Homework #3, due MONDAY, MARCH 13, 11:59pm

1. **Mosaic plots** [10 points] Use a mosaic plot (or plots) to display the `Punishment` dataset in the `vcd` package. What patterns of association do you see between the conditioning (independent / explanatory) variables and the response (dependent) variable? Discuss single factor patterns as well as two and three factor interactions.

   Notes / advice:

   a) The description on p. 142 of patterns visible in Figure 7.9, which shows the results of an arthritis treatment study, will be helpful. Here single factors are discussed (ex. the relationship of treatment and improvement), as well as two factor interactions (sex / treatment pairs and improvement).

   b) The labeling and ability to control directional splits is better with the `mosaic()` function in the `vcd` package, but `geom_mosaic` (`ggmosaic` package) is also an option if you prefer the `ggplot2` framework.

   c) Pay attention to the order of the variables in the formula, for example `"Improved ~ Treatment1 + Sex"` in the `Arthritis` dataset example. The variable to the left of the tilde is the response variable. Add conditioning variables one at a time to see the effect each addition has on the plot. With `geom_mosaic`, this information—response vs. conditioning variables and order of variables--is specified in the aesthetic mapping. See `?geom_mosaic` for more details.

   d) Pay attention as well to the ordering of directional splits specified with the `direction` parameter (`vcd`). Note that the last one listed (`"h"` in the `Arthritis` example) specifies the direction of the split (horizontal or vertical) for the response variable, somewhat counterintuitively since it appears first in the formula. The other `"h"`s and `"v"`s specify the directions of the splits for the conditioning variables, in the order in which they appear in the formula. (They are both `"v"` in this example.)

   e) Note that the fill colors are linked to the levels of the response variable only – choose them carefully. Think about the order of the levels of the response variable (which affects location in the plot) and appropriate color choices for the level or levels that represent an effect. For example, with the `Arthritis` data, `"Marked"` improvement is intentionally more prominent than `"Some"` improvement, which in turn is more prominent than `"None"` in the plot.

   f) Finally, note that both `vcd` and `ggmosaic` have `mosaic()` functions, so watch out for conflicts. Either use `vcd::mosaic()` or don't load both packages at the same time.

2. **Heatmaps** [20 points] Use heatmaps to explore the sample sales dataset `"WA_Sales_Products_2012-14.csv"` (88475 rows x 11 columns) found here:

   https://www.ibm.com/communities/analytics/watson-analytics-blog/sales-products-sample-data/

   Include 3 heatmaps in your assignment and describe the patterns you found. You may sample, subset and summarize as you wish. (The numbers of rows and columns in your heatmaps will have no effect on your grade.) Explain choices regarding the ordering of rows / columns and scaling. Choose a perceptually uniform color scale. (Note: if you scale the data yourself, rather than setting a `scale` parameter, be aware the the base `scale()` function returns a matrix. Either convert the result to a vector or write your own rescaling function.)

3. **Missing values** [20 points] Use the techniques in Chap. 9 (to be discussed in class) to visually explore patterns of missing data in the College Scorecard Data found here: https://collegescorecard.ed.gov/data/
   Do not choose the large green "download all data" button but rather, the first choice under "featured downloads": **Scorecard data 5 MB CSV**
   (7703 rows x 122 columns)

   Sample and/or subset as you wish. Include 3 graphs that show missing patterns. Describe the patterns you see. Choose the missing data patterns of a single variable (or single group of variables) to discuss in more detail, drawing on insight gleaned from the documentation, in addition to the graphs. Discuss possible causes for the missing data patterns, and potential biases which might result from these patterns. The documentation is available here: https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf