

White Wine Quality Analysis

A statistical approach

Anna Del Savio, 2097098
Francesco Tomaselli, 2089207
Inês Jesus, 2073570

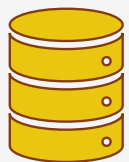
May 2023





Introduction

Goal: Which characteristics of the wine give it quality?



Dataset



Exploratory
Data Analysis



Models



Conclusion

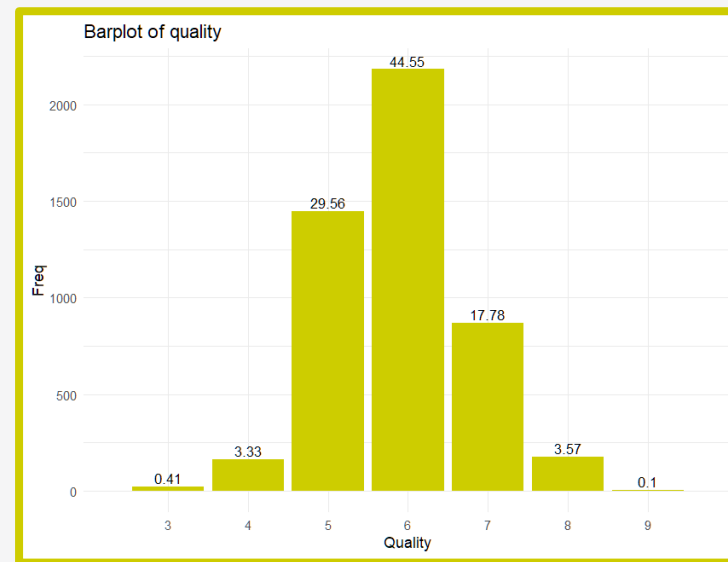


Dataset

- 4898 observations
- No missing values

1. Fixed acidity
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

12. Quality

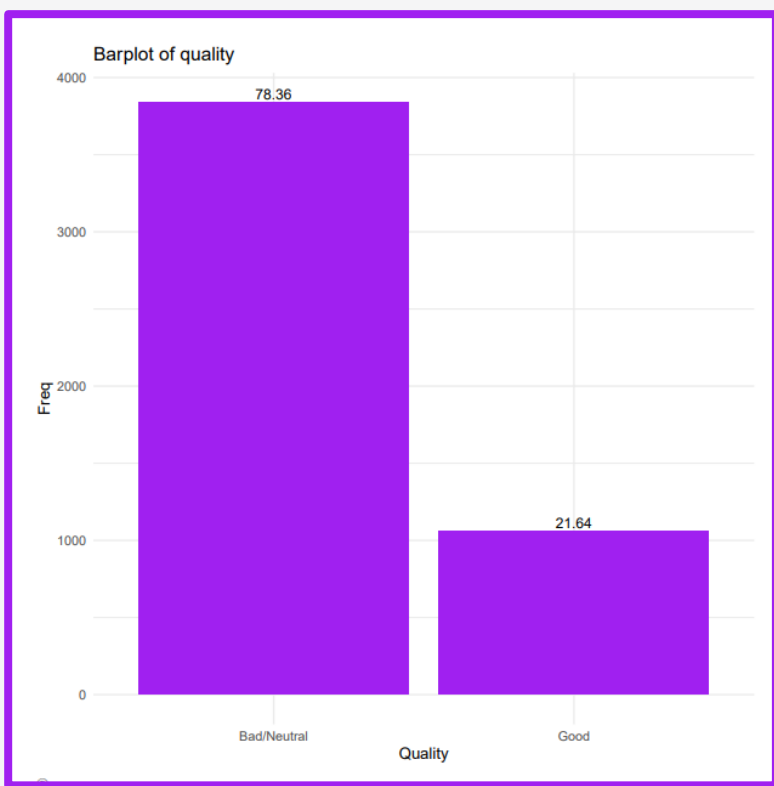


3	4	5	6	7	8	9
20	163	1457	2198	880	175	5



Dataset

(Imbalanced data problem)

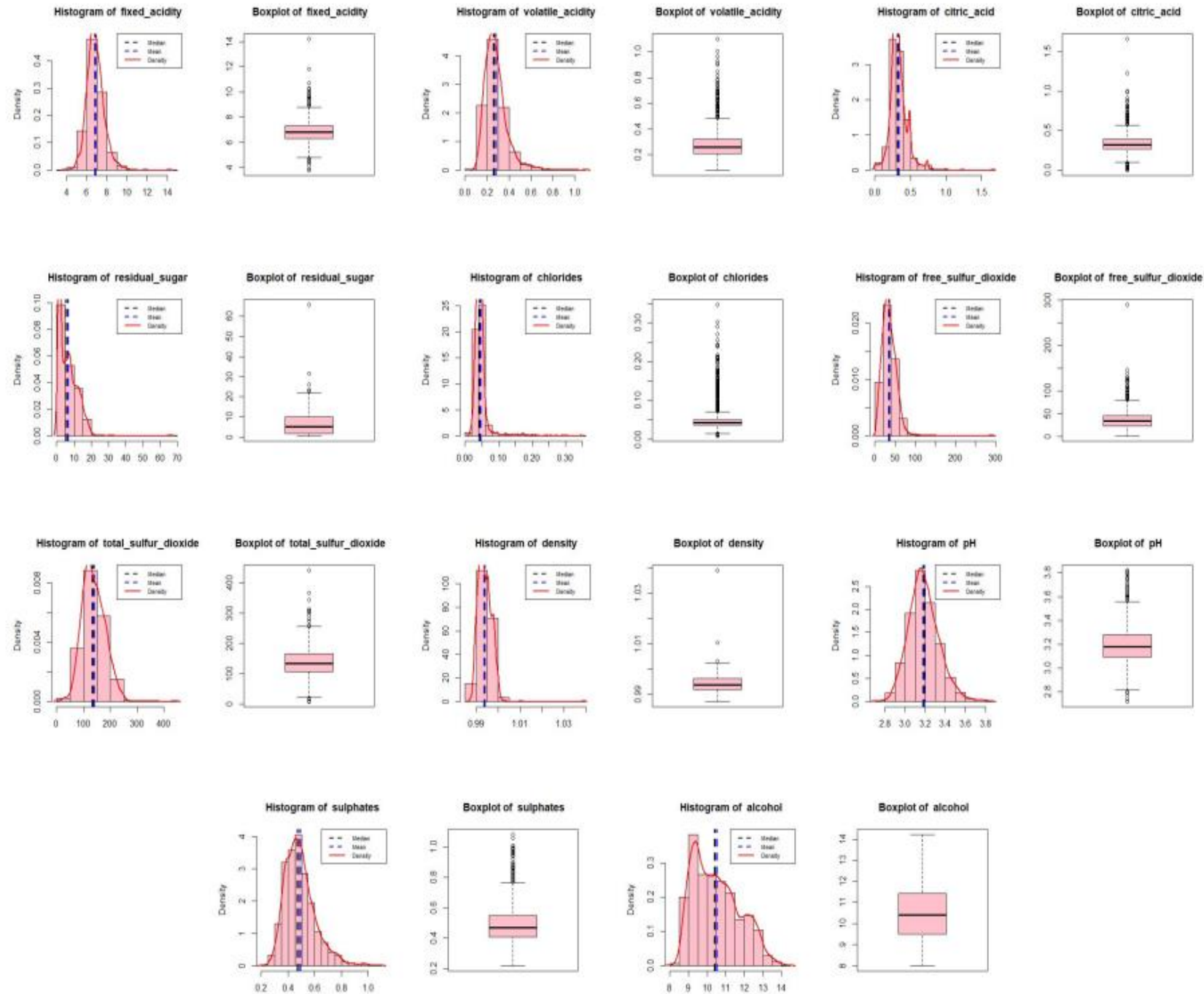


Bad/neutral quality (0)	Good quality (1)
3838	1060

- Every good wine, we have 3.62 bad/neutral quality wine
- **Imbalanced data**
- Accuracy is untrustable
- Use surrogate diagnostics: specificity, sensitivity and AUC

Exploratory Data Analysis (Univariate)

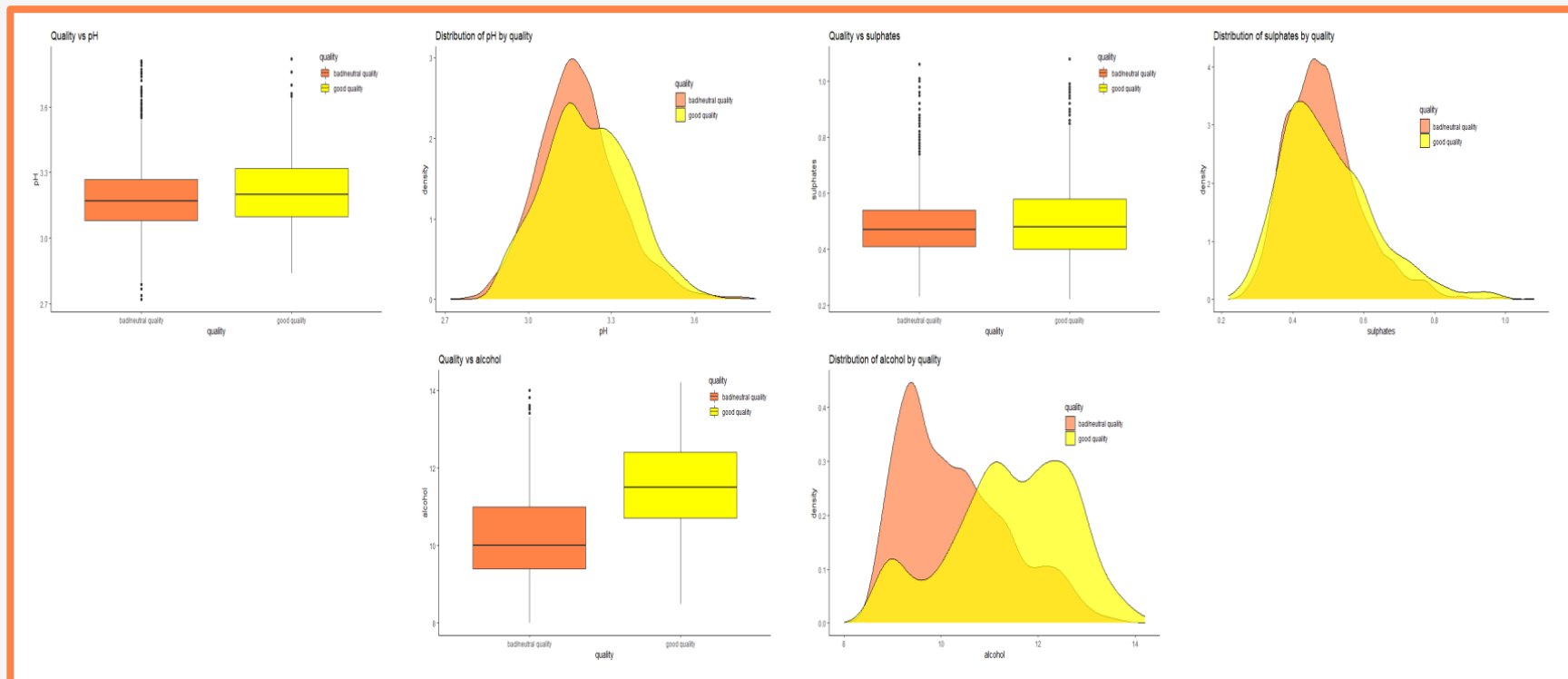
- Unimodal
- Not normally distributed
- Right-skewed
- A lot of outliers





Exploratory Data Analysis (Bivariate)

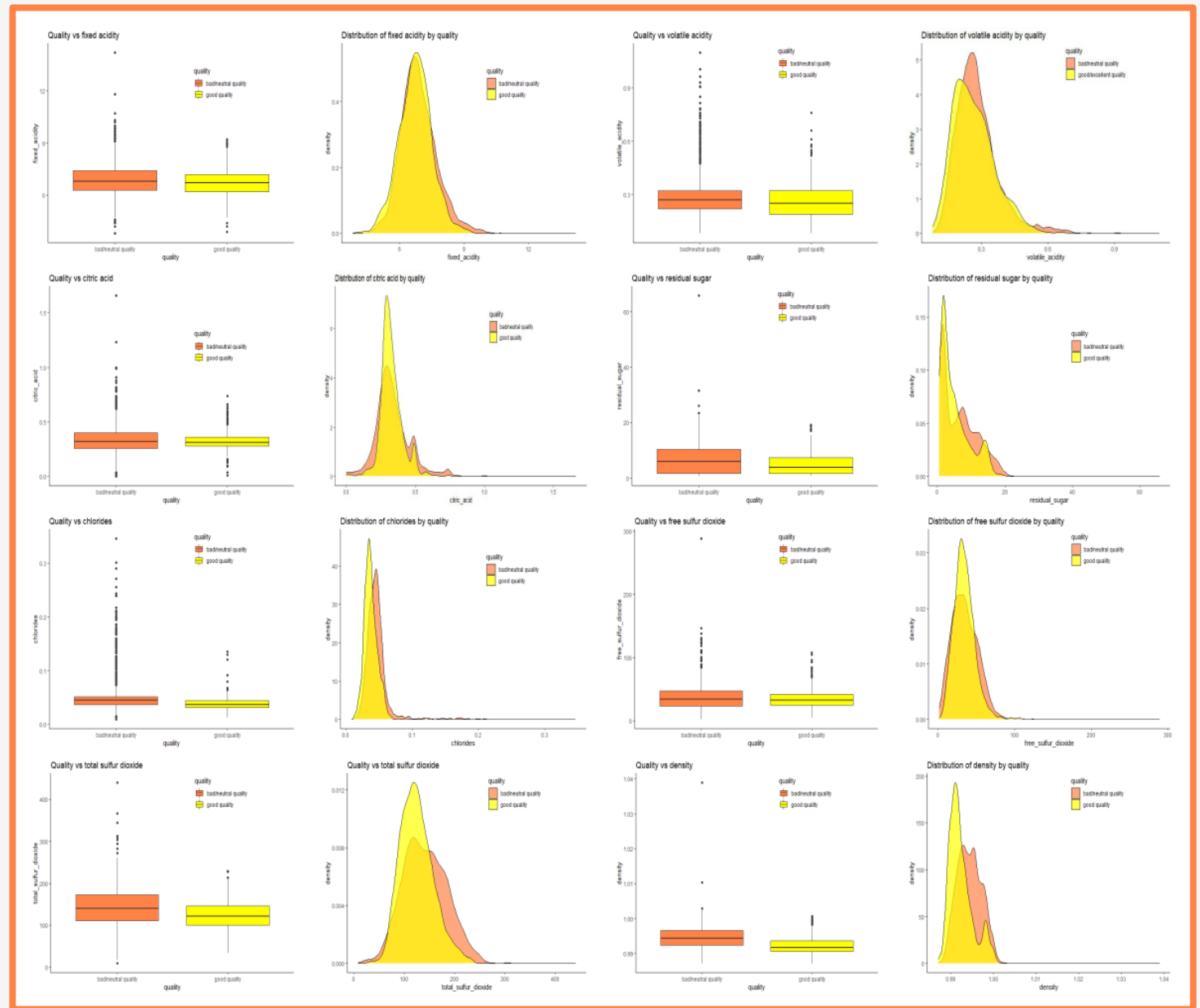
- Mean higher for “good quality” class
- Difference in mean is more notable for variable *alcohol*



- Mean slightly higher for “bad/neutral quality” class

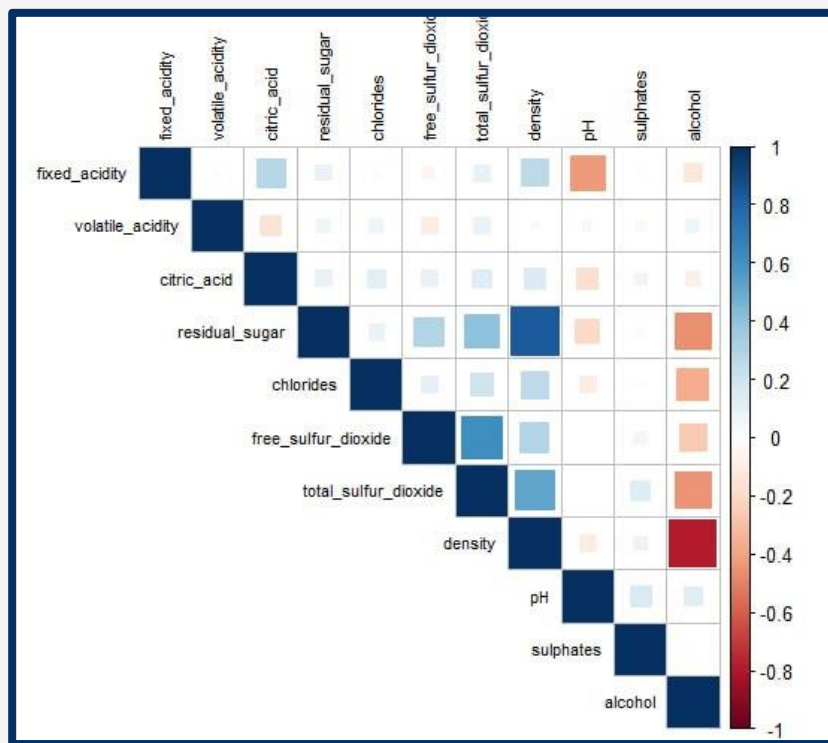
- More outliers for “bad/neutral quality” class

- Still not normally distributed when divided by classes

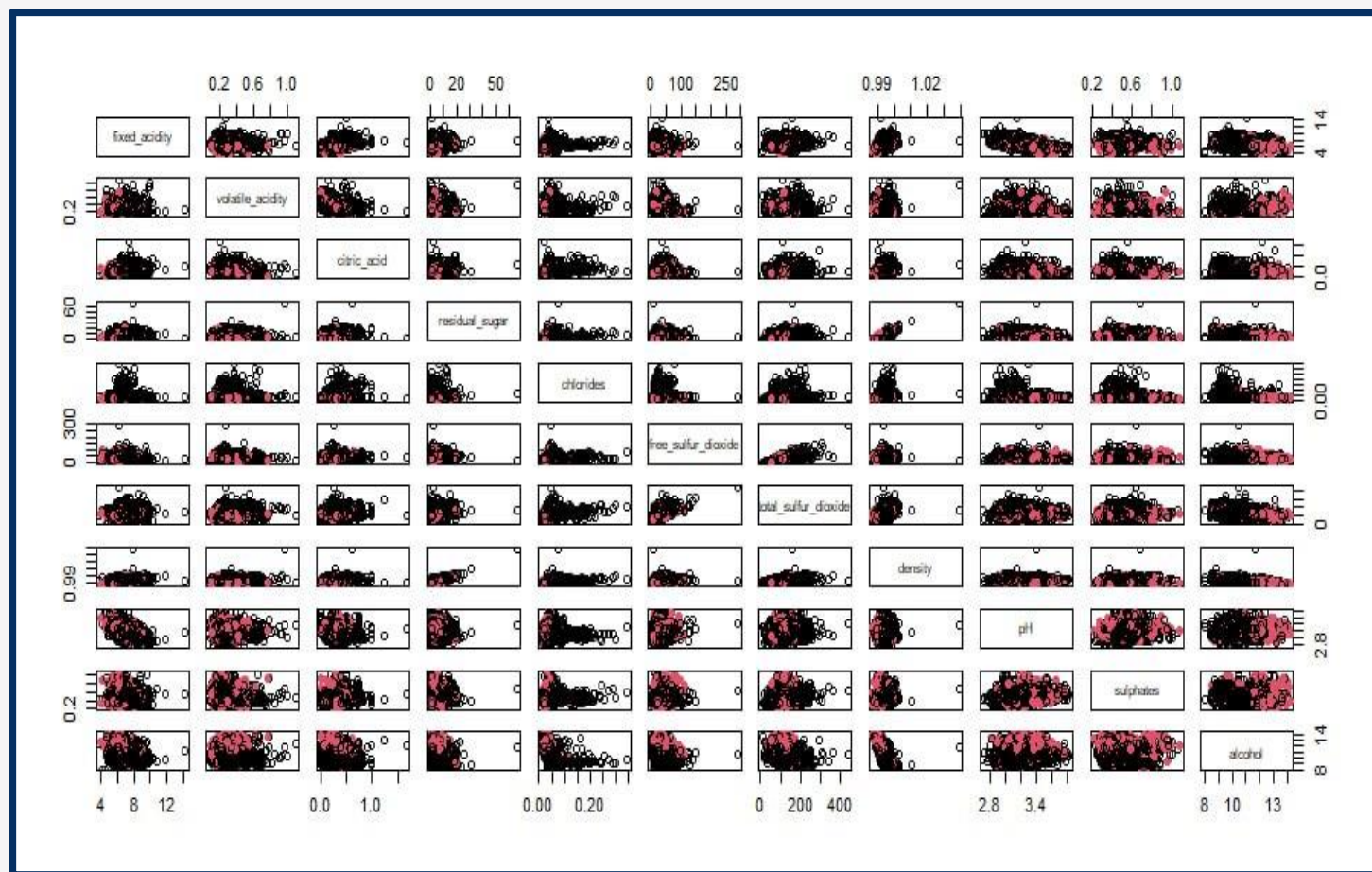




Exploratory Data Analysis (Correlation)



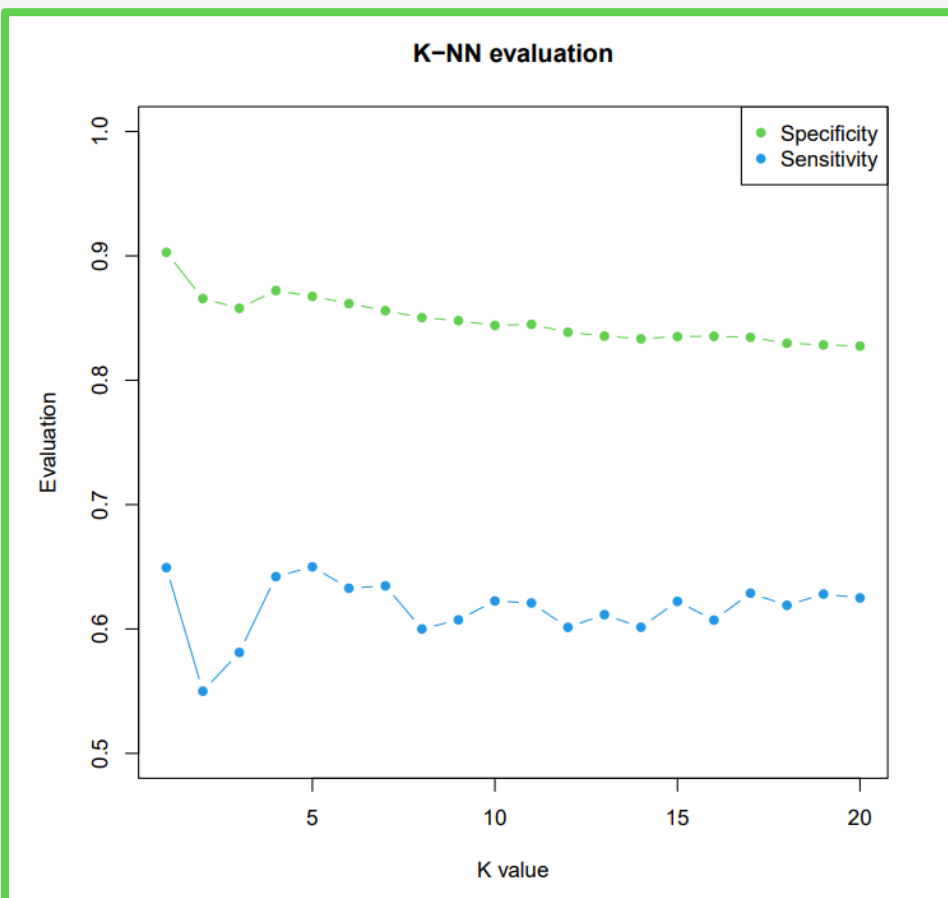
- Multicollinearity
- Remove *density* variable





Models

(K - Nearest Neighbours)



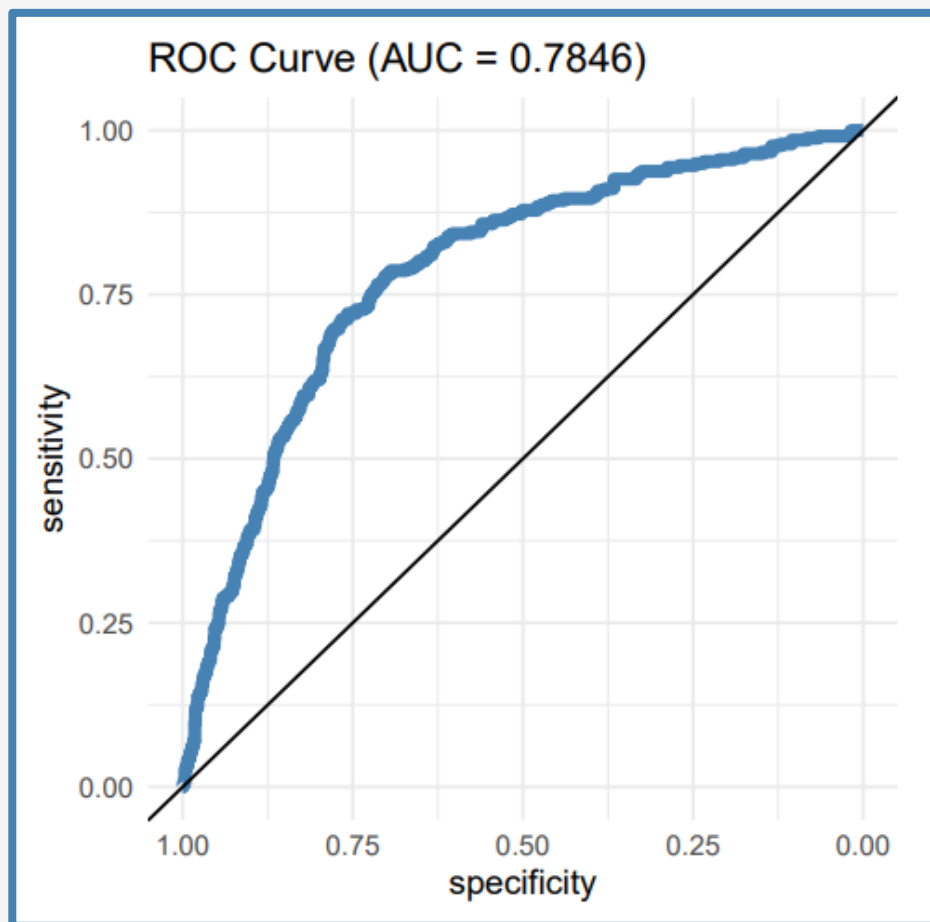
- 80% of observations to train the model and 20% to test it
- Best model turns out to be the **1-NN**

Specificity	Sensitivity
90.3%	64.9%



Models

(Naive Bayes)



- Assumption of independence between features not respected
- Even so, relatively good fit with:

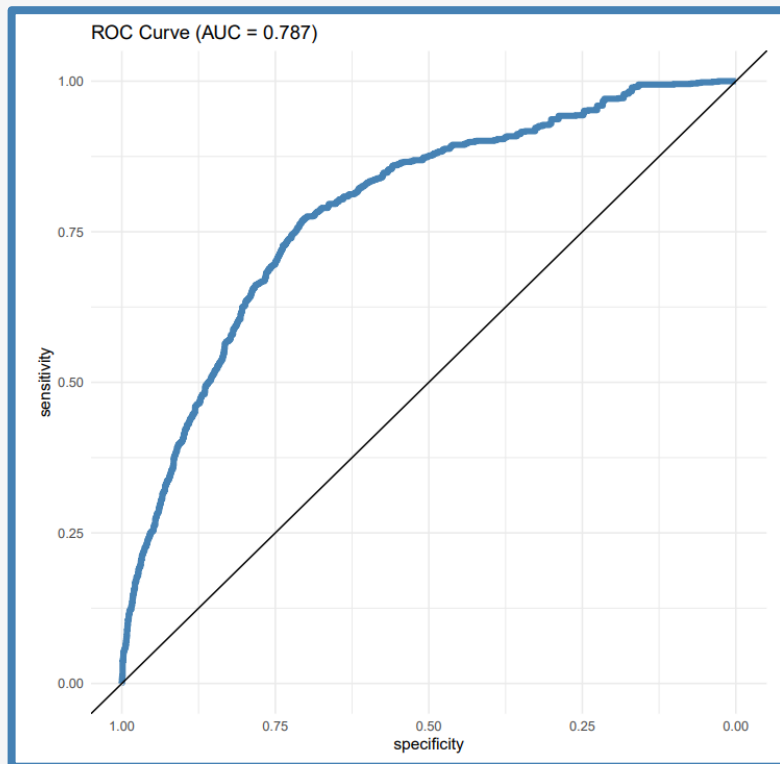
Specificity	Sensitivity
70.3%	77.7%



Models

(Logistic Regression)

$$\begin{aligned} \text{logit}(\hat{\mu}) = & -13.04 - 3.94 * \text{volatile_acidity} - 0.76 * \text{citric_acid} + 0.06 * \text{residual_sugar} \\ & - 17.93 * \text{chlorides} + 0.01 * \text{free_sulfur_dioxide} - 0.003 * \text{total_sulfure_dioxide} \\ & + 1.07 * \text{pH} + 1.27 * \text{sulphates} + 0.87 * \text{alcohol} \end{aligned}$$



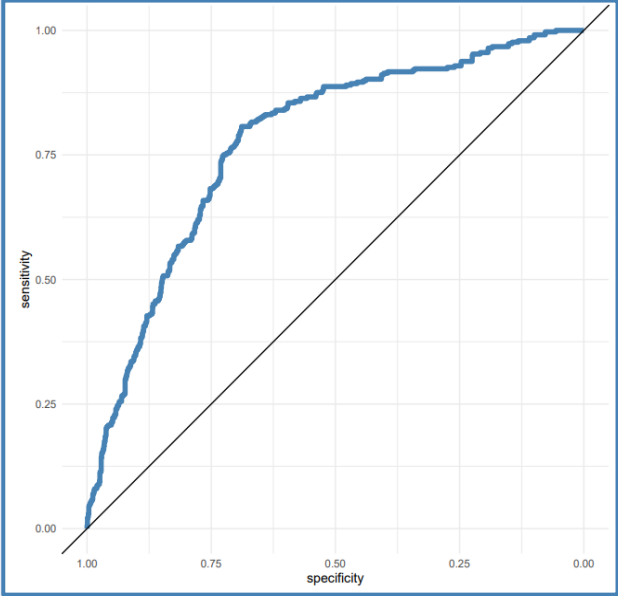
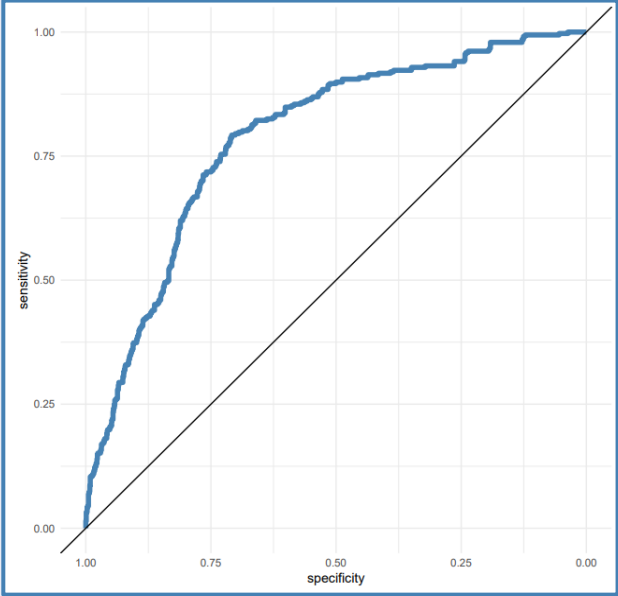
- Feature selection using forward and backward stepwise and AIC in both directions produce the same results
- Good fit with:

Specificity	Sensitivity
70.2%	77.3%



Models

(Penalised Logistic Regression)

	Lasso Regression	Ridge Regression
Specificity	68.9%	70.3%
Sensitivity	80.7%	79.2%
AUC	0.777	0.785
ROC curve plot		

Lasso keeps *volatile acidity, chlorides, free sulfur dioxide, sulfates and alcohol*

Conclusion



Good fit for all models except for K-Nearest Neighbours



Logistic Regression allows us to understand how the quality likely varies based on each wine characteristic



It is possible that different type of features, like year of production and grape type, could bring better classification results

