ASSIGNMENT 6

Wine Chemical Analysis Using Hierarchical Clustering

This dataset is adapted from the Wine Data Set originally available on the UCI Machine Learning Repository. It has been modified for unsupervised learning. The dataset contains results from a chemical analysis of wines produced in a specific region of Italy, derived from three different cultivars. Each record represents a unique wine sample, analyzed for 13 distinct chemical constituents that contribute to the wine's characteristics and quality.

Attributes

1. **Alcohol**: The percentage of alcohol in the wine, indicating its strength and flavor profile.
2. **Malic Acid**: A natural organic acid found in wine, contributing to its tartness and overall acidity level.
3. **Ash**: The residue remaining after the combustion of the wine sample, representing the inorganic material present, which can affect taste and quality.
4. **Alcalinity of Ash**: A measure of the alkaline content of the ash, influencing the wine's pH and taste profile.
5. **Magnesium**: The concentration of magnesium in the wine, which can impact fermentation and flavor complexity.
6. **Total Phenols**: The total amount of phenolic compounds, which contribute to the wine's color, flavor, and mouthfeel.
7. **Flavanoids**: A subset of phenolic compounds known for their antioxidant properties and influence on flavor and color.
8. **Nonflavanoid Phenols**: Phenolic compounds that are not classified as flavonoids, contributing to the wine's overall profile.
9. **Proanthocyanins**: Compounds responsible for astringency and color in red wines, affecting both flavor and aging potential.
10. **Color Intensity**: A measure of the depth of color in the wine, indicating its richness and potential aging ability.
11. **Hue**: The overall color quality of the wine, influenced by its chemical composition.
12. **OD280/OD315 of Diluted Wines**: The optical density measured at two wavelengths (280 nm and 315 nm), used to assess the concentration of phenolic compounds in the wine.
13. **Proline**: An amino acid present in wine that can influence flavor and the wine's aromatic profile.
14. **Alcohol_Level**: This categorical variable classifies the alcohol content in the wine into three distinct levels, providing insights into its strength and potential impact on flavor:
    - Low" indicates wines with an alcohol content below 12.5%, typically lighter in body and flavor.
    - "Medium" represents wines with an alcohol content between 12.5% and 13.5%, offering a balanced profile in terms of strength and taste.
    - "High" includes wines with an alcohol content above 13.5%, usually stronger and more full-bodied.
15. **Ash_Content**: A categorical variable describing the ash content level in the wine. It is classified based on the amount of inorganic residue remaining after combustion:
    - "Low" indicates an ash content below 2.3, associated with wines that may have a lighter mineral profile.
    - “Moderate" represents an ash content between 2.3 and 2.6, indicating a typical balance of mineral content.
    - "High" signifies an ash content above 2.6, potentially contributing to a richer taste and greater complexity.
16. **Color_Intensity_Group**: This variable categorizes the depth of the wine's color, which can be an indicator of its richness and aging potential:

- "Light" corresponds to color intensity levels below 4.0, often found in lighter-bodied wines with a more delicate appearance.
- "Moderate" indicates color intensity between 4.0 and 6.0, suggesting a well-balanced visual depth.
- "Dark" refers to color intensity levels above 6.0, typically found in wines with a richer, fuller appearance, often suggesting greater aging potential and concentration.

Relevant libraries: e1071, dplyr, tinytex, cluster, dbscan, fpc, ggplot2.

Answer the following Three questions (7 parts)

Question 1: Hierarchical Clustering Using AGNES and DIANA

Q1-A) [30 Points] Use the wine dataset to perform hierarchical clustering using both AGNES and DIANA.Plot dendrograms for both AGNES and DIANA methods.

Hint:

#First exclude categorical variables and use scale() to normalized your data.

```
agnes() #compare single, complete, ward, and average.
print(Hierarchical.single$ac)
pltree()
rect.hclust()
```

#Do the same thing using diana()

Q2-B) [20 Points] Compare the agglomerative coefficient (AC) of AGNES with the divisive coefficient (DC) of DIANA. Which method provides a more cohesive clustering structure based on the coefficients?

~~~~~~~~~

Question 2: Visualize Categorical Variables

Q2-A) [30 Points] For each of the categorical variables, namely, Alcohol_Level, Ash_Content, Color_Intensity_Group, calculate the median of the corresponding numeric variable (Alcohol, Ash, Color_Intensity).

Hint: `Median.Recommendation <- wines %>% group_by(Alcohol_Level) %>% summarise(Median_Alcohol = median(Alcohol, na.rm = TRUE))`

Q2-B) [30 Points] Create bar plots to visualize these median values across different levels of the categorical variables. Use ggplot2 to plot these visualizations with appropriate labels and titles.

Hint: `ggplot(Median.Recommendation, aes(x = Median_Color_Intensity, y = factor(Color_Intensity_Group) , fill = factor(Ash_Content), …)`

Q2-C) [20 Points] Interpret the bar plots: What do the median values suggest about the distribution of the numeric variables across the different categories? What insights can you derive regarding the characteristics of the wines?

~~~~~~~~~

Question 3: Determine Appropriate Clustering Technique

Q3-A) [20 Points] Considering the results from AGNES and DIANA, which hierarchical clustering technique is more appropriate for analyzing the wine dataset?

- Compare the results obtained from AGNES and DIANA, focusing on the number of clusters formed and the quality of separation.
- Discuss any meaningful patterns or insights observed in the clusters.

Q3-B) [20 Points] Would another clustering approach, such as K-Means or DBSCAN, be more suitable for this dataset? Explain your reasoning based on the dataset properties and clustering results.

Note that Explanations can be included in HTML as comments using #. Also, there is no unique way to develop your code; provided partial codes (pseudocodes) just serve you as hints and you don't need to use them.

- ❖ Submit your report in HTML format using R-Markdown. In your answers, please include your corresponding questions numbers, namely, Q1-A, Q1-B, etc.