CMSE 201 Data Analysis Project

As one part of your semester, you'll turn in a full-fledged analysis of a publicly available dataset. This is your chance to find and understand data that matters to you.

You'll submit that final analysis as a Jupyter notebook.

Steps to Creating your Analysis

You'll need to:

1. Identify a public dataset you care about. (Start with the resources below; course staff can also help you with this.)
2. Make sure you can load the data into a Jupyter notebook using Pandas. (This step may involve some data cleaning and transformation.)
3. Conduct exploratory data analysis to get a sense of the data.
4. Pose three interesting questions about the dataset and answer them using whatever combination of plots or computations you see fit.
5. Submit your final project, including your questions, code, plots, and answers.

Resources for finding and selecting datasets:

- Data.gov - Federal
- Data.gov - State of Michigan
- Kaggle
- Data.world

Important Deadlines and Things You'll Hand In

Wednesday, Oct. 12 or Wednesday, Oct. 13 - in-class discussion

You'll bring in to class ideas for two different public datasets, as well as some questions that you think you might be able to answer with those datasets. In class, you'll share those datasets and questions with members of your group, who will give you feedback on them. You'll then choose the one you think is more promising to share with the entire class.

Friday, October 14 - One-Page Project Description

You'll hand in a description of the dataset you've selected, which will be no more than one page. Your description should include:

- What the data is.
- Why you chose it.
- What you're interested in about it.
- What sorts of questions do you want to answer with this data.

Include a link to the dataset and any other relevant web page, and any supplemental information that you think may be helpful in us giving you feedback.

Friday, October 28 - Preliminary Analysis Notebook

You'll hand in a Jupyter Notebook with the beginnings of your analysis. Your notebook needs to be able to:

- Load your chosen dataset
- Produce at least one visualization you can explain
- Include **three** questions you hope to answer using the data
- A preliminary answer to **one** of the three questions you're trying to address (meaning, analysis and plotting completed, but not necessarily as nice as you might want it to be).

Friday, November 11 - Final Analysis Due

You'll hand in a Jupyter Notebook with your final analysis. It should contain both code and explanation that leads your readers through:

- The dataset you chose
- Why you chose it
- Your three research questions and what you found by exploring the datasets
- Any and all computations and visualizations you created

You will be graded on your research questions (more challenging and interesting questions will receive more points) and the way in which you answer the questions (more sophisticated analysis and informative/aesthetically appealing plots will receive more points).