This is a systematic tutorial for the web scraping and processing of the Covid-19 data of New York State. QA steps are also covered. Email Tophey at [hz1954@nyu.edu](mailto:hz1954@nyu.edu) for any questions.

The routine is weekly work, for QA is implemented weekly. The more efficient way would be doing all of the work monthly. The methods are introduced with with three sections: 1. The weekly routine, 2. Additional code setting for a new month. 3. Bi-weekly resetting of total deaths.

We talk about all the steps within the directory of *scrape\_NY\_COVID\_clean* folder. e.g. “./input” means the *input* folder within *scrape\_NY\_COVID\_clean.*

Section 1. The weekly routine

1.1 Update the url links to store the NY press releases.   
Google sheet for url links:

<https://docs.google.com/spreadsheets/d/1hHsqMuyxUbYBi74ddTmoQJ3OfQpU1wRsUBu1mg_6Mnw/edit#gid=415383079>

NY press release source: <https://www.governor.ny.gov/keywords/coronavirus>. See also in the “Go here” in the Google sheet.

Save the google sheet, download it, and put it in the directory *./input,* replace the old file.

1.2. Open file *./scrape\_NY\_COVID\_clean.Rproj*. Now in the R project, click open file *./scripts/main\_file.R*. Run the code from the beginning to comment “Switch point 1”.

1.3. Open file *./scripts/scrape\_”the month you are working on”.R*. Run everything.

1.4. Switch back to ./*scripts/main\_file.R*. Run the code from “Switch point 1” to comment “Switch point 2”. Or you can skip the code for previous months under comment “Data combining” except the first time that this file is used.

1.5. Switch to ./scripts/*new\_death\_nyc.R*.

For the first time, run everything in it.

For future use, only need to run all the code below comment “## 07/01/2021 to 07/31/2021”. The QA work for 5 NYC boroughs is no longer ours. Thus, running or editing the code in this script file is actually useless now. But I realized the best solution is to run it anyway, so that our data does not conflict with the previous versions. No further editing is needed in this file.

1.6. Switch back to *./scripts/main\_file.R*. Run the code from comment “Switch Point 2” to the end.

1.7. Open *./scripts/edting\_v2.R.* Run everything. The updated data is saved in *./output/NY\_press\_releases\_0320\_to\_present\_V2.xlsx.*

*1.8 Open ./output/NY\_press\_releases\_0320\_to\_present\_V2.xlsx.* Add the updated dates to our data in Google Drive.

<https://docs.google.com/spreadsheets/d/1HTwDf74xT4PFaoRwCU5Wnz74V39v3RTm/edit#gid=534330739> Location: COVID-19 Data Project / (2) Track – Daily COVID-19 #s / Team Resources / Internship 2.0 Regions / Northeast / QA Only Historical Data Sheets

Routine finished, yeah!!

Section 2. Additional code setting for a new month.

When we start working on the automation of a new month, some coding is needed. The following steps are based on Section 1 above.

2.1. Finish Step 1.1. In step 1.2, add a line of code under comment “separate links to each month”, and change the numbers so that they match our new month. Do the same to the code under comment “save only the links”. The finish Step 1.2.

2.2. In Step 1.3, run everything in *./scripts/scrape\_”the previous month”.R* to make sure the previous month’s work is finished. Copy everything in this file, and paste them to a new empty .R file. Change the numbers so that they match our new month. Run it. Now Step 1.3 is finished.

2.3 Before running any code in Step 1.4, in *./scripts/main\_file.R*, find comments like “#release\_07\_2021 includes data 07/01/2021 to 07/31/2021” and add a correlated chunk of code for our new month. Run the code for both the previous month and the new month.

2.4 (This step is mistakenly forgotten in the video!!!)

In ./script/main\_file.R, find the code below comment “#combine data all the months”. Add code

%>% cbind(release\_N\_2021)

(“N” is the new month you are working on.) Then finish all the steps left in Section 1.

Section 3. Bi-weekly resetting of total deaths.

This resetting procedure is used to keep our total death data of non-NYC counties from deviating too far from the data in the data entry sheet. The total deaths in the data entry sheet are directly obtained from sources. Ours are the cumulative sums of all the new deaths and the total deaths of a previous date in the data entry sheet.

When it is roughly two weeks since the last resetting, we need some coding in *editing\_v2.R*.

3.1 Finish everything step in Section 1 and Section 2 (if Section 2 is needed). Or you can choose to finish everything except Step 1.7.

3.2 Open the current data entry sheet of northeast from Covid -19 *Data Project (2021)/Track – Daily Covid-19 #s/*  in Google Drive.

<https://docs.google.com/spreadsheets/d/10NtcPW5dT_mpJVLJCl7ZEHOjYTxzXl7LJMSpaOsiVS0/edit#gid=311031188>

This data entry sheet be replaced regularly by a new file in Google Drive. When this happens, you will find the data entry sheet you QAed last week in position: *Covid -19 Data Project (2021)/Track – Daily Covid-19 #s/QA Team/Historic Team Entry Sheets/All Historic Team Entry Sheets*, and it’s renamed “HISTORIC…”

3.3 In the current data entry sheet, copy the most recent total deaths column, from “Albany County, New York” to “Yates County, New York”. Then open local file input/ny\_death\_#entry\_reset.xlsx and paste the copied data in a new column. Write the date on that column. ATTENTION: our data has a one-day lag compared to the data entry sheet. So e.g., if we copied and pasted the total deaths of 07/10/2021 from the data entry sheet, we write “2021-07-11” in input/ny\_death\_#entry\_reset.xlsx. Make sure the format of dates is consistent in input/ny\_death\_#entry\_reset.xlsx. Save the file.

3.4 In the context above, open ./scripts/editing\_v2.R. Find comment “## Include all reference dates for resetting”, add “2021-07-11” to c("2021-02-13", "2021-03-19", "2021-04-03", "2021-04-18", "2021-04-30", "2021-05-15", "2021-05-29", "2021-06-11", "2021-06-26").

3.5 In ./scripts/editing\_v2.R, find the most recent comment that looks like “## 06/26/2021. Reset the total deaths of Non\_NYC counties after (including) 06/26/2021. Copy the code from the comment to “rm(ny\_v2\_temp)” and paste it to a new .R file. Change the dates so they match the new reference date for resetting. Put the code back in ./scripts/editing\_v2.R below last “rm(ny\_v2\_temp”.

3.6 Run everything in editing\_v2.R. Done!!!