Capstone Project: Task Description

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 a real-world dataset on a topic of your interest (see section 2)
- Identify a story worth telling (section 3)
- Tell this story using visualizations that are insightful, truthful, and well-designed. Upload a PDF file with the visual story, a companion Jupyter Notebook that includes code and additional explanations, and further files needed to reproduce your work (section 4)

Allowed tools, technologies and languages are described in section 5. And the grading criteria are specified 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 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 dataset on a topic that you are **100% interested** in
- Choose a dataset that has a sufficient **size** (number of rows and columns) and **diversity** (e.g. time dimension, categorical data, numeric data, geographic data). You may also combine it with data (from different sources) to make it larger and richer. If your dataset is simplistic, small and its contents obvious, then it may not be suited for this capstone project. Conversely, the data should not be so large that it is difficult to handle it in a Jupyter notebook. Feel free to ask me for feedback on the suitability of your data choice.
- If you put together 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.

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
- OECD
- Spotify (Retrieval of own data, listening history, etc.)
- Spotify Developer API (metadata on songs, artists, ...)
- UNHCR (Refugee Data)
- Uppsala Conflict Data Program (UCDP)
- World Bank Open Data
- World Happiness Report
- World Health Organization (WHO)
- Stackoverflow Developer Survey

In addition, you may find interesting datasets on platforms such as Google Dataset Search or Kaggle.

- Note that a platform such as Kaggle is by itself never a credible data source - because it is never the original source of information, but just a platform. Data can be uploaded by anyone, and documentation is often missing or incomplete. Have you ever seen a newspaper article referencing Kaggle as a source?
- This is a problem, because if the data source is unknown or not trustworthy, what can we then hope to learn from the data? Garbage in, garbage out, ...
- My recommendation is: 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, or (2) at least make and document your own quality/plausibility checks.
- Being truthful about the data and its source is important, and will be considered in the grading (see below)

3. How to find a good story?

Your task is to tell an insightful, truthful, and well designed story using visualizations. A **story** must have a **core insight** and/or even have a **call for action** for your target audience.

This means that you need to identify such a story in the beginning of your work by exploring your data. This exploratory part can be challenging and time consuming. "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." (Nussbaumer Knaflic, 2015)

"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." (ibid)

Important: Your final submission should NOT be exploratory, but rather explanatory!

Only focus on the two pearls, not on the 100 oysters. Drop everything that is unrelated to your actual story. Even if you put a a lot of efforts into your initial exploratory data analysis: Resist the temptation to show it to your audience.

4. What are the deliverables?

Submit a zip file containing the following:

(1) A PDF file that includes your complete story

- It should contain a minimum of 3 visualizations.
- The visualizations should have some variety (e.g. not only line charts)
- Good visualizations are "worth a 1000 words": optimize each visualization to make it understandable, beautiful and memorable. Finetune and annotate the visualizations so that they are self-explanatory.
- Apart from the visualizations, the story must also be told in text form. The text may contain additional context information and explanations.
- You have a lot of latitude in how exactly your file looks like. You can make a graphic-heavy one-page handout, or a multi-page article with more textual information (using a data journalistic style).
- Make deliberate choices regarding the composition of title, text and visualizations and make consistent design choices for the entire document.

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

- You can think of this Jupyter Notebook as a companion file of the actual visual story. The interested reader can consult the Notebook to obtain the code necessary to reproduce your work, as well as additional and more technical explanations.
- Explain your data, processing steps, assumptions, etc. in Markdown form. Code-related explanations can be made in the form of code comments.
- If you have a time consuming text collection and cleaning part, it may make sense to store this part in a separate notebook.

(3) Any further resource needed to reproduce your work

- Data
- Images
- Helper scripts
- ...

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 out using Python. The benefit of this programming-based approach is that your main work will be reproducible and easily extendable. The choice of Python visualization packages is yours.
- For final optimizations to your visualizations (adding or finetuning annotations, highlighting individual items, etc.), you can save your visualizations in svg format and use a vector editing program (e.g. Inkscape). Note that these edits won't be reproducible.
- For the creation of your final PDF document, you can use standard software (Word, Pages, Latex, ...). Also 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:

- **Story**: Your story is interesting and well communicated. You provide your audience with the core information needed to understand the story. The audience takes away concrete insights and/or a call for action.
- **Truth**: Your entire work - from data collection to cleaning to visual representation - is done with scientific integrity. Seek for truth: Use data from a credible source. State your sources. Do not manipulate or misrepresent the data to mislead the audience.
- **Design**: Your visualizations are understandable, beautiful and memorable and tailored towards the needs of your audience: You use the right type of visualization for your data and question. You follow best practices of visual design (e.g. contrast, repetition, alignment, proximity). You avoid typical pitfalls (e.g. chart junk, misleading axes). You design your visualizations with the nature of human perception in mind.
- **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/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.