# Practical Machine Learning

# Exploratory Data Analysis Report

Sanjeeb Humagain

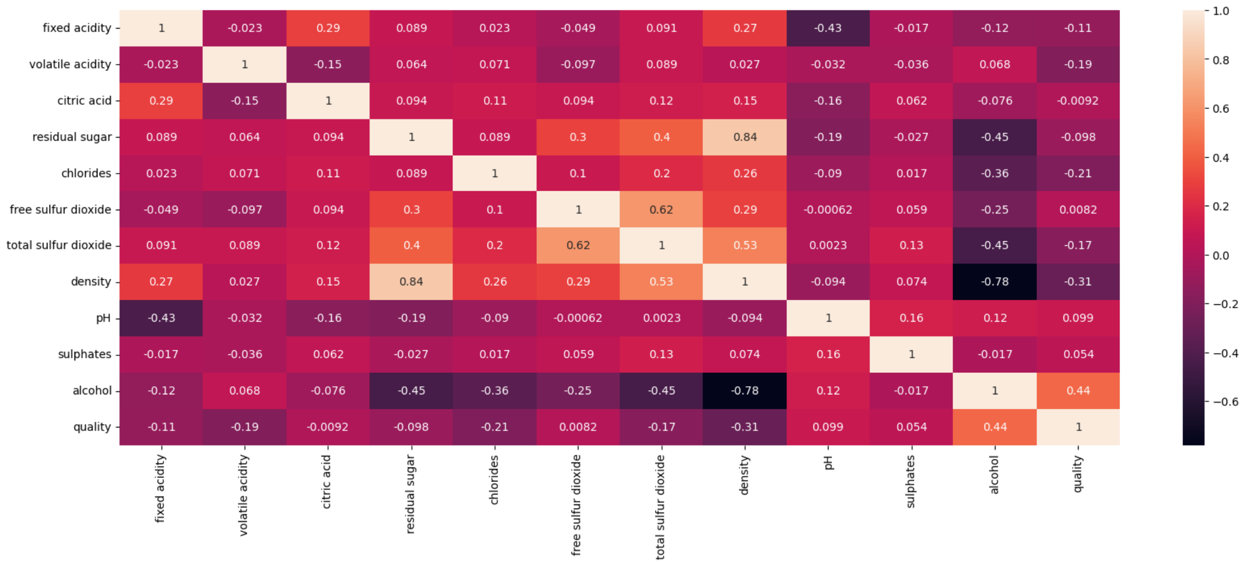
April 9, 2024

## Introduction:

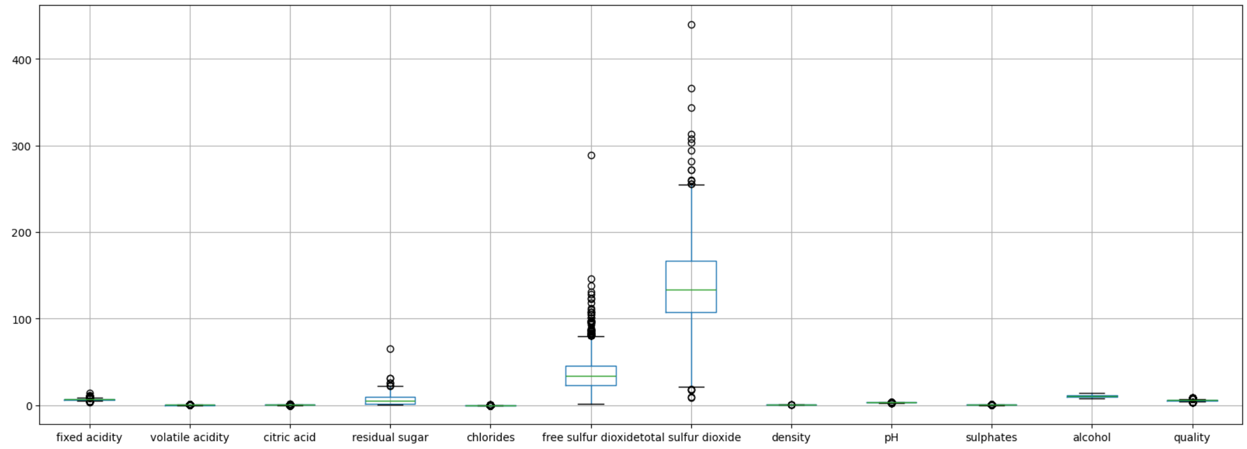
Exploratory data analysis (EDA) is used to check the pattern of data and prepare it before feeding it into any Machine Learning model to improve the performance of the model. Sometimes, it would be difficult to comprehend the nature of data until we visualize it. Analyzing the data would help to decide whether data cleaning is needed or not. Data preprocessing is crucial to transform the raw data into machine readable format. It involves steps such as, performing analysis, filtering the data, transformation, and encoding the data. Improper data preprocessing will degrade the quality and performance of the model. EDA is used to conduct an analysis on raw data to see if there are any missing values, visualize the data distribution, see the correlation between features and remove unnecessary data and features from the training data set. The goal is to make the data suitable for the ML model which will increase the model performance.

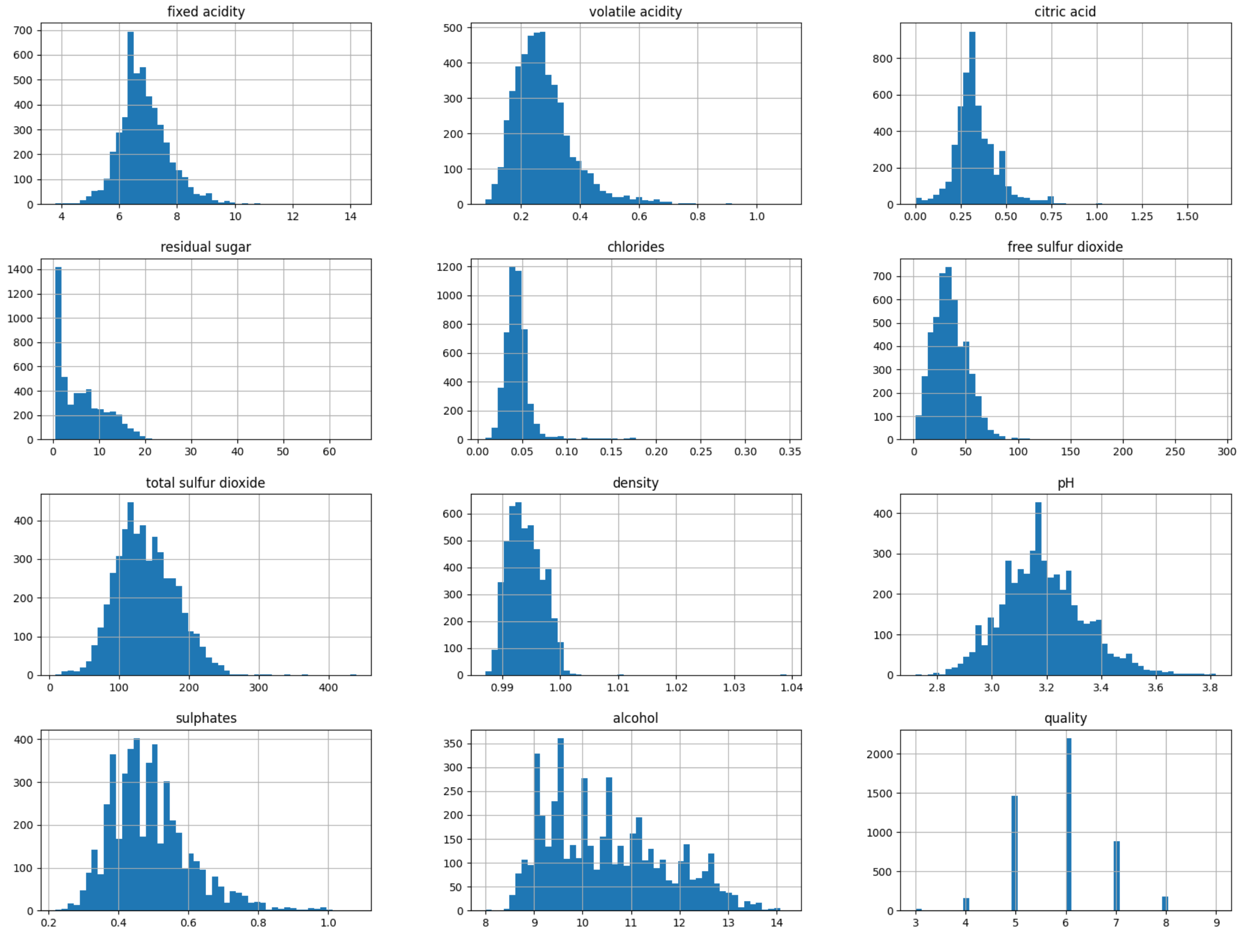
## Results of the Analysis:

The Wine Quality data of red wine was used for this analysis which has 12 features and 4898 samples. The data does not have missing values which we could see by using info or describe function of pandas. Describe function gives the important statistical information of each column of the data set. Additionally, using the isnull() function we can see the number of missing values in each features if they exist. This would help in data cleaning process. Additionally, the correlation of between the features was also visualized using heatmap function. Which helped to get the understanding of how the features are correlated. We could remove some features if found uncorrelated and unnecessary. Also, the box plot helped to visualize the outlier which should be removed to make the performance of the ML model better. There were functions which we could use to remove some features, sort on the basis of some features that would help to understand the nature of features and compare them. Some of the images of the model is attached below.



| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fixed acidity | 1 | -0.023 | 0.29 | 0.089 | 0.023 | -0.049 | 0.091 | 0.27 | -0.43 | -0.017 | -0.12 | -0.11 |
| volatile acidity | -0.023 | 1 | -0.15 | 0.064 | 0.071 | -0.097 | 0.089 | 0.027 | -0.032 | -0.036 | 0.068 | -0.19 |
| citric acid | 0.29 | -0.15 | 1 | 0.094 | 0.11 | 0.094 | 0.12 | 0.15 | -0.16 | 0.062 | -0.076 | -0.0092 |
| residual sugar | 0.089 | 0.064 | 0.094 | 1 | 0.089 | 0.3 | 0.4 | 0.84 | -0.19 | -0.027 | -0.45 | -0.098 |
| chlorides | 0.023 | 0.071 | 0.11 | 0.089 | 1 | 0.1 | 0.2 | 0.26 | -0.09 | 0.017 | -0.36 | -0.21 |
| free sulfur dioxide | -0.049 | -0.097 | 0.094 | 0.3 | 0.1 | 1 | 0.62 | 0.29 | -0.00062 | 0.059 | -0.25 | 0.0082 |
| total sulfur dioxide | 0.091 | 0.089 | 0.12 | 0.4 | 0.2 | 0.62 | 1 | 0.53 | 0.0023 | 0.13 | -0.45 | -0.17 |
| density | 0.27 | 0.027 | 0.15 | 0.84 | 0.26 | 0.29 | 0.53 | 1 | -0.094 | 0.074 | -0.78 | -0.31 |
| pH | -0.43 | -0.032 | -0.16 | -0.19 | -0.09 | -0.00062 | 0.0023 | -0.094 | 1 | 0.16 | 0.12 | 0.099 |
| sulphates | -0.017 | -0.036 | 0.062 | -0.027 | 0.017 | 0.059 | 0.13 | 0.074 | 0.16 | 1 | -0.017 | 0.054 |
| alcohol | -0.12 | 0.068 | -0.076 | -0.45 | -0.36 | -0.25 | -0.45 | -0.78 | 0.12 | -0.017 | 1 | 0.44 |
| quality | -0.11 | -0.19 | -0.0092 | -0.098 | -0.21 | 0.0082 | -0.17 | -0.31 | 0.099 | 0.054 | 0.44 | 1 |




fixed acidity
0 100 200 300 400 500 600 700
4 6 8 10 12 14
volatile acidity
0 100 200 300 400 500
0.2 0.4 0.6 0.8 1.0
citric acid
0 200 400 600 800
0.00 0.25 0.50 0.75 1.00 1.25 1.50
residual sugar
0 200 400 600 800 1000 1200 1400
0 10 20 30 40 50 60
chlorides
0 200 400 600 800 1000 1200
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
free sulfur dioxide
0 100 200 300 400 500 600 700
0 50 100 150 200 250 300
total sulfur dioxide
0 100 200 300 400
0 100 200 300 400
density
0 100 200 300 400 500 600
0.99 1.00 1.01 1.02 1.03 1.04
pH
0 100 200 300 400
2.8 3.0 3.2 3.4 3.6 3.8
sulphates
0 100 200 300 400
0.2 0.4 0.6 0.8 1.0
alcohol
0 50 100 150 200 250 300 350
8 9 10 11 12 13 14
quality
0 500 1000 1500 2000
3 4 5 6 7 8 9