Information Visualisation
Project Assignment

Prof. Dr. Beat Signer (bsigner@vub.be)
Yoshi Malaise (ymalaise@vub.be)

Academic Year 2021-2022

Deadlines

1. **Monday, 21th of February at 23:59**: Email to assistant with three group members (Section 2)
2. **Sunday, 27th of February at 23:59**: Canvas upload with top three datasets (Section 3.3)
3. **Thursday, 21th of April**: Intermediate presentation (Section 7)
4. **Week of May 23**: Final presentation (Section 7)
5. **Sunday, 29th of May at 23:59**: Canvas upload with all deliverables (Section 7)
6. (Optional) Monday, 30th of May: Email concerning possible point distributions (Section 2)

Note: Deadlines are in Europe/Brussels timezone

1 Goal

The goal of this project is to process and visualise a large dataset using the theoretical knowledge that you will obtain during this course. During the lectures you will learn various principles, guidelines and techniques that will help you decide on your visualisation. As this is a project counting for 40% of your final grade, you are expected to choose a reasonably sized dataset with an **interactive** visualisation that helps your **target user** to get a good exploration and/or presentation of your data.

2 Groups

Groups are formed with **three** students. The project will be discussed during the very first introduction exercise with the deadline for handing in the composition of your group set to the 21th of February at midnight.

You are expected to work in group towards your final deliverables. Any problems concerning group members should be reported as soon as possible to both the assistant and professor. At the end of the project (before you are graded) you have the possibility to give away a maximum two points to your other group members. This is usually used in case of unequal task distribution and is completely optional. When sending the mail about this point distribution, make sure that you put all group members in CC. During the oral exam, we will also assess your knowledge on the project.

3 Choosing a Dataset

We expect groups to use a dataset-oriented approach towards their final visualisation. This means that you should first find a dataset that you like and then define the tasks needed to gradually process and visualise this dataset.

Datasets can be taken from anywhere but **have to approved by us**, meaning you have to choose an existing dataset that is big enough to be a challenge. We advice that you decide on a **topic** first, then find three potential datasets about that topic. Once you have your main dataset you will have to decide the **target user** and tasks that you are going to create a visualisation for. This is important as you will have to motivate why this user needs this visualisation and evaluate your final result with those users.

3.1 Sources

The following list is a helpful set of sources where you can find public datasets. These sources including a few more domain-specific ones will be discussed during exercise 1:

- Kaggle: https://www.kaggle.com/datasets
- Zenodo: https://zenodo.org/
- Google Dataset Search: https://datasetsearch.research.google.com/
- DataMed: https://datamed.org/

Of course, you can pick any source you want as long as the dataset is publicly available.

3.2 Requirements

The following list contains some soft requirements set for the dataset. This list helps you decide on the three datasets and our method of approving the datasets:

- **Size:** You can combine multiple datasets together, but your main dataset should be reasonably sized (both in the amount of data that is available and the information that is represented).
- **Real data:** The data that your dataset contains should be real data. Generated data that is not based on any real source are not allowed (e.g. randomly generated sales).
- **Meta-information:** Related to the "real data", the dataset that you use should contain information about how the information was obtained. On the dataset sources that we will discuss this is usually the case.

3.3 Approval

Once you have selected **three** potential datasets you have to upload a description on Canvas (does not have to be a report, can be a text file). For each of the three datasets you should provide:

- The source of the dataset (where can we download it)
- In case the above link does not contain meta-information on how the data was retrieved, make sure that you specifically mention it
- A short summary of the dataset
- Target users (who are you going to create a visualisation for)
  Note: It is important for your final evaluations that you design the project for a specific users
- Potential initial ideas you have (e.g. do you want to combine or compare data, do you already have an initial idea to visualise it ...)

We will respond as soon as possible with our feedback on each dataset.

4 Preprocessing the Dataset

Preprocessing a dataset is part of the assignment and inevitable with any dataset. This processing could be to combine multiple datasets together or rearranging the data before creating visualisations. You are not required to perform this preprocessing in the same framework that you use for the visualisation. While many frameworks offer this capability, it is often easier to choose a specific framework depending on the type of data that you are dealing with. During the exercises you will learn dataset preprocessing using Python.

Note: The main goal of the project is to visualise information. While the preprocessing is graded, it should not be the main aspect of your project.

5 Visualising the Dataset

Once you have chosen a dataset and prepared it, you can start to visualise it using any framework that is capable to do so. Keep in mind that *creativity is awarded*.

It is of utmost importance that you correctly apply the theoretical knowledge and guidelines from the lectures, including the use of interactive visualisations.

A small list of frameworks:

- R (Studio): Used by academics for stationary visualisation and preprocessing. The Shiny[1] package can be used to create interactive web interfaces.
- D3.js: Commonly used visualisation framework on the web, can be used on top of MapBox or Leaflet
- Plot.ly: Visualisation framework that works for R, D3.js and Python
- Leaftlet.JS: Useful for geospatial visualisation
- MapBox: Similar to leaflet
- Tableau: Well known visualisation framework https://www.tableau.com/
- Python: Commonly used visualisation framework by academics and programmers. You can use frameworks such as Bokeh for interactive visualisations.

6 Evaluating the Visualisation

The validation and evaluation of your solution is an important aspect of any academic project, including your visualisation. Information about this evaluation will be discussed in the lectures and exercises.

[1] https://shiny.rstudio.com

7 Presentations and Final Report

During this course you will have to perform two presentations. Both presentations will be held **online** through BigBlueButton on Canvas. Timeslots for these presentations will be published as an announcement on Canvas. Every team member is expected to participate in the **live** video presentation. BigBlueButton supports the uploading of slides and screensharing.

The first presentation is an interim progress presentation on the **21th of April**. During this progress presentation you have **15 minutes** to present your chosen dataset, the steps you have taken towards processing and visualising the dataset and a GANTT chart of your future progress. Make sure to mention why you have chosen the dataset and for which task(s) you are creating a visualisation and how you are brining the data to the target user. Immediately following the presentation you will receive feedback on your progress and the provided GANTT chart.

The final presentation **during the week of May 23th** should be a complete overview of your project. Similar to the first presentation you discuss your dataset, the preprocessing performed on these datasets, the visualisation(s) you have made and the evaluations. For this presentation you have **20 minutes** with an additional 10 minutes for questions. After the presentation you will receive potential feedback of things that are important to outline in your final report (such as aspects that were not clear from your presentation).

The deadline for the final report and visualisation deliverables (dataset, source code of preprocessing and visualisation) is on **Sunday the 29th of May at midnight** by uploading it on Canvas. If your deliverables are too large to upload to Canvas, you can also upload them on the Sharepoint of the VUB. Make sure that you provide a report with the scientific quality expected from a Master course (e.g. use of references, VUB layout, writing style, ...). The report should have a minimum of 10 pages (with images but without possible appendices) describing your complete solution (**including setup instructions on how we can run and view your result**).

8 Referencing

Like any academic project, all deliverables including your report, images and source code are checked for plagiarism. Make sure that you always reference sources in both your source code, report **and your final visualisation**. This also includes the dataset(s) that you have chosen.

9 Deliverables

Your final deliverables should consist of:

- Interactive visualisation
  - Demo video showcasing the functionalities
  - Source code of the visualisation
  - Source code of the preprocessing
  - (optional) Website URL of the live version
- Interim and final presentations
- Evaluation of your solution with target users
- A report of your visualisation, evaluation and preprocessing

10 Examples

Throughout the lectures you will see various real-world examples. The following list are several examples that represent the quality that we are after:

- Coronavirus 2019
  Provides a useful dashboard of information about infections, deaths and recoveries. Also notice how the visualisation is accompanied by additional information containing the sources. Note that this visualisation also has several mistakes, such as its slow loading time.
- Codex Atlanticus
  An interactive visualisation that makes it easy to see what pages of the Codex Atlanticus contain certain information. The visualisation is good in clearly visualising how much information is included on each page.
- Wind and Words (Game of Thrones)
  This interactive visualisation shows conversations and grammar between multiple characters in the Game of Thrones series.
- SelfieCity
  This interactive visualisation proves that you do not need textual or relational datasets. On this site you can see what type of selfies are taken around the world including interactive comments that show that certain countries have more smiles on selfies than others for example. The site also contains some findings based on their datasets that are also visually represented.