Project II —Clustering

Group 18
Jiapeng Sun
Hao Xu
2021/5/7

Task 1

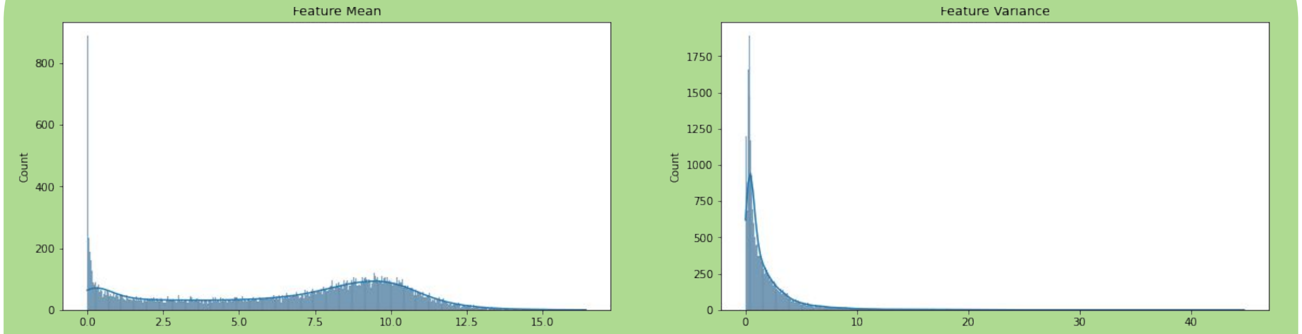
Step 1 – Missing value checking

By checking, there is no missing value in the data

Step 3 – constant features checking

By checking, there are 267 features with 0 variance.

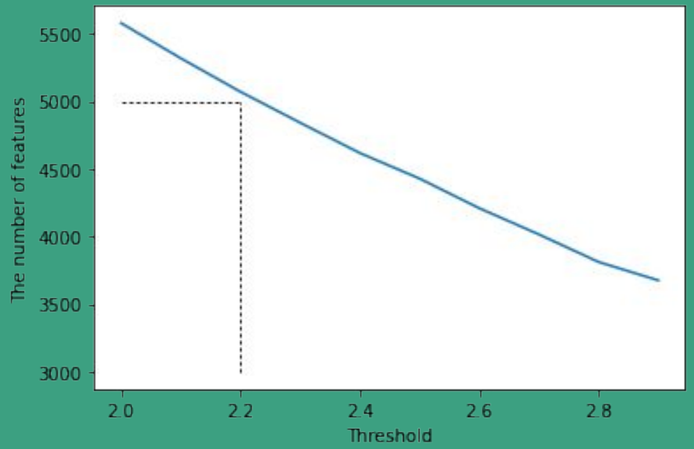
Step 2 - means and variances of the features



The dataset has 20531 features. These two figures shows the mean and variance values for each feature. Here, we can see that the distribution of mean values is pretty even, from 0 to 15. For variance, however, most are nearly 0 and the number drops sharply when the variance increases.

Task 2

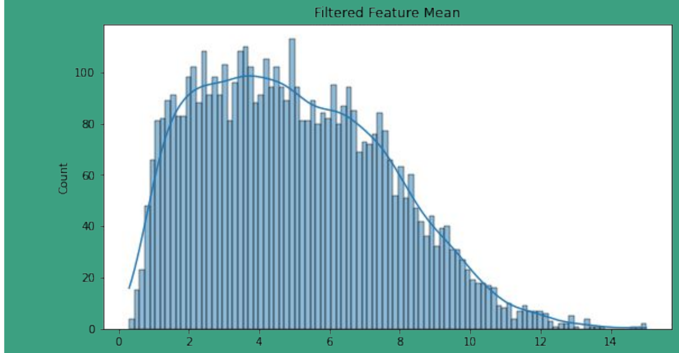
Step 1 – using variance filtering to reduce the number of features

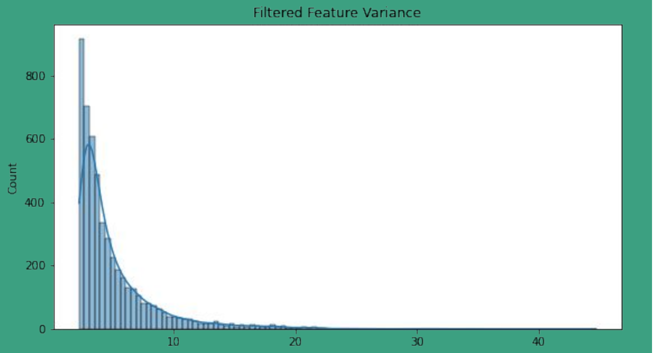


Here, we test the value of threshold from 2 to 3. We can find that the 5000 is located at around 2.2. Thus, we set the threshold as 2.2 and the number of features in fact is 5071.

Task 2

Step 2 – Check the means and variances for each feature again





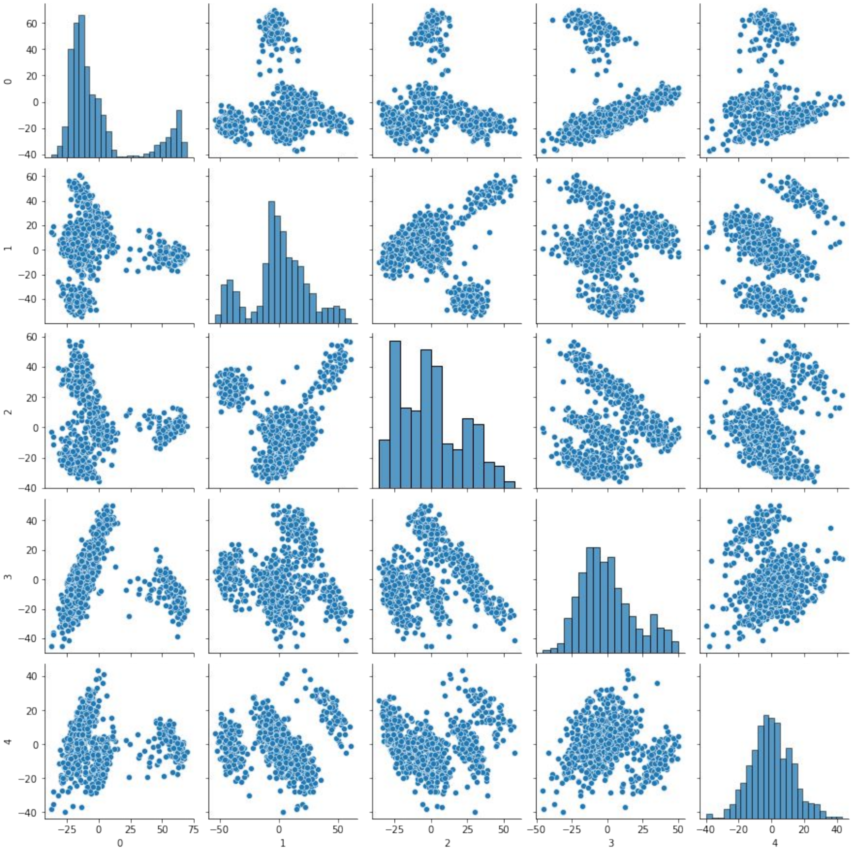
Now, we can see that the distribution of means is much more centered. And we need to center and standardize the data before applying PCA, in order to improve the performance.

Task 3

Firstly, we applied the PCA (n_components=50) on the filtered data.

The figure here is the pair plot of the first 5 components. We can see that there might be 2 to 4 clusters:

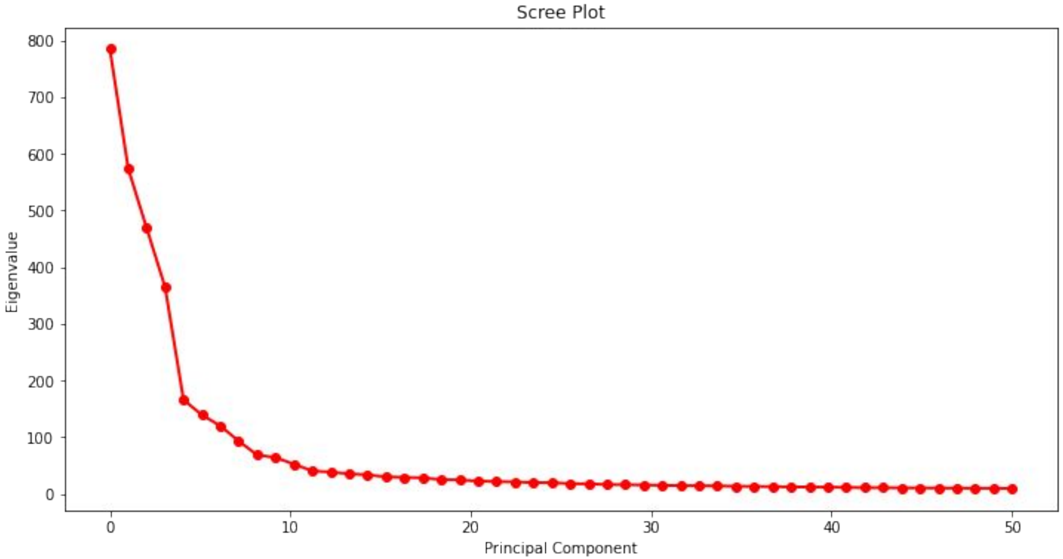
- Component 3 to 0 shows likely 2 clusters
- Component 2 to 1 shows likely 3 clusters
- Component 3 to 1 shows likely 4 clusters



Task 3

The figure here is the scree plot. We see there is an elbow at around 5.

It shows a potential that we may keep only first 5 components from the PCA, so we will perform PCA again before next steps.



Task 4.1 - K-means

Evaluation:

- **Silhouette Score**

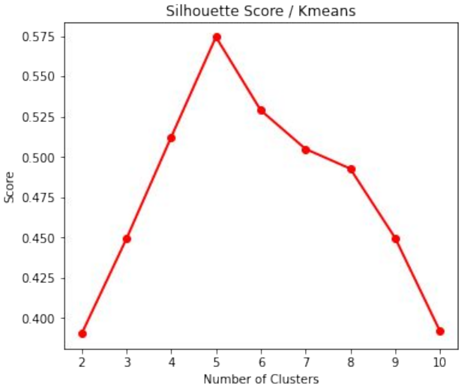
  The higher, the better

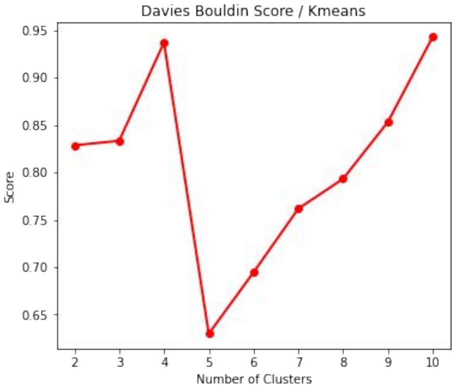
- **Davies Bouldin Score**

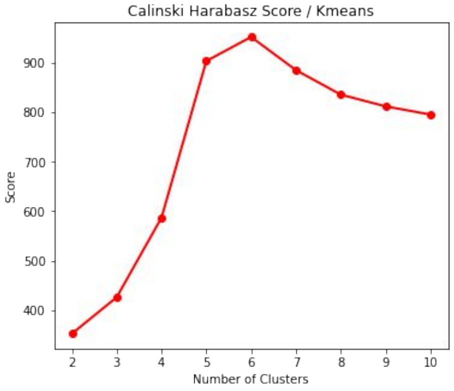
  The lower, the better

- **Calinski Harabasz Score**

  The higher, the better





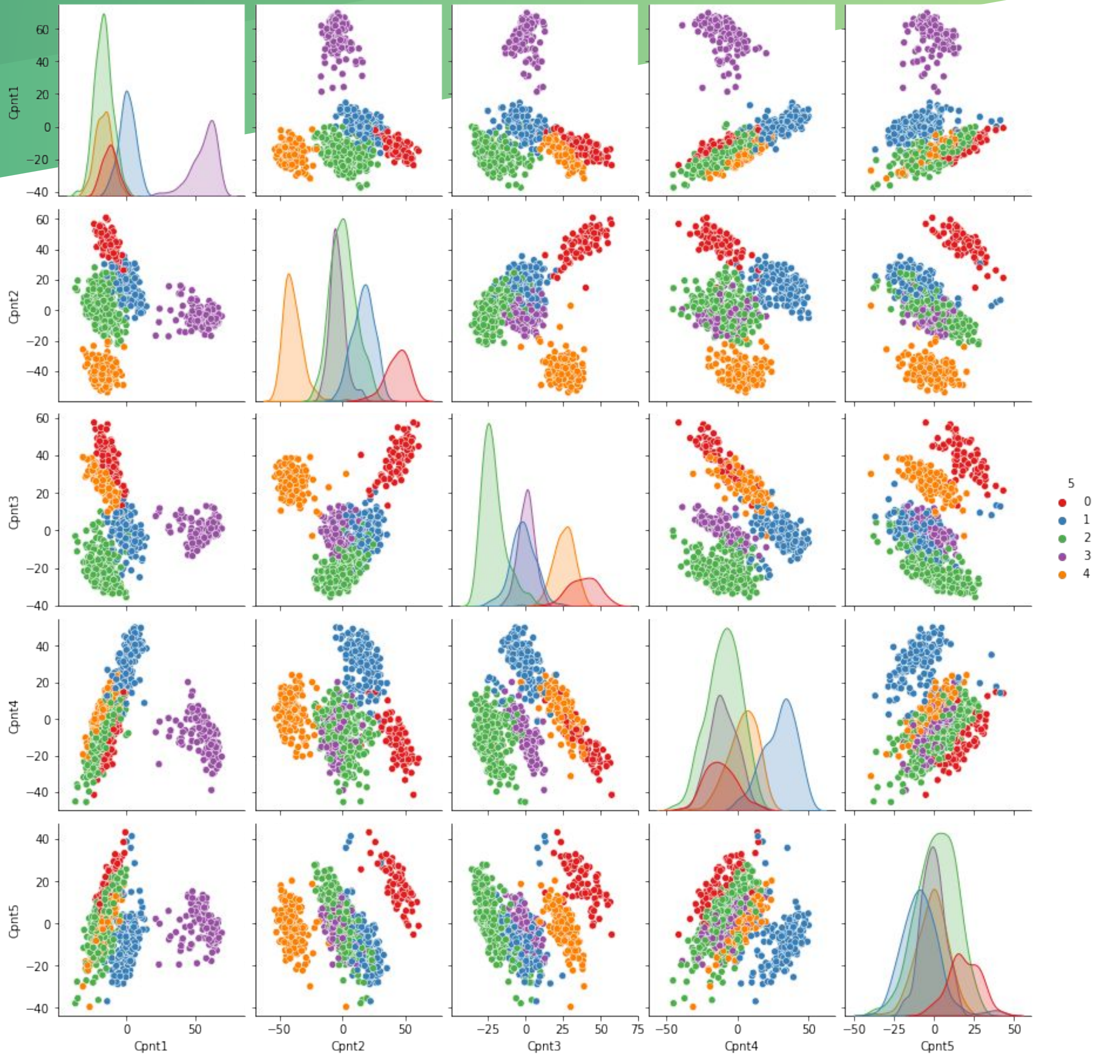


We tested K = 2 to 10

Considering the three indexes comprehensively

We choose 5 as the final clustering number

Task 4.1 - K-means



Here is the pair plot of the data by all 5 components.

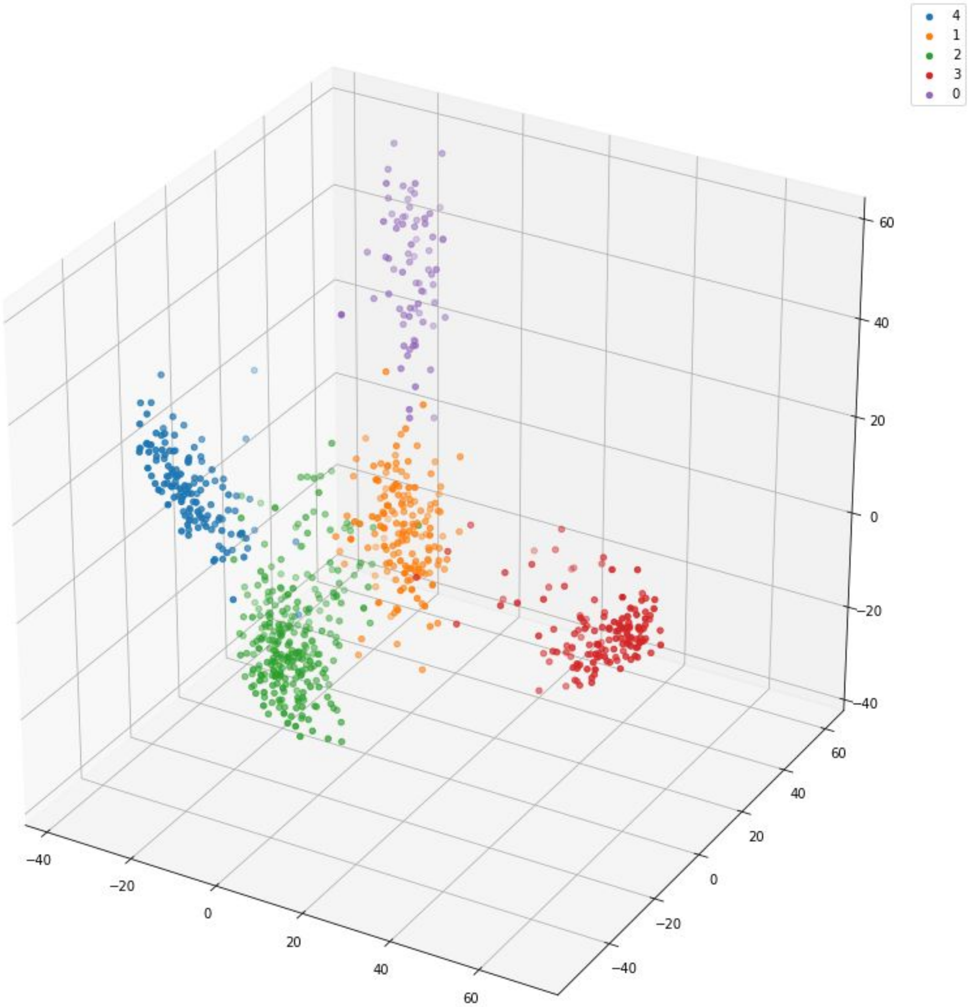
Different clusters are labeled in different colors.

Task 4.1 - K-means

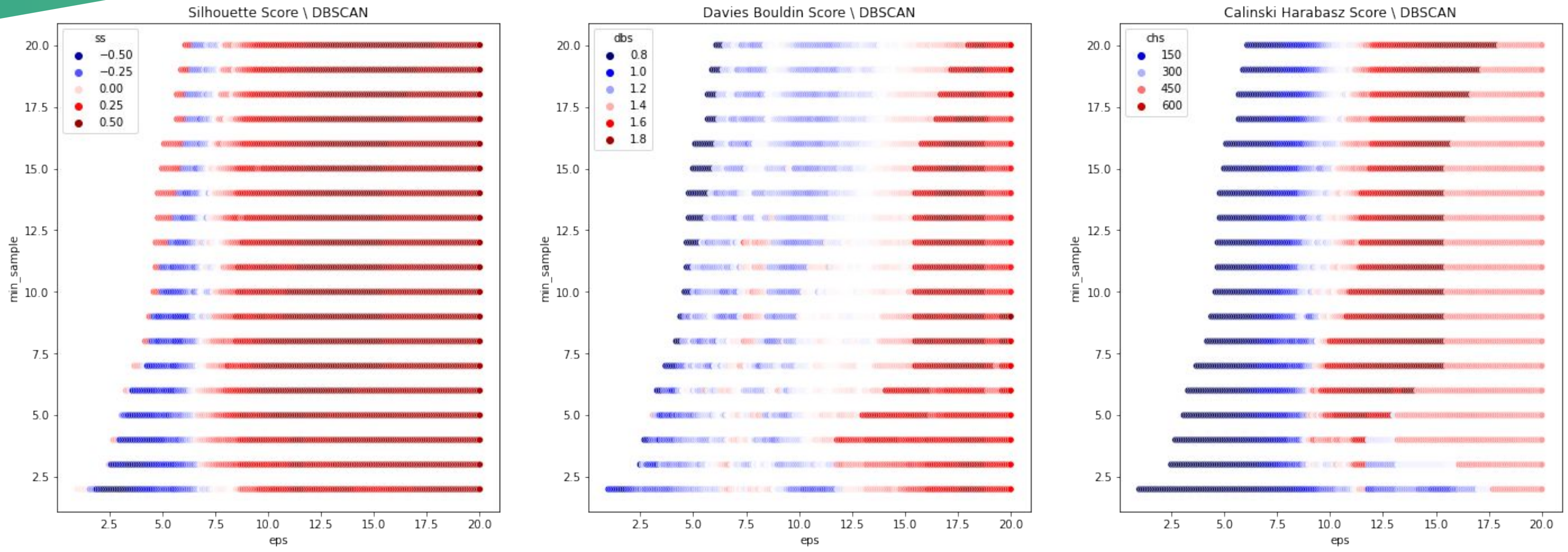
Here is the 3D plot of the data by the first three components.

Different clusters are labeled in different colors.

From the plot we can see the five clusters.



Task 4.2 - DBSCAN



We tested eps = 0 to 20 with step 0.1 and min_sample = 2 to 20

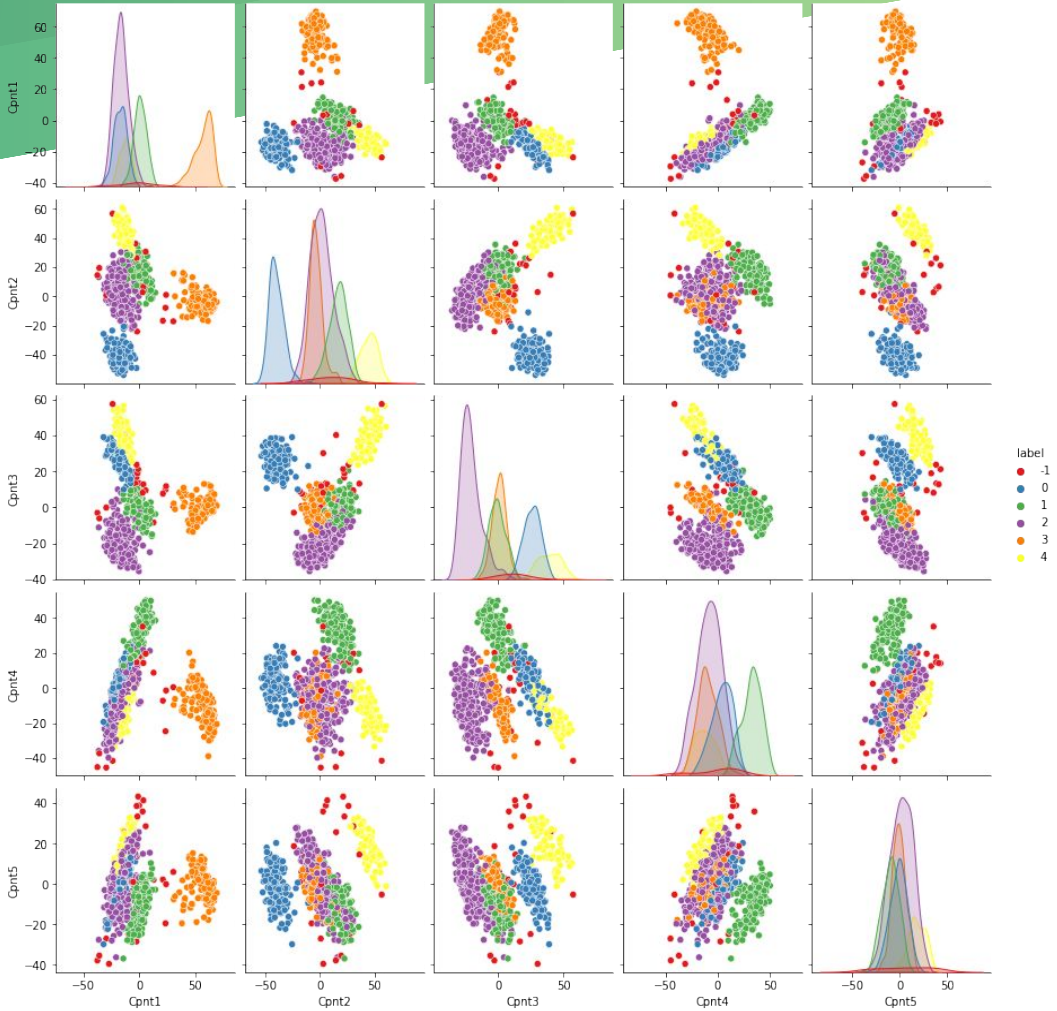
Considering the three indexes comprehensively.

Here we choose eps=15, min_sample=7 as the final clustering parameter

Task 4.2 - DBSCAN

Here is the pair plot of the data by all 5 components.

Different clusters are labeled in different colors, with -1 as the label for outliers.

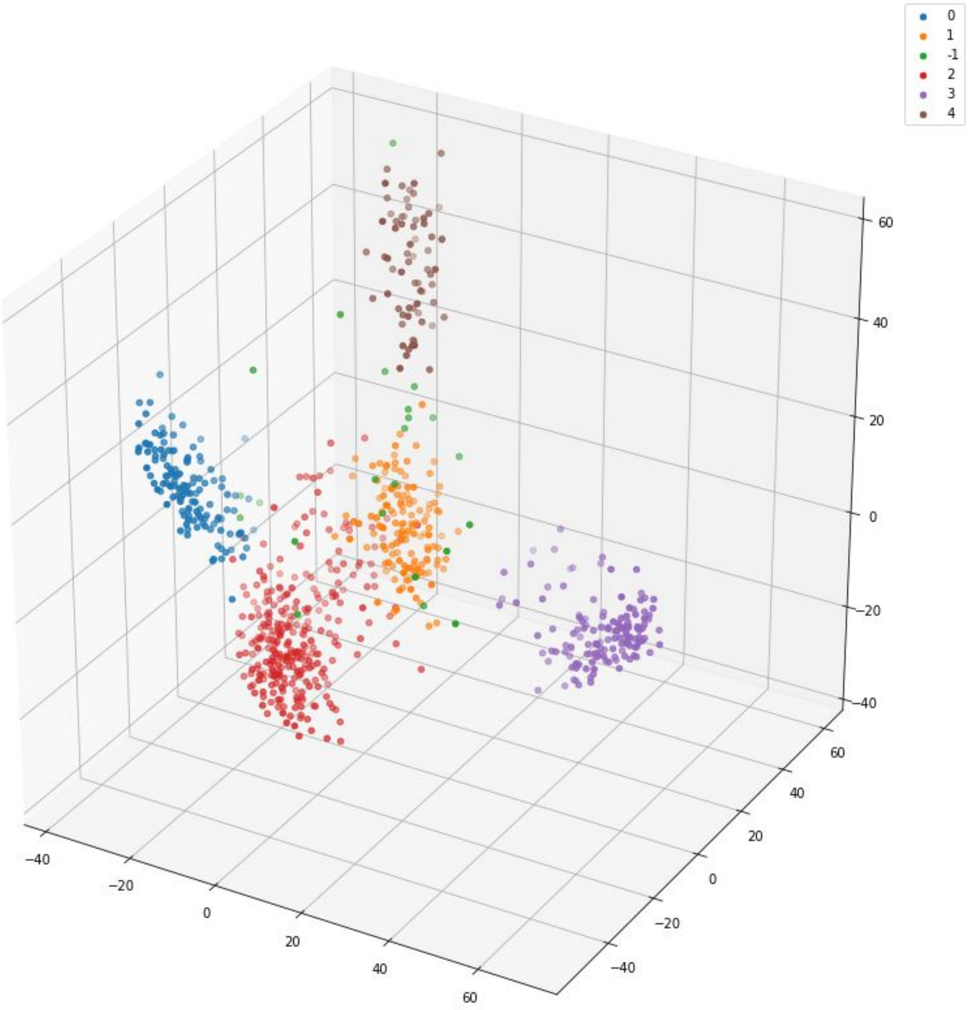


Task 4.2 - DBSCAN

Here is the 3D plot of the data by the first three components.

Different clusters are labeled in different colors, with -1 as the label for outliers.

From the plot we can see the five clusters.

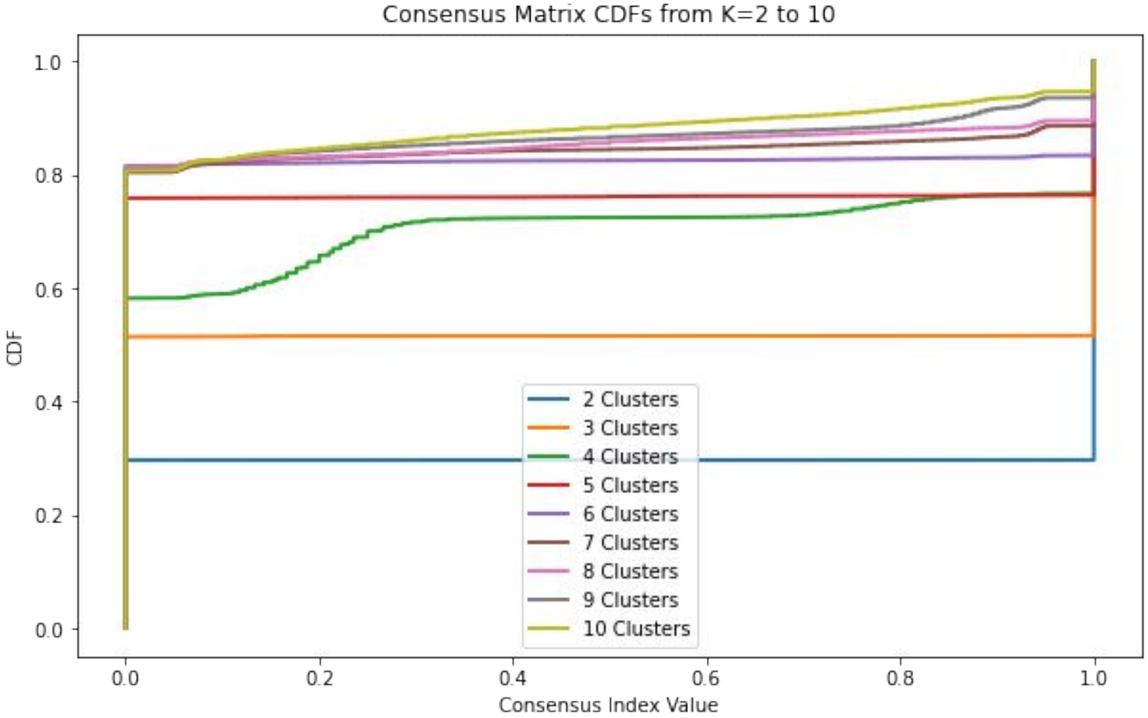


Task 5.1 - eCDF

We compute the Consensus Matrix with:

- **M = 25**
- **K = 2 to 10**

We calculate the eCDF for each matrix, and made the plot on the left.

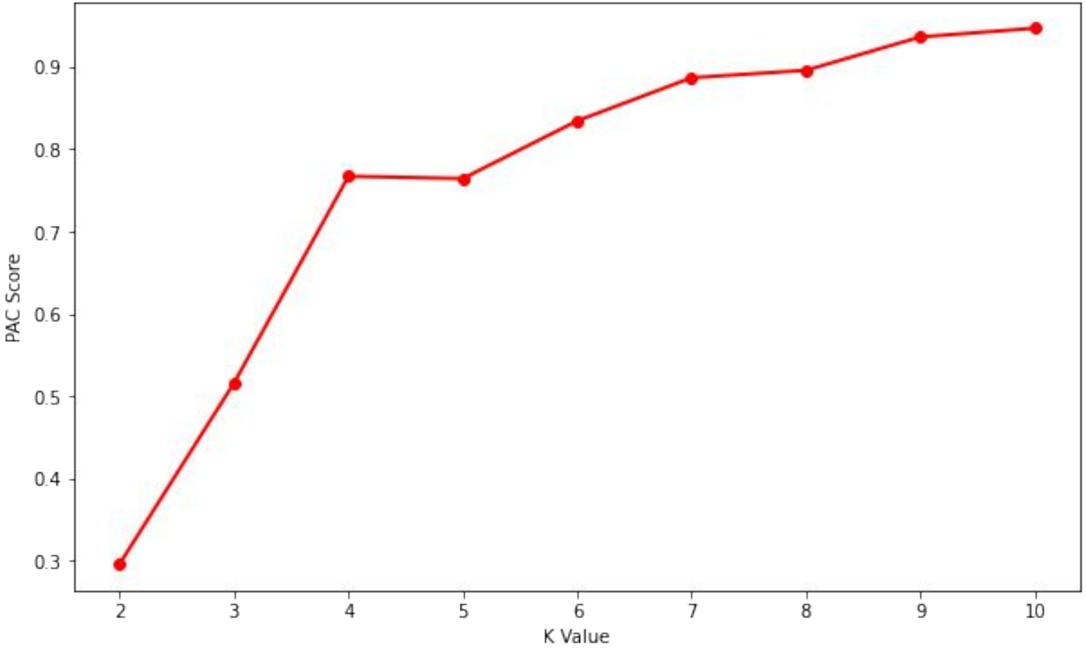


Task 5.2 - PAC Score

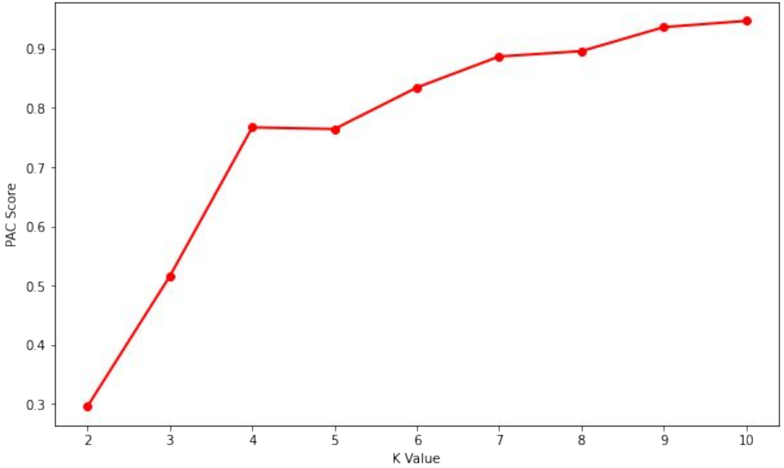
We compute the PAC score for each K based on eCDF with threshold:

- q1=0.01
- q2=0.99

We made the plot on the left.



Task 5.3 - Cluster Count Selection



- Since PAC score shows the instability, so we would like it the lower the better.
- From the plot above we see that under K-means method, PAC score grows with the number of clusters.
- It make sense that when K is small like 2 we got a nice PAC score since the model from is really sample, which means more stable, that's why the plot grows fast in the beginning. Due to that we would like focusing more on the slowly rising part of the plot.
- If selecting K value only depends on PAC score, we would like select 4 or 5, since they are the lower than others when at the slowly rising part of the plot.
- This selection result that we may pick K=4 or 5 agrees with our results from Task 4

Thanks for Reading !