MA615/415 Midterm Project

Assignment

In this assignment your goal is to acquire a dataset, clean, explore, and prepare the dataset for analysis. The datasets below have been selected for this assignment. However, you may choose another dataset as long as you get my permission for the one you choose.

Prepare a presentation of your data including a description of the dataset, a description of what you did to clean and organize the data, and summary statistics, visualizations, and notes that reflect your exploration of the data. Make you include notes on the source of the data and what you know about how it was collected and assembled, and how it is maintained. Be sure that you download and examine any metadata that is available for your dataset.

Submit your material as a link your submission in your github. You may consult with other students as you work on this project with the following conditions:

1. Only consult with students in this class.
2. Acknowledge any student you consult, describe their contribution to your results.
3. Prepare your own presentation in your own words.

Datasets

The datasets listed below offer interesting data cleaning opportunities.

County-level oil and gas production

Automobile fuel economy
Use "Datasets for All Model Years (1984-2017)."
You will see that you can pick one of two zipped files. You will need the documentation that goes with them.

Job patterns for minorities and women
Download `2009-EEO-1-Job-Patterns-Data-ZIP.zip` and use `YEAR09_CBSA_NAC3.txt`

Consolidated state performance report
This is a site about academic achievement assessment in school districts all over the U.S.
You will see that there are four csv files available for download on this site. You don't need to present all of them. Onc

2001 residential finance survey

National Data Buoy Center
The NOAA Buoy Data Center is a gateway to the data from a worldwide network of data collection buoys.
Find a buoy that has Standard Meteorological Data going back to at least 1980.
You can see information about each buoy and get access to its data by clicking on the buoy on the map.
You will see that the data from each year is stored in a single file.
You will have to read the data one year at a time, combining the observations from 1980 to the present.

Data sources

Government Datafiles	catalog.data.gov/dataset
MBTA	www.mbta.com
Federal Reserve	Federal Reserve Data
SBA	SBA Data
SBA	SBA: Understand your market
Kaggle	Kaggle datasets
US Vital Statistics	Data from the National Center for Health Statistics
re3data	Registry of Research Data Repositories

Grading criteria:

- Cleaning steps are organized in stages with well-organized R scripts which make appropriate use of data wrangling and subsetting tools.
- Code is well-commented, making implementation clear without being verbose.
- A written presentation is prepared in Markdown or Sweave. This may be all that you submit. Make sure you submit the source file.
- Final data is tidy.
- The presentation includes a description of the dataset and the steps taken to clean, summarize, explore, and organize it. The description should be sufficient for the process to be reproduced. If you use random numbers, be sure to use set.seed().
- The presentation includes well-labeled graphics.
- The presentation anticipates questions that might be answered with further analysis.
- Git has been used for version control.