Semester project

CMSE 402, Data Visualization Principles and Techniques
Spring 2024

Purpose: The purpose of this semester project is for you to integrate many of the visualization skills and tools that you have learned this semester to make an evidence-based argument in a subject area that you care about.

Overview: One of the primary uses for data visualizations is to provide information for decision-making. This is true in business, research, and many other fields. In this project, you are going to create a series of data visualizations to convince someone to take evidence-based action on a subject that you care about. The subject can pertain to your degree program, research interests, or personal interests, and the person can be hypothetical – a research collaborator, a lawmaker or policy maker, an employer, etc. You will present these figures as part of a digital poster (made using PowerPoint, Keynote, Prezi, Canva, etc.) or data visualization dashboard (e.g using Tableau or PowerBI; or you could experiment Prezi's dashboard functionality as well) that you will share with the rest of the class at the end of the semester. You have a great deal of flexibility to define the project according to your interests and needs, but it must include (I) one or more complex datasets, (II) a clearly-articulated argument, supported by a set of research questions, and (III) a handful of carefully chosen visualizations that answer the questions you have decided on.

Deadlines: There are several deadlines for this project. Each deadline is associated with a graded component of the project, and the general purpose of this timeline is to help you to keep on track and to get feedback from your colleagues at critical points in the project's development. The deadlines are as follows, **make sure you pay attention to the appropriate days/dates given the section you're in**:

- March 13/14 (Wednesday/Thursday) – Two project ideas are due **before the start of class**.
- March 29 (Friday), end of day – Written check-in regarding your dataset(s) and analysis thus far.
- April 3/4 (Wednesday/Thursday) – Draft versions of data visualizations (for peer feedback) due **before the start of class**.
- April 15/16 (Monday/Tuesday) – Draft of poster (for peer feedback) due **before the start of class**.
- April 21 (Sunday), end of the day – Final version of project is due.
- April 26 (Friday) end of the day – Final write-up is due.

Project details: The project is composed of several components, as detailed below. The central point is that you need to create **at least four data visualizations** that have been carefully chosen to show the most information possible, as clearly as possible, and to answer the questions you have chosen to explore. These data visualizations should use as many of the data visualization types and techniques that you have learned about as is practical, and should be tailored to be effective for use in your chosen presentation format.

- **Two project ideas:** Come up with *two substantially different ideas for projects*, including brainstorming/identifying the dataset(s) you'd use, the types of questions you might ask, and the argument(s) you might make based on the dataset(s) and questions. If you don't already have some datasets in mind, please refer to the list at the end of this document. Summarize these two different project ideas in a PDF, text, or Markdown file, and submit them to the repository. You will present these ideas to your group members in class, receive feedback from them regarding your project ideas, and choose one of the ideas as your final project to move forward. **Note:** before submitting this part of the assignment, make sure to download and examine the dataset(s) you are interested in to make sure it has the type of information you think it does, and that it'll be relatively easy to work with! The document describing your project ideas should be put in the directory `project_ideas`. **Do not commit your datasets to the repository!** If you're sick of having Git complain about your datasets, you can create a `.gitignore` file that tells Git you do not want to commit those files.

- **Written check-in:** In this part of the project, you should submit a brief written document (in PDF, text, or markdown format) that explains which of the two projects you decided to pursue, and what progress you have made so far in analyzing the data. Also, list the final version of the questions you have decided to ask of the data. This should be put in the directory `written_checkin`. You will receive feedback on this from your instructor on this aspect of the project.

- **Draft of visualizations:** In this part of the project, you are submitting drafts of your data visualizations in order to get feedback from your peers in class. Note that it's fine to create an animation or an interactive plot, but **keep in mind what your chosen format will be for your final presentation of your results.** If you are making a digital poster, you likely will not be able to easily embed interactive visualizations and thus should make primarily static visualizations. That said, you could always supplement your static presentation with an interactive or animated visualization. If you're using a data visualization dashboard, it maybe easier to create interactive visualizations (but you may find the limitations of the dashboard visualizations to be overly constraining or frustrating to customize.

  The draft visualizations should be put in the directory `draft_visualizations`, and you will refer to them in class.

- **Draft of presentation:** In this part of the project, you are creating a draft version of your digital poster or data visualization dashboard. This draft should be as complete as possible and should have all of the text and visualizations that you think are appropriate.

  Note that, when making a digital poster, there are "standard" components that comprise a research poster (information can be found here and here), and you can use PowerPoint as a poster template if you're not familiar with other poster-making software. You could also try your hand at the Poster 2.0 approach that popped up here and there on the internet a few years back. You can also use Keynote, Prezi, or Canva to make your poster, but regardless of the software you use, it should contain much of the same content.

  If you choose to make a visualization dashboard, you should still be striving to convey similar information as one might include in the poster, but you should also be thinking about how the ability to interact with your dashboard might allow someone to enhance their understanding of your results as well. If your final dashboard is just as valuable as a static image of the same content, then you probably haven't leveraged the true functionality that the tool has to offer. You can see a great example of a Tableau dashboard here.

Your final presentation product (whether a poster or a dashboard) needs to be understandable as a standalone object, but you should also think about how you would use it as a tool if you were giving a brief presentation to another person. You should think about how much information you can convey in your chosen format as well as the relative sizes of the visualizations and their supporting text such that your audience can readily extract information from your poster. If you make the figures or text too small, it might be challenging for someone to digest your message when viewing your presentation!

The presentation draft should be put in via the directory `draft_presentation`.

- **Final version of presentation:** This should be put in the directory `final_presentation`, along with a text description of the feedback that you received during the peer review session and the changes you made as a result.
- **Final writeup:** The final writeup is a brief summary of what you did: what dataset(s) did you use, what questions did you ask, and ultimately what was the argument you made? What do you feel you learned as a part of this project? What went as expected, and what did not? It only needs to be a page or two long – the point is to summarize and reflect on your experience, not to present all of the material again. Note that this should be in PDF, plain text, or markdown format, and should be put in the directory `final_writeup`.

Handing in the project components: Turn in all files relating to the semester project using the GitHub classroom repository where you found these instructions. Make sure to turn in all of your code (to create your draft and final figure versions), your figures (both draft and final versions), your poster and slides, and your writeup by pushing your commits by the specified deadlines. When I send out feedback (which will be returned as a text file in the relevant subdirectory), if you're working with your repository on the command line, you can get the feedback file by first committing all of your changes to the repository, then typing "`git fetch`" and then "`git merge`". You should also be able to view the feedback file through the GitHub web interface, if you prefer.

If you're looking for links to possible data sources, scroll to the next page!

Possible places to find datasets:

- dataquest.io meta-list of places to find datasets
- Kaggle – lots of datasets on miscellaneous topics.
- Data Hub
- Open Data Inception – 2600+ open data portals around the world
- Large list of public domain datasets from "awesomedata" on GitHub
- data.gov – enormous database of datasets
- US census datasets (you can also use their new tool for exploring their datasets and you can get more fine-grained data like this one on the same website)
- public datasets on Google Cloud
- Amazon AWS public datasets
- StackExchange open data topic
- Quora answers the question "Where can I find large public datasets?"
- fivethirtyeight.com datasets (from articles on FiveThirtyEight)
- All open-source data from Buzzfeed News
- r/datasets on reddit
- R datasets – an archive of datasets distributed with the R language (mostly small datasets, though)
- City of Chicago data portal