

# Assignment 3: Exploratory Data Analysis Using Multivariate Visualization Techniques

## Introduction

For this assignment, I have chosen to explore a dataset featuring physicochemical tests of red Vinho Verde wine samples. I aimed to uncover patterns and correlations within the data that could potentially influence the quality of wine, focusing specifically on the variables of alcohol content, sulphates, and pH levels.

## Dataset Overview

The dataset I utilized comprises 1,599 red wine samples, each detailed by 11 features and 1 target variable, which is the quality of the wine. My analysis concentrated on the multivariate relationships between the chosen variables and how they may affect the quality.

## Analysis Methodology

I adopted several multivariate visualization techniques to delve into the data:

### 1. Correlation Heatmap

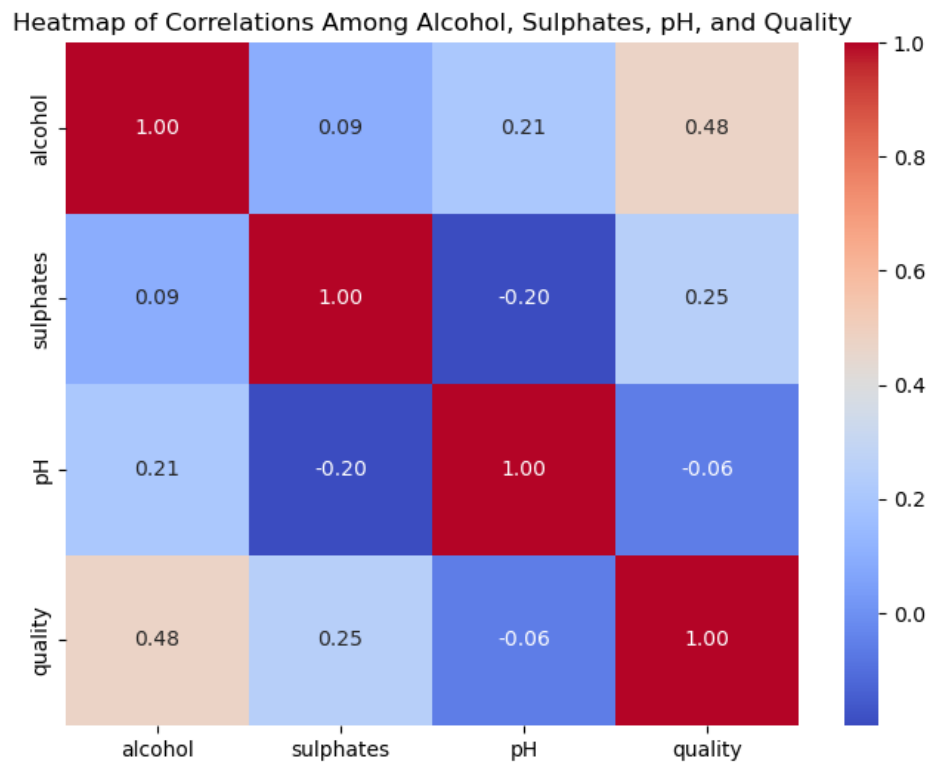
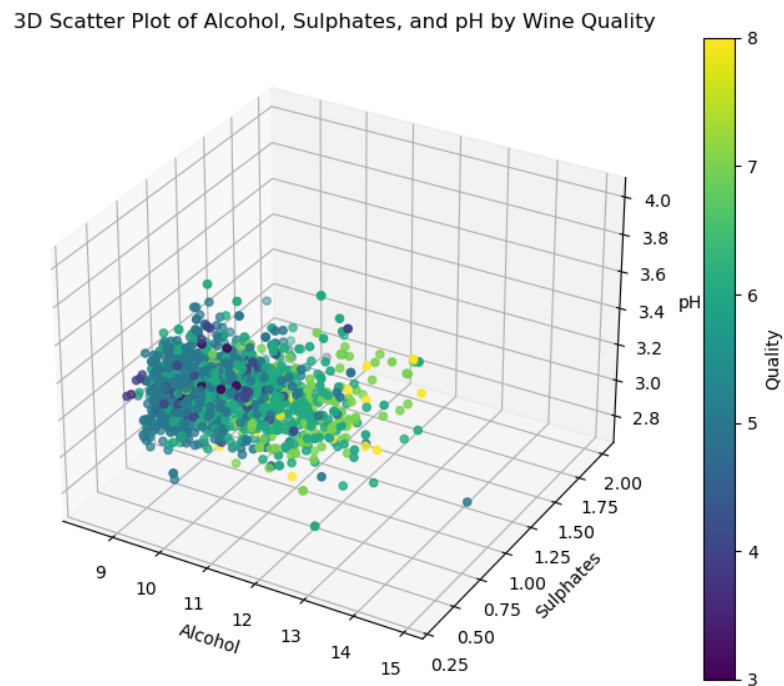


Figure 1: Heatmap

My first step was to create a heatmap of correlations to understand the linear relationships between the variables at a glance. I discovered a notably strong positive correlation between alcohol content and wine quality, indicating that wines with higher alcohol percentages tend to have higher quality ratings. Sulphates followed suit with a modest positive correlation, whereas pH levels showed a minimal negative correlation with quality, suggesting other factors may play more significant roles in determining wine quality.

## 2. 3D Scatter Plot



*Figure 2: 3D Scatterplot*

Next, I crafted a 3D scatter plot to visually interpret the complex interactions between alcohol content, sulphates, and pH, with the varying colors denoting the quality of wine. This visualization clearly showed that the best-rated wines often had higher levels of alcohol and sulphates, regardless of their pH level, thus reinforcing the initial insights from the correlation heatmap.

## 3. Trellis Plot

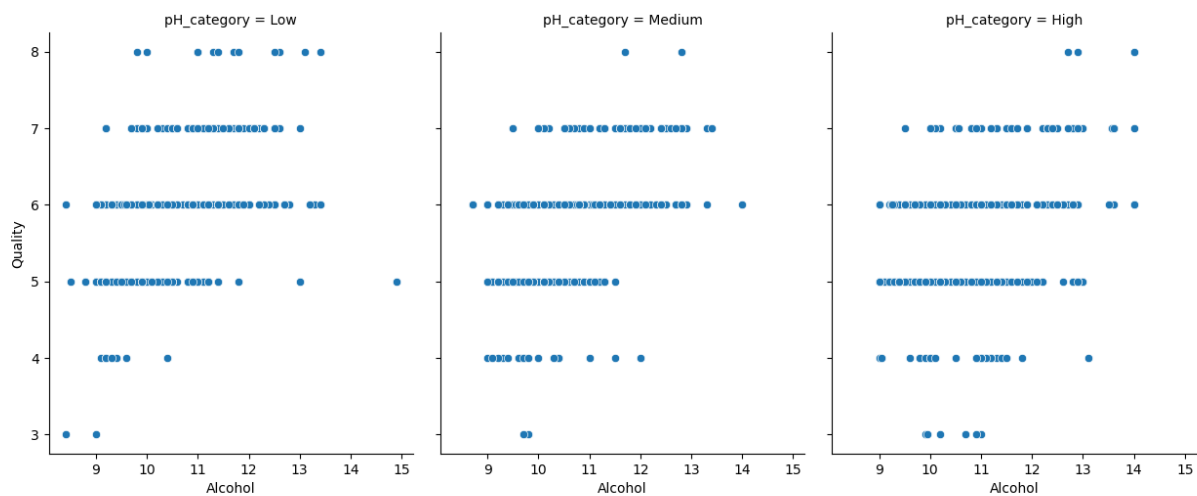


Figure 3: Trellis Plot

To further explore these relationships, I utilized a trellis plot to examine how the connection between alcohol and quality holds up across different pH levels. The pattern remained consistent—wines with higher alcohol content generally had higher quality scores. Notably, this trend was most distinct within the medium pH category, hinting at a possible interaction between alcohol levels and pH in determining wine quality.

## Conclusions

Through my exploratory data analysis, I found compelling evidence that alcohol content is a significant predictor of wine quality, with sulphates also contributing to the overall rating. Although pH levels did not exhibit a strong direct relationship with quality, their interaction with alcohol content seemed to influence quality outcomes. These findings suggest that understanding the balance of these variables could be crucial in optimizing wine production processes and enhancing quality control measures.

In conclusion, my analysis offers a data-driven perspective on the intricate dynamics between wine composition and quality, providing a foundation for further investigation into wine science and production techniques.

## Reference

Dataset retrieved from <https://archive.ics.uci.edu/dataset/186/wine+quality>