Python final project – Wine

Dekel Menashe 311224117

Intro

Overview:

The project subject is Wine.

I got a dataset that contain a wine parameters and ingredients of the wine and the target is the wine quality.

compare several classification algorithms to predict wine quality which has a score between 0 and 10. from the dataset we will find out what makes a good wine by using the data source.

The data set of the project contain a physicochemical tests of wines by multi variables

That based on sensory data will score a quality between 0 and 10.

The dataset I got for the project contain the following:

14 columns and 6498 rows.

1 – sample the id of each row.

2 - fixed acidity: numeric, continuous - base level of acidity

3 - volatile acidity: numeric, continuous - varying level of acidity

4 - citric acid: numeric, continuous - measure of citric acid

5 - residual sugar: numeric, continuous - level of sugars

6 - chlorides: numeric, continuous - levels of chloride

7 - free sulfur dioxide: numeric, continuous - free particles, sulfor dioxide

8 - total sulfur dioxide: numeric, continuous - total sulfor dioxide

9 - density: category - liquid density rating

10 - pH: numeric, continuous - pH level

11 - sulphates: numeric, continuous - sulphate count

12 - alcohol: numeric, continuous - percentage of alcohol

13 - kind: category - red or white wine

Output variable (based on sensory data):

14 – quality: numeric (score between 0 and 10)

There are 2 categorical columns: density and kind.

-density have the categories: very high, high, medium, low.

-kind have the categories: white, red.

The remaining 12 contain numeric values.

Initial Data Analysis

I started by checking the data, I printed the first 5 rows, and checked that the data can load and is readable.

	sample	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	kind
0	0	7.0	0.27	2.16	20.7	0.045	165.0	170.0	very high	3.00	0.45	8.8	6	white
1	1	6.3	0.30	2.04	1.6	0.049	136.0	132.0	medium	3.30	0.49	9.5	6	white
2	2	8.1	0.28	2.40	6.9	0.050	152.0	97.0	high	3.26	0.44	10.1	6	white
3	3	7.2	0.23	1.92	8.5	0.058	167.0	NaN	high	3.19	0.40	9.9	6	white
4	4	7.2	0.23	1.92	8.5	0.058	171.0	186.0	high	3.19	0.40	9.9	6	white

Next I checked the describe command to show statistic on the data and checks for trends and irregularities:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
count	6497.000000	6425.000000	6497.000000	6497.000000	6497.000000	6497.000000	6298.000000	6432.000000	6497.000000	6497.000000	6497.000000
mean	7.215307	0.339553	1.864781	5.443235	0.056034	146.881484	115.586615	3.235340	0.531268	10.491801	5.818378
std	1.296434	0.164185	0.889913	4.757804	0.035034	25.742084	56.478534	0.436573	0.148806	1.192712	0.873255
min	3.800000	0.080000	0.000000	0.600000	0.009000	59.000000	6.000000	2.720000	0.220000	8.000000	3.000000
25%	6.400000	0.230000	1.350000	1.800000	0.038000	129.000000	77.000000	3.110000	0.430000	9.500000	5.000000
50%	7.000000	0.290000	1.860000	3.000000	0.047000	147.000000	118.000000	3.210000	0.510000	10.300000	6.000000
75%	7.700000	0.400000	2.380000	8.100000	0.065000	164.000000	155.000000	3.320000	0.600000	11.300000	6.000000
max	15.900000	1.580000	9.960000	65.800000	0.611000	357.000000	440.000000	13.210000	2.000000	14.900000	9.000000

Then I checked for empty values and found out that there is some missing values:

```
sample                  0
fixed acidity           0
volatile acidity       72
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide  199
density                69
pH                     65
sulphates               0
alcohol                 0
quality                 0
kind                  183
dtype: int64
(6497, 14)
```

I removed unnecessary columns, found only one that are unnecessary, the sample that is like an ID for each data but it is taking space and we don't need it.

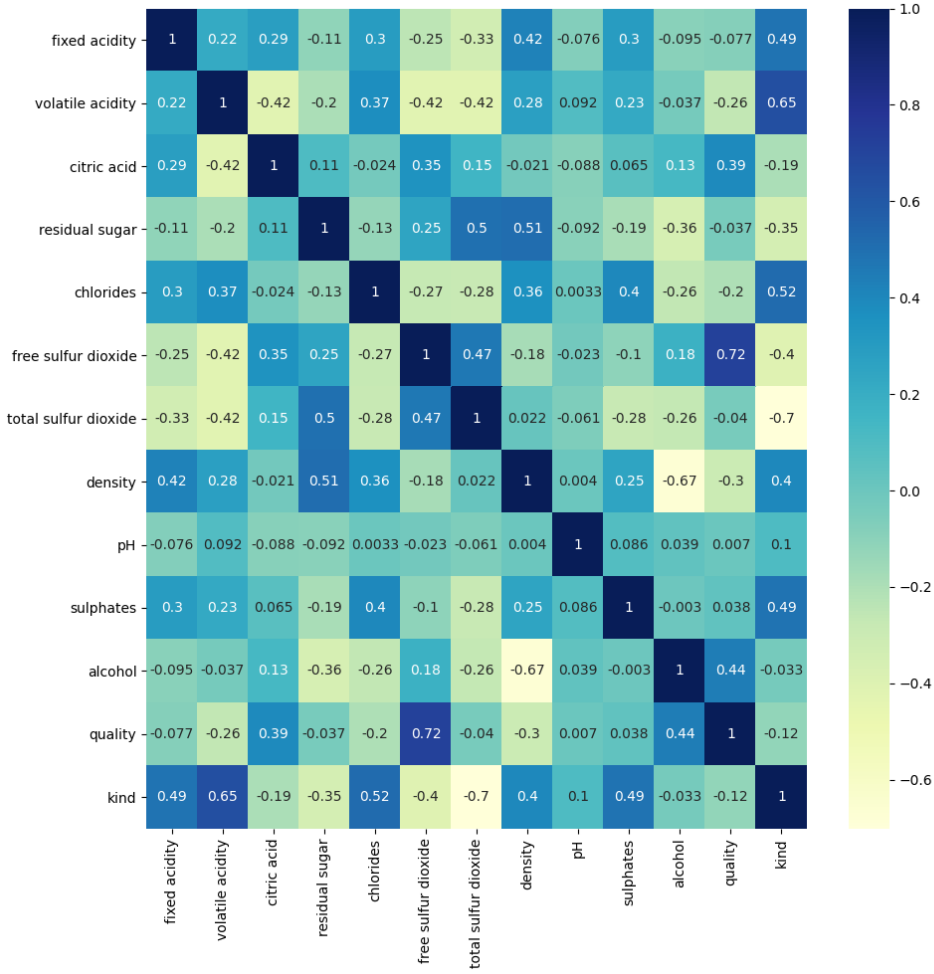
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	kind
0	7.0	0.27	2.16	20.7	0.045	165.0	170.0	very high	3.00	0.45	8.8	6	white
1	6.3	0.30	2.04	1.6	0.049	136.0	132.0	medium	3.30	0.49	9.5	6	white
2	8.1	0.28	2.40	6.9	0.050	152.0	97.0	high	3.26	0.44	10.1	6	white
3	7.2	0.23	1.92	8.5	0.058	167.0	NaN	high	3.19	0.40	9.9	6	white
4	7.2	0.23	1.92	8.5	0.058	171.0	186.0	high	3.19	0.40	9.9	6	white

I changed all the categorial column in the dataset to numerical I had 2 one is nembers by order of the value: low is 1 and very high is 4 and the wine colors are 0 for white and 1 for red.

density
4
2
3
3
3

kind
0
0
0
0
0

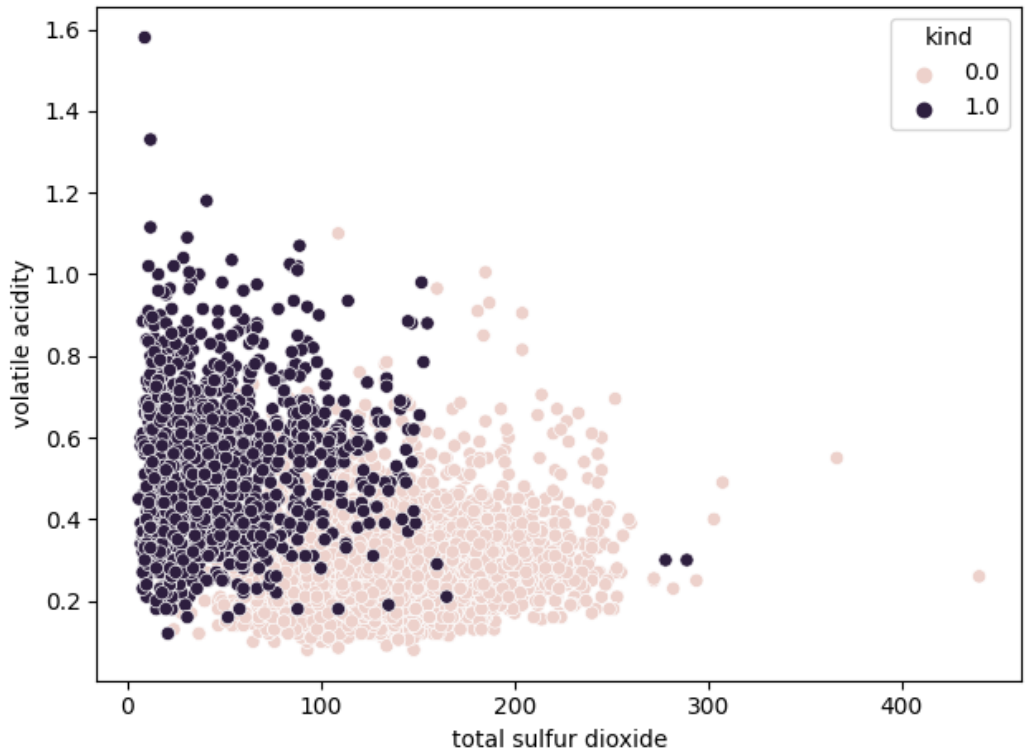
I check for column correlation with an heat map and found out:



Kind is effected by: total sulfur dioxide(-0.70), volatile acidity(0.65), chlorides(0.52), fixed acidity(0.49), sulphates(0.49).
Quality is effected by: free sulfur dioxide(0.72), alcohol(0.44).
Alcohol is effected by: density(-0.67), quality(0.44).
Sulphates is effected by: kind(0.49).
pH has no high correlation.
Density is effected by: alcohol(-0.67), residual sugar(0.52).
Total sulfur dioxide is effected by: kind(-0.7), residual sugar(0.5), free sulfur dioxide(0.47).
Free sulfur dioxide is effected by: quality(0.72), total sulfur dioxide(0.47).
Chlorides is effected by: kind(0.52).
Residual sugar is effected by: density(0.51), total sulfur dioxide(0.50).
Citric acid is effected by: volatile acidity(-0.42).
Volatile acidity is effected by: kind(0.65).
Fixed acidity is effected by: kind(0.49).

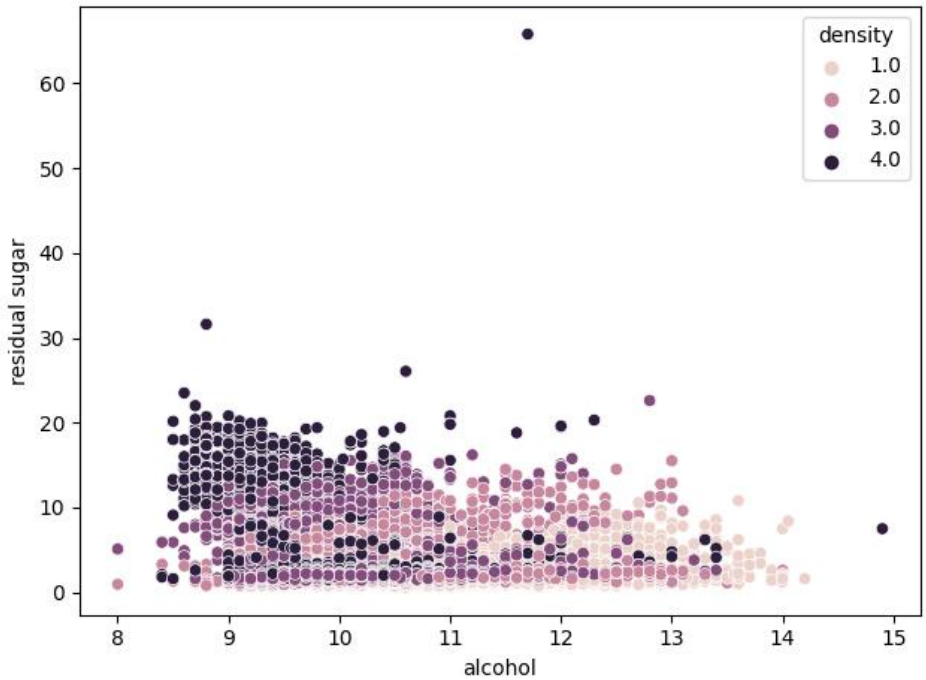
First lets handle the categorial missing values at the dataset:

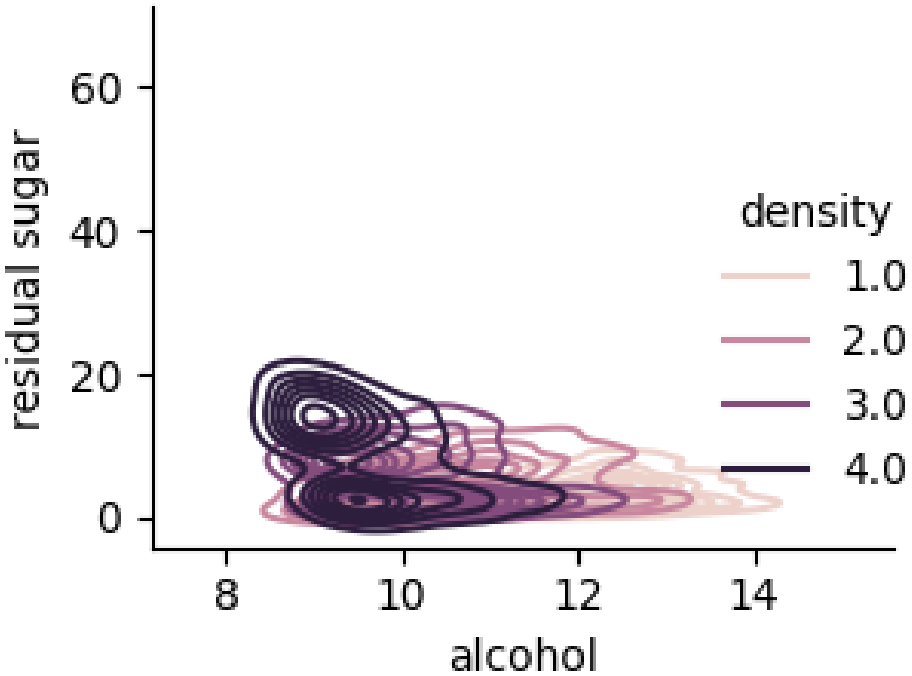
Because “kind” has high correlation with “total sulfur dioxide” and “volatile acidity” I created a scatter plot to visualize the correlation and help me fill the missing values.



as you can see the different between the colors is very pronounced so filling the missing values will be easy when taking into consideration the 2 features.

The density column has high correlation with “alcohol” and “residual sugar”
So I tried to replace the values by these correlation but it is unclear what should be the values, because its only 69 rows I decided to drop them:



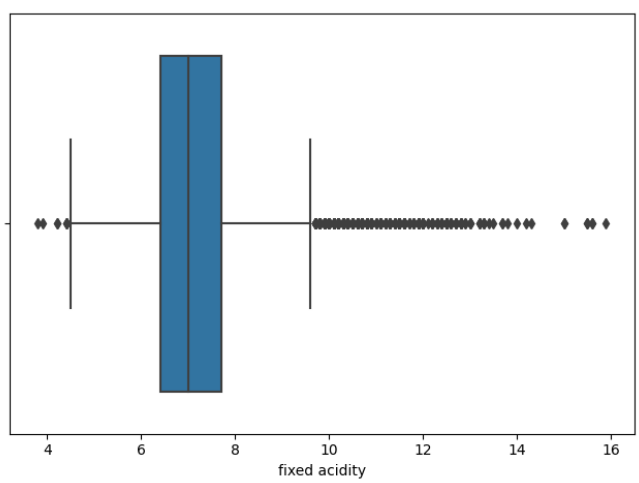


I decided to take all the missing values in the numeric column and insert the mean of the dataset column in each one instead of dropping the rows and losing data in other column.

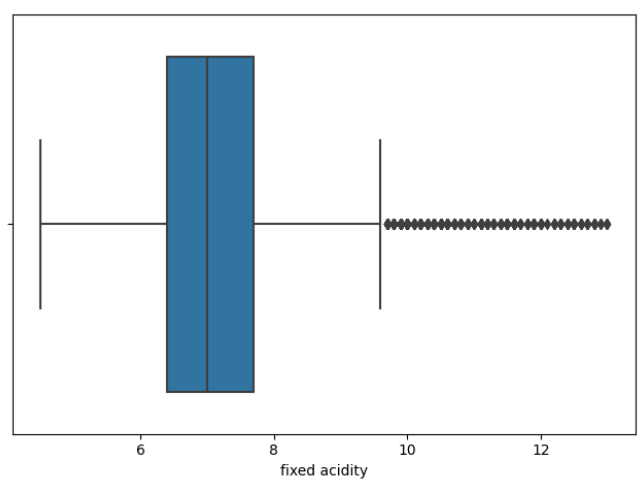
```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
kind                    0
dtype: int64
(6428, 13)
```

I searched for outlines by each column :

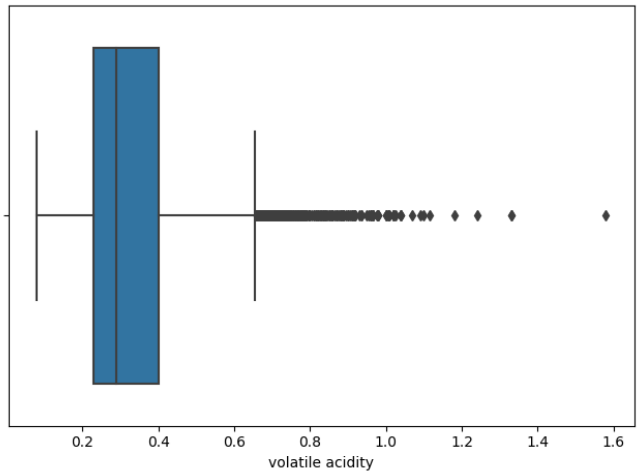
Fixed acidity before:



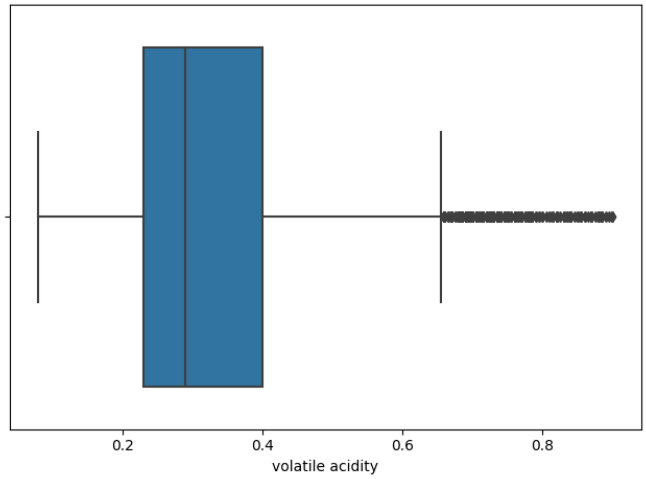
Fixed acidity after:



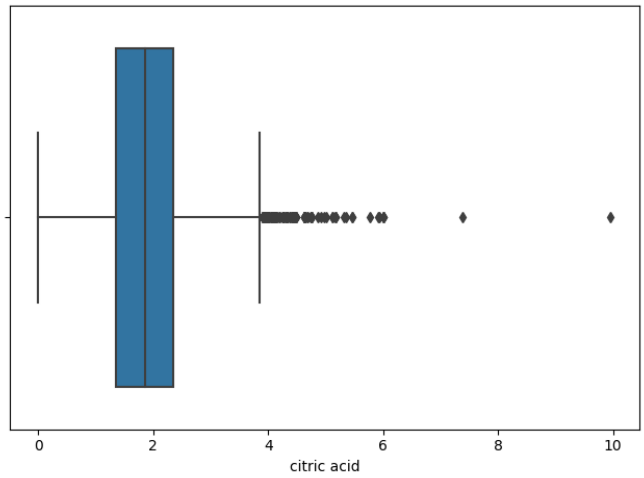
Volatile acidity before:



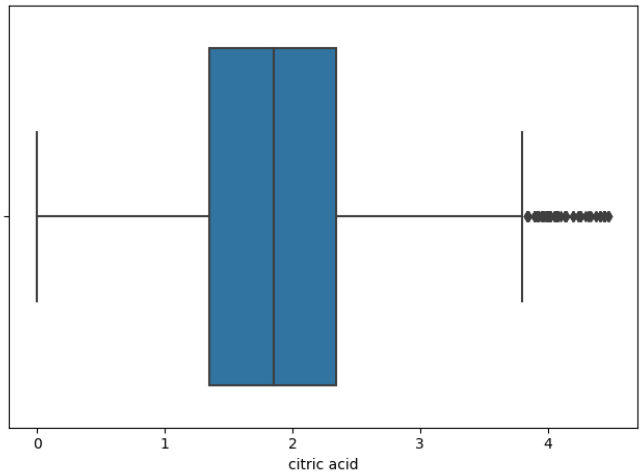
Volatile acidity after:



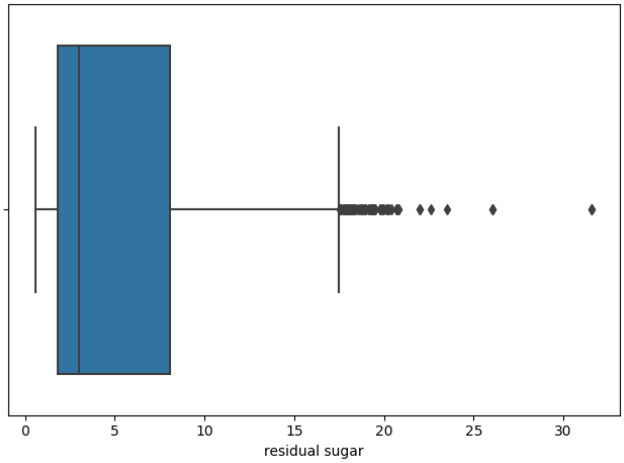
Citric acid before:



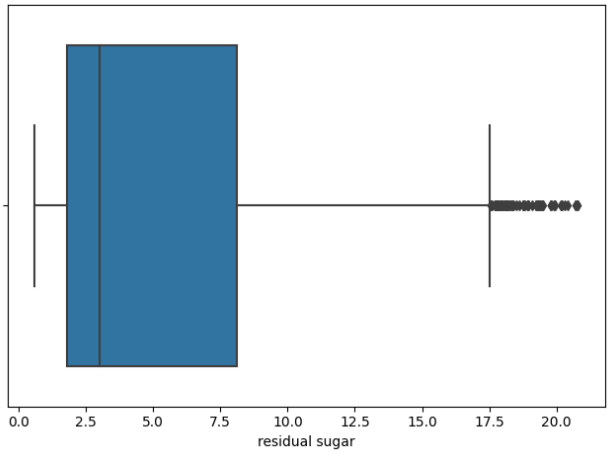
Citric acid after:



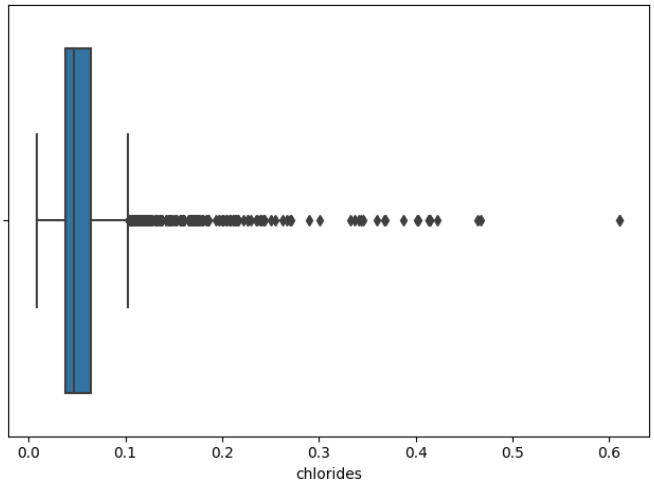
Residual sugar before:



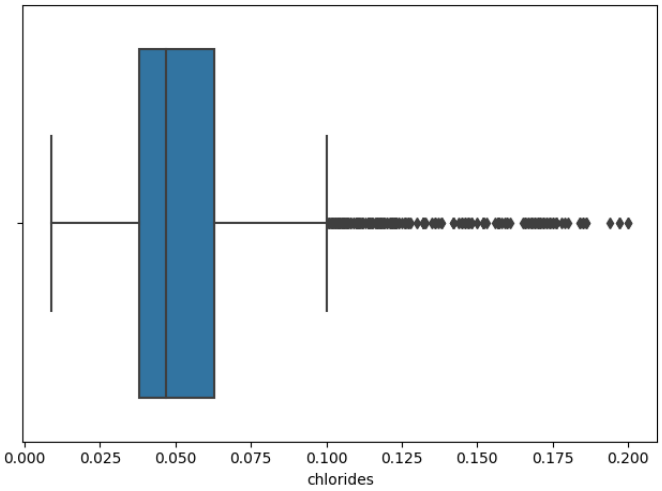
Residual sugar after:



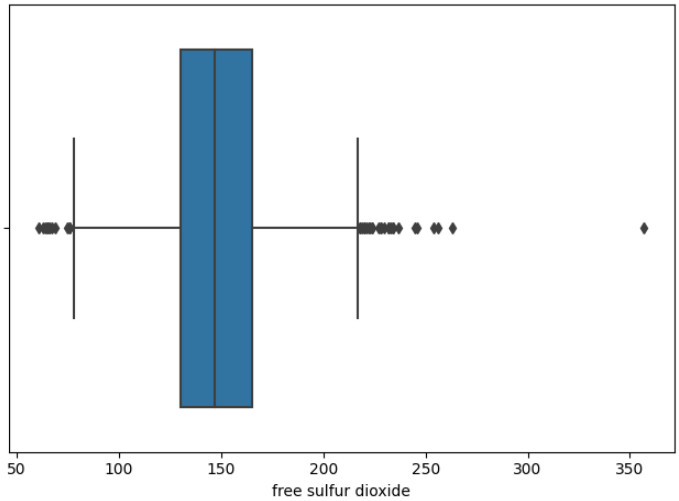
Chlorides before:



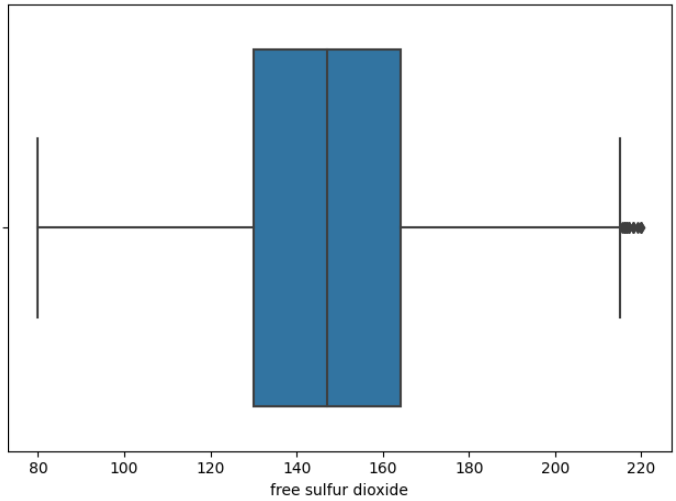
Chlorides after:



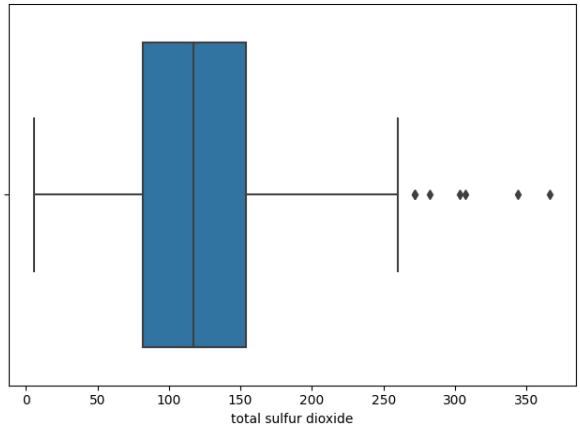
Free sulfur dioxide before:



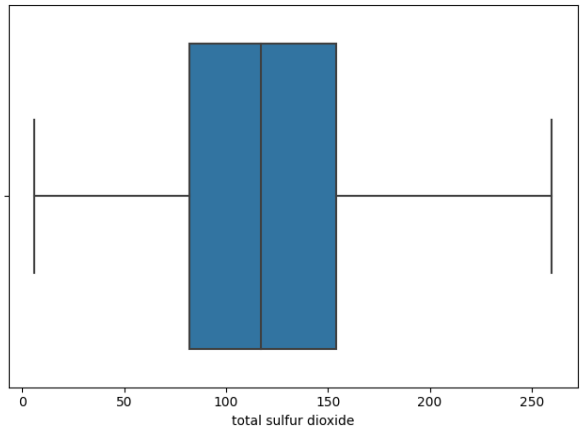
Free sulfur dioxide after:



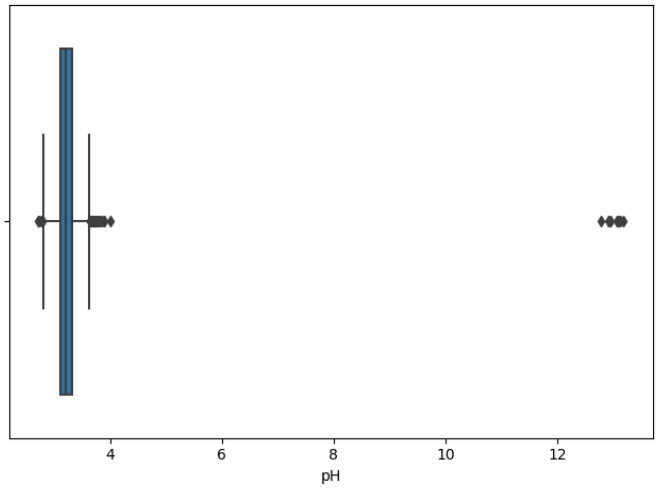
Total sulfur dioxide before:



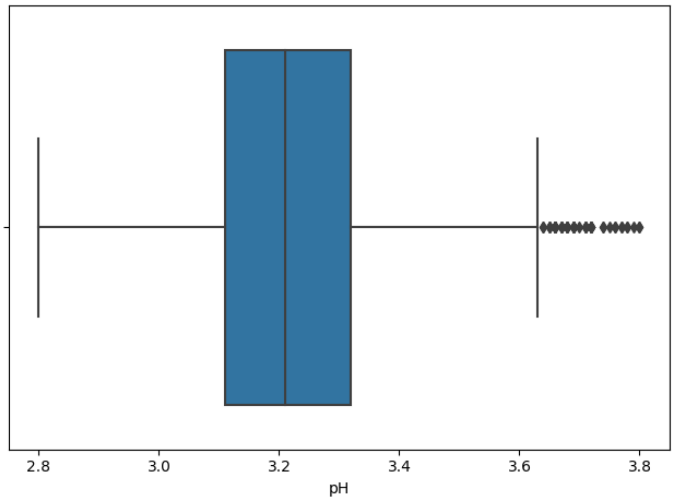
Total sulfur dioxide after:



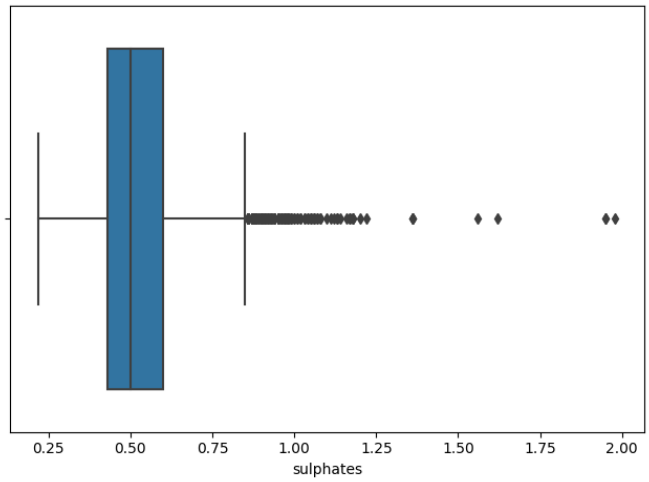
pH before:



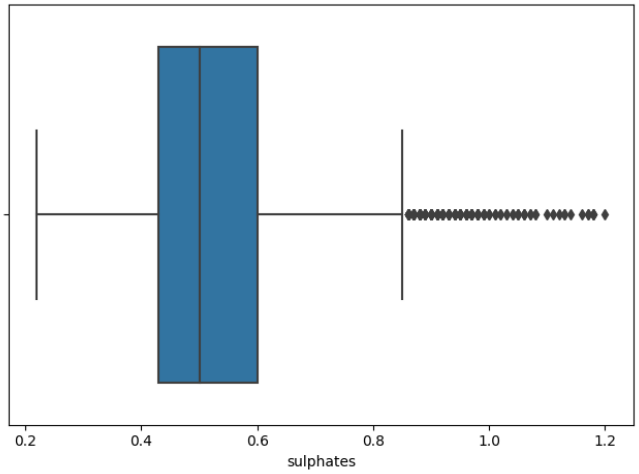
pH after:



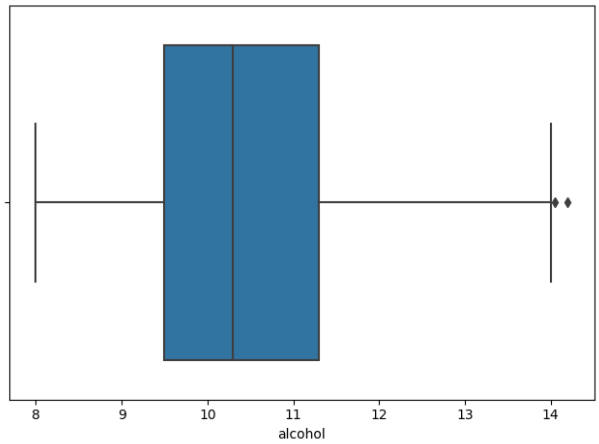
Sulphates before:



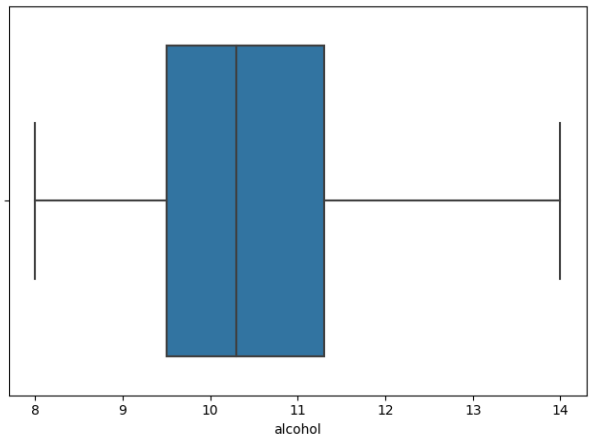
Sulphates after:



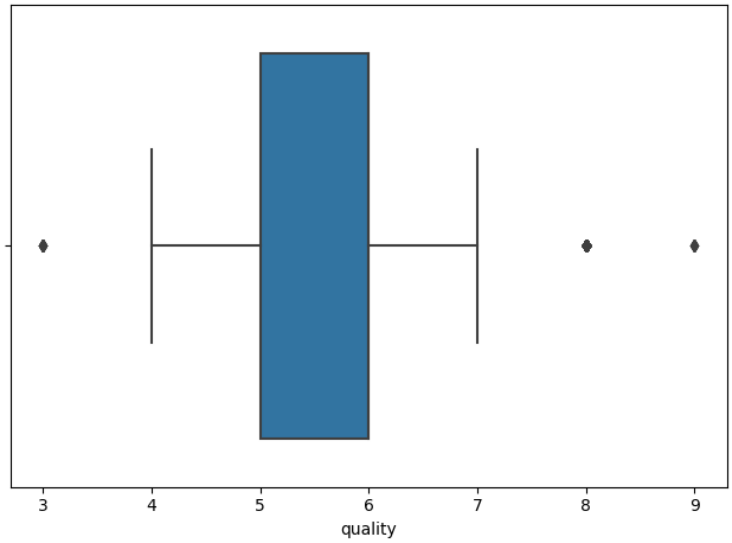
Alcohol before:



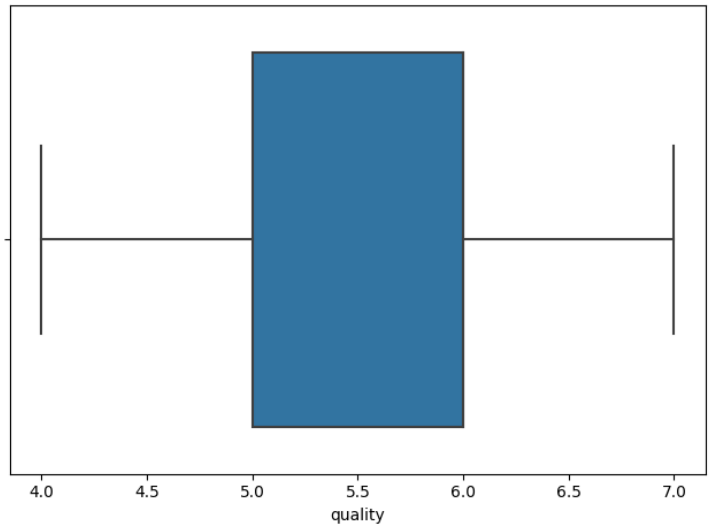
Alcohol after:



Quality before:



Quality after:

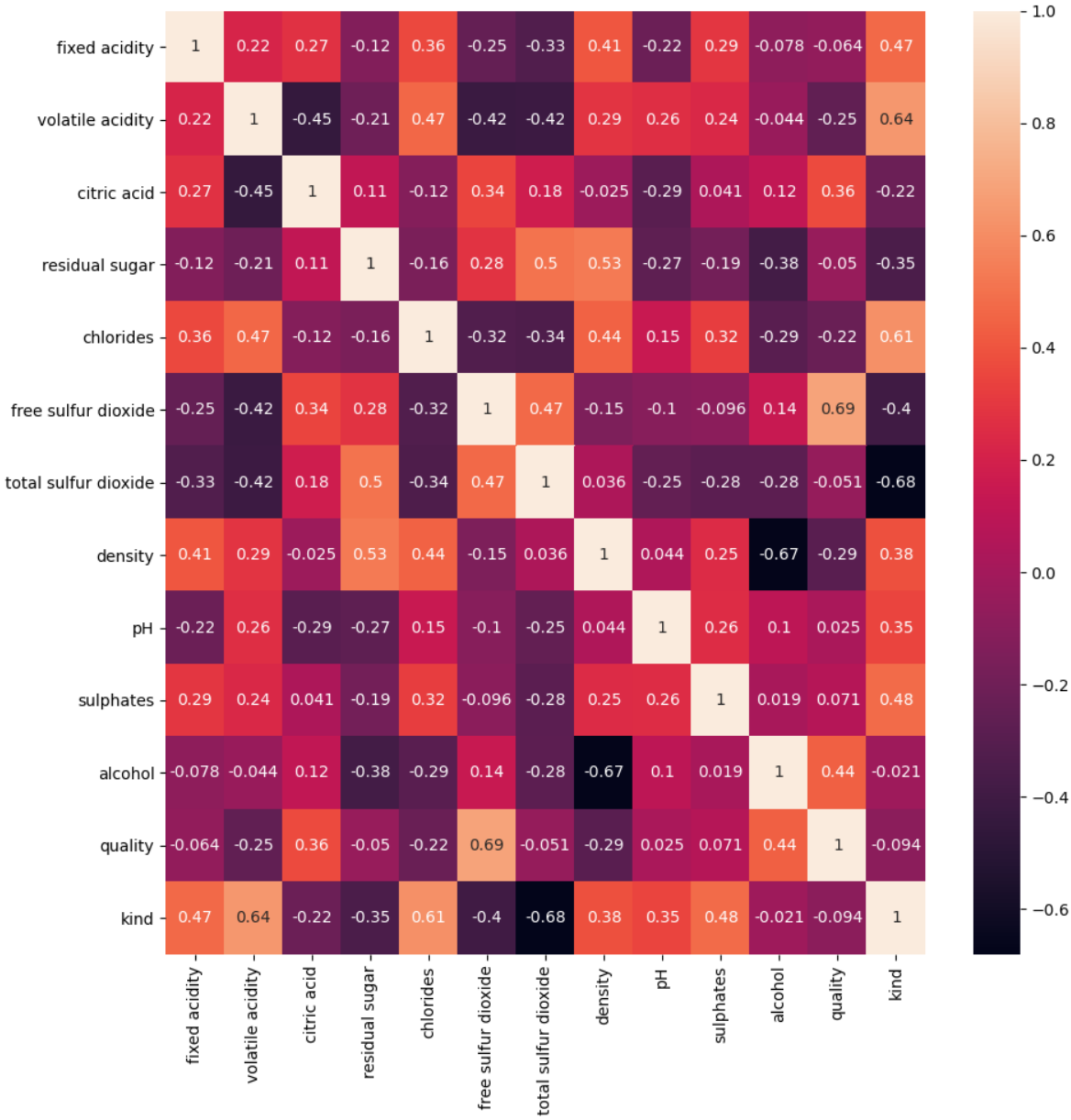


After cleaning all the dataset I left with 5988 rows and 13 column of 6497 and 14 rows which is losing 6.926% of the dataset

```
(5988, 13)
```

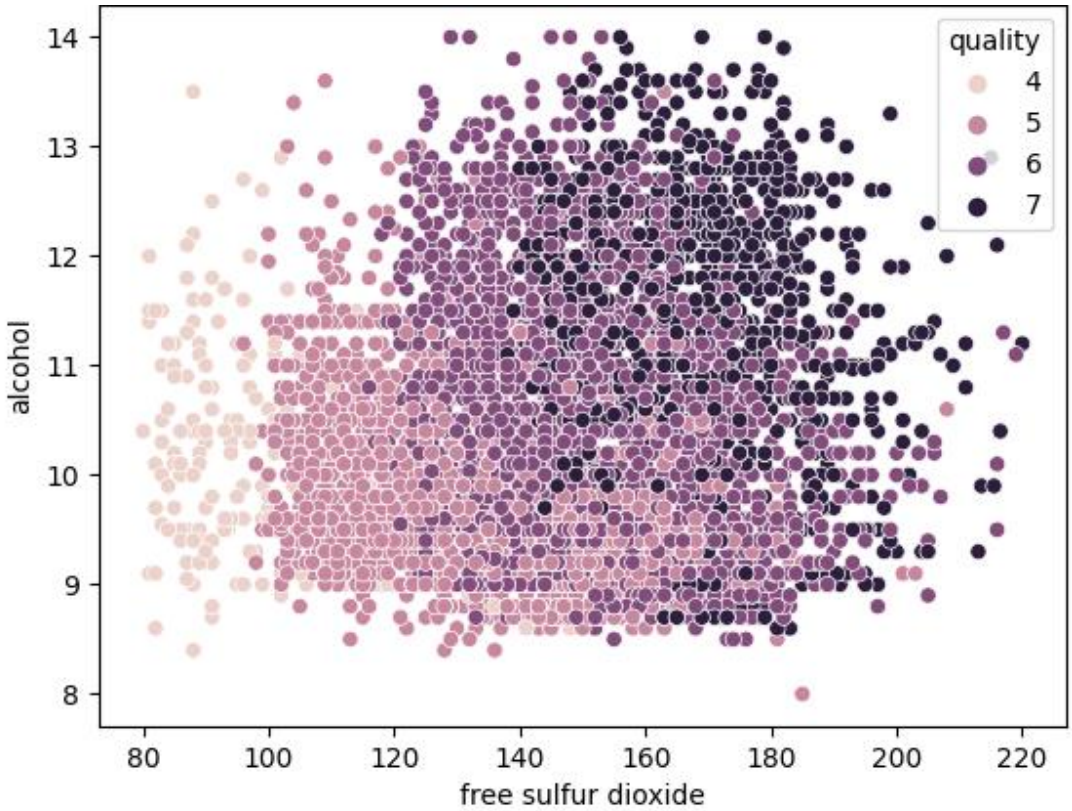
Exploratory Data Analysis

First lets check the target class (quality) correlation in a new heatmap after the data is clean.

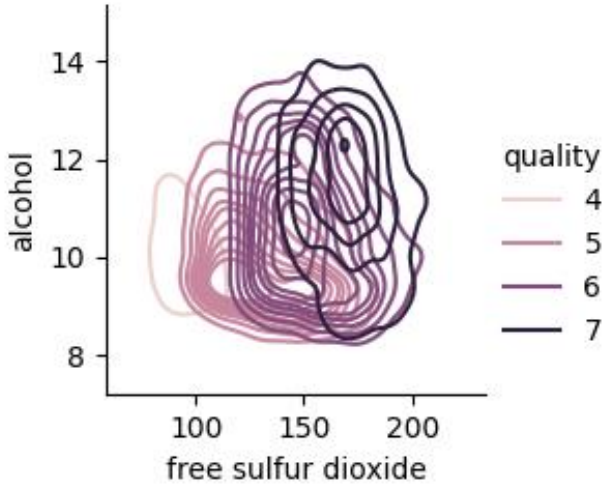


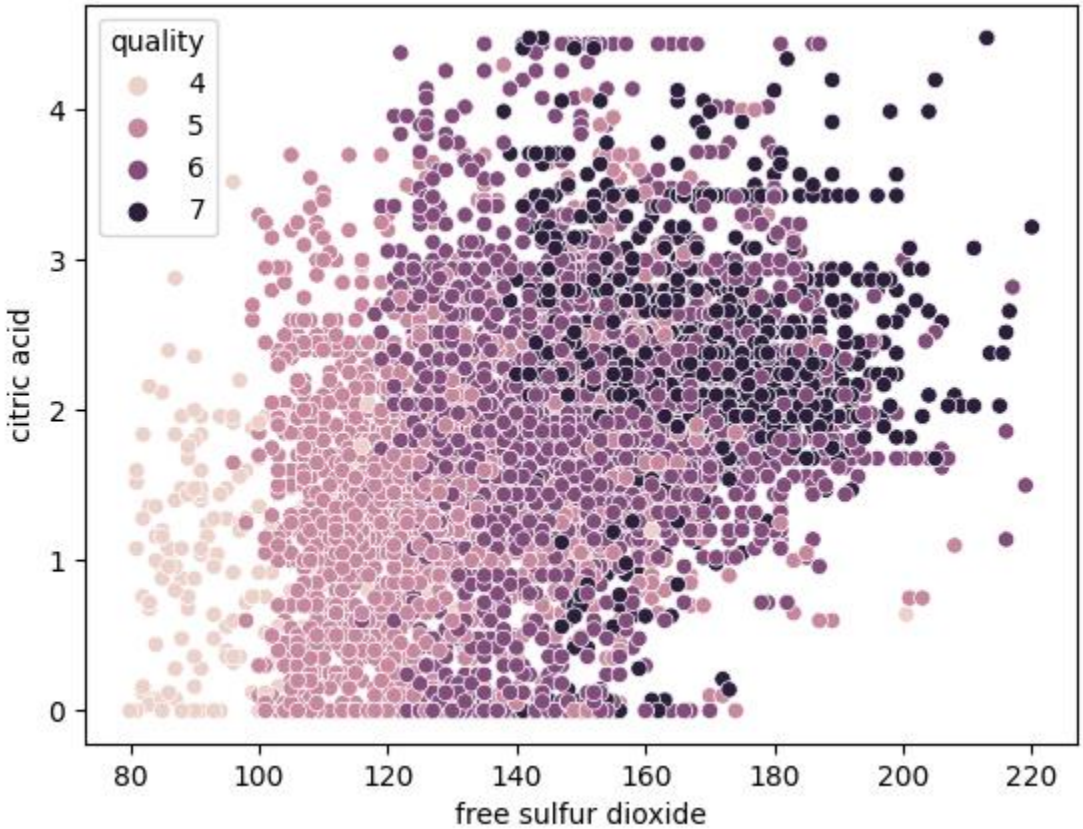
We can see that the target class quality is influenced the most by: free sulfur dioxide 0.69, alcohol 0.44 and citric acid 0.36.

I created a scatterplot to try to understand what's influence the quality of the wine.

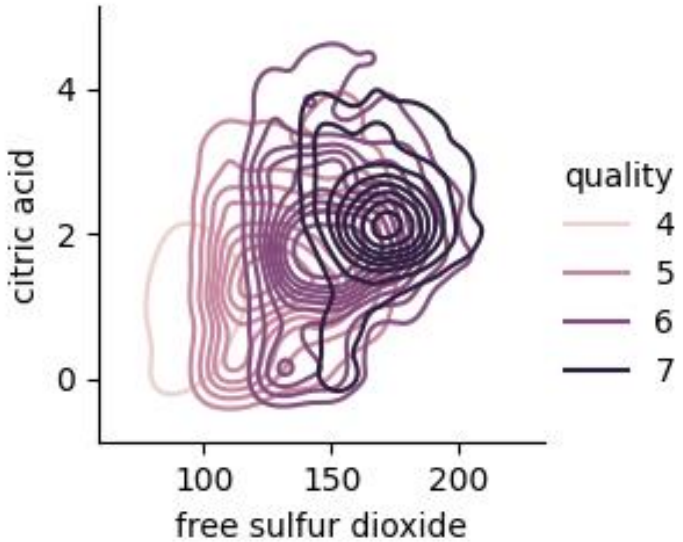


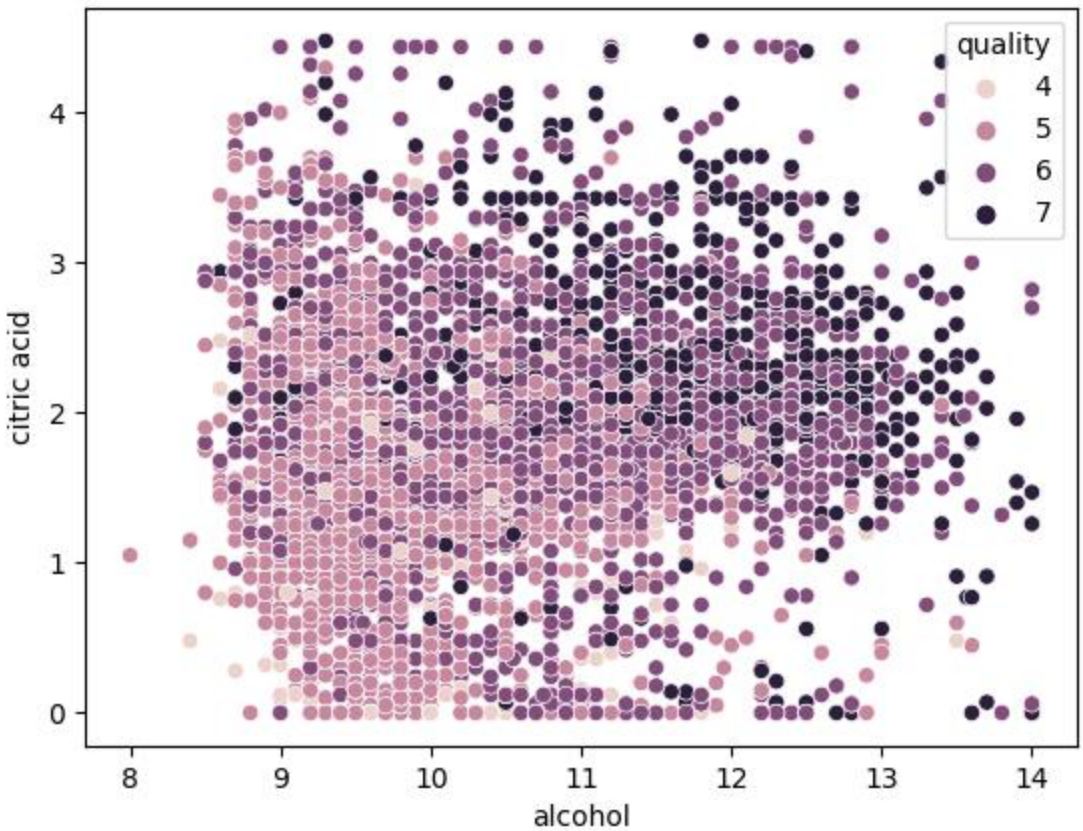
From the first scatterplot we can see clearly in most cases that the higher the free sulfur dioxide and alcohol the higher the wine quality is.
When the free sulfur dioxide is low(80-120) then the quality is either 4 or 5 and when its higher the quality improves and the alcohol bears more weight



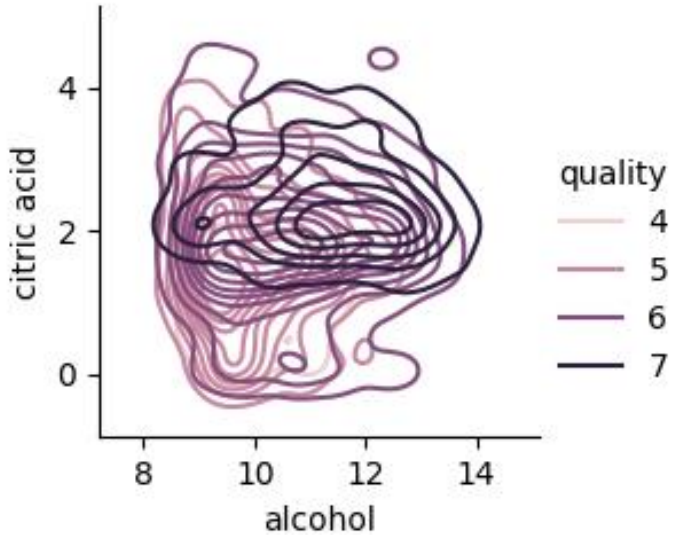


From this plot we can see again that the higher the free sulfur dioxide the higher the quality is, but the citric acid is influencing only when the quality is high (6 and above). From free sulfur dioxide of 140 and above when the citric acid in most cases is more then 1.8 the quality of the wine is 7.





The 3rd plot is not clear as the first 2 but in most cases when the alcohol is more than 10 and the citric acid is more than 1.5 the wine quality is 6 or 7, when the alcohol is more than 10.6 and the citric acid is less then 0.5 the wine quality is still 6 or 7 in most of the cases.



The 2 columns that effect quality the most are **alcohol(0.69)** and **free sulfur dioxide(0.44).**

I wanted to improve the correlation by joining columns.

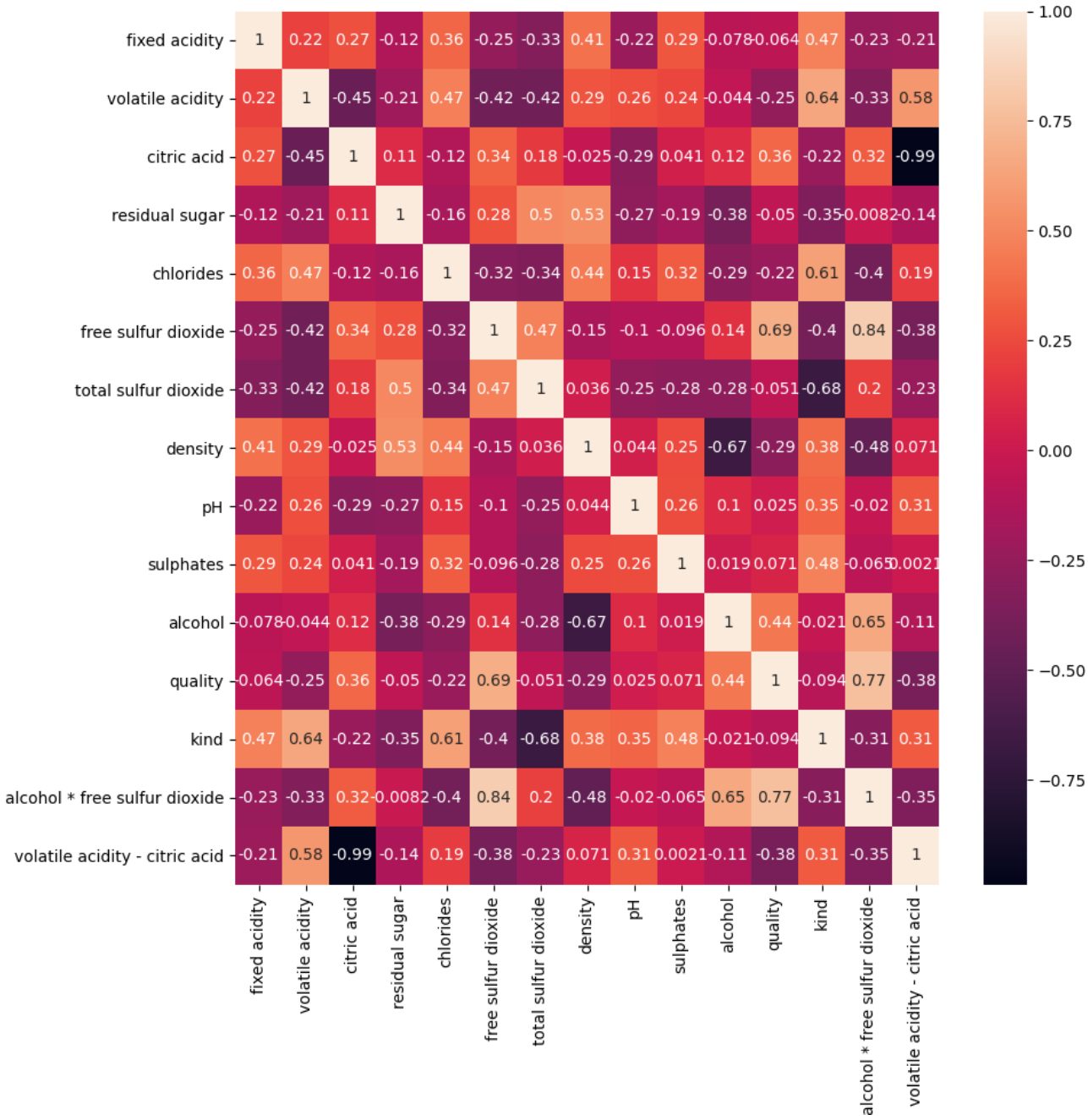
I took the 6 column that effect the quality the most and joined the by these steps:

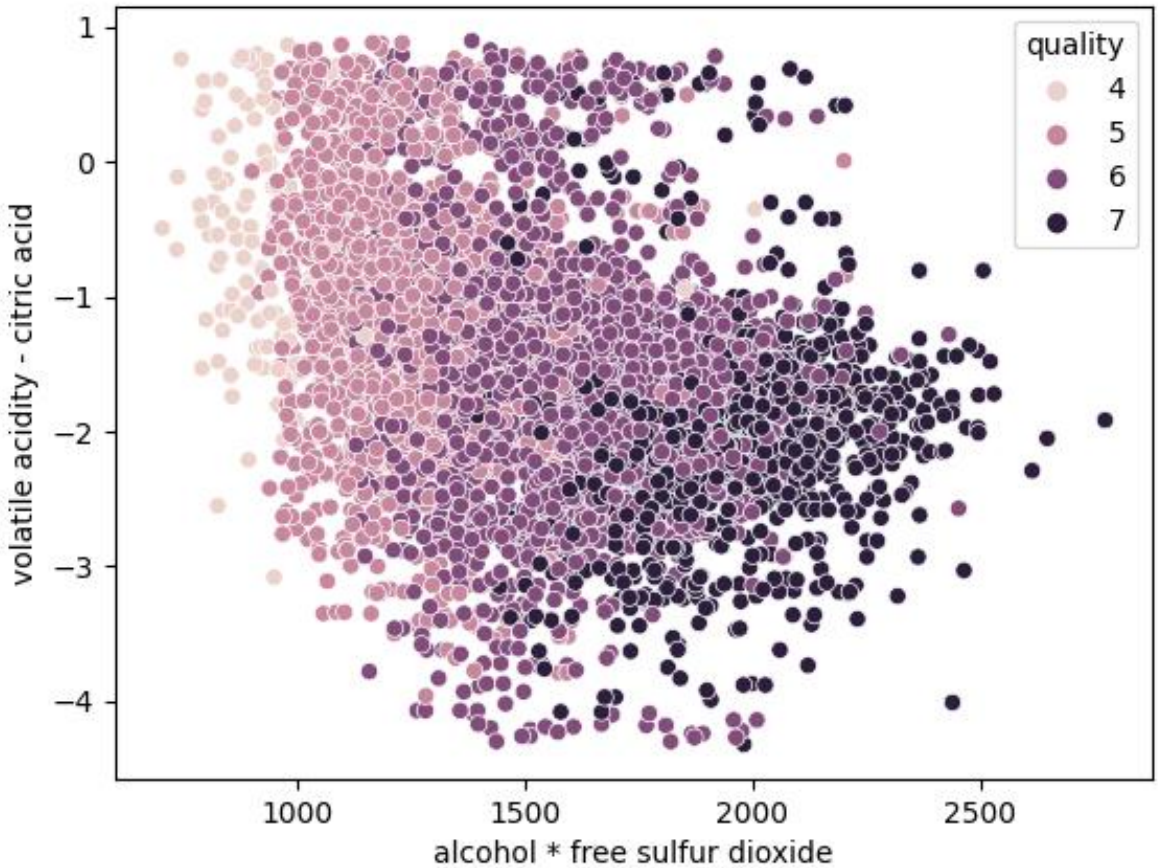
- I joined **alcohol** and **free sulfur dioxide** by multiplicate them.
- I joined **volatile acidity** and **citric acid** by finding the difference.

At the end I got to columns with improved correlation:

alcohol and **free sulfur dioxide** with **0.77** correlation to quality.

volatile acidity and **citric acid** with **0.38** correlation to quality.





As we can see from this scatterplot the alcohol * free sulfur dioxide has its most influence on all the range of the quality.
The volatile acidity – citric acid has no influence on the low quality wine, for getting a 7 grade on the quality in most cases the quality is 7 when the volatile acidity is 2.1 or lower.

Classification Model

1. After handling with the data and correlation, the 2 parameters I choose for the Gaussian Naïve Bayes are the join correlation I made earlier.
   alcohol multiply free sulfur dioxide
   volatile acidity difference citric acid
   lets check the accuracy of this 2 columns I choose
   the results are:

```
alcohol * free sulfur dioxide & volatile acidity - citric acid
0.6917084028937117
              precision    recall  f1-score   support

           4      0.875     0.233     0.368        60
           5      0.684     0.750     0.716       597
           6      0.693     0.710     0.701       834
           7      0.695     0.618     0.654       306

    accuracy                          0.692      1797
   macro avg      0.737     0.578     0.610      1797
weighted avg      0.696     0.692     0.687      1797
```

the accuracy rate is 0.696 so the parameters I choose earlier really are the correct one.

2. Building a decision tree.
   First I took the original data, dropped the empty values rows.
   Dropped the target column, quality and building a tree as we can see below:

	precision	recall	f1-score	support
3	0.375	0.300	0.333	10
4	0.826	0.603	0.697	63
5	0.811	0.820	0.816	650
6	0.776	0.791	0.783	823
7	0.706	0.744	0.725	320
8	0.780	0.516	0.621	62
9	0.000	0.000	0.000	1
accuracy			0.775	1929
macro avg	0.611	0.539	0.568	1929
weighted avg	0.776	0.775	0.774	1929

The decision tree has 0.776 accuracy rate.
Except quality rate 5 all of the rates are below 0.7, the edges cant get a good predictions because they are outliers and with almost no examples to test.

Then I took the clean dataset, this dataset contains the 2 columns I changed by joining them throw their relations with other columns and getting the best correlation.
For improving the tree decision and the display of these 2 trees I changed the max depth of them to 8 which did improved the result blow:

	precision	recall	f1-score	support
4	0.861	0.517	0.646	60
5	0.812	0.822	0.817	597
6	0.811	0.838	0.824	834
7	0.827	0.794	0.810	306
accuracy			0.815	1797
macro avg	0.828	0.743	0.774	1797
weighted avg	0.815	0.815	0.813	1797

The accuracy rate of the clean dataset tree is 0.815 which is 0.039 better.
The prediction for a higher quality wine (6&7) is above 0.8 in the clean data set compare to 0.77 and 0.70 in the first dataset which is a big improve.

Summary

In conclusion the decision of getting a good win quality depends on many parameters where the most important one is the Free sulfur dioxide when the higher it get the higher the wine quality is, when there are no exceptions on the low-quality wine. When the Citric acid is more then 2 and the Free sulfur dioxide is high the wine is in quality 7 in most cases so if we want a high-quality rate 7 wine these to can really make the different.

On this project one of the hardest things to do was the decide what to do with empty values on the dataset. what to do with outliers, how many to remove so the data will by influence in a positive way.
Understanding the data and the importance of each column and what does it means. It was hard for me to choose what to do in each step and decide what to fix or remove because I wanted the best results I can get and on the other side not doing any harm to the data or changing the ending decision.

I learn a lot about wines and about how to handle dataset, step by step.
Next time I will have a project with handling data the firs thin I'm going to do is read about the subject of the data because its can really help later on in the project, on this one I just stated and thought my day to day knowledge will be enough because I don know a little bit about wine, but after doing some mistakes and not understanding the data I realized I need to read more about it before continuing the project. If I were doing that from the start it would of save me a lot of time.

At the end I'm happy with the results of the project and what I've learned from it. Thank you for reading grab a glass of wine (my recommendation with a lot of free sulfur dioxide) and have fun.
Dekel Menashe 311224117