Mini Project 2: Data Exploration and Engineering

Objective

The objective of this assignment is to enable you to build and train skills in business 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`.
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. Check if the values of the attributes are normally distributed.
6. Plot diagrams that visualize the differences in red and white wine samples. Use as many diagrams as appropriate. Use the diagrams as a support for answering the following questions:
   a. what do diagrams show exactly?
   b. which type of wine has higher average quality, how big is the difference?
   c. which type of wine has higher average level of alcohol?
   d. which one has higher average quantity of residual sugar?
   e. do the quantity of alcohol and residual sugar influence the quality of the wine?
7. Which other questions might be of interest for the wine consumers and which of wine distributers?
8. Split the aggregated data into five subsets by binning the attribute pH. Which subset has 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 wine attribute has the biggest influence on the wine quality? Which has the lowest? Are there any attributes, apart from the wine quality, which are highly correlated?
10. Do you get the same correlation results when you analyze the red and the white wine data sets separately?

Prepare the data for further analysis

11. Explore the feature ‘residual sugar’. Does it contain outliers? On which rows of the data frame are they found? Remove those rows.
12. Remove the attributes with the lowest correlation to the wine quality and any one highly correlated to another independent attribute.
13. Transform the categorical data into numeric, applying appropriate coding methods.
14. Print out ten random rows from the final dataset as a prove of concept.

Note

The project brings 20 study points.

Have fun!

the instructor