Applied Data Science

Clustering and Fitting

Kaggle Dataset link: [Red Wine Quality (kaggle.com)]

GitHub Link: Rajith24699/ADS-Clustering-and-Fitting (github.com)

Student Name: Rajith Narasimha Murthy

Student ID:23004934

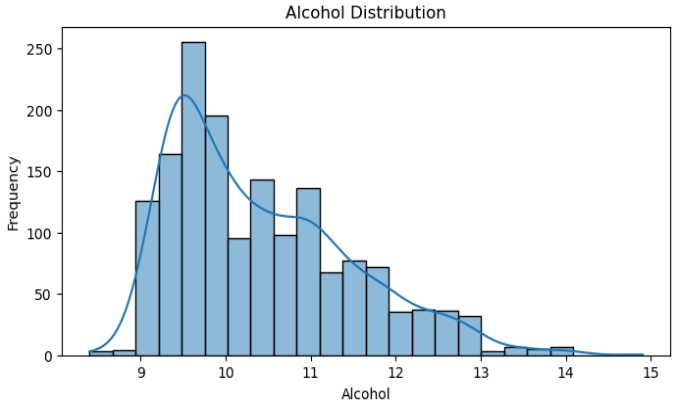
Introduction:

The wine dataset provides a comprehensive collection of attributes and quality ratings for red wine samples. It encompasses various chemical properties of wines.

Exploring this dataset can provide valuable insights into the relationship between chemical composition and the perceived quality of red wines.

Data Visualization:

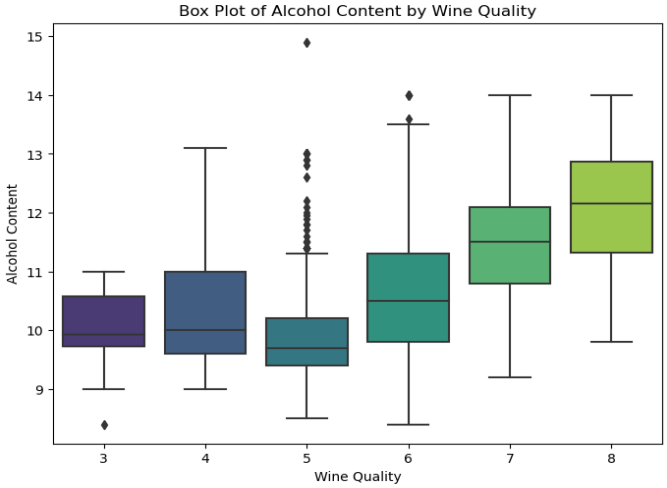
➢ Histogram for the 'alcohol' distribution:



The graph displays the distribution of alcohol content in various samples. The x-axis represents the alcohol content, ranging from 9 to 15. The y-axis represents the frequency, indicating how often each alcohol content appears in the dataset. The highest frequency occurs around 250 at an alcohol content of approximately 10.

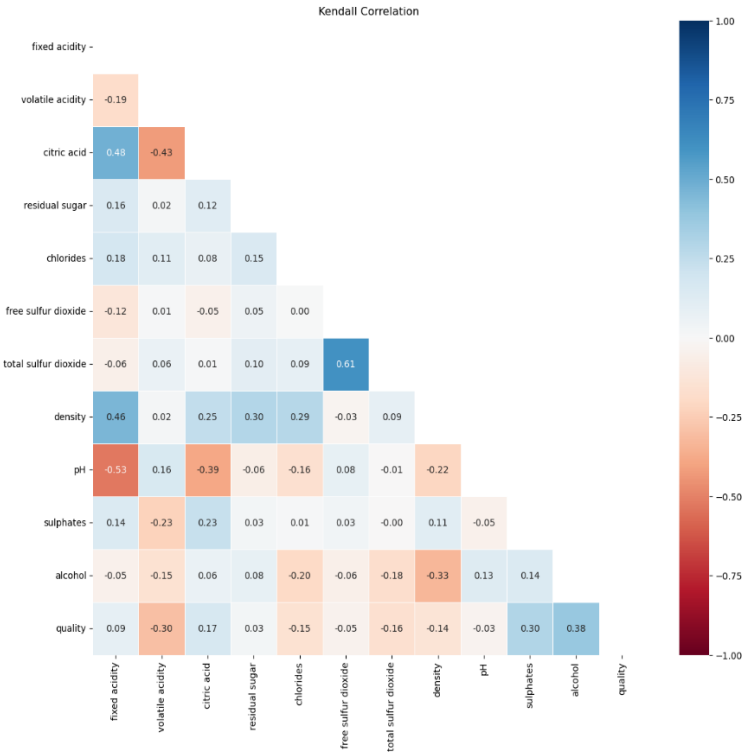
The majority of samples have alcohol content centred around 10. There is a gradual decline in frequency as alcohol content deviates from this central value. Fewer samples have extremely low or high alcohol content.

➢ Box Plot of Alcohol Content by Wine Quality



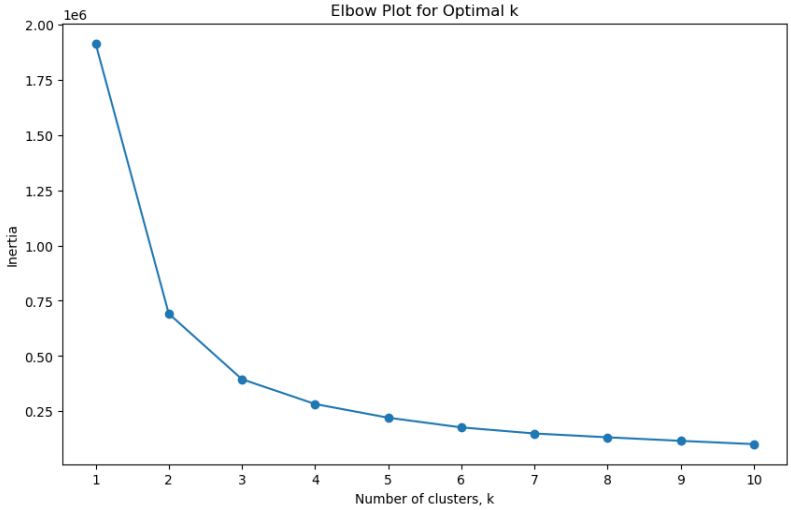
The box plot displays the distributi on of alcohol content for different wine quality ratings. Each box repr esents the interquartile range (IQR) of alcohol content for wines of a specific quality rating. The line insi de each box indicates the median alcohol content.Whiskers extend to show the range within 1.5 times IQ R, Points outside this range plots as individual dots, representing outliers.

➢ Heatmap



The heatmap represents the Kendall Correlation between different variables related to wine quality. Each square in the heatmap shows the correlation between the variables on each axis. Correlation values range from -1 (strong negative correlation) to +1 (strong positive correlation). Darker colors indicate stronger correlations.Citric Acid has a strong negative correlation with pH (-0.39). This suggests that as citric acid content increases, pH tends to decrease.Density has a moderate positive correlation with residual sugar (0.29) and total sulfur dioxide (0.29).Alcohol has moderate positive correlations with quality (0.38).Many other features exhibit weaker correlations, as indicated by lighter colors

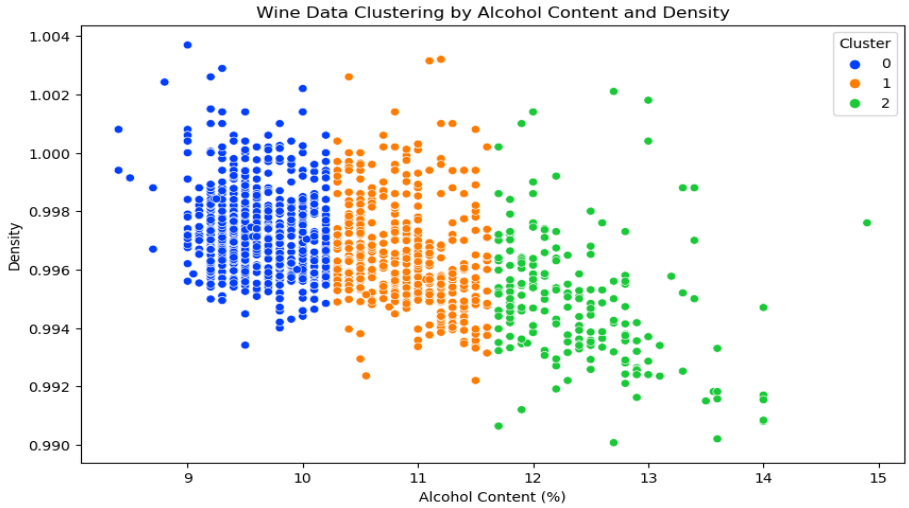
➢ Elbow Plot



The x-axis is labeled “Number of clusters, k” and ranges from 1-10.
The y-axis is labeled “Inertia” and has values ranging approximately from 0 to 2.00e6
Blue line connects dots representing different inertias for each respective cluster number.
There’s a sharp decline in inertia from 1 to around 4 clusters beyond this point, reduction in inertia is minimal

The inertia (also known as the within-cluster sum of squares) measures how tightly the data points within each cluster are grouped. As we increase the number of clusters (k), the inertia tends to decrease. The “elbow” point represents an optimal value for k, where adding another cluster doesn’t significantly improve the fit to the data. In this plot, it appears that k=4 could be an appropriate choice for the number of clusters.
When performing k-means clustering, consider selecting 4 clusters based on this elbow plot. However, further domain knowledge and validation are necessary to confirm the optimal k value.

➢ Clustering:



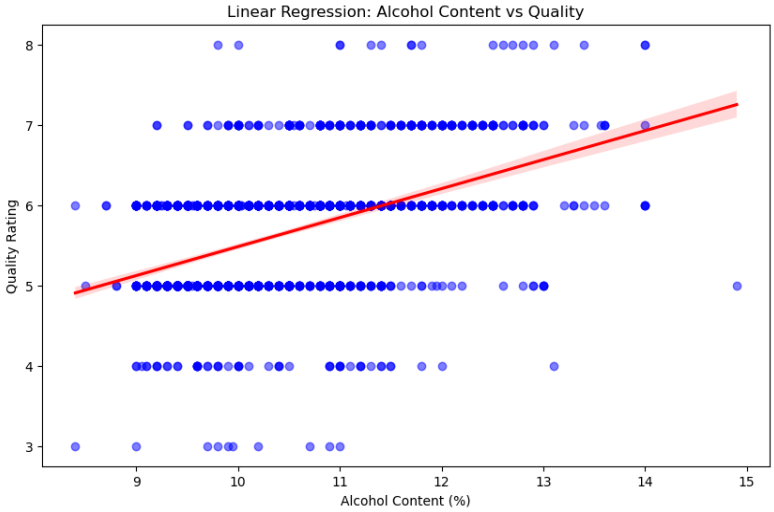
The scatter plot displays wine data clustered by alcohol content and density. There are three distinct clusters labelled as 0, 1, and 2.

Different clusters represent distinct wine profiles, which can guide production decisions

Cluster Analysis:

- Cluster 0 (Blue): Wines in this cluster have higher density (approximately 0.996 to 1.004) and lower alcohol content (around 9% to 11%). These wines are likely sweeter and less alcoholic.
- Cluster 1 (Orange): Wines in this cluster have medium density (around 0.994 to just above 1.000) and medium alcohol content (approximately between 10% to nearly 13%). These wines could be balanced in terms of sweetness and alcohol content.
- Cluster 2 (Green): This cluster consists of wines with lower density (roughly from about 0.994 to 0.998) and higher alcohol content (from about 12% up to almost 15%). These wines are likely drier and more alcoholic.

➢ Linear Regression:



The scatter plot illustrates how the alcohol content (%) relates to the quality rating of wines. Each blue dot represents a wine sample, showing different combinations of alcohol content and quality.
The red line across the plot represents the best-fit line, summarizing the overall trend between these two variables. Generally, there's an upward trend in quality as alcohol content increases.
However, there's notable variability in quality ratings at each alcohol level. The red line helps estimate how quality tends to change with varying alcohol content.

- **The Mean Squared Error (MSE):** value of **0.4995** indicates the average squared difference between the actual quality ratings and the predicted quality ratings on the test set.
- **The R-squared (R^2):** score of **0.2356** represents the proportion of variance in the quality ratings that is explained by the linear regression model.

In this analysis, we explored the wine dataset through data visualization, clustering, and model-fitting techniques. The clustering analysis identified distinct clusters based on wine attributes, while the regression model provided insights into the relationship between alcohol content and wine quality.