Course Name: Statistical learning for big data
Course Code: MVE440
Project 1

Devosmita Chatterjee
Personnummer: 910812-7748

June 11, 2021

Contents

Question 1: Clustering

Firstly, the required libraries are imported. Then, the input data csv file 'Q1_X.csv' is loaded and imported it into dataframe.

1 Methods and Results

1.1 Exploratory Data Analysis (EDA)

An exploratory data analysis is performed with the given data in order to understand the dataset by summarizing their main characteristics, either statistically or visually. Data size, data type, missing data, duplicate data are listed in table 1.

Table 1: *The table presents the exploratory data analysis.*

Exploratory Data Analysis							
Data	**Row size n**	**Column size p**	**Data type**	**Missing values**	**Duplicate rows**	**Duplicate columns**	**Constant columns**
Q1_X.csv	560	974	Numeric	0	0	0	98

In the above result, 98 constant columns/features are dropped for building machine learning models. The reason is that the constant features won't affect the predictor that is, it do not add any variation in the data and thus, removing them also saves computational time.

From exploratory data analysis, it is found that the number of features is large relative to the number of observations ($p > n$) in the dataset, known as the "Curse of Dimensionality", which causes certain machine learning algorithms struggle to train effective models. This motivates to perform dimensionality reduction of the data.

1.2 Principal Component Analysis (PCA) for dimensionality reduction

Principal Component Analysis (PCA) is chosen as the desired dimensionality reduction technique. PCA projects each data point onto only the first few principal components to obtain lower dimensional data, while preserving maximum variance of the data. Simply, PCA transforms a large set of features into a smaller one, that still retains most of the information of all the features.

1.2.1 Data Standardization

Data standardization of a feature means to scale the observations of the feature with mean 0 and standard deviation 1 given by the formula

$$Z = \frac{X - mean(X)}{sd(X)}$$

Before applying PCA, the features are standardized because it will give more importance to the features having higher variances than the features with low variances, while identifying the right principal components. Therefore, the given data is standardized around mean 0 and with standard deviation 1.

1.2.2 Performing PCA on the standardized data

Now, PCA is performed on the standardized data with number of components = 10. Then, the individual explained variances and the cumulative explained variances are calculated and plotted.

Figure 1 shows the percentage of variance (y-axis) captured depending on the number of principal components (x-axis). The graph shows that 10 principal components are able to explain around 80% of the variance in the dataset.

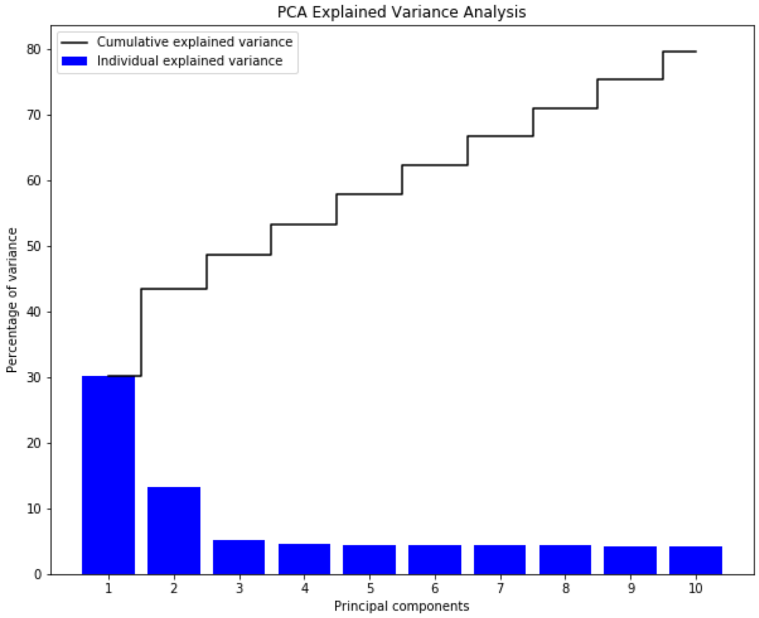

Figure 1: *The figure shows the individual explained variances and the cumulative explained variances.*

According to the previous graph, 3 principal components is a reasonable choice which explains around 50% of the variance in the dataset. Then, the pairwise comparison of the 3 principal components is performed visually. Figure 2 shows that the pairwise plot of principal components is used to explore the data.

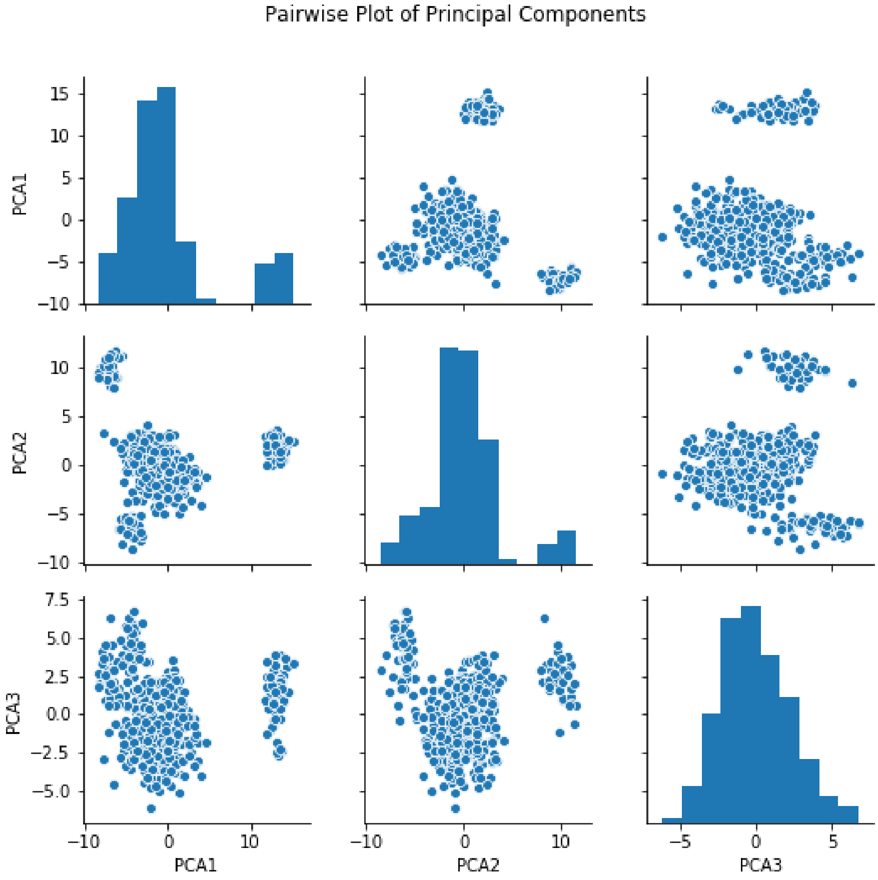

Figure 2: *The figure shows the pairwise plot of principal components.*

1.3 Clustering using k-means based on principal components

1.3.1 Finding an optimal k using Elbow method

For k-means clustering, Elbow method is often used to find the optimal number of clusters in the dataset.

In Elbow method, to choose an optimal k for k-means clustering, the Within Cluster Sum of Squares (sum of the squared distance between each data point in the cluster and the cluster centroid) is plotted for different number of clusters. The plot shows the trend that the error decreases as the number of clusters increases. The optimal number of clusters is selected where there is a sharp and steep fall of the distance (or, an elbow occurs).

The relationship between the number of clusters and the Within Cluster Sum of Squares is shown in figure 3. The graph shows that there is a sharp fall of distance at k = 4. The optimal number of clusters using Elbow method is 4. Still, there is a confusion to pick the best value of k.

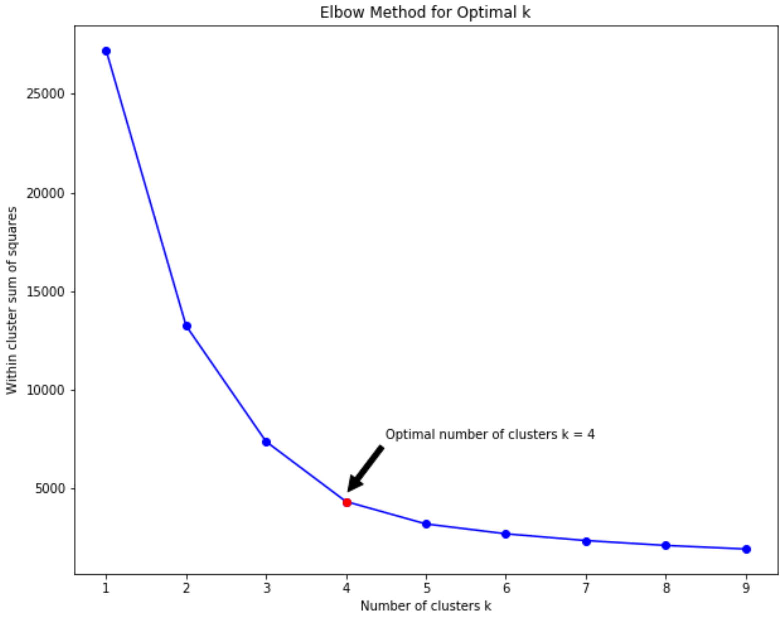

Figure 3: *The figure shows the optimal number of clusters using Elbow method.*

For higher dimensional data, it is better to employ the Silhouette method than the Elbow method to find an optimal k.

1.3.2 Finding an optimal k using Silhouette method

The silhouette plot displays a measure of how close each data point in one cluster is to data points in the neighboring clusters. The range of the silhouette coefficient is between [-1,1]. The Silhouette coefficient of +1 indicates that the data point is far away from the neighboring clusters. The Silhouette coefficient of 0 indicates that the data point is on the decision boundary between two neighboring clusters. The Silhouette coefficient < 0 indicates overlapping clusters. Figure 4 shows that the optimal number of clusters using silhouette coefficient method is 3, or 4.

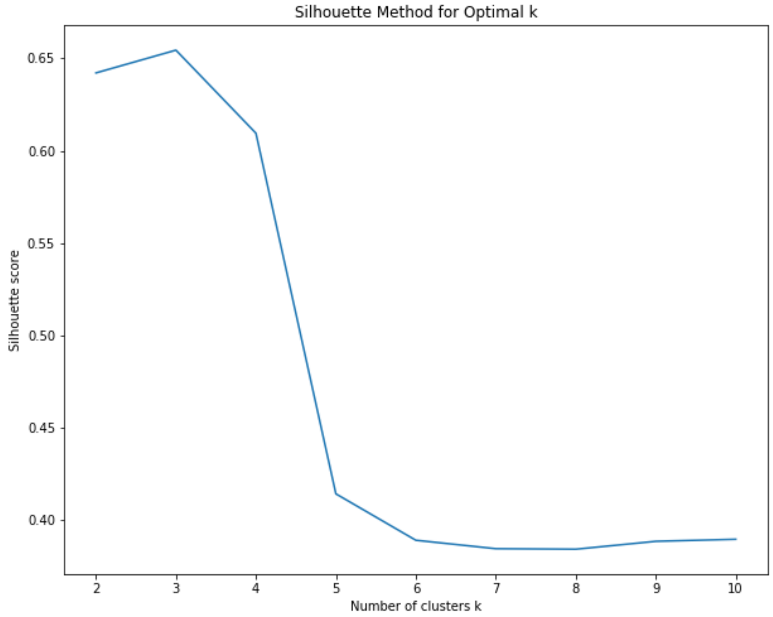

Figure 4: *The figure shows the optimal number of clusters using Silhouette method.*

1.3.3 K-means clustering on principal components

Figure 5 shows that K-means clustering is performed on principal components with 4 clusters.

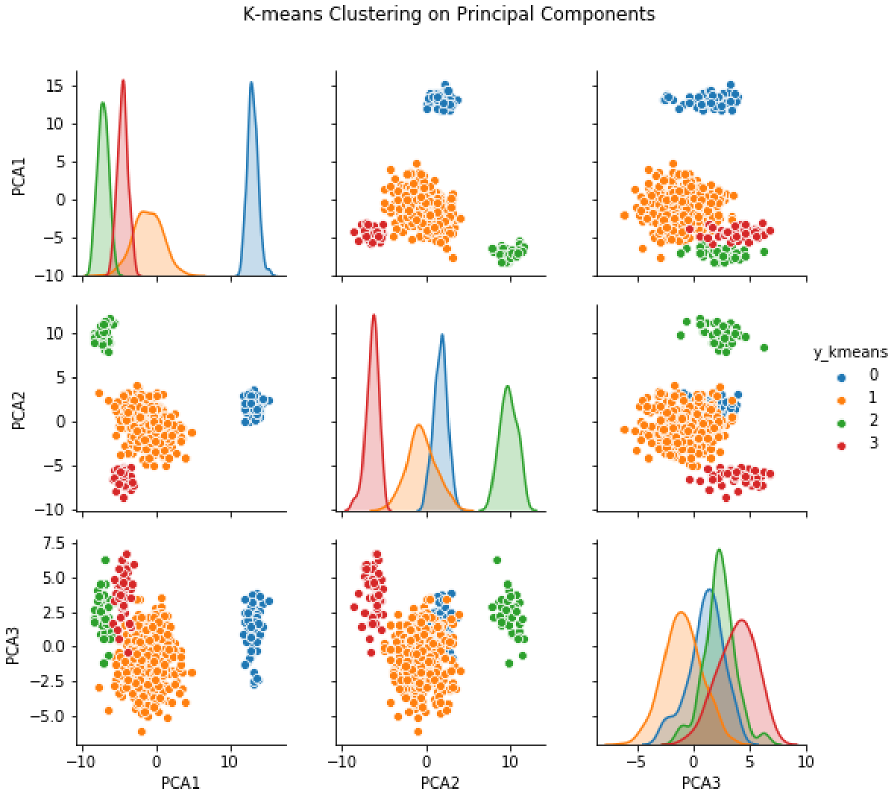

Figure 5: *The figure shows that K-means clustering is performed on principal components with 4 clusters.*

1.4 Clustering using Gaussian Mixture Models (GMM) based on principal components

1.4.1 Minimizing the Bayesian Information Criterion (BIC) for optimal number of clusters k

Figure 6 shows that the optimal number of clusters by minimizing the BIC is 4.

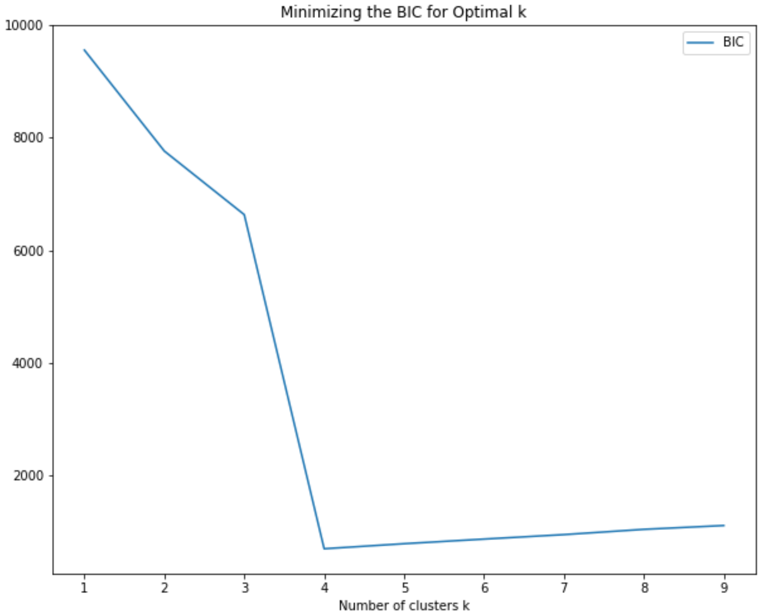

Figure 6: *The figure shows that the Bayesian Information Criterion (BIC) is minimized to obtain the optimal number of clusters k.*

1.4.2 Minimizing the Akaike Information Criterion (AIC) for optimal number of clusters k

Figure 7 shows that the optimal number of clusters by minimizing the AIC is 4.

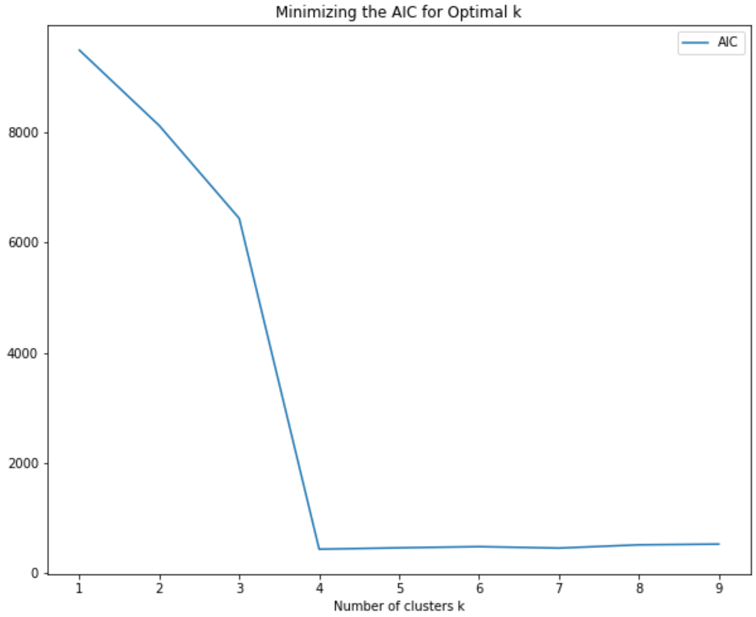

Figure 7: *The figure shows that the Akaike Information Criterion (AIC) is minimized to obtain the optimal number of clusters k.*

1.4.3 GMM clustering on principal components

Figure 8 shows that GMM clustering is performed on principal components with 4 clusters.

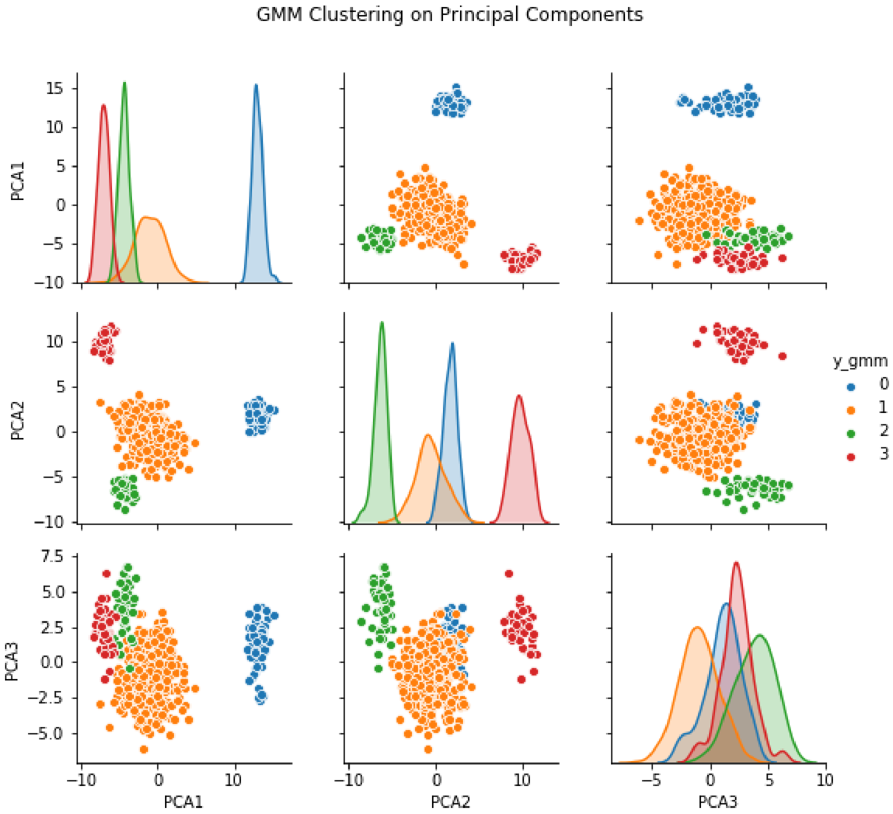

Figure 8: *The figure shows that GMM clustering is performed on principal components with 4 clusters.*

1.5 Clustering using Density Based Spatial Clustering of Applications with Noise (DB-SCAN) based on principal components

1.5.1 Finding the parameter epsilon by plotting K-distance graph

Figure 9 shows the k distance graph. eps is chosen by using a k-distance graph where k = minPts-1 nearest neighbor.

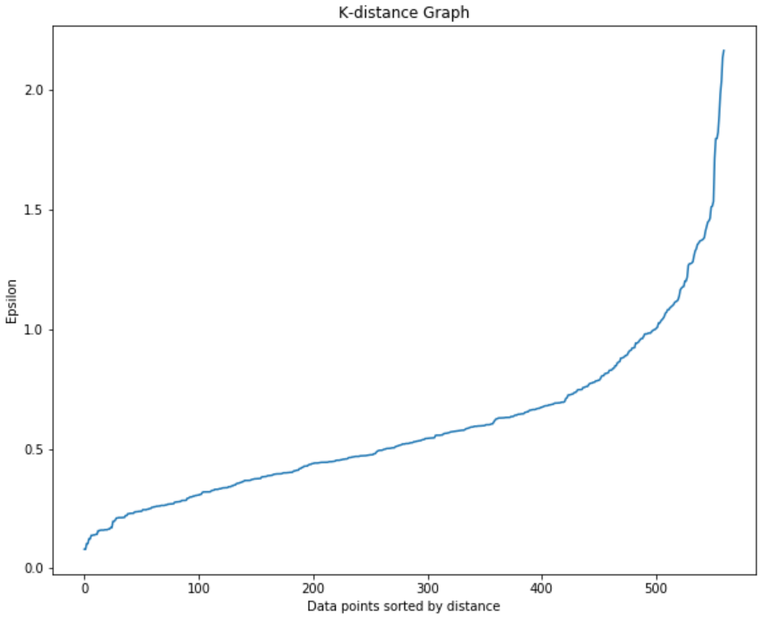

Figure 9: *The figure shows the k-distance graph.*

1.5.2 DBCAN clustering on principal components

Figure 10 shows DBSCAN clustering on principal components. minPts = 2*D where D is the number of dimensions in the data set.

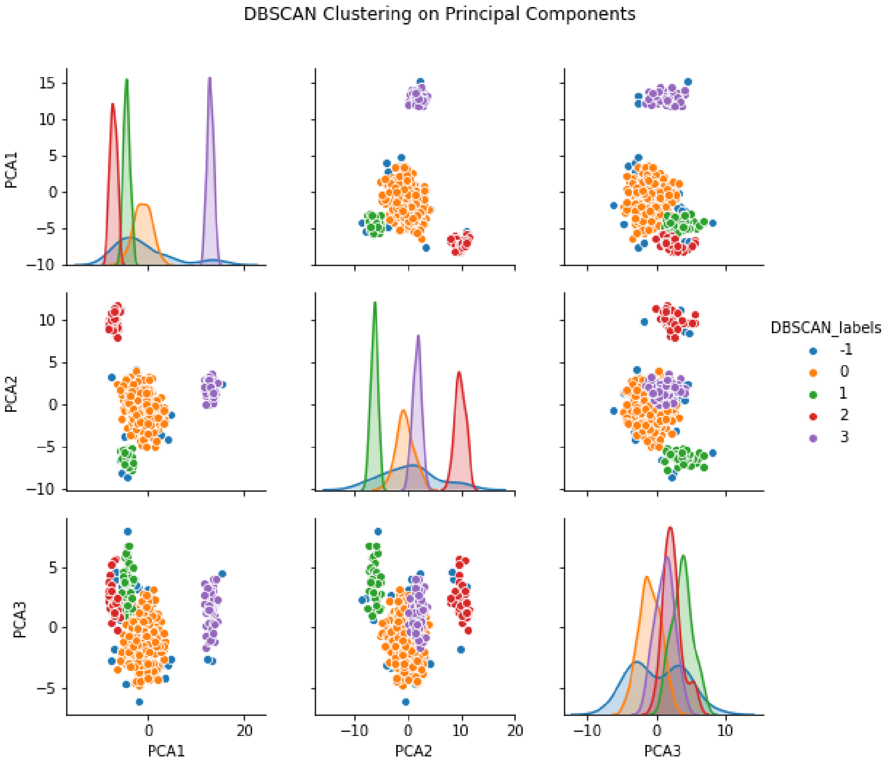

Figure 10: *The figure shows DBSCAN clustering on principal components.*

2 Discussion

In the section, I discuss about the following results:-

1. From exploratory data analysis of the given dataset, it is found that there are constant columns. A machine learning model is given by the mathematical equation Y = f(X) in which Y = dependent variable and X = independent variables. Thus, machine learning models estimates values of Y given values of X. When a whole column of X is constant, then the relationship between Y and X is meaningless because it doesn't add any variation in the data. Moreover, it increases computational time. Hence, the constant columns are removed from the data.

2. PCA is a powerful dimensionality reduction method. PCA reduces the dimension of the data from $p = 974$ to $p = 3$ while preserving maximum variance of the data.

3. K-means is a distance based clustering method. Thus, it can handle well separated clusters. It is a simple algorithm, which works well when the clusters are approximately round, but it can not deal with noisy, elongated, or partially overlapping clusters.

4. GMM is a probabilistic approach to clustering which works similarly like k-means, but it assumes input variables are a mix of Gaussian distributions. As a result, it can deal much better with elongated, and overlapping clusters, but not noisy clusters.

5. DBSCAN, is a density based clustering algorithm. It works well for data which contains clusters of a similar density. Since DBSCAN groups together data points which are densely packed together, it can discover well-defined clusters even from very irregular-shaped distributions, noise (like most real life data). But, DBSCAN is more computationally intensive and requires tuning of parameters.

6. From the figures, 4 clusters seams to be reasonable.