EDAV Fall 2019 PSet 3

Read *Graphical Data Analysis with R*, Ch. 6, 7

Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class.

Data cleaning: for many of the questions, you will have to clean up the data more than in past assignments. Labels do not have to perfect but they have to be legible. Often it is helpful to shorten or abbreviate labels: this can be done before plotting or at times within the plot functions. You may make any changes to the data as appropriate before plotting, including renaming column names and the like. Be sure though to include all adjustments in your scripts.

1. Coal Emissions

Data: https://datadiscovery.nlm.nih.gov/Environmental-Health/TOXMAP-EPA-Clean-Air-Markets-2016-Coal-Emissions/n97u-wtk7

(OK to manually download `.csv` file)

(a) Using `parcoords::parcoords()`, create an interactive parallel coordinate plot of `SO2`, `NOx`, `CO2`, `Heat Input` and `Gross Load`. Each line in the plot should represent one coal plant. Note any patterns.

(b) Create a second interactive parallel coordiante plot, with the lines colored by state. Include only the states with more than 15 plants. Do any new patterns emerge? Explain briefly.

(c) What do `SO2`, `NOx` and `CO2` stand for? Briefly describe why each is harmful.

2. Planets

Using **rvest**, scrape the data from this table: https://nssdc.gsfc.nasa.gov/planetary/factsheet/ (hint: `html_table` is helpful). Remove `MOON` data. Then, using `GGally::ggparcoord()`, create two parallel coordinate plots of the numerical variables in the dataset, one colored by the value of `Ring System?` and one colored by planet name. In both cases, each line should represent one planet. Use `coord_flip` so the variable names are easier to read. Describe any patterns you find.

3. Heart Disease

Data: four data frames that begin with `heart_disease` in the **ucidata** package

Packages: You may use **vcd** or **ggmosaic**.

(a) Create three mosaic plots, each involving two categorical variables from `heart_disease_cl` and interpret the plots. (You may reuse variables, for example X ~ Y and X ~ Z).

(b) Combine the four heart disease datasets and create a mosaic plot showing chest pain by sex and location. Describe any patterns.

4. District 3 Elementary Schools

Recently, there has been much debate about the lack of racial and economic diversity at Manhattan District 3 elementary schools, part of a larger and long-standing controversy about iniquities in the New York City public school system as a whole.

The *New York Times* article, "Rezoning Plan to Remake 3 Upper West Side Schools Will Proceed, City Says," (https://www.nytimes.com/2016/11/10/nyregion/rezoning-plan-for-3-upper-west-side-schools-will-proceed-city-says.html) (2016-11-10) identifies the 11 elementary schools in Manhattan District 3.

For this question, we will analyze parent survey results for these schools.

Data: https://www.schools.nyc.gov/about-us/reports/school-quality/nyc-school-survey

(a) Choose one of the likert style questions from the 2019 parent survey and use a diverging stacked bar chart to show results for the 11 schools identified in the article referenced above.

(b) Choose a question that was asked in 2014 and 2019 and compare results for the three schools discussed most in the article: P.S. 199, P.S. 191, and P.S. 452. You may use two separate diverging stacked bar charts or combine all the information in one.

(c) Interpret your findings of (b) in light of the reputations of the schools as presented in the article. Are they surprising or what you would have expected?