Instructions for Identifying Datasets for 36-315 Final Projects, Spring 2022

Ron Yurko

Contents

Sources **1**

Data Requirements **1**

Graph Requirements **2**

For the final group project, you get to pick your own dataset. First, I will provide some sources where you can find datasets online (but you may use any source you want for finding datasets). Then, I will discuss requirements for the dataset, as well as guidelines on the type of graphs that you should make for the project. These requirements and guidelines will help you decide (as a group) which dataset to choose for your project. The dataset you choose should be something the group is excited about, because you'll be working with this dataset for the remainder of the class. Your project will culminate into a 15-20 minute presentation and a public-facing HTML file.

See here for examples of projects that 36-315 students did last year. Looking at these past projects will give you a good idea of what a "good dataset" is for a 36-315 project, and what will be expected of you. That said, I do NOT recommend that you use the same dataset that a previous 36-315 team used.

Sources

Here are some repositories with many, many datasets to choose from:

- RStudio #TidyTuesday Project
- The US Government
- More links on GitHub
- FiveThirtyEight
- American Psychological Association
- Kaggle
- Google Public Data Explorer
- Stanford Statistics
- UCI Machine Learning Repository

You can use other sources - these are just suggestions.

Data Requirements

1. Your data must be contain a mix of categorical and quantitative variables and be complex enough that you can create at least 8 interesting graphs (so datasets with only a few variables will not work).
2. You CANNOT use any of the datasets that were used in any previous assignments in this course or any other course you have taken. You must use a dataset that everyone in your group has never worked with before.

Graph Requirements

In the group project itself, you will have some requirements on the graphs you must use, which will certainly influence your dataset choices:

1. Each group is required to make **at least two** of the following types of graphs:

- Classic EDA graphs (e.g., side-by-side plots, facetted plots, scatterplots). In other words, "first half of semester" graphs.
- Dendrograms, PCA, MDS, or other clustering-based graphics.
- Choropleth maps, heat maps, or other map-type graphics.
- Time series plots.
- Graphs relating to text analysis.

In other words, you can NOT just turn in 8 classical EDA graphs. You are required to make at least one graph that we will be learning about in the second half of the semester.

(Note: We will be covering all of these visuals throughout the rest of the semester.)

2. Each group can have no more than two graphs that show a single variable (e.g. one-variable bar charts, histograms, density estimates).

3. You can **NOT** have more than 3 of the same type of graph. For example, don't just use 5 smoothed density plots.

4. **Your graphs should tell a somewhat cohesive story.** Come up with some general questions you want to answer with your dataset. **You should be able to come up with at least three interesting, overarching questions for your dataset.** You'll use your graphs to walk the viewer through a comprehensive analysis of those questions, as well as demonstrate your findings and conclusions.