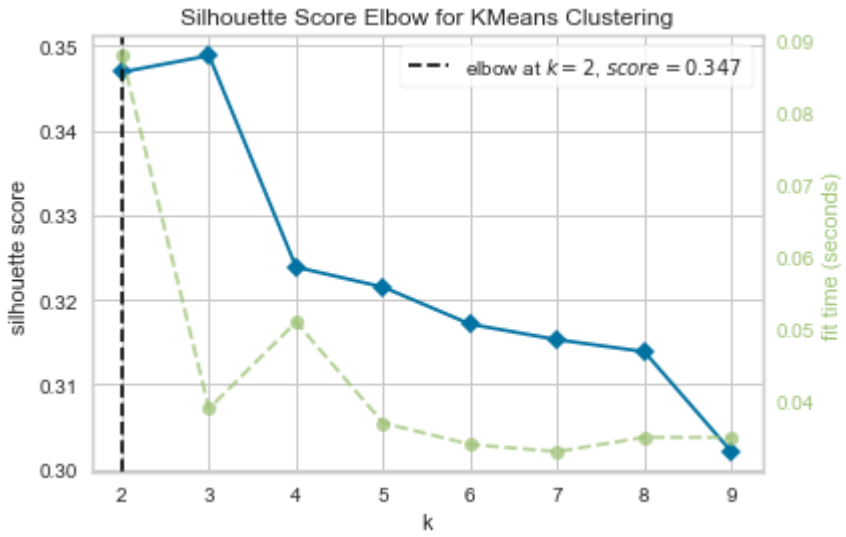
People died from diabetes event ratio

3. Metrics Distortion score



Distortion: It is calculated as the average of the squared distances from the cluster centres of the respective clusters. It suggest us to form 4 clusters, highest score is at 4, making a Elbow

4. Silhouette score



Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique.
We calculate centre point distance -1 to all the points within cluster and from that same centre point we calculate the distance nearest cluster data points also we average them and Use below method to calculate coef.

We now define a silhouette (value) of one data

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$

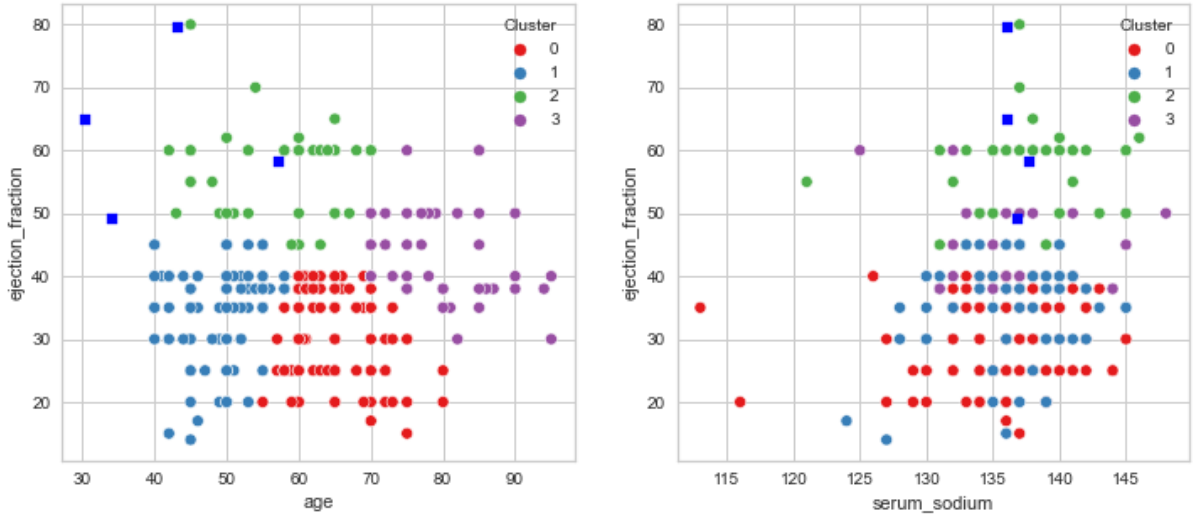
The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.

The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster

Since the choice of method is unsupervised learning, so we do not have any label data to form cluster, for k-means we provide k value, as per k value it will gonna form a cluster

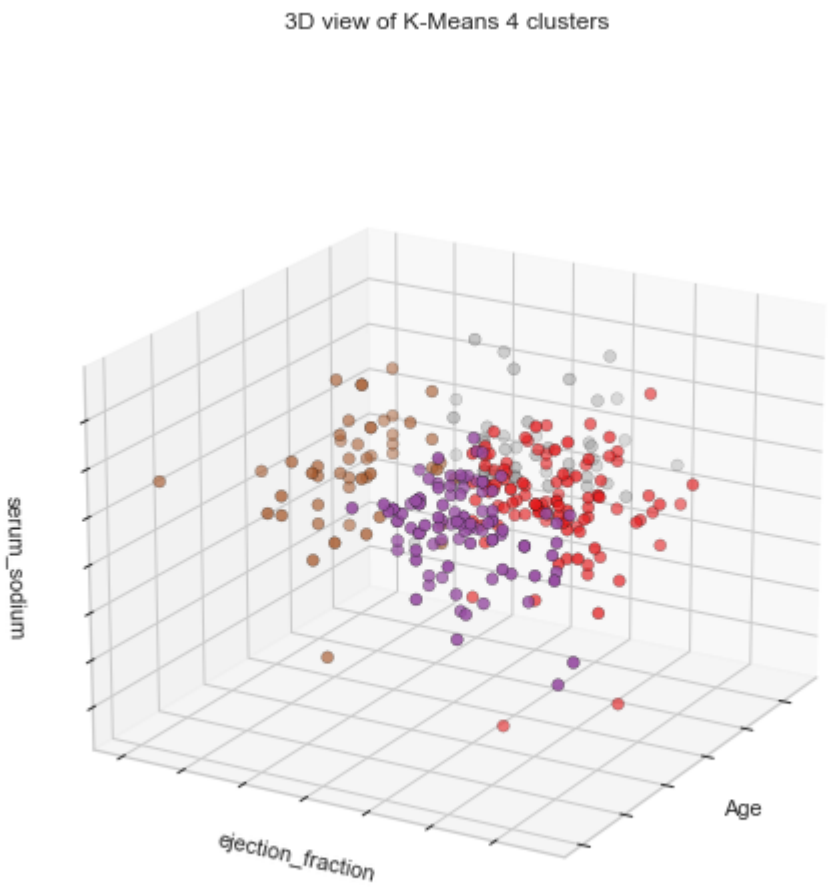
4 forming cluster using K Means, where K is 4:

2D View:



Silhouette coef = 0.3243696465249645

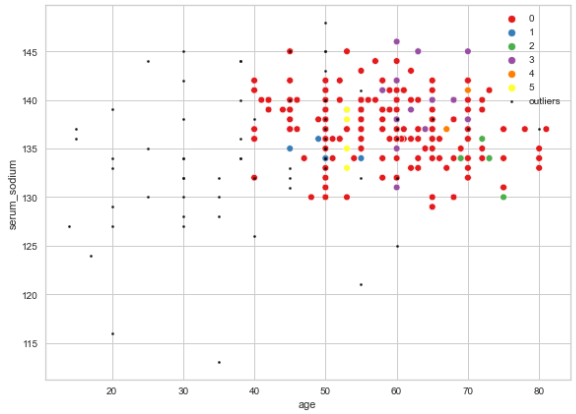
3D view:

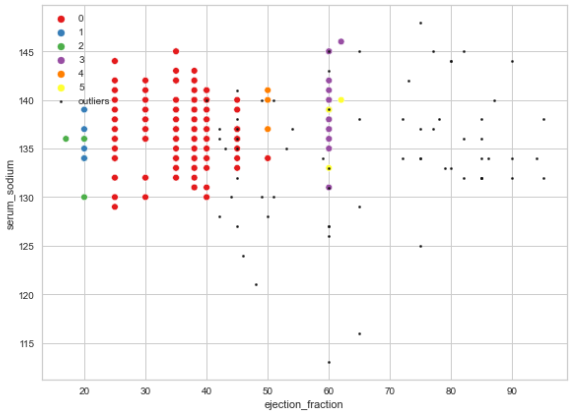


DBSCAN, it’s a density based clustering works on epsilon and minimum sample in cluster to form a cluster

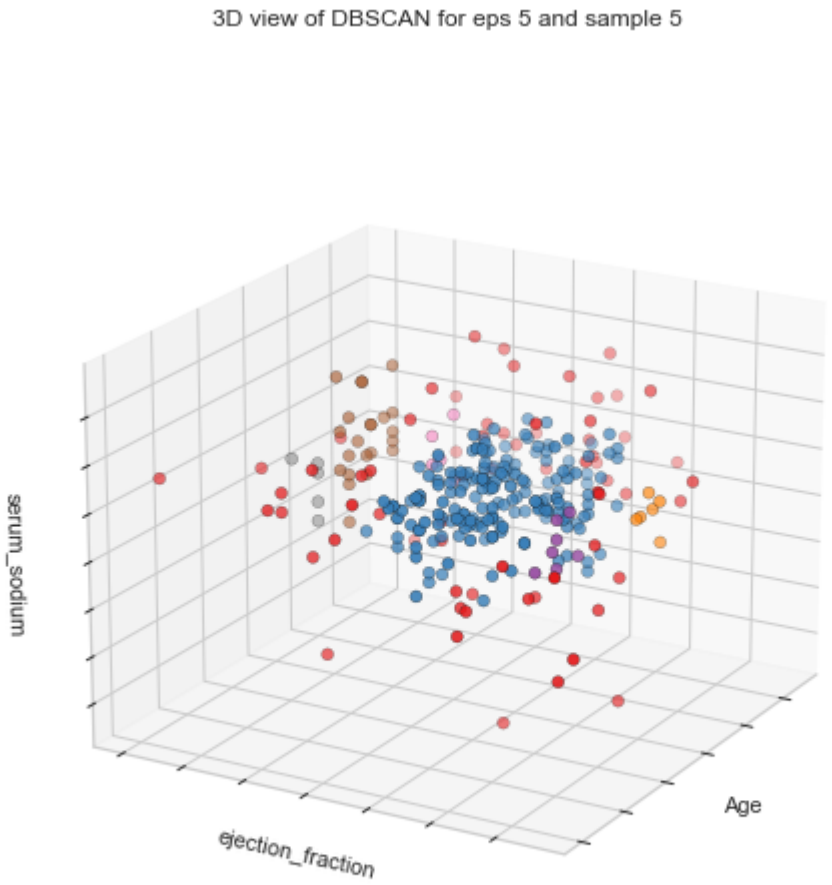
As you can see a epsilon value is 5 which means the radius or cluster is 5 and minimum samples means that it should have at least 5 samples, minimum 5 samples to form a cluster

2D View:



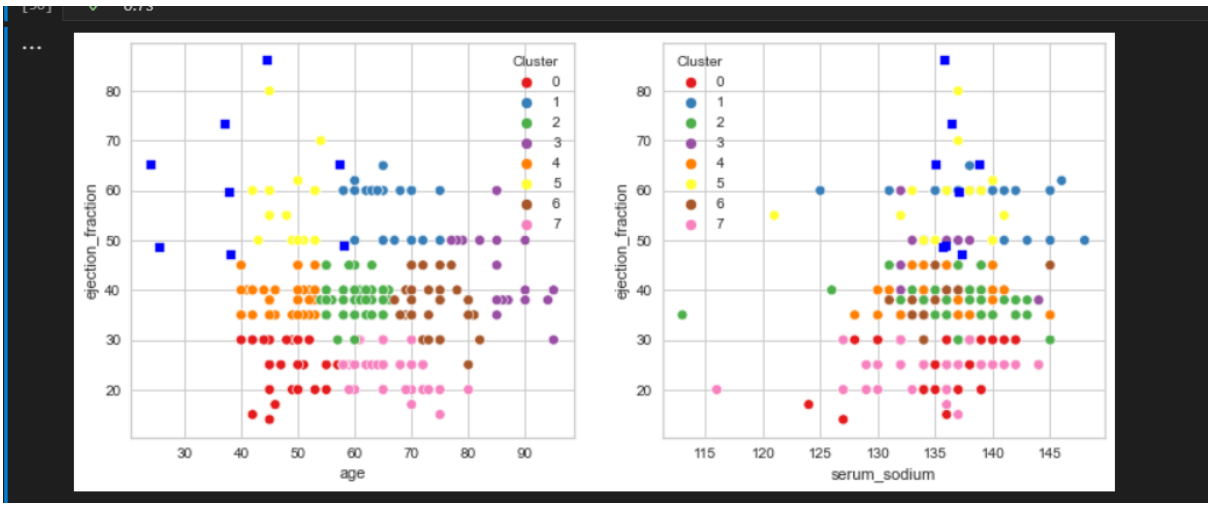


Silhouette coef: -0.02
3D View:

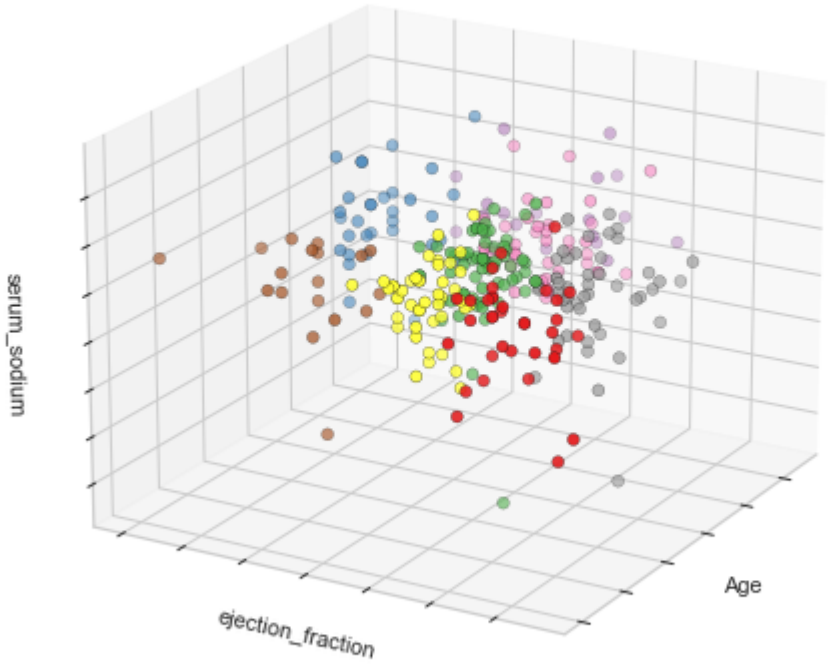


Experiments:

With K-means form 8 clusters

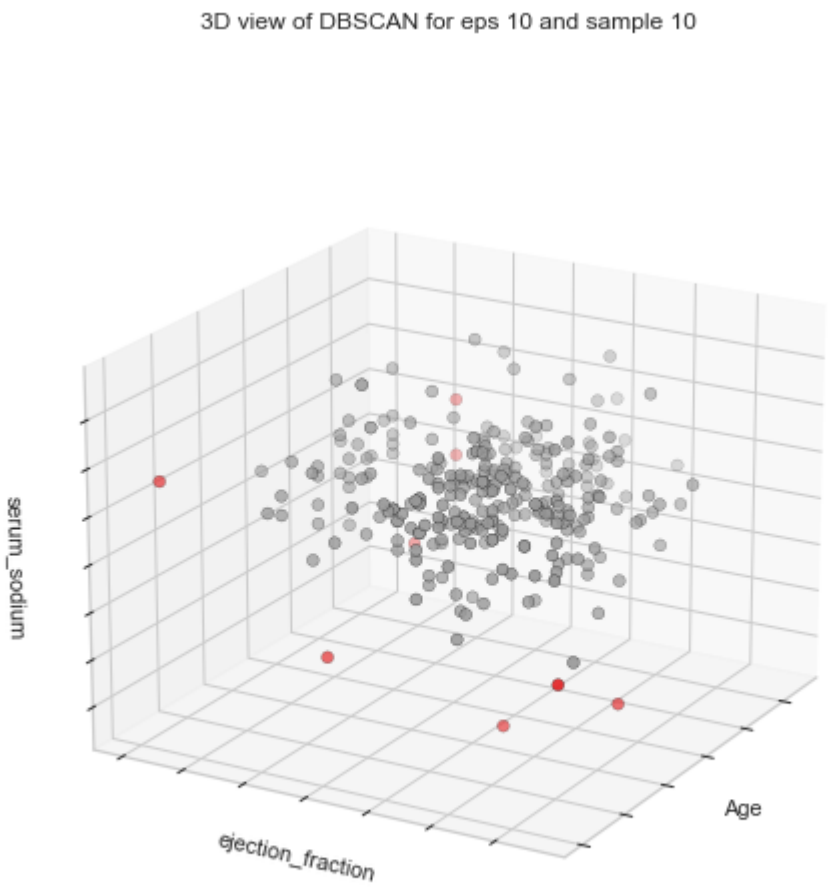


3D view of K-Means 8 clusters



8 clusters in 3D Pane

DBSCAN:
EPS: 10 and Sample Size 10, it formed a 1 big cluster and -1 stands for outliers

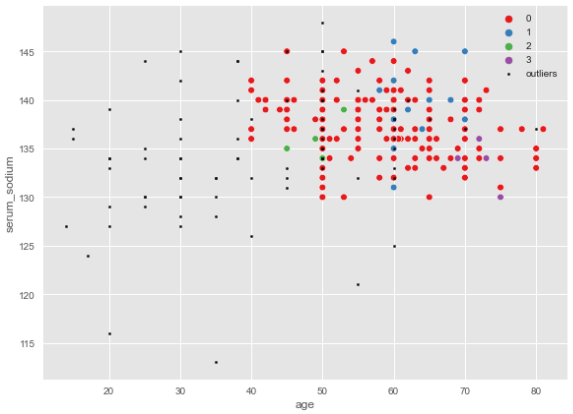


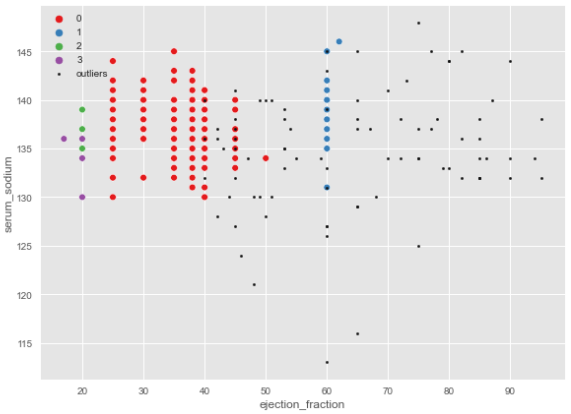
Best Silhouette Score so far on DBSCAN:

Eps =5 and minimum samples=6

Coef = 0.010936652241343991

Total Outliers in DBSCAN using following parameter we get 85





3D view of DBSCAN for eps 5 and sample 6