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 - a complex category that depends on multiple numeric and non-numeric parameters, such as content of alcohol and sugar, flavor, geographical origin, production technology, and human taste. The goal is to reveal insights explaining these dependencies.

Tasks

Load and Clean the Data

1. Load wine data from the provided in `wine-data.zip` archive source files into Python data frames.
2. Clean the data in both, applying the techniques learned earlier.
3. Aggregate the two sources into one, still keeping the identity of each wine sample's type - "red" or "white".
4. Search the web for alternative public data sources in the wine quality domain, such as documents, images, video. Ingest and store the found content into your local data lake for further processing and analysis.

Explore the Data

5. Explore the features of the three data frames from pp.1-3 above. Identify proper dependent and the independent variables. Visualise the findings of the exploration as appropriate.
6. Transform the data as necessary to prepare it for exploratory analysis and visualization. Transformations may include encoding the categorical data into numeric and discretization of the continuous data.
7. Calculate the descriptive statistics of the numeric data. Check whether the distribution of the values of the attributes is normal.
8. 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?
    9. Discuss which other questions might be of interest for the wine consumers and which of wine distributers.
10. 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?
11. Search for correlation between the normally distributed dependent and independent variables. Create a correlation matrix and a heat map and explore 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

12. Explore the features for outliers? Which feature contains outliers? Which of the observations of this feature is outlier? Remove those observations.
13. Remove the attributes, which aren't correlated with the wine quality, as well as the attributes that are highly correlated with another independent attribute.
14. Apply data transformation as preparation for further analysis:
    a. scaling, normalization and standartisation of the numeric variables
    b. dimensionality reduction by PCA (Principle Component Analysis).

Create and Deploy Interactive Application

15. Use Streamlit to build an application, which allows interactive data loading, analysis, and visualization of the outcomes. Apply various 2D and 3D data visualization techniques.
16. Extend your application with useful information, extracted from the alternative public sources, collected in p.4 above. Apply AI techniques for operating with text and visuals.

Note

This is a two-week group project. The solution brings 30 study points to each contributer.

Have fun!

the instructor