Project 1
DATA 110
Saidi

Project 1

For your first project, you will use some of the techniques you learned during the first 6 weeks to create a data visualization appropriate for that dataset, then write about the data and what the visualization shows. Here are your steps in detail:

1. Find a dataset (be sure to note the source of the data). You can refer to the list of possible sources in Blackboard → Course Resources. You may also select from other sources. ***You MUST get my approval for your choice of dataset BEFORE you begin working on the project.***
   Dataset submission is due by 11:59 pm on Friday, October 13th.

   What type of dataset is appropriate for this project?
   Your dataset should **include both quantitative and categorical variables**. **It should have AT LEAST 4 variables.** Sets that include dates can be helpful for plotting time series information, but this is not a requirement.

2. Start a new markdown/quarto document for your project.

3. **Before you load the data, at the beginning of your markdown document,** write a brief introduction that describes your dataset topic and the variables (you may need to define them for your audience), and establish what you plan to explore. You also MUST identify the source for your dataset (* Note: Kaggle is not a source – it is a repository)

4. Include subtitles and provide detailed comments on ALL chunks about that chunk's action or use to help your audience understand your intentions.

5. Load the necessary libraries and dataset.

6. Perform any necessary cleaning.

7. Explore the data with at least one data visualization, though you will likely have a few trivial plots as you explore the dataset at the outset. The visualization must include the following components:

   - Meaningful labels for axes
   - Meaningful title
   - Caption for your data source
   - At least 2 colors for distinguishing categorical groups
   - Change the default ggplot theme
   - Some sort of legend to make sense of colors, shapes, and sizes that describe any variables.

   Some suggestions for visualizations include side-by-side box plots, histograms, bargraphs, scatterplots, treemaps, heatmaps, alluvials or streamgraphs. The type of data you use will help determine which visualization you should use.

8. **At the end of your document,** write a second brief essay (***incorporated directly into your Markdown file***). The essay should describe:
   a. How you cleaned the dataset up (be detailed and specific, using proper terminology where appropriate).

b. What the visualization represents, any interesting patterns or surprises that arise within the visualization.
c. Anything that you might have shown that you could not get to work or that you wished you could have included.

9. Submit your data set so I can download it along with your completed project in Markdown. Knit/render your Markdown and either publish it in Rpubs or Github. The completed project should include your name, the topic as your title, the process you went through cleaning and exploring the data via comments and subtitles, the final visualization with labels, titles, and a legend, and the essay. Submit this Project in the Project Dropbox by **11:59 pm by Tuesday, October 17th.**

* *Special note: Start on this early. I am willing to help you find a dataset if you come to me early for help. If you get stuck with coding or other software, you can also contact me for help, as long as it is far ahead of the due date.*

Be prepared to present your project in class on Wednesday, October 18th. **You will be allowed to speak for no more than 2 minutes, so you must prepare in advance what you intend to speak about.**

Rubric

Evaluation	Criterion	Points Allotted
Sophisticated	Project was submitted on time. The dataset is appropriate for the project. The work is focused and clearly organized. Data is sourced. Comments and subtitles are included. Graphs show something important about the data, axes and titles are labeled, and legends are included. The language is precise and ideas are clearly and correctly communicated to the audience. All requirements are answered thoroughly and correctly. 1. Create one properly labeled and titled data visualization that includes all required elements (listed above) 2. Fully answer the intro and ending essay, based on requirements. 3. Submit final version as HTML, Word or pdf from knitted Markdown either in Rpubs or Github 4. Submit project on time. 5. Submission has been edited for grammar/punctuation/sentence structure.	100%
Acceptable	The dataset is appropriate for the project. Data is sourced. Graphs for the most part show something about the data, axes and titles are labeled, and legends are included. The language is relatively clear and communicated to the audience. Most requirements are answered thoroughly and correctly. One of the above steps above are omitted. Or two or three steps are underdeveloped.	80%
Developing Competence	The dataset is generally appropriate for the project. The visualization may be somewhat unfocused or underdeveloped but it does have some coherence. Problems with the use of language occasionally interfere with the audience's ability to understand what is being communicated. Not all requirements are answered, and/or not all answers are correct, or not all requirements are met	60%
Inadequate	Project was not submitted on time. The project has at least one serious weakness. The dataset may not be appropriate for the assignment. The visualization may be underdeveloped. Problems with the use of language seriously interfere with the audience's ability to understand what is being communicated. Not all requirements are answered, and/or not all answers are correct.	40%