Module 4

This week we will be looking at Enterococcus levels in the Hudson River, using data from the organization Riverkeeper (http://www.riverkeeper.org/).

Background: Enterococcus is a fecal indicating bacteria that lives in the intestines of humans and other warm-blooded animals. Enterococcus (“ Entero”) counts are useful as a water quality indicator due to their abundance in human sewage, correlation with many human pathogens and low abundance in sewage free environments. The United States Environmental Protection Agency (EPA) reports Entero counts as colonies (or cells) per 100 ml of water.

Riverkeeper has based its assessment of acceptable water quality on the 2012 Federal Recreational Water Quality Criteria from the US EPA. Unacceptable water is based on an illness rate of 32 per 1000 swimmers.

The federal standard for unacceptable water quality is a single sample value of greater than 110 Enterococcus/100 mL, or five or more samples with a geometric mean (a weighted average) greater than 30 Enterococcus/100 mL.

Data: I have provided the data on our github page, in the folder https://github.com/charleyferrari/CUNY_DATA608/tree/master/lecture4/Data. I have not cleaned it – you need to do so.

This assignment must be done in python. It must be done using the ‘bokeh’, 'seaborn', or 'pandas' package. You may turn in either a . py file or an ipython notebook file.

Questions:

- Create lists & graphs of the best and worst places to swim in the dataset.

- The testing of water quality can be sporadic. Which sites have been tested most regularly? Which ones have long gaps between tests? Pick out 5-10 sites and visually compare how regularly their water quality is tested.

- Is there a relationship between the amount of rain and water quality? Show this relationship graphically. If you can, estimate the effect of rain on quality at different sites and create a visualization to compare them.