Assignment 2

20CS30039 - Rahul Mandal
20CS10051 - Sailada Vishnu Vardhan

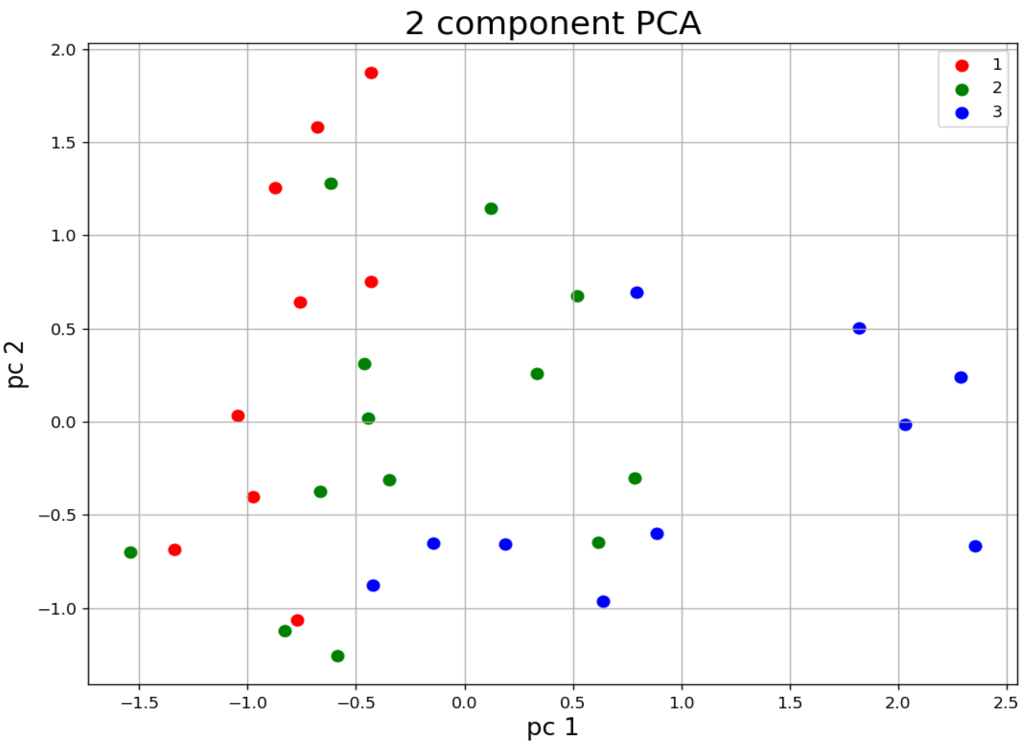
Question 1:

The given data has some missing values with ‘?’ in it, these are taken care of by filling them with the mean of that column.

Principal Components Analysis (PCA):

Principal Component Analysis is an unsupervised learning algorithm, It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. Principal Components are the newly transformed features.

To achieve co variance more than 95%, we have to select a number of components at least 21.
But for visualizing the graph we have taken the number of components as 2.
Below is the graph of the plot and pc1, pc2 are the two principal components. 1,2,3 are the outputs.



The calculated Co-Variance for above plot is = 0.28086641456381367
But the taken value by considering n_components = 21 is 0.9589751623955571

K-means Clustering:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process.

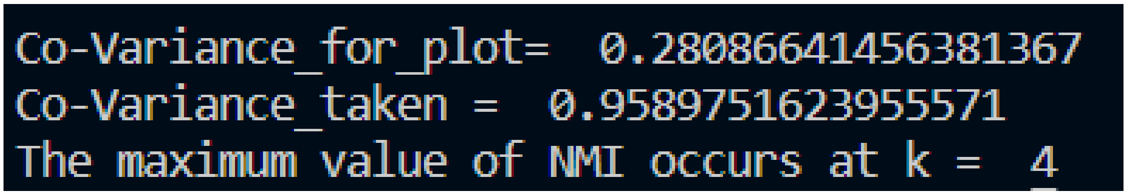
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

Determines the best value for K center points or centroids by an iterative process.Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

The functions mainly used here are `cluster_recal(X, Y_pred, centroids, k)` **`and`** `centroids_recal(centroids, clusters, k),` we call these functions to recalculate centroid and clusters until there is no change in the centroids positions.

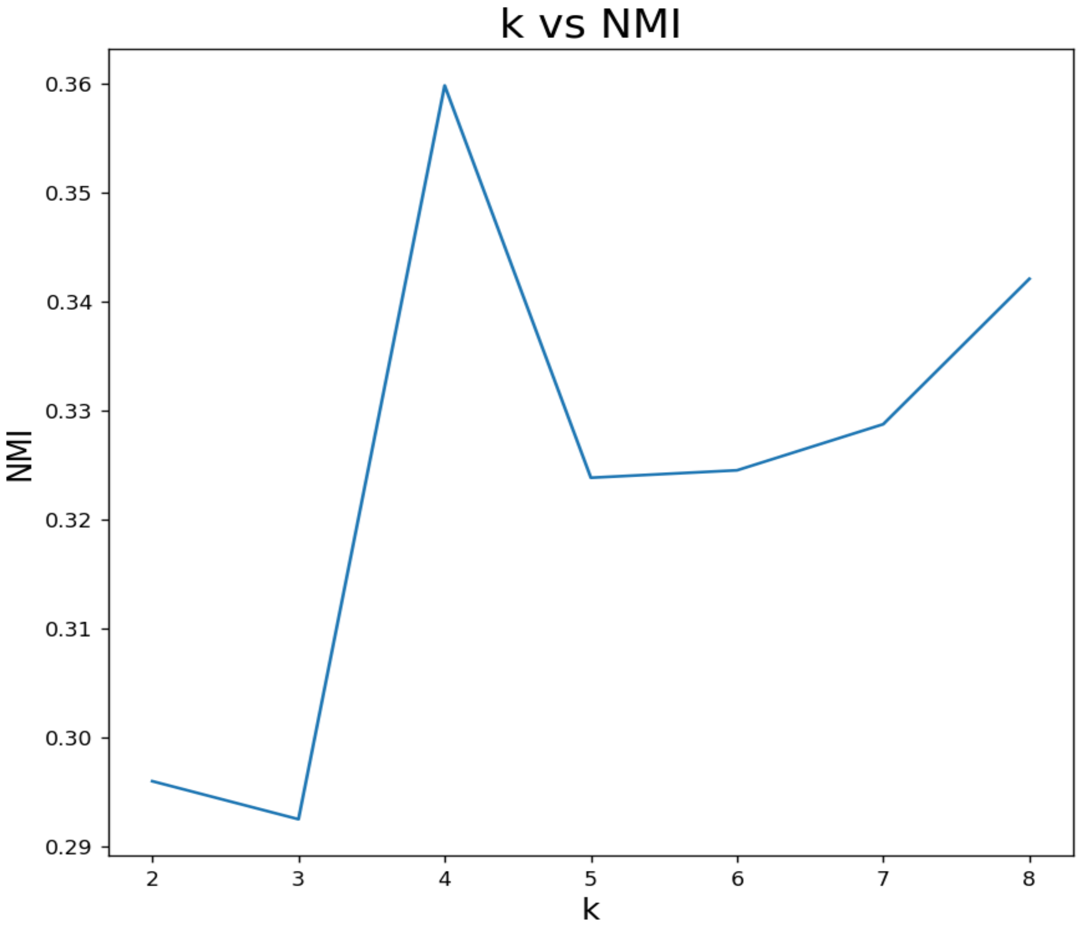
For determining the quality of clustering Normalized mutual information (NMI) is used, Since it's normalized we can measure and compare the NMI between different clusterings having different number of clusters.

The final results printing on the terminal are as below



```
Co-Variance_for_plot=  0.28086641456381367
Co-Variance_taken =  0.9589751623955571
The maximum value of NMI occurs at k =  4
```

Below is the plot of K vs NMI


k vs NMI
NMI
k