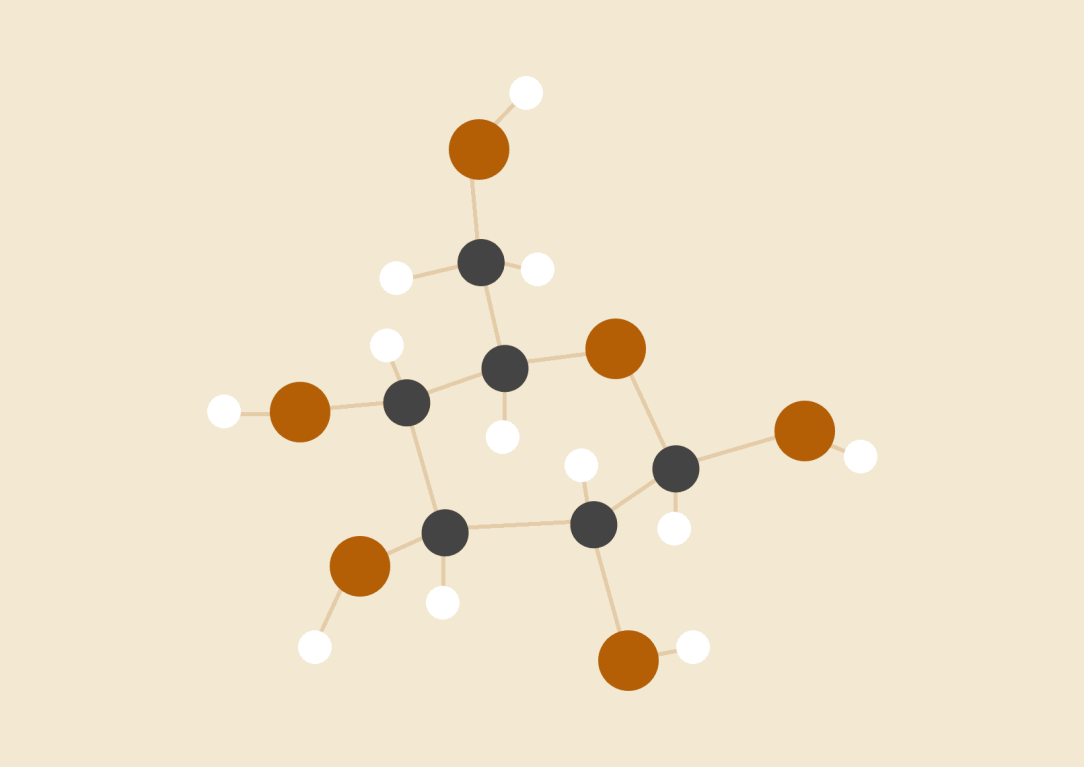
DS5004-Machine Learning for Data Science

Assignment # 01

Clustering



Name: Muhammad Usman

Roll no: 21P-8743

21.03.2022

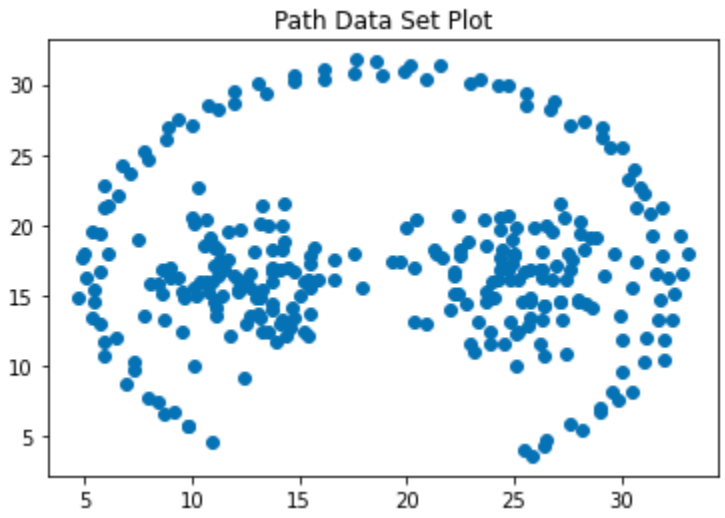
Clustering

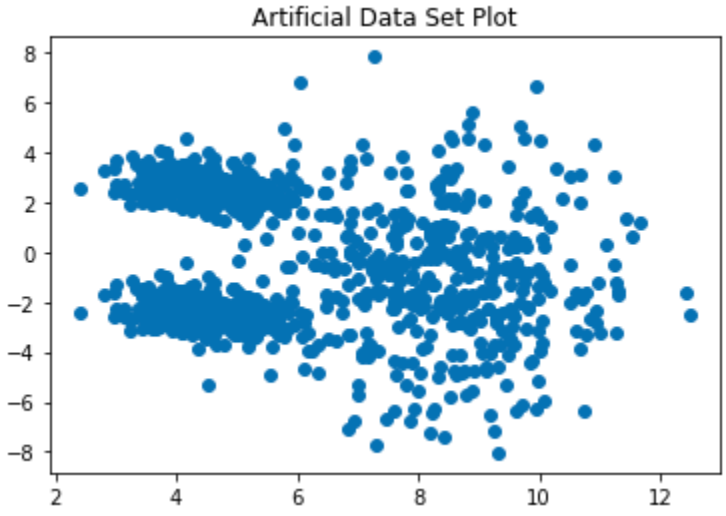
I have implemented 4 clustering algorithms (K Means, K Mode, K Medoid) using built -in library (sklearn) and one algorithm without any built in library functions as asked on the provided Data Sets (Path.csv,Spiral.csv,Artificial.csv).

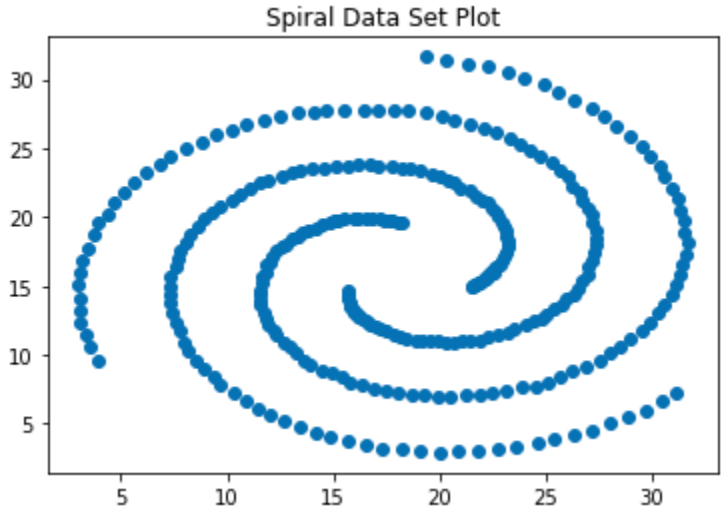
To find the optimal value for k (number of clusters) I use the Elbow Method.

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k especially for K Means.

Data Sets Visualization



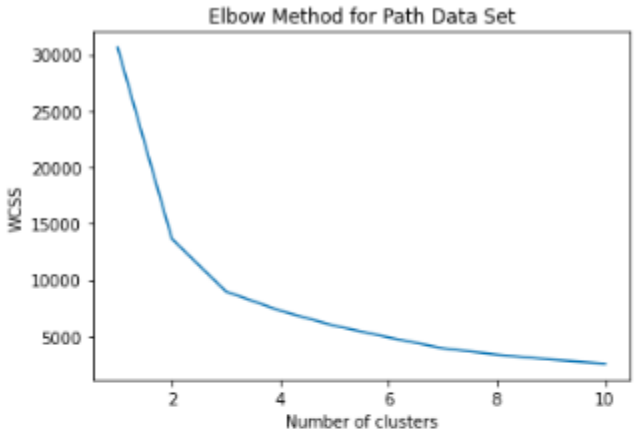

Artificial Data Set Plot
8
6
4
2
0
−2
−4
−6
−8
2
4
6
8
10
12

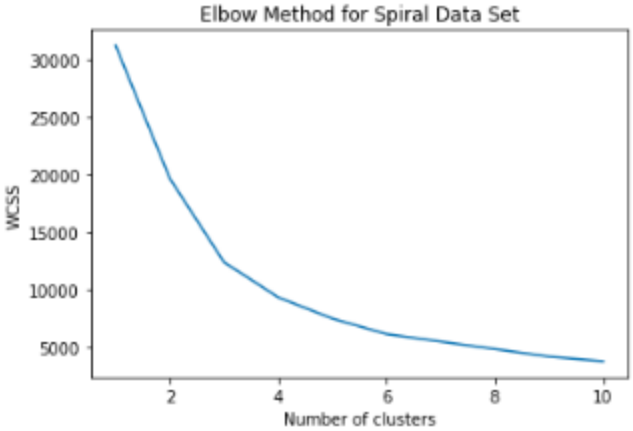

Spiral Data Set Plot
30
25
20
15
10
5
5
10
15
20
25
30

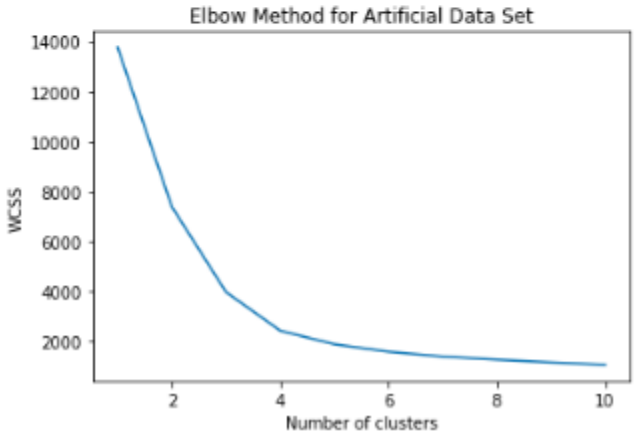
Elbow Method Visualization

The below chart visualizes the relationship between the number of clusters and within Cluster Sum of Squares (WCSS) then we select the number of clusters where the change in WCSS begins to level off.

We can see a bend at K=3 in the above graph indicating 3 is the optimal number of clusters.

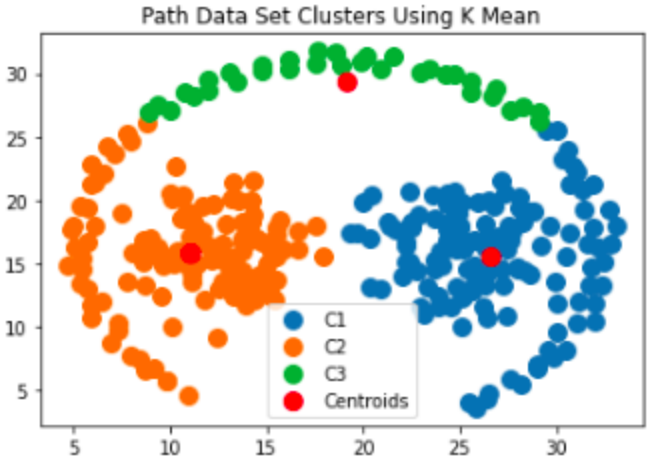


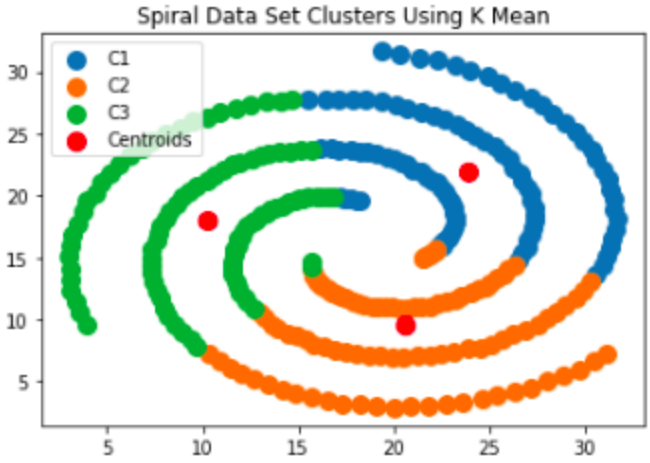


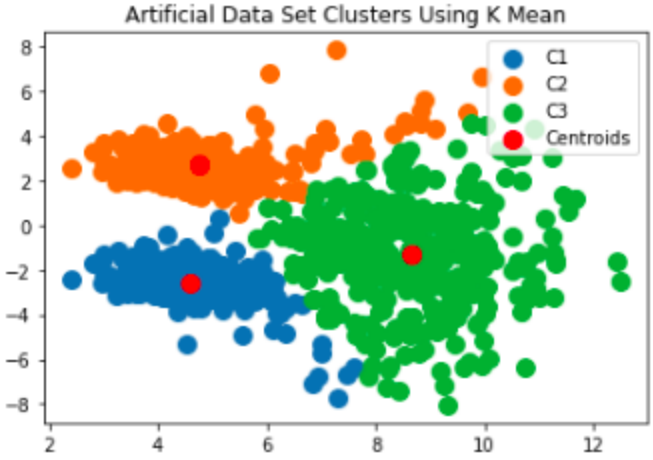


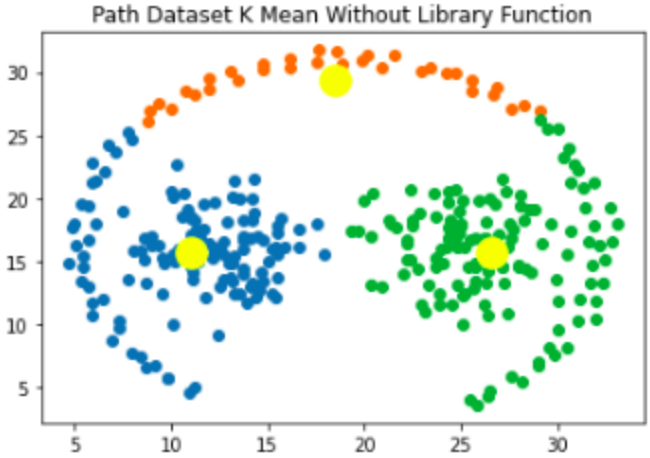
K Mean Clustering

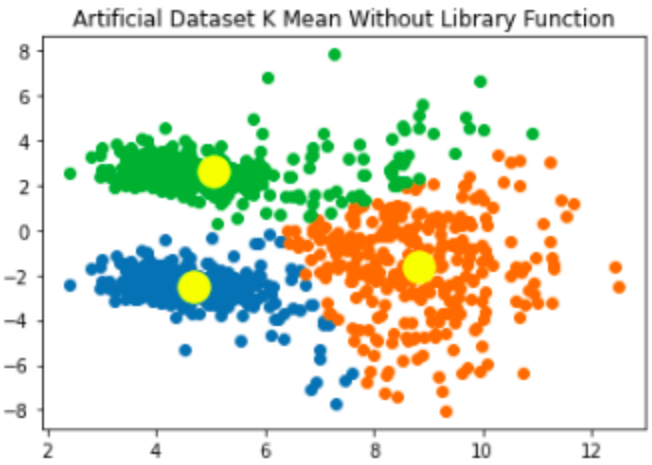
After applying K Means Clustering on the provided data sets the result is shown below.

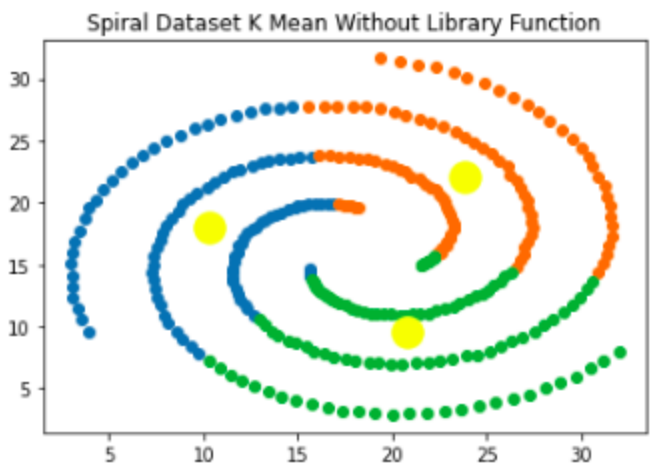






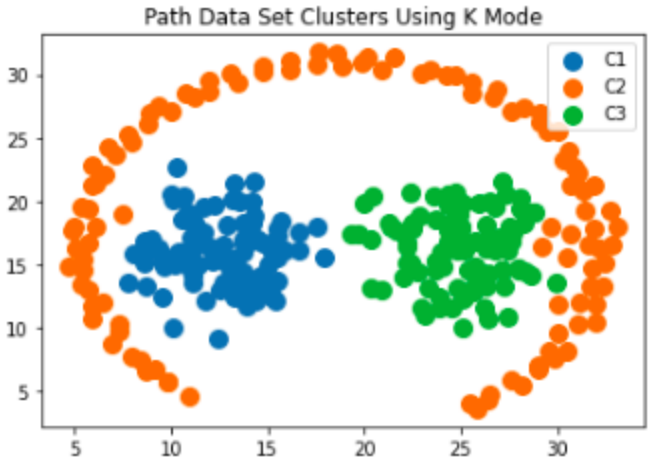


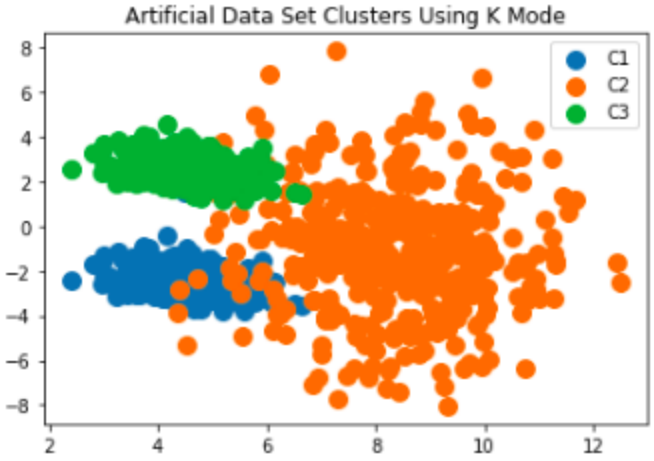


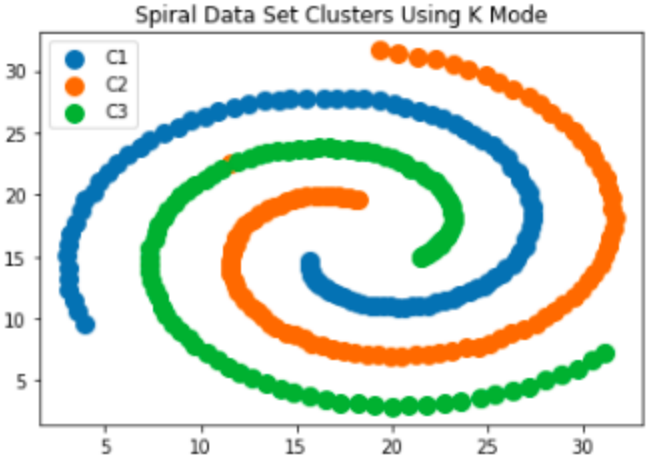


K Mode Clustering

After applying K Mode Clustering on the provided data sets the result is shown below.

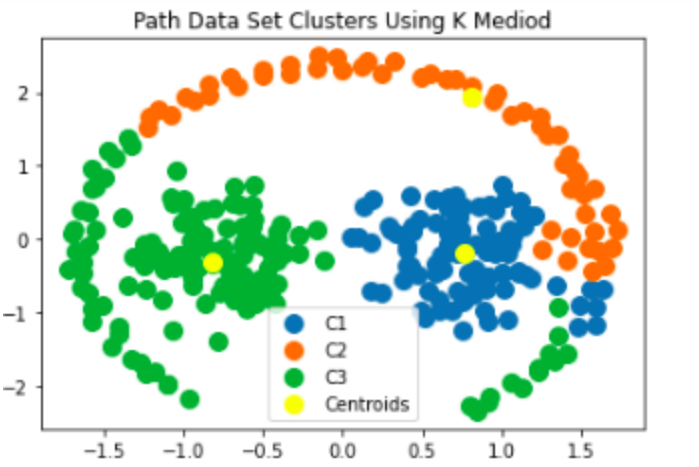


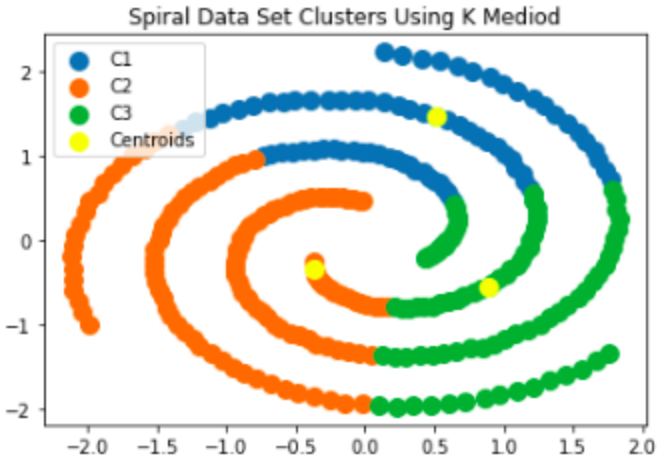


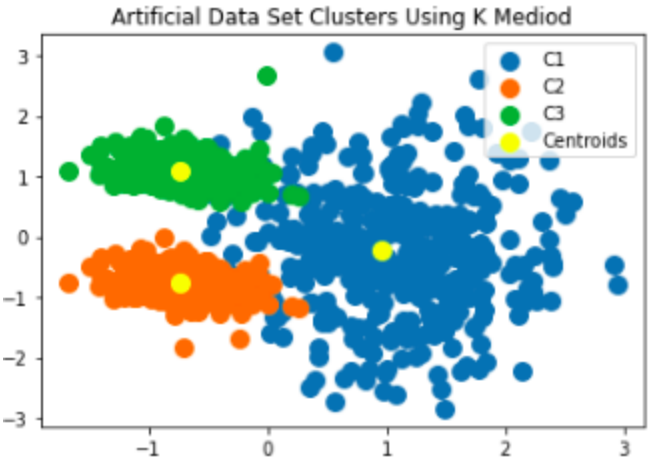


K Medoid Clustering

After applying K Medoid Clustering on the provided Data Sets result is shown below.







Clustering Evaluations:

If we would like to evaluate the clustering by just visualizing the above figures, it is evident that K Mode has performed well clustering as compared to other's techniques especially in the case of Spiral Data Set because K Mean and K Mode couldn't cluster well on irregular(spiral) based shapes.

But there are some main metrics used to evaluate the performance of Clustering algorithms, to rigorously have a set of measures.

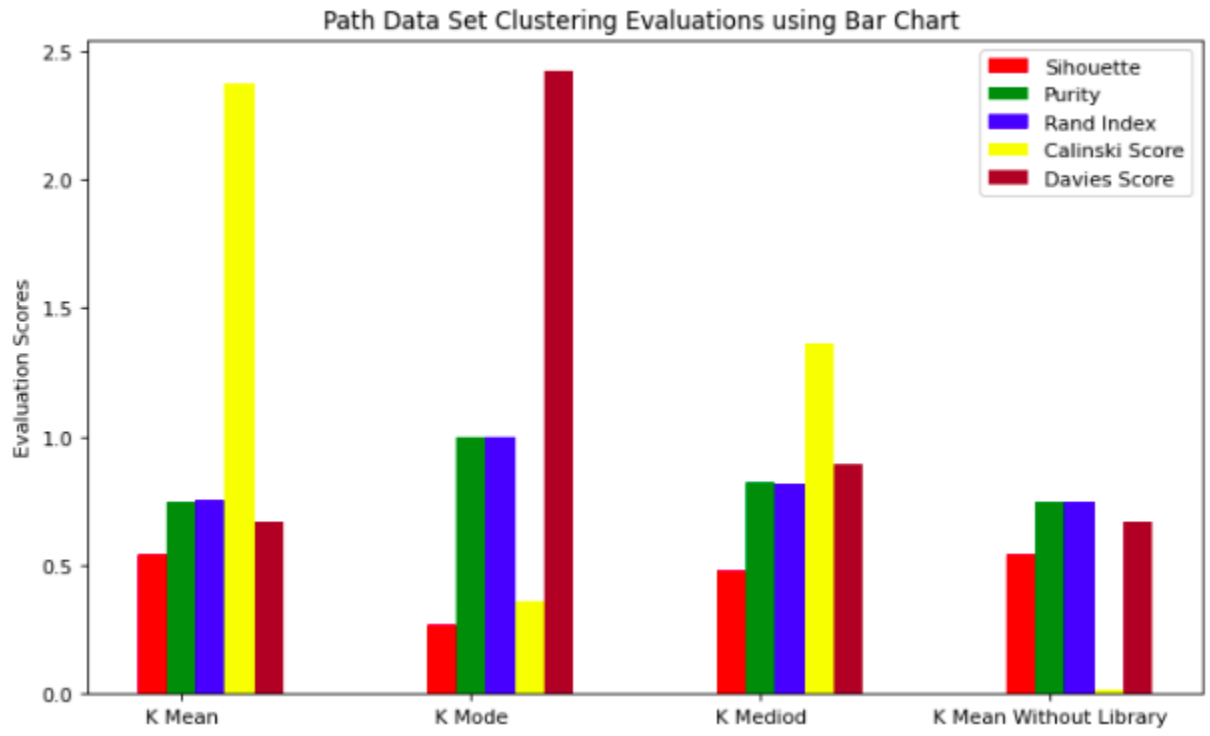
Few of them that i have used for Evaluations are:

1. **Silhouette Coefficient**
2. **Purity**
3. **Rand Index**
4. **Calinski-Harabasz Index**
5. **Davies-Bouldin Index**

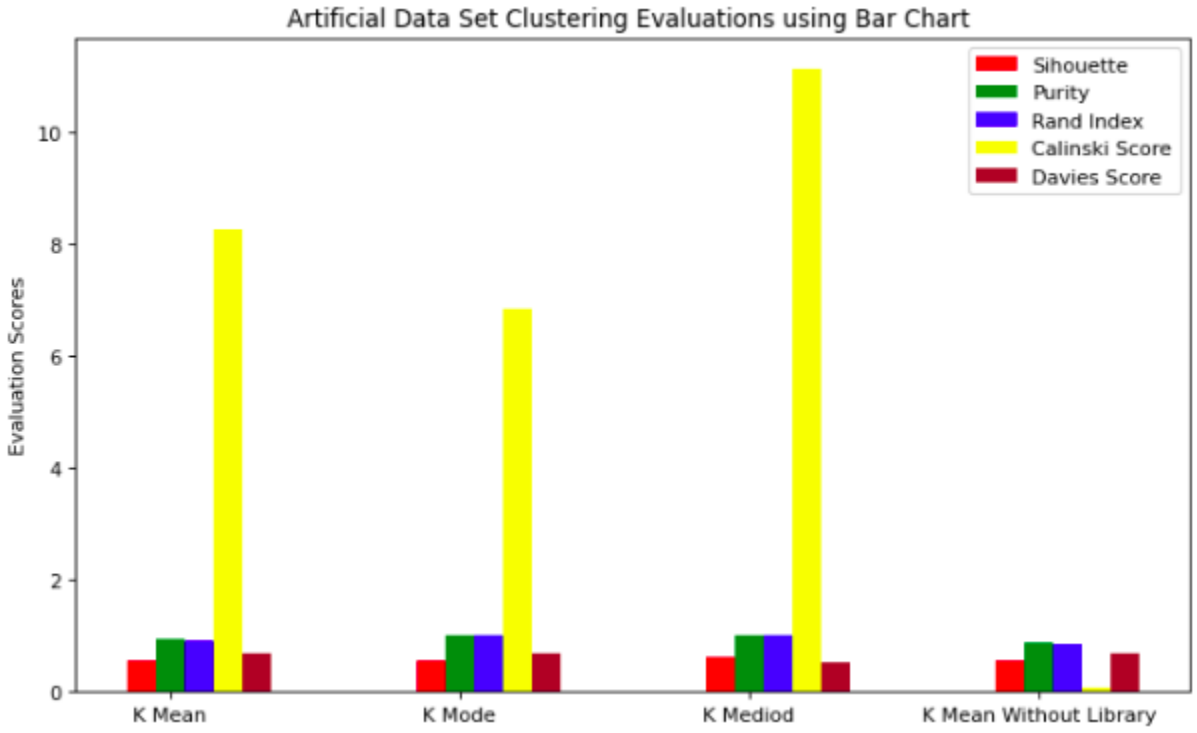
I have used K (number of clusters) is equal to 3 in all the clustering algorithms for all the provided datasets which is giving the best score.

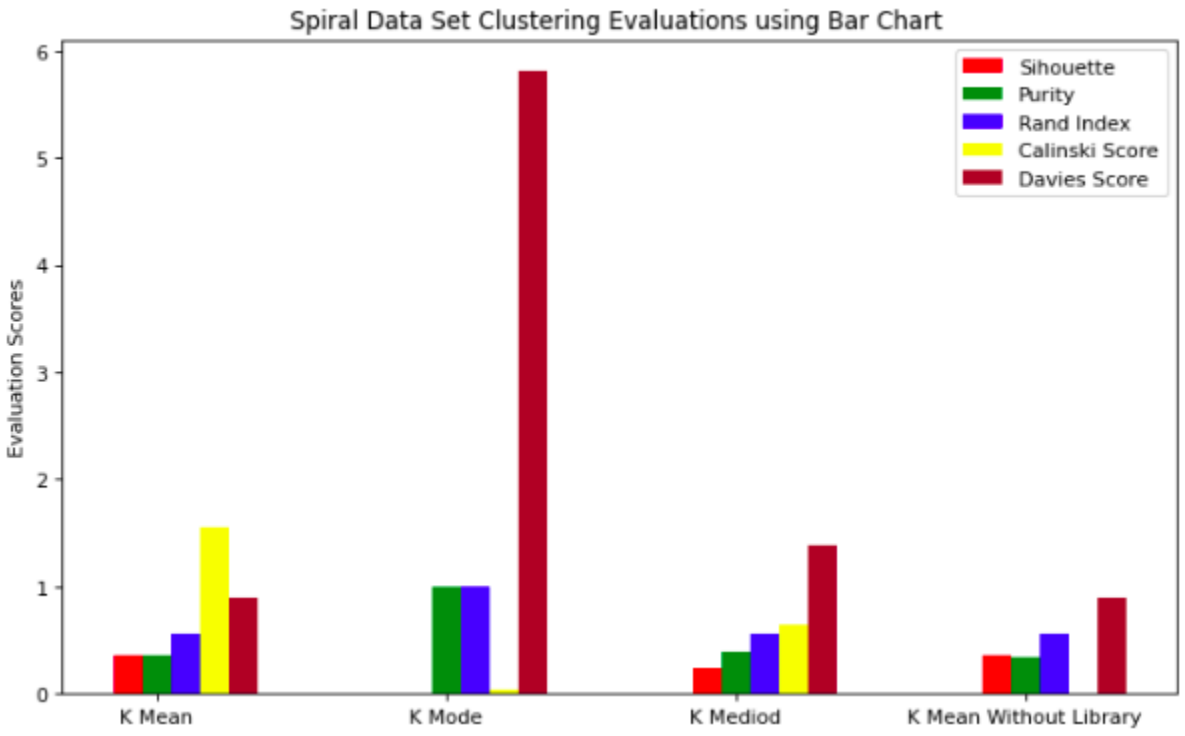
Evaluations Results

Evaluation results could be seen in the below bar charts.



If we analyze the above bar graph, k mode has a relatively good score(purity,rand index etc) as compared to other algorithms.





Evaluations Results in Tabular Format

Path Data Set

Evaluations Techniques	K Mean	K Mode	K Medoid	Clustering Without Library Functions (K Means)
Silhouette Coefficient	0.541	0.226	0.479	0.540
Purity	0.745	1.0	0.822	0.743
Rand Index	0.7497	0.960	0.812	0.749

Calinski-Harabasz Index	356.44	50.51	160.32	357.80
Davies-Bouldin Index	0.671	2.351	1.177	0.686

REFERENCES

1. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6
2. https://machinelearningmastery.com/clustering-algorithms-with-python/
3. https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203
4. https://www.kaggle.com/saptarsi/kmedoid-sg