Concepts and Technologies of AI – 5CS037

"A journey through Exploratory Data Analysis on Wine Quality."

University ID: 2332244

Lecturer: Mr. Siman Giri

Tutor: Mr. Ronit Shrestha

Submitted by: Naomi Thing

Submitted on: 2023/12/24.

Acknowledgement:

I would want to sincerely thank Mr. Siman Giri, our module leader and our lecturer for his invaluable guidance and constant support throughout the project. We also extend our special gratitude to our devoted tutor, Mr. Ronit Shrestha, for his aid and wisdom. This exploratory voyage into wine quality data analysis has been greatly shaped by your mentorship.

Table of Contents

1. Data Selection

1.1 Overview: Unveiling Wine Quality Insights

The extracted dataset from Kaggle was updated by M Yasser H 2 years ago. This dataset presents an engrossing story about the red varieties of Portuguese “Vinho Verde” wine. It describes in detail the various chemical concentrations in the wine and how they affect its overall quality. This dataset invites investigation, presenting the problem as a combination of regression and classification work with balanced but ordered classes – an arrangement where average wines outnumber good or bad ones.

Exploratory Questions:

- How do different chemical components correlate with each other?
- Is it possible to pinpoint the precise compounds that most significantly affect wine quality, and how do their concentrations affect the final impression?

1.2 Inspection: Unveiling Key Observations

We decided to focus our investigation on the wine quality dataset. A brief analysis was carried out to evaluate the information and its components, offering an overview of the dataset’s complexities.

The first inspection figured out that there were 1143 rows and 13 columns. The columns were of `int`, `float`, and `object` datatypes.

Columns representing their attributes:

- **Fixed acidity:** The fixed acidity of the wine.
- **Volatile acidity:** The volatile acidity of the wine.
- **Citric acid:** The citric acid content in the wine.
- **Residual sugar:** The residual sugar content in the wine.
- **Chlorides:** The chloride content in the wine.
- **Free sulfur dioxide:** The total amount of sulfur dioxide present in the wine.
- **Total sulfur dioxide:** The total amount of sulfur dioxide present in the wine.
- **Density:** The density of the wine.
- **pH:** The pH level of the wine.
- **Sulphates:** The sulphate content in the wine.

- **Alcohol:** the alcohol content of the wine.
- **Quality:** the quality rating of the wine.
- **Id:** An identifier for each record in the dataset.

Preliminary Findings:

- There are no missing values in the visible dataset.
- The "quality" column shows the dataset may be used for accessing wine quality.
- A highlighted value in the "volatile acidity" column suggests potential outliers or points of interest.

The foundation for more data cleansing and exploratory research was laid by these first discoveries.

2. Data Cleaning and statistical analysis

2.1 Data Cleaning

Following the first data inspection, the dataset underwent the following refinement steps.

Handling missing values: replaced null values with the mean values for specific columns, including: 'citric acid', 'chlorides', 'sulphates', and 'alcohol'. Confirmed the replacement to ensure no remaining null values in the dataset.

Removing duplicate rows: duplicate rows were identified and removed by using the `drop_duplicates()` method. Remaining duplicates after removal were checked, which returned '0' using `ndt.duplicated().sum()`.

Modifying and Creating new Columns: 'Id' was dropped, and then the column was checked to make sure there were no duplicate values left.

2.2 Summary Statistics

Using the `describe()` method, descriptive statistics were obtained for the numerical features. The `var()` function is used to find the variance of a numerical column, and although this function does not display these numbers, the range can be found by deducting the minimum from the maximum. Prior to utilizing the `mode()` method to determine the mode, we must first use the `.unique()` function to extract the unique values from our quality column for the categories column. At this point, the procedure is complete.

3. Data Visualization and Exploration

3.1 Univariate Analysis

- Figure 1

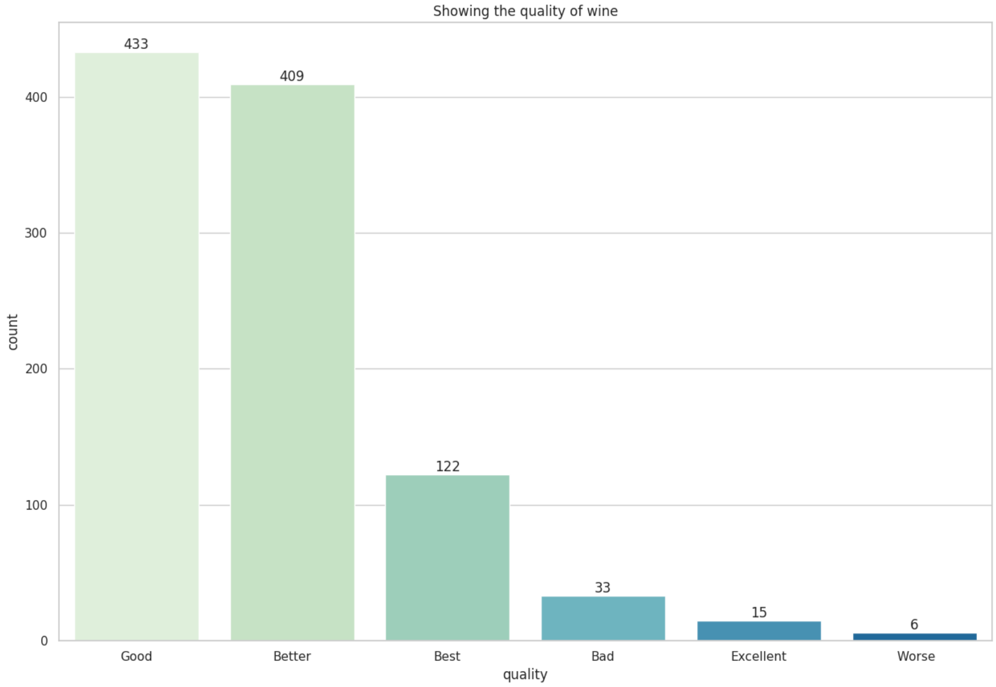


Figure 1: Bar plot of quantity

In the figure above, only a small percentage of wines are scored as “Excellent” or “Worse”, according to the univariate analysis of the wine quality dataset. Most wines are classified as “Good” or “Better”. Compared to the ‘Good’ and ‘Better’ quality categories, the ‘Best’ quality category has fewer wines, and ‘Bad’ wines are rarer. Most wines according to the chart, are either ordinary or slightly above average in quality. Understanding consumer preferences and directing winemaking production standards could both receive help from this information.

- **Figure 2**

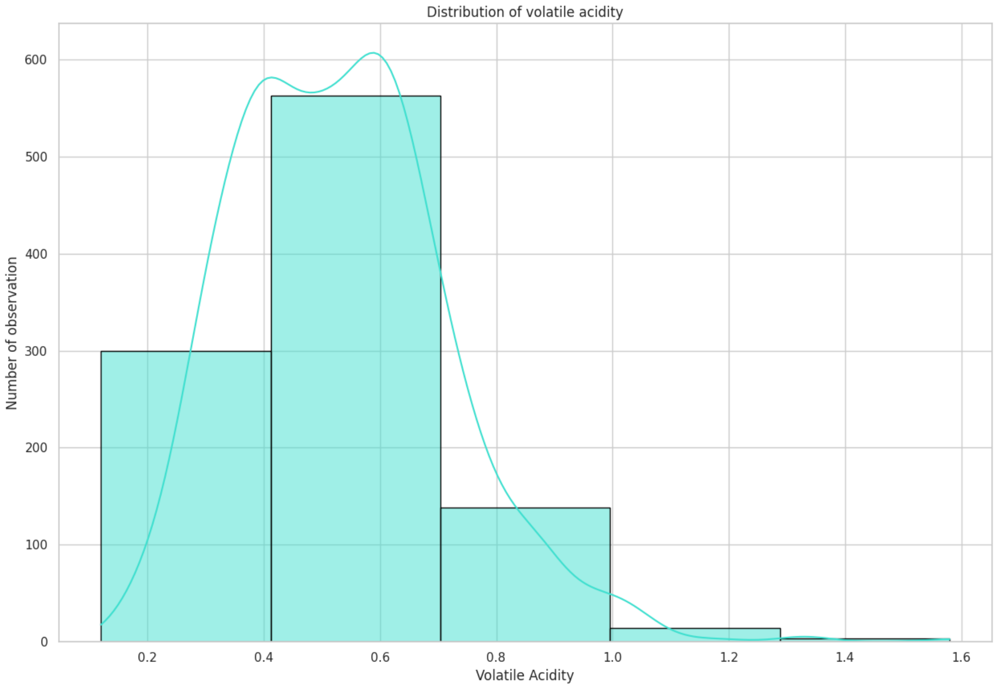


Figure 2: Histogram for volatile acidity

Some noteworthy findings appear from the univariate analysis of this second graphic, which shows the distribution of volatile acidity. The volatile acidity frequency distribution is depicted by the histogram, which also shows concentrations around lower values, with a peak occurring shortly before 0.4. The data distribution is smoothed out using the KDE curve, which displays a right-skewed pattern with a long tail that extends to higher values.

- **Figure 3**

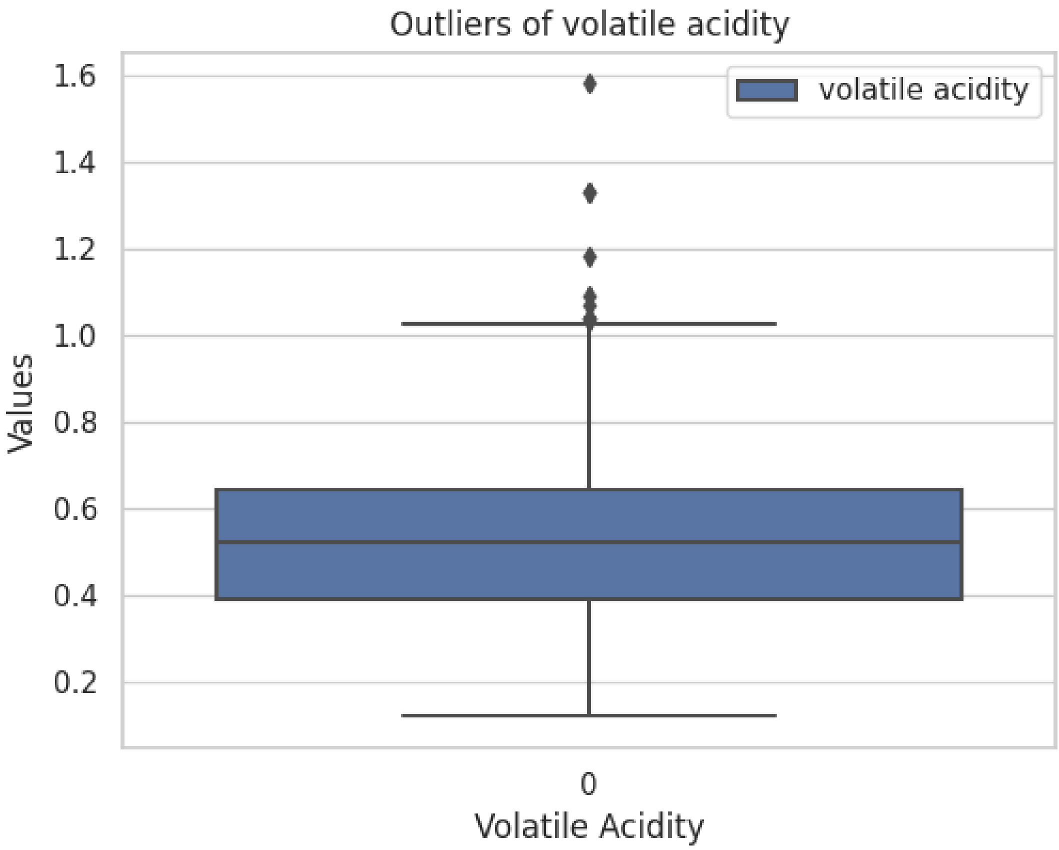


Figure 3: Outliers of volatile acidity

The key characteristics of volatile acidity are emphasized by the boxplot analysis. The median, which is found between 0.4 and 0.5, denotes a tightly clustered central tendency, while the interquartile range is quite narrow. A few outliers stick out above the upper whisker, suggesting notable departures from the general pattern. The data range is shown by the boxplot's whiskers, which highlight a possible right skew depending on where the median is located. The interquartile range and total range, which include outliers, are useful tools for assessing variability. These metrics are especially useful when examining wine quality evaluations and chemical process control.

3.2 Bi-variate Analysis

- Figure 1

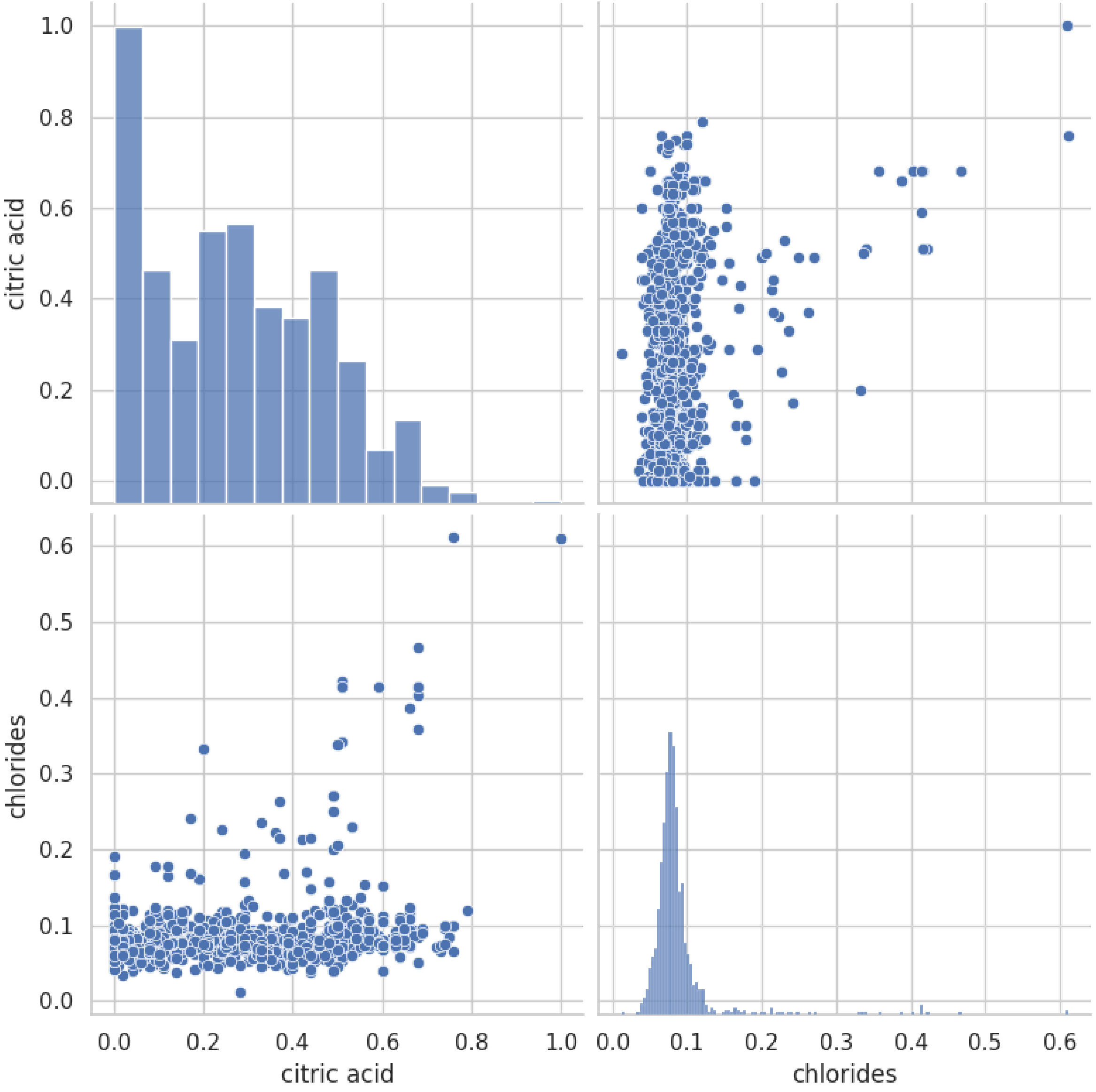
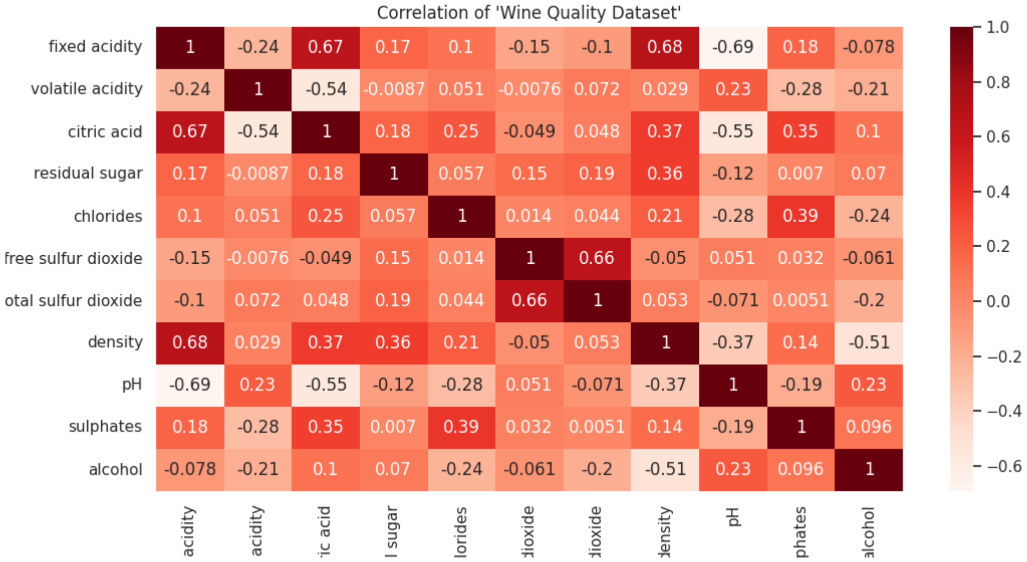


Figure 4: Bar graph and scatter plot

The wine quality dataset’s pair plot analysis reveals important tendencies. The ‘citric acid’ bar graph displays a distribution that is right skewed, showing a low level that is commonly present, with a center of 0.0 to 0.4. The ‘chlorides’ bar graph has a pattern that is tilted to the right, showing a decreased content, especially in the range between 0.0 and 0.1. ‘Citric acid’ and ‘chlorides’ have scatter plot that shows a scattered relationship rather than a strong linear link, showing that wines with lower citric acid levels have larger chloride ranges. There are noticeable errors, particularly with high “chlorides” and “citric acid” readings. In conclusion, the right-skewed distributions of both variables and the lack of a distinct linear pattern in their relationship highlight the significance of more statistical measures for a complete understanding.

- **Figure 2**



	acidity	acidity	ic acid	l sugar	lorides	ioxide	ioxide	ensity	pH	phates	lcohol
fixed acidity	1	-0.24	0.67	0.17	0.1	-0.15	-0.1	0.68	-0.69	0.18	-0.078
volatile acidity	-0.24	1	-0.54	-0.0087	0.051	-0.0076	0.072	0.029	0.23	-0.28	-0.21
citric acid	0.67	-0.54	1	0.18	0.25	-0.049	0.048	0.37	-0.55	0.35	0.1
residual sugar	0.17	-0.0087	0.18	1	0.057	0.15	0.19	0.36	-0.12	0.007	0.07
chlorides	0.1	0.051	0.25	0.057	1	0.014	0.044	0.21	-0.28	0.39	-0.24
free sulfur dioxide	-0.15	-0.0076	-0.049	0.15	0.014	1	0.66	-0.05	0.051	0.032	-0.061
otal sulfur dioxide	-0.1	0.072	0.048	0.19	0.044	0.66	1	0.053	-0.071	0.0051	-0.2
density	0.68	0.029	0.37	0.36	0.21	-0.05	0.053	1	-0.37	0.14	-0.51
pH	-0.69	0.23	-0.55	-0.12	-0.28	0.051	-0.071	-0.37	1	-0.19	0.23
sulphates	0.18	-0.28	0.35	0.007	0.39	0.032	0.0051	0.14	-0.19	1	0.096
alcohol	-0.078	-0.21	0.1	0.07	-0.24	-0.061	-0.2	-0.51	0.23	0.096	1

Figure 5: Heatmap of Wine quality dataset correlation

The figure shows a heatmap that illustrates the relationship between different physiochemical characteristics in a "Wine Quality Dataset." The correlation strength is shown by color gradients, which go from dark red (correlation of 1) to white (correlation of 0). Variables rising together are shown by positive correlations (closer to 1), whereas negative correlations (closer to -1) suggest in inverse relationship. Notable relationships include a significant negative association and a strong positive correlation between citric acid and volatile acidity and fixed acidity, respectively. Understanding the interrelationships in the dataset with the help of this visual tool is crucial for predictive modelling and data analysis.

4. Conclusion

To sum up, our investigation into the Wine Quality Dataset uncovered several fascinating trends. Uneven distributions were revealed by univariate analysis, particularly in the volatile acidity and chloride concentration. While the heatmap displayed attribute correlations, bivariate analyses using pair plots and boxplots revealed intricate linkages. Outliers demand closer examination, and although visualizations supplied some early clarity, more statistical analysis could improve accuracy. These discoveries lay the groundwork for next investigations and highlight the importance of statistical and data visualization techniques in interpreting the subtleties of wine quality.