Proposal: Corona virus vs H1N1 flu within the united states.

Websites we will be using:

<https://en.wikipedia.org/wiki/2009_flu_pandemic_in_the_United_States>

<https://www.worldometers.info/coronavirus/country/us/>

What we will be webscraping:

Table on Wikipedia showing h1n1 cases/hospitalization/death rate of the H1N1 virus within each state

Table on worldometrics showing corona virus cases total cases/ total death/ new cases/ new death/ active cases within each state

Possibly a timeline of h1n1

Data cleaning:

Remove hospitalization line of h1n1 table

Remove active cases, new cases new deaths and sources from corona table

Put only numbers in h1n1 (remove sup id and NLU from each cell)

Using Structured database such as SQL:

Why use a structured database instead of Mongodb:

Structured dataset attributes itself better to MySQL such as pgAdmin

Lower amount of data makes sense for MySQL

Starting out:

* For our project, the goal was to compare covid19 to another recent pandemic.
* The most recent other pandemic was back in 2009 with the H1N1 flu.
* While looking for info on the web we were able to find a table on Wikipedia showing a breakdown of H1N1 cases and deaths by states.
* Additionally we found a similar table of current confirmed cases for Covid19 on a website called worldometers.info.

Data collection:

* Using python and pandas we were able to scrape both sites to get the tables using pd.read\_html and converted them into DataFrames, making sure to change the column names to work within SQL.

Data cleanup:

* The first step for cleaning up we had to do was drop unnecessary columns, such as hospitalizations in the H1N1 table and new cases, new deaths and active cases within the Covid19 table.
* Next we had to get rid of the superscript on the Wikipedia table which was used to link the info to which article it was taken from. (Useful but not necessary for what we needed)
* Additionally we had to get rid of any other info that wasn’t the number for each datapoint (ex: “/NLU” & “wave 2”)
* Once that was done, we had to remove all commas from numbers that were above 1000 so that they could be added to SQL as INT.
* Finally, we changed the way some of the names of states were spelled so that they would match so we could have a clean dataset, this included places such as U.S Virgin Islands and District of Columbia.
* Export to Csv using df.to\_csv.

SQL:

* We used SQL and pgAdmin 4 for our database.
* The data we were taking as already incredibly structured and relational so using SQL made more sense and the data set itself was rather small.

Findings:

* From what we gathered within less than 3 months of Covid19’s introduction into the US it has more confirmed cases and nearly as many confirmed deaths than over a year of the H1N1 “pandemic”. We can conclude that not only will is spread faster than H1N1 but also kill many more people. If we look at New York for example, in 2009 during the H1N1 pandemic, we can see that there were 2,738 confirmed cases and 206 confirmed deaths. During the Covid19 pandemic, New York had 52,000 confirmed cases and 728 deaths at the time of our web scraping on Saturday and has since jumped to 67,000 confirmed cases and over 1,300 total deaths.
* As a interesting data point California which is number 3 in confirmed cases for Covid 19 was affected most by H1N1 in 2009 with 10,000+ cases and 600+ deaths.