Wine Quality Evaluation

Yula Ko, Farzana Manjra, Natalia Gomez,
Victoria Wayda, JiaRui (Jesse) Shao

Data Diggers

Introduction

Determine what makes a “good” wine

We plan to determine which chemical properties make a wine 'good’ by utilizing various methods. We believe the target audience that may find this useful could be anyone else who is looking to determine the quality of wine based on these variables.

Overview

01 Intro to this data

Visualize the data 02

03 Logistic Regression

K-NN 04

05 Classification

Conclusion 06

INTRODUCTION

TO THIS DATA

Introduction

The Variables

Fixed Acidity	Volatile Acidity	Citric Acid
Residual Sugar	Chlorides	Free Sulfur Dioxide
Total Sulfur Dioxide	Density	pH
Sulphates	Alcohol	**Quality**

- 12 Variables
- 1,599 Observations
- Quality = Response Variable

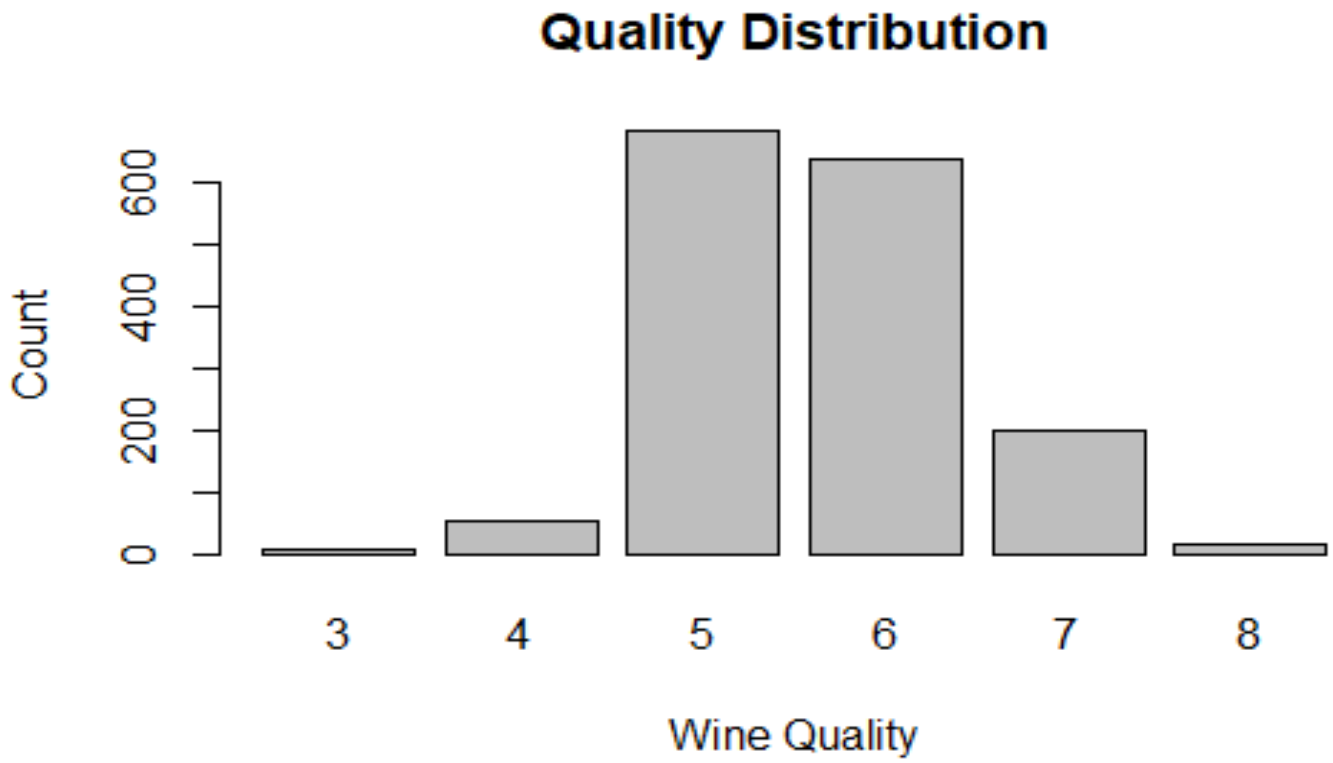
Data Description

- Fixed Acidity - most acids in wine that do not evaporate easily
- Volatile Acidity - amount of acetic acid, unpleasant taste
- Citric Acid - preservative that can add a fresh taste
- Residual Sugar - amount of sugar after fermentation stops
- Chlorides - amount of salt
- Free Sulfur Dioxide - prevents oxidation
- Total Sulfur Dioxide - preservative that can affect the taste
- Density - dependant on alcohol and sugar content
- pH - how acidic the wine is
- Sulphates - contributes to SO2 levels, preservative, fresh taste
- Alcohol - the percentage of alcoholic content
- **Output Variable(Y-response): Quality**

VISUALIZATION

OF THE DATASET

Bar Chart



Good: 855

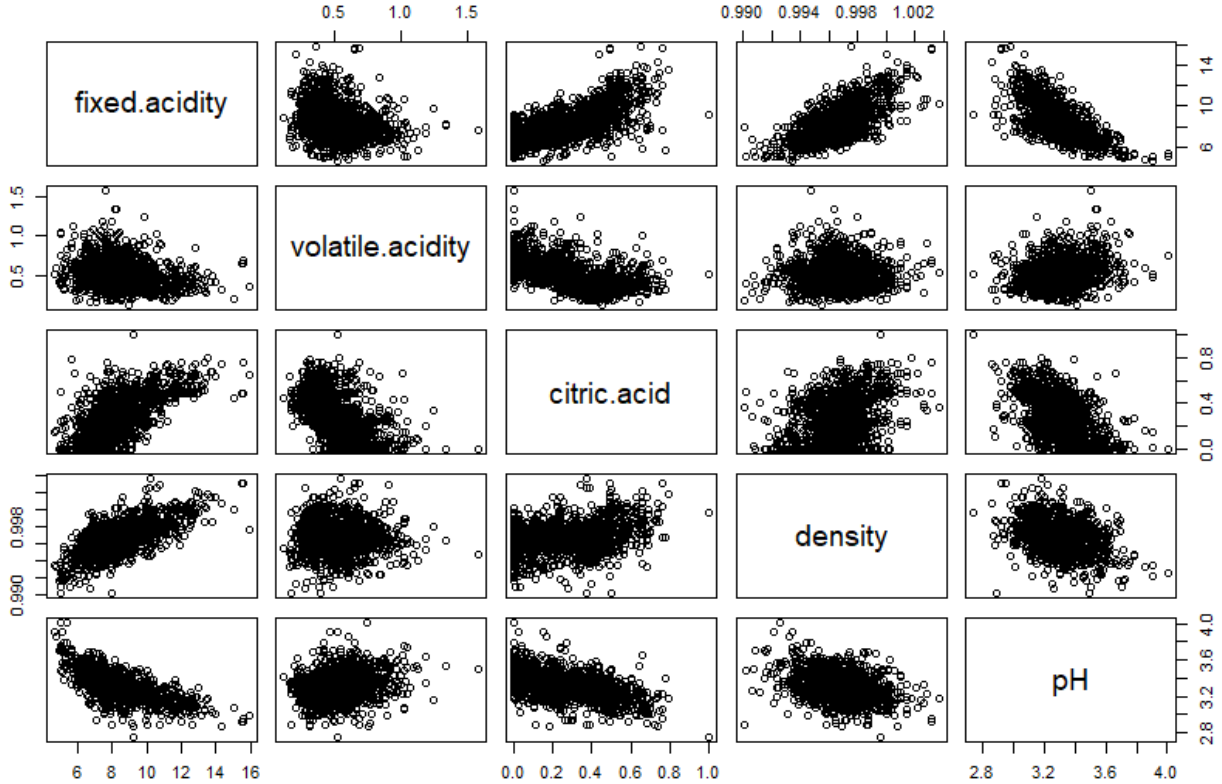
Bad: 744

Summary Statistics

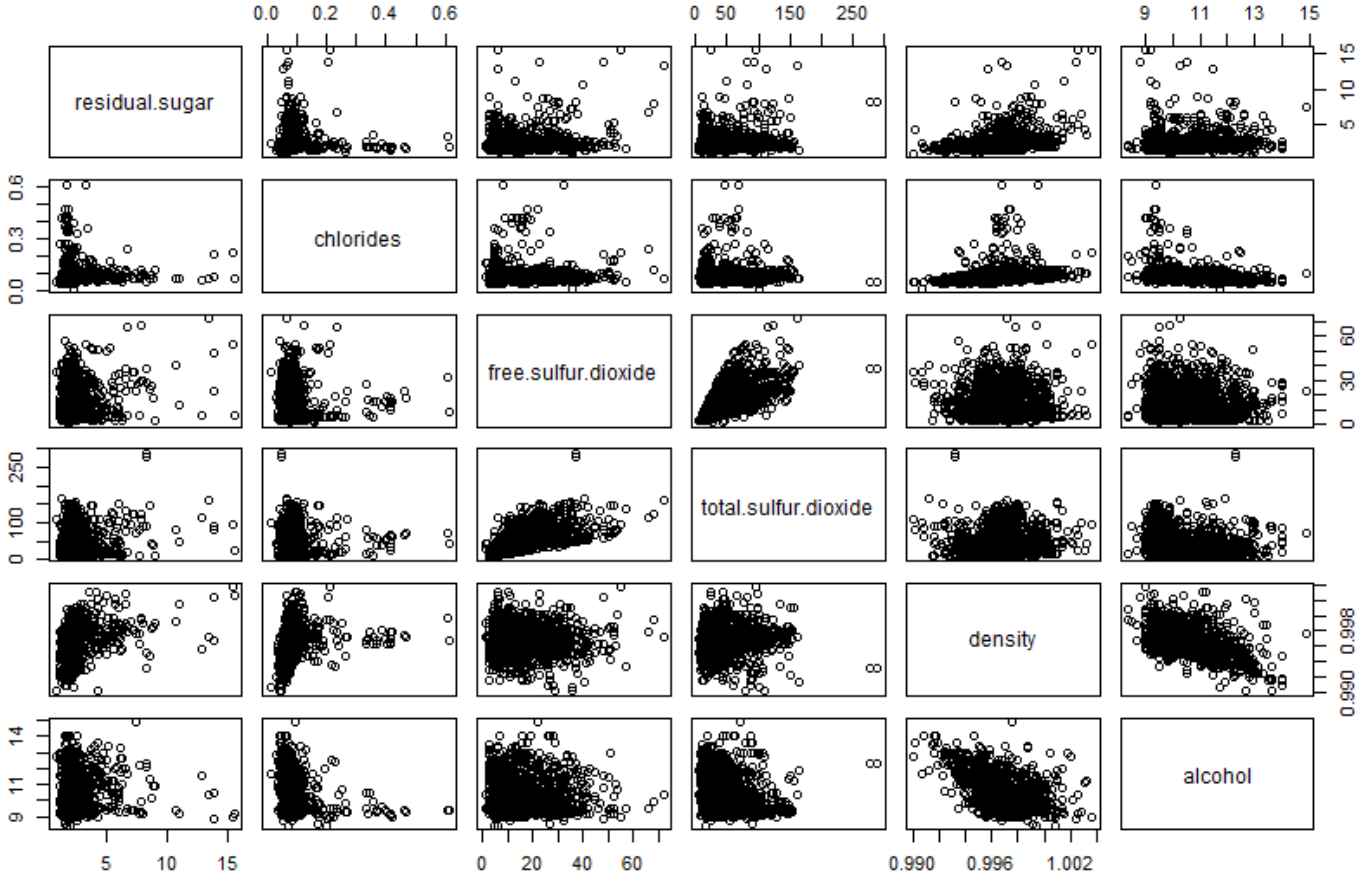
	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide
Mean	8.32	0.53	0.27	2.54	0.09	15.87
Min	4.60	0.12	0.00	0.90	0.01	1.00
Max	15.90	1.58	1.00	15.50	0.61	72.00
Median	7.90	0.52	0.26	2.20	0.08	14.00
SD	1.74	0.18	0.19	1.40	0.05	10.46

	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol
Mean	46.47	0.99	3.31	0.66	10.42
Min	6.00	0.99	2.74	0.33	8.40
Max	289.00	1.00	4.01	2.00	14.90
Median	38.00	0.99	3.31	0.62	10.20
SD	32.89	0.00	0.15	0.17	1.07

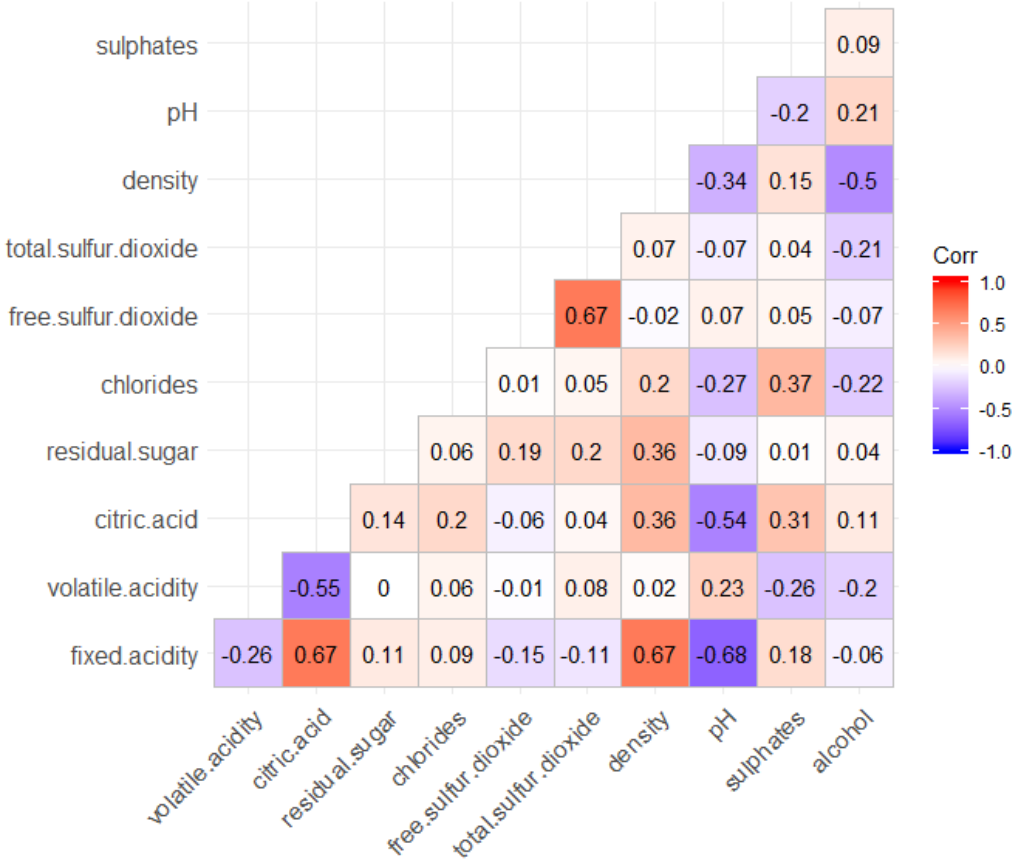
Correlation



Correlation

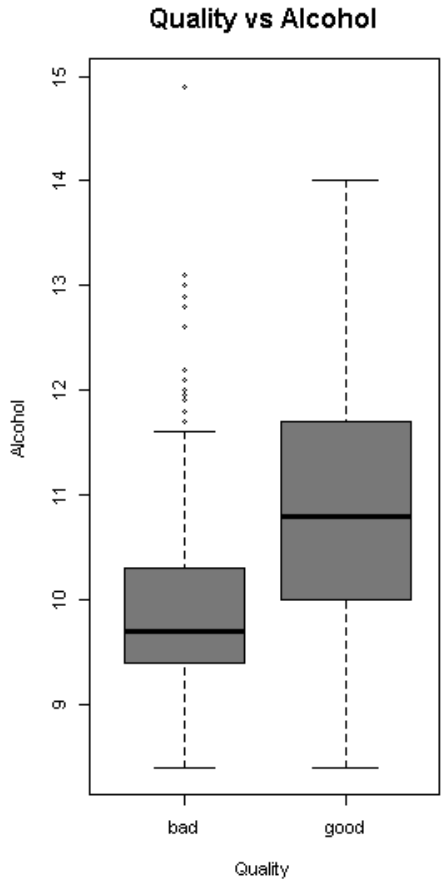


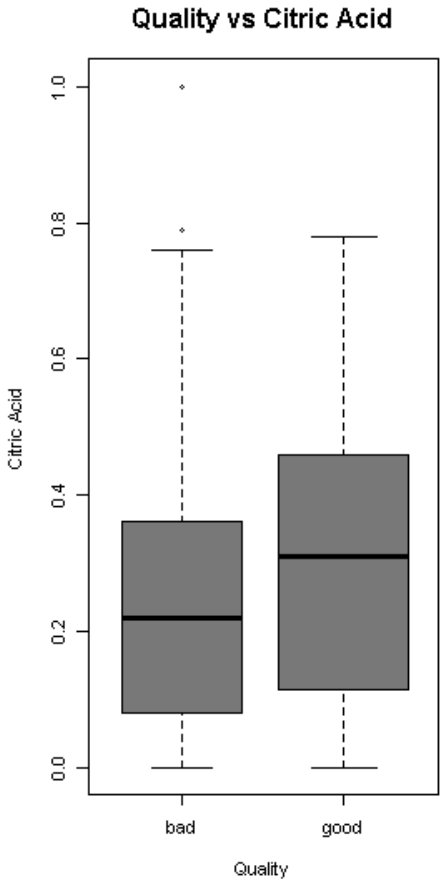
Correlation Matrix Heatmap

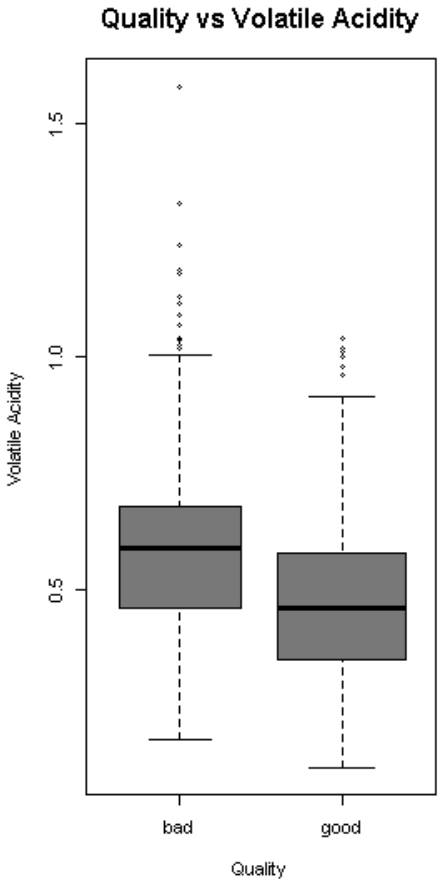


- Red = positive correlation
- Blue = negative correlation
- Numbers closer to 1 or -1 mean higher correlation , 0 means no correlation
- Above .70 or below -.70 means highly correlated

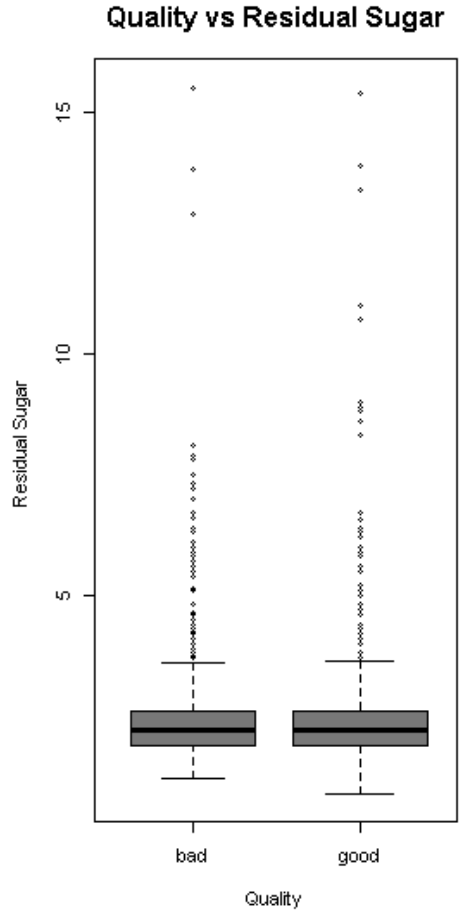
Box Plots

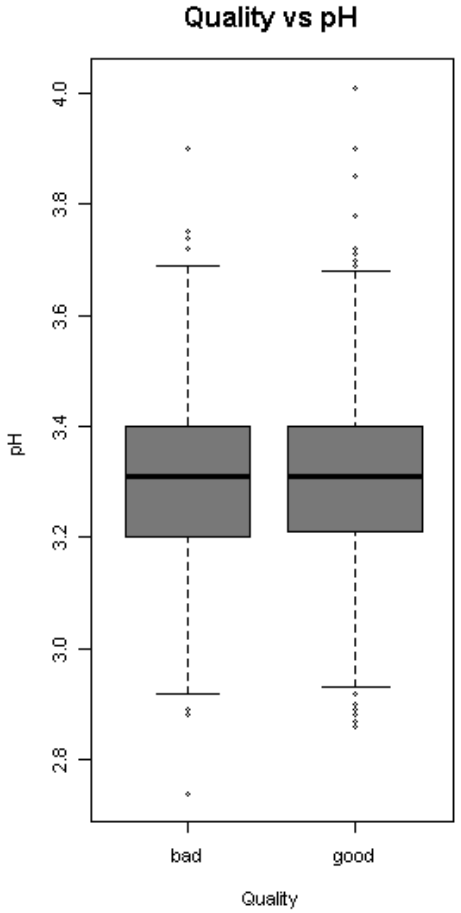

Quality vs Alcohol
Alcohol
9
10
11
12
13
14
15
bad
good
Quality

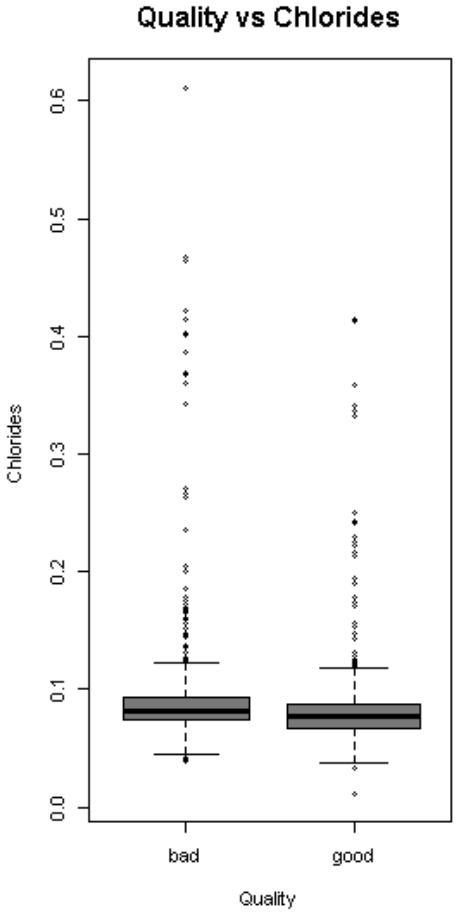

Quality vs Citric Acid
Citric Acid
0.0
0.2
0.4
0.6
0.8
1.0
bad
good
Quality


Quality vs Volatile Acidity
Volatile Acidity
0.5
1.0
1.5
bad
good
Quality

Box Plots


Quality vs Residual Sugar
Residual Sugar
5
10
15
bad
good
Quality


Quality vs pH
pH
2.8
3.0
3.2
3.4
3.6
3.8
4.0
bad
good
Quality


Quality vs Chlorides
Chlorides
0.0
0.1
0.2
0.3
0.4
0.5
0.6
bad
good
Quality

LOGISTIC REGRESSION

Logistic Regression

Model 1

Coefficients	P-value
(Intercept)	0.58890
Alcohol	< 2e-16 ***
Fixed Acidity	0.16736
Volatile Acidity	1.79e-11 ***
Citric Acid	0.02354 *
Residual Sugar	0.30351
Chlorides	
0.01259 *	Free Sulfur
Dioxide 0.00698 **	
Total Sulfur Dioxide	1.29e-08 ***
Density	0.53024
pH	0.59717
Sulphates	6.36e-10 ***

Model 2

Coefficients	P-value
(Intercept)	< 2e-16 ***
Alcohol	< 2e-16 ***
Volatile Acidity	1.24e-11 ***
Citric Acid	0.45448
Chlorides	0.00431 **
Free Sulfur Dioxide	0.00566 **
Total Sulfur Dioxide	6.48e-10 ***
Sulphates	2.16e-10 ***

Model 3

Coefficients	P-value
(Intercept)	< 2e-16 ***
Alcohol	< 2e-16 ***
Volatile Acidity	5.60e-15 ***
Chlorides	0.00202 **
Free Sulfur Dioxide	0.00281 **
Total Sulfur Dioxide	9.95e-11 ***
Sulphates	2.47e-10 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Holdout Method

Model 1

	True Bad	True Good
Predicted Bad	105	45
Predicted Good	27	143

- Accuracy = 77.50%
- Recall = 76.06%

Model 2

	True Bad	True Good
Predicted Bad	101	41
Predicted Good	31	147

- Accuracy = 77.50%
- Recall = 78.19%

Model 3

	True Bad	True Good
Predicted Bad	101	40
Predicted Good	31	148

- Accuracy = 77.81%
- Recall = 78.72%

10-Fold Cross-Validation

	Accuracy
Model 1	74.29%
Model 2	74.61%
Model 3	74.63%

Accuracy for the third model is slightly higher. That indicates the third model is better than the first model and the second model

The Fitted Model

$$\text{logit}(\pi(y=1|x) = -8.14 + 0.86\,\text{alcohol} - 2.9\,\text{volatile.acidity} - 4.42\,\text{chlorides} + 0.02\,\text{free.sulfur.dioxide} - 0.02\,\text{total.sulfur.dioxide} + 2.71\,\text{sulphates}$$

Coefficient Interpretation:

0.86 tells us when increasing alcohol by one unit, the log odds of good quality is expected to increase by 0.86, with all other predictors held fixed.

good=1

bad=0

Odds Ratio

(Intercept)	2.995201e+98
Alcohol	2.361693e+00
Volatile Acidity	5.524713e-02
Chlorides	1.202188e-02
Free Sulfur Dioxide	1.024149e+00
Total Sulfur Dioxide	9.826471e-01
Sulphates	1.496742e+01

Interpretation:

- The odds of alcohol equals 2.36
- It means that: holding all other independent variables at a fixed value, for a one unit increase in alcohol, the odds of wine quality good increase by a factor of 2.36;
- Or we can expect to see about 136% increase in the odds of quality good.

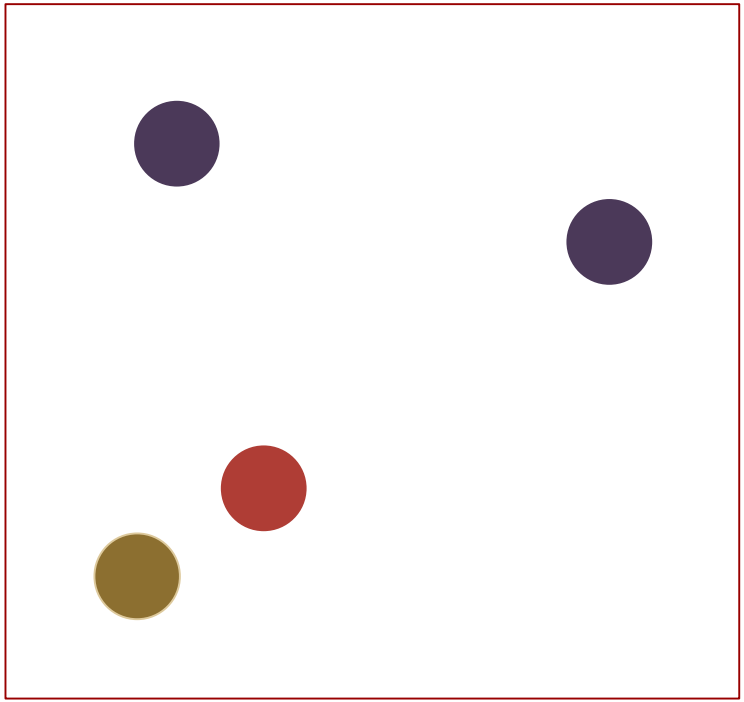
K-NN

K NEAREST NEIGHBORS

K NEAREST NEIGHBORS

K-NN

- The output is qualitative
- An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors



K-NN

Confusion Matrix

- K=1

	True Bad	True Good
Predicted Bad	89	39
Predicted Good	58	141

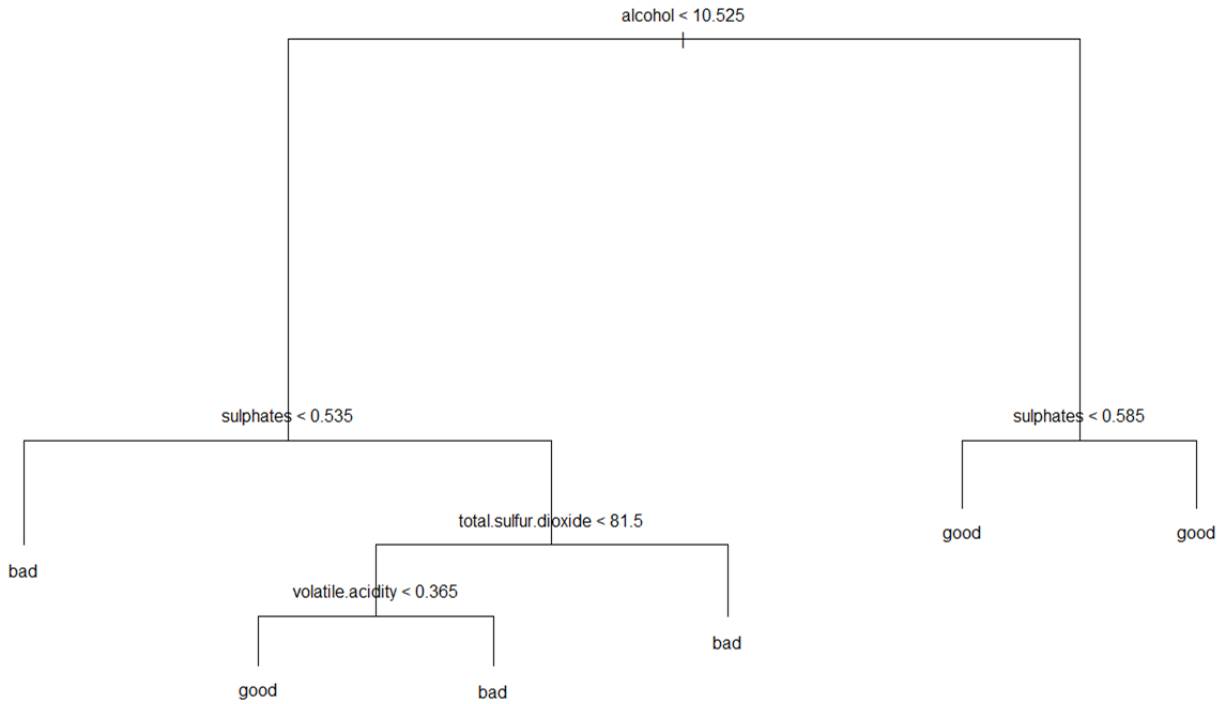
- Accuracy = 75.91%
- Recall =78.33%

CLASSIFICATION

DECISION TREE & RANDOM FOREST

Decision Tree

Decision Tree



Decision Tree

Confusion Matrix

	True Bad	True Good
Predicted Bad	107	68
Predicted Good	25	120

- Accuracy = 70.94%
- Recall = 63.83%

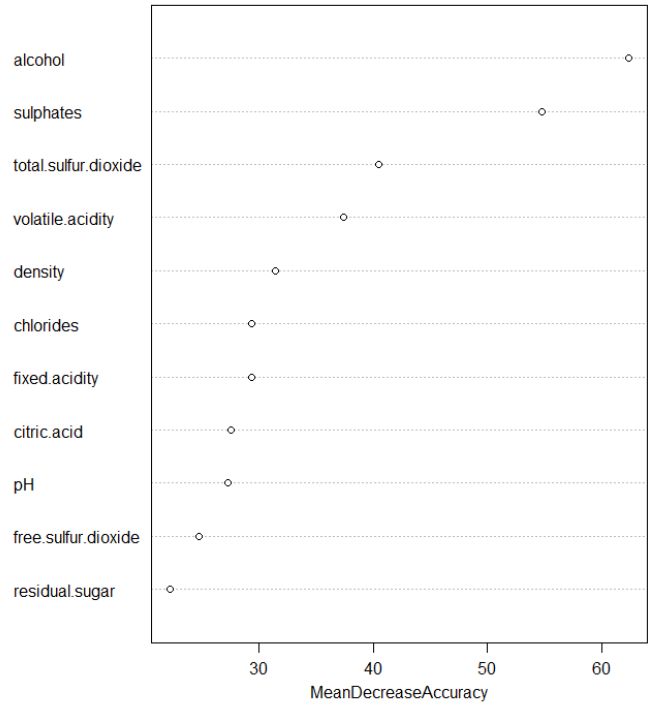
Random Forest

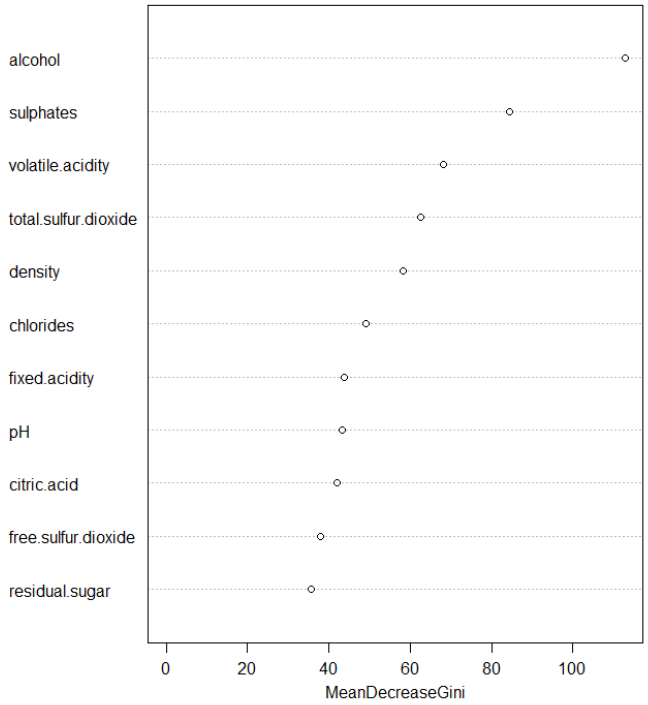
Random Forest

- Ensemble Algorithm
- Reduce chances of over-fitting
- Higher model performance or accuracy
- Accuracy = 83.75%

Random Forest

Variable importance





CONCLUSION

HOW TO DETERMINE A GOOD WINE

Conclusion

Logistic Regression

1

Hold-out: 77.81%
Cross Validation: 74.63%

K-NN

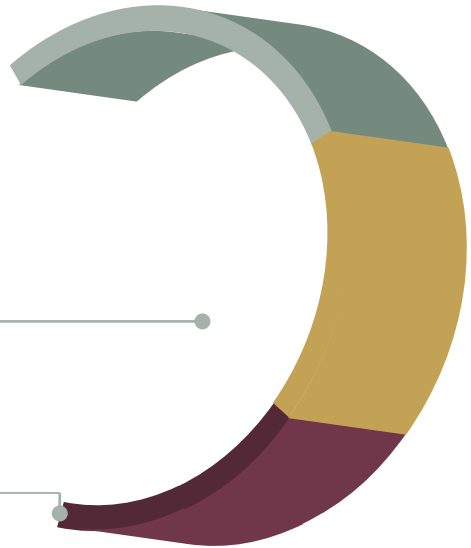
75.91%

2

Classification

3

Decision Tree: 63.83%
Random Forest: 84.06%



Thanks for listening

Q & A