Course project description

For “Data processing and visualization”, IE500417

Data set

An example data on CO_2 and Greenhouse Gas Emissions is provided CO_2 Source Data from “Our World in Data”. It is from 1850 up to current data, updated yearly, all countries and regions. Please use the complete dataset.
Description of the columns is here. As well as example code. You can find your own dataset related to your question.
In addition, you should **find another data set which can be linked to CO_2 cases**. We don’t give any examples here to not limit your creativity. There should be plenty of opportunities to get open data sets related to weather, economic indicators, business activity, country policies, etc. You can even import some social network data, such as tweets or posts. Be creative here.

Task

Create a data analysis platform with interactive visualization for this data set to help users understand it. The visualization should be usable to answer several questions, and explore the data in several perspectives, for example:

- Who is the target user and what is the context it is used for?
- Apply different interactive visualization methods to describe the evolution of greenhouse emissions. Make sure the users can compare spatial-temporal features.
- Analyze potential correlations between *Total greenhouse gas emissions* and possible analytical factors. (Based on your own question, for example, do the warm countries have more to emissions, does life expectancy have an impact toward greenhouse emissions?)
- What is your prediction for greenhouse emission evolution in the next 3 months (can be trend, sequence or on your own choice)? How do you evaluate your results? Use different prediction methods and compare your predictions.
- Can you find other factors which could affect greenhouse emissions? For example, temperature, weather, even stock prices, or social media emotions? Find new datasets to prove your points.
- Users should be able to do exploratory analysis with the visualization to answer the final question: How to fight greenhouse emissions?

Note: you may need to do some data cleaning and pre-processing/processing before visualizing it.

Deliverables

As a result of the project, you should deliver:

- The visualization tool - either as a set of files with instructions on how to run it, or a URL to a website where the tool is available for trial.
- A 5-10 minute intro video of the tool, explaining the main functions, and reflection questions, as shown below.

All these deliverables should be included in the final portfolio delivered in the Inspera system. **Deadline for Inspera submission: December 9th 12:00 (noon, not midnight)**.

Requirements of the visualization tool

It must support two main features:

1. Filter the data according to the different combinations of attributes. It can be based on your own choice, for example, to show only countries where the median age of the population is over 40 years and GDP per capita is less than 5000.
2. Explore the relationships between attributes with expressive figures.
3. Incorporate another dataset (of your choice) into the visualization. This is a very open part of the project - you can find any data set where you expect to see a connection with CO_2 data. This part is not mandatory, strictly speaking, i.e., you can get a grade without it. But this part is worth 30% of the grade for the project. I.e., without this step the maximum grade you can get for the project is 70%.

Questions for reflection (should be covered in your video submission)

In addition to your visualization tool, you should also cover this in your 5-10 minute video, where you reflect on several things. The following answers will have an **impact on the grade**:

1. Describe the theoretical principles that you followed when working on this assignment. This is an important part where you show that your choices were not random. Describe how you chose colors and chart types; how did you design the views and layouts. We have discussed a lot of theory in the course, this is the place where you connect the theory to practical exercise. You don't have to re-cite long chapters. Assume the teacher (and evaluators) knows the theory. You should show how you considered the theory during your design and implementation process.
2. Describe some hypotheses and findings that one can see using your tool. Here you should not go deep into machine learning, statistics etc. You should rather show how your tool is useful to get insight.
3. The assumptions you made about the unclear attributes in the data.

Reflection is also your chance to justify some of your choices that are counterintuitive but probably make sense if you explain them to the readers (examiners) properly.

In addition, reflect on the following questions (these are used only for course feedback and improvement, will not be graded):

4. What were the most difficult things?
5. How much time did it take? Try to estimate it in hours.
6. Do you feel there is something missing? What other things should be included in the assignment?

Grading

This assignment has a level of detail different from the previous. The main reason is to allow you to work creatively. Visualization is something which cannot be described in strict rules. However, there are some principles that will be considered when grading your assignment, therefore you should think about that when working on your solution:

1. (15%) Does your solution cover the user's needs? Can it help to answer relevant research questions? You may have a nice and pleasant visualization, but if it does not allow the user to do his/her work, the value is limited.
2. (10%) Can you as a user give some examples of interesting results that you can find with the tool? Any hypotheses that you can either confirm or disprove? This should be done in the presentation and your reflection report. I.e., your tool has more value if you can show in practice that something useful can be uncovered.
3. (15%) Do you follow the theoretical principles? We have discussed different design rules, you should consider some of them in this assignment:
    - User interface design rules, including choice of colors, and responsiveness constraints.
    - Type of charts based on data type and focus
    - Layouts and connection between charts.
4. (15%) Your reflection answers on the applied theoretical principles and assumptions about attribute types in your solution.
5. (15%) Usability of the system. This part involves some subjectivism. However, that is a natural part of visualizations - they are always consumed by humans who have some subjective biases. The following aspects will be graded:
    - Is the system intuitive to use? Is it easy to understand the controls and navigate? Example of bad interface: a table with plain text entries which actually has sorting feature, but the user is not anyhow informed that the sorting is there (no hover effect, no arrows or buttons, no hints).
    - Is it consistent with other systems and *normal* conventions? Bad consistency example: buttons with the same icons mean different things on different pages; drop-down menu that requires double click.
    - Does it look nice. Subjective opinion of the examiners :). Worth about 5%.
6. (20%) How rich is the functionality and how finished the *product* looks? Does it explore different questions? This is rather hard to formalize. Nevertheless, a tool that allows to explore 3-4 different hypotheses and has filtering by 10 different parameters will get more points here compared to a tool that allows filtering only by season and shows a single chart.
7. (10%) How dynamic the system is? The most basic case: you have created visualizations based on the exact data you got. This is considered only a partial solution – it is usable only once. Your system should be dynamic – we could update values in the CSV files (without changing data format - the columns), run the tool again and it should always reflect the newest data.

In addition, there are some constraints that can lead to a reduced grade:

8. Lack of interaction. Ability to explore the data interactively can reveal much more insight than static images. If you provide a tool that has only static images without interaction, you get a penalty of -20%.
9. Inaccurate calculations. If you have either mathematical or logical bugs in the calculations, it may lead to reduction of grade up to -20% based on severity of the issue.

Error reporting

If you find any errors in the assignment description or dataset, let the teachers know ASAP: di.wu@ntnu.no.

Suggested tools

There are many possible ways how you can implement the visualization. Choose one which you like best. If you are not familiar with the tools, make sure you reserve enough time to familiarize with them! Expect that the familiarization may require 2-3 weeks (20+ hours)!
Some suggested Python-based tools:

- Plotly (https://plot.ly/python/) - interactive charts, not possible (or very challenging?) to get connection between views, but you can work in the familiar environment of Jupyter notebooks. Easy to use and learn.
- Dash (https://plot.ly/products/dash/) - interactive dashboards, with connection between the views. You create a web-application in Python. To learn Dash, take this tutorial: https://dash.plotly.com/
- Bokeh (https://bokeh.pydata.org/en/latest/) - similar to Dash, but not based on Plotly.

Javascript-based tools (you can do data-preprocessing in Python and visualize the results in HTML+Javascript):

- Google Charts (https://developers.google.com/chart/) - easy to use, not the most flexible and extensive library, but rich enough. See a tutorial on Google Charts here.
- D3.js - probably the most flexible and rich Javascript-based data visualization library. Hard to grasp at first. See a tutorial (incomplete) on D3.js charts here.

You can choose any other tool that allows you to create interactive visualizations that satisfy all the project requirements.