DAT-330 Assignment 1

Anne Konicki

Question 1:

Part A

Identify two attributes from the dataset that are most suitable for performing cluster analysis.

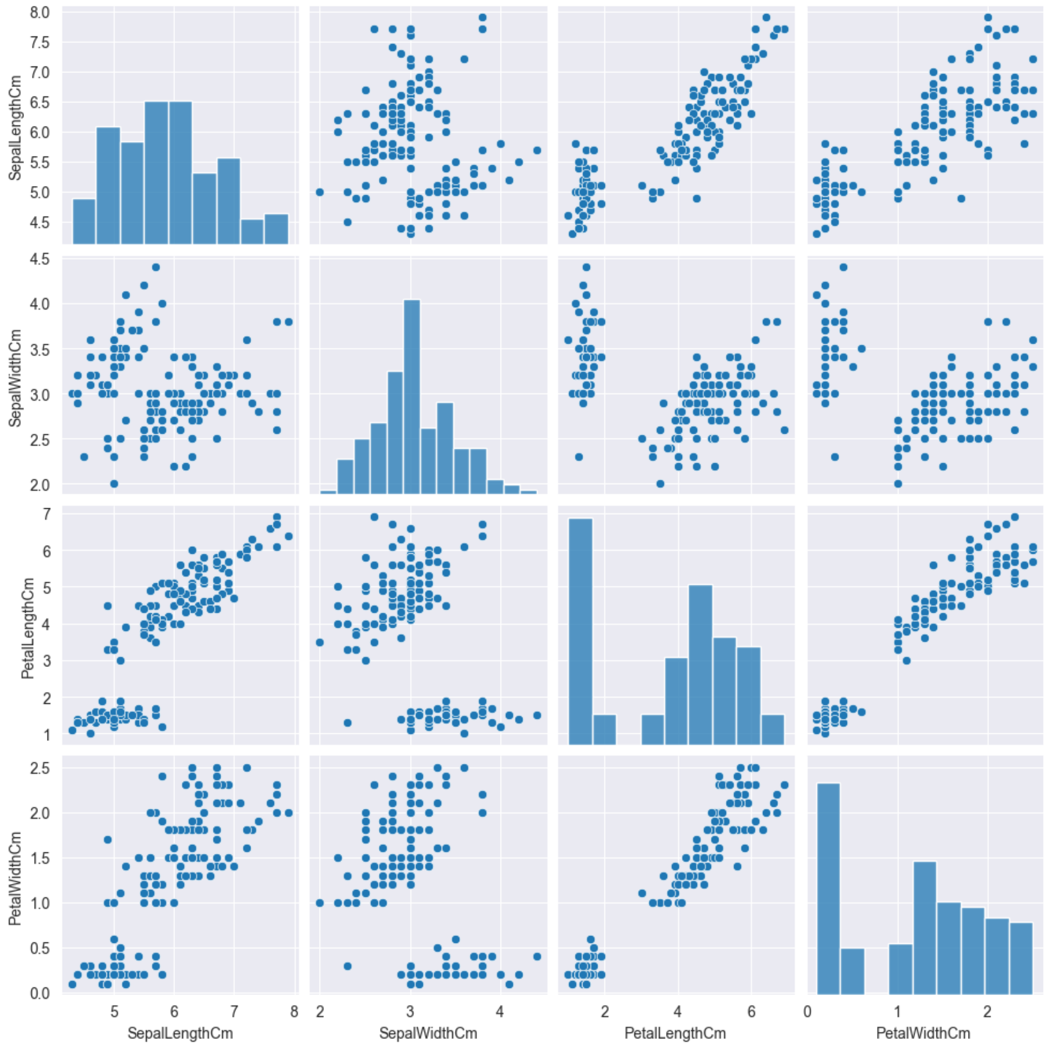

Figure 1 – All attributes plotted against one another

Based on the graph above, I wanted to graph attributed that would have some less obvious clusters. For example, the PetalWidthCm (X) and PetalLengthCm (Y) graph has two obvious clusters, one in the bottom left and one in the top right. The attributes I chose to cluster were the SepalLengthCm and SepalWidthCm attributes on the X and Y axis respectively.

Part B

Perform Cluster Analysts using the k-means, Agglomerative, and DBSCAN clustering methods on the dataset.

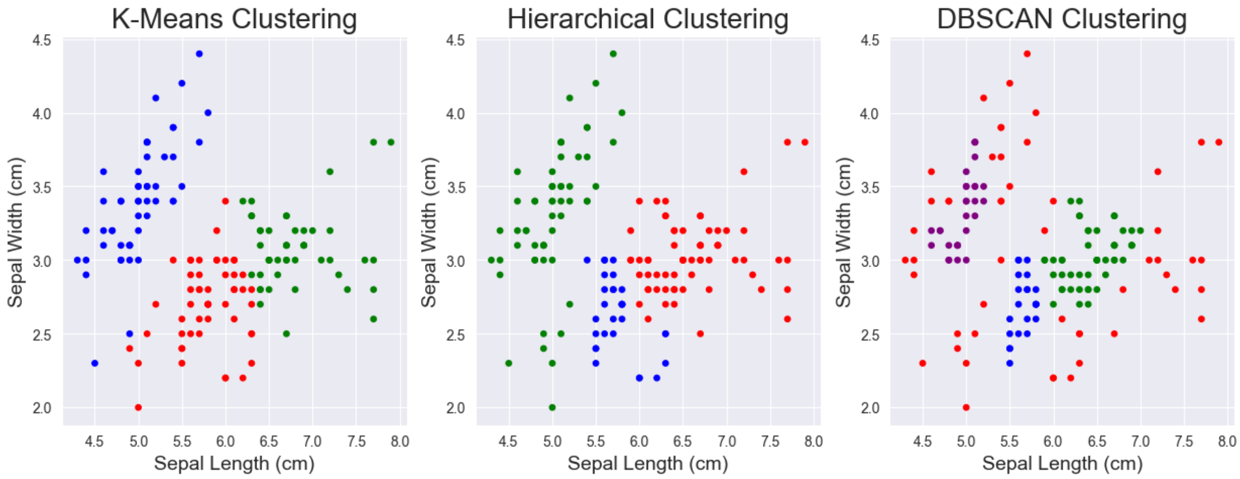

Figure 2 – K-Means, Agglomerative, and DBSCAN Clusters

The process for clustering each set of data was the same as each method used in class, creating an instance of the class made for the clustering method and using the ".fit()" function on the dataset. Three clusters were chosen for the k-means and agglomerative clustering, as 2 of the clusters were easy to see with the naked eye and would be the same between both methods. By choosing 3 clusters, the differences between the two clustering methods would be seen.

DBSCAN does not use the number of clusters parameter, but rather an "eps" parameter, used to get the distance to other points a reference point could consider in its cluster, and the "min_samples" parameter for determining how many points need to be nearby a reference point for them to be clustered together. After a period of trial and error, I had settled on the values of .15 for the eps parameter and 5 for the min_samples parameter. These two parameters seem to do a fair job of detecting outliers and clustering the data, as increasing the distance would change where the line is drawn by the blue and green clusters. In the DBSCAN data, the red cluster is reserved for outliers and should not be considered an actual cluster.

Part C

Identify which clustering method and parameter combination produced the best cluster separation.

Based on the graphs used above for each cluster, I believe that DBSCAN produced the best cluster separation. DBSCAN and Agglomerative clustering both have similar clusters in the larger bottom right section of the graph, with the line drawn between the clusters being nearly the same line. In addition, DBSCAN had also filtered for the outliers, giving cleaner more congested clusters rather than broad sweeping clusters that cover larger segments of data. Had more data been present, it's possible that Agglomerative clustering would have produced the best cluster separation.

Part D

Evaluate if any clustering method failed to produce meaningful clusters and explain why.

Based on the separation between the red and green clusters in the k-means clustering from figure 2 being drastically different than the other two methods used, k-means clusters would not be the best separation. While I do not think that k-means had necessarily "failed" to produce meaningful clusters, the clusters that k-means did produce could have been far cleaner. The imaginary line drawn between the red and blue clusters on the k-means graph is a near identical separation as the green and blue clusters on the agglomerative clusters, save for a few outliers, which keeps the k-means clusters from being considered a failure.

Question 2

Prove that the DBSCAN clustering method works best on the above data over k-Means and Agglomerative clustering methods.

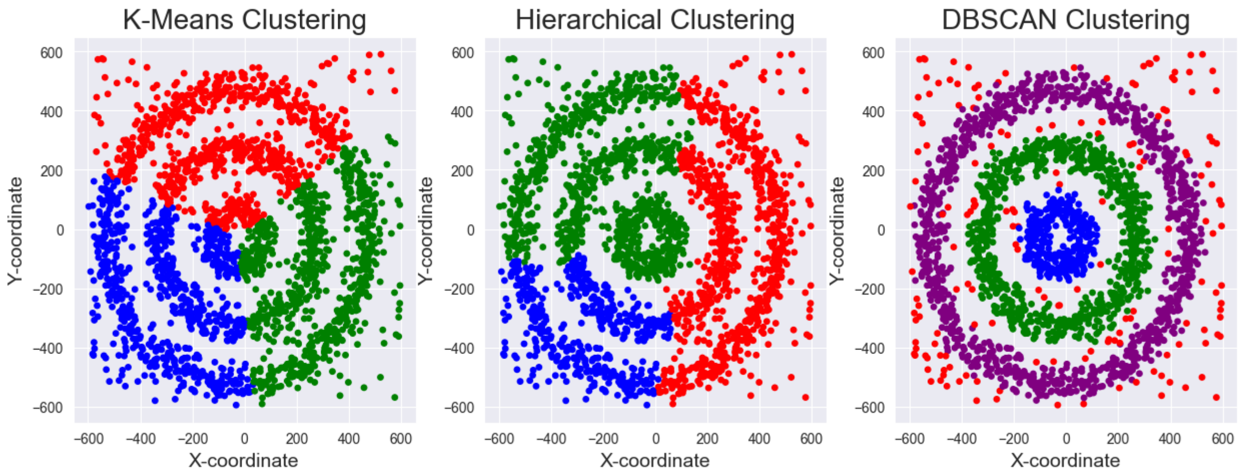

Figure 3 – Concentric Circle data clustered by K-Means, Agglomerative, and DBSCAN methods

The DBSCAN clustering method was able to cluster the concentric circles together, while the other two methods were unable to do so. The k-means method ad simply divided the cluster into thirds, while the agglomerative cluster had gotten the central circle in one cluster, but the rest of the clusters are in 25%, 25%, and 50% sections.

The DBSCAN clustering method had taken trial and error in order to get the parameters right, however once chosen reliably clusters the data into the concentric circles and outliers. Because of this, the DBSCAN clustering method is preferred for the above data than either the k-means or the agglomerative clustering methods