**Week 3 Homework:** (5 points)

#### Visually Describe a Dataset

The purpose of this exercise is to find a dataset that is interesting to you and then use what you have learned so far to visually describe a dataset with R plots. This means showing a few (3) single-dimension plots as well as at least one multidimension plot. One of the hardest parts of this assignment is going to be finding the right data. You will turn in a PDF file with the plots and your R script. Make sure your plots are vector graphics (saved as a PDF file). Each plot should be on its own page. Note that work done in this assignment can be used in both the Work-in-Progress report (coming up) and the Final Poster.

#### Data

You need to find a substantial dataset for this project. Your dataset must score at least 100 in the formula below:

(NumberOfColumns \* 4) \* (NumberOfRows/100) >= 100 Examples:

A dataset with 5 columns and 10,000 rows would score: (5 \* 4) \* (10,000/100) = 2000 A dataset with 17 columns and 3,000 rows would score: (17 \* 4) \* (350/100) = 238

A dataset with 5 columns and 500 rows would score: (5 \* 4) \* (500/100) = 100

The point of this is to make sure you have a large enough dataset to work with. You need enough columns to look at how different variables might related. You need enough rows to show that you can work with a decent-sized dataset.

Provide the following information about your data:

* Description – what does the data set represent, 1 or 2 sentances
* Provide the out of the str function on your data set
* Provide the calculation notes above using your data set, specifically: (NumberOfColumns \* 4) \* (NumberOfRows/100)

#### Descriptive Plots

Your data may have any types. For example, you might have a mix of continuous, text, and categorical data columns, or you might have several columns of just continuous data. The type of the data should dictate what kinds of plots you make.

To visually describe the data, first, make some plots in R that show how the data are distributed. For continuous data, this might be histograms, box plots, or density plots. For categorical data, a good choice is a bar chart that shows the frequencies of different categories. A pie charts works for this, too. Include at least three of these single-dimension plots. If you tell me your data have only three columns, I will probably say to find a different dataset. Next, make a multidimension plot using at least two columns, though as you have seen in labs, some two-dimension plots can take three different columns to make.

Indicate which plots are distribution oriented ie single dimension and which plot is multidimensional.

Note that in this assignment the plots and their context text should be all I need to get a sense of what your data are about. The plots don’t need to be “pretty” unless you want them to be, but they do need context items that help me understand your data. All plots should have a title and possibly context text under the title so I know what the data are about. Include titles and axis labels in the plots. All plots should list the source in the lower-right corner in small (but readable) text.

#### Format of the PDF File

The file should be a single PDF file. The first page should include a title, your name, the source of the data, and a table of contents of the plots. Include your plots on the following pages. You can use functions like par(mfrow = …) to have more than one plot per page.

Upload this PDF file with the file name: DataReport-<your name>.pdf Also upload your R script.