HOMEWORK 02

MATH 295: DATA SCIENCE II

Assigned: 2014-09-16
Due: 2014-09-23

Instructor: James P. Bagrow
bagrow.com/ds2/

Instructions Submission, showing your work, etc. are as per Homework 1. Replace `'HW01'` with `'HW02'` in your filenames (email me `HW02_<yourname>.zip`, for example).

Project Tycho is an initiative at the University of Pittsburgh to clean, standardize, and release historical disease records for the United States. This is a very challenging, labor-intensive task, and their work is ongoing. So far they have built an online API for these data.

- Explore the website, make an account so you can access their API, and **request an API key**.

P1. Write and run a script to download the **Level 2 data** for **state**-level cases and deaths of the following diseases: measles, whooping cough, scarlet fever, diphtheria, influenza, mumps, rubella, chickenpox. Save these data to files in `data/`. Be sure to consult their API documentation (Project Tycho Level 2 Data API Help). Describe any problems you may have encountered along the way. Show your work.

WARNING: Do not share your API key, especially with me! Place it in a separate, one-line text file and structure your download script so it loads that string from that file. When you zip up your directory to submit your homework, do not include that text file.

P2. Plot disease levels over time. You can do this per state, or aggregated by region or the entire US. Ensure that your plots are **readable** and not just a blob of time series. Provide a brief interpretation of these plots, potentially with summary statistics and citations as needed.

P3. An important public health problem is to **forecast** the levels of a disease, so that resources such as medicine and medical supplies can be stockpiled. This was particularly important in the era before mass vaccination. Polio was infamous for having "good" and "bad" years, and these were relatively hard to predict.

Can you think of a way to predict the future levels of a disease (numbers of cases) from the past or present levels? Introduce a **basic statistical argument** for summarizing the **predictability** of a disease. Apply this argument to each downloaded disease. You may work with aggregated time series for the entire US if you wish. Then apply this methodology to the Polio disease data from Homework 1. How predictable are these diseases compared with Polio?

- You don't have to actually *predict* the levels of a disease, you only need to find a way to demonstrate if one disease is more "predictable" than another.
- You have immense freedom in your "basic statistical argument". The key is that you must demonstrate a reasonable and logical argument. The precise argument, and level of rigor, is very much up to you. Further, your approach does not need to be complicated. Simple is often best.

Bonus. Can you estimate a disease's **incidence rate** over time? If so, try to do it. If not, explain why not.