Assignment 2

Intelligent data analysis
DV1597

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

April 28, 2024

1 Introduction

For this assignment, you are working in groups of two (2) students. The objective of this assignment is to perform explanatory data analysis on the provided datasets. More information about this data and what Python packages you are allowed to use in this assignment is also listed below. The result from this assignment should be an interactive Jupyter Notebook and a written report (PDF format) that you submit via the course page on Canvas before the deadline (which is also stated on Canvas).

2 The assignment

This assignment involves importing a few datasets from an external file to a suitable Python data structure. It may include handling various data cleaning tasks and transformations, as well as aggregating the data in different ways. Finally, you are expected to perform data-driven analyses to answer the questions described in Section 3 and, if necessary, create graphical graphs.

2.1 Datasets

The *European Centre for Disease Prevention and Control* (ECDC)[1] gathers the datasets that you will use in this assignment. The data is publicly available and contains information about the coronavirus pandemic, such as the number of COVID-19 cases and deaths, hospitalization and Intensive Care Unit (ICU) admission rates and current occupancy, and vaccination across EU/EEA. More information on the data collection process can be found on the ECDC website[2]. There are three (3) datasets you need to work with, which can be found on the course page in Canvas (it is important that you use exactly that file for this exercise). The datasets are as follows:

1. `COVID-19_daily_number_of_new_cases_and_deaths.csv`
2. `COVID-19_vaccination.csv`
3. `COVID-19_hospital_and_ICU_admission_rates.csv`

[1] ECDC: https://www.ecdc.europa.eu/en

[2] ECDC Data Collection: https://www.ecdc.europa.eu/en/covid-19/data-collection

2.2 Allowed programming language, packages, and tools

This assignment should be solved using Python (version 3) and packages available in Pip3 that are possible to execute on Linux, macOS, and Win64. Note that the resulting notebook you submit for the examination should be of the type Jupyter Notebook.

2.3 Grades

This assignment is graded with the following grades: `A, B, C, D, E, Fx, F`.

3 Questions and examination

Your task is to (i) perform explanatory data analysis on the given datasets and (ii) answer the provided questions in Sections 3.1 and 3.2. More specifically, you should inspect the datasets and perform some initial visualizations to identify the relationship between the variables. In addition, you may need to perform statistical analysis. Apart from your written answers in the report, you also need to include statistical tests, tables, and plots (or more advanced visualizations) to ground your answers in the data. All the code for the statistical tests, plots, etc., should be available in your notebook.

3.1 Mandatory questions

The mandatory questions are presented below. **Note**, it is important that you motivate every step and the choice of methods, statistical tests, and visualization techniques when addressing these questions.

1. Since 2020 till 2022, which top 10 countries reported the most number of cases of COVID-19 in quarters per year (Q1-Q4 in each year)? Do you find these numbers relatively high compared to the country's population? Do you see any connection among these countries regarding their geographical locations and the period in which most cases are observed? Discuss your observations.

2. Visualize each country's total number of cases and deaths on a map plot using their geographical locations in 2020, 2021, and 2022. Discuss your observations.

3. What are the top 3 popular vaccine brands that have been used across the EU/EEA? Can you observe the same popularity at the country level? Are there any exceptions?

4. Considering the previous question, which target groups mainly received these vaccine brands in each country?

5. Which countries are the most skeptical towards the first dose of the COVID-19 vaccine? Do you think this matter had any impact on the hospitalization level?

6. Rank all EU/EEA countries, based on their vaccinated population under age 18 for the **first** dose of the COVID-19 vaccine, which countries have the most and least vaccinated people under age 18 in regards to their total populations?

7. Which countries have the oldest vaccinated population for the **second** dose of the COVID-19 vaccine in regards to their total population?

8. Which countries' health care was most affected by the coronavirus pandemic in 2020 compared to others? How are these statistics in 2022?

3.2 Mandatory questions for grades A and B

If you aim for an A or B grade in the course, you must also answer three additional questions. First, specify the three (3) most interesting questions you think are worth investigating. One (1) of the questions must be based on the combination of at least two of the provided datasets. Motivate why you think those questions are worth exploring. Note that the chosen questions should not be the same as the questions in Section 3.1. Furthermore, the questions need to provide some valuable knowledge, e.g., for the public regarding COVID-19 and its effects on EU/EEA countries.
Note it is important to motivate every step and the choice of methods, statistical tests, and visualization techniques when addressing these questions.

4 Report and notebook to hand in

Your submitted report and notebook must be as self-explained as possible. Ensure that the **report** has sections and subsections that follow a logical order. You may consider providing an introduction to the problem you are investigating. For the question sections (3.1 and 3.2), you may group the questions based on their topics for each section while providing your answers, followed by identified results in terms of plots, tables, etc., and your interpretations. You may finalize your report with a discussion section and possible future directions for expanding the report, e.g., those questions you are interested in studying but do not have enough time or relevant data.
If you prefer to write your report using a specific template, we suggest the "Springer Lecture Notes in Computer Science." The template is available in both LaTeX and Microsoft Word formats. The page limit for the report is 6 to 10 pages, using the suggested template.
Concerning the **code**, use markup cells in the notebook to create sufficient sections/subsection headings, as well as text descriptions before the cells containing 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 plot you create, which involves (i) giving them sufficient size to make the shown data interpretable visually, (ii) a proper heading in the plot, and (iii) not putting too many label data on either axis. **Note** that your submitted notebook should be a Jupyter Notebook that includes Python 3 code.

5 Upon submission of your report and notebook

Before uploading your report, make sure that the following are met:

1. That you have included your **names** and **email addresses** in both the report and notebook.

2. That you have answered **all questions** in section 3.1. If you aim for grade A or B, then make sure to answer the questions in Section 3.2.

3. Both the report and notebook follow logically and a well-structured **format** with written explanations.

4. That you have carefully checked both the report and the notebook for **spelling and grammatical errors**.

5. That your notebook is of type **Jupyter Notebook**.

6. That your report and notebook are 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!