Assignment 1

Intelligent Data Analysis
DV1597

Shahrooz Abghari, Bruna Palm, Martin Boldt
Department of Computer Science
Blekinge Institute of Technology, Sweden

April 14, 2024

1 Introduction

For this assignment, you need to work in groups of **two** (2) students. The objective of this assignment is to answer the questions stated in the "Questions" Section below. You will have to import and manage a specific dataset provided to you. More information about the data and Python packages that you are allowed to use in this assignment is listed below. The result from this assignment should be an interactive Jupyter Notebook that you submit via the course page on Canvas before the deadline (which is also stated on Canvas).

2 Before you start

Before starting, it is a good idea to revisit the first couple of lectures in the course and the two first exercises. Also, consult the course literature and reference materials for help with specific tasks in this assignment, which should be documented in a Jupyter Notebook submitted for examination. Your notebook should contain (i) a description of what you have done; (ii) the Python code you have used (including comments for explanation); and (iii) your detailed answers to the given questions.

3 The assignment

This assignment involves importing a dataset from an external file to a suitable Python data structure. It also involves handling various data cleaning tasks (e.g., addressing missing values) as well as aggregating the data in various ways. Finally, you are expected to perform data-driven analyses in order to answer the questions described in Section 4 and, if needed, create graphical plots.

3.1 Dataset

The dataset that you will use in this assignment is gathered by the Swedish Meteorological and Hydrological Institute (SMHI) and consists of official air temperature measures from the city of Karlskrona between *2010-03-01 (00:00 UTC) – 2024-01-01 (06:00 UTC)*. The dataset is available from the Canvas course page as a CSV file, and it is important that you **use exactly the provided file** for this exercise.

3.2 Allowed programming language, packages, and tools

This assignment should be solved using Python (version 3) and the packages available in pip3[1]. The implementation should be executable on Linux, Mac OS X, and Windows 64-bit. Note that the notebook that you submit for the examination should be of the type Jupyter Notebook[2].

3.3 Grades

This assignment is graded with the following grades: `A, B, C, D, E, Fx, F`. Please, see the next Section for a description of what is required of you for reaching each grade.

4 Questions, tasks, and examination

Below are presented the questions and tasks that need to be correctly answered to reach the different passing grades (ECTS A-E) associated with this assignment. If you, for instance, answer only the questions targeting grade E, then your assignment can not be given a higher grade than E. However, if you fail to answer some of the questions correctly, then your assignment could, of course, still be given a grade of Fx/F. To do this, you must update your notebook before resubmitting. Furthermore, if you are aiming for higher grades, for example, a grade of B, you also must answer all the questions for the lower grades (i.e., questions for grades E, D, and C). If you answer some of the questions incorrectly, this could mean that the overall grade assigned to this assignment is lower than the one you are aiming for.

4.1 Grade E

Questions and tasks that should be answered to achieve ETCS grade E:

Q1. Explain the data cleaning steps you consider for analyzing the provided dataset. You must motivate and describe why, how, and in which order you are going to apply the selected steps to the dataset. **Note** that you should consider applying your proposed data cleaning steps on the dataset to answer Q2-Q11.

Q2. Does the dataset contain any missing values? If so, how many in both absolute terms and percentages? Reflect on the number of missing values. Do you regard it as much or not?

Q3. Which strategy was used for handling the missing values? Motivate why you consider the selected strategy the most suitable for this task.

Q4. Calculate the following statistics for the air temperature values in the dataset:

- minimum value
- maximum value
- sample mean
- Q1/Q2/Q3 quartiles
- sample standard deviation
- 95% confidence interval of the mean

Q5. Plot all available air temperature data as a line plot with dates on the x-axis.

[1] pip3: https://pip.pypa.io/en/stable/
[2] Jupyter Notebook: https://jupyter.org/

Q6. Rank the overall temperature per month, sort from the coldest to the warmest. List the top **ten** (10) coldest months in the dataset. Format the output as: `YEAR/Month: temp`, e.g."2010/January: -5.0"

4.2 Grade D

Questions that should also be answered to achieve ECTS grade D:

Q7. Which distribution is suitable to consider to fit the air temperature data? How well does the data follow that distribution given some measure, e.g., p-value?

4.3 Grade C

Questions and tasks that should also be answered to achieve ECTS grade C:

Q8. Do you regard any of the air temperature measures in the dataset to be extreme values/outliers? If so, how many values? Motivate why you consider these values as anomalies.

Q9. Calculate the mean temperature for each day in the year 2023 and then plot those means using a line plot with dates on the x-axis.

4.4 Grade B

Questions and tasks that should also be answered to achieve ECTS grade B:

Q10. Rank the overall temperature per **two** (2) consecutive months starting with the warmest 2-month period first, e.g., $\frac{Y_1M_1+Y_1M_2}{2}$, $\frac{Y_1M_2+Y_1M_3}{2}$, ... $\frac{Y_1M_{11}+Y_1M_{12}}{2}$, $\frac{Y_1M_{12}+Y_2M_1}{2}$, ... After that, you should sort those means from highest to lowest and list the **ten** (10) warmest 2-month periods all together in the dataset.

4.5 Grade A

Questions and tasks that should also be answered to achieve ECTS grade A:

Q11. Calculate the 95% confidence intervals (CI) for the mean of each daily average in **Q9** and add it to the line plot that shows the mean values of each day of the year 2023. You can, for example, solve this by plotting the upper and lower bounds as lines above and below the mean line in your plot. Another more nicely looking solution is to visualize the CI as shades around the mean line in your plot.

5 Notebook to hand-in

You must make your submitted notebook as self-explained as possible. Use markup cells in the notebook to create sufficient sections/subsection headings and text explanations before the cells that contain your Python code. Comment on your Python code extensively. Any output that is presented by your Python code should be professionally formatted. This also applies to any plots you create, which involves (i) giving them sufficient size to interpret the shown data visually; (ii) proper heading in the plot; and (iii) not putting too much label data on either axis. Note that your submitted notebook should be a Jupyter Notebook that includes Python 3 code.

6 Before you submit your notebook

Before you upload your report, make sure:

1. that you have included your **names** and **email addresses** in the report.
2. that you have answered **all questions and tasks** for the targeted grade.
3. that the notebook follows logically and a well-structured **format** with written explanations.
4. that you have carefully checked the notebook for both **spelling and grammatical errors**.
5. that your notebook is of the **Jupyter Notebook** type.
6. that your notebook is written in **English**.

Failure to comply with any of the aspects above could result in a failing grade, and you have to revise the report and submit it again at a later deadline.

Good luck!