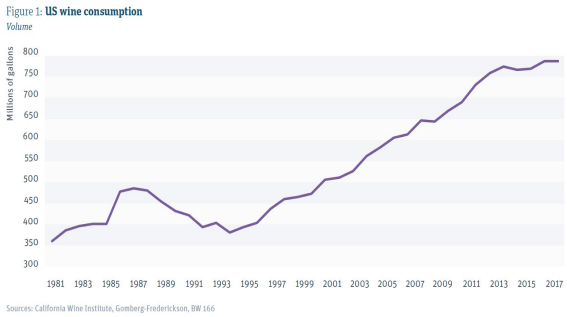
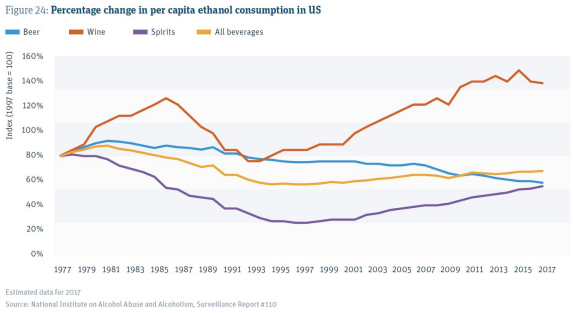
ANLY 512 Final Project
Wine Quality Research

Chenxi Liu
Yuan Liu
Qian Yi
Jingyu Zhang

Background

According to the *State of The Wine Industry Report 2019*, the United States is the largest wine-consuming country in the world, giving US producers an amazing home-court advantage. The historical data indicated that from 1994, the US wine business experienced a significant long-term increasing trend.





With solid reports and statistics, the United States per capita wine consumption performed a significant increase while spirits and beer showed relatively modest trends. So in the United States, the wine market has tremendous potential that may generate significant revenues.

The quality and taste of wine are crucial. Whether the wine meets human taste preferences will be directly related to the market sizes and market shares of these products. The goal of this project is to build machine learning models that could support wine tasting evaluations given the different physicochemical attributes. Through these models, wine manufacturers could have ideas about the key factors that they should concern. They could also generate research directions to improve wine quality and wine taste. These models are useful to improve wine production. In addition, these models can also help in target marketing by modeling customers' preference from different niche markets.

Dataset

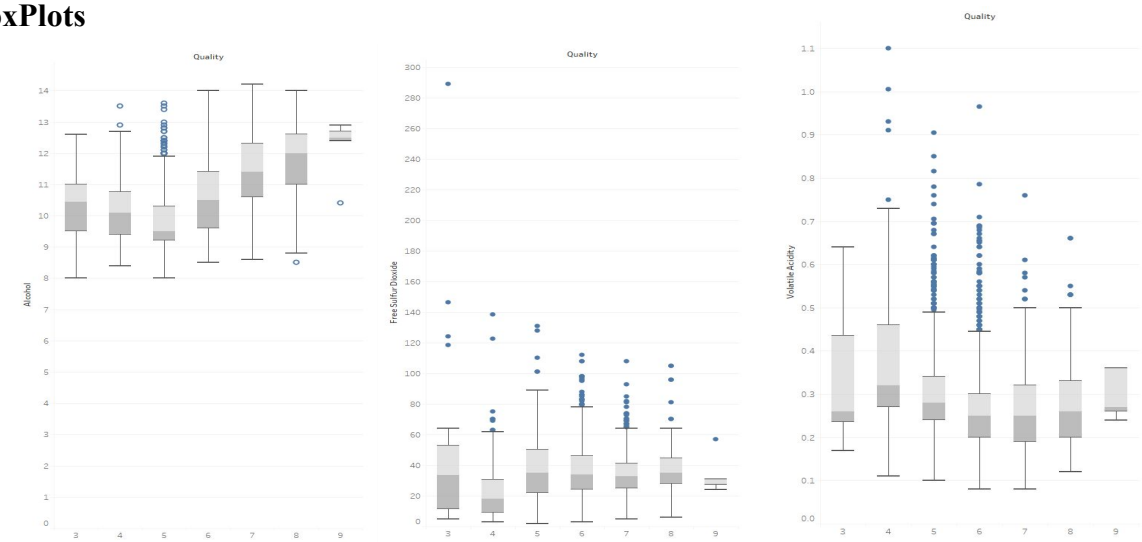
This dataset was generated from University of California Irvine Machine Learning Repository website. There are 11 predictor columns and each column represents a physicochemical attribute of the wine. The label column contains categorical levels of wine quality.

Exploratory Data Analysis

BarPlots

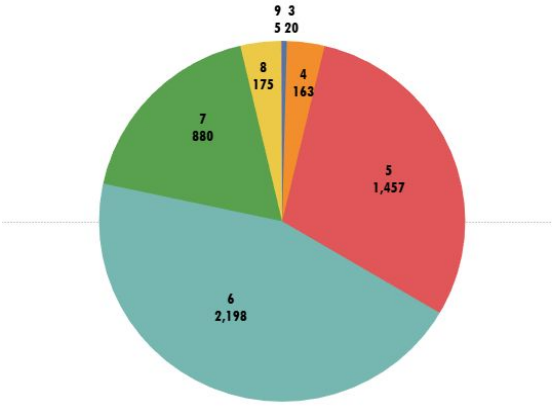
This bar plot shows the general idea of the data. It's obvious that some of the attributes vary very little for different wine quality levels, such as Density PH and average sulfate, however, there are some attributes that vary quite a bit such as alcohol and free sulfur dioxide.

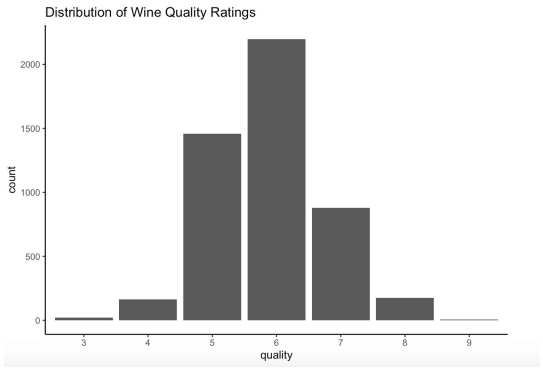
BoxPlots



For a closer look, the boxplots for the three variables have been used and these variables vary a lot based on the wine quality level. It's clear that there are some outliers within the predictors.

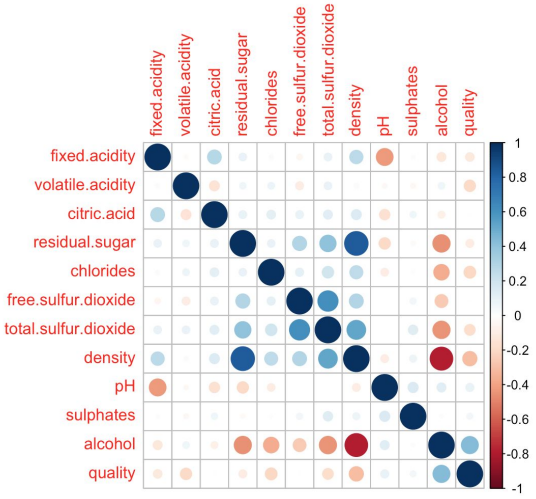
Label Distribution

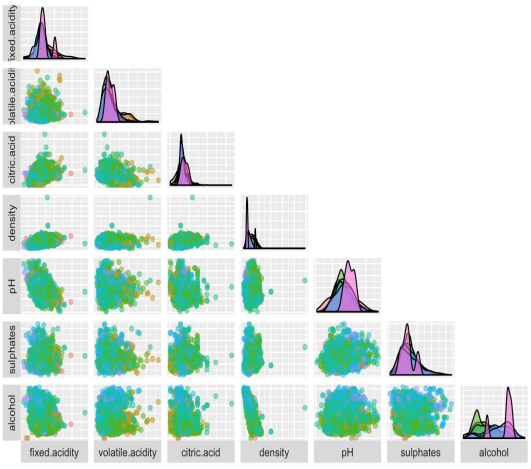




For looking at the label distribution, there were very few observations for wine quality 3 and wine quality 9. The wine quality 5, wine quality 6, and wine quality 7 show large observations. So there was a clear label imbalance problem and it might impact the model prediction negatively. Then oversampling could be used to create a more balanced label distribution in order to fix this problem.

Correlation

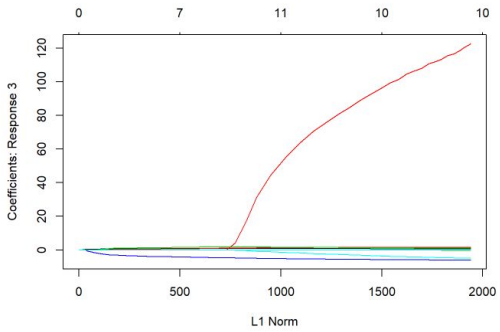


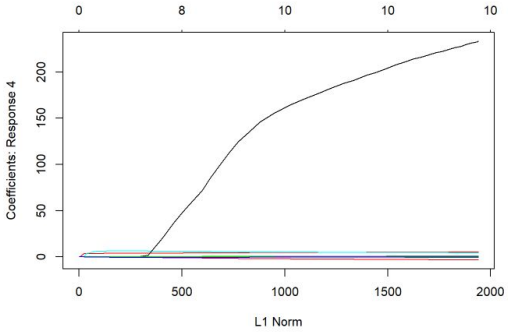


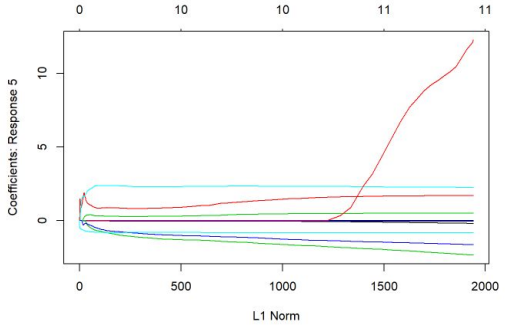
The correlation matrix can also be drawn to see the correlation between white wine's quality and other variables. The blue dot means the variables are positively correlated, and the red dot means the variables are negatively correlated. Then the graph shows that alcohol and density have a relatively strong relationship with the quality, specifically, the alcohol shows the relatively positive relation with the wine quality and the density shows relatively negative relation with the wine quality.

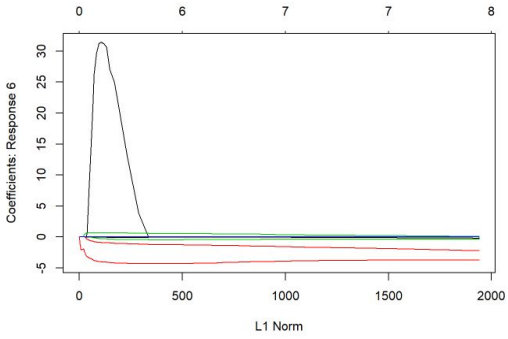
Multinomial Lasso Regression

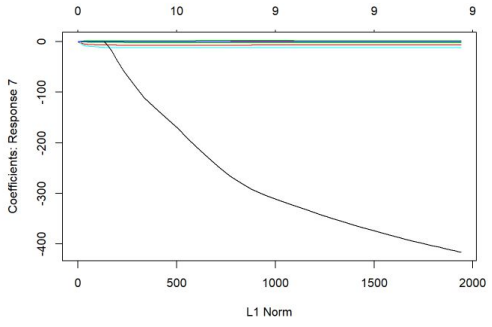
For more detailed analysis, the multinomial lasso regression is used. Ideally, the lasso regression will provide variable choice for each of the labels. (level of wine quality) In the process of building the model, the data imbalance is noticed as there are only 5 observations for the quality level 9 and 20 observations for the quality level 3. (Out of more than 3600 observations) Oversampling method is applied to solve the problem of data imbalance. After building the model, the following graphs are presented.

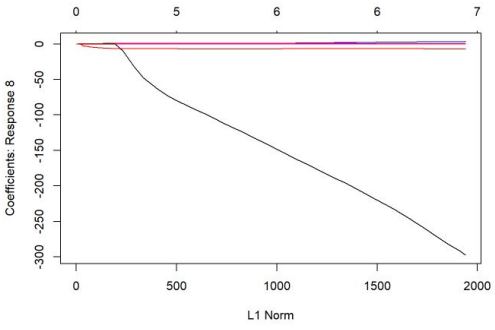


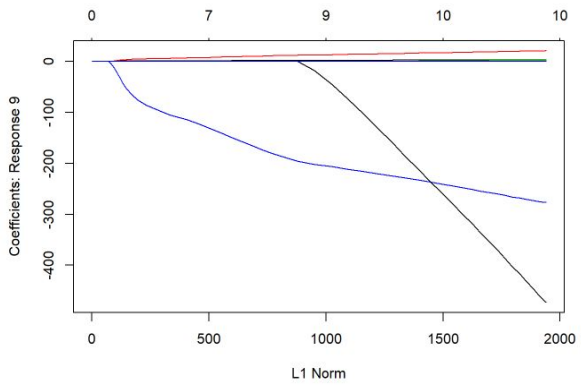




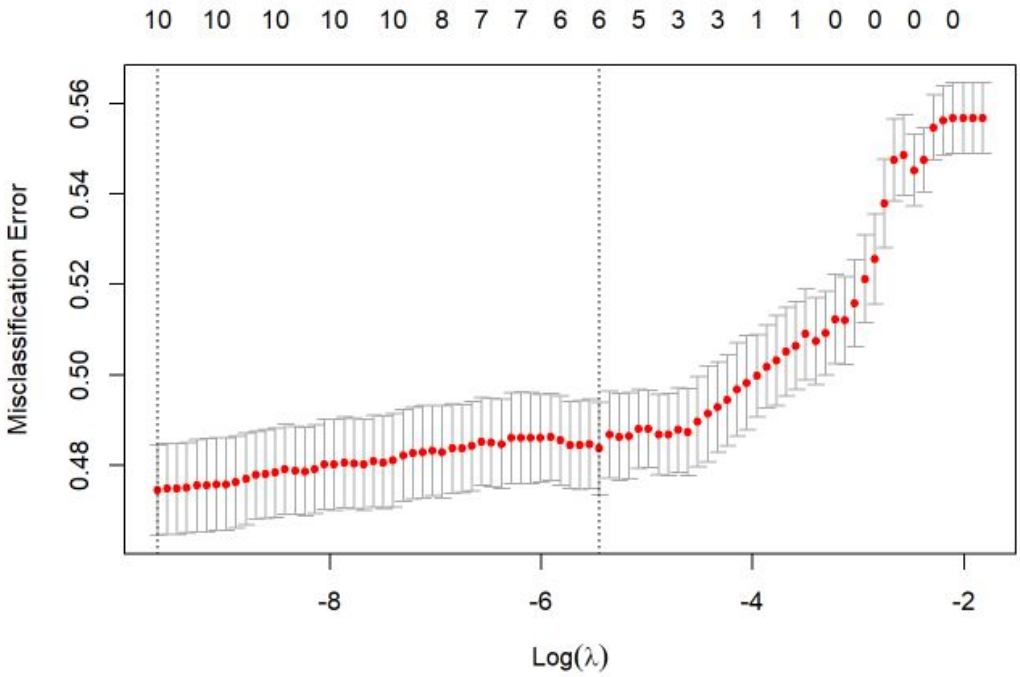








To better interpret the model, we choose the lambda with the least misclassification error on the testing data.



Then the coefficient of each predictor associated to the different level of wine quality is obtained:

```
## $`3`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)          -2.038391e+02
## (Intercept)           .
## fixed.acidity         5.598210e-01
## volatile.acidity      1.544162e+00
## citric.acid          -1.522891e-01
## residual.sugar       -1.069421e-01
## chlorides             .
## free.sulfur.dioxide   1.521637e-02
## total.sulfur.dioxide  8.130362e-03
## density               7.998550e+01
## pH                    1.017839e-01
## sulphates            -5.320139e+00
## alcohol              -2.527539e-02
##
## $`4`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)          -2.866734e+02
## (Intercept)           .
## fixed.acidity        -3.558959e-01
## volatile.acidity      4.859123e+00
## citric.acid           2.868932e-01
## residual.sugar       -2.010750e-01
## chlorides             3.441476e+00
## free.sulfur.dioxide  -5.490122e-02
## total.sulfur.dioxide -1.506094e-03
## density               1.886115e+02
## pH                   -3.441663e+00
## sulphates            -1.771338e+00
## alcohol              -6.230058e-01
##
```

```
## $`5`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)          -1.127979e+02
## (Intercept)           .
## fixed.acidity        -2.379319e-01
## volatile.acidity      7.425178e-01
## citric.acid           1.255643e-01
## residual.sugar       -6.332946e-02
## chlorides             3.119120e-02
## free.sulfur.dioxide  -8.886790e-03
## total.sulfur.dioxide  3.560308e-03
## density               1.437077e+01
## pH                   -2.868947e+00
## sulphates            -1.683779e+00
## alcohol              -8.317170e-01
##
## $`6`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)          -1.057067e+02
## (Intercept)           .
## fixed.acidity        -2.998124e-01
## volatile.acidity     -4.502309e+00
## citric.acid          -4.225328e-01
## residual.sugar        .
## chlorides            -1.419198e+00
## free.sulfur.dioxide  -1.405314e-03
## total.sulfur.dioxide  1.065751e-03
## density               .
## pH                   -2.897014e+00
## sulphates             1.816807e-01
## alcohol               1.304029e-02
##
```

```
## $`7`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)           3.380622e+02
## (Intercept)           .
## fixed.acidity         1.549179e-01
## volatile.acidity     -7.172379e+00
## citric.acid          -1.595888e+00
## residual.sugar        2.133992e-01
## chlorides            -1.158916e+01
## free.sulfur.dioxide   .
## total.sulfur.dioxide -3.274419e-04
## density              -4.615320e+02
## pH                    .
## sulphates             2.333521e+00
## alcohol               7.265307e-02
##
## $`8`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)           2.723865e+02
## (Intercept)           .
## fixed.acidity         .
## volatile.acidity     -7.156854e+00
## citric.acid           .
## residual.sugar        2.407980e-01
## chlorides             5.484369e+00
## free.sulfur.dioxide   9.851367e-03
## total.sulfur.dioxide  .
## density              -4.010822e+02
## pH                    4.629667e-02
## sulphates             .
## alcohol               4.676187e-01
##
```

```
## $`9`
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)           98.56832794
## (Intercept)            .
## fixed.acidity          2.46338102
## volatile.acidity       .
## citric.acid            0.40586922
## residual.sugar         0.37889867
## chlorides           -289.67223963
## free.sulfur.dioxide    0.06032348
## total.sulfur.dioxide  -0.03565604
## density             -287.74588347
## pH                    16.75761392
## sulphates              2.87349833
## alcohol                .
```

Based on the result above, the least important factors for each of the wine quality levels are identified. For example, for wine quality 3, the total sulfur dioxide has the most positive

coefficient which means if the total sulfur dioxide increases, the wine is more likely to be wine quality 3. Same for the wine quality 9, if the chlorides decrease, the wine is more likely to be wine quality 9 since chloride has a negative coefficient, but it does not seem to be effective for wine quality 3 since the chlorides is ignored in that model. As a result, different wine elements may affect wine taste differently.

Modeling & Analysis

K-Means Clustering

Suppose the label (label column is quality) for each wine is unknown, K-Means clustering method can be applied to group data into several clusters. Each cluster contains objects who have strong similarities.

The first step of performing K-Means clustering is to scale the dataset. In this dataset, all columns are quantitative. Scaling dataset is useful to normalize the data within a particular range. Below is the summary of all columns except the label column after the scaling.

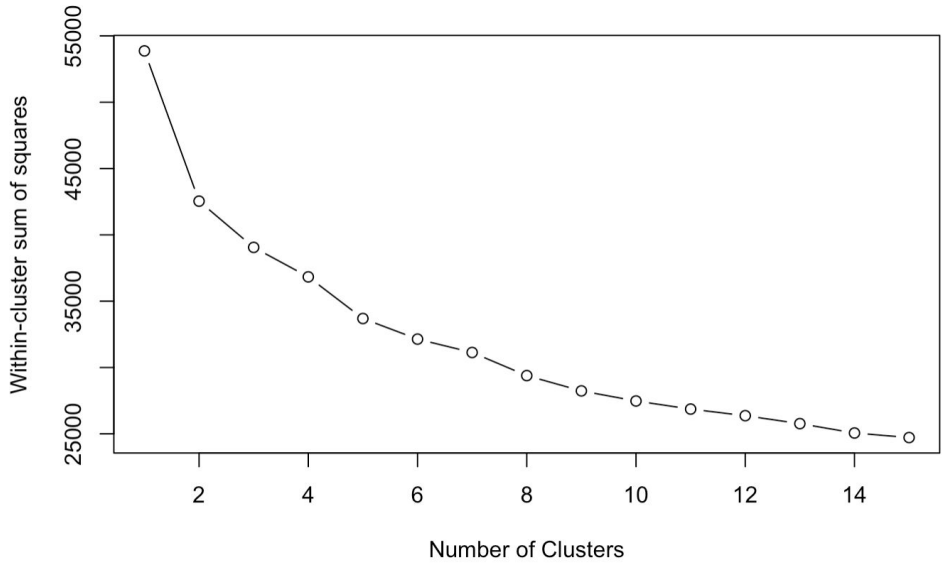
```
fixed.acidity      volatile.acidity   citric.acid      residual.sugar      chlorides
Min.   :-3.61998   Min.   :-1.9668   Min.   :-2.7615   Min.   :-1.1418   Min.   :-1.6831
1st Qu.:-0.65743   1st Qu.:-0.6770   1st Qu.:-0.5304   1st Qu.:-0.9250   1st Qu.:-0.4473
Median :-0.06492   Median :-0.1810   Median :-0.1173   Median :-0.2349   Median :-0.1269
Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
3rd Qu.: 0.52758   3rd Qu.: 0.4143   3rd Qu.: 0.4612   3rd Qu.: 0.6917   3rd Qu.: 0.1935
Max.   : 8.70422   Max.   : 8.1528   Max.   :10.9553   Max.   :11.7129   Max.   :13.7417
free.sulfur.dioxide total.sulfur.dioxide    density              pH            sulphates
Min.   :-1.95848    Min.   :-3.0439      Min.   :-2.31280   Min.   :-3.10109   Min.   :-2.3645
1st Qu.:-0.72370    1st Qu.:-0.7144      1st Qu.:-0.77063   1st Qu.:-0.65077   1st Qu.:-0.6996
Median :-0.07691    Median :-0.1026      Median :-0.09608   Median :-0.05475   Median :-0.1739
Mean   : 0.00000    Mean   : 0.0000      Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
3rd Qu.: 0.62867    3rd Qu.: 0.6739      3rd Qu.: 0.69298   3rd Qu.: 0.60750   3rd Qu.: 0.5271
Max.   :14.91679    Max.   : 7.0977      Max.   :15.02976   Max.   : 4.18365   Max.   : 5.1711
   alcohol
Min.   :-2.04309
1st Qu.:-0.82419
Median :-0.09285
Mean   : 0.00000
3rd Qu.: 0.71974
Max.   : 2.99502
```

A fundamental step for the K-Means algorithm is to determine the optimal number of clusters. Three methods can be applied here: Elbow method, Silhouette Scores method, and

In the elbow method, the total within-cluster sum of squares is used to decide how they are clustered. The smaller the total within-cluster sum of squares, the better the K-Means clustering

is. The sum of squares tends to decrease to 0 as *k* increases. The sum of squares is 0 when *k* is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster. Therefore, the goal is to choose a small value of *k* that still has a low sum of squares, and the elbow usually represents where it starts to have diminishing returns by increasing *k*.

This plot is generated by the elbow method, and it may be an elbow when *k* is 2. For sure, it is more like a fairly smooth curve, and it's unclear what is the best value of *k* to choose. Sometimes, the elbow method doesn't always work well; especially if the data is not very clustered.

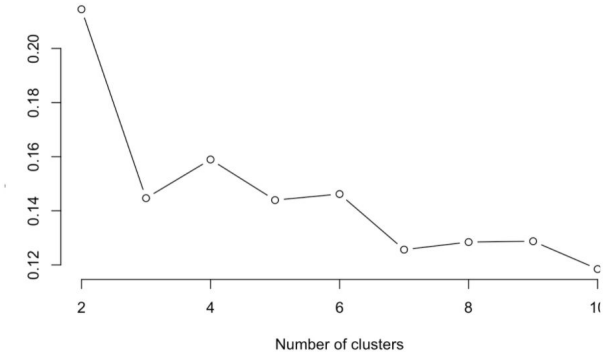


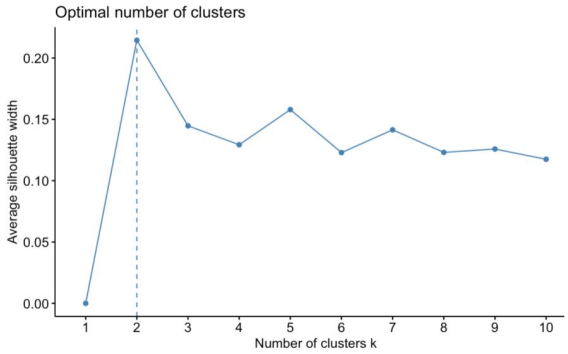
For sure, it is more like a fairly smooth curve, and it's unclear what is the best value of *k* to choose. Sometimes, the elbow method doesn't always work well; especially if the data is not very clustered.

In the Silhouette Scores method, it determines how well each data point lies within its cluster. For each instance, it calculates the mean distance to the other instances in the same cluster as y; it also calculates the mean distance to the instances of the next closet cluster as x. Based on the formula below, the silhouette score for each instance can be calculated. Ideally, if it is equal to 1, then it means this instance is in the perfect cluster; if it is -1, then it means the instance is in the worst cluster. Then taking the average of all silhouette scores of all data points to get the average silhouette score.

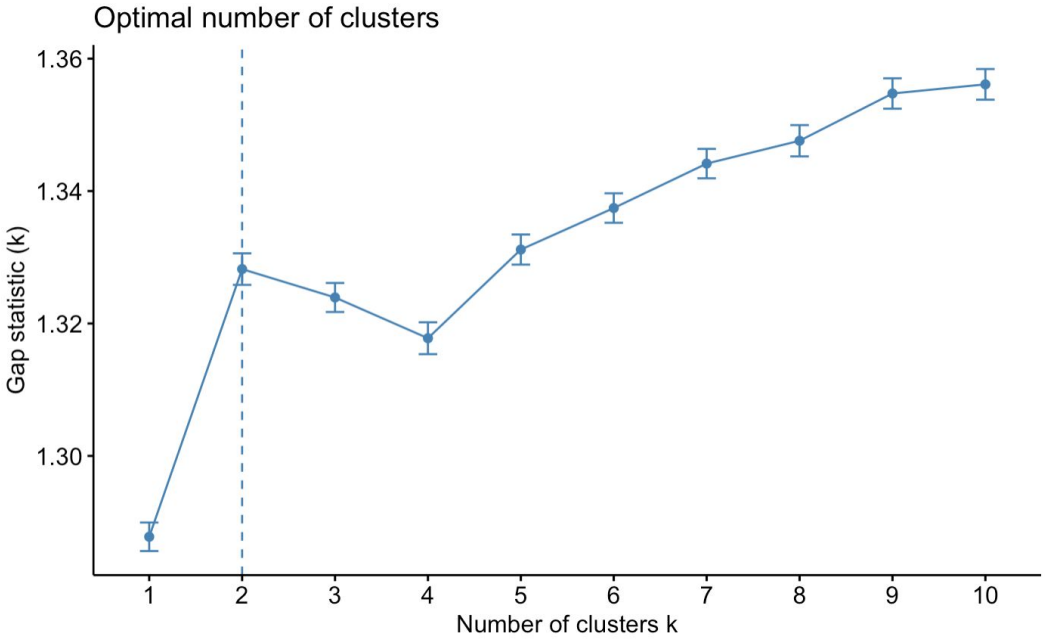
$$\textbf{\textit{Silhouette Score} = (x-y)/ max(x,y)}$$

In this case, when k is equal to 2, it has the highest average silhouette score.



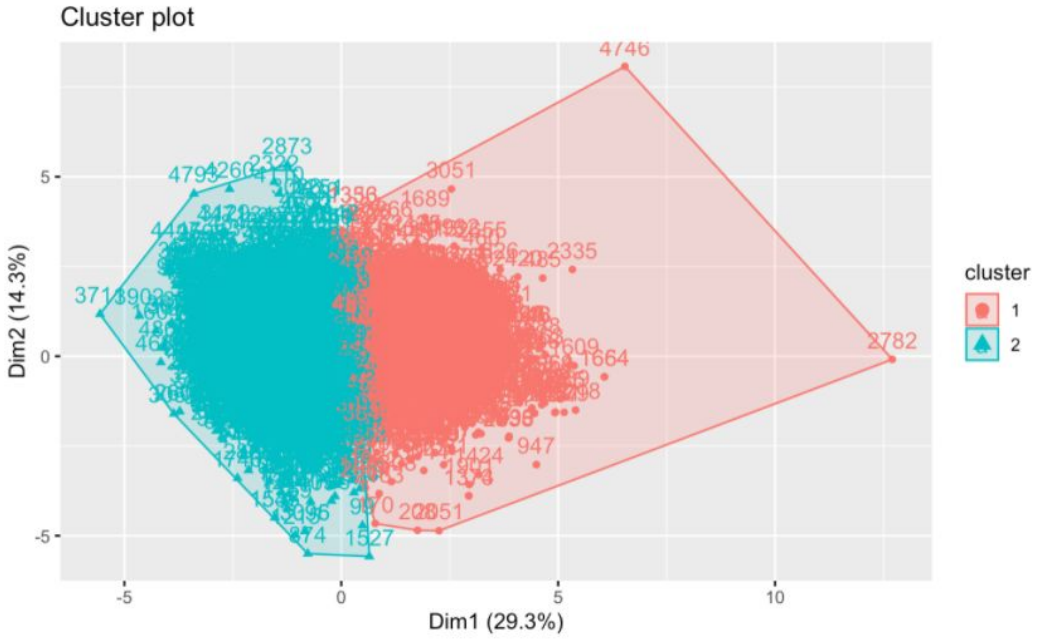


In the Gap Statistics method, it still gives the result as 2.



With these three approaches suggesting 2 as the number of optimal clusters, K-Means clustering analysis can be performed using 2 clusters. The first graph shows the size of two clusters, and the second one is the visualization of two clusters.

```
## K-means clustering with 2 clusters of sizes 1957, 2941
##
## Cluster means:
##   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## 1     0.1758942       0.04687908   0.2342012      0.8479233  0.3983139
## 2    -0.1170435      -0.03119428  -0.1558422     -0.5642251 -0.2650460
##   free.sulfur.dioxide total.sulfur.dioxide    density         pH
## 1           0.5972147            0.7656087  0.9473672 -0.2015099
## 2          -0.3973986           -0.5094513 -0.6303970  0.1340887
##     sulphates    alcohol
## 1  0.05986124 -0.7907243
## 2 -0.03983286  0.5261637
```

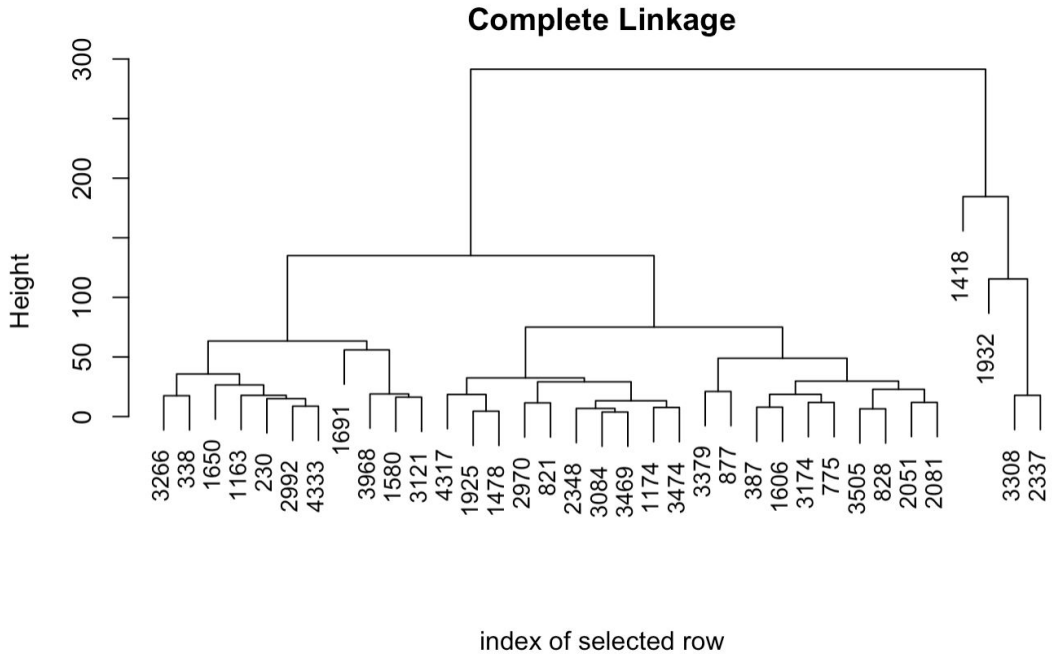


In the original dataset, there are seven groups. The quality column starts from 3 (low quality) to 9(high quality). However, K-Means clustering only gives two clusters. By going deep into the K-Means clustering results, it shows most wines with high quality in one cluster, and other wines with low quality in another cluster. Specifically, it shows wines with quality equal to 3, 4 (low quality) are all in one cluster; and wines with quality equal to 7,8,9 (high quality) are all in the other cluster. But for wines with quality equal to 5,6 (medium quality), the K-Means clustering does not do a good job. Some are in cluster A, and some are in cluster B. Therefore, the K-Means clustering can be used to separate very high quality wines or very low quality wines, but cannot separate wines with medium quality. In real life, most wines are medium quality. People always choose based on their taste, brand, and other preferences. From the wine production perspective, there are not many differences among medium quality wines, it is truly difficult to separate similar wines only based on their alcohol, volatile acidity, and other aspects.

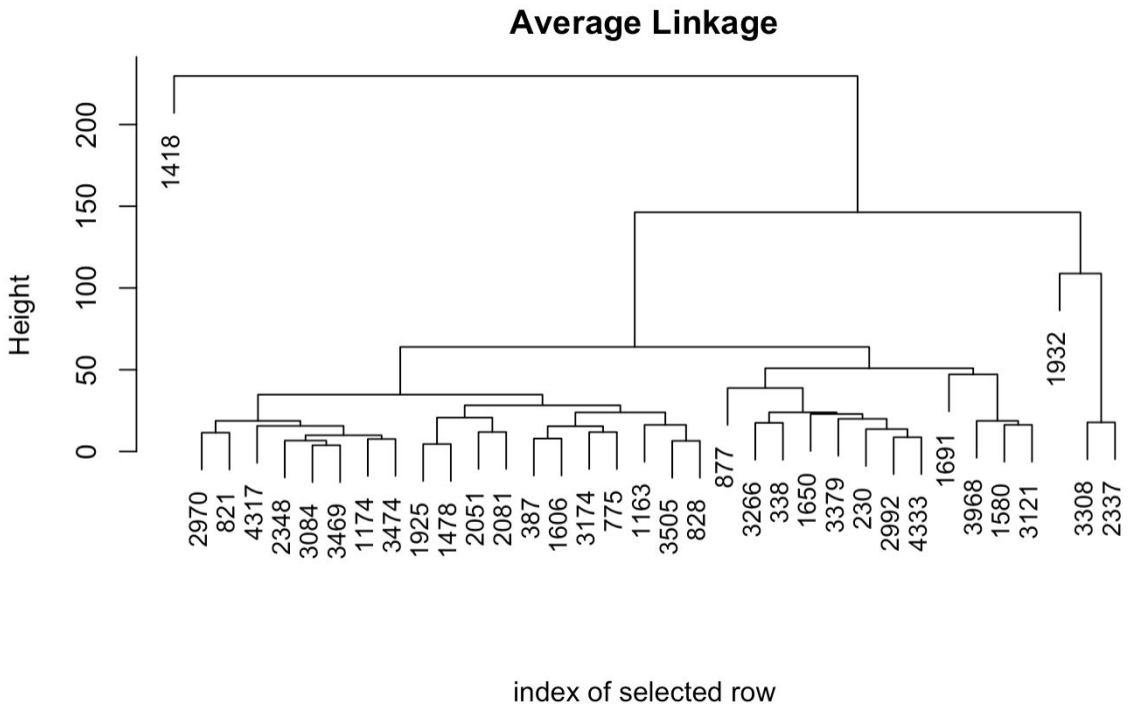
Therefore, even if K-Means clustering result is not very effective, it is still worthy to use for wine with very high quality and very low quality.

Hierarchical Clustering

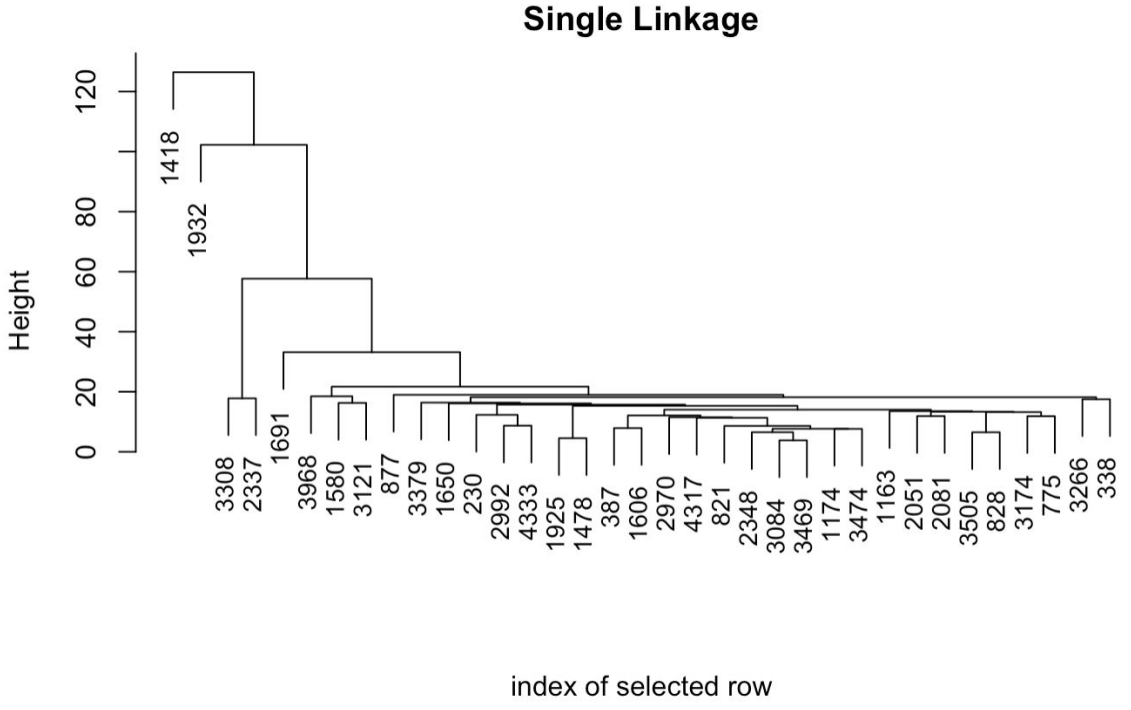
Hierarchical clustering is another clustering method to group data based on the similarities. However, one drawback of this method is that it cannot handle large dataset, but the K-Means clustering method can. The reason for this is that in hierarchical clustering algorithm, it calculates distance between every pair of points, so the time complexity is O(n^2). In order to perform the hierarchical clustering and make it easy to interpret, one way is to select several data from each original class, and then use this method to cluster them. More specifically, in the original dataset, the quality column gives each row a label, which is categorical data starting from 3 to 9. These are 7 groups. Then, 5 rows are randomly selected from each group, and then all these 35 rows become a new dataframe for hierarchical clustering.



This is the hierarchical clustering dendrogram using complete linkage, which computes all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and records the largest of these dissimilarities.



This is the hierarchical clustering dendrogram using average linkage, which computes all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and records the average of these dissimilarities.



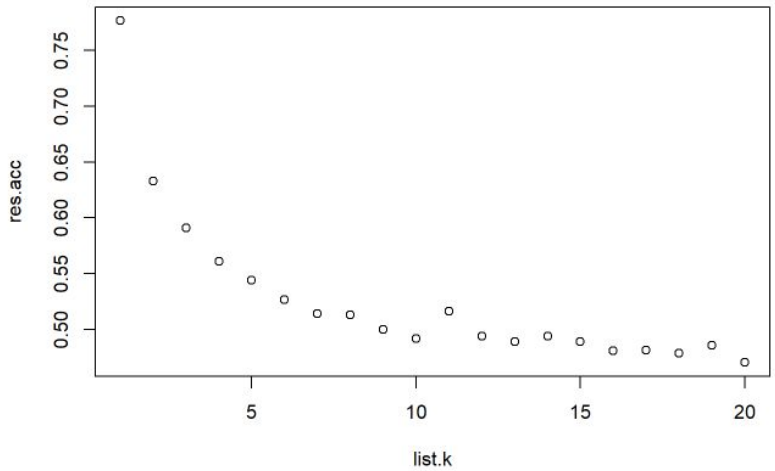
This is the hierarchical clustering dendrogram using single linkage, which computes all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and records the smallest of these dissimilarities.

Based on these three results, average and complete linkage tend to yield more balanced clusters. In complete linkage, those right four can be outliers which can be ignored. If the height is 100, then it split the dataset into four clusters. By looking at the original labels of these wines (remember the quality label starts from 3 (very low) to 9 (very high)), it is found that the distance between wines with quality 3, 4 or 5 are very small, this means they have strong similarities. And for wines with quality 7, 8, or 9, they have short distance, this means there are strong similarities among them. However, the distance between quality 3 and 9 is large, this means they have strong dissimilarities. This result is consistent with the dataset, wines with similar quality will have similar contents, and wines with huge different quality will have very different contents or ratios. Similar result in the average linkage which also gives four clusters if the height is around 50 and ignoring outliers.

Therefore, hierarchical clustering is also a good one to split the dataset into different clusters. However, no matter hierarchical clustering or K-Means clustering, they can only give cluster results instead of predicting labels. Therefore, they are very useful in the data mining process. To predict the label and find the most important factors in deciding the wine quality, other machine learning models will be made next.

K-nearest Neighbor

KNN is also used to predict the wine quality level using the given predictors. In order to choose the best K-value for the model, twenty different k-value is tested:



Based on the graph above, k=1 is chosen to be the optimal k-value for the knn model, which means that for an unknown observation n, n's classification only depends on the observation that is closest to n. The k = 1 model does make sense, since people believe a small difference in the component in wine could actually result in a big difference in terms of taste. As a result, the model with k=1 achieved 77.7% accuracy.

Support Vector Machine

Polynomial kernel:

```
Overall Statistics

               Accuracy : 0.5365
                 95% CI : (0.5106, 0.5622)
    No Information Rate : 0.4492
    P-Value [Acc > NIR] : 1.277e-11

                  Kappa : 0.2168
```

Radial kernel:

```
Overall Statistics

               Accuracy : 0.5842
                 95% CI : (0.5585, 0.6096)
    No Information Rate : 0.4492
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3184
```

Sigmoid kernel:

```
Overall Statistics

               Accuracy : 0.4458
                 95% CI : (0.4202, 0.4717)
    No Information Rate : 0.4492
    P-Value [Acc > NIR] : 0.6133

                  Kappa : 0.152
```

Linear kernel:

```
Overall Statistics

               Accuracy : 0.5256
                 95% CI : (0.4996, 0.5514)
    No Information Rate : 0.4492
    P-Value [Acc > NIR] : 2.704e-09

                  Kappa : 0.1972
```

After the K-nearest Neighbor model, the Support Vector Machine model has been used. The SVM model is a supervised machine learning model that uses classification algorithms for group classification problems. So this model was used with different types of kernel such as the linear, polynomial, sigmoid, radial.

According to the graph on the top, the accuracy for different kernels are 0.54 for polynomial, 0.58 for radial, 0.45 for sigmoid, and 0.53 for linear. It's obvious that the radial kernel provides the highest accuracy. But this accuracy is not good enough yet.

Furthermore, for the SVM model, it has several key parameters that need to be set appropriately to achieve the best classification results for a given problem. The parameters that result in a relatively good classification accuracy for the overall wine market, may result in a poor

classification accuracy for a niche market. So the SVM model itself has some limitations. For the next step, the other models such as decision tree and random forest would be used to obtain higher accuracy.

Decision Tree

Through the early exploration, the unbalanced data distribution was noticed. However, when training a decision tree model, the small number of records in quality score 3 and quality score 9 may lead to a negative effect for the accuracy of the model. To solve this constraint, the oversampling method was applied to the data cleaning and preprocessing step.

To build the better decision tree model, the original quality scores of white wine were split into 3 categories:

(1) If the quality score lower than 5, then assign it to "Low" quality level;
(2) If the quality score between 5 and 7, we assign it to "Medium" quality level;
(3) If the quality score higher than 7, we assign it to "High" quality level

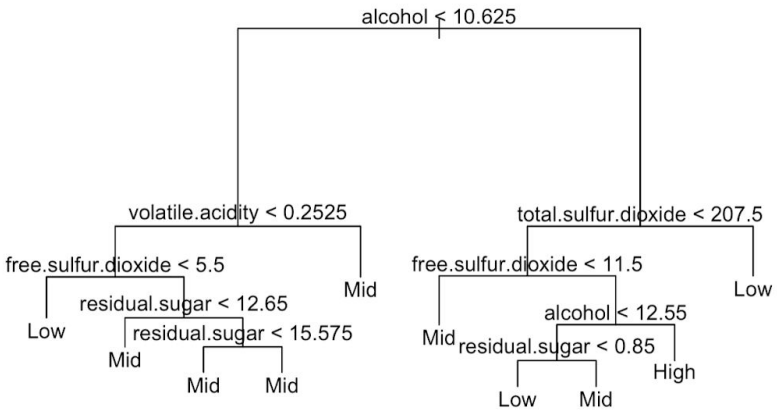
Then the original quality score column was removed since we consider the quality level as a new label column. Before we split the data into training and testing sets, we took a look at the summary of data. Now, in this dataset, we have 11 predictors and 1 categorical label column.

```
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 4.200   Min.   :0.080    Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
 1st Qu.: 6.300   1st Qu.:0.210    1st Qu.:0.2700   1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
 Median : 6.800   Median :0.260    Median :0.3100   Median : 5.000   Median :0.04300   Median : 34.00
 Mean   : 6.867   Mean   :0.277    Mean   :0.3325   Mean   : 6.282   Mean   :0.04519   Mean   : 36.05
 3rd Qu.: 7.300   3rd Qu.:0.320    3rd Qu.:0.3800   3rd Qu.: 9.700   3rd Qu.:0.05000   3rd Qu.: 46.00
 Max.   :14.200   Max.   :1.100    Max.   :1.6600   Max.   :65.800   Max.   :0.34600   Max.   :289.00
 total.sulfur.dioxide    density             pH          sulphates         alcohol      quality_degree
 Min.   : 10.0        Min.   :0.9874   Min.   :2.740   Min.   :0.2200   Min.   : 8.00   High:1091
 1st Qu.:107.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   Low : 275
 Median :134.0        Median :0.9937   Median :3.180   Median :0.4700   Median :10.40   Mid :3532
 Mean   :139.3        Mean   :0.9940   Mean   :3.189   Mean   :0.4919   Mean   :10.54
 3rd Qu.:168.0        3rd Qu.:0.9960   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
 Max.   :440.0        Max.   :1.0390   Max.   :3.810   Max.   :1.0800   Max.   :14.20
```

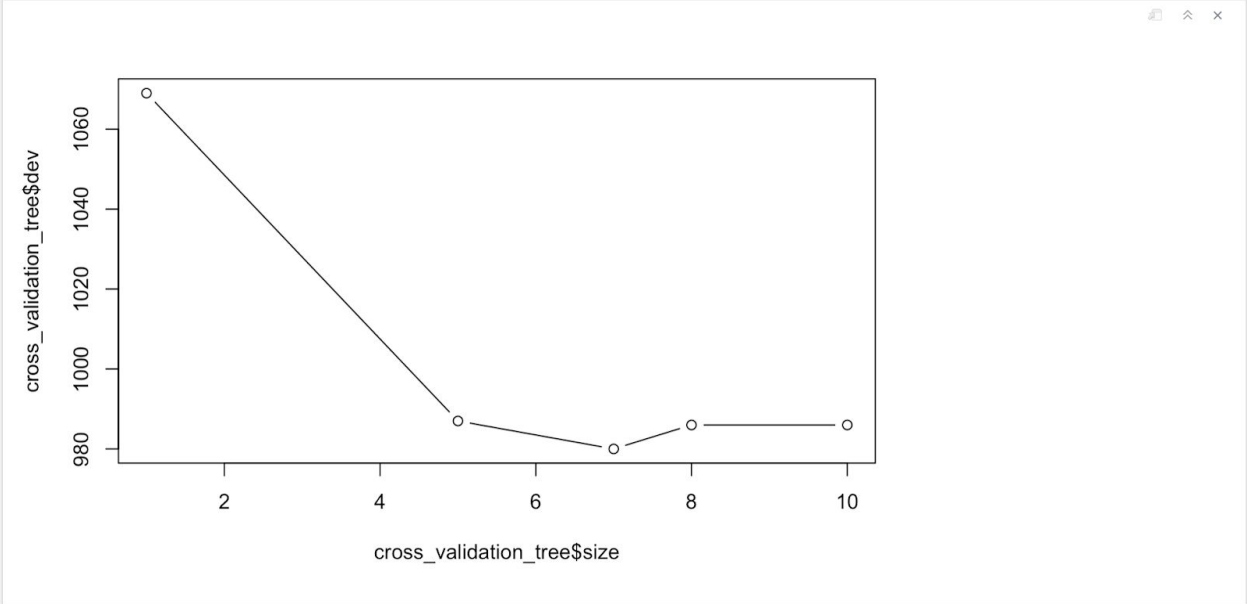
After this, 80% of data were randomly selected as the training set and 20% of data left composed as the testing set.

```
Classification tree:
tree(formula = quality_degree ~ ., data = training_set)
Variables actually used in tree construction:
[1] "alcohol"             "volatile.acidity"    "free.sulfur.dioxide" "residual.sugar"      "total.sulfur.dioxide"
Number of terminal nodes:  10
Residual mean deviance:  1.153 = 4507 / 3908
Misclassification error rate: 0.2422 = 949 / 3918
```

A decision tree model was trained by using the training data. According to the results, this decision tree used 5 variables to do the prediction and there are 10 terminal nodes in this tree.

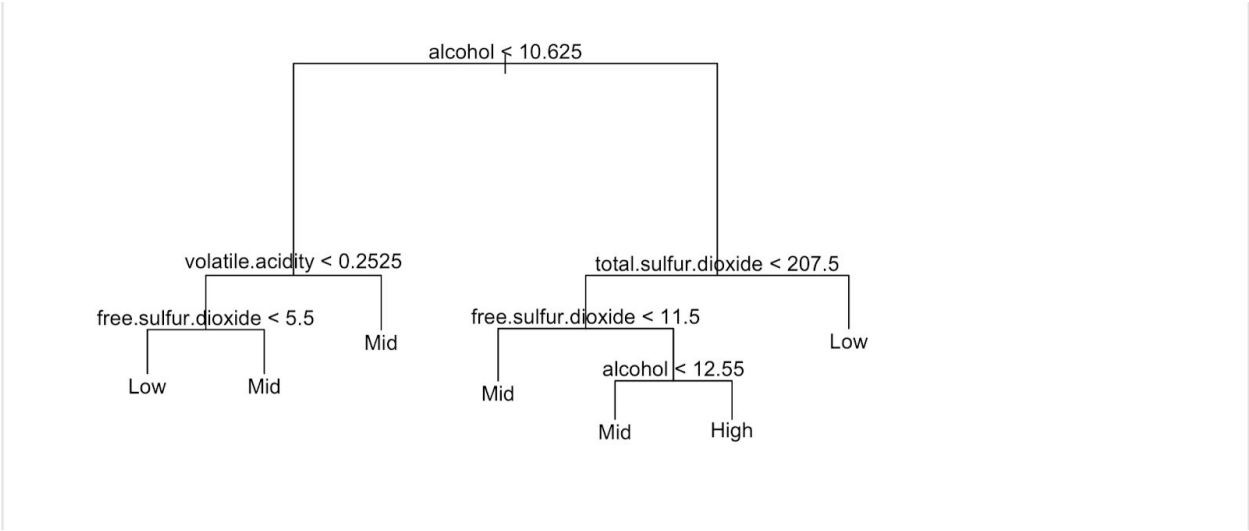


However, this tree seems complicated with 10 terminal nodes. The pruning method was applied to simplify the tree. The cross validation technique was used to identify the best number of terminal nodes that would have the smallest classification error.



According to this plot, 7 is the best number of terminal nodes. With above information, the pruned tree was built with testing accuracy achieved 0.72, which indicated that the pruned tree

was a relatively good model.

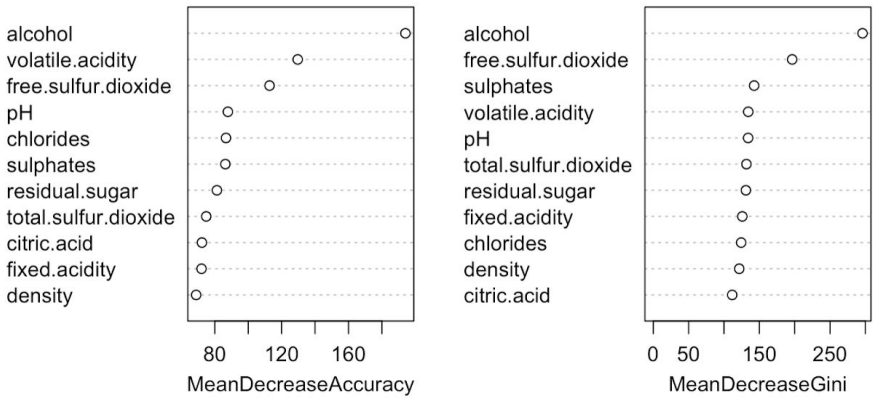


From this model, a clear observation was provided that the most important predictor is alcohol, followed by volatile acidity total sulfur dioxide and free sulfur dioxide. According to these results, deeper analysis and explanation will be provided in a random forest approach.

Random Forest

With previous decision tree results, 4 out of 11 predictors were included in the decision tree. Thus, random forest models were built with a number of subset variables equal to 4 and the number of trees equals to 500. The model evaluation was performed by using the testing set. According to the results, the random forest accuracy reached 0.91, which was a significant improvement compared with the decision tree model. A bagging approach was also applied to reduce the variance of this statistical learning method. Through this process, the testing set accuracy improved to 0.92.

To find the most important variables to predict the wine quality, importance measures were plotted. According to the graph below, the most important predictors are alcohol, volatile acidity, free sulfur dioxide and PH.

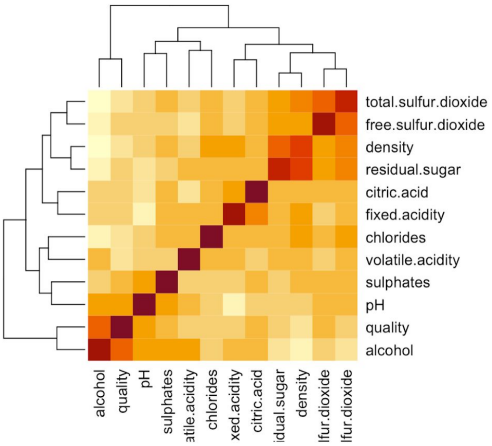


These results make a lot of sense. According to some research online, the wine with higher alcohol content will have a fuller, richer taste, while a lower alcohol wine will taste lighter; The volatile acidity can add fruit smelling like raspberry, passion fruit, or cherry-like flavors; The free sulfur dioxide could improve the taste and retain the wine's fruit flavors and freshness of aroma. Thus, the 3 most important variables are highly related to the taste of wine.

Generalized Additive Model

The Generalized Additive Model was applied for this dataset to predict the wine quality. Since the wine quality label is categorical data, a **logistic regression** of GAM was performed. To better perform the logistic regression, the quality scores were split into two categories: Above Average and Below Average.

Through the previous exploration in the random forest model, the most important variables were detected. Those variables will also be considered in the Generalized Additive Model. However, the collinearity problems should be considered in this situation. Thus, the scatter plots and a heatmap were drawn to observe the relationship between predictors.



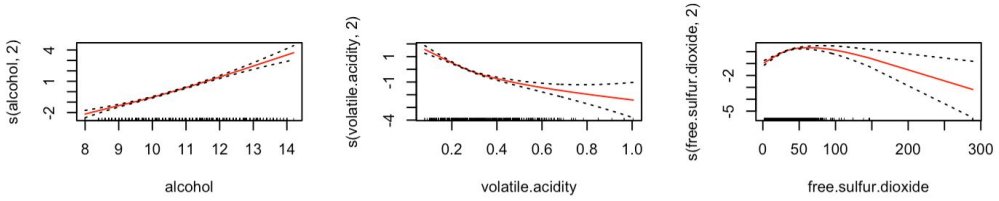
According to the scatter plots and heatmap, the strong correlation between total.sulfur.dioxide and free.sulfur.dioxide and correlation between density and residual sugar were discovered. Thus, the variable total.sulfur.dioxide and density will be removed to avoid collinearity problems.

Two models were created:

(1) GAM model with only 3 most important variables from random forest results

The degree of freedom = 2 has highest testing accuracy:

The brief visualization of the GAM model is shown below. The accuracy of the model is 73.97%



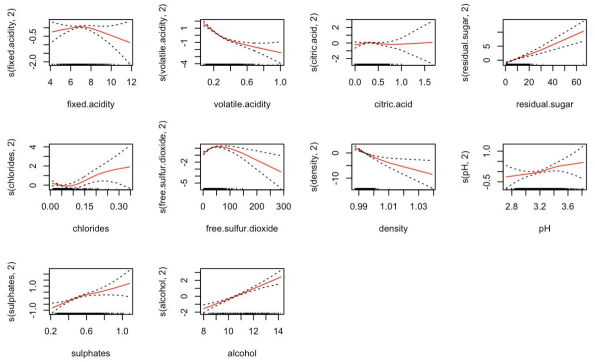

```
Anova for Parametric Effects
                          Df Sum Sq Mean Sq F value    Pr(>F)
s(alcohol, 2)              1  425.9  425.88 442.642 < 2.2e-16 ***
s(volatile.acidity, 2)     1  186.3  186.33 193.664 < 2.2e-16 ***
s(free.sulfur.dioxide, 2)  1   28.7   28.73  29.857 4.942e-08 ***
Residuals               3911 3762.9    0.96
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                        Npar Df Npar Chisq    P(Chi)
(Intercept)
s(alcohol, 2)                 1     5.917   0.01499 *
s(volatile.acidity, 2)        1    31.114 2.433e-08 ***
s(free.sulfur.dioxide, 2)     1   119.203 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(2) GAM model with 10 variables

The degree of freedom = 2 has highest testing accuracy:

The brief visualization of the GAM model is shown below. The accuracy of the model became lower than the previous model. Besides, some coefficients are not significant, which means this GAM model is not good enough although the GAM model is flexible.




```
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq  F value    Pr(>F)
## s(fixed.acidity, 2)       1   12.2  12.198  12.8356 0.0003442 ***
## s(volatile.acidity, 2)    1   95.5  95.533 100.5264 < 2.2e-16 ***
## s(citric.acid, 2)         1    0.8   0.781   0.8213 0.3648455
## s(residual.sugar, 2)      1   12.7  12.690  13.3536 0.0002613 ***
## s(chlorides, 2)           1   29.7  29.693  31.2454 2.430e-08 ***
## s(free.sulfur.dioxide, 2) 1    0.1   0.122   0.1284 0.7201280
## s(density, 2)             1  259.2 259.215 272.7634 < 2.2e-16 ***
## s(pH, 2)                  1   68.6  68.646  72.2335 < 2.2e-16 ***
## s(sulphates, 2)           1   56.1  56.079  59.0100 1.972e-14 ***
## s(alcohol, 2)             1   70.1  70.065  73.7271 < 2.2e-16 ***
## Residuals              3897 3703.4   0.950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                          Npar Df Npar Chisq    P(Chi)
## (Intercept)
## s(fixed.acidity, 2)            1     8.638 0.0032930 **
## s(volatile.acidity, 2)         1    36.371 1.630e-09 ***
## s(citric.acid, 2)              1     8.898 0.0028548 **
## s(residual.sugar, 2)           1     3.819 0.0506872 .
## s(chlorides, 2)                1    17.510 2.858e-05 ***
## s(free.sulfur.dioxide, 2)      1    88.972 < 2.2e-16 ***
## s(density, 2)                  1    12.085 0.0005083 ***
## s(pH, 2)                       1     2.404 0.1210085
## s(sulphates, 2)                1     4.252 0.0392118 *
## s(alcohol, 2)                  1     0.966 0.3256614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of the GAM model, an important information could be extracted: Although many physicochemical attributes will directly affect the wine taste and wine quality, the key indicators are still alcohol, volatile acidity and free sulfur dioxide, which will play the majority role of identifying the wine quality score.

Results & Conclusion

The goal of this project is to build machine learning models that could support wine tasting evaluations. To achieve this goal, the exploratory data analysis and multiple unsupervised learning as well as supervised learning models were provided. From the exploratory analysis, the data imbalance issue was detected. To solve this issue, the oversampling method was applied.

In the early stage of the research, the Multinomial Lasso Regression offered useful insight into the data that different elements might impact the likelihood of the wine quality classification differently. For example, chlorides have the most negative coefficient for wine quality 9, which means fewer chlorides makes the wine more likely to be wine quality 9. However, it did not seem to be a factor in the wine quality 3 model. So for wine manufactures, when they want to focus on a niche market with a certain quality level of wines, they could apply the similar method to find the most related elements.

In unsupervised learning, K-Means clustering and hierarchical clustering were performed. It showed that these two methods were useful to split very good quality and very bad quality wines, but not very effective on medium quality wines. The reason for this is that the quality for medium quality wines were hard to justify, and the ratio of contents were very similar. Therefore, in the future, to deal with a lot of wine with unknown quality, these two methods can be applied to separate data into groups, and for each group, a label prediction model can be performed to get the final result.

In supervised learning, the K-Nearest Neighbor, SVM, Decision Tree, Random Forest and GAM models were included.

For the K-Nearest Neighbor model, k = 1 model was selected due to the high accuracy. The choice of k = 1 as the optimal KNN model provided insight that small and subtle differences in the wine component could actually result in a huge deviation in terms of the wine taste. Therefore, the unknown observation was classified only based on one other observation that was closest.

The SVM model did not provide good accuracy, while the SVM model itself had some limitations. Even if the parameters that result in a relatively good classification accuracy for the overall wine market, may result in a poor classification accuracy for a niche market. Thus, the SVM model was not recommended for this case.

Among all the models, the Random Forest model has the highest accuracy with 92%. Through the importance measures plot, three most important factors in wine quality classification were noticed.

- **Alcohol**
- **Volatile acidity**
- **Free sulfur dioxide**

According to research, the wine with higher alcohol content will have a fuller, richer taste, while a lower alcohol wine will taste lighter; The volatile acidity can add fruit smelling like raspberry, passion fruit, or cherry-like flavors; The free sulfur dioxide could improve the taste and retain the wine's fruit flavors and freshness of aroma. Thus, the three most important variables are highly related to the taste of wine.

The GAM model provided insights that alcohol, volatile acidity and free sulfur dioxide still played important roles in evaluating the wine quality, while adding too many predictors may affect each other and yield a lower the model accuray. Thus, wine manufactures should focus on

the three most important variables instead of adding too many factors.

To conclude, the wine manufacturers should focus on alcohol, volatile acidity and free sulfur dioxide to improve the wine quality and wine taste. Of course, wine quality consists from the producers' perspective as well as from the consumers' perspective of objective, relative, and subjective factors. So these models can also help in target marketing by modeling customers' preference from different niche markets.

Reference

The State of Wine Industry Report 2019
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwju_7HplYLpAhVHlXIEHVKuCugQFjAAegQIAhAB&url=https%3A%2F%2Fwww.svb.com%2Fglobalassets%2Flibrary%2Fimages%2Fcontent%2Ftrends_and_insights%2Freports%2Fwine_report%2Fsvb-2019-wine-report&usg=AOvVaw3cgDwApE7569K-TzNx25sb

Dataset Reference:
https://archive.ics.uci.edu/ml/datasets/Wine+Quality