Systems with Machine Learning

Task 2: Data Preparation

Topic: Wine Quality Analysis

1. Data Description

The dataset contains chemical descriptions of 6499 Portuguese “Vinho Verde” wines. There are 4899 entries for white wine, and 1600 entries for red wines. The description of this dataset can be obtained from the UCI website. They are described as follows:

	ATTRIBUTES	DESCRIPTION
1	**pH**	To measure ripeness
2	**Density**	Density in gram per cm3
3	**Alcohol**	Volume of alcohol in %
4	**Fixed Acidity**	Impart sourness and resist microbial infection, measured in number of grams of tartaric acid per dm3.
5	**Volatile Acidity**	Number of grams of acetic acid per dm3 of wine
6	**Citric Acid**	Number of grams of citric acid per dm3 of wine
7	**Residual Sugar**	Remaining sugar after fermentation stops
8	**Chlorides**	Number of grams of sodium chloride per dm3 of wine
9	**Free Sulfur dioxide**	Number of grams of free sulfites per dm3 of wine
10	**Total Sulfur dioxide**	Number of grams of total sulfite (free sulphite + bound)
11	**Sulphates**	Number of grams of potassium sulfate per dm3 of wine
12	**Quality**	Target variable, 1-10

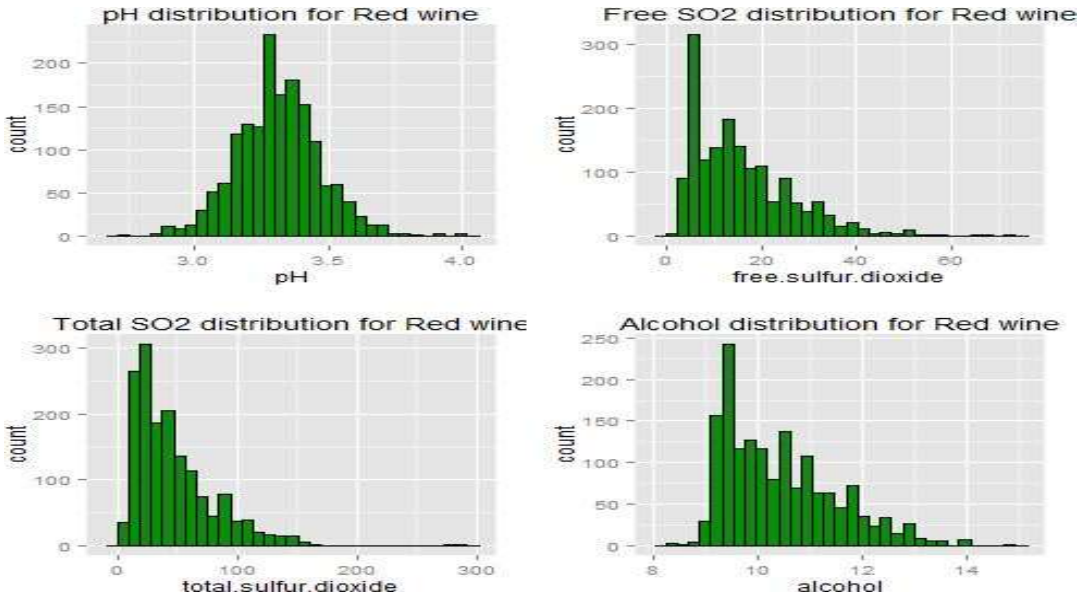
The source of the data is taken from the UCL Machine Learning Repository https://archive.ics.uci.edu/ml/index.php., provided by Paulo Cortez, from the University of Minho, Portugal.

2. Data Collection

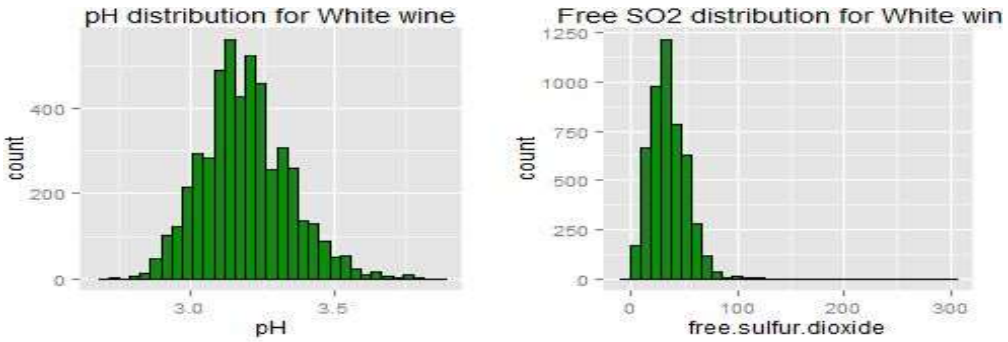
Observations from data sample

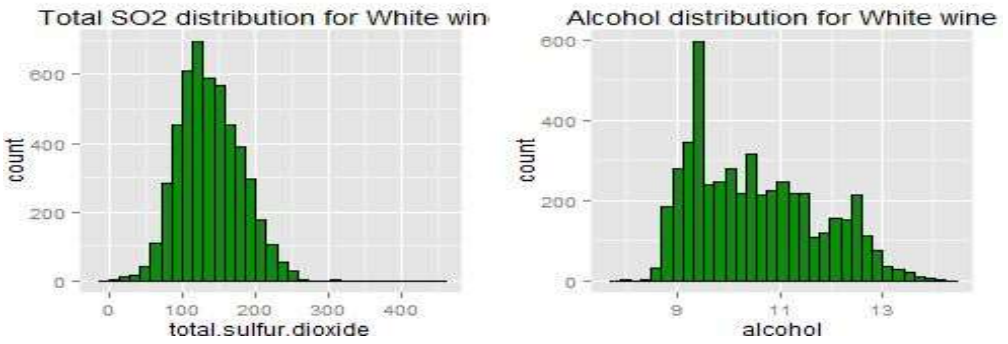
1. There is a big range for sulfur.dioxide (both Free and Total) across the samples.
2. The alcohol content varies from 8.00 to 14.90 for the samples in the dataset.
3. The quality of the samples range from 3 to 9 with 6 being the median.
4. The range for fixed acidity is quite high with minimum being 3.8 and maximum being 15.9,
5. pH value varies from 2.720 to 4.010 with a median being 3.210.

A distribution analysis of the key parameters pH, Total SO2, Free SO2 and Alcohol is explored across Red and White wine as follows:



1. The pH value seems to display a normal distribution with major samples exhibiting values between 3.0 and 3.5
2. The free sulfur dioxide seems to be between the 1-100 count with peaking around 50 mark
3. The total sulfur dioxide seems to have a spread between 0 and 300 and exhibiting a peak around 150.There seems to be a lower peak exhibited around the 50 mark.
4. The alcohol content seems to vary from 8 to 14 with major peaks around 10 with a lower count between 13 and 14.





The spread of the quality analysis for the Red and White types are given as below

The spread for the quality for both Red and White seems to exhibit similar normal distribution except for the fact that White wine distribution exhibit a peak quality around quality rating of 6 while Red wine exhibit a peak quality rating of approx 5.Only White wines seems to have been rated with a quality of 9 from the given sample.

3. Analyses of Data Relations and Correlations

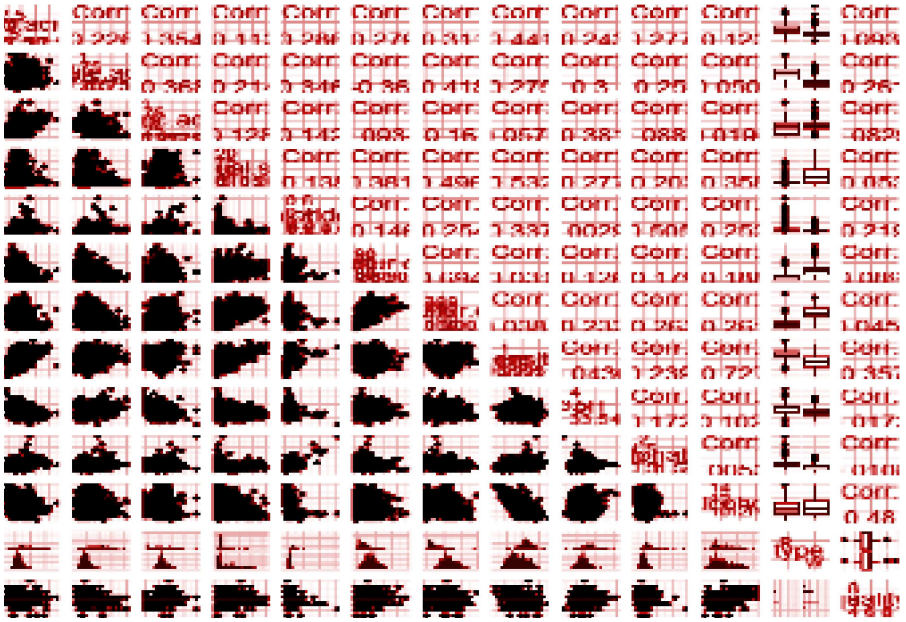
Analyze relations and correlations within the data (data groups, similar data, label/class correlations, sources, etc.) Present some examples.

The two most important features among all 12 attributes are Sulphur dioxide (both free and total) and Alcohol. LAST Volatile acidity contributes to acidic tastes and has a negative correlation to wine quality. SECOND The most important factor to decide the quality of wine is alcohol, higher concentration of alcohol leads to a better quality of wine and lower density of wine.

The dataset appears very clean, with no missing data and clear structure. All variables are numeric. The range of independent variables varies greatly, so when building the model we will normalize them to be within the same range. Next step we will check the pairwise relationship of each variable.

Running Scatterplot matrices

A scatterplot matrix is derived to understand the overall variable behaviour and correlations



Scatterplot output indicates the following correlation behaviour

Free SO2: Noticeable positive correlation with Total SO2 and Residual sugar Negative correlation with pH sulphates and Alcohol.

Total So2: Positive correlation between free so2 and residual sugar Negative Correlation with pH,Sulphates and Alcohol

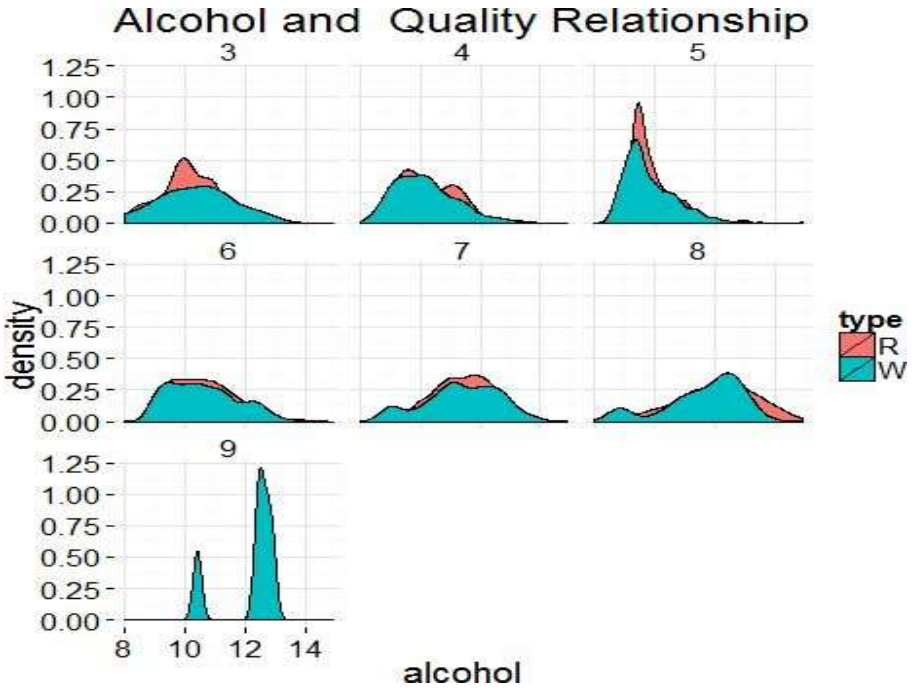
pH: Positive correlation with Sulfated,Alcohol and Volatile Acidity Negative correlation with Total and Free SO2,Residual sugar,citric acid,acidity(volatile and Fixed)

Alcohol: Positive correlation with pH and quality NEGATIVE Correlation with density,total and free so2,chlorides

Quality: positive correlation with alcohol negative correlation with density,chlorides,volatile acidity.

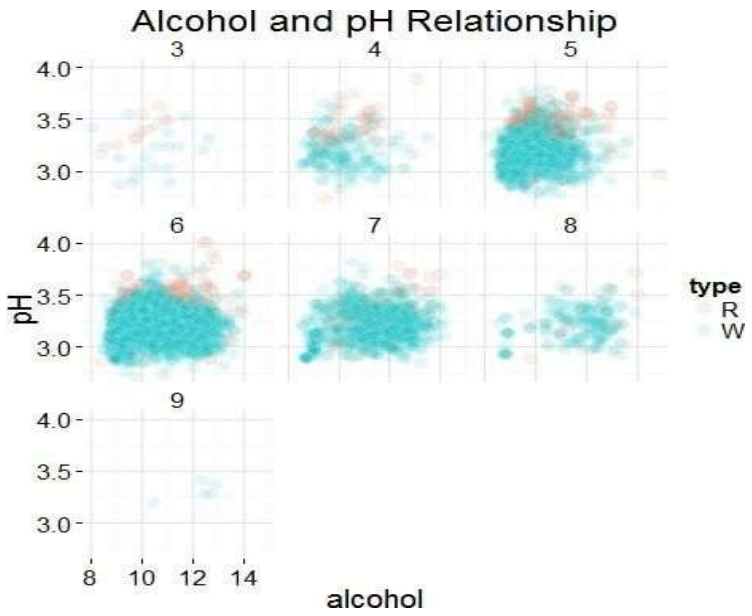
Red and White wine types: significant disparity in the mean ,median and q-q values for residual sugar,citric acid. fixed and volatile acidity

Alcohol: Scatterplot matrices indicate a strong positive correlation between Alcohol Content and Quality and without any bias towards the Red or White Wine.It will be interesting to see the distribution of Alcohol content across both Red and White wine



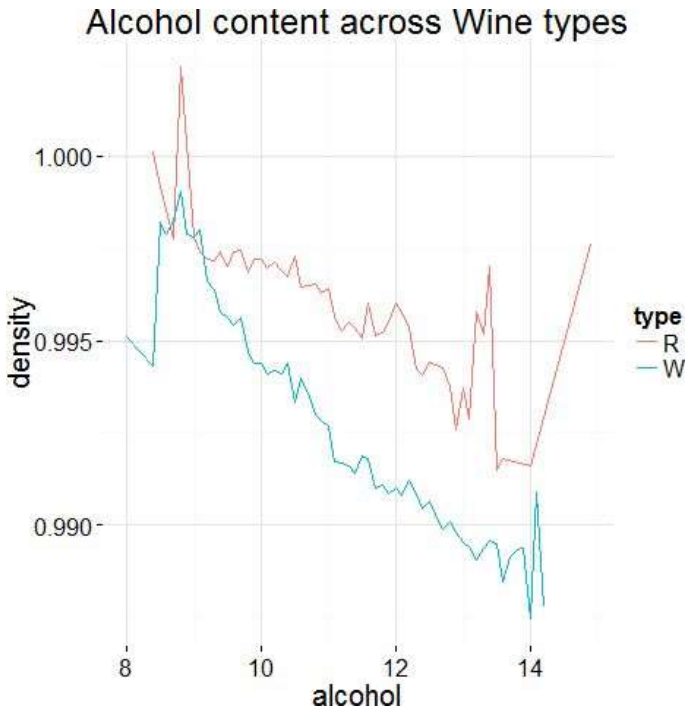
There seems to be no significant bias of the alcohol content even though there are samples with higher Alcohol content for Red wine exhibiting a higher density reading for the quality levels of 3 and 5 as compared to White Wine.

From our earlier scatterplot matrices, alcohol seems to exhibit a strong correlation with PH value.



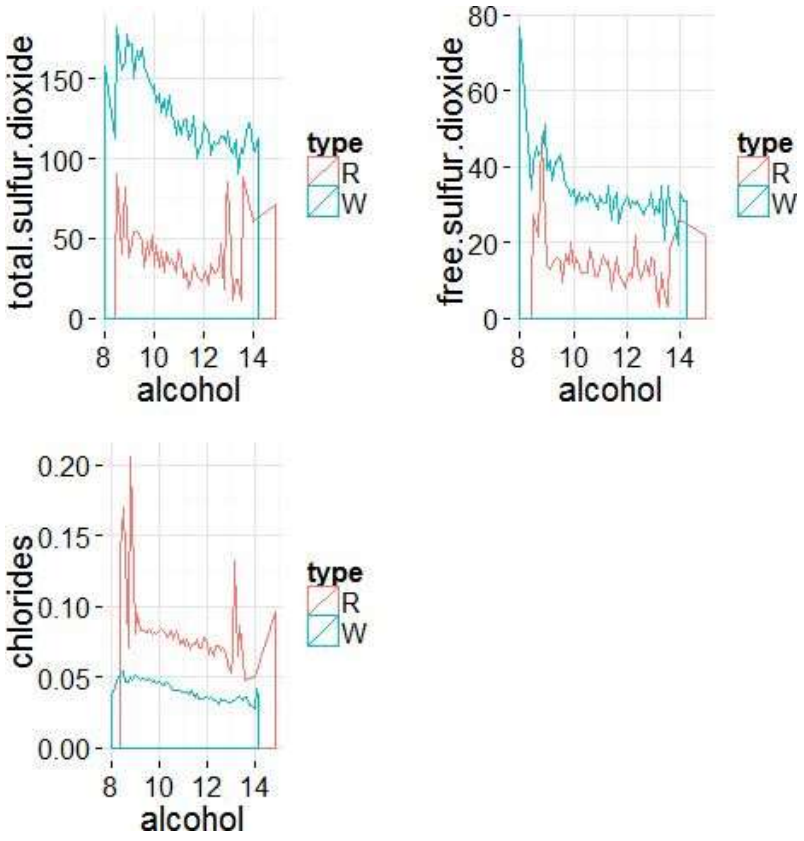
This indicates an interesting observation that as quality rating increases,the Red wine has instances of more PH value then White wine for similar Alcohol content.

Alcohol also exhibited a strong negative correlation with density and a further analysis on this is provided below.



As expected, there seems to dip in density with increase in the Alcohol content and the white wine exhibits a more prominent dip.

The negative correlation of Alcohol with Total and Free SO2 and Chlorides are analysed as below:



The observations from the above analysis are as follows:

Total SO2: White wine exhibits higher Total SO2 contents than Red wine across all Alcohol level Total SO2 content decreases with Alcohol content for White wine

Free SO2: Again White wine exhibits higher Free SO2 levels across all Alcohol content though the unit difference between Red and White wine seems to be lower as compared to the Total SO2 difference

The Free SO2 content decreases as the alcohol content increases for White wine.

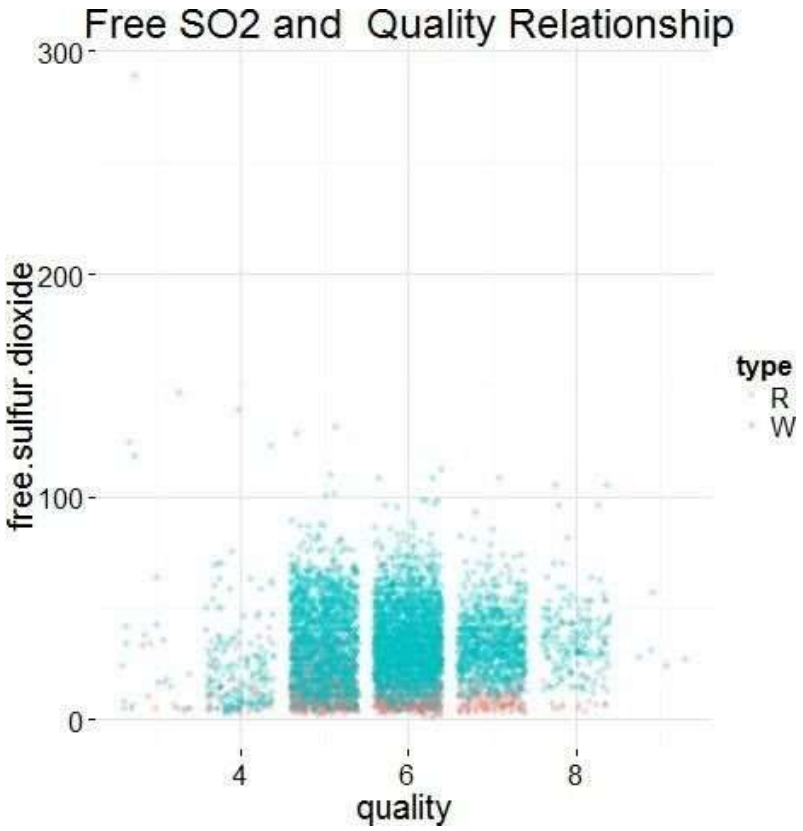
Chloride: Red wine indicated a higher chloride content than white wine with increasing Alcohol content

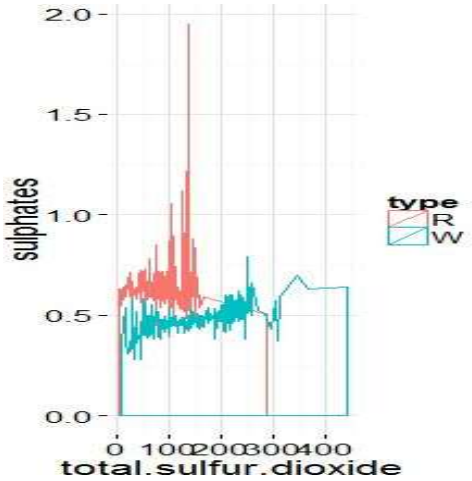
The Chloride content is quite high at lower Alcohol content between 8 and 9 but then exhibits steady reduction till Alcohol content level of 13 before a spike.

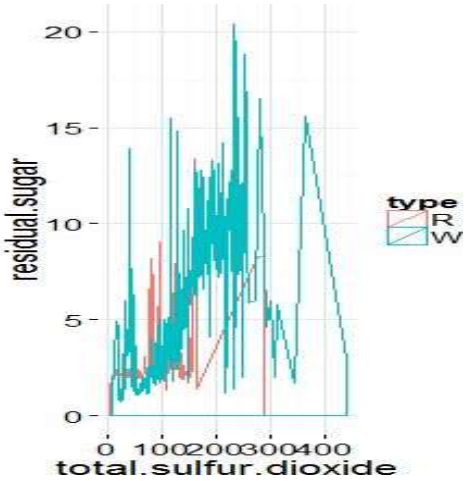
White wine exhibits lower Chloride levels across Alcohol content levels and holds a steady pattern throughout.

Sulphur Dioxide: Usage of SO2 in Wines has been a topic of discussion for a long time due to the health related issues.It will be interesting to see the distribution of SO2 across Red and White wine and their final impact on quality.

Analysis of Free SO2 across the Red and White wine is provided below:







White wine seems to exhibit a total SO2 level higher than 280 units

Residual Sugar

White wine exhibits high level of Residual sugar around 250 unit mark for Total SO2 as compared to Red wine and generally the quantity of Residual sugar seems to be on

higher after the 150 unit level for Total SO2

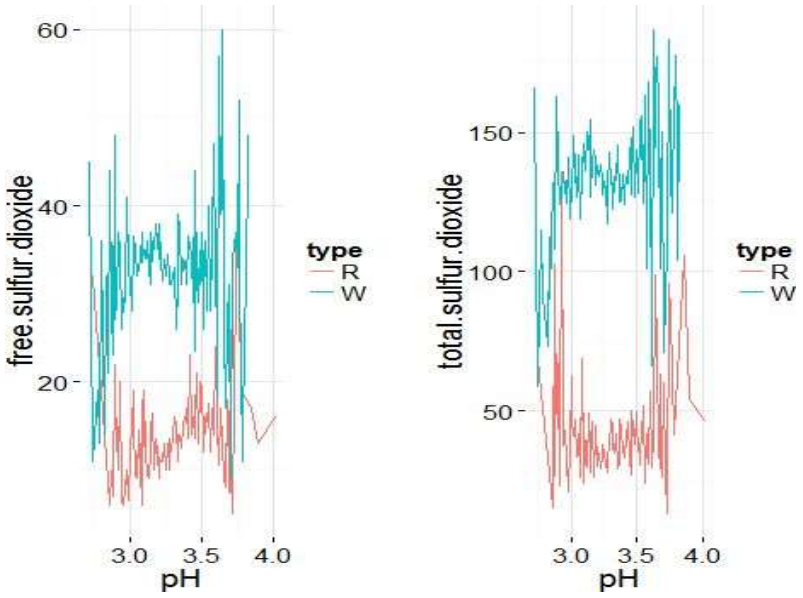
The relationship of Sulphate and Residual Sugar is analysed as below:

Sulphate
sulphate level is quite high for the red wine as compared to white wine. Red wine do not exhibit a Free SO2 level beyond 70 units

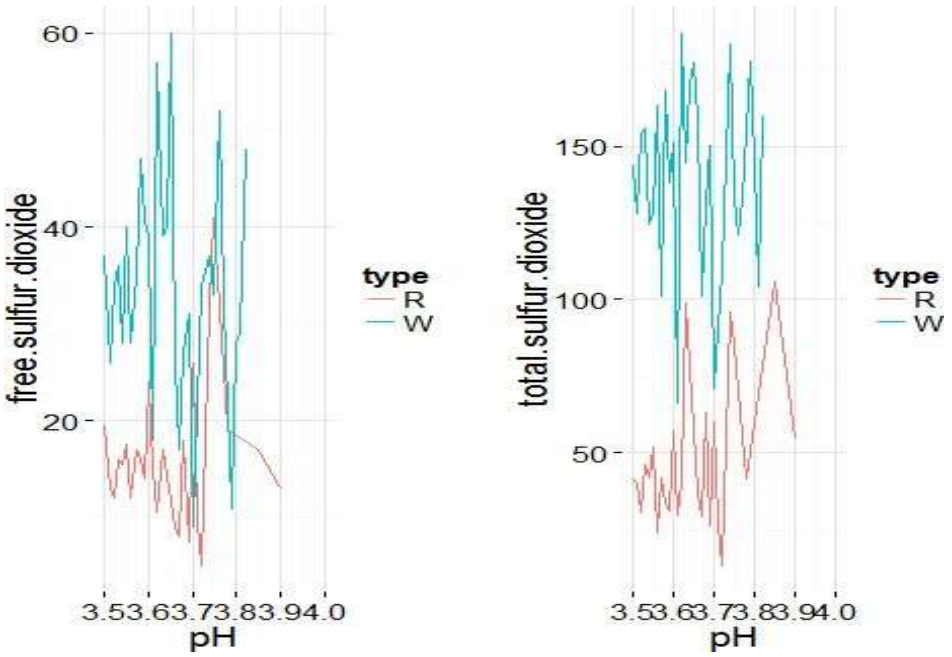
Residual Sugar
White wine exhibits a higher level of Residual sugar and has peaks around 150 mark.

Final Plots

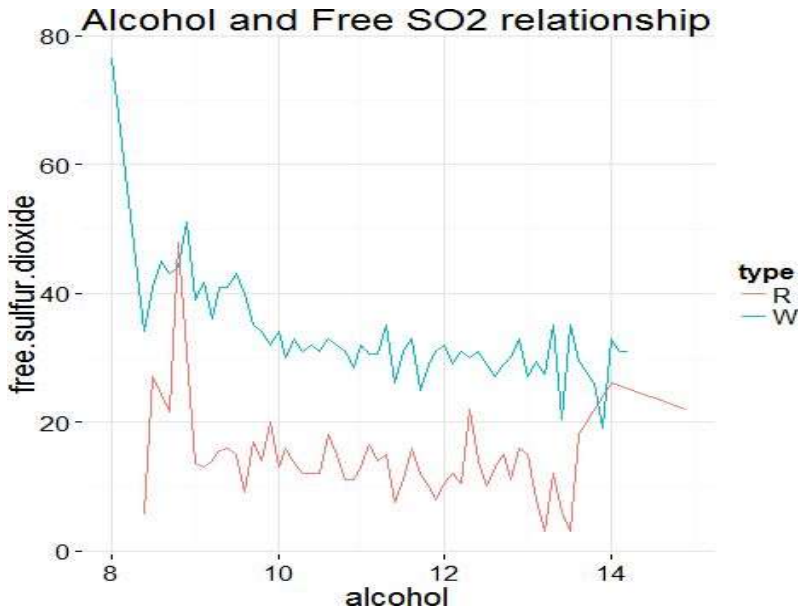
A final comparison is done between the Red and White wine to understand the difference between the two variants for the parameter of Total and Free SO2 and the PH values



The above plot indicates that white wine does exhibit higher SO2 components as compared to Red Wine for similar pH values across all pH values within the sample.There seems to be higher variation for both SO2 values in both Red and White wines between a pH value of 3.5 and 4.0.A closer look at these pH interval is given below.



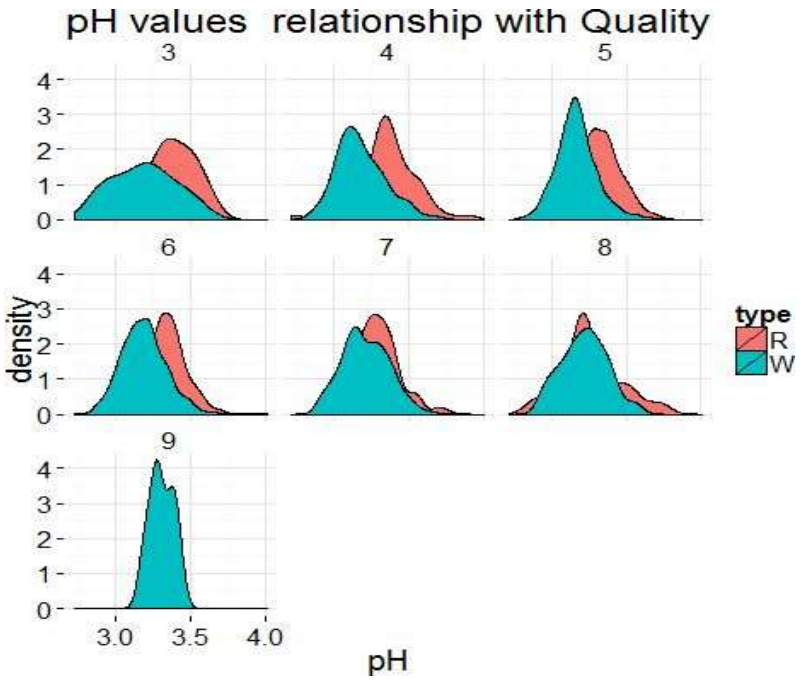
The above analysis plot indicate a high peak for free SO2 of unit 60 for a pH value of 3.65 while a high peak for Red wine for a pH value of 3.75 for 41 units.In case of Total so2,the peak of around 180 units for White at a pH level around 3.62 while Red wine exhibits a peak of around 105 units at a pH level of 3.85.Also it is observed that only Red wines in the sample has a Ph value beyond 3.85 and the Total and Free SO2 levels at this level is low.



The above plot indicates that for the same alcohol content,the content of free SO2 is higher for white wine than Red wine and also the free SO2 decreases quite significantly with increase in the alcohol content

Final Plot Three

Since pH value is considered to be a key contributor in determining the quality of wine,an analysis plot is created for both variants as given below:



From the above analysis plot,there doesn't seem to be any specific relations between pH values and quality in terms of the spread.However the Red wine tends to exhibit a higher pH value density then wine for quality rating till 7 while quality rating of 8 has more similar values of density. The quality rating of 9 exhibits a narrower spread for pH value between 3.1 and 3.6.

Summary

The analysis performed on the sample dataset can be summarised as below:

pH value is considered an important parameter when determining the quality of the Wine.The analysis over the samples however indicate that there is no specific values of pH which provides bias for quality ratings and a higher density of Red Wine samples did indicate a higher PH values as compared to White wine samples for the same quality ratings.These pH value however was found to be optimum between a value of 3.0 and 3.5 .A pH value of higher than 3.5 tends to exhibit a higher SO2 values which can be concern for people with concerns of health issues with SO2.Samples with higher alcohol content did exhibit lower SO2 counts and also White wine samples exhibited a higher level of SO2 components as compared to Red wine for the same level of Alcohol.

Some of the learnings from the analysis were as follows:

1. The understanding that Red Wine generally exhibits more SO2 properties than white

wine seems to be not true as per the samples considered.The analysis proves that White wine exhibits a higher level of SO2 properties.

2. It always seemed that pH value was a key factor in determining the quality of the wines but from the analysis ,it seems that pH value does not exhibit any patterns which can be utilized as akey deterministic variable for wine quality testing by sensory analysis.

3. From the samples analyzed,the wines with higher Alcohol content exhibited lower SO2 content as compared with samples with lower Alcohol content. 4.For the buyer conscious of the sugar content in the wines,White wine exhibits more residual sugar and we have seen spikes in the residual sugar for certain ranges of the Free and Total SO2 primarily with White wine.

A limitation of the current analysis is that the current data consists of samples collected from a specific portugal region. It will be interesting to obtain datasets across various wine making regions to eliminate any bias created by any specific qualities of the product,

4. Noise and Errors

None of the dataset cells is empty and they are filled with the proper data. We assume no errors occurred during the measuring process because we don't know what equipment was used.

5. Difficulties or Important data

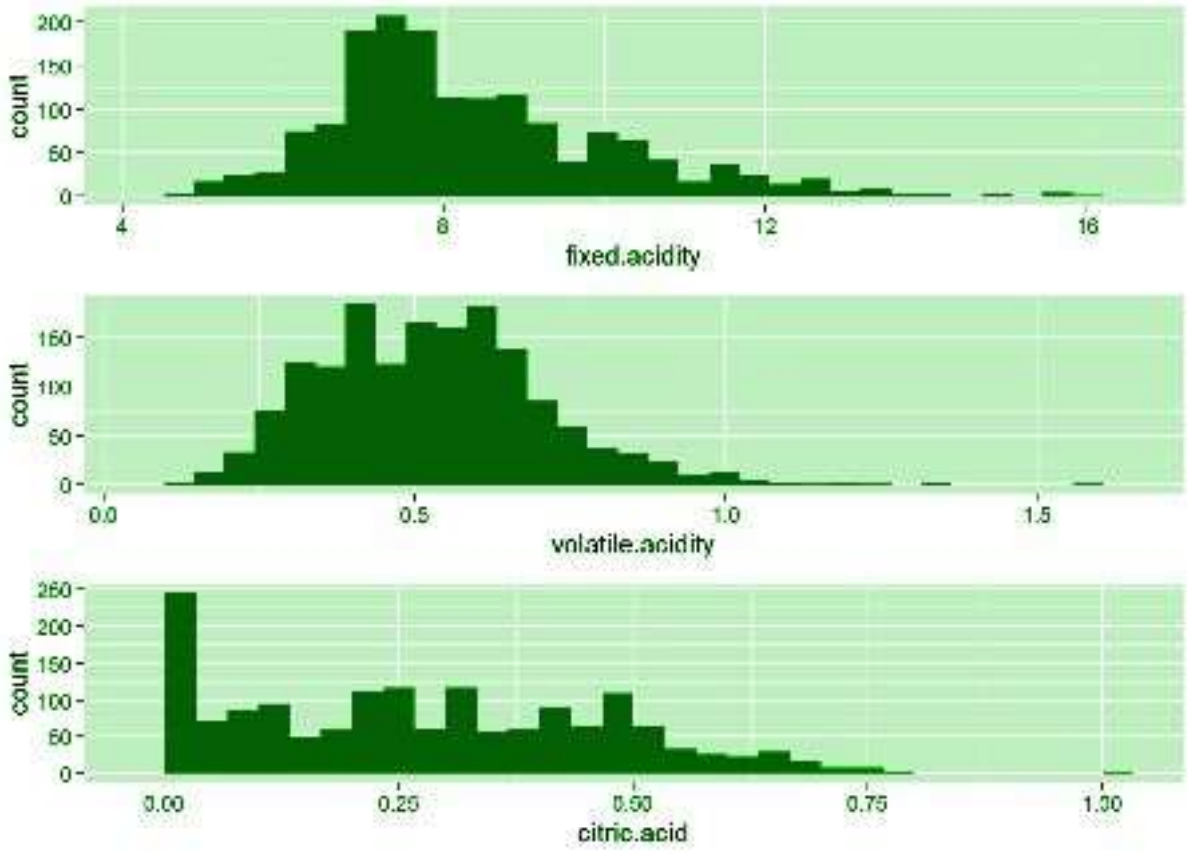
It appears that acid, sugar, and alcohol levels are the most important features when tasting and deciding wine quality. In particular, the balance among these factors to give a harmonized overall taste seems to be the main concern.

Citric acid is a bit unusual in that it displays an overall uniform distribution while having a huge peak at the lower level. This indicates citric acid level can be a very useful feature in the following analysis.

Aside from that, for the other distributions, it seems some sort of combination between the variables may be necessary to further explore the relationship between wine quality and different criteria - since the distributions vary from variable to variable.

We have two types of wines, and we need to train the neural network separately for each of those datasets.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
3.00	5.00	6.00	5.63	6.00	8.00

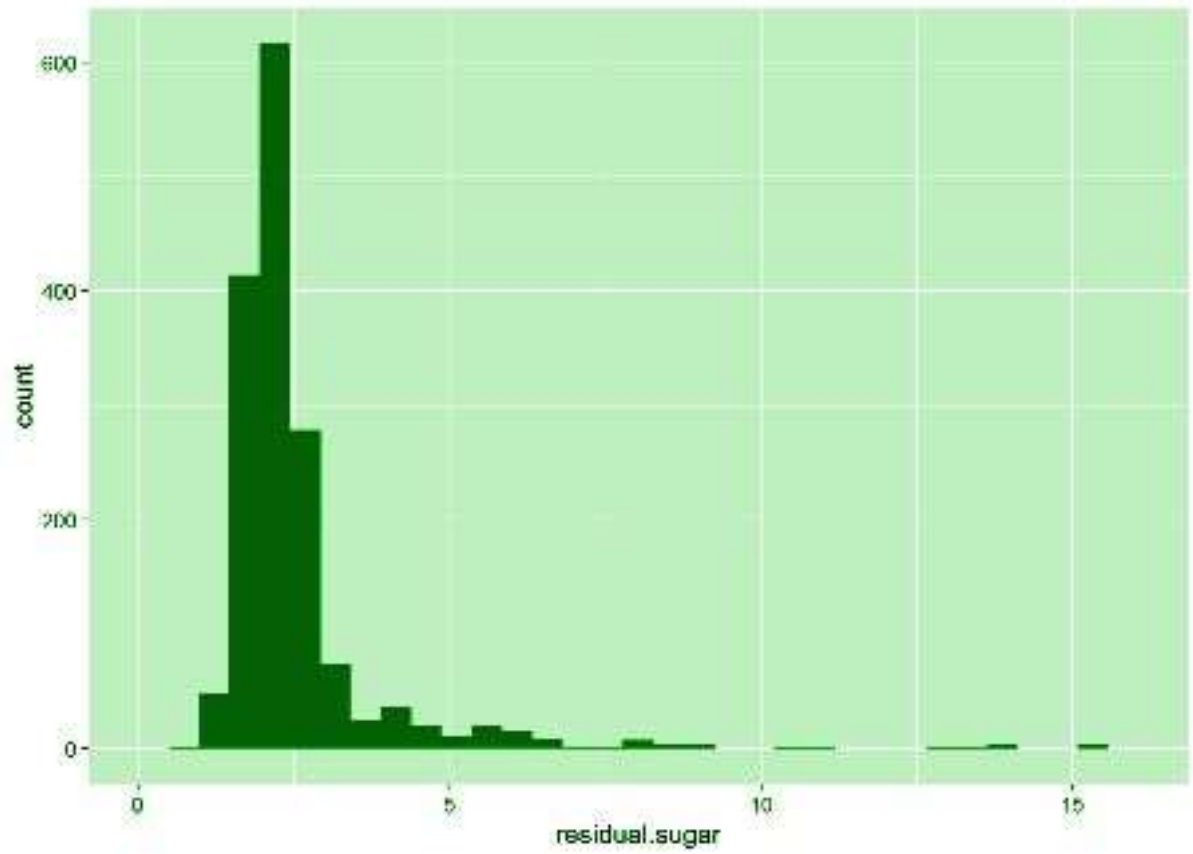


The above plots show that while both fixed and volatile acidity exhibiting somewhat normal distribution, citric acid is more uniform with a peak at the lower end. The following is some summary stats for these above variables.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
4.60	7.10	7.90	8.32	9.20	15.90

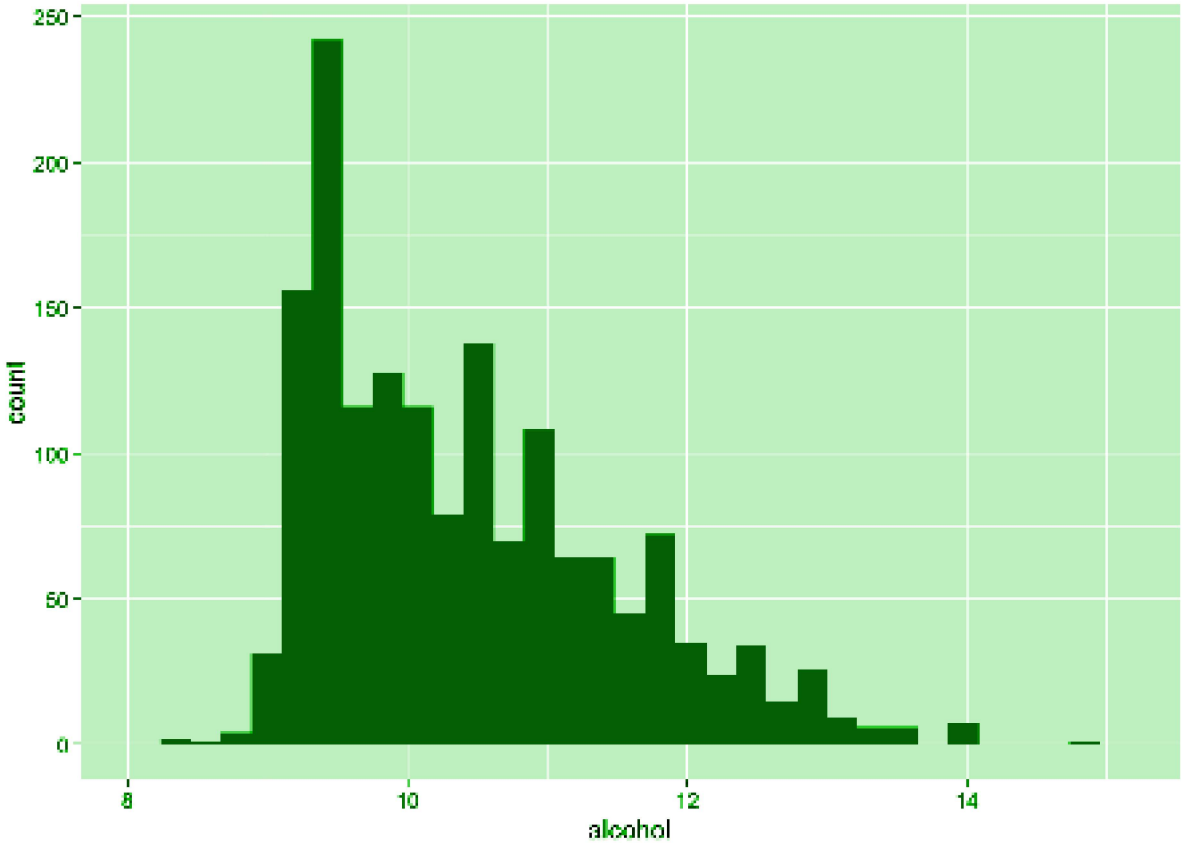
Min	1st Quartile	Median	Mean	3rd Quartile	Max
0.12	0.39	0.52	0.52	0.64	1.58

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0.00	0.09	0.26	0.27	0.42	1.00



Min	1st Quartile	Median	Mean	3rd Quartile	Max
0.90	1.90	2.20	2.54	2.60	15.50

The previous is the histogram and summary stats of residual.sugar. As it shows, the distribution is unimodal, nearly normal and right skewed. It seems that there are outliers in the higher end, i.e. high residual sugar levels. This may potentially be the wines that have higher quality or otherwise



Min	1st Quartile	Median	Mean	3rd Quartile	Max
8.40	9.50	10.20	10.42	11.10	14.90

The above plot shows alcohol distribution and the summary stats. Although it is not strictly unimodal, it does exhibit some strend as the alcohol level goes up - the count decreases.

6. Data Representation

Input variables will be represented as a vector (list) of parameters in particular order (based on physicochemical tests):

1. fixed acidity (tartaric acid - g / dm^3)
2. volatile acidity (acetic acid - g / dm^3)
3. citric acid (g / dm^3)
4. residual sugar (g / dm^3)
5. chlorides (sodium chloride - g / dm^3)
6. free sulfur dioxide (mg / dm^3)
7. total sulfur dioxide (mg / dm^3)
8. density (g / cm^3)
9. pH
10. sulphates (potassium sulphate - g / dm^3)
11. alcohol (% by volume)
12. type of wine(red or white)

Output variable will be represented as a integer value (based on sensory data):

1. quality (score between 0 and 10)

7. Data Normalization

The dataset describes each wine with 11 attributes and finally a quality grade. Each of those parameters has a different scale. E.g. fixed acidity for the white wines varies between 3.8 and 14.2 but sulfates vary between 0.22 and 1.08. It is a big difference. In order not to prioritize any of those attributes, each column will be scaled to the 0-1 scale using the min-max normalization algorithm. In such a case, the biggest value in each column will equal 1 and the lowest will equal 0, all other values will lie in between, relatively placed to the original distribution.

It can be achieved using a ***scikit-learn*** Python library. In the preprocessing package we can find many useful scalers. ***MinMaxScaler*** is exactly what we need.

```
from sklearn.preprocessing import MinMaxScaler
import pandas as pd

URL = 'http://[...]'

data = pd.read_csv(URL)
```

```
scaler = MinMaxScaler()
normalized_data  = scaler.fit_transform(data)
```

Basic statistics of normalized data may be found below:

Basic Red Wines

	MIN	MAX	AVG	STDEV	MEDIAN	MODE
fixed acidity	0	1	0.33	0.15	0.29	0.23
volatile acidity	0	1	0.28	0.12	0.27	0.33
citric acid	0	1	0.27	0.19	0.26	0.00
residual sugar	0	1	0.11	0.10	0.09	0.08
chlorides	0	1	0.13	0.08	0.11	0.11
free sulfur dioxide	0	1	0.21	0.15	0.18	0.07
total sulfur dioxide	0	1	0.14	0.12	0.11	0.08
density	0	1	0.49	0.14	0.49	0.52
pH	0	1	0.45	0.12	0.45	0.44
sulfates	0	1	0.20	0.10	0.17	0.16
alcohol	0	1	0.31	0.16	0.28	0.17

White Wines

	MIN	MAX	AVG	STDEV	MEDIAN	MODE
fixed acidity	0	1	0.29	0.08	0.29	0.29
volatile acidity	0	1	0.19	0.10	0.18	0.20
citric acid	0	1	0.20	0.07	0.19	0.18
residual sugar	0	1	0.09	0.08	0.07	0.01
chlorides	0	1	0.11	0.06	0.10	0.10

free sulfur dioxide	0	1	0.12	0.06	0.11	0.09
total sulfur dioxide	0	1	0.30	0.10	0.29	0.24
density	0	1	0.13	0.06	0.13	0.09
pH	0	1	0.43	0.14	0.42	0.38
sulfates	0	1	0.31	0.13	0.29	0.33
alcohol	0	1	0.41	0.20	0.39	0.23

8. Data Augmentation

In our case data augmentation is not applicable. It would be extremely hard to extend the dataset with artificial records as we don't know what parameters are crucial during the assessment of wine quality. Randomly generated values might even diminish the effectiveness of the neural network.

9. Data Partitioning

For a base of the partitioning ratio, we will split the dataset into three groups:

1. TRAIN 70%
2. VALIDATION 15%
3. TEST 15%

It may be achieved with the ***train_test_split*** function that can be found in the ***model_selection*** package of **scikit-learn** library. We are going to create 3 splits.

- SPLIT 1 - original_data (TRAIN = 70%, VAL = 15%, TEST = 15%)
- SPLIT 2 - normalized data (TRAIN = 70%, VAL = 15%, TEST = 15%)
- SPLIT 3 - VAL is a subset of TRAIN (TRAIN = 85%, VAL = 15%, TEST = 15%)

```
from sklearn.model_selection import train_test_split

# SPLIT 1
x_train, x_val_test, y_train, y_val_test = train_test_split(
   data_original, qualities_original, train_size=0.70)

x_val, x_test, y_val, y_test = train_test_split(
   x_val_test, y_val_test, train_size=0.50)

# SPLIT 2
x_train, x_val_test, y_train, y_val_test = train_test_split(
   data_normalized, qualities_normalized, train_size=0.70)
```

```
x_val, x_test, y_val, y_test = train_test_split(
   x_val_test, y_val_test, train_size=0.50)

# SPLIT 3
x_train, x_test, y_train, y_test = train_test_split(
   data_normalized, qualities_normalized, train_size=0.85)

_, _, x_val, y_val = train_test_split(
   x_train, y_train, test_size = 15/85)
```

Splits in numbers

As ***train_test_split*** function splits the dataset randomly we are not capable of creating reproducible subsets of the data. The only statistics that won't change are quantities of each of subsets.

SPLIT 1 / 2:

Red Wines:

SPLIT 1 / 2				
	TRAIN	**VAL**	**TEST**	**TOTAL**
Ratio	70%	15%	15%	**100%**
Rows	1119	240	240	**1599**

White Wines:

SPLIT 1 / 2				
	TRAIN	**VAL**	**TEST**	**TOTAL**
Ratio	70%	15%	15%	**100%**
Rows	3429	735	735	**4898**

SPLIT 3:

Red Wines:

SPLIT 3				
	TRAIN	**VAL**	**TEST**	**TOTAL**
Ratio	85%	15%	15%	**115%**
Rows	1359	240	240	**1839**

White Wines:

SPLIT 3				
	TRAIN	**VAL**	**TEST**	**TOTAL**
Ratio	85%	15%	15%	**115%**
Rows	4163	735	735	**5633**