Class Notes: Week 2

Three Important Concepts for Today

Concept 1. Data can be usefully organized into tables with “cases” and “variables.” In “tidy data,” every case is the same sort of thing, e.g. a person, a car, a year, a country in a year.

We talked about data tables, cases, variables, etc. in Week 1.

Concept 2. Data graphics can be constructed easily when each case corresponds to a “glyph” (mark) on the graph, and each variable to a graphical attribute of that glyph such as x- or y-position, color, size, length, shape, etc. Such data is called “glyph-ready.” (The same is true for more technical presentations of data, e.g., models, predictions, etc. — once the data are set up with appropriate cases and variables, the presentation is straightforward.)

Concept 3. When data are not yet in glyph-ready form, you can transfigure them into glyph-ready form. Such transfigurations are accomplished by performing one or more of a small set of basic operations on data tables: the so-called data “verbs.”

Today’s Agenda

a. Introduce some software and commands that …

   1. make it easy to access data tables and see how they are structured: `data()`, `help()`, `names()`, `nrow()`, `str()`, `summary()`, `head()`
   2. let you easily map variables onto graphical attributes for glyph-ready data: `mScatter()`, `mBar()`, `makeWorldMap()`, `mUSMap()`
   3. implement two of the data verbs. `group_by()`, `summarise()`

b. Explore some basic graphical choices:

   1. The format of graphs: major categories of glyph: points, bars, shapes.
   2. The nomenclature for different parts of graphics: frame, scale, guide.
   3. Effective mappings from variables to graphical attributes.

c. See what makes data tables glyph-ready or not, and how the data verbs can be used to transfigure data tables into glyph-ready data.

Today’s Work

1. Discussion on deconstructing graphs and glyphs. Link to worksheet.
2. Create an Rmd file, `InClass-2-XXX.Rmd`. Work through the topics under “Software and Commands,” putting your answers into the Rmd file.

Software and Commands

Three (unrelated) examples:

- `NHANES`
- `Minneapolis2013`

- `CountryData`

1. What is the setting for the data? That is, what are they about?
2. How many cases are there?
3. How many variables are there? What are their names?
4. Pick out three of the variables and say whether
    - the variable is quantitative or categorical
    - if categorical, what are some levels of the variable
    - if quantitative, what are the units of measurement of the variable.
5. Describe, in everyday terms, what kind of thing cases represent in each of the data tables.

Mapping Variables to Graphical Attributes

You're going to make some simple graphics.

NHANES To speed things up, make a subset of just 2000 cases from `NHANES`:

```
Small <- sample_n( NHANES, size=2000 )
mScatter( Small )
```

Notice that the argument to `mScatter()` is a data table.

Make a graph of height against age, height against weight, etc. Use one or more other graphical attributes such as color, size, etc. Find an relationship that interests you.

CountryData The `mWorldMap()` function makes it easy to construct country-by-country maps. It takes three arguments:

1. a data table
2. a character string specifying the variable that identifies the country
3. a character string specifying the variable whose level will be represented by color.

For example, here's a map of the number of deaths in each country (per 1000 inhabitants per year):

```
mWorldMap( CountryData, key="country", fill="death" )
```

Make that map and comment on the pattern it shows.

Make a map of some other variable of interest to you and comment on what it shows.

Bar Charts

A bar chart is a simple and limited form of graph It represents a number as the length of a bar.

Consider the Minneapolis 2013 election data. Here's a bar chart that might be used to show the election results:

This graph reflects the following data table (only part of which is shown):

```
Source: local data frame [6 x 2]

               First votes
1       BETSY HODGES 28935
2        MARK ANDREW 19584
3        DON SAMUELS  8335
4         CAM WINTON  7511
5 JACKIE CHERRYHOMES  3524
6           BOB FINE  2094
```

Compare the `Minneapolis2013` data table and the data table printed above.

1. Do they have the same number of cases?
2. Do the cases in the two tables represent the same sort of thing?
3. Do the two tables have any variable(s) in common?
4. Speculate on how the two tables are related to one another.

Data verbs for summarizing and grouping

`summarise()` : Find an expression involving `summarize()` and `NHANES` that will produce the following.

- number of people (cases) in `NHANES`
- total weight of all the people in `NHANES` (silly)
- mean weight of all the people in `NHANES`

`group_by()` : repeat the above, but calculating the results group-by-group for:

- males versus females
- smokers and non-smokers
- people with and without diabetes
- break down the smokers versus smokers further, by sex
- break down the people with diabetes further, by smoking

Zip Codes

Make a scatter plot of the `ZipGeography` data. Use latitude, longitude and time zone to set position and color. By choosing the right combination, you should be able to construct a plot whose meaning is immediately obvious to anyone familiar with US geography.