Chiranjeev
P21CS007
chiranjeev.1@iitj.ac.in

Assignment - 5
K-means and GMM

01-11-2021

I. Dataset1 Description:

a. **Name of Dataset1 :** Wine dataset of Sklearn (link - **[Dataset link]**)
b. **About Dataset1:** The wine dataset is a classic and very easy multi-class classification dataset.

c. **Dataset1 features :**

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

d. **Data Pre-processing and Features Selection:**
Found that there are not any Not Available (NA) values in the dataset.
Removing the attribute (Name of Wine - categories) because the techniques to be used are Unsupervised Learning techniques.

II. Techniques Used:

Approach 1: PCA followed by Kmeans/GMM

A.

A.1. K-means & PCA

1. First, standardize the data using Normal standardization.

2. Then transform the data into 2-Dimensions with the help of Principal Component Analysis (PCA) technique.

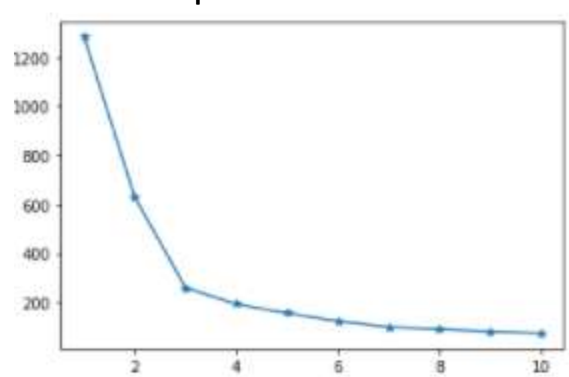
3. Now we will apply K-means technique to this 2-Dimensional transformed data.

4. Create regions for each cluster.

Analysis and Visualization:

Used Silhoutte scores and Elbow method to decide the final number of clusters to be taken for the dataset.

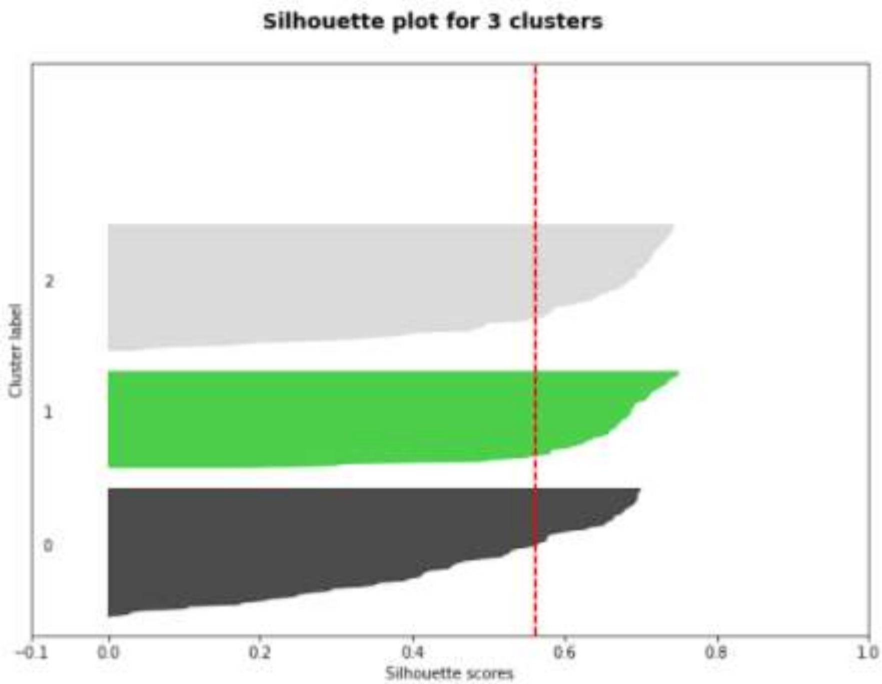
Elbow method plot:



Silhouette scores for various number of clusters:

```
num_clusters: 2  silhouette_score: 0.4649140908920152
num_clusters: 3  silhouette_score: 0.5610505693103247
num_clusters: 4  silhouette_score: 0.4914213395710318
num_clusters: 5  silhouette_score: 0.4559244619913197
num_clusters: 6  silhouette_score: 0.4483651644133675
num_clusters: 7  silhouette_score: 0.4219428974727114
num_clusters: 8  silhouette_score: 0.4105709291196553
num_clusters: 9  silhouette_score: 0.3832977328888657
num_clusters: 10  silhouette_score: 0.37840104215399895
```

Silhouette score plot for final cluster = 3:



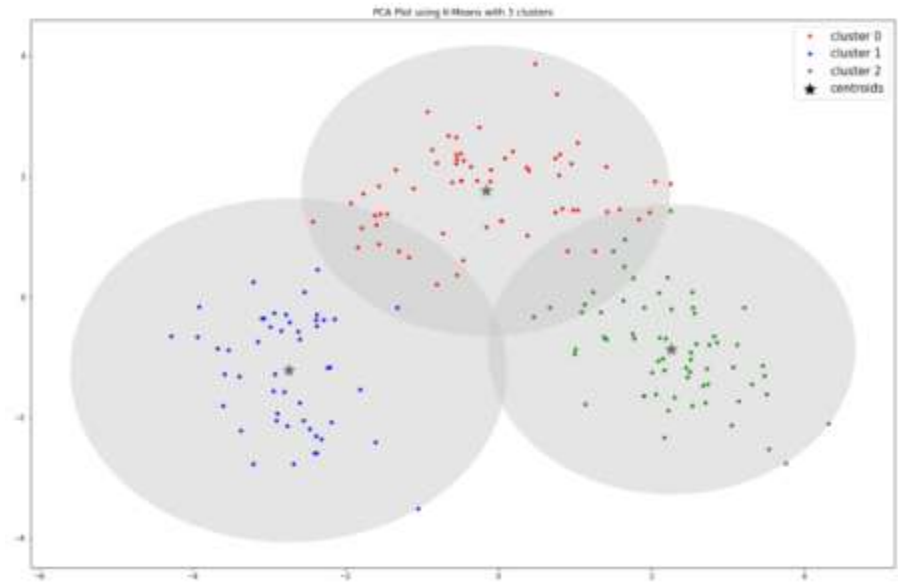
About Silhouette Coefficient:

Silhouette Coefficient is calculated using the mean intra-cluster distance "***a***" and the mean nearest-cluster distance ***"b"*** for each sample.

Silhouette Coefficient for a sample is: $\frac{(b-a)}{\max(a,b)}$

- Best value for Silhouette score is 1 and the worst value is -1
- Values near 0 indicate overlapping clusters
- Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar

Clusters Obtained from PCA followed by Kmeans



Aanalysis and Visualization:

- On Visualizing the Silhouette scores for various number of clusters in range 1 to 10 and on observing thickness of each cluster; for K=3 we found out that Silhouette score is highest (0.56105) and also, visualized that each cluster has almost similar thickness for all 3 clusters.
- On visualizing Elbow method also we found out that elbow occurs at 3 number of clusters.
- On visualizing clusters above we can see 3 clusters are nicely separated and capture the data distribution nicely.

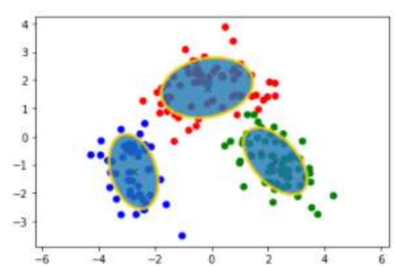
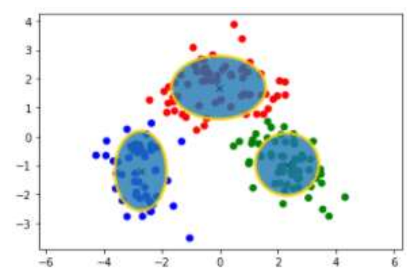
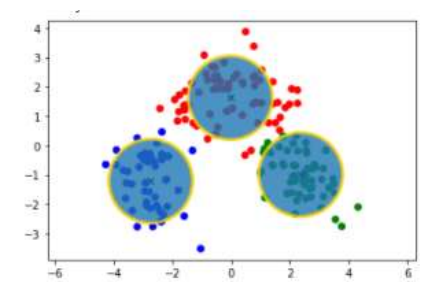
A.2. <u>GMM & PCA</u>

1. First, standardize the data using Normal standardization.

2. Then transform the data into 2-Dimensions with the help of Principal Component Analysis (PCA) technique.

3. Now we will apply GMM techniques to this 2-Dimensional transformed data by keeping final number of clusters same as we obtained from k-means and suing those cluster centroids here for initialization. Drawn ellipses and spherical regions for each cluster. Tried with various types of covariance matrices.

Aanalysis and Visualization:

Types of Covariance Matrix		
Full	**Diagonal**	**Identity**
Shapes : Ellipses inclined at various angles	Axis Aligned Ellipses	Spherical
		
As we can visualize from above 3 shapes obtained from different types of Covariance matrices, spheres generated from that Identity Covariance matrix can capture comparatively better data distribution as compared to others.		

Approach 2: Kmeans/GMM followed by PCA

B.

B.1. K-means & PCA

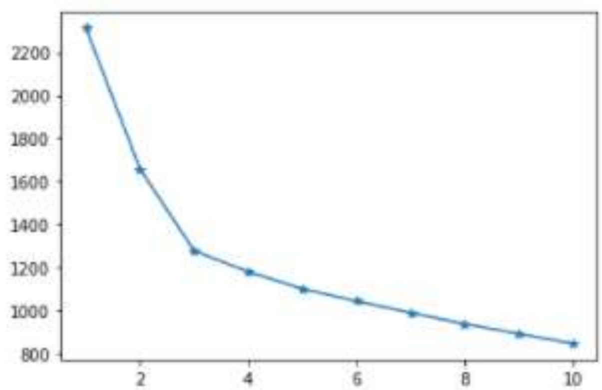
1. First, standardize the data using Normal standardization.

2. Then will use K-means and then used PCA to transform data along with clusters coloured with different colours.

Analysis and Visualization:

Used Silhoutte scores and Elbow method to decide the final number of clusters to be taken for the dataset.

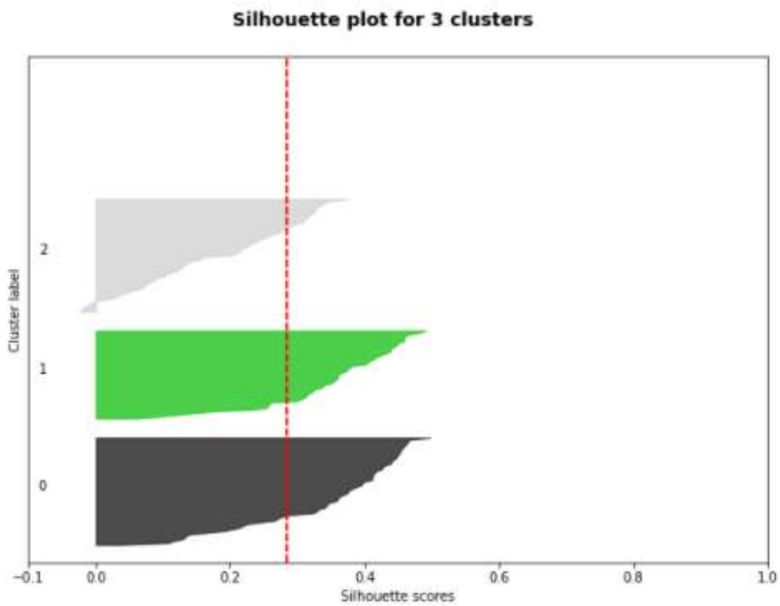
Elbow method plot:



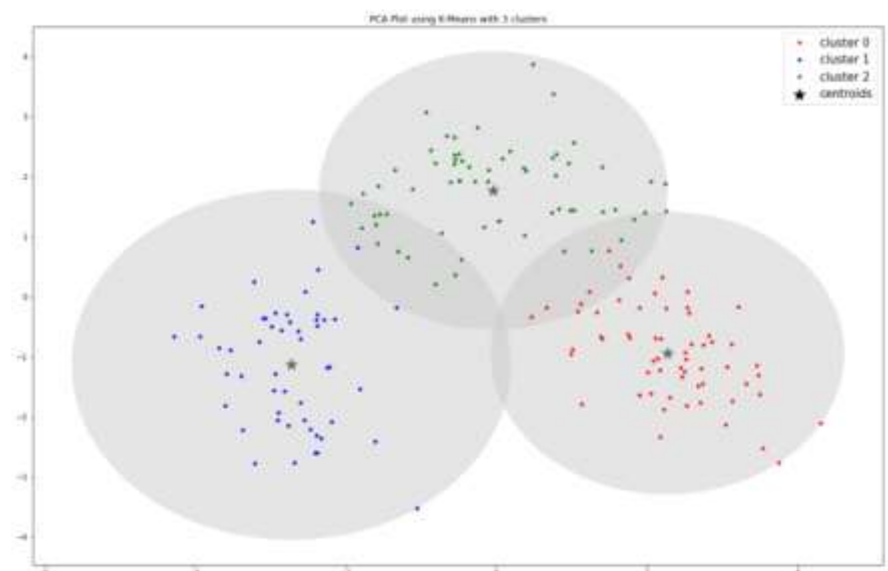
Silhouette scores for various number of clusters:

```
num_clusters: 2  silhouette_score: 0.25931695553182554
num_clusters: 3  silhouette_score: 0.2848589191898987
num_clusters: 4  silhouette_score: 0.24419555236115403
num_clusters: 5  silhouette_score: 0.23469284086426176
num_clusters: 6  silhouette_score: 0.19548548243786448
num_clusters: 7  silhouette_score: 0.16661003127824417
num_clusters: 8  silhouette_score: 0.14295550417594152
num_clusters: 9  silhouette_score: 0.1433301890321741
num_clusters: 10  silhouette_score: 0.13966690414947042
```

Silhouette score plot for final cluster = 3:



<u>Clusters Obtained from Kmeans followed by PCA</u>



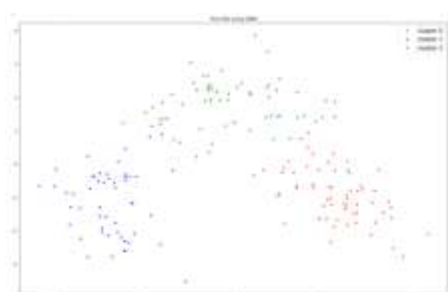
- On Visualizing the Silhouette scores for various number of clusters in range 1 to 10 and on observing thickness of each cluster; for K=3 we found out that Silhouette score is highest (0.2848) and also, visualized that each cluster has almost similar thickness for all 3 clusters.
- On visualizing Elbow method also we found out that elbow occurs at 3 number of clusters and after that Elbow method plot almost becomes stagnant.
- On visualizing clusters above we can see 3 clusters are nicely separated and capture the data distribution nicely.

B.2. <u>GMM & PCA</u>

1. First, standardize the data using Normal standardization.

2. Then we will use GMM and then used PCA to transform data along with clusters coloured with different colours.

3. We will keep final number of clusters same as we obtained from k-means and using those cluster centroids here for initialization. Tried with various types of covariance matrices:

Types of Covariance Matrix		
Full	**Diagonal**	**Identity**
		

Dataset 2: Breast Cancer Dataset

I. Dataset2 Description:

a. **Name of Dataset2 :** Breast Cancer Wisconsin (Diagnostic) Dataset (link - Dataset link).

b. **About Dataset2:** Features in the data are computed from a digitalized image of a fine needle aspirate (FNA) of breast mass that describe characteristics of the cell nuclei present in the image in the 3-dimensional space.

c. **Dataset1 features :**

1. id
2. diagnosis
3. radius_mean
4. texture_mean
5. perimeter_mean
6. area_mean
7. smoothness_mean
8. compactness_mean
9. concavity_mean
10. concave points_mean
11. symmetry_mean
12. fractal_dimension_mean
13. radius_se
14. texture_se
15. perimeter_se
16. area_se
17. smoothness_se
18. compactness_se
19. concavity_se
20. concave points_se
21. symmetry_se
22. fractal_dimension_se
23. radius_worst
24. texture_worst
25. perimeter_worst
26. area_worst
27. smoothness_worst
28. compactness_worst
29. concavity_worst
30. concave points_worst
31. symmetry_worst
32. fractal_dimension_worst

d. Data Pre-processing and Features Selection:

- Feature 'id' is dropped as for classification task id is not an attribute of breast.
- Found that there are not any Not Available (NA) values in the dataset.
- Removing the attribute (Name of Wine - categories) because the techniques to be used are Unsupervised Learning techniques.

II. Techniques Used:

Approach 1: PCA followed by Kmeans/GMM

A.

A.1. K-means & PCA

1. First, standardize the data using Normal standardization.

2. Then transform the data into 2-Dimensions with the help of Principal Component Analysis (PCA) technique.

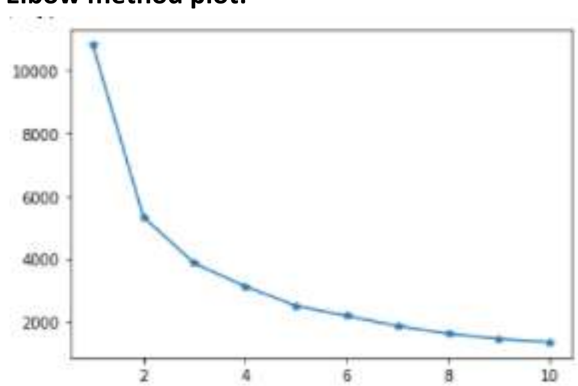
3. Now we will apply K-means technique to this 2-Dimensional transformed data.

4. Create regions for each cluster.

Analysis and Visualization:

Used Silhoutte scores and Elbow method to decide the final number of clusters to be taken for the dataset.

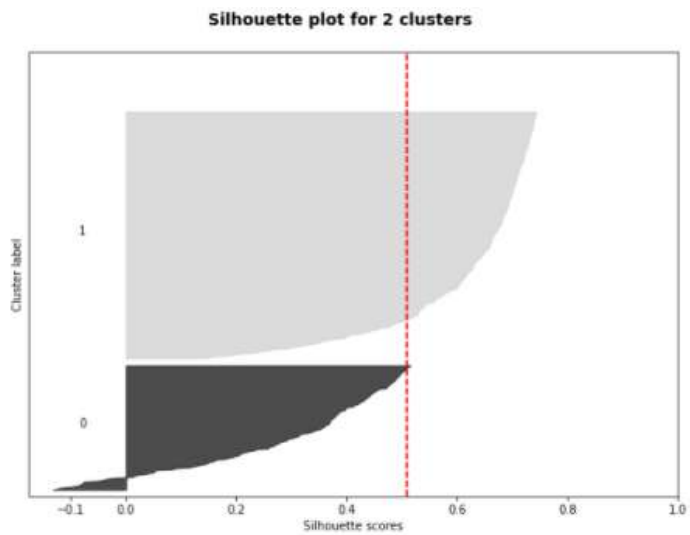
Elbow method plot:



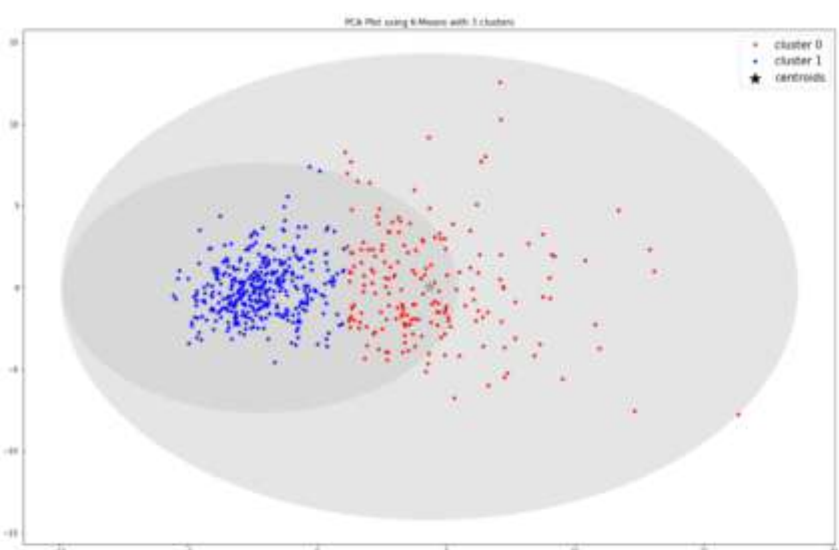
Silhouette scores for various number of clusters:

```
num_clusters: 2  silhouette_score: 0.5084690190672095
num_clusters: 3  silhouette_score: 0.47667244607401915
num_clusters: 4  silhouette_score: 0.4656018483069037
num_clusters: 5  silhouette_score: 0.36344947291702684
num_clusters: 6  silhouette_score: 0.3571164311820635
num_clusters: 7  silhouette_score: 0.3657077589480557
num_clusters: 8  silhouette_score: 0.3724505845972132
num_clusters: 9  silhouette_score: 0.3407102072607699
num_clusters: 10  silhouette_score: 0.33804979895724946
```

Silhouette score plot for final cluster = 2:



Clusters Obtained from PCA followed by Kmeans



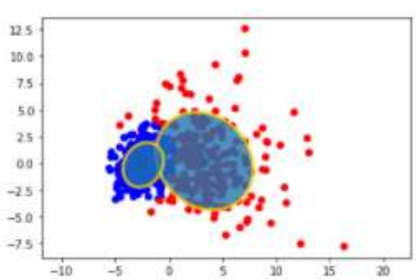
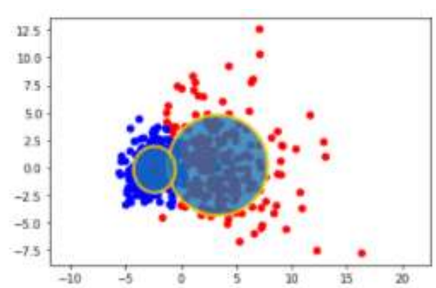
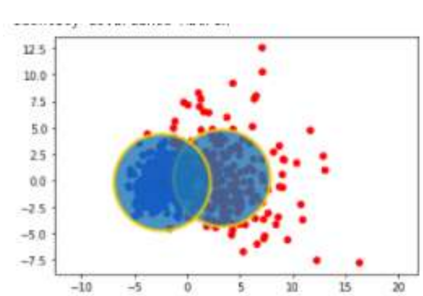
Aanalysis and Visualization:

- On Visualizing the Silhouette scores for various number of clusters in range 1 to 10 and on observing thickness of each cluster; for K=2 we found out that Silhouette score is highest (0.5084).
- On visualizing Elbow method also we found out that elbow occurs at 2 number of clusters.
- On visualizing clusters above we can see 2 clusters are nicely separated and capture the data distribution nicely.

A.2. <u>GMM & PCA</u>

1. First, standardize the data using Normal standardization.
2. Then transform the data into 2-Dimensions with the help of Principal Component Analysis (PCA) technique.
3. Now we will apply GMM techniques to this 2-Dimensional transformed data by keeping final number of clusters same as we obtained from k-means and suing those cluster centroids here for initialization. Drawn ellipses and spherical regions for each cluster. Tried with various types of covariance matrices.

Aanalysis and Visualization:

Types of Covariance Matrix		
Full	**Diagonal**	**Identity**
Shapes : Ellipses inclined at various angles	Axis Aligned Ellipses	Spherical
		
As we can visualize from above 3 shapes obtained from different types of Covariance matrices, spheres generated from that Identity Covariance matrix can capture comparatively better data distribution as compared to others.		

Approach 2: Kmeans/GMM followed by PCA

B.

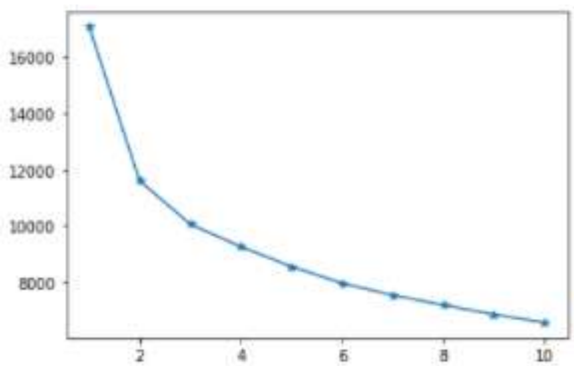
B.1. K-means & PCA

1. First, standardize the data using Normal standardization.
2. Then will use K-means and then used PCA to transform data along with clusters coloured with different colours.

Analysis and Visualization:

Used Silhoutte scores and Elbow method to decide the final number of clusters to be taken for the dataset.

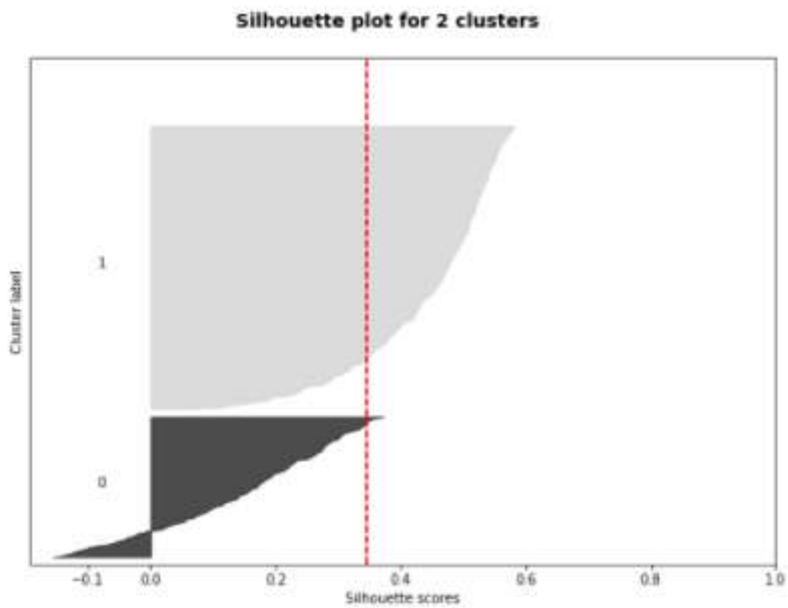
Elbow method plot:



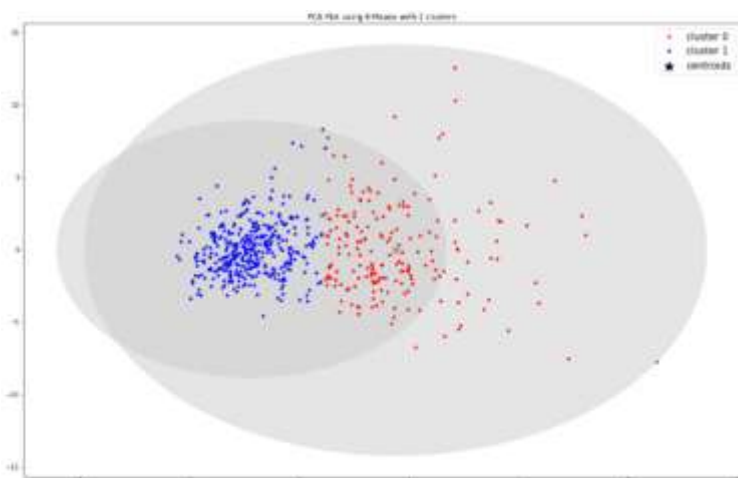
Silhouette scores for various number of clusters:

```
num_clusters: 2  silhouette_score: 0.3449740051034408
num_clusters: 3  silhouette_score: 0.3143840098608098
num_clusters: 4  silhouette_score: 0.2814748685818915
num_clusters: 5  silhouette_score: 0.17564864774698707
num_clusters: 6  silhouette_score: 0.1623342984960757
num_clusters: 7  silhouette_score: 0.15380731419772567
num_clusters: 8  silhouette_score: 0.13143186913523333
num_clusters: 9  silhouette_score: 0.14846949703298873
num_clusters: 10  silhouette_score: 0.1392799660081245
```

Silhouette score plot for final cluster = 2:



<u>Clusters Obtained from Kmeans followed by PCA</u>

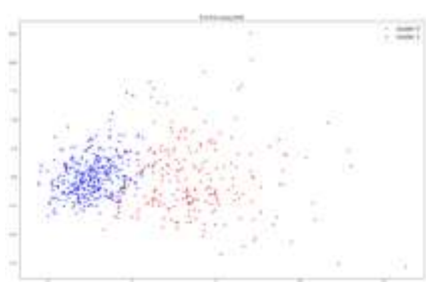
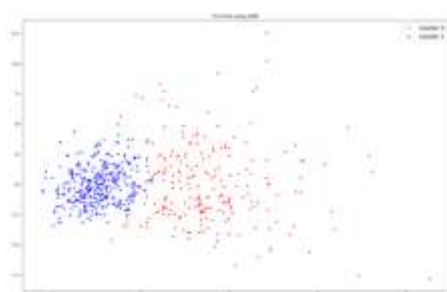
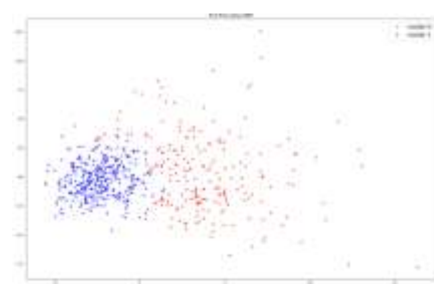


- On Visualizing the Silhouette scores for various number of clusters in range 1 to 10 and on observing thickness of each cluster; for K=2 we found out that Silhouette score is highest (0.3449).
- On visualizing Elbow method also we found out that elbow occurs at 2 number of clusters and after that Elbow method plot almost becomes stagnant.

- On visualizing clusters above we can see 2 clusters are nicely separated and capture the data distribution nicely.
- Through silhouette score plots we can observe the negative portions which say that some samples have been assigned wrong cluster.

B.2. <u>GMM & PCA</u>

1. First, standardize the data using Normal standardization.
2. Then we will use GMM and then used PCA to transform data along with clusters coloured with different colours.
3. We will keep final number of clusters same as we obtained from k-means and using those cluster centroids here for initialization. Tried with various types of covariance matrices:

Types of Covariance Matrix		
Full	**Diagonal**	**Identity**
		

Conclusion:

1. Analysed that higher the silhouette score tells better the number of clusters to take.

2. Observed and noticed that K-means followed by GMM can produce better results as K-means can help in initializing centres for GMM model