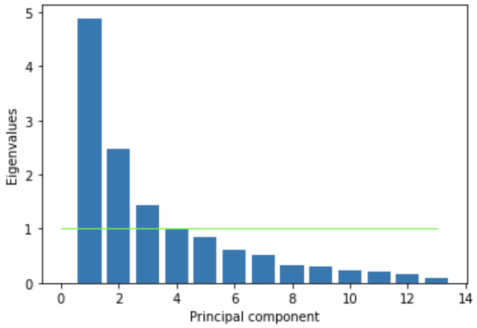
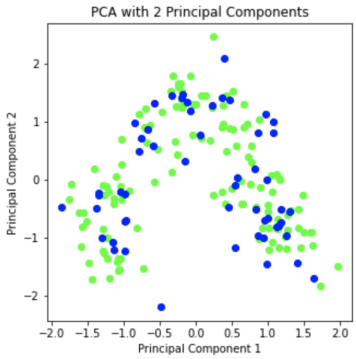
Question 1

1. I first split the data into train and test sets with a 70/30 split. Then, I used Standard Scaler to prepare the data before doing PCA. Then, I did PCA and generated the eigenvalues for each of the principal components. Finally, I graphed my eigenvalues and found how many of said values were above 1.
2. I prepared the data using Standard Scaler to make sure that running a PCA was possible. I graphed the eigenvalues to visually be able to easily see how the eigenvalues compared to one another and which were significant for consideration.
3. I found that three of the principal components have eigenvalues greater than 1. Furthermore, I found that 37.329648% of the variance is explained by the first component and 18.818926% of the variance is explained by the second component.
4. Since the first two principal components explain over 50% of the variance, these two principal components are more important than the other eleven principal components combined when it comes to explaining the variance.

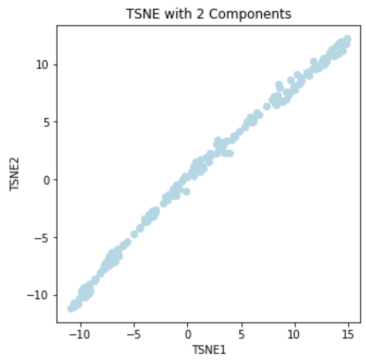


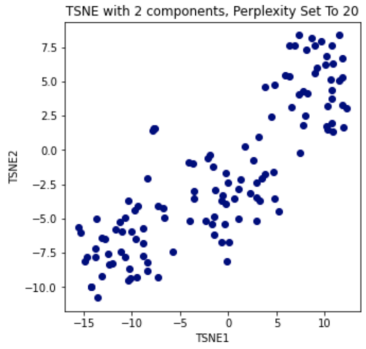


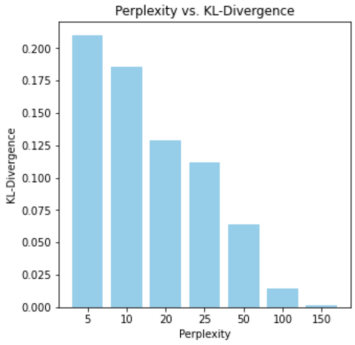
Question 2

1. I first split the data into train and test sets with a 70/30 split. Then, I used Standard Scaler to prepare the data before doing a t-SNE analysis. Then, I did a t-SNE analysis on the data and fit it. Then, I plotted the t-SNE with two components. Then, I made an array with values for Perplexity ranging from 5 to 150, doing several t-SNE analyses with these values set as Perplexity, plotting Perplexity against the KL-Divergence on a bar graph. Finally, I plotted the 2D component with Perplexity set to 20.
2. I prepared the data using Standard Scaler to make sure it was ready for t-SNE analysis. I ran a for loop throughout an array of Perplexity values ranging from 5 to 150 to see how KL-Divergence changes when changing the Perplexity. Finally, I graphed Perplexity vs. KL-Divergence to visualize how KL-Divergence changes when Perplexity varies and also plotted the 2D Component when Pexplexity is not set to 20 and when it is to see what changes.
3. I found that KL-Divergence decreases as Perplexity increases. I also found that the t-SNE with 2 components with Perplexity set to 20 shows a negative relationship.

4. The negative linear relationship and clustering shown in the graphs indicates that a higher Perplexity is more suitable for this dataset.

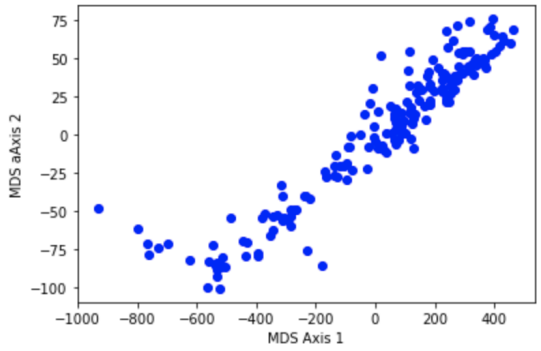






Question 3

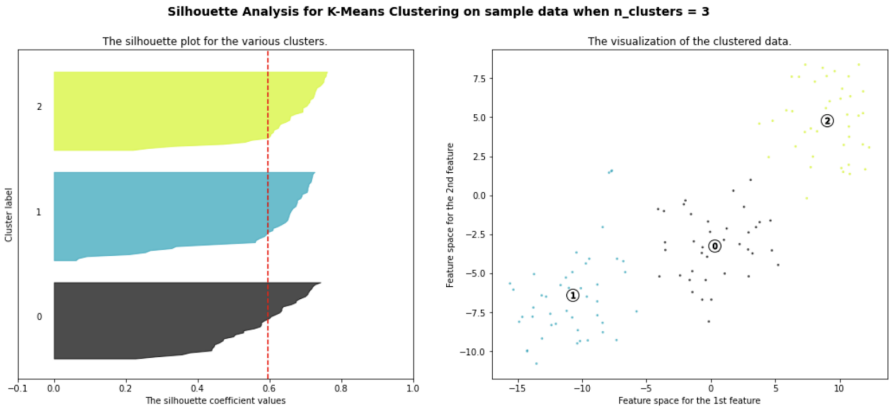
1. I first used pairwise distances in the DataFrame to create a square matrix. Then, I ran an MDS on the data and used 2-dimensional embedding. Finally, I calculated the stress score.
2. I used pairwise distances to get the distances between all the points in the data. I ran an MDS and used 2-dimensional embedding as the question stated to do so.
3. I found a stress score for the embedding to be 0.018010074536651818. I also found there to be a general positive relationship between the axes as shown on the graph.
4. The low stress score shows that this model is a good fit for the data. Furthermore, the plotted solution is similar to t-SNE as a linear relationship is shown between the two components.



Question 4

1. First, I used the 2D t-SNE method. Then, I made an array with cluster values ranging from 2 to 6 that I looped through, constantly applying the Silhouette method to find the optimal number of clusters. I then used K-Means with the optimal number of clusters. Finally, I calculated the total sum of the distance of all the points.

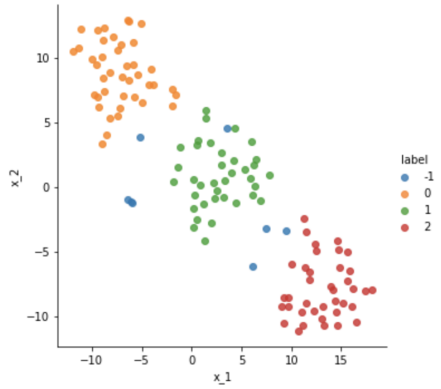
2. I ran a loop throughout an array of cluster values to find the best number of clusters to use. I used the Silhouette score method to find the optimal number of clusters since this is a more accurate method to do so as opposed to eyeballing.
3. I found the best silhouette score to be 0.5954112 and the sum of the distances of all the points to their respective cluster centers is 1495.9091796875.
4. The highest Silhouette score (as opposed to the Silhouette scores of other cluster numbers) of 0.5954112 means that three clusters is the optimal number of clusters in this dataset. Furthermore, the sum of the distances is relatively normal and all the clusters can be easily identified visually.



Question 5

1. First, I used the 2D t-SNE method. Then, I ran a DBScan several times with different epsilon values and different minPoints.
2. I ran a DBScan with different epsilon values and different minPoints to find the ideal values for both of these.
3. I found the best epsilon value to be 2 and the best minPoints value to be 4 with there being 103 core points with 3 clusters and 8 unclassified points when these parameters are set.
4. An epsilon value of 2 and a minPoints value of 4 produces the ideal result as it shows 3 distinct clusters, minimizes unclassified points, and maximizes core points.

```
Number of core points: 103
Number of clusters: 3
Number of unclassified points: 8
```



Extra Credit 1

I think there are three different kinds of wine because previous questions have indicated three clusters to be the optimal number for this dataset. I think these three wines differ in the features of alcohol, ash alkalinity, ash, and stilbene as these are the principal components that made the biggest difference in my PCA.

Extra Credit 2

As someone who loves to drink wine, it was really interesting to learn about wine through this dataset. Something new I learned that's worth noting is that there is a moderate positive correlation between alcohol content and color intensity as they have a correlation coefficient of 0.5463641950837039. This means that the higher the alcohol content, the more concentrated (darker) the color of the wine is – I didn't know that there was a notable correlation between these two things.