Summary

In the Term Project I analyzed two wine’s datasets. One data set contains data about red wine and another data set contains data about white wine. Originally these datasets contained 12 variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality. The quality variable contains values that are characterized the quality of wine in integers varying from one to ten. I decided to create a new variable called quality\_binary. This variable contains only two values 0 and 1. 0 corresponds quality of wines from 0 to 5, and 1 corresponds quality of wines from 6 to 10. Other words zero indicates a low quality and one indicates a high quality of wine.

Research questions:

1. Is there a statistically significant correlation between pH and the citric acid concentration for the white wine
2. Is there a statistically significant correlation between pH and the sulphates concentration for the white wine
3. I chose 5 variables for both the red wine and the white wine- citric acid, residual sugar, pH, total sulfur dioxide, and sulphates to build a logistic regression model predicting the quality of wines (low quality or high quality).
4. Finally, I evaluated whether the residual sugar concentration of the red wine statistically significant differs from that of the white wine.

As a part of my EDA I calculated the mean, maximum, minimum, median, 1st quartile, 3d quartile, and standard deviation for each variable. I also created a histogram and a boxplot for each variable. I performed the normal test for all variables and It turned out that all of them are not normally distributed. Because the normal distribution assumption of Pearson correlation was violated I used the Spearman correlation coefficient to calculate correlation between pH and the citric acid concentration for white wine and between pH

and the sulphates concentration for white wine. The correlation between pH and the

citric acid concentration is moderate (-0.548). I used the permutation method to calculate p value and the confidence interval. P value is close to zero that indicates that the correlation is statistically significant. The 90% confidence interval is (-0.041, 0.041). The correlation between pH and the sulphates concentration is very low (-0.08). I used the

permutation method to calculate p value and the confidence interval. P value is 0.002

that indicates that the correlation is statistically significant. The 90% confidence interval

is (-0.042, 0.042). There are obvious outliers for sulphates and I decided to filter out values that are higher than 1.2. In this settings p value is 0.05 that is on boundary of statistical significance.

I also run the logistic regression to make a model of wine quality using 5 variables-citric acid, residual sugar, pH, total sulfur dioxide, and sulphates. The residual sugar concentration is not statistically significant for both red wine and white wine. The effect of the

total sulfur dioxide is the smallest for both the red wine and the white wine. For both

the white wine and the red wine pH, sulphates and citric acid positively affect the

quality of wine. Sulphate concentration has the strongest effect for both wines’ quality.

Accuracy for the white wine model is 0.68, and accuracy for the red wine model is 0.64.

I used the permutation method to evaluate statistical significance the difference in

means (3.85) between residual sugar concentration of the red wine and the white wine. The p value is close to zero that indicates that that difference is statistically significant . The confidence interval is (-0.231, 0.211).