Project Milestone 1

Project Milestone 1

Astrid Fuentes, MS

Bellevue University

DSC 540: Data Preparation

Prof. Benjamin Schneider

January 8, 2021

Introduction

Throughout the term, we will be working on a project that has 5 milestones, with each milestone being due every two weeks, beginning in Weeks 3 & 4. The goal of the project is to build on the materials covered during each 2-week period. This week, we are working on milestone. We need to Select the data we want to work with.

- Select 3 different data sources that have different file types of information – and the data will need to have a relationship between them, or such relationship will need to be created.
- Select one of each of the following types of datasets – and a minimum of 1000 rows across all datasets. Each dataset should have a minimum of 10 columns/variables.
- Types of dataset needed: CSV/Excel/PDF or another flat file source, website with table formatted data, API.

For this project, I have decided to work with covid19 data. Given the high impact this pandemic has had over all of us. I am confident that I can find plenty data.

Data Source 1

The Centers for Disease Control and Prevention (CDC) is a great source of health data. I was able to find the "Provisional COVID-19 Death Counts by Sex, Age, and week" (Provisional_COVID-19_Death_Counts_by_Sex__Age__and_Week.csv) data set containing deaths involving coronavirus disease 2019 (COVID-19) are reported to NCHS by sex and age group and week ending date. This data is categorized as National Center for Health Statistics (NCHS) data and is considered public.

This data set was created on May 15, 2020 and is updated weekly. Number of deaths reported in are the total number of deaths received and coded as of the date of analysis, and do not represent all deaths that occurred in that period. Data during this period are incomplete because of the lag in time between when the death occurred and when the death certificate is completed, submitted to NCHS and processed for reporting purposes. This delay can range from 1 week to 8 weeks or more.

The "Provisional COVID-19 Death Counts by Sex, Age, and week" csv file contains 8 columns and 1909 rows. I plan on adding more data to it to make it 10 columns. The current columns in the file are: Data as of, State, MMWR, Week, End Week, Sex, Age Group, Total Deaths, COVID-19 Deaths.

Data Source 2

For the table style website, I decided to use Worldometer. This is a reference website that provides counters and real-time statistics for diverse topics. It is part of the Real Time Statistics Project, and managed by international developers, researchers and volunteers.

This website has been tracking covid19 cases since it started about a year ago in China. It shows a table structured data set that allows us to see number of cases and deaths, among other information for each country, state, county, etc. around the world. For this project, I will be using United States data and include study specific data for states like Florida and California.

The tale has the following fields: Country/State/County, Total Cases, New Cases, Total Deaths, New Deaths, Total Recovered, Active Cases, Tot Cases/1M pop, Deaths/1M pop, Total Tests, Tests/1M pop, Population, Source, Projections.

It is important to mention that when the table is showing rows by state, there are not over 1000 rows. However, some states like Florida and California, allow you to look at data by county. In this case, I can extract information for specific states to make the dataset over 1000 rows.

Data Source 3: API

For the API for this project, I selected The Covid tracking project API. The API's documentation has a complete codebook which I have included below.

Fields

date

Field type: integer

Date
Date on which data was collected by The COVID Tracking Project.

dateChecked

Field type: string

Deprecated. This is an old label for *lastUpdateEt*.

death

Field type: integer

Deaths (confirmed and probable)
Total **fatalities with confirmed OR probable COVID-19 case diagnosis** (per the expanded CSTE case definition of April 5th, 2020 approved by the CDC). In some states, these individuals must also have COVID-19 listed on the death certificate to count as a COVID-19 death. When states post multiple numbers for fatalities, the metric includes only deaths with COVID-19 listed on the death certificate, unless deaths among cases is a more reliable metric in the state.

Returns nullif no data is available

deathIncrease

Field type: integer

New deaths

Daily increase in death, calculated from the previous day's value.

Returns nullif no data is available

hash

Field type: string

A hash for this record

hospitalized

Field type: integer

Deprecated. Old label for hospitalizedCumulative.

Returns nullif no data is available

hospitalizedCumulative

Field type: integer

Cumulative hospitalized/Ever hospitalized

Total number of individuals who have **ever been hospitalized with COVID-19**. Definitions vary by state / territory, and it is not always clear whether pediatric patients are included in this metric. Where possible, we report patients hospitalized with confirmed or suspected COVID-19 cases.

Returns nullif no data is available

hospitalizedCurrently

Field type: integer

Currently hospitalized/Now hospitalized

Individuals who are **currently hospitalized with COVID-19**. Definitions vary by state / territory, and it is not always clear whether pediatric patients are included in this metric. Where possible, we report patients hospitalized with confirmed or suspected COVID-19 cases.

Returns nullif no data is available

hospitalizedIncrease

Field type: integer

New total hospitalizations

Daily increase in *hospitalizedCumulative*, calculated from the previous day's value.

Returns nullif no data is available

inIcuCumulative

Field type: integer

Cumulative in ICU/Ever in ICU

Total number of individuals who have **ever been hospitalized in the Intensive Care Unit with COVID-19**. Definitions vary by state / territory, and it is not always clear whether pediatric patients are included in this metric. Where possible, we report patients in the ICU with confirmed or suspected COVID-19 cases.

Returns nullif no data is available

inIcuCurrently

Field type: integer

Currently in ICU/Now in ICU

Individuals who are **currently hospitalized in the Intensive Care Unit with COVID-19**. Definitions vary by state / territory, and it is not always clear whether pediatric patients are included in this metric. Where possible, we report patients in the ICU with confirmed or suspected COVID-19 cases.

Returns nullif no data is available

lastModified

Field type: string

Deprecated. Old label for lastUpdateET.

negative

Field type: integer

Negative PCR tests (people)

Total number of **unique people with a completed PCR test that returns negative**. For states / territories that do not report this number directly, we compute it using one of several methods, depending on which data points the state provides. Due to complex reporting procedures, this number might be mixing units and therefore, at best, it should only be considered an estimate of the number of people with a completed PCR test that return negative.

Returns nullif no data is available

negativeIncrease

Field type: integer

Increase in *negative* computed by subtracting the value of *negative* for the previous day from the value for *negative* from the current day.

Returns nullif no data is available

onVentilatorCumulative

Field type: integer

Cumulative on ventilator/Ever on ventilator
Total number of individuals who have **ever been hospitalized under advanced ventilation with COVID-19**. Definitions vary by state / territory, and it is not always clear whether pediatric patients are included in this metric. Where possible, we report patients on ventilation with confirmed or suspected COVID-19 cases.

Returns nullif no data is available

onVentilatorCurrently
Field type: integer

Currently on ventilator/Now on ventilator
Individuals who are **currently hospitalized under advanced ventilation with COVID-19**. Definitions vary by state / territory, and it is not always clear whether pediatric patients are included in this metric. Where possible, we report patients on ventilation with confirmed or suspected COVID-19 cases.

Returns nullif no data is available

pending
Field type: integer

Pending
Total number of **viral tests that have not been completed** as reported by the state or territory.

Returns nullif no data is available

posNeg
Field type: integer

Deprecated. Computed by adding *positive* and *negative* values.

Returns nullif no data is available

positive
Field type: integer

Cases (confirmed plus probable)
Total number of **confirmed plus probable cases** of COVID-19 reported by the state or territory, ideally per the August 5, 2020 CSTE case definition. Some states are following the older April 5th, 2020 CSTE case definition or using their own custom definitions. Not all states and territories report probable cases. If a state is not reporting probable cases, this field will just represent confirmed cases.

Returns nullif no data is available

positiveIncrease
Field type: integer

New cases

The daily increase in API field *positive*, which measures **Cases (confirmed plus probable)** calculated based on the previous day's value.

Returns nullif no data is available

recovered

Field type: integer

Recovered

Total number of **people that are identified as recovered from COVID-19**. States provide very disparate definitions on what constitutes a "recovered" COVID-19 case. Types of "recovered" cases include those who are discharged from hospitals, released from isolation after meeting CDC guidance on symptoms cessation, or those who have not been identified as fatalities after a number of days (30 or more) post disease onset. Specifics vary for each state or territory.

Returns nullif no data is available

states

Field type: integer

States

Only available in national records. The number of states and territories included in the US dataset for this day.

total

Field type: integer

Deprecated. Computed by adding *positive*, *negative*, and *pending* values.

Returns nullif no data is available

totalTestResults

Field type: integer

Total test results

At the national level, this metric is a summary statistic which—because it sums figures from states reporting tests in **test encounters** with those reporting tests in **specimens** and in **people**—is an aggregate calculation of heterogeneous figures. Therefore, it should be contextualized as, at best, an estimate of national testing performance.

In most states, the totalTestResults field is currently computed by adding positive and negative values because, historically, some states do not report totals, and to work around different reporting cadences for cases and tests. In Colorado, Delaware, the District of Columbia, Florida, Hawaii, Minnesota, Nevada, New York, North Dakota, Rhode Island, Virginia, Washington, and Wisconsin, where reliable testing encounters figures are available with a complete time series, we directly report those figures in this field. In Alaska, America Samoa, Arizona, Arkansas, California, Georgia, Indiana, Kentucky, Maryland, Massachusetts, Missouri, Nebraska, New Hampshire, Ohio, Oregon, Texas, Utah, Vermont, and Wyoming, where reliable specimens figures are available with a

complete time series, we directly report those figures in this field. In Alabama, Idaho, and South Dakota, where reliable unique people figures are available with a complete time series, we directly report those figures in this field. We are in the process of switching all states over to use directly reported total figures, using a policy of preferring testing encounters, specimens, and people, in that order.

Returns nullif no data is available

totalTestResultsIncrease

Field type: integer

New tests

Daily increase in *totalTestResults*, calculated from the previous day's value. This calculation includes all the caveats associated with Total tests/*totalTestResults*, and we recommend against using it at the state/territory level.

Returns nullif no data is available

Conclusion

Unfortunately, as is expected with public health data sets, data is anonymized so this data lacks any sort of patient identifier that could be the typical key field. Despite the lack of participant identifiers, my data sources are related by a common field, the location. If appropriate, I would like to use location (state and/or county) as key to build relationships between my data sets. I would also like to find additional demographics information that I can use to find insights about how covid19 affects different races, ethnicities, etc.

I can think of the following questions I would like to answer with this data set:

1. Are the number of deaths higher in male patients than female?
2. Are seniors' death rate higher than non-senior?
3. Examine the overall trend of the time series and how some of the variables might be contributing to that trend.
4. Are number of deaths higher in certain states during a particular time-period. Possible causes?

In order to be able to accomplish this, I will need to work on completing my data sets, cleaning the data, renaming some of the variables to make them more meaningful, identify outliers, bad data, etc.

References

Center for Disease Control and Prevention: Provisional COVID-19 Death Counts by Sex, Age, and week. https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-W/vsak-wrfu/data

Worldometer: Coronavirus. https://www.worldometers.info/coronavirus/country/us/

The Covid Tracking Project. https://covidtracking.com/data/api