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.

You are to write a Python function `simpleDDA` that takes

- the imput 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;

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 and identify your work on parts (A), (B), and (C).

(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

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

(B) 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`).

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

(C) 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.

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

Marking Scheme

A total of 10 marks wil be allocated, with 2 marks for each of

- (A)
- (B)
- (C)
- Testing
- Notebook layout and documentation

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), and (C).