Exercise 8

EDA ON WINE DATASET

Aim:

To perform EDA on Wine Data Set.

Colab Link

https://colab.research.google.com/drive/12oe8dqaFgn3Yz4S8nbgW0aYkbXszK8kk?usp=sharing

import numpy as np

import matplotlib.pyplot as plt

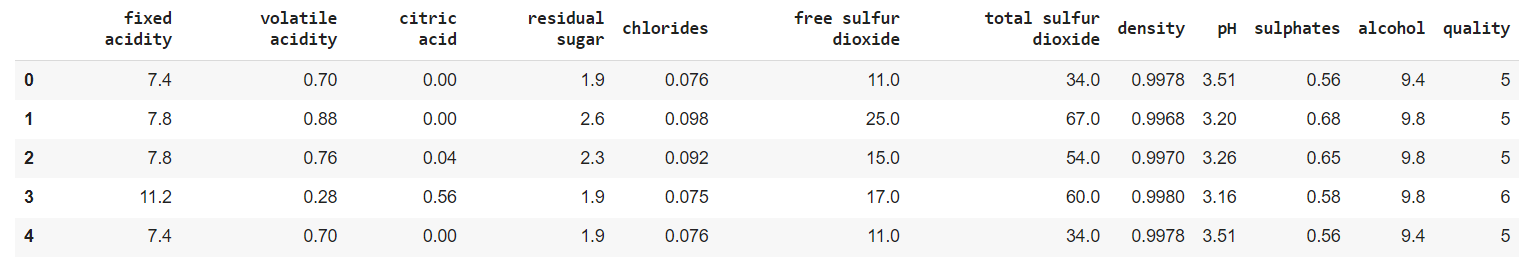
import pandas as pd

import seaborn as sns

%matplotlib inline

df=pd.read\_csv("/content/winequality-red.csv")

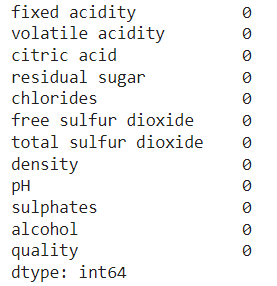
df.head()



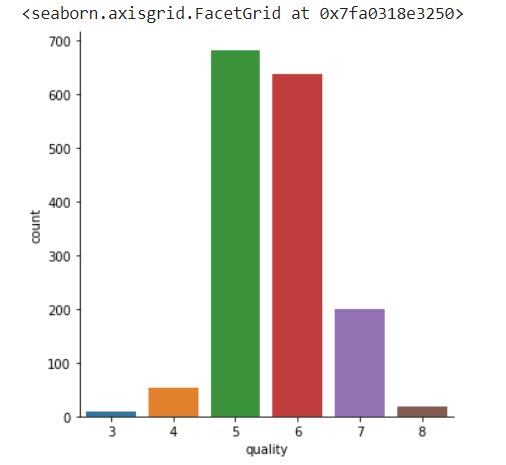
df.shape

df.columns

print(df.isna().sum())



sns.catplot(x='quality',data=df,kind='count')



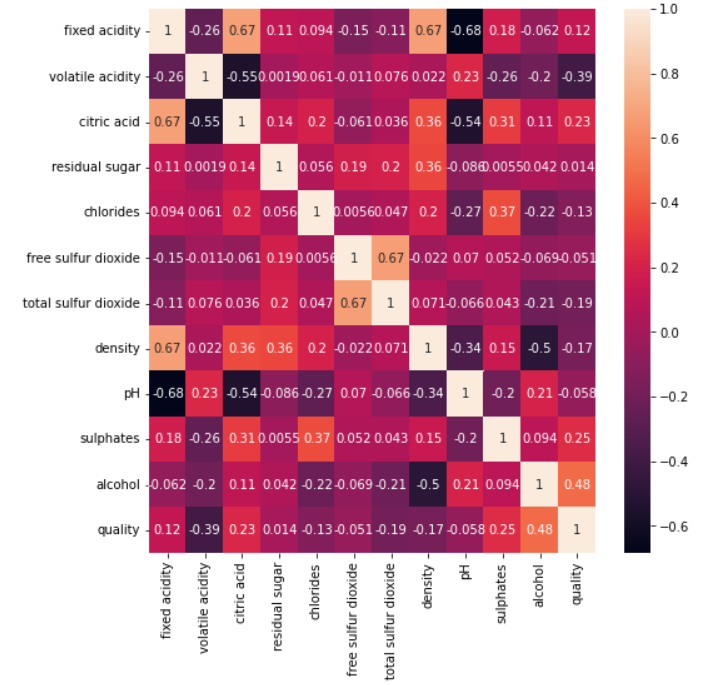
*CONCLUTION:*

*Quality has a high number of values in categories 5, 6 and 7.*

*Only a few observations are there for the categories 3 & 8*

.figure(figsize=(8,8))

sns.heatmap(df.corr(),color='rgb', annot=True)

**

*CONCLUTION:*

* density has a strong positive correlation with residual sugar and citric acid, whereas it has a strong negative correlation with fixed acidity.
* pH & fixed acidity has negative correlation.
* density & fixed acidity has positive correlation.
* citric acid & fixed acidity has positive correlation.
* citric acid & volatile acidity has negative correlation.
* free sulphur dioxide & total sulphur dioxide has positive correlation.

plt.figure(figsize=(10,12))

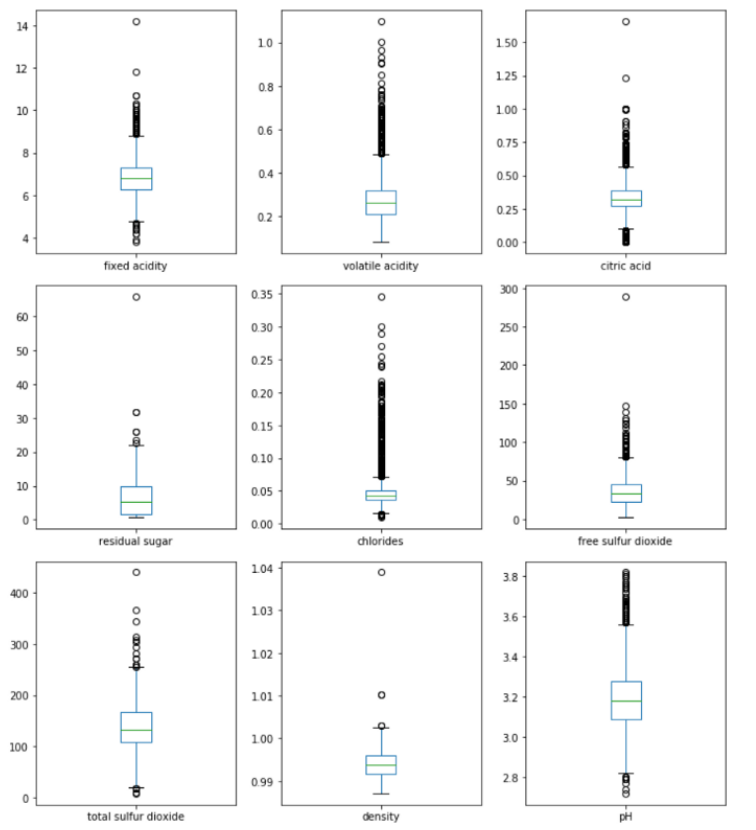
for i, col in enumerate(list(df.columns.values)):

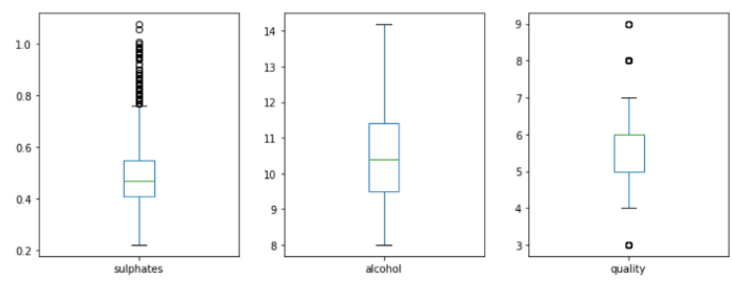
    plt.subplot(4,3,i+1)

 df.boxplot(col)

plt.grid()

plt.tight\_layout()





CONCLUTION:

Except for alcohol, all other features have outliers.

# CODE TO REMOVE OUTLIERS

def mod\_outlier(df):

        df1 = df.copy()

        df = df.\_get\_numeric\_data()

        q1 = df.quantile(0.25)

        q3 = df.quantile(0.75)

        iqr = q3 - q1

        lower\_bound = q1 -(1.5 \* iqr)

        upper\_bound = q3 +(1.5 \* iqr)

        for col in df.columns:

            for i in range(0,len(df[col])):

                if df[col][i] < lower\_bound[col]:

                    df[col][i] = lower\_bound[col]

                if df[col][i] > upper\_bound[col]:

                    df[col][i] = upper\_bound[col]

        for col in df.columns:

            df1[col] = df[col]

        return(df1)

df = mod\_outlier(df)

# create box plots

fig, ax = plt.subplots(ncols=5, nrows=3, figsize=(16,8))

index = 0

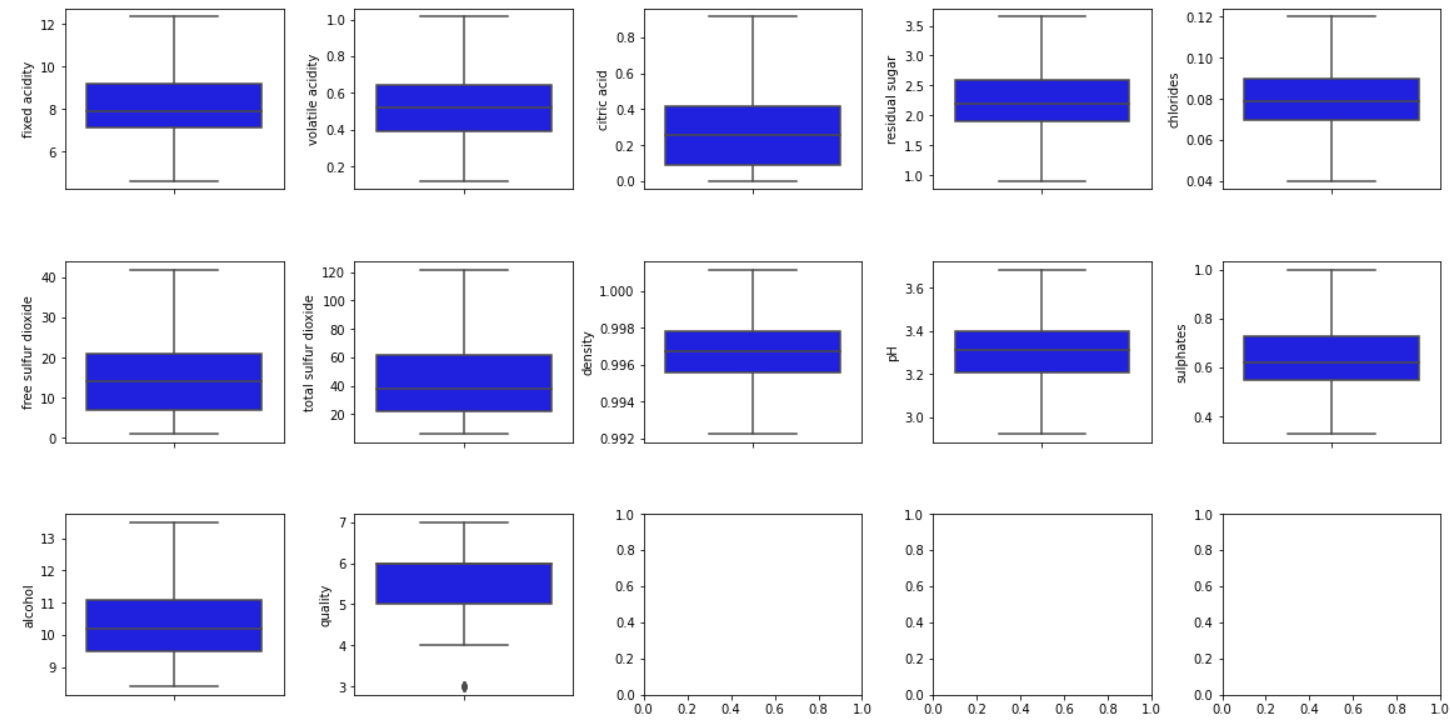
ax = ax.flatten()

for col, value in df.items():

    sns.boxplot(y=col, data=df, color='b', ax=ax[index])

    index += 1

plt.tight\_layout(pad=0.5, w\_pad=0.7, h\_pad=5.0)



CONCLUTION:

All the ouliers are removed.

plt.figure(figsize=(10,12))

for i, col in enumerate(list(df.columns.values)):

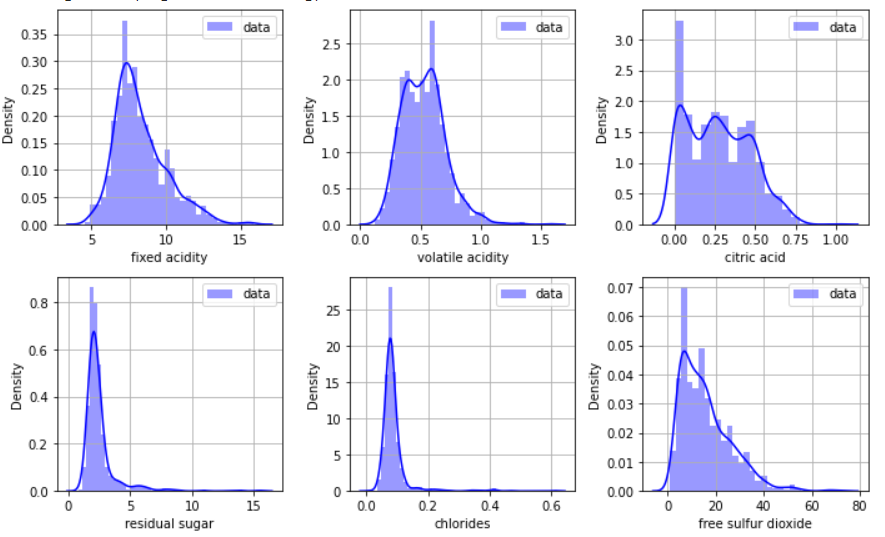
    plt.subplot(4,3,i+1)

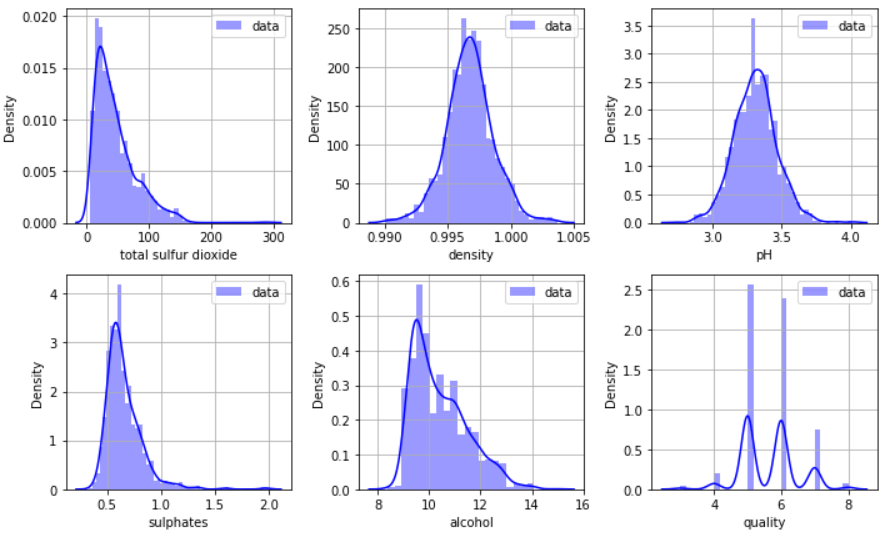
    sns.distplot(df[col],color='b',kde=True,label='data')

    plt.grid()

    plt.legend(loc='upper right')

    plt.tight\_layout()

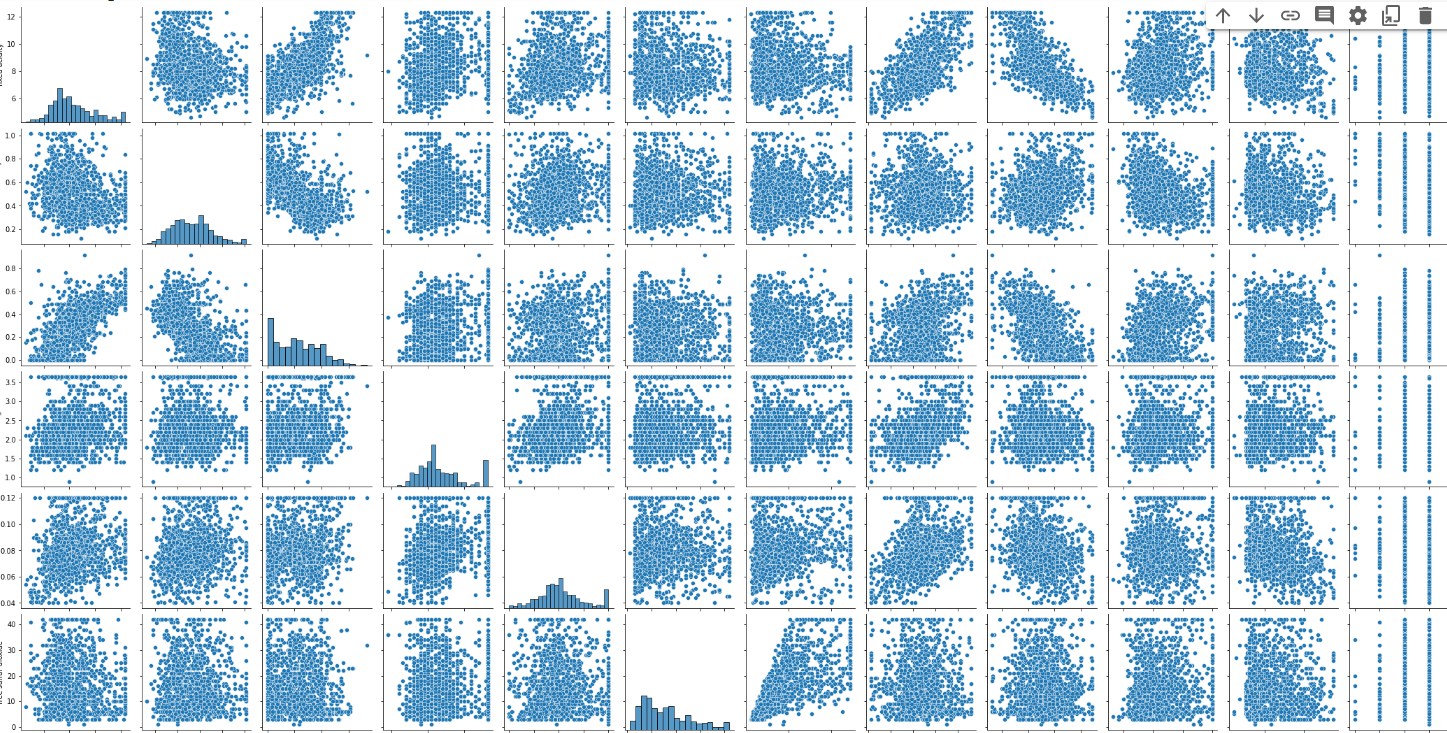


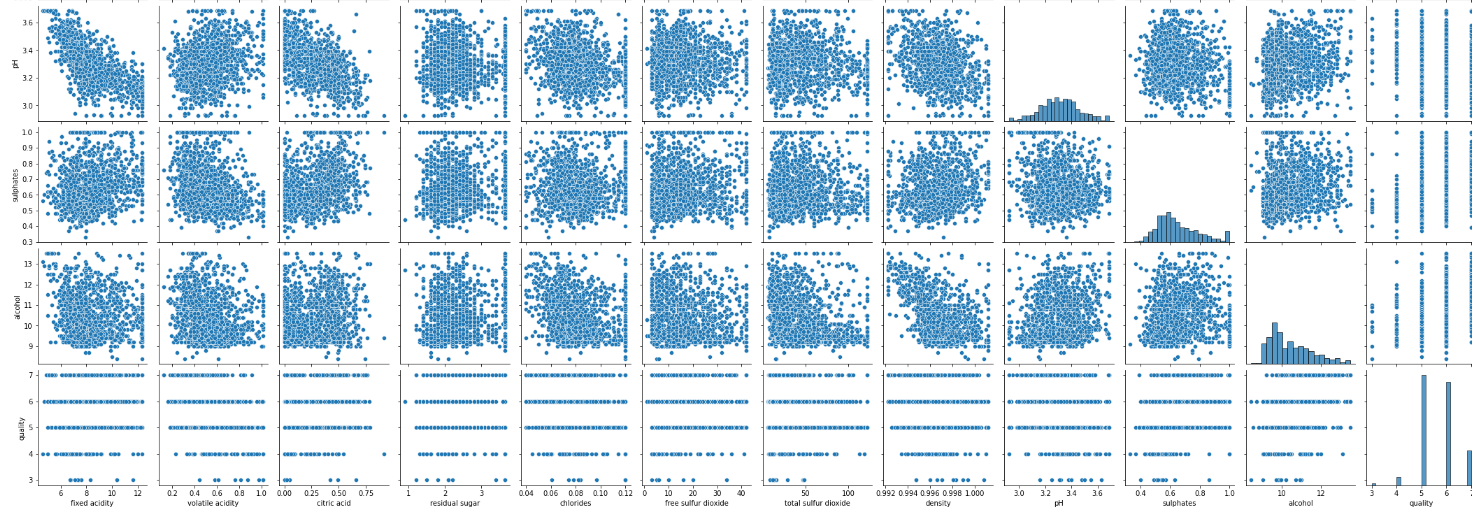


CONCLUTION:

* pH and density feature is approximately normally distributed.
* Remaining all independent features are positively skewed.

sns.pairplot(data=df,kind='scatter')

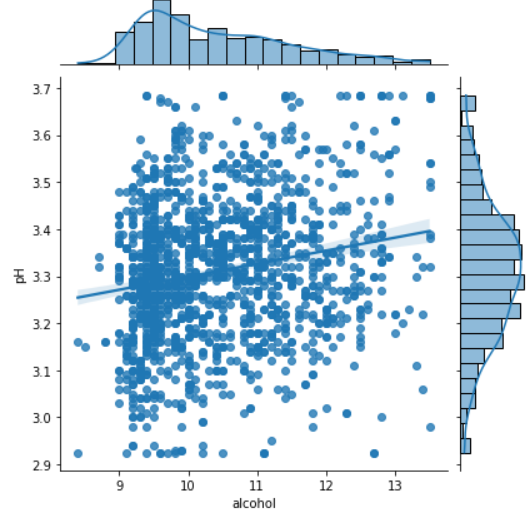




CONCLUTION:

Above pairplot shows the relationship among various features for all the possible pair of features.

sns.jointplot(x='alcohol',y='pH',data=df, kind='reg')



CONCLUTION:

Alcohol is weakly positively related to the pH values.

coor=round(cor(df),1)

ggcorrplot(corr,hc.order=TRUE,

           type="lower",

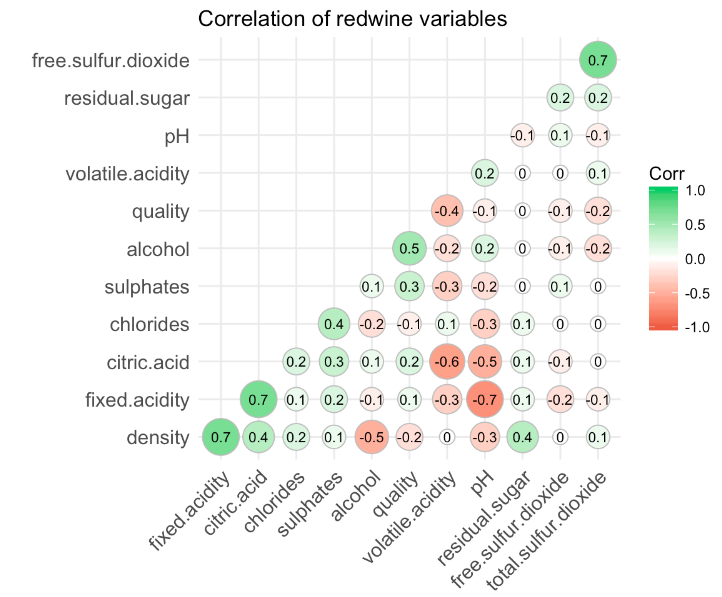
           lab-'TRUE',

           lab\_size=3,

           method='circle',

           colors=c("tomato2","white","springgreen3"),

           title="Correlation of redwine variables"



Insight:

1. Alcohol: the amount of alcohol in wine

2. Volatile acidity: are high acetic acid in wine which leads to an unpleasant vinegar taste

3. Sulphates: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant

4. Citric Acid: acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)

5. Total Sulfur Dioxide: is the amount of free + bound forms of SO2

6. Density: sweeter wines have a higher density

7. Chlorides: the amount of salt in the wine

8. Fixed acidity: are non-volatile acids that do not evaporate readily

9. pH: the level of acidity

10. Free Sulfur Dioxide: it prevents microbial growth and the oxidation of wine

11. Residual sugar: is the amount of sugar remaining after fermentation stops.

RESULT:

Thus EDA on wine Data Set was successfully performed.