Project Instructions

MAI-BB-1-WS2024 ILV Data Engineering

G. Brandmayr

2024-09-12

Overview

The project in Data Engineering counts x% of your final grade[1]. The project weighs 100 points and consists of 3 tasks:

1. Choose your dataset and group in Moodle (0 points)
2. Create a data exploration and analysis notebook (70 points)
3. Present your project (30 points)

Before you start implementing, read this document entirely.

1 Dataset Choice

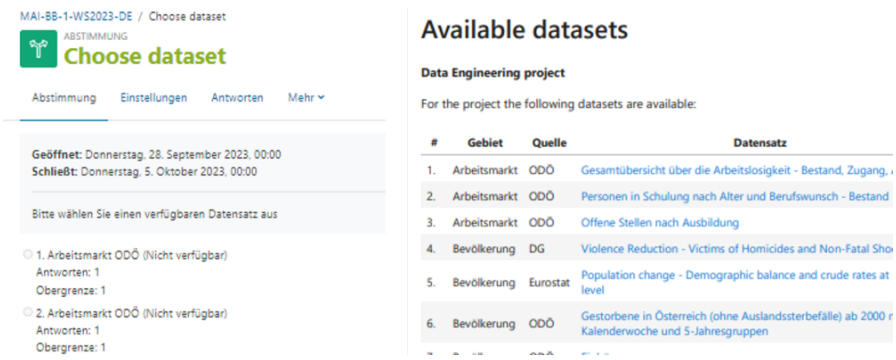

Figure 1: **Left** Moodle choice dialog. **Right** Dataset listing.

In Moodle (Fig. 1) select a dataset in the "Choose dataset" dialog. The project is performed in groups of size N_G and the maximum number of students per group cannot be exceeded. The datasets and their links are provided in Moodle via the document "Available Datasets" (Fig. 1 right).

[1] See grading guide on Moodle for x.

2 Data Exploration and Analysis

This task contains the major steps of a data science project in a single jupyter notebook. The goal is to identify and answer N_G **non-trivial hypotheses** - 1 per group member - quantitatively, i.e., based on data, and export the corresponding artifacts (figures, etc.) for the project presentation (the next task).

Follow the instructions in the jupyter notebook `prj01.ipynb`.

Each group should **submit a ZIP archive**, comprised of the notebook, the sources and results. The extracted notebook should run and have all required local resources available from the extracted archive. This applies in particular for directories (don't use absolute paths).

3 Poster Presentation

The outcome artifact of your project will be a poster. It should be built around the **hypotheses**. Unlike the exploration notebook it should be **concise**, compact and not cluttered with implementation details. Only essential information should be contained, but unnecessary information, redundancy and fill words/empty words should be avoided to make the best use of the limited space.

There is a poster template provided on Moodle, including some good examples and a bad example. Choose a telling title. E.g., prefer "Median income increases with highest earned educational degree" vs. "Group 7 Data Engineering Project". The **authors** (= group members) must be listed in the order of their hypotheses appearance.

The artifact should contain the following sections:

Introduction to the problem. Why do you perform this investigation, what do you answer? Due to space restrictions long elaborations on literature and citations such as [1] should be avoided (maybe 1 citation if important).

Dataset Description Cite the source of your data, e.g., like this [2]. It should include the date when you retrieved it. Briefly describe it qualitatively and quantitatively (number of rows, columns, etc.). Explain the subset you use and the characteristics of the most important variables. Also mention, what you didn't use for the analysis and why. You may consider using a figure (e.g., box- and stripplot, etc.) or a table.

Hypothesis n The most important section, $n = 1, 2, \ldots, N_\mathrm{G}$. Treat each hypothesis separately, covering the following aspects:

1. **Hypothesis** statement - possibly as section title.
2. **Method** of investigation. How did you answer the hypothesis?
3. **Result** - built around the artifacts (figures, tables) exported from the notebook, which may look like Fig. 2. In a poster—opposed to a report—**do not**

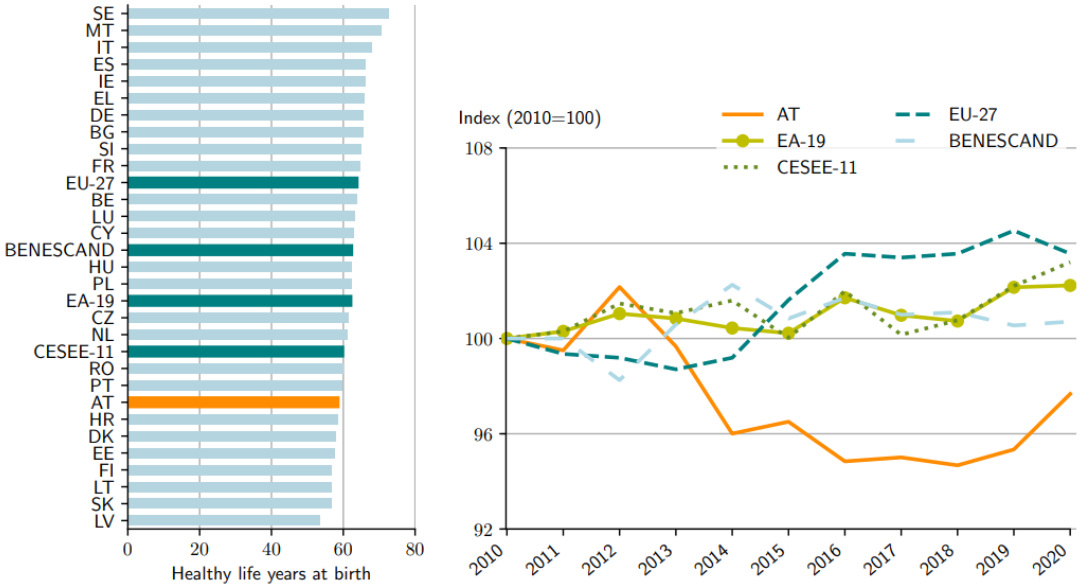

Figure 2: Healthy life years at birth. **Left** 2020 **Right** 2010-2020 Index (2010=100)

use figure captions. As rule of thumb, try to get along with 1 figure (use sub-figures) per hypothesis. Due to space restrictions **cross references**, such as this: "Fig. 2", should be omitted. Utilize the space with appropriate layout of figures.

Method and hypothesis statement should be brief, with a clear emphasis on result. No need for subheadings or many paragraphs. Save space!

Conclusion Interpret the results and explain gained insights.

References Use at most 2-3 references. It's recommended to **not use** a full sized section, but rather a space saving method.

You have $2 + N_{\mathrm{G}}$ min to present the poster. To present the poster each speaker needs to zoom in on the respective section. Take care for readability and according font size (of course you are free to modify the template).

It is an art in itself to create good posters. Good posters avoid excessive text and speak through compelling figures[2]. Let your figures tell your message! You cannot afford blabla (excessive text)—you will simply run out of time without having anything useful said. Thus, be clear what your result is.

Criteria

The criteria for the presentation are:

[2]Such figures are also what differentiate good reports and scientific articles, although it is less apparent.

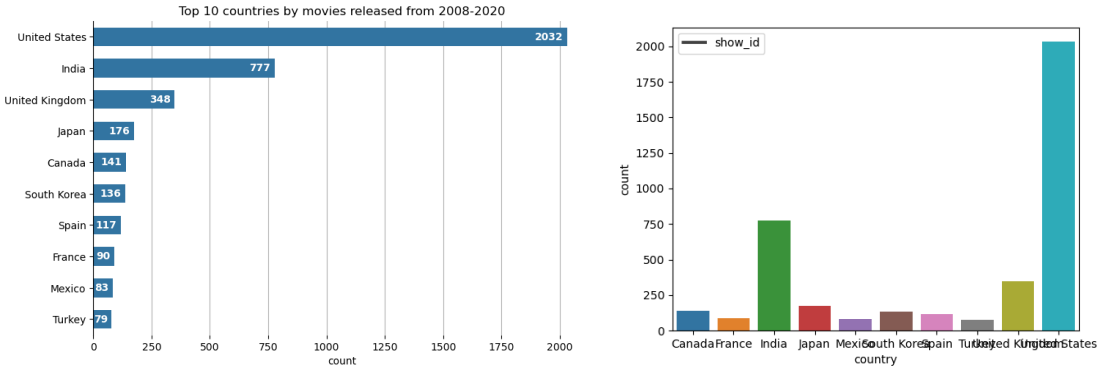

Figure 3: **Left** Horizontal bar chart with useful bar arrangement, readable labels and values. **Right** Vertical bar chart with useless bar arrangement, false legend, useless colors, missing title, unreadable tick labels and unnecessary x-label.

- importance of sections:
  1. Hypotheses 70%
  2. Dataset Description 10%
  3. Conclusion 10%
  4. Introduction 10%
- correct content
- argumentation based on facts (i.e., data)
- overarching story-line "connecting the dots"
- useful figures (compare Fig. 3 left and right) and tables
- concise language (grammar, spelling - use spellchecker)

Avoid:

- misusing the page limit for poor, insufficient content. It is easy to write a short bad report. A long, good report is hard. A shorter, good report is harder.
- deeply nested or many headings - they consume a lot of space. The `description` environment may be used instead.
- table of content, figures or the like - this is not a thesis.
- too much white space around figures. Use proper layout - e.g., place two plots side-by-side as shown in Fig. 2 and 3.
- common truths

- hear say
- plagiarism
- self-plagiarism or redundancy (copy paste)
- immediate discussion in the results section
- exploration style figures, i.e., missing labels, title or units. Programming column names (e.g., "VAR_GrpAge10_19"), etc., see Fig. 3 right.

References

[1] Peter J Bickel, Eugene A Hammel, and J William O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation." In: *Science* 187.4175 (1975), pp. 398–404.

[2] Eurostat European Union. *Healthy life years from birth*. 2023. URL: https://ec.europa.eu/eurostat/databrowser/view/HLTH_HLYE/default/table?lang=en (visited on 09/12/2023).