Section 2: Exploratory Data Analysis and Data Preparation

Dataset Overview and Initial Analysis

We begin our analysis with a dataset containing 1,599 wine samples and 11 physicochemical features, along with a target variable representing wine quality. The features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. An initial analysis revealed that the dataset contains no missing values, indicating a complete dataset ready for further exploration.

The distribution of feature values is summarized in Table A1 in the appendix, which provides detailed statistical measures for all features. Additional visualizations showing the distribution of each feature can be found in the appendix (Figure A2), allowing us to understand symmetry, potential outliers, and other characteristics that complement the statistical summary.

Target Variable Analysis and Feature Correlations

Wine quality serves as the target variable, taking discrete values from 3 to 8. Understanding the distribution of wine quality is essential for addressing the challenges inherent in classification tasks, particularly given the class imbalance observed in the dataset. Additionally, analyzing feature relationships is critical for identifying multicollinearity and understanding the underlying data structure that may influence model performance.

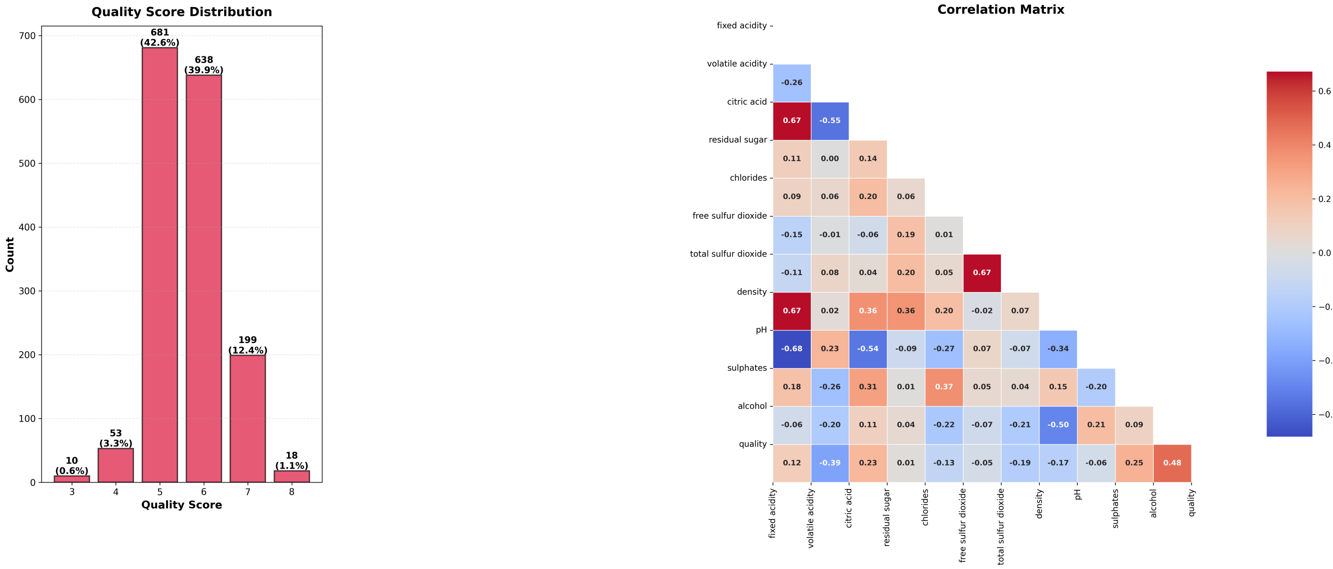


Figure 1: Quality Distribution and Correlation Matrix

Figure 1: Left: Quality Score Distribution (bar chart); Right: Correlation Matrix showing pairwise relationships between all features

The bar chart (left) reveals that this variable exhibits class imbalance, with quality levels 5 and 6 being the most frequent, respectively, while there are very few wines with low quality (such as 3) or extremely high quality (such as 8). The order of magnitude of the differences is very significant, as in our dataset there is a probability of 0.63% of having a wine with quality 3, while the probability is more than 68 times greater for quality 5 (42.59%). Similarly, quality 6 has a probability of 39.90%, which is more than 63 times greater than quality 3.[1]

The correlation matrix (right) provides a comprehensive view of pairwise linear relationships between all physicochemical features, enabling us to assess both feature-feature interactions and feature-target associations simultaneously. Notable pairwise correlations between features include strong positive correlations between fixed acidity and citric

[1] As a team, we read the codebook and documentation of the data (Cortez et al., 2009; UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/wine+quality) and although in the description of the dictionaries the source states that quality can take values from 0 to 10, in our dataset we only found possible values from 3 to 8.

acid (0.67), fixed acidity and density (0.67), and total sulfur dioxide and free sulfur dioxide (0.67). Strong negative correlations are observed between fixed acidity and pH (-0.68), citric acid and volatile acidity (-0.55), and citric acid and pH (-0.54). These relationships indicate potential multicollinearity that may influence model performance. By examining the quality row in the correlation matrix, we identify the strongest associations with wine quality. Alcohol content shows the strongest positive correlation (0.48), followed by sulphates (0.25) and citric acid (0.23). Conversely, volatile acidity exhibits the strongest negative correlation (-0.39), indicating that higher levels are associated with lower quality scores. Other features showing moderate negative correlations include total sulfur dioxide (-0.19) and density (-0.17). While quality is an ordinal variable, we compute Pearson correlation coefficients given its natural ordering, which allows us to capture linear trends. These correlations provide an initial indication of feature importance, though the classification models will ultimately determine the true discriminative power of each feature.

Feature Engineering and Normalization

Prior to model training, it is essential to prepare the data appropriately. Machine learning algorithms, particularly Support Vector Machines (SVM) and Artificial Neural Networks (ANN), are sensitive to the scale of input features. Features with larger numerical ranges can dominate the learning process, potentially biasing the model toward those features. Therefore, we apply feature normalization to ensure all features contribute equally to the model's learning process.

Our data preparation pipeline consists of two main steps: (1) outlier detection and treatment, and (2) feature standardization. Initial analysis revealed the presence of significant outliers across multiple features, particularly in residual sugar, chlorides, and sulphates. Outliers can significantly affect the mean and standard deviation used in standardization, potentially distorting the transformation. We identify outliers using the Interquartile Range (IQR) method, which is robust to extreme values. Outliers are capped (rather than removed) to preserve the sample size, ensuring we retain all 1,599 samples for model training. Subsequently, we apply standardization (Z-score normalization), which transforms features to have zero mean and unit variance according to the formula:

$$z = \frac{x - \mu}{\sigma}$$

where z is the standardized value, x is the original value, μ is the mean, and σ is the standard deviation of the feature.

The normalized dataset[2] contains all original features standardized to have zero mean and unit variance, with outliers appropriately treated. This dataset will serve as the baseline for all classification models (SVM, ANN, and Random Forest), ensuring that feature scaling does not bias the learning process and enabling fair comparison across different algorithms.

Feature Interactions

The correlation analysis revealed several strong pairwise relationships between features ($|r| > 0.5$), suggesting that these variables may interact in ways that influence wine quality. While linear models can capture individual feature effects, they may miss how the combination of features together affects the outcome. For example, the effect of fixed acidity on wine quality may depend on the level of citric acid present. To address this, we create multiplicative interaction features for the strongly correlated pairs identified in the correlation matrix. Specifically, we create interaction terms for: fixed acidity × citric acid (0.67), fixed acidity × density (0.67), total sulfur dioxide × free sulfur dioxide (0.67), fixed acidity × pH (-0.68), citric acid × volatile acidity (-0.55), citric acid × pH (-0.54), and alcohol × density (-0.50). The dataset with interactions[3] contains all original standardized features plus these seven interaction terms, expanding the feature space from 11 to 18 features. This dataset allows us to evaluate whether explicitly modeling feature interactions improves classification performance compared to the baseline normalized dataset.

[2] This dataset is available in our GitHub repository (https://github.com/modie25/ml_f25_project) for replication purposes.

[3] This dataset is also available in our GitHub repository (https://github.com/modie25/ml_f25_project).

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables called principal components. These components are ordered by the amount of variance they explain in the data. PCA serves two purposes in our analysis: (1) visualization of high-dimensional data in lower-dimensional space (2D/3D), and (2) as an alternative dataset for model training, allowing us to compare classification performance with and without dimensionality reduction.

We apply PCA to the normalized dataset to ensure that all features contribute equally to the principal components. The scree plot and cumulative explained variance plot (see Figure A3 in the appendix) reveal how many components are needed to capture the majority of the variance in the data. The analysis shows that 7 components are required to retain 90% of the variance, while 9 components capture 95% of the total variance. This suggests that dimensionality reduction is feasible, as most information can be preserved with fewer components than the original 11 features. The visualization in 2D and 3D space helps us understand the separability of different quality classes in the reduced-dimensional space.

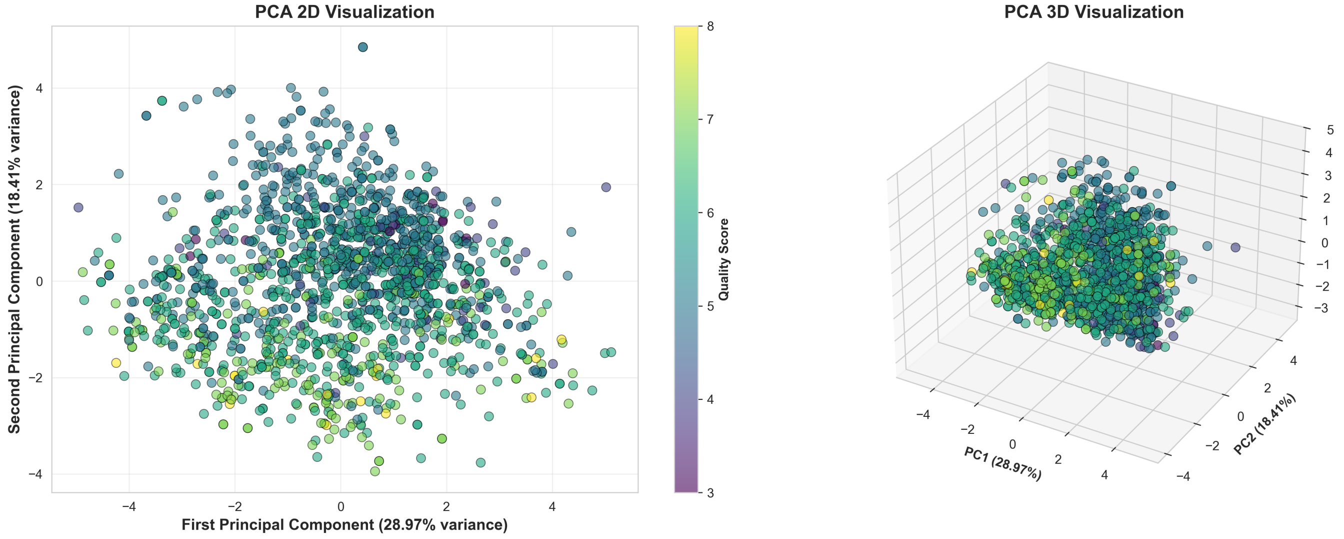


Figure 2: PCA Visualization - Left: 2D projection showing first two principal components; Right: 3D projection showing first three principal components

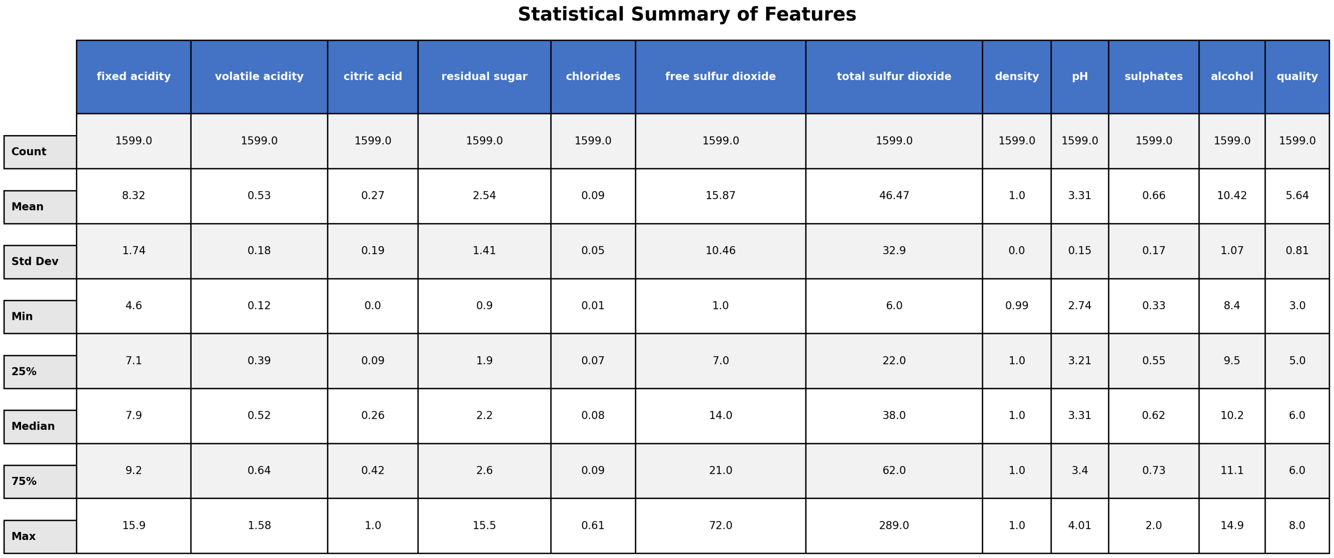
The 2D and 3D projections show that different quality classes are largely overlapping, with no clear linear separation between quality levels. This indicates that the first few principal components capture variance related to physico-chemical properties but do not directly correspond to quality discrimination, suggesting that non-linear classification methods may be necessary to effectively distinguish between quality classes. The PCA-transformed dataset[4], created using principal components that retain 95% of the total variance, will be used as an alternative input for classification models. This allows us to compare performance with the normalized dataset and assess whether dimensionality reduction improves or hinders classification accuracy.

[4] This dataset is also available in our GitHub repository (https://github.com/modie25/ml_f25_project).

Appendix: Additional Visualizations

A1. Statistical Summary Table

The following table provides comprehensive descriptive statistics for all features in the dataset, including count, mean, standard deviation, minimum, maximum, and quartiles.



Statistical Summary of Features

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
Count	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0
Mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.0	3.31	0.66	10.42	5.64
Std Dev	1.74	0.18	0.19	1.41	0.05	10.46	32.9	0.0	0.15	0.17	1.07	0.81
Min	4.6	0.12	0.0	0.9	0.01	1.0	6.0	0.99	2.74	0.33	8.4	3.0
25%	7.1	0.39	0.09	1.9	0.07	7.0	22.0	1.0	3.21	0.55	9.5	5.0
Median	7.9	0.52	0.26	2.2	0.08	14.0	38.0	1.0	3.31	0.62	10.2	6.0
75%	9.2	0.64	0.42	2.6	0.09	21.0	62.0	1.0	3.4	0.73	11.1	6.0
Max	15.9	1.58	1.0	15.5	0.61	72.0	289.0	1.0	4.01	2.0	14.9	8.0

Figure 3: Statistical Summary of Features

Table A1: Statistical Summary of Features

A2. Feature Distributions

The following visualization provides a comprehensive view of the distribution of each feature, including symmetry, potential outliers, and other characteristics that complement the statistical summary table.

Figure A2: Distribution of all features showing histograms for each physicochemical property

A3. PCA Variance Analysis

The scree plot and cumulative explained variance plot provide detailed information about the variance explained by each principal component and the number of components needed to retain a specified percentage of the total variance.

Figure A3: PCA Variance Analysis - Left: Scree plot showing explained variance per component; Right: Cumulative explained variance

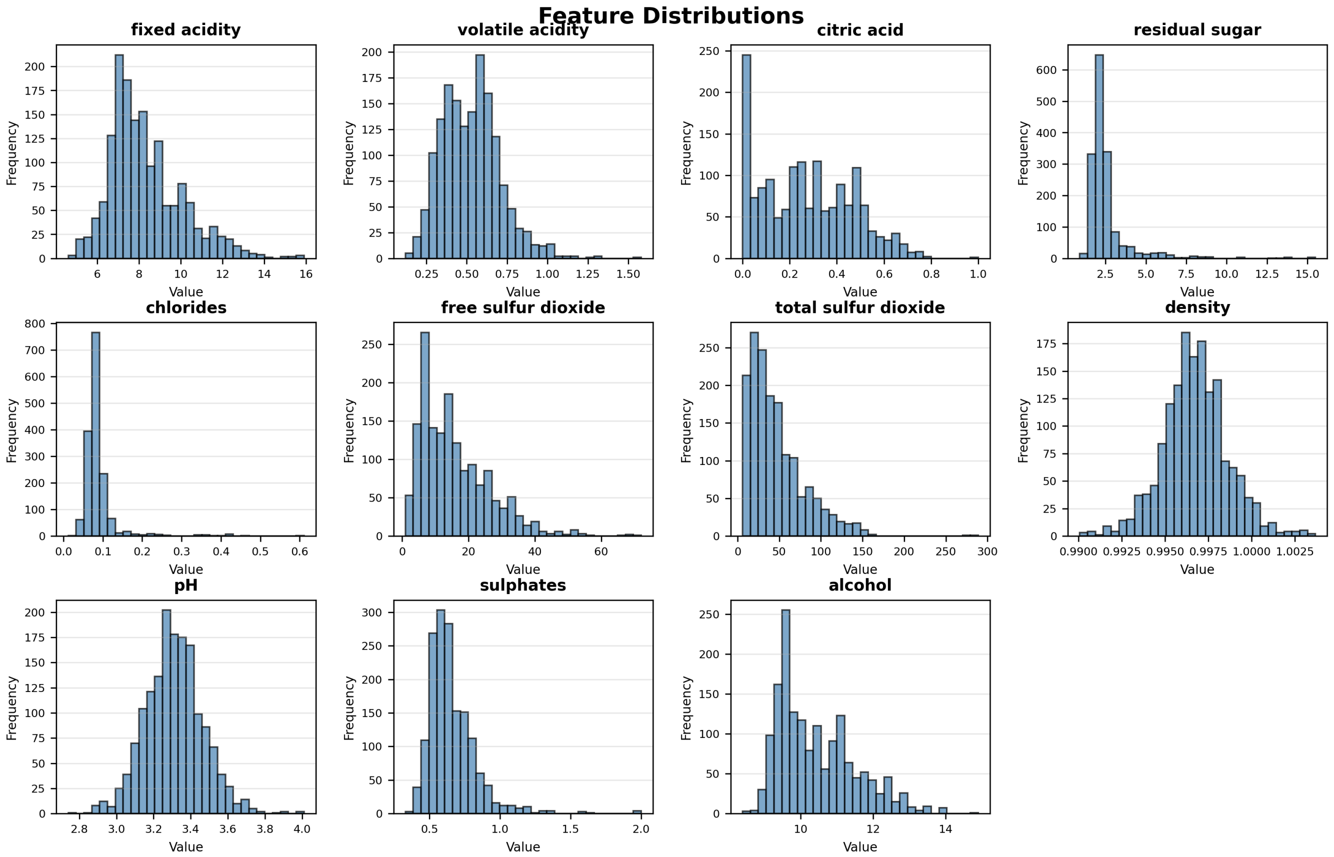


Figure 4: Feature Distributions

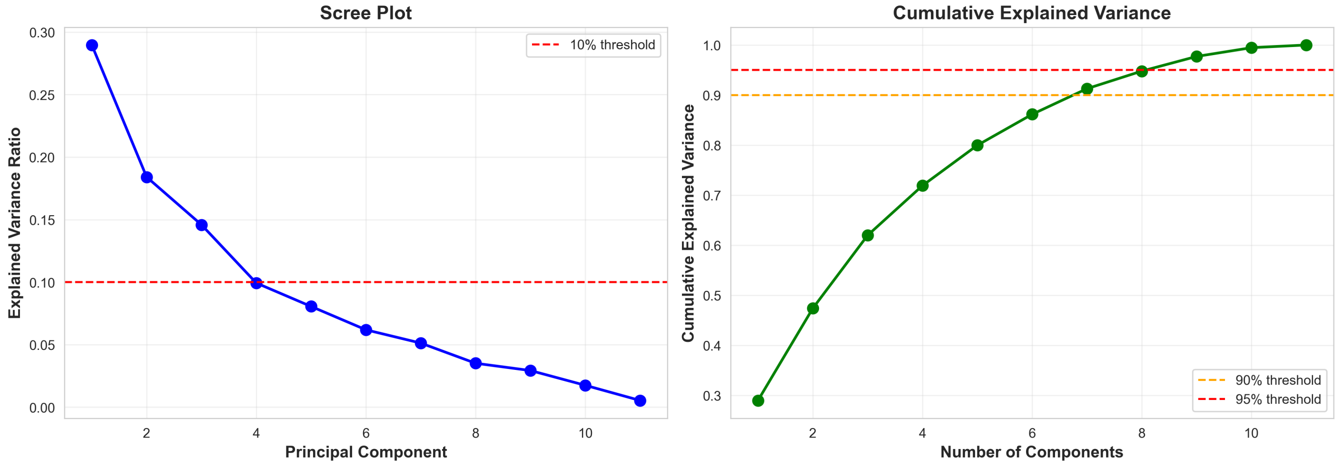


Figure 5: PCA Variance Analysis