Zach Khan

Miniproject 1: R

**Dataset Description: Wine Quality Dataset**

The dataset I chose to analyze is the wine quality dataset that I found on Kaggle from the UCI Machine Learning Repository. This dataset contains physicochemical attributes of white wine samples from the Vinho Verde region of Portugal. The dataset consists of 4,898 instances with 12 attributes such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality score. The quality score is an integer value ranging from 3 to 9 which represents the sensory rating given by the wine experts who did the tastings. The goal is to look at the relationships between different physicochemical properties and the quality of the wine.

**Analysis and Results**

1. Basic Statistical Summary

A basic statistical summary was generated to examine the distribution of each feature. The summary revealed that alcohol content ranged from 8.0 to 14.2, and the quality scores had a mean of 5.88, indicating that most wines were of average quality. This helped to understand the overall spread and central tendency of the data.

2. Check for missing values

A check for missing values was performed, confirming that there were no missing values which made sure that the dataset was complete and ready for the other operations.

3. Distribution of Wine Quality Scores

The distribution of wine quality scores was visualized using a bar chart, showing that most wines were rated between 5 and 6 (very few samples receiving extreme scores). This suggested that the dataset was pretty balanced but slightly skewed toward average quality wines.

4. Correlation Analysis

A correlation analysis was conducted to see if there were any relationships between variables. From that I found that alcohol had the highest positive correlation (0.44) with quality which could mean that wines with higher alcohol content tended to receive better ratings. And on the other end, density had a negative correlation (-0.31) with quality, meaning that denser wines were generally rated lower.

5. Box plots for key features

Box plots were made to see the variability in different features which showed that features like volatile acidity and chlorides had several outliers.

6. Identify potential outliers using IQR method

Because of operation 5, an outlier detection analysis using the Interquartile Range (IQR) method was performed. I found that citric acid, volatile acidity, and chlorides had a significant number of outliers. This highlights variations when the wine was made or inconsistencies with the measurement.

7. Quality prediction analysis

A linear regression model was built to find the relationship between physicochemical attributes and wine quality. The model got an R-squared value of 0.28 which mean that the features explained only 28% of the variance in wine quality. So while physicochemical properties do influence quality, other factors like sensory attributes and winemaking techniques play a role.

8. Feature importance analysis

Analysis of feature importance showed that density, volatile acidity, pH, sulphates, and alcohol were the most influential features. The negative impact of density and volatile acidity on quality suggested that wines with high acidity and density were generally rated lower.

9. Density plots for key features by quality rating

Density plots were created for the key features across different quality ratings. This showed that higher-quality wines tended to have higher alcohol content, while lower-quality wines had more variation in volatile acidity. The plots lined up with the correlation findings and helped reinforce the significance of certain attributes in determining wine quality.

10. Skewness, Kurtosis, and Spread in Wine Quality

The mean quality score was 5.88, with a standard deviation of 0.89, indicating moderate spread among ratings. The skewness (0.16) showed a slight rightward tilt, meaning more wines had higher ratings than expected in a normal distribution. The kurtosis (3.22) showed that the distribution was slightly more peaked than a normal distribution, indicating that extreme values were slightly more frequent. The quality range was 6 (from 3 to 9), indicating that there is a wide range of wine qualities in the dataset. These statistical measures helped validate previous findings from the correlation and regression analyses.

Summary

The most significant features were ranked based on their correlation with wine quality, with alcohol (0.44) showing the strongest positive correlation and density (-0.31) having the strongest negative correlation. With those results, increasing alcohol content while having lower density could contribute to higher wine quality ratings.

Output  
Packages, libraries, and read dataset

A computer code with blue text

Description automatically generated

Operation 1  
A screenshot of a computer

Description automatically generated  
Operation 2

A close-up of a white background

Description automatically generated

Operation 3

A close-up of a computer code

Description automatically generated  
A graph of quality score

Description automatically generated  
Operation 4

A close-up of a computer code

Description automatically generated

A close-up of a chart

Description automatically generated

Operation 5

A computer code with blue text

Description automatically generated

A group of different types of sulfur

Description automatically generated with medium confidence

Operation 6

A screenshot of a computer code

Description automatically generated

Operation 7

A screenshot of a computer

Description automatically generated

Operation 8A screenshot of a computer

Description automatically generated

Operation 9

A close-up of a math equation

Description automatically generated

A graph of colored lines

Description automatically generated

Operation 10

A computer screen shot of a computer code

Description automatically generated

Other

A screenshot of a computer code

Description automatically generated

