DCF Week 5 Warm-Up for Assignment

Data and Computing Fundamentals

Do this work in an Rmd file named `Week-5-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.

You will use four data tables in this exercise:

1. `ZipGeography` in the `DCF` package.
2. `Restaurants` which you must load into R.
3. `Cuisines` . . . . . ditto . . .
4. `ViolationCodes` . . . . . ditto . . .

Read in (2), (3), and (4) with these commands:

```
load( url( "http://tinyurl.com/m4o4n2b/DCF/ViolationCodes.rda" ) )
load( url( "http://tinyurl.com/m4o4n2b/DCF/Cuisines.rda" ) )
load( url( "http://tinyurl.com/m4o4n2b/DCF/Restaurants.rda" ) )
```

Do this now, before reading on. It will take about 2 minutes for the last one. Then, while you're waiting, read the rest of this activity. You'll know it's working if `Restaurants`, `Cuisines`, `ViolationCodes` show up in your account.

How's the Food?

Government agencies have increasingly been putting data in publicly accessible places. For example, in India, the http://attendance.gov.in/ website tracks the attendance at work of government employees.[2]

New York City publishes many datasets, including health inspections of restaurants. That's what you're going to work on now.

The data table `Restaurants` contains information about each health violation. (See the introduction for how to access the data table.) Note that the `DBA` variable contains the name of the restaurant.

- What is the meaning of a case in this data table? You can look at the data table with this command: `View( Restaurants )`.
- How many cases are there?
- Some of the entries in the
- How many distinct restaurant names are there?
- Are there any restaurants with the same name but with multiple branch locations? Select one or more variables that plausibly identify a unique branch.
- Which restaurants (individual branches) have the most violations?

[1] `XXX` should be replaced by your personal ID, e.g. your initials.
[2] See this New York Times article.

The data table `ViolationCodes` contains a description of the different types of violations and whether they are critical. `Cuisines` gives the meaning for the `CUISINECODE` variable.

- How many critical violations were reported? Non-critical?
- Which restaurants have the greatest number of "critical" violations? Produce a table showing the number of critical and non-critical violations for each restaurant.
- What's the most common cuisine type in the whole city? In each borough? In each zip code?

The data table `Cuisines` details the code for each restaurant's cuisine type.

- What kind of cuisine has the most violations per restaurant, on average?

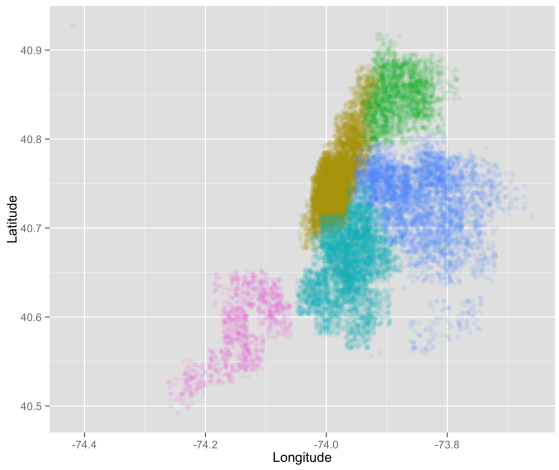
You can create a table where the case is "an individual restaurant branch" with this statement:

```
RR <- Restaurants %>%
  group_by( PHONE ) %>%
  filter( row_number(PHONE)==1 )
```

Get the latitude and longitude of each zip code from `ZipGeography`. Plot out the location of each restaurant, using the borough for color. Use the aesthetic

```
position=position_jitter( width=0.02, height=0.02 )
```

to spread out the restaurants a bit. Play with `alpha=` to get a nice graphic. Which borough corresponds to each borough number?



- Choose 3 or 4 cuisines of interest to you. Plot out the location of each restaurant of that type as an additional layer on the previous graphic.
- Find the violations that are most likely to be repeated for a given restaurant branch. (If a violation appears twice or more for a given restaurant branch, the violation extended over more than one inspection period.)
- Make an informative plot of the distribution of `SCORES` for each letter `CURRENTGRADE`.
- Do the same for the distribution of `SCORES` by `ACTION`.