Exercise-2: Exploratory Data Analysis

Bimal Pandey

Exploratory Data Analysis:

Exploratory Data Analysis refers to the crucial process of conducting preliminary investigations on data to discover patterns, identify anomalies, test hypotheses, and verify assumptions using summary statistics and graphical representation. It is a good habit to first examine the data and try to extract as many insights as possible. EDA is all about making sense of the data at hand before getting filthy with it. Data handlers utilize exploratory data analysis (EDA) to study and investigate data sets, as well as describe their key properties, frequently using data visualization techniques. EDA assists in determining how to best modify data sources to obtain the answers they require, making it easier for them to detect patterns, identify anomalies, test a hypothesis, or confirm assumptions.

EDA on Red Wine Quality Analysis: The Red Wine dataset consists of 1599 observations and 12 characteristics, out of which 11 are input variables and the remaining one is output variable. Here, the data have only float and integer values (only for the target variable) and there are no null/missing values. The describe() function returns the count, mean, standard deviation, minimum, 25%, 50%, 75%, and maximum values and the qualities of data. The duplicate records are removed using data.drop_duplicates(inplace=True).

Input Variables:

- fixed Acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

Output variable:

- quality

Used libraries and modules: To start with, I imported the following libraries and loaded the dataset, and the original data are separated by ";" in the given data set.

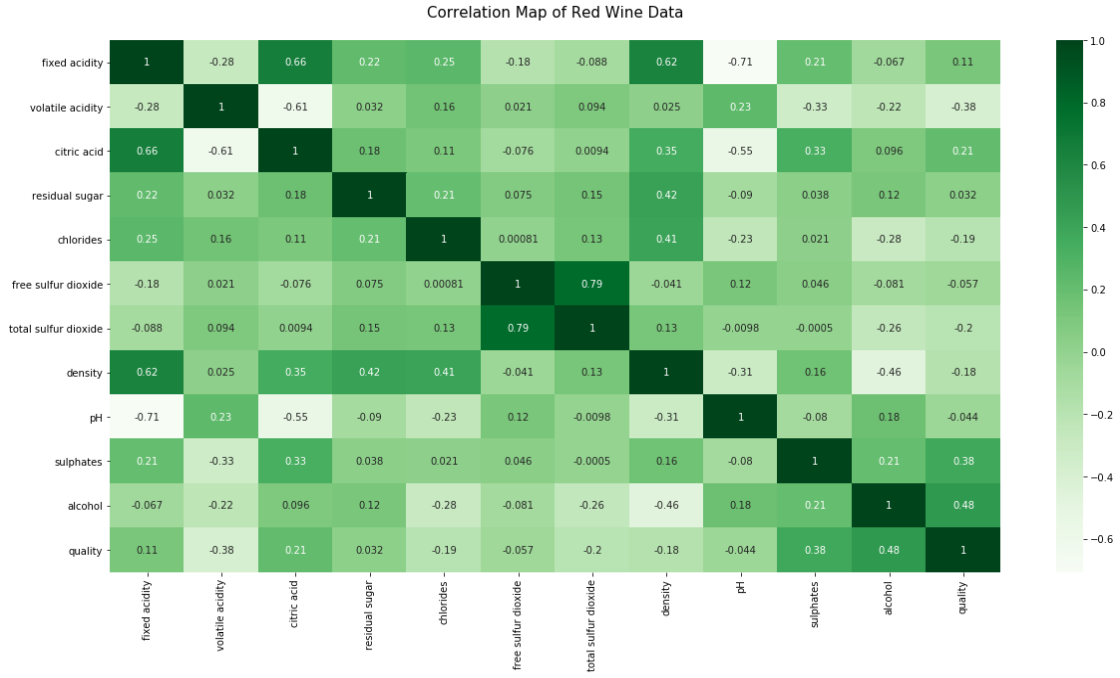
- Numpy: It will provide support for efficient numerical computation
- Pandas: It is a convenient library that supports data frames. Working with pandas will bring ease to many crucial data operations.

- Seaborn: It is a visualization library based on Matplotlib which provides a high-level interface for drawing attractive statistical graphics.
- Sklearn: It is a Python library for data mining, data analysis, and machine learning.
- Matplotlib: It provides a MATLAB-like plotting framework

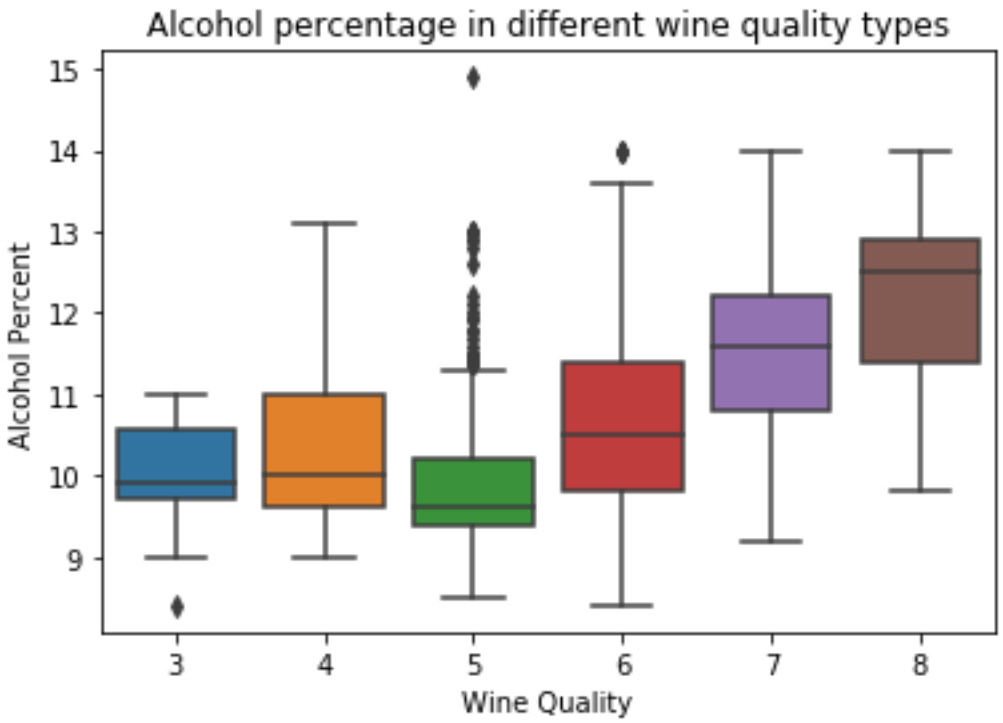
Most of the data points are of quality 5 and very few correspond to quality levels 6 and 7. So, we can say that the dataset is an example of an imbalanced dataset. From the correlation map and values obtained by printing the correlation response of all input variables with quality, we can say that alcohol, sulphates, citric acid, fixed acidity, and residual sugar have a maximum correlation with the response variable quality. Here, we can also see that "density" has a significant positive association with "residual sugar" but a strong negative correlation with "alcohol". Similarly, “free sulfur dioxide” has almost no correlation with “quality”. This means that they must be thoroughly examined for extensive pattern and correlation exploration and they should be focused for future investigation solely on these four variables.

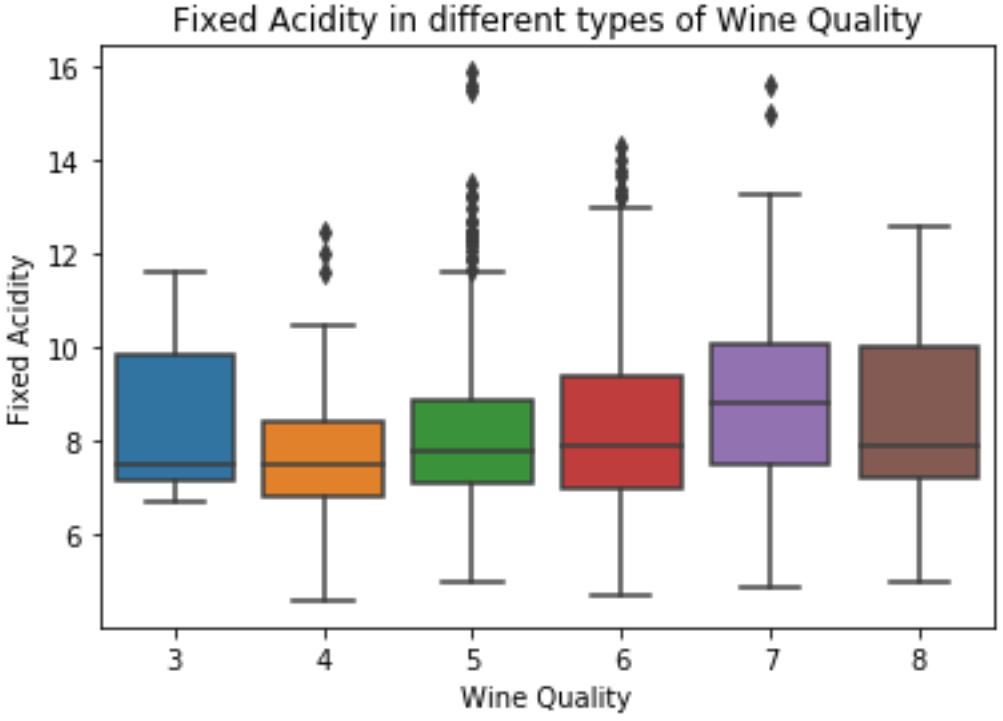
The box plot is used for further analysis of these variables for different wine quality types. A box plot (or box-and-whisker plot) depicts the distribution of quantitative data in a form that allows for comparisons between variables. The box represents the dataset's quartiles, while the whiskers extend to illustrate the remainder of the distribution. The box plot (also known as the box and whisker diagram) is a standardized method of depicting the distribution of data based on the five-number summary.

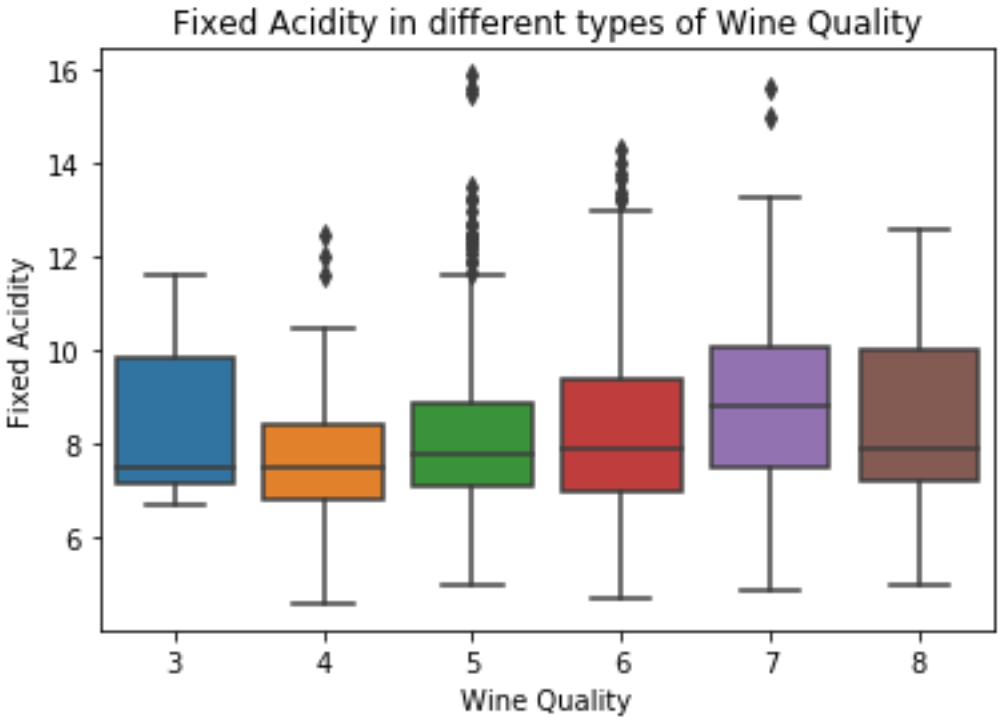
The correlation of different input variables for the output response variable quality is shown in the following map:



The box plot for the highly correlated input variables are shown below:


Alcohol percentage in different wine quality types
Alcohol Percent
15
14
13
12
11
10
9
3
4
5
6
7
8
Wine Quality


Fixed Acidity in different types of Wine Quality
Fixed Acidity
16
14
12
10
8
6
3
4
5
6
7
8
Wine Quality


Fixed Acidity in different types of Wine Quality
Fixed Acidity
16
14
12
10
8
6
3
4
5
6
7
8
Wine Quality