Deliverable 1: Dataset Selection, Business Question, and Variable Explanation

STUDENT'S NAME

Vandana Rathore
Aditya Pradeep Waghmode
Vinay Anand Kuppurao

Based on the [wine quality dataset](#), we can formulate a prediction and a classification type business question:

==**Prediction Question:**== "How accurately can we predict the overall quality score of a wine based on its chemical attributes, such as acidity levels, alcohol content, and residual sugar?"

==**Classification Question:**== "Can we classify wines into different quality categories (e.g., low, medium, high) based on their chemical attributes, and what attributes are the most influential in making this classification?"

Plan for Answering the Questions:

1. For the prediction question, we will use regression techniques, such as linear regression, to predict the numerical wine quality score. To predict the overall quality score of wine accurately using the Wine Quality dataset, we'll need to consider which variables are crucial for answering this prediction question. The dataset provides various chemical attributes of wines and their associated quality scores, which range from 3 to 9. Here are the variables that might be crucial for predicting wine quality:

   a) **Fixed Acidity**: The fixed acidity level can influence the taste and stability of the wine. It's an essential variable to consider for predicting quality.
   b) **Volatile Acidity**: High volatile acidity levels can lead to off flavors, so this is crucial for assessing wine quality.
   c) **Citric Acid**: Citric acid can contribute to the freshness and flavor of the wine. It may positively impact wine quality.
   d) **Residual Sugar**: The amount of residual sugar can affect the sweetness of the wine. Different wine styles require various levels of residual sugar.
   e) **Chlorides**: Chloride content can influence the taste, with high levels potentially making the wine taste salty. This can affect the overall quality score.
   f) **Free Sulfur Dioxide and Total Sulfur Dioxide**: Sulfur dioxide is used as a preservative and can impact the wine's aroma and taste. Balancing these levels is essential for wine quality.
   g) **Density**: Density is related to the concentration of alcohol and sugar, affecting the wine's body and mouthfeel.
   h) **pH**: The pH level influences the wine's stability and taste, making it an important variable to consider.
   i) **Alcohol**: Alcohol content can significantly impact wine quality. Finding the right balance is key.
   j) **Sulfates**: Sulfates can act as antioxidants and antimicrobial agents, impacting the wine's aroma and taste.

These chemical attributes are vital for predicting wine quality, as they are key factors that influence the overall taste, aroma, and balance of the wine.

2. For the classification question, we will employ classification algorithms such as logistic regression to categorize wines into red wine or white wine based on their characteristics. Here are the variables that could be crucial for answering this classification question: -

   (a) **Fixed Acidity**: Red and white wines can have different levels of fixed acidity, with red wines typically having slightly higher values. This could be a relevant feature for classification.
   (b) **Citric Acid**: Citric acid levels can vary between red and white wines, with white wines generally having higher citric acid content. This feature can contribute to the classification.
   (c) **Residual Sugar**: White wines typically have more residual sugar than red wines. This difference in residual sugar content can be a significant distinguishing factor.
   (d) **Chlorides**: Chloride levels can differ, and this might be a useful feature for classification. Red wine often has higher chloride levels.

Variable Explanation: In the wine quality dataset, there are several variables, each with a specific role in the wine quality assessment. Here are explanations for some of the crucial variables:

Input variables (based on physicochemical tests):

1. **Fixed Acidity:** This variable represents the fixed acidity of the wine and is measured in grams per liter (g/dm³). It is an important factor in wine taste and affects the overall quality. Its values typically range from around 4 to 16 g/dm³.
2. **Volatile Acidity:** Volatile acidity is another measurement of acidity, but it indicates the amount of acetic acid in the wine. It is also measured in g/dm³ and plays a crucial role in wine quality. Values typically range from 0.12 to 1.58 g/dm³.
3. **Citric Acid:** Citric acid content in g/dm³ provides freshness and flavor to the wine. It can positively influence the wine quality, with values ranging from about 0 to 0.76 g/dm³.
4. **Residual Sugar:** Residual sugar represents the amount of sugar left in the wine after fermentation. It is measured in g/dm³ and significantly impacts the wine's sweetness and quality. Values can vary from 0.9 to 15.5 g/dm³.
5. **Chlorides:** This variable measures the amount of salt in the wine, expressed in g/dm³. It can affect the wine's overall taste and is usually found in the range of 0.01 to 0.61 g/dm³.
6. **Free Sulfur Dioxide:** Free sulfur dioxide, measured in parts per million (ppm), is an important preservative in wine. It influences the wine's stability and quality. Typical values range from 1 to 72 ppm.
7. **Total Sulfur Dioxide:** Similar to free sulfur dioxide, this variable measures the total sulfur dioxide content in ppm. It is an essential factor for wine quality and preservation, with values ranging from 6 to 289 ppm.
8. **Density:** Density is measured as the density of the wine in g/cm³. It can indicate the wine's alcohol content and sweetness. Values are generally between 0.99 and 1.04 g/cm³.
9. **pH:** This variable measures the pH level of the wine, affecting its taste and stability. It typically ranges from 2.74 to 4.01.

10. **Sulphates:** Sulphates, measured in g/dm³, are additives that can contribute to wine quality and preservation. Values range from about 0.33 to 2.00 g/dm³.
11. **Alcohol:** The alcohol content of the wine is given as a percentage, which can significantly influence the wine's body, mouthfeel, and overall quality. Typical values range from 8.4% to 14.9%.

Output variable (based on sensory data):

12. **Quality:** This is the output variable, representing the overall quality score of the wine, given on a scale from 0 to 10. Higher values indicate better wine quality.

Understanding these variables is crucial for analyzing and modeling the wine quality dataset, as they play a significant role in assessing and predicting the quality of the wines based on their physicochemical characteristics.