Unveiling the Secrets of Wine Quality

Final Project of the course Data Science

Mentor: Marc

Zofia & Fengyu

Table of content:

Motivation

1. Cultural Significance
2. Enhancing Wine Production
3. Analytical Depth

Background

A group of researchers: Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and Jose Reis.

Collected Time: May 2004 to February 2007
Quality Rating: Median of three sensory assessors
Chemical Features: iLab

Modeling wine preferences by data mining from physicochemical properties(2009)

-> Result: Support Vector Machine: Alcohol, Citric acid and residual sugar.

Research question:

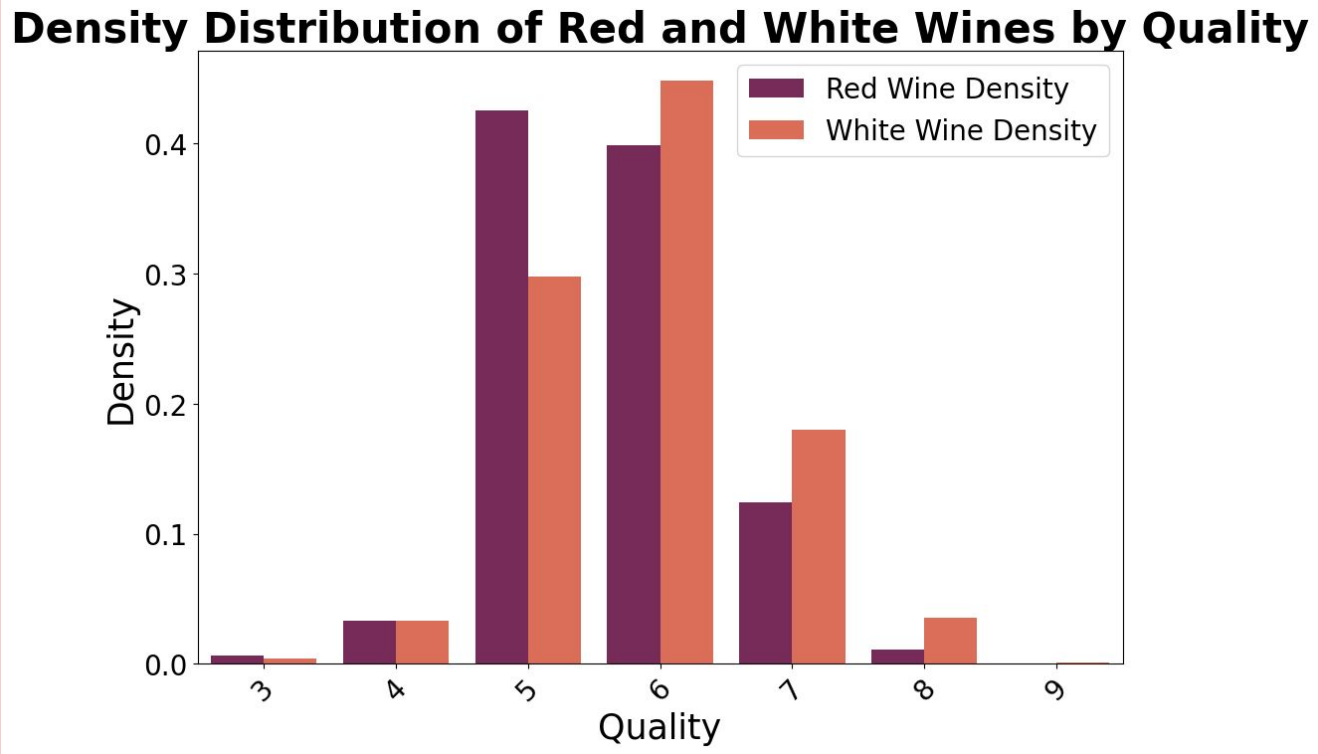
🍇Which features contribute the most to predict good and poor quality of wine?

🍇What is the recipe for a good and poor wine?

Wine Quality distribution

- Red Wine 🍷 (**1599** data points) and
- White Wine 🥂 (**4898** data points)

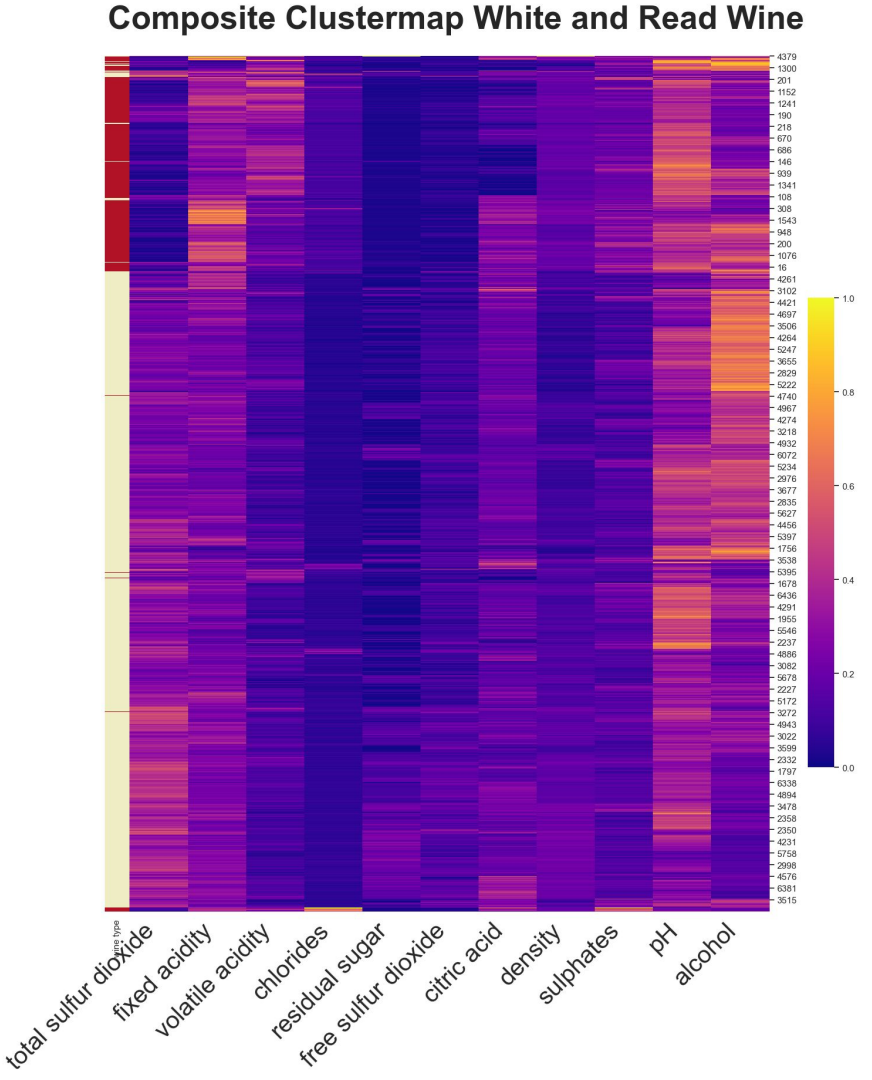
Data spans various **quality** levels from 0 to 10.



Dataset

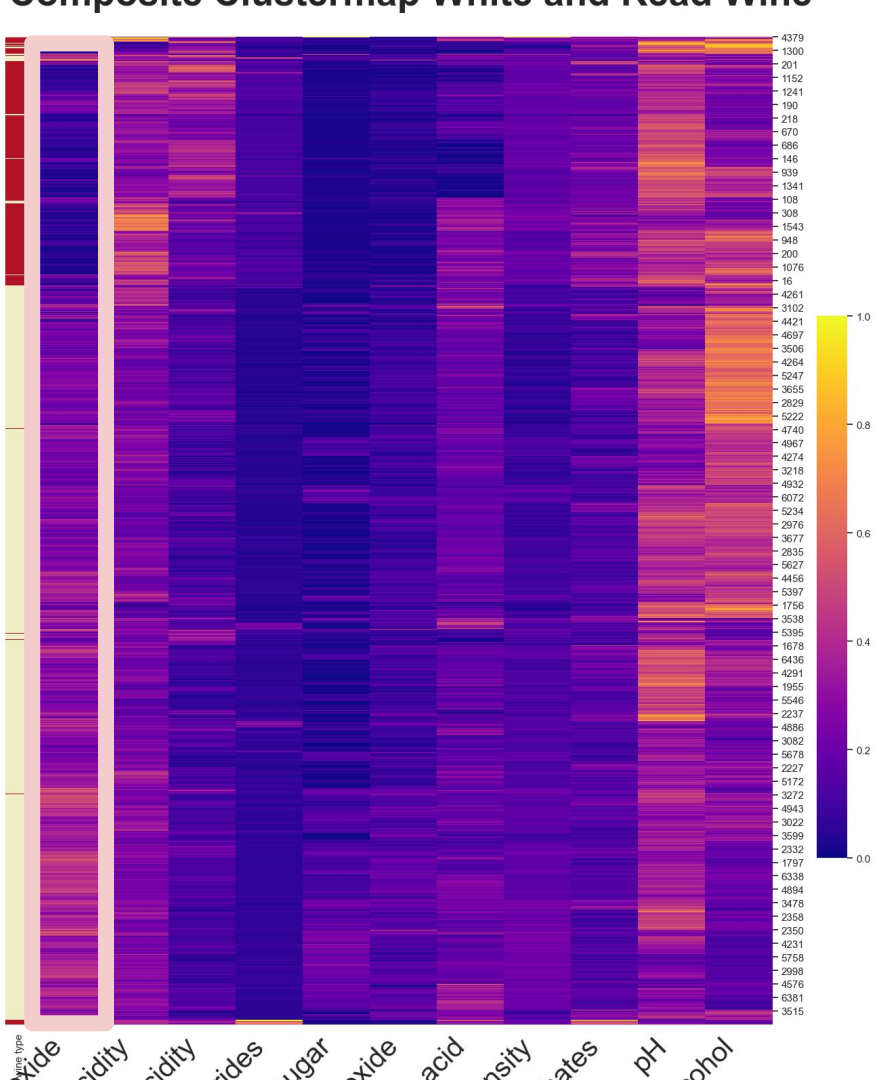
- Red Wine 🍷 (**1599** data points)
  and
- White Wine 🥂 (**4898** data points)

Includes **physicochemical properties** of wine and quality scores.



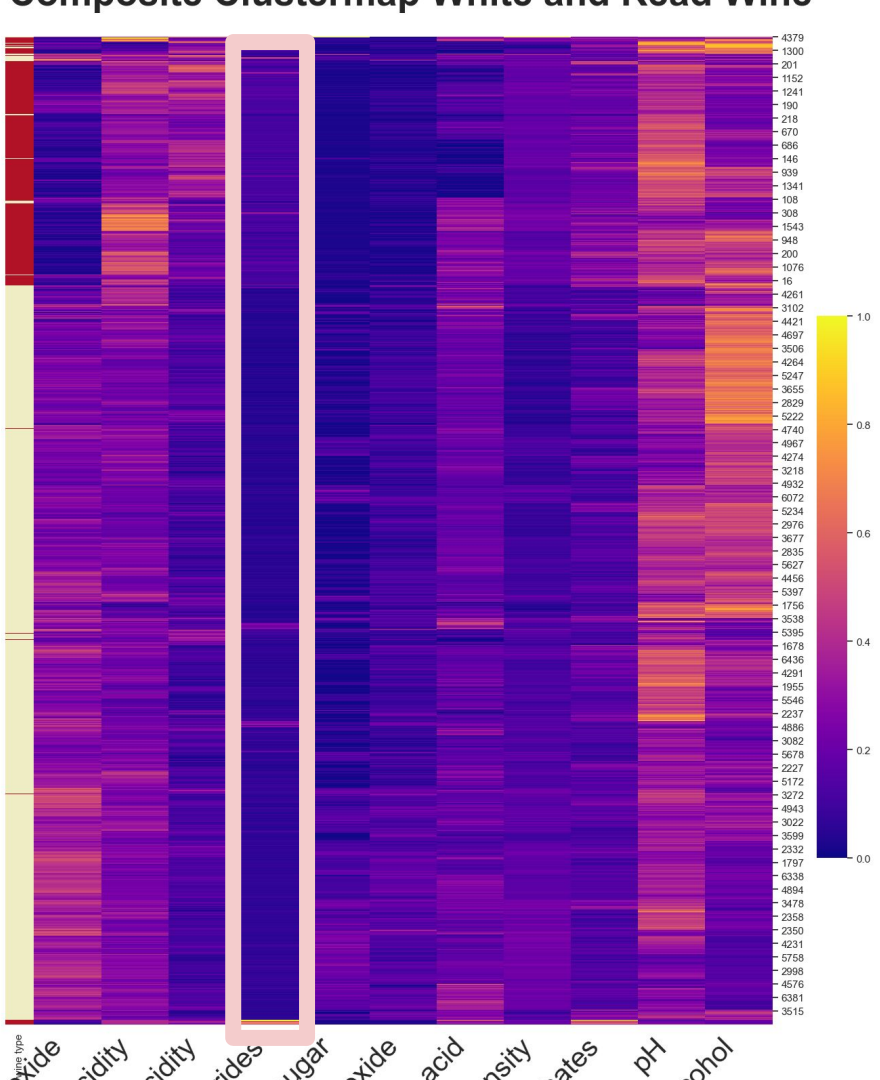
Dataset

Total Sulfur Dioxide



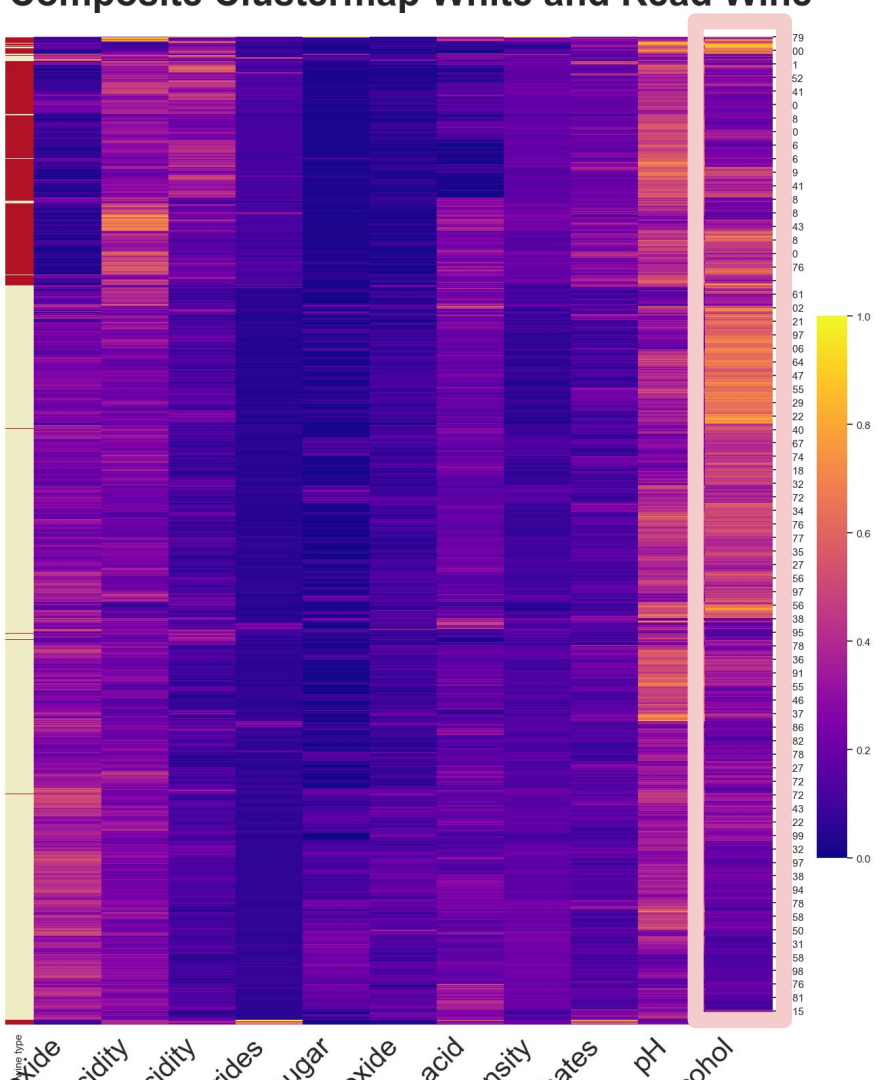
Dataset

Chlorides

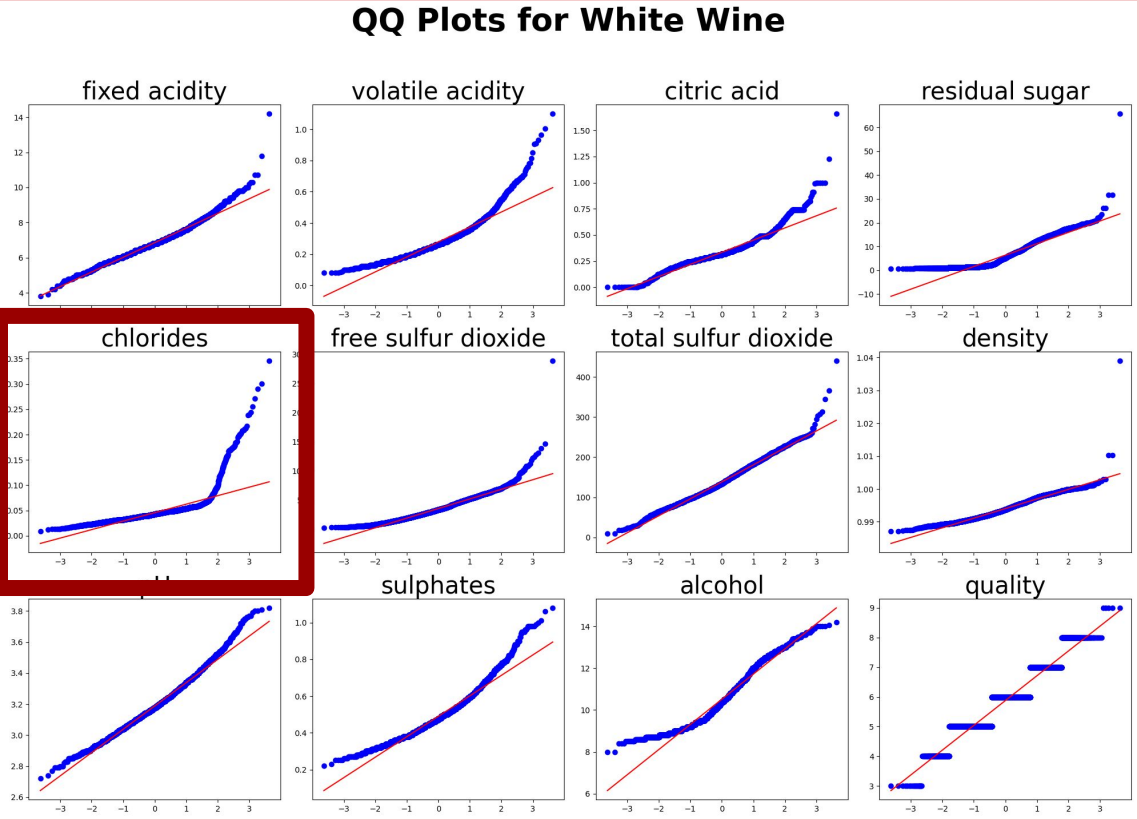


Dataset

Alcohol

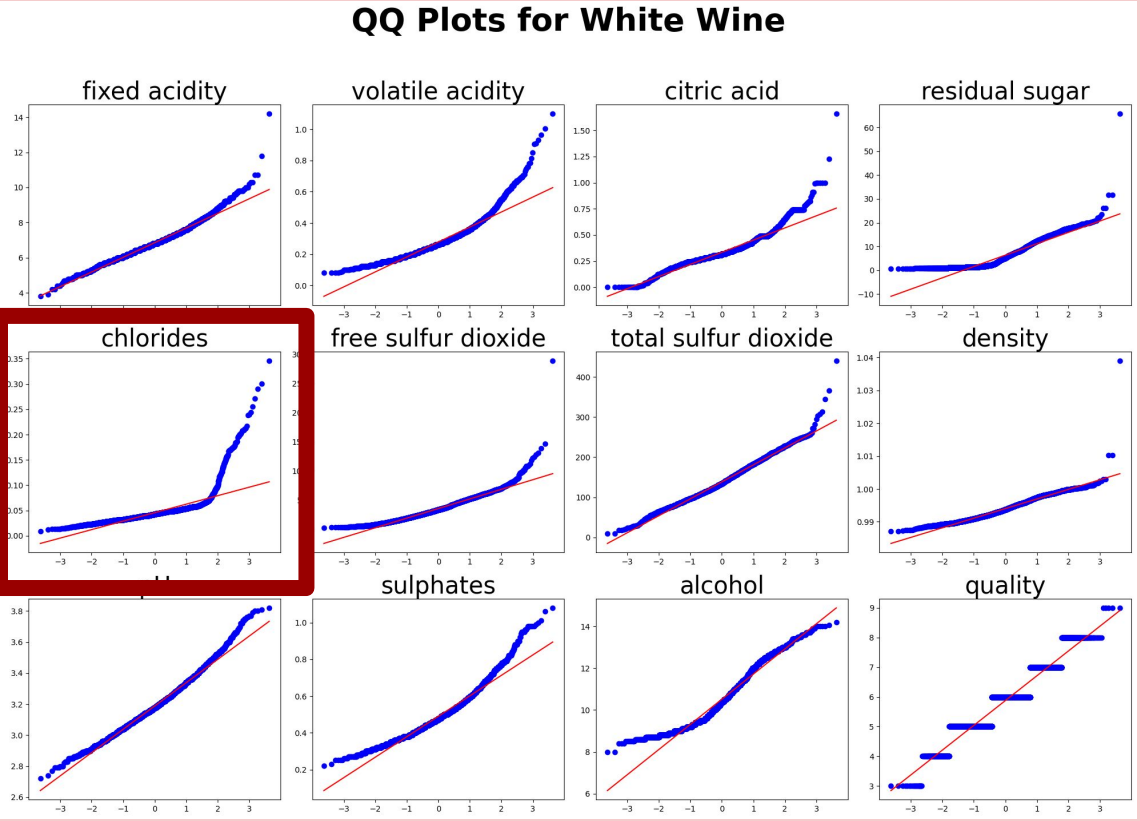


Transformation

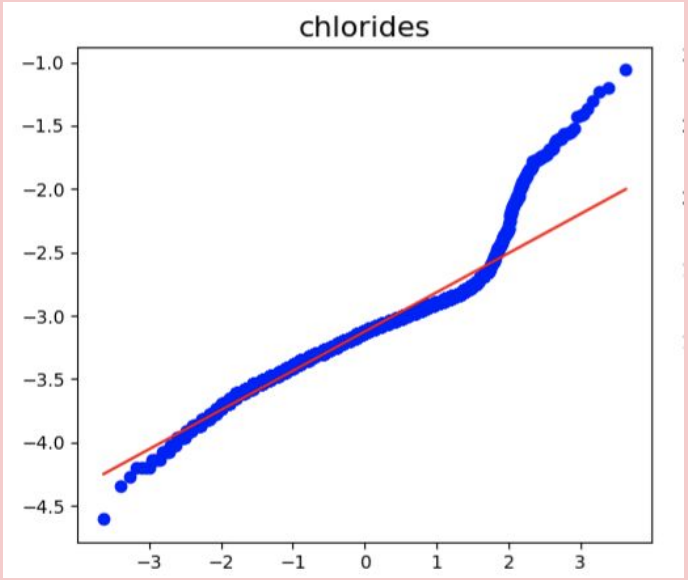


Skewness Coefficient of Chlorides: 5.02

Transformation



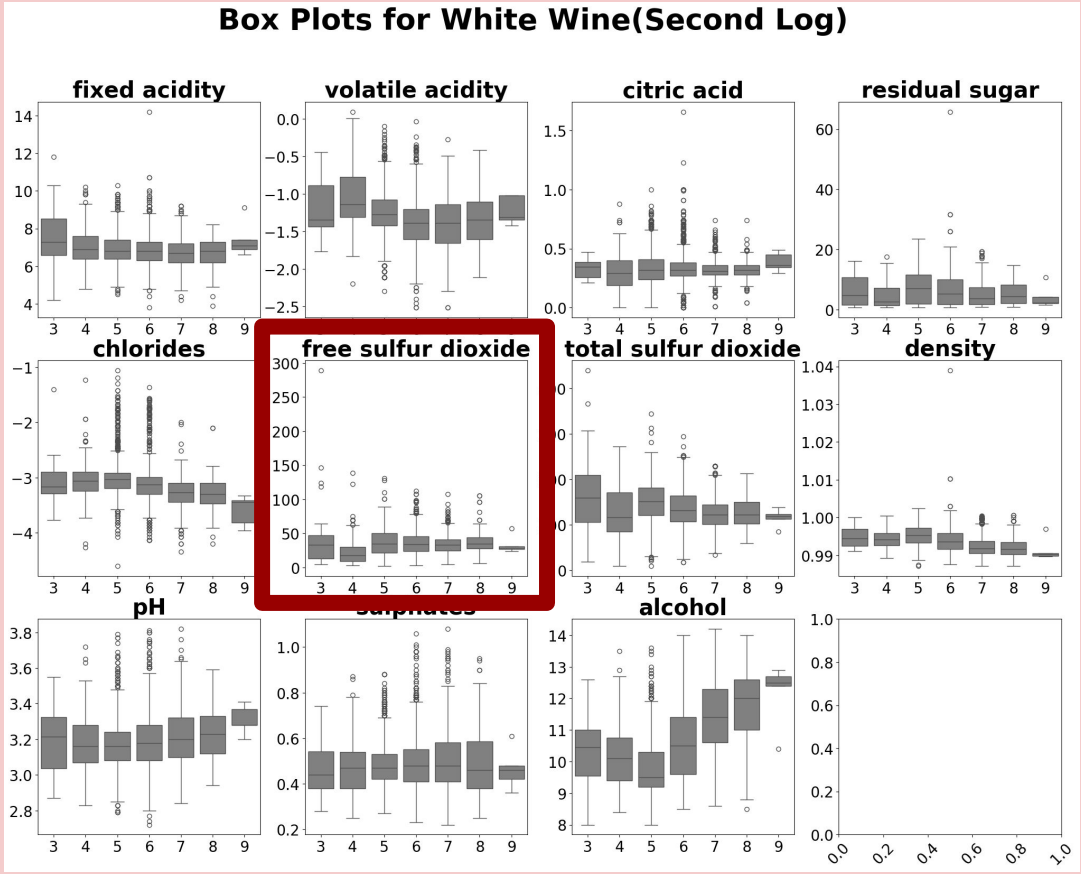
Skewness Coefficient of Chlorides: 5.02



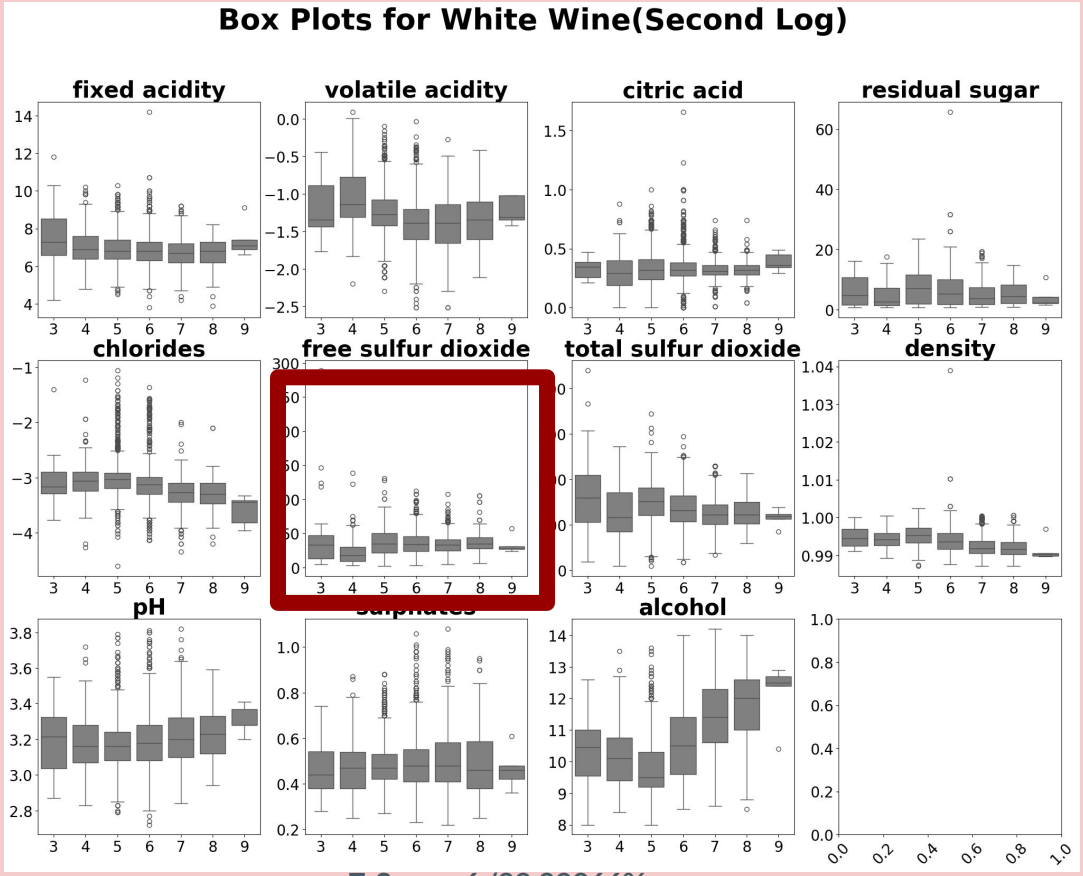
Skewness Coefficient of Chlorides(Log): 1.28

White Wine: Chlorides, Volatile Acidity
Red Wine: Residual Sugar, Chlorides

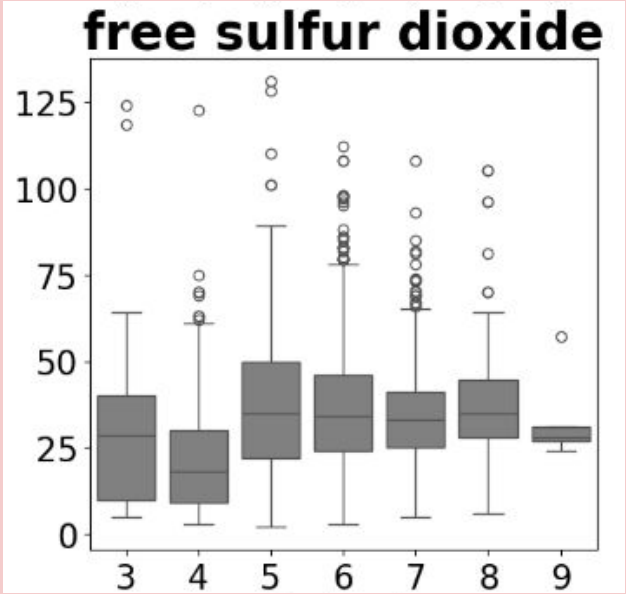
Outlier Analysis



Outlier Analysis



Z-Score: 6 /99.99966%

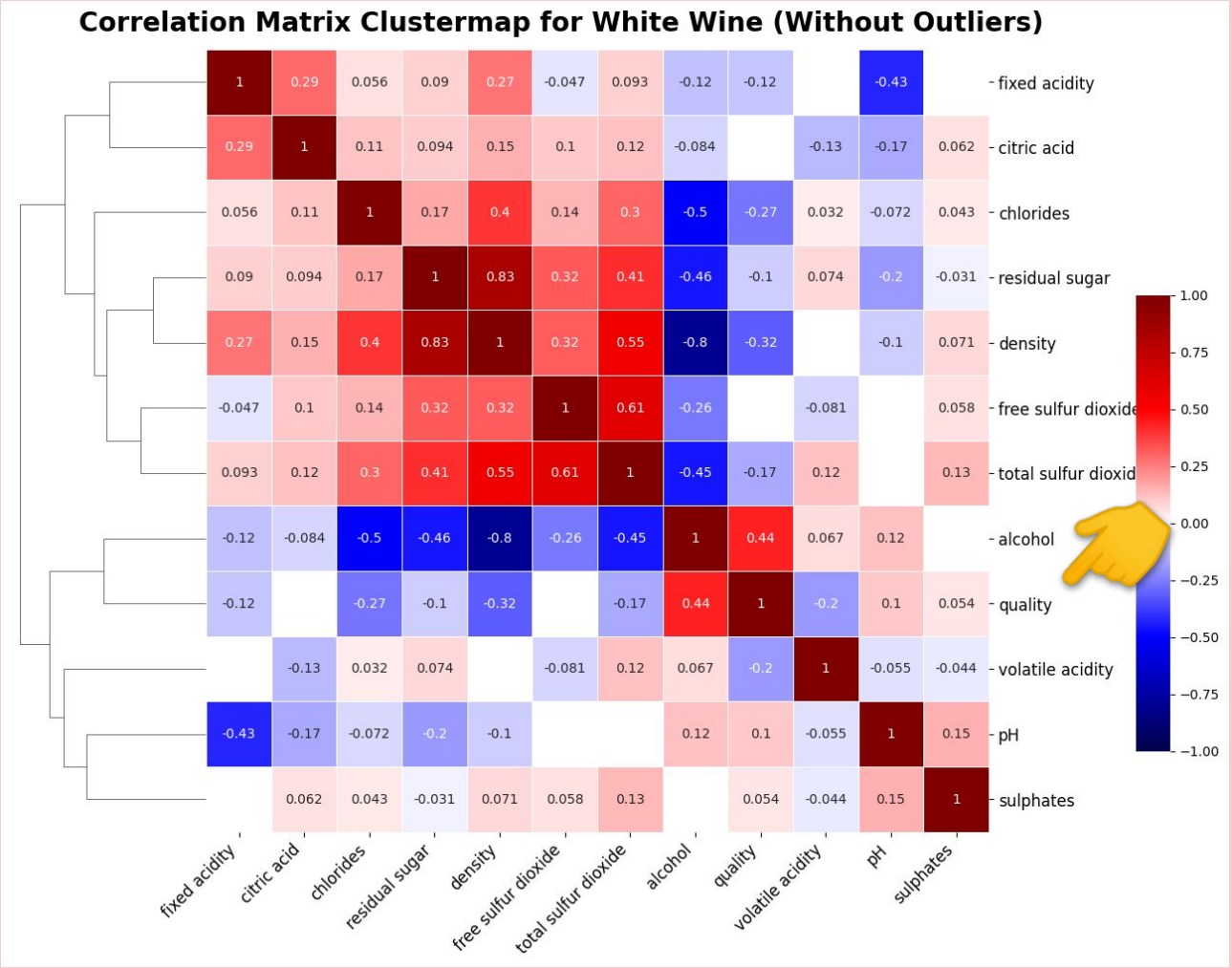


White Wine: 10 outliers; Free Sulfur Dioxide (3 outliers), Citric Acid(2 outliers); 1 outlier: Fixed Acidity, Residual Sugar, Chlorides, Total Sulfur Dioxide, Density

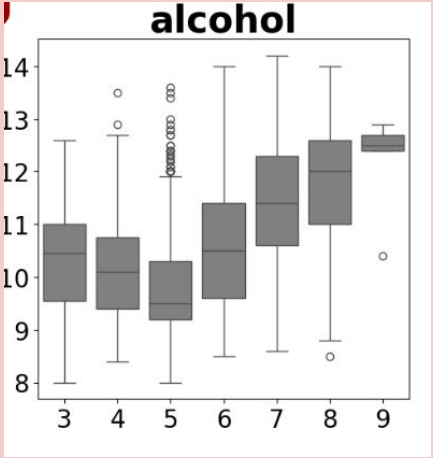
Red Wine: 8 outliers; Sulphates (4 outliers); Total Sulfur Dioxide(2 outliers); Chlorides(2 outliers).

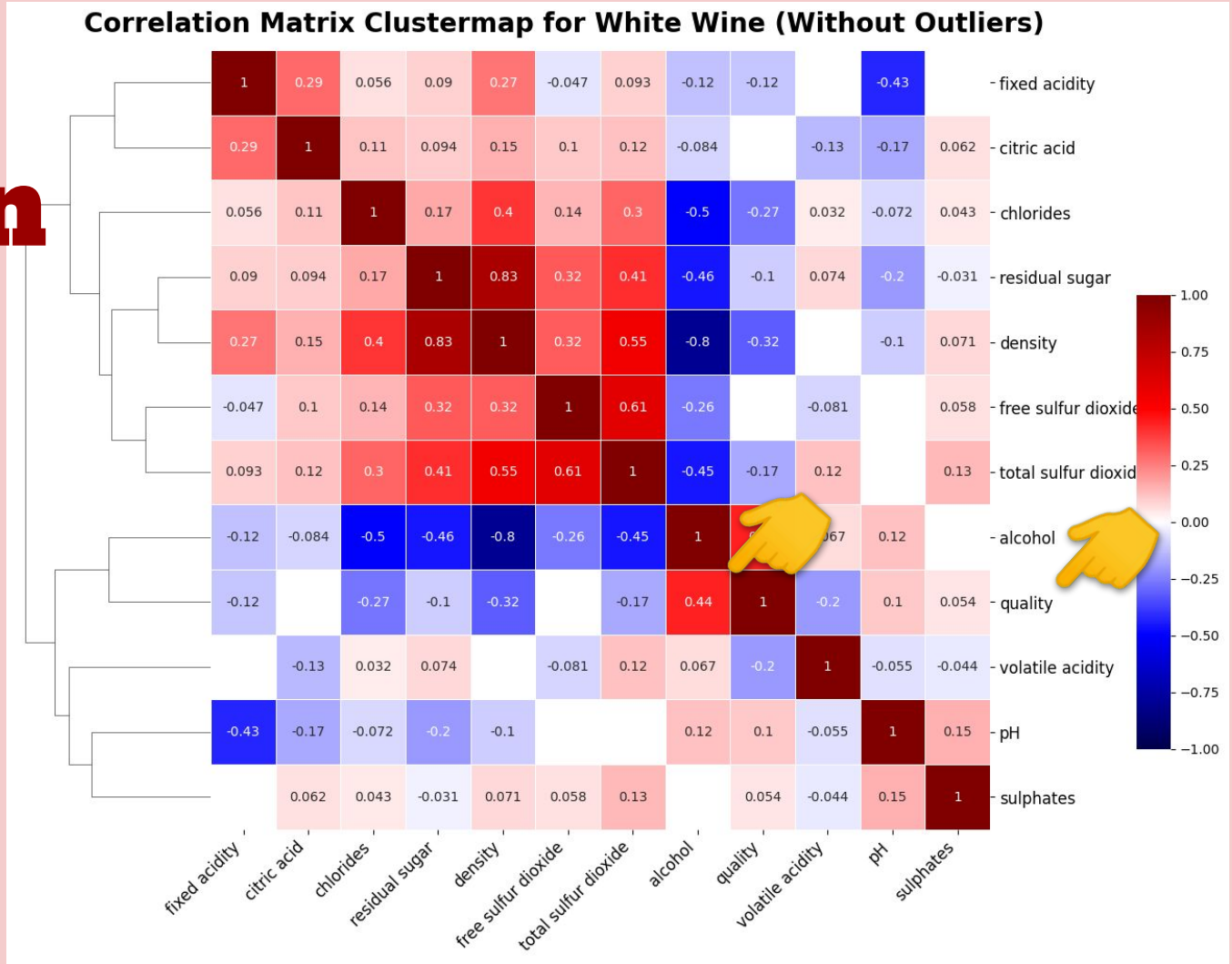
Feature Selection

What are the important features of wine quality?



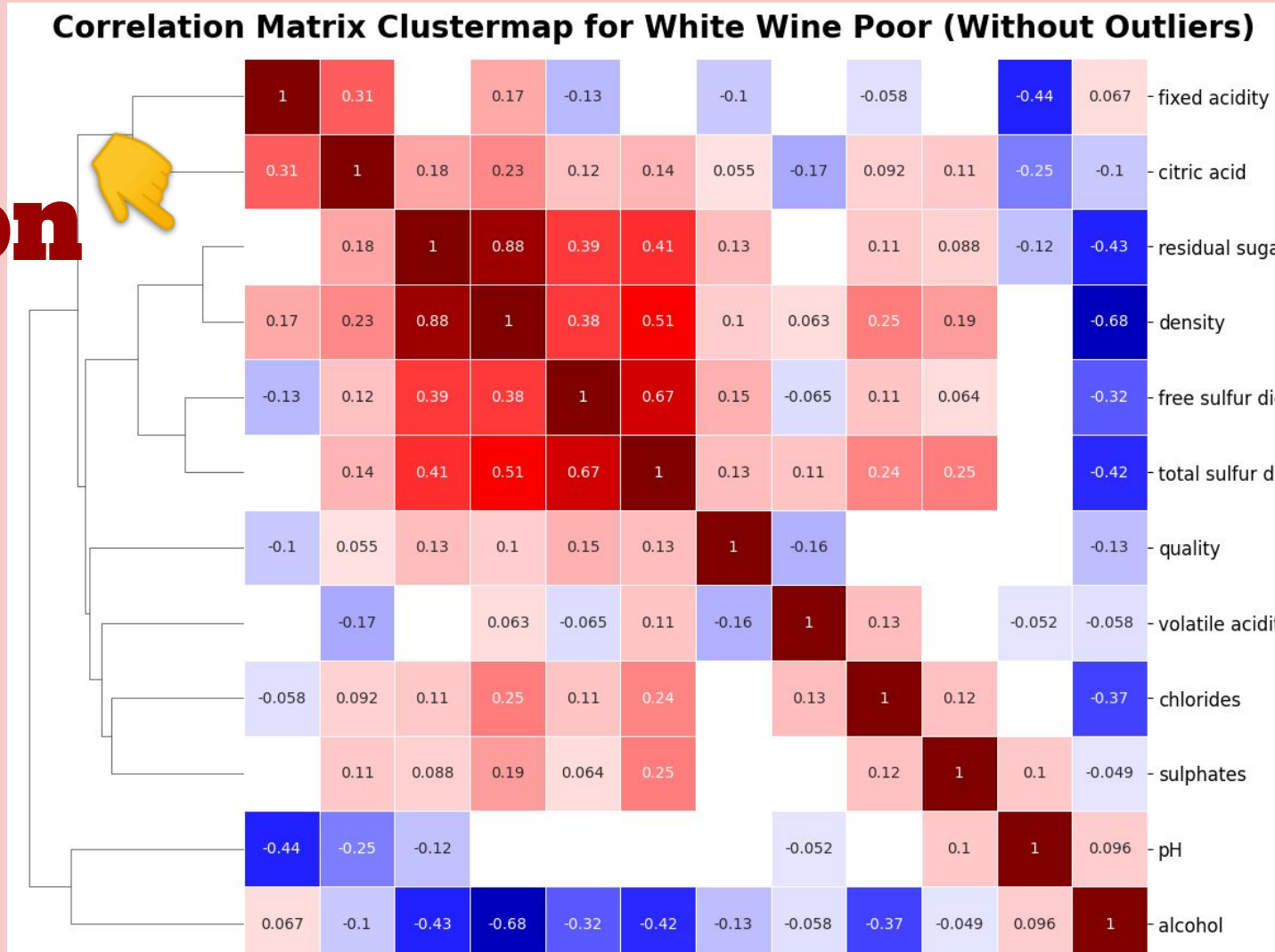
Feature Selection





Feature Selection

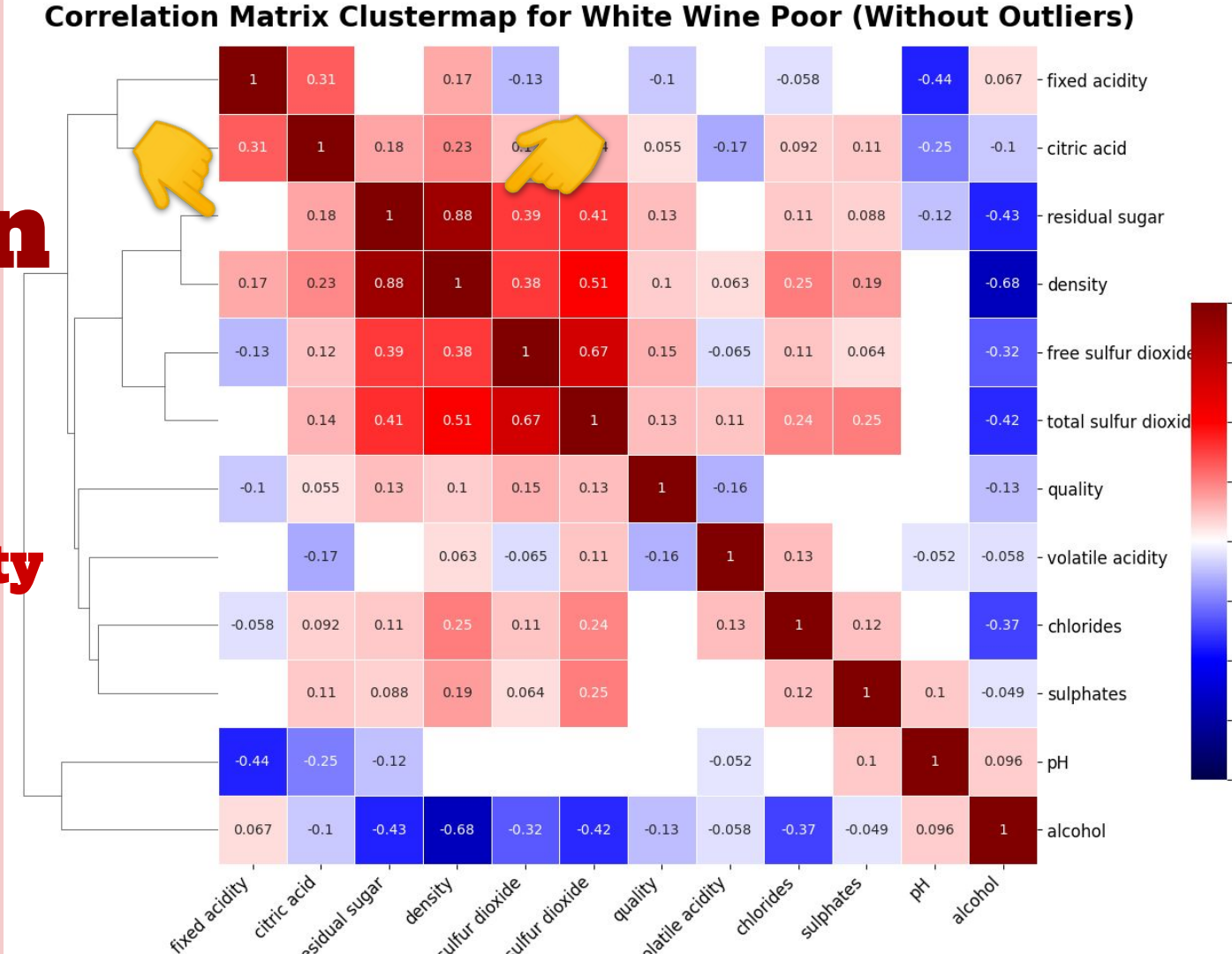
Is there high correlation between predictor variables?



Feature Selection

Multicollinearity

Threshold: 0.6



Model - Piecewise linear regression

Datasets:

- White Poor Wine (N=1639)
  - **R^2 = 0.078, F(8, 1630) = 17.34, p < .001**
  - **Volatile Acidity,** Free Sulfur Dioxide, Residual Sugar, Alcohol, **Total Sulfur Dioxide, F**ixed Acidity, Density, Citric Acid.
- White Good Wine (N=3249)
  - **R^2 = 0.118, F(8, 3240) = 62.01, p < .001**
    Alcohol, Density, Chlorides, Total Sulfur Dioxide, **Residual Sugar,** pH, Fixed Acidity
- Red Poor Wine (N=741)
  - **R^2 = 0.109, F(8, 732) = 11.20, p < .001**
  - **Volatile Acidity, Total Sulfur Dioxide,** pH, Citric Acid, Alcohol, Sulphates, Density, Residual Sugar.
- Red Good Wine (N=851)
  - **R^2 = 0.0.218, F(8, 842) = 33.52, p < .001**
  - Alcohol, volatile acidity, sulphates, chlorides, total sulfur dioxide, **residual sugar,** pH

Results

White Poor Wine

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.078
Model:	OLS	Adj. R-squared:	0.074
Method:	Least Squares	F-statistic:	17.34
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	5.67e-25
Time:	09:45:41	Log-Likelihood:	-591.66
No. Observations:	1639	AIC:	1201.
Df Residuals:	1630	BIC:	1250.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	57.0498	12.422	4.593	0.000	32.685	81.415
volatile acidity	-0.1803	0.027	-6.591	0.000	-0.234	-0.127
free sulfur dioxide	0.0002	0.001	0.249	0.804	-0.001	0.001
residual sugar	0.0222	0.005	4.837	0.000	0.013	0.031
alcohol	-0.0809	0.017	-4.675	0.000	-0.115	-0.047
total sulfur dioxide	0.0008	0.000	2.958	0.003	0.000	0.001
fixed acidity	-0.0170	0.012	-1.384	0.167	-0.041	0.007
density	-52.0447	12.437	-4.185	0.000	-76.439	-27.651
citric acid	0.0822	0.066	1.244	0.214	-0.047	0.212

Red Poor Wine

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.109
Model:	OLS	Adj. R-squared:	0.099
Method:	Least Squares	F-statistic:	11.20
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	5.10e-15
Time:	08:49:53	Log-Likelihood:	-206.41
No. Observations:	741	AIC:	430.8
Df Residuals:	732	BIC:	472.3
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.9600	10.808	-0.459	0.646	-26.178	16.258
volatile acidity	-0.5329	0.078	-6.836	0.000	-0.686	-0.380
total sulfur dioxide	0.0016	0.000	4.447	0.000	0.001	0.002
pH	-0.2705	0.094	-2.868	0.004	-0.456	-0.085
citric acid	-0.2440	0.097	-2.509	0.012	-0.435	-0.053
alcohol	0.0009	0.020	0.046	0.963	-0.038	0.040
sulphates	0.0208	0.078	0.265	0.791	-0.133	0.174
density	11.1070	10.798	1.029	0.304	-10.092	32.306
residual sugar	-0.0610	0.042	-1.446	0.148	-0.144	0.022

Similar: volatile acidity, total sulfur dioxide

Results

Good White Wine

Best Model:

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.118
Model:	OLS	Adj. R-squared:	0.116
Method:	Least Squares	F-statistic:	62.01
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	5.47e-84
Time:	01:54:18	Log-Likelihood:	-2715.8
No. Observations:	3249	AIC:	5448.
Df Residuals:	3241	BIC:	5496.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	116.8419	20.723	5.638	0.000	76.211	157.473
alcohol	0.0230	0.026	0.900	0.368	-0.027	0.073
density	-114.9411	20.973	-5.480	0.000	-156.063	-73.820
chlorides	-0.0822	0.039	-2.120	0.034	-0.158	-0.006
total sulfur dioxide	0.0003	0.000	1.116	0.265	-0.000	0.001
residual sugar	0.0521	0.008	6.553	0.000	0.037	0.068
pH	0.6748	0.102	6.620	0.000	0.475	0.875
fixed acidity	0.1030	0.021	4.829	0.000	0.061	0.145

Omnibus:	571.594	Durbin-Watson:	1.479

Good Red Wine

Best Model:

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.218
Model:	OLS	Adj. R-squared:	0.211
Method:	Least Squares	F-statistic:	33.52
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	2.81e-41
Time:	01:53:45	Log-Likelihood:	-497.80
No. Observations:	851	AIC:	1012.
Df Residuals:	843	BIC:	1050.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.0896	0.363	14.023	0.000	4.377	5.802
alcohol	0.1361	0.015	9.185	0.000	0.107	0.165
volatile acidity	-0.2645	0.103	-2.563	0.011	-0.467	-0.062
sulphates	0.6418	0.111	5.807	0.000	0.425	0.859
chlorides	-0.1992	0.054	-3.665	0.000	-0.306	-0.093
total sulfur dioxide	-0.0022	0.001	-3.392	0.001	-0.003	-0.001
residual sugar	0.1180	0.045	2.641	0.008	0.030	0.206
pH	-0.3419	0.110	-3.121	0.002	-0.557	-0.127

Omnibus:	134.696	Durbin-Watson:	1.669

Similar: volatile acidity, total sulfur dioxide

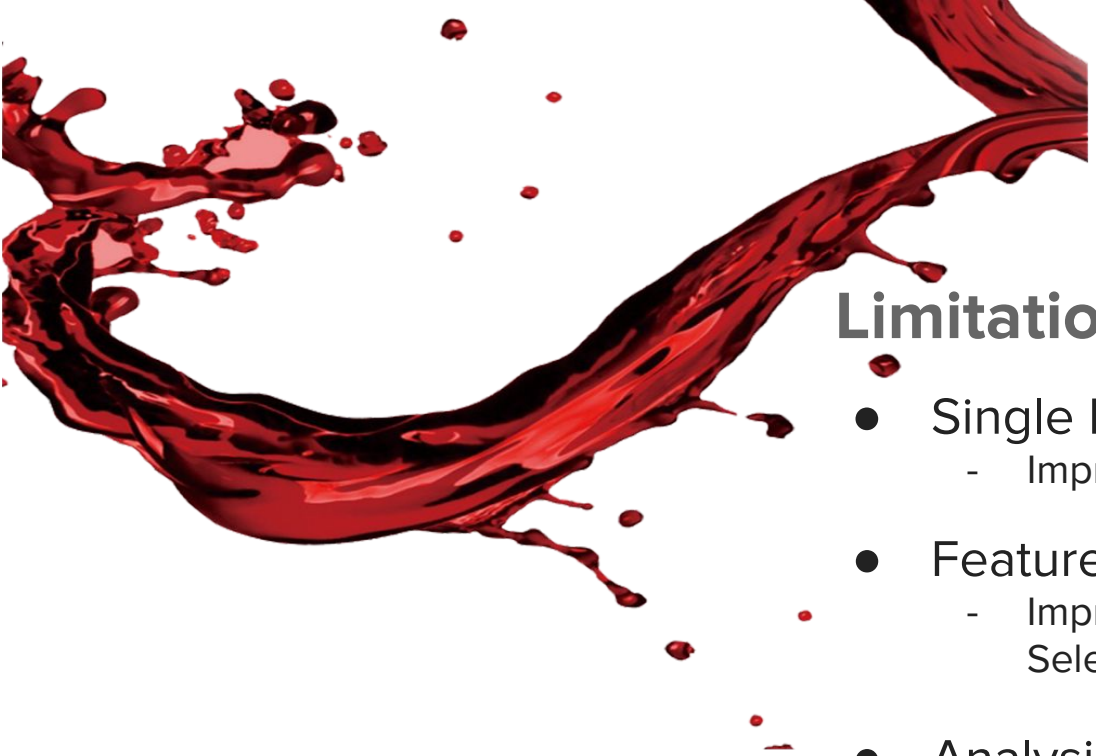
Recipes

Important Features

⬇

Mean +/- STD

- White Poor Wine
  - **Volatile acidity:** 0.20 - 0.42
  - **Free sulfur dioxide:** 16.18 - 53.87
  - **Fixed acidity:** 6.08 - 7.85
- White Good Wine
  - **Alcohol:** 9.60 - 12.09
  - **Residual sugar:** 1.22 - 10.86
  - **Volatile acidity:** 0.17 - 0.35
- Red Poor Wine
  - **Volatile acidity:** 0.41 - 0.77
  - **Total sulfur dioxide:** 17.86 - 91.21
  - **Citric acid:** 0.06 - 0.42
- Red Good Wine
  - **Alcohol:** 9.75 - 11.96
  - **Sulphates:** 0.55 - 0.83
  - **Chlorides:** 0.05 - 0.12

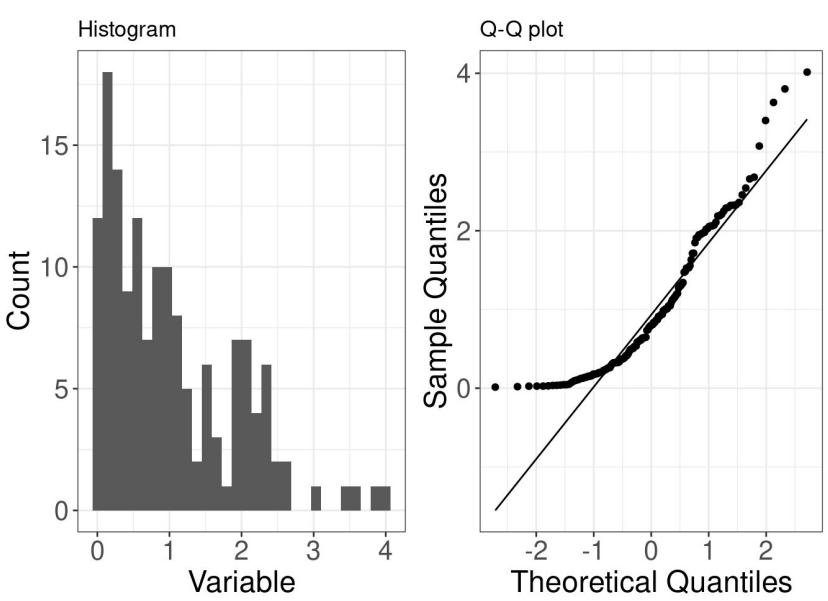


Discusion

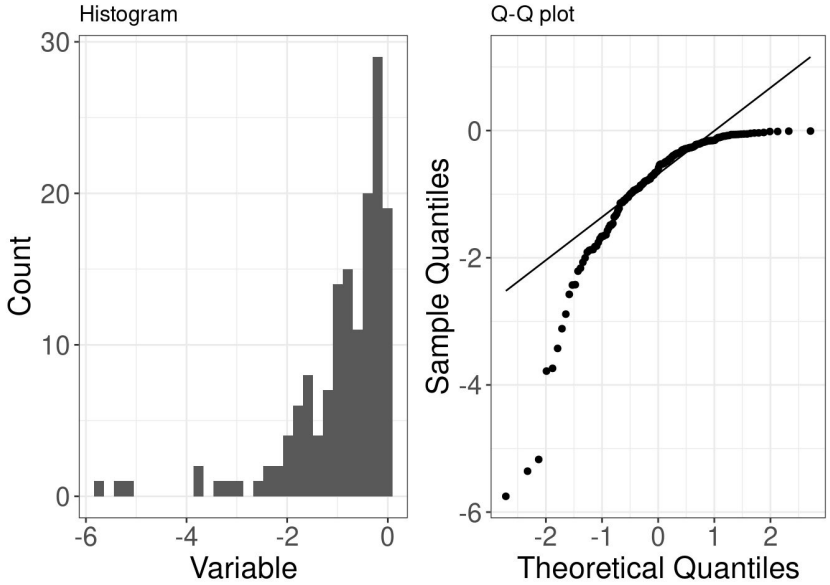
Limitations:

- Single Model Approach
  - Improvement: Compare Multiple Models
- Feature Selection
  - Improvement: Explore Alternative Feature Selection Methods (eg., PCA)
- Analysis of only psychochemical data
  - Improvement: Expand Dataset for Generalization (information about weather, temperature, year, region)

Appendix

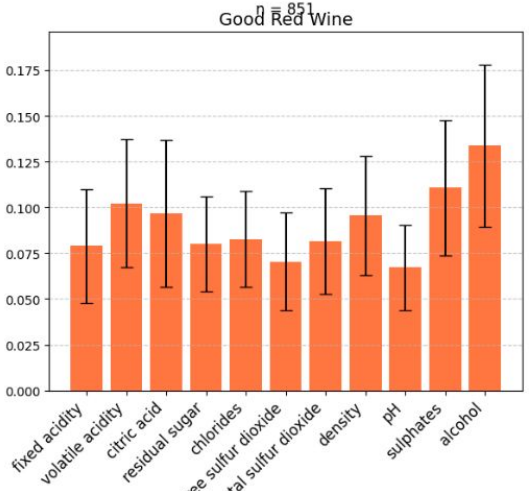


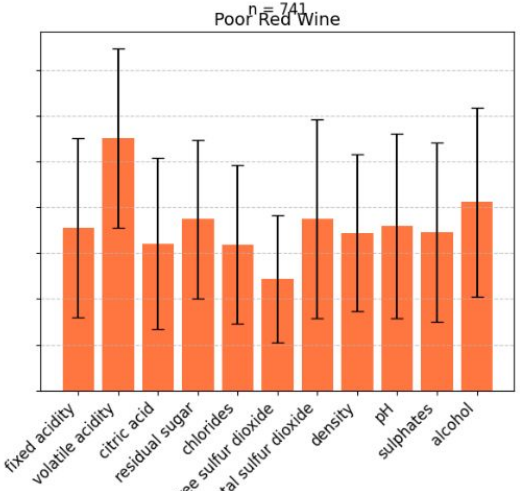
Right-skewed data(Positive)

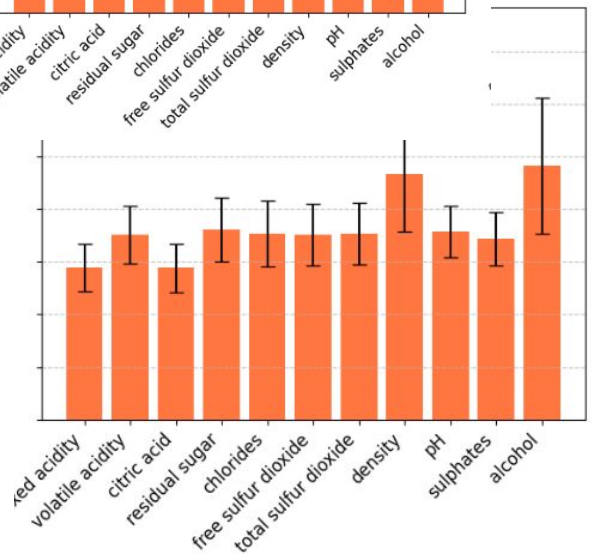


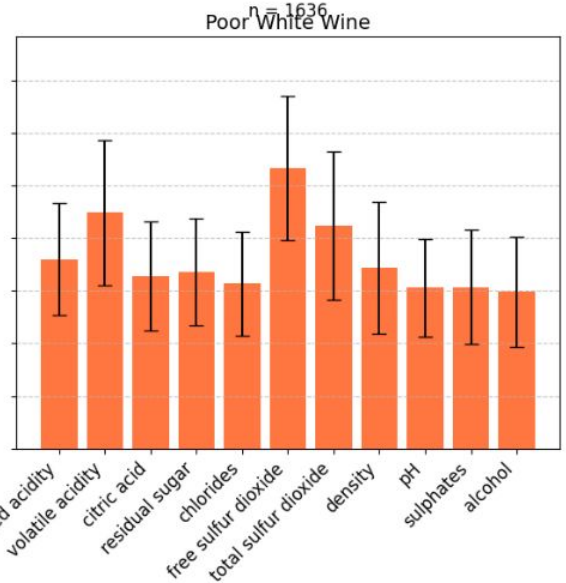
Left-skewed data(Negative)

Appendix


n = 851
Good Red Wine
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol


n = 741
Poor Red Wine
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol


xed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol


n = 1636
Poor White Wine
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol