Assignment 1: Data Pre-processing

Objective

The objective of this assignment is to enable you to build and train skills in data collection, exploration and preprocessing.

Tasks

Load the data

1. Load wine data from the two source files `winequality-red.xlsx` and `winequality-white.xslx`, which you can find in the Data Science repository on Github: https://github.com/datsoftlyngby/soft2022spring-DS/tree/main/Data.
2. Join the two files in one, but keeping the identity of each source file by adding a column "type", in which stays either "red" or "white".

Explore the data

3. Explore the general parameters of the new file:
   - number of rows and columns
   - type of data in each column
   - descriptive statstics of the numeric data (count, mean, min, max, std, quantiles)
4. Plot diagrams that visualize the differences in red and white wines. What do your diagrams show? Can you tell which type of wine has higher average quality? Which type of wine has higher average level of alcohol? Which one has higher average quantity of residual sugar?
5. Split the data into five subsets by binning the attribute pH. Identify the subset with the highest density? What if you split the data in ten subsets?
6. Use the function `corr()` to create a correlation matrix of all data and investigate it. Can you tell which vine attribute has the biggest influence on the wine quality. Do you get the same results when you analyze the red and white wine data sets separately?

Prepare the data for further analysis

7. Search the data for allocating
   a. missing values
   b. cells with a value of '0' (zero)

   Replace these values with the average value of their column.
8. Explore the feature 'residual sugar'. Is there any outlier (a value much different from the rest)? On which row is it found? Remove that row.
9. Identify the attribute with the lowest correlation to the wine quality and remove it.
10. Finally, transform categorical data into numeric and print out the start and the end of the preprocessed data frame.

Note

We recommend using IPython Jupyter notebook format.
Use as many and different diagrams as you find appropriate.
You can submit this assignment as a teamwork.
Upload it to Peergrade and/or show it in class by 1st March.

Have fun!
the instructors