Kiana Ocean (kocean@andrew.cmu.edu)
Joe Liang (wjliang@andrew.cmu.edu)
Xiaoye Shao (xiaoyes@andrew.cmu.edu)

README
Covid-19 Community Impact Platform
Data Focused Python (95888-A1) Final Project

1. Prior to running our app, in addition to the default Anaconda packages, you must have the **plotly** Python library installed.

   To install plotly using conda, run the following command in your terminal:

   ```
   % conda install -c plotly plotly=4.11.0
   ```

   Alternatively, to install using pip, run the following in your terminal:

   ```
   % pip install plotly==4.11.0
   ```

2. A comprehensive list of the packages required to run the application are as follows. We assume these are installed with Anaconda based on the documentation found here: https://docs.anaconda.com/anaconda/packages/pkg-docs/

   - beautifulsoup4 4.9.3
   - certifi 2020.6.20
   - chardet 3.0.4
   - cycler 0.10.0
   - idna 2.10
   - kiwisolver 1.2.0
   - lxml 4.5.2
   - matplotlib 3.3.2
   - numpy 1.19.2
   - pandas 1.1.2
   - Pillow 7.2.0
   - pip 20.2.2
   - plotly 4.10.0
   - pyparsing 2.4.7
   - python-dateutil 2.8.1
   - pytz 2020.1
   - requests 2.24.0
   - retrying 1.3.3
   - setuptools 49.6.0.post20200814
   - six 1.15.0
   - soupsieve 2.0.1
   - urllib3 1.25.10
   - wheel 0.35.1
   - wincertstore 0.2

3. Download the DFP-CCIP folder to the desired location on your computer.

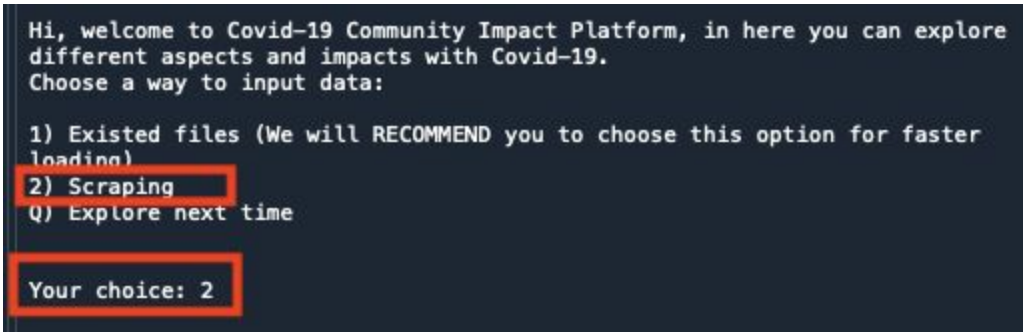
4. DFP-CCIP contains app.py, the main file you will use to run the application. DFP-CCIP also contains three sub-folders:
   a. **data:** Raw data files used in the application are stored here. When the scraping portion of the app is run, the scraped data will be saved here.
   b. **dfp_ccip:** Contains the supporting modules required to run the application. Three subfolders contain modules to perform the following functions:
      i. **factory:** Scrape and combine data

ii. **validation:** Validate user inputs
iii. **ccip_utils:** Analysis and visualization of the scraped data

c. **output:** If you choose to download data from the application, it will be saved here by default if you do not specify an output path.

5. To run the application, open app.py (DFP-CCIP/app.py) and execute the program.

6. The **first time** you run our application, you will need to input the data through the scraping function. To do this, choose option 2 in the initial welcome menu. The scraping process will take approximately 2 minutes.



Once the scraped data is saved to your computer, the subsequent times you run our application you can bypass the scraping process by selecting option 1 from the welcome menu.

7. You will be given the option to view data at the county, state, or country (United States) level. If you choose county or state, you will be prompted to enter the name of the location for which you would like to view data. The state and county names you enter must match exactly (ignoring case) to the state and county names in our data. (E.g., If you're looking for St. Louis County, Missouri, "St Louis" will be an invalid county name, but "St. Louis" will be valid, as will "st. louis").

8. In addition to a table summarizing the key facts for your specified location, you will be able to view data on four separate aspects of the impact of Covid-19 in your location of interest.
    a. Health: Counts of Covid-19 cases and deaths
    b. Economy: Unemployment rates
    c. Health + Demographic: Population adjusted counts of Covid-19 cases and deaths
    d. Health + Economy: Covid-19 cases and unemployment rates over time

9. If you want to view data for specific dates, please ensure you follow the instructions on how to format the dates properly, as outlined in the application. Note the following:
    a. For the health related data, data is daily and available from February 1st. Dates should be entered in M D format, e.g., "3 2" for March 2nd or "3" for March.

b. For the economic data, unemployment data is monthly and pulled for the past 14 months. Note there is about a 1 month lag from when data is collected and published. For example, as of October 16, 2020, the latest month of data available is August 2020. Indicate the month for which you would like to view data in form MM/YYYY. If you would rather view a time series chart of all available data, enter "N/A" or simply press enter. If the month you select is outside of the 14 month range of available data, you will be shown data for either the first or last month of available data. For example, if 01/2017 is entered, you would be shown data for 07/2019. If 10/2020 is entered, you would be shown data for 08/2020.

10. After you are shown a chart or map, you will be given the option to download the data used to generate the plot. You may choose the path to download the data or it will be downloaded to the output folder by default.

11. There are two .txt files within the data folder that contain state names and their corresponding abbreviations. The sources for those files are https://gist.github.com/bensie/130828 and https://www.800florals.com/care/state-abbreviations.asp.

12. Another file, demo_state.txt contains population data by state, and is from https://www.infoplease.com/us/states/state-population-by-rank. This data is used in place of the scraped demographic data in the health functions due to the lack of population data in our merged data in many counties in the early months of the pandemic. In all other functions we use the demographic data we scraped.

13. **Important caveat regarding data accuracy:** Because the Covid-19 case counts are updated on a daily basis and are being compiled from many different sources by a number of journalists, there may be some inaccuracies within the data. Additionally, because we are merging county-level data across several sources, there could be some inconsistencies in the way county-level data is collected and counted by each respective source. Ideally, if we were to continue development of this platform, we would identify such inconsistencies to ensure accuracy of the data.

Appendix A

This flow chart outlines the general flow of our application.