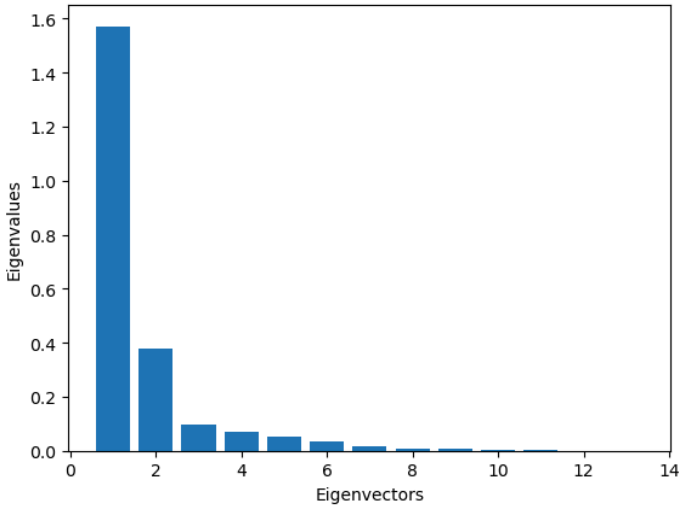
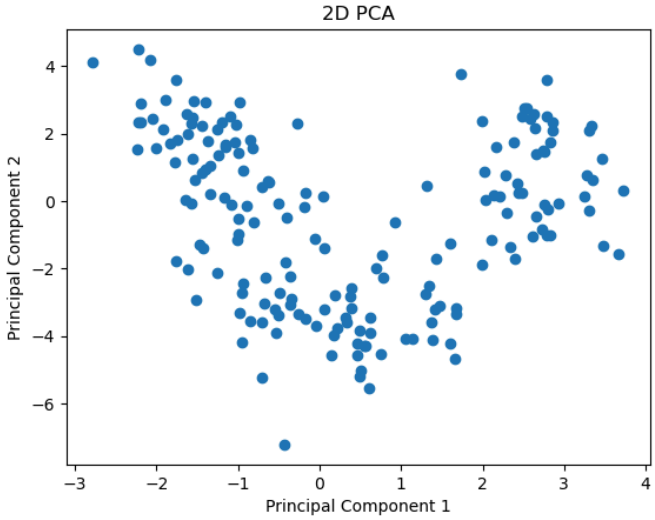
1. Do a PCA on the data. How many Eigenvalues are above 1? Plotting the 2D solution (projecting the data on the first 2 principal components), how much of the variance is explained by these two dimensions, and how would you interpret them?

I standardized the dataset and generated the covariance matrix. I then ran PCA. This gave me the following bar graph of eigenvectors and eigenvalues. There is one axis with an eigenvalue above 1.



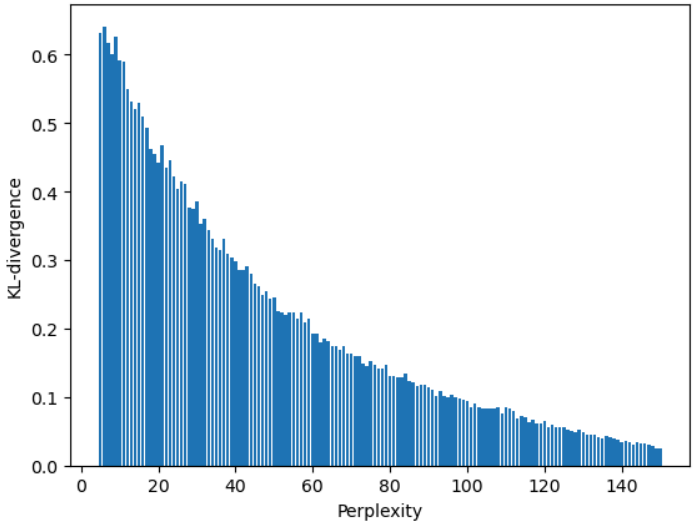
This graph shows that a large amount of the variance is accounted for by the first two principal components.



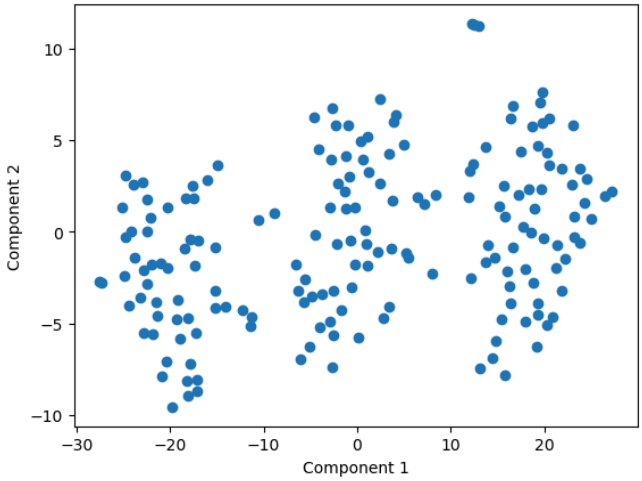
These two principal components account for 86.9758% of the variance in the data. The first principal component is largely composed of columns 2, 4, and 8, which I interpreted as acidity. This is likely flawed as I lack the domain knowledge to interpret this properly. The second principal component is largely composed of columns 1, 10, and 13, so I interpreted this as strength.

2. Use t-SNE on the data. How does KL-divergence depend on Perplexity (vary Perplexity from 5 to 150)? Make sure to plot this relationship. Also, show a plot of the 2D component with a Perplexity of 20.

I ran t-SNE on the standardized data with 2 dimensions, as the previous PCA showed that two components accounted for the majority of the data. Graphing the Perplexity against the KL-divergence gives the following graph. This shows that as Perplexity increases, KL-divergence decreases, as expected.



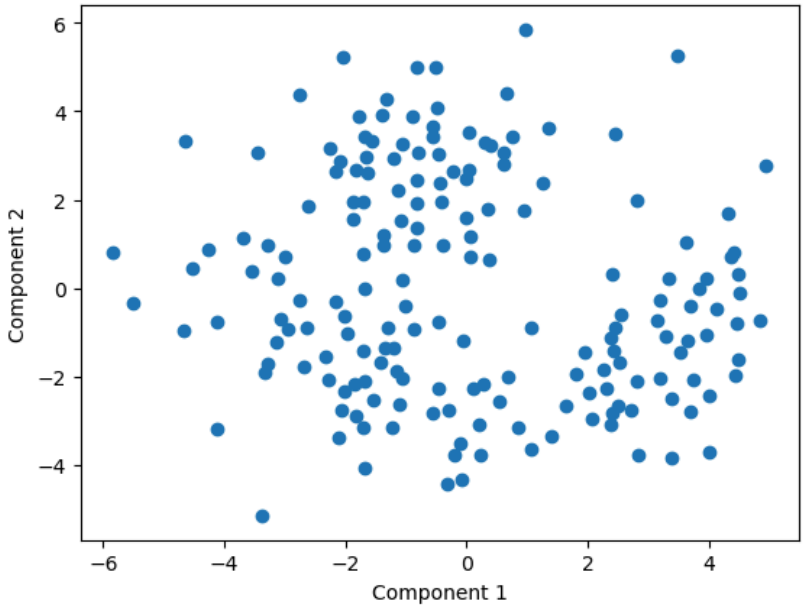
The 2D component with a Perplexity of 20 gives the following plot.



There appear to be three clusters.

3. Use MDS on the data. Try a 2-dimensional embedding. What is the resulting stress of this embedding? Also, plot this solution and comment on how it compares to t-SNE.

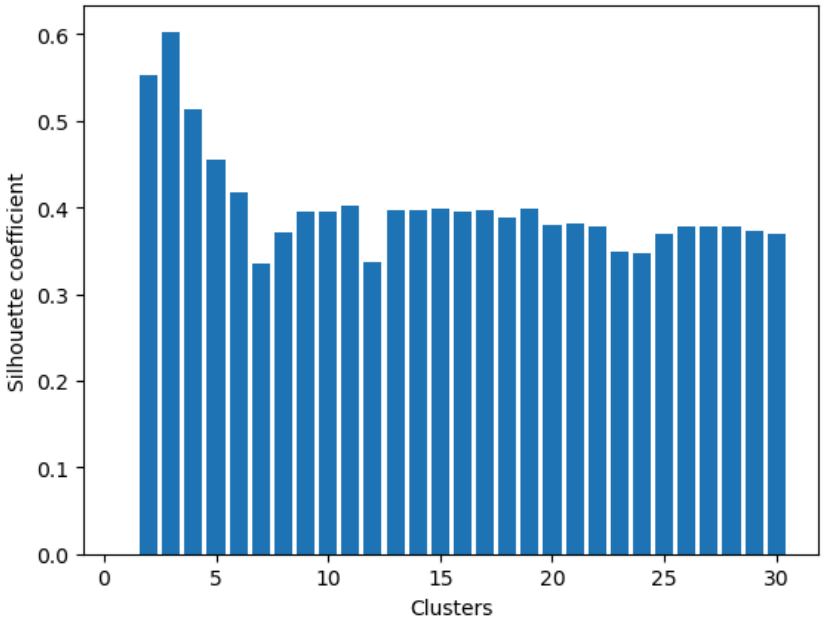
Running MDS to create a 2D embedding with Euclidean distance as dissimilarity. gives a stress of 21192.94. This is very high, which shows that a 2D space does not capture the data well at all, which does not make too much sense as other methods of dimension reduction have shown otherwise.



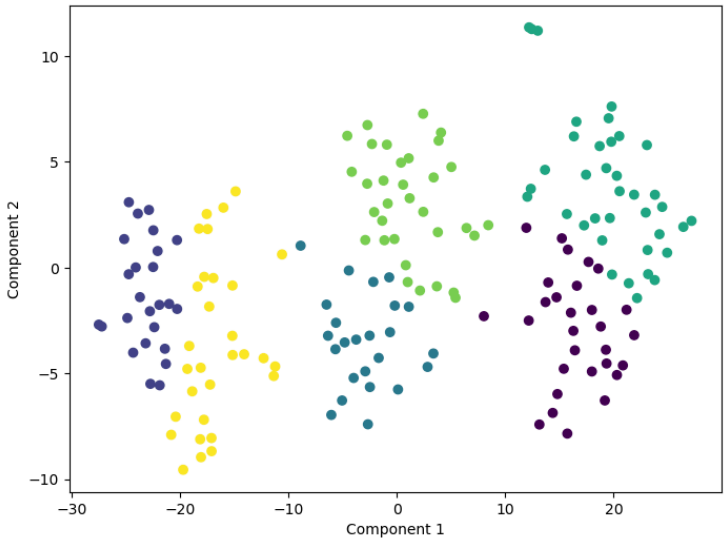
This graph is very different from the t-SNE graph. This is expected, as t-SNE and MDS preserve different aspects of the original dataset, and t-SNE does not preserve the global structure of the data well at low dimensions.

4. Building on one of the dimensionality reduction methods above that yielded a 2D solution (1-3, your choice), use the Silhouette method to determine the optimal number of clusters and then use kMeans with that number (k) to produce a plot that represents each wine as a dot in a 2D space in the color of its cluster. What is the total sum of the distance of all points to their respective clusters centers, of this solution?

I chose to use t-SNE. Doing the silhouette method yielded the following graph:



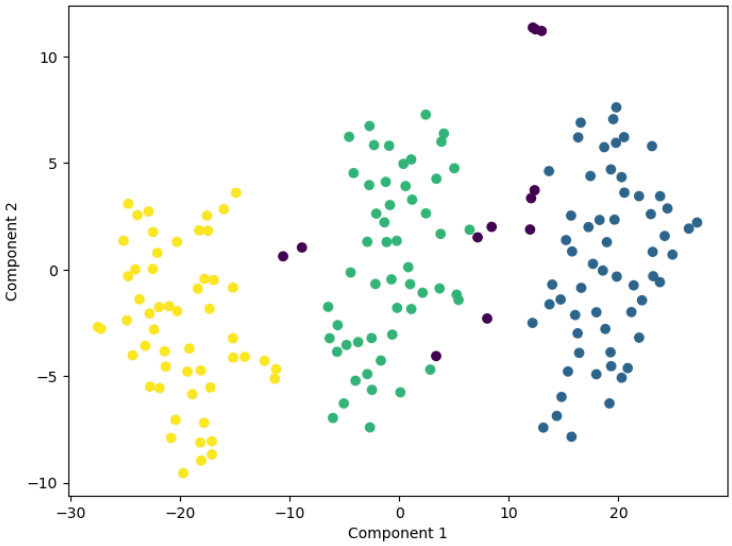
The range of k I used was 2-30, because that seemed like a reasonable range for types of wines to me. The minimum silhouette score was 0.33520, from the kMeans with 6 clusters.



This resulted in the above graph. However, due to the tendency of kMeans to get stuck in local minima, running this again resulted in different clusters almost every time. This gave a total sum of the distance of all points to their respective clusters' center of 3207.7637, which is fairly high.

5. Building on one of the dimensionality reduction methods above that yielded a 2D solution (1-3, your choice), use dBScan to produce a plot that represents each wine as a dot in a 2D space in the color of its cluster. Make sure to suitably pick the radius of the perimeter ("epsilon") and the minimal number of points within the perimeter to form a cluster ("minPoints") and comment on your choice of these two hyperparameters

As I lack the knowledge to pick a good minPoints value, I chose the natural log of the number of samples, which rounded down to 5. I also do not have the domain knowledge to pick a proper epsilon, so I just tried different values of epsilon until I produced a plot that visually seemed most correct. This produced the following graph, showing that there are 4 clusters, although the 4th cluster seems to just be outliers.



Extra Credit:
a) Given your answers to all of these questions taken together, how many different kinds of wine do you think there are and how do they differ?

I think that there were 6 types of wine, as the silhouette method on kMeans yielded the lowest silhouette coefficient for 6 clusters. I think they differ in acidity and strength.