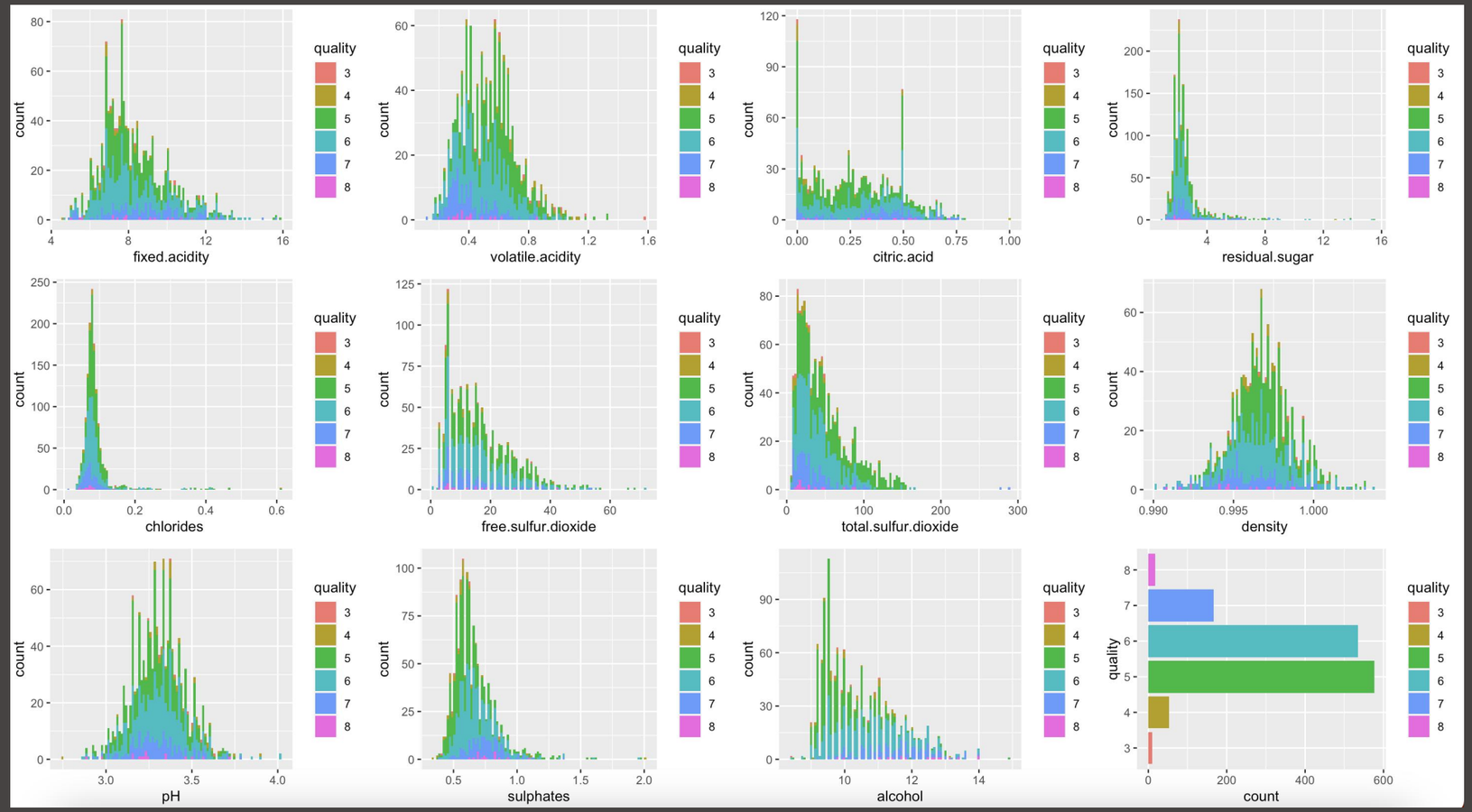
Analysis of Red Wine Quality Data Set

Date: 23/11/2023

Introduction

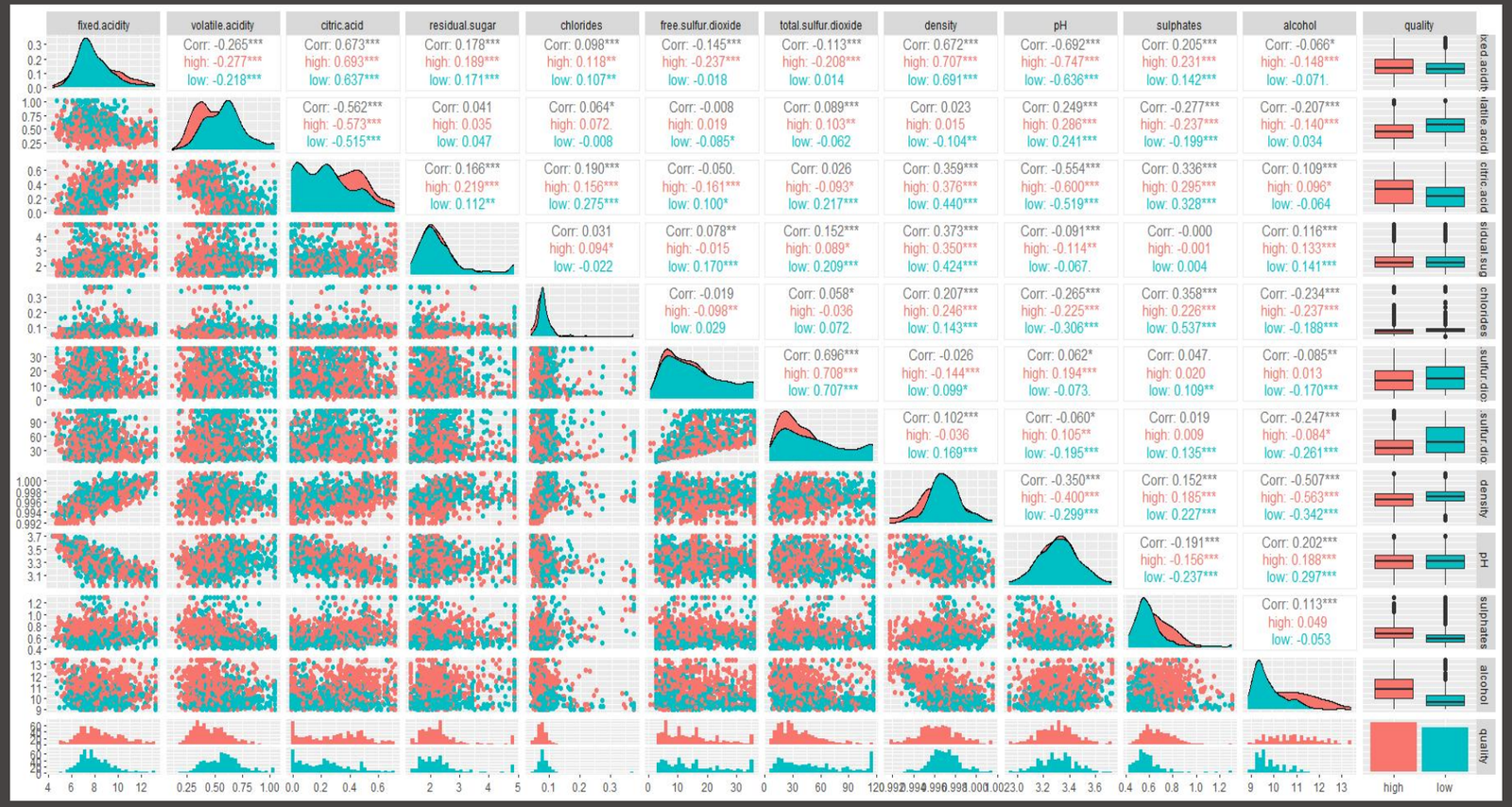
- Dataset Red Wine Quality
  (https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009)
- 11 attributes
- All x attributes are numerical and continuous, y attribute is categorical
- 1,599 records, no missing values, 240 records were duplicated and removed

Preliminary Analysis - Histograms



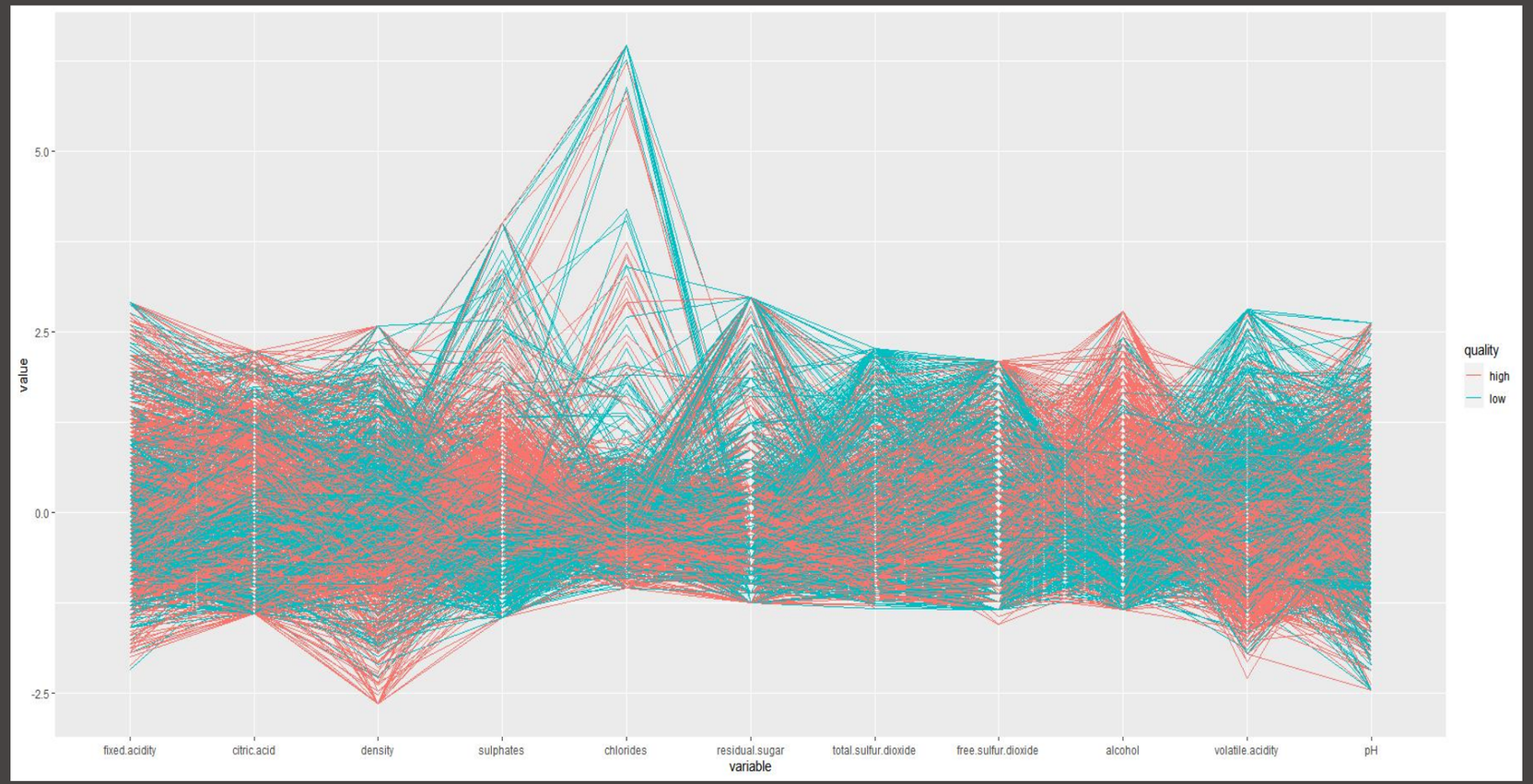
- Histograms show outliers and mostly well spread data
- Residual sugar, chlorides, and sulphates are more centered
- Class imbalanced - most data belongs to levels 5 and 6

Preliminary Analysis - Pairwise comparison of data



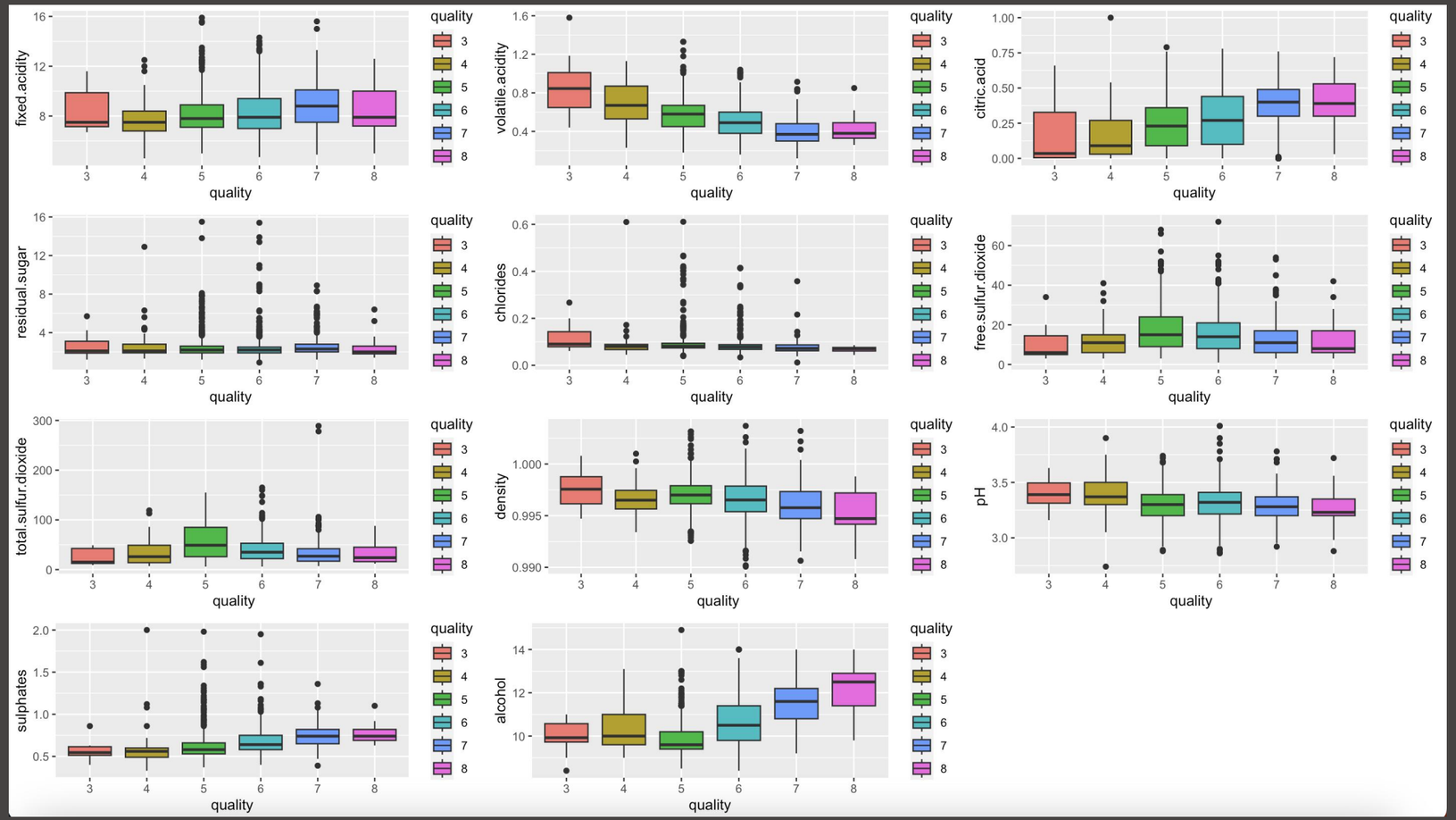
- The alcohol content can distinguish good wine from bad wine very well
- Fixed. acidity has a strong relationship with density and pH

Preliminary Analysis - parallel coordinates plot



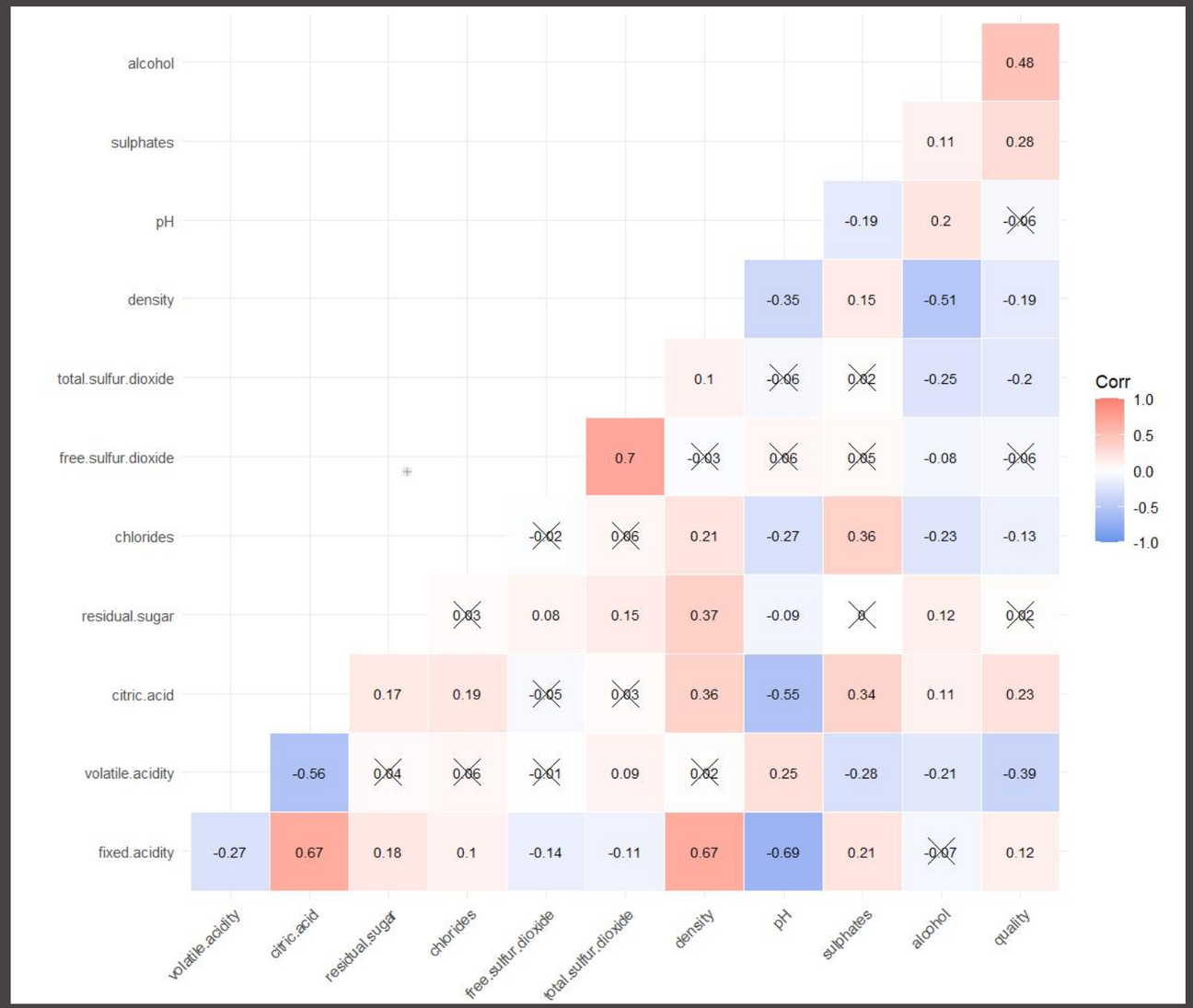
- There is a clear layering in sulphates, alcohol, and volatile.acidity environment
- There are obvious outliers in chlorides

Preliminary Analysis - Boxplot



- Many outliers

Preliminary Analysis - Correlation



- Highest correlation between:
  - Fixed acidity-pH (0.69)
  - Fixed acidity-citric acid (0.67)
  - Fixed acidity-density (0.67)
  - Free sulfur dioxide-total sulfur dioxide (0.67)

Data Mining - Training Data

- Fix class imbalance
- Split 70% for training data and 30% for test data
- Methods:
  - Decision Tree
  - K-Nearest Neighbors
  - Naïve Bayes Classifiers
  - Linear Support Vector Machines

Total data:
High 52.91%
Low 47.09%

	quality	n
1	high	719
2	low	640

Training data:
High 52.94%
Low 47.06%

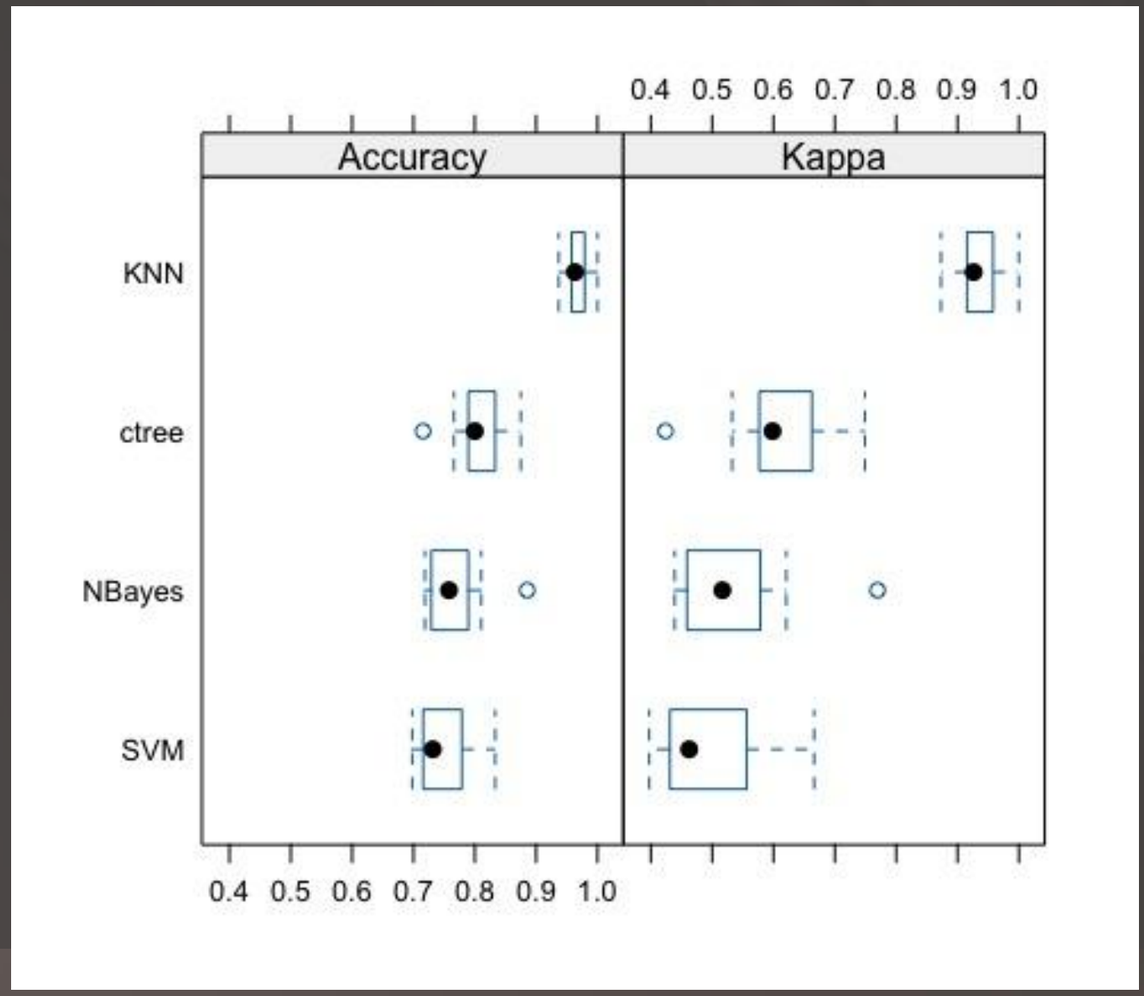
	quality	n
1	high	504
2	low	448

Test data:
High 52.83%
Low 47.17%

	quality	n
1	high	215
2	low	192

Data Mining - Comparison

- K-Nearest Neighbors had the most accurate prediction



Data Mining - Comparison

- K-Nearest Neighbors had the highest accuracy and kappa

```
Call:
summary.resamples(object = resamps)

Models: ctree, KNN, NBayes, SVM
Number of resamples: 10

Accuracy
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
ctree  0.7157895 0.7900219 0.8000000 0.8044905 0.8328947 0.8750000    0
KNN    0.9368421 0.9578947 0.9633224 0.9695061 0.9790559 1.0000000    0
NBayes 0.7187500 0.7303856 0.7578947 0.7698955 0.7894737 0.8854167    0
SVM    0.6979167 0.7165296 0.7315789 0.7457503 0.7710526 0.8333333    0

Kappa
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
ctree  0.4239838 0.5788667 0.5984407 0.6070636 0.6622141 0.7490196    0
KNN    0.8730512 0.9155556 0.9263642 0.9387022 0.9578634 1.0000000    0
NBayes 0.4389610 0.4616598 0.5166371 0.5399122 0.5784799 0.7696335    0
SVM    0.3974026 0.4319920 0.4625434 0.4911636 0.5410239 0.6662321    0
```

Data Mining - Train Data

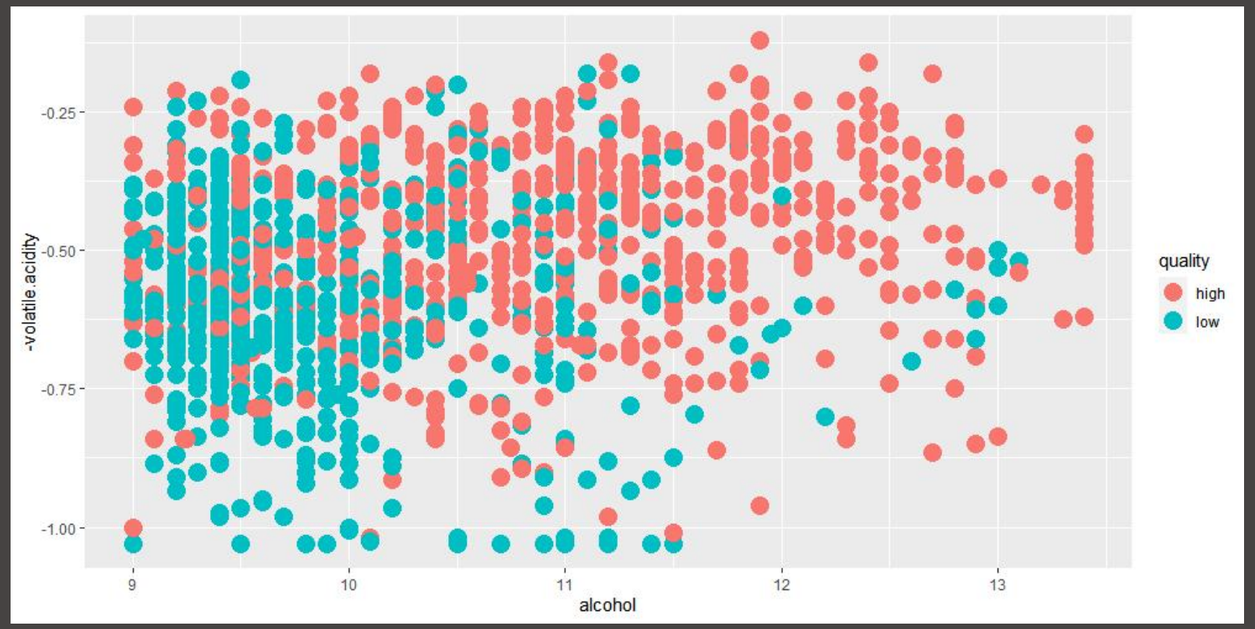
- Performance is best when k is equal to 1
- 68.3% accuracy in the test set

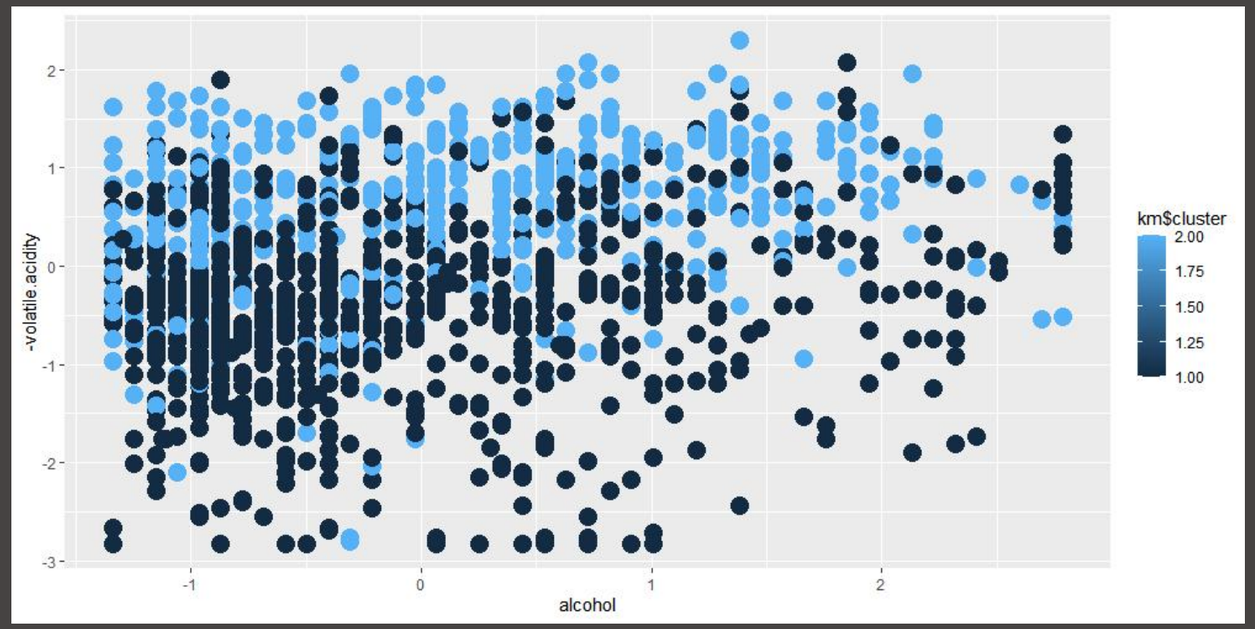
```
k-Nearest Neighbors

952 samples
 11 predictor
  2 classes: 'high', 'low'

Pre-processing: scaled (11)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 857, 857, 857, 857, 856, 858, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9684868  0.9367852
   2  0.8171601  0.6321567
   3  0.8193202  0.6358651
   4  0.7856689  0.5692691
   5  0.7877961  0.5722389
   6  0.7846930  0.5657909
   7  0.7793311  0.5552953
   8  0.7646382  0.5263586
   9  0.7698684  0.5368630
  10  0.7740461  0.5458925

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 1.
```

Cluster Analysis - After Scale

- Choose volatile. acidity and alcohol as the x-axis and y-axis

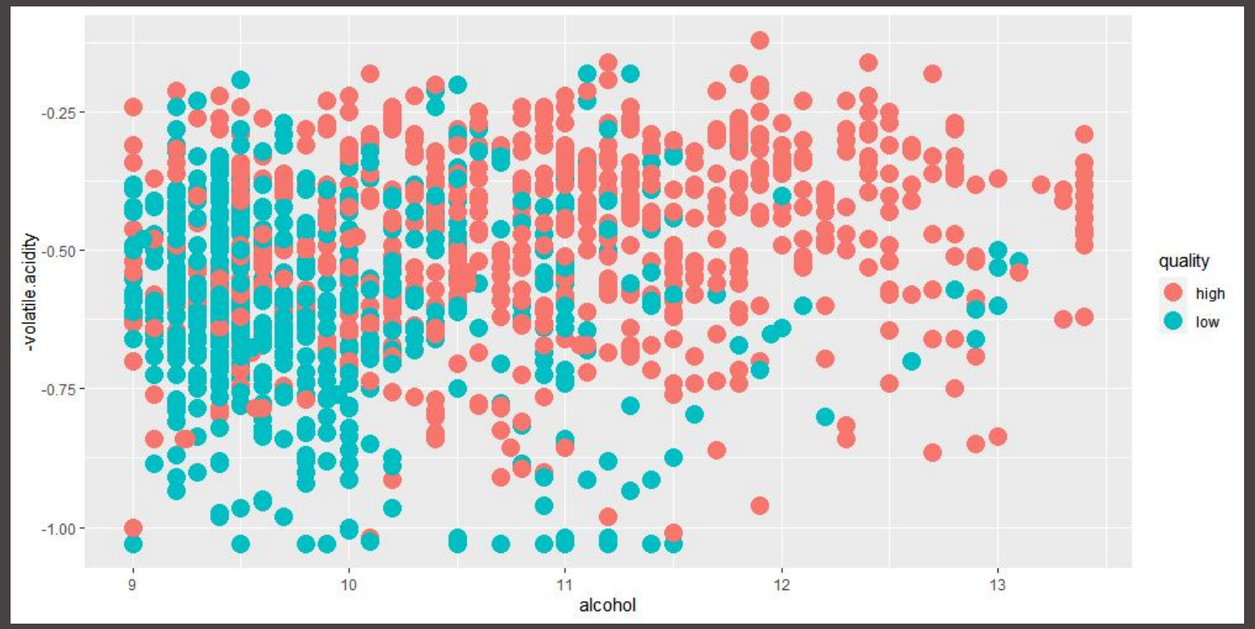
K-means Clustering - Classification based on km results

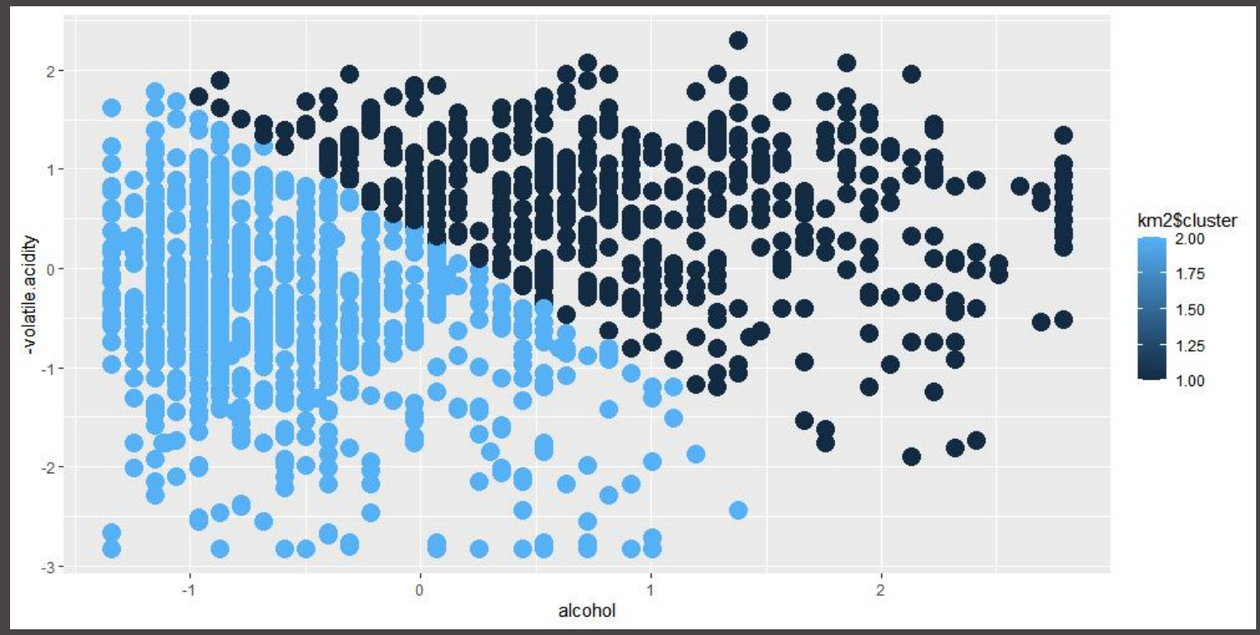




- Consider All Factors

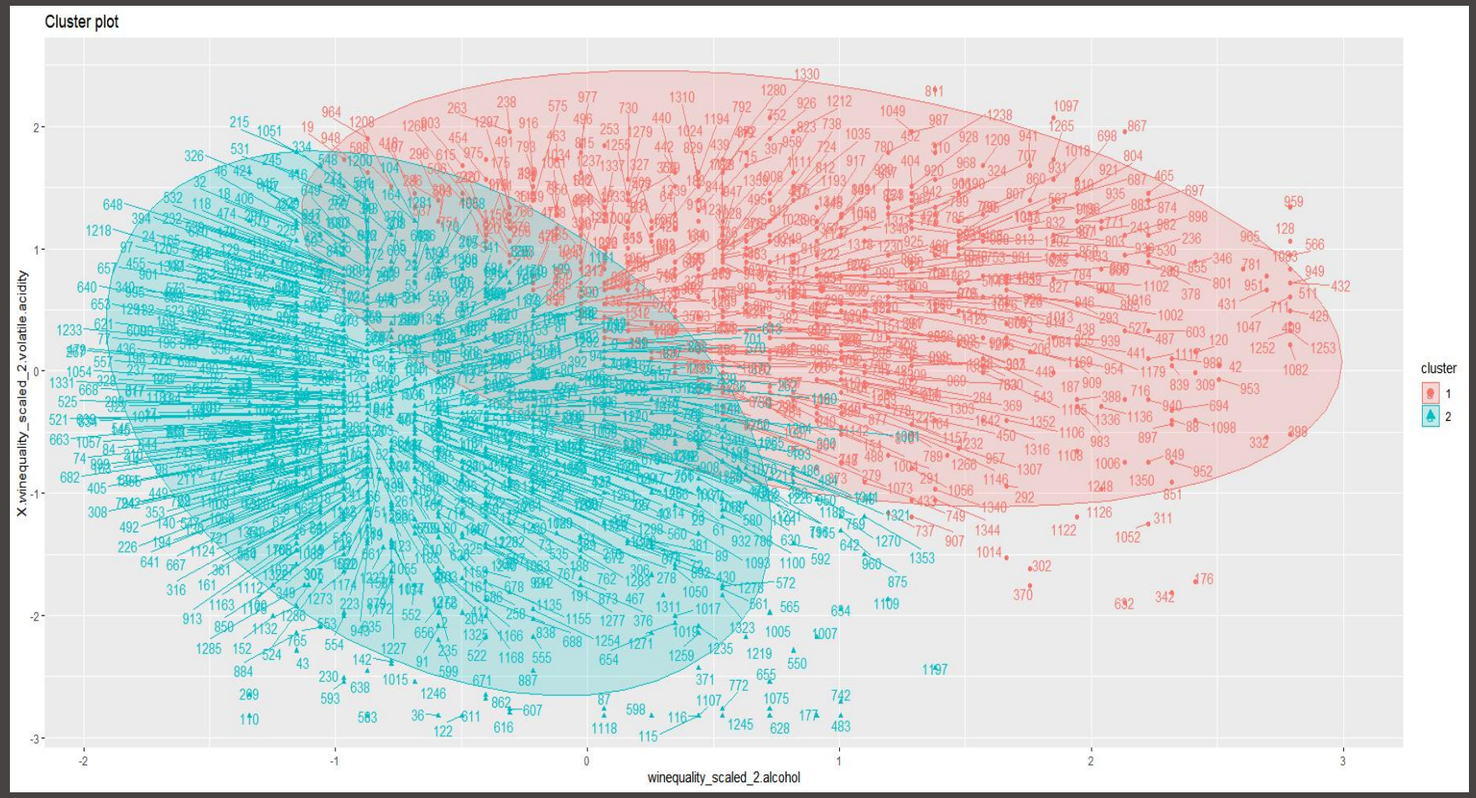
K-means Clustering - Classification based on km results





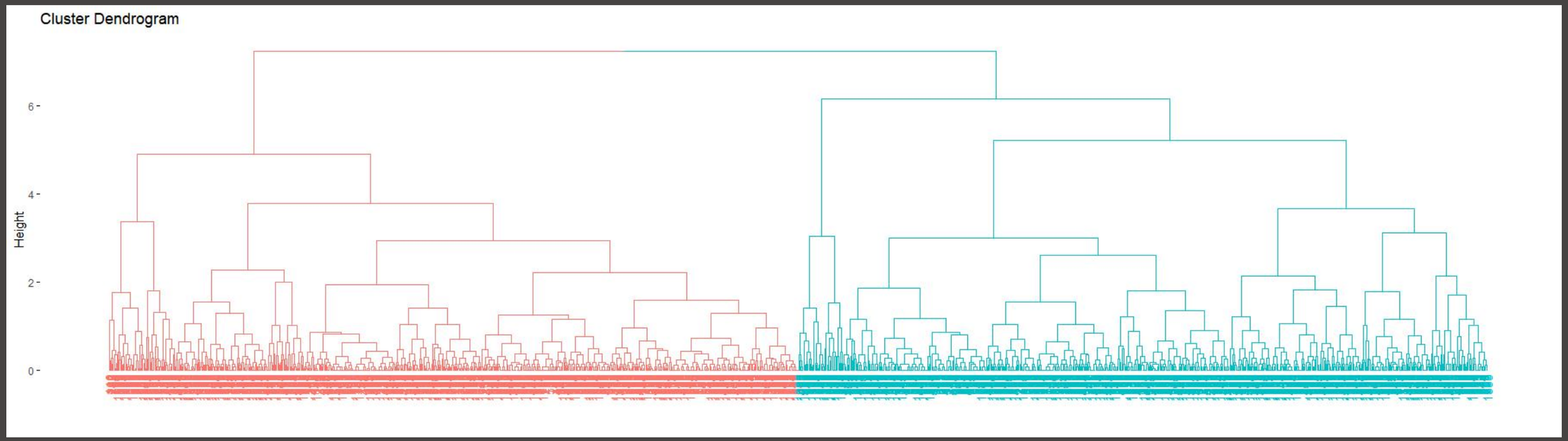
- Consider only two Factors
- The accuracy of km prediction is 58.4%

Cluster Analysis - Feature Visualization



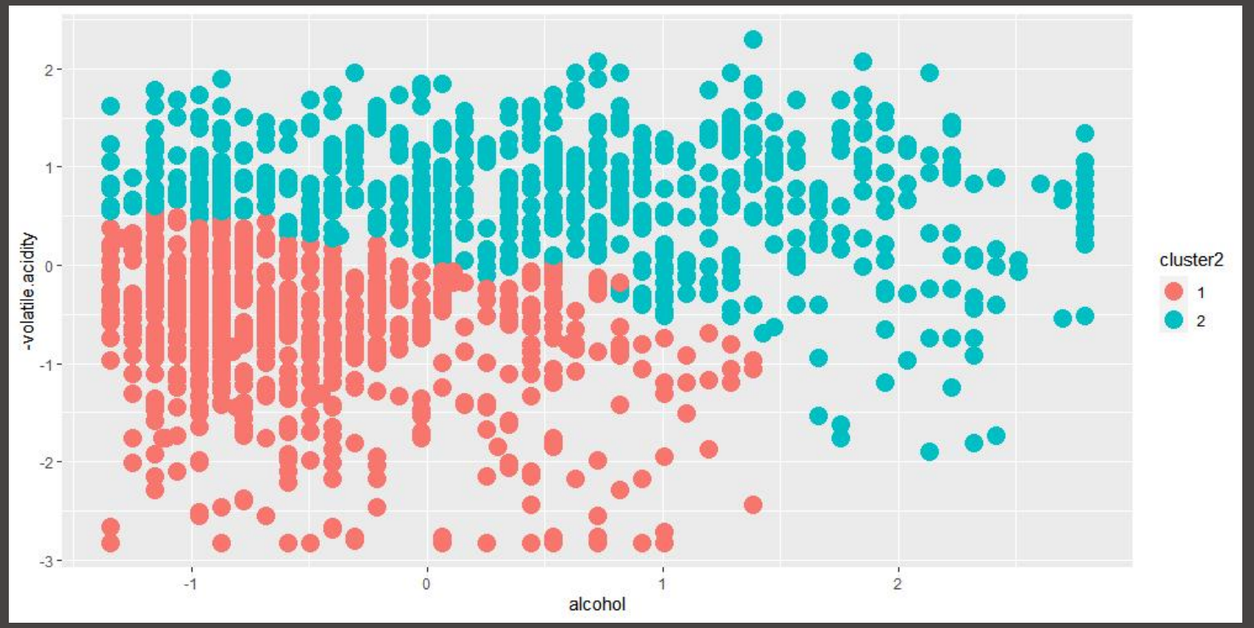
- Most of the distinctions are obvious

Hierarchical Clustering



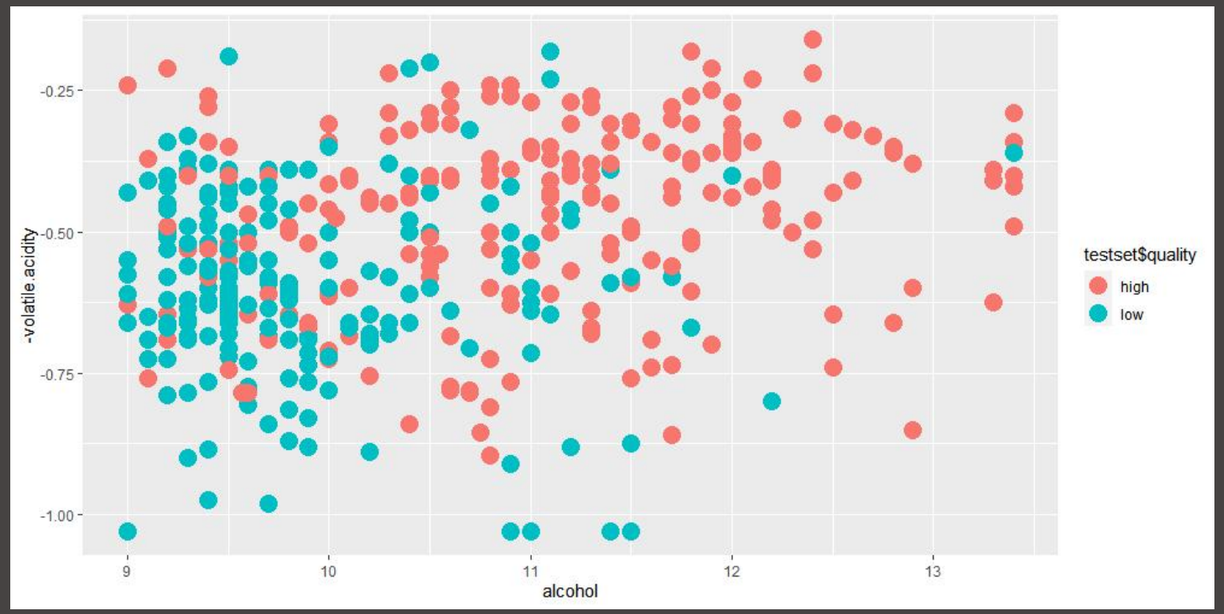
- The color distribution is very consistent with the proportion of the original data
- (52.9%:47.1%)
- The accuracy of Hierarchical Clustering is 69.2%

Hierarchical Clustering



- The Hierarchical Clustering results are represented in a scatter plot

K-means Clustering - Classification based on km results



Model Tuning - Results

Alcohol and volatile acidity

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  165  55
      low    50 137

               Accuracy : 0.742
                 95% CI : (0.6966, 0.7839)
    No Information Rate : 0.5283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4817

 Mcnemar's Test P-Value : 0.6963

            Sensitivity : 0.7674
            Specificity : 0.7135
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.7326
             Prevalence : 0.5283
         Detection Rate : 0.4054
   Detection Prevalence : 0.5405
      Balanced Accuracy : 0.7405

       'Positive' Class : high
```

Alcohol, volatile acidity, and pH

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  144  62
      low    71 130

               Accuracy : 0.6732
                 95% CI : (0.6253, 0.7186)
    No Information Rate : 0.5283
    P-Value [Acc > NIR] : 2.034e-09

                  Kappa : 0.346

 Mcnemar's Test P-Value : 0.4879

            Sensitivity : 0.6698
            Specificity : 0.6771
         Pos Pred Value : 0.6990
         Neg Pred Value : 0.6468
             Prevalence : 0.5283
         Detection Rate : 0.3538
   Detection Prevalence : 0.5061
      Balanced Accuracy : 0.6734

       'Positive' Class : high
```

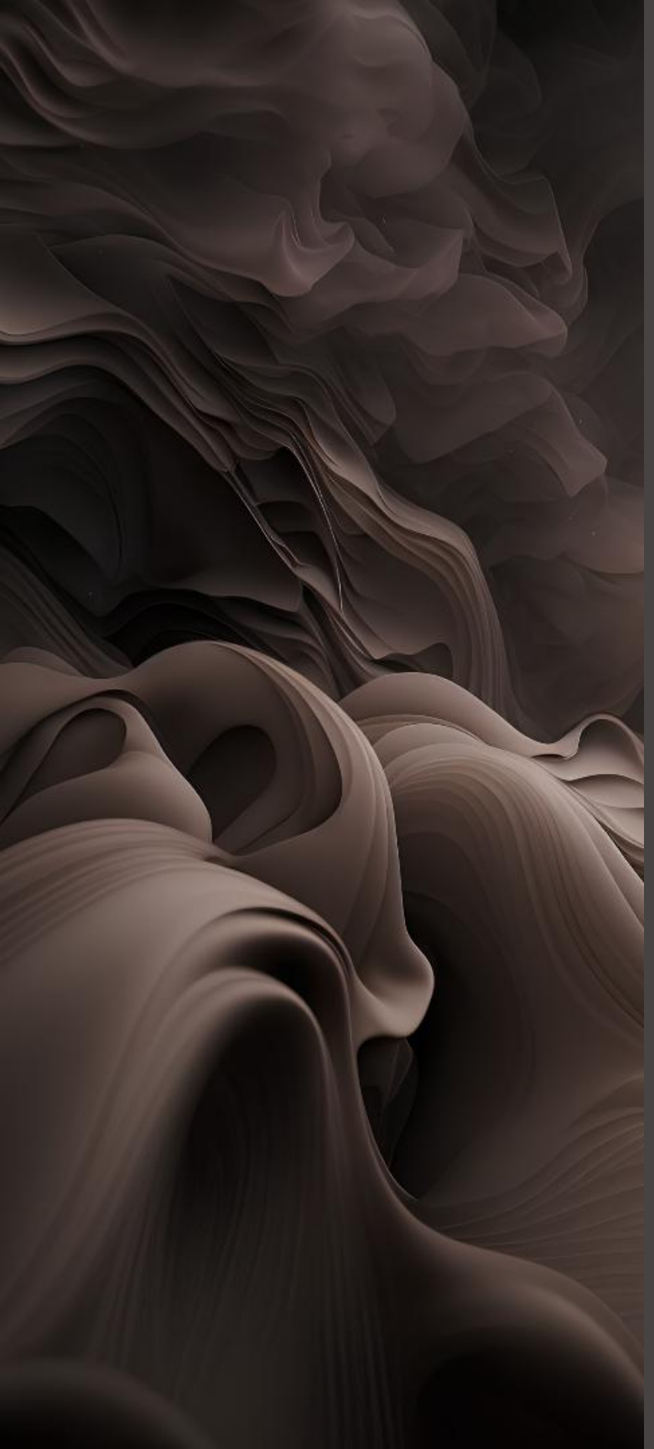
Fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, sulphates, and alcohol

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  157  73
      low    58 119

               Accuracy : 0.6781
                 95% CI : (0.6303, 0.7233)
    No Information Rate : 0.5283
    P-Value [Acc > NIR] : 5.755e-10

                  Kappa : 0.3515

 Mcnemar's Test P-Value : 0.2213

            Sensitivity : 0.7302
            Specificity : 0.6198
         Pos Pred Value : 0.6826
         Neg Pred Value : 0.6723
             Prevalence : 0.5283
         Detection Rate : 0.3857
   Detection Prevalence : 0.5651
      Balanced Accuracy : 0.6750

       'Positive' Class : high
```

Data Mining - Test Data

- Predictions have high False Positive and False Negative
- Kappa is low

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  170  53
      low    45 139

               Accuracy : 0.7592
                 95% CI : (0.7146, 0.8)
    No Information Rate : 0.5283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5158

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.7907
            Specificity : 0.7240
         Pos Pred Value : 0.7623
         Neg Pred Value : 0.7554
             Prevalence : 0.5283
         Detection Rate : 0.4177
   Detection Prevalence : 0.5479
      Balanced Accuracy : 0.7573

       'Positive' Class : high
```



Data set improvement suggestions：

1. More data of 8 or 3 high-quality wines are required, and an extremely small number will seriously affect the prediction of this quality wine, even if the method of increasing the cost of error is used, it is impossible to make the expected correction for such a small number of data categories.

2. Some data outside the normal range need to be corrected.

3. As can be seen from the chart, there is a lot of noise in the red wine quality data set, and no obvious difference in location can be obtained in the classification process, operations such as feature selection and data conversion are required

The findings after data analysis:

1、From preliminary data processing to later accuracy comparison, two factors can achieve such high accuracy, confirming that most of the data in the red wine dataset is noise or has minimal impact

2、Through the analysis of various indicators of red wine, it can be predicted that the probability of this is a bottle of good wine is more than 70%, indicating that a bottle of good wine and a bottle of bad wine do have a relatively obvious difference in the content of various ingredients, and different contents of ingredients will match different quality red wine.

3、The data set is relatively complete and does not require too much data processing at the beginning.

4、Good red wine generally has higher alcohol concentration and lower volatile acidity.

Suggestions after data analysis:

1、Cooling temperature reduces the evaporation rate of alcohol. Place the red wine in the refrigerator to cool for a period of time, then slowly raise it to the desired temperature, which can increase the alcohol content of the red wine.、

2、During the brewing process, appropriate acidification treatment can be carried out as needed, and the acidity in the wine can be balanced by adding an appropriate amount of acid to reduce the volatile acid content.

3、Different grape varieties produce different levels of acidity, and choosing low acidity grape varieties can reduce the content of volatile acids.