Assignment #1: Data Exploration
Spring 2016
CS3943/9223
Prof. Rumi Chunara
Total: 30 points

All questions must be completed in R. Implement and comment your code so that anyone reading the file can reproduce the code easily (e.g. set the file path once at the beginning of the script where it can be easily changed). Save the code as an R markdown file, and upload it to NYU classes.

1. Data Exploration (15 points total).

Google Flu Trends (GFT) relays flu search activity based on aggregated Google Search query data. Queries related to flu were determined by training a model that select which combination of all Google queries matched trends over time of the Centers for Disease Control and Prevention flu time series (You can read more in the paper: Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2009): 1012-1014. which is in the Resources/Research Papers section for the class on NYU Classes).

The project is no longer active but results are reported here: https://www.google.org/flutrends/about/ and standardized to make data more comparable across regions. The 'baseline' level for each region (0) is its average flu search activity, measured over many seasons. Activity levels for each region represent how much flu search activity differs from that region's 'baseline' level. The following problem involves selecting and using GFT data, along generating relevant questions by which to explore a data set. To do this question, first download the U.S. data and load it into R.

a) Comparing HHS region and state values
    - What is an HHS region? (Yes you will have to look it up) (1 point).
    - How do the HHS region values compare to the state values (describe and show the analysis you choose along with a written answer) (2 points).
    - Group the cities by state and examine how those grouped values compare to state values. Use descriptive statistics. How have you decided to handle missing data and why does that make sense? (3 points).
    - Design two relevant metrics that would be informative from this data. Examples of potential metrics include: Quantiles, mean, median, variance, and max, and these across the various geographic segments. Be selective. Think about who could use the data and what will be important to track over time. Also, what will summarize the data, but still be useful. Describe why you chose your metrics and show how you calculated them. (4 points).
    - Find the population of the states (give a complete citation/credit for your source), and create a comparison of population vs. peak flu trend value for the most recent year. Describe how you have decided do the

comparison, show your calculations and report your findings. (5 points).

2. Simple Data Analysis (5 points total).

Here is an exercise requiring you to combine different pieces of information to make some inference from the data.

For this question, download the flu data for all of the countries.

- Plot the center latitude for the country versus peak week of flu in the most recent year of data. Is there any relationship? In your response remember to credit your source for the latitude information.

3. Web Scraping (10 points).

Many times data we want to use is not in a ready to download format. This question gives you some experience in scraping data from websites.
Here is an example of some data that is available only on in HTML table format:
http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6401a4.htm?s_cid=mm6401a4_w
The table shows influenza status and vaccination status for different segments of the U.S. population. Take a few minutes to read and interpret the data.

a) Read the Vaccine Status data from the table on the above website into an R data frame. There are many packages to use, I suggest you try the XML package which has useful functions such as htmlParse() to read in HTML documents and readHTML_Table().

b) Find another example of a table somewhere on the web to load into R (Reminder, everyone must complete this assignment independently including finding a unique table to download). Provide the link to where the table is found along with your code.