## Introducción

In our dataset “Wine Data Set”  the information comes from a chemical study applied to three different varieties of wine classes, these three types of classes come from the same region in Italy, in our dataframe we have a total of fourteen variables of which 13 are numeric and one is categorical ( the wine class), our dataframe does not contain any null value and has a total of one hundred seventy-eight rows, of which are divided into 59 of class 1, 71 of class 2 and finally we have 48 of class 3.

Of our 13 remaining variables we have the second column that is Alcohol, this refers to the grades that the wines contain, the next column contains the malic acid which helps determine the acidity of a wine, this variable is measured in g / l, the next variable is the Ash that is the content of ceizas that our wine has is measured in g / l, then we have Alcalinity of ash which is the sum of the ammonium cations that are mixed in the organic acids of the wine, The magnesium column is measured in mg / l, and so on as you will see in the images, our proline column is the one with the highest numerical value.

## Univariate Analysis

**Variables:**

**1)Class**

**2) Alcohol**

**3) Malic acid**

**4) Ash**

**5) Alcalinity of ash**

**6) Magnesium**

**7) Total phenols**

**8) Flavanoids**

**9) Nonflavanoid phenols**

**10) Proanthocyanins**

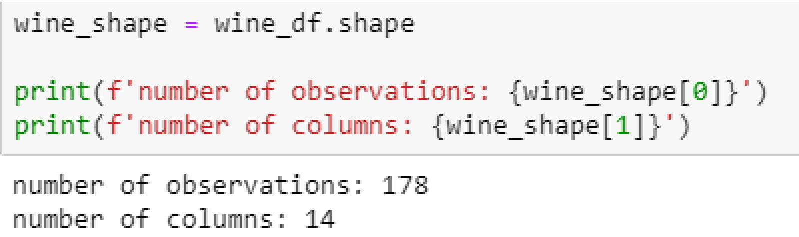
**11)Color intensity**

**12)Hue**

**13)OD280/OD315 of diluted wines**

**14)Proline**

To start with our univariate analysis we first obtained the size or shape of our dataframe, as previously stated we have 14 columns and 178 observations or rows.

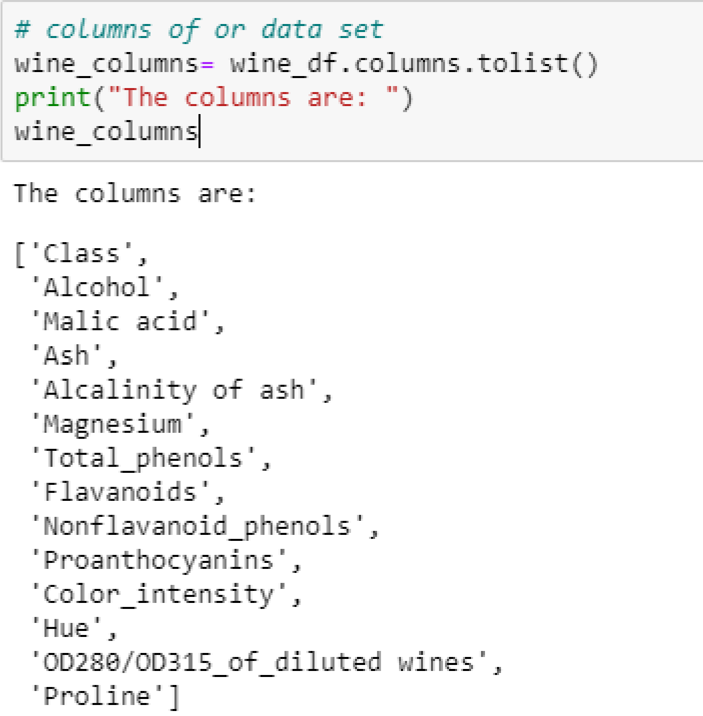


```
wine_shape = wine_df.shape

print(f'number of observations: {wine_shape[0]}')
print(f'number of columns: {wine_shape[1]}')
```

```
number of observations: 178
number of columns: 14
```

we obtained our total of columns in the form of a list, so in this way, we confirm that they are well and if we need to use them later as strings we already have them available.



```
# columns of or data set
wine_columns= wine_df.columns.tolist()
print("The columns are: ")
wine_columns
```

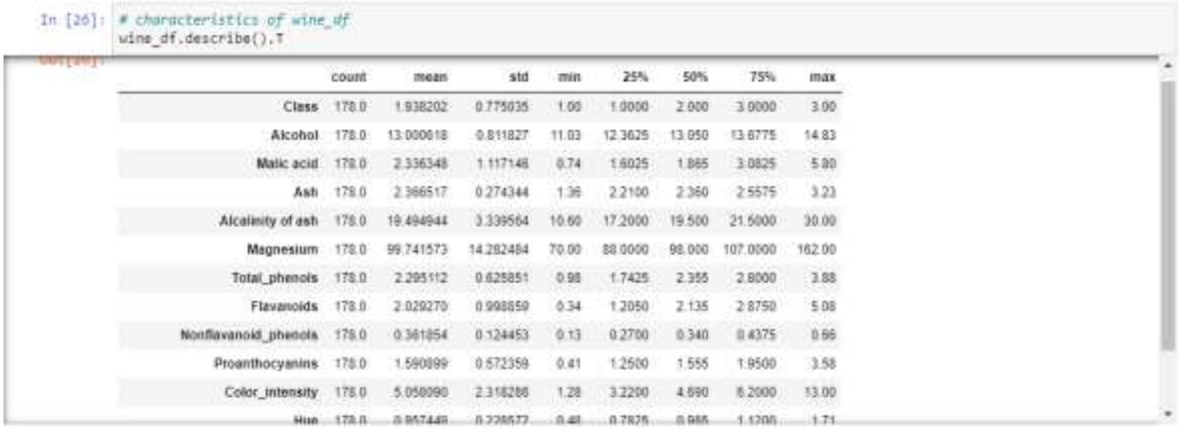
```
The columns are: 

['Class',
 'Alcohol',
 'Malic acid',
 'Ash',
 'Alcalinity of ash',
 'Magnesium',
 'Total_phenols',
 'Flavanoids',
 'Nonflavanoid_phenols',
 'Proanthocyanins',
 'Color_intensity',
 'Hue',
 'OD280/OD315_of_diluted wines',
 'Proline']
```

we obtain the important statistics of the dataframe, but in this case we use the transpose to make it easier for us to visualize the statistics, although this point can be subjective.

With the help of this we can realize that variables are the ones with the highest mean and standard deviation, we can also see the maximum or the greatest value of your row and the minimum value that our row has.

As you can see we also have three sections that say 25%, 50% and 75% this represents our quartiles of the data

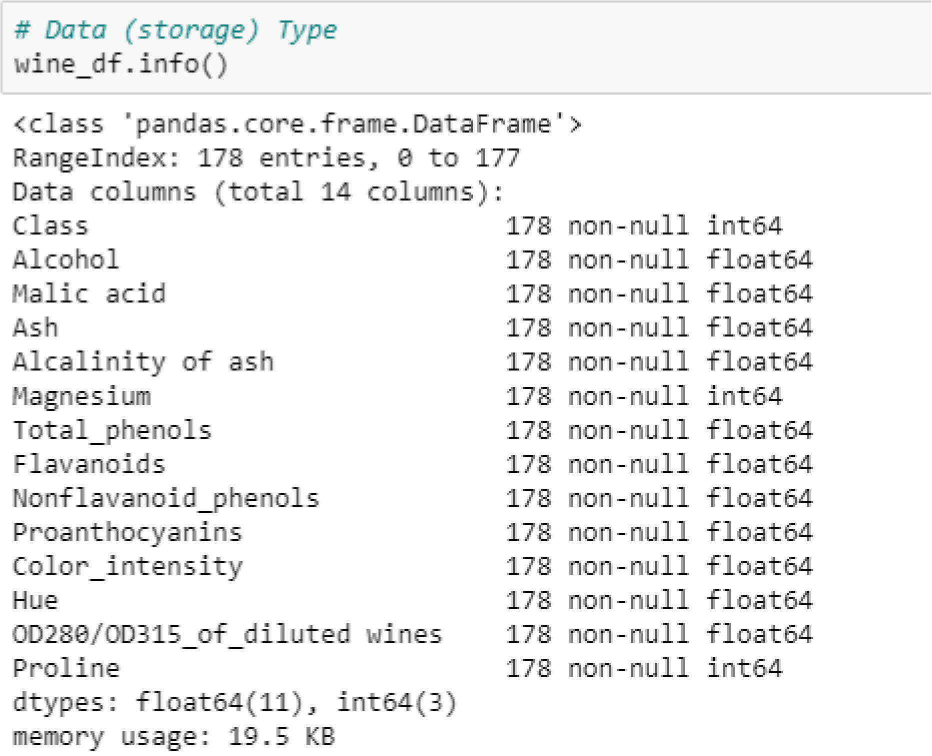


```
In [26]: # characteristics of wine_df
         wine_df.describe().T
Out[26]:
```

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Class | 178.0 | 1.938202 | 0.775035 | 1.00 | 1.0000 | 2.000 | 3.0000 | 3.00 |
| Alcohol | 178.0 | 13.000618 | 0.811827 | 11.03 | 12.3625 | 13.050 | 13.6775 | 14.83 |
| Malic acid | 178.0 | 2.336348 | 1.117146 | 0.74 | 1.6025 | 1.865 | 3.0825 | 5.80 |
| Ash | 178.0 | 2.366517 | 0.274344 | 1.36 | 2.2100 | 2.360 | 2.5575 | 3.23 |
| Alcalinity of ash | 178.0 | 19.494944 | 3.339564 | 10.60 | 17.2000 | 19.500 | 21.5000 | 30.00 |
| Magnesium | 178.0 | 99.741573 | 14.282484 | 70.00 | 88.0000 | 98.000 | 107.0000 | 162.00 |
| Total_phenols | 178.0 | 2.295112 | 0.625851 | 0.98 | 1.7425 | 2.355 | 2.8000 | 3.88 |
| Flavanoids | 178.0 | 2.029270 | 0.998859 | 0.34 | 1.2050 | 2.135 | 2.8750 | 5.08 |
| Nonflavanoid_phenols | 178.0 | 0.361854 | 0.124453 | 0.13 | 0.2700 | 0.340 | 0.4375 | 0.66 |
| Proanthocyanins | 178.0 | 1.590899 | 0.572359 | 0.41 | 1.2500 | 1.555 | 1.9500 | 3.58 |
| Color_intensity | 178.0 | 5.058090 | 2.318286 | 1.28 | 3.2200 | 4.690 | 6.2000 | 13.00 |
| Hue | 178.0 | 0.957449 | 0.228572 | 0.48 | 0.7825 | 0.965 | 1.1200 | 1.71 |

we use the function info () in our dataset this in order to have information that (describes) can not provide us, for example what type of variables we have, in our case our variables are integer (int) and floating (float).

We can also observe the amount of memory it is occupying on our computer.

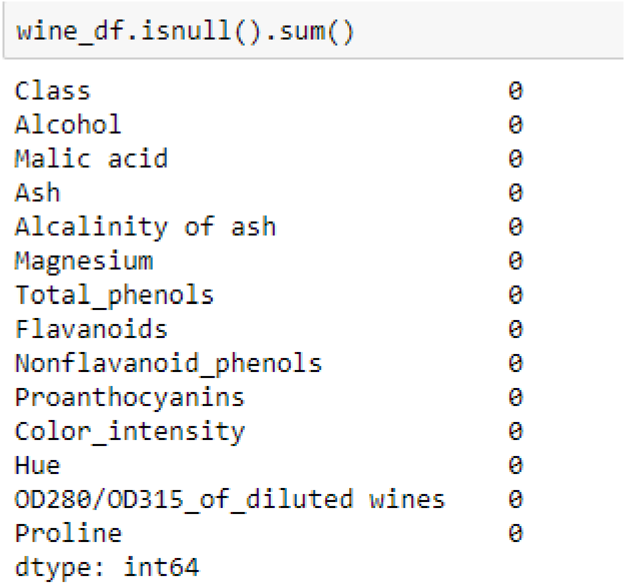


```
# Data (storage) Type
wine_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
Class                           178 non-null int64
Alcohol                         178 non-null float64
Malic acid                      178 non-null float64
Ash                             178 non-null float64
Alcalinity of ash               178 non-null float64
Magnesium                       178 non-null int64
Total_phenols                   178 non-null float64
Flavanoids                      178 non-null float64
Nonflavanoid_phenols            178 non-null float64
Proanthocyanins                 178 non-null float64
Color_intensity                 178 non-null float64
Hue                             178 non-null float64
OD280/OD315_of_diluted wines    178 non-null float64
Proline                         178 non-null int64
dtypes: float64(11), int64(3)
memory usage: 19.5 KB
```

Now we proceed to find the null values of the dataframe, but in this case we do not have any, that is good since we will not have to make any modifications.

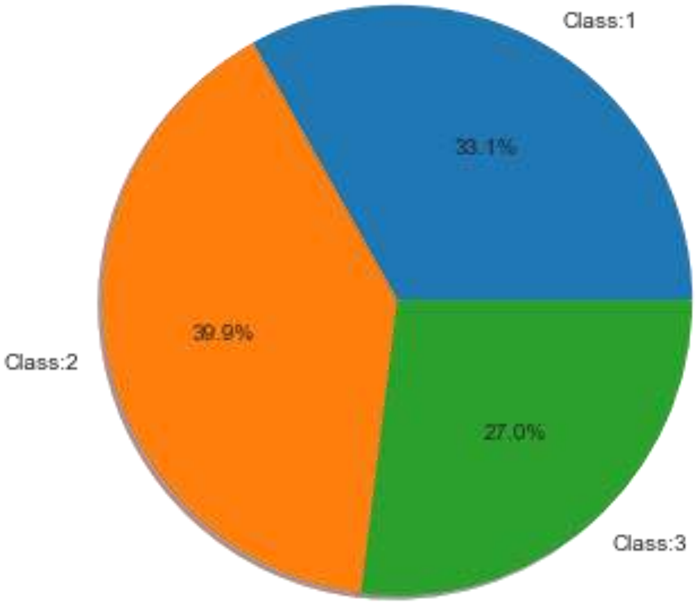
This point is important because if we had null values when making predictions or graphs we can have them badly.



```
wine_df.isnull().sum()
```

```
Class                           0
Alcohol                         0
Malic acid                      0
Ash                             0
Alcalinity of ash               0
Magnesium                       0
Total_phenols                   0
Flavanoids                      0
Nonflavanoid_phenols            0
Proanthocyanins                 0
Color_intensity                 0
Hue                             0
OD280/OD315_of_diluted wines    0
Proline                         0
dtype: int64
```

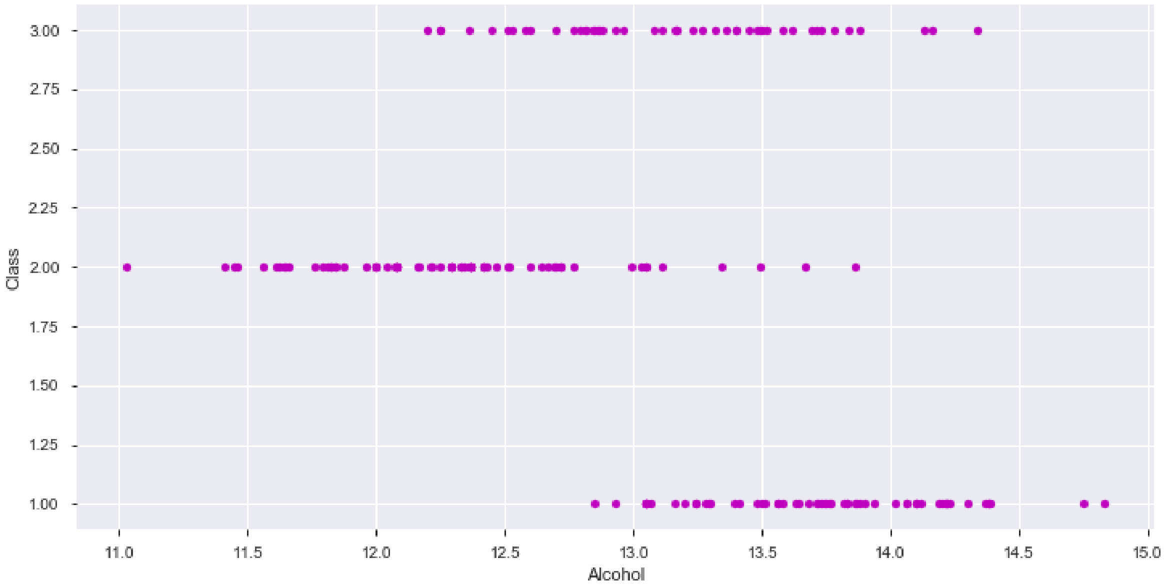
In the following pie chart we can see what percentage of our data we have, of the wine class 1 constitutes 33.1%, class 2 constitutes 39.9% and finally class 3 constitutes 27.0% of our total data.



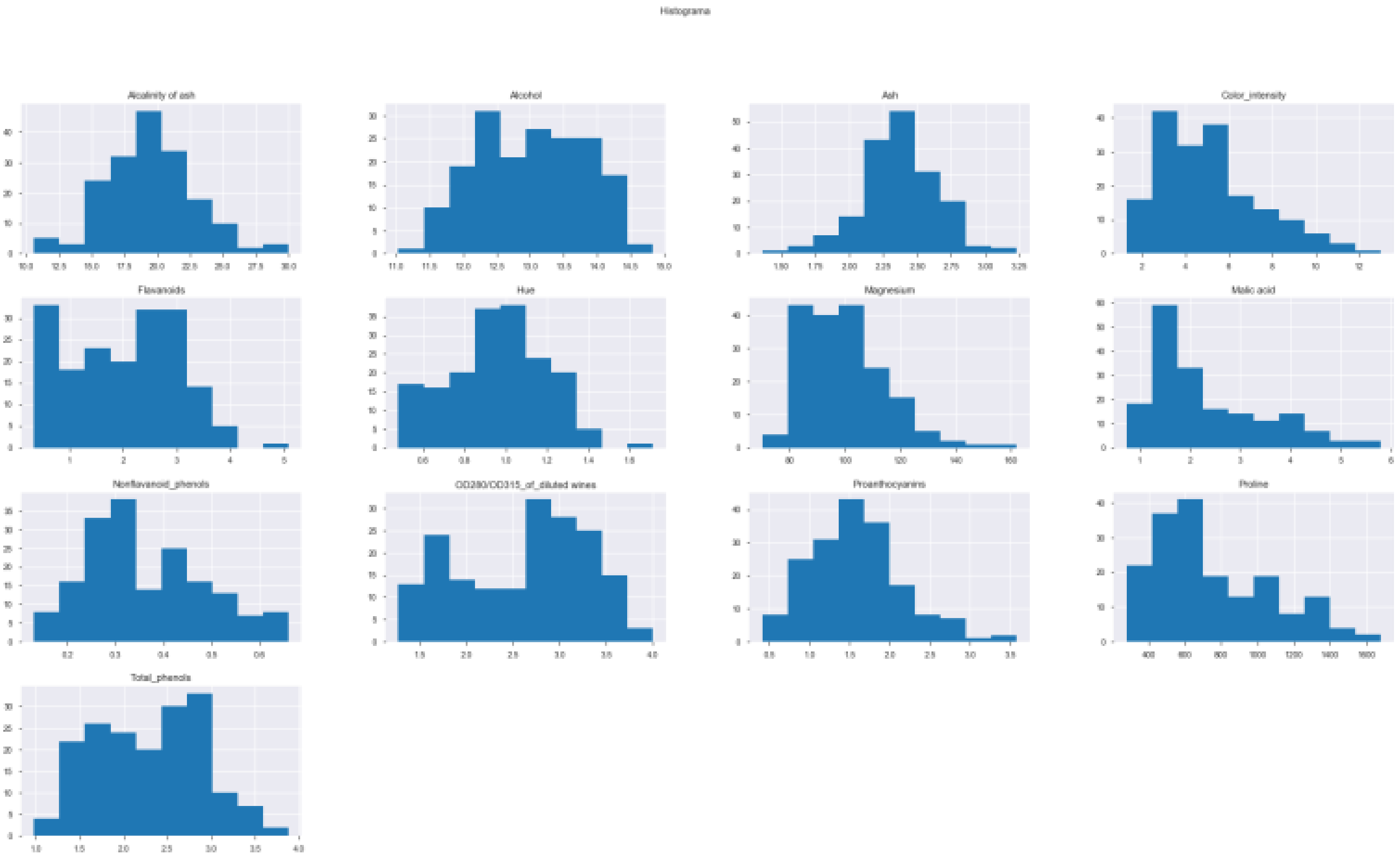
In the following graph we can quickly visualize how most of the grade of alchol is from class 1 from 13 to 14 degrees and it has atypical values that are greater than 14.5 degrees of alchol.

from class 2 we can infer the tendency of the degree of alcohol ranges from 11.5 to 12.5 having atypical values at 11 degrees and greater than 13.5 degrees of alcohol.

from class 3 we can infer that the tendency of alchol degrees is from 12.5 to approximately 13.7, having atypical values greater than 14.0 degrees of alchol.



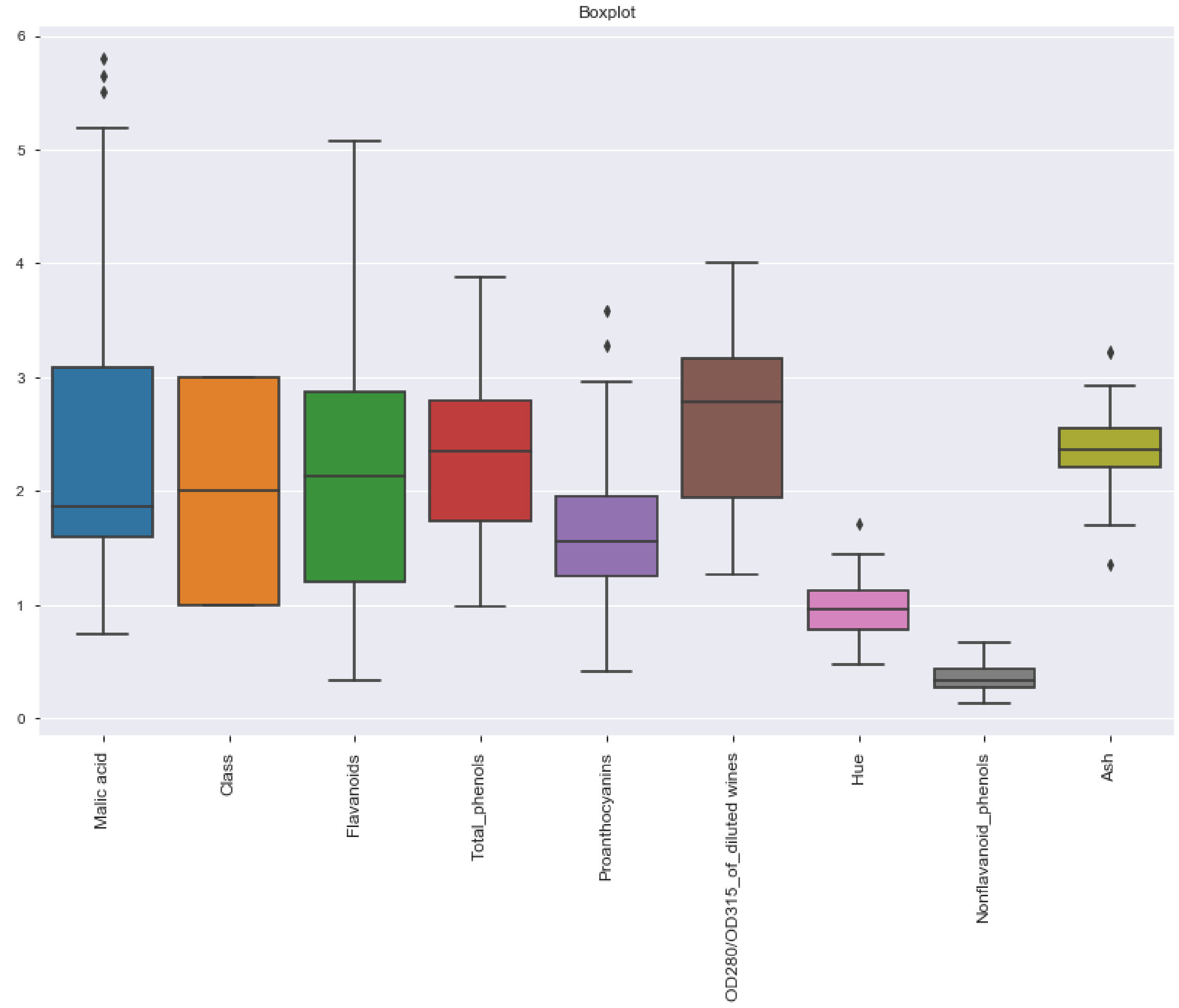
With the help of the following graphs we can see the behavior, the distribution that our variables have, we can realize that the variables that tend to have a normal distribution are the variable Ash, Alcalinity of ash and Alcohol.



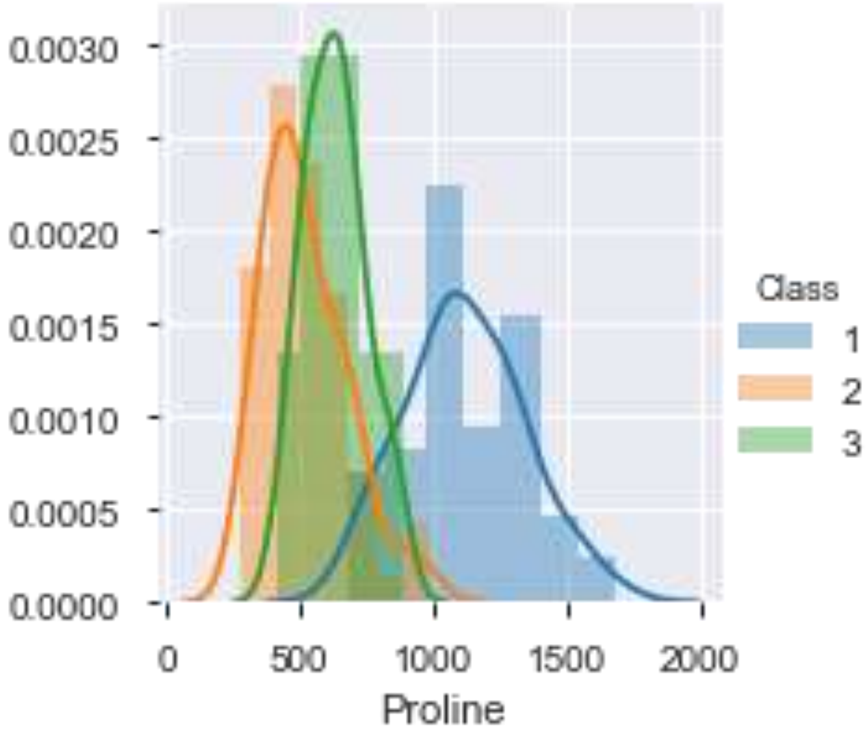
With the support of the graphic boxplots we can see what are the atypical values, the median, the quartiles, the range, etc. For example, if we read the boxplot of the variable "malic acid" we can see that its median is less than 2 degrees, quartile 1 is less than approximately 2 ° 1.7, quartile 3 is slightly greater than 3 °.

We can see that atypical values are greater than 5 degrees of alcohol.

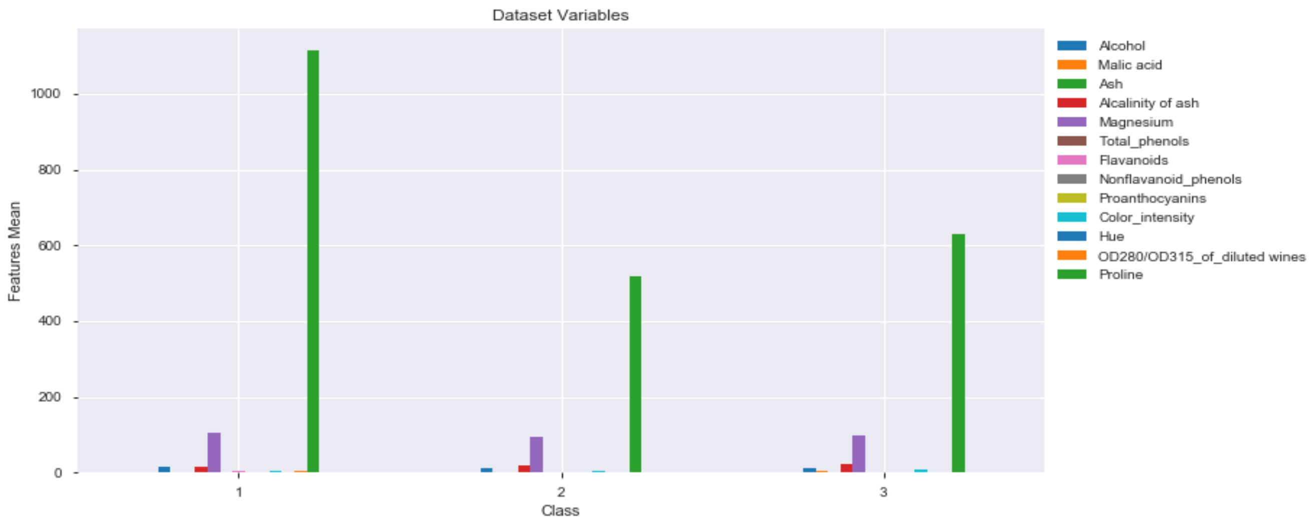
In this way we can analyze each one of our boxplost.



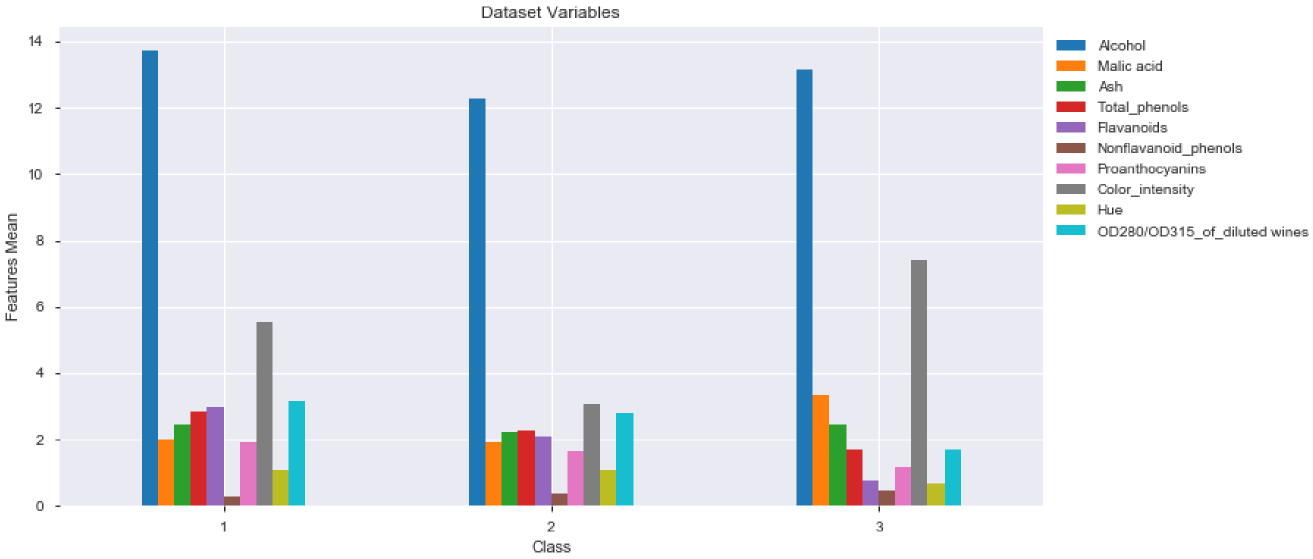
We use the Seaborn distributions, of all the graphs obtained the ones that best serve us is that of "Proline", we can realize that as long as the proline is greater than 1000 it is the kind of wine 1.



In the following graph we can see the two variables: proline and magnesium that has a higher

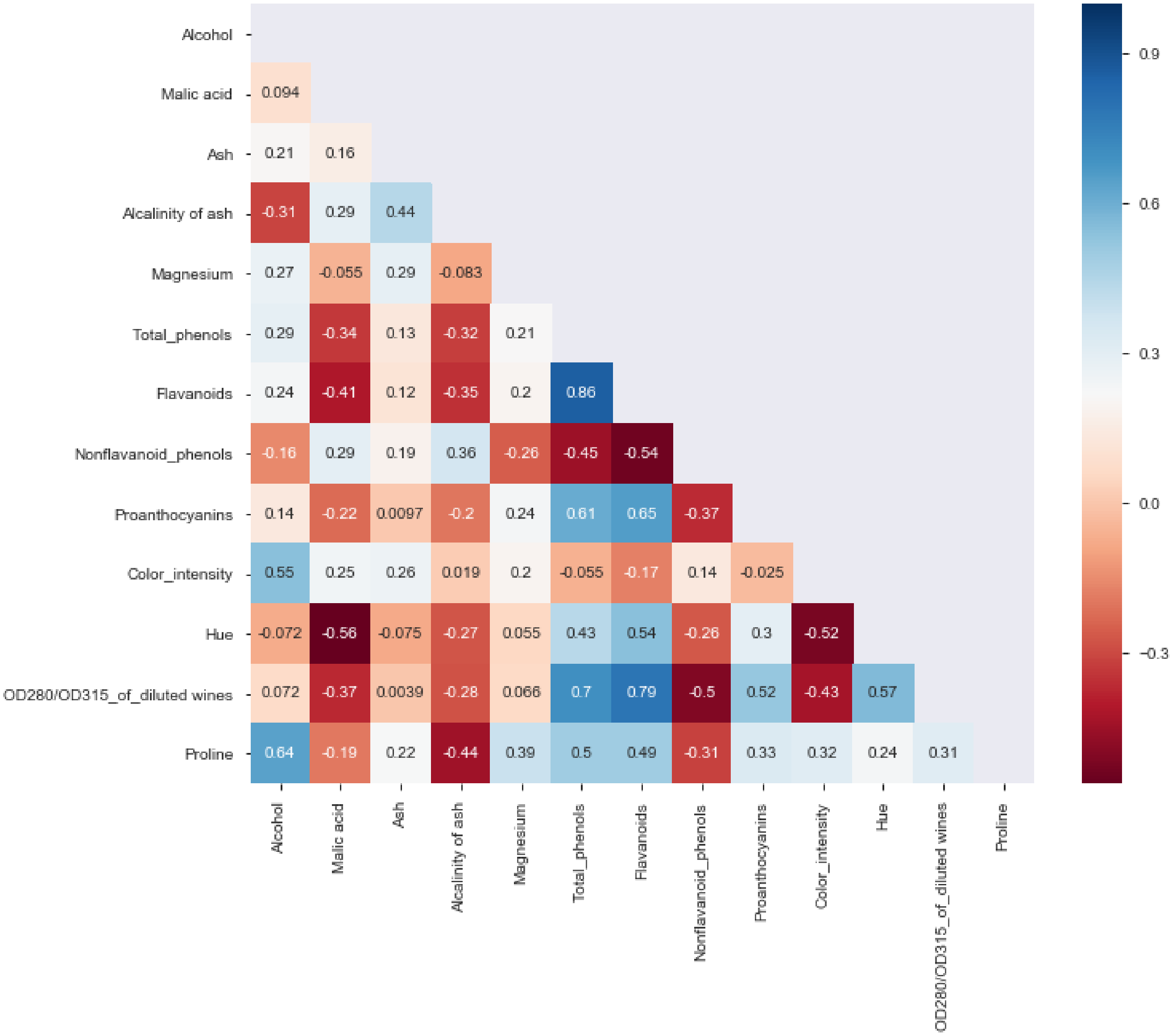


Contrary to the previous graph in this graph we can see the three variables that have the lowest average, this time it is the variables Nonflavanoid_phenols, Hue and Flavanoids.

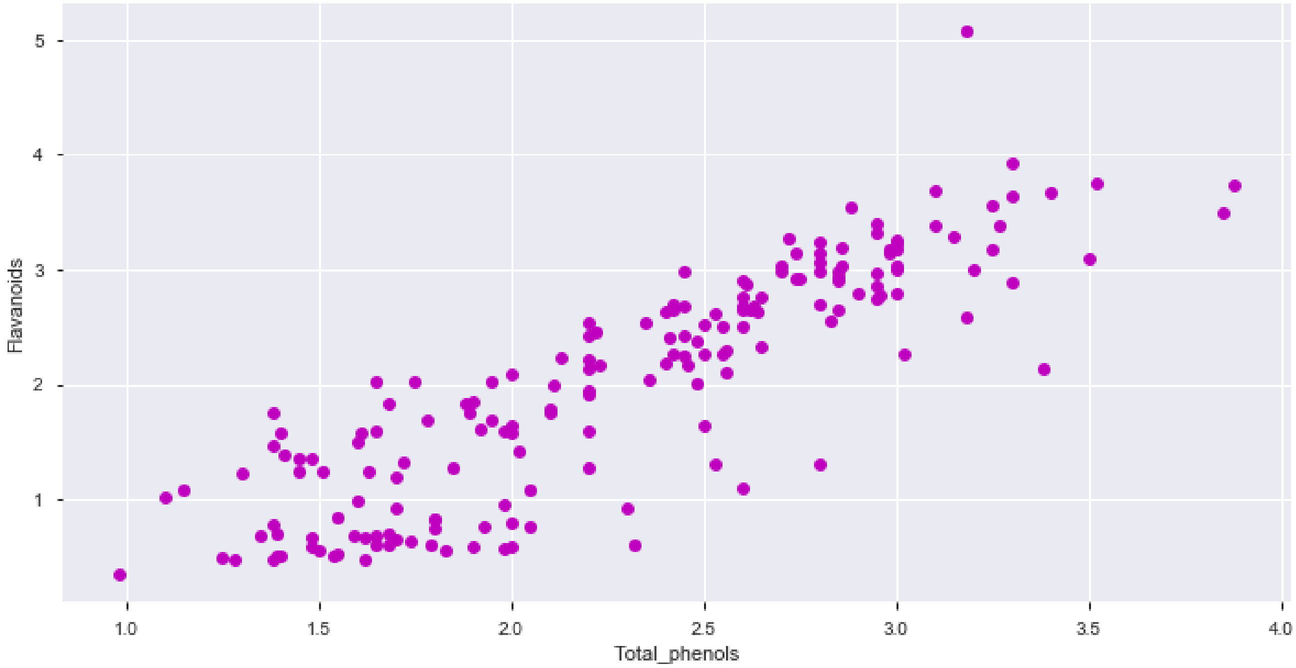


## Multivariate Analisys

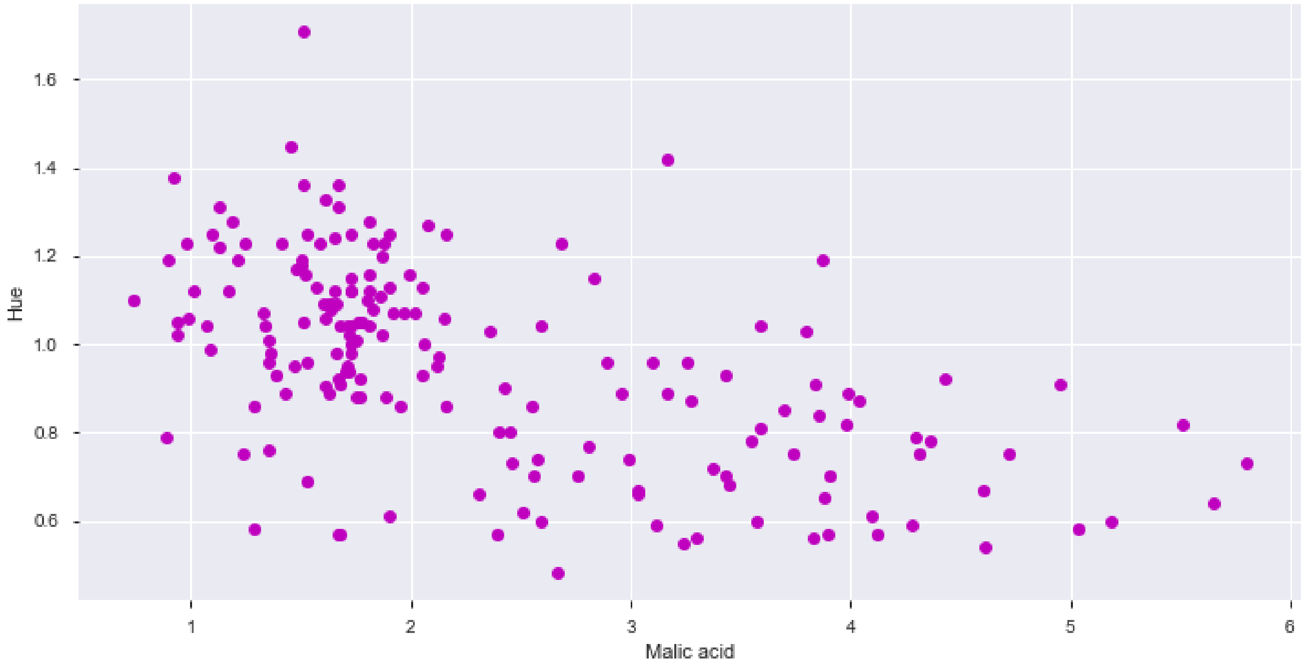
With our correlation matrix we can see which are the variables that have a positive correlation and which have a negative correlation.



In our case, the variables that have a higher positive correlation are Total_phenols and Flavanoids, we can make a graph for corroborate it.



variables with greater negative correlation:



## Conclusión

In conclusion, we can be sure that the positive correlations are true, because they are very reactive substances, composed of more than one phenol group (benzene ring with a hydroxide group -OH-) per molecule, found in parts of the vine and its fruits

These substances are complex and consist of hundreds of different molecules, which pass into the wine during processing, being determining to establish their quality.

These substances contain antioxidant, anticarcinogenic and anti-inflammatory properties. Then as the phenols increase the flavonoids too and vice versa they will do which makes sense.

The negative correlation is also correct, because the acidic acid is greater when the grapes are not yet ripe, instead the shade of wine is greater when the grapes are already ripe.

According to Vinetur (THURSDAY, DECEMBER 29, 2016) Malic acid is one of nature's most abundant acids, present in many vegetables and fruits, especially when they are green.

The hue indicates the degree of evolution of the wine, that is, its old age. The young red wine almost always maintains a vivid tone, between purples and rubies, but when it ages, the red tones accentuate and turn towards the brick, tile and brown, until crowning in the ocher and amber.