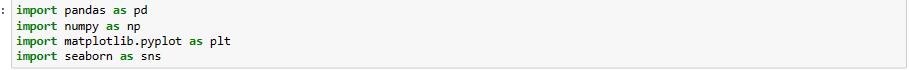
**BHARATINTERN TASK#2 – WINE QUALITY PREDICTION**

**TASK#2 – WINE QUALITY PREDICTION USING SIMPLE LINEAR REGRESSION(SLR)**

**Step1-** Importing All The required libraries of Python in Jupyter Notebook.



# Step2- Understanding the Dataset

**Features Included in the Dataset:** The dataset consists of various features or columns, each providing specific information about wine.

**Range of Values for Each Feature:** To gain insights into the dataset's characteristics, we will examine the range of values for each feature. This involves calculating the minimum, maximum, mean, and standard deviation for numeric features. This step helps identify any potential outliers or data inconsistencies.

**Handling Missing Values and Outliers:** During our analysis, I also pay close attention to missing values and outliers. Missing data can be problematic, as it may affect the quality of our predictions. I will decide on an appropriate strategy for handling missing values, such as imputation or removal. Additionally, identifying and addressing outliers is essential, as they can distort our models' predictions. Outliers may be handled through techniques like trimming or transformation.

# Step3- Analysing Features in the Dataset

# Wine is an alcoholic beverage made from fermented grapes. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide, and heat. It is a pleasant tasting alcoholic beverage, loved celebrated. It will definitely be interesting to analyse the physicochemical attributes of wine and understand their relationships and significance with wine quality classifications.

**Attribute Description:**

**fixed acidity** : Primary fixed acids found in wine are tartaric, succinic, citric, and malic.

**volatile acidity:** Gaseous acids present in wine.

**citric acid :** It is weak organic acid, found in citrus fruits naturally.

**residual sugar:** Amount of sugar left after fermentation.

**chlorides :** Amount of salt present in wine.

**free sulfur dioxide :** So2 is used for prevention of wine by oxidation and microbial spoilage.

**total sulfur dioxide :** Total So2 in wine

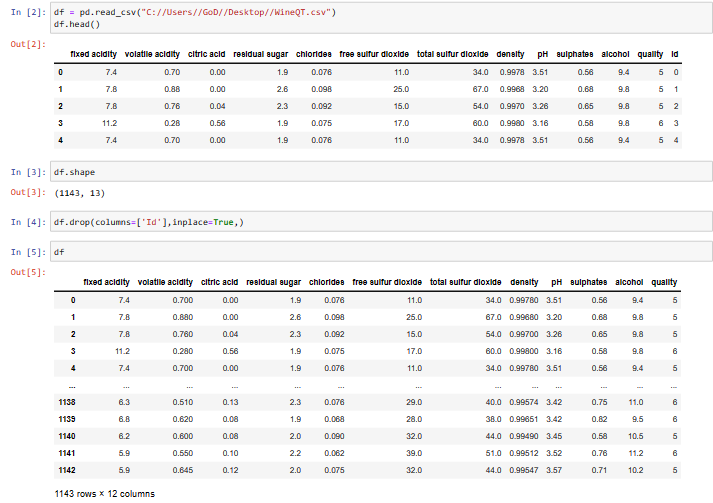
**density :** Describes the density of the wine.

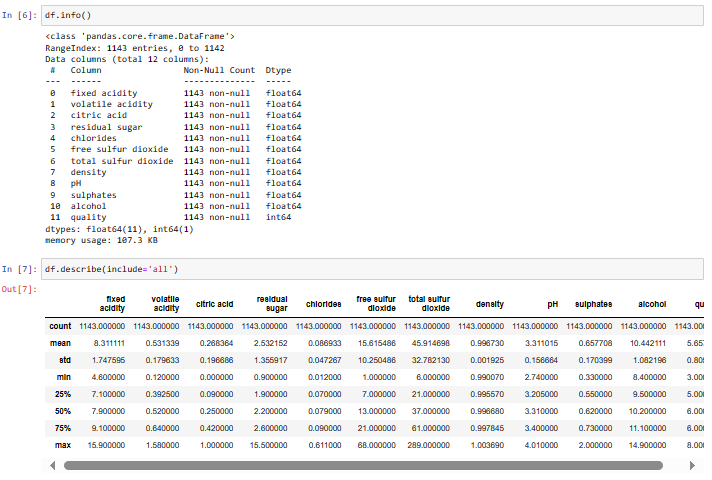
**pH :** pH is used for checking acidity.

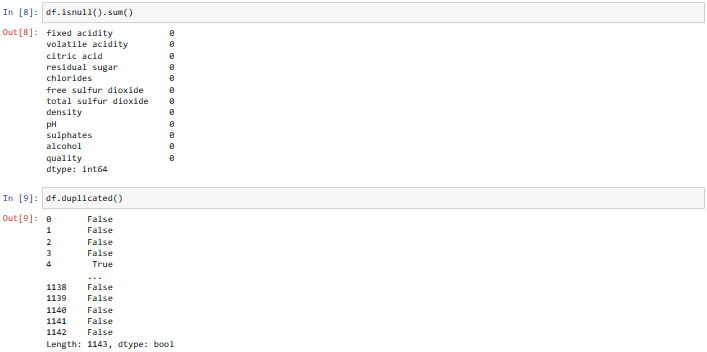
**sulphates :** Added sulfites preserve freshness and protect wine from oxidation, and bacteria.

**alcohol :** Percent of alcohol present in wine.

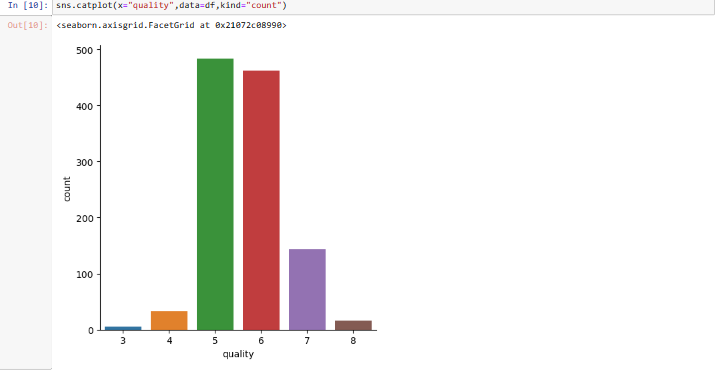
Each feature plays a significant role in influencing wine quality in the beverages markets. By thoroughly examining these features, I aim to uncover valuable insights that will inform my predictive modeling efforts. I tried to explore the relationships between features, identify key variables that impact wine quality, and assess the overall data quality. This analysis will serve as a foundation for data preprocessing and modeling steps, ensuring that I build robust and accurate predictive models for wine quality.

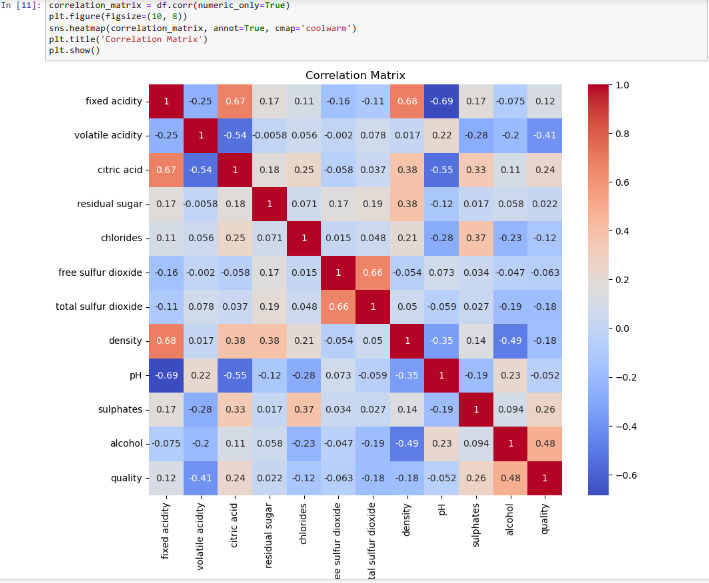






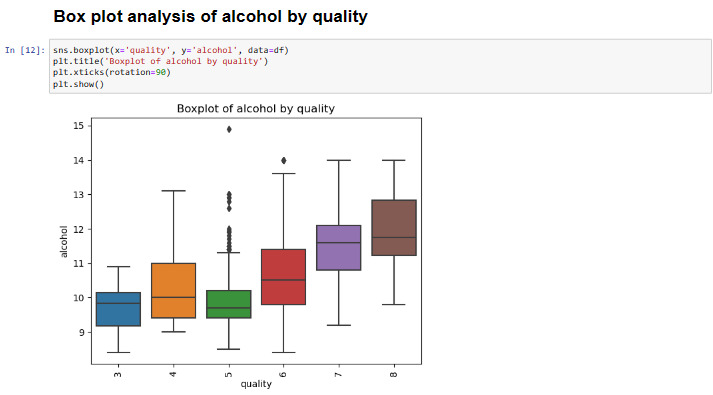
The **catplot** method is used to plot categorical plots.This function gives users access to a number of axes-level functions that illustrate the connection between numerical data and one or more category variables. The underlying axes-level function is chosen by the kind parameter. Here y-axis is kind which takes count of wine quality and x-axis represents various qualities of wine.



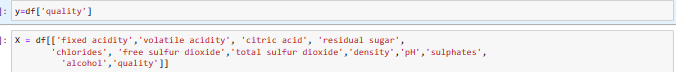


#### **Step4-** Boxplot Analysis by Quality

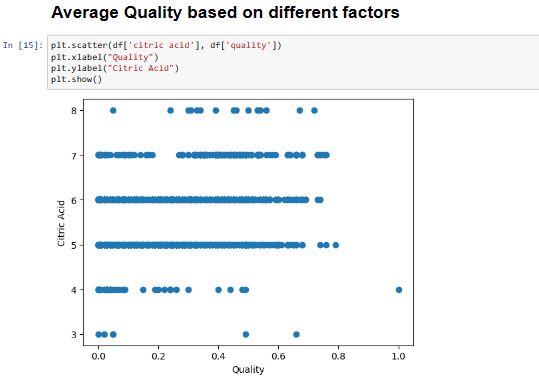
To gain insights into the distribution of the number of alcohol brands of different qualities, I created a boxplot. The boxplot provides a visual representation of the data's central tendency and spread. I plotted the number of alcohol brands on the y-axis and the quality on the x-axis. Each box represents a quality wise distribution of alcohol counts.

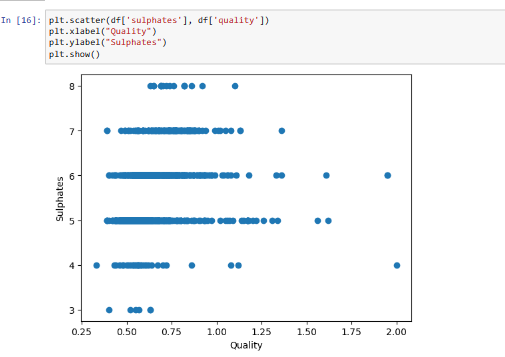


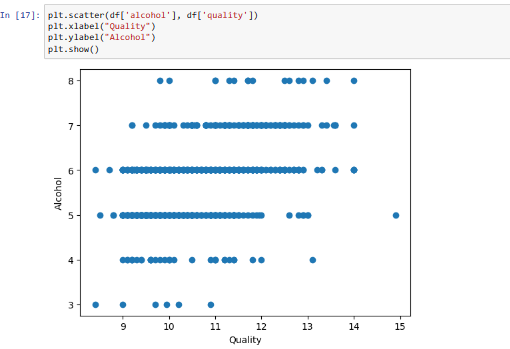
**Step5-** Splitting Dataset into X and y for training and testing purpose.



**Step6-** Finding the Average Alcohol Quality based on different factors







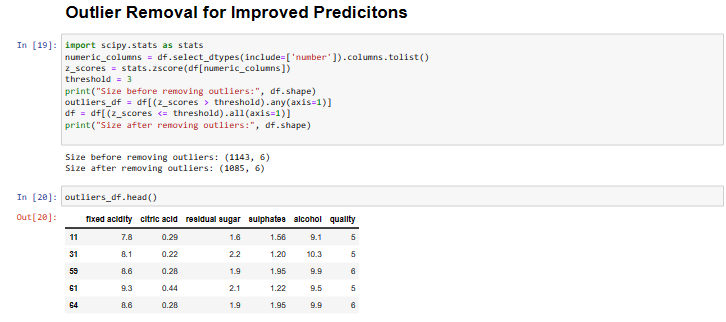
**Step7-** Feature Significance Analysis

Feature Significance analysis is all about selecting the most important features out of all the features for more accurate and precise prediction of the data so that there are no dominance of anyone feature over others. I have removed total of 6 features which I thought would be of no need for my analysis of quality.



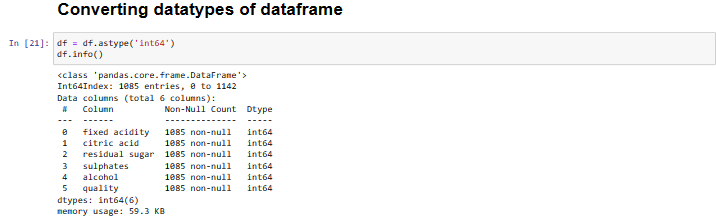
**Step8-** Outlier Removal for better prediction

Outlier removal is one of the most important and compulsory task of a data analysis project. Outliers are those data points which are very far from the mean of the dataset and does not underlie within a particular pattern. It is essential to remove those. Total of 58 Outliers were removed after cleaning.

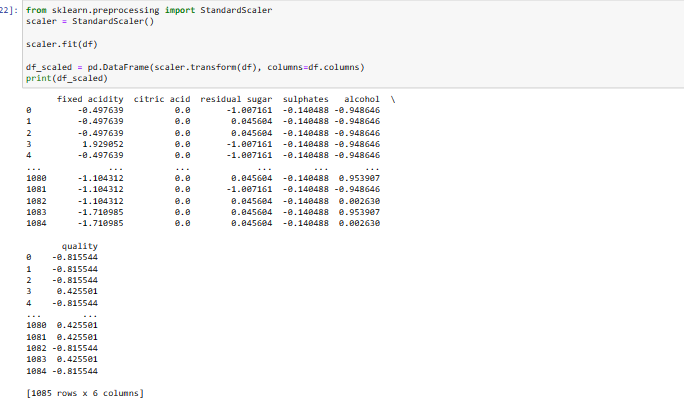


**Step9-** Converting all the datatypes into integers.

Converting all the datatypes to integers so that they can be easily fit into the Regression model for machine learning and prediction of wine quality.



**Step10-** Scaling all the features



**Step11-** Implementing Linear Regression model for wine quality prediction

Splitting the dataset into X(Independent) variables and y(Dependent) variable where y is the quality that I want to predict and X are all the factors on which y depends. Training and testing them in the Simple Linear Regression Model(SLR), evaluating the output, accuracy, Mean Squared Error(MSE) and R2 Score, Visualized the actual output and predicted output using scatterplot. R2 Score is 1 which is maximum which means our model is almost nearly 100% accurate and Alcohol is the factor on which quality mostly depends on.

