**MACHINE LEARNING – PROJECT PROPOSAL**

**WINE QUALITY PREDICTION**

Contents

[Background 2](#_Toc25941302)

[Problem 2](#_Toc25941303)

[Input 2](#_Toc25941304)

[Output 3](#_Toc25941305)

[Datasets 3](#_Toc25941306)

[Project Design 3](#_Toc25941307)

[Data Exploration 4](#_Toc25941308)

[Model Construction 9](#_Toc25941309)

**CP322 Jiayao Pang 194174300**

# Background

Nowadays, wine appears more frequently and its production process has been improved a lot. Due to this and many other reasons, the quality of wine varies a lot as well. And this confused the wine enterprises about how to determine the price of their productions. In other words, how the quality varies based on the wine’s different features.

# Problem

With 11 given values of different features of the wine, we predict the quality of it and then it’s easy for us to determine the price.

# Input

The 11 features can be read from the dataset file using *pandas.read()*.

1 - fixed acidity (I modified it to a categorical feature.)

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

The value standing for quality ranged [0, 10]

# Datasets

I downloaded the training and testing dataset from the website [*https://archive.ics.uci.edu/ml/datasets/Wine+Quality*](https://archive.ics.uci.edu/ml/datasets/Wine+Quality)

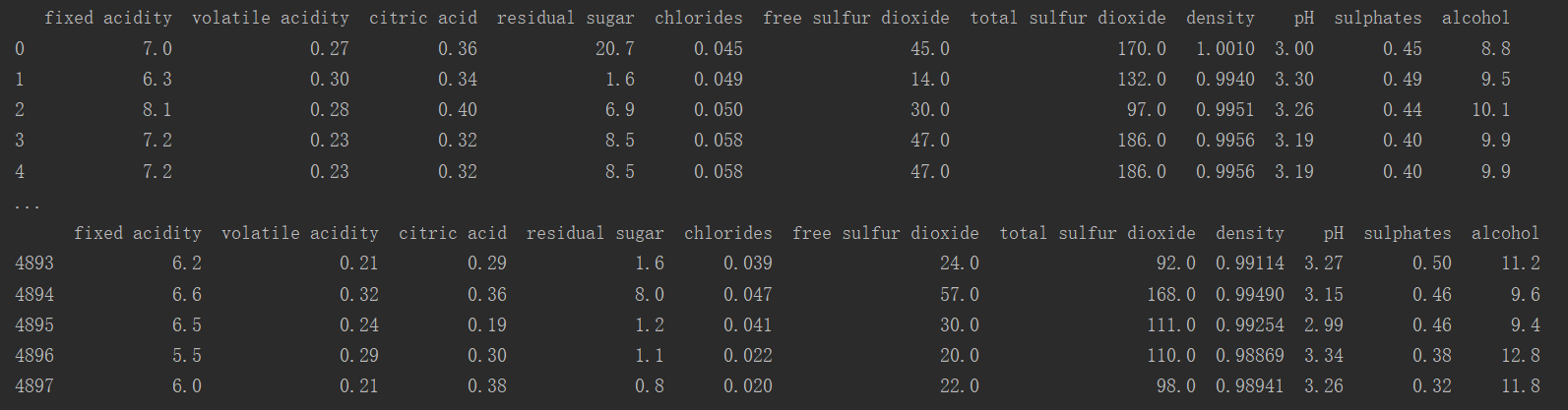
Note: I modified feature ‘fixed acidity’ to a categorical one.

# Project Design

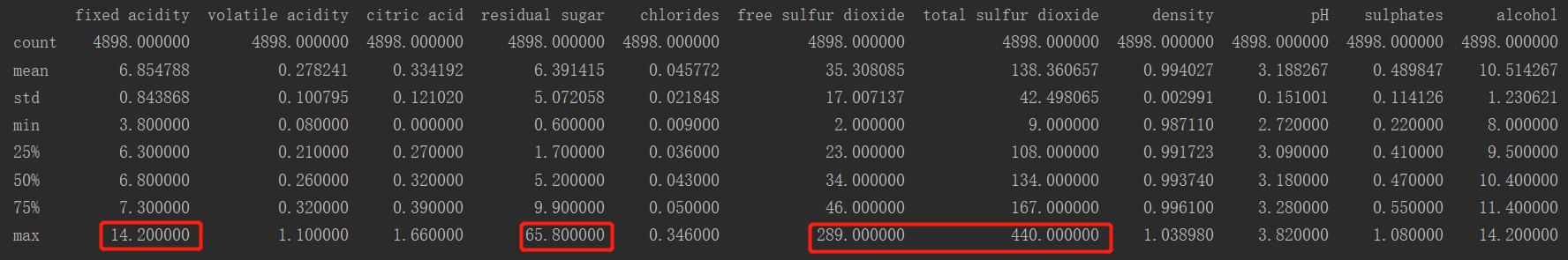
Source code of my project can be seen on [*https://github.com/PJYGit/ML-FinalProject*](https://github.com/PJYGit/ML-FinalProject)

## Data Exploration

Obviously, before the data exploration I need to construct the ABT and the data quality report.

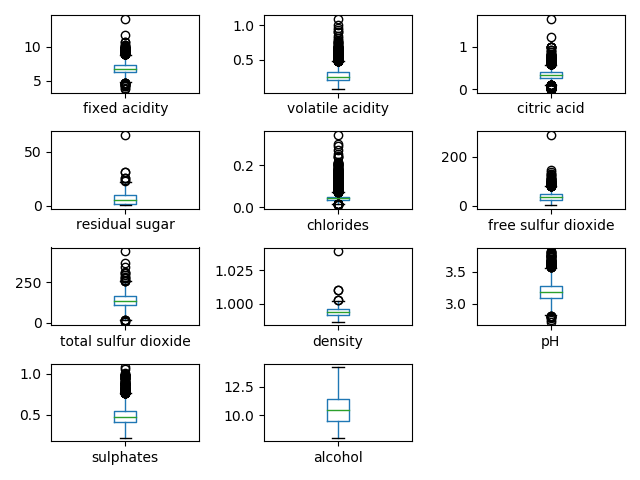


a short view of the ABT

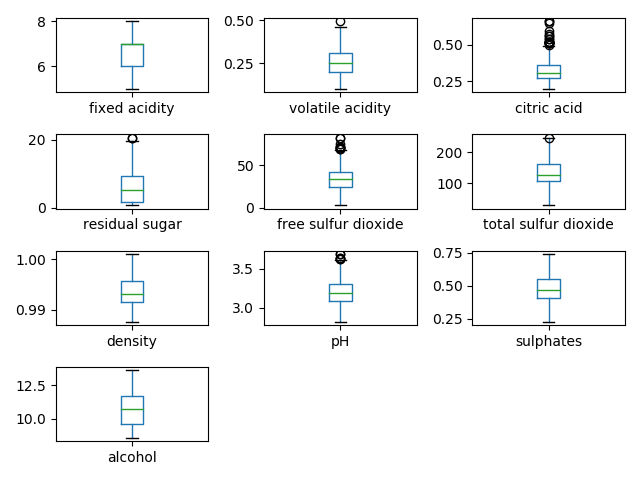


data quality report

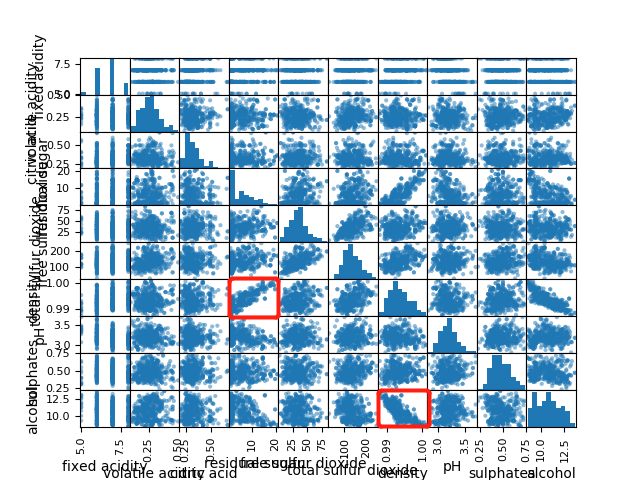
From the data quality report, we can easily notice that surprisingly, there is no missing value. But there are some outliers. So, I draw the box plots for every descriptive feature.



With the view of all the box plots, it’s pretty important for me to handle the outliers. I just deleted all the outliers out of the normal range. But for feature ‘chlorides’, the values are almost all outliers and the range of them is so small. I just delete this feature. After dealing with the outliers, we can see that the box plots are much better.

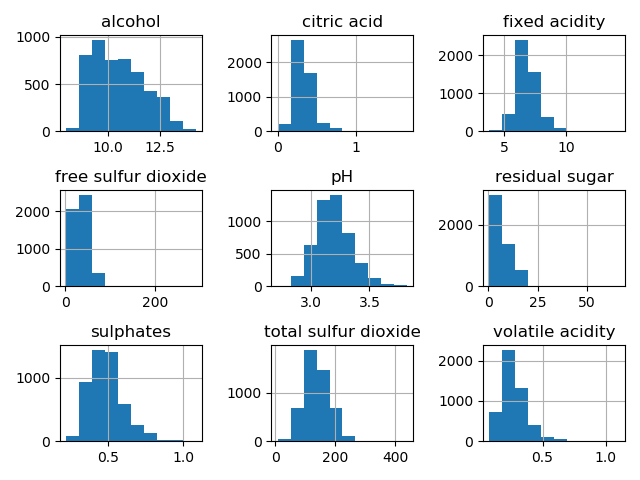


After dropping the feature ‘chlorides’, there are still 10 kinds of different descriptive features. A high dimension is not good for model construction. Then I draw the scatter plots paying attention to the relationships between different features.

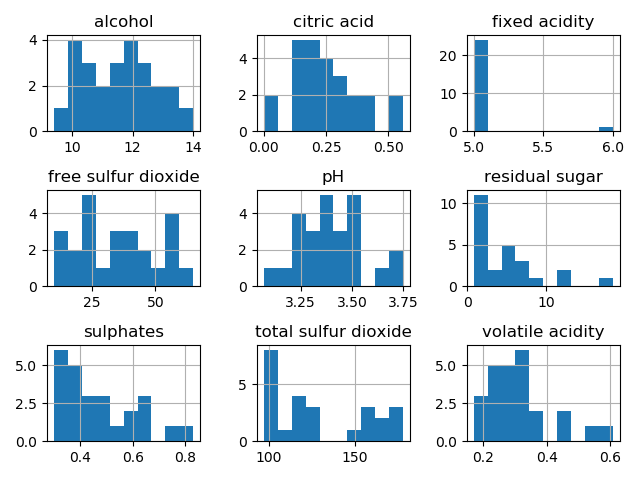
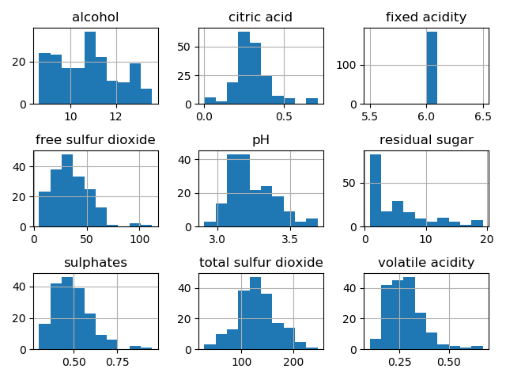


It’s easy to notice that there is a strong linear relationship between feature ‘residual sugar’ and ‘density’ and also between feature ‘alcohol’ and ‘density’. For the goal of lower dimension, I dropped the feature ‘density’.

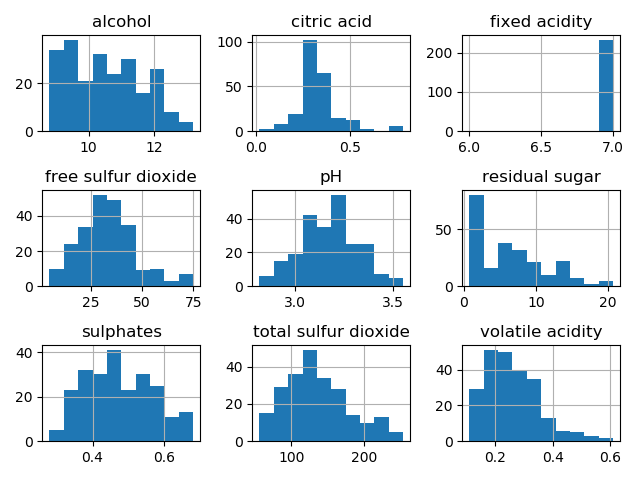
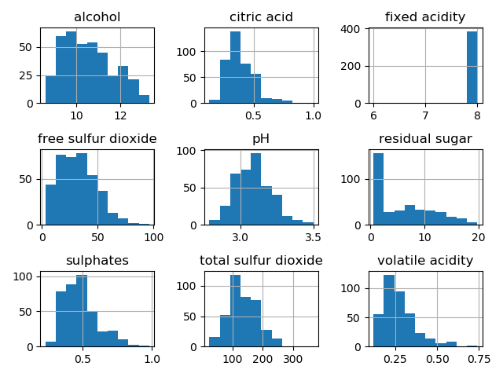
The relationship exploration above is only for continuous features. For the relationships between categorical and continuous features, I draw several histogram for different levels (5 levels in total) of the only categorical feature ‘fixed acidity’.



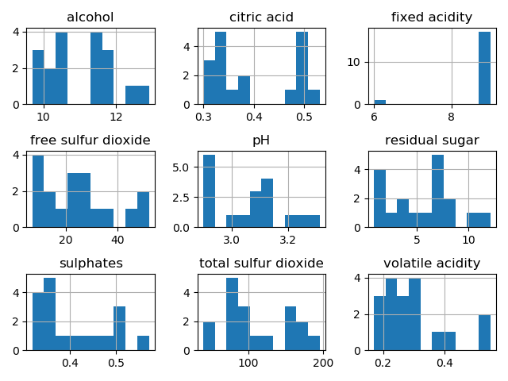
original

fixed acidity = 5 fixed acidity = 6

fixed acidity = 7 fixed acidity = 8



fixed acidity = 9

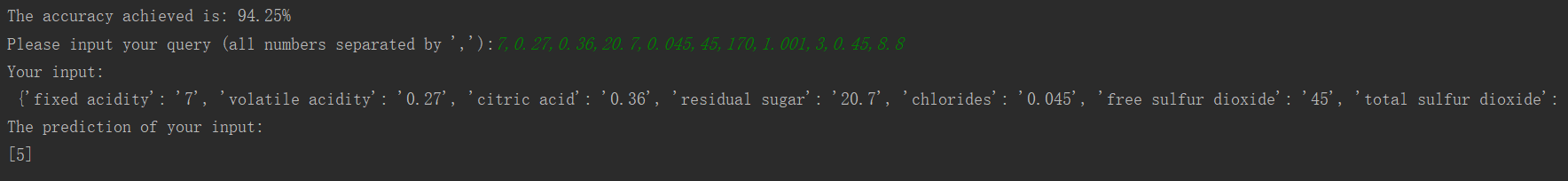
It seems like the feature ‘fixed acidity’ has relationship with every other feature. But since it’s the only categorical feature and there may be some unforeseen consequence after deleting it, I just leave it there.

## Model Construction

After the data exploration, I divided the whole dataset into 2 parts which are training dataset (80%) and test dataset (20%). The test dataset can’t be seen while training the model.

For this project, I chose the way of error-based learning, more specifically, it’s the linear regression with gradient descent. Cause in my opinion, this model suits more with high dimension training.

After the training, the highest accuracy of the prediction can be 96.2% (with an acceptable error of ± 1). Normally, it would be more than 91%.



output of the code