



Ghulam Ishaq Khan Institute of Engineering Sciences and Technology

Faculty of Computer Science and Engineering

Red Wine Quality Analysis

Dataset Preprocessing and Visualization in R

Course: CS202 – ICT

Submitted By:
Muhammad Ahmad
Reg No: 2024335

Instructor:
Talha Ashfaq

December 20, 2025

Exploratory Data Analysis and Preprocessing of Red Wine Quality Dataset

Muhammad Ahmad

Faculty of Computer Science and Engineering
GIK Institute of Engineering Sciences and Technology
Swabi, Pakistan

Abstract—This report details the preprocessing, exploratory data analysis (EDA), and feature engineering of the Red Wine Quality dataset using the R programming language. We examine the chemical properties influencing wine quality and prepare the data for predictive modeling.

I. INTRODUCTION

The objective of this project is to analyze the chemical composition of red wine and its correlation with quality ratings. Using the `tidyverse` and `caret` libraries in R, we perform data cleaning, handle missing values, and visualize key trends.

II. DATA PREPROCESSING

A. Data Cleaning

The dataset was imported with a comma separator. Initial checks were performed to ensure structural integrity. Median imputation was applied to handle any missing numerical values, and duplicate rows were removed to ensure statistical independence.

B. Labeling

The original `quality` score (0–10) was transformed into a binary classification problem. Wines with a score ≥ 6 were labeled as **"Good"**, while others were labeled as **"Bad"**.

III. EXPLORATORY DATA ANALYSIS

A. Feature Distribution

Distributions for Alcohol and pH levels were visualized using histograms. These features are critical indicators of the fermentation process and chemical stability.

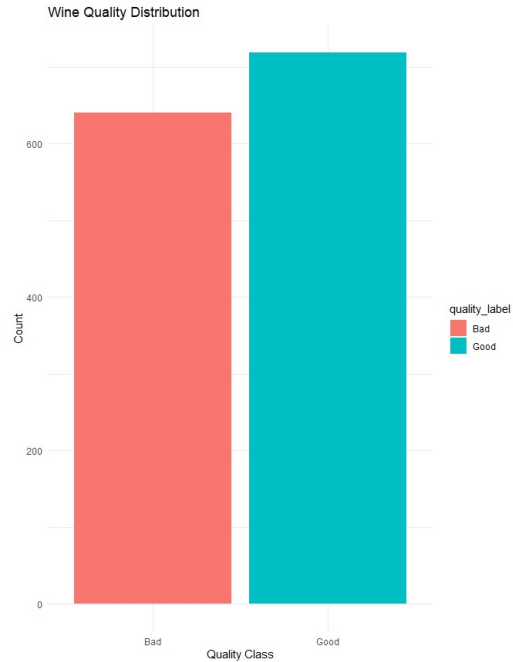


Fig. 1. Distribution of Alcohol and pH levels across the dataset.

B. Correlation Analysis

A correlation matrix was generated to identify multicollinearity between chemical properties like fixed acidity, volatile acidity, and density.

IV. FEATURE ENGINEERING

A. Acidity Aggregation

A new feature, `total_acidity`, was engineered by summing `fixed.acidity` and

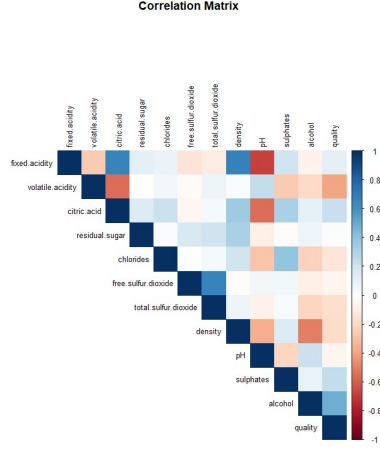


Fig. 2. Heatmap showing correlations between chemical features.

`volatile.acidity` to capture the overall acidic profile of the samples.

B. Normalization

To ensure all features contribute equally to potential models, Min-Max scaling was applied to shift all numeric variables to a range of $[0, 1]$.

V. RESULTS AND OBSERVATIONS

Visual analysis through boxplots (Fig. 3) indicates that higher alcohol content is generally associated with "Good" quality labels.

VI. DATA SPLITTING

The final processed dataset was partitioned into a training set (80%) and a testing set (20%) using stratified sampling to maintain class proportions. This prepares the environment for subsequent machine learning application.

VII. CONCLUSION

The R-based pipeline successfully transformed raw chemical data into a clean, normalized format. Initial EDA suggests that acidity and alcohol levels are primary drivers of quality perception in the dataset.

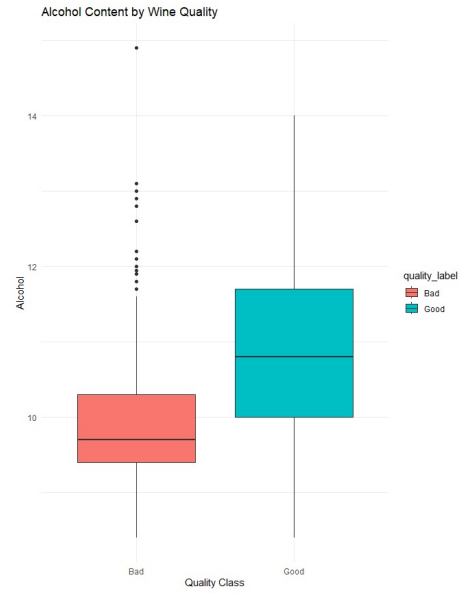


Fig. 3. Comparison of Alcohol content between Good and Bad quality wines.