Homework 1

Foundations of Data Science TA team

released: 27/02/2024; due date: 07/03/2024 (14:00)

Introduction

In this homework, you'll have a chance to get back into programming with Python. For this, we will look at data on the quality and cost of care in hospitals in the United States of America. After this homework, you should be able to

- describe data using `pandas`' built-in methods;
- identify missing values using `pandas`' built-in methods, and
- formulate hypotheses on reasons for missing values, and
- build on these hypotheses to replace missing values where possible;
- use `seaborn` for basic visualisation.

1 Getting an overview

When encountering a new data set, the first step is to get an overview. For this purpose, answer the following questions:

1. How many rows and columns are in the data?
2. How many distinct hospitals have information reported in the data set? Hint: What information do you have to consider to uniquely identify a hospital?
3. How many states are represented in the data? Does that match your expectation? Explain.
4. Which hospital is the most expensive for hip and knee procedures? How much more expensive is this hospital compared to the second most expensive?
5. Which hospital is the least expensive considering the sum of average costs of all treatment categories reported in the data? Does this low price have an effect on the quality reported? Explain your reasoning.

2 Missing data

1. Are there any values missing in the data? If yes, which column(s) are affected?

2. If there are column(s) with missing data, propose a way of handling the missing entries (for each column). Note: There is no need to actually implement the handling of missing values. Your script file still has to contain evidence of how you arrived at your answer(s), e.g. how you identified or summarised missing values.

3 Basic visualisation

1. Plot the distribution of the cost for treating pneumonia. Save the figure under `output/pneumonia-cost.png`. Don't forget to add units, axis labels and a plot title.
2. Plot the distribution of the cost for treating heart failure stratified by the quality of hospitals in this domain. Note: There are multiple ways to achieve this; you only have to make sure that all relevant information is represented in the plot. Save the figure under `output/heart-failure-cost-vs-quality.png`.

Deliverables

You are asked to submit a report (one to three pages, must be in PDF format) and the `.py` file completed by you to solve the homework. The script file should contain all steps necessary to arrive at the answer to individual questions. Do not hardcode values, which are printed to the terminal. Your answers should be computed in the script and still be correct if the data changed. The report can be organised in sections similar to this question sheet and should include all plots you are asked to create or which you decide to create to help answer a question. Where necessary, add an interpretation to the plot. All questions, for which you should explictly include the answer in your report are highlighted. The Python script submitted has to run without having to make any modifications. Do not use absolute filepaths. We provide you with a `.zip` file which contains all the necessary files and the directory structure required. You can assume that the working directory during execution of your submitted script is the `hw01/code/` directory in which `skeleton.py` is stored. Make sure that you save the plots you are asked to generate to the `hw01/output/` directory. Once you have completed the homework you can submit the same directory, including the completed script, your report, and all plots as a `.zip` file again.

Data Description

The data set you work with in this homework contains information on hospitals in the United States in which Medicare patients received treatment. The data set contains information on

- hospital name and location,
- ratings of hospital performance in general across multiple dimensions (Mortality, Safety, etc.) relative to the national average,
- and information on cost, quality and value for multiple categories of treatments (heart attack, heart failure, hip and knee surgery, pneumonia).

Python environment

To be able to solve this homework, you need to have a working Python 3.11 installation and the following packages installed in the environment in which you execute the script:

- `matplotlib` (3.8)
- `numpy` (1.26)
- `pandas` (2.2)
- `scipy` (1.12)
- `seaborn` (0.13)

The versions specified were used to create the homework. You do not have to use exactly the same version of the respective packages but if you encounter problems related to a particular package, we strongly suggest you install the version specified here as a first step and check if the issue persists.