RED WINE QUALITY

***CS 4661***

***FALL-2018***

A close up of a bottle and a glass of red wine

Description generated with very high confidence

**Team Members:**

Kruti Shah (306650284)

Ravi Amin (306598765)

Riddhiben Patel (306612701)

Rutviben Patel (306019563)

Smitkumar Kaushikkumar Patel (306587208)

**DATA-SET:** wineQualityReds.csv

### Data set link : <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

**Project Description**

* It shows relation between quality and other variables of wine. We want to do transformation to see if we can increase correlation coefficient between them. Used stepwise variable selection method to choose best predictor of wine quality.

**Goal**

* Our focus is to see how each chemical component influences the quality of wine (0 'very bad' to 10 'very excellent'). The usage of this analysis will help to understand whether by modifying the variables, it is possible to increase the quality of the wine on the market.

**RED WINE DATASET INFORMATION**

In this project we do Analysis of **Red Wine Data** which contains 1,599 red wines with 12 variables on the chemical properties of the wine.

**Input Variables:**

* Fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
* Volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
* Citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
* Residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
* Chlorides: the amount of salt in the wine
* Free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; - it prevents microbial growth and the oxidation of wine
* Total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at - free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
* Density: the density of water is close to that of water depending on the percent alcohol and sugar content
* pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
* Sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
* Alcohol: the percent alcohol content of the wine

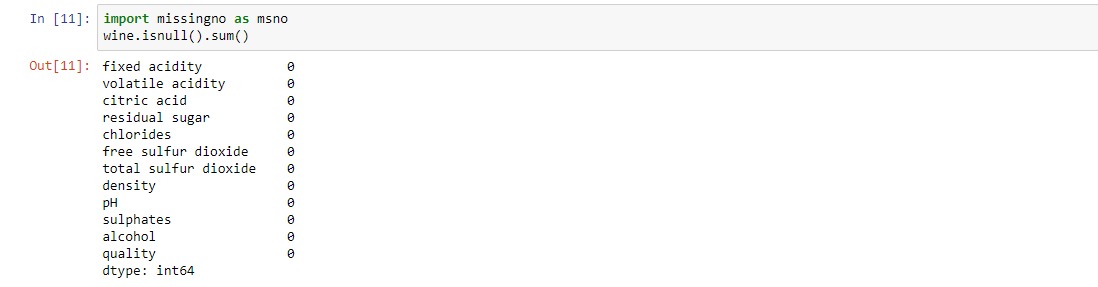
**Output variable:**

* Quality (score between 0 and 10)

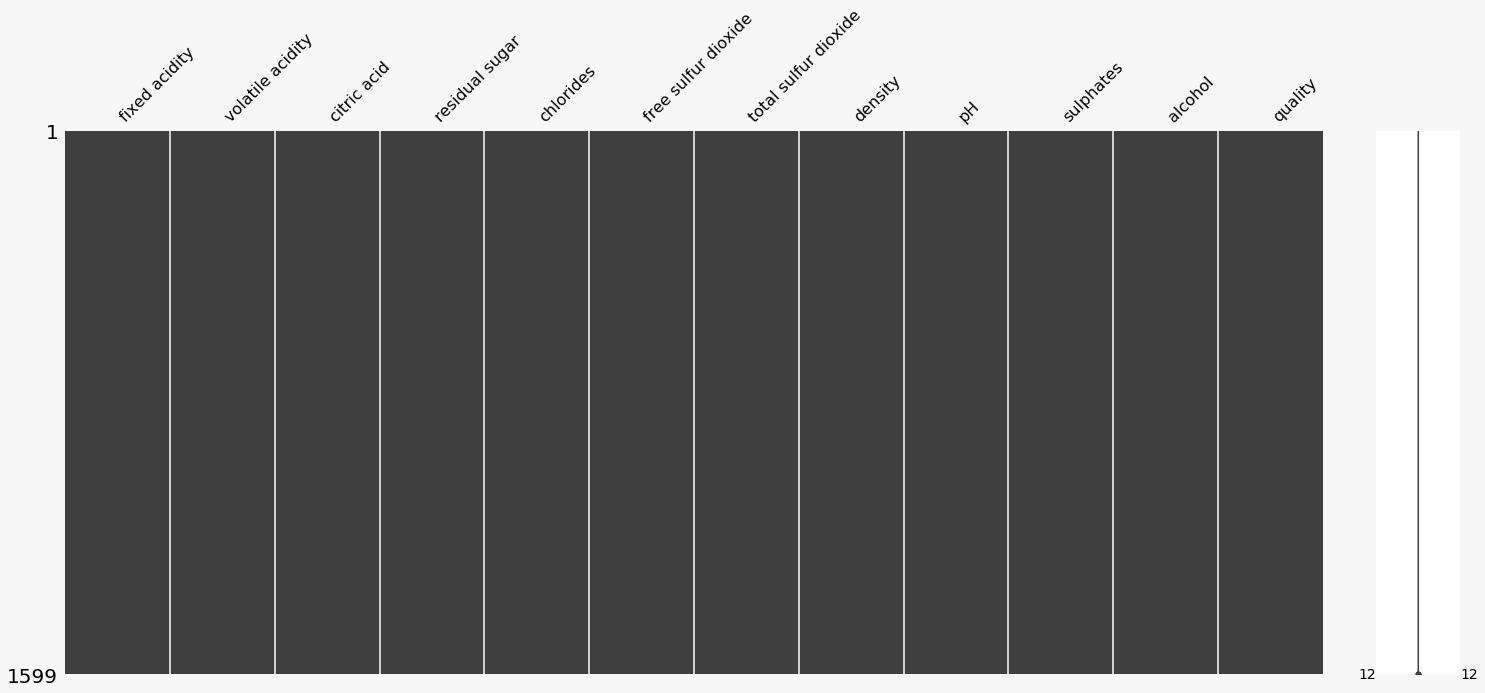
***What we found from the Analysis of this dataset?***

* For the whole data set most of the people gave rating 5 and 6.
* Nobody gave rating 0, 1, 2, 9, 10. This might be because most of the people randomly choose the rating 5 and 6 And surprisingly no body rated 9 and 10 means the wine quality might not be good in reality.
* I first thought that acidity has predictive capability as quality increases with increase value of citric acid and decreases with increased value of volatile acidity.
* For residual sugar nobody gave rating 3 and 8 for the value greater than 6.8.
* May be only one people gave rating 4 for residual sugar value greater than 6.8.
* Most of the rating 5 falls below the alcohol value 11.
* Most of the rating 7 lies above the alcohol value 11.
* Rating 4, 6 are randomly distributed.
* The interesting fact is for the total sulfur dioxide value from 99 to 153 people gave rating 5 except of some outliers.
* People gave high rating for low value of pH.
* No people rated 8 for having chloride value greater than 0.121.
* For sulphate value greater than 0.94 people did not give rating 3.
* May be only one people gave rating 8. Most of the people gave rating 4.
* Density showed predictor for quality as it has trend. For higher value of density, quality is low and for lower value of density, quality is high.
* The linear model gave me seven final variables (volatile acidity, log10(chlorides), free sulfur dioxide, total sulfur dioxide, pH, log10(sulphates), alcohol) for prediction of quality of wine.
* There might be other variables (which are not present in our data) we need to consider for wine quality prediction.

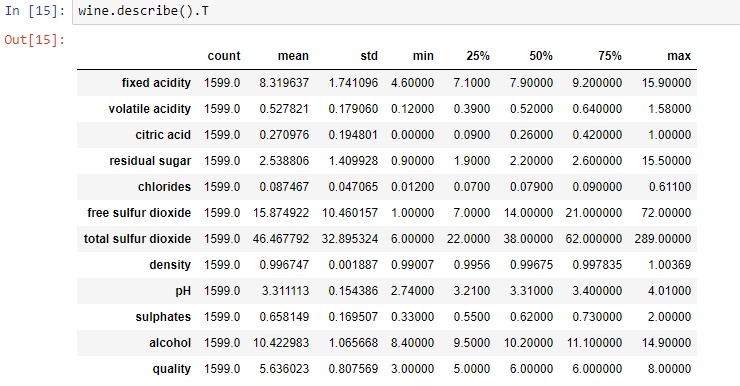
### **Checking the null/missing value in the Dataset [Cleaning Dataset]**



* This function counts the columns which contain null value but here Data is preprocessed and cleaned with dummy and null values.
* Bar-Chart representation depicts that there is no null values in data



### **Statistical information for Dataset**



* The summary of Red Wine dataset looks perfect, there is no visible abnormality or negative values in data.

### **Learn about Target Vector (Output Column)**

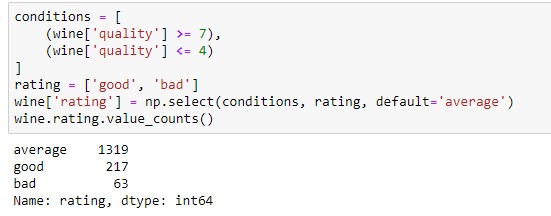
* Target Vector is QUALITY. Nobody gave rating 0, 1, 2, 9, 10. This might be because most of the people randomly choose the rating 5 and 6 And surprisingly no body rated 9 and 10 means the wine quality might not be good in reality.



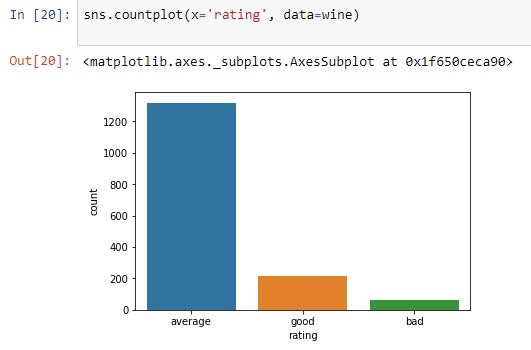
* The above distribution shows the range for response variable (*quality*) is between 3 to 8.

### **Converting numerical value to categorical value of Target Variables**

* We divide the Wine Quality into 3 Categories:
* bad: 1-4
* average: 5-6
* good: 7-10

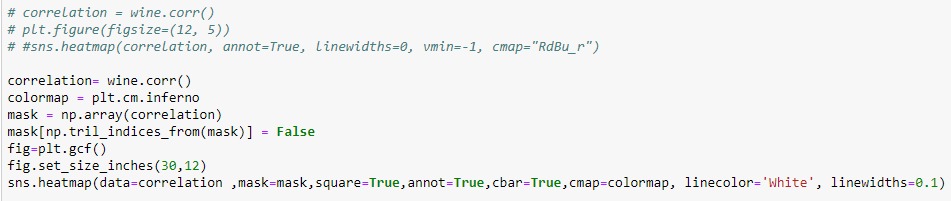


* BAR-CHART REPRESENTATION:

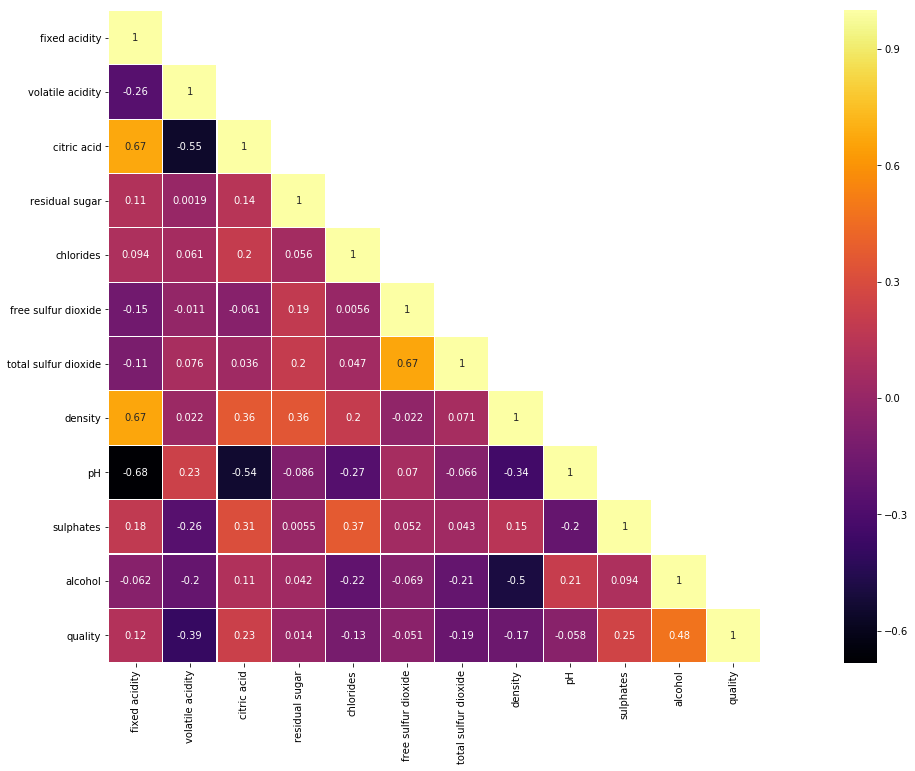


**DATASET ANALYSIS**

### **Correlation between features/variables:**



### Below figure shows how different variables(Factors) are affecting the wine Quality?



### Most affecting Factors are:

* **Alcohol**
* **Volatile acidity**
* **Sulphates**
* **Critic Acid**
* Least affecting Factors:
* **Residual sugar**
* **Free Sulphur Dioxide**
* **Ph**

### Positive Correlated Factors:

### Alcohol

### Sulphates

### Citric acid

### Fixed acidity

### (all the factors are in decreasing order ***Most to least)***

### Negative Correlated Factors:

### volatile acidity

### total sulfur dioxide

### density

### chlorides

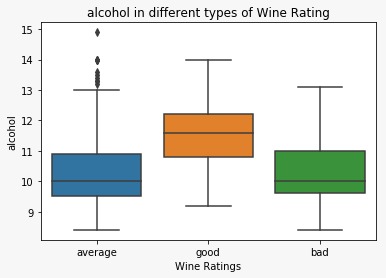
### (all the factors are in decreasing order ***Most to least)***

**Analysis for How Different Factors affect Wine-Quality:**

#### Analysis of alcohol percentage Vs wine quality:

#### 

#### Box-plot representation:



* In figure we can see that Good Wine contain more alcohol.

#### Analysis of sulphates Vs wine ratings:

#### 

#### Box-plot representation:

#### 

* From figure we can say that in Good Wine contain more Sulphate.

#### Analysis of Citric Acid Vs wine ratings:

#### 

#### VIOLINPLOT REPRESENTATION:

#### .

#### We can say that with the Raise of Citric Acid the Wine Quality Gets Improve.

#### Analysis of fixed acidity & wine ratings:

#### 

#### Box-plot representation:

#### 

#### With Increase Feature Fixed Acidity we get good wine Ratings.

#### Analysis of pH Vs wine ratings:

#### 

#### SWARMPLOT REPRESENTATION:

#### 

### **Linear Regression**

* Graphs for different quality ratings shows a linear regression between residual\_sugar & alcohol in red wine



A screenshot of a computer

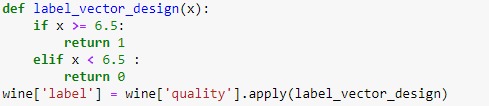
Description generated with very high confidence

* The linear regression plots above for different wine quality ratings (bad, average & good) shows the regression between alcohol and residual sugar content of the red wine.
* We can observe from the trendline that, for good and average wine types the residual sugar content remains almost constant irrespective of alcohol content value. Whereas for bad quality wine, the residual sugar content increases gradually with the increase in alcohol content.
* This analysis can help in manufacturing the good quality wine with continuous monitoring and controlling the alcohol and residual sugar content of the red wine.

**Classification Techniques for Predicting Accuracy**

### For that we divide wine quality in label vector 1(good) and 0(bad)

* 1 (good) quality >= 6.5
* 0 (bad) quality < 6.5



* We use following algorithms to find the accuracy

1. KNearestNeighbors
2. Logistic Regression
3. Random Forest Classifier
4. Decision Tree Classifier

* Accuracy Without Normalization

A screenshot of a social media post

Description generated with very high confidence

A screenshot of a cell phone

Description generated with very high confidence

* Random Forest Classifier Give the best Accuracy: **90.50%**
* Representation of Accuracy Using Graph:

A screenshot of a cell phone

Description generated with very high confidence

A close up of a logo

Description generated with very high confidence

* Accuracy after Normalization

A screenshot of a cell phone

Description generated with very high confidence

A screenshot of a social media post

Description generated with very high confidence

A screenshot of a cell phone

Description generated with very high confidence

* We found that after normalization Accuracy remain same so, Data is already

Normalized

Responsibility of Team Members:

**Kruti Shah:**

Data Analysis for How Different Factor Affect Wine Quality

Find Accuracy: Random Forest, (BEFORE NORMALIZATION)

Decision Tree Algorithm, (BEFORE NORMALIZATION)

Presentation

**Ravi Amin:**

Correlation Between Features And Variables,

Find Accuracy: Random Forest (AFTER NORMALIZATION)

Decision Tree (AFTER NORMALIZATION)

Documentation

**Riddhiben Patel:**

Data Analysis to Check for The Null/Missing Values

Statistical information for Dataset

Learn about Target Vector

Find Accuracy: Logistic Regression (BEFORE NORMALIZATION)

**Rutviben Patel:**

Correlation Between Features and Variables,

Find Accuracy: KNN (AFTER NORMALIZATION)

LOGISTIC (AFTER NORMALIZATION)

Presentation

**Smitkumar Kaushikkumar Patel:**

Converting numerical value to categorical value of Target Variables

Data Analysis for How Different Factor Affect Wine Quality,

Correlation Analysis

Find Accuracy: LINEAR REGRESSION,

KNN (BEFORE NORMALIZATION)

Documentation