Karen Bates

SI 507 Final Project Data Checkpoint

**GitHub Link to Repository:**

<https://github.com/karbates/final-project-covid19.git>

**Data Sources**

**Kaiser Family Foundation**

* URL: <https://www.kff.org/other/state-indicator/adults-at-higher-risk-of-serious-illness-if-infected-with-coronavirus/>
* Format: HTML

I scraped the page for the percent of the adult population that is at an increased risk for severe illness or death should they contract COVID-19 due to their health status. After scraping, I organized the information in a dictionary and wrote the dictionary to a csv. The csv was then used to create a dataset in the database. I was originally planning to collect this information from each state’s webpage so I do have caching implemented. However, it isn’t very useful since the information I scraped is static and all on one page.

There are only 102 pieces of data pulled from the tags (50 states plus one territory and 51 percentages) that are zipped into 51 rows. While it’s small, I felt that it was important information that may help contextualize each states’ situation so I wanted to include it. If I am able to, I also discovered tables on the Kaiser Family Foundation website that are downloadable so I might download them since scraping is impractical. I will then add the data (percent of adults in the state that are obese and number of hospital beds in a state) as columns to existing dataset showing the at risk population.

Screenshot of cached data for at-risk population from KFF webpage:

**A picture containing water, people, white

Description automatically generated**

**COVID Tracking API**

* URL(s):
  + <https://covidtracking.com/api/states/daily>.json
  + <https://covidtracking.com/api/states/info>.json
* Format: JSON

I set up a function to call the ‘states daily’ extension of the COVID Tracking API to collect data on a certain state (when the state is input as a parameter). This information is then cached and will be formatted to display the most pertinent information to the user in an HTML page using a Flask App. The ‘states info’ extension of the API returns more static information about a state including it’s respective health department and affiliated websites, and the state health department’s Twitter handle. I’ve downloaded the csv file for the data and read it into the database as its own dataset.

The COVID Tracking team began collecting data from Washington State on 2/28/2020. There is one record for each state since the first day that state reported a confirmed case of COVID-19. As of 4/19/2020, calling the “states daily” filter and including all states, returns 2,508 dictionaries with information by date and state. When one state is searched, the number of dictionaries returned is around 50, depending on when the first cases were found in the state. The “states info” file contains about 50 rows with 9 columns, one for each state and a few territories may be included as well.

Below is a work in progress function to get data from COVID Tracking API using the COVID\_CACHE and the make\_request\_using\_cache function that it references. The output currently isn’t saving in a cache and I’m not sure how to get the caching to work since the data is updated regularly. I will probably need to find a way to use the date in the unique key as well.

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**Twitter**

* URL: <https://api.twitter.com/1.1/search/tweets.json>
* Format: JSON

I’m using the Twitter API to pull recent Tweets from the CDC. I may also call for Tweets from individual state health departments since that information is available from the COVID-19 Tracking database or the NIH as they are also a trusted source of information. If a user opts to see the Tweets, they will be displayed alongside the COVID-19 case counts and other information returned to a user after they select a specific state. CDC Tweets will be shown regardless of the state selected. The Tweets are being cached and a request is being made through the cache if it is available, instead of calling the Twitter API every time.

The maximum number of Tweets Twitter returns is 500 but it’s also limited to Tweets from the last 30 days. I’m estimating that the number of records returned is between 100 and 500 for Tweets from the CDC.

Screenshot of cached Tweets:

**A close up of text on a black background

Description automatically generated**

**Database**

The database will consist of two tables with health-related information on states and COVID-19 case counts. The two tables with both have the state name that could act as a Primary Key/Foreign Key. But ideally, I will be able to use the FIPS code for each state which is a standardized number associated with each state. One table does not currently have the FIPS code yet so I need to ensure I can add it as a column first.

The first table will be based on information coming from the Kaiser Family Foundation website. It will contain the FIPS code (if I can add it), the state name, and the percentage of the state’s population that is at-risk. To build the table out and make it more substantial, I’m also hoping to add information from downloaded csv files since I’m not able to scrape it from the pages. The added information will include the percent of adults that are obese in the state (since obesity is suspected to be a strong indicator of negative health outcomes) and the number of hospital beds in a state. The second table will contain static data from calling the COVID-19 API. It will not include the number of cases or other dynamic information but it will contain the state FIPS code, the state two letter abbreviation, one or two (if more than one is listed) websites for the state’s health department and health resources, and the state’s Twitter account.

Below is a screenshot of the table containing the percent of a state’s population that is at-risk, titled “AtRiskPopulation” and a table with state information relating to COVID-19 titled “StateInfo.”

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a social media post

Description automatically generated**

**Interaction and Presentation**

Information will be presented to the user through a Flask App. The user will have the option to select a certain state from a drop down menu. This will take the user to a page that contains information on that specific state including the state’s percentage of at-risk population, the number of confirmed COVID-19 cases, the total number of recovered cases, and the total number of deaths in that state.

When selecting a state, the user will also have options to select one or more media items related to that state, recent COVID-19 related headlines (national), and recent Tweets by the CDC.

I am also tentatively hoping to include a graph of the confirmed cases for that state using Plotly. However, I’m still working through how I could compile the dates and confirmed cases so they can be graphed. Another option I am considering exploring is a visualization to put the state’s percentage of at risk population in context, relative to other states.