The task is to Predict the quality of Red Wine.

# DATA PREPARATION AND UNDERSTANDING

One of the first steps i engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated.

## 

## UNDERSTANDING THE DATA

* I loaded the data given into the environment and tried to understand the basic appearance of the data. Our dataset has 1599 rows × 12 features including the target variable.
* The data had 11 float type variables and 1 integer type variables.
* Data was summarized to find basic statistical metrics.

**MISSING VALUES:**

This dataset had no missing values.

fixed acidity 0

volatile acidity 0

citric acid 0

residual sugar 0

chlorides 0

free sulfur dioxide 0

total sulfur dioxide 0

density 0

pH 0

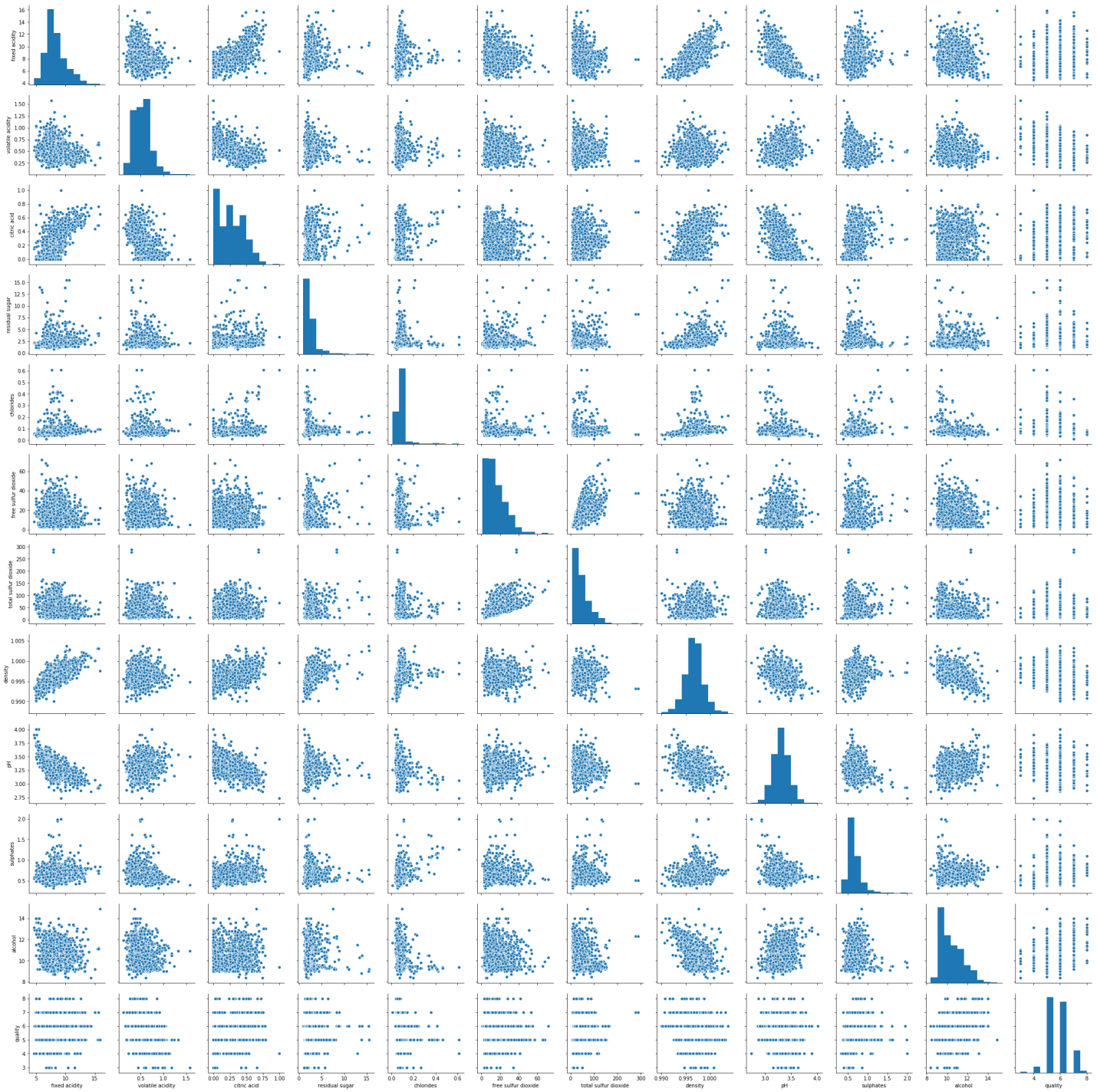
sulphates 0

alcohol 0

quality 0

good 0

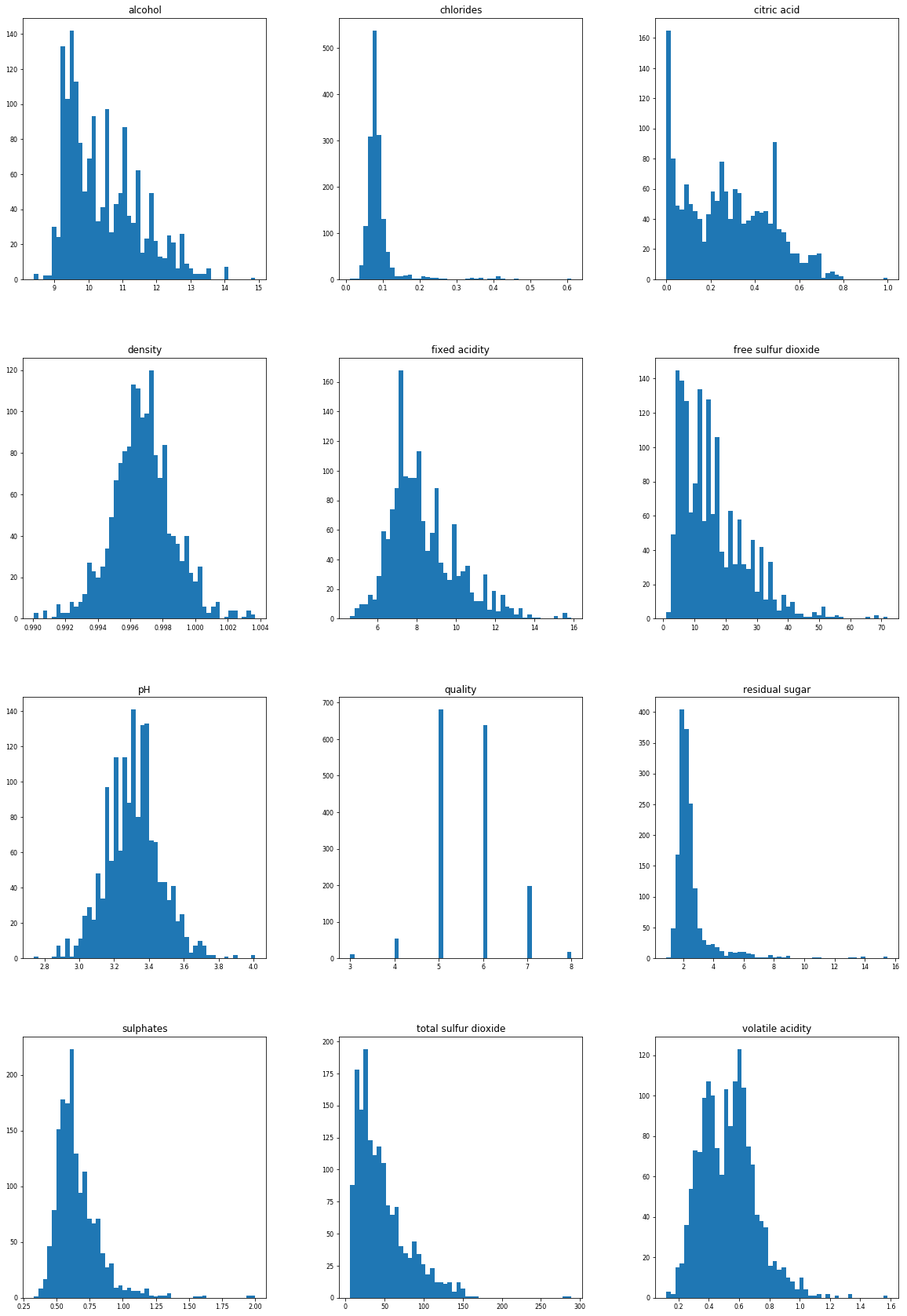
**PAIRPLOT TO SEE THE DISTRIBUTION:**

****

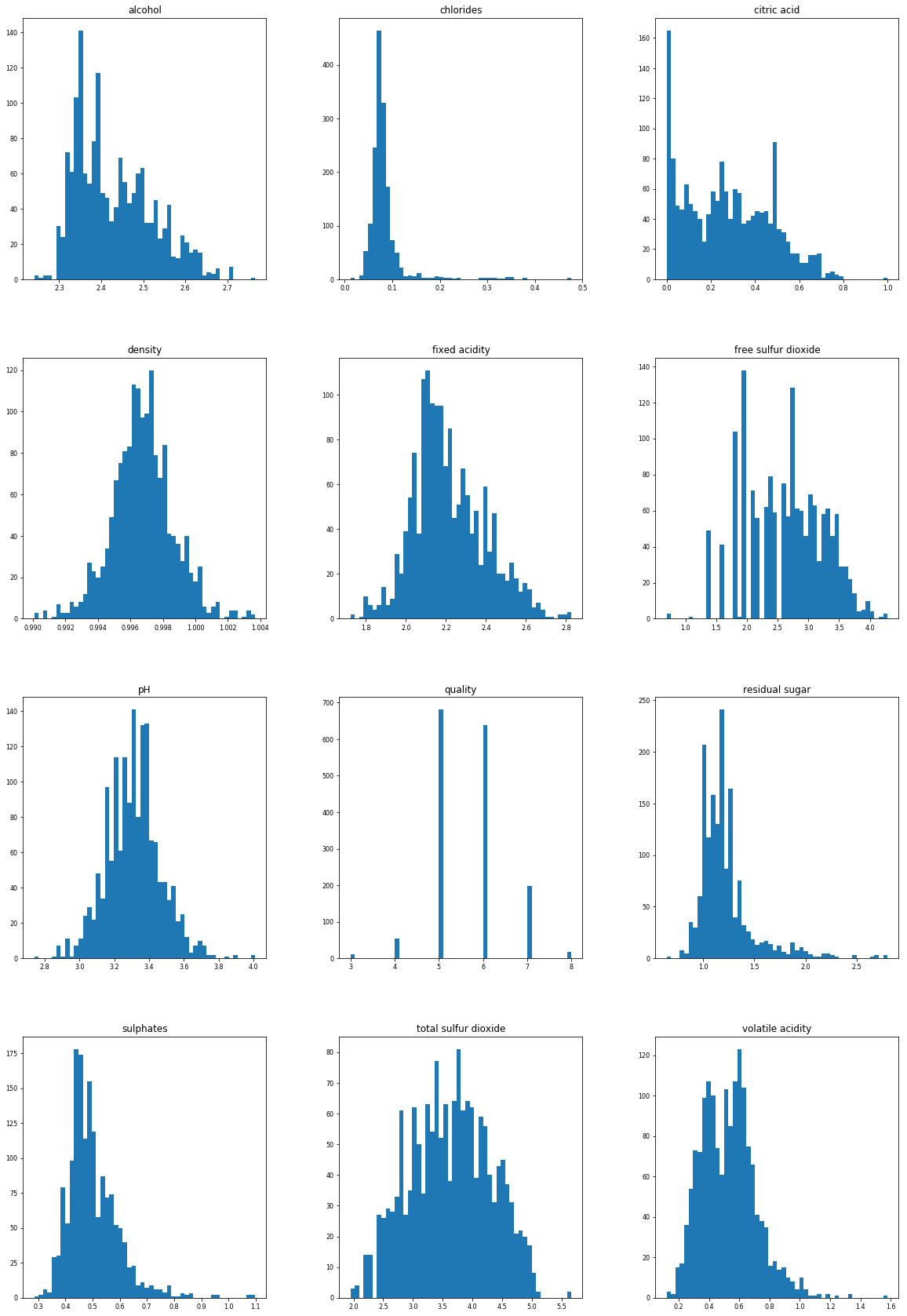
**SKEWNESS:**

Data skewness affects the symmetry of the distribution. Any distribution with skewness value between -0.5 and 0.5 is considered fairly normal. Distributions with values beyond them are considered skewed. I measured the skewness of each variables. For the variables with very high skewness values, we applied the log transformation to bring them to a fairly normal distribution.

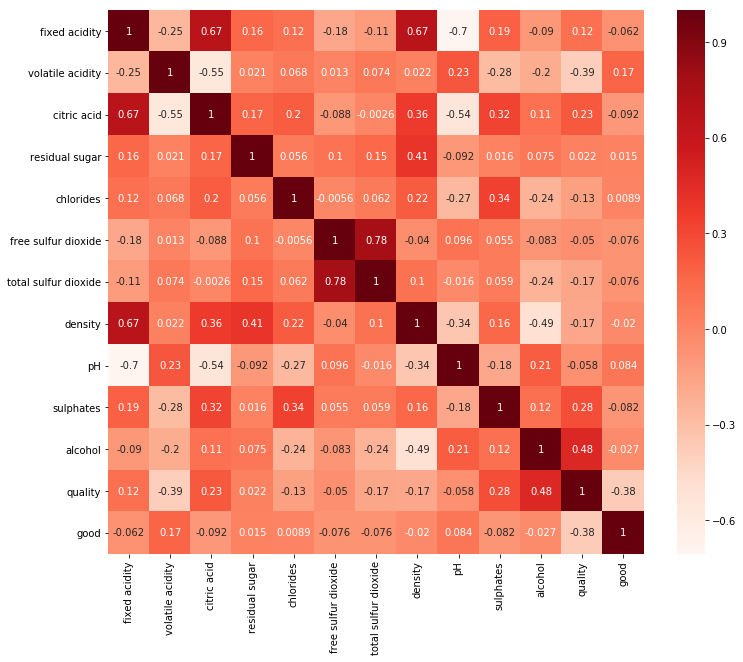
BEFORE SKEWNESS:



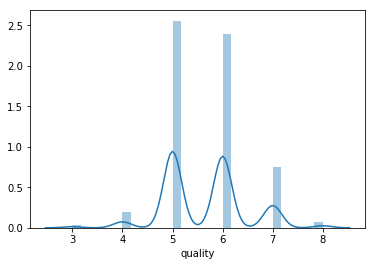
AFTER SKEWNESS:



CORRELATION:



As it is a multiclass problem we will use classification technique because the target variable is in levels so it is better to use classification technique. The target variable was QUALITY which has 6 levels and those which are 5 and 6 I put them in 1 category and the rest in 0 category.



def isgood(quality):

if quality == 5 & 6:

return 1

else:

return 0

redwine['good'] = redwine['quality'].apply(isgood)

So I used Random Forest, Decision Tree and Logistic Regression which gave me the accuracy of 96%,95% and 97% respectively through which I came to know that Logistic Regression is the best method to be used in this model as it gives a accuracy of 97%.