# Telling Stories with Data Final Project

PUBHLTH 460: Spring, 2021
taught by Prof. Nicholas Reich

## Overview of the project assignment

For the project you will create, in small groups, a thorough analysis of one of several large datasets recommended by the instructor. The goal is to tell a story using data about the topic of your dataset. The project will have components completed by the group and by each individual student.

Each group is required to hand in the following items on Moodle by 11:59pm Eastern time on Wednesday, May 5, 2021:

- A ten-slide presentation summarizing the key findings.
- A 500 word technical write-up describing the details of your group's analysis.
- A video of a no more than 10 minute presentation. (exact submission details TBD.)

Additionally, each individual student will hand in a separate detailed write-up that describes the analyses that they contributed to their group analysis (see details below).

The group portion of the project will be worth 60% and the individual part 40% of the final grade.

## Choices of a primary dataset and motivating question

- Opioid data from the CDC Wonder database
  - Use your analytical and data visualization skills to recommend measures that could take steps toward resolving or reducing the national opioid crisis.
- COVID-19 data from COVIDcast
  - Use your analytical and data visualization skills to recommend measures that could reduce the impact of the COVID-19 pandemic in the US.
  - Use your analytical and data visualization skills to identify community-level risk-factors for high case, hospitalization, and death rates in the US.
- Data from the HELP study, on physical/sexual violence and substance abuse (see `data-stories-2021/data/HELP` on Google Drive for more info and data)
  - Use your analytical and data visualization skills to identify risk factors related to physical and sexual abuse that are associated with outcomes related to substance abuse.
- Data from the Framingham Heart Study, a landmark epidemiological study on heart health in human populations (see `data-stories-2021/data/FRAMINGHAM` on Google Drive for more info and data)
  - Use your analytical and data visualization skills to identify risk factors associated with poor heart health outcomes.
- If you have a particular dataset that your group is interested in, you may petition the instructor to allow you to use it. In general, datasets should be associated with human health in some way and should arise from a real-world setting.

Teams are encouraged to use external data sources to supplement their primary analyses.

## Guidelines for the group write-up

**Overall outline:** Your team will create a story that contains a compelling central narrative. Your job is to convince the reader that your story is important and interesting. You must convey an understanding of the broader societal context of the data you are analyzing. Using sophisticated methods is not a pre-requisite for a strong project. As we have seen throughout the semester, using appropriate and simple visualization techniques is often the key to telling an effective story. While a regression or other modeling analysis may be a valuable supplement to your story, it is not a required element of the project.

**Working as a group:** Your group should assemble an outline of the key elements of the story that you want to tell and which team member will be responsible for each element. The general idea is that each element should focus on one key observation or insight about the dataset. Tell a short, compelling story with a small number of elements. The elements should complement each other and together tell a coherent story about your dataset. Elements could be data visualizations, regression analyses, integrations from other datasets, or some other quantitative piece of the story.

**Evaluation:** As a team, you are encouraged to pay close attention to the evaluation rubric provided for the project (see below).

## Guidelines for the individual write-up

Each member of the group will serve as the lead on one or more elements that tell a story about the assigned dataset. In addition to incorporating these elements into the final group deliverables, each individual student will produce a separate write-up (including up to 4 total tables and figures but no code) about their individual analyses. These individual analyses should provide more technical detail about the analyses performed as well as giving a brief introduction and conclusion to the analysis. Each write-up should stand on its own, providing tables and figures as necessary.

The individual data analysis write-ups will be due Friday May 7th at 5pm ET, to be handed in on Moodle. Individual analyses should be handed in as HTML files knitted using RMarkdown. No code should be displayed in these reports, however, figures resulting from your analyses should be dynamically created.

## Grading

Your project grade makes up 25% of your final grade for the class. The grading rubric for the group portion of the project will be evaluated based on the Rubric (see below).

**Participation:** To evaluate individuals' contributions to the group, I will be using the following approach to evaluate each of your contributions to the project. Each student will be given 100 points to allocate among your teammates (excluding yourself). The more points you give to a teammate, the more you are indicating they contributed to the project. You cannot allocate the same number of points for any two team members. I reserve the right to intervene to correct gross imbalances in allocations if necessary. The number of points that you receive from your teammates will be summed, divided by 100, and then used as a multiplier on the final grade for the 60-point group component of the project.

**Example:** Your group receives 50/60 points for the "final product produced by the group". You have three teammates who give you scores of 35, 40 and 30, respectively. Therefore, you receive a total of 105 points from your teammates. So your final "group" grade is (50/60) * (105/100) = 0.875 = 52.5/60.

## Grading Rubric

There are five main criteria on which your group project will be evaluated (60 total points):

- Narrative (20 points)
- Evidence and methods (15 points)
- Graphics (15 points)
- Organization (10 points)

And two main criteria on which your individual components will be evaluated (40 total points):

- Overall quality of analysis (30 points)
  - Rubric: correct implementation and interpretation of method(s) used, appropriate use of equations to show what methods/models have been used (if applicable), appropriate use of graphics/tables to support central results, succinct summary of key results.
- Clarity and presentation (10 points)
  - Rubric: clear statement/summary of goals and central results, clear and accurate description of methods/models used, use of figures rather than text to illustrate central ideas, figures dynamically generated within the RMarkdown file, figure limit adhered to.

For judging each criteria there are point ranges for Poor, Fair, and Exceptional.

These criteria have been adapted from the 2018 Public Health Data Challenge Judging Criteria document.

## Narrative (20 points)

Your presentation should provide a narrative that engages the audience and holds their interest. This is done through constructing a strong theme that is driven with compelling facts and a connection to the overarching context of your datasest. Your points for this section will be based the work in your presentation.

| Point Range | 1-4 Poor | 5-15 Fair | 16-20 Exceptional |
|---|---|---|---|
| Attributes | Poor Strength of Story<br>- Did not convince audience of importance/impact<br>-Presentation disconnected with intended audience<br>- Does not demonstrate an understanding of the city's context uninteresting<br>- Does not display enthusiasm for their project<br>- analysis lacks applications | Fair Strength of Story<br>- convinced audience of importance/impact, but could be stronger<br>-Presentation moderate connection with intended audience<br>- Demonstrates a moderate understanding of the city's context interest<br>- Displays moderate enthusiasm for their project<br>- analysis has few applications | Exceptional Strength of Story<br>- convinced audience of importance/impact<br>- "wow factor"<br>-Presentation well connected with intended audience<br>- Demonstrates a strong understanding of the city's context<br>- Clear communication Interesting<br>- Displays strong enthusiasm for their project<br>- analysis has clear applications |

## Evidence & Methods (15 points)

The group presentation and overall narrative must be supported by the data through properly conducted, in-depth analysis and carry objectively correct interpretations of evidence. The score for this criteria will be based on the technical write-up and your presentation.

| Point Range | 1-3 Poor | 4-12 Fair | 13-15 Exceptional |
|---|---|---|---|
| Attributes | Poor Evidence<br>- Statements not backed by data<br>- Statements refuted by data<br><br>Poor Depth<br>- Only very simple data attributes plotted<br><br>Poor Methods<br>-inadequate/superficial or improper/overly-simplistic methods used<br>- Incorrect interpretations of modeling/graphical items | Fair Evidence<br>- Questionable statements about the data<br>- Some statements not supported by data<br><br>Fair Depth<br>- Relatively simplistic data relationships visualized<br><br>Fair Methods<br>- questionable methods used<br>- Lacks sophistication<br>- Some incorrect interpretations of modeling/graphical items | Great Evidence<br>- Statements backed by data<br>- Legitimate findings<br><br>Great Depth<br>- Complex analysis and data relationships visualized<br><br>Great Methods<br>- Appropriate/sophisticated methods used properly<br>- Correct interpretations of modeling/graphical items |

## Organization (10 points)

Your presentation of the materials should be organized in a logical progression for the audience to follow. Additionally it will be important to use terminology and phrasing that are clearly understandable. Your visual materials used in the presentation should be well organized so that an audience could easily follow along. You are limited in the number of slides present and you should imagine a time limit on an oral presentation, so don't overly complicate/clutter your visual materials. Same with the technical write-up, you have limited space, so it is important to be clear and succinct in your presentation of methods. All materials should contain correct spelling and grammar.

| Point Range | 1-2 Poor | 3-8 Fair | 9-10 Exceptional |
|---|---|---|---|
| Attributes | Poor Progression<br>- Difficult to follow train of thought<br>- Results presented out of order<br><br>Poor Lexicon<br>- Use lots of unclear jargon, phrases or terms that would be unfamiliar to audience<br><br>Clutter and Complication<br>- Huge volume of words<br>- Poorly organized items<br>- Visual does not match presentation<br>- Many typos | Fair Progression<br>- Fairly clear train of thought<br>- Results presented in reasonable order<br><br>Fair Lexicon<br>- Use little unclear jargon, phrases or terms that would be unfamiliar to audience<br><br>Little Clutter and Complication<br>- reasonable volume of words<br>- organized items<br>- Visual matches presentation<br>- Few typos | Great Progression<br>- Intuitive order to presentation materials<br>- Covered all important topics<br><br>Great Lexicon<br>- Use no jargon that would be unfamiliar to audience<br>- clear/concise language used.<br><br>No Clutter and Complication<br>- Concise/clear wording<br>- Well organized items<br>- Visuals complement presentation<br>- Few typos |

## Graphics (15 points)

The use of graphical aesthetics (color, shape, size, position, orientation, etc.) are appropriately used given the context of the data. The scales associated with these attributes should also be contextually appropriate. The graphics used contribute to the strength of the narrative and improve the ability to connect to the data in meaningful ways.

| Point Range | 1-3 Poor | 4-12 Fair | 13-15 Exceptional |
|---|---|---|---|
| Attribute | Poor Depth<br>- Only very simple data attributes plotted<br>Poor Impact<br>- Data visualization did not contribute in a meaningful way to the presentation<br><br>Poor Aesthetics<br>- Use continuous graphical aesthetic to display categorical data element, or vice-versa<br>- Distractingly poor graphical Choices<br><br>Poor Scaling<br>- numeric scales improper for the context of the data | Fair Depth<br>- Relatively simplistic data relationships visualized<br><br>Fair Impact<br>- Data visualizations contribute moderately to the presentation<br><br>Fair Aesthetics<br>- Generally avoids using continuous graphical aesthetic to display categorical data element, or vice-versa<br><br>Fair Scaling<br>- numeric scales proper for the context of the data | Great Depth<br>- Complex data relationships visualized<br><br>Great Impact<br>- Data visualizations contribute in a major way to the presentation<br><br>Great Aesthetics<br>- Appropriately uses graphical Aesthetics<br><br>Great Scaling<br>- numeric scales proper for the context of the data |