DCF Week 4 Warm-Up Exercises

Data and Computing Fundamentals

Do this work in an Rmd file named `Week-4-Warmup-XXX.Rmd`.[1] Rather than typing commands at the console, type them into a chunk and run that chunk in the console. (If you're not sure what this means, ask! There is a keyboard shortcut that makes it easy.) When the chunk does what you want, compile the Rmd document to HTML. Then move on to the next task and repeat the cycle: compose, get it working, compile to HTML.

Cities of the World

The data table `WorldCities` (in the `DCF` package) identifies cities around the world that have large populations or are large for their region.

- Check the table for plausibility: is it possibly what it is claimed to be?. For instance ... What's the total number of people represented? Explain why or why not the data pass this plausibility test. Create another plausibility test and describe it. It can be very simple. If you can, implement it and state whether the data table passes the test.
- How many cities larger than 100,000? Larger than 1,000,000?
- Make a scatterplot of the latitude and longitude of cities larger than 100K.
  - Decide what variables to map to the x and y aesthetics.[2]
  - Use the size of the dot to show the city's population. In other words, map the variable population to the `size` aesthetic.
  - Use transparency, called `alpha`, to handle overplotting. Alpha can run from zero to one: zero is completely transparent (a.k.a. invisible); one is completely opaque. You will be *setting* `alpha` the same for every city. Recall that in `ggplot` graphics, variables are *mapped* to aesthetics using the `aes()` function. In contrast, aesthetic properties that are the same for every case are *set* outside the `aes()` function. In a typical use, the `ggplot()` command will look like

    ```
    ggplot( data=???, aes( x=???, y=??? ))
    ```

    The layers of the plot will be used like this:

    ```
    geom_point( alpha=???, aes( size=??? ) )
    ```

    where, of course, you will replace the `???` with appropriate variables or constants or data tables.
- When you have your plotting commands complete, use those commands to make another graphic, but add this expression to govern the `size` attribute: `+ scale_size_area()`. This will make the *area* of the dot proportional to the value of the variable mapped to it. Without `scale_size_area()`, the *diameter* of the dot is proportional to the variable. Explain which scale, area or diameter, you think is most informative. (Include both graphics in your Rmd file along with your explanation.)
- Create a data table `BiggestByCountry` that has the one biggest city in each country.
- Plot the locations of `BiggestByCountry` as another layer in your graphic. Make them red.
- Add to the graphic the names of the cities from `BiggestByCountry`. Hint: use `geom_text()`. Set the `size=2`. Remember, *setting* is different from *mapping* a variable. You'll use the `label=` aesthetic to represent the city names.

[1] `XXX` should be replaced by your personal ID, e.g. your initials.
[2] Remember, "aesthetic" is being used in its original sense: how things are perceived.

- Find the countries where the biggest city is more than 5M people

  The resulting table will have a couple of dozen cases. Display as output in your report all the cases but just these variables: city name, country, and population.

- List by name all of the data verbs you used in this work. To jog your memory, here are first letters of seven important data verbs: A, F, G, J, M, S, S.