Karen Bates

SI 507 Final Project Data Checkpoint

GitHub Link to Repository:

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

## Data sources (1/2 - 1 page)

* For each
  + origin, including URLs for data and documentation
  + format(s) (e.g., JSON, CSV, HTML)
  + how you accessed the data, and whether caching was used
  + summary of data
    - # records available (OK to estimate)
    - # records retrieved (OK to estimate)
    - description of records, including important fields/attributes of each record for the purpose of and what they represent
  + evidence of caching - show that you used caching to retrieve data where appropriate (e.g., for data access via APIs or scraping). The evidence could be one or more of the following:
    - a snapshot of a chunk of your cache file
    - indication of the parts of your code that implement caching
    - a link to a brief video showing the difference between running your data access program with and without caching

## Database (1/2 - 1 page)

* Database schema (SQL CREATE TABLE statements indicating table names, fields, data types, and constraints)
* Any foreign key-primary key relations among your tables
* Screenshots showing some of the data in each of your tables

## Interaction and Presentation Plans (1/2 page)

* High-level, plain-English description of the user-facing capabilities of your project—what options does the user have for selecting and displaying data?
* Interactive and presentation technologies you plan to use (e.g., Flask, Plotly, command line prompts)

**Database**

The database will consist of two tables with health-related information on states and COVID-19. 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 have the FIPS code with it yet so I need to ensure I can add it as a column first.

The first table will be primarily based on information coming from the Kaiser Family Foundation website. It will contain the FIPS code, the state name, and the percentage of the state’s population that is at-risk. 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 things like the total hospital bed capacity in the state and the state’s health department website.

**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. 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.

|  |  |  |
| --- | --- | --- |
| **Component** | **Requirement** | **Points** |
| **Project code** | Valid GitHub repo link is provided and contains data access code | 5 |
| **Data sources** | Origin and documentation for each data source is provided accurately | 5 |
|  | Access techniques are clearly described | 5 |
|  | Caching is used where appropriate and evidence is provided | 5 |
|  | Data summary is provided and relevant data fields are described | 5 |
| **Database** | Schema and screenshots are provided and schema makes sense given the description of data source records and relevant fields | 5 |
|  | Relations are accurately described and make sense | 5 |
| **Interaction/**  **Presentation** | Plans for application capabilities and interactive/presentation technologies are described clearly and make sense | 5 |
|  | **Total** | **40** |

For my final project, I would like to leverage Health API: COVID-19 and the Kaiser Family Foundation webpages with state-by-state statistics on the percentage of high risk adults in a state. My objective is to create an interactive program that allows a user to enter a country or state for which they would like to see the total confirmed cases, deaths, and recoveries from COVID-19, using the Health API. If the user entered a state, they will be asked if they would like to see the percentage of adults in that state that are classified as high risk due to underlying health concerns. To find this information, the program will crawl and scrape the Kaiser Family Foundation’s website which has a page with this information for each state. Finally, the user can either enter a new country or state to see their respective stats on COVID-19 cases.

A few other ideas I may consider implementing, are a graph of the COVID-19 cases for a state or country that is searched, depending on how I can implement caching and leverage the “last update” key in the results. Additionally, because I enjoyed using the News API and it could help to contextualize the results, I am considering returning the prior day’s headlines related to COVID-19 for the country that was searched in the first step.

The Health API: COVID-19 is a new-to-me (and probably everyone) API that does not require any authorization. I have successfully called the API to return information. I have used the Kaiser Family Foundation website for reading articles but never for scraping and I could not find any information regarding their T’s & C’s and crawling and scrapping their pages. I successfully returned the soup from one of the state pages I would like to work with but was having some difficulty extracting the exact text I want from the table. This project has a challenge score of 11 points since the API is new to me and the crawling will involve navigating multiple pages.

Health API: COVID-19 Link:

<https://health-api.com/#data-source>

Kaiser Family Foundation State Information:

* Potential beginning point to build a dictionary of links for each state like the National Parks Project: <https://www.kff.org/statedata/>
* Each state also has a page with the state’s health state which contains a link to the page that has stats on adults at higher risk of serious illness: <https://www.kff.org/state-category/health-status/?state=MI>
* Full link for the table containing the desired information: <https://www.kff.org/other/state-indicator/adults-at-higher-risk-of-serious-illness-if-infected-with-coronavirus/?currentTimeframe=0&selectedRows=%7B%22states%22:%7B%22michigan%22:%7B%7D%7D%7D&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>