Group 4
Huan He | Qijia Zheng | Kelsey Strachan

Wine Quality Exploratory Data Analysis

Section I: Statistical Analysis

Assessing wine quality is crucial within the viticulture industry to optimize production and revenue. In this exploratory data analysis (EDA), we examine two related wine datasets to uncover notable patterns and insights. The wine quality dataset utilized in this analysis was developed by researchers from the Department of Information Systems and R&D Centre Algoritmi at the University of Minho in Guimarães, Portugal.[1] This institution collaborates with the Viticulture Commission of the Vinho Verde region (CVRVV) in Porto. The analysis dataset includes physicochemical and sensory data derived from two distinct varieties of Portuguese "Vinho Verde" wine - red and white - sourced from the Vinho Verde wine region in the northwest of Portugal. Spanning from May 2004 to February 2007, this dataset ensures a comprehensive representation of the variability in wine production and quality over different seasons and vintages.[2] As per the researchers, the data collection process involved three key stages: physicochemical tests, sensory evaluations, and data recording. While factors such as seasonal fluctuations, subjective assessments, and testing inconsistencies may have impacted the data, the dataset remains a valuable resource for analyzing wine quality and developing predictive models.

The full wine dataset contains two sub-datasets - red wine and white wine. The white wine dataset includes 4,898 complete records, whereas the red wine dataset comprises 1,599 fully observed entries, both exhibiting no missing data points. Each data set is characterized by 11 physicochemical features and a quality score from 0 (very bad) to 10 (excellent). The features included are `fixed acidity`, `volatile acidity`, `citric acid`, `residual sugar`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `density`, `pH`, `sulphates`, `alcohol`, and `quality`. Both datasets share identical data types. All features are of type `float64`, with the exception of `quality` which is of type `int64`. The red wine dataset contains 240 duplicated rows, while the white wine dataset has 937 duplicated rows. The duplicated rows are considered independent sample collection points and will remain in the analysis. Summary statistics were generated to further examine the distribution and patterns in both datasets (Table 1).

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

Table 1. Data Summary Statistics per Wine Type

Attribute(Units)	Red Wine				White Wine			
	Min	Max	Mean	Std	Min	Max	Mean	Std
fixed acidity (g/dm^3)	4.60	15.90	8.32	1.74	3.80	14.20	6.85	0.84
volatile acidity (g/dm^3)	0.12	1.58	0.53	0.18	0.08	1.10	0.28	0.10
citric acidity (g/dm^3)	0.00	1.00	0.27	0.20	0.00	1.66	0.33	0.12
residual sugar (g/dm^3)	0.90	15.50	2.54	1.41	0.60	65.80	6.39	5.07
chlorides (g/dm^3)	0.01	0.61	0.09	0.05	0.01	0.35	0.05	0.02
free sulfur dioxide (mg/dm^3)	1.00	72.00	15.87	10.46	2.00	289.00	35.31	17.01
total sulfur dioxide (mg/dm^3)	6.00	289.00	46.47	32.90	9.00	440.00	138.36	42.50
density (g/cm^3)	0.99	1.00	1.00	0.001	0.99	1.04	0.99	0.003
pH	2.74	4.01	3.31	0.15	2.72	3.82	3.19	0.15
sulphates (g/dm^3)	0.33	2.00	0.66	0.17	0.22	1.08	0.49	0.11
alcohol (vol.%)	8.40	14.90	10.42	1.07	8.00	14.20	10.51	1.23
quality	3.00	8.00	5.64	0.81	3.00	9.00	5.88	0.89

Through EDA, we observe varying means, minimums, and maximums across both the red and white datasets. The mean values for `fixed acidity`, `volatile acidity`, and `citric acid` are comparable between the two wine types, showing similarities in physicochemical properties. However, there is a notable difference in `residual sugar`, with white wine having a mean of 6.39 g/dm^3 and red wine at 2.54 g/dm^3, which may be influenced by production methods (I.e., fermentation processes or sugar content in grape varieties) or product specifications.[3] We also observe a broad range in `fixed acidity` for both wine types, with red wine ranging from a minimum of 4.6 g/dm^3 to a maximum of 15.9 g/dm^3, and white wine ranging from 3.8 g/dm^3 to 14.2 g/dm^3. This may suggest that this feature could influence taste and the overall quality, as acidity plays a crucial role in the flavour profile. The spread of `free sulfur dioxide` shows a minimum of 1 mg/dm^3, maximum of 72 mg/dm^3, and a mean of 15.9 mg/dm^3 for red wine. For white wine, the range is broader, with a minimum of 2 mg/dm^3, maximum of 289 mg/dm^3, and a mean of 35.3 mg/dm^3. The values for `total sulfur dioxide` are relatively similar to `free sulfur dioxide`. White wine has a minimum of 9 mg/dm^3, a maximum of 440 mg/dm^3, and a mean of approximately 138.4 mg/dm^3, while red wine shows a minimum of 6 mg/dm^3, a maximum of 289 mg/dm^3 and a mean of 46.5 mg/dm^3. The marked difference between the two types may indicate differences in preservation techniques. For features `density`, `pH`, `sulphate`, and `alcohol`, the minimum, maximum, and mean values are similar between the two types.

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

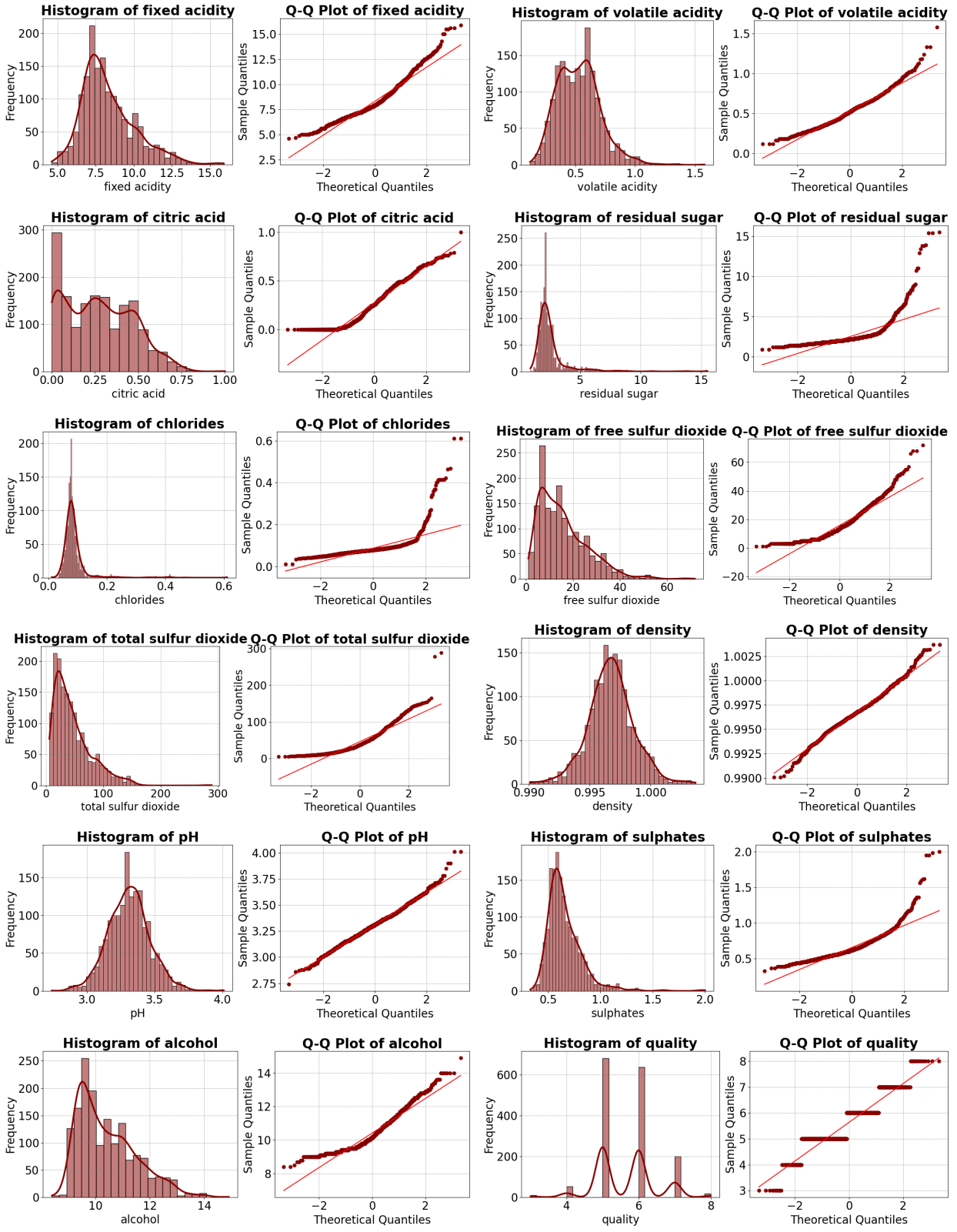

Figure 1. Red Wine Feature Distributions and Q-Q Plots

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

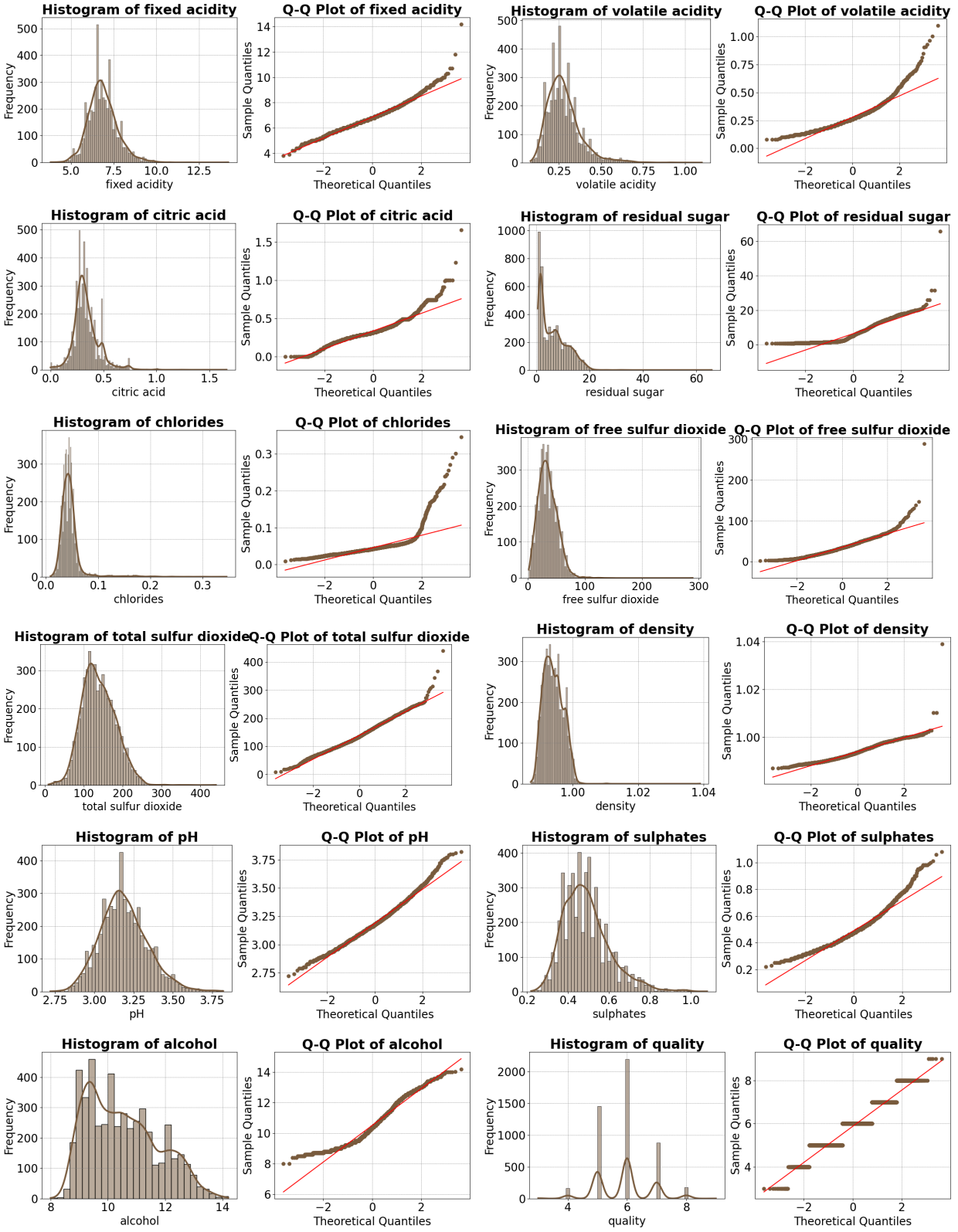

Figure 2. White Wine Feature Distributions and Q-Q Plots

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

Preliminary red and white wine feature distributions were briefly explored in the proposal, but are expanded to include clearer distribution patterns alongside Q-Q Plots (Figure 1, Figure 2). For red wine (Figure 1), `fixed acidity`, `volatile acidity`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `sulphates` and `alcohol` seem to have distributions that are more or less normal, as the Q-Q plots for these features show deviations. `Citric acid`, however, displays a flatter distribution with several modes, suggesting a multimodal distribution, and its Q-Q plot indicates deviations from normality, particularly in the tails. `Residual sugar` also shows some deviation from normality, as indicated by its Q-Q plot. The `density` and `pH` values are the features that appear to be most normally distributed, with their Q-Q plots showing a straight alignment along the reference line.

For white wine (Figure 2), while some features like `fixed acidity`, `chlorides`,` free sulfur dioxide`, `pH`, and `sulphates` are approximately normally distributed, others like `volatile acidity`, `citric acid`, `residual sugar`, `total sulfur dioxide`, `density`, and `alcohol`show deviations from normality. Features such as `citric acid` and `volatile acidity` display multiple peaks in their histograms, which implies that the data consists of several subgroups or clusters. This kind of distribution is often indicative of mixed populations or varying conditions within the dataset. We see a widespread distribution of data points for the `alcohol` feature within the white dataset, whereas we see a peak at ~9, with a gradual decline as the `alcohol` content increases for red (Figure 1, 2).

While both red and white wines share fundamental chemical attributes, key differences arise due to distinct production methods and sensory goals. Acidity profiles differ markedly: red wines typically exhibit higher `fixed acidity` and lower `volatile acidity`, aligning with longer maceration times and tannin extraction. `Residual sugar` is substantially lower in red wines, reflecting their dominance in dry styles, whereas white wines show greater variability and extreme sweetness potential. `Sulfur dioxide` usage is significantly higher in white wines, likely due to lighter pigmentation's increased susceptibility to oxidation. `Density` is marginally lower in white wines, consistent with lower `residual sugar` and `alcohol` content.

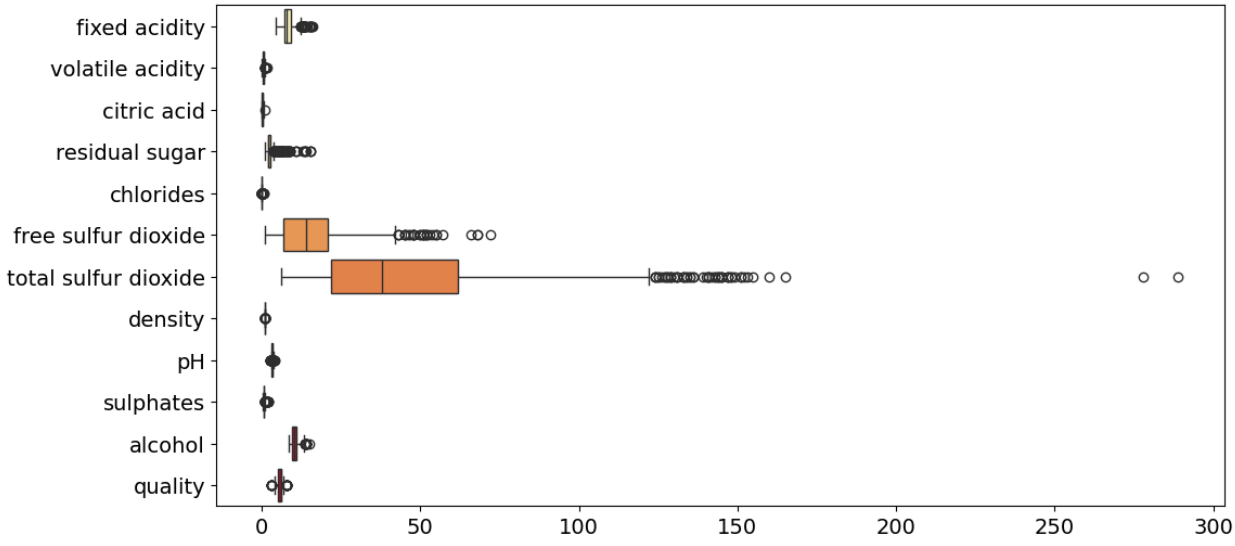

Figure 3. Red Wine Box Plot Analysis

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

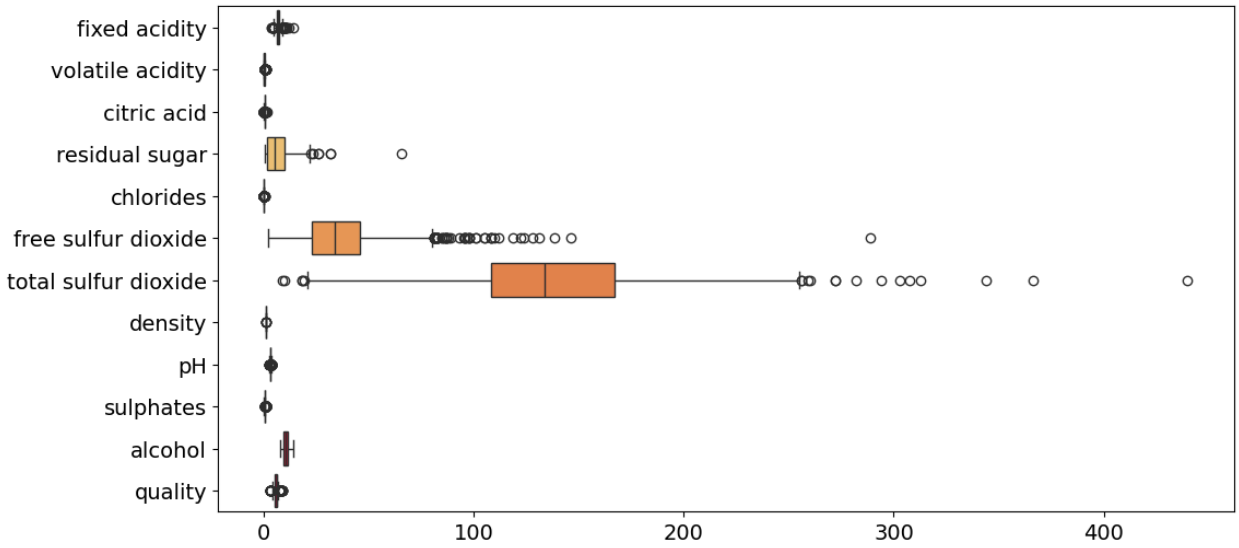

Figure 4. White Wine Box Plot Analysis

Figure 3 and 4 systematically visualize the distributional characteristics and central tendencies of 1 key chemical properties and quality metrics in red and white wine data. For red wine (Figure 3), features like `fixed acidity`, `chlorides`, `sulphates` and `alcohol` show stable medians, `volatile acidity` and `citric acid` display low spread, `pH` and `density` approximate normality, while `residual sugar` and has extreme outliers. `Free sulfur dioxide` and `total sulfur dioxide` display high variability and extreme outliers. For white wine (Figure 4), most features share similar distribution as red wine data, except that `residual sugar` has wider spread and `total sulfur dioxide` has less outliers.

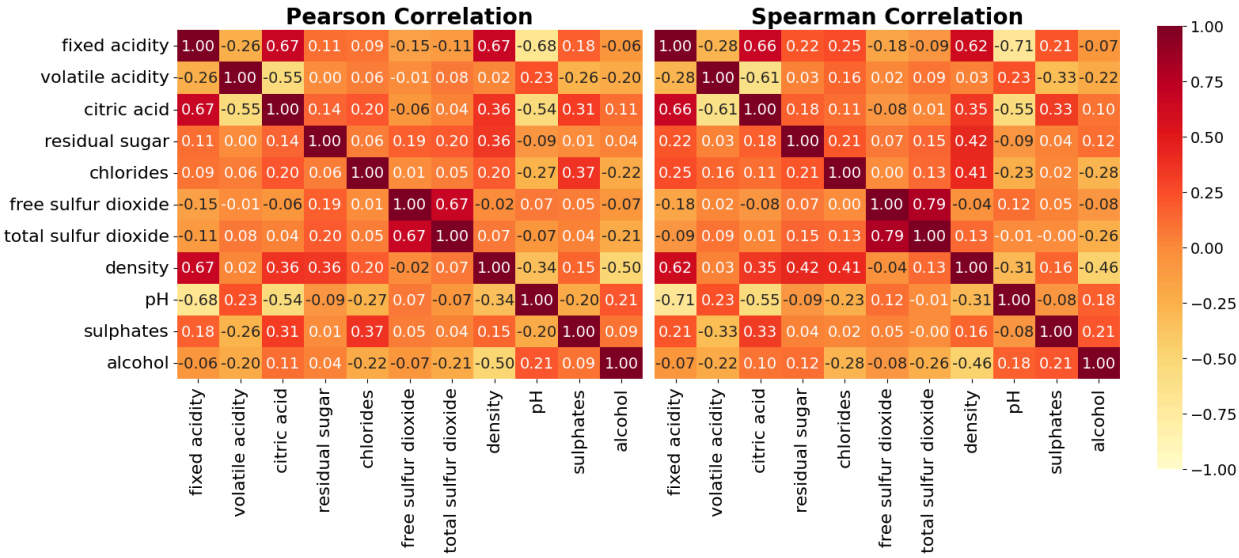

Figure 5. Pearson (Left) and Spearman (Right) Correlation Analysis of Red Wine Features

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

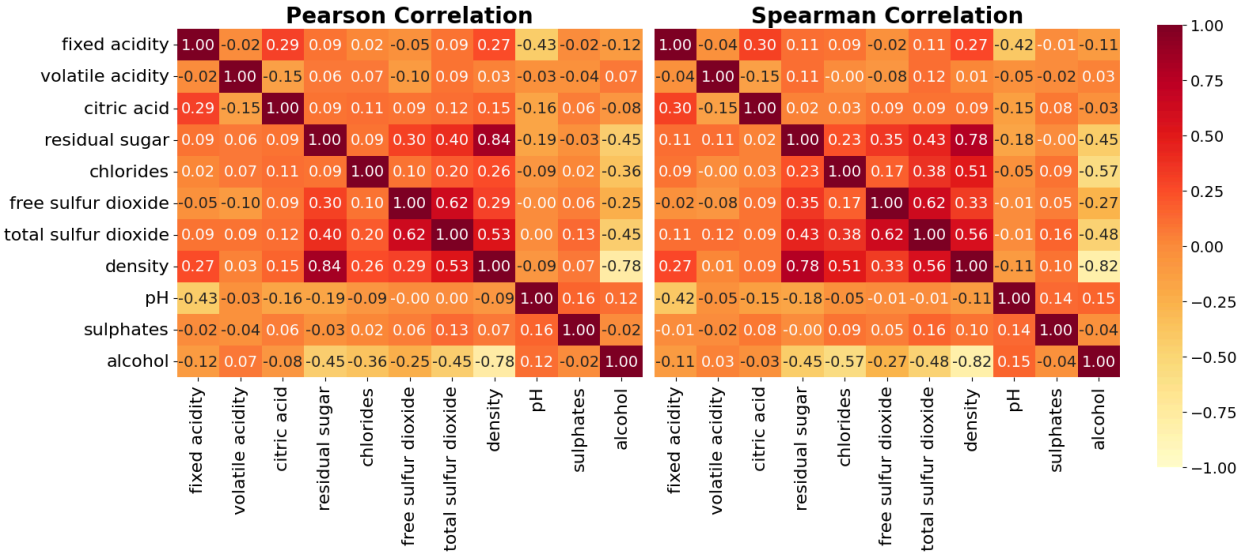

Figure 6. Pearson (Left) and Spearman (Right) Correlation Analysis of White Wine Features

Figure 5 and 6 present the correlation matrices of red and white wine features using both Pearson and Spearman correlation coefficients. Pearson correlation measures the linear relationship between variables, whereas Spearman correlation captures monotonic relationships, making it more robust to non-linear associations.

For red wine (Figure 5), from the Pearson correlation matrix (left), notable relationships include a strong positive correlation between `fixed acidity` and `citric acid` (~0.67) and between `free sulfur dioxide` and `total sulfur dioxide` (~0.67). Conversely, `pH` and `fixed acidity` show a strong negative correlation (~-0.68), suggesting that as `fixed acidity` increases, `pH` tends to decrease. `Density` also exhibits a moderate positive correlation with `fixed acidity` (~0.67). The Spearman correlation matrix (right) similarly reflects these trends, though some correlations appear stronger, such as the relationship between `free sulfur dioxide` and `total sulfur dioxide` (~0.79). Spearman correlations also highlight some non-linear relationships that may be less apparent in the Pearson analysis, particularly for variables like `pH` and `fixed acidity`, where a more pronounced negative correlation is observed.

For white wine (Figure 6), from the Pearson correlation matrix (left), notable relationships include a strong positive correlation between `density` and `residual sugar` (~0.84) and between `free sulfur dioxide` and `total sulfur dioxide` (~0.62). Conversely, `density` and `alcohol` show a strong negative correlation (~-0.78), suggesting that as `density` increases, `alcohol` tends to decrease. Spearman correlations also highlight some non-linear relationships that may be less apparent in the Pearson analysis, particularly for variables like `density` and `chlorides`, where a more pronounced positive correlation is observed.

Section II: Scientific Questions

Through our analysis, our group seeks to **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.

We also aim to isolate significant **differences in the physicochemical properties** between white and red wines to evaluate how 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.

Lastly, we are interested in whether certain physicochemical features are **more strongly correlated** with wine quality, and if they contribute to the perceived quality of wine. We seek to identify the most influential features and provide insight into improving wine quality.

Section III: Analysis Techniques for Prediction

Our group aims to use following statistical analysis techniques to answer our research questions:

1. Kernel Density Estimation (KDE) & Histograms
    - Use of KDE and histograms will help visualize the distribution of wine quality and each physicochemical property for both red and white wines. This will provide insights into whether these distributions are unimodal or multimodal and how different features vary between wine types.
2. Descriptive Statistics & Hypothesis Testing (t-tests or Mann-Whitney U test)
    - Through comparing the means and variances of physicochemical properties between red and white wines, this will help identify significant differences in composition. If the data is normally distributed, a t-test will be used; otherwise, a Mann-Whitney U test will be applied.
3. Correlation Analysis (Pearson, Spearman)
    - To determine which physicochemical features are most strongly correlated with wine quality, we will compute both Pearson and Spearman correlation coefficients.
4. Generalized Linear Models (GLMs)
    - Use of GLMs will allow us to model wine quality as a function of physicochemical features while handling different types of relationships.
5. Generalized Additive Models (GAMs) & Splines
    - GAMs extend GLMs by allowing nonlinear relationships between physicochemical properties and wine quality. This is useful since features like acidity or sulfur dioxide may have threshold effects. Splines can capture smooth variations in these relationships.
6. Gaussian Mixture Models (GMMs)

- GMMs can help model the underlying distribution of wine quality and detect potential clusters of high- and low-quality wines. This is particularly useful if wine quality follows a multimodal distribution.

7. Principal Component Analysis (PCA) & Feature Importance via Random Forests
   - PCA will be used to reduce dimensionality and identify the most important physicochemical features. Additionally, Random Forest feature importance analysis will help determine which factors contribute most to wine quality.
8. Predictive Modeling (Random Forest)
   - To build a predictive model for wine quality, we will test machine learning models like Random Forest. The model will be evaluated using cross-validation to ensure robustness.

Appendix

Extracts of both datasets being used for this analysis can be found below in Figure 5 and Figure 6, but full datasets can be accessed via: https://archive.ics.uci.edu/dataset/186/wine+quality[1]

	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

Figure 7. 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

Figure 8. White Wine Data Extract

Group 4
Huan He | Qijia Zheng | Kelsey Strachan

References

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine quality," *UCI Machine Learning Repository*, 2009. [Online]. Available: https://doi.org/10.24432/C56S3T.
2. P. Cortez, F. Almeida, and T. Matos, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547-553, 2009.
3. A. Luque, M. Mazzoleni, F. Zamora-Polo, A. Ferramosca, J. R. Lama, and F. Previdi, "Determining the importance of physicochemical properties in the perceived quality of wines," Dept. de Ingeniería del Diseño, Escuela Politécnica Superior, Universidad de Sevilla, Seville, Spain, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10287348