Get Data on the PDF

Takes a Kentucky PDF of Covid data and saves as a link. I utilize a python library called tableau-py which can take the coordinates of a table in the pdf and transform into dataframe. First thing the scraper gets is the date. One thing to note is if the PDF format changes, this scraper may no longer work because defined coordinates do not match the table.

Get Total Deaths and Total Cases

The next dataframe the pdf scraper gets is the confirmed cases, probable cases, total deaths, confirmed deaths, probable deaths...etc There is some added transformation involved to get the date on this table as well as to get total cases on the table.

Race of Cases Where Race is Known

Next is a data frame of cases by race by percentage. The unknown row is calculated based on how much of race is known from the pdf, then does a calculation.

Ethnicity of Cases Where Ethnicity is Known

Next is the ethnicity by cases by percent. The unknown is a calculated field based on the % of cases known as indicated by the pdf.

Race of Deaths Where Race is Known

Next is the race of deaths by percentage then calculated by multiplying percent by total deaths. The Unknown row is a calculated field based on the deaths by race which are unknown. Kentucky has been sometimes inconsistent whether they include Native Hawaiian or Other Pacific, and/or American Indian or Alaska Native.

Ethnicity of Deaths Where Ethnicity is Known

Next is the ethnicity by deaths by percent. The approach is similar to ethnicity in cases where ethnicity is known.

Converting New Scraped Covid Data into Same Format as CRDT Table

In order to upload the data to CRDT and for easy comparison between old data and newly scraped data, these cells are transforming new data to the same format as the CRDT table. One thing to note is under total deaths we add confirmed deaths and probable deaths together.

Data Validation

This was to simulate the data validation check in the CRDT data entry
Does four checks (if not zero will highlight red):

- Difference in sum of cases by race v. total confirmed cases
- Difference in sum of deaths by race v. Total confirmed deaths
- Difference in sum of cases by ethnicity v. Total confirmed cases
- Difference in sum of deaths by ethnicity v. Total confirmed deaths

Create a csv for today’s scraper data

On your computer if you would like to create a new file csv with today’s newly scraped data as the first row. If you already have a file on your computer with historical indiana data, you can avoid this cell.

Add today’s scraper data to historical data

If created a csv in the past and would just like to append new data to it. Use this cell.

Future Updates

I will continue to work on the data quality checker for Kentucky over my winter break and into next semester.