REPORT TITLE:
"Exploratory Data Analysis and Clustering of Wine Quality Dataset"

NAME:
Madina Kumabayeva

ID:
30406A, DSE Student 1st year

Introduction:

This report presents an in-depth analysis of the Wine Quality Dataset, focusing on unsupervised learning techniques to explore natural groupings and patterns. The dataset, which includes various physicochemical properties and quality ratings of wine samples, offers a unique opportunity to understand the factors that influence wine quality through clustering.

Description of the Dataset:

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

The dataset consists of measurements of different wine samples from the north of Portugal. It includes both physicochemical properties such as alcohol percentage, acidity, sugar level, and sensory quality rated by experts. Each entry represents a separate wine sample, making the dataset suitable for unsupervised learning to uncover patterns or groupings without using the quality ratings as predefined labels. The white wine quality dataset contains a comprehensive collection of physicochemical properties and quality ratings of white wines. The dataset consists of 4898 entries, each representing a unique wine sample. The data is organized into 12 columns, detailing various chemical measurements and the corresponding quality rating of each wine.

Data Structure and Variables. The dataset includes the following variables:

1. Fixed Acidity (float64): The concentration of non-volatile acids in the wine, measured in grams per liter (g/L). This variable typically affects the taste and stability of the wine.
2. Volatile Acidity (float64): The concentration of volatile acids in the wine, measured in grams per liter (g/L). Higher levels can lead to an unpleasant, vinegar-like taste.
3. Citric Acid (float64): The concentration of citric acid, measured in grams per liter (g/L). It contributes to the freshness and flavor of the wine.
4. Residual Sugar (float64): The amount of sugar remaining after fermentation, measured in grams per liter (g/L). It can impact the sweetness of the wine.
5. Chlorides (float64): The concentration of chlorides (salt) in the wine, measured in grams per liter (g/L). Higher levels can negatively affect the taste.

6. Free Sulfur Dioxide (float64): The amount of free sulfur dioxide, measured in milligrams per liter (mg/L). It acts as an antioxidant and antimicrobial agent in the wine.
7. Total Sulfur Dioxide (float64): The total amount of sulfur dioxide, measured in milligrams per liter (mg/L), including both free and bound forms. It helps in preserving the wine.
8. Density (float64): The density of the wine, measured in grams per cubic centimeter (g/cm^3). This variable can be used to infer the alcohol content and sugar levels.
9. pH (float64): The pH level of the wine, indicating its acidity or alkalinity. The pH scale ranges from 0 to 14, with lower values being more acidic.
10. Sulphates (float64): The concentration of sulphates, measured in grams per liter (g/L). Sulphates can contribute to the wine's stability and enhance its flavor.
11. Alcohol (float64): The alcohol content of the wine, measured as a percentage (%). This variable is crucial for determining the wine's strength and overall quality.
12. Quality (int64): The quality rating of the wine, typically on a scale from 0 to 10. This subjective measure is provided by wine experts and reflects the overall sensory experience of the wine.

Data Preprocessing:

- Scaling and Normalization: All features were scaled to have zero mean and unit variance to ensure that no single variable would dominate the distance calculations used in clustering.
- Handling Missing Values: Missing values were imputed using the median value for each feature, ensuring that the dataset was complete for analysis.

Clustering Analysis:

- K-Means Clustering: Applied to segment the data into groups based on similar chemical properties. The optimal number of clusters was determined using the Elbow Method, which suggested a clear bend at three clusters, indicating a natural grouping within the data.
- Silhouette Analysis: Conducted to assess the quality of the clusters formed by K-Means. The silhouette scores indicated moderate separation and cohesion within the clusters, with an average score suggesting that the clusters are reasonably well-defined but could be improved.

Visualizations:

- Cluster Summary Plot: Shows the average value of each physicochemical property for each cluster, providing insights into the distinguishing characteristics of each group.
- PCA Scatter Plot: Reduces the dimensionality of the data to two principal components and visualizes the clusters in this new space, highlighting the separation among them.

- Silhouette Plot: Each sample’s silhouette score is plotted, helping to visually assess how well each object has been clustered.

The clustering analysis revealed distinct groupings within the wines based on their physicochemical properties, suggesting intrinsic patterns that relate to wine quality. These findings could help in targeting specific quality improvement strategies for wine production or in segmenting wines into quality-based categories more effectively.

Regression Analysis:

Linear Regression: Performed to predict the quality rating of the wine based on its physicochemical properties. The regression model provided insights into which features were most influential in determining wine quality.

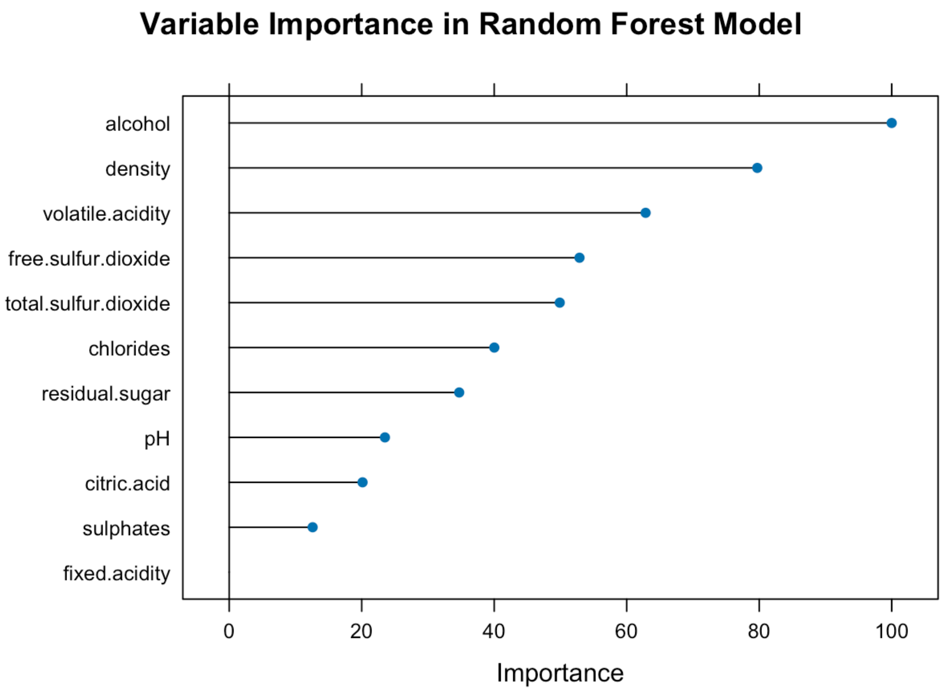
Model Evaluation: The performance of the regression model was assessed using metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics indicated the model's accuracy and the extent of prediction errors.

Feature Importance: The regression coefficients were analyzed to determine the relative importance of each feature. Features like alcohol content and residual sugar were found to have significant impacts on the predicted quality.
The regression analysis complemented the clustering findings by quantifying the relationships between physicochemical properties and wine quality. This approach can aid in developing predictive models for wine quality assessment and in identifying key factors for quality improvement.

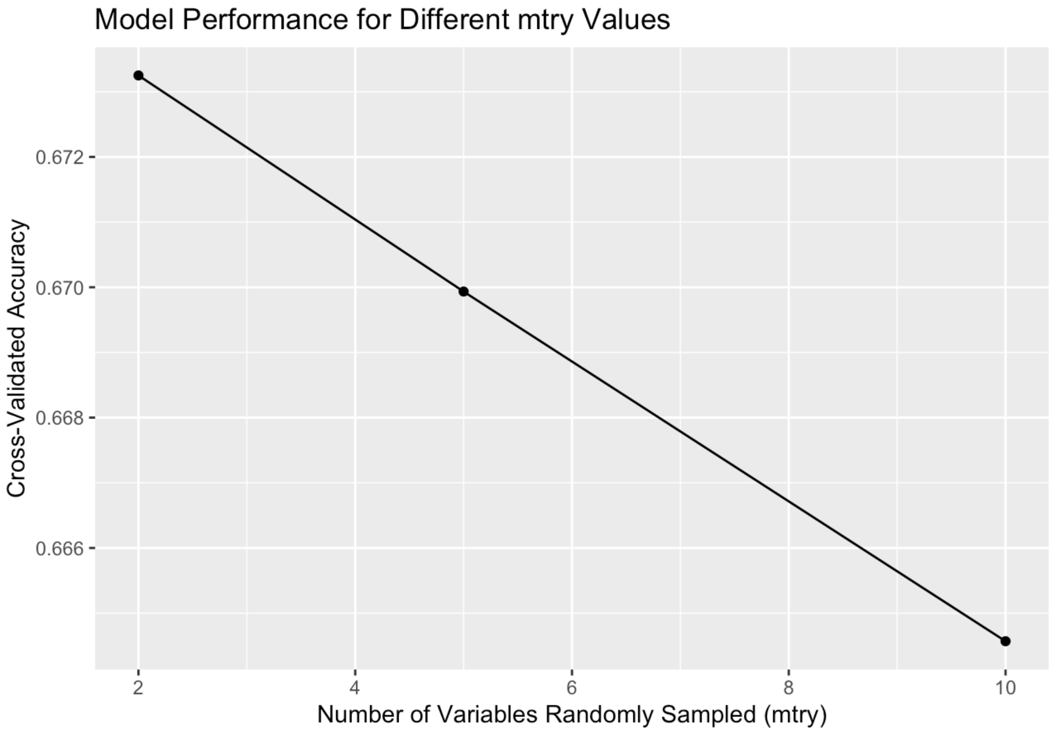
The project created in Rstudio applies various statistical learning techniques, including Random Forest classification, clustering with K-Means, and linear regression. Here is a detailed description of each technique:

Random Forest Classification

The Random Forest algorithm is used for classification in this code to predict the quality of white wine based on its physicochemical properties. The process starts with loading the data and converting the quality column into a factor for classification. The dataset is then split into training and test sets, and a Random Forest model is trained using 10-fold cross-validation to ensure robust model performance. Predictions are made on the test data, and the model's performance is evaluated using a confusion matrix, which provides metrics such as accuracy, sensitivity, and specificity. The model is further tuned by adjusting the mtry parameter, which controls the number of variables considered at each split in the decision trees of the forest. Variable importance is assessed to understand which features most influence the quality predictions, and the results are visualized through plots showing model performance and confusion matrices.



"Variable Importance Plot" from a Random Forest model, commonly used in machine learning to evaluate the significance of different predictors in the model.



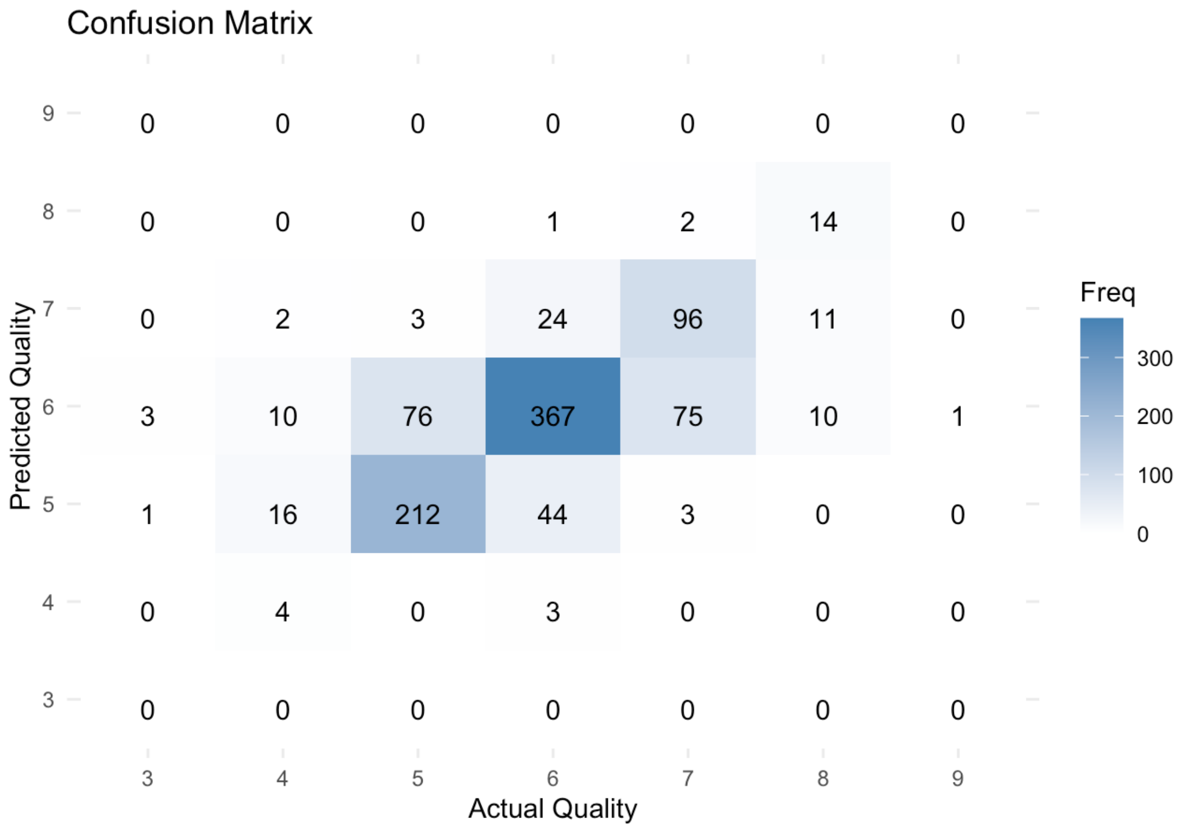
The plot titled "Model Performance for Different mtry Values" illustrates the relationship between the number of variables randomly sampled (mtry parameter) in a Random Forest model and the model's cross-validated accuracy. This graph uses a line chart to depict how the accuracy of the model changes as the mtry value increases from 2 to 10.

In the graph, the x-axis represents the number of variables randomly sampled (mtry), while the y-axis shows the cross-validated accuracy of the model. Each data point on the plot represents the average accuracy achieved by the model at each specific mtry value during cross-validation.

From the plot, it is observed that the model's accuracy decreases steadily as the number of variables sampled increases. The highest accuracy is achieved at the lowest mtry value of 2, indicating a cross-validated accuracy of approximately 0.672. As the mtry value increases to 10, the accuracy declines to about 0.666.

This trend suggests that sampling fewer variables at each split in this Random Forest model leads to higher accuracy. It can be inferred that for this specific dataset and model setup, including more variables in the decision process at each node may introduce noise or less relevant information, thus decreasing the model's predictive performance.

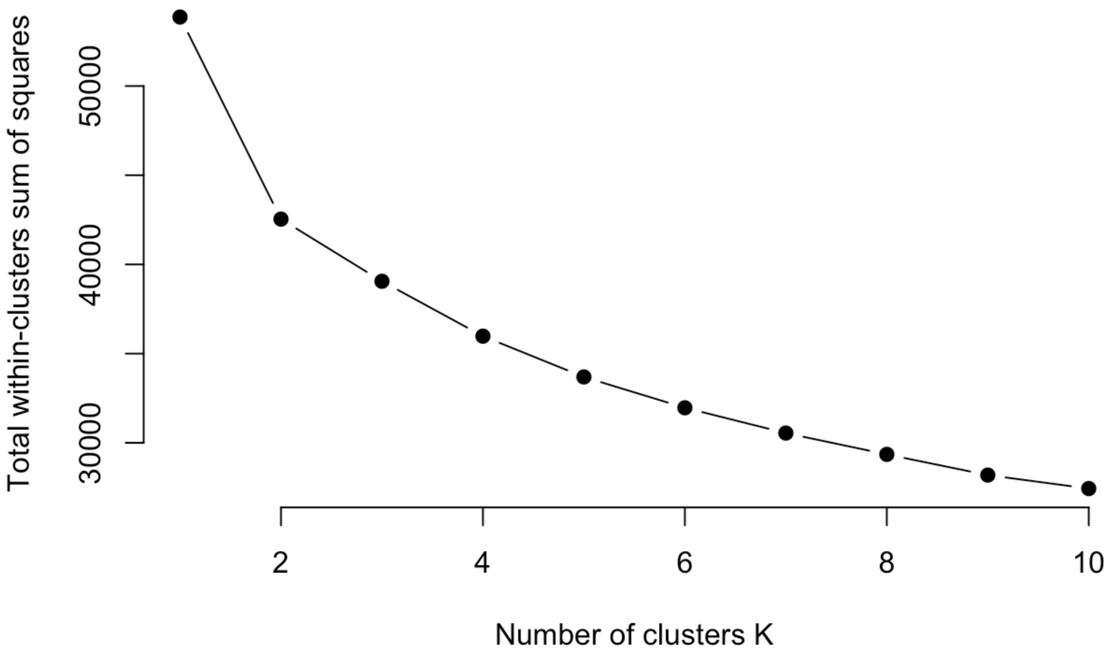
The purpose of this plot is to aid in tuning the mtry parameter of the Random Forest model to optimize its accuracy. By evaluating how changes in this parameter affect performance, one can select the most appropriate mtry value that maximizes the model's accuracy. This is crucial for model tuning in machine learning, where the goal is to find the best parameter settings that yield the most accurate predictions on new, unseen data.



```
Confusion Matrix and Statistics

          Reference
Prediction   3   4   5   6   7   8   9
         3   0   0   0   0   0   0   0
         4   0   4   0   3   0   0   0
         5   1  16 212  44   3   0   0
         6   3  10  76 367  75  10   1
         7   0   2   3  24  96  11   0
         8   0   0   0   1   2  14   0
         9   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.7086
                 95% CI : (0.679, 0.7369)
    No Information Rate : 0.4489
    P-Value [Acc > NIR] : < 2.2e-16
```

Clustering with K-Means

Clustering is performed to identify natural groupings within the wine data based on chemical properties. The quality column is removed for clustering purposes, and the data is scaled to ensure each feature contributes equally to the distance calculations. The optimal number of clusters is determined using the Elbow Method, which involves plotting the within-cluster sum of squares for different numbers of clusters and identifying the "elbow point" where the rate of decrease sharply slows. K-Means clustering is then applied with the optimal number of clusters, and the results are analyzed by calculating the mean of each variable within each cluster. Visualizations, such as bar plots of cluster means and PCA scatter plots, help to interpret the clusters. Silhouette analysis is also conducted to evaluate how well the data points have been assigned to clusters.

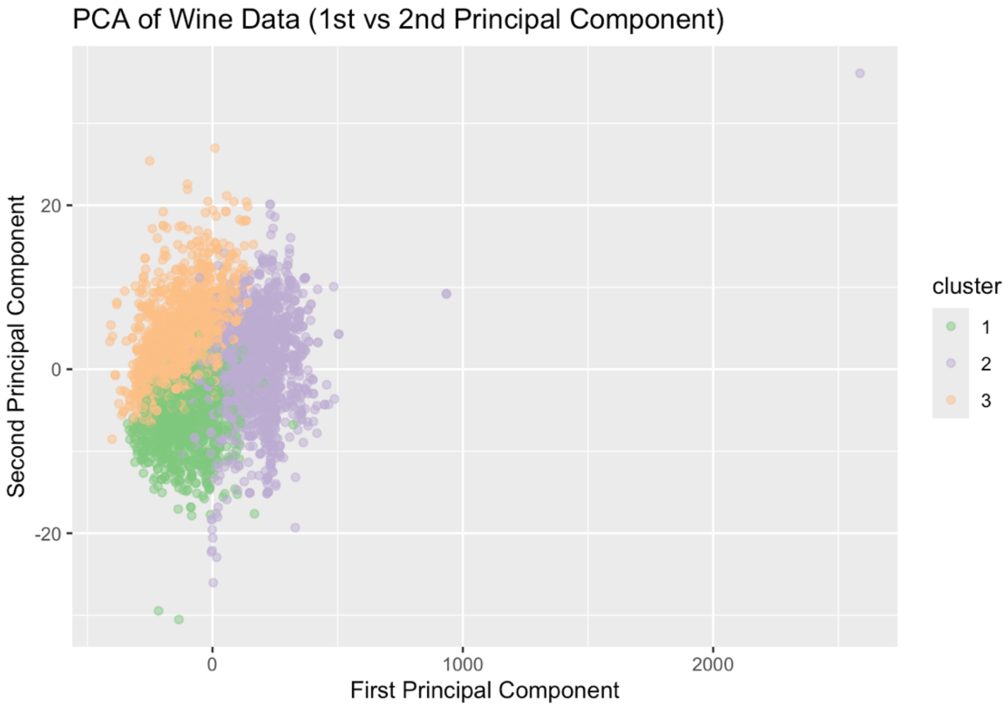


The plot displayed is known as an "Elbow Plot," commonly used in k-means clustering to help determine the optimal number of clusters to use in data analysis. This graph shows the total within-cluster sum of squares (WSS) against the number of clusters K. Each point on the graph represents the WSS value corresponding to a specific number of clusters, ranging from 2 to 10. Analysis of the Plot:

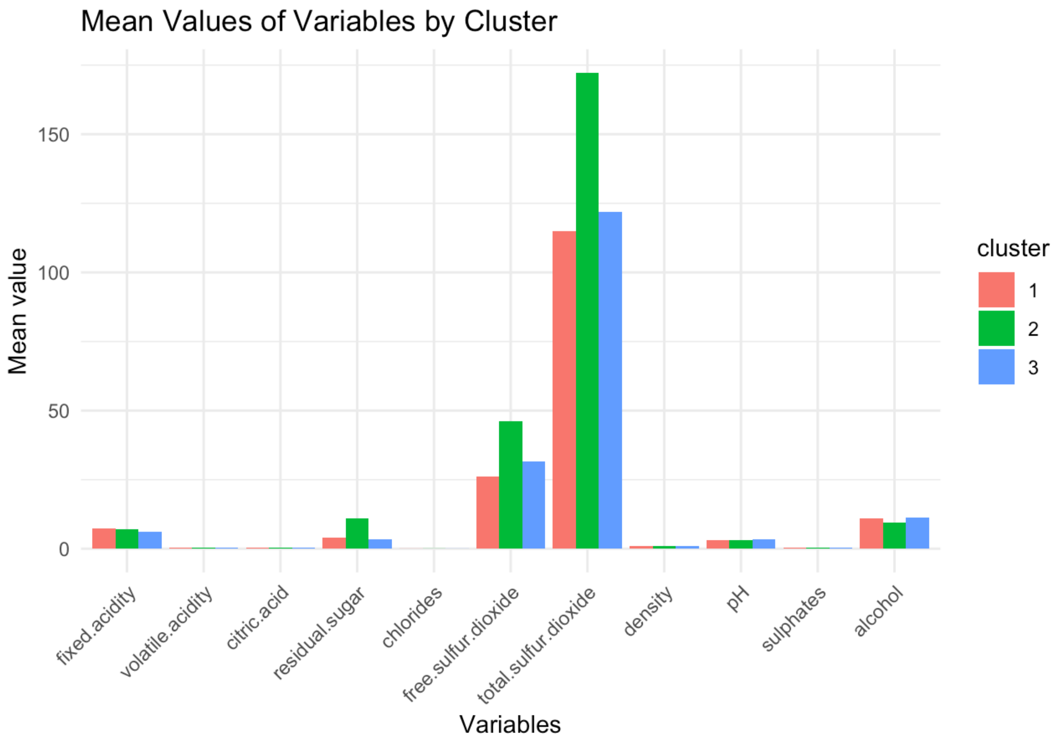
- X-axis: Represents the number of clusters (K), varying from 2 to 10.
- Y-axis: Shows the total within-cluster sum of squares (WSS), a measure of variance within each cluster.

- Data Points: Each black dot marks the total WSS for each number of clusters. The line connecting these points helps visualize the rate of decrease in WSS as the number of clusters increases. Observations from the Plot:
- Sharp Decrease and Leveling Off: The plot features a sharp decrease in WSS as the number of clusters increases from 2 to 4, indicating significant gains in cluster tightness and separation by increasing the number of clusters. After K=4, the decline in WSS slows down considerably, indicating diminishing returns in cluster separation from additional clusters.
- Elbow Point: The "elbow" of the plot, which is the point where the rate of decrease in WSS shifts from being steep to more gradual, appears to be at K=4. This point is typically considered when selecting the optimal number of clusters, as it suggests a balance between maximizing variance explained and minimizing the number of clusters. Purpose of the Plot:

The primary purpose of this elbow plot is to provide a visual tool to help in selecting the most appropriate number of clusters for k-means clustering. The elbow method looks for a point where the improvements in minimizing WSS (indicating tighter and better-separated clusters) begin to diminish. In practical terms, choosing the number of clusters at the elbow can lead to a more meaningful clustering by avoiding overfitting with too many clusters and underfitting with too few. Based on the elbow plot, selecting four clusters might be optimal for this dataset as it represents a point where additional clusters do not significantly improve the tightness of the clustering. This insight is crucial for effective cluster analysis, as it influences the granularity and usefulness of the derived clusters in subsequent data analysis and decision-making processes. Using the optimal number of clusters ensures that the clustering model is both efficient and informative, capturing the inherent structure of the data without unnecessary complexity.

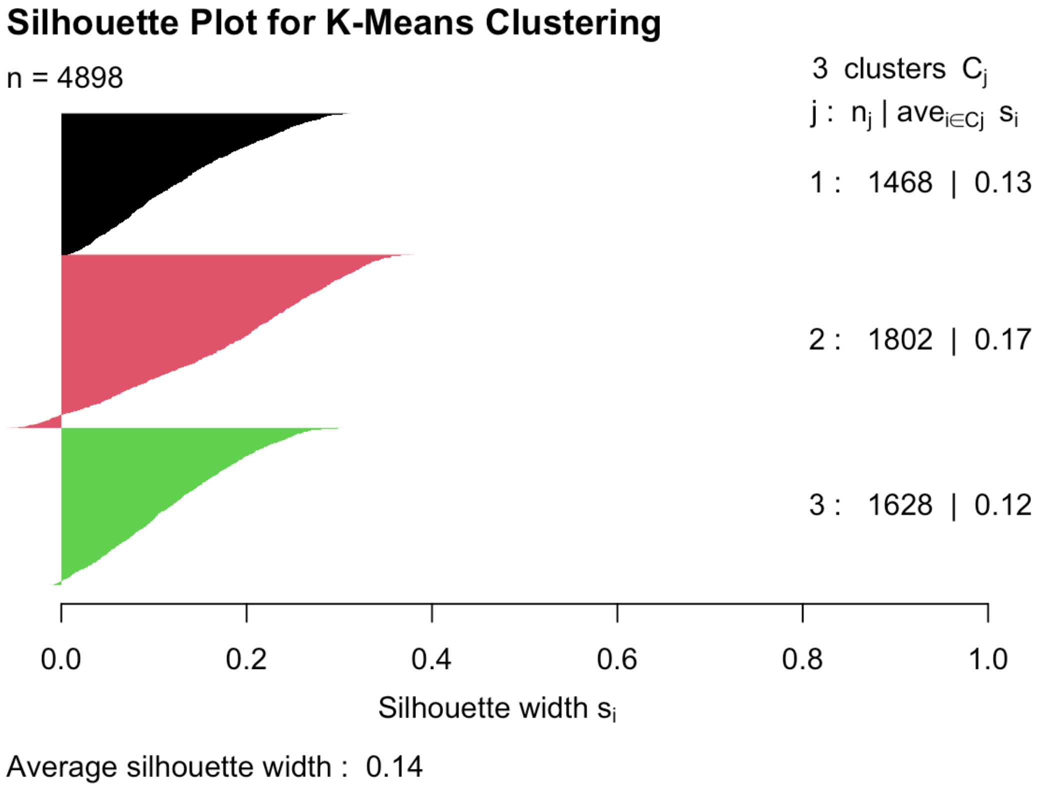


PCA of Wine Data (1st vs 2nd Principal Component)," illustrates the results of a principal component analysis (PCA) on a dataset concerning wine characteristics. The plot visualizes the wine samples in a reduced dimensional space defined by the first and second principal components, which are derived to capture the maximum variance in the dataset.

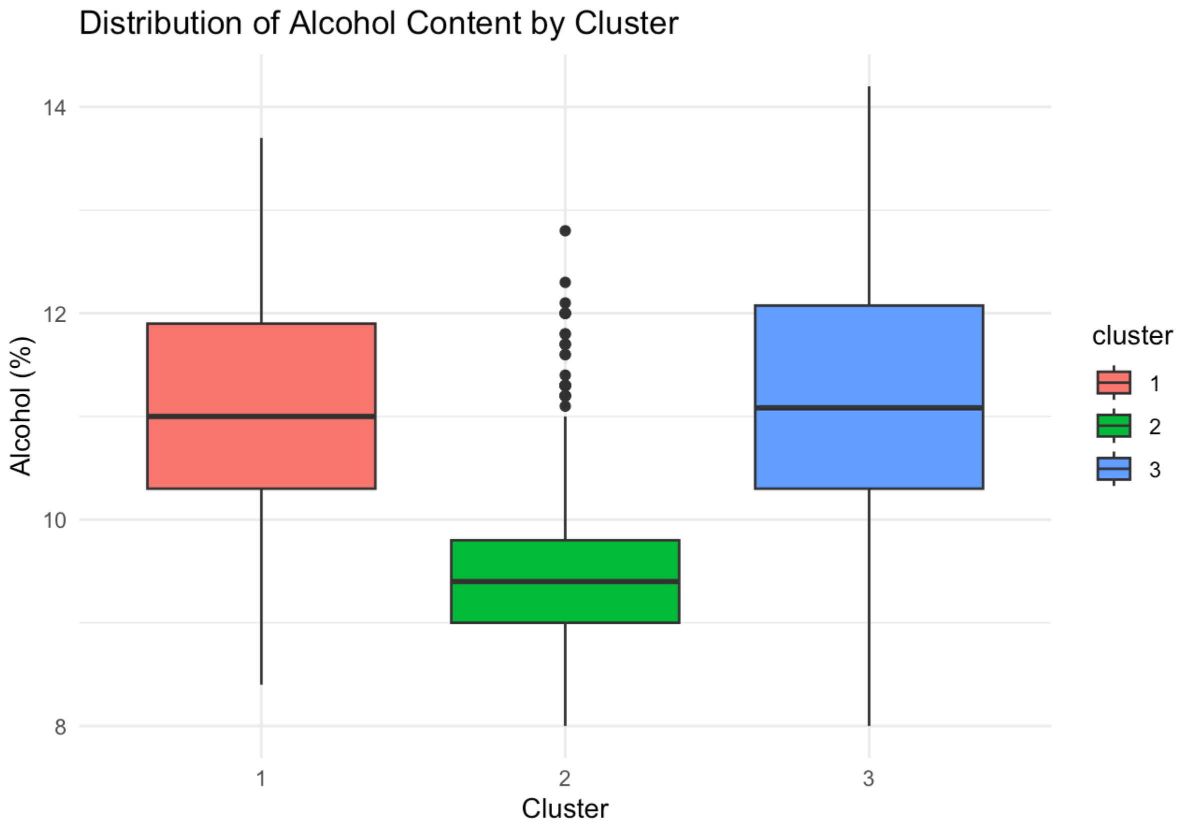


The provided plot, titled "Mean Values of Variables by Cluster," is a bar chart that displays the mean values of various physicochemical variables of wine across three different clusters. This chart allows for a visual comparison of the average levels of each variable like fixed acidity, volatile acidity, and alcohol content, among others, segmented by cluster. Analysis of the Plot:

- Variables: The x-axis lists the physicochemical properties of wine such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.
- Mean Values: The y-axis measures the mean values of these variables within each cluster, allowing us to observe which variables are more pronounced in each cluster.
- Clusters: The clusters are color-coded—red for cluster 1, green for cluster 2, and blue for cluster 3—making it easy to distinguish which cluster each bar represents. Observations from the Plot:
- Cluster 2 (Green): This cluster is notably characterized by significantly higher mean values of total sulfur dioxide compared to the other clusters. It also shows relatively high mean values for free sulfur dioxide.
- Cluster 1 (Red): Exhibits a higher mean level of residual sugar compared to the other two clusters.
- Cluster 3 (Blue): Shows a moderate level of total and free sulfur dioxide but is lower compared to cluster 2. It has the lowest mean levels of residual sugar. Purpose of the Plot: This plot is essential for understanding how different clusters identified within the wine dataset vary according to the mean levels of important physicochemical properties. Such an analysis is crucial for:
- Identifying Patterns: Understanding how different physicochemical characteristics are distributed across the clusters can help in identifying patterns or profiles typical of each cluster.
- Profile Analysis: For example, a cluster with high sulfur dioxide might correspond to wines that are preserved longer, while another with high residual sugar levels might correspond to sweeter wines.
- Targeted Actions: The insights derived from such a plot can guide specific actions, such as tailoring marketing strategies to different segments of wine consumers or adjusting production techniques to enhance wine quality.
- Data Insights: It provides a comprehensive overview of the dataset in a clustered format, making it easier to handle complex multivariate data for further statistical analysis or reporting. In summary, the "Mean Values of Variables by Cluster" plot is a powerful visualization tool that summarizes complex data into actionable insights, facilitating deeper understanding and targeted analysis of wine characteristics grouped by similar properties.

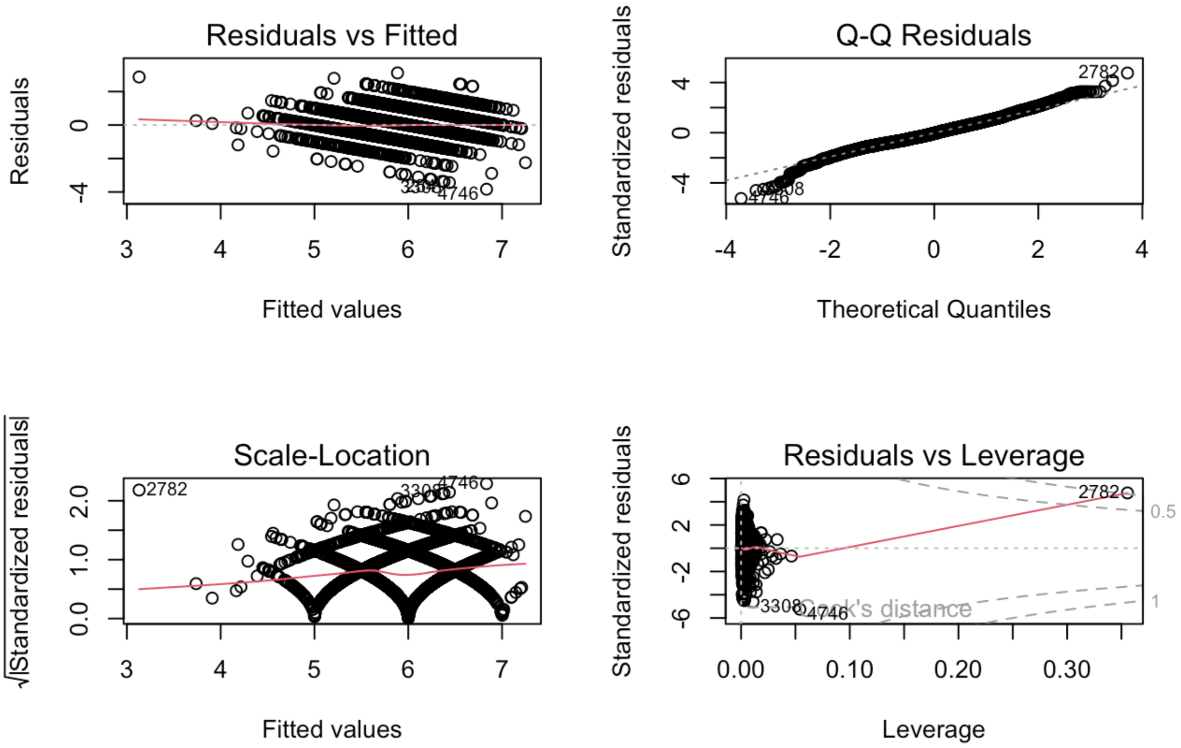


This silhouette plot indicates that the clusters are reasonably well-defined but not ideal, as indicated by the average silhouette width of 0.14. Cluster 2 shows the highest average silhouette width, suggesting that it is the most distinct and well-defined among the three clusters, whereas Clusters 1 and 3 have lower values, indicating a lesser degree of separation from neighboring clusters. This type of analysis is critical for confirming the appropriateness of the clustering approach and for refining the clustering parameters.



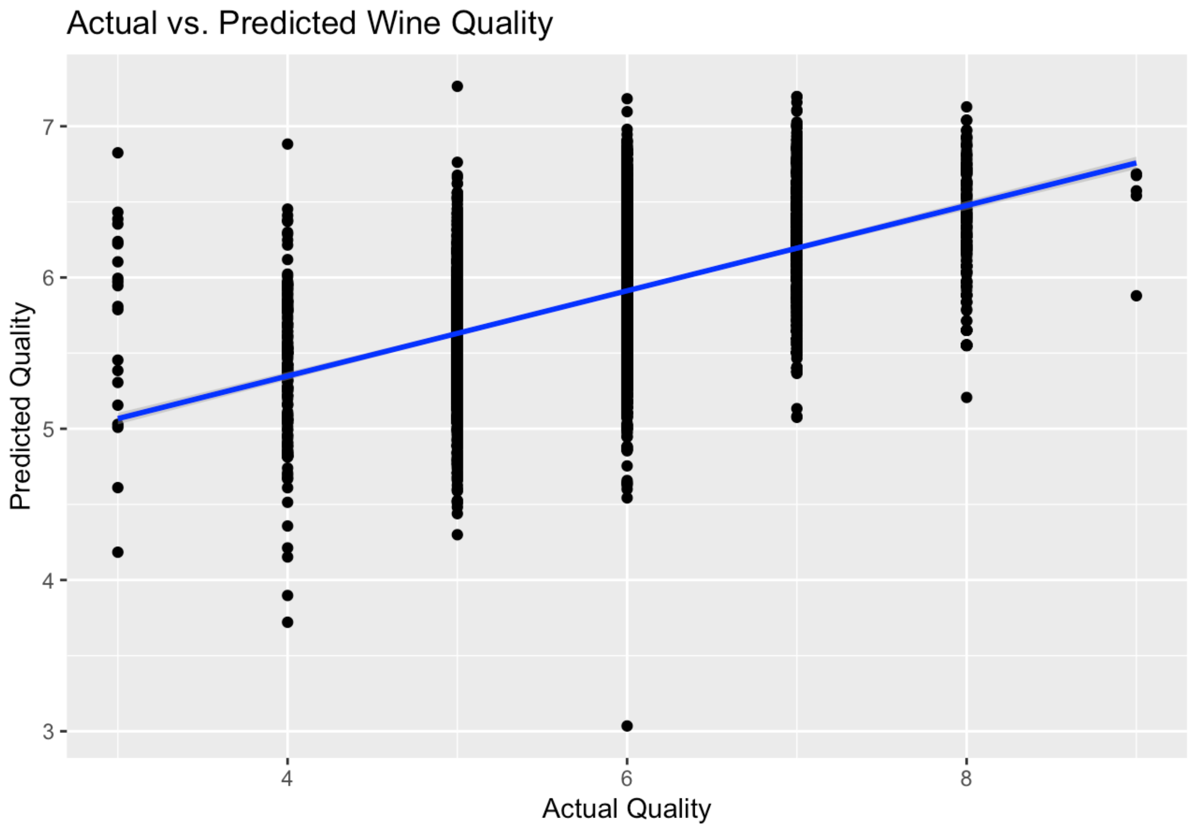
Linear Regression

Linear regression is used to model the relationship between the physicochemical properties of the wine and its quality. The data is first checked for missing values, scaled, and then a full linear regression model is fitted using all predictors. Diagnostic plots are generated to assess the model's assumptions and performance. Stepwise regression is applied to improve the model by selecting the most significant predictors based on the Akaike Information Criterion (AIC). Additionally, the caret package is used to train a linear regression model with 10-fold cross-validation, providing a robust assessment of model performance. Predictions are made using the stepwise model, and actual versus predicted values are plotted to visualize the model's accuracy. This analysis helps in understanding which chemical properties significantly influence wine quality and how well the model can predict quality based on these properties.



The diagnostic plots collectively suggest that while the model generally meets the assumptions of linear regression, there are indications of some issues such as potential outliers and non-normality in the residuals, particularly evident in the tails of the Normal Q-Q plot and the identified influential points in the Residuals vs Leverage plot. These diagnostics are crucial for validating the assumptions underlying linear regression and for guiding potential improvements in the model, such as transformations of variables or exclusion of outliers to achieve a more robust and reliable model fit.

```
Residual standard error: 0.7514 on 4886 degrees of freedom
Multiple R-squared:  0.2819,     Adjusted R-squared:  0.2803
F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```



This plot is used to visually assess the accuracy and reliability of the predictive model used for estimating wine quality. Key purposes include:

Model Evaluation: It helps in evaluating how well the predictive model corresponds to actual outcomes, with the trend line providing a quick visual assessment of overall model accuracy.

Identification of Patterns: Observing the spread and density of points can reveal biases or systematic errors in the model, such as consistently overestimating or underestimating the quality for certain ranges of actual quality.

Outlier Detection: Identification of outliers, where predictions greatly differ from actual values, can indicate data points that are either anomalous or potentially influential in model training, warranting further investigation.

In summary, this plot is essential for analyzing the performance of a regression model used for predicting wine quality. It provides a clear visual representation of the relationship between predicted and actual values, facilitating the identification of trends, biases, and outliers in the model's predictions. This analysis is crucial for refining the model and improving its predictive accuracy in future iterations.

The comprehensive analysis encompassing various statistical and machine learning techniques on the wine quality dataset has provided insightful findings that elucidate the characteristics influencing wine quality and the effectiveness of predictive modeling. Through detailed exploratory data analysis, regression modeling, clustering, and principal component analysis, the report has identified significant patterns and relationships within the dataset.

1. Exploratory Data Analysis (EDA): Initial EDA revealed key variables affecting wine quality and established baseline understandings, such as the influence of alcohol content and acidity on the quality ratings. Visualizations such as box plots and variable importance plots highlighted the critical predictors within the dataset.
2. Regression Analysis: The regression models, particularly the linear regression, demonstrated the capability to predict wine quality based on physicochemical properties. However, the residual plots indicated the presence of some outliers and potential heteroscedasticity, suggesting the need for further model refinement or the exploration of non-linear models.
3. Clustering: The application of k-means clustering, supported by the elbow method and silhouette analysis, effectively segmented the wines into distinct groups based on their chemical properties. This classification has potential implications for market segmentation and targeted wine production strategies.
4. PCA Analysis: Principal Component Analysis effectively reduced the dimensionality of the dataset, providing a clear visualization of the clustering and revealing the inherent structure of the data. The PCA plot underscored the separation between clusters, validating the clustering approach and offering a graphical representation of the dataset's complexity.
5. Model Evaluation**: Throughout the analyses, various models were evaluated using appropriate metrics. The actual vs. predicted plots for the regression models provided a straightforward assessment of model accuracy, showing good predictive performance but also room for improvement in handling outliers and extreme values. The findings from this analysis underscore the complexity of wine quality assessment and the potential of analytical approaches to uncover underlying patterns. While the models used have shown good predictive power, there are opportunities for improvement, particularly in addressing outliers and enhancing model robustness. Further research could explore more sophisticated machine learning models, such as ensemble methods or neural networks, which might capture the nuances of the data more effectively. Additionally, the clustering results can be leveraged by winemakers and marketers to tailor their products more closely to consumer preferences or to optimize the wine-making process for specific clusters of wine. Overall, this analysis not only provides a foundation for further research but also offers practical insights that could inform both production and marketing strategies in the wine industry.