Class project

Due December 9, 2020

2020-09-08

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

In this class, we have talked through many aspects of analysis in R. To guide the tour through R, we introduced tutorials via Swirl or Rstudio.cloud Primers. We added to this topics on Reproducible Research, version control, and Good Programming Practices. In this project, we need to tie it all together. Here, we will work in small teams to join the data revolution. If nothing else, 2020 has been the year of the Data Scientist. Between Covid-19 data, population income or other social data, and election data, we have been locked at home with nothing between us and a fun and perhaps informative data analysis except a keyboard. In this project, the challenge will be to choose a large and multivariate dataset, pull the data into R, and them munge it as appropriate into a tidy dataset. From here, do something interesting with it!

We will work in small teams, determined by class size. Choose one of the three topics listed below, alter to fit your interest. Submit a project proposal for approval. This should be submitted through a new GitHub repo with all team members and myself listed as collaborators.

Project 1: COVID-19

We are awash in case load data. We also have data on hospital beds, ICU beds, first responder counts, population demographics, etc. Create a dashboard combining COVID case load data with some other demographic to create an informative graphic. The graphic should contain a time component, should allow fitering in some interesting way and create meaningful summary statistics. Here are a couple of possible data sources:

https://covidatlas.com/data
https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals

Project 2: Social disparity

There are a ton of reports of social disparity in the U.S. Create a dashboard combining racial population statistics with other data such as income, health or other factor. This dashboard could include a time component and should include a spatial component. Ideally, it would involve a map and plots of the data. Question: does the time series data accurately depict what the underlying data show? Are there other features to the data that might suggest other trends that might be interesting to follow up on? Here are a couple of papers and data sources:

https://www.racialequitytools.org/resourcefiles/lui.pdf
https://www.federalreserve.gov/econres/notes/feds-notes/recent-trends-in-wealth-holding-by-race-and-ethnicity-evidence-from-the-survey-of-consumer-finances-20170927.htm
https://www.pgpf.org/blog/2019/10/income-and-wealth-in-the-united-states-an-overview-of-data
https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc/pinc-01.html

Project 3: Election 2020

This is an Presidential election year in the U.S. Here, you should create a dashboard of political interest. Ideally, the dashboard will contain a map and summary plots of the data. What data can you find that

might suggest one candidate over the other, locally, regionally, nationally? Is there a time component to the data showing enthusiams change? Is there social data available that might show trends? A few of the many possible data sources:

Twitter: https://towardsdatascience.com/understanding-political-twitter-ce3476a38377
https://web.archive.org/web/20160628093159/http://www.nsd.uib.no/macrodataguide/country2.html?id=840&c=United%20States%20of%20America https://www.cpds-data.org/

Rubrics for grading (passing > 8 pts):

1. Version control – git – all member make commits? (2 pts)
2. Reproducible Research adherance (2 pts)
3. Good Programing Practices followed (2 pts)
4. Overall project creativitity (2 pts)
5. Overall project execution (2 pts)