Data Analytics Assignment- Wine Data analysis

We need to do the following with respect to the dataset provided

- Disclosing the wine quality dataset
- Analyzing red wine
- Analyzing white wine

https://archive.ics.uci.edu/ml/datasets/wine+quality

Questions

1. Load the pandas library and create two different data frames, namely, df_red for holding the red wine dataset and df_white for holding the white wine dataset.

2. Find the name of the available columns.

3. Let's see some sample data from the red wine data frame. check the entries between the 100th and 110th rows.

4. Check the data types for each column.

5. Describe the data frame to get more descriptive information.

6. Find which of the columns have null values.

7. We will continue analyzing the red wine dataset. First, we will start by exploring the most correlated columns. Second, we will compare two different columns and observe their columns.

First, start with the quality column of red wine. Plot a graph of your choice to do the same.

8. Let's next find out which of the columns from the red wine database are highly correlated.

You need to confirm whether the following is true

- Alcohol is positively correlated with the quality of red wine.
- Alcohol has a weak positive correlation with the pH value.
- Citric acid and density have a strong positive correlation with fixed acidity.
- pH has a negative correlation with density, fixed acidity, citric acid, and sulfates.

Use pair plot, heatmap, etc to draw these correlations.

9. Let's see how the quality of wine varies with respect to alcohol concentration. Use a boxplot for the same.

Can you draw this conclusion: The higher the alcohol concentration is, the higher the quality of the wine.

10. Let's also see the correlation between the alcohol column and pH values. Use joint plot and Pearson regression for the same.

11. Create a correlation function that can give a correlation between any two columns. Name it get_correlation. Now find the correlation between alcohol and PH using this function.

12. Let's start with white wine analysis now. Load the white wine data frame.

13. Find the average quality of both red wine and white wine.

14. Let's add a new attribute, wine_category, to both data frames.

15. let's find out what are the unique values of the column quality in both types of wines.

16. Although the quality column is numerical, here, we are interested in taking quality as the class. To make it clear, let's convert numerical values into categorical values.

17. To do so, we need a set of rules. Let's define a set of rules:

quality_label - low if value <= 5, medium if value>5 and value <=7 and high if value > 7

Write the code to add this quality_label column using the rule described above.

18. Count the number of values in each category of wine for column quality_label.

19. Concatenate both red and white wine data frames now.

20. Reshuffle the rows for randomization.

21. Let's use the combined data frame and group them using the columns, alcohol, density, pH, and quality.

Hint: - create a subset of attributes that we are interested in. Then, create three different data frames for low-quality wine, medium-quality wine, and high-quality wine. Finally, concatenate them.

22. How will you do a univariate analysis for numerical data . Apply the logic to do univariate analysis.

23. Do a multivariate analysis for all the columns using heatmap.

24. Draw a count plot for wine_category.