

**Institute of System Science**

SA4108: SA50 - Mobile Application Development

Batch: SA50

Machine Learning Continuous Assessment

BY

Chua Khiong Kiat A0055647X

Contents

[1. Linear Regression / Time Series Forecasting 3](#_Toc41076970)

[1.1. Introduction 3](#_Toc41076971)

[1.2. Data Dictionary 4](#_Toc41076972)

[1.3. The Aim of Analysis 5](#_Toc41076973)

[1.4. General Information of the Data 6](#_Toc41076974)

[2. Classification Techniques 7](#_Toc41076975)

[2.1. Introduction 7](#_Toc41076976)

[2.2. Data Dictionary 8](#_Toc41076977)

[2.3. The Aim of Analysis 10](#_Toc41076978)

[2.4. General Information of the Data 11](#_Toc41076979)

[2.5. Data Exploration 12](#_Toc41076980)

[2.6. Checking for Null Values 13](#_Toc41076981)

[2.7. Filling of the Row Data 13](#_Toc41076982)

[2.8. Getting information on the correlation list after filling null 13](#_Toc41076983)

[2.9. Logistic Regression 15](#_Toc41076984)

[2.9.1. Base Comparison 15](#_Toc41076985)

[2.9.2. Sensitivity Analysis 15](#_Toc41076986)

[Observation 16](#_Toc41076987)

[Discussion 16](#_Toc41076988)

[2.10. K-NN Classification 17](#_Toc41076989)

[Base Comparison 17](#_Toc41076990)

[Sensitivity Analysis 17](#_Toc41076991)

[Observation 17](#_Toc41076992)

[Discussion 18](#_Toc41076993)

[2.11. Decision Trees 19](#_Toc41076994)

[Base Comparison 19](#_Toc41076995)

[Sensitivity Analysis 19](#_Toc41076996)

[Observation 20](#_Toc41076997)

[Discussion 21](#_Toc41076998)

[2.12. Neural Network 23](#_Toc41076999)

[Base Comparison 23](#_Toc41077000)

[Sensitivity Analysis 23](#_Toc41077001)

[Observation 23](#_Toc41077002)

[Discussion 24](#_Toc41077003)

[2.13. KMeans 25](#_Toc41077004)

[2.13.1. Read the dataset 25](#_Toc41077005)

[2.13.2. Find the optimum number of clusters for k-means 25](#_Toc41077006)

[2.13.3. Create a k-means classifier with clusters = 3 and polt 26](#_Toc41077007)

[2.13.4. Plotting the centroids of the clusters 26](#_Toc41077008)

[2.13.5. View the relationship between quality and free sulfur dioxide/total sulfur dioxide(Actual Categorization) 27](#_Toc41077009)

[2.14. K-means 3D 29](#_Toc41077010)

[2.14.1. Read the dataset 29](#_Toc41077011)

[2.14.2. Find the optimum number of clusters for k-means 29](#_Toc41077012)

[2.14.3. create a k-means classifier with clusters = 3 and polt 30](#_Toc41077013)

[2.14.4. Plotting the centroids of the clusters 30](#_Toc41077014)

[2.14.5. View the relationship between quality and free sulfur dioxide/total sulfur dioxide/density (Actual Categorization) 30](#_Toc41077015)

[2.14.6. Divided quality into range 31](#_Toc41077016)

[2.14.7. View the relationship between quality in range and free sulfur dioxide/total sulfur dioxide(Actual Categorization) 32](#_Toc41077017)

[2.15. DBSCAN 33](#_Toc41077018)

[3. Appendix 34](#_Toc41077019)

# Linear Regression / Time Series Forecasting

## Introduction

## Data Dictionary

## The Aim of Analysis

## General Information of the Data

# Classification Techniques

## Introduction

The dataset was downloaded from the UCI Machine Learning Repository.

The link from the file is as follow:

<https://www.kaggle.com/rajyellow46/wine-quality>

File Name: winequalityN.csv

The dataset contains information related to red and white variants of the Portuguese "Vinho Verde" wine. The reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant.

The dataset contains null values which needs to do data cleaning to ensure a proper set of data can be used.

## Data Dictionary

|  |  |  |
| --- | --- | --- |
| S/N | Data Column Title | Description |
| 1 | Wine Type |  |
| 2 | Fixed Acidity | Fixed acids include tartaric, malic, citric, and succinic acids which are found in grapes (except succinic). This variable is usually expressed in g(tartaricacid)dm3 in the dataset. |
| 3 | Volatile Acidity | These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid though other acids like lactic, formic and butyric acids might also be present. Excess of volatile acids are undesirable and lead to unpleasant flavor. In the US, the legal limits of volatile acidity are 1.2 g/L for red table wine and 1.1 g/L for white table wine. The volatile acidity is expressed in g(aceticacid)dm3 in the dataset. |
| 4 | Citric Acid | This is one of the fixed acids which gives a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness. It's usually expressed in gdm3 in the dataset. |
| 5 | Residual sugar | This typically refers to the natural sugar from grapes which remains after the fermentation process stops, or is stopped. It's usually expressed in g/dm3 in the dataset. |
| 6 | Chlorides | Chloride concentration in the wine is influenced by terroir and its highest levels are found in wines coming from countries where irrigation is carried out using salty water or in areas with brackish terrains. This is usually a major contributor to saltiness in wine. It's usually expressed in g(sodiumchloride)dm3g(sodiumchloride)dm3 in the dataset. |
| 7 | Free Sulfur Dioxide | This is the part of the sulphur dioxide that when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. They are also known as sulfites and too much of it is undesirable and gives a pungent odour. This variable is expressed in mgdm3 in the dataset. |
| 8 | Total Sulfur Dioxide | This is the sum total of the bound and the free sulfur dioxide ( SO2 ). Here, it's expressed in mgdm3 . This is mainly added to kill harmful bacteria and preserve quality and freshness. There are usually legal limits for sulfur levels in wines and excess of it can even kill good yeast and give out undesirable odour. |
| 9 | Density | This can be represented as a comparison of the weight of a specific volume of wine to an equivalent volume of water. It is generally used as a measure of the conversion of sugar to alcohol. Here, it's expressed in gcm3 . |
| 10 | pH | Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. You might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic. |
| 11 | Sulphates | These are mineral salts containing sulfur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. They are connected to the fermentation process and affects the wine aroma and flavor. Here, it's expressed in g(potassiumsulphate)dm3 in the dataset. |
| 12 | Alcohol | Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine. We interpret alcohol using many different taste receptors which is why it can taste bitter, sweet, spicy, and oily all at once. Your genetics actually plays a role in how bitter or sweet alcohol tastes. Regardless, we can all sense alcohol towards the backs of our mouths in our throats as a warming sensation. |
| 13 | Quality | Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts. |
| 14 | Quality Group (Derived from Quality) | The Quality Group has been split with the following ranges to define good (0-4),average (4-7) ,bad (7-10) |

## The Aim of Analysis

There are a few aims to the current analysis:

* Identify the various use of different machine learning model based on supervising techniques
* 3 outputs have been identified to be used as a machine learning target.
  + Type
  + Quality
  + Quality Group (Derived from Quality)
* Classification Methods include:
  + Logistic Regression
  + K-NN Classification
  + Decision Trees
  + Neural Network
* Each Classification method will be accompanied with a 2 Unsupervised Techniques
  + PCA
  + Pearson Correlation

A separate analysis with unsupervised learning using k-Means and Agglomerative Clustering will be used to find a good clustering pattern.

## General Information of the Data

**Type**: Two types of wines such as red wine and white wine.

**Fixed acidity**: Fixed acids include tartaric, malic, citric, and succinic acids which are found in grapes (except succinic)

Acids are one of the fundamental properties of wine and contribute greatly to the taste of the wine, Acidity in food and drink tastes tart and zesty. Tasting acidity is also sometimes confused with alcohol. Wines with higher acidity feel lighter-bodied because they come across as “spritzy”. Reducing acids significantly might lead to wines tasting flat. If you prefer a wine that is richer and rounder, you enjoy slightly less acidity.

**Volatile acidity**: These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid though other acids like lactic, formic and butyric acids might also be present. Excess of volatile acids are undesirable and lead to unpleasant flavour.

**Citric acid**: This is one of the fixed acids which gives a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness.

**Residual sugar**: This typically refers to the natural sugar from grapes which remains after the fermentation process stops, or is stopped.

**Chlorides**: Chloride concentration in the wine is influenced by terroir and its highest levels are found in wines coming from countries where irrigation is carried out using salty water or in areas with brackish terrains.

**Free sulfur dioxide**: This is the part of the sulphur dioxide that when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. They are also known as sulfites and too much of it is undesirable and gives a pungent odour.

**Total sulfur dioxide**: This is the sum total of the bound and the free sulfur dioxide. This is mainly added to kill harmful bacteria and preserve quality and freshness. There are usually legal limits for sulfur levels in wines and excess of it can even kill good yeast and give out undesirable odour.

**Density**: This can be represented as a comparison of the weight of a specific volume of wine to an equivalent volume of water. It is generally used as a measure of the conversion of sugar to alcohol.

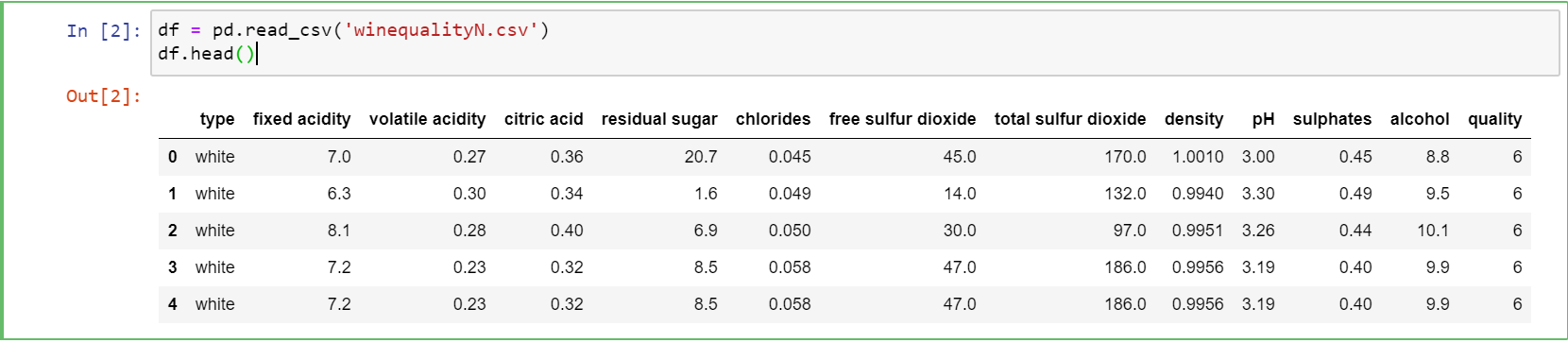
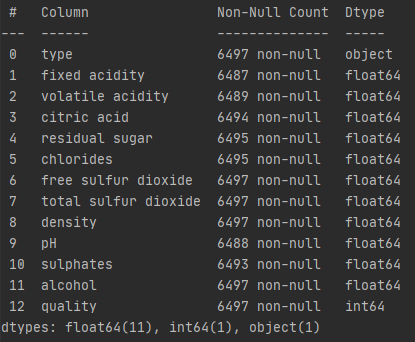
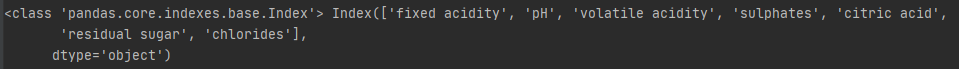
**pH**: Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. You might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.

**Sulphates**: These are mineral salts containing sulfur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. They are connected to the fermentation process and affects the wine aroma and flavour.

**Alcohol**: It's usually measured in % vol or alcohol by volume (ABV).

**Quality**: Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.

## Data Exploration

* Importing Packages  
  
* Reading the CSV file  
  
* Getting the information from the csv  
  
* Getting Column Head  
  

## Checking for Null Values

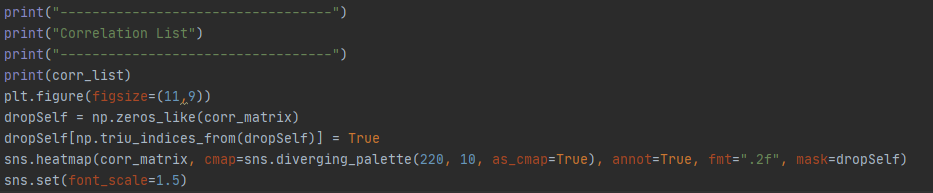
* Getting the number of columns that have null values and the percentage with respect to the total number rows  
  
* Checking NULL values in data.



## Filling of the Row Data



## Getting information on the correlation list after filling null



## Logistic Regression

### Base Comparison

Model 1: To Predict Wine Type

Model 2: To Predict Quality

Model 3: To Predict Quality Group

Assumptions:

1. Duration of model is based on purely model fitting
2. Column expressed in array from 0 – 13  
   [0,1,2,3,4,5,6,7,8,9,10,11,12,13] is [“Wine Type”, ….., “Quality”, “Quality Group”]



### Sensitivity Analysis

#### Sensitivity Analysis 1

#### Sensitivity Analysis 2

### Observation

#### Sensitivity Analysis 1

* It was noted that comparing models that utilize PCA has overall much lesser time required to train the model.
* Comparing Model 2 and 3, model 3 has an estimated 33% reduction in model fitting duration after binning the quality into 3 groups.
* It is noted that Model 2 which has 10 outputs (Quality Class) have a lower accuracy (estimated. 48% to 55%).
* The accuracy is better/equal when comparing Model with PCA and without PCA (this is true for all 3 models)

### Discussion

#### Sensitivity Analysis 1

* It is noted that with feature engineering (e.g. PCA), PCA is able to reduce training time significantly as this is consistent as lesser data is being passed through to the model (11 columns of data compared to 3 columns).
* It is noted that PCA in general can increase accuracy (most of the time), this is especially true when the base training model has a high accuracy to start with (e.g. Model 1 and Model 3).
* It is evident that model training with PCA is beneficial when the base model inherently has a high accuracy value. Given the example in Model 1 and 3, the duration required to train the model differs as much as 500%. Given that we only use 6000 datapoints, during the event that 6 million datapoints is required. The training will be increased exponentially. Hence the conclusion that using feature engineering on Machine Learning Models is very important.

## K-NN Classification

### Base Comparison

Model 1: To Predict Wine Type

Model 2: To Predict Quality

Model 3: To Predict Quality Group

Assumptions:

1. Duration of model is based on purely model fitting
2. Column expressed in array from 0 – 12  
   [0,1,2,3,4,5,6,7,8,9,10,11,12] is [“Wine Type”, ….., “Quality”]
3. All columns considered except for Wine Type & Quality
4. For Model 3, Quality Group is represented using data binning for quality scores [0(Low: 0 to 4),1 (Mid: 4 to 7) ,2 (High: 7 to 10)]

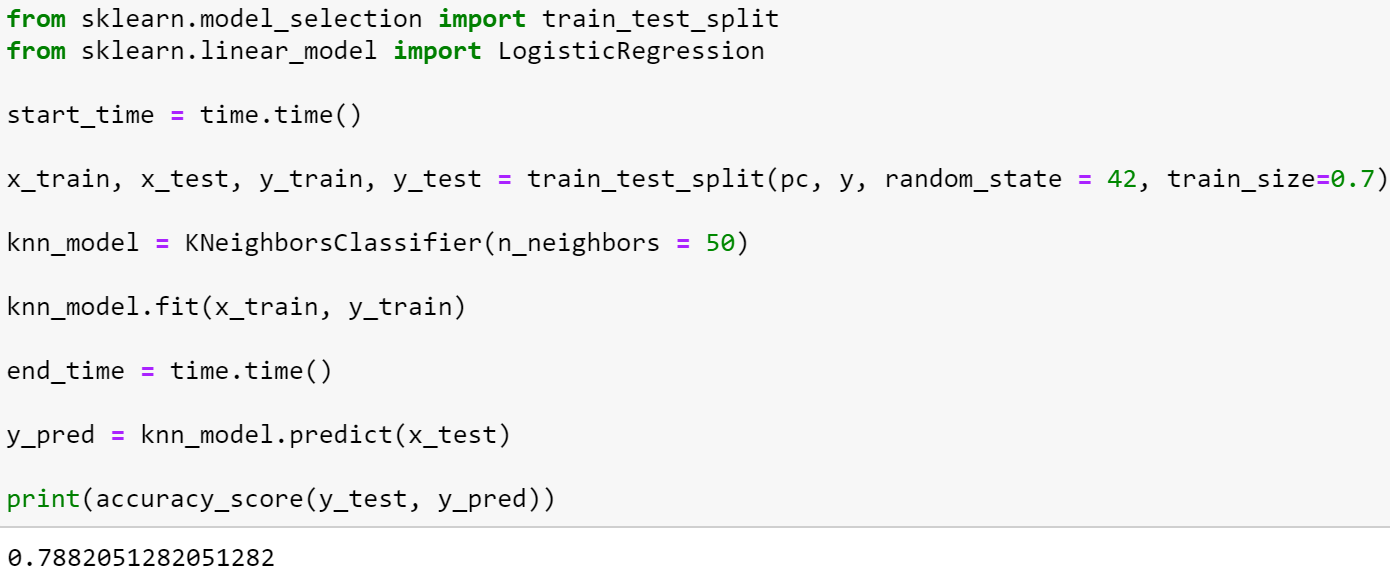
|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Column Input | [1,2,3,4,5,6,7,8,9,10,11] | [1,2,3,4,5,6,7,8,9,10,11] | [1,2,3,4,5,6,7,8,9,10,11] |
| Column Output | [0] | [12] | [12] |
| K = number of neighbours | 50 | 50 | 50 |
| Duration | 0.014006853103637695 | 0.10392022132873535 | 0.015992164611816406 |
| Accuracy | 0.9415384615384615 | 0.46 | 0.7543589743589744 |

### Sensitivity Analysis

### Observation

Our initial model only has an accuracy of 46% when predicting quality. After data engineering, accuracy of predicting quality is 75.4%.

Utilizing Principal Component Analysis for dimension reduction, we have improved accuracy of 78.8%



### Discussion

## Decision Trees

### Base Comparison

Model 1: To Predict Wine Type

Model 2: To Predict Quality

Model 3: To Predict Quality Group

Assumptions:

1. Duration of model is based on purely model fitting
2. Column expressed in array from 0 – 13  
   [0,1,2,3,4,5,6,7,8,9,10,11,12,13] is [“Wine Type”, ….., “Quality”, “Quality Group”]
3. All column inputs have been considered except wine type, quality, quality group.
4. Decision tree criterion is ‘Gini’.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Column Input | [1,2,3,4,5,6,7,8,9,10,11] | [1,2,3,4,5,6,7,8,9,10,11] | [1,2,3,4,5,6,7,8,9,10,11] |
| Column Output | [0] | [12] | [13] |
| Max Depth | 3 | 3 | 3 |
| Duration | 0.0114 | 0.0089 | 0.01 |
| Accuracy | 94.9% | 53.64% | 93.33% |

### Sensitivity Analysis

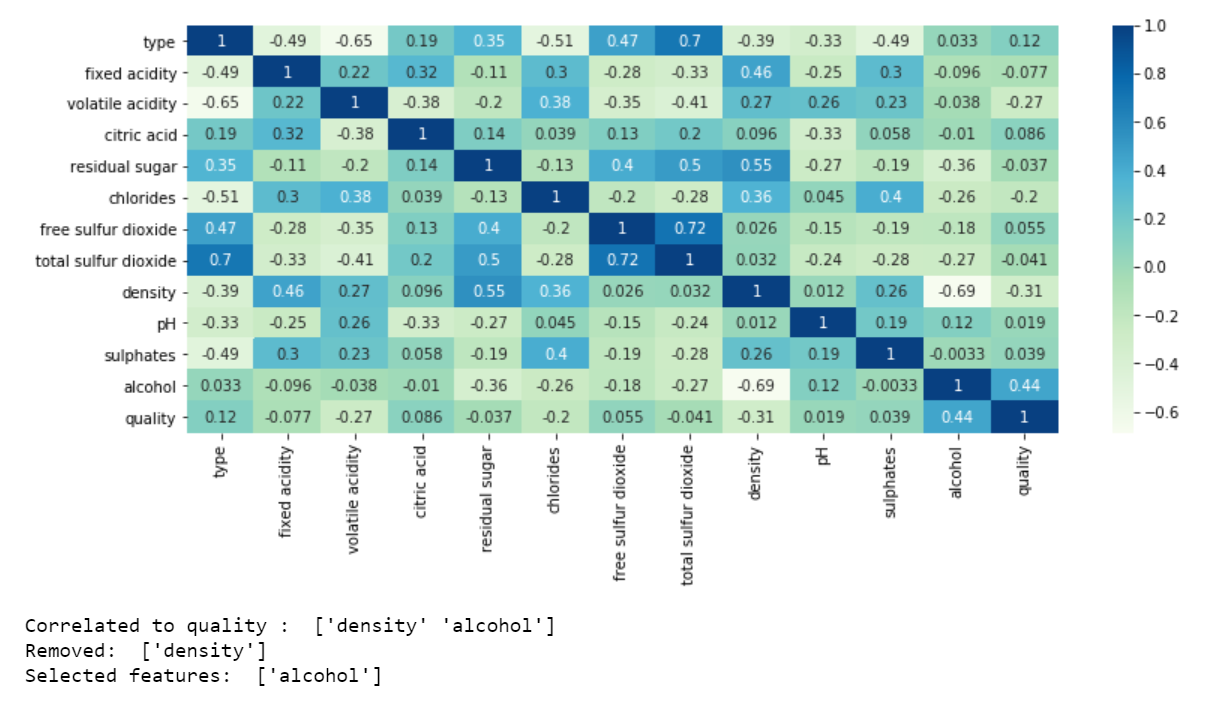
### Observation

### Discussion

#### Feature Selection by Pearson Correlation Matrix

By Using Pearson Correlation, we can know

* ‘quality’ is correlated with alcohol.
* ‘Free sulfur dioxide’ and ‘total sulfur dioxide’ is highly correlated (0.72 in our confusion matrix) and they can have almost the same effect on an independent variable.
* Therefore, when the two features are highly correlated, we can drop one of them in our feature set.
* Correlation of ‘total sulfur dioxide’ and ‘residual sugar’ is 0.5 and they have moderate correlation.
* Correlation of ‘density’ and ‘residual sugar’ is 0.55 and they also have moderate correlation.
* In conclusion, our target variable is correlated with alcohol and low correlation with other features.

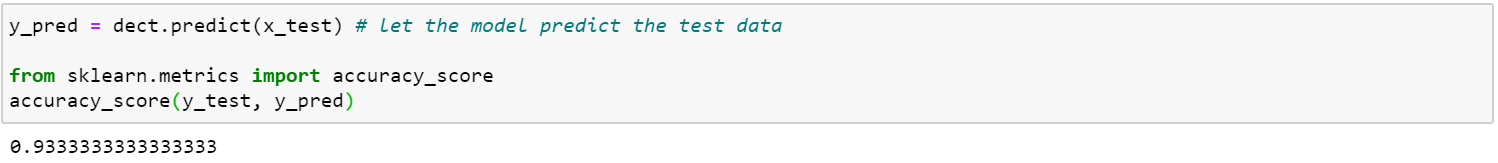


We can also drop one of the features from ‘free sulfur dioxide’ and ‘total sulfur dioxide’ since they are highly correlated.

The accuracy that results after dropping one of them makes no difference in our models.

As mentioned below, the accuracy is 0.93333.





#### Impact of Data Engineering and Feature Engineering

In our dataset, we use two types of data engineering.

* The first one is categorical data engineering. By using categorial data engineering, we map ‘type’ category to a unique value such as white = 0 and red = 1. By using label encoding, we can easily visualize the relationship between the features and the target in our plots and matrix.
* The second one is data cleansing. Since our dataset contains missing data and NaN, we fill with a value of mean value on NaN column.

For the feature engineering, we use feature selection and feature extraction techniques.

* By using Pearson correlation matrix, we can know which features is correlated to each other.
* Since ‘free sulfur dioxide’ and ‘total sulfur dioxide’ are highly correlated with each other, removing one of these features doesn’t affect on our accuracy results.
* Moreover, dropping ‘sugar’ or ‘total sulfur dioxide’ will give better accuracy results since these are loosely correlated with our dependent variable ‘quality’.

## Neural Network

### Base Comparison

Model 1: To Predict Wine Type

Model 2: To Predict Quality

Model 3: To Predict Quality Group

Assumptions:

1. Duration of model is based on purely model fitting
2. Column expressed in array from 0 – 13  
   [0,1,2,3,4,5,6,7,8,9,10,11,12,13] is [“Wine Type”, ….., “Quality”, “Quality Group”]
3. All column inputs have been considered except wine type, quality, quality group.
4. Only 1 Hidden Layer of Activation Mode “Sigmoid” has been done for the base comparison for all data type.

Results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Column Input | [1,2,3,4,5,6,7,8,9,10,11] | [1,2,3,4,5,6,7,8,9,10,11] | [1,2,3,4,5,6,7,8,9,10,11] |
| Column Output | [0] | [12] | [13] |
| Dense Value | 500 | 500 | 500 |
| Duration | 12.161257028579712 | 12.11571193 | 11.330935001373291 |
| Accuracy | 0.9830769300460815 | 0.558461547 | 0.9215384721755981 |

### Sensitivity Analysis

#### Sensitivity Analysis 1

Graph below shows a summary of Models 1,2,3 are affected by the dense values.

Assumptions:

1. Column expressed in array from 0 – 13  
   [0,1,2,3,4,5,6,7,8,9,10,11,12,13] is [“Wine Type”, ….., “Quality”, “Quality Group”]
2. All column inputs have been considered except wine type [0], quality [12], quality group [13].

#### Sensitivity Analysis 2

### Observation

### Discussion

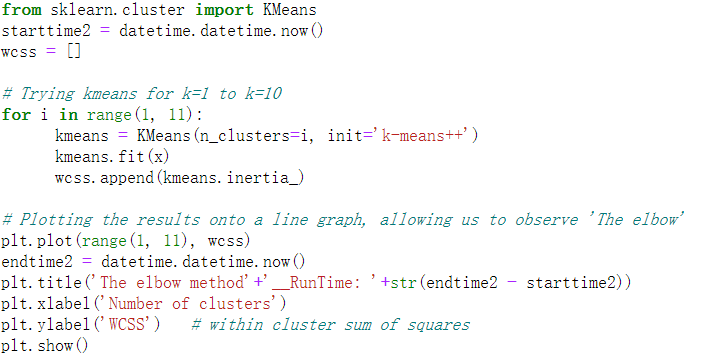
## KMeans

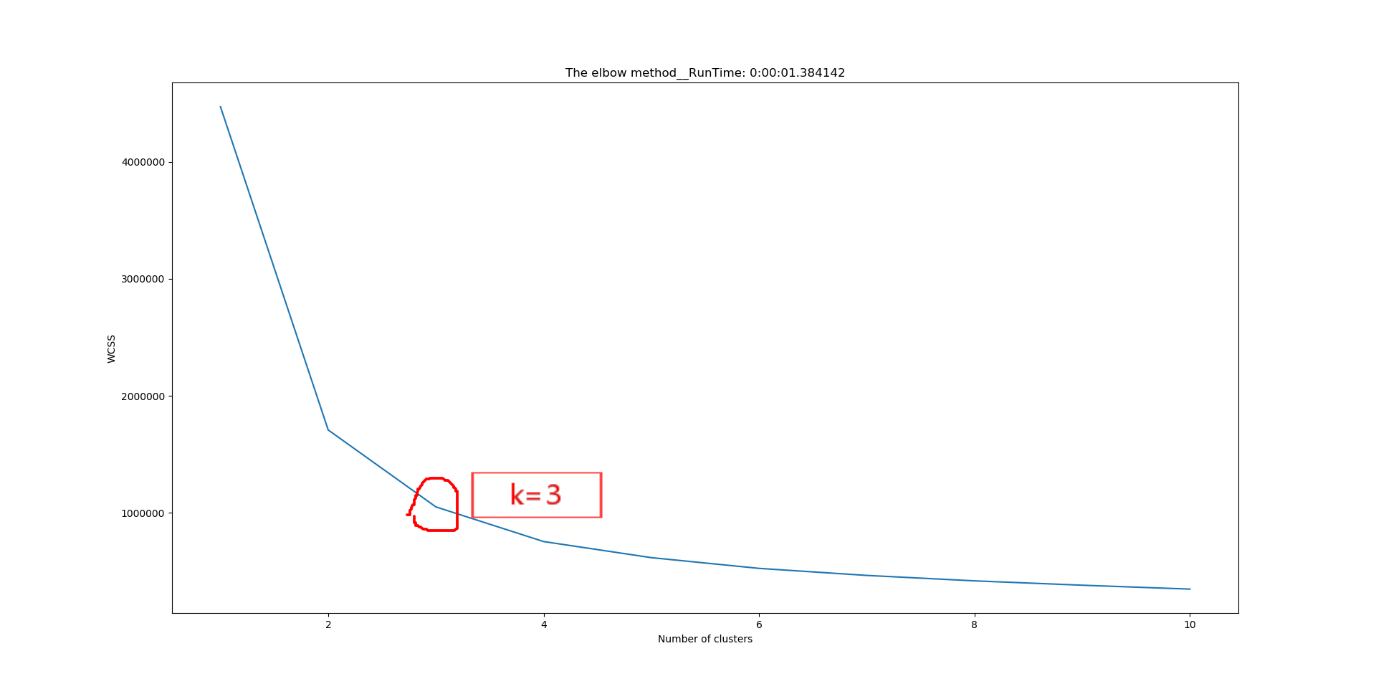
### Read the dataset



read the data in rows 0-2000, columns 6(free sulfur dioxide) and 7(total sulfur dioxide)

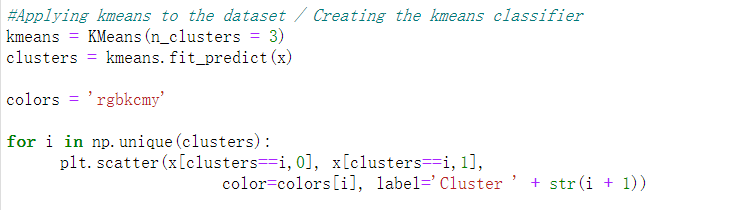
### Find the optimum number of clusters for k-means



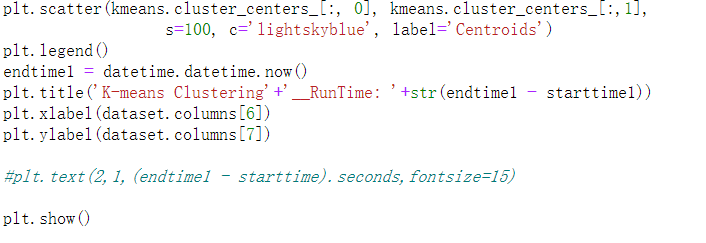


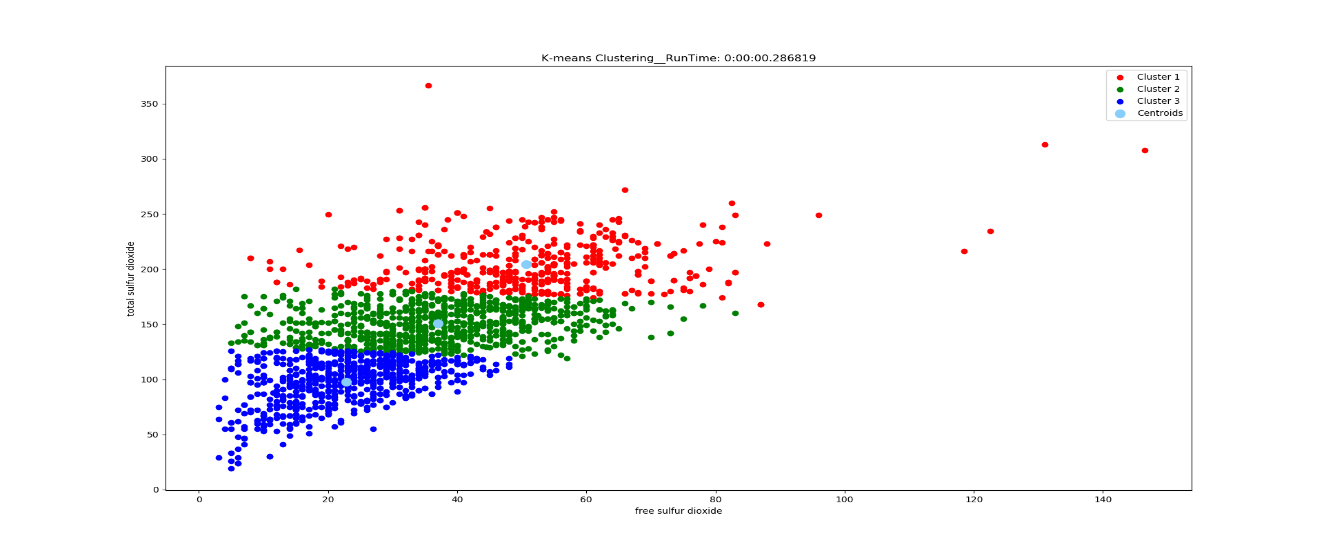
* It can be seen from the figure that the running time of the elbow graph is 1.384142 seconds.
* We can think that when K = 3, the elbow line starts to be gentle without rapid descending, so we would set the cluster value to 3.

### Create a k-means classifier with clusters = 3 and polt



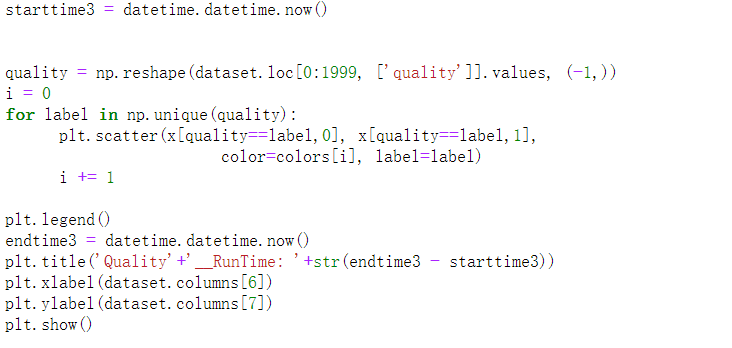
### Plotting the centroids of the clusters

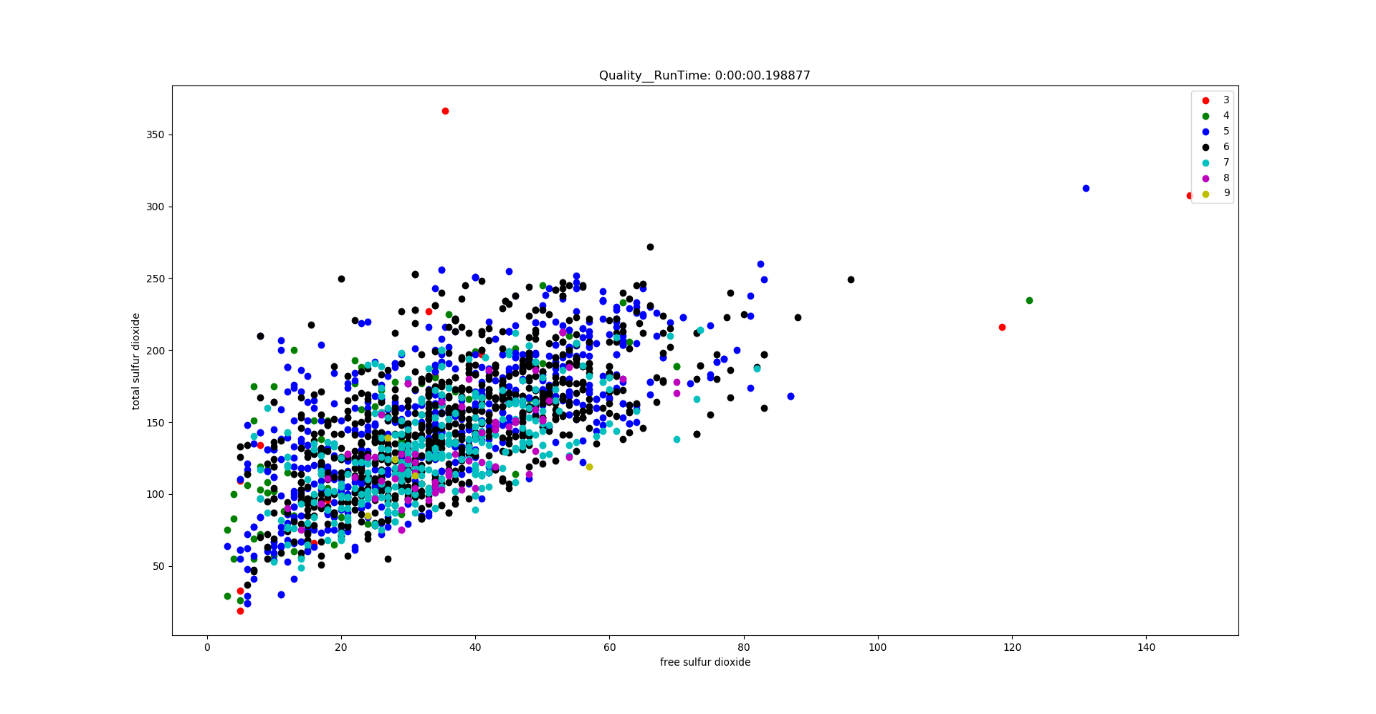




* The figure on title shows the running time is 0.286819 sec. And it contains the parameter relationship between the column 6 (free sulfur dioxide) and 7 (total sulfur dioxide).
* Calculated by the k-means algorithm, k-means cannot give us accurate conclusions because it shows no data plot group in the diagram, but through this pic we might can consider using linear regression to find the trend.

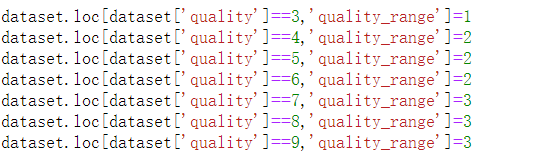
### View the relationship between quality and free sulfur dioxide/total sulfur dioxide(Actual Categorization)



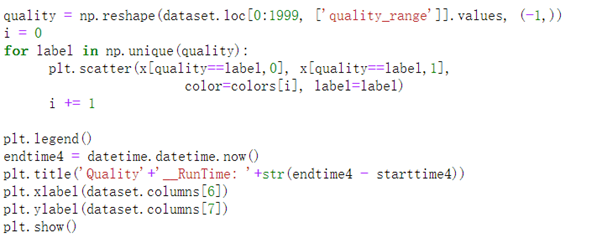


* The running time is 0.198877 sec.
* From the dataset we found that the quality of the wine ranges from 3 to 9. But it seems that there is no obvious relationship between quality and free sulfur dioxide/total sulfur dioxide.

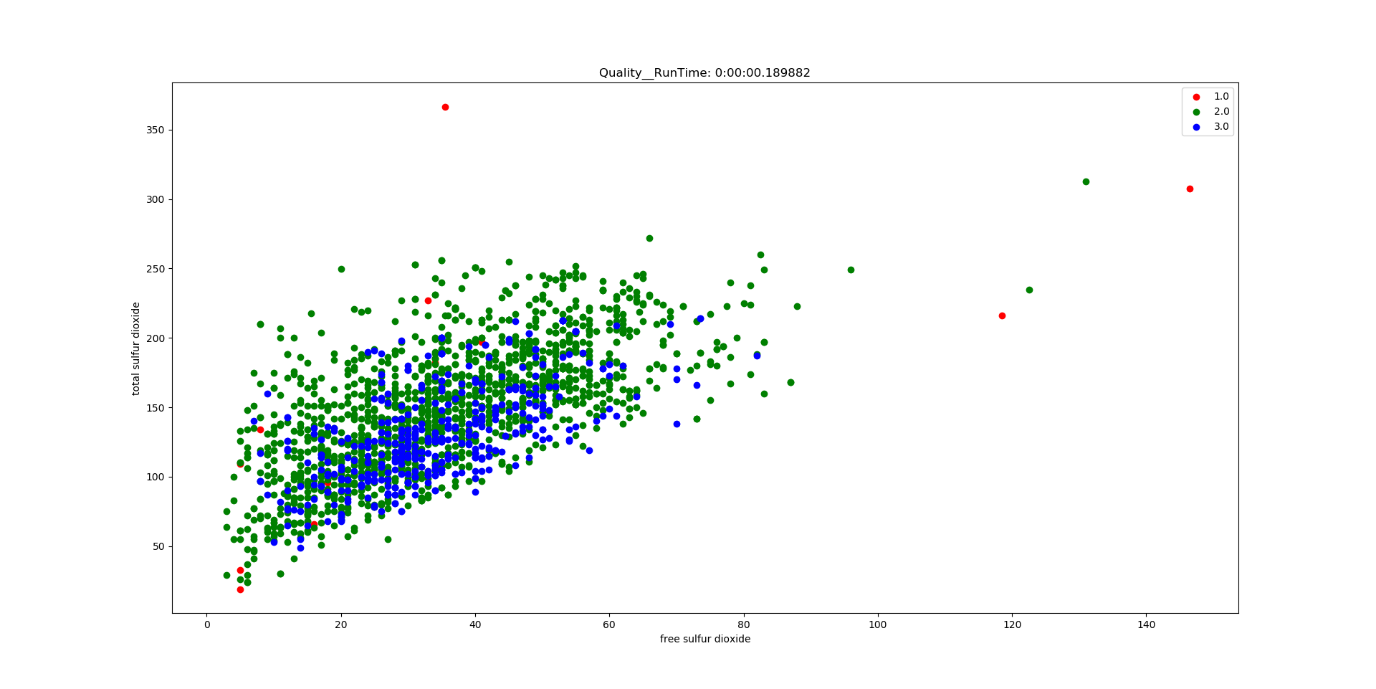
**6. Divided quality into range**



**7.View the relationship between quality in range and free sulfur dioxide/total sulfur dioxide(Actual Categorization)**



We try to find some extra info from using K-means algorithm. Now we are binning the quality from 3-9 into 3 ranges, quality [3-4) to range 1; quality [4-7) to range 2, quality [7-9) to range3.



When we use the ” quality ranges” to color the plot, the figure changes easier to be known, for example the plot of quality range 1 is discrete from the main group of data, and the other plots are concentrated in a group around 5-80 in X-label and 40-250 in Y-label.

But all the plot of quality range 2 and 3 are mixed together, we might need other way to find the regular.

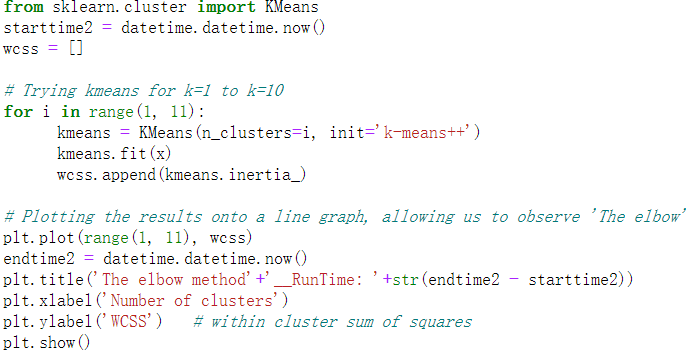
## K-means 3D

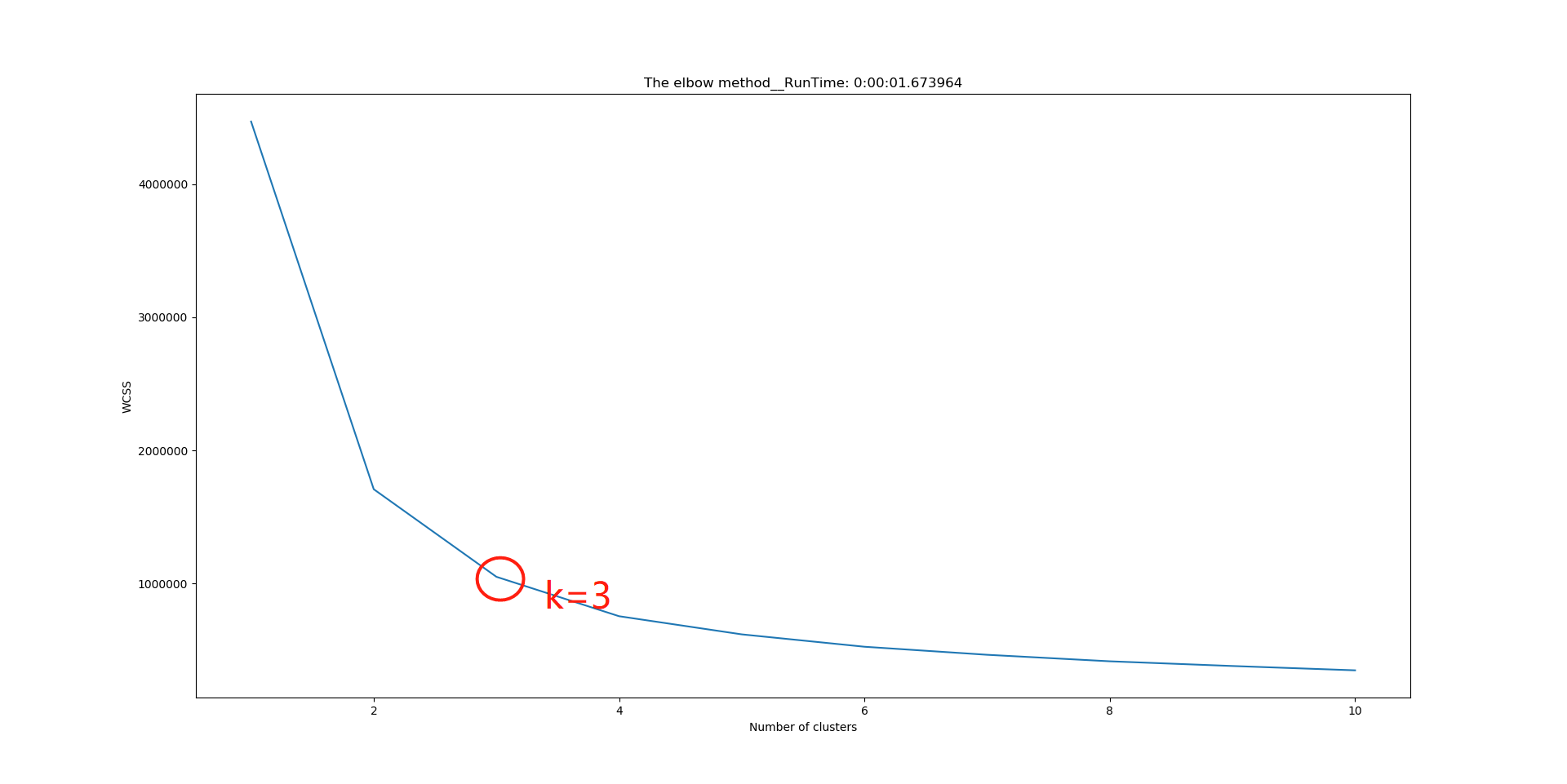
### Read the dataset



read the data in rows 0-2000, columns 6(free sulfur dioxide) and 7(total sulfur dioxide)and8(density)

### Find the optimum number of clusters for k-means

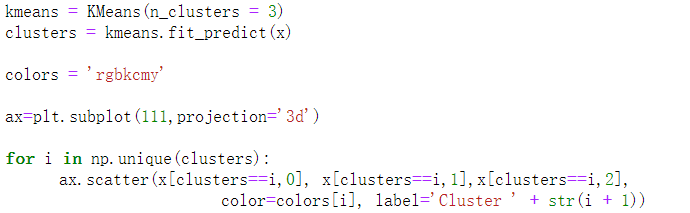




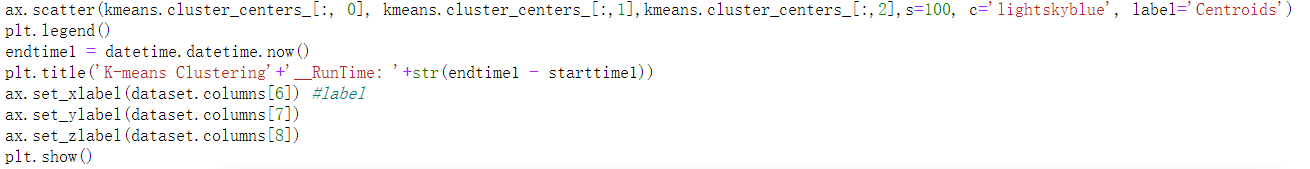
It can be seen from the figure that the running time of the elbow graph is 1.673964 seconds.

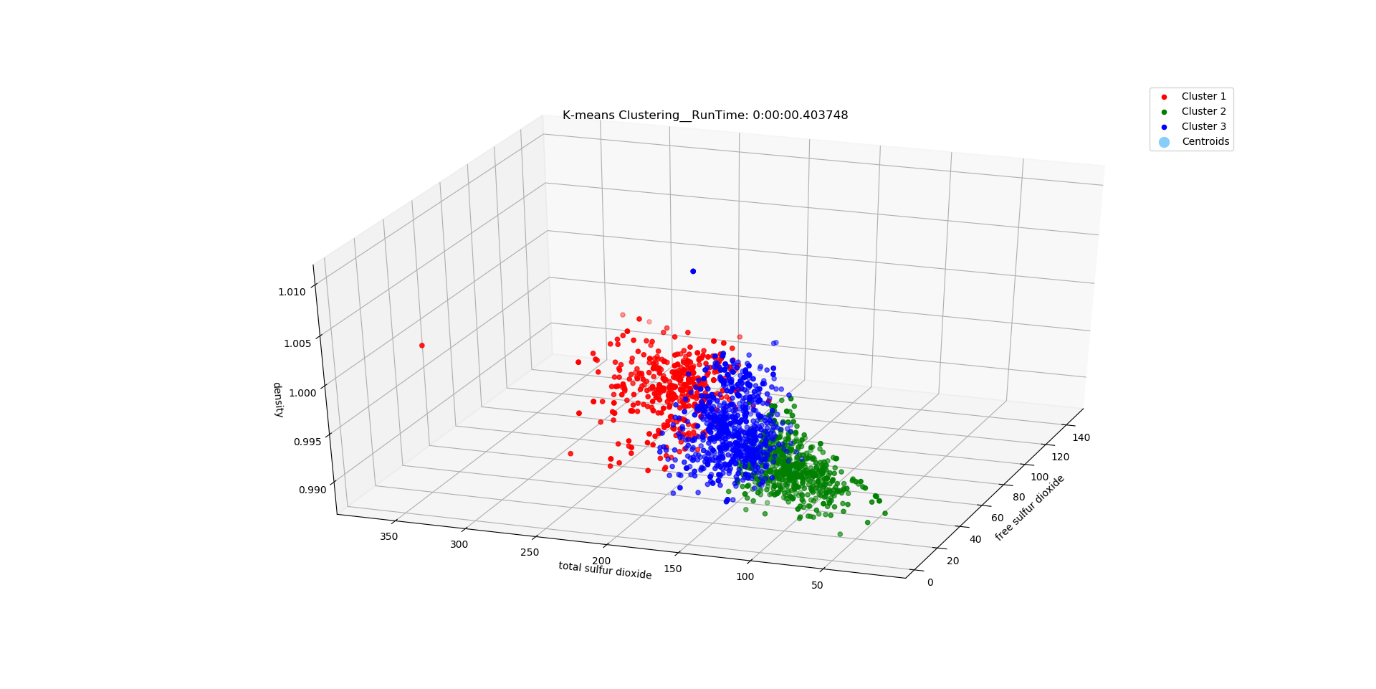
We can think that when K = 3, the elbow line starts to be gentle without rapid descending, so we would set the cluster value to 3.

### create a k-means classifier with clusters = 3 and polt



### Plotting the centroids of the clusters

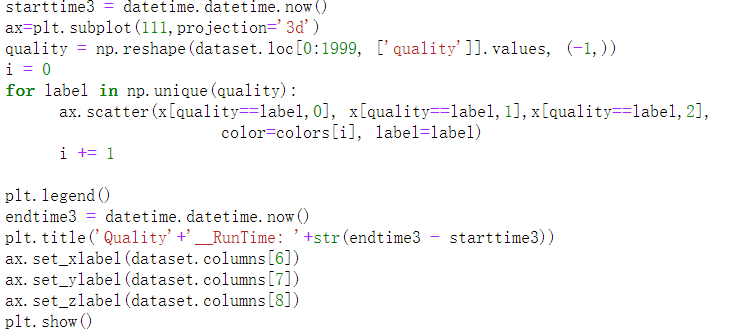


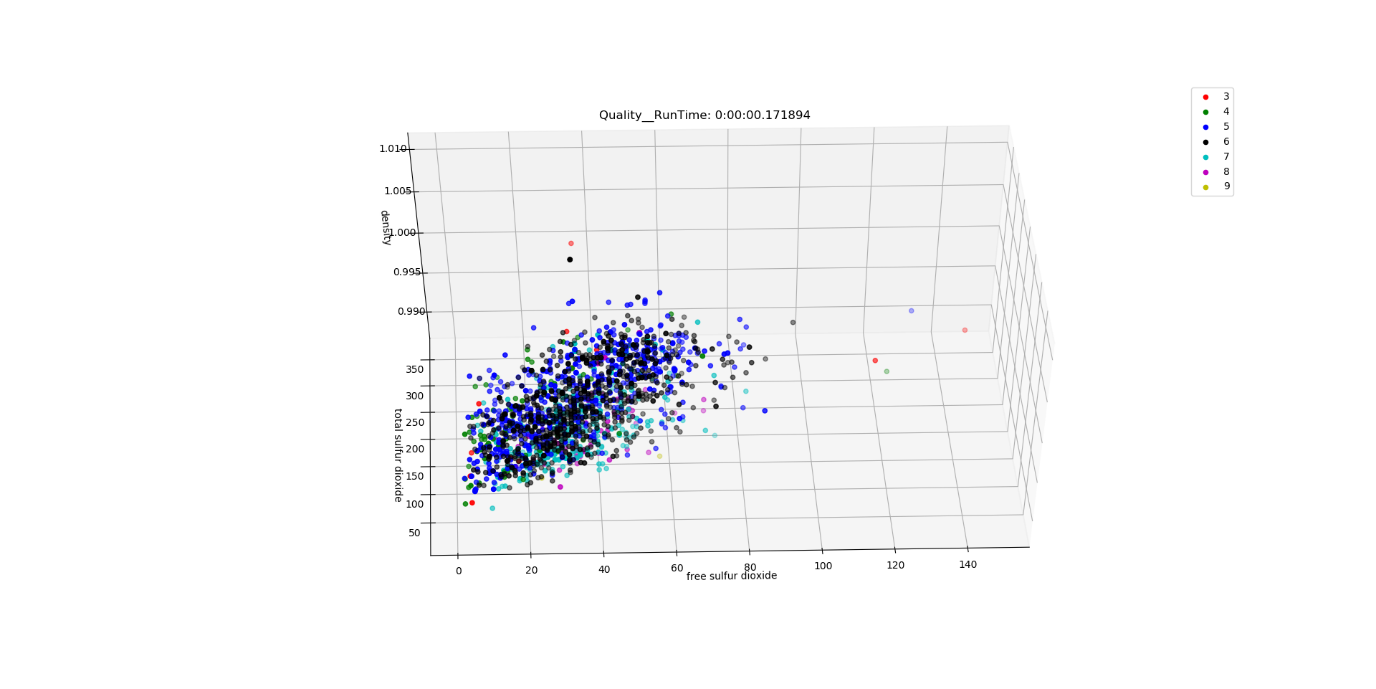


The figure on title shows the running time is 0. sec. And it contains the parameter relationship between the column 6 (free sulfur dioxide) and 7 (total sulfur dioxide)and 8(density)

Calculated by the k-means algorithm, We tried to add a variable, but still could not find data plot group in the diagram.

### View the relationship between quality and free sulfur dioxide/total sulfur dioxide/density (Actual Categorization)

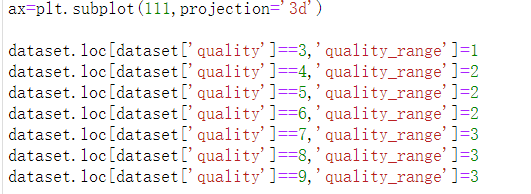




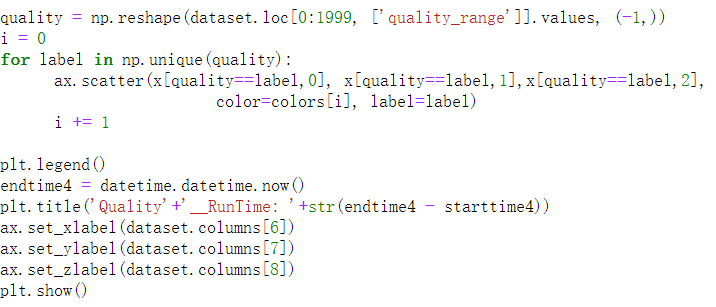
The running time is 0.171894 sec.

From the dataset we found that the quality of the wine ranges from 3 to 9. But it seems that there is no obvious relationship between quality and free sulfur dioxide/total sulfur dioxide/density.

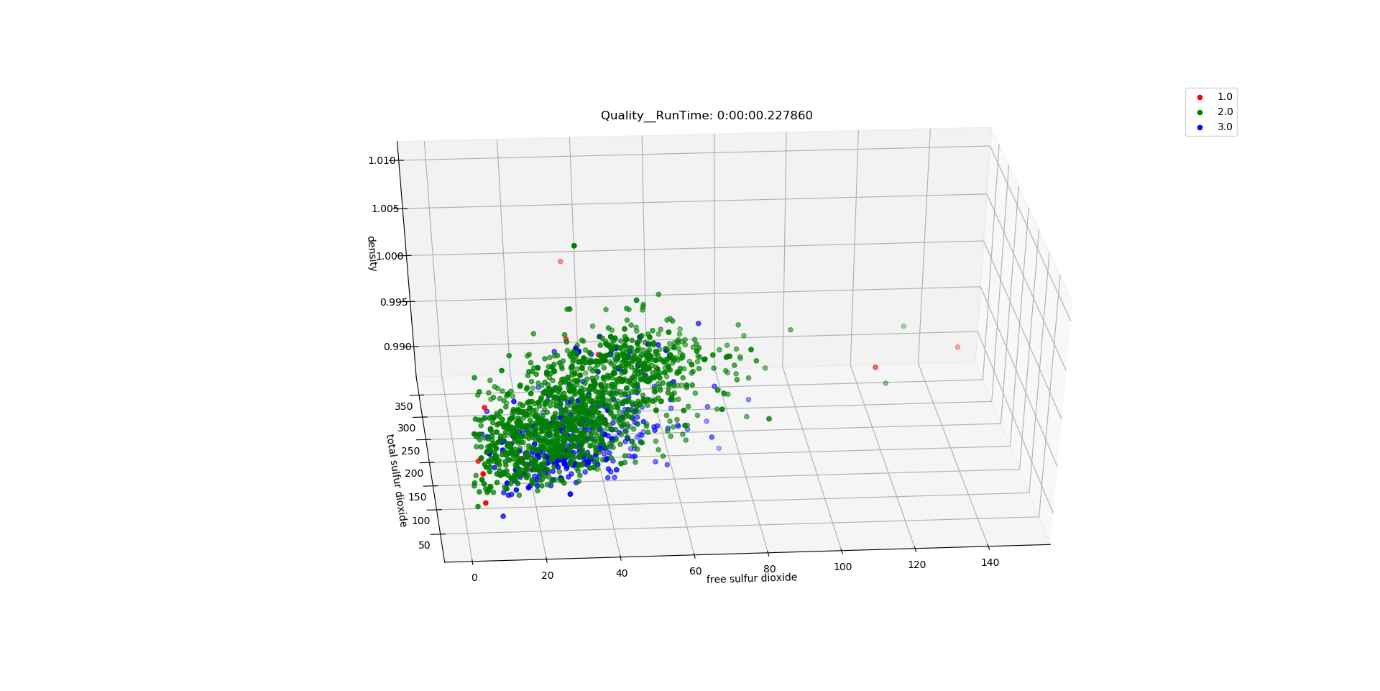
### Divided quality into range



### View the relationship between quality in range and free sulfur dioxide/total sulfur dioxide(Actual Categorization)



We try to find some extra info from using K-means algorithm. Now we are binning the quality from 3-9 into 3 ranges, quality [3-4) to range 1; quality [4-7) to range 2, quality [7-9) to range3.



When we use the ” quality ranges” to color the plot, the figure changes easier to be known, for example the plot of quality range 1 is discrete from the main group of data, and the other plots are concentrated in a group around 5-80 in X-label and 40-250 in Y-label and 0-1.1 in z-label.

But all the plot of quality range 2 and 3 are mixed together, we might need other way to find the regular.

## DBSCAN

# Appendix