Experiment 1: Exploratory Data Analysis

Pravesh Ganwani
T.E. I.T.
Batch B
Roll No. 2018140021

Aim:
To perform exploratory data analysis on a dataset using python libraries

Problem Statement:
Choose a dataset of your choice from any of the following Repository Links, download it:

1. Kaggle: https://www.kaggle.com/
2. UCI Machine Leaning Repository: https://archive.ics.uci.edu/ml/index.php

Perform EDA on the chosen dataset. This must cover the following aspects:

1. Data statistics – pandas
   Number of data samples, number of features, number of classes, number of data samples per class
2. Data Cleaning – pandas
   Removing of missing values, conversion to numbers etc.
3. Data visualisation – pandas/matplotlib/seaborn
   Work with any 3 different plots of pandas/matplotlib and 3 different plots of seaborn on the dataset and write your observations on the basis of the plots.

Tool/Language:
Programming language: Python
Libraries: NumPy, pandas, matplotlib, seaborn

Code with visualization graphs:

1) **Dataset Chosen:** Red Wine Quality
2) **Dataset Description:** This dataset is related to red variants of the Portuguese "Vinho Verde" wine.
   a) Input variables (based on physicochemical tests):
      - fixed acidity (most acids involved with wine or fixed or non-volatile)
      - 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 flavour to wines)
      - residual sugar (the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/litre)
      - chlorides (the amount of salt in the wine)
      - free Sulphur dioxide (the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulphite ion)
      - total Sulphur dioxide (amount of free and bound forms of S02)

- 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))
- sulphates (a wine additive which can contribute to Sulphur dioxide gas (S02) levels, which acts as an antimicrobial)
- alcohol

b) Output variable (based on sensory data):

- quality (score between 0 and 10)

3) Code:

```
# Importing Dataset to Google Colab
from google.colab import files
uploaded = files.upload()

# Importing Libraries
import numpy as np
import pandas as pd
import io
import matplotlib.pyplot as plt
import seaborn as sns

# Reading the dataset
df = pd.read_csv(io.BytesIO(uploaded['winequality-red.csv']))
df.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

```
# Exploratory Data Analysis
df.shape  # Gives no. of rows and no. of features/columns
```

```
(1599, 12)
```

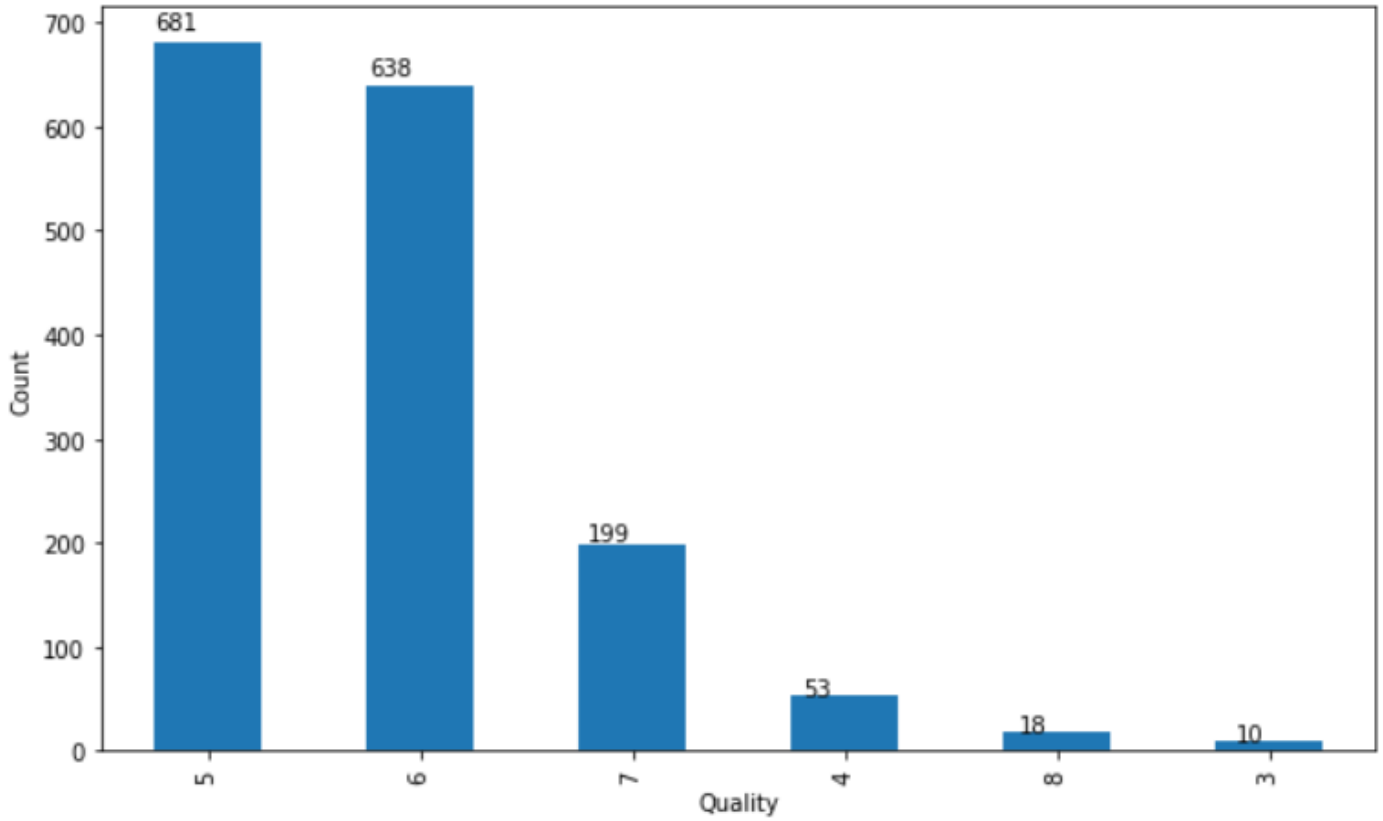
```
df.info() # Entire information about the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

```
df.describe() # Gives a brief description about the dataset
```

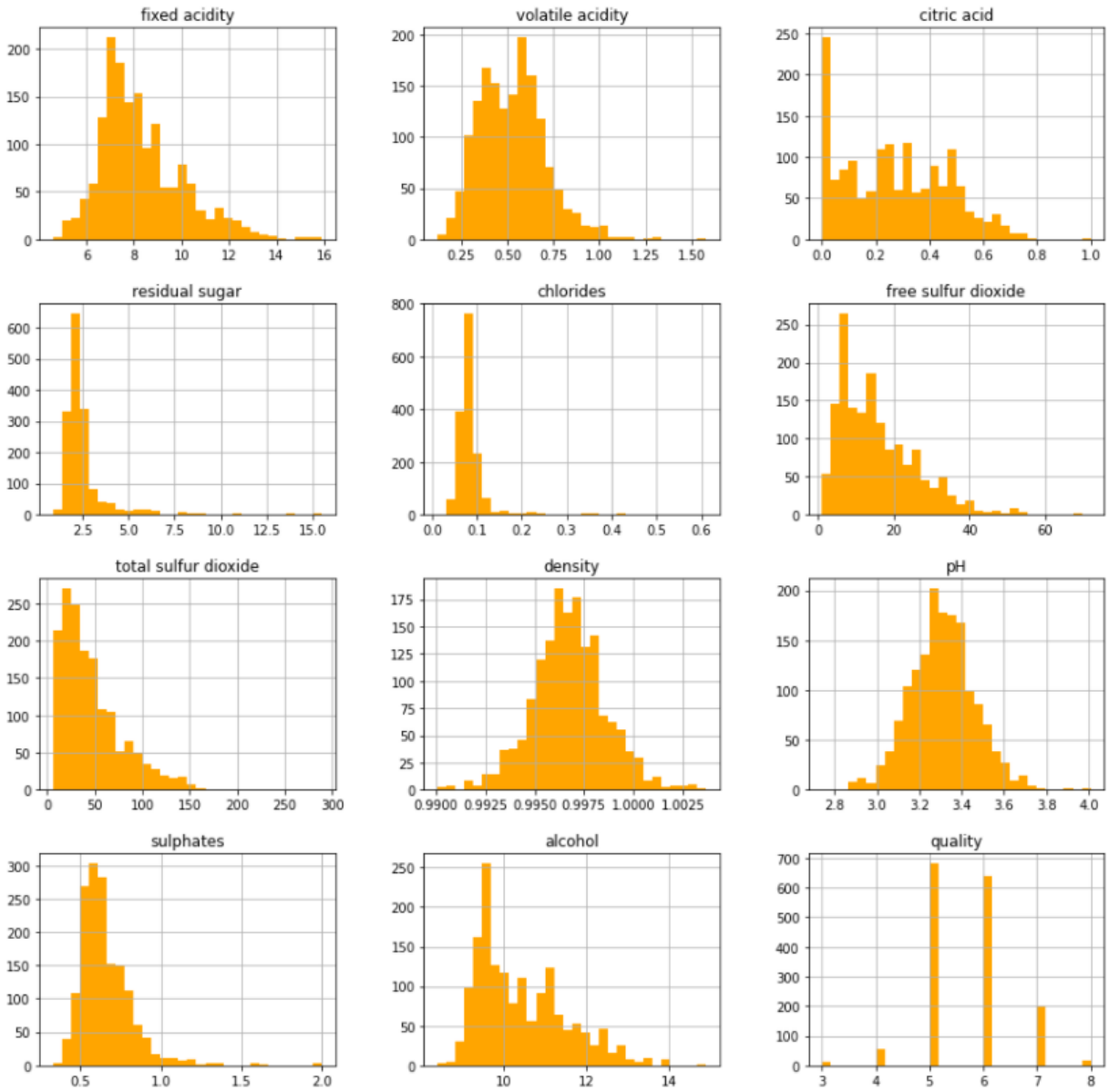
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

```
# Visualizations using Matplotlib and Pandas

# Graph 1 - Plotting bar graph of count of unique values of quality

fig = plt.figure(figsize=(10,6))
quality_counts = df['quality'].value_counts()
ax = quality_counts.plot.bar()
ax.set_xlabel('Quality')
ax.set_ylabel('Count')
for p in ax.patches:
    ax.annotate(str(p.get_height()), (p.get_x() * 1.02, p.get_height
() * 1.02))
```



Observation: The above graph shows count of unique values under quality table. The graph concludes that we have a large set of wines labelled under Quality Rank '5', whereas there are very few sets of wines labelled under Quality Rank '3'.

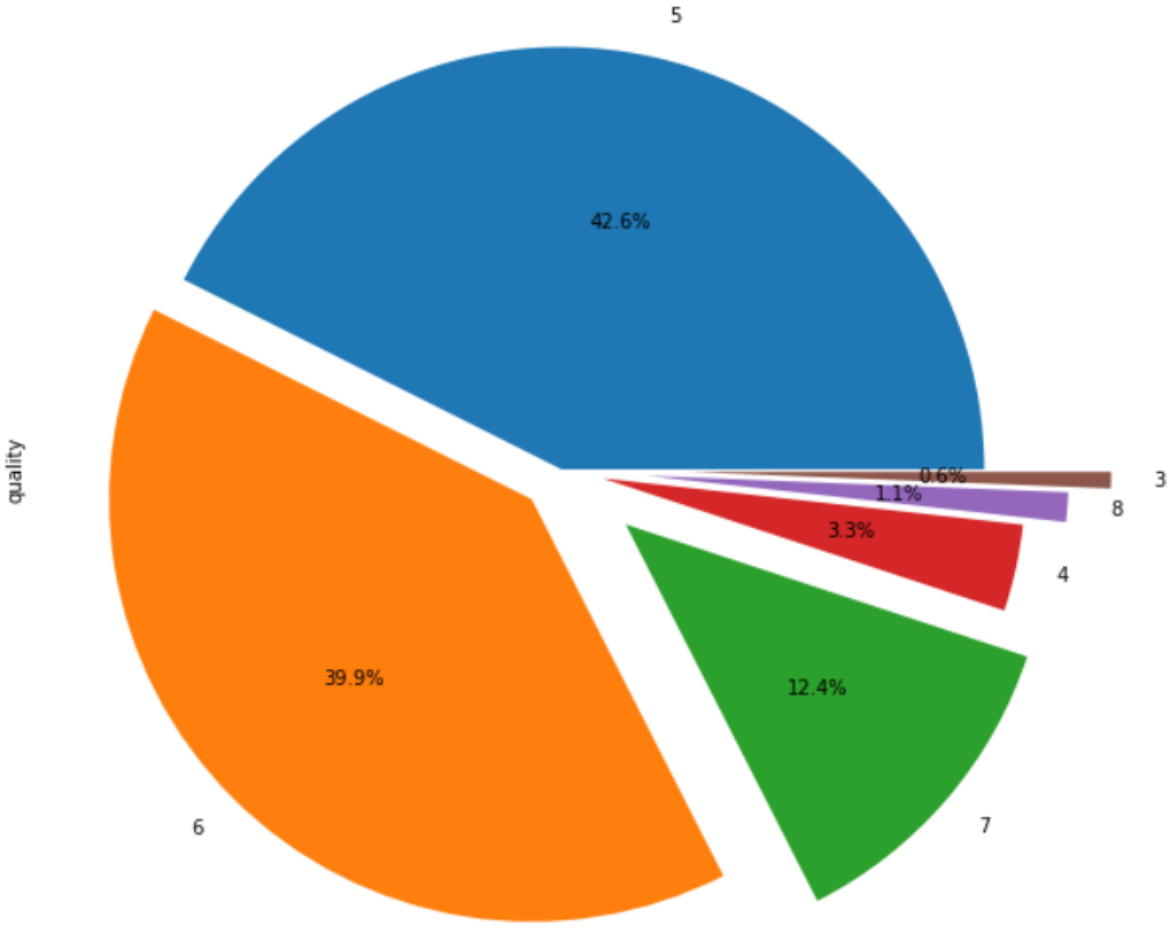
```
# Graph 2 - Plotting histogram for all columns of the dataset

df.hist(bins = 30, figsize=(15,15), color= 'orange');
```



Observations: The above graphs are histograms of each and every column of the dataset which shows the variation in distribution of the dataset for all the input variables. This will help us to understand the distribution of data.

```
# Graph 3 - Plotting a pie chart of count of unique values of quality

df['quality'].value_counts().plot.pie(autopct="%1.1f%%", explode=[0,0.1,0.2,0.1,0.2,0.3], figsize=(10,10))
```

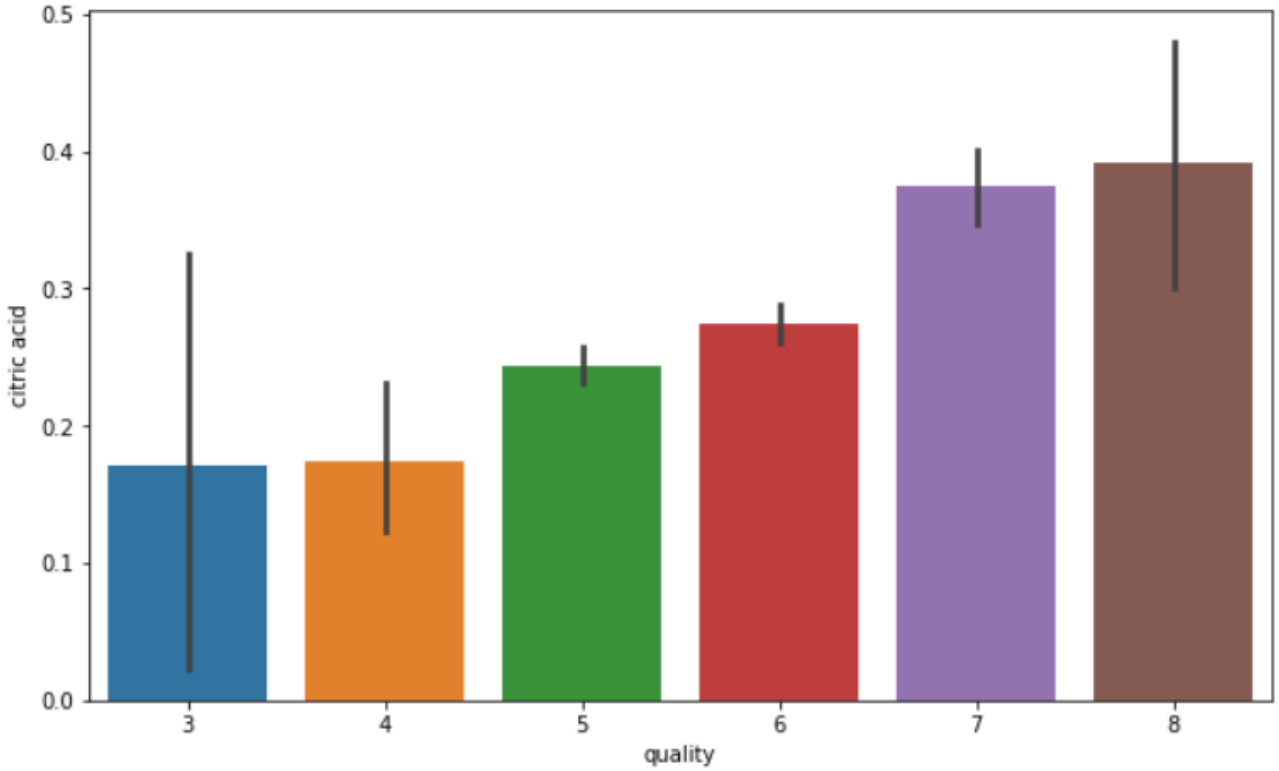
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a85478240>
```



Observations: The above pie chart is based on the count of unique values of quality column. It shows us the distribution of the dataset across the quality column. This will help to understand the percentage share of each type of quality of wine.

```
# Visualizations using Matplotlib and Seaborn

# Graph 4 - Plotting a bar plot of citric acid vs the quality of wine

fig = plt.figure(figsize=(10,6))
sns.barplot(x='quality', y='citric acid', data=df)
```

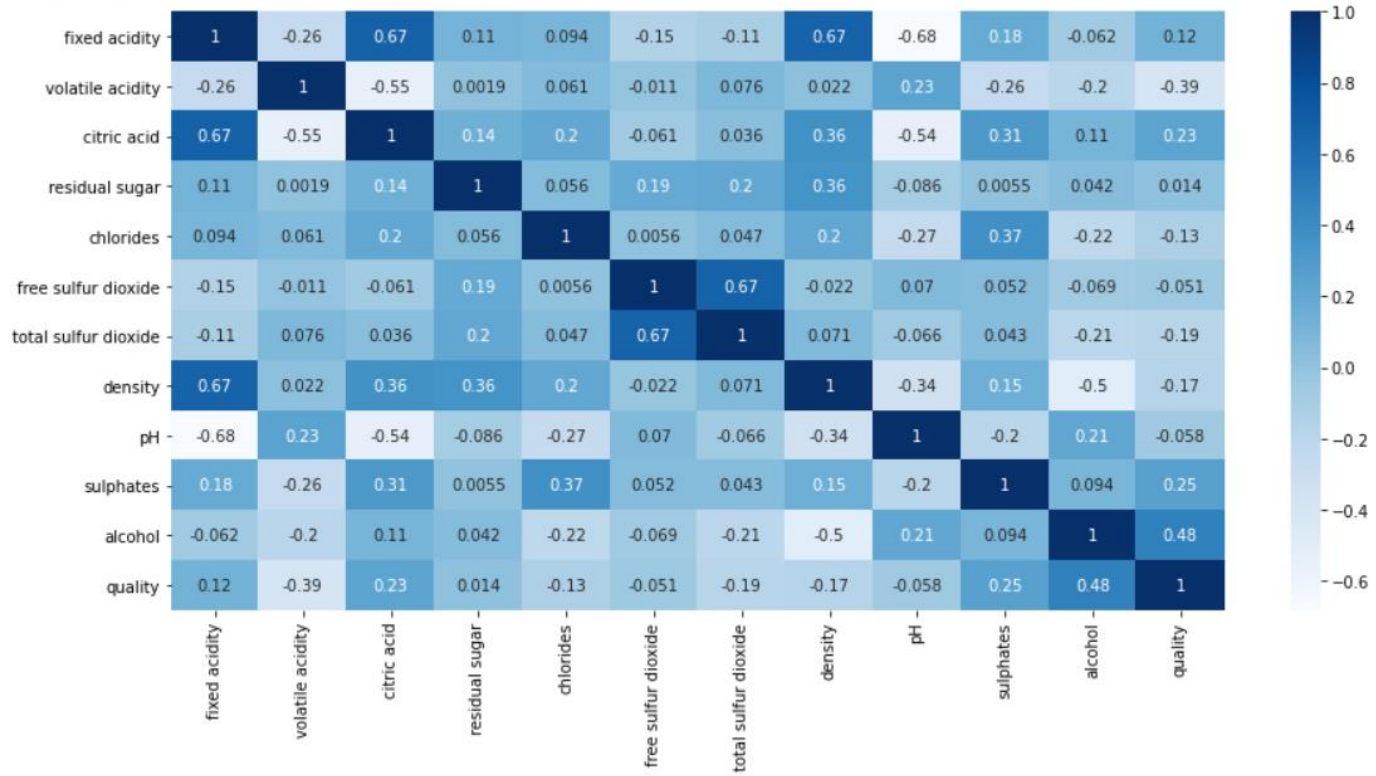
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a9056a7b8>
```



Observations: The above bar plot shows how quantity of citric acid varies the quality of the wine. With the help of the above graph, we can conclude that as the citric acid quantity in a wine increases, it increases its quality.

```
# Graph 5 - Plotting a heatmap for the correlation between the features of the dataset

fig = plt.figure(figsize=(15, 7))
sns.heatmap(data=df.corr(), annot=True, cmap="Blues") # df.corr() is the correlation between features for the dataframe
```

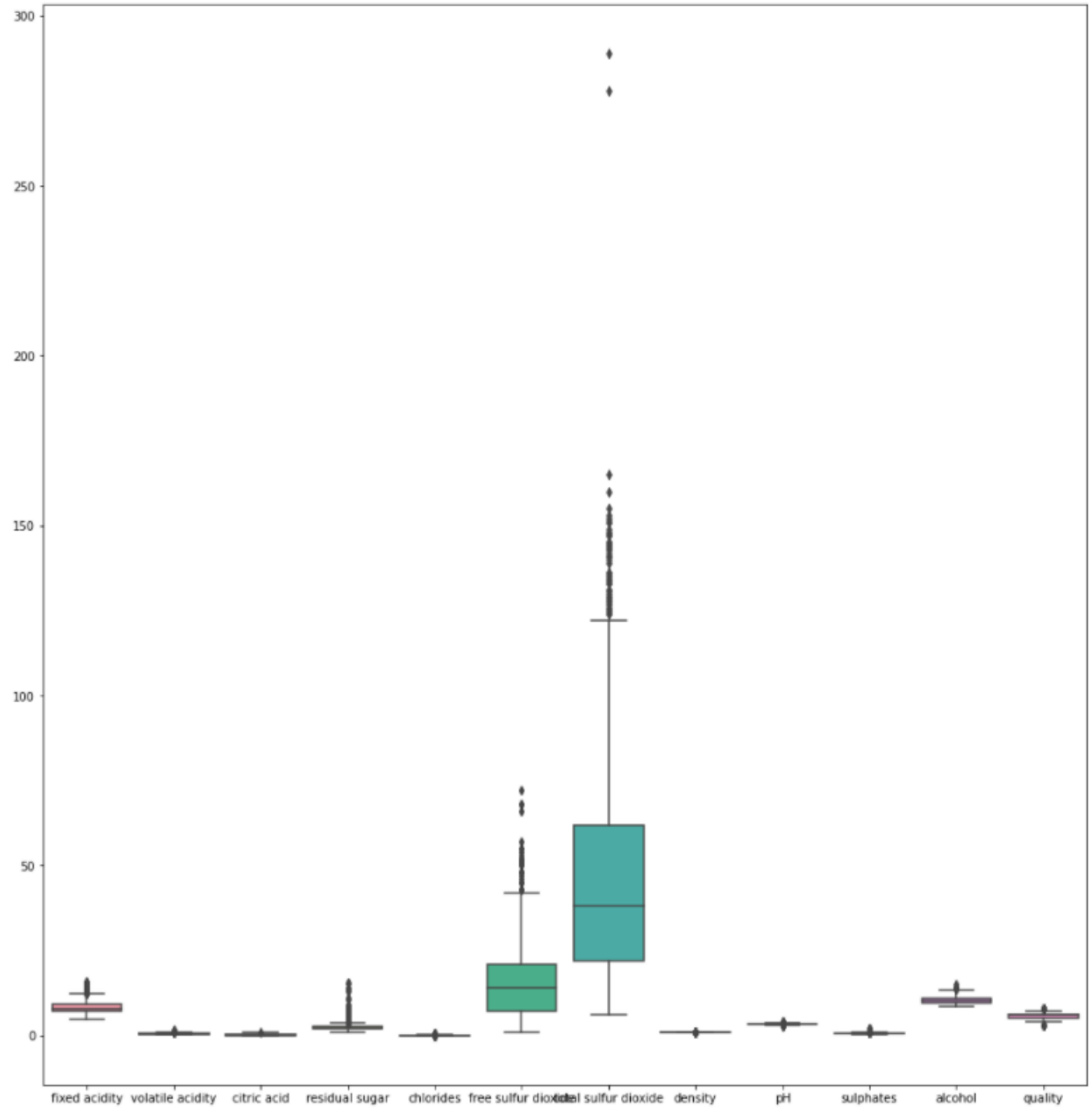
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a7c50a6d8>
```



Observations: The above heatmap plot indicated the relations between different features of the dataset. We can conclude that no 2 features are very closely related. However, we come to know that the quality of the wine depends greatly on the alcohol content in it (alcohol correlation is at 0.48 in the heatmap).

```
# Graph 6 - Plotting a boxplot for various features of the dataset to identify the outliers

plt.figure(figsize=(24,10))
sns.boxplot(data=df)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f8a7942f358>



Observations: The above boxplot indicates the distribution of various features. This will help us identify outliers and remove them during data filteration.

Conclusion:

Thus, we can see that libraries such as matplotlib, pandas and seaborn are really essential to help us interpret the data before creating a model for it. This will help us understand the type of data the model will be dealing with, the type of inputs and the expected outputs as well as correlation between various features of the dataset.