Data Analysis - Wine Quality (UCI)

DIETSCH Tanguy
ROSINA Maxime
IZARD Léonore
BERTHET Alexis

ISEP - Institut Supérieur d'Électronique de Paris
II. 2313 - Data Analysis: Project
2020/2021

Summary

Introduction

For our data analysis final project, we chose to work on the Wine Quality dataset. This dataset contains oenological objectives and sensitive data. The dataset was created in 2009 by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis in order to “Modeling wine preferences by data mining from physicochemical properties.”.

The aim of this project will be to treat and analyse this dataset in order to extract and identify the main objective criteria that makes a wine taste good.

You can find all the resources of the project on this github repository.

I - Presentation

I.A - Dataset comprehension

The dataset comprehension is the first part of our analysis. We had to understand the diverse files composing the core of the dataset in addition to the different attributes used. Later we will start some light analysis in order to better understand our dataset.

I.A.1 - Description

The Wine Quality dataset is composed of two different files: winequality-red.csv and winequality-white.csv. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. We know that the classes are ordered and not balanced. We also know that there is no missing value.

There are eleven input attributes and one output attribute in both dataset. We do not know if the eleven input attributes are all relevant. The red wine dataset contains 1599 instances and the white wine 4898.

Let's take a closer look at the different attributes.

I.A.2 - Attributes

The eleven input attributes are:

- ❏ fixed acidity
- ❏ volatile acidity
- ❏ citric acid
- ❏ residual sugar
- ❏ chlorides
- ❏ free sulfur dioxide
- ❏ total sulfur dioxide
- ❏ density
- ❏ pH
- ❏ sulphates
- ❏ alcohol

The output attribute is the quality of the wine. This score is given between 0 (poor quality) and 10 (very high quality). However, we do not have any information on the brand or the price.
Let's define more precisely all the different attributes.

First, acidity is a measurement of the quantity of acid present in a wine or must. Acidity is described as three different entities, titratable (fixed), volatile, and total. They contribute greatly to the taste of wine

Fixed acidity (g.100 mL-1)
Fixed acidity is a measurement of the total concentration of titratable acids and free hydrogen ions present in your wine. The predominant fixed acids found in wines are tartaric, malic, citric, and succinic. Their respective levels found in wine can vary greatly but in general one would expect to see 1,000 to 4,000 mg/L tartaric acid, 0 to 8,000 mg/L malic acid, 0 to 500 mg/L citric acid, and 500 to 2,000 mg/L succinic acid.

Volatile acidity (g.100 mL-1)
Volatile acids are different from fixed acids in that they cannot be measured through a titration but must be quantified using a steam distillation process. In this a sample of wine is exposed to steam which in turn encourages the volatile acids to leave the wine.
Volatile acids are produced through microbial action such as yeast fermentation, malolactic fermentation, and other fermentations carried out by spoilage organisms.

Citric acid (g/dm3)
Citric acid is often added to wines to increase acidity, complement a specific flavor. It can be added to finished wines to increase acidity and give a “fresh” flavor. The disadvantage of adding citric acid is its microbial instability.

Residual sugar (g.L-1)
The residual sugar is the quantity of sugar left in the wine after the fermentation process. To summarize, a dry wine contains from 0 to 4 grams of sugar per liter, a semi-dry wine from 4 to 12 grams per liter, a semi-sweet wine from 8 to 45 grams per liter and a sweet wine contains more than 45 grams per liter.

Chlorides (g.L-1)
Wine contains from 2 to 4 g L–1 of salts of mineral acids, along with some organic acids, and they may have a key role on a potential salty taste of a wine, with chlorides being a major contributor to saltiness.

Free sulfur dioxide (ppm)
Free sulfur dioxide is a measure of the amount of SO2 that is not bound to other molecules, and is used to calculate molecular SO2.

Sulfur Dioxide is used throughout all stages of the winemaking process to prevent oxidation and microbial growth. Excessive amounts of SO2 can inhibit fermentation and cause undesirable sensory effects.

Total sulfur dioxide (ppm)

Total sulfur dioxide (SO2) is a measure of both the free and bound forms of SO2. Bound SO2 refers to SO2 molecules that are bonded to other compounds, primarily aldehydes, pyruvate, and anthocyanins. Sulfur Dioxide is used throughout all stages of the winemaking process to prevent oxidation and microbial growth. Excessive amounts of SO2 can inhibit fermentation and cause undesirable sensory effects.

Density (kg/m3)

Density is the mass per unit volume of wine or must at 20°C. It is expressed in. grams per milliliter, and denoted by the symbol ρ 20°C.

PH

PH is the measure of the degree of relative acidity versus the relative alkalinity of any liquid, on a scale of 0 to 14, with 7 being neutral. Winemakers use pH as a way to measure ripeness in relation to acidity. Low pH wines will taste tart and crisp, while higher pH wines are more susceptible to bacterial growth. Most wine pH's fall around 3 or 4; about 3.0 to 3.4 is desirable for white wines, while about 3.3 to 3.6 is best for reds.

Sulphates (ppm)

A “sulphate” is a molecule containing sulfur as it is a key component in its molecular structure. Most sulfites found in wine are sulfur dioxide molecules and sulfite ions.

Alcohol (%)

It is the quantity of alcohol contained in wine (generally between 4,5% and 16%).

Our objective in our analysis is to understand which criteria impact the sensitive quality of the wine, in order to answer our main question:

How can we make excellent red and white wine ?

I.B - Dataset treatment (missing value)

We do not have any missing value on our both datasets. We do not need to do a first dataset treatment.

I.C - Primary Analysis

I.C.1 - Analysis of quality

We ran a first analysis on the quality output attribute, starting with all wines.

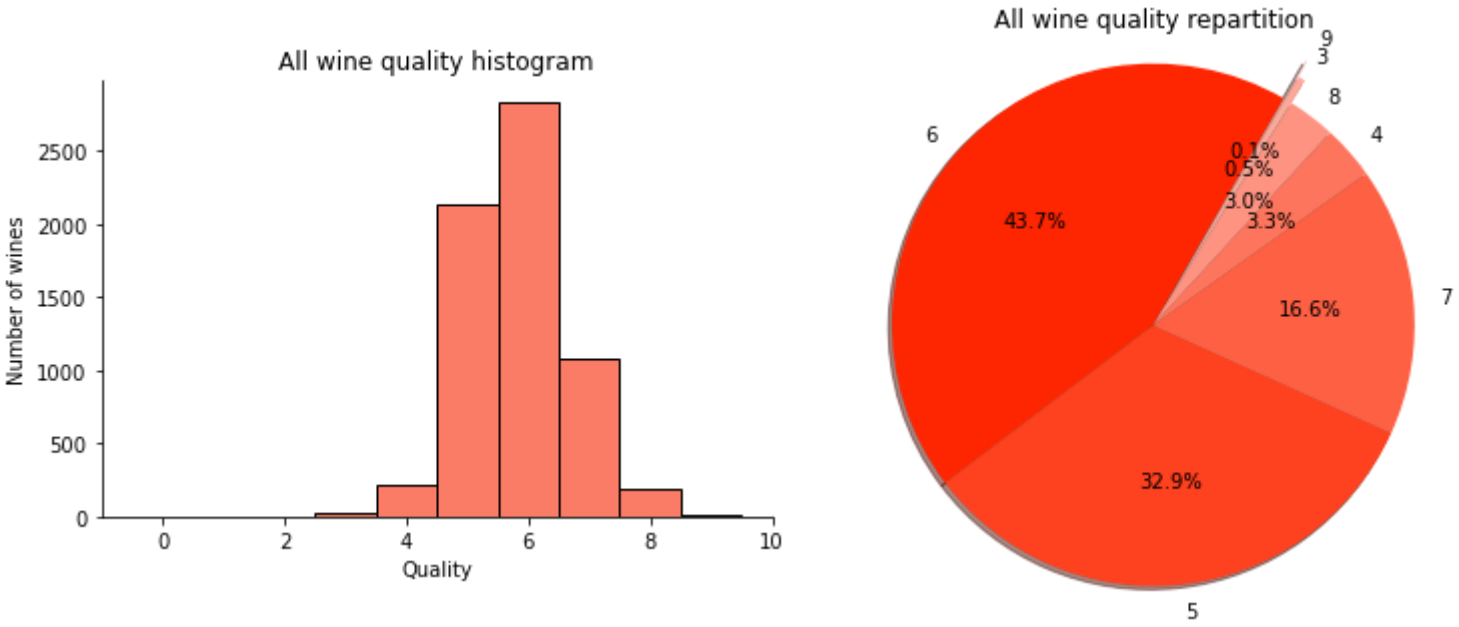

Figure I.C.1.0: All wines quality histogram and pie chart

These graphs show that the repartition of the quality between all wines follow a normal distribution. We see that there is no wine that is very bad (between 0 and 2) and no one that is very excellent (10). The majority of the wine's quality is between 4 and 7 (96.5%). These data do not surprise us. When we go to a supermarket and we take a random wine we often have a medium quality wine.

Now, we will take a closer look at the two datasets that are interesting to us: the red and white wines.

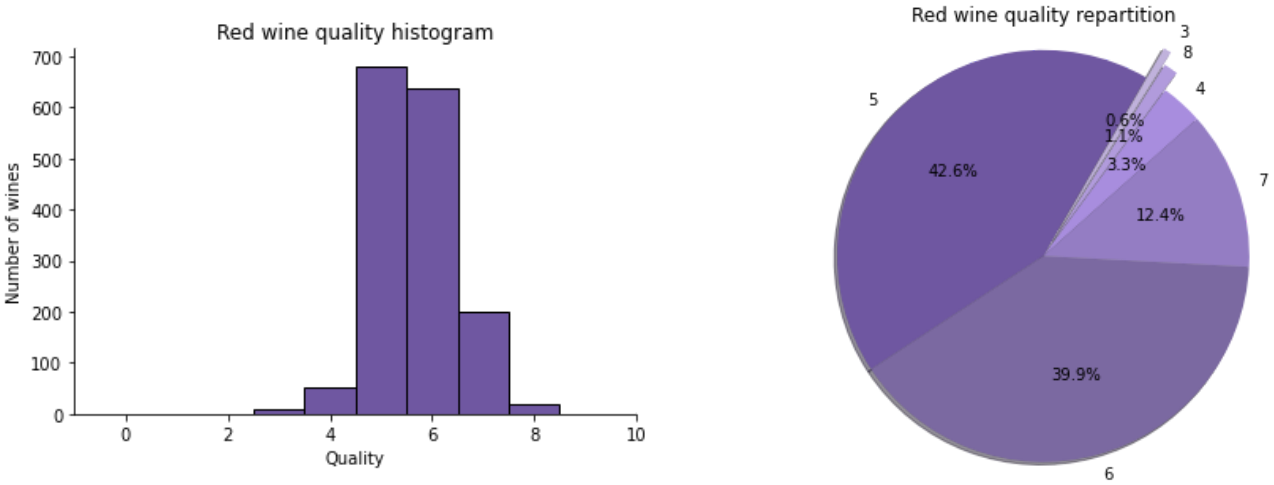

Figure I.C.1.1: Red wine quality histogram and pie chart

In these graphics we can see that the red wines are also following a normal distribution. This time, the majority of red wine is between 5 and 7 quality (94.9%). We can see that there is no red wine that has a quality superior than 8.
In contrast to the distribution of all wines, red wines have a distribution more centred around the average.
Again the results do not surprise us. It is often said that it is easier to have a better taste in mouth with white wine than red wine because red wine can be chilled. That could be one hypothesis explaining that there are no very high quality red wines in contrast to white wines.

Now, we will observe the white wine distribution.

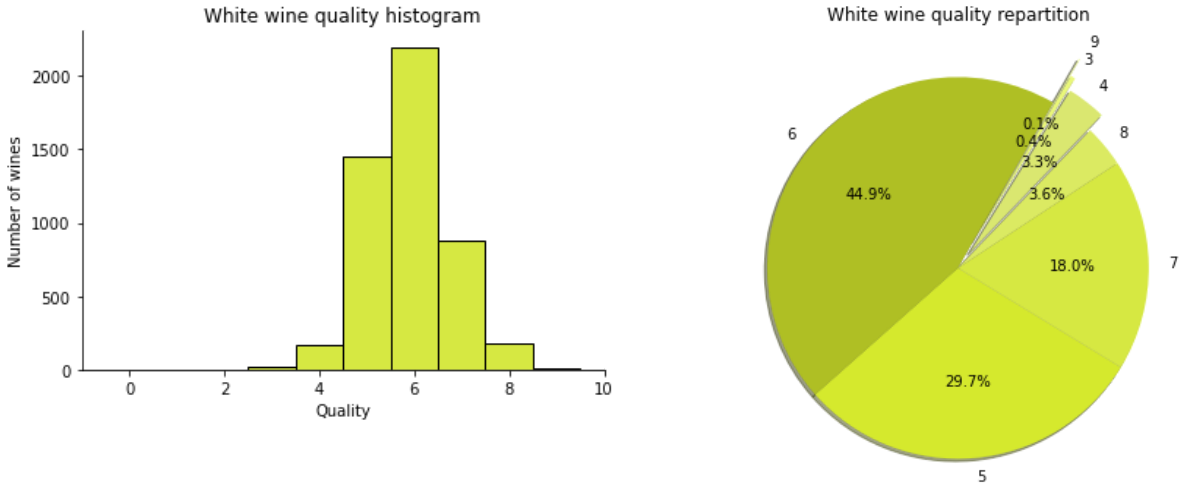

Figure I.C.1.2: White wine quality histogram and pie chart

The white wine distribution also follows a normal distribution, a little less average centered than the red wines's one. This time 96.2% of the values are between 5 and 8. We can see a small difference between the red wines and the white wines quality. It seems that white wines are of better quality in general.

After we have quickly seen the repartition of the wine by their quality, we will focus on another important point in the treatment of data: the outliers and the different regulations that are in vigor for wines.

I.C.2 - Outliers and regulations

Outliers

We ran an Z-scores analysis to detect any outlier in our datasets.
Z-scores allow us to quantify the unusualness of an observation (if the data follows a normal distribution). Z-scores are the number of standard deviations above and below the mean that each value falls.
All Z-score below 3 were kept.

Red Wines

Before

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000

After

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
count	1451.000000	1451.000000	1451.000000	1451.000000	1451.000000	1451.000000

White Wines

Before

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000

After

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
count	4487.000000	4487.000000	4487.000000	4487.000000	4487.000000	4487.000000

We removed 148 outliers for the red wine dataset and 411 for the white wine dataset.

Moreover, we need to check if there are any regulations or any recommendations for some attributes.

Regulations & recommendations

After some research, we found out that these regulations and recommendations are in place.

Volatile acidity limitations

Volatile acidity limitations are a maximum of 0.98 g·l-1 for red wines and a maximum of 0.88 g·l-1 for white wines.

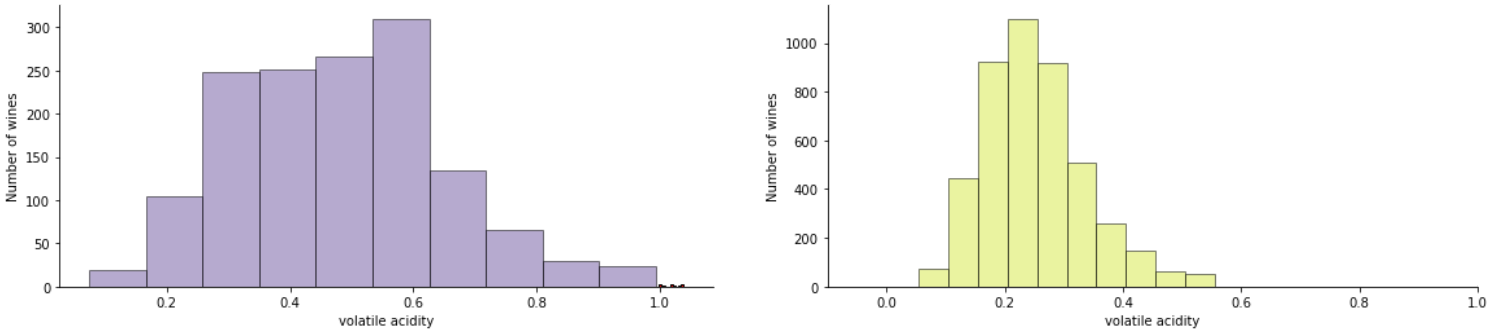

Figure I.C.2.1: Volatile acidity limitations

We can see on the two figures that the very important majority of the wines are following the volatile acidity limitations.

Alcohol recommendations

Alcohol is not limited but the recommendations for wines are to have alcohol between 4.5% and 16%.

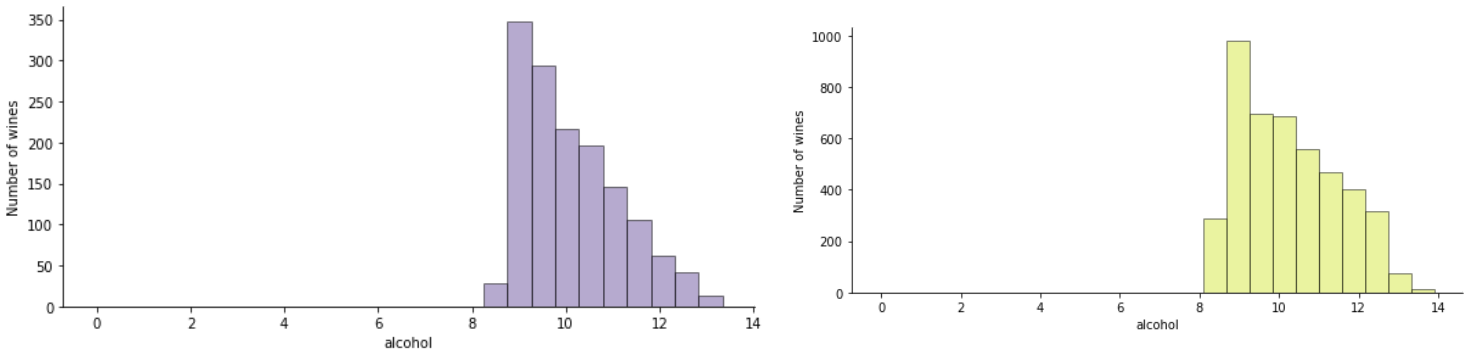

Figure I.C.2.2: Alcohol recommendations

We can also see that we have no wine out of regulation regarding alcohol limitations.

Wine total sulfur dioxide limitations

The EU has set a legal limit for total SO2 of 150 mg/litre in red wines and 200 mg/litre in white wines.

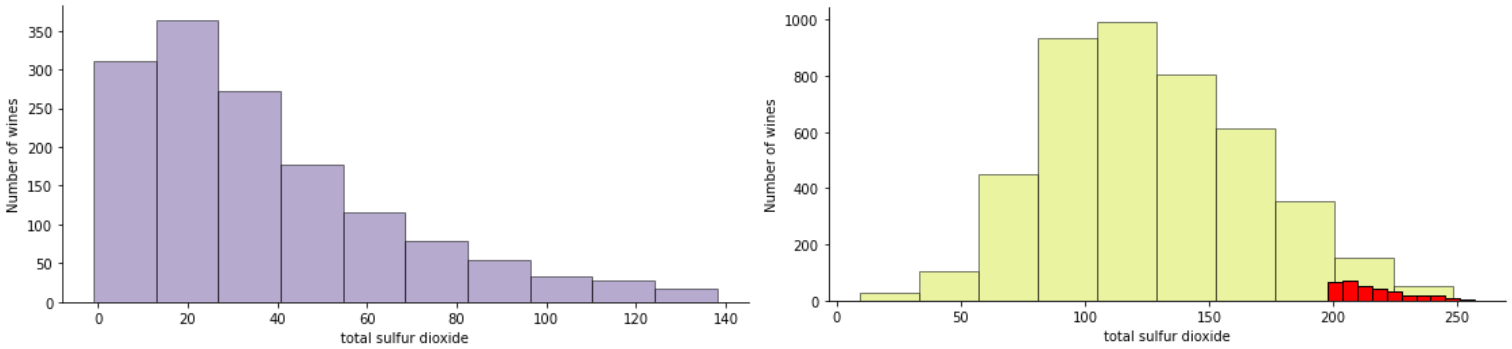

Figure I.C.2.3: Total sulfur dioxide limitations

This time we can clearly see that some of the white wines are not following the EU rules about total sulfur dioxide.

Maybe these wines were not from the EU. However, we do not have many wines that are not following the EU limitations, so we choose to keep them.

Conclusion

After we had resolved the outliers problem, we could try our data to the wine regulations that we had found. Only some wines did not follow the regulations, and we chose to keep them because we think they might follow their country regulations outside of the EU.

II - Observation

Now that we have realised our first observations and that we have defined the different attributes we will see in this part the difference observations that we can make with our datasets.

II.A - Analysis of quality depending on attributes

In this part we want to understand how the attributes are evolving with red, white, and all wines. In order to understand better we plot 3 curves for each attribute. Each curve is the average of the attribute for each quality step.

Fixed acidity

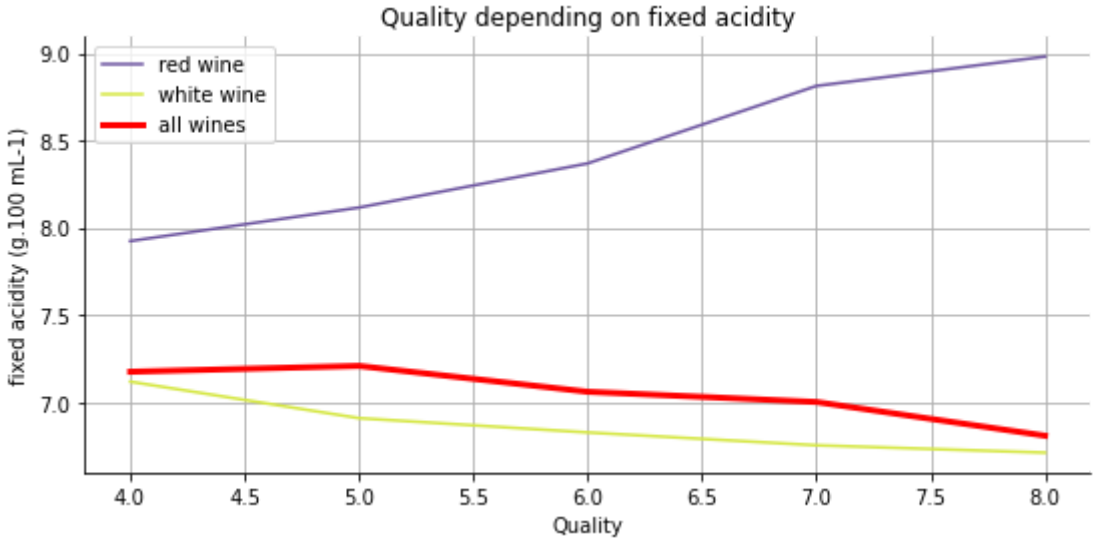

Figure II.A.1: Wine quality depending on fixed acidity

In this first graph we can see that red and white wine are reacting very differently for the fixed acidity.
In fact, for the red wine, we can see that the acidity is increasing with the quality of wines. In contrast, for the white wine, the fixed acidity drops below for the high quality wines.

Now let's focus on the volatile acidity.

Volatile acidity

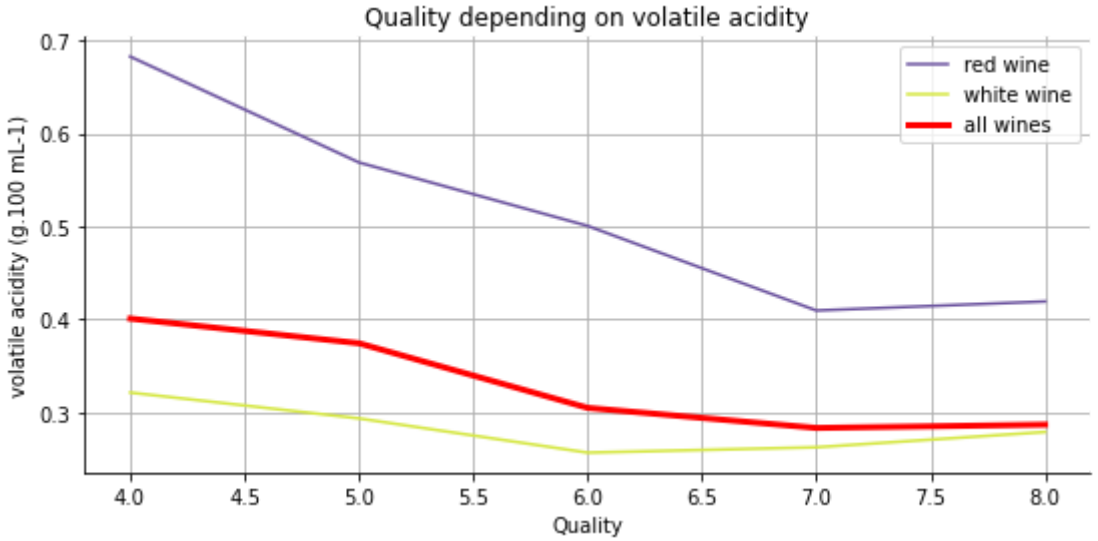

Figure II.A.2: Wine quality depending on volatile acidity

Here the curves are following the same model. Bad red wines begin with a volatile acidity around 0.9 g.100mL-1. Then, the volatile acidity drops for the red wine with a higher quality and sticks around 0.4 g.100mL-1 for the high quality wines.
White wines are a little more constant, the volatile acidity is always around 0.3 g.100mL-1.

Citric acid

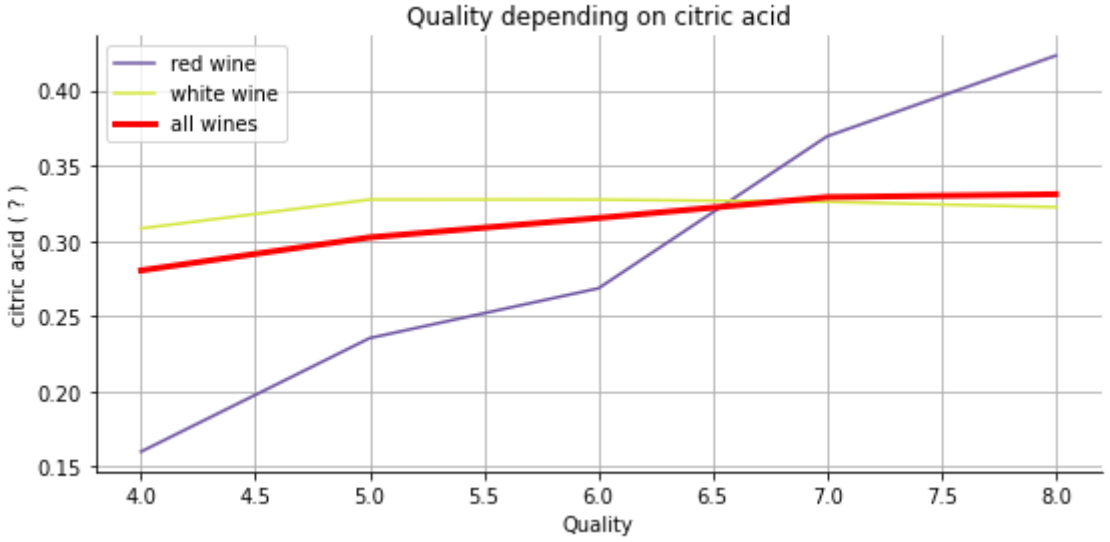

Figure II.A.3: Wine quality depending on citric acid

This graphic is very interesting because we can see that, for the red wines, the quality is increasing as the citric acid is, reaching a value above 0.4. However, for the white wines, the value of citric acid is stagning around 0.33.

Residual sugar

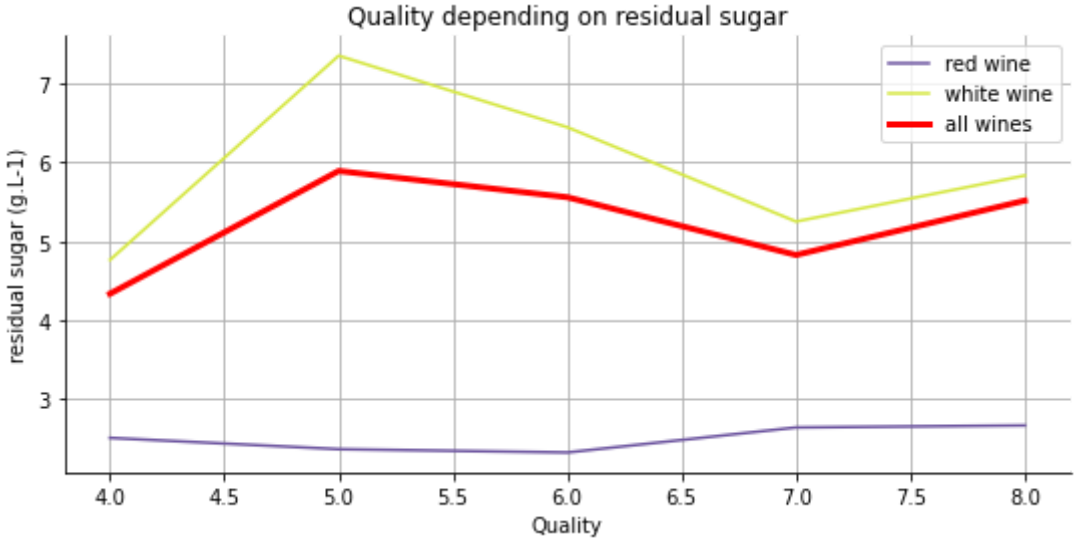

Figure II.A.4: Wine quality depending on residual sugar

This time the value of residual sugar for the red wine is stagning around 2.5 g.L-1. The value of residual sugar for white wines rides a roller coaster. It begins at 5 for a 4 quality wine, reaching 7.3 for a 5 quality wine, then dropping around 5 for a good quality (7). Finally, the value of residual sugar for high quality (8) white wine is around 6.

Chlorides

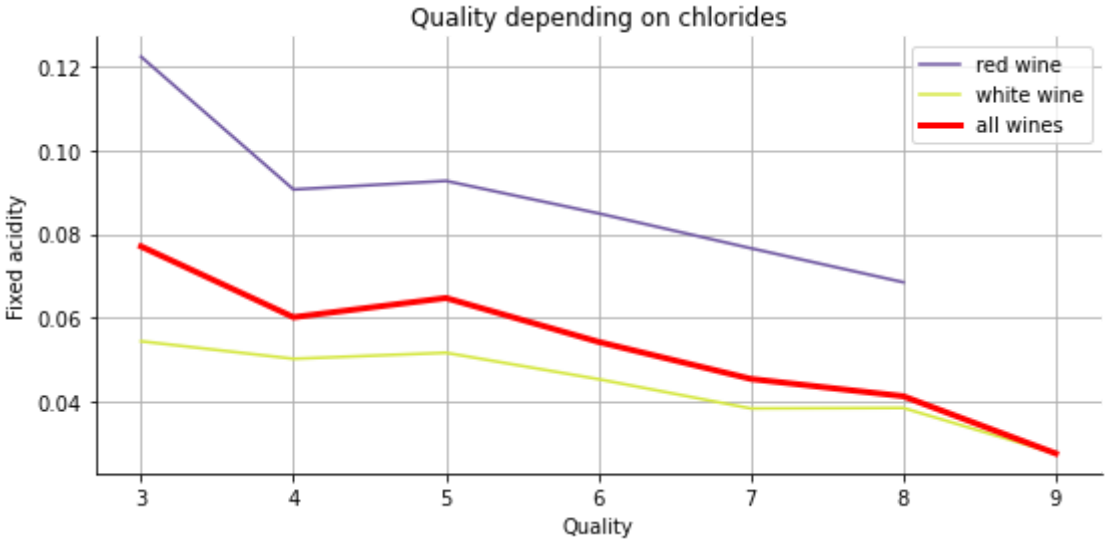

Figure II.A.5: Wine quality depending on chlorides

Red and white wines chlorides are both following the same model: more the fixed acidity decreases more the quality increases.

Free sulfur dioxide

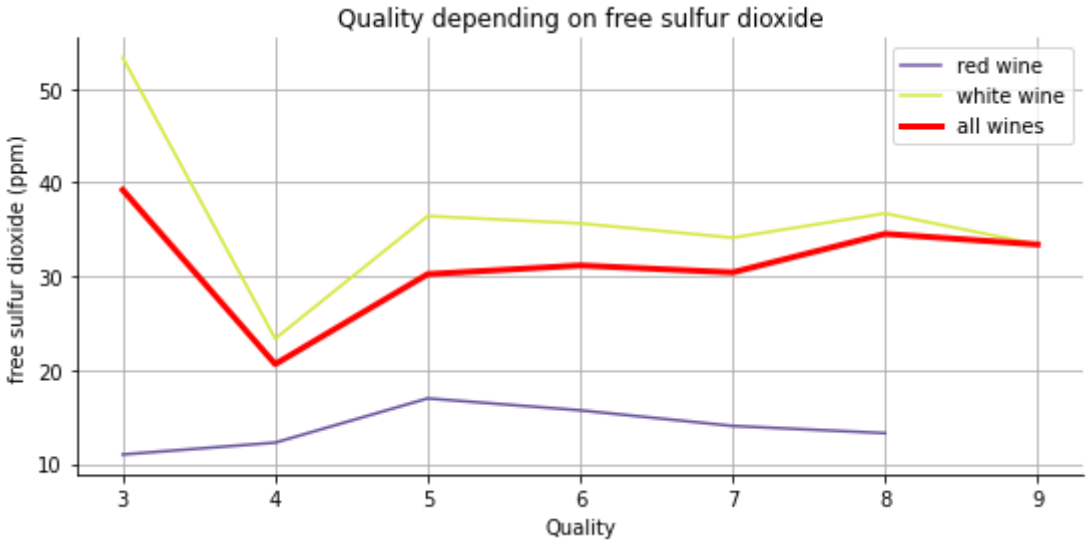

Figure II.A.6: Wine quality depending on free sulfur dioxide

Free sulfur dioxide is very stable regarding all quality of red wines. In contrast, it is very unstable for white wine. Good white wines have a free sulfur dioxide around 35ppm.

Total sulfur dioxide

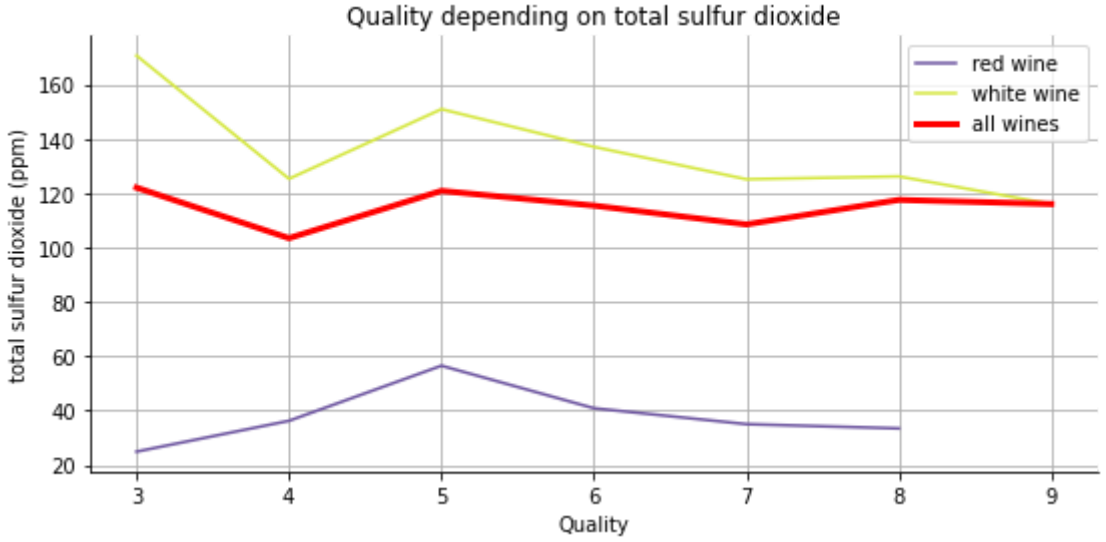

Figure II.A.7: Wine quality depending on total sulfur dioxide

Total sulfur dioxide seems to be stagnant for red wine around 37ppm, despite a little spike for 5 quality wines.

For white wines it is a little less simple to say, but it is slowly decreasing, finally reaching 120ppm for 8/9 quality wine.

Density

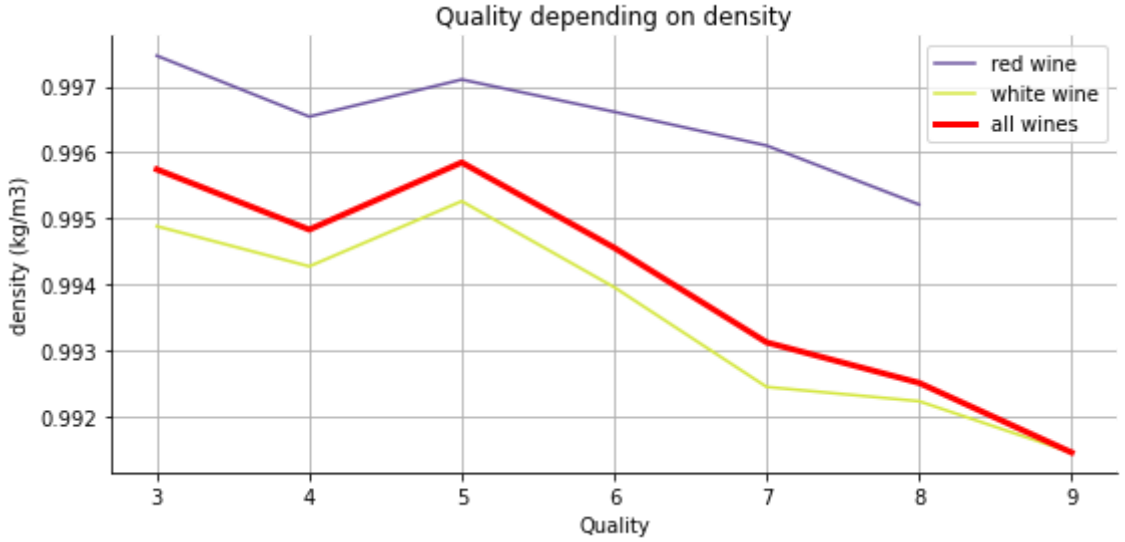

Figure II.A.8: Wine quality depending on density

For both red and white wines, quality is decreasing as the quality is increasing.

pH

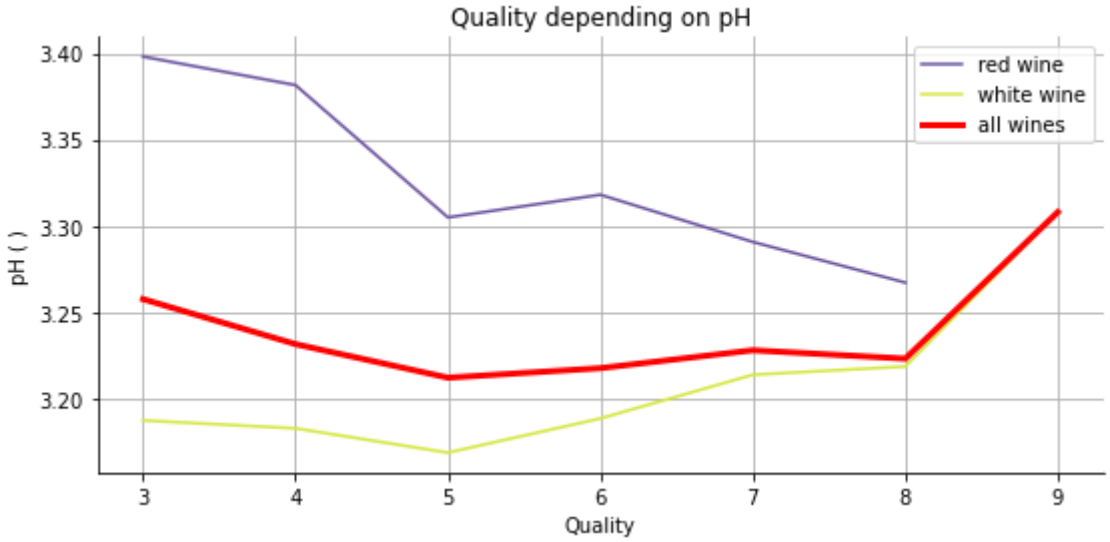

Figure II.A.9: Wine quality depending on pH

Red and white wines seem to have an opposed relationship with the pH. For white wines, a higher pH seems to mean a better quality wine, and for the red wine is the opposite.

Sulphates

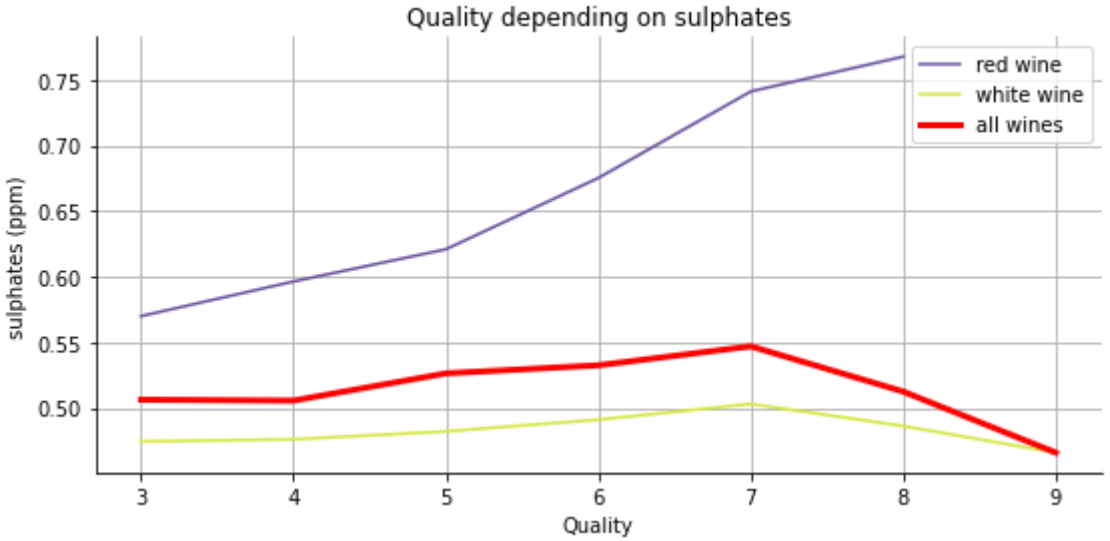

Figure II.A.10: Wine quality depending on pH

Again, red and white wines seem to have a completely different comportement regarding sulphates and quality. Red wine quality increases with sulphates, in contrast of white wine quality.

Alcohol

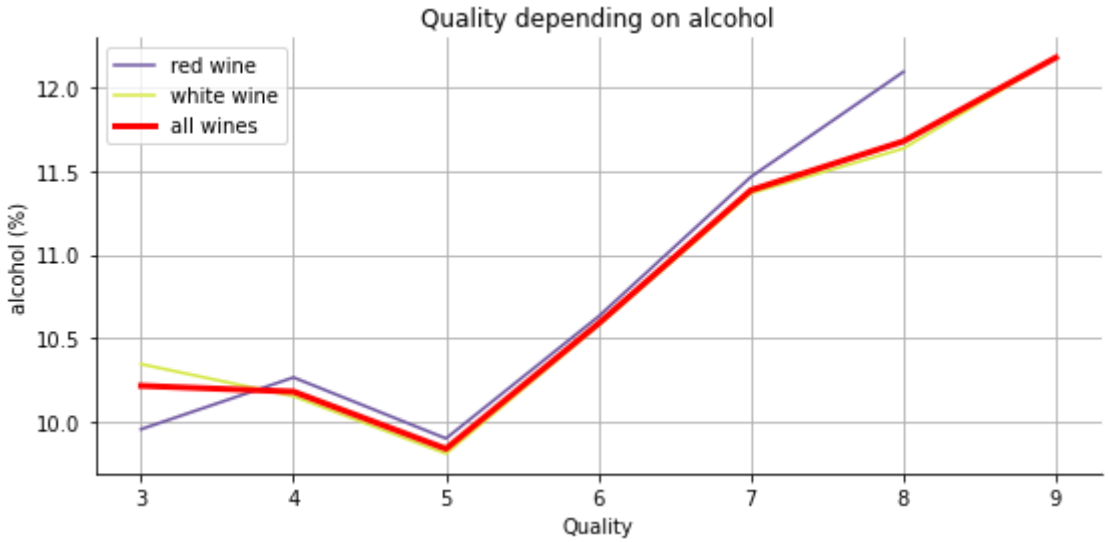

Figure II.A.11: Wine quality depending on alcohol

This time, both red and white wines have the same comportement, as the alcohol increases, the quality rises.

Conclusion

We can do a little conclusion of everything that we saw in this part. We can make some hypothesis between correlated attributes with quality.

Red wine:

- ❏ Alcohol (positively correlated)
- ❏ Sulphates (positively correlated)
- ❏ Chlorides (negatively correlated)
- ❏ Citric acid (positively correlated)
- ❏ Volatile acidity (negatively correlated)

White wine:

- ❏ Alcohol (positively correlated)
- ❏ Density (negatively correlated)
- ❏ Chlorides (negatively correlated)

II.B - Correlation between attributes

In this part we want to understand which attributes are correlated to quality. First, we take a look at the red wine correlation matrix, then we will take a look at the white wine one's.

II.B.1 - Correlation matrix for red wines

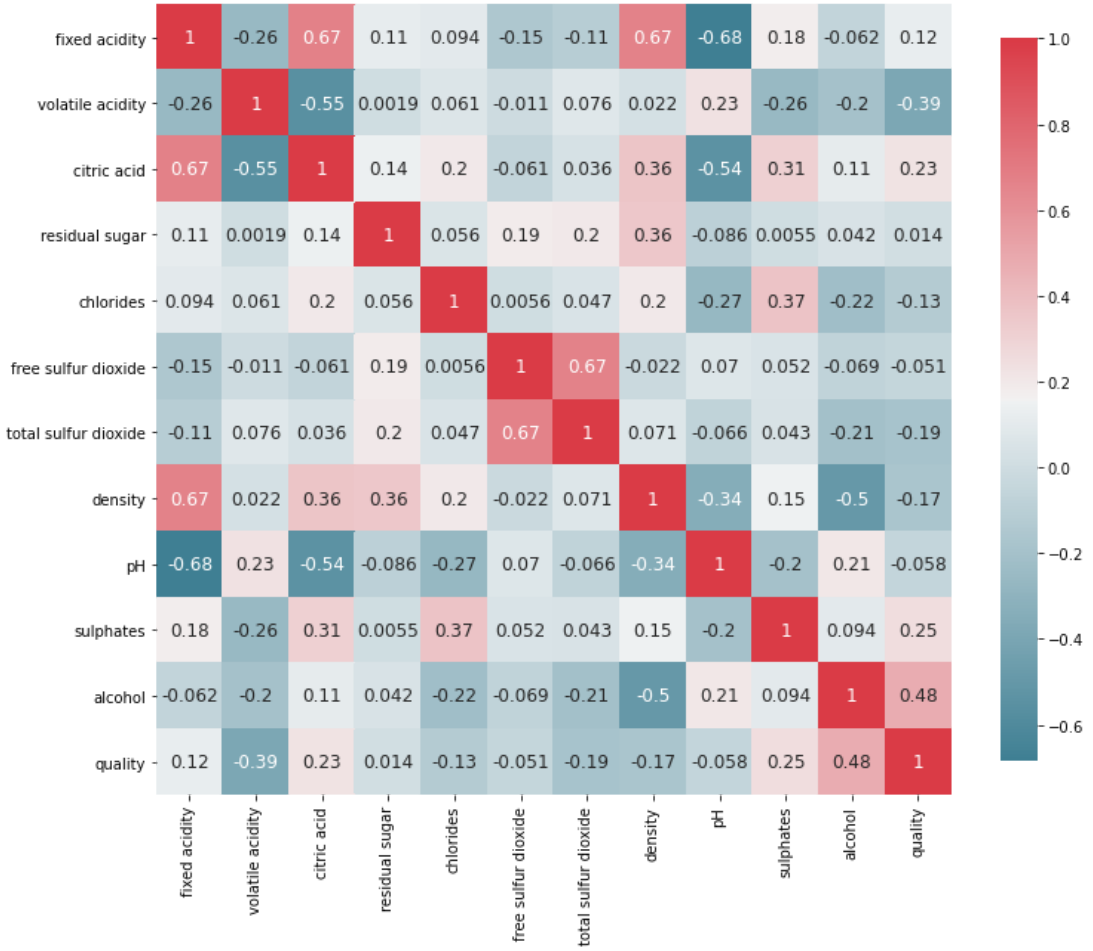

Figure II.B.1.1: Red wines correlation matrix

We first plot the red wines correlation matrix. In spite of the colours, it is a little difficult to understand fully which attribute is linked to others. Moreover, we are seeking the attributes related to the quality factors.
In order to do so, we just made the correlation matrix (below) between each attribute and the quality.

Attribute	Quality
Fixed acidity	0.145163
Volatile acidity	-0.353443
Citric acid	0.243999
Residual sugar	0.061482
Chlorides	-0.108787
Free sulfur dioxide	-0.071202
Total sulfur dioxide	-0.237745
Density	-0.167568
pH	-0.082164
Sulphates	0.386567
Alcohol	0.501501
Quality	1.000000

Figure II.B.1.2: Red wines correlation matrix for quality

We can see from the matrix above that the major quality correlated attributes for the red wines are:

- ❏ Alcohol (0.501501)
- ❏ Sulphates (0.386567)
- ❏ Citric acid (0.243999)
- ❏ Total sulfur dioxide (-0.237745)
- ❏ Density (-0.167568)

The correlation matrix allows us to say that alcohol is the most quality correlated factor for red wines. Alcohol, sulphates and citric acid are attributes positively correlated with quality, in contrast to Total sulfur dioxide and density that are negatively correlated with quality.

Now let’s take a closer look to the correlation matrix of the white wines.

II.B.2 - Correlation matrix for white wines

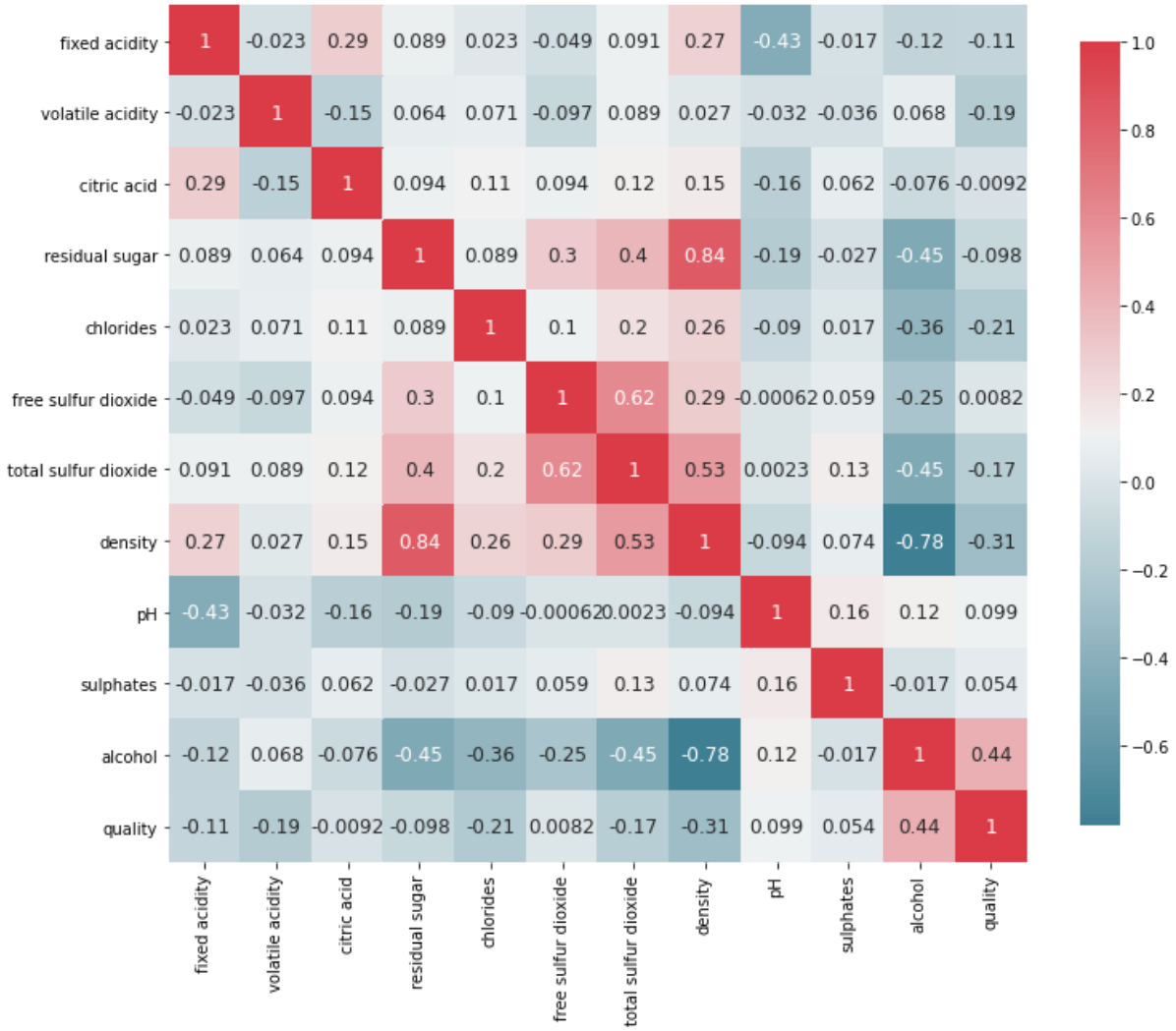

Figure II.B.2.1: White wines correlation matrix

Same as the red wines, we need to plot the correlation matrix of the different attributes with quality.

Attribute	Quality
Fixed acidity	-0.091097
Volatile acidity	-0.141278
Citric acid	0.003629
Residual sugar	-0.105612
Chlorides	-0.295800
Free sulfur dioxide	0.019293
Total sulfur dioxide	-0.171196
Density	-0.314034
pH	0.094761
Sulphates	0.037146
Alcohol	0.445076
Quality	1.000000

Figure II.B.2.2: White wines correlation matrix for quality

We can see from the matrix above that the major quality correlated attributes for the white wines are:

- ❏ Alcohol (0.445076)
- ❏ Volatile acidity (-0.141278)
- ❏ Chlorides(-0.295800)
- ❏ Total sulfur dioxide (-0.171196)
- ❏ Density (-0.314034)

The correlation matrix allows us to say that alcohol is also the most quality correlated factor for white wines. Alcohol is the only attribute positively correlated with quality, in contrast to Total sulfur dioxide, density, chlorides and volatile acidity that are negatively correlated with quality.

Conclusion

In conclusion, we choose to keep the 5 majors attributes related to quality.

- ❏ Red wines: Alcohol, Sulphates, Citric acid, Density, Total sulfur dioxide
- ❏ White wines: Alcohol, Density, Volatile acidity, Chlorides, pH

III - Visualization

III.A - PCA 2 & 3 components

First, for a simplified visualisation of data, we split the wines in two categories: a “good” wine category, containing wines with a note >= 6, and a “bad” wine categorie, for wines with a note smaller than 6.
After performing a PCA, we print the raw PCA data to visualise if there are remarkable clusters or not.
For the white wines we have:

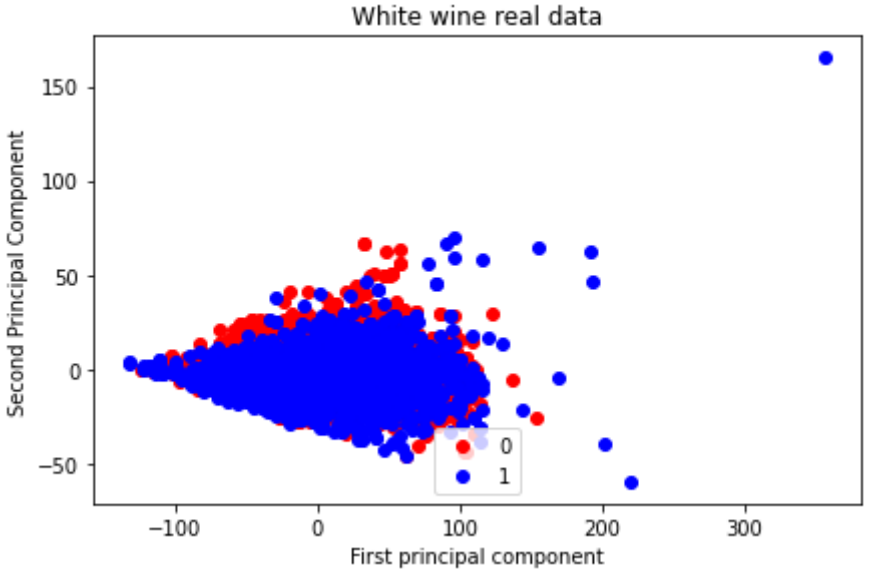

Figure III.A.1: PCA in 2 dimensions of white wines

And for red ones we have:

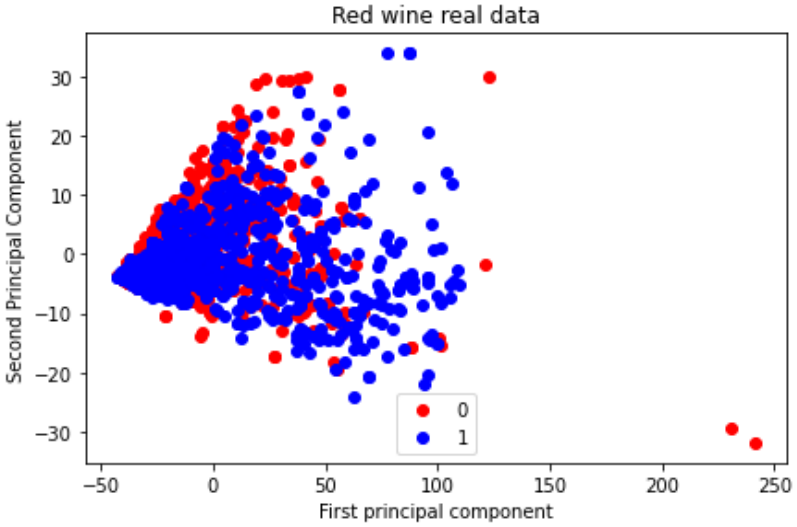

Figure III.A.2: PCA in 2 dimensions of red wines

In the two cases, we can distinguish any cluster. We can try a 3D PCA to see if good wines and bad wines split into two clusters:

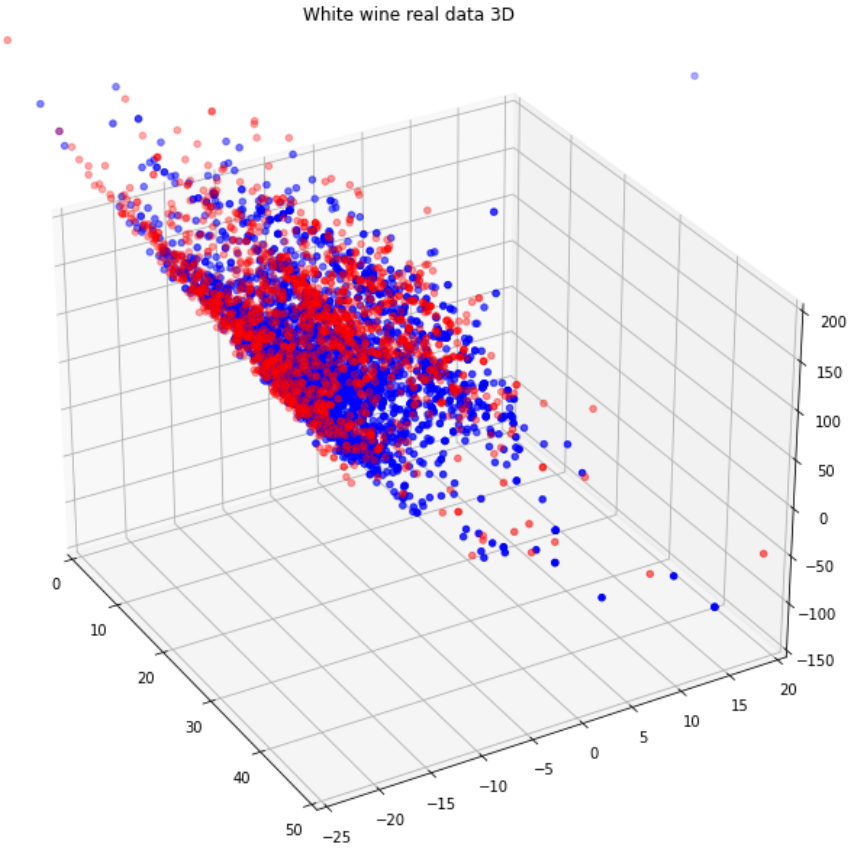

Figure III.A.3: PCA in 3 dimensions of white wines

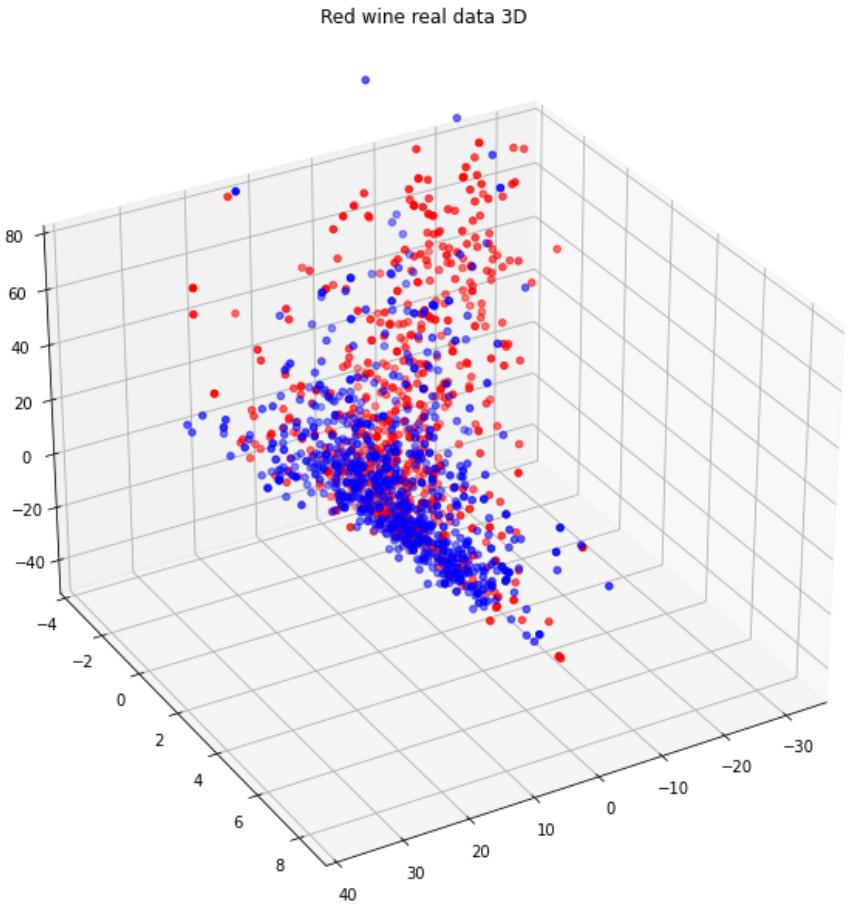

Figure III.A.4: PCA in 3 dimensions of red wines

If there is still no clear cluster, we can see in both cases that good or bad wines are more frequent in some points of the ACP.
Let's perform a K-mean clustering to study the clusterability of the dataset.
For the white wine, we obtain this two results, in 2D and in 3D:

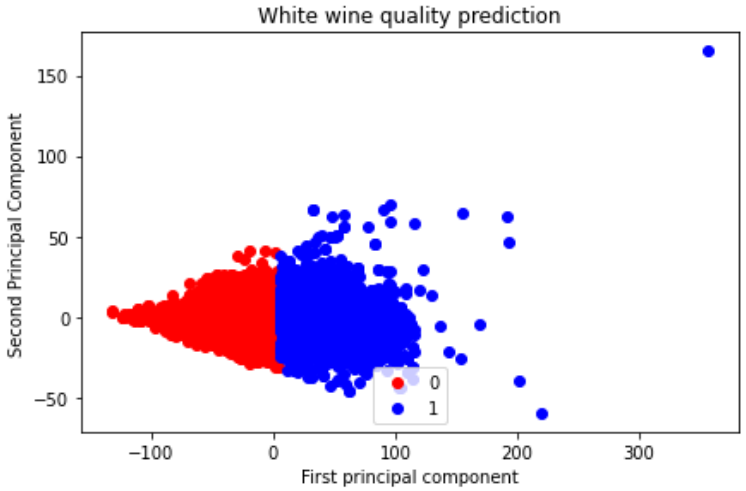

Figure III.A.4: Clustering in 2 dimensions of white wines

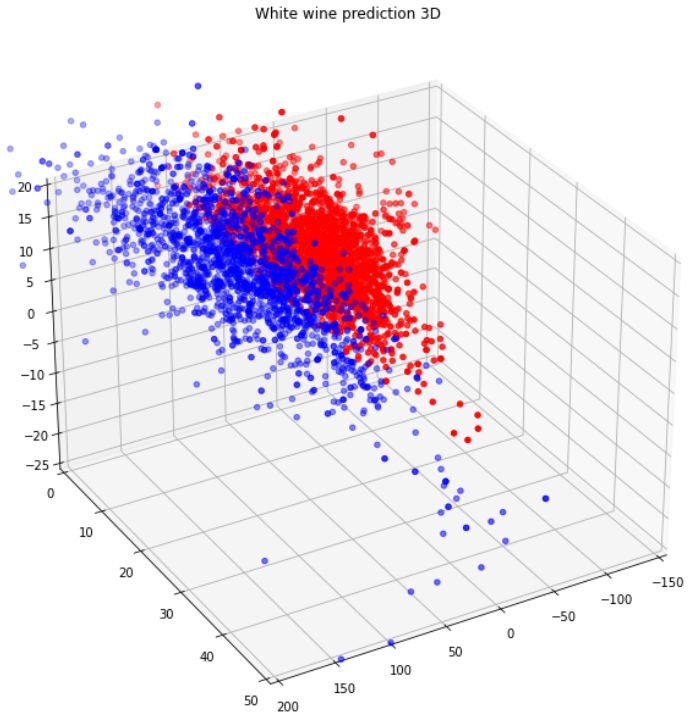

Figure III.A.5: Clustering in 3 dimensions of white wines

To have a clue about the precision of our clusters, we compute the percentages of true prediction. We got the result:

```
The clustering precision with the K-mean algorithm for white wine is of 60.14699877501021  %
```

The percentage isn't very high. By selecting randomly the category of each wine, the precision would be 50%. This quite low score is logical because, as sayed before, there are no visual clusters.

And for red wines, we got the following results:

*

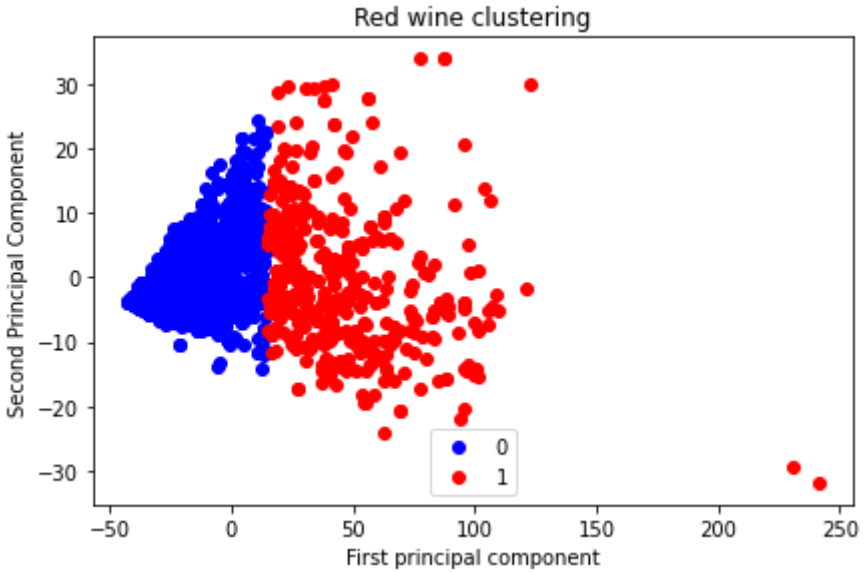

Figure III.A.6: Clustering in 2 dimensions of red wines

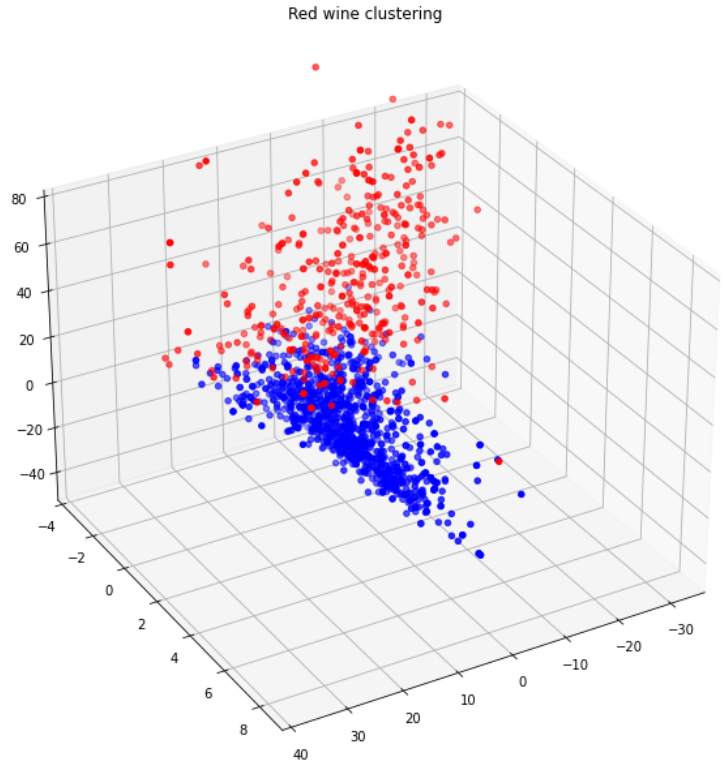

Figure III.A.6: Clustering in 3 dimensions of red wines

Here, we got this precision percentage:

```
The clustering precision with the K-mean algorithm for red wine is of 61.225766103814884  %
```

He is near the previous one and can be explained by the same way.

III.B - Boxplots depending on selected attributes

In order to better understand the repartition of our data depending on a selected attribute and the quality we plot several boxplots.

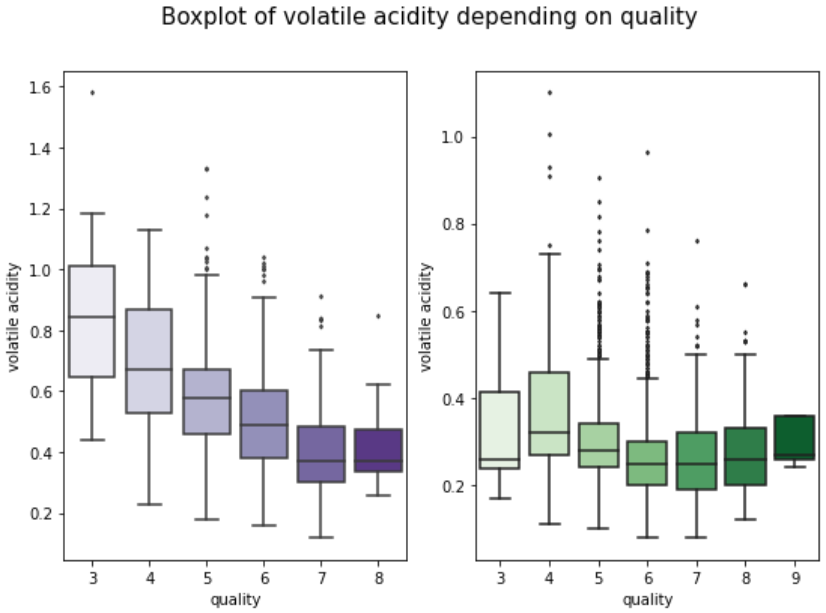

Figure III.B.1: Boxplot of volatile acidity depending on quality

We are interested in the right graph that is associated to white wine. We can clearly see some outliers. However, what is more interesting is that white wines with high quality have a volatile acidity around 0.3 g.100 mL-1.

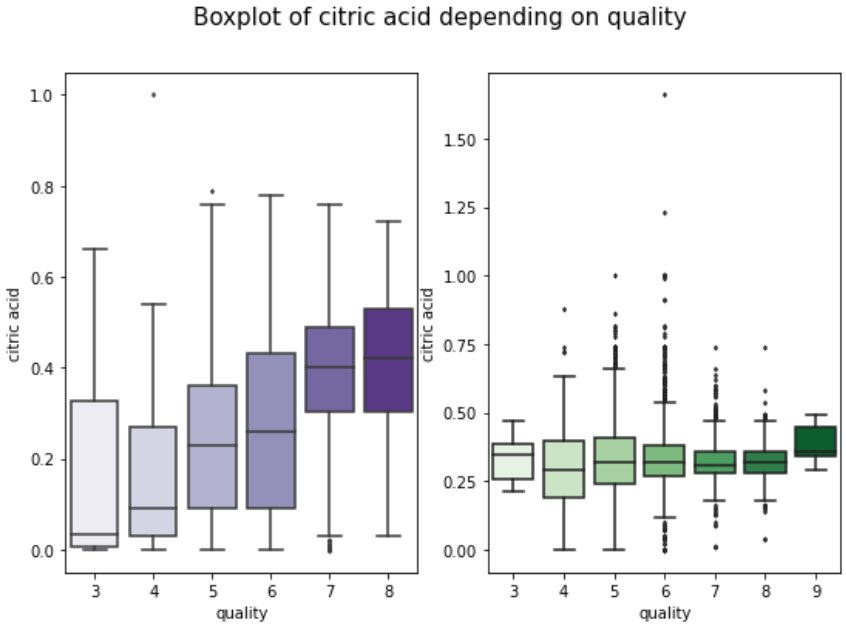

Figure III.B.2: Boxplot of citric acid depending on quality

For Citric acid we will focus on the left graph. We can clearly see that there are few outliers and that the value of citric acid for high quality wine is high, around 0.4 g/dm3.

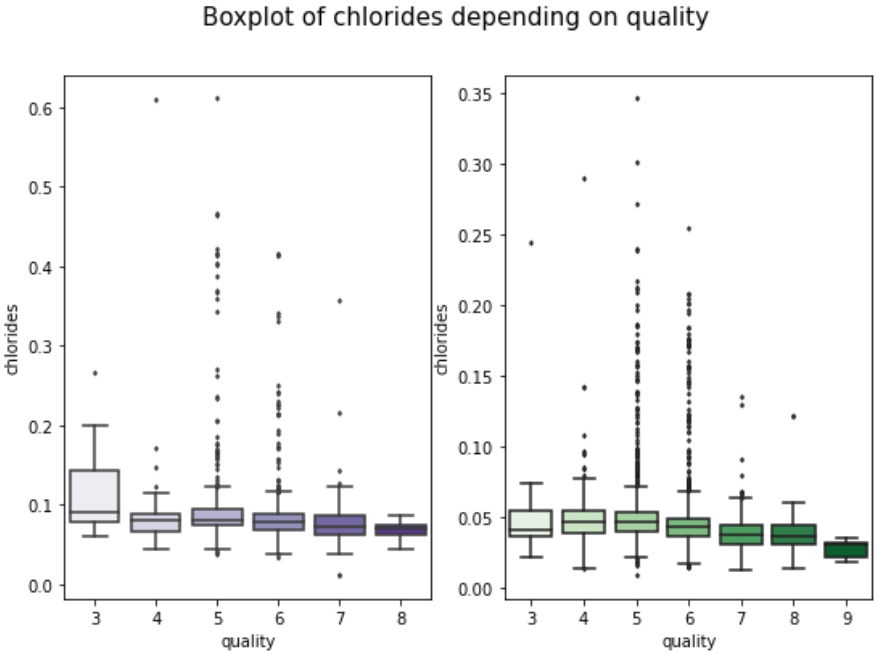

Figure III.B.3: Boxplot of chlorides depending on quality

Chlorides are more important for red wine, so we will focus on the right graph. We can clearly see a lot of outliers for middle quality white wines. However, higher the quality of the wine, lower the chlorides, around 0.04 g.L-1.

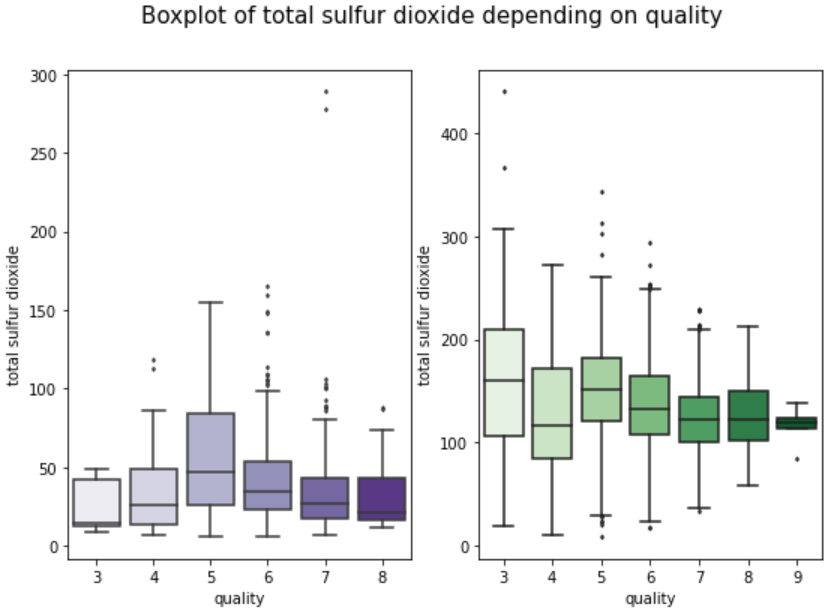

Figure III.B.4: Boxplot of total sulfur dioxide depending on quality

Total sulfur dioxide is an attribute that we keep for red wine. It is difficult to make some deductions from our graph, because the level in total sulfur dioxide for bad and high quality red wine is more or less the same.

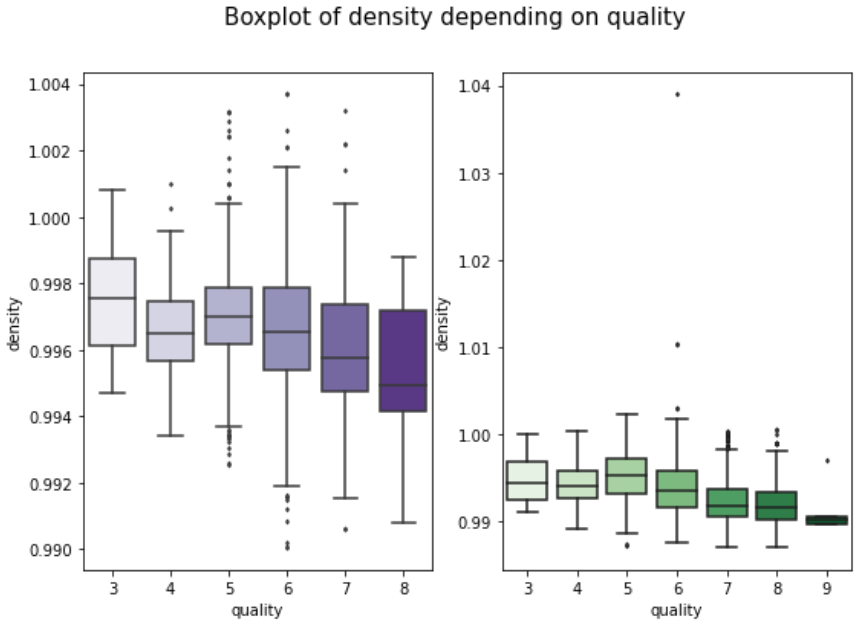

Figure III.B.5: Boxplot of density depending on quality

Density is kept for both wines. We can see a clear difference between density in red and white wines. However, lower the density is higher the quality of white wines is. It is the same for red wine, but not at the same scale.

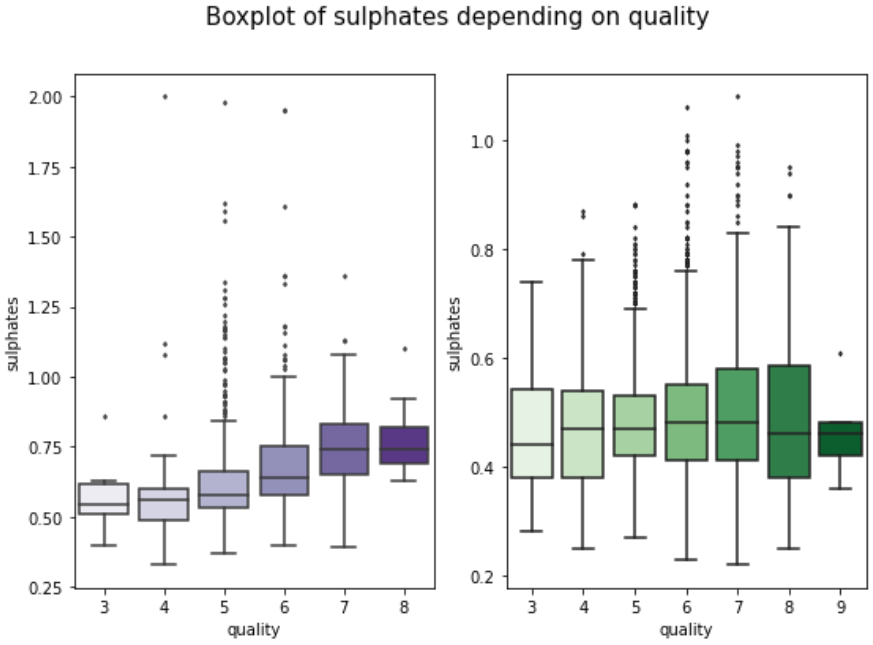

Figure III.B.6: Boxplot of sulphates depending on quality

Sulphates are for red wines. We can see that there are a lot of outliers for the middle quality wines. However, we can clearly see that the quality improves when the sulphates are higher.

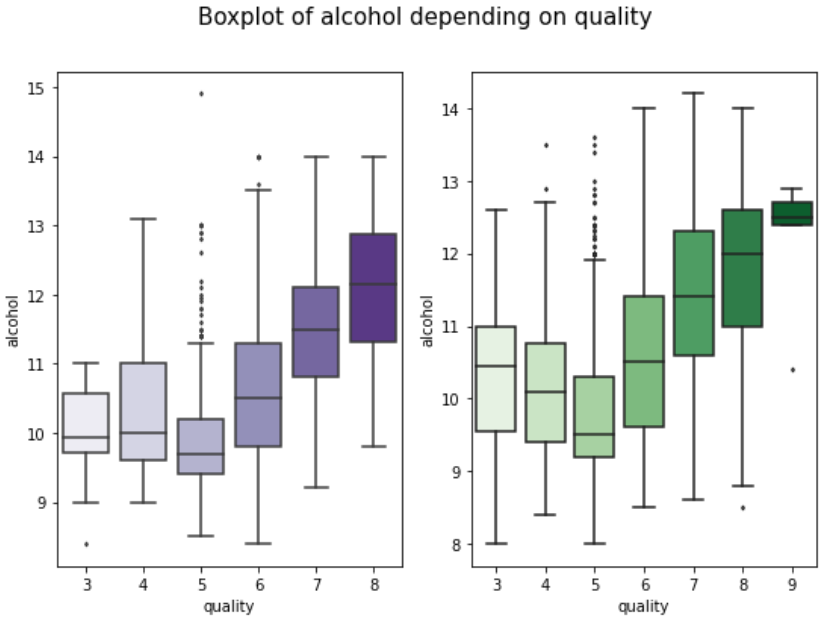

Figure III.B.7: Boxplot of alcohol depending on quality

Finally, alcohol is an important attribute for both wines. We can see that the two graphs follow the same curve, showing that the higher the rate of alcohol is, the better the wine taste.

Conclusion

The more important things to keep in mind are that the different attributes that we choose to keep really have an impact on the quality of the different wines. Alcohol and density, commun to the two dataset, are really important.

III.C - Chi-squared test

We are using a chi-2 test in order to test the dependence on the type of wine and the quality of wine (red wine or white wine).
The null hypothesis stands that the quality of wine and the type of wine is independant.

With the chi-2 test, we are going to find the p-value. If the p-value is smaller than 0.05, we can reject the null hypothesis. If not, we can't reject it.

The output is the following :

```
chi2 between  wineType  and  quality is:  (33.11103636947641, 9.982377417250757e-06, 6,
array([[7.38340773e+00, 2.26165923e+01],
```

So, the p-value of the test is 9,98 e-06 (because it's the second value of the output).
We find that the p-value is smaller than 0.05. Consequently, we can reject the null hypothesis and conclude on
the existence of a relationship (dependence link) between the quality of wine and the type of wine.

To find the strength of the dependency between the quality and the type of wine, we are going to use the Cramer test on datas. We use this test because the shape of the cross table we have is rectangle.

III.D - Cramer's V

Before calculating Cramer's coefficient for every column from our dataset, we concatenate together white and red wines and create a type column to test the correlation between all other columns and the kind of wine.
We do a double loop to calculate every possible combination of columns. Here are some interesting results:
First, we can see that the taste acidity of a wine is related to his pH:

```
The cramer coefficient between  pH  and  fixed acidity is:  0.7214702482747769
```

Other interestings things can be observed about the acidity: for example, a wine with a higher alcohol percentage will be more acid:

```
The cramer coefficient between  alcohol  and  citric acid is:  0.8415018242655289
```

One last interesting thing is that the type of wine is weakly correlated to the quality:

```
The cramer coefficient between  quality  and  wineType is:  0.15856936301353985
```

The red and white wines can't be exclusively distinguished by their quality.

IV - Predictions

Now that we have analyzed our dataset we want to be able to predict good wines. In order to do so we try different models.

IV.A - Naive Bayes

First we tried the Naive Bayes model which gives us the results below.

Red wines

	precision	recall	f1-score	support
0	0.92	0.86	0.89	309
1	0.42	0.57	0.48	54
accuracy			0.82	363
macro avg	0.67	0.72	0.69	363
weighted avg	0.85	0.82	0.83	363

White wines

	precision	recall	f1-score	support
0	0.88	0.77	0.82	875
1	0.44	0.63	0.52	247
accuracy			0.74	1122
macro avg	0.66	0.70	0.67	1122
weighted avg	0.78	0.74	0.76	1122

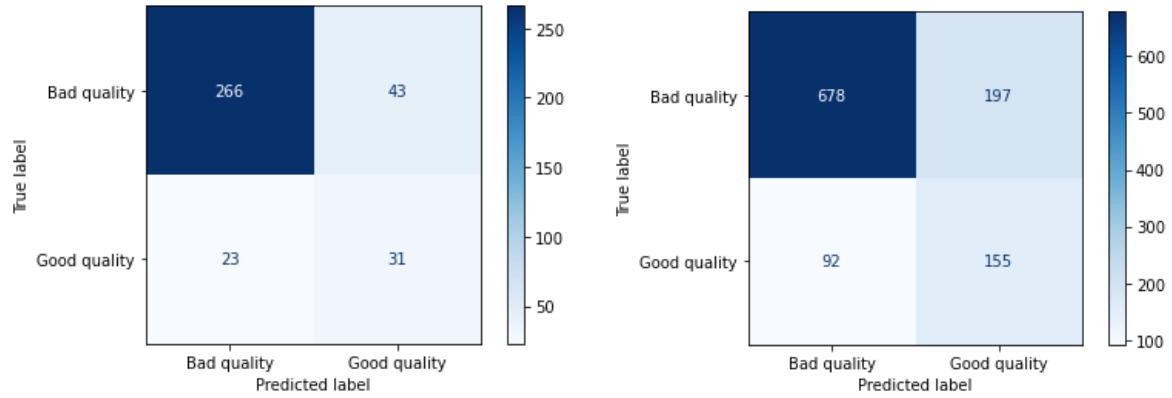

Figure IV.A: Results of the Naive Bayes Model

For both red and white wines, we can see that the Naive Bayes model does not really well predict good wine, only 42% or 0.44%.

IV.B - Decision Tree

After, we tried the Decision Tree model which gives us the results below.

Red wines

	precision	recall	f1-score	support
0	0.91	0.93	0.92	309
1	0.55	0.50	0.52	54
accuracy			0.87	363
macro avg	0.73	0.71	0.72	363
weighted avg	0.86	0.87	0.86	363

White wines

	precision	recall	f1-score	support
0	0.90	0.88	0.89	875
1	0.59	0.65	0.62	247
accuracy			0.83	1122
macro avg	0.75	0.76	0.75	1122
weighted avg	0.83	0.83	0.83	1122

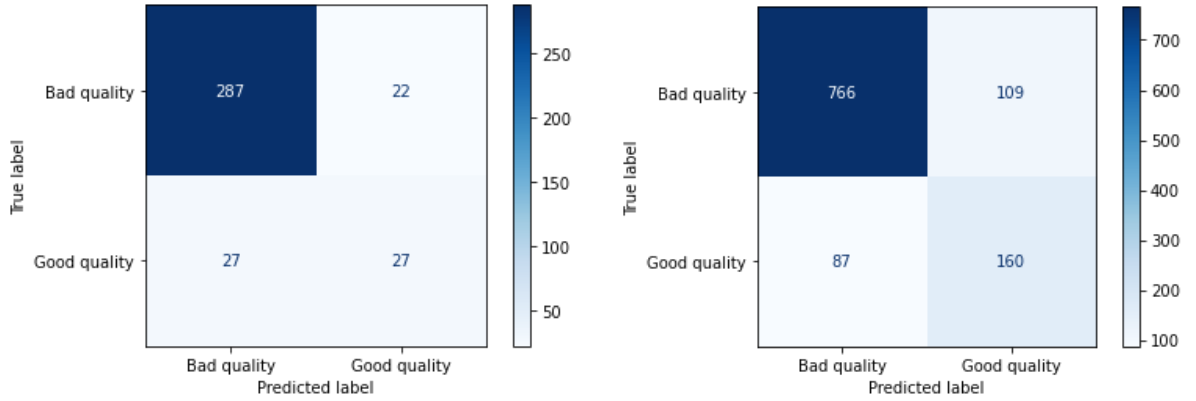

Figure IV.B:Results of the Decision Tree Model

Decision tree is a bit better, with a precision rate between 55% and 59% for predicting good wines.

IV.C - Random Forest

Then, we ran the Decision Tree model which gives us the results below.

Red wines

	precision	recall	f1-score	support
0	0.93	0.97	0.95	309
1	0.75	0.56	0.64	54
accuracy			0.91	363
macro avg	0.84	0.76	0.79	363
weighted avg	0.90	0.91	0.90	363

White wines

	precision	recall	f1-score	support
0	0.89	0.95	0.92	875
1	0.75	0.57	0.65	247
accuracy			0.86	1122
macro avg	0.82	0.76	0.78	1122
weighted avg	0.86	0.86	0.86	1122

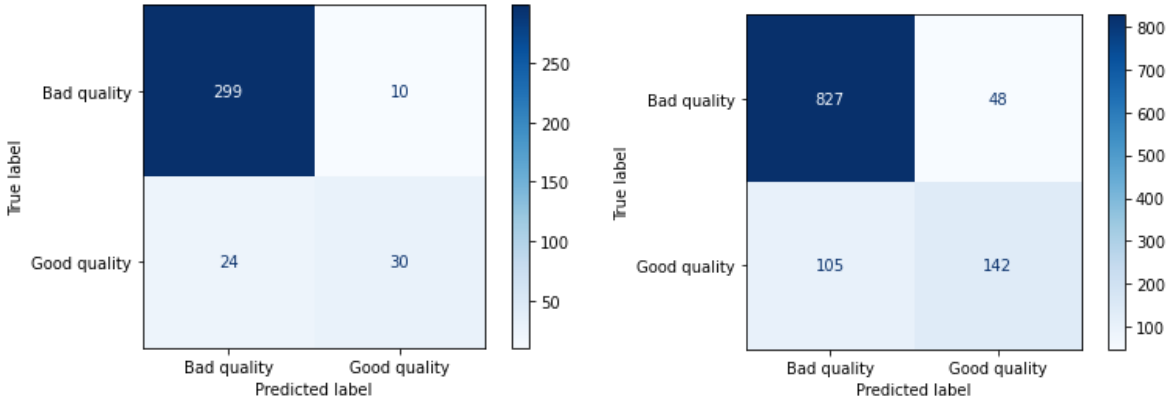

Figure IV.C:Results of theRandom Forest model

Random forest is a lot better than our two previous models. With a precision rate between 89% and 93% for predicting bad quality wines and a precision rate of 75% for good wine. The random forest will certainly be our preferred model.

IV.D - K Nearest Neighbors

Then, we ran the K Nearest Neighbors model which gives us the results below.

Red wines

	precision	recall	f1-score	support
0	0.89	0.95	0.92	309
1	0.52	0.31	0.39	54
accuracy			0.85	363
macro avg	0.70	0.63	0.65	363
weighted avg	0.83	0.85	0.84	363

White wines

	precision	recall	f1-score	support
0	0.86	0.91	0.88	875
1	0.60	0.47	0.53	247
accuracy			0.81	1122
macro avg	0.73	0.69	0.71	1122
weighted avg	0.80	0.81	0.81	1122

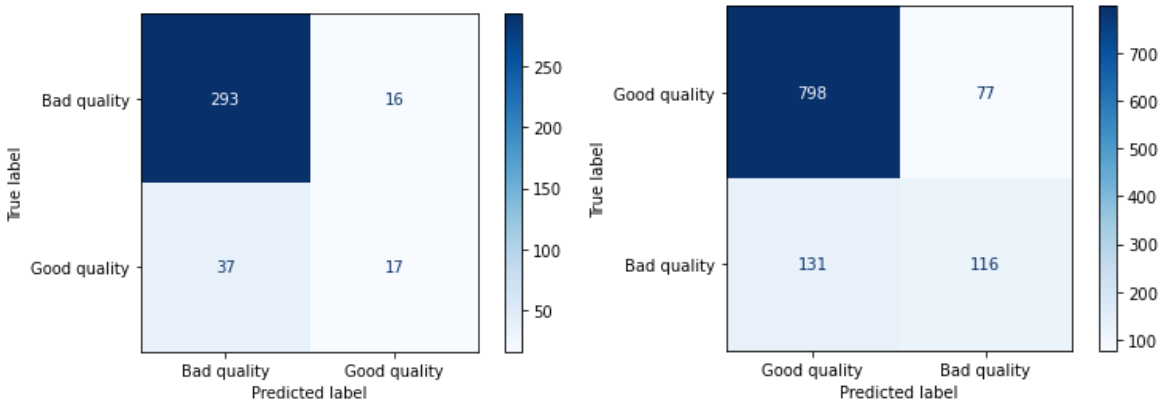

Figure IV.D:Results of the K Nearest Neighbors

The KNN model does not give us more satisfying results than the Random Forest algorithm.

IV.E - Support-Vector Machine

Then, we ran the Support-Vector Machine model which gives us the results below.

Red wines

	precision	recall	f1-score	support
0	0.89	0.97	0.93	309
1	0.63	0.31	0.42	54
accuracy			0.87	363
macro avg	0.76	0.64	0.67	363
weighted avg	0.85	0.87	0.85	363

White wines

	precision	recall	f1-score	support
0	0.83	0.96	0.89	875
1	0.67	0.31	0.43	247
accuracy			0.81	1122
macro avg	0.75	0.63	0.66	1122
weighted avg	0.80	0.81	0.79	1122

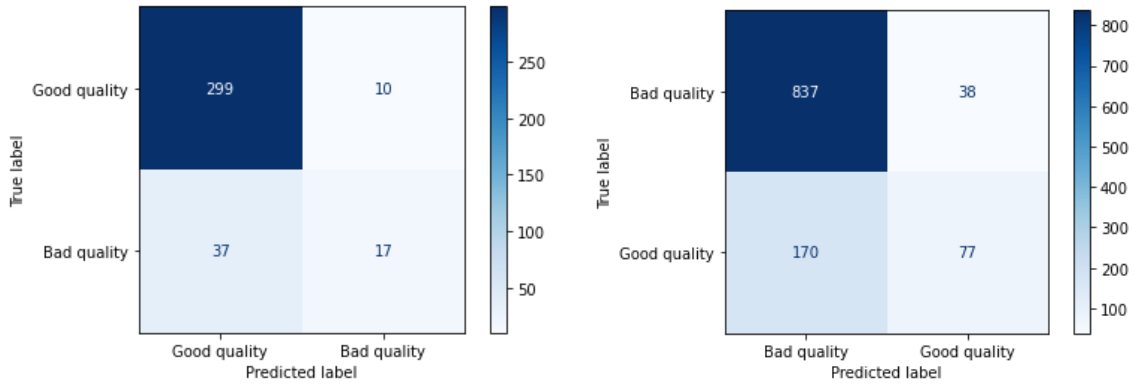

Figure IV.E:Results of the Support-Vector Machine

SVM could be our 2nd prefered algorithm with a good precision between 83% and 89%% for bad wines and 63% and 67% for good wines.

IV.F - Logistic regression

Finally, we ran the Logistic regression model which gives us the results below.

Red wines

	precision	recall	f1-score	support
0	0.90	0.95	0.92	309
1	0.56	0.37	0.44	54
accuracy			0.86	363
macro avg	0.73	0.66	0.68	363
weighted avg	0.85	0.86	0.85	363

White wines

	precision	recall	f1-score	support
0	0.82	0.93	0.87	875
1	0.52	0.26	0.35	247
accuracy			0.78	1122
macro avg	0.67	0.60	0.61	1122
weighted avg	0.75	0.78	0.76	1122

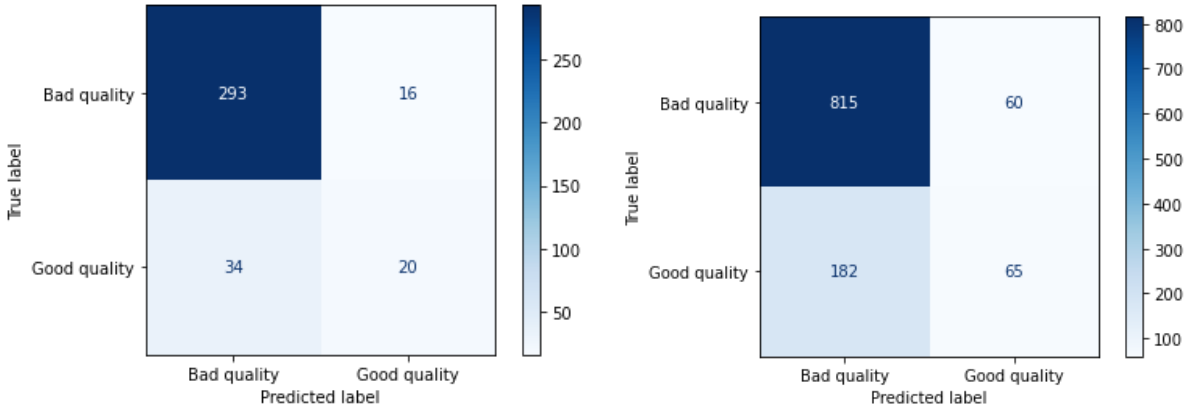

Figure IV.F: Results of the Logistic regression

Finally, logistic regression does not provide us good results, only a maximum of 56% of precision for good wines.

IV.G - What model do we choose ?

With the precedents results we can dress the following table:

Model	Naive Bayes	Decision tree	Random forest	K-Nearest Neighbors	Support-Vector Machine	Logistic regression
Precision (red wines good quality)	0.42	0.55	**0.75**	0.52	0.63	0.56
Precision (white wines good quality)	0.44	0.59	**0.75**	0.60	0.67	0.52

Figure IV.G.1: Comparison between models

The model that predicts the best good quality wines for both red and white wines is the random forest model with a precision of predicting good quality wines for both red and wines of 0.75.
We will apply it to understand which features are the most important.

Red wine White wine

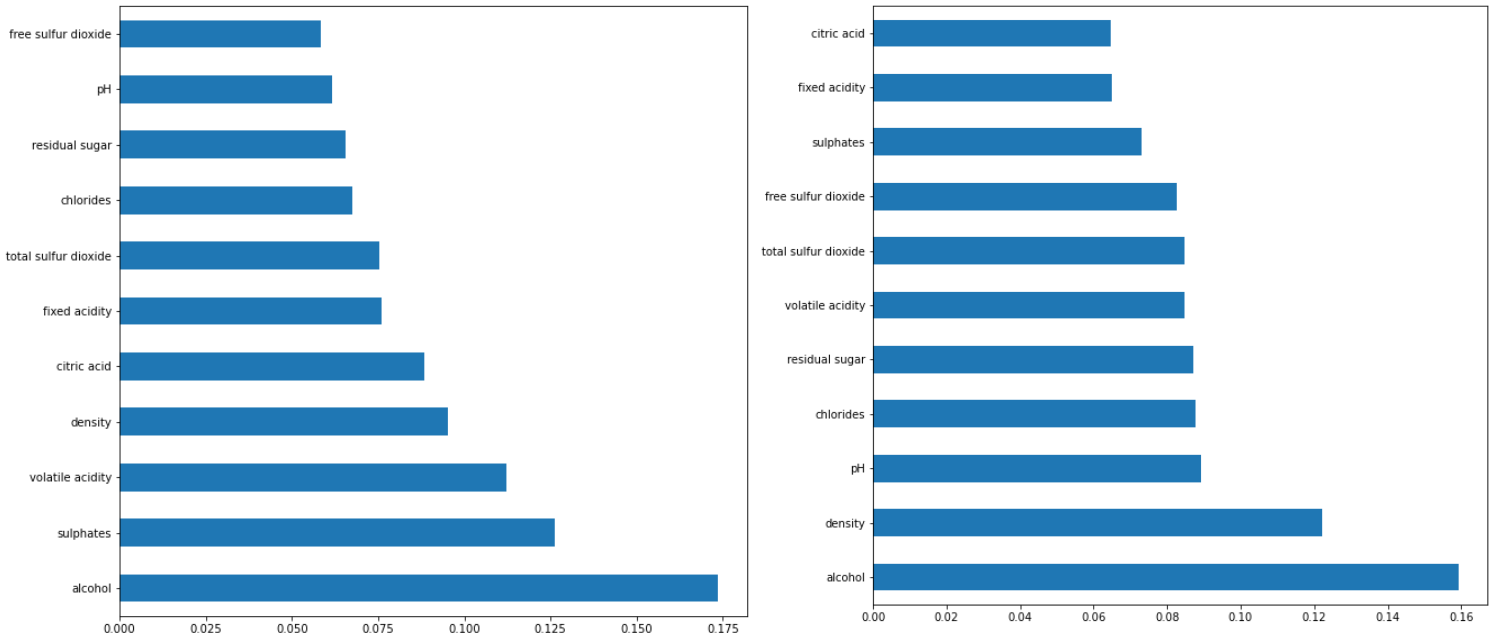

Figure IV.G.2: Features importance for red and white wines

By applying the random forest model we can clearly see what are the more important attributes for red and white wines.

Red wine: Alcohol, Sulphates, and Volatile acidity
White wine: Alcohol, density, pH

Red wines

Good quality Bad quality

	alcohol	sulphates	citric acid	volatile acidity	density
count	201.000000	201.000000	201.000000	201.000000	201.000000
mean	11.518491	0.743035	0.373731	0.410423	0.996056
std	0.940970	0.121829	0.192116	0.145665	0.001977
min	9.500000	0.470000	0.000000	0.120000	0.991570
25%	10.800000	0.650000	0.310000	0.310000	0.994730
50%	11.600000	0.740000	0.400000	0.370000	0.995720
75%	12.200000	0.820000	0.490000	0.500000	0.997320
max	13.600000	1.130000	0.760000	0.915000	1.002200

	alcohol	sulphates	citric acid	volatile acidity	density
count	1250.000000	1250.000000	1250.000000	1250.000000	1250.000000
mean	10.244627	0.626432	0.247960	0.541044	0.996815
std	0.919154	0.123685	0.184985	0.164956	0.001647
min	8.500000	0.330000	0.000000	0.160000	0.991500
25%	9.500000	0.540000	0.080000	0.420000	0.995772
50%	10.000000	0.600000	0.240000	0.540000	0.996800
75%	10.900000	0.680000	0.400000	0.645000	0.997815
max	13.500000	1.160000	0.790000	1.040000	1.002100

Figure IV.G.3: Comparison between good and bad quality wines

We can do a simple comparison between low and high quality red wines. We can clearly see a lot of differences for all means and median of the different attributes except density.

White wines

Good quality

	alcohol	density	pH	chlorides	residual sugar
count	1016.000000	1016.000000	1016.000000	1016.000000	1016.000000
mean	11.408842	0.992445	3.210531	0.037846	5.343652
std	1.252739	0.002799	0.154395	0.009791	4.322212
min	8.500000	0.987110	2.840000	0.012000	0.800000
25%	10.700000	0.990537	3.100000	0.031000	1.800000
50%	11.500000	0.991740	3.200000	0.036000	3.950000
75%	12.400000	0.993700	3.320000	0.044000	7.600000
max	14.200000	1.000600	3.640000	0.091000	19.250000

Bad quality

	alcohol	density	pH	chlorides	residual sugar
count	3471.000000	3471.000000	3471.000000	3471.000000	3471.000000
mean	10.282401	0.994413	3.181461	0.044727	6.730539
std	1.093141	0.002786	0.139418	0.011767	5.081763
min	8.400000	0.987940	2.790000	0.014000	0.600000
25%	9.400000	0.992200	3.090000	0.037000	1.700000
50%	10.100000	0.994300	3.170000	0.044000	6.200000
75%	11.000000	0.996540	3.270000	0.051000	10.450000
max	14.000000	1.001960	3.630000	0.110000	20.800000

Figure IV.G.3: Comparison between good and bad quality wines

We can do the same comparison for withe wines. We can clearly see a lot of differences for all means and median of the different attributes, especially for chlorides.

Conclusion

In conclusion to this part we can create a table with the theoretical most important values to have if we want to create an excellent wine (red or white).

Inevitably, each attribute has a role to play in the taste of wine, but we can say that the following attributes have a bigger role to play than the other one.

Attributes	Red Wine	White Wine
Alcohol (%)	11.51	11.4
Density (kg/m3)	0.996	0.992
pH	-	3.21
Chlorides (g.L-1)	-	0.038
Residual Sugar (g.L-1)	-	5.344
Sulphates (ppm)	0.743	-
Citric Acid (g/dm3)	0.374	-
Volatile acidity (g.100 mL-1)	0.41	-

Figure IV.conclusion: Theoretical mean of different attributes for excellent wine

IV.H - Improving the model

In order to improve the model we will do an Hyperparameter Tuning over our Random Forest model. The goal is to choose the best parameters for our model in order to be very accurate.
To do so, we apply a Grid Search algorithm to our model and obtain the following result:

Without Grid Search

Red wines

	precision	recall	f1-score	support
0	0.93	0.97	0.95	309
1	0.75	0.56	0.64	54
accuracy			0.91	363
macro avg	0.84	0.76	0.79	363
weighted avg	0.90	0.91	0.90	363

White wines

	precision	recall	f1-score	support
0	0.89	0.95	0.92	875
1	0.75	0.57	0.65	247
accuracy			0.86	1122
macro avg	0.82	0.76	0.78	1122
weighted avg	0.86	0.86	0.86	1122

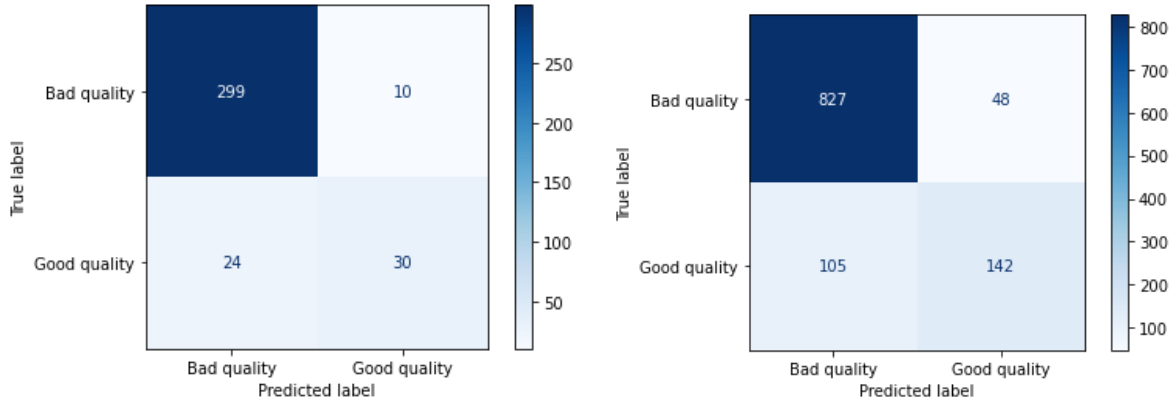

Figure IV.H.1:Results of the Random Forest model without grid search algorithm

With Grid Search

Red wines

	precision	recall	f1-score	support
0	0.91	0.97	0.94	309
1	0.72	0.43	0.53	54
accuracy			0.89	363
macro avg	0.81	0.70	0.74	363
weighted avg	0.88	0.89	0.88	363

White wines

	precision	recall	f1-score	support
0	0.87	0.95	0.91	875
1	0.74	0.49	0.59	247
accuracy			0.85	1122
macro avg	0.80	0.72	0.75	1122
weighted avg	0.84	0.85	0.84	1122

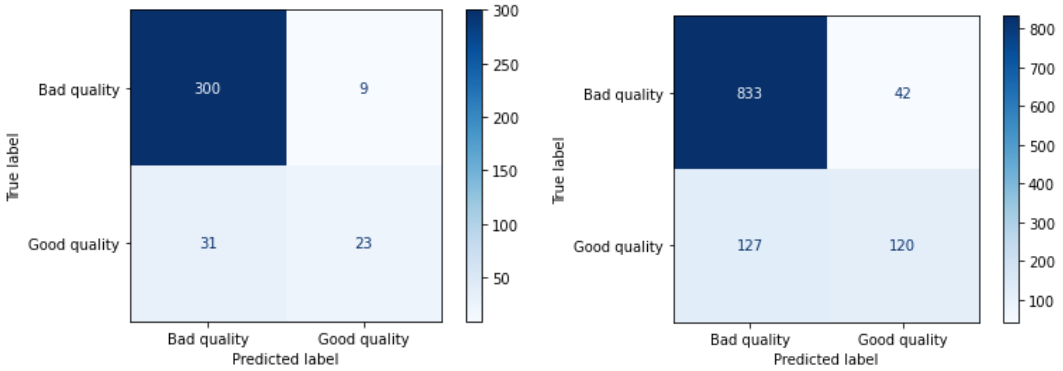

Figure IV.H.2: Results of the Random Forest model with grid search algorithm

We can see that our Hyperparameter Tuning over Random Forest did not really perform well. For both red and white wines we lose precision. However, we were able to predict more bad quality wines that were really bad quality wines but less high quality wines that were really high quality wines.

Conclusion

To conclude this part, we think that we could perform the grid search algorithm better in order to have more precise results.

Conclusion

To conclude the project, we have seen all the different attributes that can contribute to the quality of the Vinho Verde wine. We wanted to have the theoretical numeric values of the attributes that are making a wine taste good.

Vinho Verde Red Wines:

- Alcohol: around 11,5%
- Density: around 0.996 kg/m3
- Sulphates: around 0.743 ppm
- Citric acid: around 0.374 g/dm3
- Volatile acidity: around 0.41 g.100 mL-1

Vinho Verde White Wines:

- Alcohol: around 11,4%
- Density: around 0.992 kg/m3
- pH: around 3.21
- Chlorides: around 0.038 g.L-1
- Residual Sugar: around 5.344 g.L-1

However, it is important to keep in mind that the attributes are not the only factors that are making a wine taste good. The tools, the techniques, the means of transportation or conservation time are also factors which can improve or decrease the quality of a wine.

In addition, we could improve our theoretical results by:

- balancing the data: our data is very unbalanced between red and white wines, and between good and bad quality wines
- managing the outliers: we managed the outliers by using a Z-score method, but we can improve it
- finding the best parameter for our model by using algorithms
- using larger dataset
- applying our model to other wine types

Ressources

Outliers
Link 1

Regulations and recommendations on wine
Link 1
Link 2
Link 3

Description on the differents attributes
Link 1
Link 2
Link 3
Link 4
Link 5
Link 6
Link 7
Link 8
Link 9
Link 10
Link 11
Link 12
Link 13

PCA on Wine Quality Dataset
Link 1

Several Classification Techniques Wine Quality Dataset
Link 1
Link 2
Link 3

Slides
Link 1

Hyperparameter tuning
Link 1


isep
École d'ingénieurs du numérique