**Explanation for the result**

Here is the description of the factor: Alcohol, Volatile acidity, Sulphates Citric Acid, Total Sulfur Dioxide, Density, Chlorides, Fixed acidity, pH, Free Sulfur Dioxide, and Residual sugar. We use this data frame to test and predict the quality of the wine.

Before I compile the data, I check the distribution of the quality level of wines, and results are showing that the most of the wine’s quality are level 5 and 6, but they also have some outliers. Then I check the relationship between quality and other factors and found that volatile acidity and chloride are negatively related to quality, which means high quality wine have less volatile acidity and chloride in it. Citric acid, sulphates and alcohol are positively related to quality.

After that, I use box plot to check the median and range that to identify the quality of the wine, and their relationship are paired with the graph I found above. But since the graph is about the median and the range to test the wine and they also have the maximum and minimum value, as well as the outliers, then it does not mean decisive. Take the plot graph below as an example, there do exist high quality wine with low alcohol.

图表, 箱线图

描述已自动生成

And the correlation heatmap also give the intuition of the relationship between quality and other feature variables.

After I check the relationship here, I start to train the data and test it by machine learning, it could predict the quality of the wine. And I used two model here to see which one is better fit for the data frame. And after the comparison I prefer the results generated by Gaussian model since Gaussian model has higher accuracy on predict the quality of the wine. (Gaussian on the left and Bernoulli on the right).

手机屏幕的截图

描述已自动生成 手机屏幕的截图

描述已自动生成