Assignment 2: Exploratory Data Analysis

In this assignment, you will identify a dataset of interest and perform exploratory analysis to better understand the shape & structure of the data, identify data quality issues, investigate initial questions, and develop preliminary insights & hypotheses. Your final submission will take the form of a report consisting of annotated and/or captioned visualizations that convey key insights gained during your analysis process.

Step 1: Data Selection

First, pick a topic area of interest to you and find a dataset that can provide insights into that topic. To streamline the assignment, we've pre-selected a number of datasets included below for you to choose from (see the Recommended Data Sources section below).

However, if you would like to investigate a different topic and dataset, you are free to do so. If working with a self-selected dataset and you have doubts about its appropriateness for the course, please check with the course staff. Be advised that data collection and preparation (also known as *data wrangling*) can be a very tedious and time-consuming process. Be sure you have sufficient time to conduct exploratory analysis, after preparing the data.

After selecting a topic and dataset – *but prior to analysis* – you should write down an initial set of **at least three questions** you'd like to investigate. These questions should be clearly listed at the top of your final submission report.

Part 2: Exploratory Visual Analysis

Next, you will perform an exploratory analysis of your dataset using a visualization tool such as Vega-Lite or Tableau. You should consider two different phases of exploration.

- In the first phase, you should seek to gain an overview of the shape & stucture of your dataset. What variables does the dataset contain? How are they distributed? Are there any notable data quality issues? Are there any surprising relationships among the variables? Be sure to perform "sanity checks" for any patterns you expect the data to contain.

- In the second phase, you should investigate your initial questions, as well as any new questions that arise during your exploration. For each question, start by creating a visualization that might provide a useful answer. Then refine the visualization (by adding additional variables, changing sorting or axis scales, filtering or subsetting data, *etc.*) to develop better perspectives, explore unexpected observations, or sanity check your assumptions. You should repeat this process for each of your questions, but feel free to revise your questions or branch off to explore new questions if the data warrants.

Final Deliverable

Your final submission should take the form of a sequence of images – similar to a comic book – that consists of **8 or more** visualizations detailing your most important insights.

- Your "insights" can include surprises or issues (such as data quality problems affecting your analysis) as well as responses to your analysis questions. Where appropriate, we encourage you to include annotated visualizations to guide viewers' attention and provide interpretive context. (See this page for some examples of what we mean by "annotated visualizations.")
- Each image should be a visualization, including any titles or descriptive annotations highlighting the insight(s) shown in that view. For example, annotations could take the form of guidelines and text labels, differential coloring, and/or fading of non-focal elements. You are also free to include a short caption for each image, though no more than 2 sentences: be concise! You may create annotations using the visualization tools of your choice (see our tool recommendations below), or by adding them using image editing or vector graphics tools.
- Provide sufficient detail such that anyone can read your report and understand what you've learned without already being familiar with the dataset. For example, be sure to provide a clear overview of what data is being visualized and what the data variables mean. To help gauge the scope of this assignment, see this example report analyzing motion picture data.