# Student: Christian Campos

# SID: 862080812

# Wine Quality

The dataset is tabular data.
This dataset contains information about physicochemical tests such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.

## Data Set Information:

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. Vinho verde is a unique product from the Minho (northwest) region of Portugal. Medium in alcohol, is it particularly appreciated due to its freshness (specially in the summer). Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available. Therefore, there is no data about grape types, wine brand, wine selling price, etc.
These datasets can be viewed as classification or regression tasks.
The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones).
Outlier detection algorithms could be used to detect the few excellent or poor wines.
Also, we are not sure if all input variables are relevant.
So it could be interesting to test feature selection methods.

Data set information and datasets are taken from the source provided below.
Source (https://archive.ics.uci.edu/ml/datasets/Wine+Quality)

In [1]:

```
import pandas as pd
import numpy as np

df_red_wine_quality = pd.read_csv("winequality-red.csv",sep=";")
df_white_wine_quality = pd.read_csv("winequality-white.csv",sep=";")
```

In [2]:

```
df_red_wine_quality.head()
```

Out[2]:

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcoho |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 |
| **1** | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.9968 | 3.20 | 0.68 | 9.8 |
| **2** | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.9970 | 3.26 | 0.65 | 9.8 |
| **3** | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.9980 | 3.16 | 0.58 | 9.8 |
| **4** | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 |

In [3]:

```
df_white_wine_quality.head()
```

Out[3]:

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcoho |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 7.0 | 0.27 | 0.36 | 20.7 | 0.045 | 45.0 | 170.0 | 1.0010 | 3.00 | 0.45 | 8.8 |
| **1** | 6.3 | 0.30 | 0.34 | 1.6 | 0.049 | 14.0 | 132.0 | 0.9940 | 3.30 | 0.49 | 9.5 |
| **2** | 8.1 | 0.28 | 0.40 | 6.9 | 0.050 | 30.0 | 97.0 | 0.9951 | 3.26 | 0.44 | 10.1 |
| **3** | 7.2 | 0.23 | 0.32 | 8.5 | 0.058 | 47.0 | 186.0 | 0.9956 | 3.19 | 0.40 | 9.9 |
| **4** | 7.2 | 0.23 | 0.32 | 8.5 | 0.058 | 47.0 | 186.0 | 0.9956 | 3.19 | 0.40 | 9.9 |

# Analysis Wine Quality Dataset Variables

I'm checking the red and white wine variables datatypes.
Whether or not their variables are quantitative or categorical or both.

In [4]:

```
df_red_wine_quality.dtypes
```

Out[4]:

```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
dtype: object
```

In [5]:

```
df_white_wine_quality.dtypes
```

Out[5]:

```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
dtype: object
```

Both red and white quality dataset variables are quantitative.
Red and wine quality variable is an int64.
Let's check if there's any missing datas in the red and white wine dataset variables.

In [6]:

```
df_red_wine_quality.isnull().any()
```

Out[6]:

```
fixed acidity           False
volatile acidity        False
citric acid             False
residual sugar          False
chlorides               False
free sulfur dioxide     False
total sulfur dioxide    False
density                 False
pH                      False
sulphates               False
alcohol                 False
quality                 False
dtype: bool
```

In [33]:

```
df_white_wine_quality.isnull().any()
```

Out[33]:

```
fixed acidity           False
volatile acidity        False
citric acid             False
residual sugar          False
chlorides               False
free sulfur dioxide     False
total sulfur dioxide    False
density                 False
pH                      False
sulphates               False
alcohol                 False
quality                 False
dtype: bool
```

There are no missing data in both red and white wine datasets.

# Statistics

We can measure the center of a variable with the mean(average) and the median (the middle value).
A measure of center gives us information about the "typical" value of a variable
The mean represents the typical value
We also need to measure the spread of the variable.
I'm only going to analysis "quality" variable of the red and white wines.

# Red Wine Quality

The quality score between 0 and 10.
I assumed the quality score between 0 and 4 is poor quality.
The quality score between 5 and 7 is normal quality.
The quality score between 8 and 10 is excellent quality.

Let's use the describe function on the red wine quality variable to show us a glimpse of the quality scores.

In [8]:

```
df_red_wine_quality["quality"].describe()
```

Out[8]:

```
count    1599.000000
mean        5.636023
std         0.807569
min         3.000000
25%         5.000000
50%         6.000000
75%         6.000000
max         8.000000
Name: quality, dtype: float64
```

In [9]:

```
df_red_wine_quality["quality"].mean()
```

Out[9]:

```
5.6360225140712945
```

The typical red wine quality score is around $5.6 \approx 6$.
A normal wine quality score.

In [10]:

```
df_red_wine_quality["quality"].min()
```

Out[10]:

```
3
```

The lowest red wine quality score is 3.
It's not zero which is great but not a good score.

In [11]:

```
df_red_wine_quality["quality"].median()
```

Out[11]:

```
6.0
```

The mean is around same quality score of the median value.
The median is also another resaonable value for the "typical" quality score.
The mean tells us that the red wine quality score is higher or less than median typical quality.
Let's look at the highest quality score of the wine quality selections.

In [12]:

```
df_red_wine_quality["quality"].max()
```

Out[12]:

```
8
```

The highest quality score is not far from the mean which is good.
The quality score is range from 0 - 10

In [13]:

```
df_red_wine_quality.loc[df_red_wine_quality["quality"].idxmax()]
```

Out[13]:

```
fixed acidity            7.9000
volatile acidity         0.3500
citric acid              0.4600
residual sugar           3.6000
chlorides                0.0780
free sulfur dioxide     15.0000
total sulfur dioxide    37.0000
density                  0.9973
pH                       3.3500
sulphates                0.8600
alcohol                 12.8000
quality                  8.0000
Name: 267, dtype: float64
```

The MAD is a mean of the absolute differences and the mean represents the "typical" value.

In [14]:

```
df_red_wine_quality["quality"].mad()
```

Out[14]:

```
0.6831779242889846
```

MAD is saying that the "typical" quality is 0.68 far away from the average quality.
Let's look the variance of the red wine quality.

In [15]:

```
df_red_wine_quality["quality"].var(ddof=0)
```

Out[15]:

```
0.6517605398308277
```

However, the variance units are different than standard deviation.
To correct the units of variance, we take the square root to obtain a more interpretable measure of spread, called the standard deviation

In [16]:

```
df_red_wine_quality["quality"].std()
```

Out[16]:

```
0.807569439734705
```

Standard deviation tells about the concentration of the data around the mean of the data set.
A low measure of Standard Deviation indicates that the data are less spread out.
A high value of Standard Deviation shows that the data in a set are spread apart from their mean average values.
Standard deviation is expressed in the same units as the data.
Let's look at the quantile.

In [17]:

```
df_red_wine_quality["quality"].quantile(.75)
```

Out[17]:

```
6.0
```

Quality red wine score is 6 between the median and the highest value but not the maximum.

# White Wine Quality

Just like the red wine quality dataset.
We're going to use statistics functions to give us glimpse of the white wine dataset.

In [18]:

```
df_white_wine_quality["quality"].describe()
```

Out[18]:

```
count    4898.000000
mean        5.877909
std         0.885639
min         3.000000
25%         5.000000
50%         6.000000
75%         6.000000
max         9.000000
Name: quality, dtype: float64
```

In [19]:

```
df_white_wine_quality["quality"].mean()
```

Out[19]:

```
5.87790935075541
```

The typical white wine quality is around $5.8 \approx 6$.

In [20]:

```
df_white_wine_quality["quality"].min()
```

Out[20]:

```
3
```

The lowest white wine quality score is 3. Just like the same quality score from the red wine.

In [21]:

```
df_white_wine_quality["quality"].median()
```

Out[21]:

```
6.0
```

White wine median quality score is 6.Just like the same quality score from the red wine.

In [22]:

```
df_white_wine_quality["quality"].max()
```

Out[22]:

```
9
```

The white quality max is higher than the red wine quality.

In [23]:

```
df_white_wine_quality.loc[df_white_wine_quality["quality"].idxmax()]
```

Out[23]:

```
fixed acidity             9.100
volatile acidity          0.270
citric acid               0.450
residual sugar           10.600
chlorides                 0.035
free sulfur dioxide      28.000
total sulfur dioxide    124.000
density                   0.997
pH                        3.200
sulphates                 0.460
alcohol                  10.400
quality                   9.000
Name: 774, dtype: float64
```

The white wine number 774 selection has the highest quality score of 9.

In [24]:

```
df_white_wine_quality["quality"].mad()
```

Out[24]:

```
0.6707927052833292
```

MAD is saying that the "typical" quality is 0.67 far away from the average quality.
Let's look the variance of the white wine quality.

In [25]:

```
df_white_wine_quality["quality"].var(ddof=0)
```

Out[25]:

```
0.78419554751975
```

To correct the variance units, we use the standard deviation.

In [26]:

```
df_white_wine_quality["quality"].std()
```

Out[26]:

```
0.8856385749678312
```

White wine quality is more spread than the red wine quality spread.

In [27]:

```
df_white_wine_quality["quality"].quantile(.75)
```

Out[27]:

```
6.0
```

Quality white wine score is 6 between the median and the highest value but not the maximum. Just like the red wine quality score.

# Visualizations

Graphics can help us understand how the values of a quantitative variable are distributed.

# Histograms

The standard visualization for a single quantitative variable is the histogram.
A histogram sorts the values into bins and uses bars to represent the number of values in each bin.

Red Wine dataset is shown in red and the white wine shown in white.

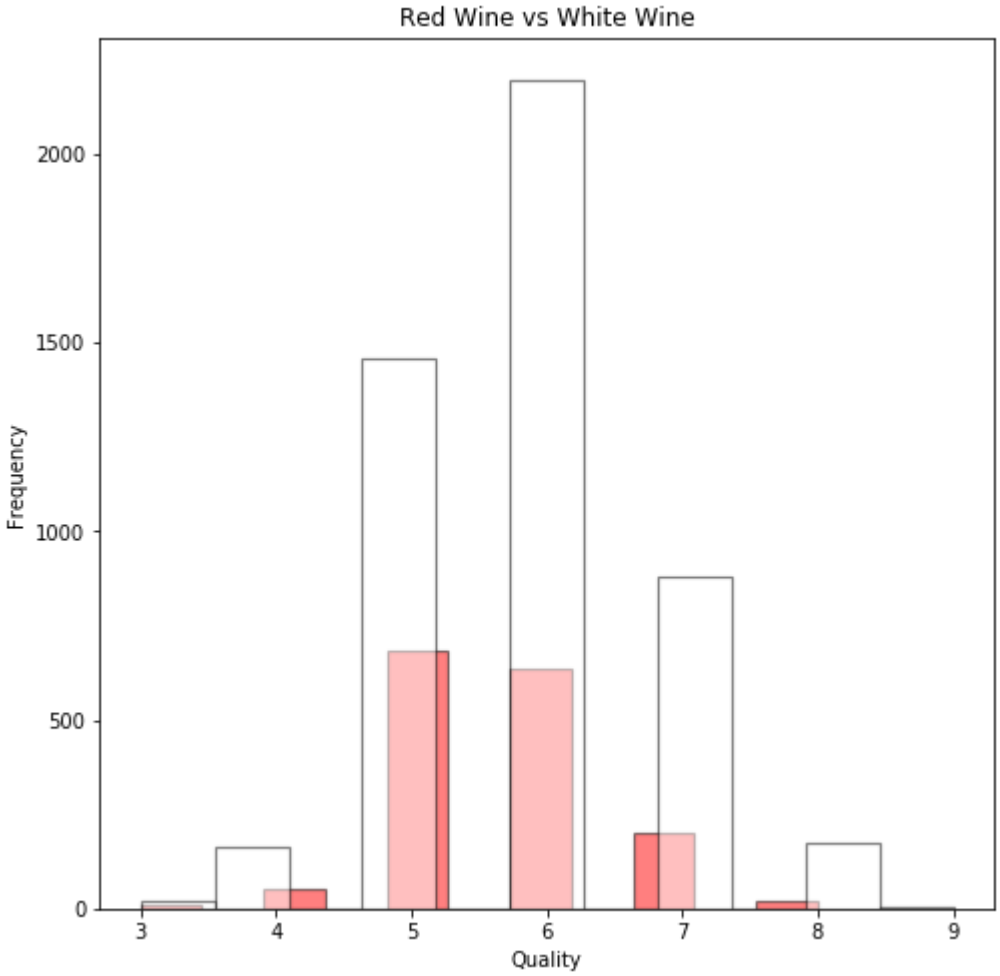
In [28]:

```
df_red_wine_quality["quality"].plot.hist(bins = 11,title="Red Wine vs White Wine",color =
"red",
                            figsize = (8,8),edgecolor='black', linewidth=1.2,alpha=.5).se
t_xlabel("Quality")

df_white_wine_quality["quality"].plot.hist(bins = 11,color = "white",
                            figsize = (8,8),edgecolor='black', linewidth=1.2,alpha=.5)
```

Out[28]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7ecd5f58d0>
```



We can visualizations the number of instances in both red wine and white wine datasets.
Both datasets shows that the average lies in 5 and 6 wine quality score which was described in the statistics section.
The white wine has more instances than the red wine.
We need more (instances) on the red wine data.
We should at least have similar instances.

# Densities

Another way to visualize the distribution of a quantitative variable is by plotting its density.
A density plot turns the jagged histogram into a smooth curve, allowing the user to focus on the general shape of the distribution.

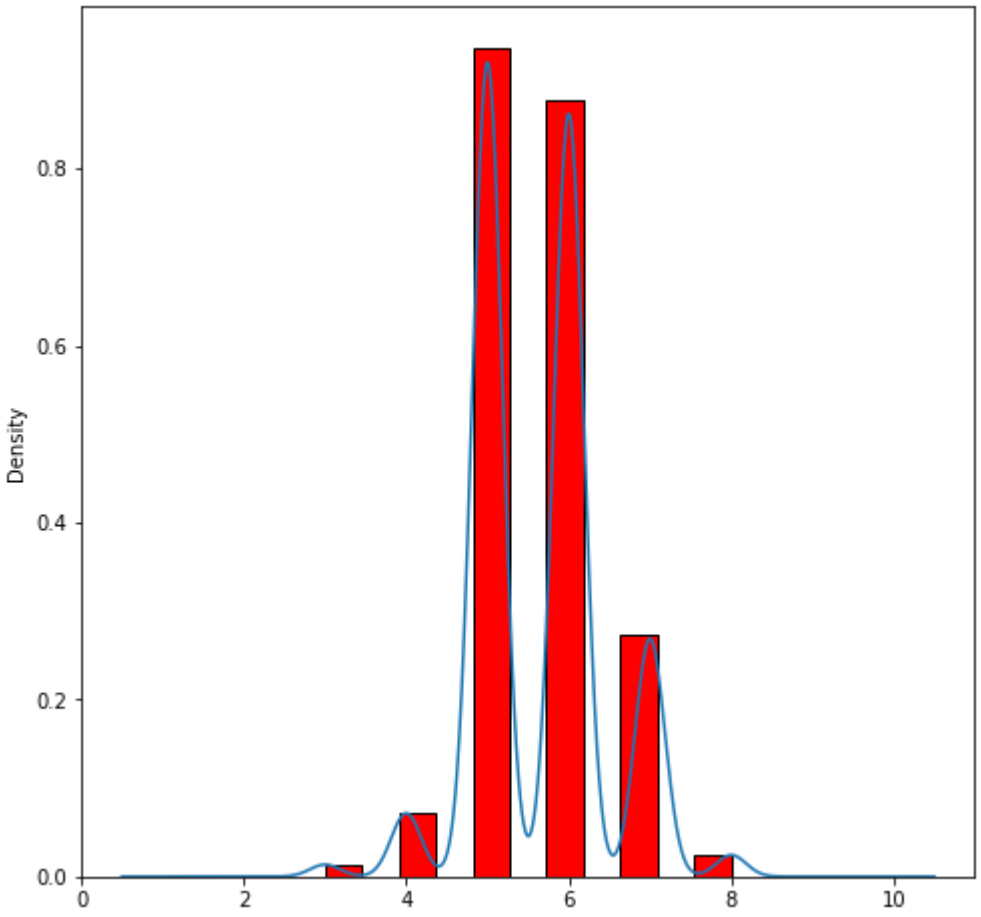
## Red Wine Quality Density

In [29]:

```
df_red_wine_quality["quality"].plot.hist(bins=11, density=True,color = "red", figsize = (8
,8),edgecolor='black')
df_red_wine_quality["quality"].plot.density()
```

Out[29]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7ecaa7e898>
```



We can tell that there's more normal quality red wine (5 and 6) than lower and excellent quality red wines.
It looks like the data is a normal distribution.
It has no skewed.

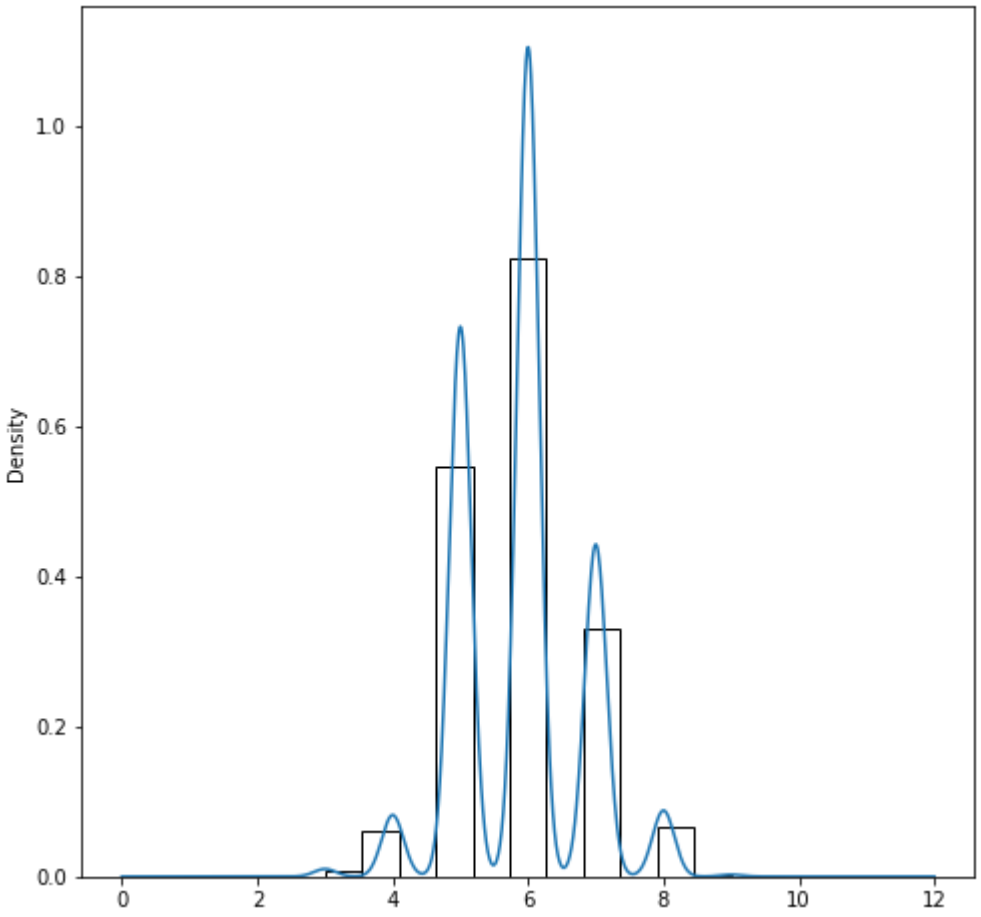
# White Wine Quality Density

In [30]:

```
df_white_wine_quality["quality"].plot.hist(bins=11, density=True,color = "white", figsize
= (8,8),edgecolor='black')
df_white_wine_quality["quality"].plot.density()
```

Out[30]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7ec1385cf8>
```




We can tell that there's more normal quality white wine (5 and 6) than lower and excellent quality white wines.
Just like the red wine.
It looks like the data is a normal distribution.
It has no skewed.

# Boxplot

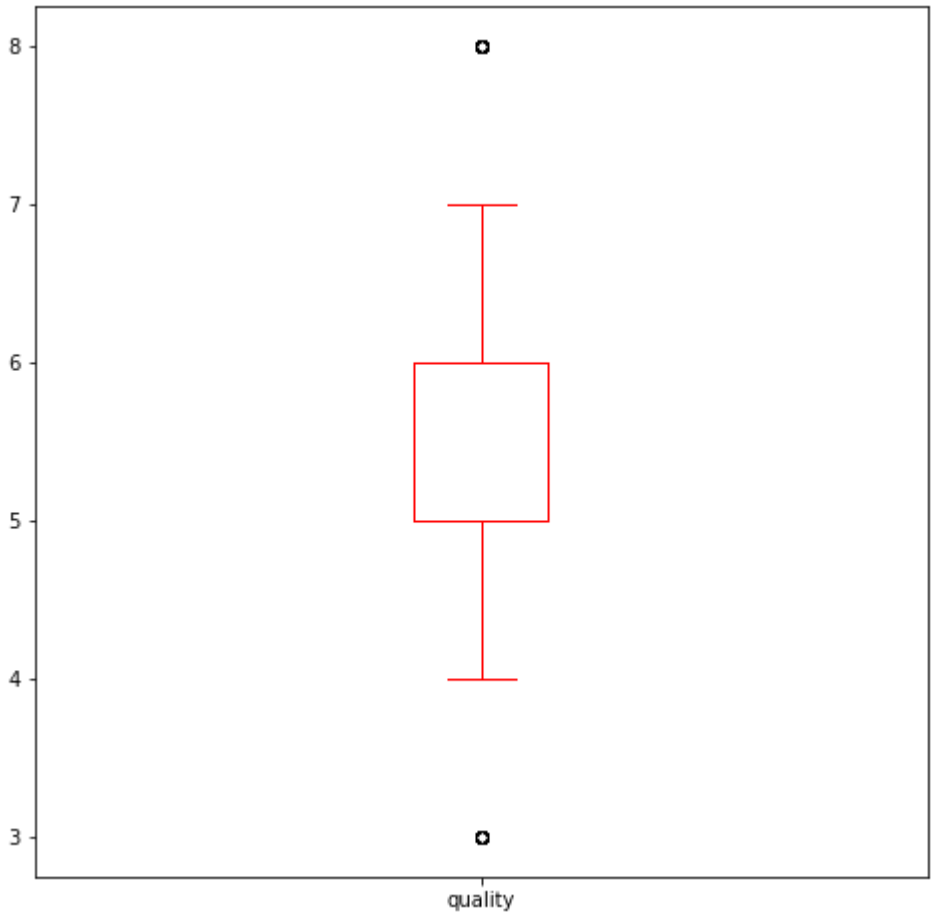
A box plot is another way to visualize the distribution of a quantitative variable.
The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.
A segment inside the box shows the median and "whiskers" above and below the box show the locations of the minimum and maximum.

In [31]:

```
df_red_wine_quality["quality"].plot.box(color = "red",figsize=(8,8))
```

Out[31]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7ec13220f0>
```



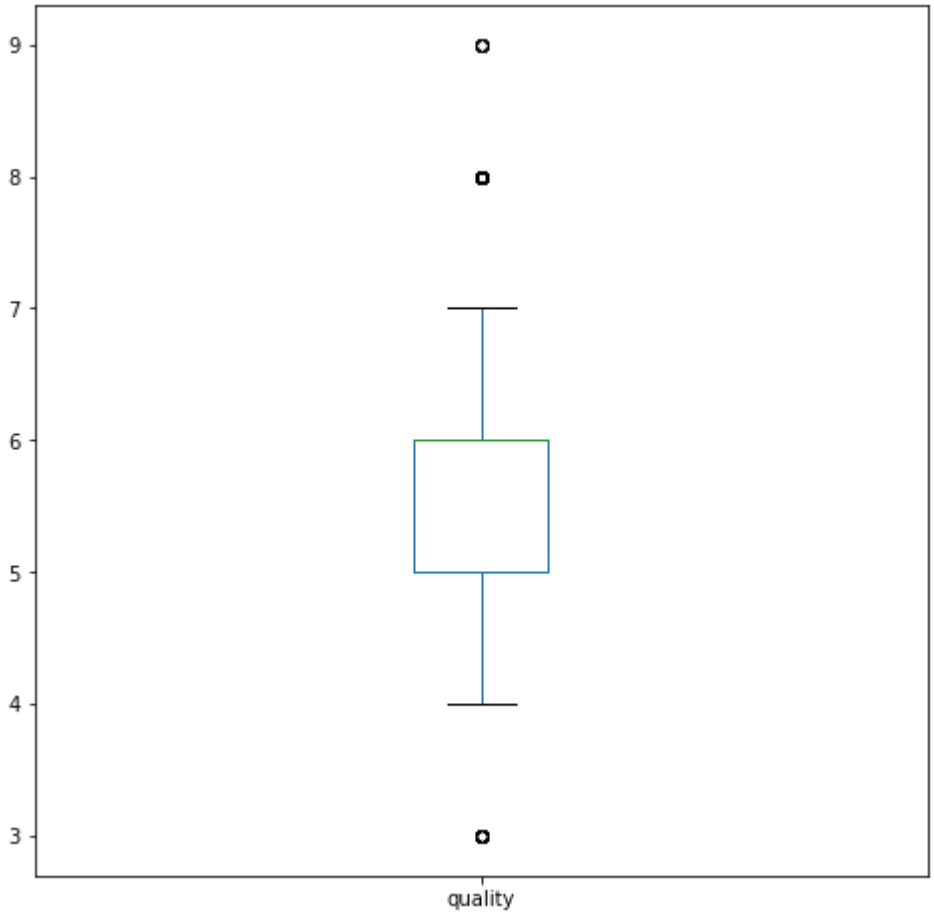

We can see the low and excellent red wine quality scores are locating in the data.
We can also see that there are only one low quality score which is great.
However, we can see that there's one excellent quality score.

In [32]:

```
df_white_wine_quality["quality"].plot.box(figsize=(8,8))
```

Out[32]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7ec1283a20>
```




We can see the low and excellent white wine quality scores are locating in the data.
We can also see that there are only one low quality score which is great.
However, we can see that there's only two excellent quality scores.

# Conclusion

The wine quality dataset goal is to model the wine quality based on physicochemical tests for Vinho Verde wine products quality business.
Both red and wine datasets shows the average normal quality scores.
In knowing this, Vinho Verde company can modify their physicochemical tests to achieve an average of excellent quality wines to produce.
Due to privacy and logistic issues , there is no data about grape types, wine brand, wine selling price, and etc.
We don't know which grape types and wine brands are in the instances.
It's harder to determine whether a wine brand quality is better than other wine brands.
Maybe the different type of grapes can affect the wine quality scores based on their physicochemical properties.
As a result, we need more data attributes such as the grape types , wine brand, wine selling price, and etc.
We need more number of instances as well.