Homework 1

Foundations of Data Science TA team

released: 25/02/2025; due date: 16/03/2025 (23:59)

Introduction

This homework contains two parts. In the first part, you will have a chance to get back into programming with Python, as you will look at data on the quality and cost of care in hospitals in the United States of America. In the second part, you will use different visualisation techniques to explore data from a cohort of pediatric patients with suspected appendicitis admitted with abdominal pain to Children's Hospital St. Hedwig in Regensburg, Germany, between 2016 and 2021.

After this homework, you should be able to

- describe data using `pandas`' built-in methods
- identify missing values using `pandas`' built-in methods
- formulate hypotheses on reasons for missing values
- build on these hypotheses to replace missing values where possible
- explore a dataset with different visualisation techniques using `matplotlib` and `seaborn`
- identify the appropriate graph type based on your data and question at hand
- derive practical and actionable insights from data.

1 Recapitulation of Python for Data Science

1.1 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 of the U.S. 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 least 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.

1.2 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.

1.3 Basic visualisation

1. Plot the distribution of the cost for treating pneumonia. Save the figure under the path `output/131_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 the path `output/132_heart-failure-cost-vs-quality.png`.

2 Data Visualization

2.1 Data exploration

In order to get an overview of a new dataset, it is a good practice to perform an initial exploratory data analysis. For this purpose, answer the following questions:

1. Missing values exploration:
   (a) Create a function that generates a new dataframe to hold the total number and the percentage of missing values per column.
   (b) Calculate the number of missing values per variable and the missing variables per patient. Hint: Consider the `transpose` method of `pandas`.
   (c) Visualise the missing values per variable using a `barplot` and comment on your findings. Save the figure as `211c_mv_variables_barblot.png` in the output folder.
   (d) Visualise the top ten patients that have the most missing variables using a `barplot`. What do you observe? Save the figure as `211d_mv_patients_barblot.png` in the output folder.
2. Visualise combinations of continuous and categorical variables: Visualise the `age` and the `height` of the patients by sex, as subplots of the same figure, using the appropriate graph type. Justify your choice and comment on your findings. Are your findings in alignment with your expectations? Save the figure as `212_age_height_graph.png` in the output folder. Hint: Think of the data types to choose the appropriate graph and consider the study population to comment on the findings. *Note: More than one graph types may be appropriate.*

3. Visualise categorical variables: Visualise two potential target outcomes, i.e., `Severity` and `Diagnosis`, as subplots in the same figure, using the appropriate graph type. Justify your choice and comment on your findings. Save the figure as `213_diagnosis_severity_graph.png` in the output folder.

2.2 Data Understanding

1. Visualise the `Body_Temperature` by `Diagnosis` for all patients, using a `histogram`, a `boxplot` and a `violin plot`. Present the three graphs as subplots in the same figure. Comment on your findings. Are there any outliers? Can they be explained? Save the figure as `221_body_temperature.png` in the output folder.
2. Visualize six selected numerical variables (as defined in `skeleton.ipynb`) using `boxplots`. Present the results as subplots in the same figure. Comment on your findings. Which variable has the most outliers? Is there a valid explanation for them? Save the figure as `222_numerical_boxplots` in the output folder.

2.3 Explore potential risk factors

1. Visualise `CRP` against `WBC_Count` by `Diagnosis` and `Severity` using `scatterplot` graphs. Present the two graphs as subplots in the same figure. Save the figure as `231_scatterplot.png` in the output folder. Hint: The data points on each graph should have different colors based on Diagnosis and Severity respectively. *Note: In most cases, appendicitis is presented with increased inflammation markers, such as CRP and WBC count.*
2. Create a new variable combining the values of `Contralateral_Rebound_Tenderness` and `Lower_Right_Abd_Pain`. You may name the new variable `Contralateral_Rebound_Tenderness-Lower_Right_Abd_Pain`. Then, visualize the pairwise distribution of counts for `Diagnosis` against this new variable using a `heatmap`. Comment on your findings. Save the figure as `232_heatmap.png` in the output folder. Hint: Consider the pandas `crosstab` method for creating the new variable. *Note: Pain in the lower right abdomen is a common symptom of appendicitis and contralateral rebound tenderness is indicative of inflammation of the abdomen (more specifically in the peritoneum) that can be caused by acute appendicitis among other causes.*

Deliverables

You are asked to submit a **report** (two to five pages, **must be in PDF format**, acknowledgements of AI tools used do not count to page limit), and both the **`.py` file** and the **`.ipynb` file** completed by you to solve the homework. The script file and the jupyter notebook 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 and the Jupyter notebook submitted have 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 first 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**: name, city, state, state abbreviation
- **Ratings of Hospital Performance**: ratings across multiple dimensions (Mortality, Safety, etc.) relative to the national average (below, same as, above national average), as well as an overall rating from 1 (worst) to 5 (best),
- **Information on Cost, Quality and Value**: multiple categories of treatments (heart attack, heart failure, hip and knee surgery, pneumonia).

The second dataset you work with in this homework contains information on paediatric patients with suspected appendicitis abdominal pain to Children's Hospital St. Hedwig in Regensburg, Germany, between 2016 and 2021 [1]. The dataset at hand consists of a subset from the original data and contains information on:

- **Patient Demographics** (5 variables): age, height, weight, BMI, sex.
- **Physical Examination** (6 variables): Alvarado Score (AS), Paedriatic Appendicitis Score (PAS), lower right abdominal pain, contralateral rebound tenderness, nausea, body temperature.
- **Ultrasound (US) information** (4 variables): US performed, US identifier, appendix appeared on US, appendix diameter.
- **Lab tests** (2 variables): white blood cell (WBC) count, C-reactive protein (CRP) levels.
- **Disease outcomes** (2 variables): diagnosis and severity.

A more detailed description of the data variables from the second dataset can be found in table 1.

Python environment

To be able to solve this homework, you need to have a working Python 3.13 installation and the following packages installed in the environment in which you execute the script:

- `matplotlib` (3.10)
- `numpy` (2.2)
- `pandas` (2.2)
- `scipy` (1.15)
- `seaborn` (0.13)
- `jupyter` (1.1)

The versions specified were used to create the homework. 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.

References

[1] Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Ozkan, E., Chin-Cheong, K., Paschke, A., Zerres, J., Denzinger, M., Niederberger, D., Wellmann, S., Knorr, C., & Vogt, J. E. (2023). *Regensburg Pediatric Appendicitis Dataset (1.01)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7669442

Category	Variable	Description
Demographics	age	Patient's age in years.
	weight	Patient's weight in kg.
	height	Patient's height in cm.
	BMI	Patient's BMI in kg/m^2.
	Sex	Patient's gender (by birth).
Physical Exam	Alvarado_Score	Alvarado score (AS): Patient's score according to scoring system.
	Paedriatic_Appendicitis_Score	Paediatric Appendicitis Score (PAS): Patient's score according to the scoring system.
	Lower_Right_Abd_Pain	Right iliac fossa pain detected on palpation.
	Nausea	Feeling of sickness/ejection of contents from stomach through the mouth.
	Contralateral_Rebound_Tenderness	Pain of the contralateral side is felt on the release of pressure over the abdomen.
	Body_Temperature	Patient's temperature in ℃ (rectum or auditory canal).
Ultrasound Info	US_Performed	If an abdominal ultrasonography was performed or not.
	US_Number	Ultrasound identifier.
	Appendix_on_US	Detectability of the vermiform appendix during sonographic examination.
	Appendix_Diameter	Maximal outer diameter of the appendix.
Lab tests	WBC_Count	Number of white blodd cells in a unit volume of blood inflammation parameter ($\times 10^3$).
	CRP	C-reactive protein, elevated in case of inflammation, infection, or injury (mg/L).
Disease outcomes	Diagnosis	Patient's diagnosis, histologically confirmed for operated patients. Conservatively managed patients were labelled as having appendicitis if they had an AS or PAS of ≥ 4 and an appendix diameter of ≥ 6 mm.
	Severity	Severity of appendicitis: uncomplicated: subacute/ catharral, fibrosis; phlegmonous or complicated: gangrenous, perforated, abscessed.

Table 1: Category and description of the patient variables[1]. Information adapted from the UCI Machine Learning Repository