CPSC 583 INFORMATION VISUALIZATION

Data Approval

The first part of your project consists of deciding on a data set that you want to work with in your project. It is important to get this part right, as this will dictate what you will be working on in your course work project for the remainder of the term. Most importantly, the data should be something that you care about. That speaks to you (perhaps look at the slides in Lecture 2).

There are several resources in the inspiration module on D2L as well.

SIZE

The point about size is important. With too small a data set, you might end up not having something relevant to show, or lacking variance. On the other hand, you don't want to work with too big a data set either. Doing so, you will struggle to decide what parts of the data set to focus on in your visualization work and consequently spend too much time on this.

Thus, I am asking you to choose a limited and tightly scoped set of data. I am doing so mostly for your own sake. I want you to work with the visualization part of the course and steer you in direction that helps you to show that you have got that part down.

- Not too small, not too big.
- For example, 200 rows and 5 columns might be ok.
- Less or more could be ok too (the right data set can be 1000 rows – short answer; it depends).
- Ask if in doubt.

CONSIDERATIONS

Many data sets have can have empty cells, i.e. missing values. One question that arises when that is the case, is how to represent this meaningfully. Often, people resort to showing empty values in the same way that zeros are shown, and when using summary statistics (for example, average and total) just leaving them out. These solutions rarely address the issue well, so you might consider how this could be handled in your concrete situation.

The point here is that missing data might be just as, or perhaps even more important than the other parts of the data. And that there might be different underlying causes for missing data.

For example, for a cross-country healthcare data, it might be the case that the process of data collection differs between the countries. By visualizing this data set, it might be clear that countries with a more developed healthcare system have higher rates of something that you would expect less of in a more developed system. Such issues might occur because these healthcare systems are better at collecting the information and not because it happens more frequently. However, the opposite might also be true (that in fact, it happens more often).

SUBMISSION

Each student selects one data set. Submit the concrete data file (csv or excel) to the D2L Dropbox. If you find a need to submit in another format, then please ask for permission.

Each student also submits one pdf document. This document contains:

- A short description (~200 words) of each of the data sets (what does the rows/columns mean; how is the data collected; what is the source, etc.) and why you think it is interesting to work with.
- A short discussion (100-500 words) of the pros and cons of working with each of the data sets. This should both discuss technical and semantic considerations. For example, a data set might only contain one type of data (say, nominal), which might be considered a con. Likewise, a data set might say something about an interesting topic, but perhaps only portray the topic from one perspective, or the provenance of the data might be questionable.
- A brief concluding paragraph (~100 words) stating which of the data sets the student suggests working with and a brief argument for their choice.

What you put in the document could be part of your final project hand-in.

DEADLINE

Deadline evening on Sunday January 26 (11:59pm Calgary time).

NEXT STEPS

I will look at all of them at let you know in D2L whether it is approved.

You need approval before moving on. I might ask you to revise and resubmit the data.