Mini Project 2: Data Exploration and Engineering

Objective

The objective of this assignment is to enable you to build and train skills in data exploration and analysis by applying methods from statistics.

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/dat2024spring-bi/tree/main/data.
2. Clean the data in both files.
3. Aggregate the two files in one still keeping the identity of each wine type - “red” or “white”.

Explore the data

4. Explore the features of the original and the new files:
   a. number of rows and columns
   b. type of data in each column
5. Calculate the descriptive statistics of the numeric data. Is the data normally distributed?
6. Plot diagrams that visualize the differences in red and white wine samples. Use it as a support for answering the following questions:
   a. what exactly is shown on the diagrams?
   b. after seeing it, can you tell which type of wine has higher average quality?
   c. which type of wine has higher average level of alcohol?
   d. which one has higher average quantity of residual sugar?
7. Which other questions might be of interest for the wine consumers or distributers?
8. Split the aggregated 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?
9. Create a heat map or a correlation matrix of all data and investigate it. Can you tell which vine attribute has the biggest influence on the wine quality? Which has the lowest?

   Do you get the same results when you analyze the red and white wine data sets separately?

Prepare the data for further analysis

10. 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.
11. Identify the attribute with the lowest correlation to the wine quality and remove it.
12. Transform the categorical data into numeric.
13. Try to reduce the number of features of the aggregated data set by applying principal component analysis (PCA). What is the optimal number of components?
14. Print out ten random rows from the final dataset as a prove of concept.

Notes

Use as many and different diagrams, as they are appropriate.

You can develop and submit this assignment as a teamwork and get support from the instructor at the workshop on 13/02/24.

The deadline for submission is 19/02/24, 12:00.

Have fun!

the instructor