Mini Project 2: Data Exploration and Visualisation

Objective

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

You will be exploring the domain of wine quality.

Tasks

Load and Clean the Data

1. Load wine data from the two provided source files: `winequality-red.xlsx` and `winequality-white.xslx` into Python data frames.
2. Clean the data in both.
3. Aggregate the two sources into one, still keeping the identity of each wine sample's type - "red" or "white".

Explore the Data

4. Explore the features of the three data frames separately. Identify the dependent and the independent variables
5. Transform the categorical data into numeric, applying appropriate encoding methods.
6. Calculate the descriptive statistics of the numeric data. Check whether the distribution of the values of the attributes is normal.
7. 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 does each diagram show?
   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?
8. Discuss which other questions might be of interest for the wine consumers and which of wine distributers.
9. 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?
10. Create a correlation matrix and a heat map of all data and investigate it. 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?

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, which aren't correlated with the wine quality, as well as the attributes that are highly correlated with another independent attribute.
13. Transform the data by applying PCA (Principle Component Analysis).
14. Print out ten random rows from the final dataset as a prove of concept.

Create and Deploy Interactive Application

15. Use Streamlit to build an application, which allows interactive data loading and visualization of the exploratory analysis. Apply both 2D and 3D data visualization techniques.
16. The quality of wine is a complex category that also depends on multiple non-numeric parameters, such as geographical origin, the production technology, and human taste. Extend your application with useful public information, related to the quality of wine. Use either the web, Wikipedia, or Youtube as a source. Implement LLM for text summarization, if necessary.

Note

This is a group project. It brings 30 study points.

Have fun!

the instructor