**Predicting the Quality of Red Wine**

BY JUDELYN A. SALAMANCA

**Contents**

1. **Introduction**
   1. **Dataset**
   2. **Input Variables**
2. **Objective**
3. **Data Preparation**
   1. **Importing Libraries**
   2. **Reading Data**
   3. **Understanding Data**
   4. **Cleaning Data**
4. **Data Visualization**
   1. **Exploring Variables**
5. **Preparing Data for Modelling**
   1. **Convert to a Classification Problem**
   2. **Proportion of Good vs. Bad Wines**
   3. **Split Data**
6. **Modelling**
   1. **Model 1: KNeighborsClassifier**
   2. **Model 2: DecisionTreeClassifier**
   3. **Model 3: RandomForestClassifier**
7. **Conclusion**

# 1. Introduction

As the quarantine continues, I’ve picked up a number of hobbies and interests… including WINE. Recently, during Christmas I’ve acquired a taste for wines, although I don’t really know what makes a good wine. Therefore, I decided to apply some machine learning models to figure out what makes a good quality wine!

# 1.1 About Dataset

For this project, I used [Kaggle’s Red Wine Quality](https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009) dataset to build various classification models to predict whether a particular red wine is “good quality” or not. Each wine in this dataset is given a “quality” score between 0 and 10. For the purpose of this project, I converted the output to a binary output where each wine is either “good quality” (a score of 7 or higher) or not (a score below 7). The quality of a wine is determined by 11 input variables:

# 1.2 Input Variables

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

**Output variable (based on sensory data):**

1. quality (score between 0 and 10)

Let’s get the data and convert them to a format of a data frame for making manipulation with their easier. Data investigation is an interesting and addictive task. Take a look at your data, check the dimensionality and type of them.

**2. Objectives**

The objectives of this project are as follows

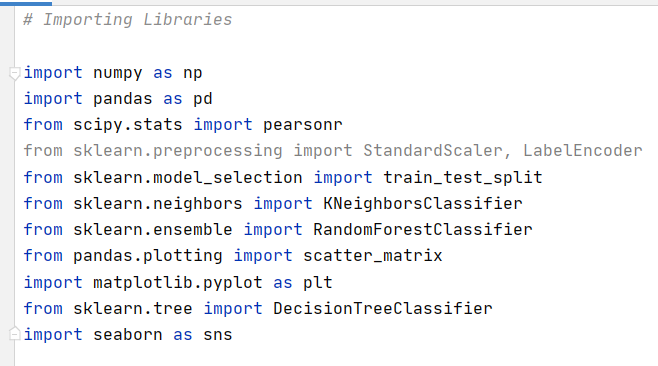
1. To experiment with different classification methods to see which yields the highest accuracy
2. To determine which features are the most indicative of a good quality wine

**3. Data Preparation**

First, I imported all of the relevant libraries that I’ll be using as well as the data itself.

So far, we have only taken a quick glance at the data to get a general understanding of the kind of data we are manipulating. Now we would be exploring data using a test set.

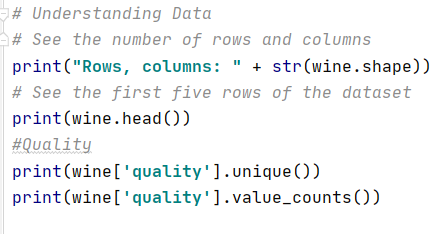
It would be interesting to take a look at the basic statistical characteristics of each numerical feature. The count, mean, min, and max rows are self-explanatory. The std row shows the standard deviation (which measures how dispersed the values are). The 25%, 50%, and 75% of rows show the corresponding percentiles.

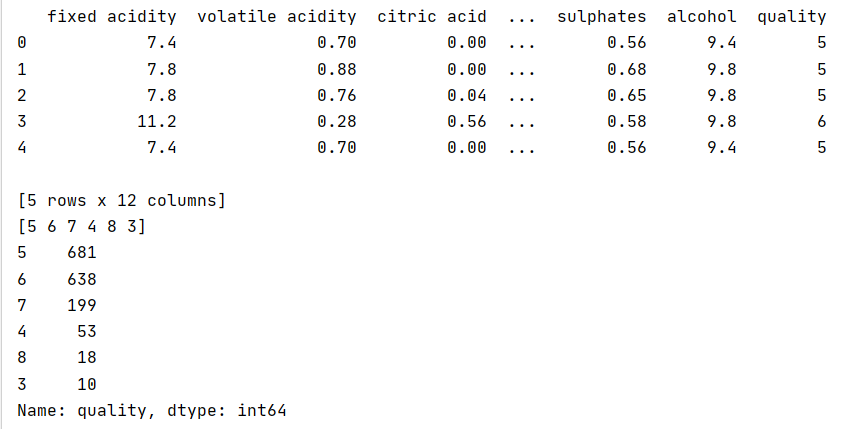
**3. 1. Importing Libraries**

**3. 2. Reading Data**

**3. 3. Understanding Data**

Next, I wanted to get a better idea of what I was working with.

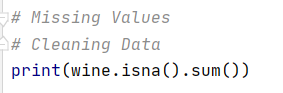
[In]

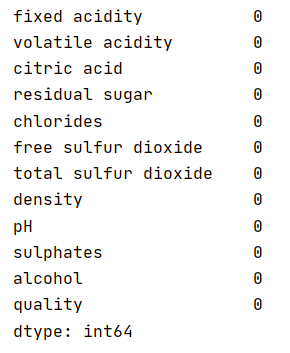
 [Out]

**3. 4. Cleaning Data**

Check missing values.

[In]



[Out]

This is a very beginner-friendly dataset. I did not have to deal with any missing values, and there isn’t much flexibility to conduct some feature engineering given these variables. Next, I wanted to explore my data a little bit more.

**4. Data Visualization**

Visualizing data is crucial for recognizing underlying patterns to exploit in the model. When the data visualizing properly it’s clear to see trends and patterns, the correlation between variables, because our brains are very good at spotting patterns on pictures. Let’s play around with different types of data visualization, parameters to make the patterns stand out.

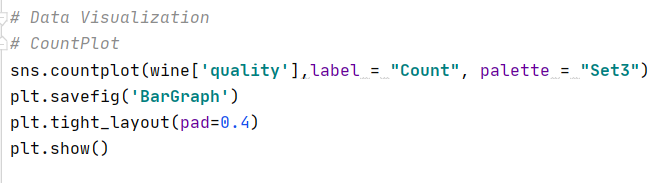
Using the following plots, we may understand data distribution for separate attributes; **for example, data distribution for attribute “alcohol” is positively skewed, for attribute “**density**” data quite normally distributed. Take attention to the wine quality data distribution. It’s a bimodal distribution and there are more wines with average quality than wines with ‘good’ or ‘bad’ quality.**

**4.1 Exploring Variables**

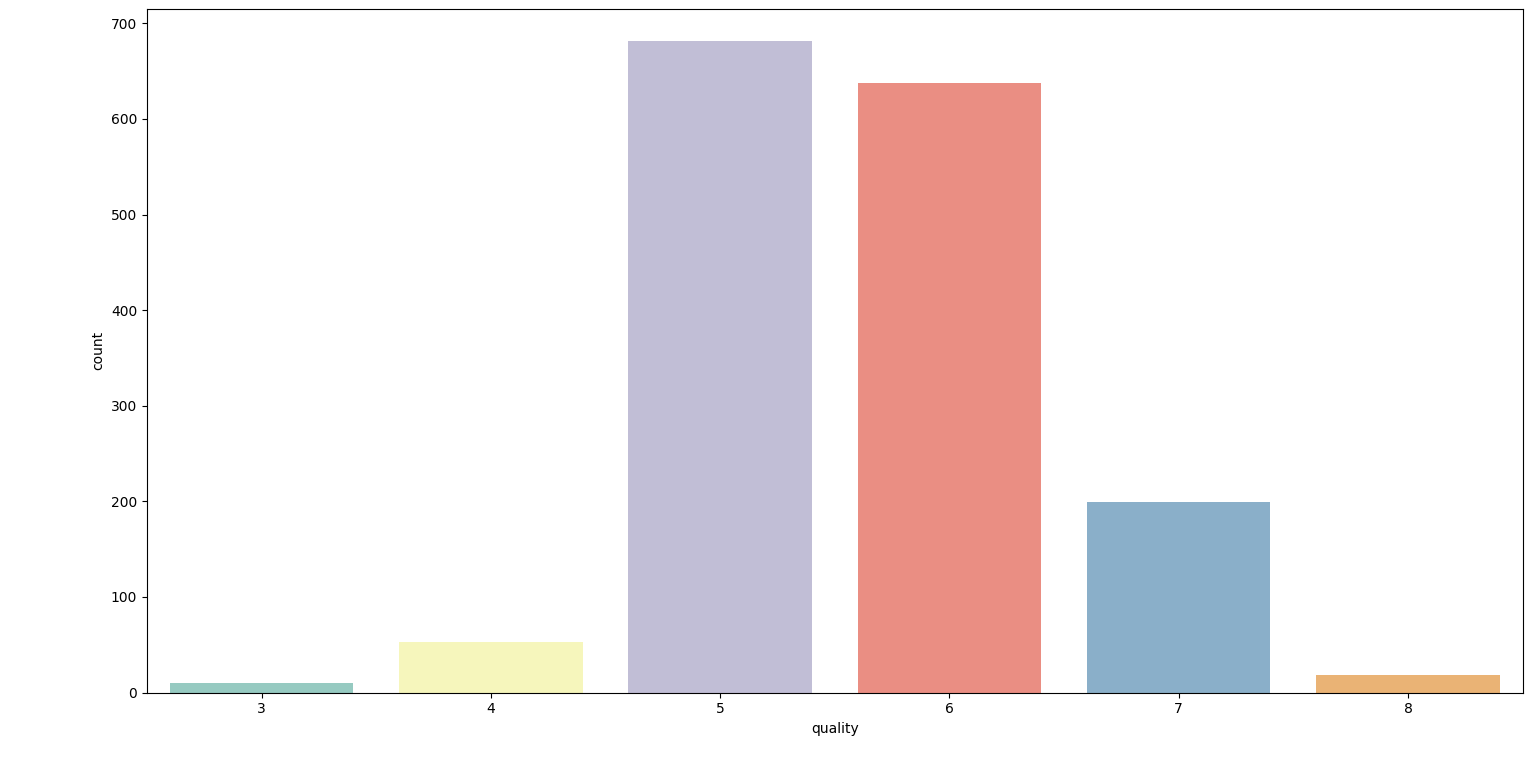
## **Count Plot of ‘quality’ variable**

First, I wanted to see the distribution of the quality variable. I wanted to make sure that I had enough ‘good quality’ wines in my dataset, you’ll see later how I defined ‘good quality’.

[In]

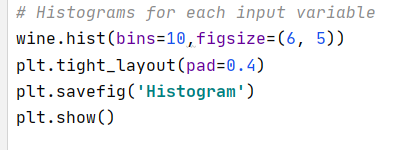


[Out]

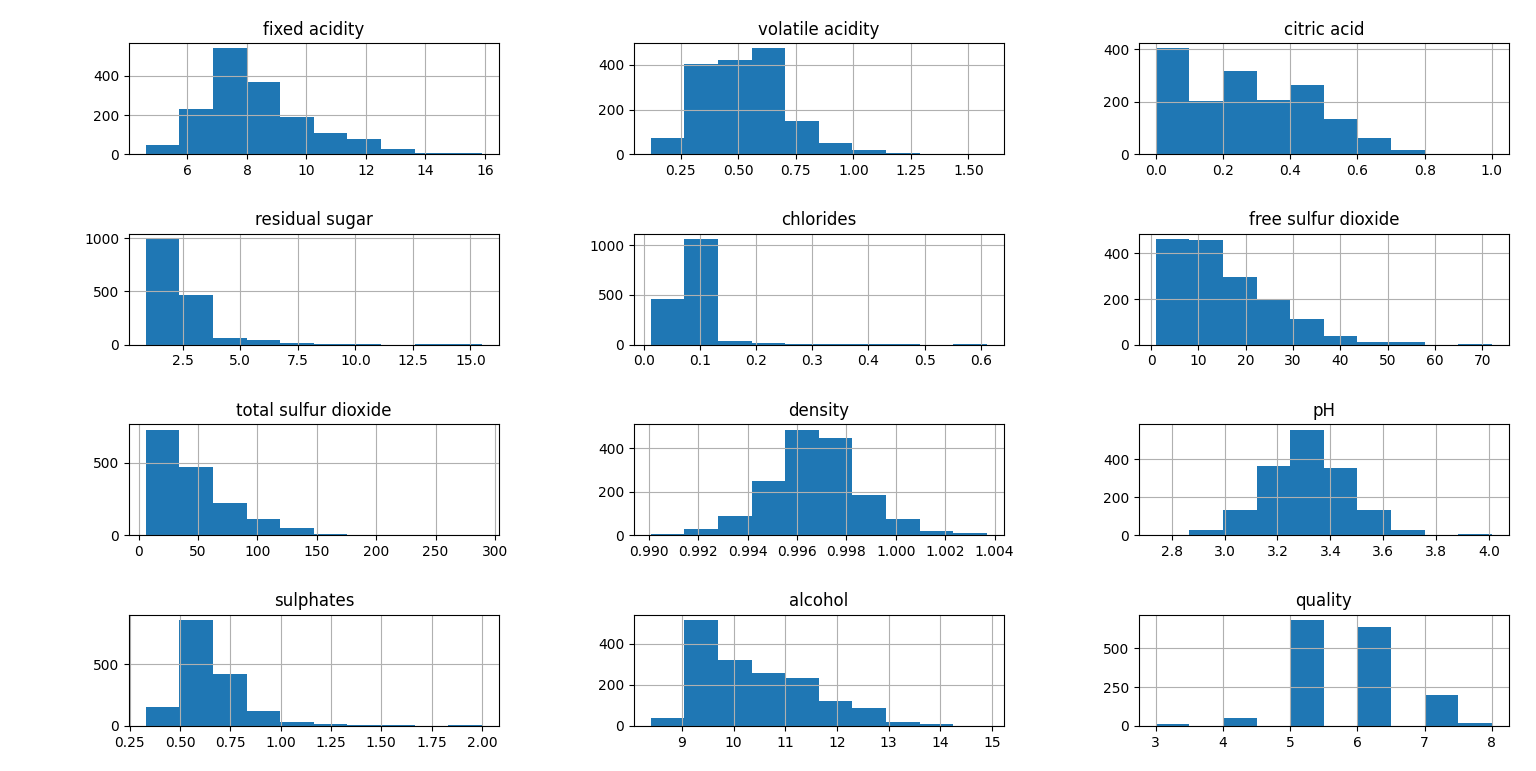


## **Histogram for each input variable**

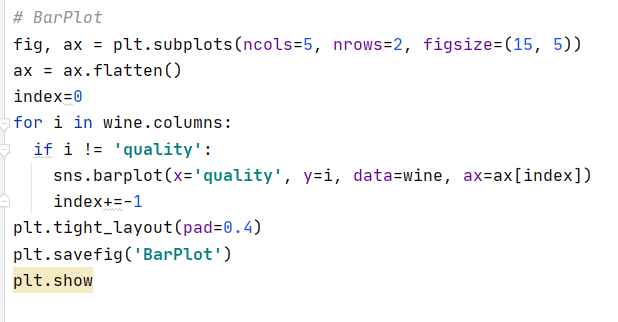
[In]

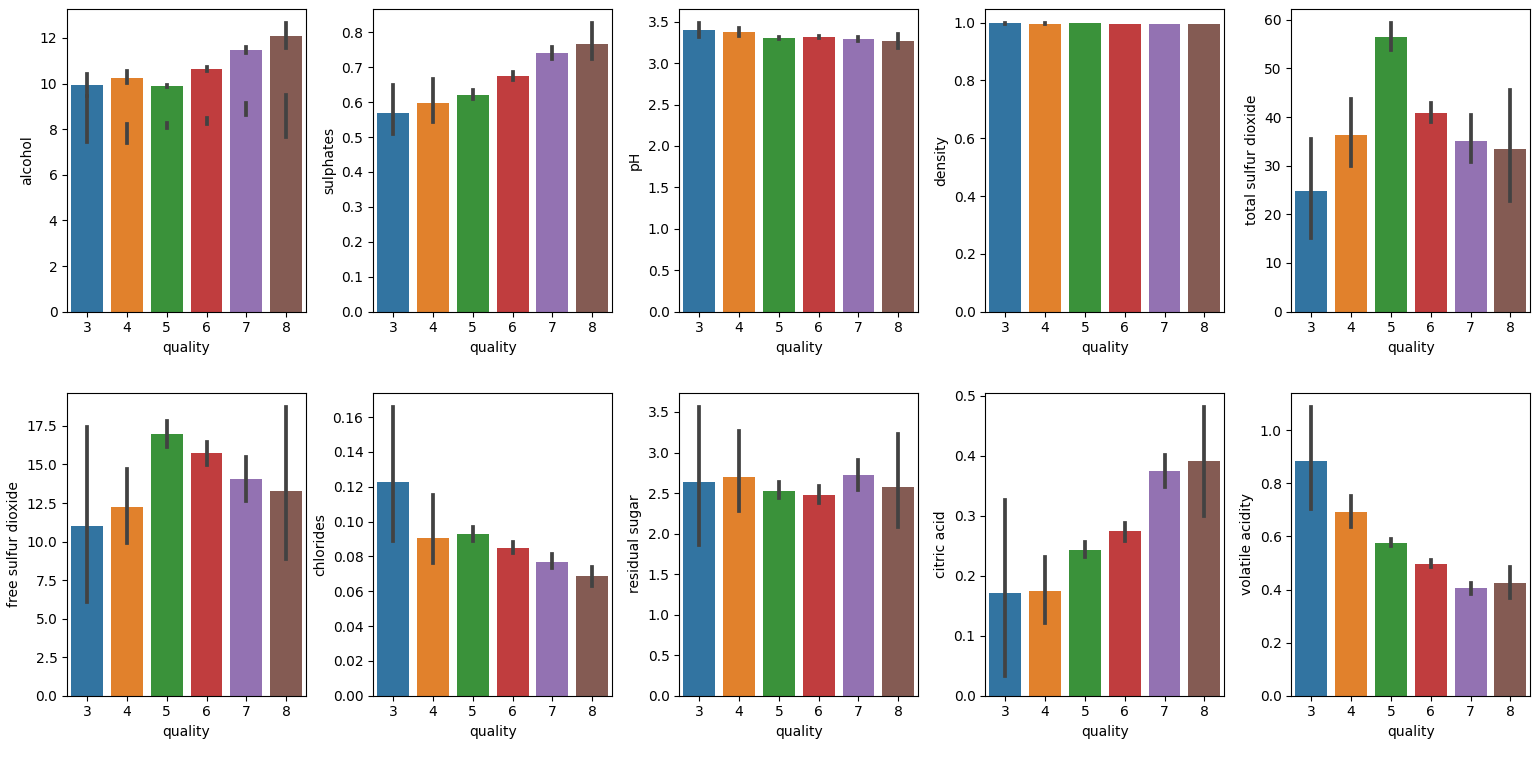


[Out]

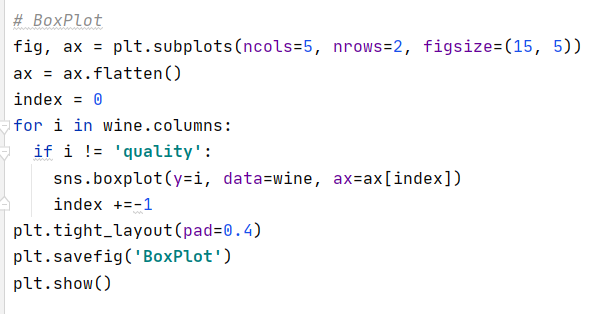


## **Bar Plot**

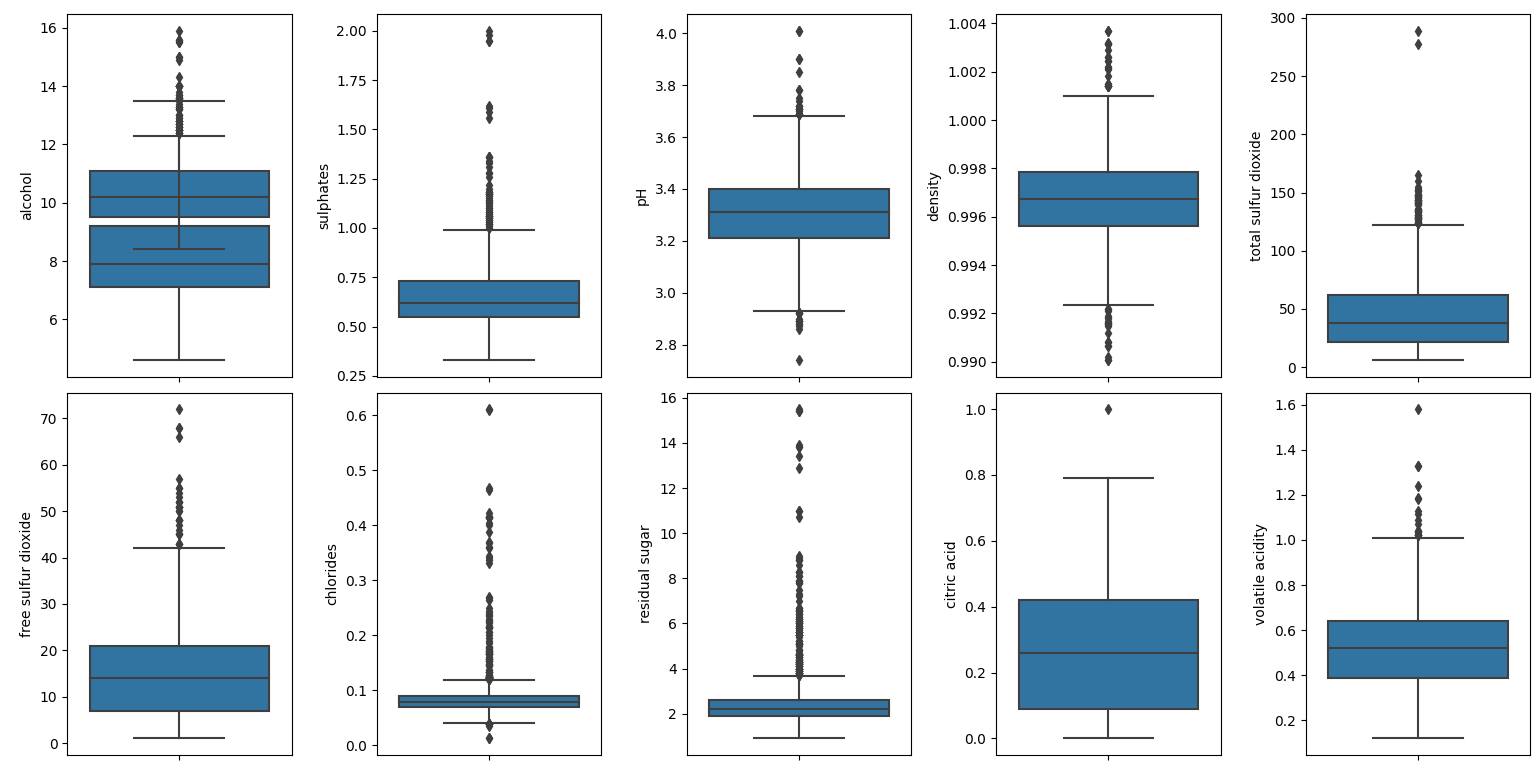
[In]

[Out]

## **Box Plot**

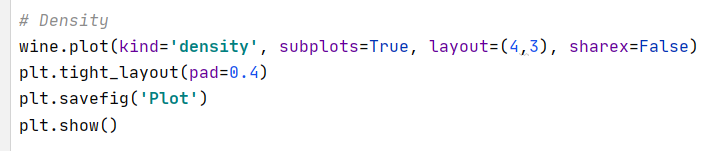
[In]

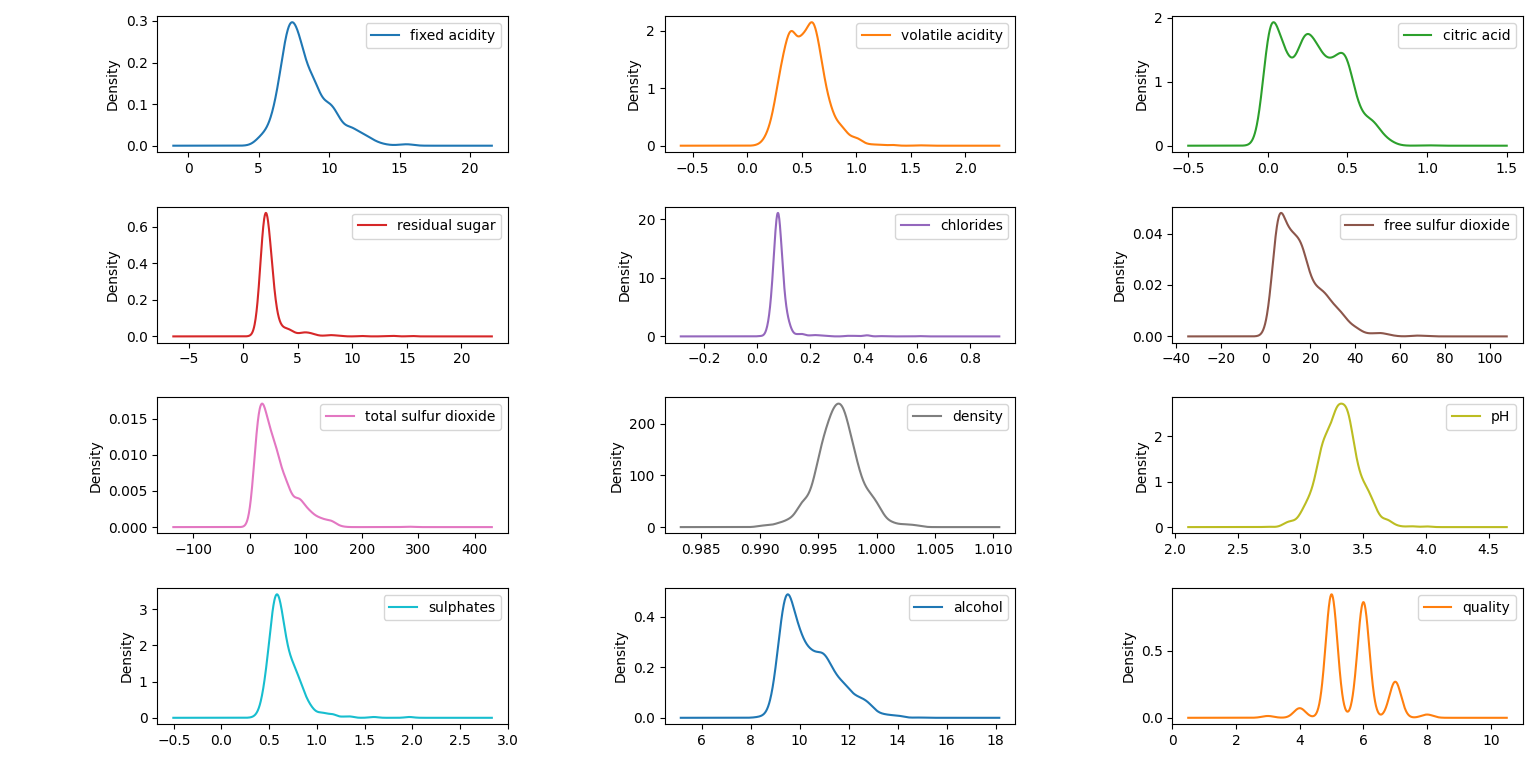
[Out]



## **Density**

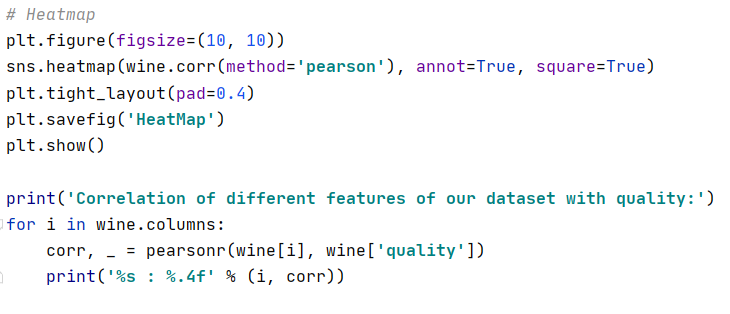
[In]

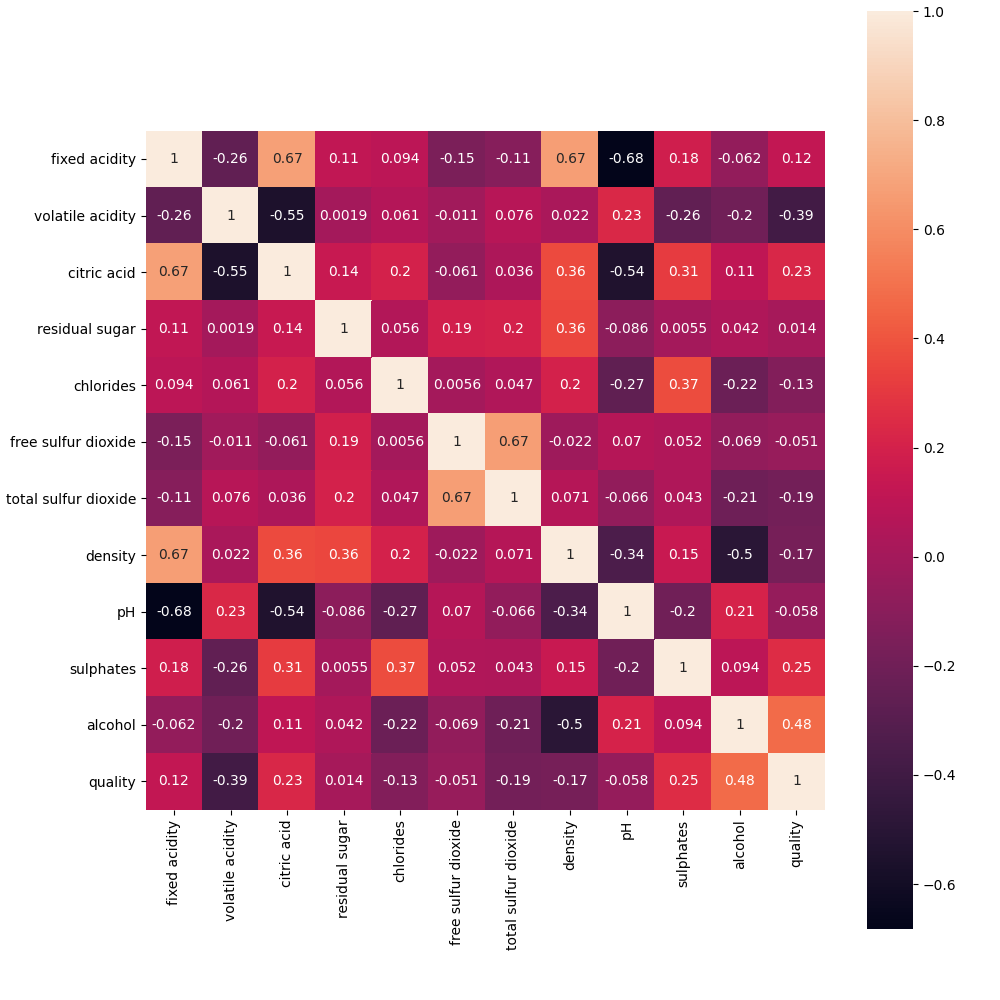


[Out]

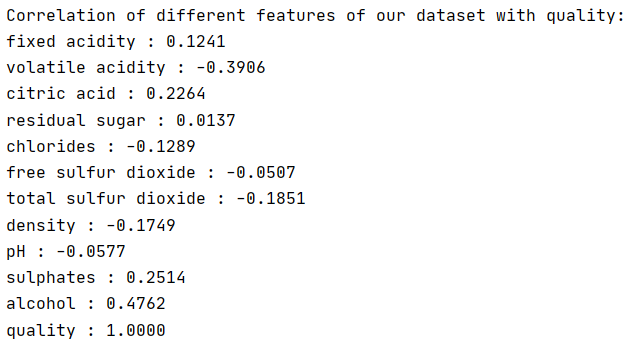
## **Heat Map**

For understanding how much each attribute correlates with the quality score of wine compute the standard correlation coefficient (also called Pearson’s r) between every pair of attributes.

[In]



[Out]

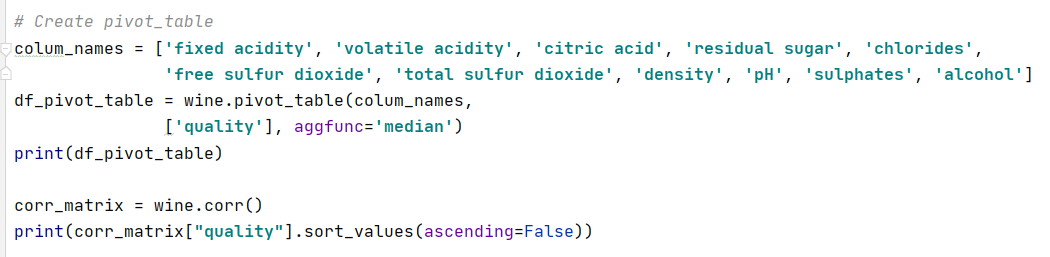


The correlation coefficient ranges from –1 to 1. When it is close to 1, it means that there is a strong positive correlation; for example, **the ‘quality’ value tends to go up when the ‘alcohol’ goes up**. When the coefficient is close to –1, it means that there is a strong negative correlation; **you can see a small negative correlation between the ‘volatile acidity’ and the ‘quality’ value**. Finally, coefficients close to zero mean that there is no linear correlation.

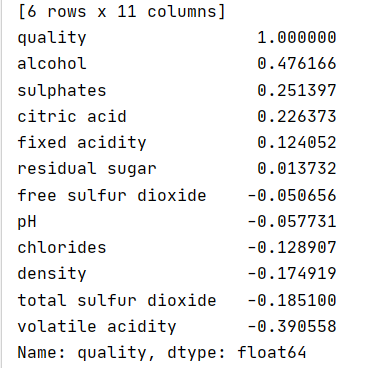
You could see more detailed information about data correlation using the correlation matrix. The correlation matrix gives us information about how the two variables interact, both the direction and magnitude.

## **Pivot Table**

Another way exploring data is an incredibly handy tool pivot table. A pivot table is a summary of your data, packaged in a chart that lets you report on and explore trends based on your information. Pivot tables are particularly useful if you have long rows or columns that hold values you need to track the sums of and easily compare to one another. So, our pivot table describes the median value each feature for each score of quality. Now we can follow trends, for example, the highest value for ‘sulphates’ tend the highest ‘quality’ score. But we can’t draw our conclusion based on correlation.

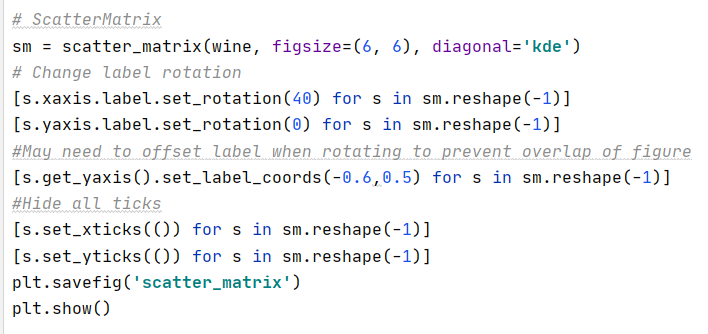
[In]

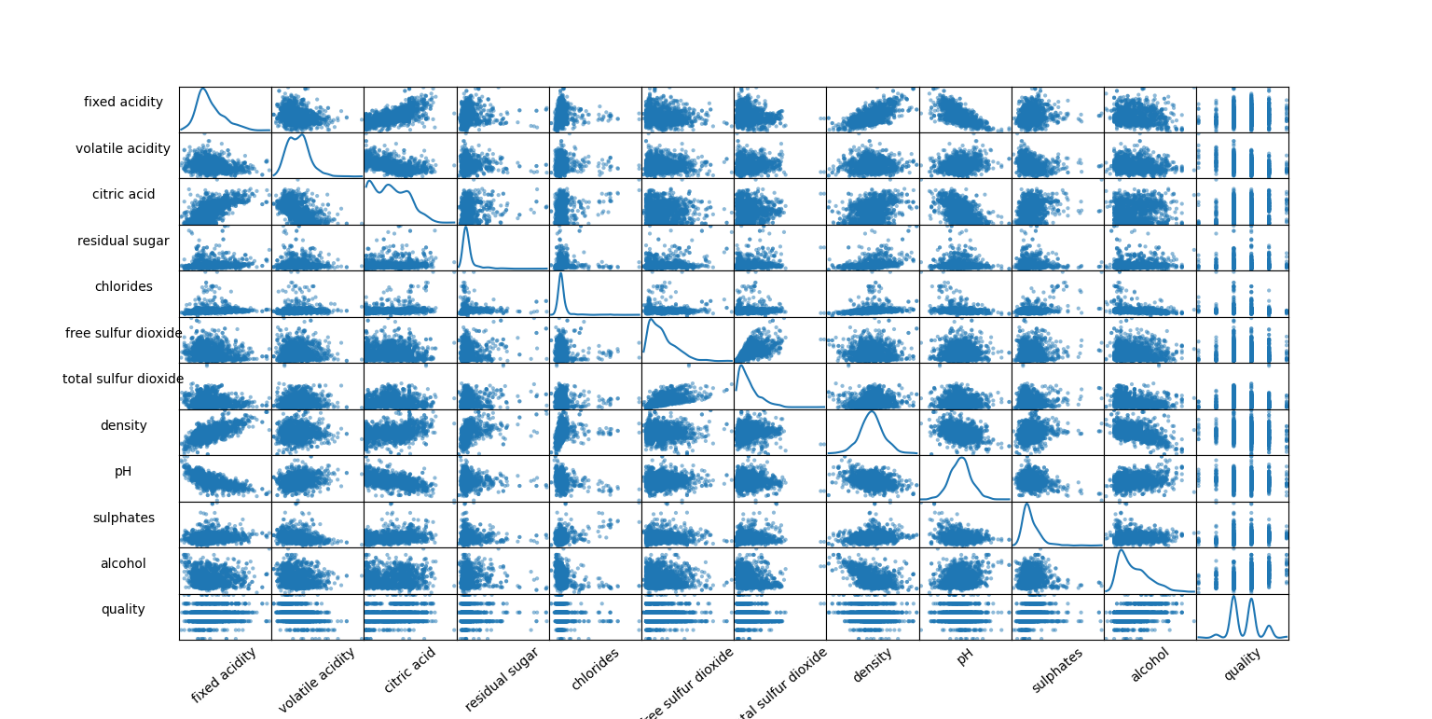
[Out]



You could visualize the scatterplot matrix for the better understanding relationship between a pair of variables. It plots every numerical attribute against every other.

## **Scatter Matrix**

[In]

[Out]

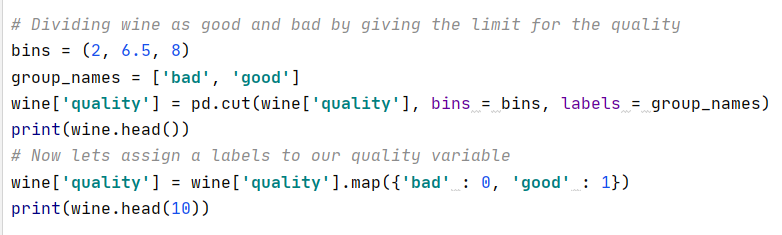
Next, I wanted to see the correlations between the variables that I’m working with. This allows me to get a much better understanding of the relationships between my variables in a quick glimpse.

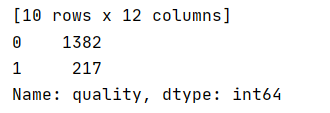
Immediately, I can see that there are some variables that are strongly correlated to **quality**. It’s likely that these variables are also the most important features in our machine learning model, but we’ll take a look at that later.

**5. Preparing Data for Modelling**

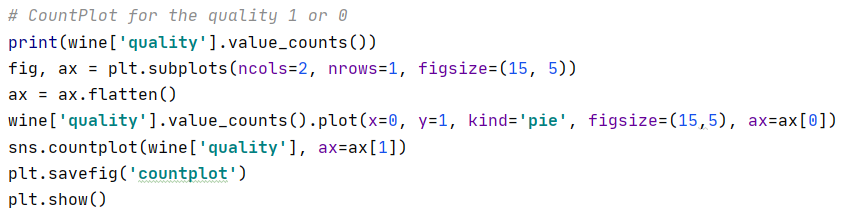
It’s time to prepare the data for our Machine Learning algorithms. In our dataset there aren’t any missing values, outliers, attributes that provide no useful information for the task. So, we could conclude that our data set is quite clean. Therefore, we won’t do any grueling data preparation, but some stuff will be needed to do. Human wine preferences scores varied from 3 to 8, so it’s straightforward to categorize answers into ‘bad’ or ‘good’ quality of wines. This allows us to practice with hyperparameter tuning on e.g., decision tree algorithms. Visualizing the graph of the number of values for each category, we could see that there are far many bad answers than good ones. Of course, machine learning algorithms operate digital values, so we assign for categorizes corresponding discrete values 0 or 1.

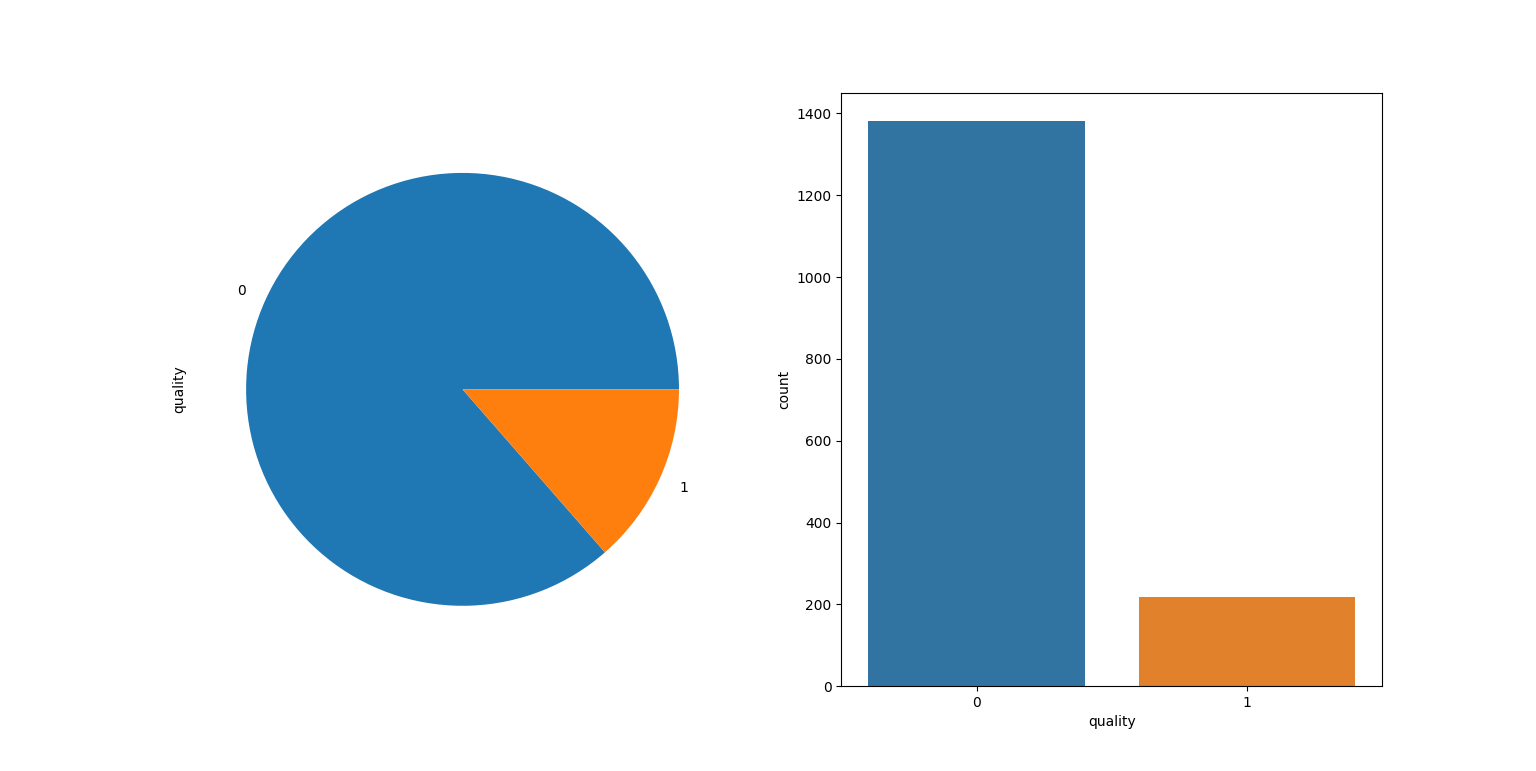
**5.1. Convert to a Classification Problem**

[In]

[Out]

**5.2. Proportion of Good vs. Bad Wines**

[In]

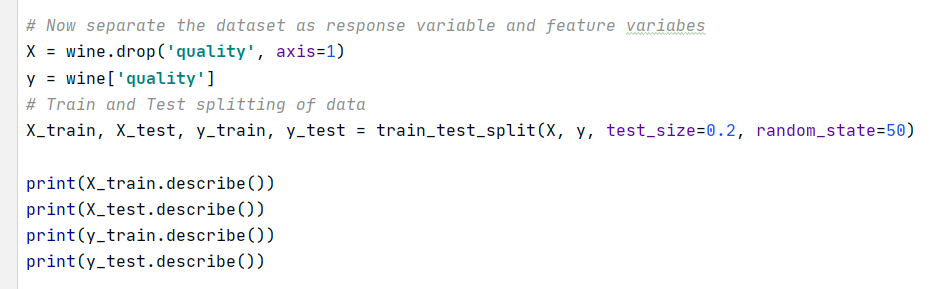
[Out]

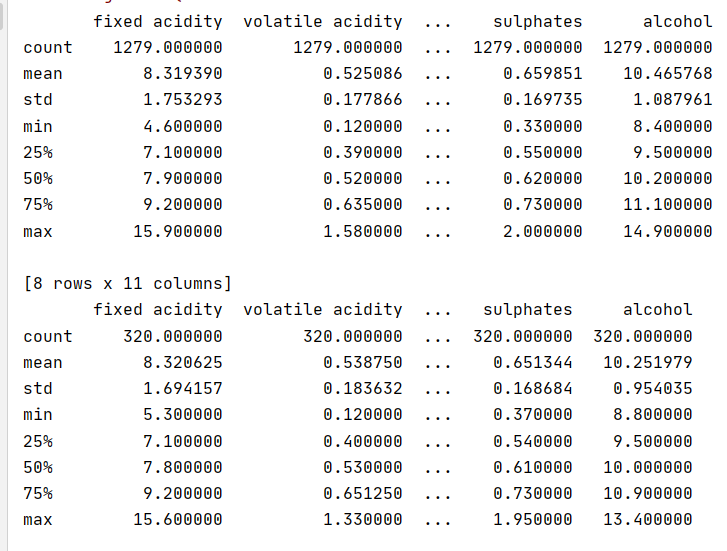
**5.2. Split Data**

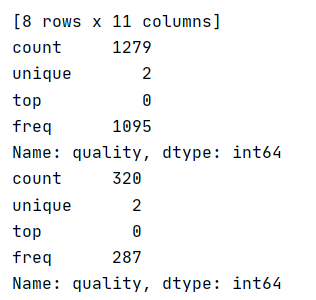
Before data investigation you should have been done with one of the most important steps is **data splitting.**The data has been split into two groups: training set 80%, test set 20%. The training set should be used to build your machine learning models. The test set should be used to see how well your model performs on unseen data.

Next, I split the data into a training and test set so that I could cross-validate my models and determine their effectiveness.

[In]



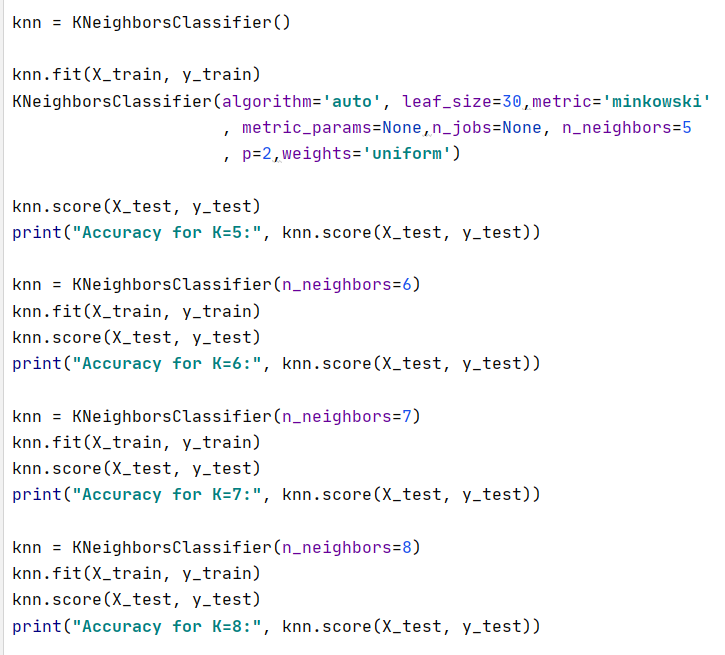
[Out]

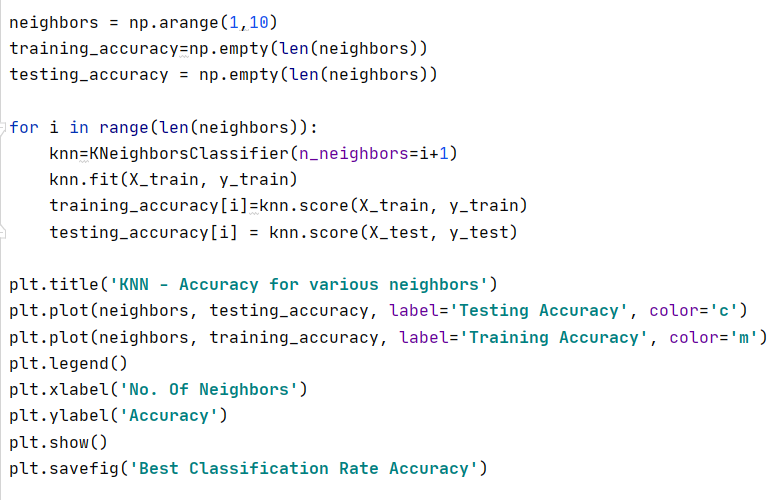


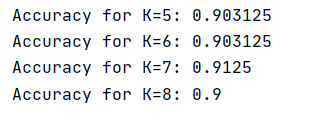
**6. Modelling**

For this project, I wanted to compare three different machine learning models: knn, decision trees, random forests. For the purpose of this project, I wanted to compare these models by their accuracy.

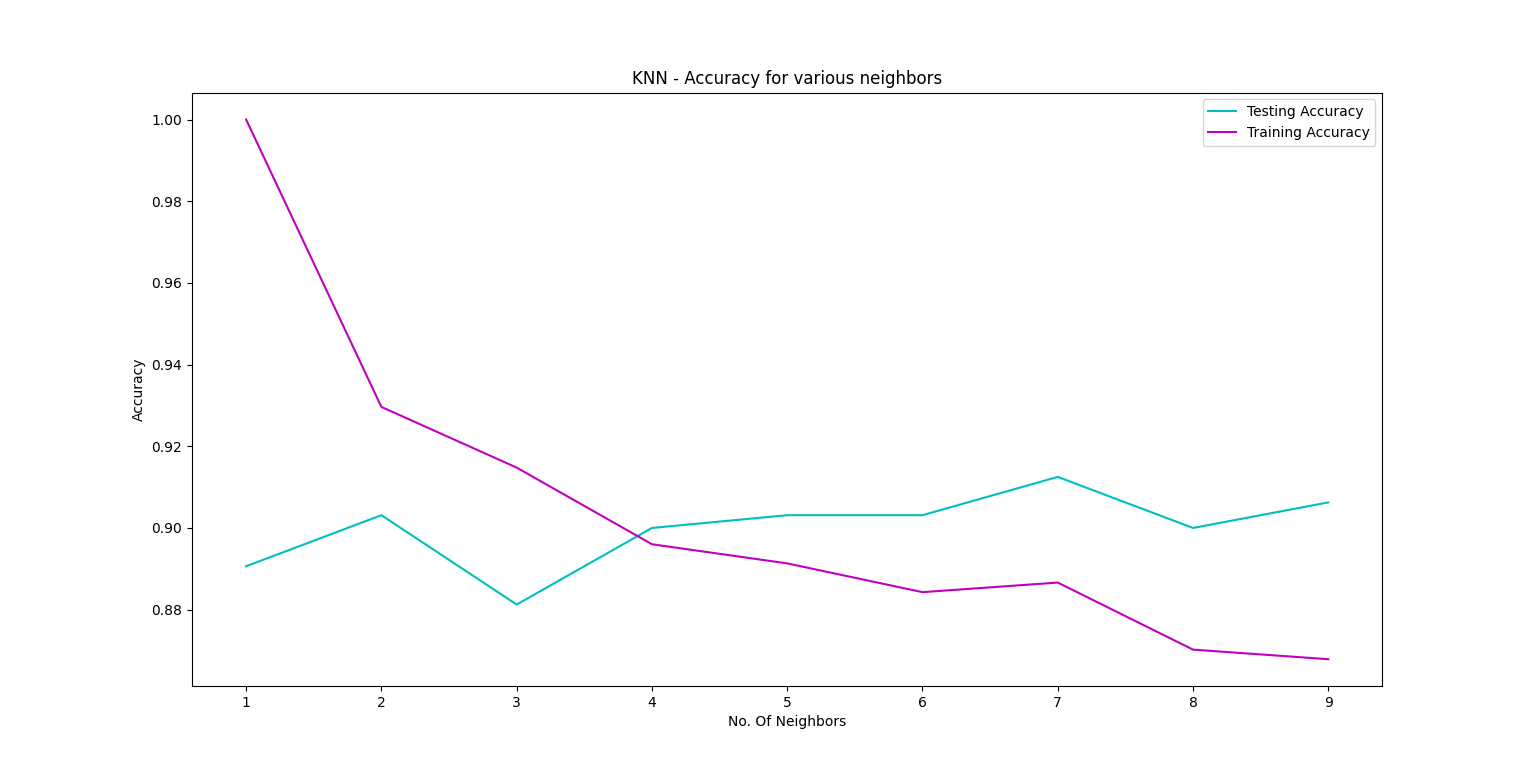
**6.1. Model 1: KNeighbors Classifier**

[In]

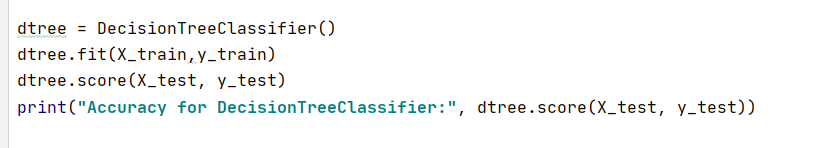


[Out]

[Out]



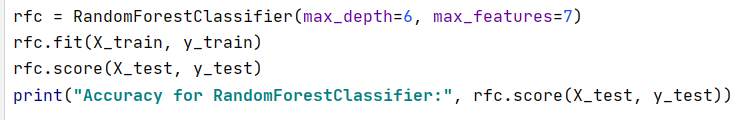
**6.2. Model 2: Decision Tree Classifier**

[In]

[Out]



**6.3. Model 3: Random Forest Classifier**

[In]

[Out]



**7. Conclusion**

**By looking into the details, we can see that good quality wines have higher levels of alcohol on average, have a lower volatile acidity on average, higher levels of sulphates on average, and higher levels of residual sugar on average. *I achieved Accuracy score for Random Forest: 94%.***