ENG 395
Spring 2021

Dataset Analysis -- 15%

- **Due Friday, Feb 26**
  - **Make your dataset selection by Friday, Feb 19**
  - **Complete a "rough draft" of your dataset biography by class on Wednesday, Feb 24**
- **Dataset biography + 900-1200 word (~3-4 page) reflection**
- **MLA citation style**
- **Turn in via "Dataset Analysis" portal on Blackboard Assignments page. You should turn in both your dataset biography and your reflection paper via this portal.**

Your dataset analysis will consist of a description of and a reflection on an extant dataset. You will select an already existing dataset to focus on; you will complete a dataset biography of that data; and you will write a 900-1200 word (~3-4 double spaced pages) reflection on this dataset.

Selecting your dataset
At the bottom of this assignment sheet, I have provided a list of datasets you may want to work with for this assignment. I recommend using this list to select your dataset. If none of the datasets below interest you and/or if you have an idea about the specific dataset you might like to use, you may also find your own dataset to use for this assignment. However, I will need to approve this choice.

No matter what dataset you use for this assignment, you should select one you want to learn more about and that you are interested in spending some time with. It should also adhere to the below criteria:

- It should be collected by someone else/an organization (i.e., not you).
- It should be "conceptually contained," meaning all of the data should "go together" in some way. For instance, the data in the UN's World Contraceptive Use dataset is composed of data from reporting countries on contraceptive use among their populations. The dataset you select should have obvious principles of inclusion.
- It should fit broadly under the remit of the arts, humanities, and/or social sciences. In other words, datasets from/geared toward answering questions about science will likely be less useful for this assignment, but talk to me if you have a specific one in mind that you think could work.
- You should be able to access the majority or the entirety of the dataset (or at least the portion you are examining).
- You should be able to find at least some documentation of how the data was collected and who collected it.
- Your dataset should be medium-sized to large-ish. What this means will vary by dataset, but generally speaking, your dataset should include hundreds of individual observations at least. However, it should not be so large that you can't download it all onto your computer or examine enough observations to understand your data. If the dataset you want to work with is large enough that examining it in its entirety is too burdensome or time-consuming, you may want to select a smaller subset of the data to use for this assignment.

- Ideally, you should be able to download the dataset and examine it. At a minimum, if you can't download the data directly, you should be able to examine it in some way. How you examine your data is up to you: this might mean reading through rows of an Excel sheet, or using a data visualization program or platform to visualize your data in various ways, or using a programming language to do this. The easiest thing is often to download your data as an Excel file or as a csv (which you can then open in Excel/Google sheets); if you have less technical experience, look for this option when deciding what dataset to select.

You will indicate which dataset you have selected for this assignment as part of response paper 3, which is due Friday, Feb 19. If you want to work with a dataset NOT included in the list below, I strongly recommend you get in touch with me sooner than this about your choice.

Completing your dataset biography

Once you have selected your dataset, you will complete a dataset biography, as described by Heather Krause in "Data Biographies: Getting to Know Your Data" (https://gijn.org/2017/03/27/data-biographies-getting-to-know-your-data/). I have provided a template for you (linked on our course website and stored in our class Google drive folder), and you will fill this template out for your selected dataset. Completing this portion of the assignment will require you to investigate what your data is, where it comes from, who collected it, how it was collected, and why it was collected. Your dataset biography is meant to help you familiarize yourself with your data in a structured way. You will start working on your dataset biography as part of response paper 3, which is due Friday, Feb 19.

In order to complete your dataset biography, you will need to be able to explore your data. There are many, many ways to explore a dataset, and the simplest (and also often the most effective, depending on your data) is just to open up the data in Excel/Google sheets and start reading. Additionally, some of the datasets linked below are part of projects that provide visualizations of the data or applications for exploring it (i.e., the Slave Voyages data). You might find these useful in completing this assignment. Additionally, on Wednesday, February 17 and Monday, February 22 we will spend time in class together learning how to use some basic functions in Tableau (https://www.tableau.com/academic/students) to explore sample datasets. Tableau is free for students.

Here is how to complete your dataset biography:

1. Download the Dataset Biography Template from our course site or our class Google drive folder (Syllabus and Assignments folder). The template was originally created by Heather Krause; I have made some small revisions for this assignment.
2. When you open the template in Excel/Google sheets, you will see that it contains 2 tabs ("general info" and "fields or variables"). The "general info" tab is the one you should fill out first. There are 2 completed rows here. These are examples showing you how you might fill the template out. You can delete them when you hand in your dataset biography.
3. Depending on the dataset you have selected for this assignment, it might make sense to fill just one row out (considering your dataset as a whole), or it might make sense to consider different parts of your dataset separately. In the template, you can see that one of the example rows is more general -- it considers the UN Violence Against Women dataset as a whole -- and one is more granular -- it considers just data from Malawi. In general, the more specific you can be, the better.

4. After you fill out the "general info" tab, move to the "fields or variables" tab. This tab is empty. Here, you should copy and paste each field or variable name from your dataset, one per column (row 1), and provide definitions of each field/variable in the cell below (row 2). The feasibility of this task will depend to some extent on the size of your dataset and how many fields/variables it includes (talk to me if you have questions about what you should include if there are lots and lots). In general, if you imagine your dataset as a spreadsheet, the "fields" or "variables" refer to the column headers (descriptions or facets of each data point in your dataset), while your dataset's "observations" are comprised by each separate row (individual data points). What I am asking you to do here is basically to provide definitions for each field/variable included in your dataset (and its metadata). Ideally, you will be able to take these definitions from the dataset's documentation.

Writing your reflection

The final step is to write a reflection about your dataset (~900-1200 words). How you structure this reflection is up to you, but you should focus it on one or two of the most salient or important issues you discovered or learned from examining your dataset in detail. In your reflection, you should seek to answer the following questions (roughly, try to answer/discuss at least one question from each numbered sets of questions below):

1. What were you surprised to learn about this dataset? Or, what is something that was not fully apparent at first glance about this dataset, but that you came to see as important as you learned more about your data? Or, what is something that you think is missing from your dataset or that was overlooked or not fully thought through when your data was collected (if you are interested in this question, you should also consider what it would take to remedy this situation: is it possible to collect what has been overlooked? Why or why not? Were the creators of this dataset aware of this limitation?)?
2. How does this thing change how you think about your dataset, or what you see as most important about your dataset?
3. What would you change about how your dataset is presented or described, and/or how it was collected, to account for this? Or, knowing what you know now about your dataset, what kinds of questions does it allow us to answer/what can we learn from this dataset?

You should relate your discussion to at least 1 of our readings from class so far.

Datasets

The datasets below have been created for and by researchers, journalists, and/or policy makers. Some of them assume knowledge of the conventions of specific research fields. As always, if you have questions about your dataset and what it contains, I am happy to discuss them with you.

- Slave Voyages data (select one):
  - Trans-Atlantic: https://www.slavevoyages.org/voyage/database
  - Intra-American: https://www.slavevoyages.org/american/database
  - African Names: https://www.slavevoyages.org/resources/names-database
  - All of the above pages include options to download data in Excel form (and/or csv). You cannot download all of the columns in the Trans-Atlantic and Intra-American datasets, but if you select Download > Excel > All results with visible columns, you will get a lot.

    - Alternatively, you can download the entire database (and previous versions), as well as the 2019 code book, here: https://www.slavevoyages.org/voyage/downloads#full-versions-of-the-trans-atlantic-slave-trade-database/0/en/. HOWEVER, note that the database is an SPSS file (.sav), which is a file format made for proprietary SPSS software. I have converted the latest SPSS file (2019 release) to a csv file and placed the csv version in our class Google drive folder, in the Syllabus and Assignments folder. You will want to download the code book from the link above in order to understand this data ('tastdb-exp-2019.csv').
- Hannah Anderson and Matt Daniels, Polygraph's Film Dialogue Dataset: https://github.com/matthewfdaniels/scripts/; sources Google doc: https://docs.google.com/spreadsheets/d/1fbcldxxyRvHjDaaY0EeQnQzvSP7Ub8QYVM2bIs-tKH8/edit#gid=1668340193
    - The data is discussed here (we will discuss this project in class): "Film Dialogue from 2,000 screenplays, Broken Down by Gender and Age," *The Pudding*, April 2016, https://pudding.cool/2017/03/film-dialogue/
- Torn Apart/Separados open data: https://github.com/xpmethod/torn-apart-open-data
    - The Torn Apart/Separados project is here (we will also discuss this project in class): https://xpmethod.columbia.edu/torn-apart/volume/1/
- Mapping Police Violence: https://mappingpoliceviolence.org/ (scroll down and click on Download Full Database)
    - Read About the Data here: https://mappingpoliceviolence.org/aboutthedata
- Atlas of Surveillance Data Library: https://atlasofsurveillance.org/library
    - Click "See Dataset" under "Atlas of Surveillance" to download their data in csv form. Click "Methodology" to learn more about how they collected their data.
- The Association of Religion Data Archive, https://www.thearda.com/Archive/browse.asp
    - There are over 1,000 datasets available through this archive; some of them are historical and some are more contemporary. I have not reviewed all of them. But in general, what you're looking for are:
        - Datasets in Excel/csv format;
        - Datasets with documentation (sometimes called a Codebook) that explains each observation/variable in the dataset or that makes it possible for you to infer their meaning;
        - Bonus: Links to research articles that use or explain the data in more detail.
- The New York Public Library's Restaurant Menu collection: http://menus.nypl.org/menus
    - Learn about the data: http://menus.nypl.org/about
    - Download the data: http://menus.nypl.org/data
- Chronicling America pre-packaged datasets: https://news-navigator.labs.loc.gov/ (look under the "Pre-Packaged Datasets" heading)
    - The pre-packaged datasets include various kinds of visual content (photos, illustrations, maps, comics, etc) from Chronicling America's Newspaper Navigator dataset, all from 1905, and their corresponding metadata. The visual content is generally included in a .zip file, and the metadata is in both json and csv format. You may want to select different kinds of visual content to complete the assignment, or you may want to focus on just one.
    - Learn about Chronicling America here: https://chroniclingamerica.loc.gov/about/

- 2014 snapshot of the Tate Collection: https://github.com/tategallery/collection
    - This repo includes metadata for ~70,000 artworks owned or jointly owned by the Tate Museum. It also includes metadata for ~3,500 artists. It was last updated in 2014. It does not include the artworks themselves.
    - A number of people have used this metadata for various applications, which you might find useful in completing this assignment. The page above includes a list, but Florian Kräutli's visualizations are a good starting point: http://research.kraeutli.com/index.php/2013/11/the-tate-collection-on-github/.
- UN Data Explorer: http://data.un.org/Explorer.aspx
    - LOTS of different kinds of datasets available here; some may be useful for this assignment and some less so. I have not reviewed all of them. In general, the data has been collected by organizations like the UN, the IMF, and the World Bank. Sometimes, in the examples I have reviewed, it can be difficult to find documentation about what each variable means. Again, you're looking for:
        - Datasets in Excel/csv format;
        - Datasets with documentation (sometimes called a Codebook) that explains each observation/variable in the dataset or that makes it possible for you to infer their meaning.
- 20th-Century American Bestsellers: http://bestsellers.lib.virginia.edu/
    - This data is not available in Excel/csv form; it is only possible to examine the data using the website interface. Still, it would be possible to complete your assignment using this dataset.
    - Learn a bit more about the data here: http://bestsellers.lib.virginia.edu/help/credits/
    - Learn about a 2003 exhibit using the data here: https://www.jstor.org/stable/20864015?seq=1#metadata_info_tab_contents
    - Learn about a 2013 adaptation of the 2003 exhibit online: https://explore.lib.virginia.edu/exhibits/show/bestsellers