Analysis of wines parameters

Executive Summary

This document presents an analysis of data concerning wines parameters and their types. The analysis is based on 6497 observations of wines, each containing specific characteristics of a wine.

Document contains data summary and descriptive statistics. After exploring the data, a **predictive model to classify wines into two categories: white (marked by 0) and red (marked by 1)**.

Initial data exploration

Individual Feature Statistics

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.52532
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.7494
min	3.8	0.08	0	0.6	0.009	1
25%	6.4	0.23	0.25	1.8	0.038	17
50%	7	0.29	0.31	3	0.047	29
75%	7.7	0.4	0.39	8.1	0.065	41
max	15.9	1.58	1.66	65.8	0.611	289

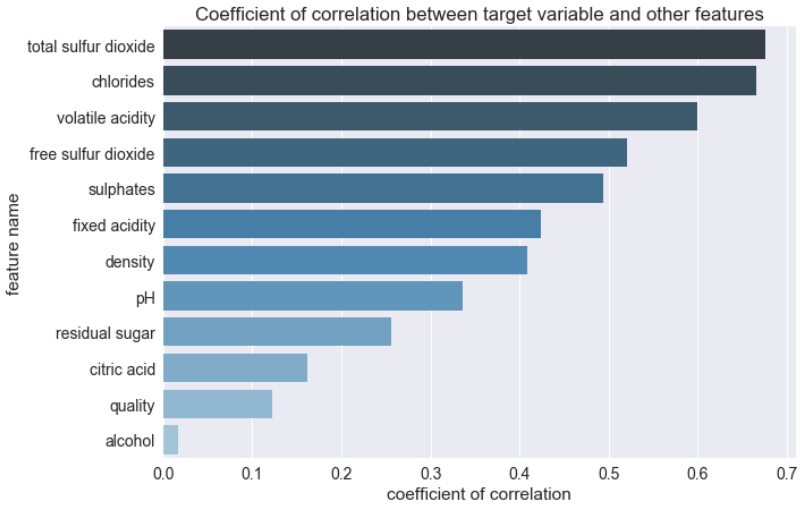
	total sulfur dioxide	density	pH	sulphates	alcohol	quality
mean	115.7446	0.994697	3.218501	0.531268	10.4918	5.818378
std	56.52185	0.002999	0.160787	0.148806	1.192712	0.873255
min	6	0.98711	2.72	0.22	8	3
25%	77	0.99234	3.11	0.43	9.5	5
50%	118	0.99489	3.21	0.51	10.3	6
75%	156	0.99699	3.32	0.6	11.3	6
max	440	1.03898	4.01	2	14.9	9

Correlation and Apparent Relationships

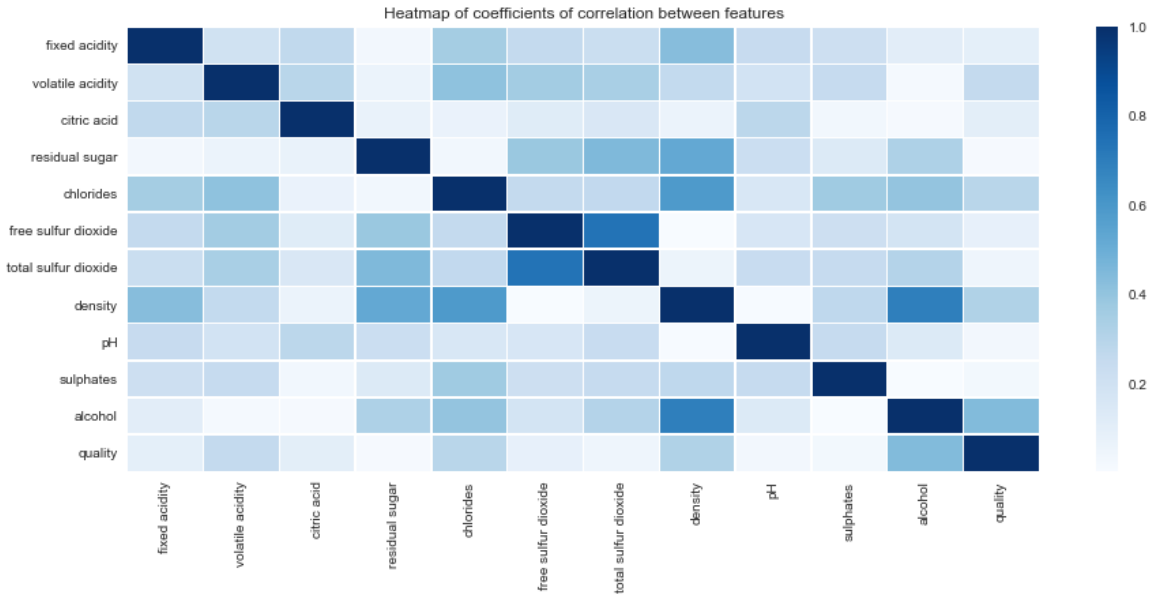
After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between wine class and the other features.

Numeric Relationships

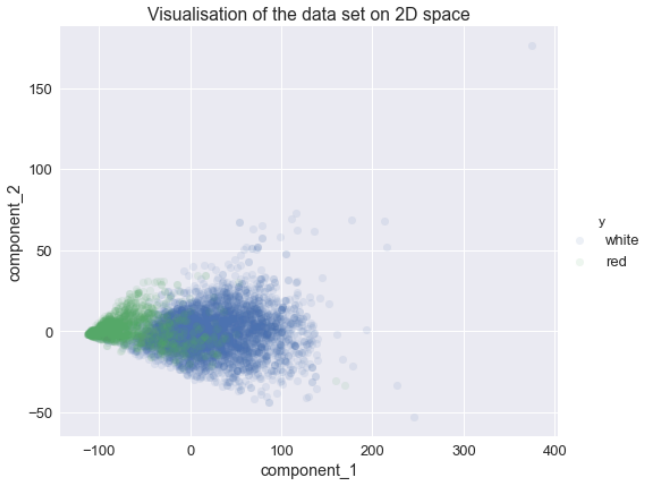
The following plot was generated initially to compare numeric features with target (wine class). The key features in this matrix are shown here:



The correlation between the numeric columns was then calculated with the following results shown in plot below:



Before started predictive analysis I shown that wines can be separate into whites and reds using available features. The graph below shows all wines with its parameters decomposed into two principal components.



Classification of wines

To build classification model I used decision tree classifier. By using numbers of techniques from the field of statistics and machine learning I was able to build high quality predictive model with quality parameters:

- Gini score: 0.983.
- Recall score: 0.962.

It means that model is able to correctly recognize over 96% of red wines.

While many factors can help indicate kind of a wine, significant features found in this analysis were:

Feature name	Importance level (higher – more important feature)
chlorides	0.625643
total sulfur dioxide	0.288745
density	0.041088
fixed acidity	0.029272
alcohol	0.015252

Summary

This analysis has shown that type of wine can be confidently predicted from its characteristics. In particular, level of chlorides, total sulfur dioxide, density, fixed acidity, and alcohol have a significant effect on the wine type.