Question1

1. What was done
To answer the question, Principal Component Analysis (PCA) was performed on the wine dataset. First, the data was standardized to have zero mean and unit variance, which is important because PCA is sensitive to the scale of the variables. Then, PCA was applied to the standardized data, and the eigenvalues associated with each principal component were computed. Finally, the data was projected onto the first two principal components, and the 2D scatter plot was created to visualize the projection.

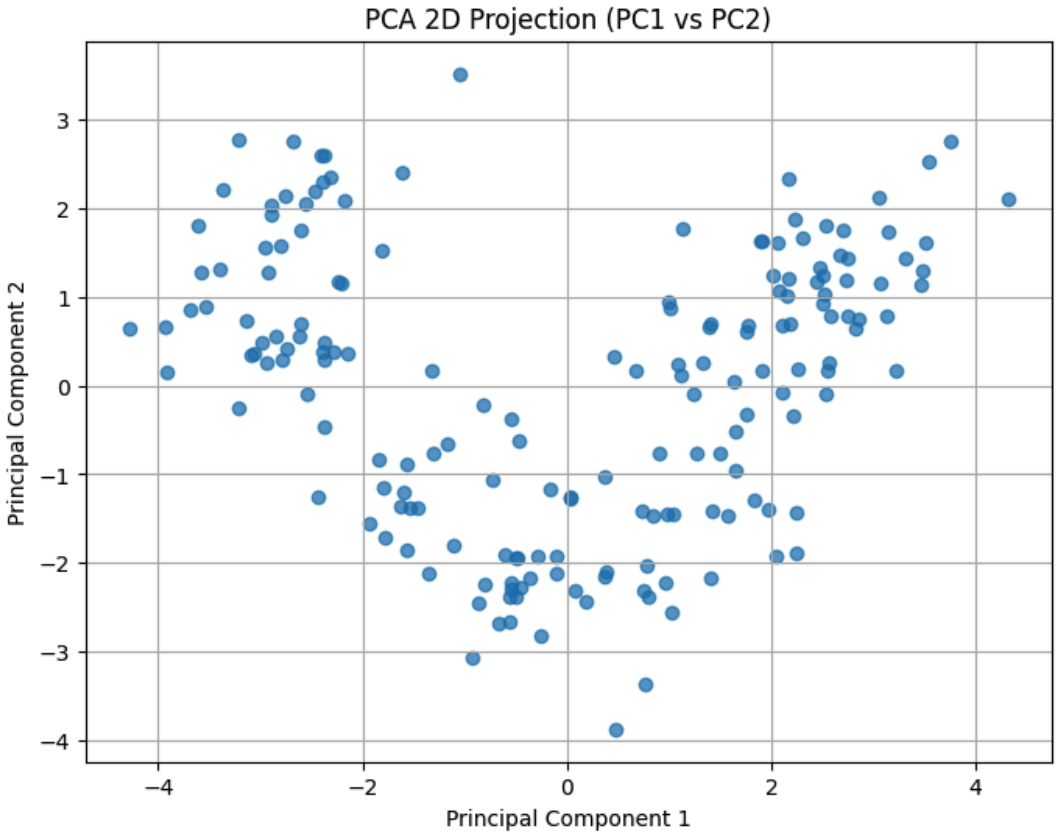
2. Why this was done
Standardizing the data was necessary because the original features (such as Alcohol, Magnesium, Proline, etc.) are on different scales, and PCA relies on variance, which would be biased toward higher-magnitude features without normalization. Using PCA helps to reduce the dimensionality while capturing the maximum variance in the data, making it easier to visualize the underlying structure of the wines.

3. What was found
The eigenvalues obtained from PCA were: [4.7324, 2.5110, 1.4542, 0.9242, 0.8580, 0.6453, 0.5544, 0.3505, 0.2905, 0.2523, 0.2271, 0.1697, 0.1039] It was found that 3 eigenvalues are greater than 1, indicating that three principal components explain more variance than an individual feature.

The variance explained by the first two principal components combined was approximately 55.41% (0.5541).

4. Interpretation of the findings
The fact that three eigenvalues are above 1 suggests that the intrinsic dimensionality of the wine dataset can be effectively reduced to three dimensions without losing much information. The first two components already capture over half of the variance, which is relatively good. In the 2D projection, we can observe that the wines tend to form loose groupings, implying that wines naturally cluster based on their chemical properties. However, some overlap remains, suggesting that full separability may require considering the third component as well.

```
All Eigenvalues:
[4.73243698 2.51108093 1.45424187 0.92416587 0.85804868 0.64528221
 0.55414147 0.35046627 0.29051203 0.25232001 0.22706428 0.16972374
 0.10396199]
Number of Eigenvalues above 1: 3
```




```
Variance explained by the first 2 components: 0.5541
```

Question2

1. What was done

To answer the question, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to the standardized wine dataset. First, the data was scaled to zero mean and unit variance. Then, t-SNE was run multiple times, varying the Perplexity parameter from 5 to 150 in steps of 5. For each run, the resulting Kullback-Leibler (KL) divergence was recorded. A plot of KL divergence versus Perplexity was created to visualize the relationship. Additionally, a 2D embedding of the dataset was generated using t-SNE with a Perplexity value of 20, and a scatter plot was produced.

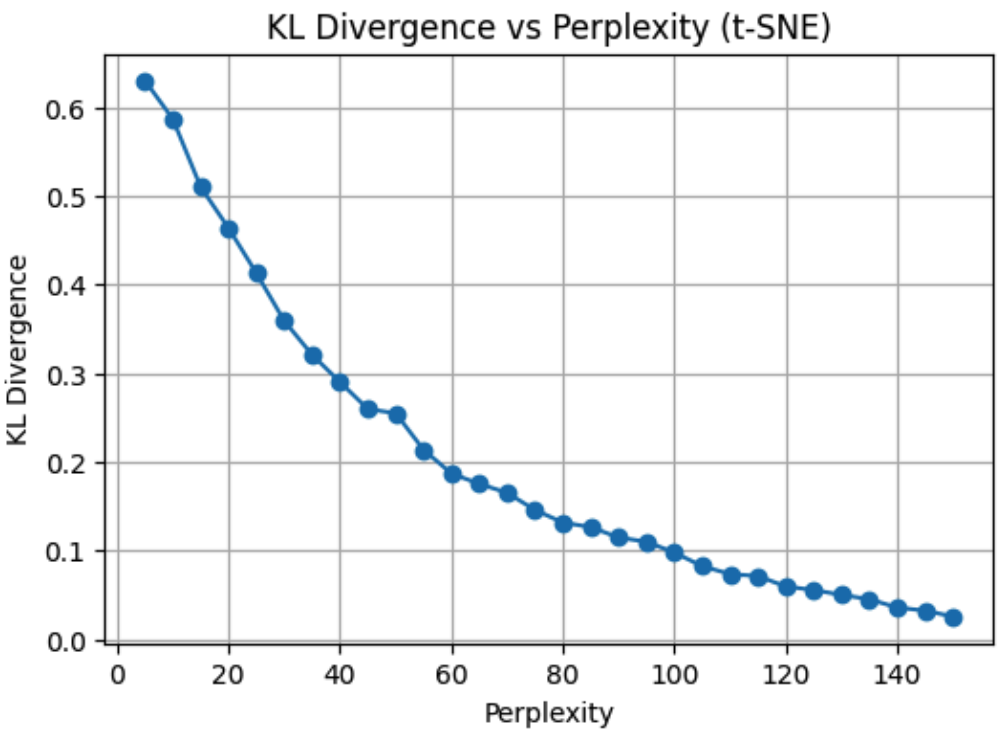
2. Why this was done

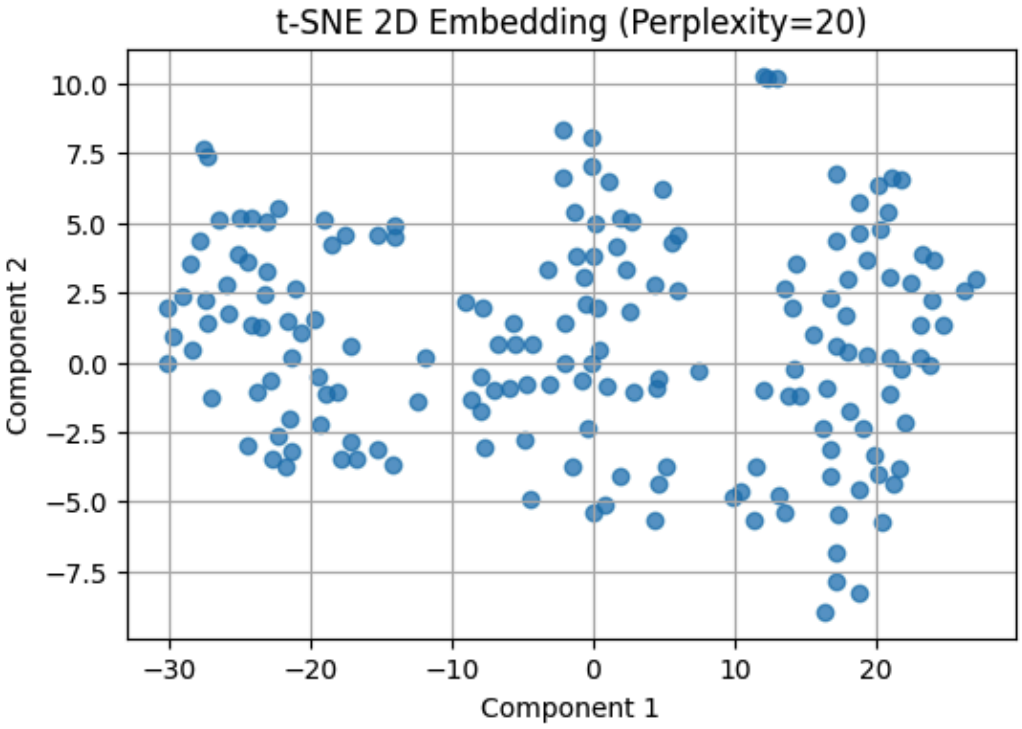
Perplexity is a key hyperparameter in t-SNE that influences the balance between local and global aspects of the data structure. It was important to systematically vary Perplexity and observe its impact on KL divergence to understand which settings lead to better preservation of data relationships. A Perplexity of 20 was specifically chosen for detailed visualization because

it is a common value that often provides a good balance for medium-sized datasets like this one (~178 samples).

3. What was found
The KL divergence generally decreased as Perplexity increased, as shown in the first plot. For example, at Perplexity 5, the KL divergence was above 0.6, but it dropped below 0.1 when Perplexity reached around 150.
The 2D t-SNE embedding with Perplexity = 20, shown below, reveals visible grouping patterns among the wine samples.

4. Interpretation of the findings
The decreasing KL divergence trend suggests that higher Perplexity values help the model better approximate the high-dimensional relationships, leading to a smoother embedding. However, very high Perplexity may cause over-smoothing, potentially hiding local structure. The 2D plot at Perplexity = 20 shows that the wine data is not randomly scattered but tends to form distinct local clusters, implying that the wines naturally group based on their chemical compositions. This supports the idea that the dataset contains meaningful internal structure, making it suitable for further clustering analysis.


KL Divergence vs Perplexity (t-SNE)
KL Divergence
Perplexity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
20
40
60
80
100
120
140


t-SNE 2D Embedding (Perplexity=20)
Component 2
Component 1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−30
−20
−10
0
10
20

Question3

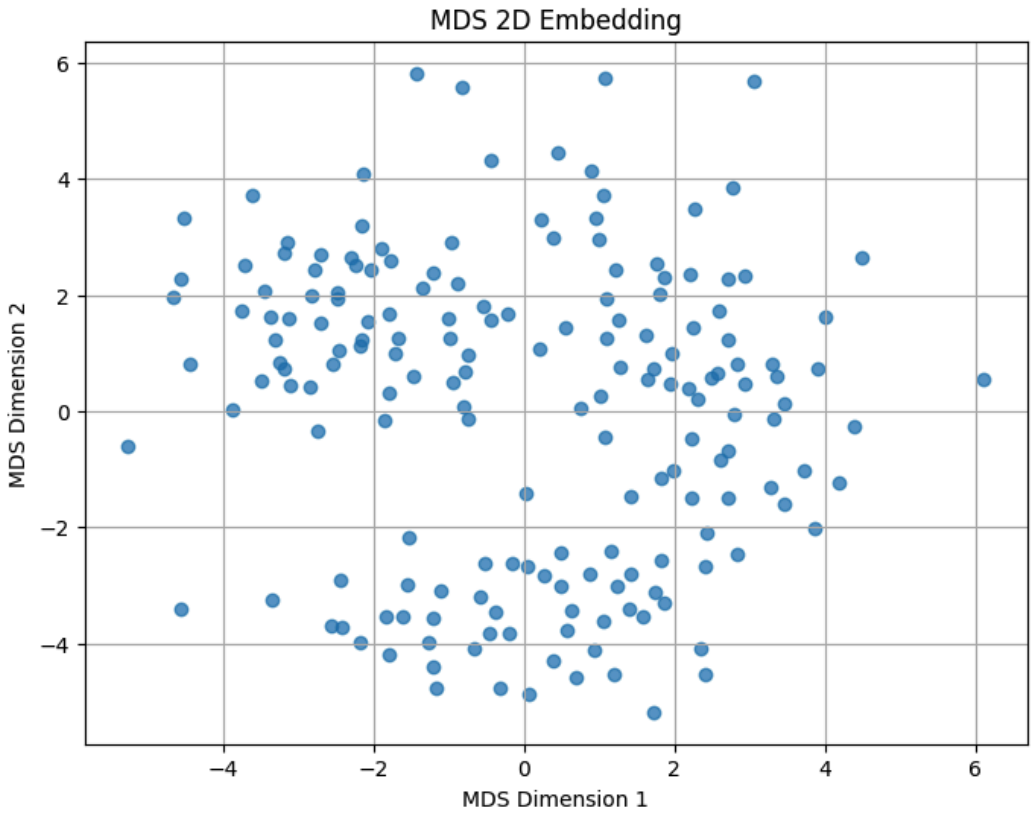
1. What was done
To answer the question, Multidimensional Scaling (MDS) was applied to the standardized wine dataset. First, the data was scaled to zero mean and unit variance. Then, MDS was performed to create a 2-dimensional embedding based on the Euclidean distances between samples. The resulting 2D embedding was visualized in a scatter plot, and the stress value—a measure of the embedding distortion—was recorded.

2. Why this was done
Standardizing the data ensures that all features contribute equally to the distance calculations used by MDS. MDS was chosen because it preserves the pairwise distances between samples as faithfully as possible in a lower-dimensional space, allowing us to visually inspect the structure of the data. Using two dimensions was appropriate to allow for direct visualization while keeping the problem computationally manageable.

3. What was found
The resulting stress from the MDS embedding was 21,818.8940.
The scatter plot below shows the 2D MDS embedding of the wine samples. Compared to the t-SNE results observed earlier, the MDS embedding shows a less distinct grouping structure. The samples appear more uniformly spread out without tight local clusters.

4. Interpretation of the findings
The relatively high stress value indicates that it was challenging to perfectly preserve the original pairwise distances in only two dimensions. Compared to t-SNE, which revealed clearer groupings, MDS appears to retain more of the global distance structure at the expense of local cluster separation. This suggests that while MDS provides a faithful distance-preserving embedding overall, t-SNE is more effective for discovering local clustering patterns within the wine dataset.

```
Resulting Stress from MDS embedding: 21818.8940
```



Question4

1. What was done

To answer the question, Principal Component Analysis (PCA) was first applied to the standardized wine dataset to reduce its dimensionality to two components. Then, the Silhouette method was used to determine the optimal number of clusters by computing the Silhouette Score for different values of k (ranging from 2 to 10) using kMeans clustering on the 2D PCA projection. Finally, kMeans was performed with the optimal number of clusters, and the resulting clustering was visualized in a 2D scatter plot.

2. Why this was done

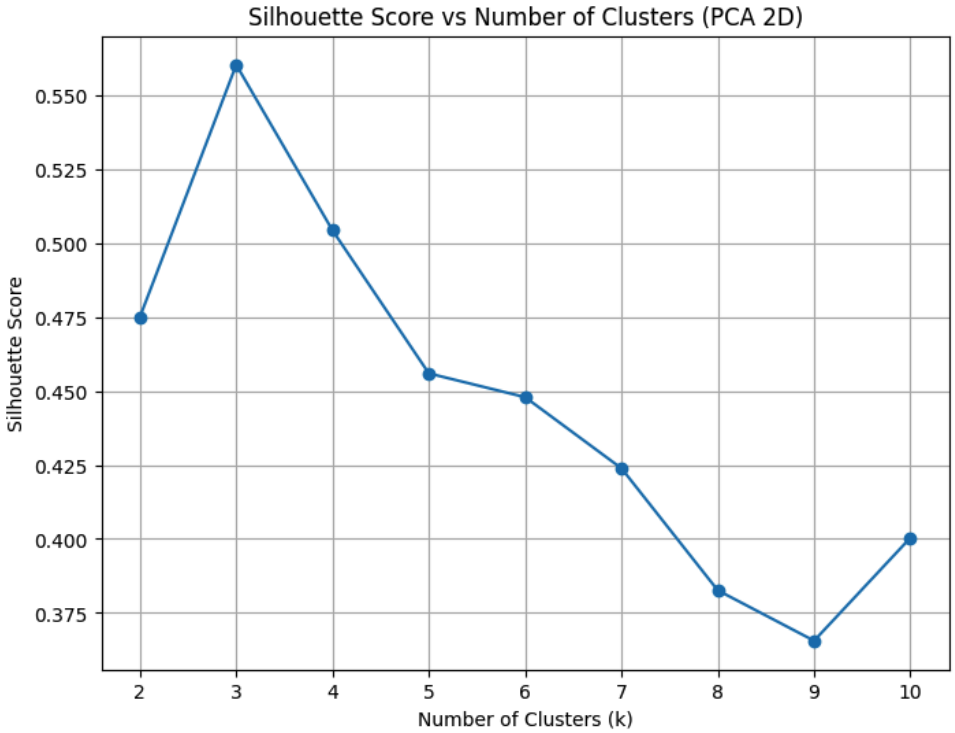
PCA was chosen to reduce the data to two dimensions to make visualization feasible while preserving as much of the variance as possible. The Silhouette Score was used because it quantitatively evaluates clustering quality by measuring how similar each point is to its own cluster compared to other clusters. Selecting the k that maximizes the Silhouette Score ensures that the clustering is both dense and well-separated.

3. What was found

The Silhouette Score peaked at k = 3, with a score of approximately 0.5541, suggesting that three clusters best represent the structure in the PCA-projected data.
The total sum of distances (inertia) between all points and their respective cluster centers for the final clustering solution was 260.0167.

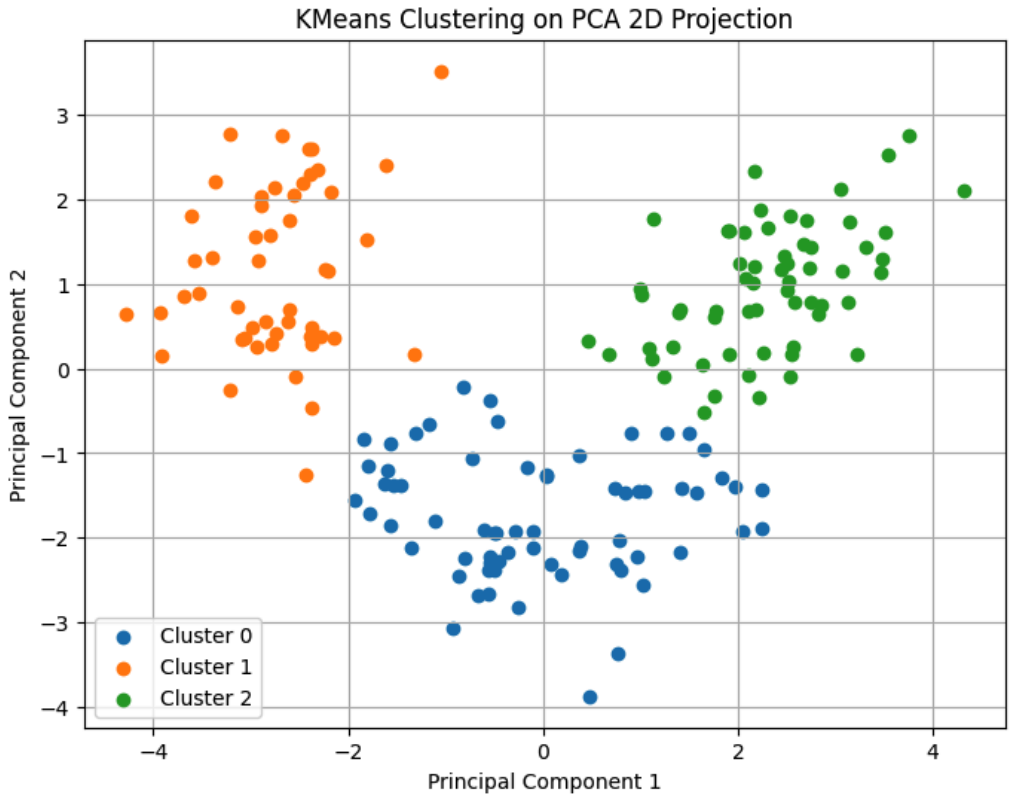
The two figures below show the Silhouette Score across different k values and the final 2D clustering result:

Silhouette Score vs Number of Clusters (k):




```
Optimal number of clusters (k): 3
```

KMeans Clustering on PCA 2D Projection:



Total sum of distances to cluster centers (Inertia): 260.0167

4. Interpretation of the findings
The results indicate that the wine samples naturally separate into three major groups based on their chemical properties when projected into two dimensions. The relatively high Silhouette Score (0.5541) suggests that the clusters are reasonably well-defined and distinct. The final kMeans clustering shows three clearly separated groups, supporting the idea that there are roughly three underlying types of wines in this dataset.

Question 5

1. What was done
To answer the question, Principal Component Analysis (PCA) was first applied to the standardized wine dataset to reduce it to two dimensions. Then, the K-distance graph was used to determine an appropriate value for the epsilon parameter (eps) in DBSCAN by plotting the distance to each point's 5th nearest neighbor and identifying the elbow point. After selecting epsilon, DBSCAN was run with different min_samples values from 4 to 10. The best min_samples value was selected based on the highest Silhouette Score. Finally, DBSCAN clustering was performed using the selected epsilon and best min_samples, and the 2D clustering result was visualized.

2. Why this was done
The K-distance method was chosen because it provides a systematic way to select the epsilon parameter, critical for DBSCAN's performance. By plotting sorted nearest-neighbor distances, the point where the graph shows a sharp increase (elbow) typically indicates a suitable epsilon. Trying multiple values of min_samples and selecting the one with the best Silhouette Score ensures that the clustering quality is maximized based on both local density and separation between clusters.

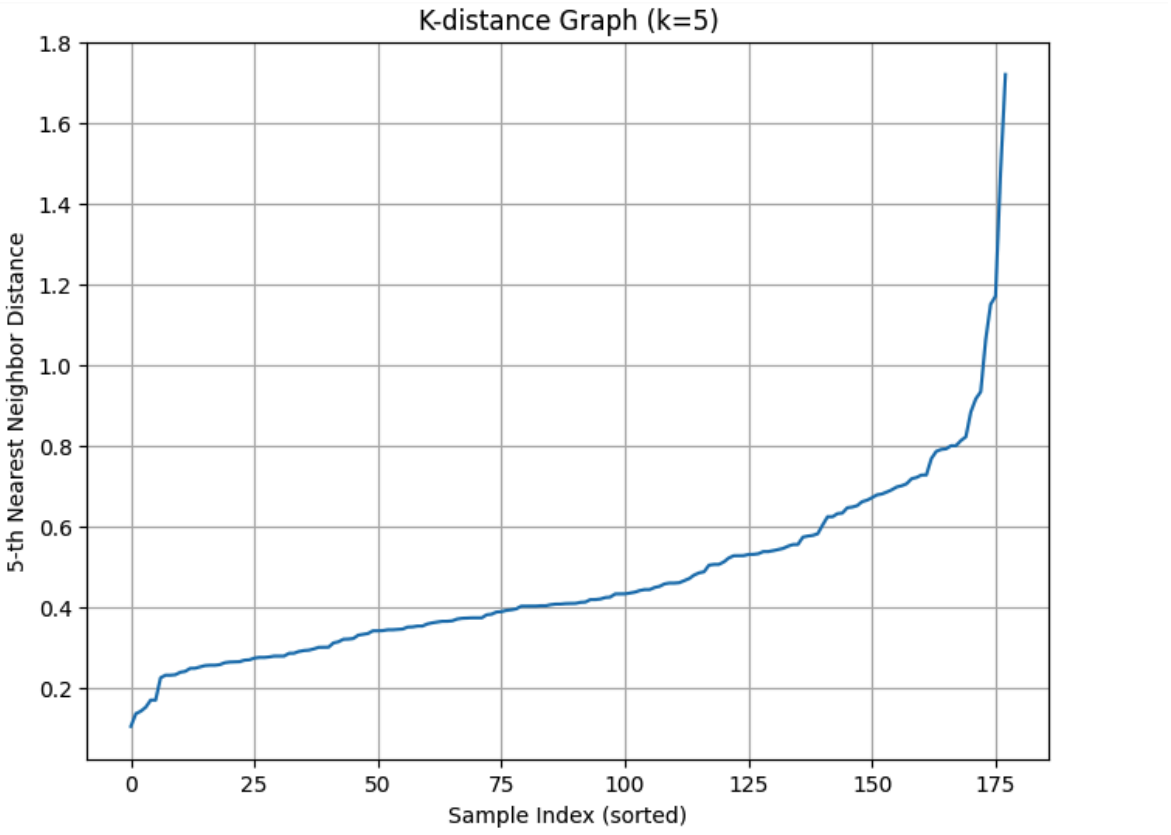
3. What was found
The selected epsilon value, estimated from the K-distance graph, was approximately 0.7271.

The best value for min_samples was found to be 10, yielding a Silhouette Score of 0.4700.

The final DBSCAN clustering produced three clusters and identified several noise points (marked with 'x').

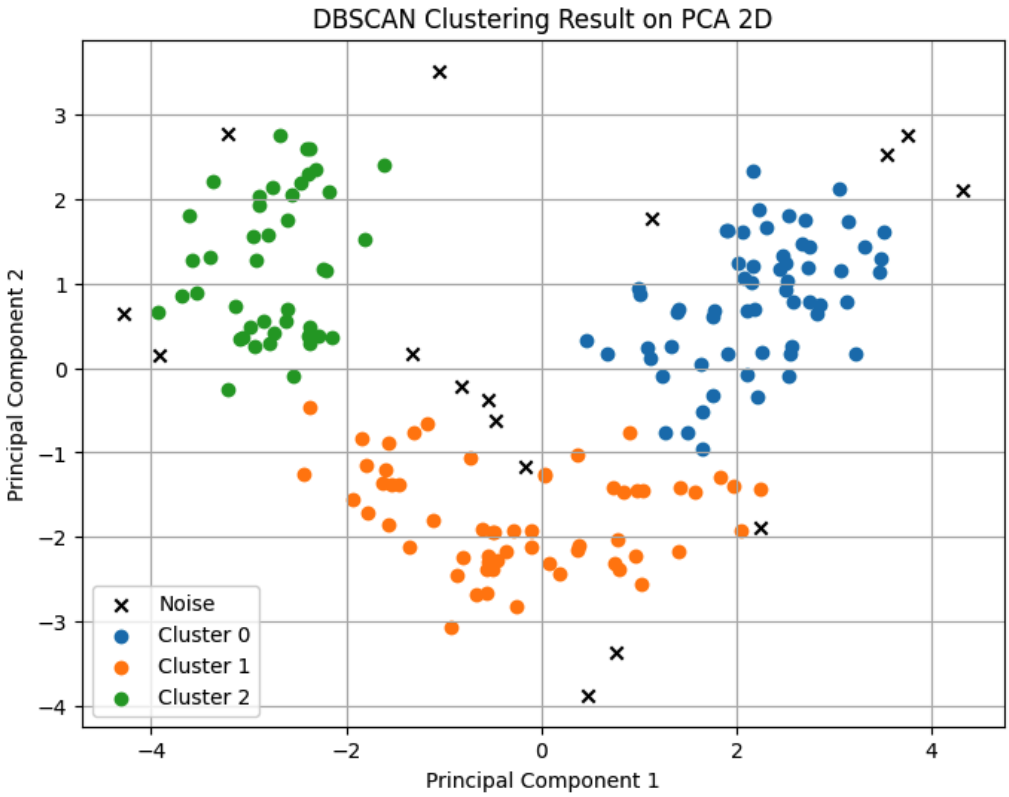
The two plots below show:

K-distance Graph (k=5):




```
Selected epsilon (approx from K-distance graph): 0.7271
Best min_samples: 10, Best Silhouette Score: 0.4700
```

DBSCAN Clustering Result on PCA 2D Projection:



4. Interpretation of the findings
The clustering results suggest that the wine dataset naturally forms three dense groupings in the 2D PCA space, aligning with earlier observations from kMeans clustering. The presence of noise points indicates that some wines are outliers and do not clearly belong to any major group, which could reflect rare or unique wine profiles. Overall, DBSCAN successfully identified meaningful clusters without requiring a predefined number of clusters, reinforcing the structure seen in previous analyses but also offering additional flexibility in handling outliers.

Extra credit

1:
1. What was done
To answer this question, the clustering results from both kMeans and DBSCAN applied to the PCA-reduced wine dataset were analyzed together. The Silhouette Scores of both clustering methods were computed to evaluate the clustering quality. Additionally, for DBSCAN, the number of noise points detected and the sizes of each cluster (including noise) were summarized. Visualizations were produced to show the distribution of cluster sizes and the 2D scatter plot of the clustered wines with noise points clearly highlighted.

2. Why this was done
Silhouette Scores provide a quantitative measure of how well the samples are clustered, with higher scores indicating denser and better-separated clusters. Analyzing the number and sizes of clusters, along with the noise points, helps to objectively determine how many different types of wine there may be and whether there are significant outliers. Visualizations make the findings more intuitive and help confirm that the clustering structure is meaningful and not an artifact of the modeling process.

3. What was found
Silhouette Score (kMeans): 0.5602

Silhouette Score (DBSCAN): 0.5787

Number of Noise Points detected by DBSCAN: 16 samples

Cluster Sizes (including noise):

Cluster 0: 60 samples

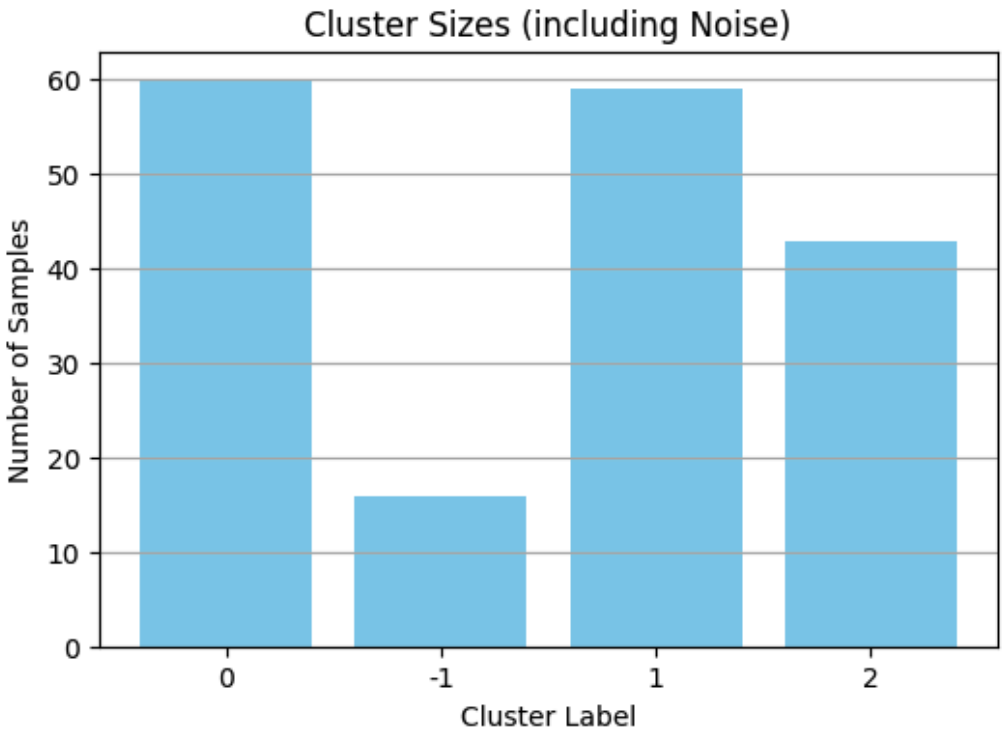
Cluster 1: 59 samples

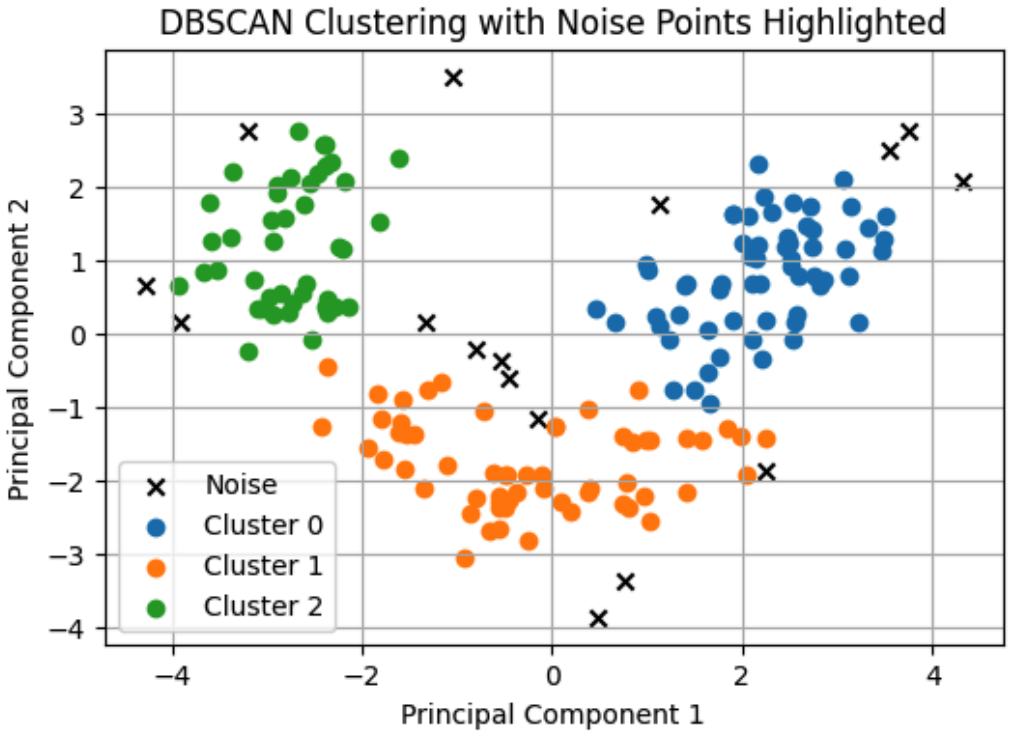
Cluster 2: 43 samples

Noise: 16 samples

The following figures illustrate the results:

Cluster Sizes (including Noise) and DBSCAN Clustering with Noise Points Highlighted:





4. Interpretation of the findings

Based on the consistent findings from both kMeans and DBSCAN clustering, it can be concluded that the wine dataset naturally separates into three main types. The high Silhouette Scores for both methods indicate that the clustering structure is robust and meaningful.

Furthermore, DBSCAN identified 16 noise points, suggesting that a subset of wines have significantly different chemical profiles compared to the main groups. These outlier wines could represent rare or specialty wines worth further investigation, both for commercial opportunities and for quality control purposes. Overall, the data-driven analysis supports the presence of three major kinds of wine with distinct characteristics.

2:
Through a series of unsupervised learning methods—including PCA, t-SNE, MDS, kMeans, and DBSCAN—we systematically explored the structure of the wine dataset. Dimensionality reduction techniques consistently revealed a natural separation among the samples, suggesting the presence of several distinct types of wines. Both clustering approaches, kMeans and DBSCAN, further supported this observation by identifying clear groupings in the data.

At the same time, DBSCAN revealed a set of samples that did not fit into any major cluster, labeling them as outliers. These wines, which diverge from the main population in terms of their chemical properties, naturally sparked my interest. Their existence within the data suggests two compelling avenues for further exploration: from a commercial standpoint, outlier wines may represent rare or unique profiles that could be positioned as premium or specialty products; from a quality control perspective, they may signal deviations from standard production processes, making early detection crucial for maintaining overall consistency and brand reputation.

The discovery of these outliers, uncovered through an objective and systematic unsupervised analysis, opens up valuable opportunities for both business development and operational quality assurance in the wine industry.