Capstone Project

1. Overview

"Graphical excellence is the well-designed presentation of interesting data - a matter of substance, of statistics, and of design [...] It consists of complex ideas communicated with clarity, precision, and efficiency." (Edward Tufte)

In spirit of this quote...

- Choose real-world data on a topic of your interest (see section 2)
- Identify a story worth telling (section 3)
- Tell this story using visualizations that are truthful, useful, and beautiful. Upload a PDF file with the main visual story, a companion Jupyter Notebook, and further files needed to reproduce your work (section 4)
- Allowed tools, technologies and languages are described in section 5.
- The grading criteria are described in section 6.

Ask me if you have questions related to the task, or to get feedback on your ideas or first results.

2. How to choose a dataset?

Choosing the "right" dataset(s) in the beginning of the course can be difficult, especially since you may not yet have a clear picture of desired properties of the dataset.

Therefore, consider the following recommendations:

- Choose a topic that you are **100% interested** in
- Your data should either have sufficient **size** (number of rows) and **variety** (e.g. time dimension, categorical data, numeric data, geographic data). Alternatively, rather than working on a single complex data set, your project can also be based on **multiple** simpler datasets that all separately contribute to your topic.
- If your dataset is too simplistic, small and its contents obvious, then it may not be suited for this capstone project.
- If you compile a data set e.g. from statistical offices, APIs, or via web scraping this may involve considerable upfront **efforts in data collection and cleaning**. Such efforts will be acknowledged in the grading (see below). Conversely, if you use an already fully cleaned data set, I expect a larger scope, complexity or innovativeness in terms of the actual visualizations and the story.

- Feel free to ask me for feedback on the suitability of your data choice.

To give you some inspiration, here are some sources that students have chosen in past semesters:

- Berkeley Earth (Temperature Data)
- Bundeskriminalamt (BKA)
- Destatis (German Federal Statistical Office)
- Deutscher Wetterdienst (DWD)
- European Centre for Disease Prevention and Control (ECDC)
- Eurostat
- Football-Data
- OECD
- Sports Reference (Football and other sports data)
- Spotify (Retrieval of own data, listening history, etc.)
- Spotify Developer API (metadata on songs, artists, ...) (some data is not longer available)
- Stackoverflow Developer Survey
- UNHCR (Refugee Data)
- Uppsala Conflict Data Program (UCDP)
- Wahlrecht.de (German Election Data)
- World Bank Open Data
- World Happiness Report
- World Health Organization (WHO)

In addition, you may find interesting datasets on platforms such as Kaggle.

- However, note that these are just platforms for data sharing. **They do not represent the original source of information**. Data on Kaggle can be uploaded by anyone, and documentation is usually missing or incomplete. This is a problem, because if the data source is unknown or not trustworthy, what can we then hope to learn from the data? Have you ever seen a newspaper article referencing Kaggle as a source?
- My recommendation: If you find an interesting dataset on Kaggle, (1) check whether it is possible to identify the underlying raw data source and then work directly with data from that source. (2) If the first option is not possible or too time consuming, then carry out some form of quality or plausibility checks of the data, and document these.
- Being transparent about the data and its source is important, and will be considered in the grading (see below)

3. How to tell a good story?

Your main deliverable should NOT be exploratory, but rather explanatory: you are asked to tell a truthful, useful, and well designed story using visualizations.

- This means that you need to **identify such a story in the beginning of your work** through an exploratory data analysis (EDA). Such a story must have one or more *core insights* and/or even have a *call for action* for your target audience.
- Once you have identified a story worth telling, **focus on just telling this story** and optimise your visualizations according to the aspects discussed in the lecture. **Drop everything from your EDA that is not relevant for your story**. Even if you put a a lot of efforts into many other aspects during your exploratory data analysis: Resist the temptation to present it to your audience, because it would distract the audience and make the insights less clear.

"You might have to open 100 oysters (test 100 different hypotheses or look at the data in 100 different ways) to find perhaps two pearls... When you are at the point of communicating your analysis to your audience, you want to be in the explanatory space, meaning you have a specific thing you want to explain, a specific story you want to tell—probably about those two pearls." (Nussbaumer Knaflic, 2015)

4. What are the deliverables?

Submit a zip file containing the following:

(1) A PDF file that includes your visual story

The binding requirements are:

- Include a minimum of 3 visualizations.
- The visualizations should have some variety (e.g. not only line charts)
- Make sure that the visualizations are self-explanatory: the reader must be able to take away the main message without reading through your entire report
- Optimize all visualizations to make them truthful, useful, and beautiful.

You have a lot of latitude in how exactly your visual story looks like:

- You can make a graphic-heavy one-pager
- Or a multi-page article with text and embedded visualizations, e.g. using a data journalistic or scientific style. The text should then contain additional contextual information and explanations.
- Make deliberate choices regarding the design (color, size, font, ...) and arrangement of all elements (title, text and visualizations).

(2) A Jupyter Notebook that includes your code and explanations

- The Jupyter Notebook serves as a companion file of the actual visual story. The reader can consult the Notebook to obtain the code necessary to reproduce your visualizations. Also it can contain additional or more technical explanations (e.g. about the data or preprocessing) that does not fit into the main project output.
- You do not need to repeat explanations that are covered in the project report.

- You can split your code into 2 notebooks, e.g. if you have a data engineering part and an analytical part.

(3) Data (and any further resource needed to reproduce your work

5. Which tools, technologies, languages can be used?

- Allowed languages for the text are **English and German**.
- The data processing and creation of data visualizations must be carried with **Python**. The choice of Python visualization packages is yours.
- For final optimizations to your visualizations (annotations, handle overlaps, highlighting), you can save your visualizations in svg format and use a vector editing software such as **Inkscape**. Note that these edits won't be reproducible (which is fine!)
- For the creation of your final PDF document, you can use standard software such as **Word**, **Pages**, **Latex**. Alternatively you may try out an open source desktop publishing software such as **Scribus**.

6. How is the project graded?

Your project is graded based on a holistic evaluation of the following 5 aspects:

Truth: Your entire workflow - from data collection to cleaning to visual representation - is grounded in scientific integrity. Use credible data sources, cite them transparently, and avoid any manipulation or framing that could mislead. Strive to present what the data actually say.

Useful: Your visual story is insightful, focused, and easy to understand. You provide the essential information needed for the audience to take away concrete insights or calls for action. You follow the principles of good visualization design (e.g., perceptual rankings, contrast, alignment) to ensure clear and intuitive communication.

Beautiful: You use layout, color, typography, etc. in a way that enhances clarity, supports the story, and creates visual harmony - without distracting from the message. Beauty serves comprehension, not decoration.

Jupyter Notebook Documentation: You demonstrate an excellent command of data manipulation and visualization libraries. Your code is clean, easy to follow and the outputs are reproducible. You provide informative explanations and comments.

Data efforts and creativity: Projects may differ strongly related to the efforts needed to collect and clean the data. If you have put a lot of effort into this part, this will be honored in the grading. Similarly, if you come up with a creative way to present your data, this will be honored as well.