Group 4
Huan He | Qijia Zheng | Kelsey Strachan

Modelling Wine Quality Distribution

Data Collection and Statistical Description

The wine quality dataset was meticulously gathered by a team of researchers comprised of Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis.[1,2] These experts are associated with the Department of Information Systems and the R&D Centre Algoritmi at the University of Minho, located in Guimarães, Portugal. They are also connected to the Viticulture Commission of the Vinho Verde region (CVRVV) in Porto. This comprehensive dataset encompasses both physicochemical and sensory data derived from two distinct types of Portuguese "Vinho Verde" wine: red and white. Vinho Verde, renowned for its refreshing taste, is a unique wine variety that enjoys particular popularity during the summer months.[2] The data was specifically sourced from the Vinho Verde wine region in the northwest of Portugal, an area celebrated for its distinctive viticultural practices and the unique characteristics of its wines.

The collection of this dataset spanned nearly three years, from May 2004 to February 2007. This extended timeframe ensured a robust representation of the variability inherent in wine production and quality over different seasons and vintages. The data collection process was methodical and multi-faceted, involving three key steps: physicochemical tests, sensory assessments, and data recording. Scientific factors such as seasonal variability, subjective tasting assessments, and minor testing inconsistencies may have influenced the dataset. Despite these considerations, the dataset provides a valuable resource for studying wine quality and developing predictive models.

The white wine data set consists of 4,898 observations, all complete with no missing values. The red wine data set contains 1,599 observations and also reports no missing observations. Both datasets share the same features and identical data types. The features included in the datasets are `` `fixed acidity` ``, `` `volatile acidity` ``, `` `citric acid` ``, `` `residual sugar` ``, `` `chlorides` ``, `` `free sulfur dioxide` ``, `` `total sulfur dioxide` ``, `` `density` ``, `` `pH` ``, `` `sulphates` ``, `` `alcohol` ``, and `` `quality` ``. All features are of type `` `float64` ``, except for `` `quality` `` which is of type `` `int64` ``. Preliminary examination of feature distributions is shown below for both white wine and red wine datasets (Figures 1 and 2).

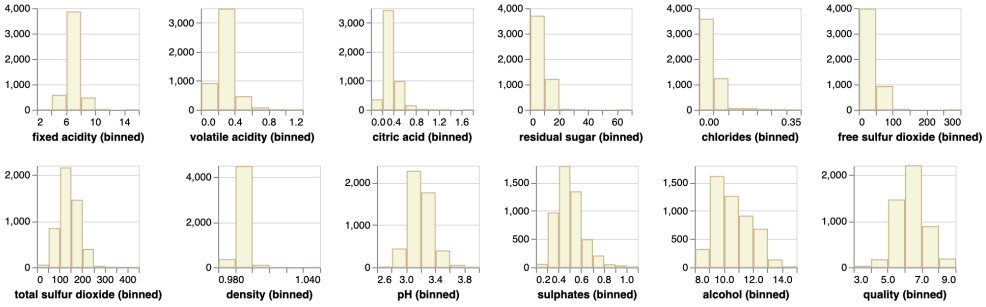

Figure 1. White Wine Feature Distribution

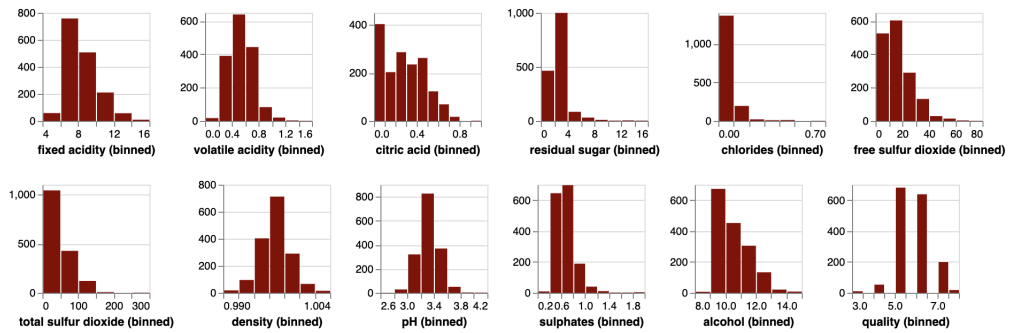

Figure 2. Red Wine Feature Distribution

Scientific Questions

Through our analysis, our group aims to answer the following questions:

1. How can we **model and predict the distribution of wine quality** based on physicochemical properties such as acidity, alcohol content, and sulfur dioxide levels? Wine quality is graded on a scale that ranges from 0 (very poor) to 10 (excellent). By analyzing the relationship between physicochemical features and wine quality, we aim to build predictive models to estimate the quality distribution for new wine samples.
2. Are there significant **differences in the physicochemical properties** between white and red wines, and how do these differences influence their quality? White and red wines are produced using different processes, which may lead to differences in their chemical composition. We aim to explore these differences and understand how production methods affect wine quality.
3. Which physicochemical features are **most strongly correlated** with wine quality, and how do they contribute to the perceived quality of wine? We seek to identify the most influential features and provide insight into improving wine quality

Appendix

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
5	7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
6	7.9	0.60	0.06	1.6	0.069	15.0	59.0	0.9964	3.30	0.46	9.4	5
7	7.3	0.65	0.00	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
8	7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
9	7.5	0.50	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.80	10.5	5

Table 1. Red Wine Data Extract

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
6	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6
7	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
8	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
9	8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6

Table 2. White Wine Data Extract

References

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Wine Quality*. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.
2. Cortez, P., Almeida, F., & Matos, T. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.