Introduction :

Wine Data analysis

Idea : to measure the importance of the attributes and their effect on the quality of Red wine & White wine and compare their effect to each type of wine
objectives : help stakeholders and wine makers to improve the quality of their products and increase their sales in the market Used data sets: Red Wine and White Wine data sets
parameters :12 numeric attributes
inputs :
1-The fixed acidity : represents the total acidity in the wine that cannot be eliminated during the aging process.
2-The volatile acidity : represents the total acidity in the wine that can be eliminated during the aging process.
3-The citric acid : content in wine represents the presence of organic acids derived from malic acid
4-The residual sugar : content represents the amount of sugar remaining in the wine after fermentation has been completed.
5-cholorides
6-The free sulfur dioxide
7-The total sulfur dioxide
8-The density : the ratio of the wine's mass to the volume of the wine.
9-The pH : represents the acidity or alkalinity of the wine.
10-sulphates
11- alcohol outputs: Quality : which we will measure the effect of each attribute on it

First step : import files
Second step : data cleaning

Before cleaning

```
> dim(df_red)
[1] 1599   12
> dim(df_white)
[1] 4898   12
```

Check duplicates

```
# check duplicates
sum(duplicated(df_red))
[1] 240
sum(duplicated(df_white))
[1] 937
```

Apply cleaning

```
df_red <- distinct(df_red)
df_white <- distinct(df_white)
```

After cleaning

```
> dim(df_red)
[1] 1359   12
> dim(df_white)
[1] 3961   12
```

There is no null values

```
> sum(is.na(df_red))
[1] 0
> sum(is.na(df_white))
[1] 0
```

```
# DATA CLEANING

# import data
df_red <- read.csv('D:\\redwine.csv')
df_white <- read.csv('D:\\whitewine.csv')
# dataframe dimensions before
dim(df_red)
dim(df_white)
# check duplicates
sum(duplicated(df_red))
sum(duplicated(df_white))
# remove duplicates
library(dplyr)
df_red <- distinct(df_red)
df_white <- distinct(df_white)
# check datatype of each column in red wine --> all ok
column_names <- names(df_red)
for (col in column_names) {
  print(is.numeric(df_red[[col]]))
}
# check datatype of each column in white wine --> all ok
column_names <- names(df_white)
for (col in column_names) {
  print(is.numeric(df_white[[col]]))
}
# check for null values
sum(is.na(df_red))
sum(is.na(df_white))
# dataframe dimensions after
dim(df_red)
dim(df_white)
```

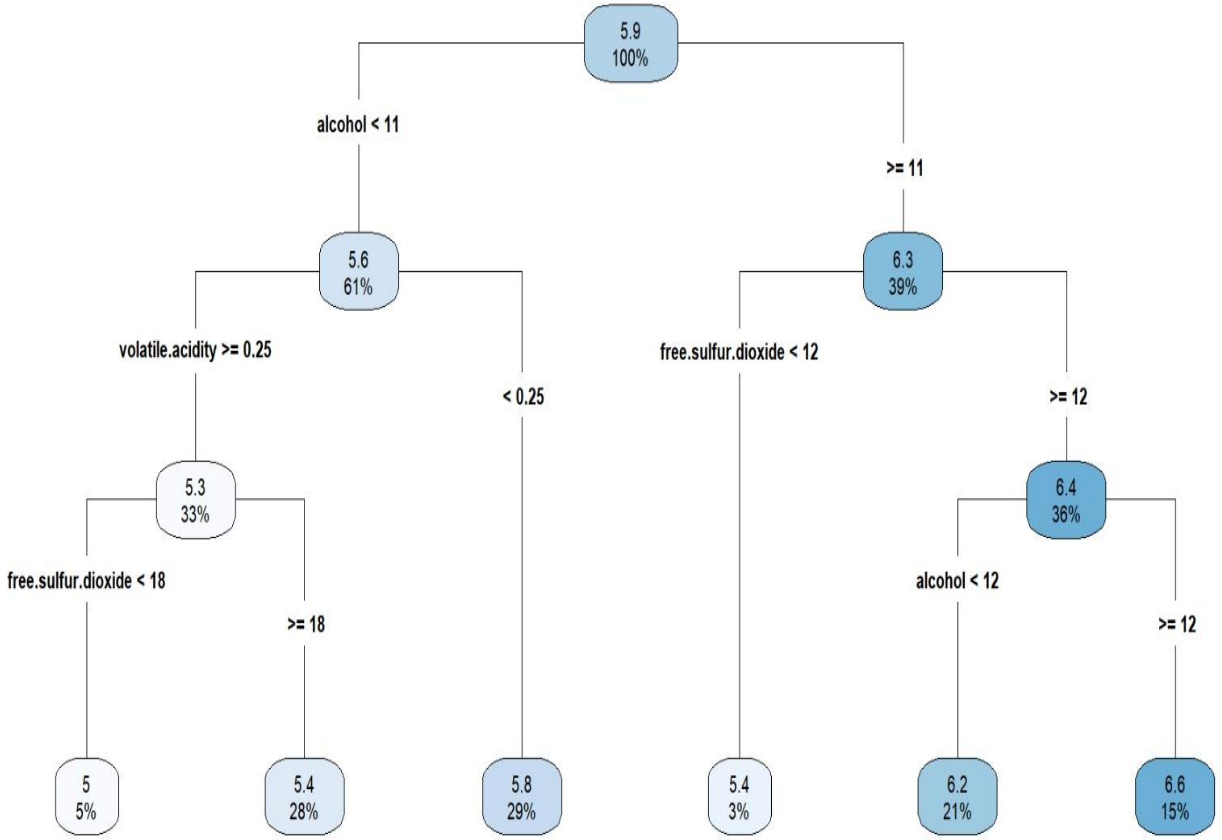
Third step : using supervised technique (decision tree)

1 - White_wine (decision tree) and their rules

```
> Wtree<-rpart(quality ~ chlorides +  volatile.acidity + fixed.acidity + alcohol+sulphates+pH+
density+total.sulfur.dioxide+free.sulfur.dioxide +residual.sugar +citric.acid,  data = whitewi
ne)
> rpart.plot(Wtree)
> whitewine_rules <- rpart.rules(Wtree)
> whitewine_rules
 quality
     5.0 when alcohol <  11       & free.sulfur.dioxide <  18 & volatile.acidity >= 0.25
     5.4 when alcohol >=       11 & free.sulfur.dioxide <  12
     5.4 when alcohol <  11       & free.sulfur.dioxide >= 18 & volatile.acidity >= 0.25
     5.8 when alcohol <  11                                   & volatile.acidity <  0.25
     6.2 when alcohol is 11 to 12 & free.sulfur.dioxide >= 12
     6.6 when alcohol >=       12 & free.sulfur.dioxide >= 12
> #identify root shape of the white_tree
> Wtree
n= 3961

node), split, n, deviance, yval
      * denotes terminal node

 1) root 3961 3141.53000 5.854835
   2) alcohol< 10.85 2433 1417.12800 5.567612
     4) volatile.acidity>=0.2525 1303  662.41750 5.349962
       8) free.sulfur.dioxide< 17.5 196   99.81633 4.969388 *
       9) free.sulfur.dioxide>=17.5 1107  529.18700 5.417344 *
     5) volatile.acidity< 0.2525 1130  621.80970 5.818584 *
   3) alcohol>=10.85 1528 1204.09400 6.312173
     6) free.sulfur.dioxide< 11.5 103  117.70870 5.359223 *
     7) free.sulfur.dioxide>=11.5 1425  986.08840 6.381053
      14) alcohol< 12.08333 850  565.55760 6.207059 *
      15) alcohol>=12.08333 575  356.75830 6.638261 *
>
```



Third step : using supervised technique (decision tree)

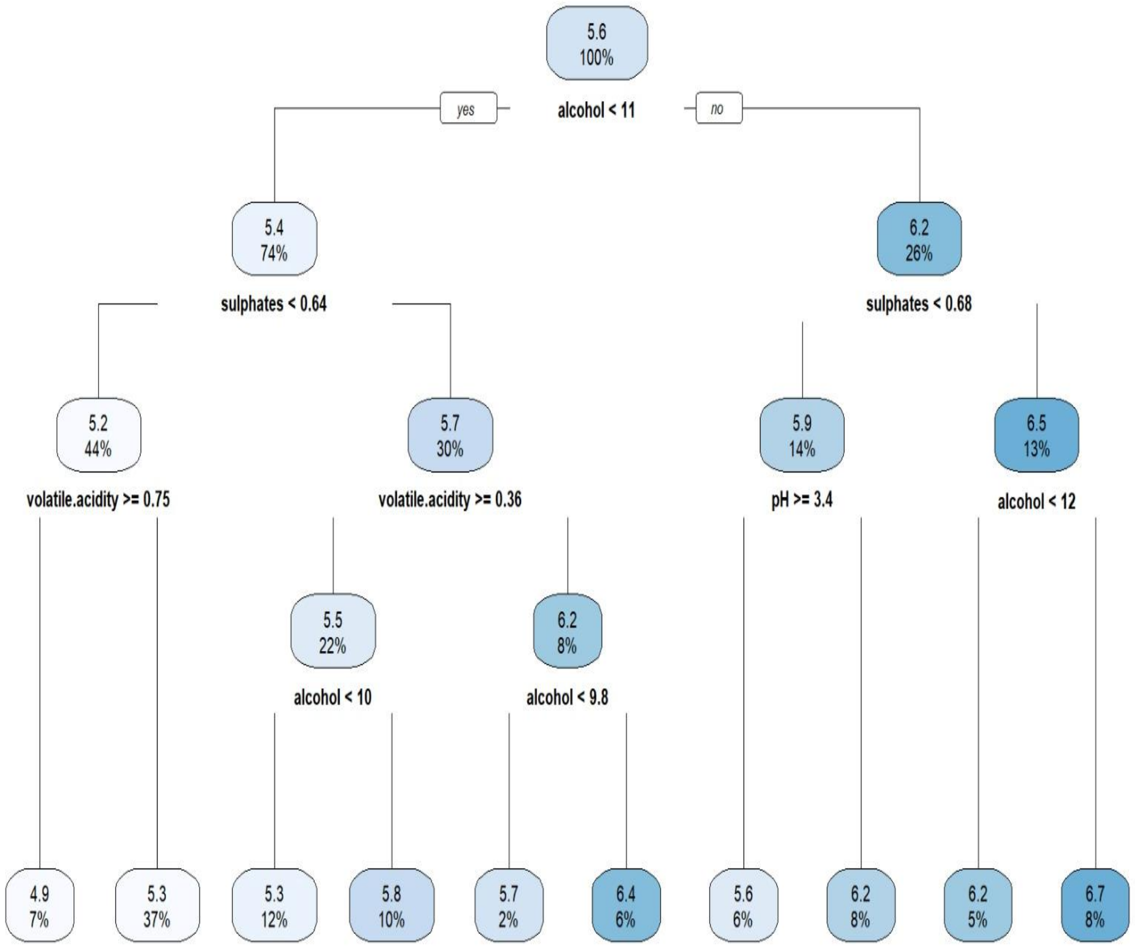
1 - Red_wine (decision tree) and their rules

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/
> Rtree<-rpart(quality ~ chlorides +  volatile.acidity + fixed.acidity + alcohol+sulphates+pH+density+total.sulfur.d
ioxide+free.sulfur.dioxide +residual.sugar +citric.acid,  data = redwine)
> rpart.plot(Rtree)
> #identify root&shap of the red wine tree
> Rtree
n= 1359

node), split, n, deviance, yval
      * denotes terminal node

 1) root 1359 921.105200 5.623252
   2) alcohol< 11.03333 1001 517.450500 5.417582
     4) sulphates< 0.635 598 244.690600 5.232441
       8) volatile.acidity>=0.7475 92  58.913040 4.891304 *
       9) volatile.acidity< 0.7475 506 173.124500 5.294466 *
     5) sulphates>=0.635 403 221.846200 5.692308
      10) volatile.acidity>=0.3625 301 135.010000 5.528239
        20) alcohol< 9.95 165  57.975760 5.321212 *
        21) alcohol>=9.95 136  61.382350 5.779412 *
      11) volatile.acidity< 0.3625 102  54.823530 6.176471
        22) alcohol< 9.75 26   7.884615 5.653846 *
        23) alcohol>=9.75 76  37.407890 6.355263 *
   3) alcohol>=11.03333 358 242.919000 6.198324
     6) sulphates< 0.675 187 116.267400 5.903743
      12) pH>=3.4 83  52.240960 5.578313 *
      13) pH< 3.4 104  48.221150 6.163462 *
     7) sulphates>=0.675 171  92.678360 6.520468
      14) alcohol< 11.55 66  32.439390 6.196970 *
      15) alcohol>=11.55 105  48.990480 6.723810 *
```

```
> redwine_rules <- rpart.rules(Rtree)
> redwine_rules
 quality
     4.9 when alcohol <  11.0         & sulphates <  0.64 & volatile.acidity >= 0.75
     5.3 when alcohol <  11.0         & sulphates <  0.64 & volatile.acidity <  0.75
     5.3 when alcohol <  10.0         & sulphates >= 0.64 & volatile.acidity >= 0.36
     5.6 when alcohol >=         11.0 & sulphates <  0.68                           & pH >= 3.4
     5.7 when alcohol <   9.8         & sulphates >= 0.64 & volatile.acidity <  0.36
     5.8 when alcohol is 10.0 to 11.0 & sulphates >= 0.64 & volatile.acidity >= 0.36
     6.2 when alcohol >=         11.0 & sulphates <  0.68                           & pH <  3.4
     6.2 when alcohol is 11.0 to 11.6 & sulphates >= 0.68
     6.4 when alcohol is  9.8 to 11.0 & sulphates >= 0.64 & volatile.acidity <  0.36
     6.7 when alcohol >=         11.6 & sulphates >= 0.68
>
```



Aim of decision tree :

1)determine which attribute plays a significant role in predicting the quality of red and white wines
2)Develop a predictive model that can classify wines into different quality categories .
3)After visual representation of the decision tree, Winemakers and stakeholders can use this information to make informed decisions about refining processes or adjusting attributes to improve wine quality
4)Understand and compare the factors affecting the quality of red and white wines separately

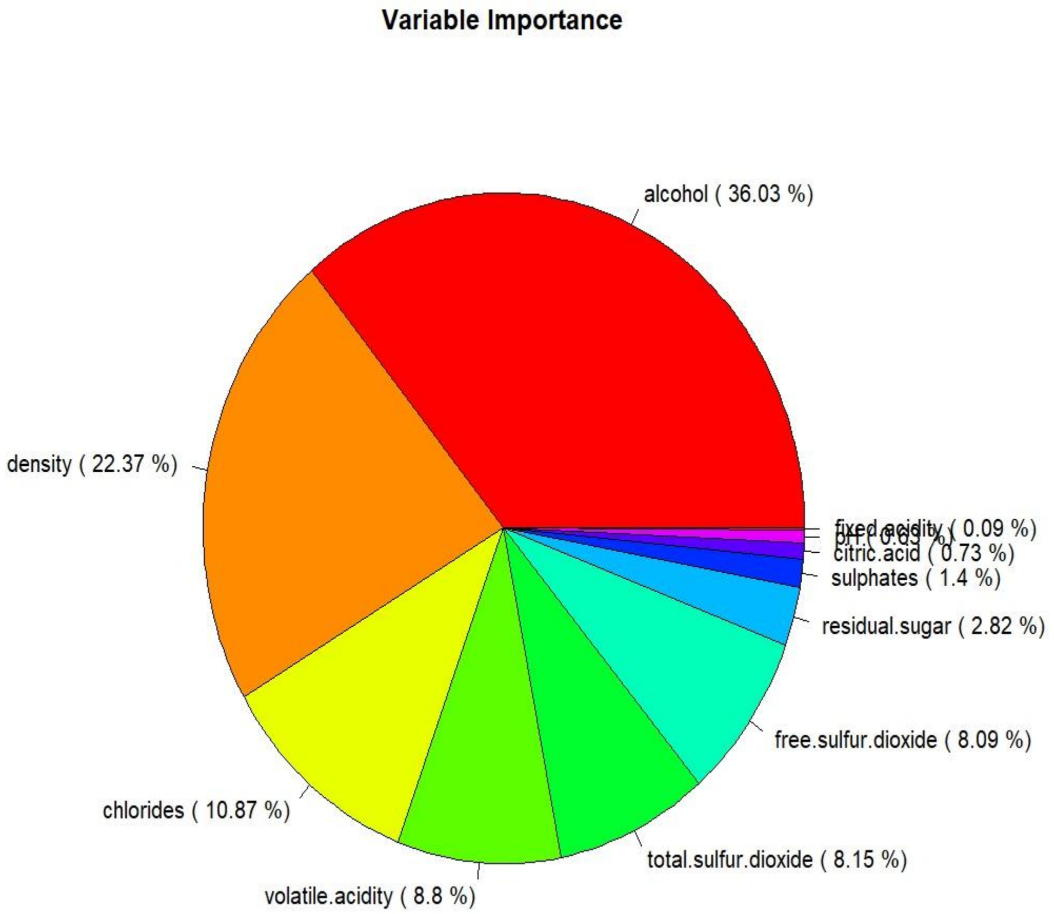
After that we can conclude the importance of each attribute from each decision tree

then using the pie chart from these trees representing the most important variables with percentage

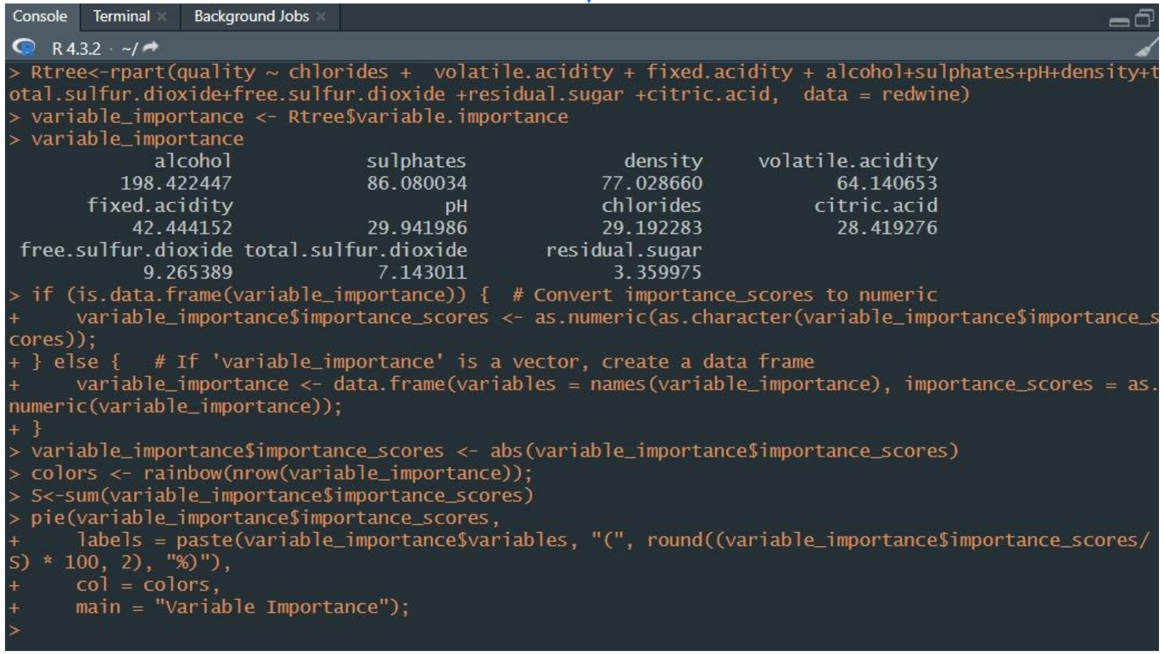
White _ wine (pie chart)

code

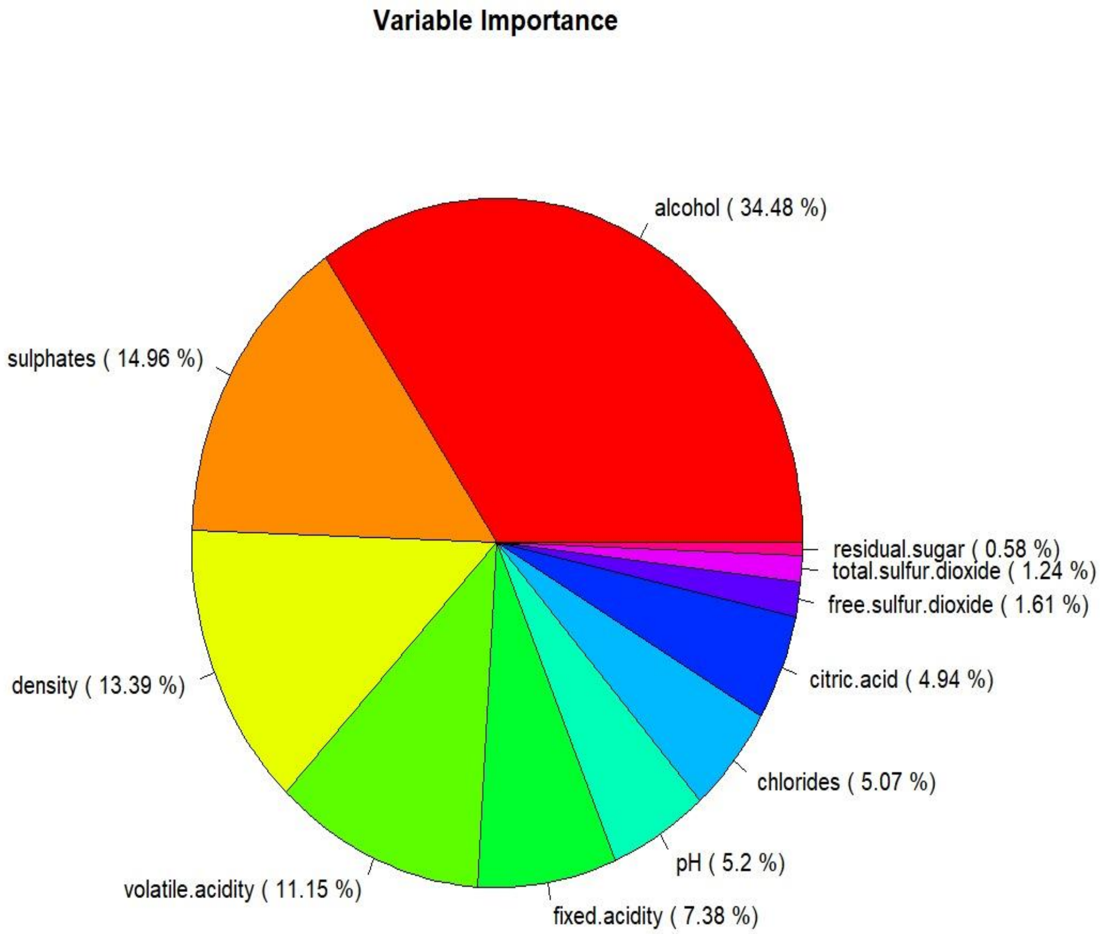
```
Console   Terminal   Background Jobs
R 4.3.2 · ~/
> Wtree<-rpart(quality ~ chlorides +  volatile.acidity + fixed.acidity + alcohol+sulphates+pH+density
+total.sulfur.dioxide+free.sulfur.dioxide +residual.sugar +citric.acid,  data = whitewine)
> variable_importance <- Wtree$variable.importance
> variable_importance
             alcohol              density            chlorides     volatile.acidity 
          595.606910           369.809314           179.676064           145.495334 
total.sulfur.dioxide  free.sulfur.dioxide       residual.sugar            sulphates 
          134.712650           133.710607            46.651587            23.155535 
         citric.acid                   pH        fixed.acidity 
           12.131023            10.467392             1.441813 
> if (is.data.frame(variable_importance)) {  # Convert importance_scores to numeric
+     variable_importance$importance_scores <- as.numeric(as.character(variable_importance$importance
_scores));
+ } else {   # If 'variable_importance' is a vector, create a data frame
+     variable_importance <- data.frame(variables = names(variable_importance), importance_scores = a
s.numeric(variable_importance));
+ }
> variable_importance$importance_scores <- abs(variable_importance$importance_scores)
> colors <- rainbow(nrow(variable_importance));
> S<-sum(variable_importance$importance_scores)
> pie(variable_importance$importance_scores,
+     labels = paste(variable_importance$variables, "(", round((variable_importance$importance_score
s/S) * 100, 2), "%)"),
+     col = colors,
+     main = "Variable Importance");
> 
```



Red _ wine (pie chart)

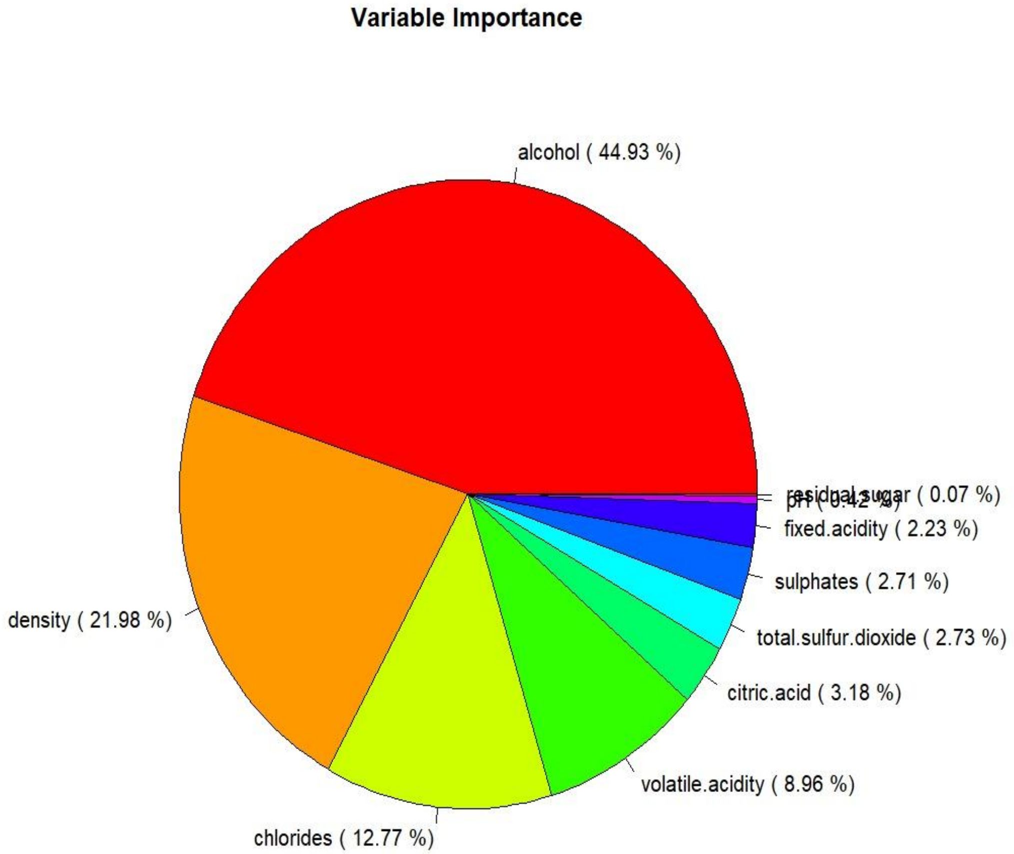


```
> Rtree<-rpart(quality ~ chlorides +  volatile.acidity + fixed.acidity + alcohol+sulphates+pH+density+total.sulfur.dioxide+free.sulfur.dioxide +residual.sugar +citric.acid,  data = redwine)
> variable_importance <- Rtree$variable.importance
> variable_importance
             alcohol            sulphates              density     volatile.acidity 
          198.422447            86.080034            77.028660            64.140653 
       fixed.acidity                   pH            chlorides          citric.acid 
           42.444152            29.941986            29.192283            28.419276 
 free.sulfur.dioxide total.sulfur.dioxide       residual.sugar 
            9.265389             7.143011             3.359975 
> if (is.data.frame(variable_importance)) {  # Convert importance_scores to numeric
+     variable_importance$importance_scores <- as.numeric(as.character(variable_importance$importance_scores));
+ } else {   # If 'variable_importance' is a vector, create a data frame
+     variable_importance <- data.frame(variables = names(variable_importance), importance_scores = as.numeric(variable_importance));
+ }
> variable_importance$importance_scores <- abs(variable_importance$importance_scores)
> colors <- rainbow(nrow(variable_importance));
> S<-sum(variable_importance$importance_scores)
> pie(variable_importance$importance_scores,
+     labels = paste(variable_importance$variables, "(", round((variable_importance$importance_scores/S) * 100, 2), "%)"),
+     col = colors,
+     main = "Variable Importance");
>
```



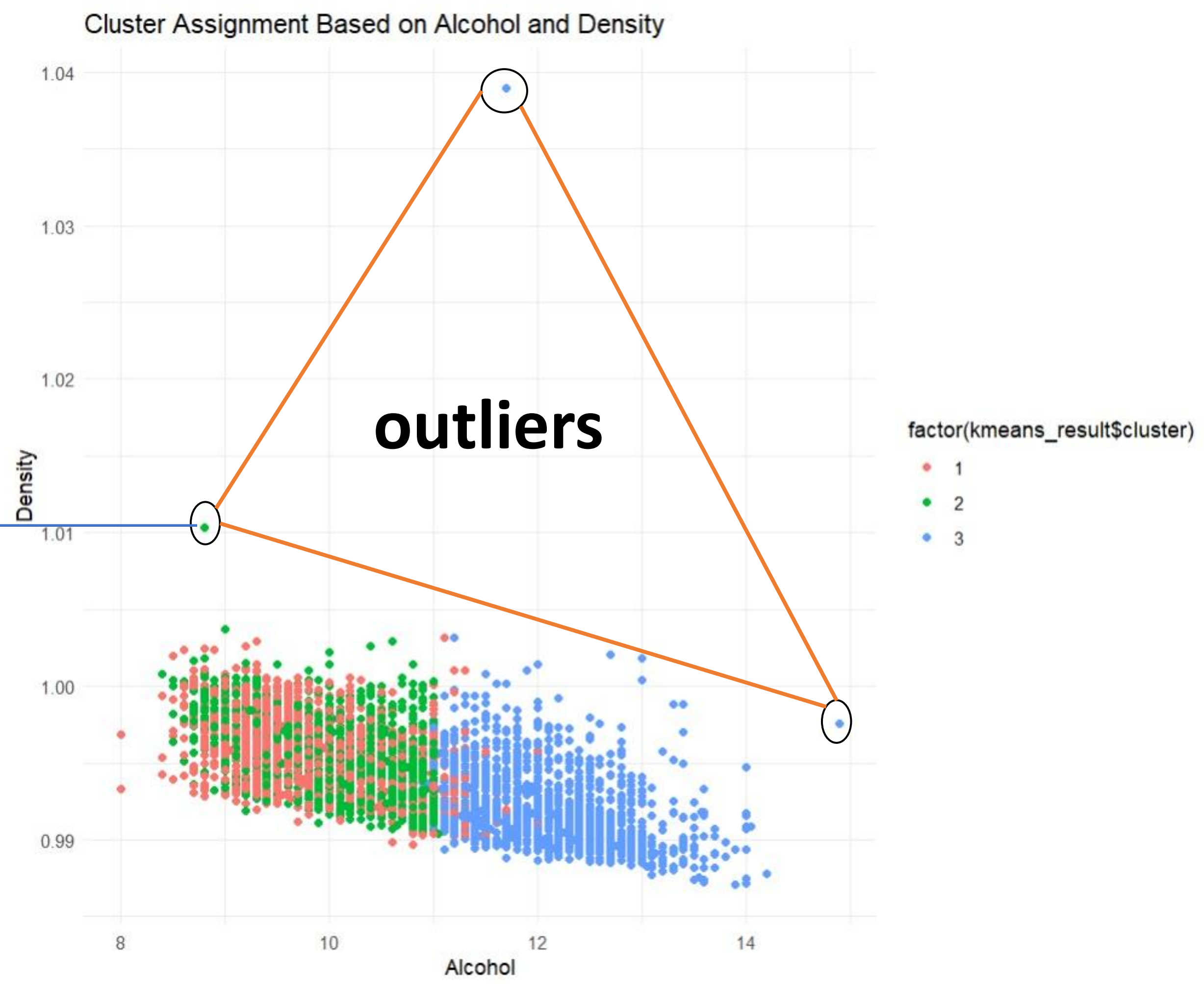
then we can form data frame which binds all wines together (ALLWINES) and apply decision tree technique on it to extract the importance variables, affect quality & represent it by pie chart

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/
> ALLWINES<-rbind(whitewine,redwine)
> winetree<-rpart(quality ~ chlorides +  volatile.acidity + fixed.acidity + alcohol+sulphates+pH+dens
y+total.sulfur.dioxide+free.sulfur.dioxide +residual.sugar +citric.acid,  data = ALLWINES)
> variable_importance <- winetree$variable.importance
> variable_importance
           alcohol              density            chlorides     volatile.acidity
        825.454481           403.829059           234.670639           164.564171
       citric.acid total.sulfur.dioxide            sulphates        fixed.acidity
         58.483688            50.219462            49.848677            41.002587
                pH       residual.sugar
          7.641388             1.293055
> if (is.data.frame(variable_importance)) {  # Convert importance_scores to numeric
+     variable_importance$importance_scores <- as.numeric(as.character(variable_importance$importance
cores));
+ } else {   # If 'variable_importance' is a vector, create a data frame
+     variable_importance <- data.frame(variables = names(variable_importance), importance_scores = a
numeric(variable_importance));
+ }
> variable_importance$importance_scores <- abs(variable_importance$importance_scores)
> colors <- rainbow(nrow(variable_importance));
> S<-sum(variable_importance$importance_scores)
> pie(variable_importance$importance_scores,
+     labels = paste(variable_importance$variables, "(", round((variable_importance$importance_scores
S) * 100, 2), "%)"),
+     col = colors,
+     main = "Variable Importance");
> 
```



After that we can pick up the most important attributes affect the quality and apply Clustering technique on it which are (alcohol & density) [unsupervised]

```
Console   Terminal   Background Jobs
R 4.3.2 · ~/
> data <- ALLWINES[, c("alcohol", "density", "quality")]
> kmeans_result <- kmeans(data, centers = 3, nstart = 20)
>
> library(ggplot2)
>
> ggplot(ALLWINES, aes(x = alcohol, y = density, color = factor(kmeans_result$clus
ter))) +
+     geom_point() +
+     labs(title = "Cluster Assignment Based on Alcohol and Density",
+          x = "Alcohol",
+          y = "Density") +
+     theme_minimal()
>
```



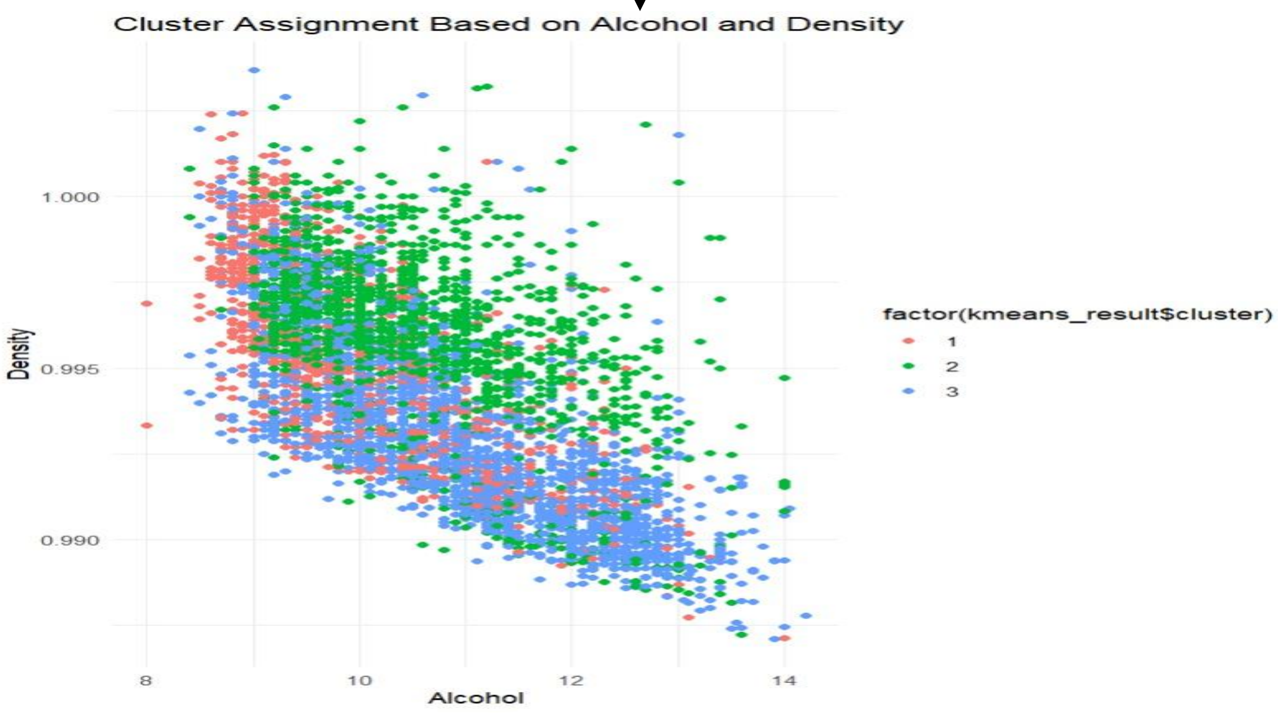
As the presence of outliers in clustering can distort cluster shapes, shift centers, and influence the overall structure, leading to biased results. Outliers may affect the performance of distance-based methods , Filtering outliers from the data before applying clustering techniques is beneficial because outliers can distort cluster shapes, shift centroids, and introduce noise, leading to biased and less reliable clustering results
, so we remove it to improve the accuracy and stability of clustering algorithm

Applying filter of outliers :

Console Terminal Background Jobs

R 4.3.2 · ~/

```
> ALLWINES <- rbind(whitewine, redwine)
> filtered_data <- subset(ALLWINES, ALLWINES$alcohol <= 14.5 & ALLWINES$density < 1.01)
> kmeans_result <- kmeans(filtered_data, centers = 3, nstart = 20)
> ggplot(filtered_data, aes(x = alcohol, y = density, color = factor(kmeans_result$cluster))) +
+     geom_point() +
+     labs(title = "Cluster Assignment Based on Alcohol and Density",
+          x = "Alcohol",
+          y = "Density") +
+     theme_minimal()
> filtered_data$cluster<- kmeans_result$cluster
> ggplot(filtered_data, aes(x = cluster, y = quality, color = factor(cluster))) +
+     geom_point() +
+     labs(title = "Cluster Assignment Based on Alcohol and Density",
+          x = "Cluster",
+          y = "Quality") +
+     theme_minimal()
>
```

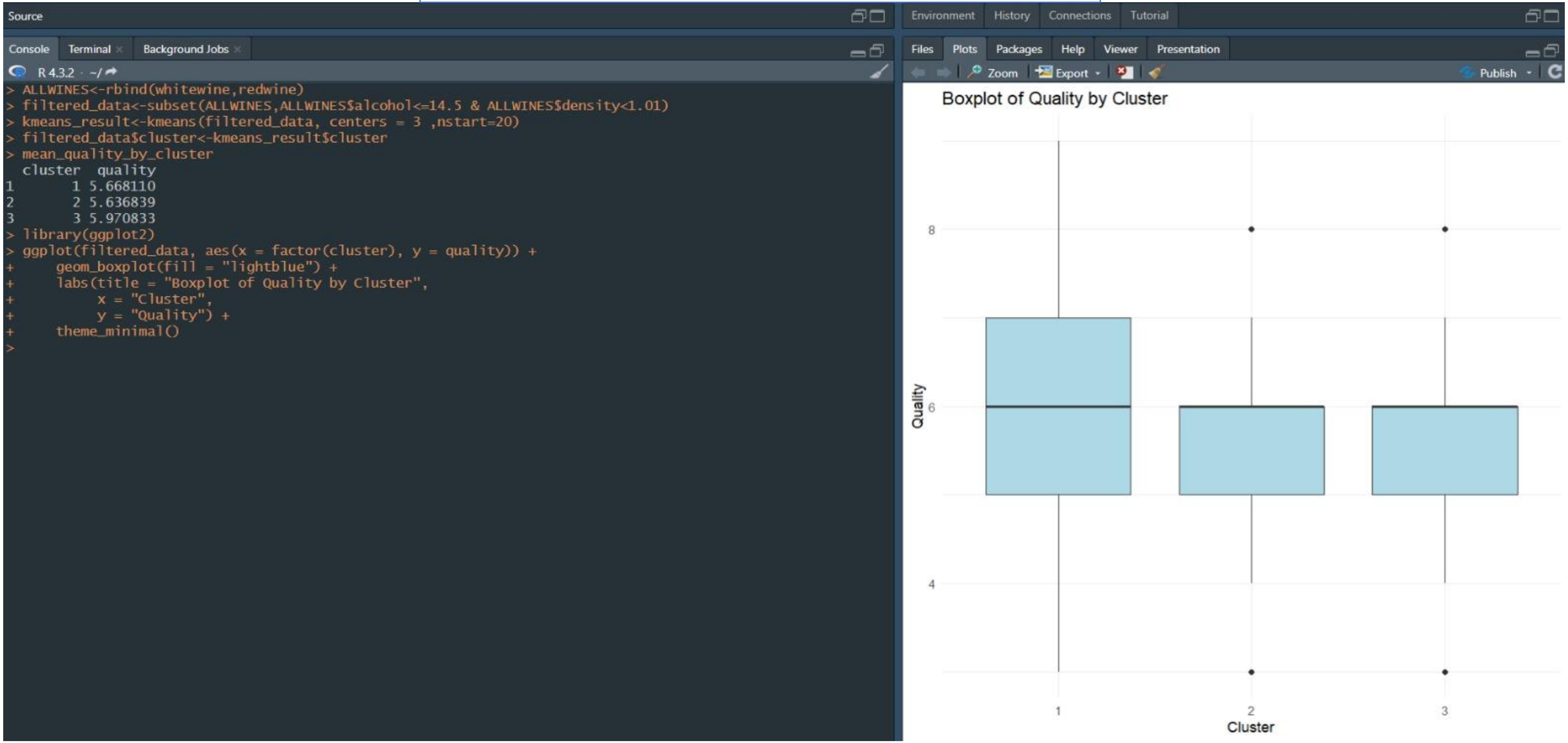


Aim of clustering:

1-Enhance Decision-Making: Facilitate data-driven decision-making for winemakers by providing insights into quality variations

2-Group wines with similar alcohol and density profiles to segment the wine market and enable targeted quality assessments for each cluster

Quality distribution by Cluster



Challenges in the dataset

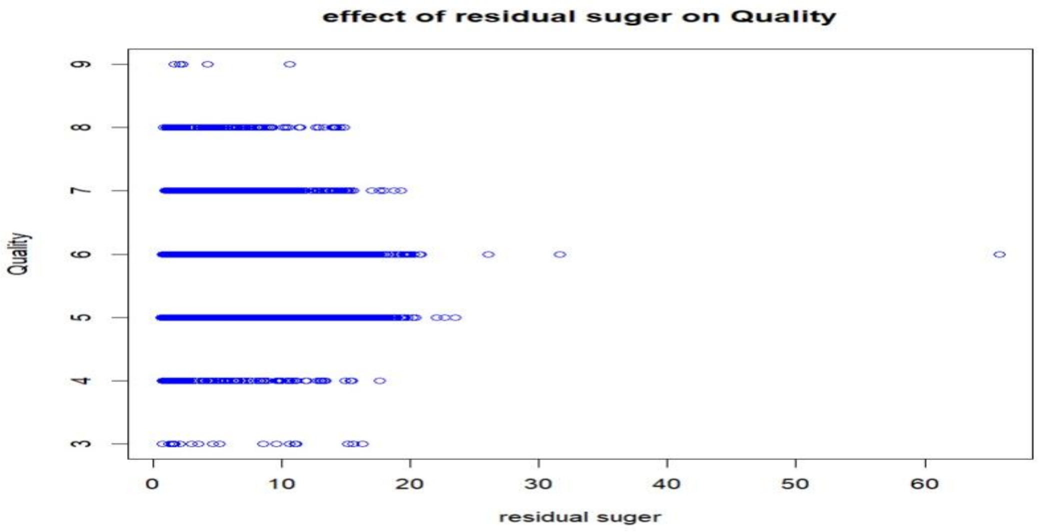
1- Outliers can significantly impact your analysis ,
It's essential to identify and decide whether to remove or handle them appropriately

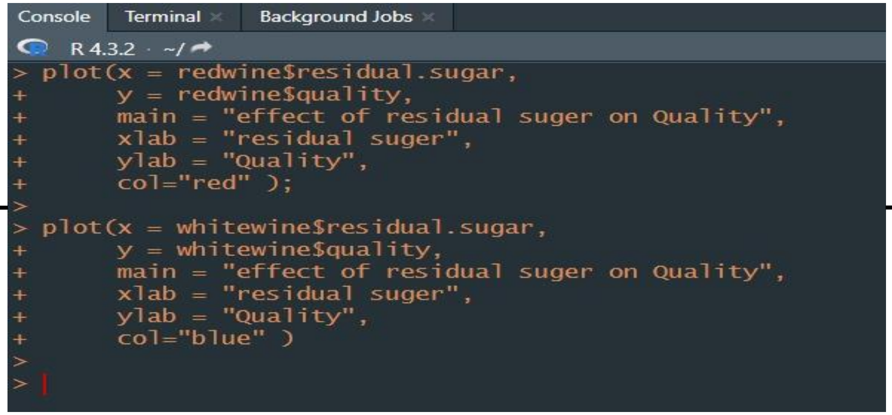
2 - Data Quality:
Ensure that the dataset is clean and accurate. Address any missing values or errors in the data

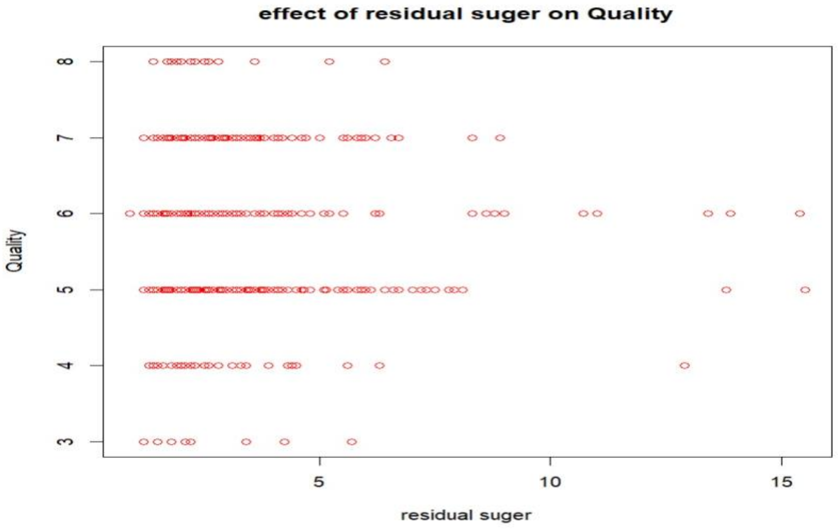
3 - dataset size is small :
We couldn't apply the machine learning techniques with high confidence

4 - Lack of Domain Expertise

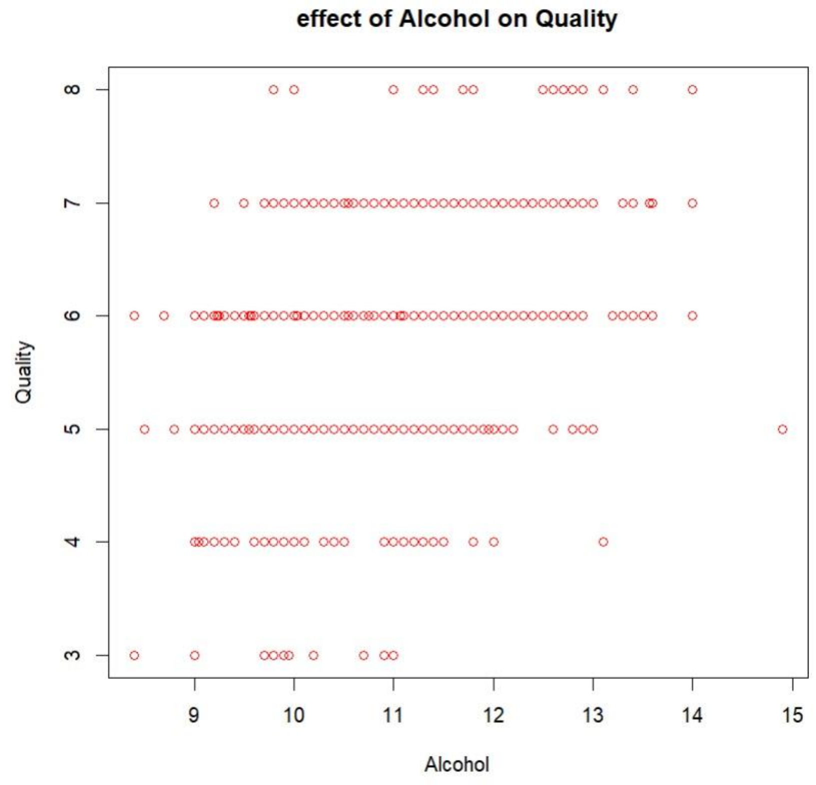
Interpretations of the results

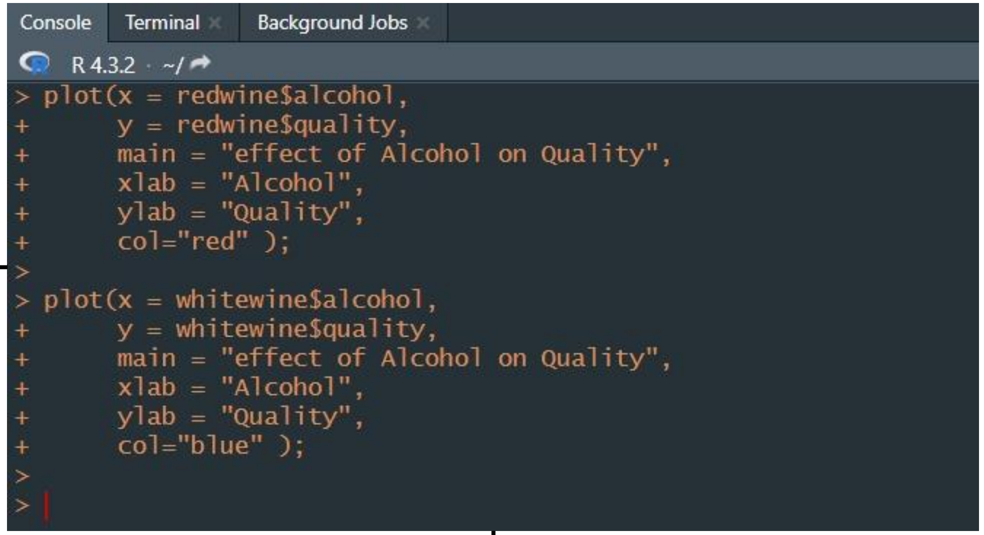




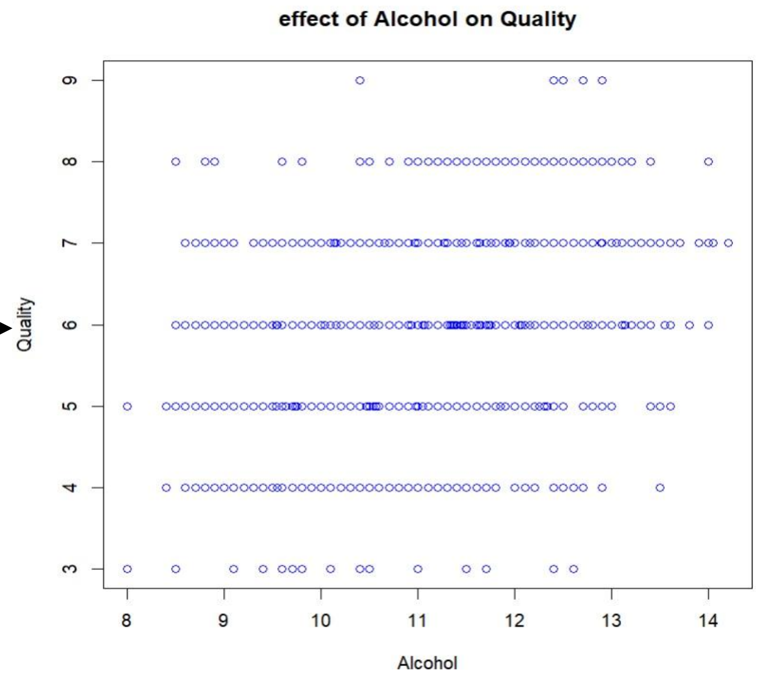


RESIDUAL SUGAR:
it has been analyzed that
ALL white wine has residual sugars in range less than 20 which don't affect the quality either than its increase or decrease
Same note in red wine all in same range distributed in different qualities
this analysis symphsizes the non importance of residual sugar on determine the quality of wine in general



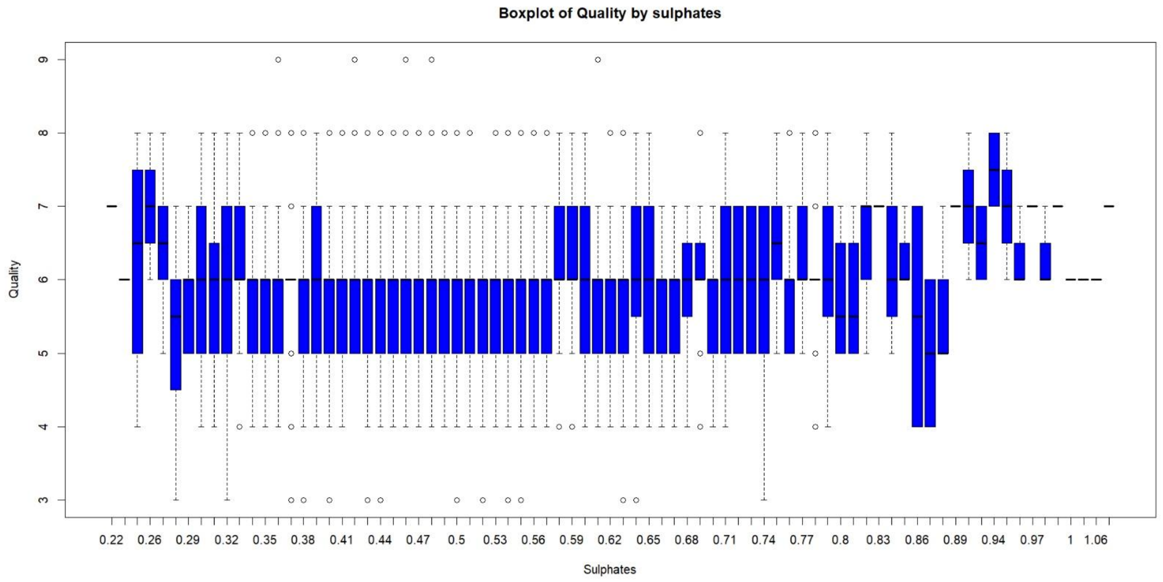


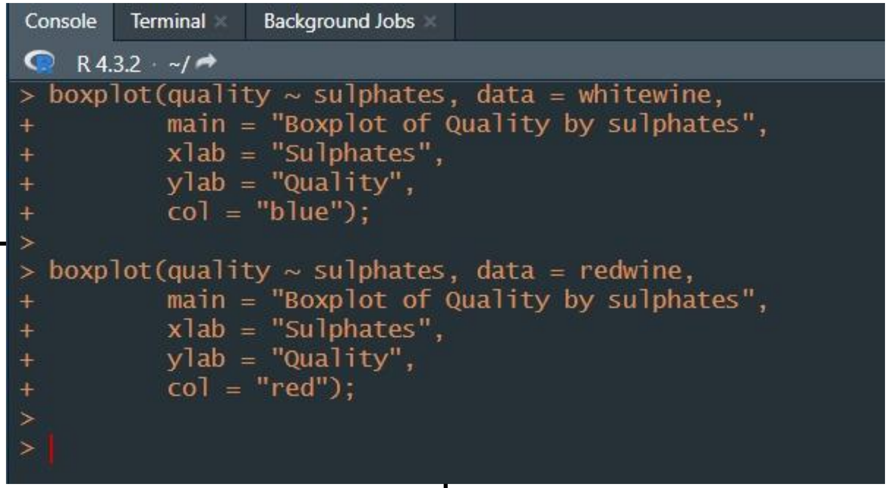
```
Console   Terminal   Background Jobs
R 4.3.2 · ~/
> plot(x = redwine$alcohol,
+      y = redwine$quality,
+      main = "effect of Alcohol on Quality",
+      xlab = "Alcohol",
+      ylab = "Quality",
+      col="red" );
>
> plot(x = whitewine$alcohol,
+      y = whitewine$quality,
+      main = "effect of Alcohol on Quality",
+      xlab = "Alcohol",
+      ylab = "Quality",
+      col="blue" );
>
>
```



ALCOHOLS:

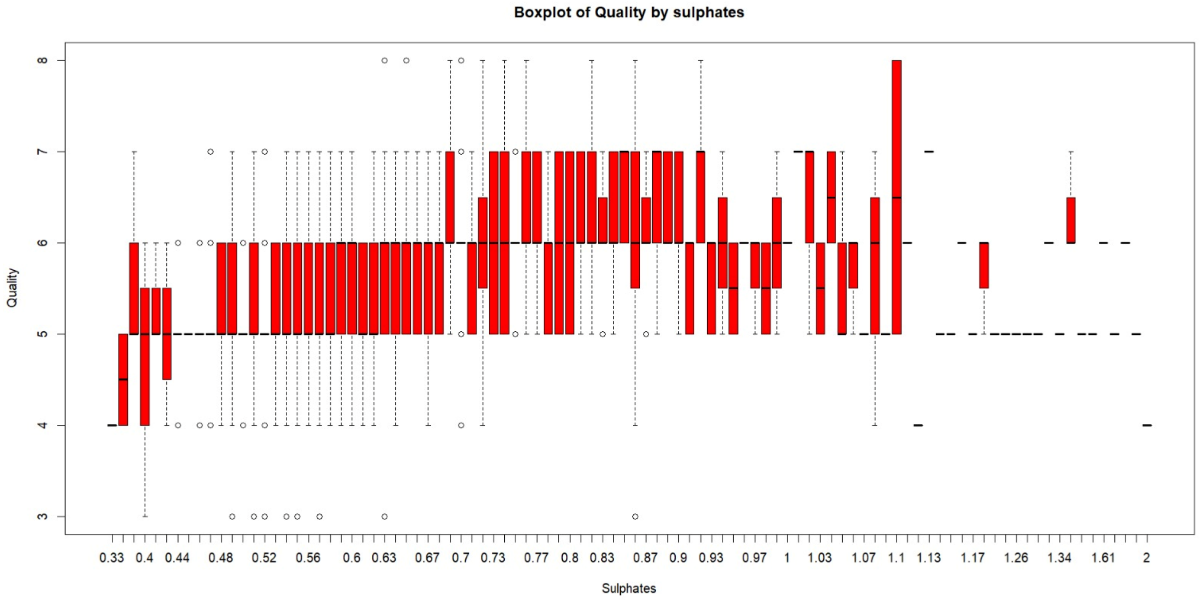
it has been analyzed that alcohol is the most important factor in controlling the quality of wine in general

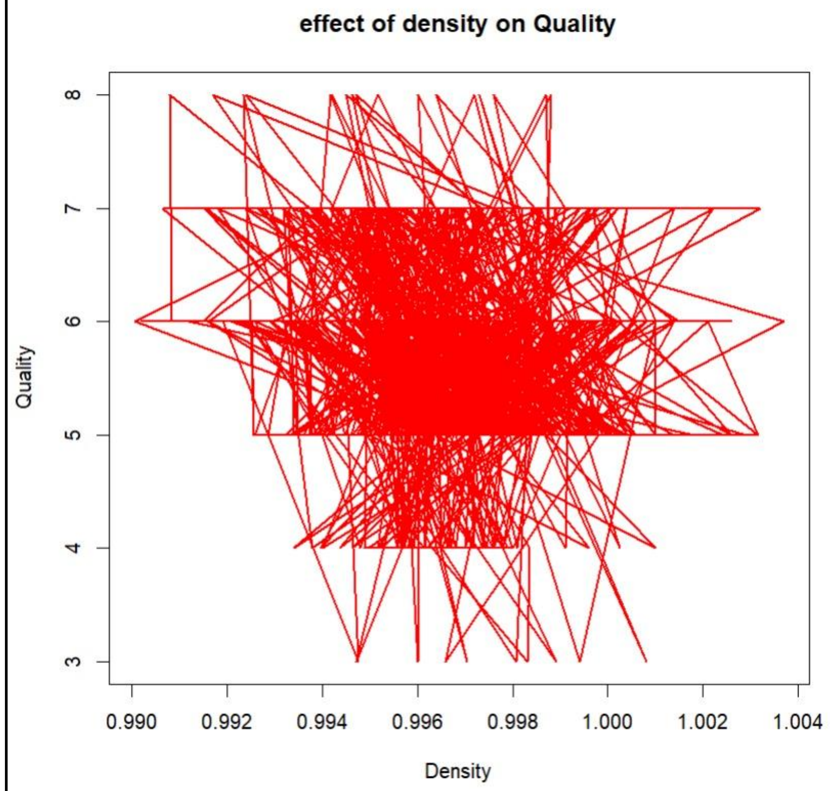




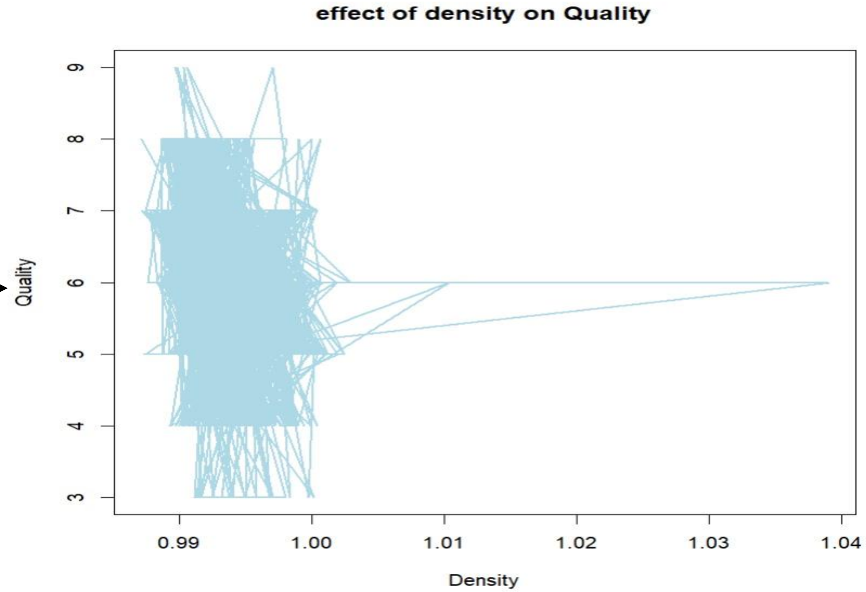
SULPHATES:

it has been analyzed that Sulphates has great effect on quality of Red wine ,direct proportional with Quality Sulphates has very less effect on the Quality of white wine






```
Console  Terminal  Background Jobs
R 4.3.2 · ~/
> plot(redwine$density, redwine$quality,
+       type = "l",
+       main = "effect of density on Quality",
+       xlab = "Density",
+       ylab = "Quality",
+       col = "red",
+       lwd = 2);
> plot(whitewine$density, whitewine$quality,
+       type = "l",
+       main = "effect of density on Quality",
+       xlab = "Density",
+       ylab = "Quality",
+       col = "lightblue",
+       lwd = 2);
>
```



DENSITY :

it has been analyzed that

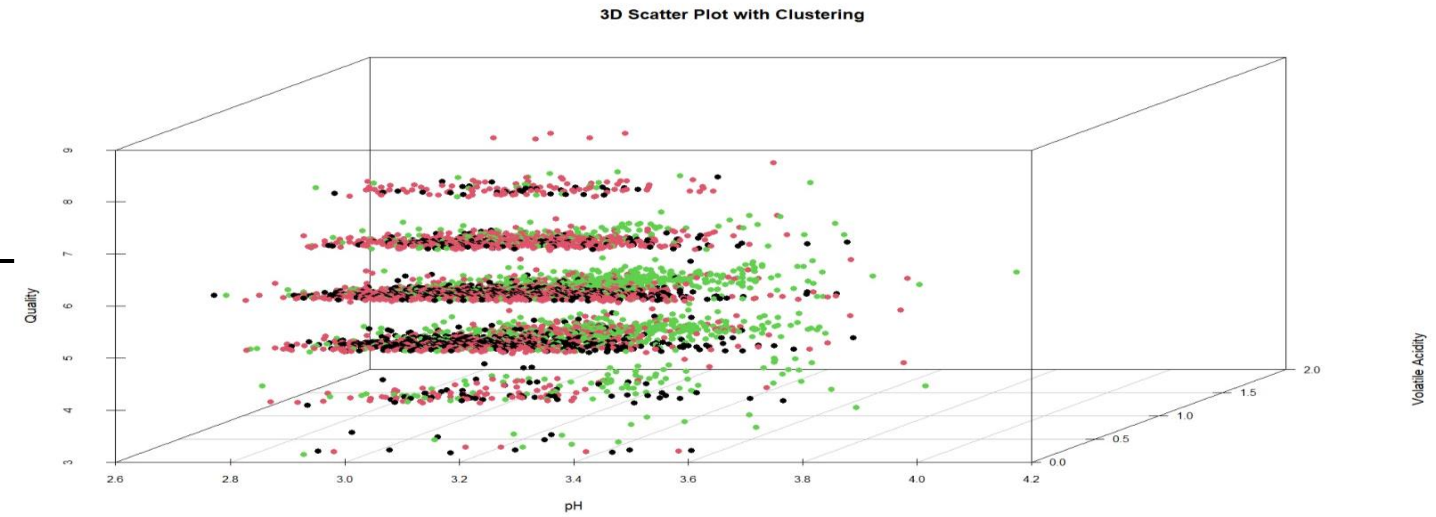
Density has very great effect on control of the quality of White wine its analyzed that all densities of wines fall in range between 0.99 &1.00 and it has been shown that if the density tends to 0.99 quality increases (fall in range 5 : 8)

Not as important as in white wine,it shows that the lines are scattered in range (0.992 : 1) interpert the weakness of the effect on the quality of Red wine

After applying clustering by all attributes to classify each type :

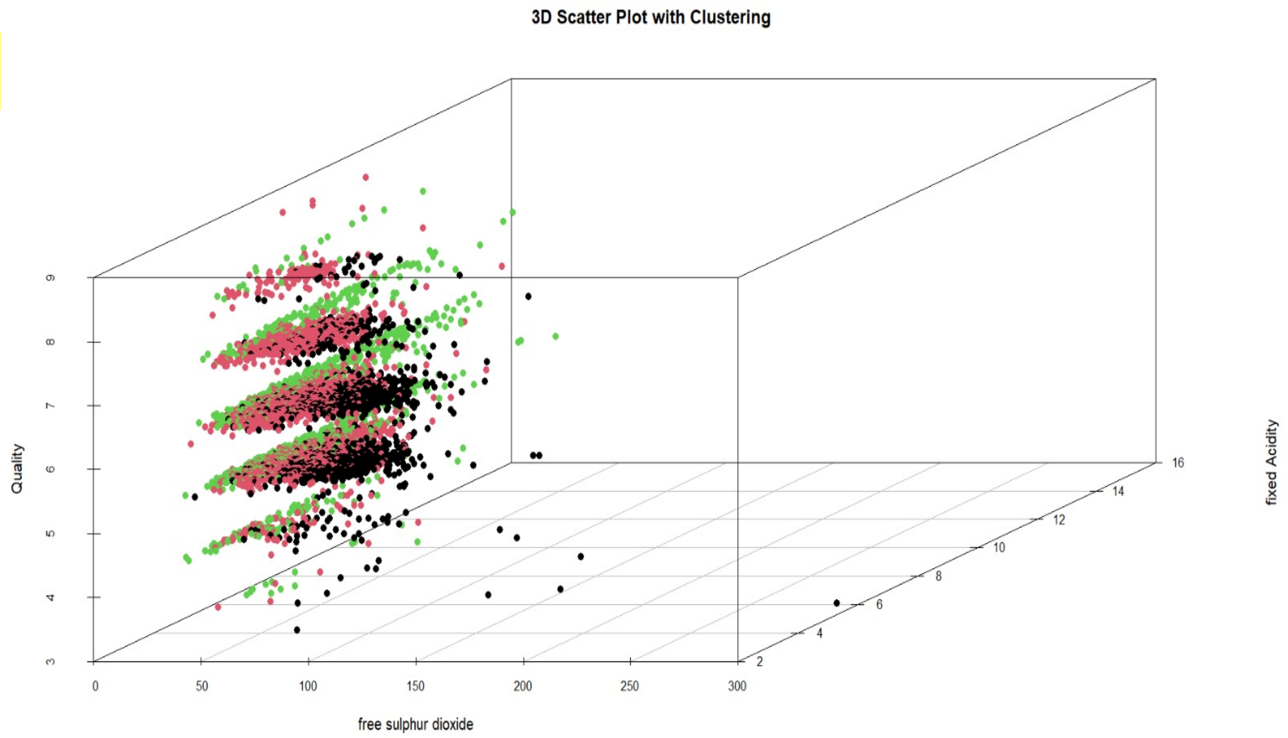
PH & volatile acidity :

it has been analyzed by the 3d modeling that volatile acidity and PH has very less effect in controlling the quality



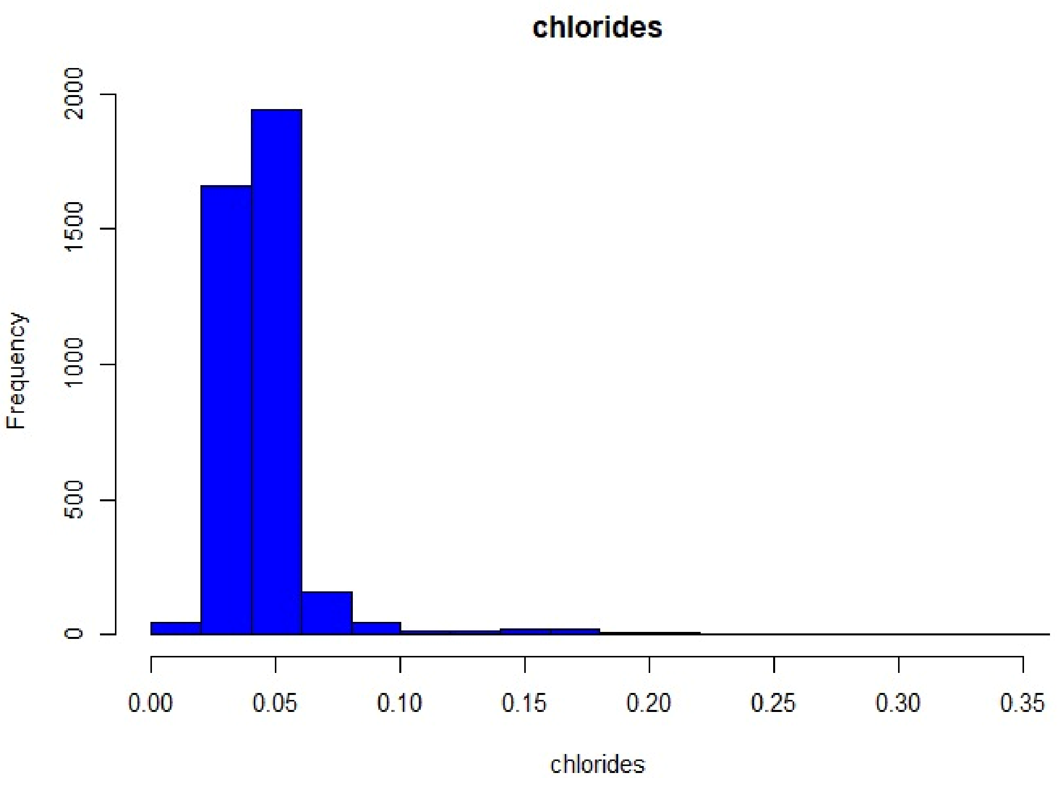

```
Console   Terminal   Background Jobs
R 4.3.2 · ~/
> attributes_for_clustering <- ALLWINES[, !(names(ALLWINES) %in% c("quality"))]
> kmeans_result <- kmeans(attributes_for_clustering, centers = 3)
> ALLWINES$cluster <- kmeans_result$cluster
> library(scatterplot3d)
> scatterplot3d(attributes_for_clustering$pH, attributes_for_clustering$volatile.acidity, ALLWINES$quality,
+               pch = 16, main = "3D Scatter Plot with Clustering",
+               xlab = "pH", ylab = "Volatile Acidity", zlab = "Quality",
+               color = ALLWINES$cluster)
> 
```

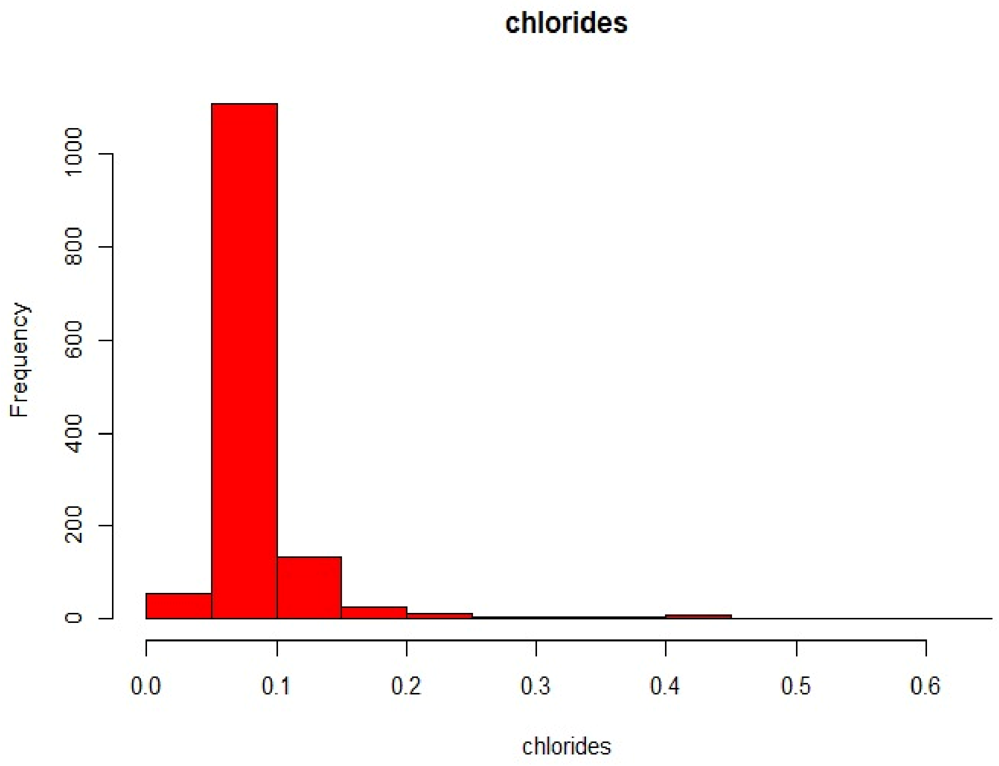
FIXED ACCIDITY & FREE SULPHUR DIOXIDE : it has been analyzed that their quantities should be in a specific ranges and not has a great influence in control of Quality of wines in general




```
Console  Terminal  Background Jobs
R 4.3.2 · ~/
> attributes_for_clustering <- ALLWINES[, !(names(ALLWINES) %in% c("quality"))]
> kmeans_result <- kmeans(attributes_for_clustering, centers = 3)
> ALLWINES$cluster <- kmeans_result$cluster
> library(scatterplot3d)
> scatterplot3d(attributes_for_clustering$pH, attributes_for_clustering$volatile.acidity, ALLWINES$quality,
+               pch = 16, main = "3D Scatter Plot with Clustering",
+               xlab = "pH", ylab = "Volatile Acidity", zlab = "Quality",
+               color = ALLWINES$cluster)
> scatterplot3d(attributes_for_clustering$free.sulfur.dioxide, attributes_for_clustering$fixed.acidity, ALL
WINES$quality,
+               pch = 16, main = "3D Scatter Plot with Clustering",
+               xlab = "free sulphur dioxide", ylab = "fixed Acidity", zlab = "Quality",
+               color = ALLWINES$cluster)
>
```

It truly reflects the real characteristics of wines :
It shows that chlorides are located in range [0.05,0.1] in Red wine
Same analyze shows that is crucial for the quantity of chlorides in White wine to locate around 0.05





Conclusion :

- Our goal was to measure the importance of each attribute on the quality of red and white wine, so, the stakeholders can improve wine quality and increase their sales.
We used a supervised technique, Decision Tree, to identify the most influential attributes in the quality of each the red and white wine.

-The concept of the “Variable Importance” was used to identify the percentage by which each attribute contributes in influencing the wine quality.

-The decision tree provided us with the important attributes upon which we can apply the clustering analysis.

-The wine data was clustered upon the alcohol and density attributes into 3 clusters. Then, the distribution of the wine quality is plotted by each cluster using the boxplot. We figured out that the three clusters almost have the same quality distribution, which is something refers to a limitation in the dataset, the small size. It leads us to unconfident results after using the machine learning techniques.

-We used the line plot to visualize the correlation between the density and the quality of each the red and the white wine. It indicated that the density is more correlated with the quality in the white wine than the red wine

-At the end, density and alcohol are the attributes that worth the focus of the stakeholders to improve the wine quality.