Mini Project 1: EDA - Data Exploration and Visualisation

Objective

The objective of this assignment is to enable you to build and train skills in business data exploration and visualization 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, and human taste. The goal is to reveal insights explaining these dependencies.

Tasks

Problem Formulation and Data Loading

1. Load wine data from the provided in `wine-data.zip` archive source files into Python data frames and get familiar to it – for example, see how much data is available, what does it contain, which are the types of the attributes, etc.
2. Formulate some expectations and hypotheses about the red and white wine that can be taken out of this data. Who could be the recipients of the analysis? Think of the needs of the wine producers, wine distributers and wine consumers – what could be useful for them to know?

Data Cleaning and Transformation

3. Clean the data from both sources, applying the wrangling techniques learned earlier, such as
   - restoring the missing data, repairing the wrong values, removing duplicates
   - converting attribute types as necessary, for example, encoding the categorical data into numeric or discretisation of continuous data
   - creating new features, removing features, renaming and reordering the data frames
4. Aggregate the two sources into one, still keeping the identity of each wine sample's type - "red" or "white". Explore the aggregated data.
5. Search the web for alternative public data sources in the wine quality domain, such as APIs, documents, images, video. Ingest and store the found content into your local data lake for further processing and analysis.
6. By this point, you have minimum three data frames: one for red wine, one for white wine, and one aggregating both. Explore and identify proper dependent and the independent variables of interest for future analysis.

Data Exploration and Analysis

7. Analyze your data by exploration of the statistical measures of the samples. Check whether their attributes are normally distributed. Compare the distributions, the descriptive statistics and measures of the three data sets and get to know how they are similar or different.
8. Examine the features for outliers. Check if removing the outliers improves the distribution and the statistics.
9. Plot diagrams that visualize the process of exploration. Show the differences and similarities between red and white wine. 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?

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? Do you get more information from it?
11. Discuss which other questions might be of interest for the wine producers, consumers and distributers. Take notes of your discussion.
12. 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?
13. Ensure you have checked for and 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. Ensure you have applied data normalisation techniques as appropriate, try scaling, normalization and/or standardization of the numeric variables. Compare and explain the effect of the different scalers on the outliers.
15. Apply statistical methods for testing your hypotheses. Estimate statistical parameters that can provide insights for it. What do they show to you?

Summary

16. Summarise and reflect on your experience. Have the results met your expectations? Have you managed to prove your initial hypotheses? Which are your main conclusions about wine, derived from this data, which of them could be shared with the related businesses?
17. Store your code in a GitHub repository and upload a link to it in Moodle. Write your narratives in a `readme.md` file. Add also any other relevant materials commenting the wine quality you have found at external public sources, as they can be used later for extending and improving the quality of your results.

Note

This is a group project and study points assignment. The solution brings 30 study points to each contributor.

Have fun!

the instructors