In-Class Assignment 19

Pulling Data from the Web

Follow the instructions below and answer the questions that follow. Add your solutions to the R script called **In-Class Assignment 19.R** and submit to Canvas by the deadline listed above. Save your file frequently to avoid losing work!

Instructions:

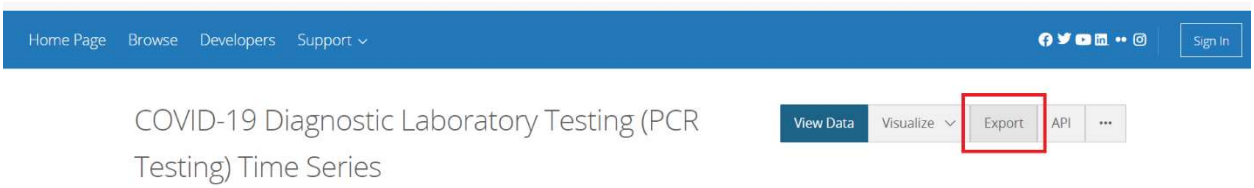
Follow the steps below to pull data from the web and provide insights based on the data you retrieved.

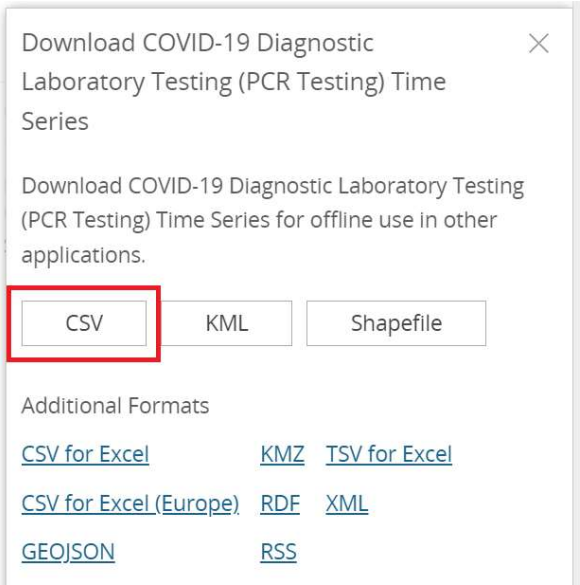
Part 1 – COVID-19 Diagnostic Laboratory Testing (PCR Testing) Time Series:

Use the read.csv() function in R to pull csv data directly from the web.

Go to the following web page: https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-icvb

Click on the **Export** button, then right-click the **CSV** button and select **Copy Link Address** to obtain the web address of the csv file. Add it to the appropriate place in the R script.





In-Class Assignment 19

Use the sqldf package to answer the following questions. **Include answers to the questions in comments under the corresponding SQL code in the R script.**

Be sure to examine the dataset to understand what is contained in it before writing your queries.

To access the data dictionary for the dataset, scroll down to the section entitled **Columns in this Dataset.**

Questions:

1. How many distinct states and FEMA regions are reported in this dataset?

2. Write a query to display the earliest reporting date for each state. Did every state start reporting on the same date?

3. Using your query from Question 2 as a subquery, find the state or territory that started reporting the latest. Give the state name and the date that state/territory started reporting. No need to account for ties.

4. What is the total number of positive, inconclusive, and negative PCR test results across all states/territories as of 12/10/21? (**NOTE:** to refer to a date in your query, use the format '2021/12/10')

5. Which state/territory had the highest number of new positive results reported in a single day? Be sure to account for ties if multiple states/territories or multiple days share the highest number. At minimum, give the name of the state/territory, the date, and the number of new positive results reported on that highest day. (**HINT**: utilize a subquery)

Part 2 – NYC DOHMH Restaurant Inspection Results:

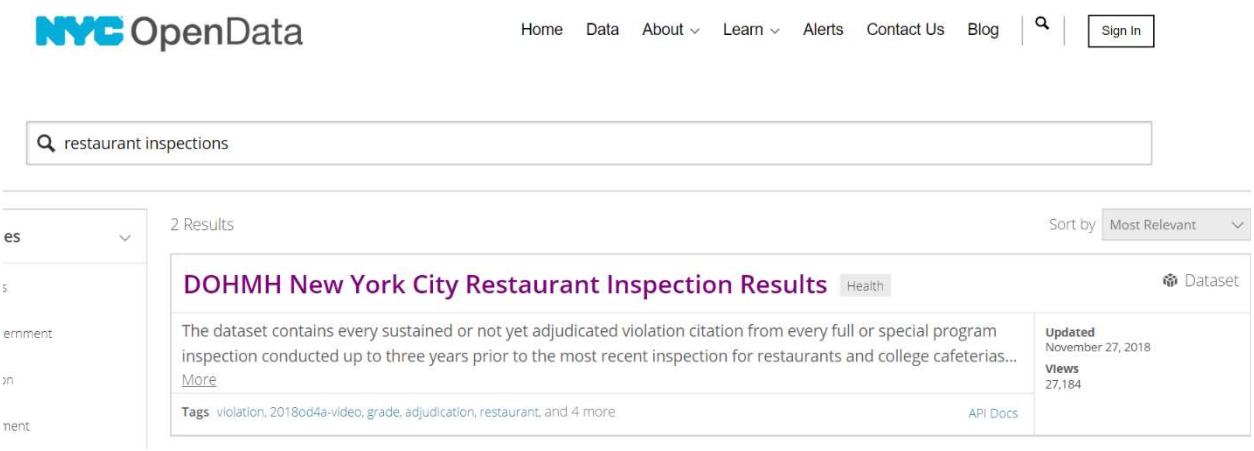
Use the GET() function from the httr package in R to pull JSON data directly from the web.

Go to the following website for NYC open data: https://opendata.cityofnewyork.us/

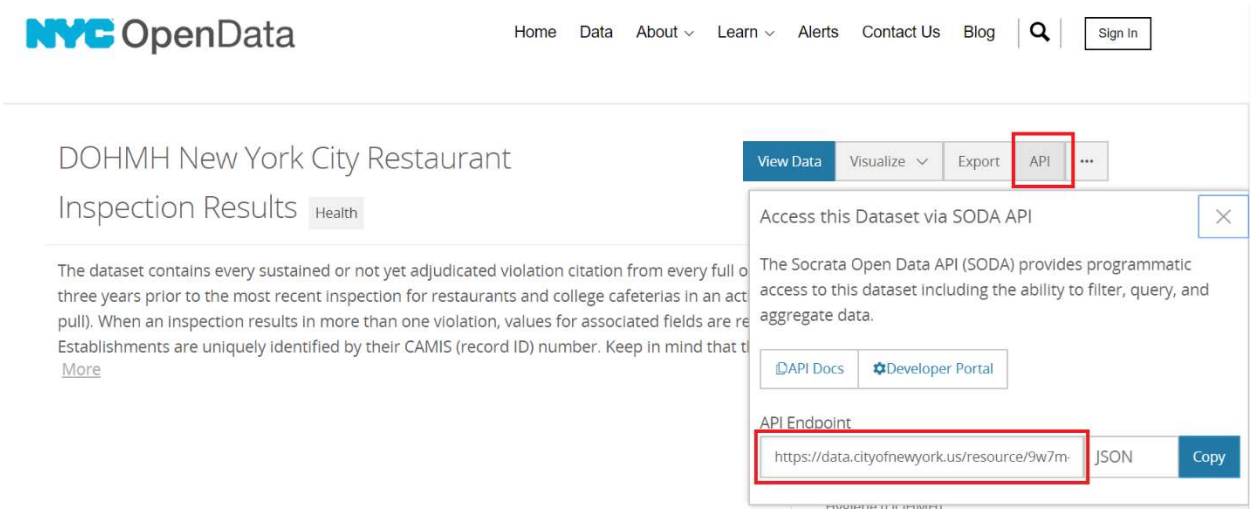
In the search bar, enter **Restaurant Inspections.** Then click the link for the first result: **DOHMH New York City Restaurant Inspection Results**

Continued on next page

In-Class Assignment 19



In the next window, click **API** and then copy the link that appears in the resulting window. This is the link you will include in the GET() function in httr.



Follow the remaining data processing steps in R to convert the JSON data into a data frame that can be analyzed in sqldf. Then use the resulting data frame to answer the following questions about restaurants in the campus neighborhood. **Include answers to the questions in comments under the corresponding SQL code in the R script.**

Continued on next page

In-Class Assignment 19

Questions:

6. How many critical violations are reported in this sample of inspections?

7. Give the name and address (building, street) of restaurant(s) with the highest number of critical violations. Account for possible ties in your results.

8. Similarly to question 8, give the name and address (building, street) of restaurant(s) with the most A grades. Account for possible ties in your results.

9. Create a data frame called **closed** containing restaurants that were indicated to be closed in the action field. The data frame should contain the restaurant name, address (building, street), inspection date, and action.

10. List the restaurants included in the **closed** data frame and order them by number of closures, from most to least. Include restaurant name and address.

11. Use SQL to answer a question of your choice about restaurant violations in the campus neighborhood.