STA 220 - Data and Web Technologies for Data Analysis - Homework 2

Working with the New York Times COVID-19 Data

As we have seen in class, COVID-19 data collected by the *New York Times* are available in a repository on Github. In this assignment we will work with these data to reproduce some of the elements of the *Times*'s reporting on COVID-19.

Exercises

1. The file `us.csv` contains aggregated data for the entire U.S. In this file, the `cases` and `deaths` columns represent *cumulative* cases and deaths due to covid. The `diff()` function can be used to compute the differences between each consecutive element of a vector, so it could be used to compute the daily numbers of cases and deaths. However, `diff()` returns a vector of length one less than the length of the original vector (e.g., diff(c(1, 3, 6, 10)) returns 2, 3, 4) and this can make it somewhat inconvenient to use when transforming columns of data frames.

   An alternative is to us the more general `filter()` function with an appropriate choice of the `filter` and `sides` arguments. Thus function can also be used to computing running averages and similar quantities.

   a. Read the file `us.csv` into R as the data frame `us` and do the following:

      - Transform the `date` column into a column of class `Date`.
      - Use `filter()` to add a column named `new_cases` containing the number of new cases reported on each date. The first value in this column will be `NA`.
      - Use `filter()` to add a column named `new_deaths` containing the number of new deaths reported on each date. The first value in this column will be `NA`.
      - Use `filter()` to add a column named `avg_new_cases` where each element represents the mean number of new cases for the previous 7 days (inclusive of the current day). The first 7 values in this column will be `NA`.
      - Use `filter()` to add a column named `avg_new_deaths` where each element represents the mean number of new deaths for the previous 7 days (inclusive of the current day). The first 7 values in this column will be `NA`.

      Note that the `filter()` function used here is `stats::filter()` from the `stats` package, which is loaded by default in R. (The `dplyr` package has a completely different `filter()` function which plays an important role in the "tidyverse". If you have problems using `filter()`, you should make sure that you do NOT have the `dplyr` package loaded. If you do, then you will need to explicitly type out `stats::filter()` to get the `stats` version.)

   b. Create a plot of daily cases similar to the one found at the top of this page. Plot only data beginning from 2020-03-01. (Note that this plot and a similar plot for deaths appear again about 1/3 of the way down the page.)

      Try to do this using the formula method of the `plot()` function with the optional arguments `type = "h"`, `col = "gray"`, and `data = us` and using the `subset` (base) argument to plot only the data for dates 2020-03-01 and after. (You may also wish to experiment with the optional argument `lwd`.)

Then use the formula interface to the `lines()` function to add the curve showing the seven-day running average. (Again, you may wish to experiment with the optional argument `lwd`.)

c. Repeat part (b) for deaths.

2. The file `us-states.csv` contains state-level data for the U.S.

   a. Read `us-states.csv` into R as the data frame `us_states` and transform the date column into a column of class `Date`.

   b. Use `subset()` to extract the data for the state of California and save it as a data frame named `California`. Be sure that the rows are correctly ordered by date, and then repeat parts 1b and 1c of this assignment for California, i.e., plot the number of daily new cases and deaths, along with their 7-day running averages.

3. The file `us-counties.csv` contains county-level data for the U.S.

   a. Read `us-counties.csv` into R as the data frame `us_counties` and transform the date column into a column of class `Date`.

   b. Use `subset()` to extract the data for Yolo County, California, and save it as a data frame named `Yolo`. Be sure that the rows are correctly ordered by date, and then repeat part 1b this assignment for Yolo County, i.e., plot the number of daily new cases along with their 7-day running average.

   Q: What do you notice when comparing the plot of daily new cases in Yolo county to the analogous plot for the state of California as a whole? What might explain what you are seeing?