Wine Quality Dataset Exploratory Data Analysis

Prepared by: Bhavya Rajpal

Date: 8-10-23

Introduction

The Wine Quality dataset is a collection of white wine samples that includes a quality score and other chemical properties. In this report, we will conduct an exploratory data analysis (EDA) on the dataset to gain insights into the qualities of white wine and understand the variables that affect their quality.

Data Cleaning

Before proceeding with the analysis, it is essential to ensure the dataset's integrity and quality. This involves cleaning the data to remove any inconsistencies, missing values, or outliers that could skew our analysis.

Dataset Loading

We began by loading the dataset from the provided URL using Pandas, creating a Data Frame for further analysis. This step allows us to work with the dataset efficiently.

Finding Missing Values

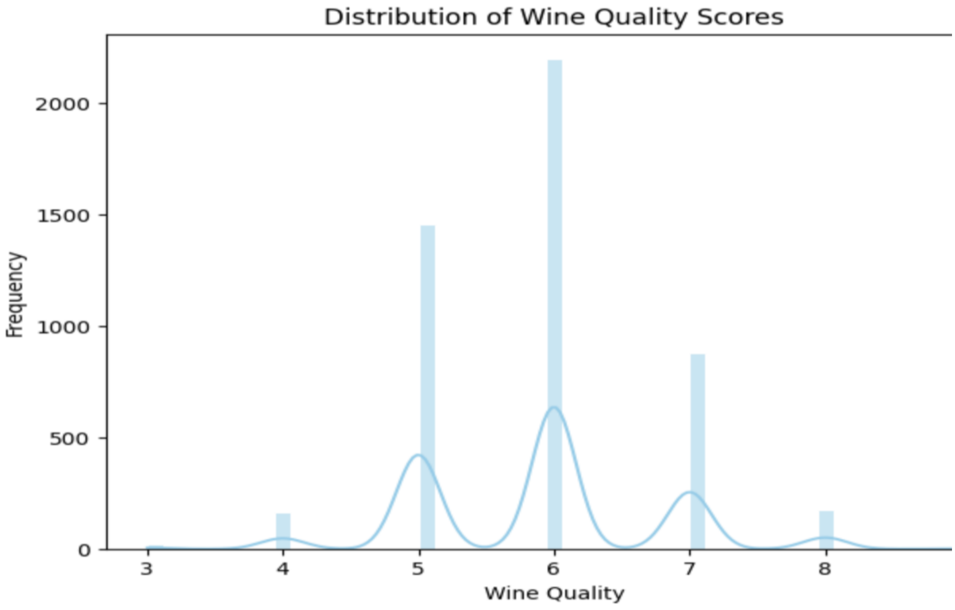
One critical aspect of data cleaning is identifying and handling missing values. Fortunately, in this dataset, there were no missing values, ensuring the data's completeness.

Analysing Exploratory Data

With the dataset cleaned and loaded, we can now delve into exploratory data analysis (EDA). This involves using visualizations and statistical analyses to uncover patterns, relationships, and insights within the data. Our analysis is organized into the following sections:

1. Distribution of Wine Quality

The first step in our analysis is to explore the distribution of wine quality scores. This information provides a fundamental understanding of the dataset.



Visualization: Histogram with KDE

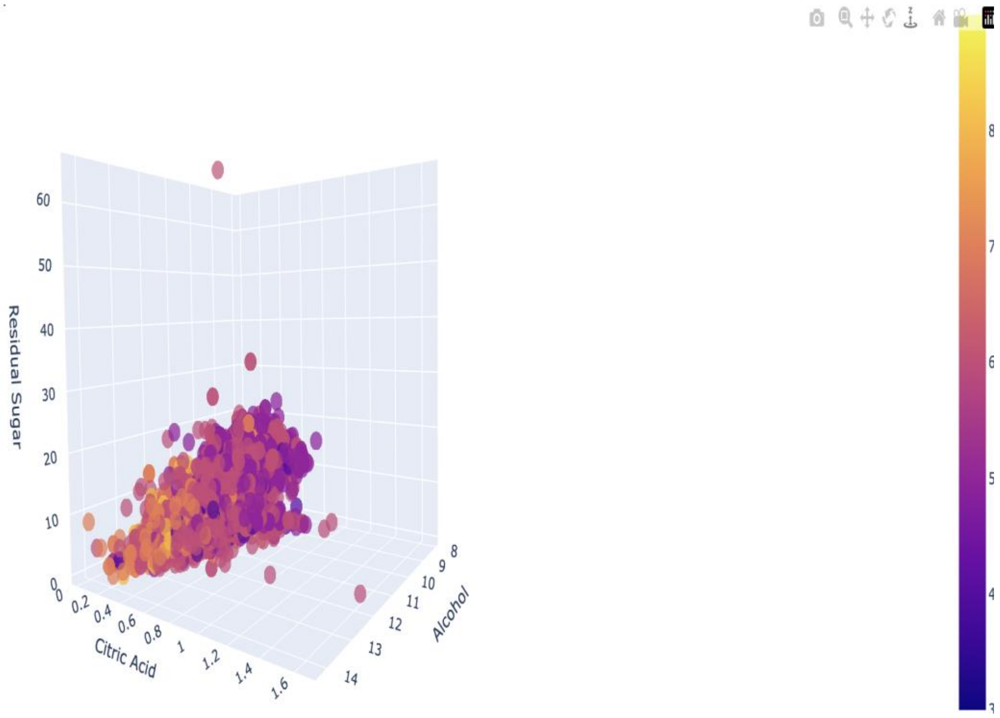
We used Seaborn to create a vibrant histogram with a Kernel Density Estimation (KDE) curve, visualizing the frequency distribution of wine quality scores. The graph shows that the majority of wines fall within the quality range of 5 to 7, with fewer wines rated extremely high (8 or 9) or low (3 or 4).

2. The Connection Between Characteristics and Wine Quality

Our next objective is to investigate how distinctive characteristics relate to wine quality. We employed a 3D scatter plot to visualize this relationship

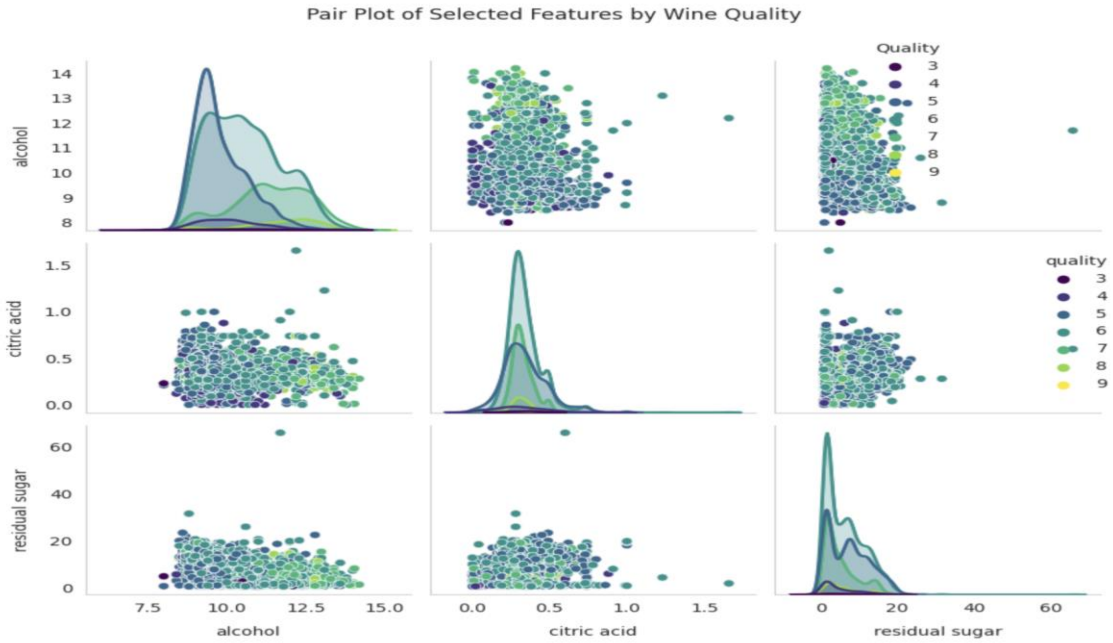
Visualization: 3D Scatter Plot

Using Plotly Express, we created a 3D scatter plot with alcohol, citric acid, and residual sugar as the axes. The color of each data point represents the wine's quality, ranging from 3 to 9. This plot allows us to visually examine the connections between these features and wine quality. We found that higher alcohol and citric acid levels are associated with better wine quality, while the relationship with residual sugar is less pronounced.



3. Selected Feature Pair Plot

To investigate correlations between specific variables and wine quality, we generated pair plots.

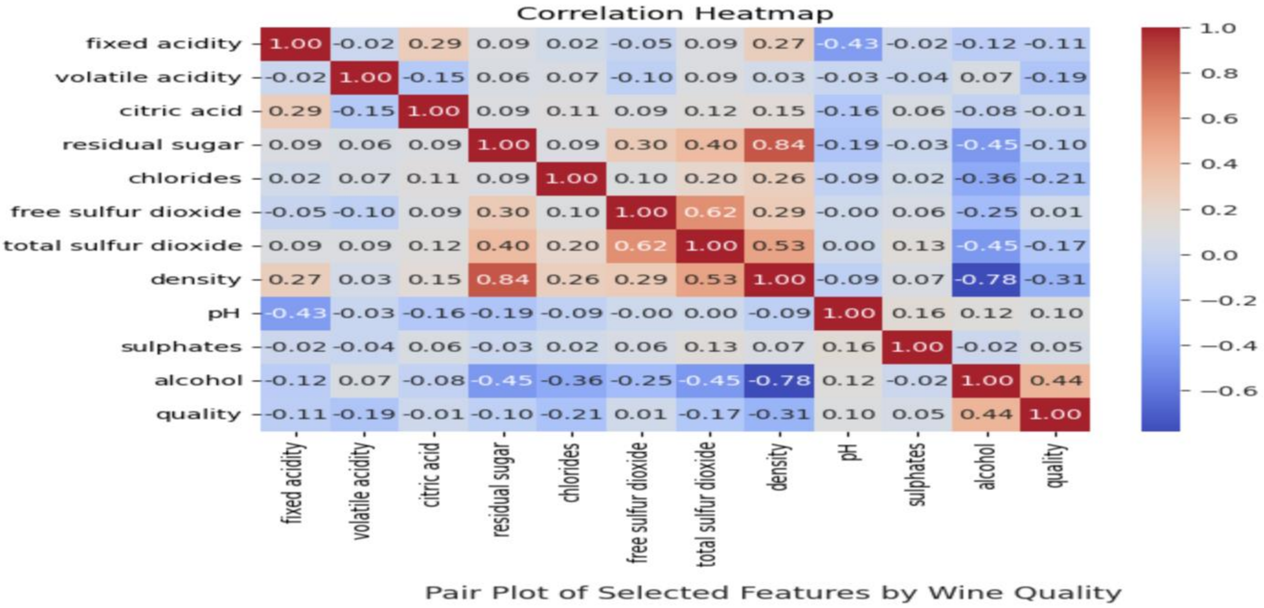


Visualization: Pair Plots

The pair plots provide a visual representation of pairwise feature distributions for various wine quality categories. By analysing these plots, we identified potential relationships and patterns. For instance, the pair plot of alcohol vs. citric acid revealed that wines with higher alcohol content tend to have elevated levels of citric acid, potentially contributing to their better quality.

4. Heatmap for Correlation

Understanding the linear relationships between distinctive characteristics is crucial. We constructed a correlation heatmap to visualize these relationships.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12	-0.11
volatile acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07	-0.19
citric acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08	-0.01
residual sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45	-0.10
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36	-0.21
free sulfur dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	-0.00	0.06	-0.25	0.01
total sulfur dioxide	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45	-0.17
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78	-0.31
pH	-0.43	-0.03	-0.16	-0.19	-0.09	-0.00	0.00	-0.09	1.00	0.16	0.12	0.10
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02	0.05
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00	0.44
quality	-0.11	-0.19	-0.01	-0.10	-0.21	0.01	-0.17	-0.31	0.10	0.05	0.44	1.00

Visualization: Correlation Heatmap

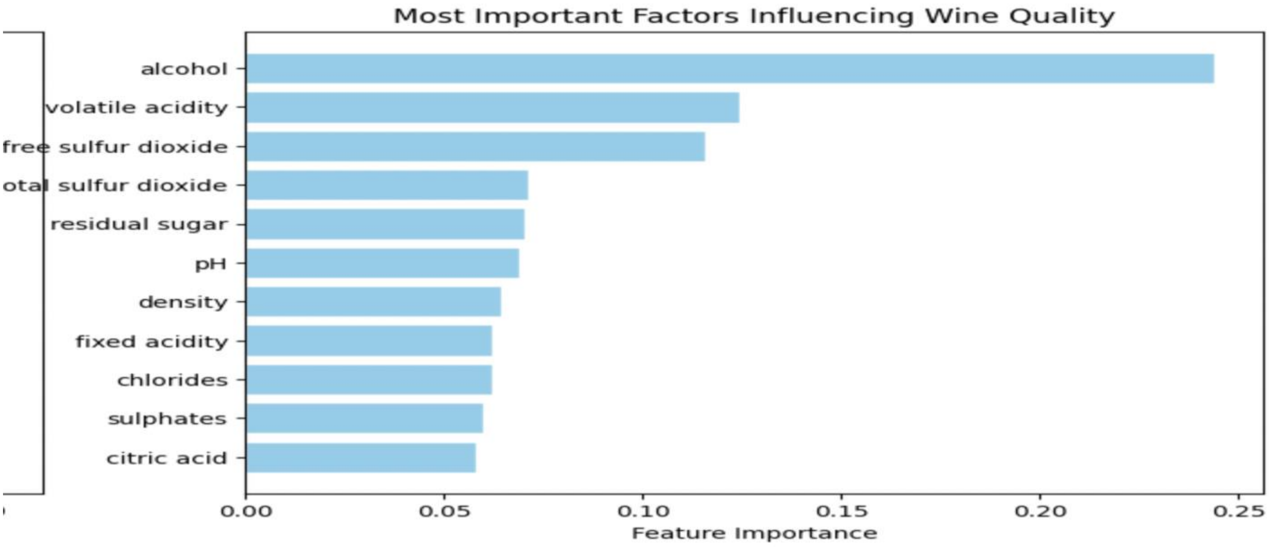
Seaborn's heatmap highlighted the strength and direction of correlations between features. Warm colors (red and orange) indicated positive correlations, while cool colors (blue) represented negative correlations. Notably, the heatmap highlighted a positive correlation between alcohol and quality, suggesting that wines with higher alcohol content tend to receive better quality ratings. Conversely, volatile acidity and quality exhibited a negative correlation, indicating that higher volatile acidity often resulted in lower quality ratings.

5. The Value of the Feature

To determine the most significant factors influencing wine quality, we employed a Random Forest Regressor model.

Feature Importance: Bar Graph

Using a Random Forest Regressor, we assessed the relative importance of each feature in predicting wine quality. The feature importance's were presented in a colourful bar graph, allowing us to identify the most critical factors. Alcohol emerged as the most influential feature, followed by volatile acidity and residual sugar.



Findings:

Our analysis revealed several significant findings:

Distribution of Wine Quality:

Many wines in the dataset fall within the quality range of 5 to 7, indicating a prevalence of moderate-quality wines.

Connection Between Characteristics and Wine Quality:

Higher levels of alcohol and citric acid are associated with better wine quality, while the relationship with residual sugar is less clear.

Selected Feature Pair Plot:

Pair plots suggest potential relationships between specific variables and wine quality, offering insights for further investigation.

Heatmap for Correlation:

The heatmap highlighted a positive correlation between alcohol and quality and a negative correlation between volatile acidity and quality.

The Value of the Feature:

Alcohol is the most influential feature in predicting wine quality, followed by volatile acidity and residual sugar.

Conclusion:

In conclusion, our exploratory data analysis of the Wine Quality dataset has provided valuable insights into the factors that influence wine quality. We have gained a better understanding of the distribution of wine quality scores, the relationship between characteristics and quality, and the most critical factors affecting wine quality. These findings serve as a foundation for further research and modelling in the domain of wine quality prediction and enhancement.

In Conclusion:

Our EDA journey through the Wine Quality dataset has been an incredible learning experience for us as college students. Our knowledge of wine quality and the underlying causes has increased as a result. Future study in this area will be well-founded on the revelations made.

This project demonstrates our commitment to data exploration and knowledge acquisition. The knowledge and understanding we gained from this analysis will unquestionably be crucial to our development as data aficionados and aspiring researchers.