ECM3420 Coursework Report

1 Introduction

Wine has been enjoyed for around 8000 years [3]. It's early methods of production have been improved to the present day. In 2019, the global wine market size was 364.25 billion dollars, it is widely consumed in western culture in different ranges of quality. Wine premiumization has promoted the wine market growth and therefore wine sold on the basis of its high quality has become more important [2].

2 Analysis objective and plan for interpretation

The quality evaluation of wine is mostly certified based on physicochemical as well as sensory tests [1]. Sensory tests are carried out by human, "wine connoisseurs". Whereas physicochemical data of wine, such as alcohol content or pH levels. Physicochemical data can be stratified and analysed to interpret what attributes of wine best determines its quality, the main objective of this project. In order to interpret this data, it must be analysed such that it can be seen what attributes are the least varied, least anomalous and most importance in terms of the wines quality as well as assessing the correlation between different attributes. Possible difficulties with the data are the fact that the wine's physicochemical properties may not individually affect the quality but in combination, does. Removing seemingly useless data may have a knock on effect on the accuracy of the data. The way to avoid this is rigorous testing of the data with and without data that may be considered for removal. Then the data will be scaled or normalized for it then to be clustered by either of the two algorithms being compared in this project.

3 Data set description and a summary of its attributes

In the dataset, winequality-red.csv, 11 physicochemical attributes are presented along with a quality rating out of 10 for 1599 different samples of wine. The 11 physicochemical attributes include:

- Fixed acidity: Otherwise known as 'non-volatile' acidity. The amount of acids that do not readily evaporate such as tartaric, malic, citric, and succinic acid. Except for succinic acid, these originate in grapes. High acidity can lead to a sour taste [4]. Volatile acidity: The amount of acetic acid in wine. High volatile acidity can lead to an unpleasant, vinegar taste [5].
- Citric acid: Found in small quantities, citric acid can add 'freshness' and flavour to wines [5]. Due to this also being a fixed acid, I predict there will be a high correlation between citric acid and fixed acidity when I execute a heatmap on the data during data exploration.
- Residual sugar: The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/litre and wines with greater than 45 grams/liter are considered sweet [5]. The dryness of wine is based on the amount of residual sugar with 'bone-dry' has the least amount of residual sugar. As this is more of a preference, I predict it will not strongly affect the quality rating of the wine.
- Chlorides: The amount of salt in the wine. High salt content is undesirable.
- Free sulfur dioxide: The free form of sulphur dioxide exists in equilibrium between molecular sulphur dioxide and bi-sulphite ion which prevents microbial growth and the oxidation of wine. [5].
- Total sulfur dioxide: amount of free and bound forms of sulphur dioxide; in high concentrations, sulphur dioxide is noticeable in the wine's scent and taste [5].

- Density: The density of water is close to that of water depending on the percent alcohol and sugar content [5]. I predict that this will, therefore, have a high correlation with the alcohol and residual sugar attributes in the correlation heatmap.
- pH: Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale [5]. I predict this will directly correlate to the other acid based attributes.
- Sulphates: The amount of sulphates which can be added to wine to contribute to sulphur dioxide levels, which acts as an antimicrobial and antioxidant [5].
- Alcohol: The amount of alcohol in the wine.

4 Summary of data exploration and actions taken for data cleaning and feature engineering.

Given these attributes, I explored what this data was and how it would relate with each other whilst having the main objective of the project in mind. For quality, the attributes that have the highest positive correlation are alcohol, sulphates, citric acid. The highest negative correlated was volatile acidity, density and total sulphur dioxide. The least correlated was residual sugar, chlorides, free sulphur dioxide, pH and fixed acidity.

The attributes with the most outliers are residual sugar, chlorides and sulphates. To find if removing any attributes is making a positive or negative impact, I run the clustering models, covered in the next section to collect their homogeneity and completeness scores. The homogeneity score of a model is a value given from 0.0 to 1.0, where the higher the score the closer the model is to having all of its clusters to contain only data points which are members of a single class of the target attribute. The completeness score of a model is a value given from 0.0 to 1.0, where the higher the score, more data points of a given class are members of the same cluster. Therefore these two score communicate how well the clustering is executed on the data. Applying this on my data as I am changing it is a quick indicator that it was a good decision, given the data exploration beforehand. As residual sugar and chlorides are not correlated to quality and have anomalous results with many outliers, these can be removed.

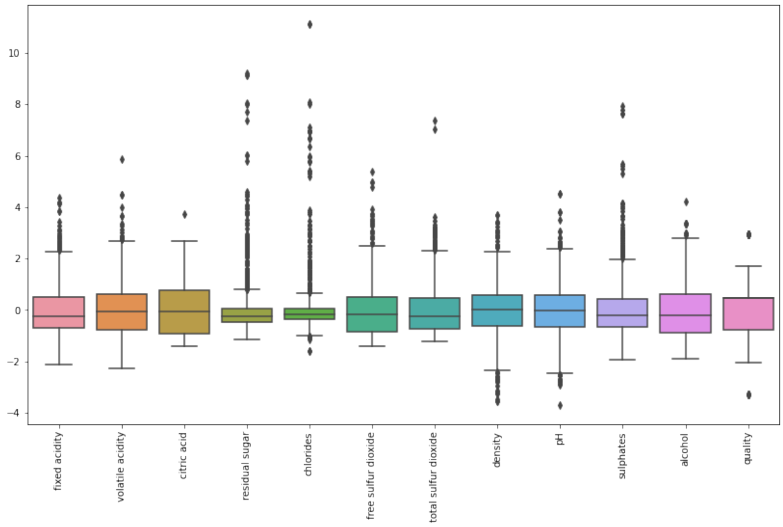

Figure 1: A histogram of each attribute in the dataset.

- Homogeneity score: 0.255899298137218 from 0.22969972607776024, therefore the clustering is closer meeting the cluster requirement.
- Completeness score: 0.17029420000391343 from 0.15468895241614275, therefore more of the members of a given quality are part of the same cluster.

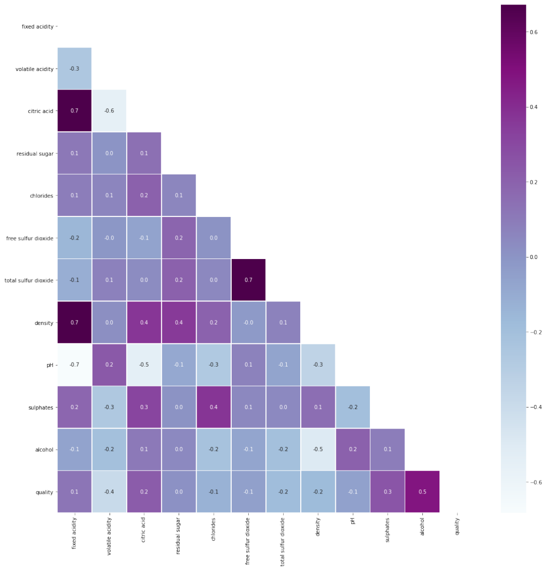

Figure 2: A heatmap of the correlation of each attribute in the dataset.

When I remove all attributes that are not correlated positively or negatively, in other words, I remove fixed acidity, residual sugar, chlorides, free sulphur dioxide and pH. Homogeneity score improved to 0.28996173881719073 and completeness score improved to 0.1951537133685739. Then to cut further weakly corelated attributes; density, total sulphur dioxide and citric acid. Improved homogeneity score to 0.3734270040047973 and completeness score to 0.2582199407375768. As sulphates had a considerable amount of outliers, removing that gives:

- Homogeneity score: 0.37563178851003587
- Completeness score 0.26077617518353335

The problem with removing this many attributes is the clustering starts to not accommodate the number of quality ratings in the silhouette score. A good middle-ground seems to be to remove fixed acidity, residual sugar, chlorides, free sulphur dioxide pH and sulphates. From my findings what surprised me the most is how pH was not a more correlating factor as pH level are a measurement of acidity or basicity

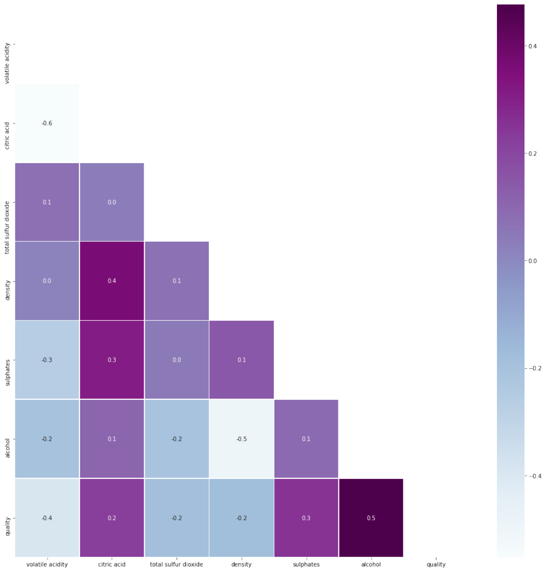

Figure 3: A heatmap of the correlation of each attribute, excluding fixed acidity, residual sugar, chlorides, free sulphur dioxide and pH, in the dataset.

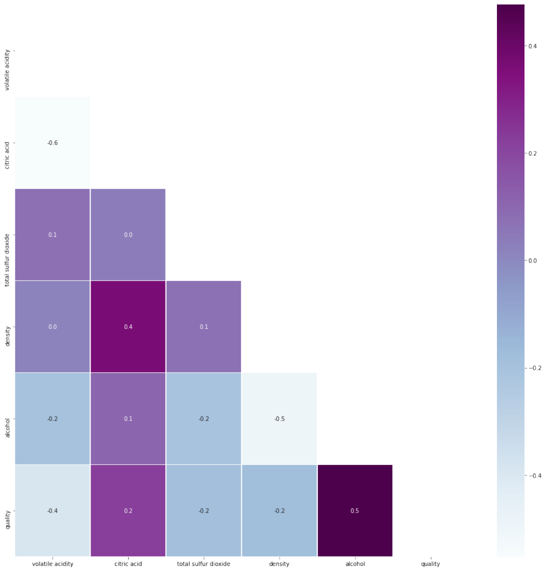

Figure 4: A heatmap of the correlation of each attribute, excluding fixed acidity, residual sugar, chlorides, free sulphur dioxide pH and sulphates, in the dataset.

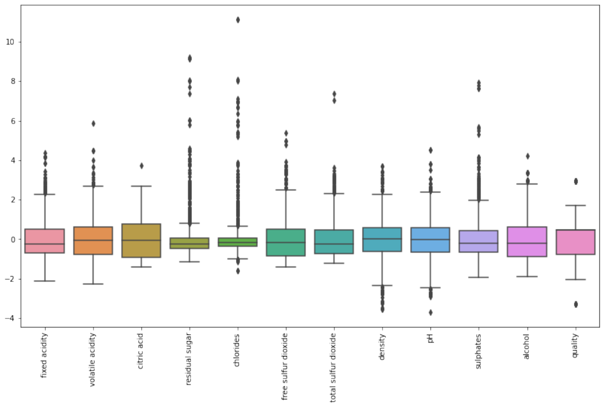

Figure 5: A histogram of each attribute, excluding fixed acidity, residual sugar, chlorides, free sulphur dioxide pH and sulphates, in the dataset.

of a sample. An acid wine will taste sour and would therefore should be ranked a lower quality. This gave me the impression that pH level would affect quality to a greater extent. The ‘quality’ column has 6 unique values; 3, 4, 5, 6, 7 and 8. I produced a bar graph for this column, figure 6, that shows that wine samples were mostly of quality 5 and 6. This may unfortunately make the clustering of the data harder as there should be two very large clusters with other very small ones. This also means that it will make it harder to cluster data together that are from the smaller quality classes as there is less data to cluster with each other. I will keep this in mind when coming across any limitations in the clustering models.

4.1 Engineering the data for clustering

Before applying the first scaling model, Kmeans, the data needs to be scaled. Due to me clustering an attribute that ranges from 3 to 8 in increments of 1, this needs to be to scale with other attributes that are float numbers that range drastically in decimal point increments. The magnitude of the attributes can be made equal with scaling so it is easily processable for clustering and no size is bias towards any attribute. This is also similar for before I apply hierarchical clustering. I will normalize the scale of attribute values for the same reason. For both models, I employ principal component analysis ,PCA, on the dataset before clustering. PCA reduces the dimensionality of the data when a dataset has too many variables to preform clustering on. The data cleansing an exploration still matters as PCA takes the given data and reduces its size with a minimal loss of information. At the end it will give two sets of

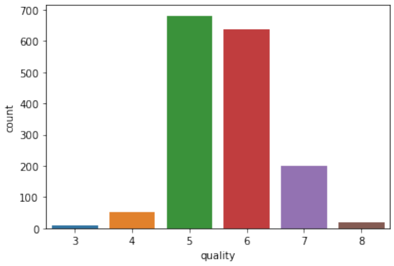

Figure 6: A bar graph showing the the amount of wine samples belonging to each rating of quality in the dataset.

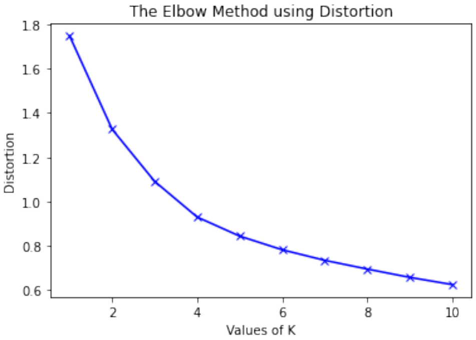

Figure 7: Graph presenting distortion against values of K.

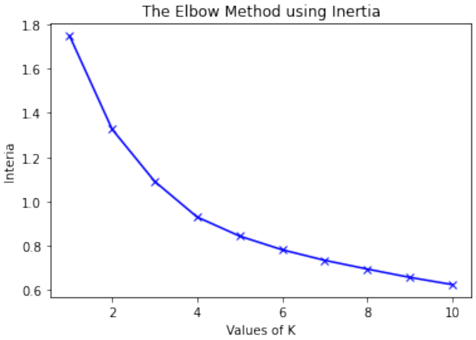

Figure 8: Graph presenting inertia against values of K.

data that a clustering model can use.

5 Model 1: K-means

In K-means clustering, clusters are produced through an iteration where the K-means algorithm chooses a random K points from the dataset, clusters the remaining points arounds the K points which become the centroids, the mean of each cluster is calculated and finally it re-clusters based on the new means. The iteration stops when clusters stop changing.

Before K-means can be executed on the data, the value of K must be decided. As stated earlier, the main objective of this project is to use the physicochemical attributes in wine to interpret its quality through clustering. Earlier from exploring the data I had found that the 'quality' column has 6 unique values; 3, 4, 5, 6, 7 and 8. Although it is a score that is meant to be from 1 to 10, as the concerned data only deals with qualities of these values, 6 clusters would be desirable for this data. Nevertheless, I must measure the data itself to find what clustering would be best for K means, finding the optimal value of K. Two elbow plots are used for this data look at K clusters, taking K as a value from 2 to 10. The first plots K against distortion; the average of the squared Euclidean distances from centroids of respective clusters. The elbow on figure 7 seems to indicate the value of K to be 3. The second elbow, figure 8 plots K against inertia; the sum of squared distances of samples to their closest centroid which seems to, also, indicate the value of K to be 3. I executed a silhouette analysis for K-means clustering iterations from K=2 to K=8.

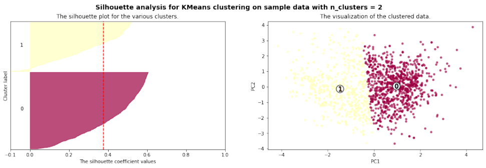

Figure 9: Silhouette analysis and Clustering where K=2

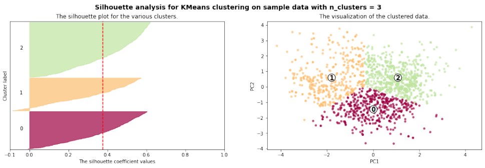

Figure 10: Silhouette analysis and Clustering where K=3

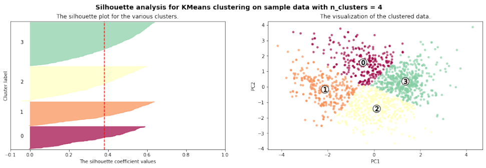

Figure 11: Silhouette analysis and Clustering where K=4

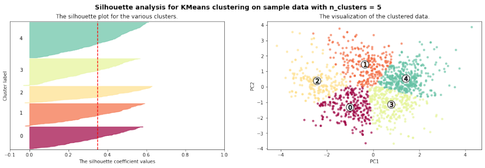

Figure 12: Silhouette analysis and Clustering where K=5

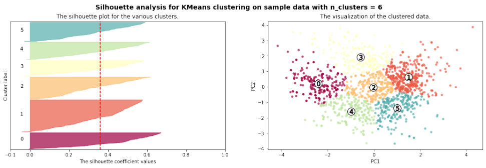

Figure 13: Silhouette analysis and Clustering where K=6

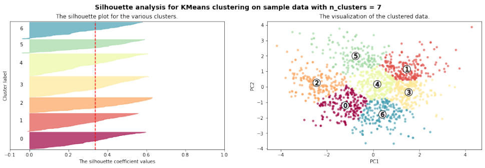

Figure 14: Silhouette analysis and Clustering where K=7

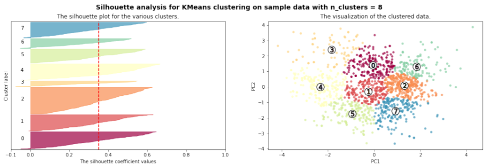

Figure 15: Silhouette analysis and Clustering where K=8

In figure 9-12, the thickness of clusters varies but all are beyond the average score. Figure 13 has a consistent thickness for each clusters and are all beyond the average score. Figure 14-15 have all clusters beyond the average score. Attributes in figure 14have similar thickness clusters but the thickness of clusters in figure 15 becomes inconsistent. As K=6 and K=7 are similar, I ran the K-means algorithm and compared the homogeneity score and completeness score.

- K=6
  - Homogeneity score: 0.296684221928499
  - Completeness score: 0.19987782063209922
- K=7
  - Homogeneity score 0.2923227748717939
  - Completeness score 0.17902337511119342

Therefore, I picked K=6. The elbow and the silhouette findings disagree. Silhouette is viewed as the more accurate model, in addition, it also agrees with 6 clusters for the 6 quality values. The scatter graph, figure 16 of 6 clusters is well shaped compared to the other graphs of K clusters. I reproduced the K-means cluster graph with the 6 clusters for analysis. The clusters of data are not very easily separated and the clusters all look of comparable sizes, which referring back to the bar graph that showed the amount of samples belonging to each measure of quality, this should not be the case. This is hence why the homogeneity score and completeness score are very low.

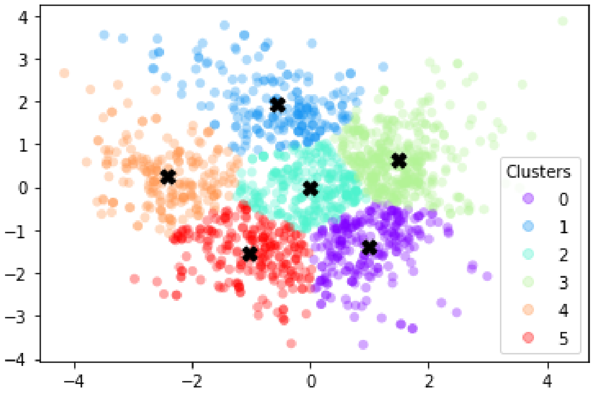

Figure 16: Scatter Graph is split into 6 clusters using K-means to represent the quality attribute in the dataset.

6 Model 2: Hierarchical Clustering

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. Hierarchical clustering groups attributes based on the similarity of the samples. Therefore most similar quality ratings will be merged and clustered together. Therefore, I naturally expect to see the hierarchical clustering produce a scatter plot with less clusters so it is therefore more general with the classification of a wine's quality. The dendrogram visualised, figure 17 from the data branches very inconsistently. Each pair of branches are different sizes from one another, therefore it makes it difficult, from this data, where to take influence for the number of clusters. A silhouette analysis is also employed before choosing the number of clusters. The amount of clusters with the highest silhouette score, in figure 18, is 2, which in the dendrogram, had the least dramatic difference in branching. Then I produced, using the information from the dendrogram and the silhouette analysis, an agglomerative cluster graph with 2 clusters, see figure 19. There is one cluster that is considerably larger than the other, which was expected from the dendrogram and the fact that it is grouping the similar qualities and there was two instances of the quality attribute that were far more common than the others, figure 6.

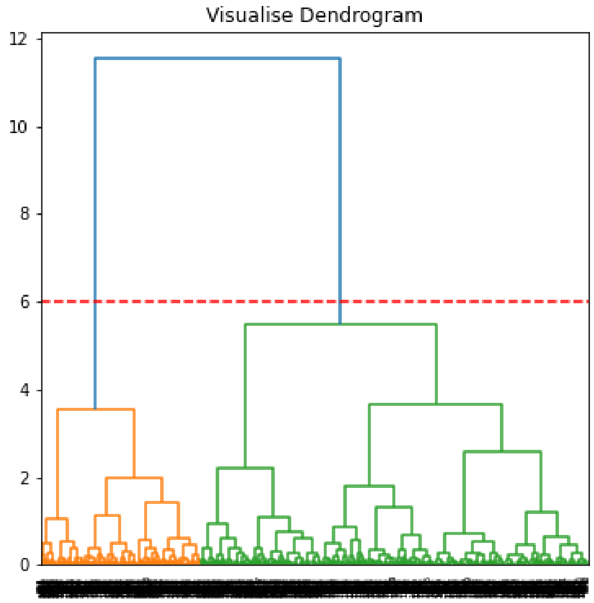

Figure 17: Dendrogram to represent the hierarchical relationship between attributes in the dataset.

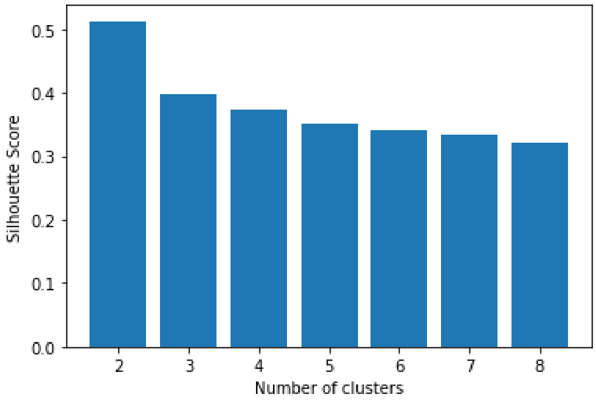

Figure 18: Bar graph presenting the silhouette scores from 2 to 8 clusters.

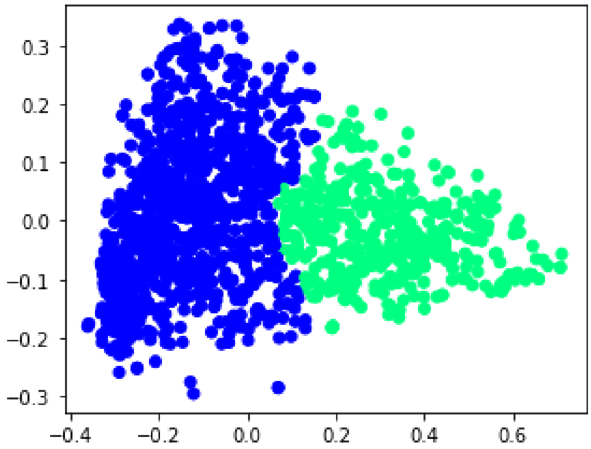

Figure 19: Agglomerative cluster graph used to represent the relationship between the quality attribute and the data.

7 Key Findings and Comparative Analysis

Both models have served to partake in the analysis of a dataset of wine samples which contains information on each sample's physiochemical properties as well as its given 'quality' rating out of ten. The clustering models were used to class wine by the quality values to find what makes a high quality wine. The K means clustering model produced a very explainable clustered scatter graph, where it clearly clustered to the six quality values available in the given dataset , figure 6. For this case the clustering algorithm successfully took the wine data set and grouped the data points into what looks like the quality values. Unfortunately, in terms of accuracy, this is not the case. Although six clusters were produced, the homogeneity score was 0.296684221928499 and the completeness score was 0.19987782063209922. These are small for these type of score for a model but I must keep in mind that it may be due to the data and these score will matter more when compared with the second model. The model also produced these six clusters with almost all equal sizes which does not follow the disparity between the quality value that shows up in the least data points and the one that shows up in the most. Hierarchical Clustering produced data that didn't seem as representative to the main objective of the project than K means did. This became evident before the clustering graph was produced, when the dendrogram was visualised it did not produce a clear answer to what would be the optimal number of clusters, therefore the silhouette analysis was produced to help. When the graph was produced, the two clusters did not look of equal size, which stated earlier, is more in line with the nature of the data. Even though it has less clustering, the homogeneity score for this model is drastically more disappointing than K-means with a score of 0.019131604474653688 and similarly with a completeness score of 0.019131604474653688. This is a clear indicator that the accuracy of clustering in hierarchical clustering is drastically less than that with K-means. The more general presentation of the data did not turn out more accurate and this data cannot explain the main objective of analysis as the classification of data into cluster one or cluster two does not classify its quality. K-means works better than hierarchical clustering for when it is already known, such as in this case, how many clusters should be chosen but the group they belong to is unknown. Hierarchical clustering is used better for data with an unknown number of clusters as it is designed to find it, although, taking this case as an example, the dendrogram was not very clear.

8 Recommended model

I would recommend K means much more as a means to analyse this dataset through clustering as it represented in the data in terms of the 'quality' attribute in the clearest way; producing a number of clusters matching the number of classes of the target attribute and clustering the data accordingly with relative success in comparison to the hierarchical clustering method. In addition to this, as I have mentioned previously, K-means clustering is known to work best when the amount of clusters needed is already known and it is not imperative to be inferred, so therefore, in this project K-means does make more sense.

9 The next steps of analysing this data

A key problem with classifying the wine samples to their quality is the unbalanced quality data, as found in data exploration , figure 6, I have reason to believe that this problem had a large problem with clustering accuracy as there isn't enough data to cluster most of the classes of quality accurately. In the future, I would recommend to attempt to process the dataset using LDA to reduce the data for clustering; a supervised learning method that would work better on the labelled data. The grouping variable, quality can be taken and LDA aims to find the maximal separation between the groups of levels of quality. As PCA is unsupervised, LDA would be a much more preferred model to employ to cleanse the data before employing clustering. In addition to employing LDA, the data itself can be made more fair before processing; collecting more diverse data, level out the data so that the amount of samples of each quality class is more equal or group the lower end of quality and the higher end of quality into two groups, low quality or high quality, allowing for a simpler classification with two clusters. Another noticeable problem was cutting out further outliers, in future analysis of the data, anomalous samples with outlier data should be taken out of the dataset if removing more columns starts to negatively impact the outcome, as explored in the third section. In the future the dataset could be clustered using the DBSCAN model. DBSCAN may be more beneficial to cluster this data as it is not sensitive to outliers, noise or varying sizes of clusters due to DBSAN using a density based definition of a cluster. For these reason, DBSCAN is known to be better at finding clusters that Kmeans and hierarchical clustering would not.

References

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

[2] fortunebusinessinsights.com. Wine market research, 2020.

[3] Patrick McGovern, Mindia Jalabadze, Stephen Batiuk, Michael P Callahan, Karen E Smith, Gretchen R Hall, Eliso Kvavadze, David Maghradze, Nana Rusishvili, Laurent Bouby, et al. Early neolithic wine of georgia in the south caucasus. *Proceedings of the National Academy of Sciences*, 114(48):E10309–E10318, 2017.

[4] Doug Nierman. Fixed acidity, 2004.

[5] F. Almeida T. Matos P. Cortez, A. Cerdeira and J. Reis. Modeling wine preferences by data mining from physicochemical properties., 2009.