EBI02156 - Technical Test

Results report

This report organizes and illustrates the analysis made on the wine datasets point by point, in the same structure as the instructions provided, with some minor modifications, specifically:

1. Load the datasets into a Pandas DataFrame

2. Perform exploratory data analysis (EDA) to get a sense of the data

3. Compute and illustrate the correlation matrix for all numerical variables.

4. Numerals 3. And 4. were joined, and correspond to creating a correlation matrix and illustrating it with a heatmap.

5. Perform VIF analysis and feature engineering.

6. Perform a multiple linear regression analysis and interpret the regression results, including coefficients, p-values, and R-squared.

7. Provide recommendations based on your analysis.

The code lines used as they appear in the executable analysisCode.py will be referenced, as well as the path of each graph and table will be shown in blue.

1. The `pandas.read_csv()` command was used to perform this step on lines 10 and 13

2. EDA:

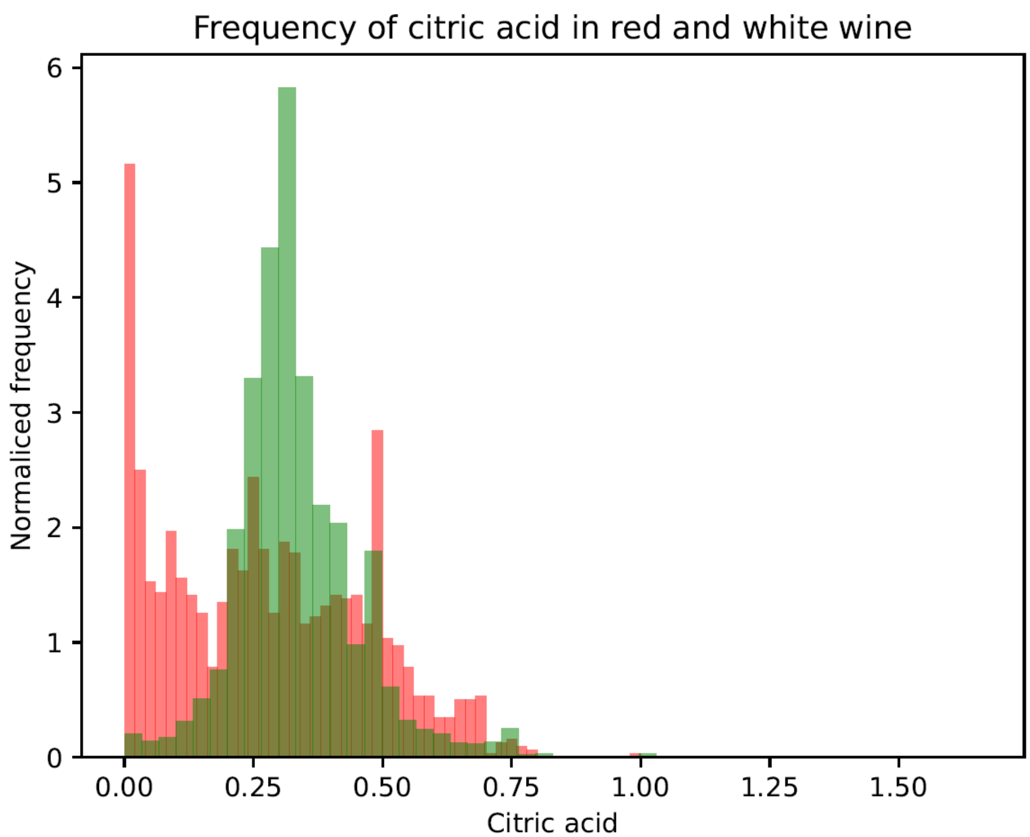
First, general statistics for the whole dataset were observed, example columns are shown:

	fixed acidity	free sulfur dioxide	total sulfur dioxide	density	color
count	6497	6497	6497	6497	6497
mean	7.215307065	30.52531938	115.7445744	0.994697	0.246114
std	1.296433758	17.74939977	56.52185452	0.002999	0.430779
min	3.8	1	6	0.98711	0
25%	6.4	17	77	0.99234	0
50%	7	29	118	0.99489	0
75%	7.7	41	156	0.99699	0
max	15.9	289	440	1.03898	1

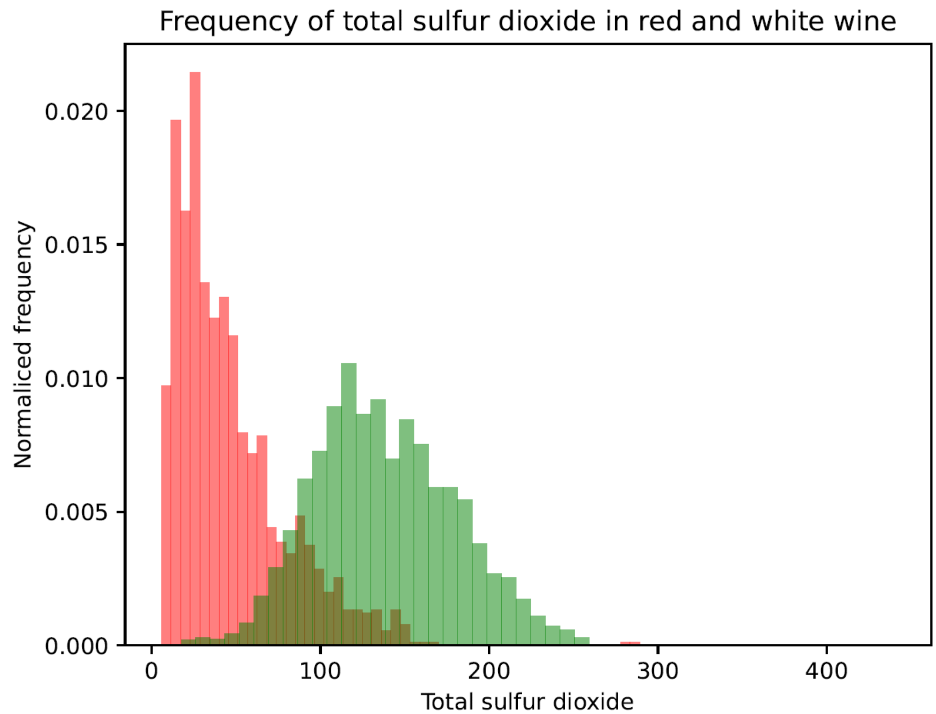
Line 22 at Plots/EDA/SummaryStatistics.csv

Datasets for red and white wine were annotated with the additional 'color' column, with 1 for red and 0 for white. The main takeaways here are that the scale of the variables is evidently different so normalization will be needed. All variables are numeric, so no transformations are needed. And finally, there are no missing data so no completion or subsampling must be performed.

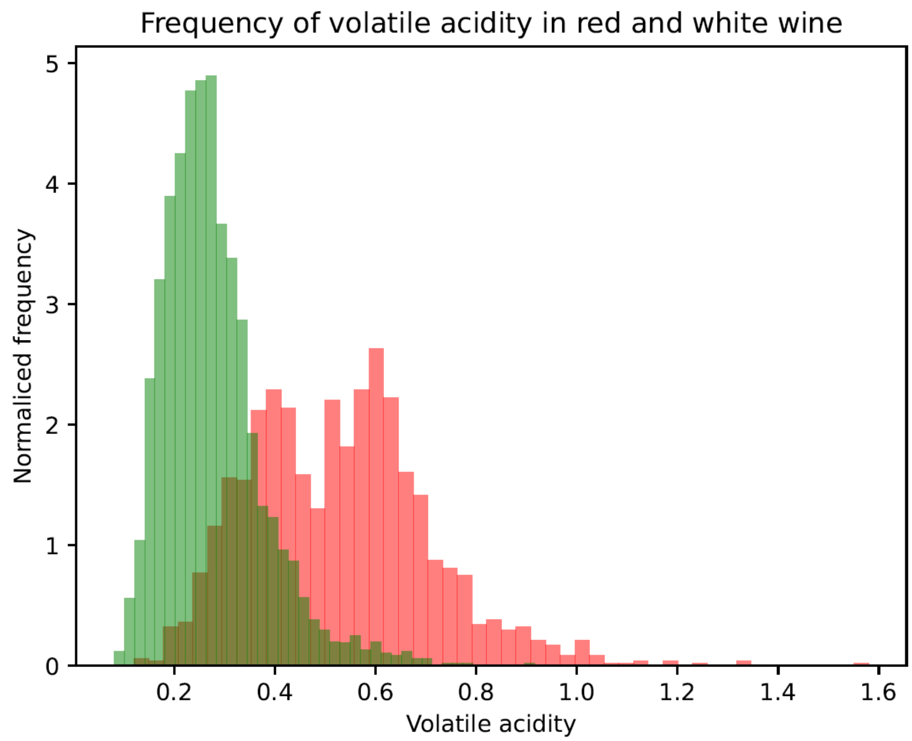
A decision has to be made whether to separate red and white wine data; a quick insight by comparing the distribution of some variables for both wines is shown below: white wine is presented in green while red wine is presented in red.



lines 29 to 35 at Plots/EDA/FrequencyHistCitricAcid.pdf



lines 29 to 35 at Plots/EDA/FrequencyHistTotalSulfurDioxide.pdf

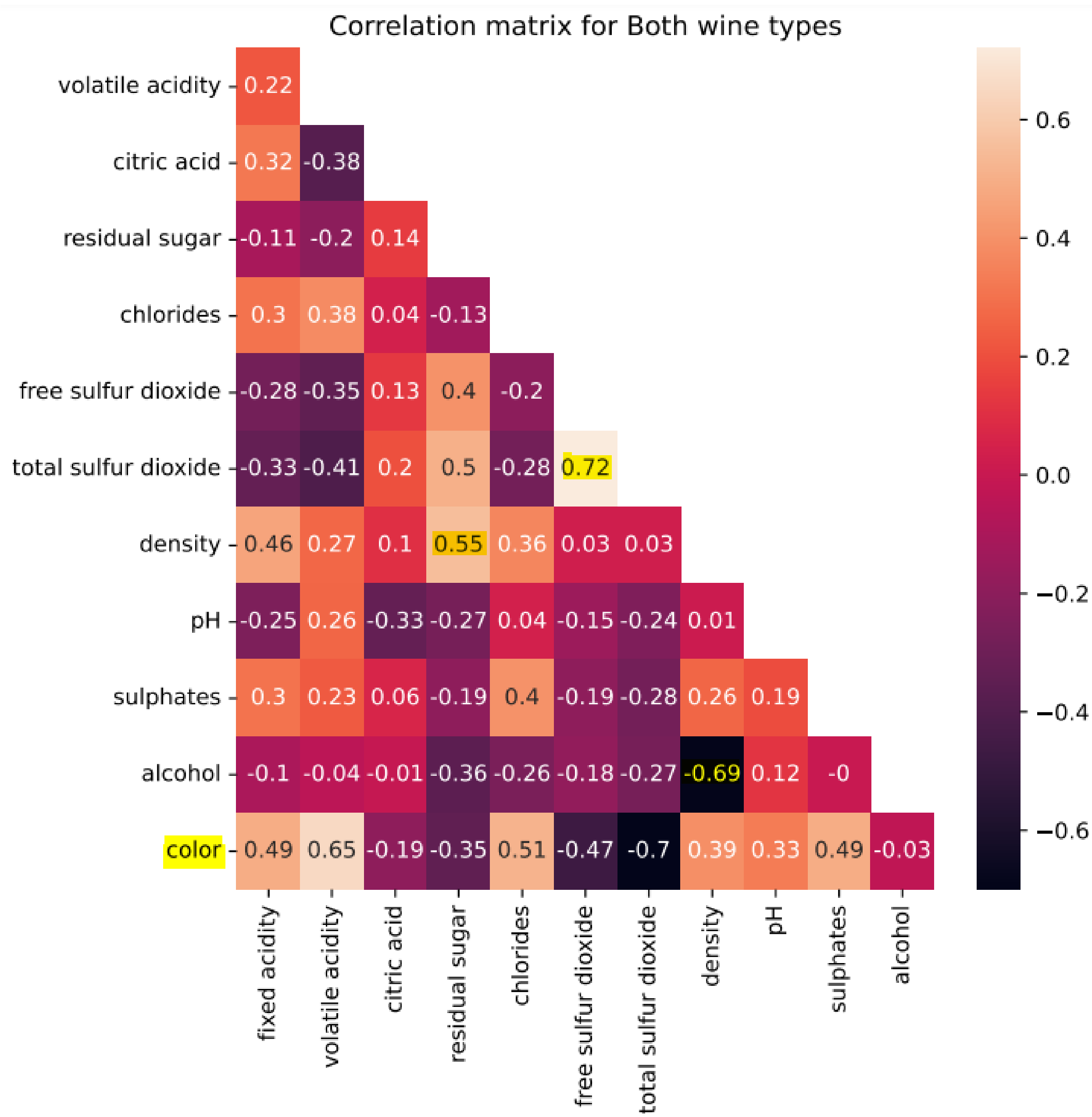


lines 29 to 35 at Plots/EDA/FrequencyHistVolatileAcidity.pdf

These plots show clear differences between the behavior of the variables for both wine types.

3. Correlation Matrix analysis, numerals **3 & 4** are included here; matrices in both pdf and csv format are available at Plots/Correlation/

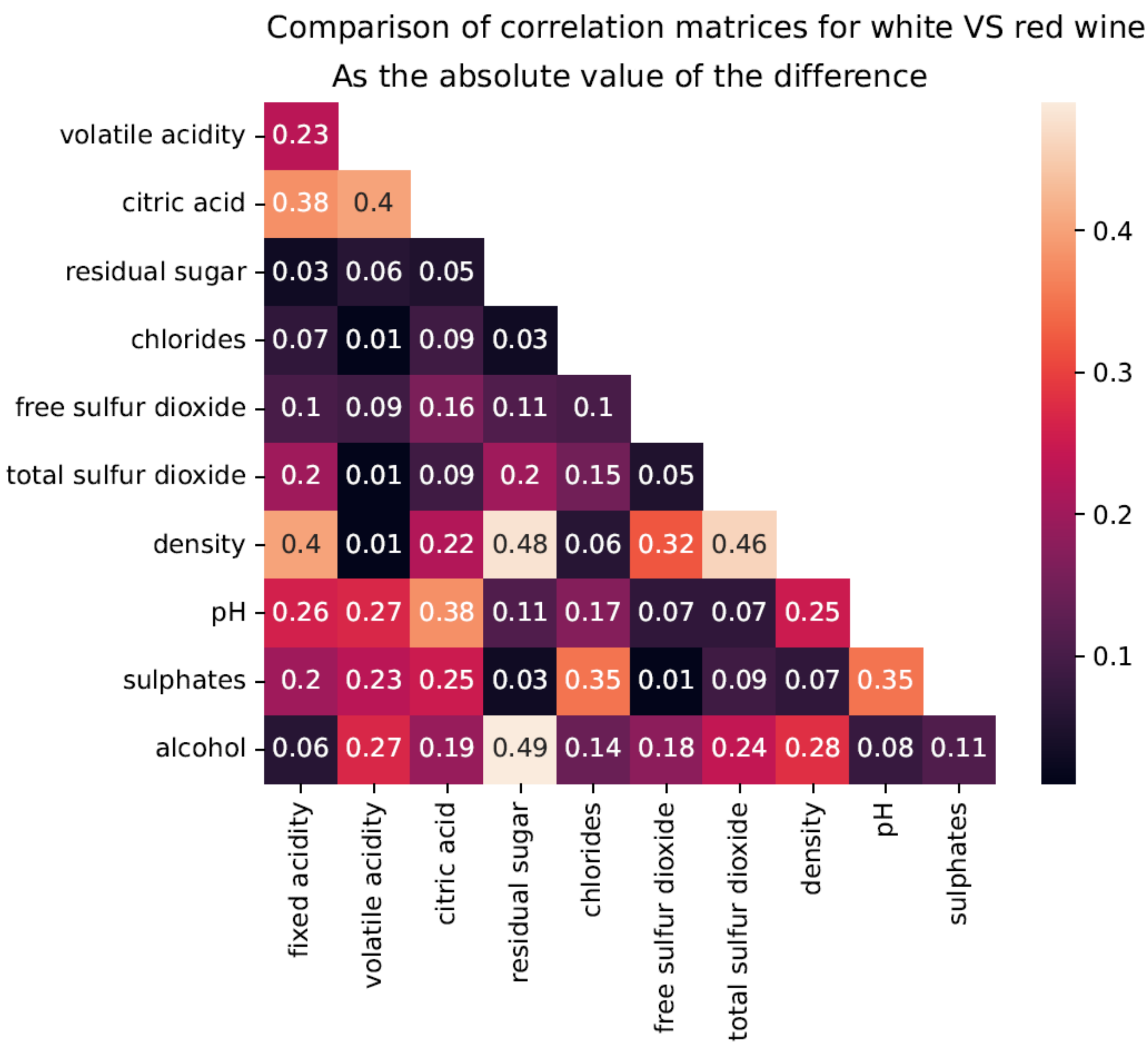
Here we have a visualization of a correlation matrix for the whole data.



Line 46 at Plots/Correlation/CorrelationMatrixBoth.pdf

We could make many conclusions looking at this matrix, but I want to focus on just a few things: High values on the last column 'color' indicates how red and white wines behave differently. The highest correlations we have are between density and alcohol, density and residual sugar, and total sulfur dioxide and free sulfur dioxide; We will take this into account for feature engineering.

I made correlation matrices for both white and red wine, which are available at Plots/Correlation/ but for comparison I'd prefer to show the absolute difference between the two of them.



Line 46 at Plots/Correlation/CorrelationMatrixComparison.pdf

If the qualities of both wines behaved similarly, we would expect to find only small values. The contrary was found, as differences bigger than 0.3 are common, so not only the value of the variables is different, but also how they interact with each other.

5. Multiple linear regression analysis was performed on the three datasets but before this I decided to make an additional analysis to evaluate variance inflation; which can alter the model in unpredictable ways.

A good rule of thumb is to avoid having values equal or higher than 5. This was accomplished by creating a new variable ‘fermentation’ as the average between 'density', 'residual sugar' and 'alcohol'; and another 'sulfur dioxide' as the average of 'free sulfur dioxide' and 'total sulfur dioxide'. csv file with values before and after feature engineering are available at Plots/VarianceInflation for all three datasets, with prefix Transformed indicating which ones represent VIF after the transformation. All were generated after variable normalization done by subtracting the mean and dividing by the standard deviation.

VIF tables after transformation are shown below:

For red and white wine:

Columns	VIF
fixed acidity	2.176578
volatile acidity	2.059113
citric acid	1.59318
chlorides	1.521999
pH	1.554657
sulphates	1.440135
color	4.187505
fermentation	1.143547
sulfur dioxide	1.834406

Line 118 at Plots/VarianceInflation/TransformedRedAndWhiteWine.csv

For white wine:

Columns	VIF
fixed acidity	1.324967
volatile acidity	1.262193
citric acid	1.159765
chlorides	1.1331
pH	1.22709
sulphates	1.114464
fermentation	1.155553
sulfur dioxide	1.21579

Line 118 at Plots/VarianceInflation/TransformedWhiteWine.csv

For red wine:

Columns	VIF
fixed acidity	3.774464
volatile acidity	3.15619
citric acid	3.109545
chlorides	1.961203
pH	2.597756
sulphates	1.86446
fermentation	1.430803
sulfur dioxide	3.507795

Line 118 at Plots/VarianceInflation/TransformedRedWine.csv

We see a very interesting pattern here, as red wine presents overall higher values than white wine, with the dataset of both having values somewhere in between, more evidence that they behave differently. With the resulting VIF values we conclude our feature engineering was successful.

6. Multiple linear regression results and analysis

Statistics for the whole model and for each variable were produced for our three datasets, all available at Plots/ModelStatistics. I'll present the reports for red and for white wine but feel free to look at the tables for the model that has both; it generally has values intermediate between the other two models.

White wine:

Statistics for the whole model:

Model:	OLS	Adj. R-squared (uncentered):	0.099
Dependent Variable:	y	AIC:	13555.98
Date:	10/6/2023 17:15	BIC:	13607.95
No. Observations:	4898	Log-Likelihood:	-6770
Df Model:	8	F-statistic:	68.32
Df Residuals:	4890	Prob (F-statistic):	7.78E-107
R-squared (uncentered):	0.101	Scale:	0.93067

Line 142 at Plots/ModelStatistics/whiteWineGeneralStatistics.csv

A rather small R-squared value indicates that the model doesn't have a great predictive power, even with a relatively high number of observations. The P value (Prob F-statistic) indicates that there is however a statistically significant relation between our input and output variables.

Statistics for each input variable:

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
fixed acidity	-0.12716	0.022418	-5.67226	1.49E-08	-0.17111	-0.08321
volatile acidity	-0.26692	0.021602	-12.3561	1.47E-34	-0.30927	-0.22457
citric acid	0.011589	0.017682	0.655414	0.512232	-0.02307	0.046252
chlorides	-0.24413	0.021298	-11.4624	4.92E-30	-0.28588	-0.20238
pH	0.062173	0.015944	3.899357	9.77E-05	0.030915	0.093431
sulphates	0.090043	0.017837	5.048086	4.62E-07	0.055075	0.125012
fermentation	0.181179	0.0288	6.291023	3.43E-10	0.124719	0.237639
sulfur dioxide	-0.15218	0.018108	-8.40392	5.60E-17	-0.18767	-0.11668

Line 145 at Plots/ModelStatistics/whiteWineFeaturesStatistics.csv

All variables show small P values (P>|t|) indicating a significant impact on wine quality; with volatile acidity and chlorides showing the biggest impact in their coefficients, both negatively impacting quality.

Red wine:

Whole model:

Model:	OLS	Adj. R-squared (uncentered):	0.324
Dependent Variable:	y	AIC:	3748.342
Date:	10/6/2023 17:15	BIC:	3791.359
No. Observations:	1599	Log-Likelihood:	-1866.2
Df Model:	8	F-statistic:	96.76
Df Residuals:	1591	Prob (F-statistic):	3.25E-131
R-squared (uncentered):	0.327	Scale:	0.60733

Line 151 at Plots/ModelStatistics/redWineGeneralStatistics.csv

A bigger R-squared indicates a stronger predictive power, as well as a lower p-value even with less than half the amount of observations of white wine; this model is clearly more promising.

Input variables:

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
fixed acidity	-0.13078	0.023813	-5.49183	4.62E-08	-0.17749	-0.08407
volatile acidity	-0.26419	0.021949	-12.0364	5.36E-32	-0.30724	-0.22114
citric acid	-0.0277	0.02491	-1.11188	0.266356	-0.07656	0.021163
chlorides	-0.1172	0.016898	-6.93584	5.85E-12	-0.15035	-0.08406
pH	-0.11903	0.02806	-4.24199	2.34E-05	-0.17407	-0.06399
sulphates	0.162876	0.018706	8.707181	7.65E-18	0.126185	0.199567
fermentation	0.944633	0.074606	12.66159	4.46E-35	0.798297	1.09097
sulfur dioxide	-0.22181	0.031561	-7.02807	3.10E-12	-0.28372	-0.15991

Line 154 at Plots/ModelStatistics/redWineFeaturesStatistics.csv

Most significant here is the Coefficient value for fermentation, the engineered variable comprised by 'density', 'residual sugar' and 'alcohol' as its high values indicates a very strong impact on quality. All variables show a significant p-value.

7. Recommendations:

- Don’t generalize for white and red wine, as they show different behaviors.
- All the variables collected show a correlation with quality, but their impact is not very strong, so keep that in mind.
- Volatile acidity and chlorides show the biggest impact on white wine; pay them special attention, nevertheless, the impact of density, residual sugar and alcohol in red wine is far greater.
- Linear models might be very easy to understand, and to draw conclusions from, but keep in mind that they are limited to linear relations, evaluating other models could provide a greater insight into the relationships present in our data.

Runtime for the analysis ranged from 4 to 6 seconds, but I must disclaim I am using a rather powerful and modern machine built with data analysis capabilities on mind, so it might take longer depending on the hardware used.