**WINE QUALITY DATA ANALYSIS REPORT**

**INTRODUCTION:**

Winemaking is an ancient craft that has been perfected over centuries, and today it is a thriving industry with a global market. Wine quality is a complex and subjective concept, influenced by a wide range of factors such as grape variety, region, soil, climate, and winemaking techniques. Understanding these factors and their impact on wine quality is crucial for winemakers, sommeliers, and wine enthusiasts alike. The data set analyzed in this report consists of the following attributes :

* fixed acidity
* volatile acidity
* citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur
* density
* pH
* quality

**PROBLEM STATEMENT:**

The objective of this data analysis report is to explore the factors that contribute to wine quality and to identify the key drivers of high-quality wines. We will analyze a dataset of wine properties, such as acidity, pH, alcohol, and residual sugar, etc to understand what makes a great wine. Specifically, we aim to address the following questions:

* What are the most important wine properties that contribute to wine quality?
* Is multiple linear regression a good model to predict the quality of wine?

By answering these questions, this report aims to provide valuable insights into the complex world of wine quality and to help winemakers, sommeliers, and wine enthusiasts make more informed decisions.

**METHODOLOGY :**

1. Raw data collection : [UCI Machine Learning Repository: Wine Quality Data Set](https://archive.ics.uci.edu/ml/datasets/wine+quality)
2. Data preprocessing :

* White wine : the data is already clean, there are no missing values. There are some attributes that showed little correlation with wine quality but removing them led to decrease in R-squared value so all the columns are kept.
* Red wine : The data is already clean as well and I did not drop any columns.

1. Performed detailed EDA (Exploratory Data Analysis) : I created graphs and tables by mainly using matplotlib and seaborn library in Python.
2. Modeling and algorithms :

* Independent and dependent variables were defined
* Train and test split of data was made for algorithm deployment
* Multiple linear regression is used because we have multiple variables that are relevant to predict the quality of wine (dependent variable).
* Ordinary least squared is used for estimating the parameters of multiple linear regression models.
* Analyzing the results.

1. Conclusions

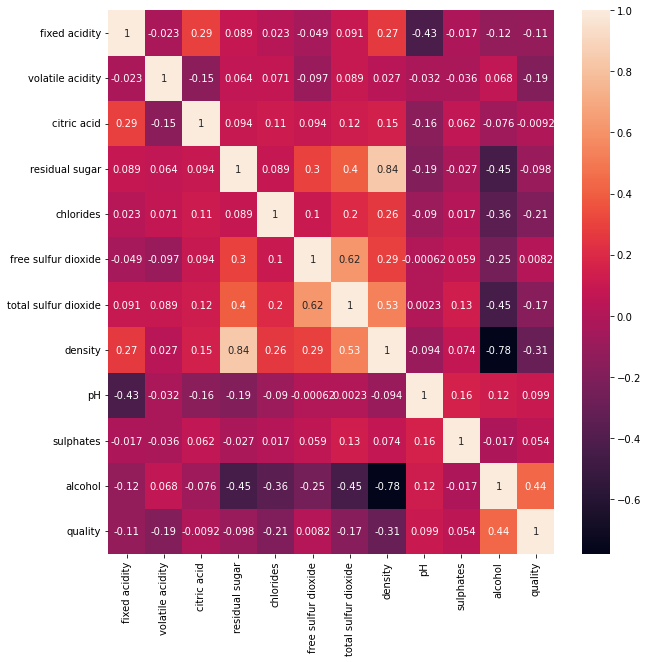
**EXPLORATORY DATA ANALYSIS :**

***For white wine :***

Graphical user interface

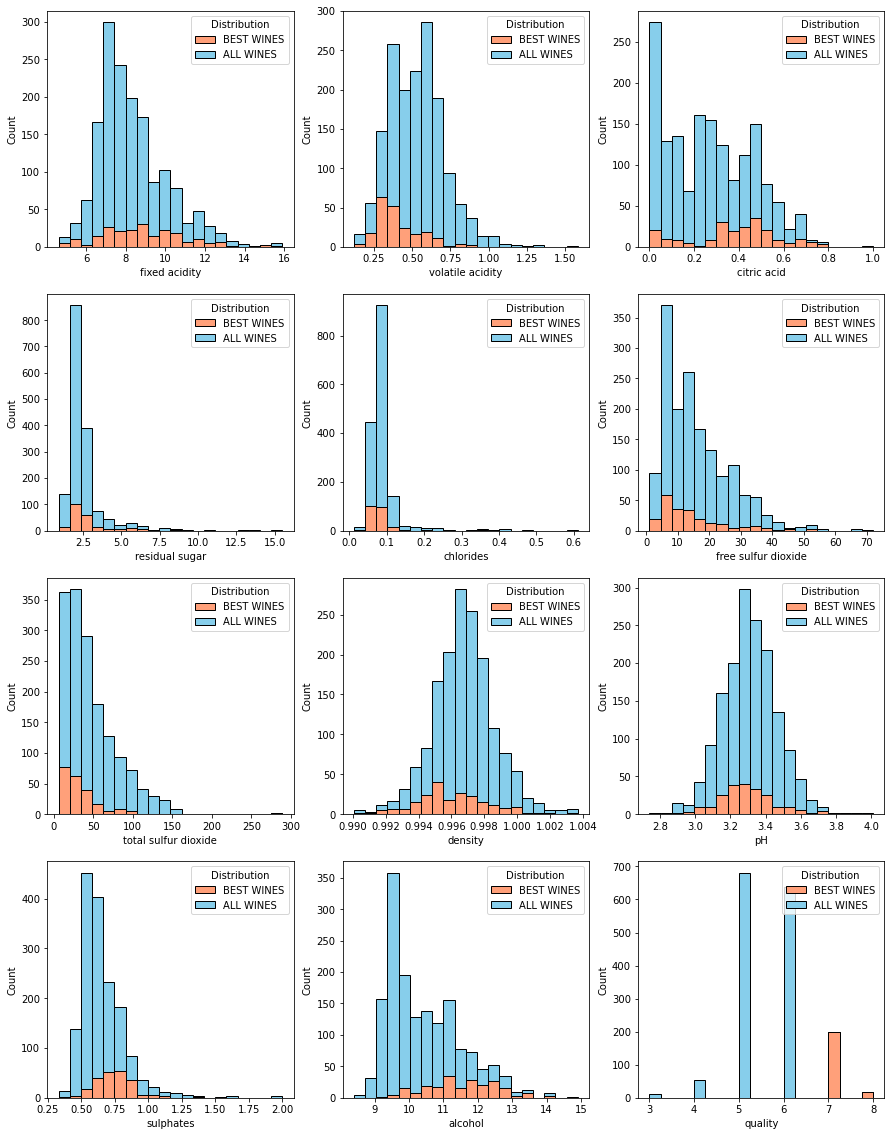
Description automatically generated

* Most of the wine are good (around 5 and 6 quality rating), just a few are bad (3 and 4) or excellent (7,8 and 9)
* The distribution of all wines and the distribution of best wines are similar except for the distribution of volatile acidity and alcohol and sulphates, with density tending to be lower (around 0.99) and alcohol tending to be higher (mostly ranging from 11 to 13 )

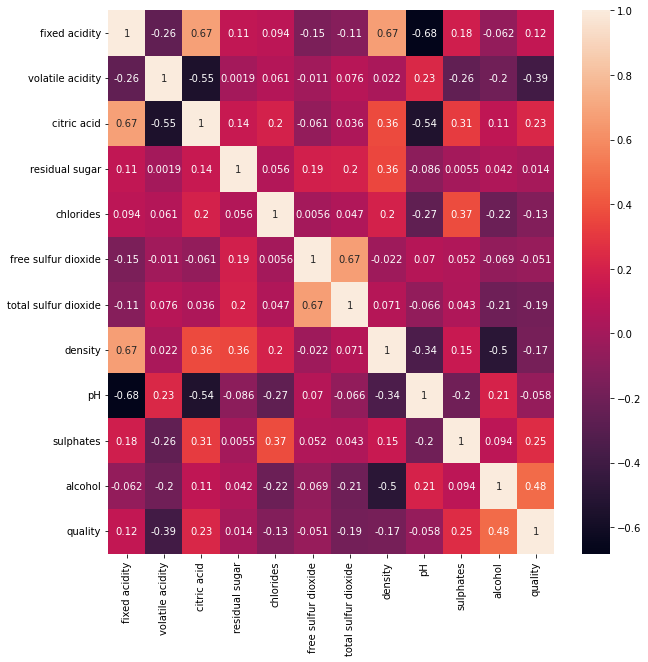


* Quality and alcohol show positive correlation( coefficient = 0.44), quality and density show negative correlation ( coefficient = -0.31).

***For red wine :***

******

* Most of the wine are good (around 5 and 6 quality rating), just a few are bad (3 and 4) or excellent (7 and 8).
* The distribution of all wines and the distribution of best wines are similar except for the distribution of volatile acidity and alcohol and sulphates, with volatile acidity tending to be lower (around 0.25) and alcohol tending to be higher (mostly around 12) and sulphates tending to be higher (around 0.75).



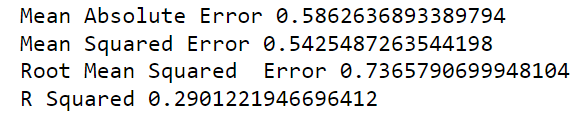
* Quality and alcohol show positive correlation( coefficient = 0.48), quality and volatile acidity show negative correlation ( coefficient = -0.39), quality and sulfates show positive correlation (coffficient = 0.25).

**MODELLING AND ALGORITHM :**

1. ***Multiple linear regression :***

Column ‘quality’ is chosen as y (dependent variable) and the other columns ranging from ‘fixed acidity’ to ‘alcohol’ as x (independent variables) .

White wine model evaluation :



Red wine model evaluation :

Text

Description automatically generated

For red wine and white wine, the R squared values are not high (0.32 and 0.29 respectively) as the linear relationship between quality (target variable) and other varibles are not that strong.

***White wine :*** Graphical user interface, application

Description automatically generated

As we can see, there is no clear linear relationship between any independent variables and the target variable. So, this is one of the reasons why the R-quared value is so low.

***Red wine :***

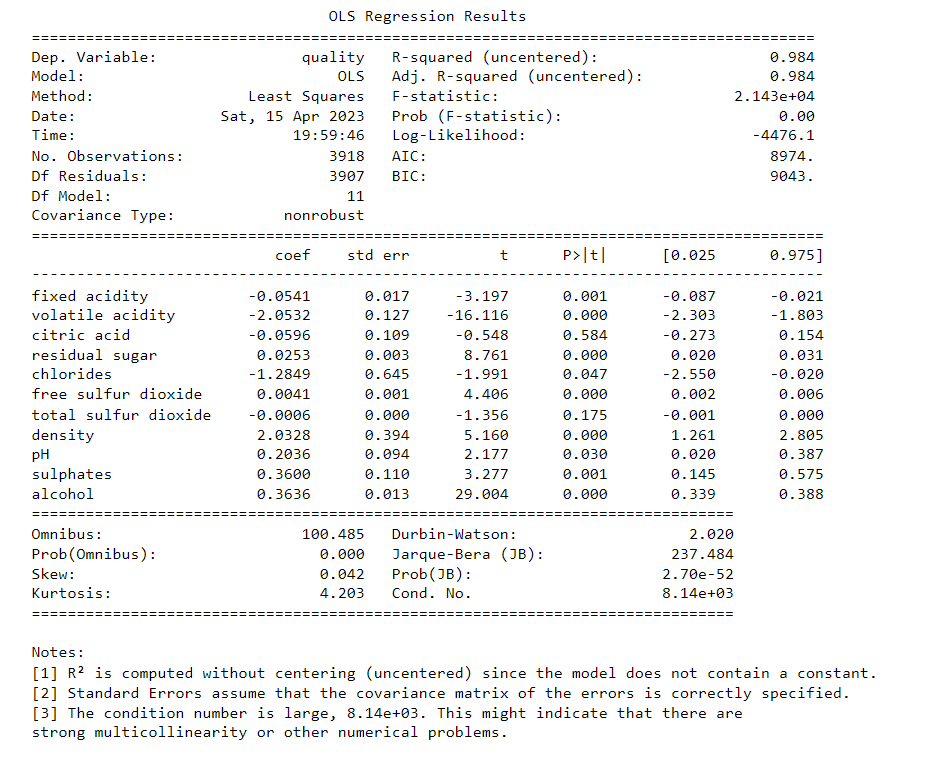
Graphical user interface, application

Description automatically generated

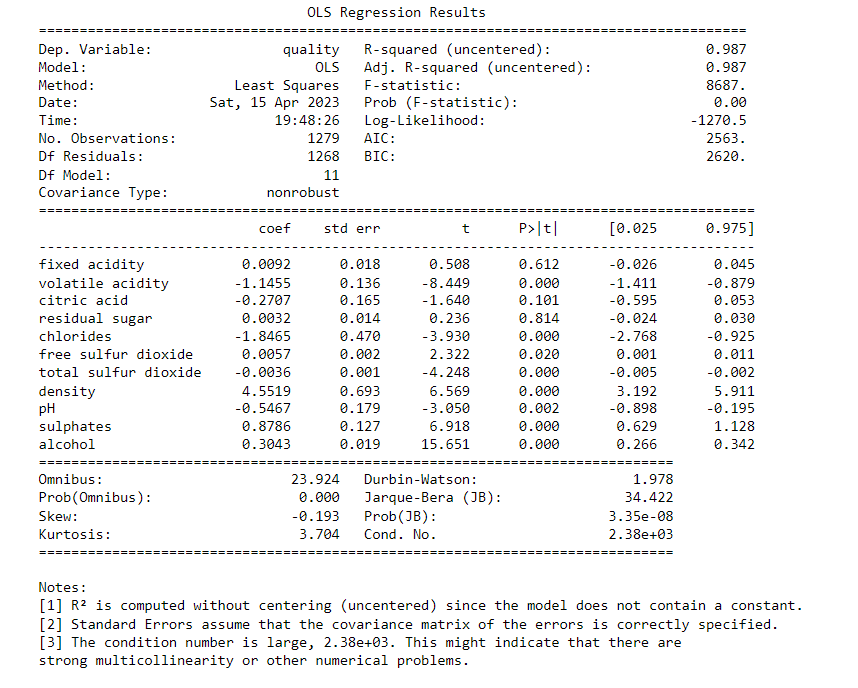
Just the same as white wine, there is no clear linear relationship between any independent variables and the target variable. So, this is one of the reasons why the R-quared value is so low.

1. ***OLS model :***

***White wine :***

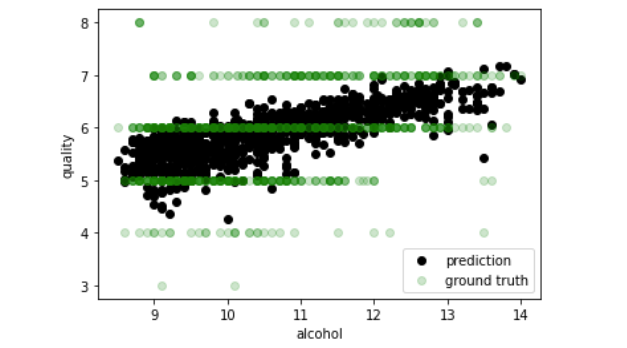


P values of fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates, alcohol are very small. Therefore, these attributes do affect the quality of wine.

***Red wine :*** 

P values of volatile acidity, chlorides, free sulfure dioxide, total sulfur dioxide, density, pH, sulfates, alcohol are very small. Therefore, these values do have impact on the quality of wine.

#### *White wine : Plot scatter of X with Y and y pred*



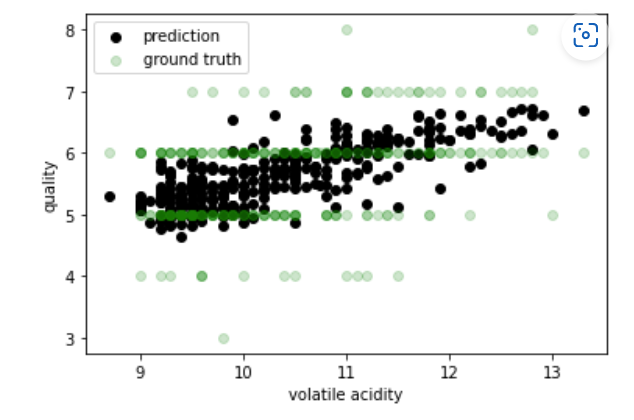
A picture containing diagram

Description automatically generated

Even though the prediction and the ground truth does not seem to fit each other, they are somewhat following the positive linear relationship between alcohol and quality or density and quality.

#### *Red wine : Plot scatter of X with Y and y pred*

#### 



Even though the prediction and the ground truth does not seem to fit each other, they are somewhat following the positive linear relationship between alcohol and quality or volatile acidity and quality.

**CONCLUSION :**

For white wine, alcohol and density are the two most important factors that influences the quality of wine. For red wine, alcohol and volatile acidity play the most important roles in the quality of wine.

Multiple linear regression is not a good model for this dataset because many of the variables depend on each other and there is no clear linear relationship between them and the quality.