EDAV Fall 2020 PSet 3

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

Grading is based both on your graphs and verbal explanations. All graphs must have informative titles and follow all best practices as discussed in class.

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 within the plot functions. Be sure to include all adjustments in your scripts.

1. penquins

We will use the `penguins_raw` dataset from the **palmerpenguins** package.

(a) Draw a parallel coordinates plot of the numeric columns in the dataset using `ggparcoord` from the **GGally** package. Choose parameters to help identify trends. What do you observe?

(b) Experiment with using color to separate factor levels of the factor variables (one at a time, not in the same graph). Include the plot which shows the most distinct clusters. Briefly describe how the clusters differ from each other. (The factor variable should not be included as an axis in the parallel coordinates plot.)

2. pulitzer

(a) Draw an interactive parallel coordinates plot of the `pulitzer` dataset in the **fivethirtyeight** package with brushing using the `parcoords()` function in the **parcoords** package.

(b) Which newspapers appear to be multivariate outliers? Briefly describe their unusal patterns.

(c) Choose one of the newspapers and research it online to gain a deeper understanding of its uniqueness. Provide a brief summary of your results. *Cite your sources by linking to the pages where you found the information.*

3. doctor visits

For this problem we will use the `DoctorVisits` data set in the **AER** package. (Note: for this package you need `data("DoctorVisits")` to load the data.)

(a) We will consider `visits` to be the dependent variable. Create a new factor column, `numvisits`, based on `visits`, that contains three levels: "0", "1", and "2+". (Tidyverse option: `forcats::fct_collapse()`). Why is this a reasonable choice?

(b) Draw a mosaic pairs plot of all factor variables in `DoctorVisits`. Based on the plot, which variables appear to have a *strong* association with `numvisits`? *medium* association? *weak or no* association? (Make your judgments relative to the other variables.)

(c) Are p-values from χ^2 (chi-squared) tests of each of the variables and `numvisits` consistent with your categorization in part (b)? Explain briefly.

(d) Draw mosaic plots of `gender`, `lchronic`, `private` and `numvisits` in stages:

- `gender` and `numvisits`
- `gender`, `lchronic` and `numvisits`
- `gender`, `lchronic`, `private` and `numvisits`

All cuts should be vertical except the last one. The last cut should be the dependent variable. Use appropriate fill colors. What patterns do you observe?

(e) Use `geom_alluvium()` from the **ggalluvial** package to display the same variables as in the last graph of part (d). What new information / perspective does the alluvial plot provide, if any?

4. likert data

(a) Find data online with at least three categorical variables whose levels form the same spectrum from one pole to its opposite. A common example is likert survey responses with categories such as "Strongly agree", "Agree", "Neutral", "Disagree", "Strongly disagree", but your choice does not need to involve likert data. **Provide a link to your data source.**

(b) If possible read your data directly from the site. If not (for example if it is in pdf form), create a data frame in your code, or read it in from a data file that you've created. Include the file with your submission. Display a table of your data in summarized form. For example:

Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree	Question
0	5	8	78	162	Academic
0	11	5	88	72	Business and industry
2	3	5	34	27	Federal, state, and local government
0	0	2	15	11	Private consultant/self-employed
2	2	5	15	10	Other

(c) Draw a diverging stacked bar chart of your data, following all best practices. You may use any package.