Objective

You must present the Task and Data Abstractions in Checkpoint I according to the What? Why? How? framework. Motivate why this domain in worth investigating, which insights you want to extract, and a convincing dataset exists for your visualization. At the end of this Checkpoint, you must have the tasks and data files for your visualization.

Requirements

Look at the course materials to understand the questions you need to answer.

Problem Domain

What is the domain you are going to tackle? Economy? Oscar winners? College student enrolment grades?

Choose a subject you find interesting and exciting. You must provide a **high-level description** to motivate the rest of the document. **Do not be too vague**. Avoid stating your project as a "visualization of COVID-19 world data". Instead, it should be about "how COVID-19 changed European commute". We want to know your scope.

Why (Task abstraction)

What tasks and what purpose?

Think about which tasks your visualization should help the users perform and which concrete questions it should give an answer to. You must provide a **set of at least five different, representative concrete questions** (and their description) organized in bullet points. We will look for the following question types: comparison, correlation, distribution, flow, geolocation (events, relationships, paths, or regions), hierarchy/inclusion, ordinal time, (cyclic/repeating) patterns, proportion, rank over time, relationships, time intervals, trends, and uncertainty. This doesn't mean you have to come up with a task of each of these types, but that you must tell us which task type your tasks are.

Please provide a description that allows for understanding the questions and their complexity. It will be reflected in your grade. Question enunciation must be as clear and concrete as possible.

Do not make questions too similar or just instances of the same meta-question. For example, both "Does the age influence the daily calory intake?" and "Does the daily number of steps decrease with age?" are trying to check how one continuous variable (age) correlates with other continuous variables (daily calory intake and daily number of steps).

If the dataset does not allow you to tackle different domain scopes, it is probably too simple, and you should choose another one. Strive for diversity, both in terms of task type and complexity. Just trivially answerable questions won't make a good project. Your grade will reflect this. On this regard, do note: you will be creating a dashboard that integrates several visualization techniques, integrating being the operative word. Don't think about these tasks/questions as something that isolated techniques can give answers to. At least some should arise from the interaction between the different parts of the dashboard ("Technique A for Task A; Technique B for Task B; etc. is *not* the way to go).

What (Data abstraction)

Which dataset(s) and how you process them?

Please tell us **which dataset you will use** and **where it is available**. Describe the dataset. Is it static or dynamic? How many variables are you dealing with? And how many items?

You can collect your data (e.g., using web scrapping or an API) if you want (it doesn't have to be a pre-made dataset). You can (probably should) also take data from several places instead of looking for this one website where you can download a zip file with exactly what you need. It may not exist and if it does it is an alarm sign: are you really looking for data to answer the questions that are relevant, or are you overfitting your questions to some "easy" dataset you found somewhere?

Still, remember you only have one week to prepare this so work smart, not hard, and do take advantage of existing datasets if they match your questions. Describe how, what sources, and what effort is involved to get **the initial dataset**, independently of where you got it from.

Parse the original data into .json or .csv format to use with D3. The initial dataset will have things in the wrong format. It will have attributes you don't care about. Some may have missing values our outliers. You may have two tables that you need to merge into a single one. You must describe how you processed the dataset (how you cleaned it, problems found and solutions, how you fixed missing values, cross-referenced different tables/datasets, etc.). If you do not need to process the data, explain why.

Describe which variables you selected from the dataset and which derived measured (**if needed**) you consider, computed, and why (based on your tasks and questions). Remember: you must visualize information and not data. Think derived measures. If you did not derive any measure, carefully justify why you didn't need any.

Having done all that, the main part of this section will, then, be the description of the final dataset that you will use in the next checkpoints/project, including:

- Description of the **dataset type** (spatial, table, field, etc.).
- Description of **each item** and **attribute** (nominal/ordinal/etc., diverging/sequential scale, etc.). Be precise. For instance, if you cover time-based variables, we want to know all features (linear/cyclic, hierarchic or not, spans or intervals or instants, etc.).
- **Semantics** (what does each attribute and item stand for). We recommend using a table to save some space. Do not forget to include ***all variables*** (yes, including the derived measures).

Mapping

Show **concrete** examples so that you can answer the questions you formulated with your data sample ("we want to know X, so if we look at attribute Y of table Z, and compare it with attribute W…"). Include ***all questions*** from the Task Abstraction. *No screenshots*! This is an important step for you (and us…) be sure that the data that you have is adequate for your goals.

Deliverables

Create a **4-page document using the provided template**. Follow the sections of the template but add more sections as you see fit. You should also **include in the submission** the files of your **final dataset** that you are going to use in the project. **Compress everything in a .zip** and submit it online, until two days before your class (ex: classes on Monday must submit until Friday end of day).

In case you're having difficulties in selecting your data source and problem domain, check out the course pages on "Data Sources" and "Visualization Ideas". Notice that some visualization domains are banned, and you cannot use them.

The way ***NOT*** to do this: looking up some vis examples (especially copy & pasteable ones…) and trying to find a subject that fits. Choose a subject you are passionate about. Choose a domain you are familiar with. Choose questions that have been keeping you awake at night. The result will be much better.

Example

You have already done something similar in the Wine Tasting Workshop. What you are doing here is very similar to our initial steps. First, we chose the domain and derived some interesting questions we wanted to answer. Then, we found the dataset. Remember the first question: Which year has the best points average? There was no attribute regarding the year. We needed to process the data and extract the "year" from the "title" attribute. We also had to clean the data. Most items had no "region_1" or a "region_2" unless they were from the USA. We also filtered the dataset to have tastings referring solely to Portuguese wines.

Another interesting question focused on whether there was a positive relationship between the wine's price and rating. The question type is pretty simple to guess: correlation. We have two quantitative variables (the price is a ratio, and the rating is ordinal), and we are trying if there is a linear relationship between them. How can we show that we have the data to answer the question? The columns "price" and "points" have the data needed for this specific question:

```
(from "dataset_wines.csv")
id,points,price,region,(…)
70163,89,50,Douro,(…)
```

These are some decisions and jobs you must make at this checkpoint. However, these are just broad strokes. We want you to explain everything in more in-depth.

Penalties

- Documents over 4 pages long: **1 grade point penalty per extra page**.
- Document uploaded after the deadline: **0.5 grade points penalty per hour of delay.**
- Document template altered (wider margins, smaller font, etc.): **1 grade point penalty.**
- Missing data files: **1 grade points penalty**.
- Incorrect file extension: **0.5 grade points per file**.

Tasks to perform during the lab

The professor will provide feedback. The grade will be made known one week later (see below). If you are not receiving feedback, then you must be peer assessing your colleagues' submissions.

Grading

Your proposal will be graded according to the following parameters:

- **Theme registration in Moodle (5%)**: you must register a valid theme in Moodle.
- **Domain identification and motivation (10%)**: clearly identify the theme you tackle and justify why it is interesting and relevant.
- **Task abstraction (35%)**: present at least five different questions, identifying their type and given concrete examples. Explain the rationale behind each question. Diversity of types and complexities will be taken into consideration.
- **Data abstraction (35%)**: present and describe the initial dataset, the data parsing/processing, and the selected and derived measures.
- **Mapping (10%)**: show that you can answer each question with data samples.
- **Peer assessment (5%)**: grade your peers' submissions in the class.

An important note on grading for this and all other Checkpoints: always ***justify*** your choices, based on the basic principles you have learned in this course (adequacy of channels to data types, human perception, etc.). Don't just describe what it is, but especially ***why*** it is as it is.

Additional Notes

After you deliver your document, your work will be graded. HOWEVER, this grade **can be improved by up to two grade points** if you correct any faults pointed out by the professor and submit a revised version of the document HIGHLIGHTING THOSE CHANGES before the beginning of the class taking place 7 days after you receive feedback in class. ***Only highlighted changes will be considered.***