Logistic Regression

**Logistic Regression**: Logistic Regression is a special case of Linear regression. Here probability of target variable is predicted, this variable is binary in nature I.e., it can be either 1(stands for success/yes) or 0(stands for failure/no).It is also known as sigmoid function where sigmoid function is

Sigmoid Function = 1/(1+e^-y)

(Here y is the dependent variable)

**Given Question:**

Develop and evaluate a logistic model to predict the quality (such as high quality and low quality) of red wines according to the several features and calculating test scores. Write your model in Python.

(Features that affect wine quality: *'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol'.)*

Please split 80% data for training and 20% data for testing.

Calculate the accuracy of the prediction.

**Given data:**

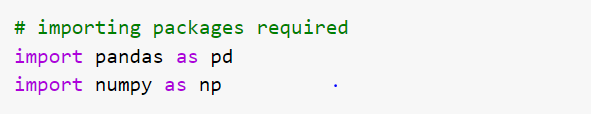
The given dataset is about the quality of wine along with the features that affect its quality. The features are as follows:

*(Fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol'.)*

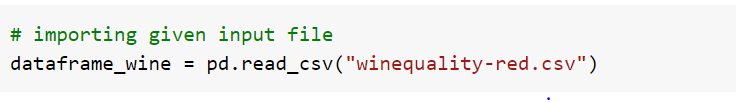
By considering above factors we are calculating the wine quality.

*https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009*

**Code:**

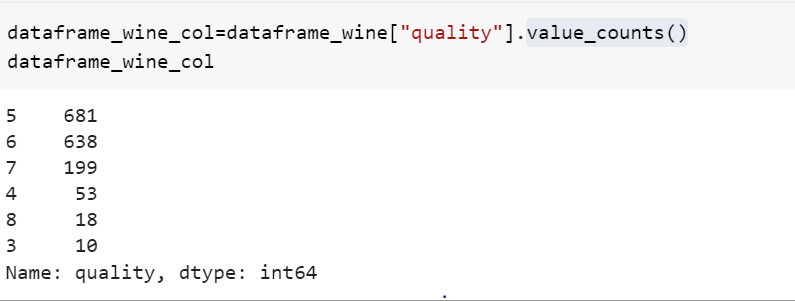


Here we are importing necessary packages.

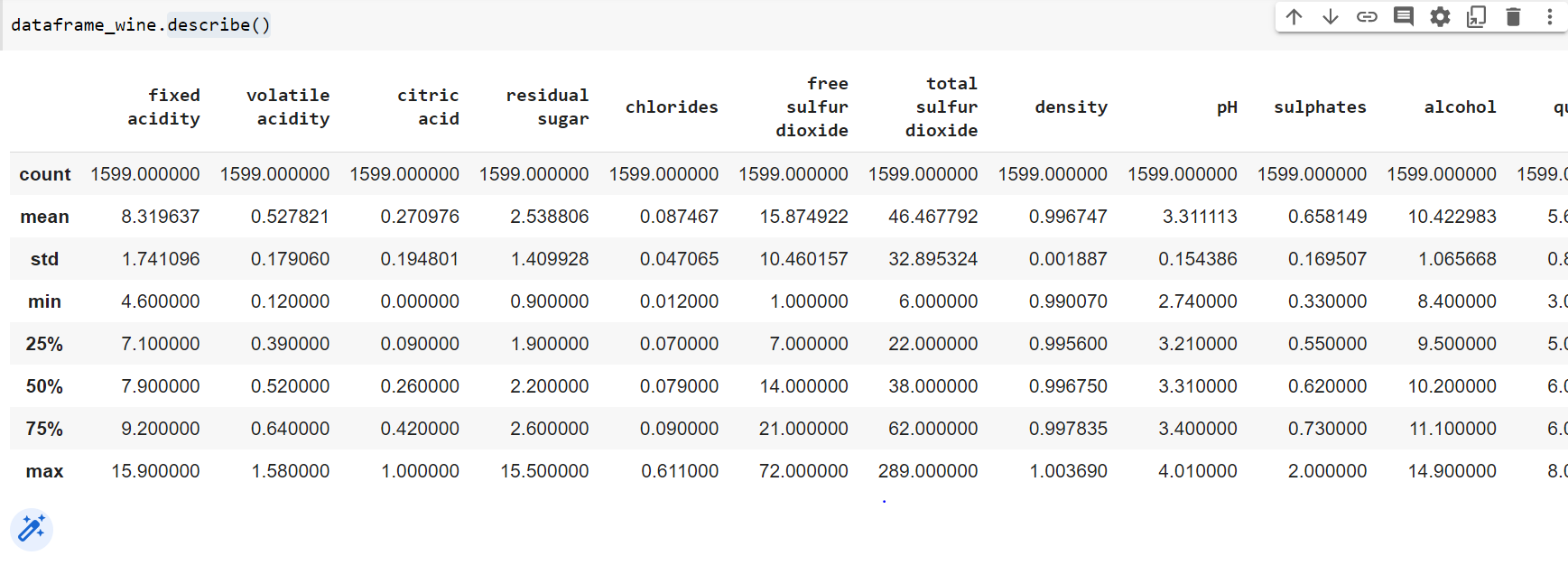


Given dataset is imported and is assigned to a variable(dataframe\_wine) to access the dataset in future.

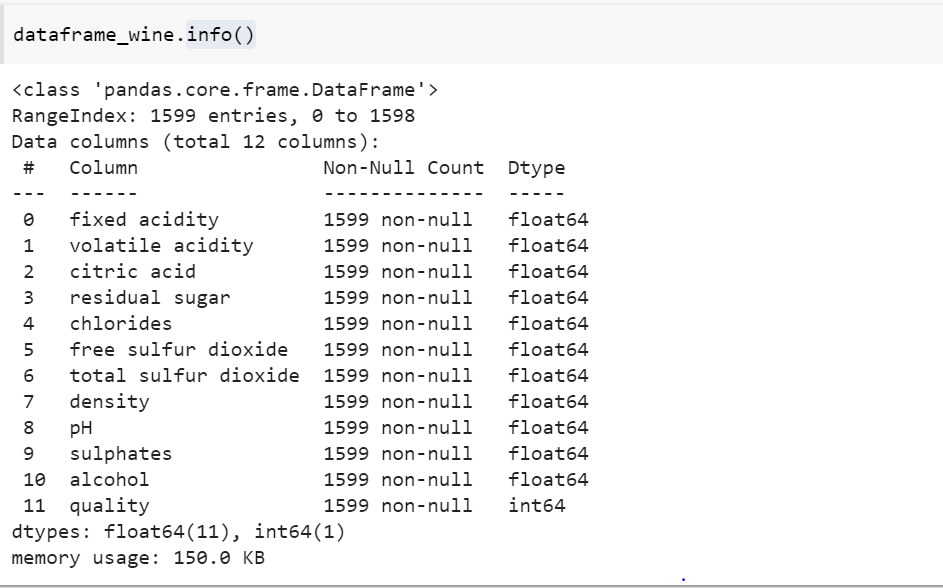
**Data Pre-Processing:**



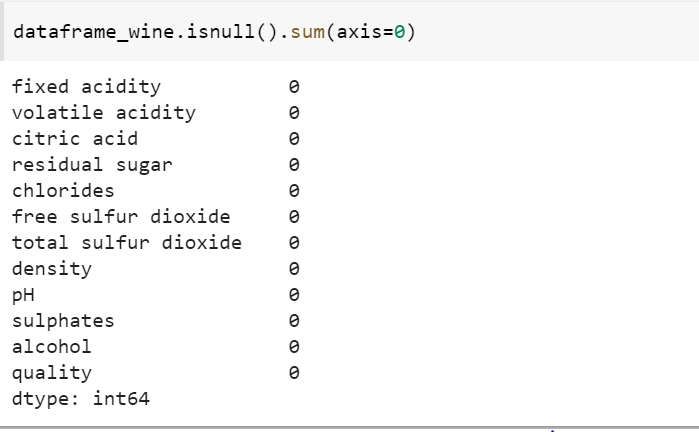
Here we are identifying and calculating the count of unique values present in the column “quality”.



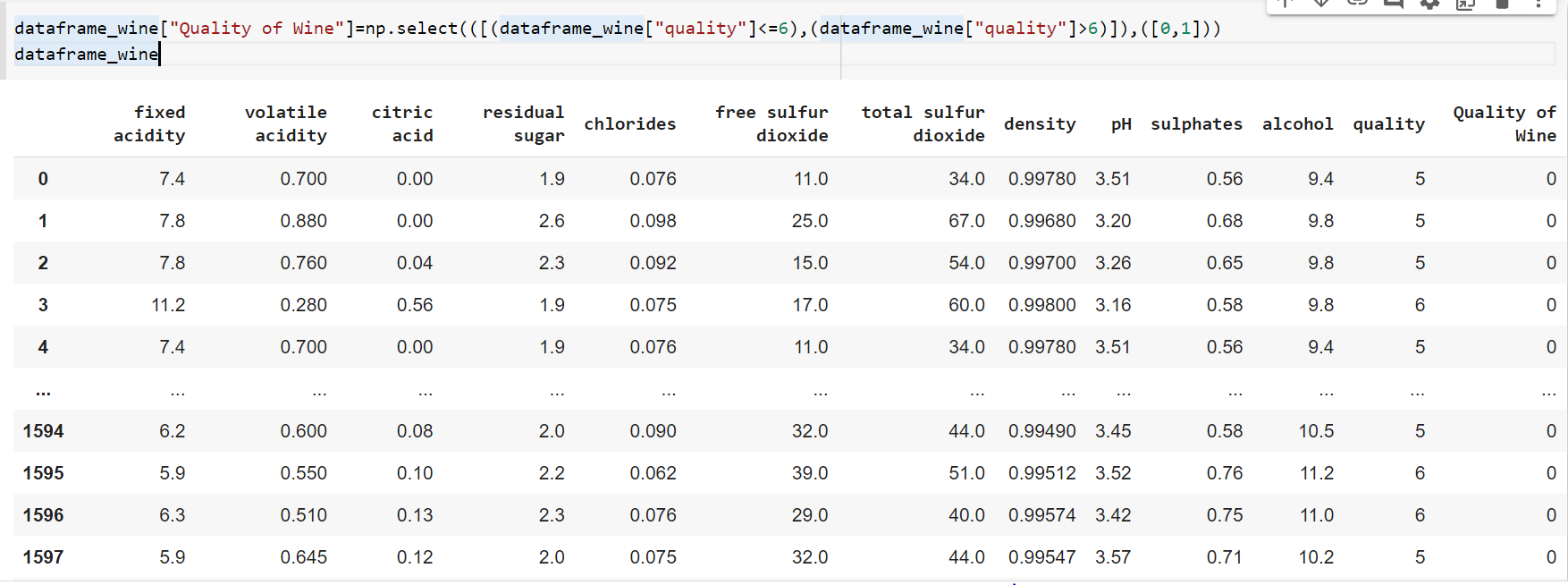
Here we are calculating and displaying the summary statistics of the given dataset.



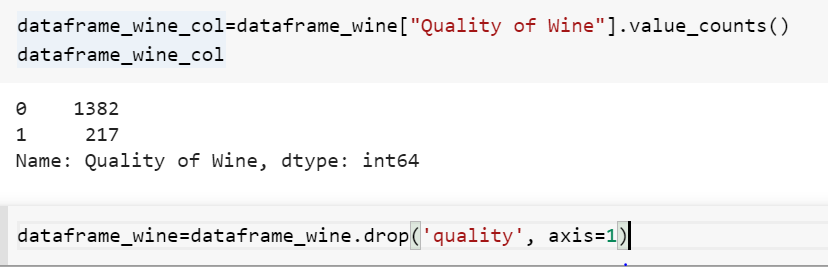
Here we are gathering the information about the dataset, there are 12 columns in which “quality” column is of int datatype and all the other columns have float as datatype.



Here we are checking if there are any null values in the given dataset.

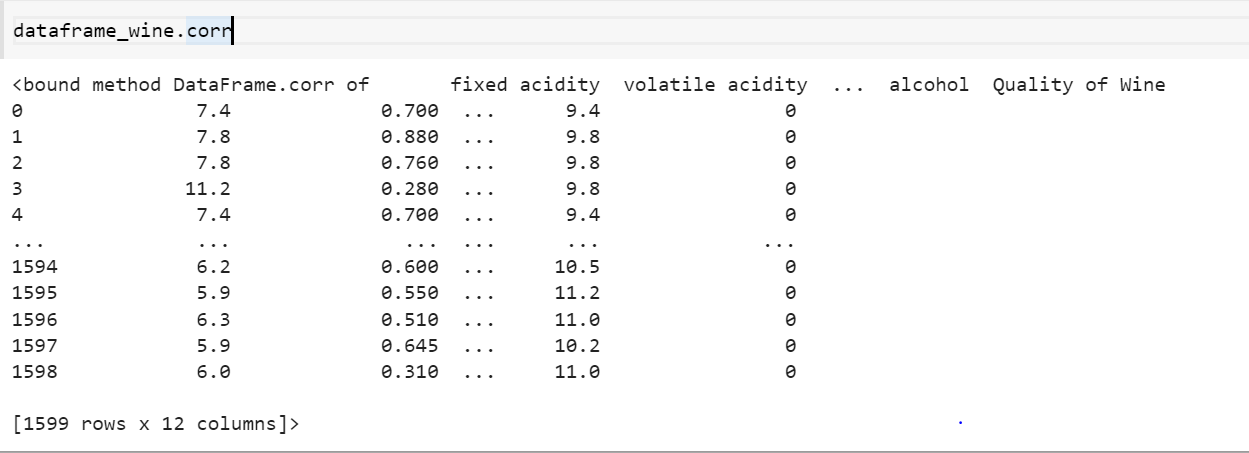


By using “quality” column, we are creating “quality of wine” where value quality less than or equal to 6 is given ‘0’ and value greater than 6 is given ‘1’.

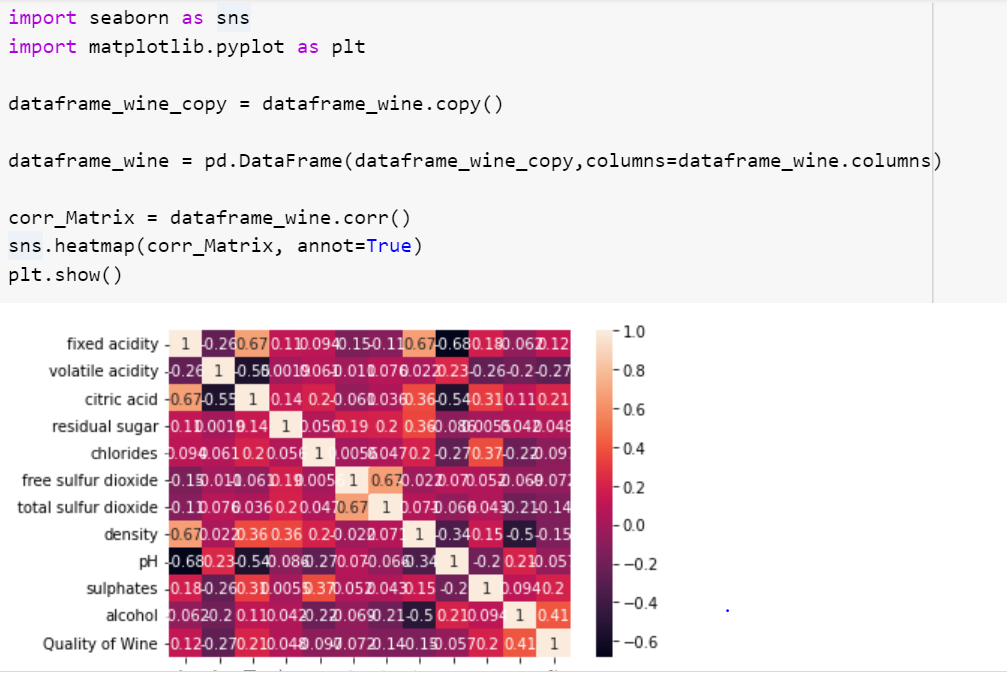


Calculating the count of unique values in the column “quality of wine” and dropping the “quality” column.

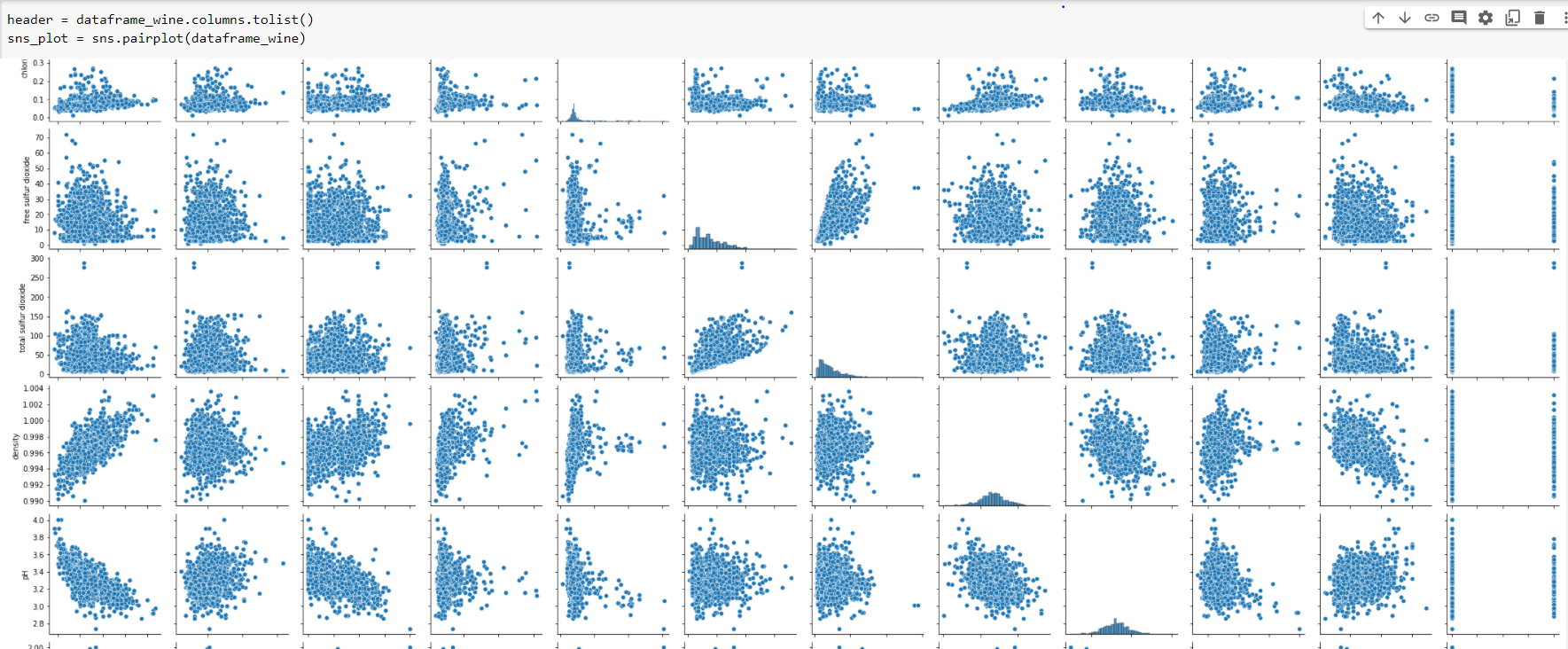
**Data Visualization:**



We are using .corr function to **find the pairwise correlation of all columns of given dataset.**



We are importing seaborn and matplotlib and we are plotting correlation matrix and heat map.



By using pairplot we are plotting pairwise relationships for the variables in each column in the given dataset.

**Splitting the data**

The given data is splitted into train dataset & test data by using sklearn package, with train\_test\_split function.

Splitted data with test size as 20%, which result in train size is 80%.

The shape of train and test data printed below.

Text

Description automatically generated

**Applying the model**

Applying the logistic regression, which is the model using to the data.

Fitting the regression with the training data set & the test data set.

The predicted values are with the test data & printed the predicted data set.

The cross validation of the data with logistic regression is tested, which in results

cross\_val\_result :-

{'fit\_time': array([0.5284667 , 0.35993075, 0.28390169, 0.25460076, 0.11091661]), 'score\_time': array([0.00239897, 0.00241494, 0.00158501, 0.00156856, 0.00156093]), 'test\_score': array([0.86875 , 0.84375 , 0.88125 , 0.846875 , 0.87774295])}

A picture containing graphical user interface

Description automatically generated

**Prediction results and test score**

printing the accuracy score with the accuracy\_score function from sklearn, the accuracy of data is 87%.

Graphical user interface, text

Description automatically generated with medium confidence

**Metrics**

The confusion matrix is [[268 8],[33 11]]

The recall score is 0.25 & other metrics are below.

Text

Description automatically generated

**Conclusion**

A screenshot of a computer

Description automatically generated with medium confidence

The accuracy of the predicted data is 87%, for the red wine

The precision for 0 is 0.89 & 1 is 0.58

The recall score for 0 is 0.97 & 1 is 0.25

The f1-score for is 0.93 & 1 is 0.35