Anthony Wilson

DSC 540 - Milestone 1

* 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)
  + <https://covid19api.com/#details> (API)
    - This API has data about the number of cases in each country, by date. So we can pull prior data since it has been reported. It looks like the data is as of January of 2020.
  + <https://washdata.org/> (CSV)
    - This is the WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP). The data measure how well the water and hygiene are within the country and region. They look at schools, health care facilities, and homes. They have the populations within each country and the percentage of coverage within each country as well as the service level. The service level is categorized by “None”, ‘limited, “basic” and “insufficient data”. “Basic” means that they have the minimum amount of coverage, which their goal is to get 100% coverage for all countries with basic coverage. Data goes back to 2000 and all the way up to 2016.
      * I plan to pull from this table <https://washdata.org/data/healthcare#!/table?geo0=region&geo1=sdg>
  + <https://news.google.com/search?q=covid-19%20cases%20symptoms%20globally&hl=en-US&gl=US&ceid=US%3Aen> (Web Site)
    - Google news - this is a central location for all news repositorires. It will have mulitple sources to pull from along with their websites to hit. From the news article links I can look for more data on this project. I plan to search for general information on symptoms, deaths, causes, etc… for countries. I want to build a repository of common themes among the different countries to try to identify which countries have shared information. I will use this to see if that helps identify the rate at which the virus spreads. I don’t know if this is overkill for my first webscrape, but we shall see.
* The relationships between them, or the relationship you will make between them:
  + The relationshiop between the three data sets will be the country names. The base or main table will be the api. I will use the countries provided there to pull infromation from the other data sets.
* What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words:

For the next milestone, I will need to decide if I am going to go to the website to pull the CSV files or if I am going to just download them. I don’t think that there will be a ton of work involved with needing to transform the data too much. I might have multiple flat files to load though. The web scrape, I had a hard time coming up with this one, just because I am not as familiar with it. I think this one will require the most work on my side. I want to scrape google/news for other news sites and try to pull data on what is happening. I want to find information around symptoms, severity, cause, death, recovered, and other items that relate to individuals within each country. I want to see if the information shared about the virus helps with slowing down the spread. My current plan for the web scrape is to create some sort of class that I can build a web scrape foundation. I will probably need to build the API foundation before the web scrape, so I can look up the countries for the web scrape search. It has been many years since I built an API pull with python, once I figured it out it didn’t seem to difficult to work with. I am thinking this will be the case.

Once I have the data gathered, merging them together should be pretty easy, I say that now. My goal, right now, is to combine them by country. I will probably use Mysql since I have used it before. The visualizations should be fairly interesting when combining keyword searches, hygiene, and COVID-19. I am more interested in the hygiene compared to the COVID-19 spread.

WHO/UNICEF. (n.d.). Home: JMP. Retrieved April 6, 2020, from https://washdata.org/

Coronavirus COVID19 API. (n.d.). Retrieved April 2, 2020, from https://covid19api.com/#details

Google. (n.d.). Retrieved April 6, 2020, from https://news.google.com/search?q=covid-19 cases symptoms globally&hl=en-US&gl=US&ceid=US:en