FINAL PROJECT – WINE QUALITY

Özgün Özerk

WINE ATTRIBUTES

In this project, 10 attributes of both red-white wines are being taken into account to predict their qualities.

11 Attributes and their brief explanation:

1- **Fixed acidity:**
Acids are major wine constituents and contribute greatly to its taste. In fact, acids impart the sourness or tartness that is a fundamental feature in wine taste. Total acidity is divided into Fixed (non-volatile) and Volatile acidity. Most common fixed acids are tartaric, malic and citric.

2- **Volatile acidity:**
Volatile acidity is mostly caused by bacteria in the wine creating acetic acid, which gives vinegar its characteristic flavour and aroma.

3- **Citric acid:**
Citric acid is one of the fixed acids, and used far less frequently than tartaric and malic due to aggressive citric flavours it can add to the wine.

4- **Residual Sugar:**

Residual sugar refers to any natural grape sugars that are left over after fermentation ceases. It carries the main role in how sweet the wine tastes.

5- **Chlorides:**
Wine contains from 2 to 4g of salts of mineral acids, along with some organic acids, and they may have a key role on a potential salty taste of a wine, with chlorides being a major contributor to saltiness.

6- **Free sulphur dioxide:**
Sulphites or sulphur dioxide is a fruit preservative widely used in dried fruits as well as wine. They are mainly used as preservatives in wines. There exist many erroneous ideas about sulphites, so to put the record straight: All wines contain sulphite. Yeast naturally produce sulphites during fermentation so there is only a rare wine which contains none. Except for the people who have allergic reaction to sulphites, the sulphite amount in wine is not harmful, in fact, there are more sulphites in fruits than there are in wine.
Secondly, there is no medical research data showing that sulphites cause headaches.

7- **Total sulphur dioxide:**
Only a proportion of the SO2 added to a wine will be effective, as an anti-oxidant. The rest will combine with other elements in the wine and cease to be useful. The part lost into the wine is said to be bound, the active part to be free

8- **Density (self-explanatory)**

9- **PH (acidity level from a scale 0 to 14)**

10- **Alcohol (amount of alcohol, given in percentage)**

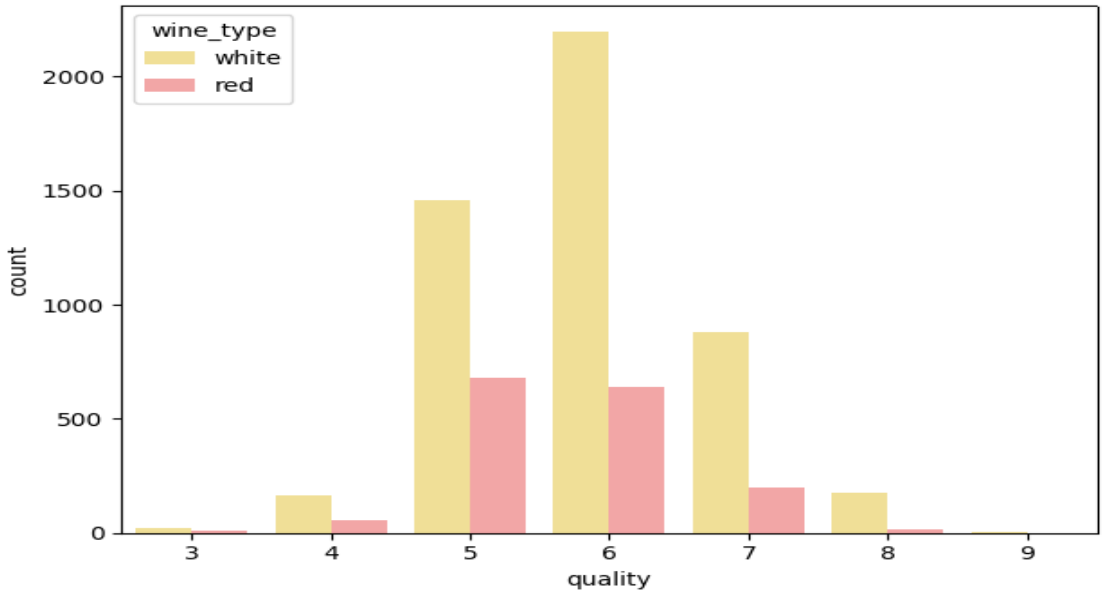


Fig 1.1 (Amount of white and red wines, and their rank of qualities)

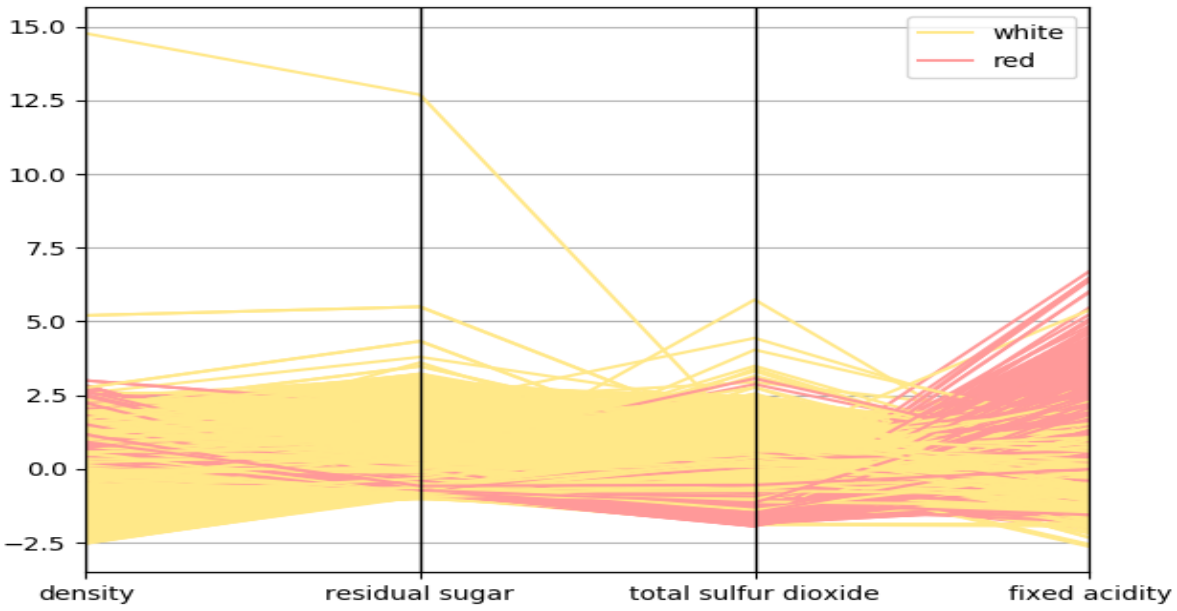


Fig 1.2 (Characteristics of red and white wine compared, regarding to *density, residual sugar, total sulphur dioxide* and *fixed acidity*)

K-MEANS (4 CENTROID) TO CLUSTER

Wine dataset does not have enough number of good and bad quality wines, huge part of the dataset consists of normal quality ones. Also, some of the attributes might have more indirect impact on the result. Though all the attributes that are explained in the upper part hold some importance regarding different aspects and one should not dismiss any of them if it is possible. Hence, the clusterization process of this dataset might not bear the best results, but still, they should give us an idea about the dataset.

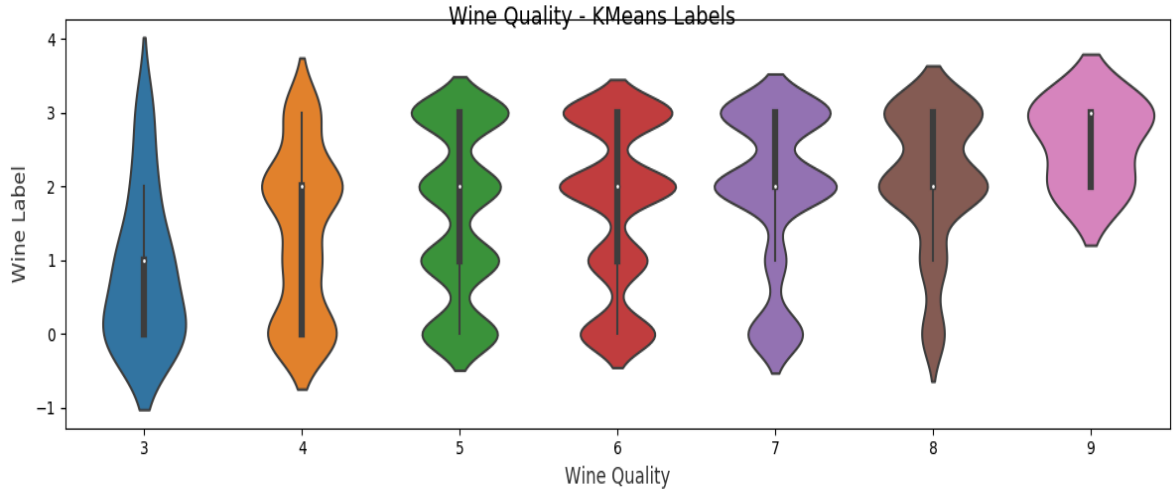


Fig 2.1 (Run 1: Violin plot for Wine Labels distribution on Wine Quality)

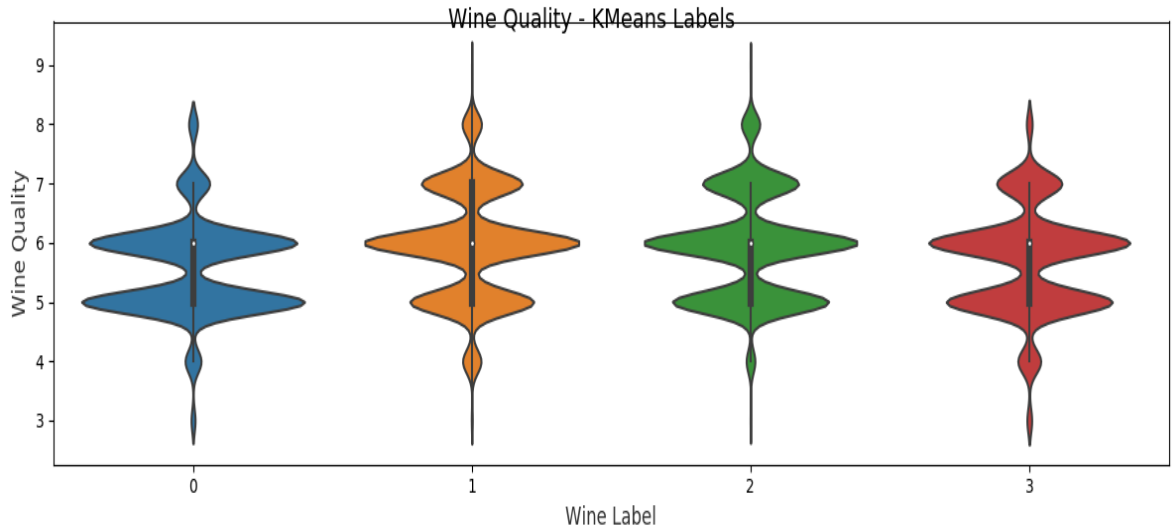


Fig 2.2 (Run 2: Violin plot for Wine Quality distribution over Labels)

Percentages for Fig 2.2:

Label 0:

- % 0,81 Quality 3
- % 3.24 Quality 4
- % 45.23 Quality 5
- % 42.16 Quality 6
- % 6.76 Quality 7
- % 1.80 Quality 8
- % 0.00 Quality 9

Label 1:

- % 0.20 Quality 3
- % 3.83 Quality 4
- % 25.77 Quality 5
- % 45.07 Quality 6
- % 21.01 Quality 7
- % 4.03 Quality 8
- % 0.10 Quality 9

Label 2:

- % 0.15 Quality 3
- % 1.76 Quality 4
- % 31.48 Quality 5
- % 43.35 Quality 6
- % 19.39 Quality 7
- % 3.72 Quality 8
- % 0.15 Quality 9

Label 3:

- % 1.03 Quality 3
- % 4.91 Quality 4
- % 35.63 Quality 5
- % 43.18 Quality 6
- % 14.00 Quality 7
- % 1.24 Quality 8
- % 0.00 Quality 9

Reason for using 4 centroids, although we have 7 different quality classes (from 3 to 9), is to cluster wines into more generalized groups, such as: low-quality, medium-quality, good-quality, very-good quality. Results for 11 centroids were too complicated to interpret or maybe even meaningless, since there might be so many relationships between 10 attributes.

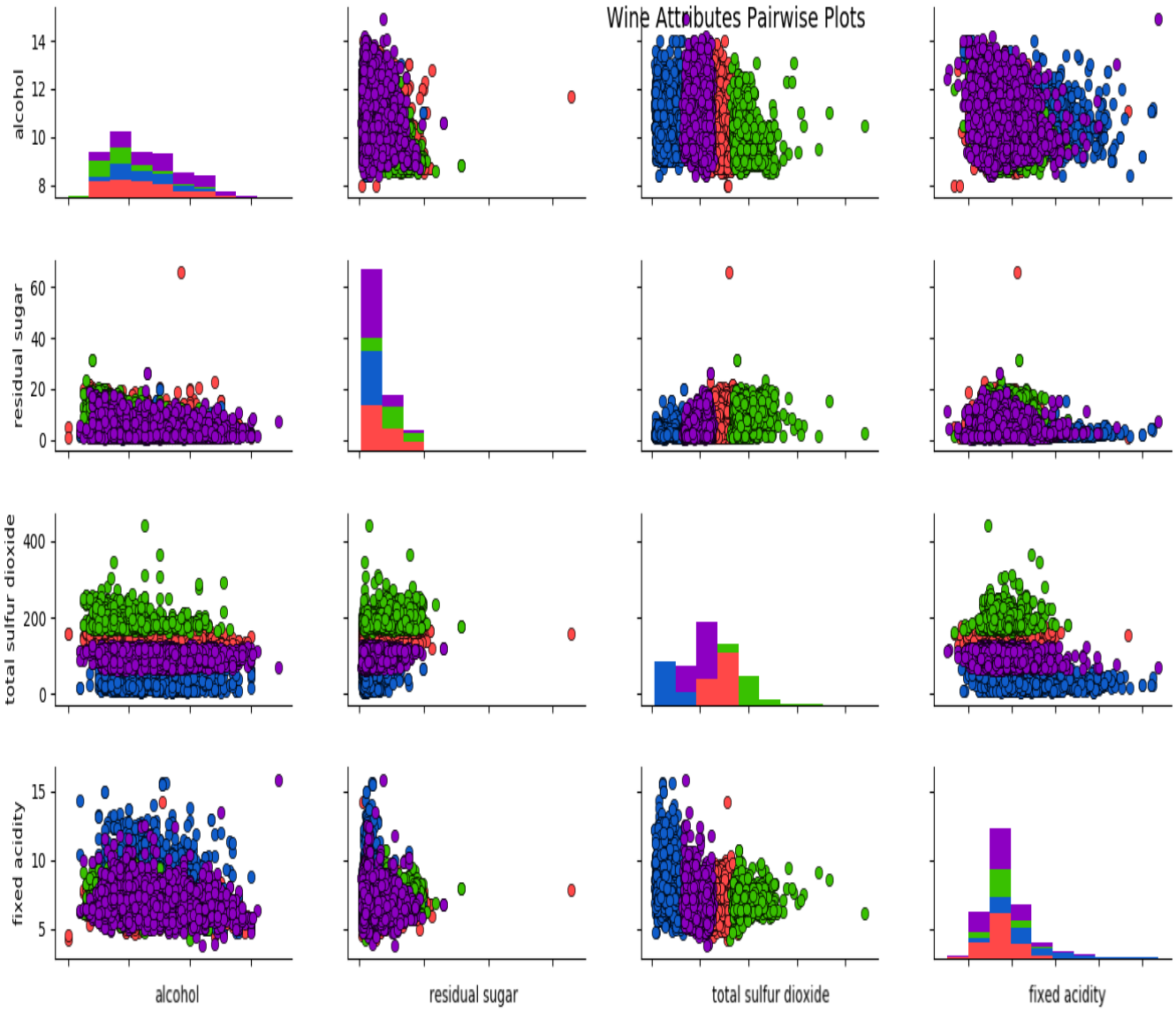


Fig 2.3 (Data representation regarding *alcohol, residual sugar, total sulphur dioxide* and *fixed acidity*

***Red-0, Blue-1, Green-2, Purple-3*)**

Interpretation of the result:

The fact that dataset has very few samples for bad and high-quality wines should not be forgotten. K-means algorithm from "*sklearn library for python 3.6"* with 4 centroids managed to separate best quality ones (quality 8-9) into label 1 and 2. And the worse and average quality ones into label 0 and 3.

From these results, it can be seen that total sulphur dioxide certainly one of the major contributors to wine quality. Amount of total sulphur dioxide should be either low or high (not moderate) to increase the possibility to obtain a better wine quality.

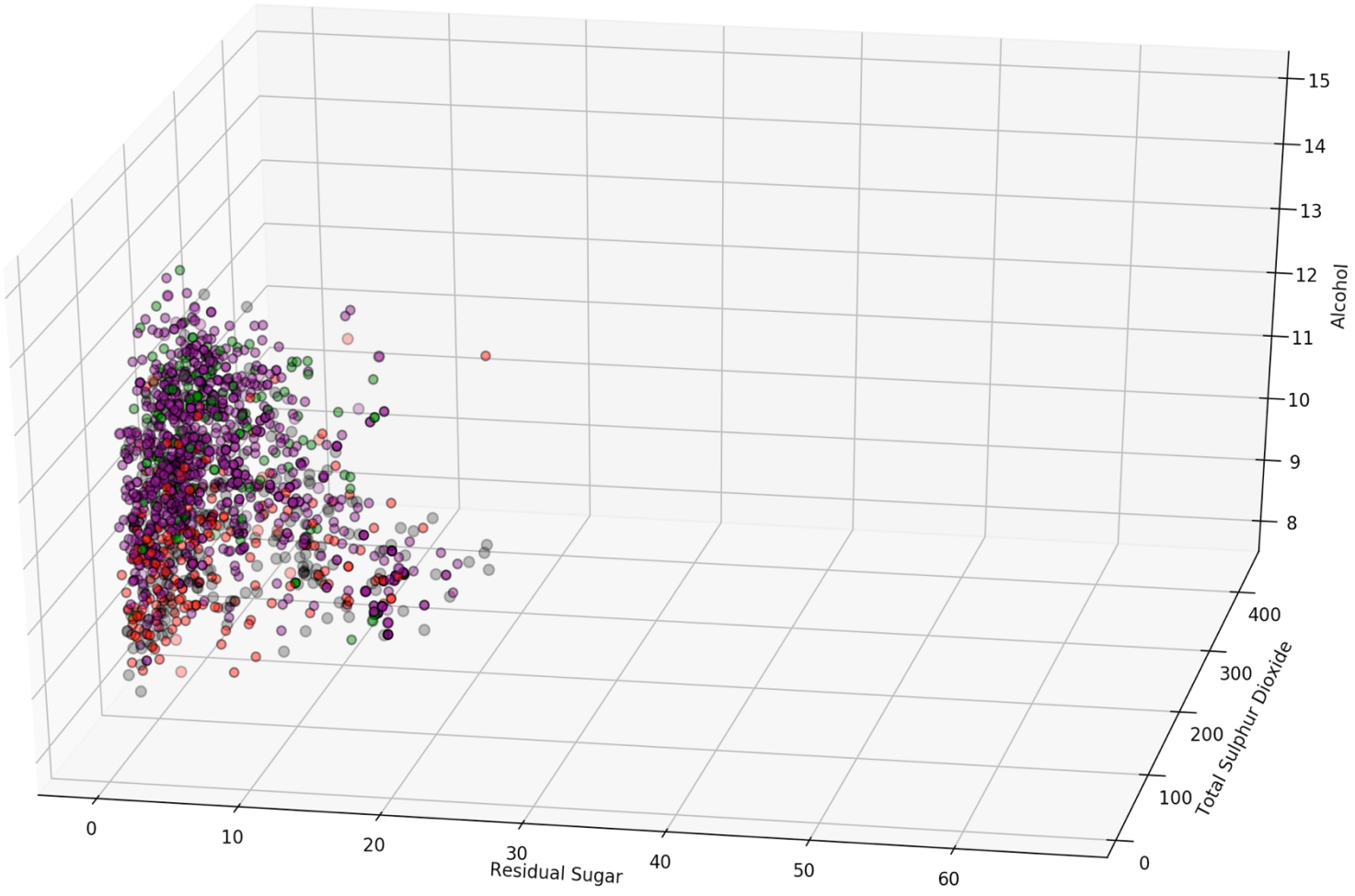
KNN

Running K-Nearest Neighbours algorithm on this dataset showed that, in guidance of previously specified 10 attributes, it is possible to predict the quality of a wine. Which can help in real life as such: Select the possible best new wine to try from a wine set that consists of wines we haven't tried yet.

	K = 3	*K = 5*	*K = 7*	*K = 10*	*K = 20*
Mean-Squared Error	0.8438	0.7415	0.7146	0.673	0.65
Variance Score					

Table 3.1 (Mean-Squared Error and Variance Score for different K values)

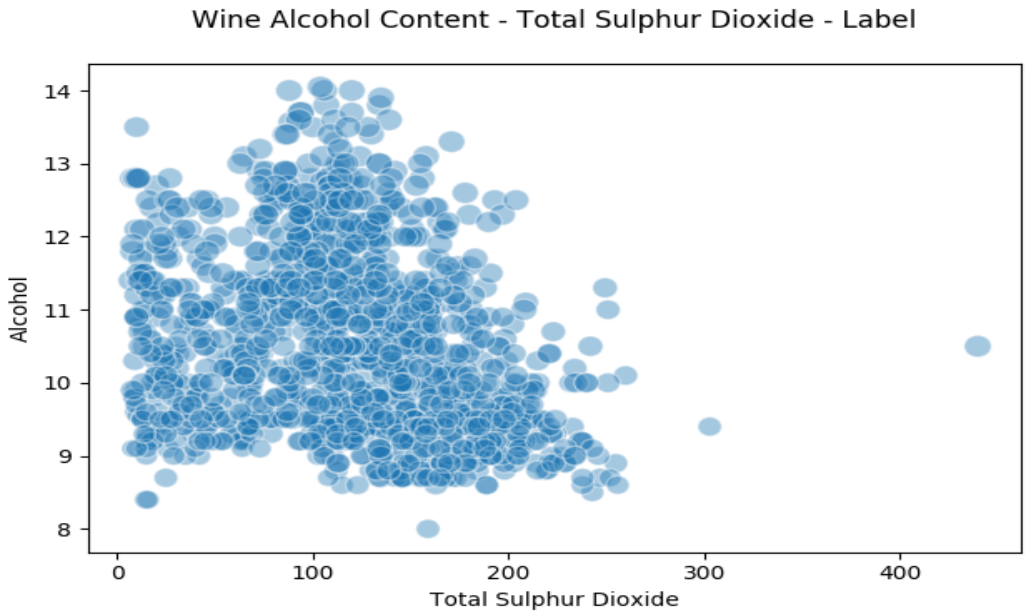
As it can be seen from the table, efficiency is not dramatically increases between K=10 and K=20 although we are doubling the computational work by doubling the K, in compare to the difference between K=5 and K=10



LINEAR REGRESSION

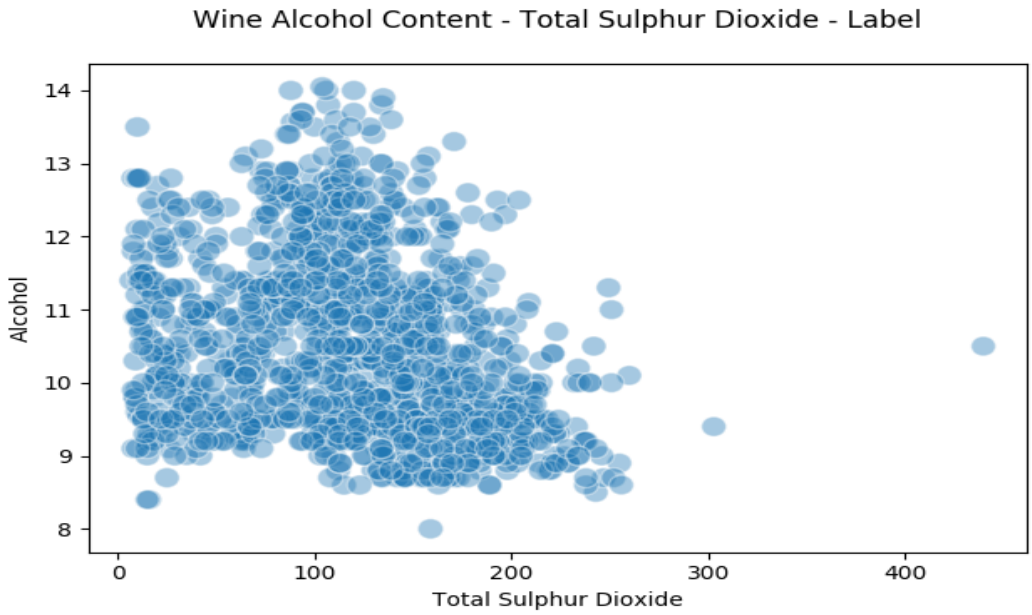
Using the knowledge we collected from K-Means *-total sulphur dioxide is being the most effective factor on determining quality and combination with fixed acidity providing the best clusterization graph-,* it will be worth to see the comparison between Linear Regression run for all the attributes and another instance of Linear Regression run for based on only total sulphur dioxide amount and fixed acidity level in wines.

All attributes:



Coefficients	0.065	0.234	0.017	0.130	0.002	0.115	0.127	-0.039	0.071
Mean-Square	0.560								
Variance Score	-1.497								

Total Sulphur Dioxide and Fixed Acidity:



Coefficients	-0.082						-0.053		
Mean-Square	0.749								
Variance Score	-101.052								

RESOURCES

Neeley, E. (2004). What's in Wine [webpage]. Retrieved from https://waterhouse.ucdavis.edu

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Walker et al., 2003; Maltman, 2013.