Unsupervised Clustering of Coffee by Flavor Scores

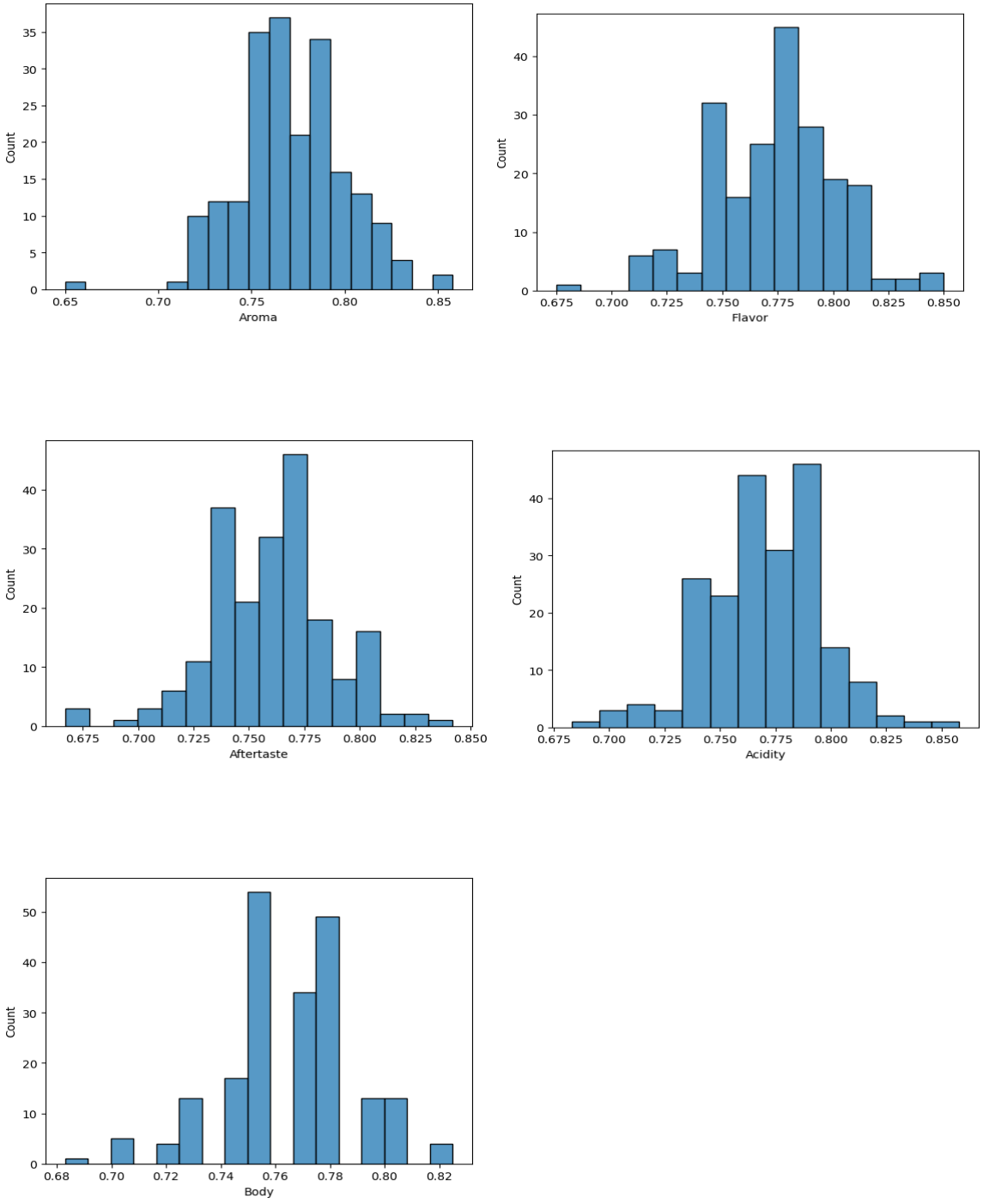
Objective:

The goal of this project is to determine how well different coffees can be clustered based on their flavor scores. By determining the optimal amount of clusters, we can determine if and how different coffees can be separated into similar taste groups. This can be useful for finding coffees that are similar to someone's favorite.

Data:

The data set used for this project was found on Kaggle. In this data set, each batch of coffee beans are given scores based on their taste. The scores include aroma, flavor, aftertaste, acidity, body, and balance. Sweetness and clean cup were included in the original data set, but were both the same value for all coffees. These attributes were scored on a 1-10 scale, and were scaled to 0-1 for training the models. All scores were approximately normally distributed (Fig 1). Lastly, an overall score for each coffee was included in the original data set, but is likely a composite score since it did not improve any of the models performance, so it was dropped.

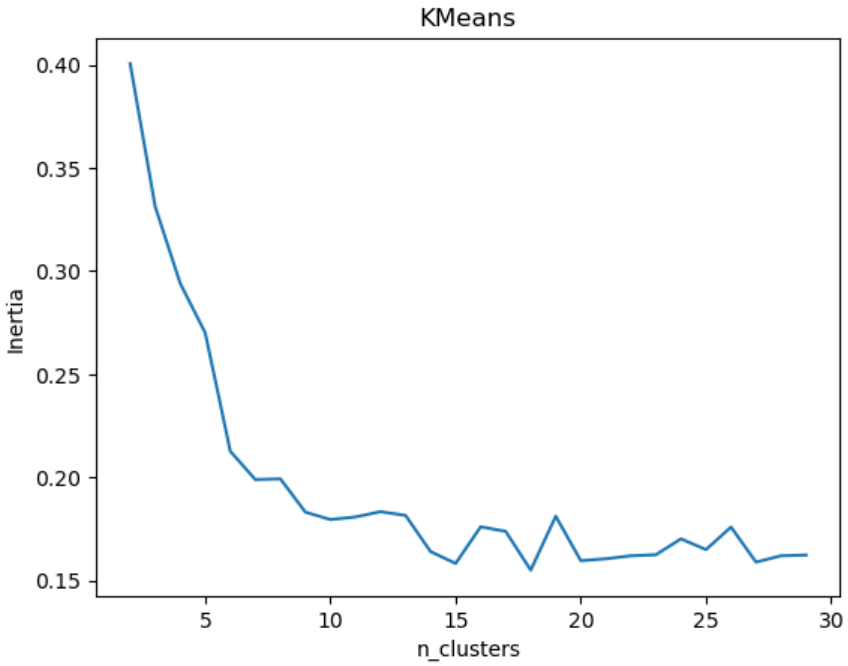
Figure 1



Models:

Three clustering models were used: KMeans, Agglomerative Clustering, and Gaussian Mixture. All three models were evaluated using silhouette and rand scores on clusters (KMeans and Agglomerative) and components (Gaussian Mixture) from 2 through 29. For KMeans, the inertia elbow method was used to find the optimal amount of clusters to be 6 (Fig 2).

Figure 2



Both the silhouette and rand scores seem to confirm that 6 clusters is optimal (Fig 3-4).

Figure 3

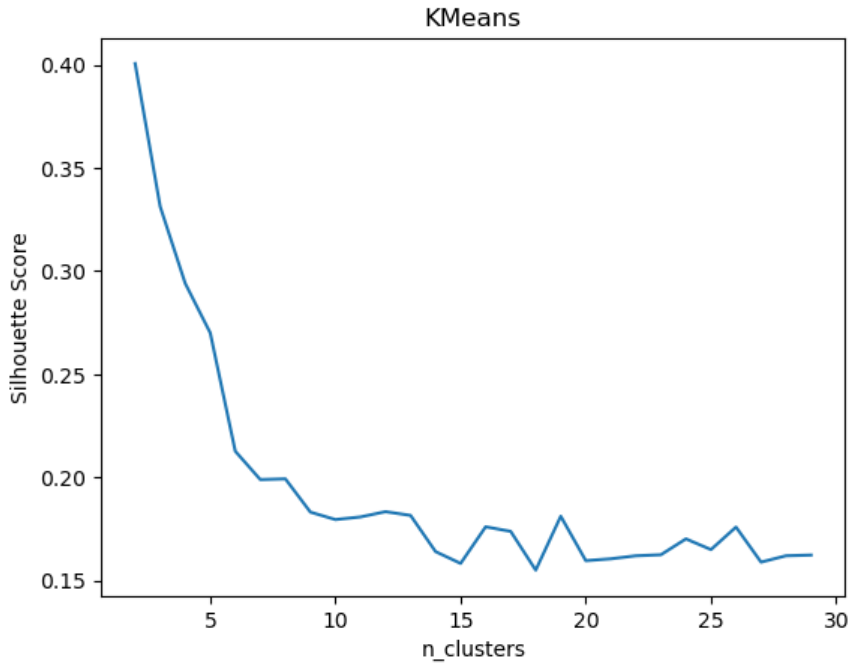
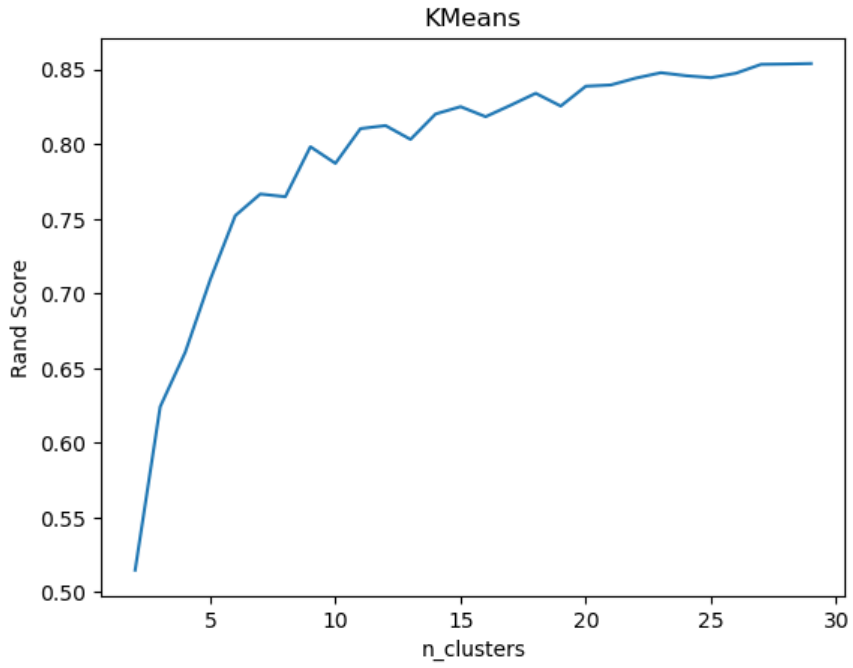


Figure 4



For Agglomerative Clustering, it seems that the optimal amount of clusters is approximately 5, which coincides with KMeans (Fig 5-6).

Figure 5

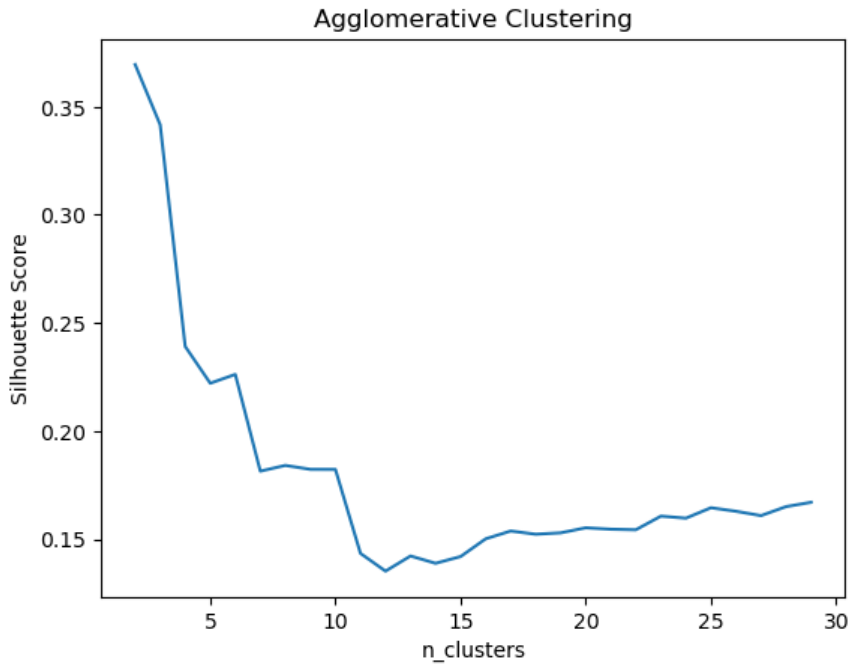
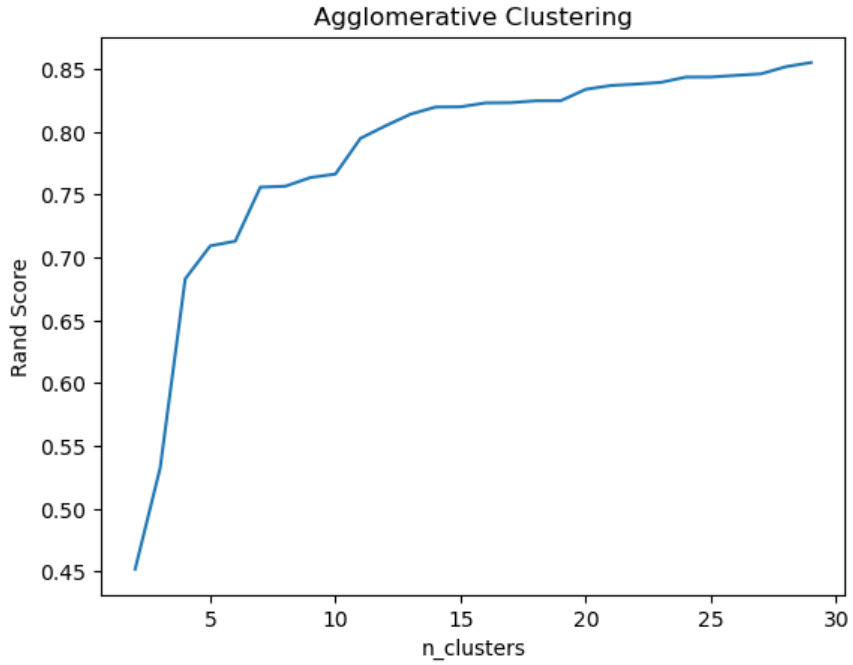


Figure 6



Lastly, for the Gaussian Mixture model, the silhouette score suggest components between 5 and 8 while the rand score suggest 5 to 9 components (Fig 7-8).

Figure 7

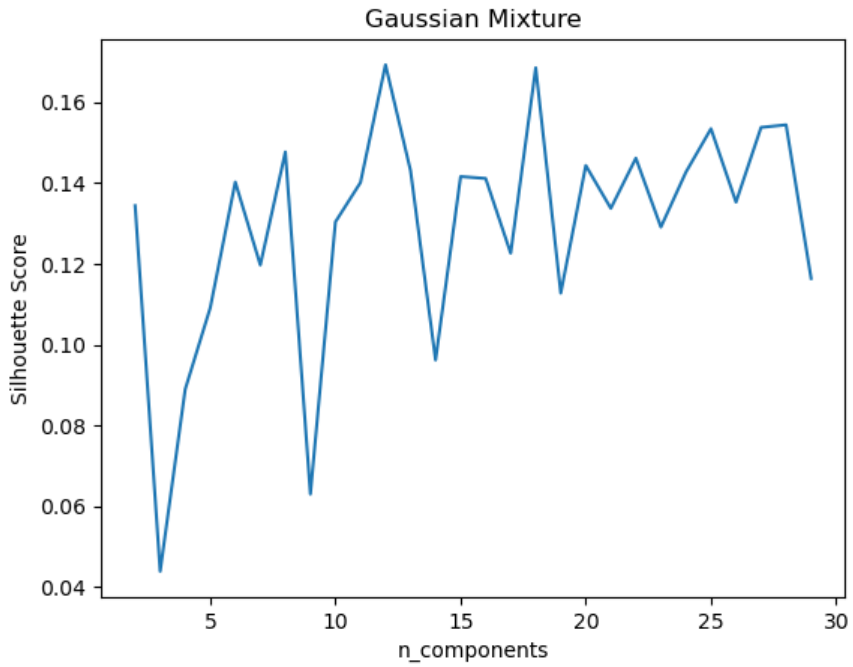
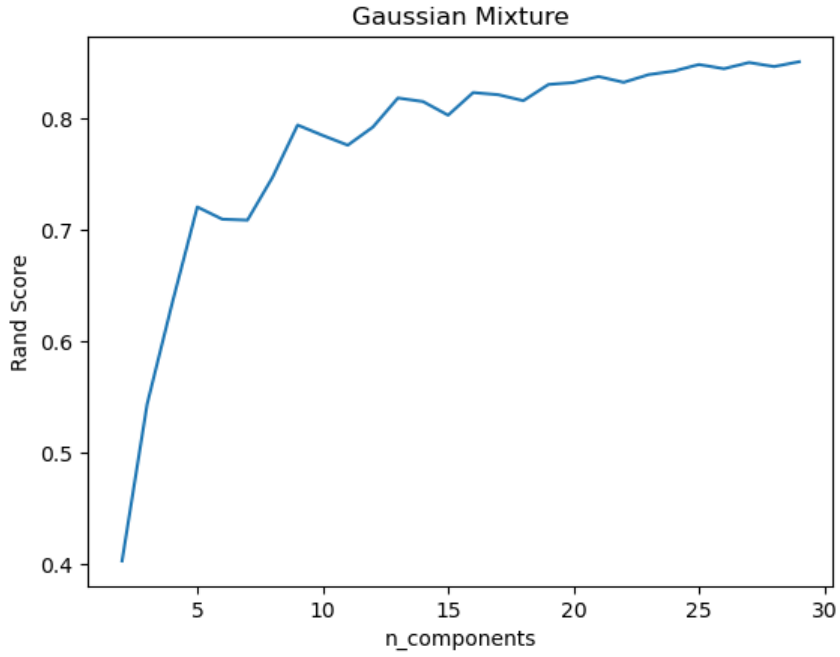


Figure 8



Summary:

From the evaluation of these models, we can determine that the coffee beans can be split into approximately 6 groups of similar taste. One problem that could be affecting how well the coffees can be clustered is the general overlap in flavor scores, as the range of scores is not very large (Fig 1). Perhaps other metrics, including flavor notes or more scientific attributes, acidity for example, could be better suited for separating the coffee groups.