Coffee Variety Clustering

Background

For the dataset required for the clustering problem, the “Coffee Quality Dataset” [1] from Kaggle was used. This dataset was taken from the Coffee Quality Institute (CQI) which is a non-profit organization that aims to increase the quality of coffee around the world. As such, this dataset contains various information on different coffee beans and their varieties worldwide. This dataset is interesting as processing it can give insights into what qualities of coffee affects its flavor and taste the most and even find interesting information like does altitude affect how good a coffee will taste etc.

Some of the important attributes that will be looked at in dataset are the following:

Altitude	Altitude in feet where the coffee beans were grown.
Aroma*	The scent or fragrance of the coffee.
Flavor*	The flavor of coffee is evaluated based on the taste, including any sweetness, bitterness, acidity, and other flavor notes.
Aftertaste*	The lingering taste that remains in the mouth after swallowing the coffee.
Acidity*	The brightness or liveliness of the taste.
Body*	The thickness or viscosity of the coffee in the mouth.
Balance*	How well the different flavor components of the coffee work together.
Uniformity*	The consistency of the coffee from cup to cup.
Clean Cup*	The score of a coffee that is free of any off-flavors or defects, such as sourness, mustiness, or staleness.
Sweetness*	Described as caramel-like, fruity, or floral, and is a desirable quality in coffee.
Total Cup Points	A total rating out of 100 of the Coffee.
Variety	The name of the coffee bean.

Attributes marked with * were mainly used, and these are called the sensory evaluations or the coffee quality scores.

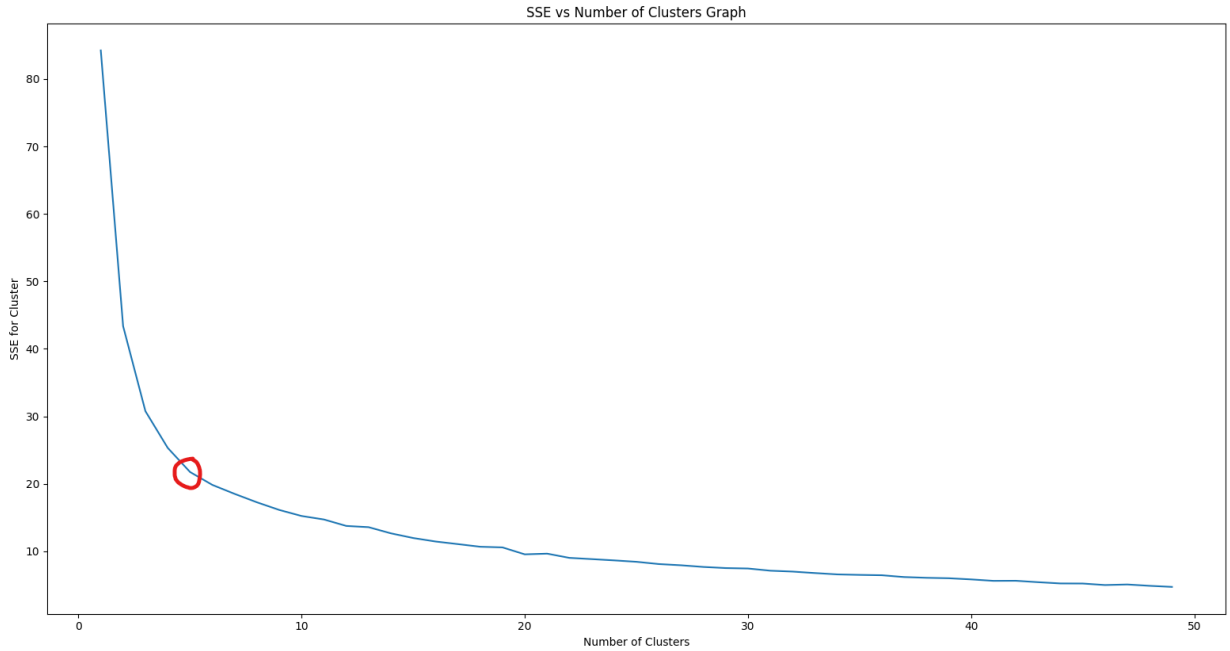
Methods

For the methods to analyze the dataset K-means clustering was used along with hierarchical dendrograms to find the various clusters that were made using the different sensory evaluations from the dataset (also known as the coffee quality scores). Afterwards, we evaluated how each of the features may or may not be related.
For K-means clustering we only look at the sensory attributes and evaluate the clusters that were created and see the percentage of each variety of the coffee bean in that cluster. Following K-means clustering, we also create 2 hierarchical dendrograms (min and max linkage) for the same 9 features and evaluate those clusters. And finally we will also look at if there is any relationship between the altitude a coffee bean is grown at and its Total Cup points using a single linkage hierarchical dendrogram.

For the pre-processing of the dataset, firstly we extract all the attributes we will use in our clustering (listed in the table above) and get rid of everything else. Afterwards, for all the values in the altitude column, if it is in a range it is converted into a single value by taking the average, if it is already a single value nothing is done, otherwise if it is anything else, then we convert it to a null value. This makes the altitude column easier to work with as now it is just integers or null values. Then we drop any rows with null values and remove rows where the coffee bean variety isn't known.

For K-means clustering the graph of sum of squared error vs number of clusters was created to find the knee of a curve or the elbow point of the best number of clusters that will represent the dataset.

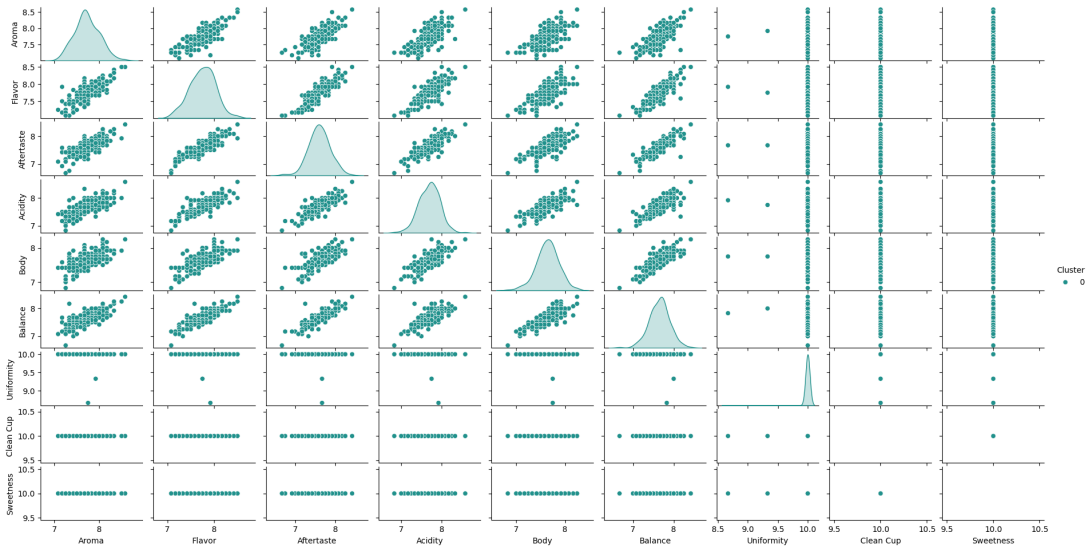
In our case, we got the following graph:



Looking at the graph, I picked 5 as the number of clusters using the elbow method.

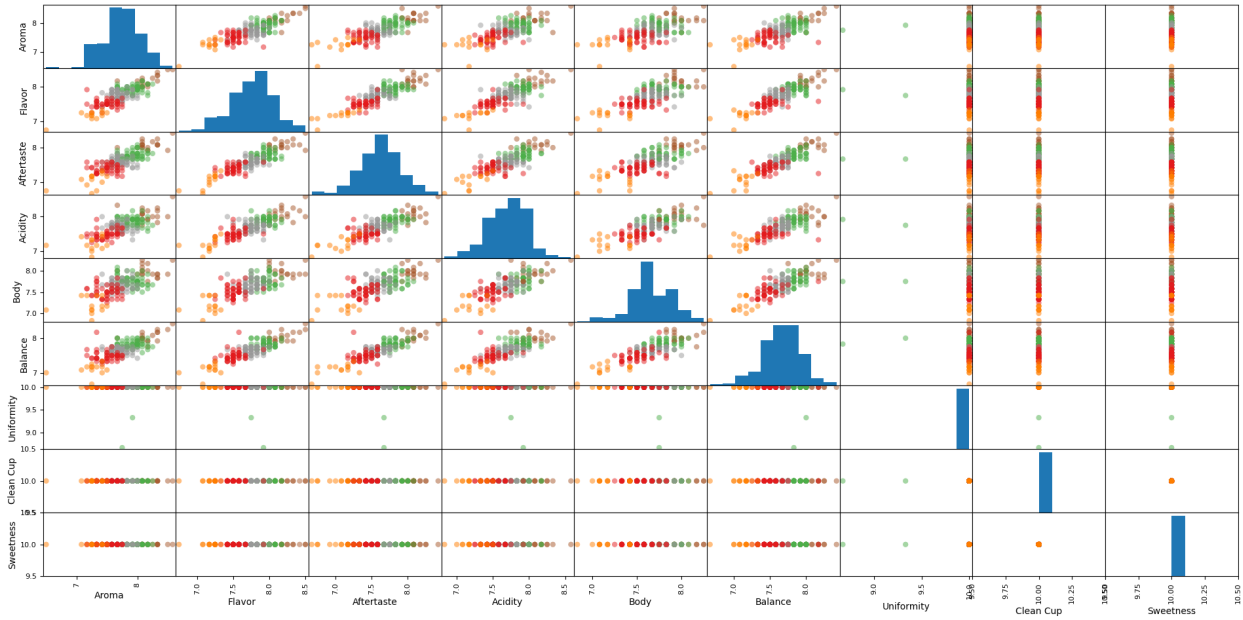
I had also attempted to use DBSCAN for my dataset [2][3]. However, after experimenting with it, I had noticed that the dataset had a somewhat uniform density where DBSCAN considered the entire dataset as

a single cluster. The image below shows my attempt with DBSCAN using the sources [2] & [3] which turned out to not be meaningful. The code is kept in the script but is commented out.



Results

For the K-means with 5 clusters, we had the following scatters plots. [4]



As expected from the clusters created with the various scatter plots of different attributes we see that there is generally positive correlation between all the attributes, suggesting that a coffee created from a coffee bean with a high aroma score will also tend to be high in other score, in other words all the sensory evaluation attributes have a strong positive correlation.

However, looking at the Variety of the coffee bean in each attribute is the most interesting as it shows which variety of coffee beans are the most similar in coffee quality scores.

```
Cluster 1 Varieties:
    Catuai              Count:7   , Percentage:15.91%
    Caturra             Count:6   , Percentage:13.64%
    Bourbon             Count:4   , Percentage:9.09%
```

```
Cluster 2 Varieties:
    Gesha               Count:15  , Percentage:28.30%
    Typica              Count:9   , Percentage:16.98%
    Caturra             Count:5   , Percentage:9.43%
```

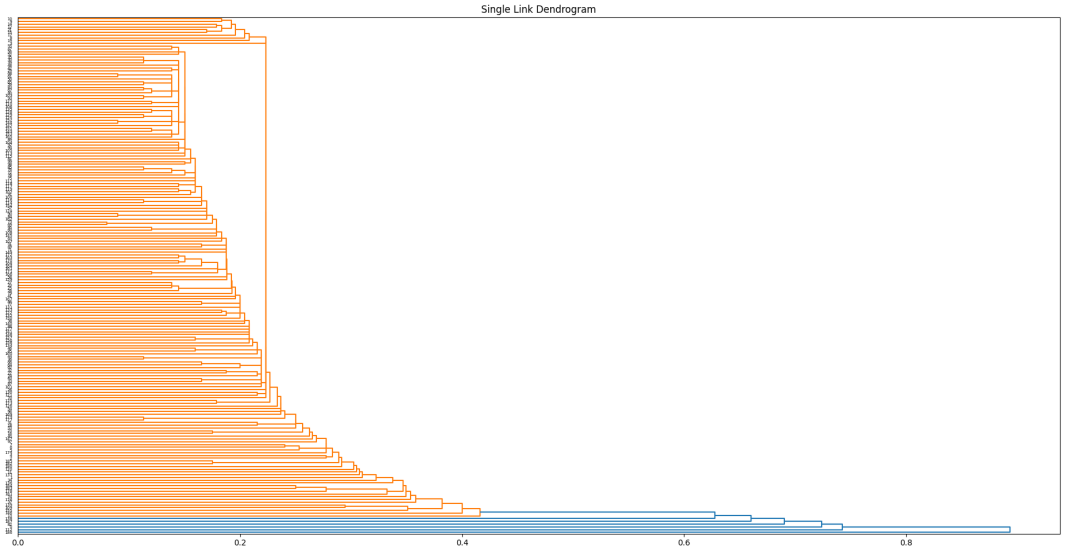
```
Cluster 3 Varieties:
    Caturra             Count:4   , Percentage:30.77%
    Catimor             Count:2   , Percentage:15.38%
    Typica              Count:2   , Percentage:15.38%
```

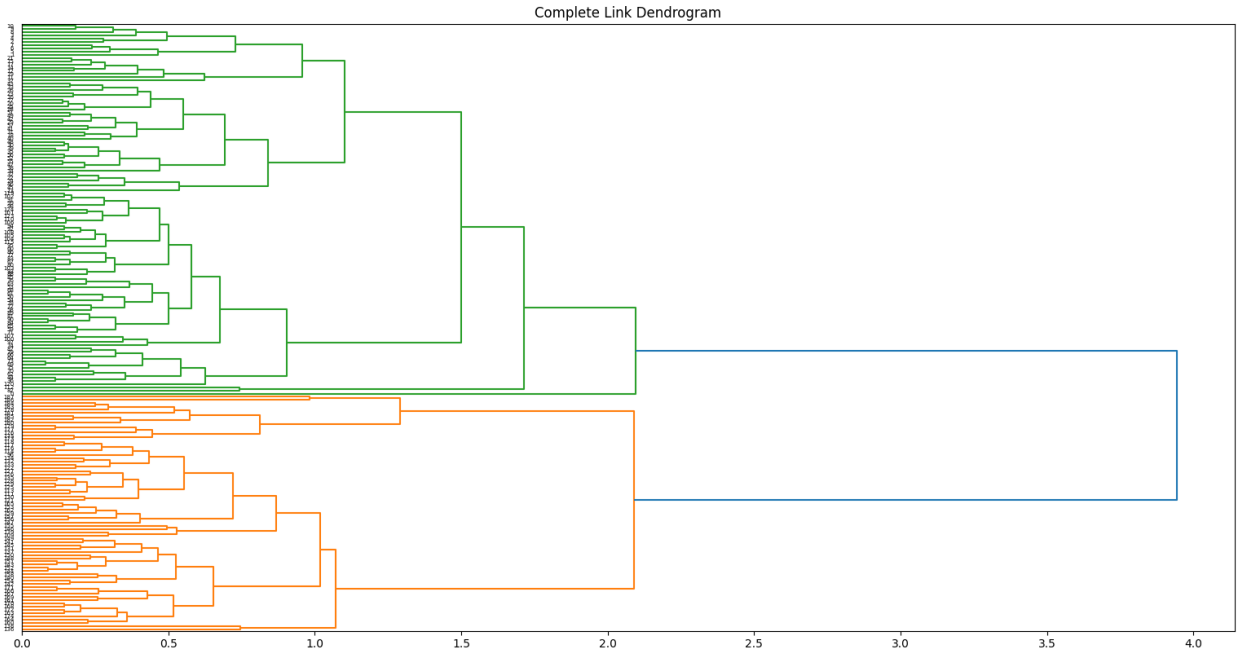
```
Cluster 4 Varieties:
    Gesha               Count:8   , Percentage:42.11%
    SL34                Count:2   , Percentage:10.53%
    Bourbon             Count:2   , Percentage:10.53%
```

```
Cluster 5 Varieties:
    Bourbon             Count:9   , Percentage:15.25%
    Caturra             Count:9   , Percentage:15.25%
    Typica              Count:8   , Percentage:13.56%
```

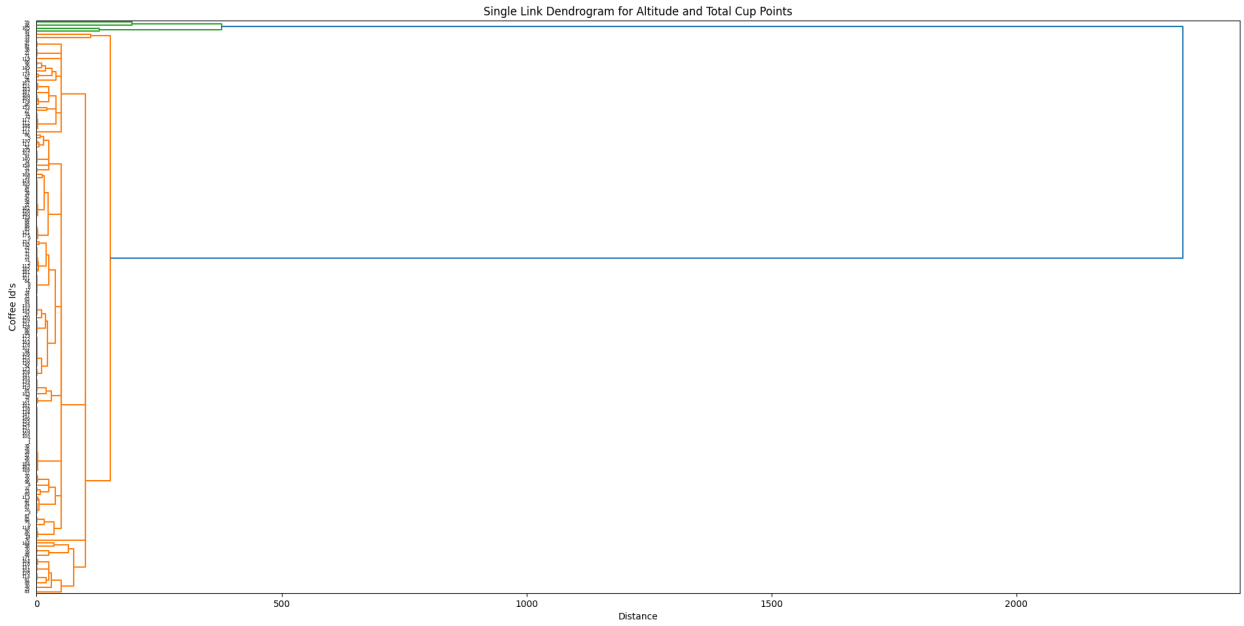
Looking at the top 3 highest counts of Coffee variety in each cluster we see which coffee beans are the most similar to each other in terms of the sensory evaluation attributes.

For the Hierarchical Dendrograms we have the following graphs:





For the single link dendrogram or the min distance, we see that the coffee attributes that are the most similar have links between them created first and that there are mainly only 2 clusters that are created with most of the coffees being in the orange cluster. However, for complete link or max distance dendrogram we see that there are 2 clusters that are of similar size being created. This is because the complete link uses the furthest 2 points of the cluster to determine the distance between them. It also shows the dissimilarity between the coffees as the distance increases in the graph.



For the min distance dendrogram for altitude vs total cup points we see that there are mainly 2 clusters,

looking closely at how closely grouped the coffee clusters are we see that there isn't a strong connection between how good a coffee is vs the altitude the coffee bean was grown at which is interesting to know.

Conclusions

Looking at the dataset and the results, there are many interesting findings which would be helpful for coffee enthusiasts or businesses that deal with coffee to know about. From our observations in the k-means clustering we see which coffee varieties are the most similar to each other in terms of taste, smell, aftertaste, etc. for all the coffee quality scores, that is if for example if you find the coffee made from the Gesha bean you might also like the coffee made from Typica. From the dendrograms we can see which coffees are closely related to each other since if the distance between is small they are very similar in sensory evaluation metrics & how the hierarchy of various coffees are merged. This sort of in depth information on the tastes or quality of coffee can be particularly useful for coffee shops as they can use this information for quality assurance or menu planning.

References

[1] Boyar, F. (2023, May). Coffee quality data(CQI May-2023). https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi

[2] Dbscan algorithm clustering in python. (2021). Engineering Education (EngEd) Program | Section. Retrieved April 12, 2022, from https://www.section.io/engineering-education/dbscan-clustering-in-python/

[3] Bedre, R. (2020, March 23). DBSCAN clustering algorithm in Python (With example dataset). RS Blog. https://www.reneshbedre.com/blog/dbscan-python.html

[4] Yerra, B. (2018, February 19). Centroid-based clustering k-means algorithm. Retrieved December 6, 2023, from https://mlbhanuyerra.github.io/2018-02-19-Clustering-K-means/