List 4: Python libraries

compiled by Ewelina Tomana

2nd December 2024

General information

All code should be written according to The Style Guide for Python Code. Each function or class should be described with a docstring (see Docstring Conventions). We recommend sticking to the Google or NumPy documentation comment styles.

If pre- or postconditions or *invariants* are not met, the code should raise an *exception*.

It is of utmost importance to maintain originality and avoid any form of plagiarism. When integrating code from external sources, it is crucial to provide precise references to the original material.

Please submit the codes for both exercise 1 and exercise 2. Test functions should be written only for exercise 1. Additionally, for exercise 2, create a report in PDF format containing diagrams and answers to the provided questions.

Exercise 1

Consider a scenario where you have a dataset representing the heights of a population. You are interested in analysing this dataset, which follows a normal distribution.

1. Use NumPy to generate a dataset of 1000 heights from a normal distribution with a mean of 170 cm and a standard deviation of 10 cm. The function should be named `generate_height_data(size=1000, mean=170, std_dev=10)`.
2. Calculate and print the mean, median, and standard deviation of the generated dataset using created function `descriptive_statistics(height_data)`.
3. Create a function `visualise_histogram(height_data)` which creates histogram of the generated data using Matplotlib.
4. Create function `calculate_percentiles(height_data)` which calculates and prints the 25th, 50th, and 75th percentiles of the dataset.
5. Create function `identify_outliers(height_data)` which identifies and prints any potential outliers in the dataset using the 1.5 IQR rule.
6. Create function `random_sampling(height_data)` which with use of NumPy randomly samples 50 heights from the dataset.
7. Formulate and test a hypothesis about the average height of the population. For example write function `hypothesis_testing(data, null_hypothesis_mean=165)` which test whether the average height is significantly different from 165 cm.
8. Write a function `calculate_probability(data, threshold_height=180)` which calculates and prints the probability of randomly selecting an individual with a height greater than 180 cm from the dataset.

Exercise 2

In this task, we will be using the dataset created based on the *Heart Disease Dataset* from the *University of California Irvine Machine Learning Repository*. The dataset was created in 1988 and includes data from four medical institutions. The dataset contains 14 columns:

1. *Age* — patient's age in years.
2. *Sex* — gender.
3. *Chest pain type* — type of chest pain experienced by the patient during diagnostic tests. The chest pain type is categorized into four categories:
   - value 0: no chest pain (*typical angina*),
   - value 1: atypical chest pain (*atypical angina*),
   - value 2: non-anginal pain (*non-anginal pain*),
   - value 3: asymptomatic (*asymptomatic*).
4. *Resting blood pressure* [mmHg] — blood pressure measured when the patient is at rest, without any physical or emotional activity.
5. *Serum cholesterol* [mg/dl] — amount of cholesterol in the blood expressed in milligrams per deciliter of blood [mg/dl].
6. *Fasting blood sugar* > 120 mg/dl — fasting blood sugar level (after at least eight hours since the last meal) expressed in milligrams per deciliter of blood [mg/dl].
7. *Resting electrocardiographic results* — refers to resting electrocardiogram (ECG) results. This column takes three possible values: 0, 1, 2, which respectively mean:
   - value 0: normal ECG result,
   - value 1: suspicion of myocardial ischemia,
   - value 2: presence of a distant abnormality in the conduction of electrical impulses in the heart.
8. *Maximum heart rate achieved* — maximum heart rate achieved during the exercise test. The result of this variable is expressed in beats per minute [bpm].
9. *Exercise induced angina* — indicates whether the patient experiences angina during the exercise test.
10. *ST depression induced by exercise relative to rest* — the value of this variable reflects the degree of ST segment depression in the electrocardiogram during the exercise test compared to the resting state. These values are expressed in millimeters [mm].
11. *Slope of the peak exercise ST segment* — describes the slope of the ST segment in the electrocardiogram during maximum exertion in the exercise test. It can take three values:
    - value 1: upsloping slope of the ST segment,
    - value 2: flat slope of the ST segment,
    - value 3: downsloping slope of the ST segment.
12. *Number of major vessels* — the number of large blood vessels (coronary arteries) supplying the heart that are narrowed or blocked. This variable is expressed as an integer and can take values from 0 to 3, depending on the number of coronary vessels with stenosis.
13. *Thal defect* — refers to the results of the thalassemia test, which is performed on patients to detect defects in hemoglobin. In the context of this dataset, the values of the *Thal defect* variable are interpreted as follows:
    - value 0: no information,
    - value 1: normal,

- value 2: reversible defect,
- value 3: fixed defect.

14. *Disease* — indicates whether the patient has heart disease.

Load the file `heart_disease_dataset.csv` into a DataFrame structure and perform the following subtasks:

1. Answer the question of whether, according to the analysed dataset, more women or men suffer from heart diseases. By what percentage?
2. Compare the average value of serum cholesterol separately for the group of women and the group of men depending on the presence of heart disease.
3. Draw a histogram of people with heart diseases. In which age range are the most affected individuals?
4. Draw a box plot for the maximum achieved heart rate during the exercise test depending on the presence of heart disease. What observations can be made based on this plot?
5. Draw a bar chart for the frequency of heart disease occurrence depending on whether the patient has angina during the exercise test. What observations can be made based on the chart?