**Team 12 SA4108 CA submission**

**Unsupervised Clustering Report**

Group members:​

​

Li Zhengyi​

Liou Cheng Xiang​

Sivakumar Jayachelvi Nivvya​

Stephanie Cheng​

Thant Htet Myet​

Wang Wang Wang

**Unsupervised Techniques**

**Problem statement**

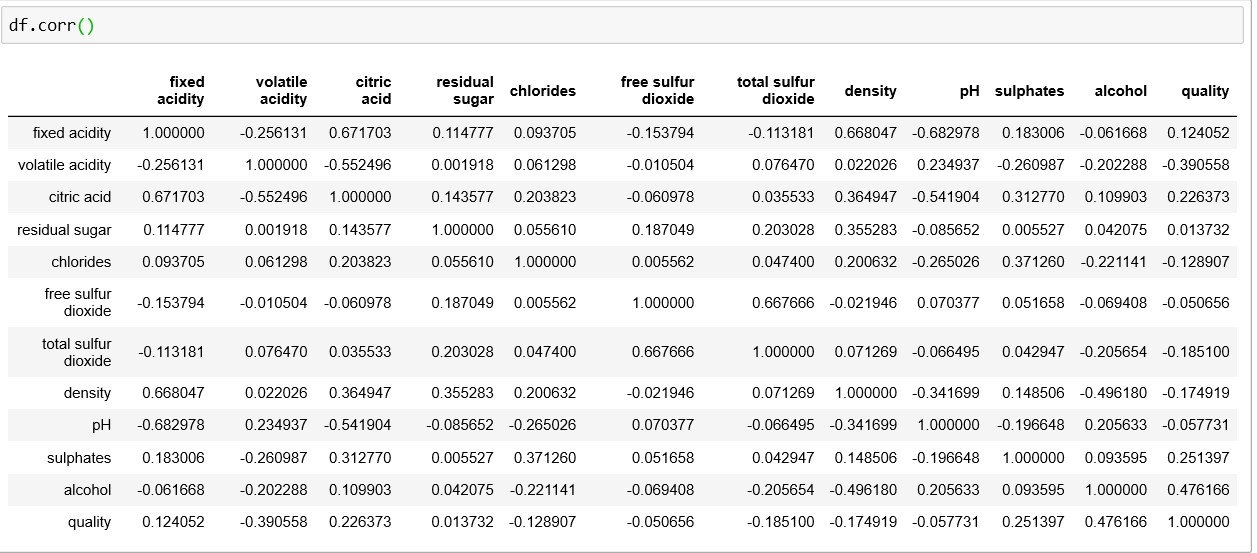
For our unsupervised clustering technique analysis, we have chosen the red wine quality dataset on Kaggle. It contains 12 variables, including the quality of the red wine. Our group wants to explore the dataset to see if we can identify clusters of samples with similar traits. We hypothesize that if there are clusters formed, these clusters should correspond to the quality of the wine. In other words, the samples in each cluster should have a similar quality score. We will then test for the performance of the Machine Learning algorithm by checking whether each cluster can capture a quality class correctly, and how many correct samples are there in a cluster. The result will verify whether our hypothesis is valid.

**Data Dictionary**

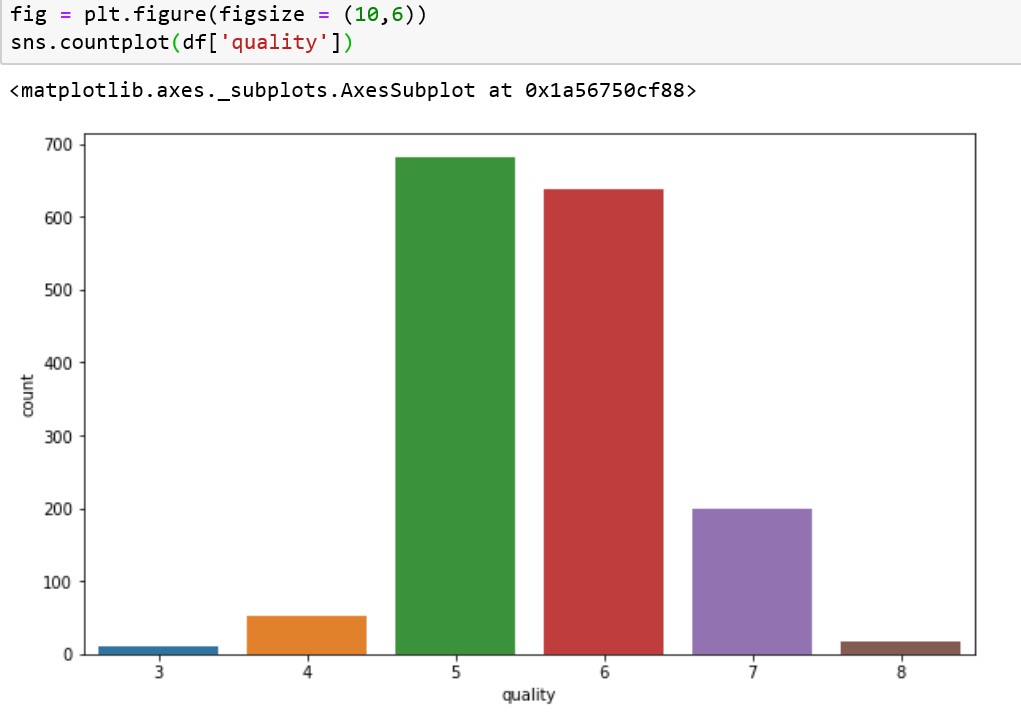
|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Data Type | Field Size | Description |
| fixed acidity | float | 2.1 | most acids involved with wine or fixed or non-volatile (do not evaporate readily) |
| volatile acidity | float | 1.3 | the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste |
| citric acid | float | 1.2 | found in small quantities, citric acid can add 'freshness' and flavor to wines |
| residual sugar | float | 1.1 | the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet |
| chlorides | float | 1.3 | the amount of salt in the wine |
| free sulfur dioxide | float | 2.1 | the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine |
| total sulfur dioxide | float | 3.1 | amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine |
| density | float | 1.5 | the density of water is close to that of water depending on the percent alcohol and sugar content |
| pH | float | 1.2 | 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 |
| sulphates | float | 1.2 | a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant |
| alcohol | float | 2.1 | the percent alcohol content of the wine |
| quality | int | 1 | output variable (based on sensory data, score between 0 and 10) |

**Data Preparation**

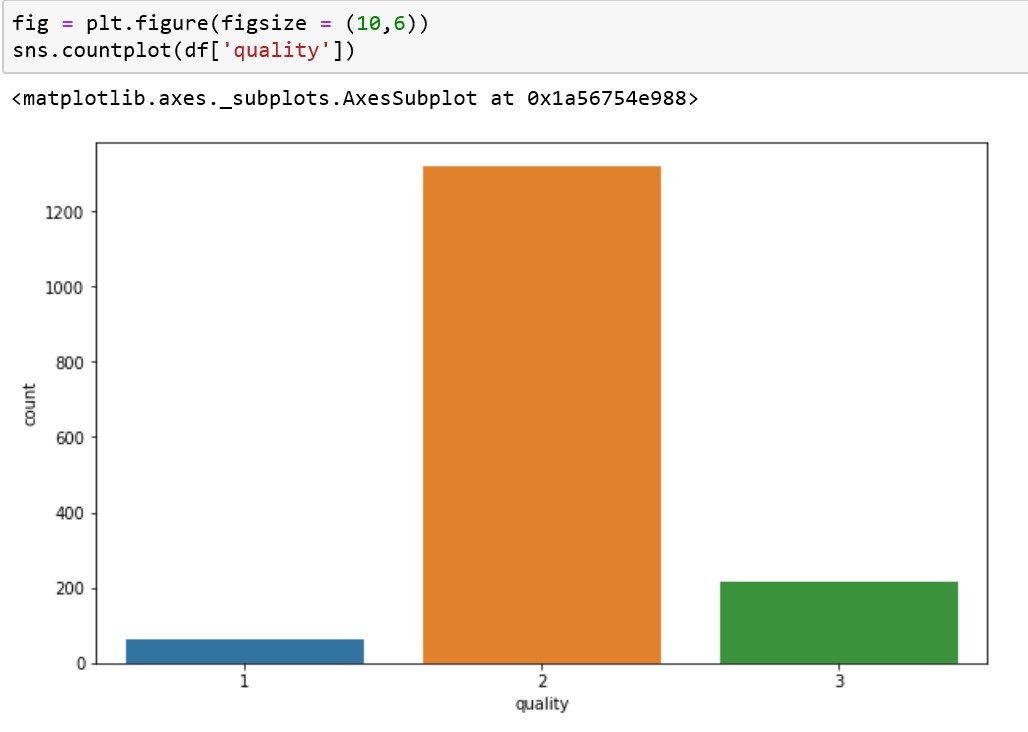
After we study the dataset, we can see that the feature set is clean and capture characteristics meaningful for the analysis. Furthermore, we tabulate the pairwise correlation score and see that there is no strong correlation between the features. Thus we do not drop any columns.



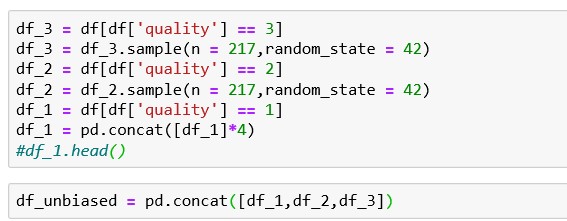
We then clean the dataset for null values by applying dropna() function on the dataframe.

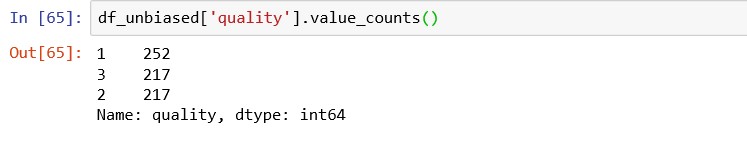
From the countplot of the ‘quality’ dataset, we can see that we have 6 different quality values. 

We then apply the binning technique to create only 3 quality classes for our analysis.



After which, we observed that we have an imbalanced dataset; thus, we apply the under-sampling and over-sampling technique to create a more balanced dataset.

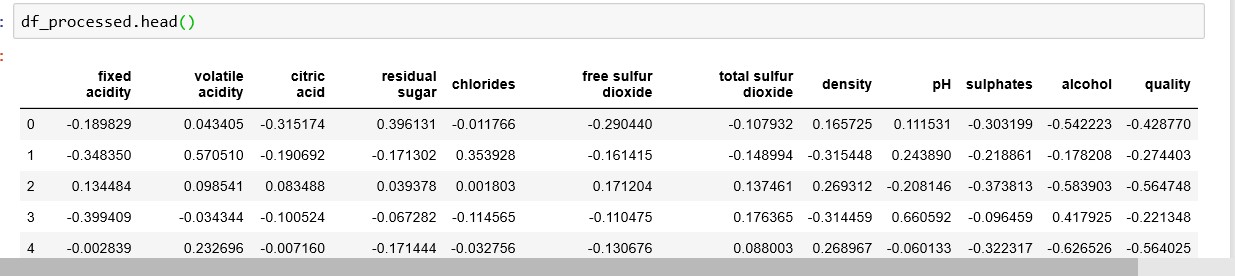




From our previous experience with the supervising techniques, we observed that scaling and normalization will improve the performance of Machine Learning algorithms. Thus we apply it to our dataset. Another observation we made when applying DBSCAN technique on the dataset is that this technique will not yield meaningful clusters if we do not do the scaling and normalization step earlier.



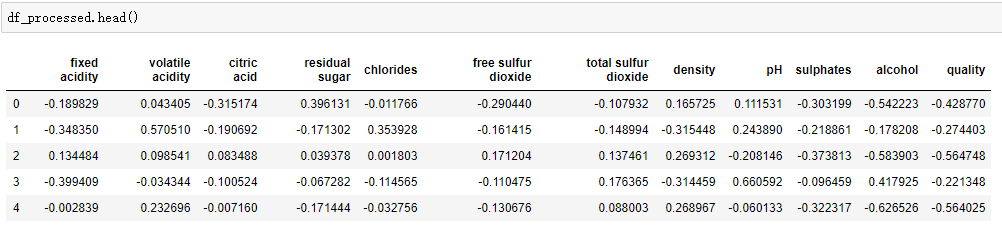
After the data preparation, we are then ready to train our DBSCAN and Kmean models. The dataset with the dependent variable and independent variables are shown below.



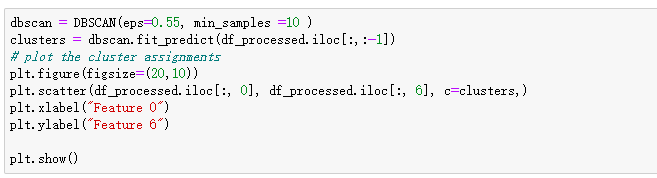
**Training Models**

**DBSCAN**

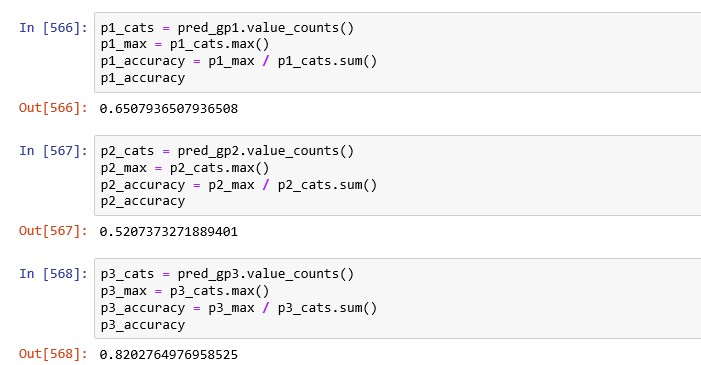
With the data ready, we proceed to train the DBSCAN model and evaluate the results. The independent variables (fixed acidity, volatile acidity, critic acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) are shown below.

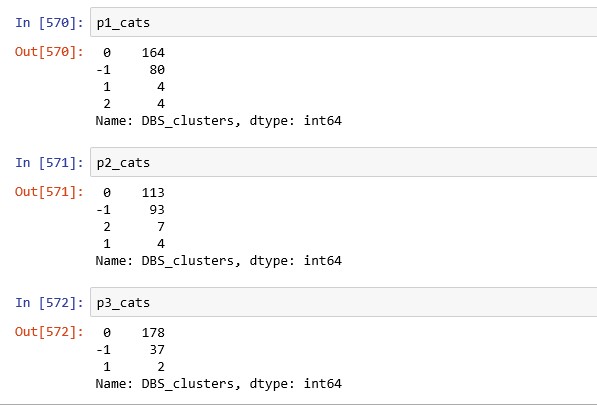


We trained our model and part of the codes are shown below. As a result, we get three clusters and the number of noise samples is 210.



In order to evaluate the result, we created a new dataframe to store the quality class of the samples and the cluster group they are assigned to by the algorithm. We then analyse their accuracy rate, and more importantly whether samples in one quality class are assigned to one specific cluster.

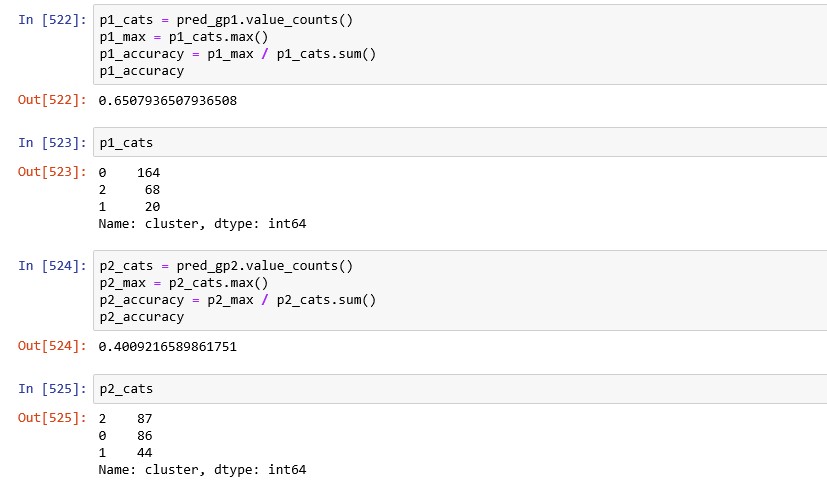


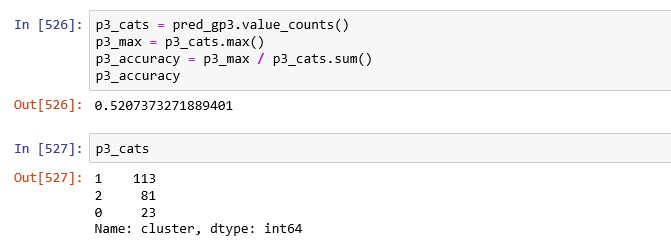


At the first look, it appears that the model has decent accuracy score of 65 %, 52 % and 82 %.However, after we check the cluster group of the sample inside each quality class, we realised that majority of the samples in quality class 1, 2 and 3 are assigned to cluster group 0. This shows that although clusters are found in this dataset with the DBSCAN algorithm, they do not correspond well to the quality class.

**K-Means Clustering**

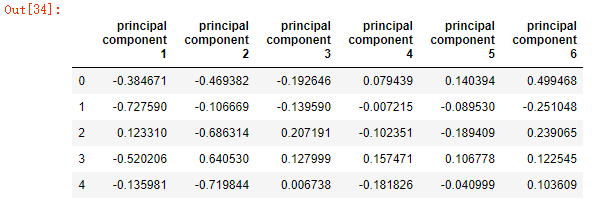
We then train the K-Means clustering model with the same dataset, and our result is as followed.





With K-means clustering, we set the K to be 3 as we expected the number of clusters to correspond to the number of quality class we defined. This time round, we are able to see that majority of the samples in each quality class are assigned to same cluster groups. However, the accuracy rate is not that high at 65%, 40% and 52% from quality class 1, 2 and 3 respectively. The results obtained do not validate our hypothesis convincingly.

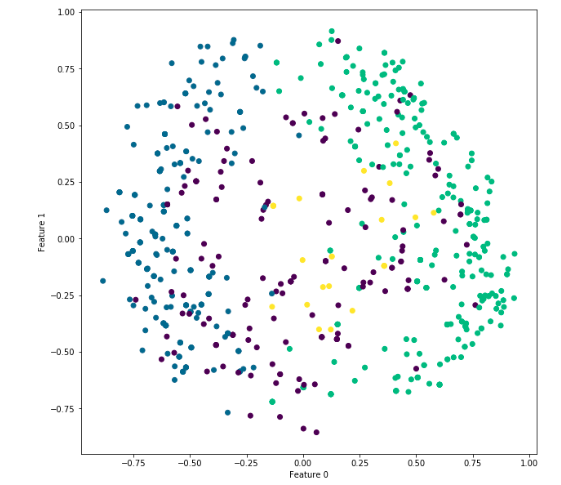
In an attempt to improve the result, we applied principal components analysis on the data. We reduced the dimension to 6D. The principal components are as shown below.



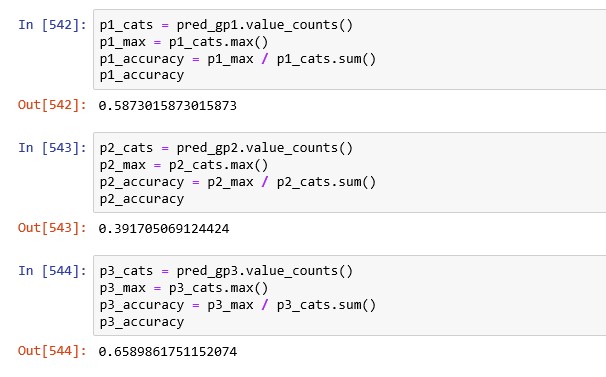
After which, we repeat the same steps above, training our model and then evaluating the result.

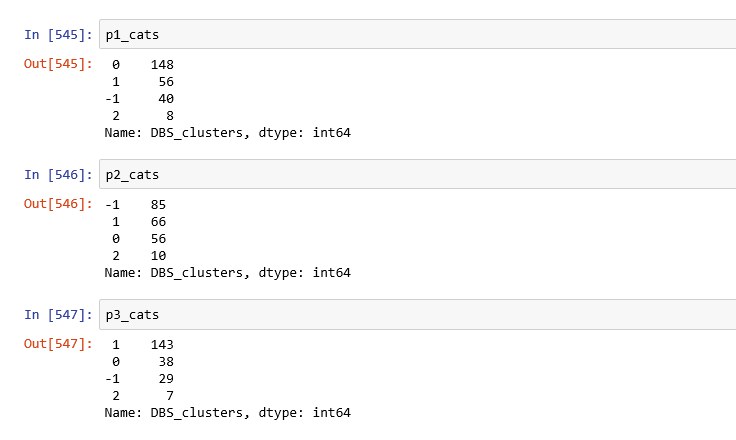
**DBSCAN (with PCA)**

For the DBSCAN, we observe that more defined clusters are formed.



We then calculate the results, as shown below:

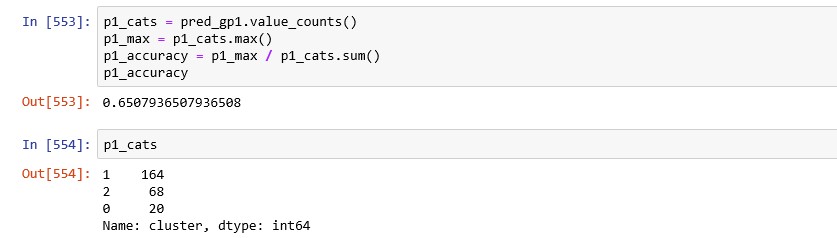


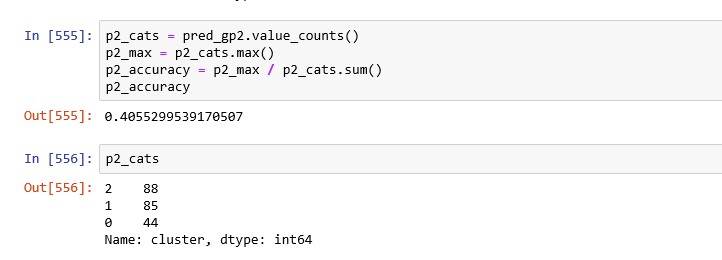


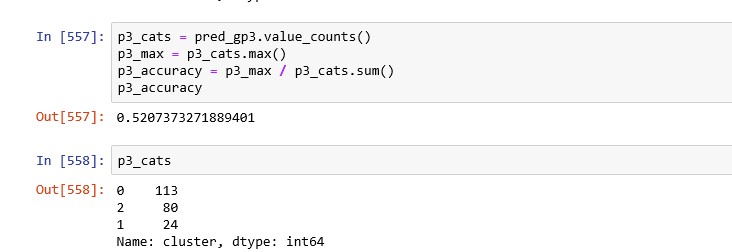
Compared to the result without PCA, we can see that majority of the samples from quality class 2 and 3 are assigned to cluster group 1. From this result, we are unable to prove that the clusters formed correspond to the quality classes very well.

**K-Means Clustering (with PCA)**

We then carry out the same analysis with K-means clustering algorithm and the result is as shown below.







Our group observed that even after doing PCA, there is only slight changes in term of samples’ cluster assignment and the accuracy score.

**Conclusion**

For the DBSCAN analysis, we understand the importance of the data normalization and data balancing. Without them, we observed that it is difficult for the model to form more than 1 cluster. However, as our results shown, although DBSCAN algorithm are able to group the samples into clusters, those clusters do not correspond well to the quality class.

Our mean accuracy for k-Means cluster algorithm with or without PCA is 55.3%. We believe this is because the clustering does not correspond to ‘quality’, which means it may correspond to other characteristics like the sweetness or color degree of red wine. Other possible reason may be we don’t group the quality in the right way.

As such, we concluded that our hypothesis that we can identify clusters in the wine sample corresponding to wine quality to be invalid.