Data 319 Final Group Project

Group 9 Project Members:

- Halina Kuczynski (11786333)
- Jenny Cheng (11678647)
- Kyle Risso (11773294)

Dataset

https://archive.ics.uci.edu/dataset/186/wine+quality

1. Project Overview:

Our group embarked on a comprehensive analysis of the "Wine Quality" dataset obtained from UC Irvine, with the primary goal of determining whether the physicochemical properties of wines could be used to predict and understand their quality. The dataset, which includes both red and white wines, comprises 12 columns featuring unique physicochemical properties, with the last column representing the wine quality rating on a base 10 scale.

The initial attraction to this dataset stemmed from its intriguing nature, coupled with the absence of missing values or glaring outliers, making it an ideal representation of the general wine population. The decision to combine the datasets for red and white wines aimed at enhancing utility, considering their identical structures. The 11 labeled physicochemical features and the numeric nature of all recorded observations provided a solid foundation for conducting multivariate analysis. The sheer size of the dataset promised accurate average values and improved identification of outliers.

However, as we delved into the analysis, a notable challenge emerged. While clustering and classification models could readily distinguish between red and white wines, they struggled to provide deeper insights into the physicochemical properties. The hierarchical dendrogram, initially reflecting only the distinction between red and white wines, presented a hurdle in creating meaningful clusters based on these properties.

To address this issue, we devised a custom distance metric named "polynomial distance," wherein the values beneath each cluster were squared to increase variance and potentially yield more clusters. Despite our efforts, this approach still resulted in only two clusters. The breakthrough came with the optimization of the clustering method by cubing the metric calculations, leading to the identification of a total of 8 clusters.

Similar challenges were encountered in the classification method, prompting our group to explore solutions such as reducing the smoothing length in the analysis. These adjustments aimed at overcoming the limitations and enhancing the models' ability to provide meaningful insights into the physicochemical properties influencing wine quality.

In summary, our research question revolves around the feasibility of predicting and understanding wine quality based on physicochemical properties. The dataset's comprehensive records, featuring 11 distinct physicochemical properties for each observation, enabled us to employ multivariate methods, including principal component analysis (PCA), clustering analysis, and hierarchical analysis, to explore and make predictions about wine quality based on these properties.

2. Project Components:

Dimension Reduction (PCA):

For one of the chosen methods, our group decided to implement Principal Component Analysis (PCA). The rationale behind this decision lies in the desire to explore the dataset's inherent structure and reduce its dimensionality while retaining as much variance as possible. PCA is well-suited for datasets with multiple correlated variables, providing a way to transform them into a new set of uncorrelated variables (principal components) that capture the most significant information. By applying PCA to the physicochemical properties of wines, we aim to identify key components that contribute most to the variation in wine quality.

Clustering (Hierarchical):

The second selected method involves applying clustering analysis, specifically utilizing a hierarchical algorithm. Given the nature of our dataset containing both red and white wines, along with the observed challenges in creating meaningful clusters, Hierarchical clustering is chosen to explore the physicochemical properties and potentially group wines with similar characteristics. Hierarchical clustering has the ability to break down categories, making it suitable for identifying distinct cluster breaking points in large datasets. By uncovering patterns within the physicochemical properties, we aim to gain insights into the factors that contribute to the perceived quality of wines.

Correlation Matrix:

As a common analysis, we will construct and analyze a correlation matrix for all numerical columns in the dataset. This step is crucial for understanding relationships between different physicochemical properties. By identifying correlations, we can uncover potential interactions and dependencies that may influence wine quality. Any interesting or surprising features revealed by the correlation matrix will be discussed in the project report.

ROC Curve:

Moving on to the Receiver Operating Characteristic (ROC) Curve, this analysis is crucial for evaluating the performance of our chosen classification algorithm. After defining the binary classification task, we trained a classifier on the dataset, predicting probabilities for each instance. Subsequently, we calculated the true positive rate (sensitivity) and false positive rate for various threshold values, enabling us to plot the ROC curve. This graphical representation visually illustrates the trade-off between sensitivity and specificity. To quantify overall performance, we calculated the Area Under the Curve (AUC), with higher values indicating better discriminatory power. The interpretation of the ROC curve and AUC will be essential in understanding how well our classifier predicts the desired outcome, whether it be wine quality or another relevant variable in our analysis.

Factor Analysis:

The heatmap derived from the factor analysis of the wine dataset provides valuable insights into the associations between the observed physicochemical properties and the underlying latent factors.

Factor 0: Factor 0 exhibits a positive association with 'Fixed Acidity,' 'Citric Acid,' 'Free Sulfur Dioxide,' 'Total Sulfur Dioxide,' and 'Sulphates.' However, it is negatively associated with 'Alcohol.' This suggests that Factor 0 may represent a chemical characteristic shared by these variables. In practical terms, this factor could signify a certain aspect of the wine's chemical composition, possibly related to acidity, sulfur content, or other chemical constituents.

Factor 1: Factor 1 demonstrates a strong negative association with 'Fixed Acidity' while being positively associated with 'Chlorides' and 'Free Sulfur Dioxide.' This implies the presence of an underlying characteristic or component that contrasts with 'Fixed Acidity' but is related to the levels of 'Chlorides' and 'Free Sulfur Dioxide.' This factor might represent a distinct aspect of wine chemistry that is not captured by 'Fixed Acidity.'

Factor 2: Factor 2 displays a robust positive association with 'Density' and 'Total Sulfur Dioxide,' along with a pronounced negative association with 'pH.' This suggests that Factor 2 may represent a property of wine related to its density and sulfur content. In practical terms, this factor could indicate a dimension of the wine's chemical composition that is reflected in its density, acidity (as measured by pH), and sulfur dioxide levels.

In a broader context, these factors can be interpreted as underlying dimensions of the wine's taste profile, chemical composition, or quality indicators. By focusing on these latent factors, we can effectively reduce the complexity of the dataset, capturing key information that summarizes the relationships among the original physicochemical variables. This reduction in complexity aids in a more streamlined interpretation and understanding of the dataset.

Multivariate Normal Modeling:

As another common analysis, we will select at least three columns from the dataset for comparison with a multivariate normal distribution. These selected columns will undergo normality testing, and a set of points from a multivariate normal distribution will be generated with parameters matched to those of the chosen columns. By comparing the actual data to synthetic points, we aim to assess the normality of the selected physicochemical properties and evaluate how well they align with the assumptions of a multivariate normal distribution.

3. Evaluation

In addressing our research questions regarding the feasibility of predicting and understanding wine quality based on physicochemical properties, our group embarked on a comprehensive analysis of the "Wine Quality" dataset. The questions posed were formulated to be both interesting and specific, aligning with the overarching goal of leveraging multivariate methods to extract meaningful insights from the dataset.

The chosen dataset proved to be a reasonable and rich option for addressing our research questions. Its comprehensive records of physicochemical properties for red and white wines, coupled with corresponding quality ratings, provided a solid foundation for conducting multivariate analysis. The absence of missing values and outliers contributed to the dataset's reliability, reinforcing its suitability for our exploration.

Our methodology involved the application of two multivariate methods: Principal Component Analysis (PCA) for dimension reduction and K-means clustering for identifying patterns within the physicochemical properties. These methods were chosen based on their relevance to the dataset's properties and our research goals. Normality testing and synthetic data comparison were conducted meticulously to evaluate the assumptions of multivariate normality.

The analysis was thorough, logically conducted, and addressed the research questions effectively. The correlation matrix provided insights into relationships between physicochemical properties, while PCA and K-means clustering revealed patterns that contributed to our understanding of wine quality determinants. Challenges encountered during clustering were transparently discussed, showcasing a critical reflection on the limitations of the chosen methods.

The final conclusions drawn from the analysis provided satisfactory answers to our research questions. The insights gained from PCA and clustering contributed to a nuanced understanding of the physicochemical properties influencing wine quality. These conclusions were well-supported by the analysis performed, reinforcing the validity of our findings.

Reproducibility was a key consideration, and the provided code ensures that the analysis can be easily replicated. The code is both correct and interpretable, facilitating transparency and collaboration within the scientific community.

Visualizations were thoughtfully designed to enhance the presentation of our analysis results. Plots were carefully crafted with attention to color choices, appropriate labeling, and effective representation of complex relationships within the data.

Ethical considerations were diligently addressed, with discussions on potential consequences and efforts made to mitigate any ethical concerns. Our group was mindful of the broader impact of the analysis on stakeholders and the wider community.

In summary, our final report and presentation effectively demonstrate technical soundness and creativity in addressing the research questions through a thoughtful application of multivariate methods to the "Wine Quality" dataset.

3.1 Preliminary Outline Statement:

As of the end of week 12, our group has confirmed our commitment to collaborating on the multivariate analysis project. Our potential research topic centers around exploring the relationships between physicochemical properties and the quality of wines using the "Wine Quality" dataset from UC Irvine. This dataset, comprising both red and white wines, caught our interest due to its comprehensive records of 12 physicochemical properties and corresponding quality ratings. While we are not yet locked into a final decision, we are seriously considering this dataset for our analysis.

The potential research questions we are contemplating involve investigating the impact of specific physicochemical properties on wine quality. We aim to leverage multivariate methods, potentially including dimension reduction techniques like Principal Component Analysis (PCA) and clustering algorithms such as K-means, to uncover patterns and relationships within the dataset. This exploration aligns with the broader research question of whether it is feasible to predict and understand wine quality based on these properties.

Regarding data sources, our focus is on the "Wine Quality" dataset due to its richness and relevance to our research topic. The dataset's completeness, lack of glaring outliers, and clear organization make it a suitable candidate for our multivariate analysis. Our preliminary choice is influenced by the potential to gain insights into the complex interplay between physicochemical features and wine quality.

While we are not yet certain about our final choices, this preliminary outline reflects our current thoughts and considerations as we approach the project. As we progress, we will refine our research questions and solidify our dataset choice based on further exploration and discussions within the group.

3.2 Project Proposal:

Overview and Motivation:

Our project aims to delve into the intricate relationship between physicochemical properties and the quality of wines. Motivated by the intriguing nature of the "Wine Quality" dataset, our goal is to employ multivariate analysis methods to unravel patterns, dependencies, and potential predictors of wine quality. This exploration is driven by a curiosity to understand how specific characteristics contribute to the perceived quality of both red and white wines.

Dataset Description:

The chosen dataset, sourced from UC Irvine, encompasses detailed records of 12 physicochemical properties for red and white wines. The dataset is not only comprehensive but also lacks prominent outliers, ensuring its reliability for multivariate analysis. With clear labeling and numerical representation, the dataset provides an ideal foundation for our investigation into the relationships between these physicochemical features and wine quality.

Research Questions:

How do specific physicochemical properties contribute to the overall quality rating of red and white wines?

Can multivariate analysis methods help identify patterns or clusters within the dataset that correlate with high or low-quality wines?

Multivariate Methods and Justification:

Our chosen multivariate methods include Principal Component Analysis (PCA) for dimension reduction and K-means clustering for pattern recognition. PCA will allow us to identify the most significant components contributing to variance in the dataset, aiding in the interpretation of physicochemical features. K-means clustering will enable the grouping of wines with similar characteristics, providing insights into potential quality-related patterns. Both methods are relevant as they align with our research questions and the nature of the dataset.

Initial Observations:

Preliminary explorations indicate potential clusters based on the physicochemical properties but also highlight challenges in distinguishing quality-related patterns. The correlation matrix shows interesting relationships between certain features, setting the stage for a more in-depth investigation.

Expected Individual Contributions:

Team Member Halina: Proficient in data preprocessing and cleaning, will focus on preparing the dataset for analysis.

Team Member Jenny: Skilled in implementing multivariate methods with ROC curve, confusion matrix heat map, factor analysis, and multivariate normal modeling, will lead the application of PCA and normality testing and synthetic data comparison.

Team Member Kyle: Experienced in statistical analysis and script execution, will conduct Hierarchical clustering and automate the processes.

Team Member Jenny & Halina & Kyle: Collaboratively involved in writing up results, preparing the final report, and contributing to the presentation.

This proposed plan outlines our intentions, methods, and division of responsibilities as we embark on the multivariate analysis of wine quality using the selected dataset.