COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

In-Class Assignment 19

Due Saturday, 12/5 by 5:30pm

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:

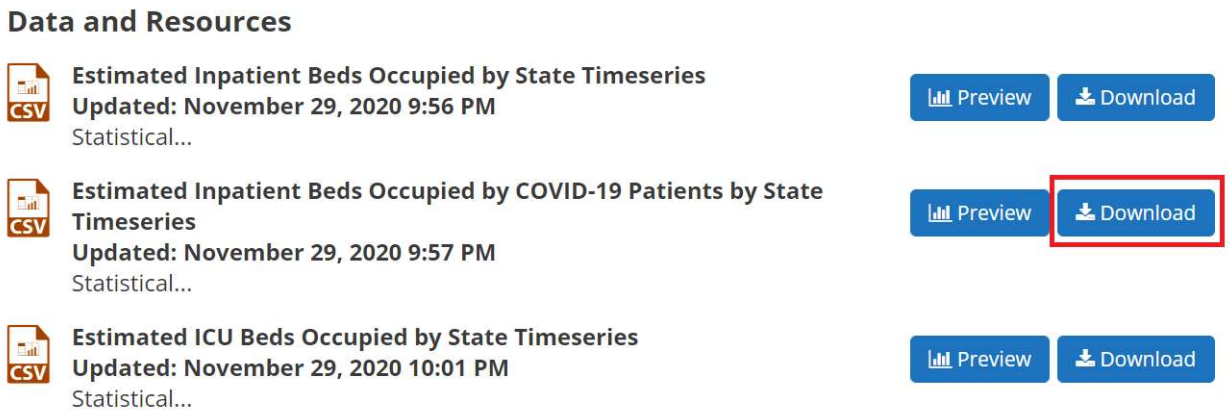
Follow the steps below to pull data from the web and provide insights based on the data you retrieved.

Part 1 – Inpatient Beds Occupied by COVID-19 Patients by State:

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

Go to the following website: https://healthdata.gov/dataset/covid-19-estimated-patient-impact-and-hospital-capacity-state

Under the **Data and Resources** section, right-click the second **Download** 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.



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.**

NOTE: when referring to variables containing a . in the name, refer to that variable name in square brackets (i.e., SELECT [variable.name] …)

To access the data dictionary for the dataset, click the **Preview** button found next to the **Download** button for the dataset.

In-Class Assignment 19

Questions:

1. Are the earliest and most recent dates reported the same for all states/territories? If so, give the dates of earliest and most recent reporting. If not, list the different dates reported by state.

2. Which state/territory had the highest percentage of inpatient beds occupied by COVID-19 patients on 11/28/2020? What percentage of inpatient beds were occupied?

3. Which date and state/territory had the highest ever percentage of inpatient beds occupied by COVID-19 patients since recording began?

4. What were the 3 worst days (by percentage) recorded for inpatient beds occupied by COVID-19 patients in New York State? Report the date, number, and percentage of beds occupied.

5. Which state/territory appears to have been most successful during the reporting period in keeping COVID-19 patients out of inpatient beds? Base on the average percentage of inpatient beds utilized over the reporting period.

6. Using the same query you wrote in #5, adjust it to show the top 10 most successful states during the reporting period. Is New York State among the top 10? If yes, give its rank and average percentage.

Part 2 – NYC DOHMH Restaurant Inspection Results:

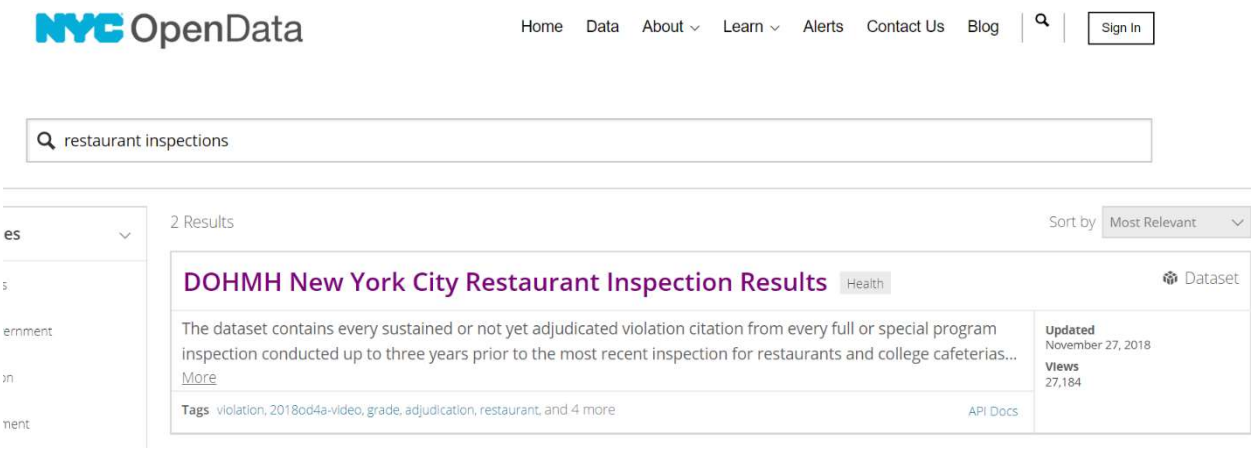
Use the GET() function from the httr package in R to pull JSON data directly from the web.

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

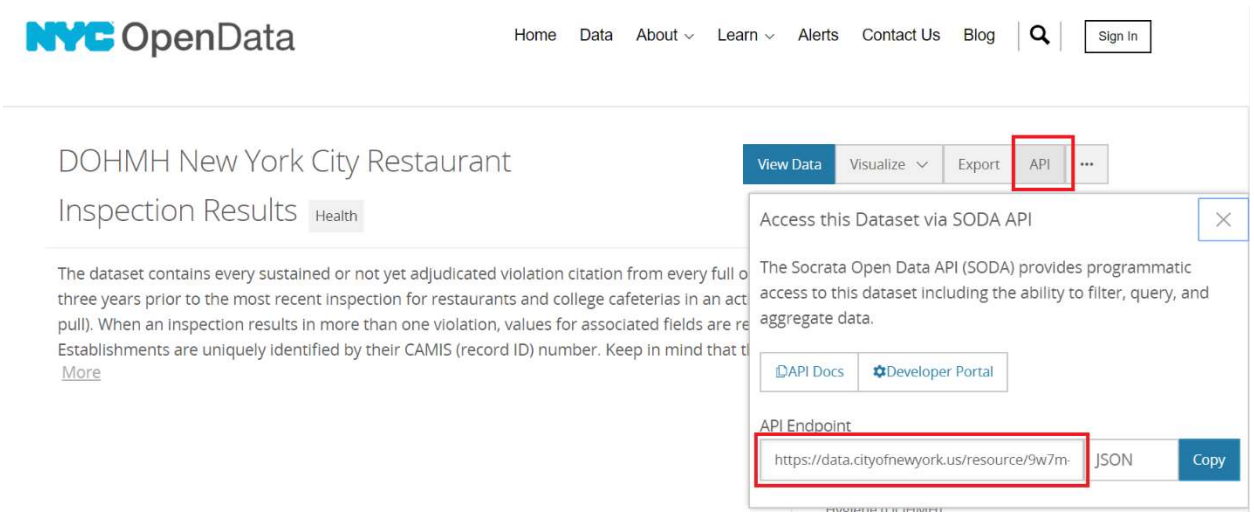
In the search bar, enter **Restaurant Inspections.** Then click the link for the first result: **DOHMH New York City Restaurant Inspection Results**

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 neightborhood. **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:

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

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

9. 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.

10. 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.

11. 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.

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