COMP 333 — Lab Assignment 2

Motivation

The purpose of this assignment is develop Python code for Descriptive Data Analysis.

It builds on the lectures of Week 3–4 and handles quantitative descriptions.

Your first task is to augment each dataframe with metadata, that is, create a pandas Series `dfWithMeta` that has two (or more) entries:

- the input dataframe `df` in tidy data format;
- a pandas Series that has a specification of the data measurement type `nominal, ordinal, interval, ratio` of each variable (column) of the dataframe; and
- for each ordinal variable, a list of the values of the data type in order;

You are to write a Python function `simpleDDA` that takes as input the above Series `dfWithMeta` and produces a dataframe `DDAdescription` that contains the required results (A), (B), and (C) below for each variable (column) of the input dataframe `df`.

Assignment

Create a Jupyter notebook using Python code and any of its libraries, but especially pandas, to write and test code to carry out the Descriptive Data Analysis tasks below in (A)–(C).

- Write your own Python function `simpleDDA()` within the notebook, and illustrate their use within the notebook.
- Structure your code and document your work.
- Test your code on at least three examples of datas. Taken together, these test examples must include at least example of each type of data measurement: `nominal, ordinal, interval, ratio`.

Organize your notebook to clearly separate & identify the work on parts (A), (B), and (C).

Show your code working on at least three examples of data.

(A) Overall Descriptions (2 marks) Your function should report, for each feature,

- number of observations
- number of entries
- number of unique values amongst the entries
- number of missing entries

(B.1) Central Tendency Descriptions (2 marks) Your function should report, for each feature,

- mode, or modes, for all data types
- median, for `ordinal, interval, ratio` data types
- mean, for `interval, ratio` data types

Use `NaN` as the result for the data types that are not relevant for the median or mean.

You should check the definition of *median* carefully. Sometimes for interval and ratio types, the median is not a value in the dataset, but the average of two values. Sometimes for ordinal types, the median cannot be a value in the dataset, so it is not defined (so use `NaN`).

(B.2) Spread Descriptions (2 marks) Your function should report, for each feature,

- number of unique values amongst the entries, for nominal data types
- range: (min,max), for `ordinal, interval, ratio` data types
- IQR: Q3-Q1, for `interval, ratio` data types
- standard deviation, for `interval, ratio` data types

Use `NaN` as the result for the data types that are not relevant.

(C) Visual Descriptions (2 marks) Have your Python function produce a grid of plots. The grid is a square grid indexed by the features of the dataframe in both dimensions. Down the diagonal is a univariate plot for each feature; and each off-diagonal grid entry is a bivariate plot for the pair of features.

Remember to choose the plot that is appropriate to the type of data in the feature, or the type of data in each of the two features.

You may begin by assuming that you have only continuous data, so that the histogram is an appropriate univariate plot, and that the scatter plot is an appropriate bivariate plot. And then work from there to improve your code to consider other types of data.

Marking Scheme

A total of 10 marks wil be allocated

- (A) 2 marks
- (B.1) 2 marks
- (B.2) 2 marks
- (C) 2 marks
- Testing 1 marks
- Notebook layout and documentation 1 marks

Remember that a notebook should indicate

- your identity (name and student number),
- the task (ie the course, the assignment number, and a short description of the task),
- the source of the input,
- a description of the input (usually accompanied by a listing of the first 10 rows of each dataframe), and
- the planned output.

Deliverable

Your deliverable is the completed ipynb notebook showing all computation and output.

Remember that your notebook should clearly identify your work on parts (A), (B.1 and B.2), and (C).