Clustering

DMPR: Lab activities for week 06

Name: Arnab Dey ID: U20210014

5) Do a single kmeans clustering

Q) What does the **inertia_** field represent?
Ans: Sum of squared distances of samples to their closest cluster center, weighted by the sample weights if provided. (taken from scikit-learn documentation).

Q) What are the centroids of the cluster?
Ans: The center of a cluster. The mean of each variable for the observations in that cluster is represented by one integer in a vector called a centroid. The multi-dimensional average of the cluster can be compared to the center. The centroids of the three clusters are:

[[-0.5083544 0.7004946 -0.2975992 0.614481 0.78115976 -0.23786062
-0.92873347 0.26173645 -0.76228786 -1.3500991 -1.314714]
[0.8674636 -0.38443145 0.2244651 0.06701774 -0.19345693 0.5969275
0.4578177 0.46987382 0.02682476 0.48141655 0.43058616]
[-0.5116155 -0.10591478 -0.0161457 -0.4971544 -0.35335493 -0.43042508
0.19229573 -0.65308934 0.50677675 0.46365282 0.48971358]]

Q) How do you map the labels back to the original data?
Ans: Use the **labels_** attribute of **KMeans** object and add a new column containing this info. The respective labels are:

[2 2 0 0 1 1 2 2 2 0 1 1 1 2 2 2 2 2 0 0 2 2 2 0 0 1 0]

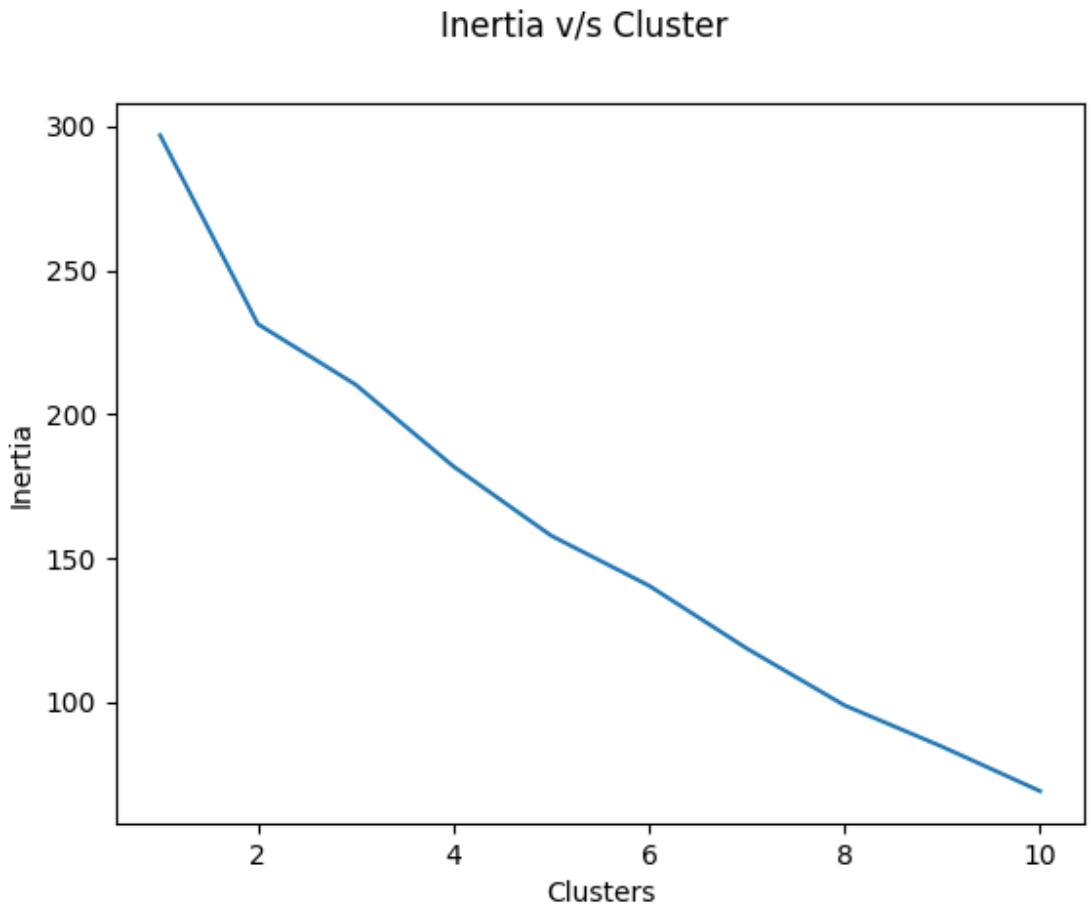
Q) Include a new column **cluster** in the original data frame **df** with the cluster label of each row and print the data frame.
Ans: Please refer to the code file **W06_P2.py**.

Q) What do you think the data points were clustered by? In other words what are the commonalities in the objects in the same cluster?
Ans: Similarities in the feature vectors.

6) Do a series of k-means clustering with k=1 to 10

Please refer to the code file.

7) Now plot the inertia against the number of clusters

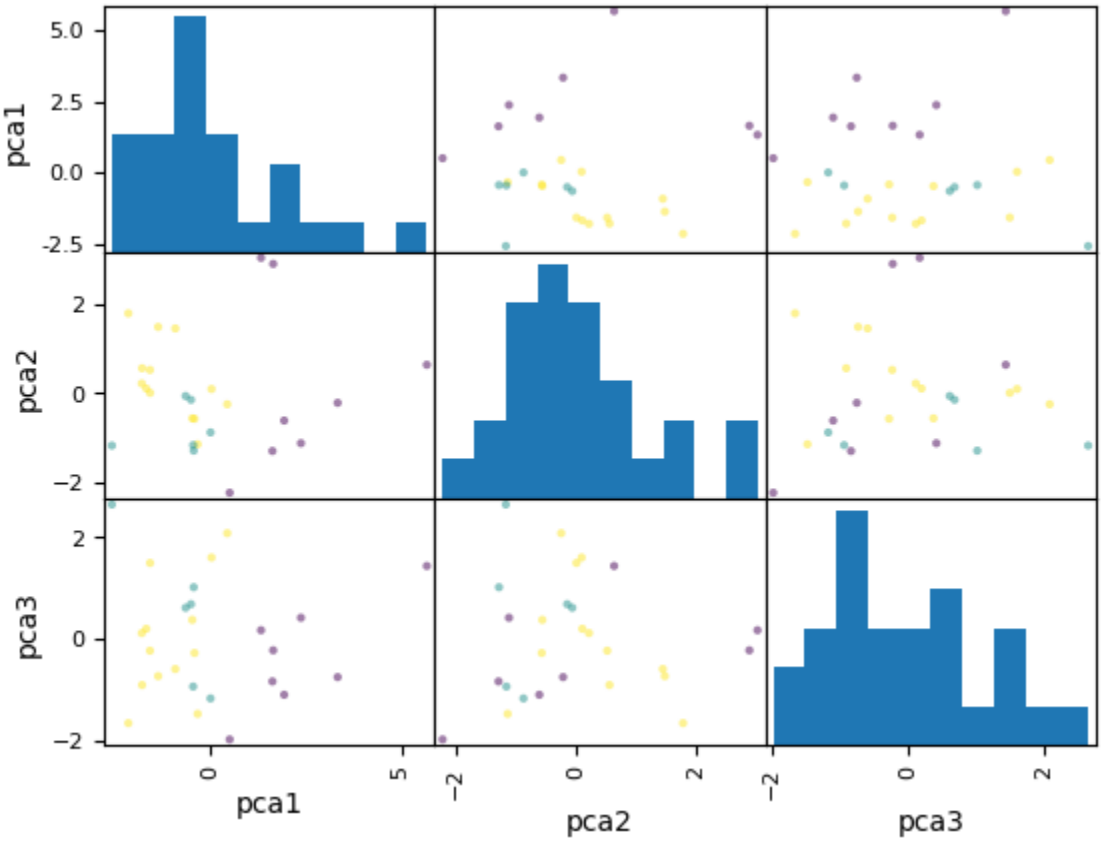


8) Use the KneeLocator class from the kneed module to find the elbow

Q) What is the number of clusters at the knee location detected by kneed? Do you agree based on your earlier plot? If not what value would you suggest?
Ans: 4. Yes, I agree.

9) Now carry out PCA to project the data into 3 dimensions and create a scatter matrix of the pairwise components

Please refer to the code file **W06_P2.py**



10) Indicate the cluster of each point by overlaying a marker with a different color