COVID-19 Data - Data Cleaning with SQL

Data Analysis using SQLite3

Count number of cases, per state, by month+year

About Dataset

COVID-19 Case Surveillance Public Use Data with Geography

Case Surveillance

This case surveillance public use dataset has 19 elements for all COVID-19 cases shared with CDC and includes demographics, geography (county and state of residence), any exposure history, disease severity indicators and outcomes, and presence of any underlying medical conditions and risk behaviors.

Currently, CDC provides the public with three versions of COVID-19 case surveillance line-listed data: this 19 data element dataset with geography, a 12 data element public use dataset, and a 32 data element restricted access dataset.

The following apply to the public use datasets and the restricted access dataset:

- Data elements can be found on the COVID-19 case report form located at www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf.
- Data are considered provisional by CDC and are subject to change until the data are reconciled and verified with the state and territorial data providers.
- Some data are suppressed to protect individual privacy.
- Datasets will include all cases with the earliest date available in each record (date received by CDC or date related to illness/specimen collection) at least 14 days prior to the creation of the previously updated datasets. This 14-day lag allows case reporting to be stabilized and ensure that time-dependent outcome data are accurately captured.
- Datasets are updated monthly.
- Datasets are created using CDC's Policy on Public Health Research and Nonresearch Data Management and Access and include protections designed to protect individual privacy.
- For more information about data collection and reporting, please see wwwn.cdc.gov/nndss/data-collection.html.
- For more information about the COVID-19 case surveillance data, please see www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html.

Updated: May 5, 2022

Data Last Updated: May 4, 2022
Metadata Last Updated: May 5, 2022
Date Created: February 3, 2021
Data Provided by: CDC Data, Analytics and Visualization Task Force
Dataset Owner: Brian Lee

What's in this Dataset?

Rows: 71.4M
Columns: 19
Each row is a Deidentified Patient Case

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

In [1]:
```
import csv
import sqlite3
import numpy as np
import pandas as pd
```

In [2]:
```
# Create SQLite3 database.
# - define connection and cursor.
connection = sqlite3.connect('covid_large_dataset.db')
cursor = connection.cursor()
```

In [3]:
```
# Create table in database
command1 = """CREATE TABLE IF NOT EXISTS covid_data(case_month TEXT, res_state TEXT
cursor.execute(command1)
connection.commit()
```

In [5]:
```
with open ('COVID-19_Case_Surveillance_Public_Use_Data_with_Geography.csv', 'r') as
    reader = csv.reader(f)
    columns = next(reader)
    command1 = 'INSERT into covid_data({0}) VALUES ({1})'
    command1 = command1.format(','.join(columns), ','.join('?' * len(columns)))
    for data in reader:
        cursor.execute(command1, data)
    connection.commit()
```

In [6]:
```
f.close()
```

In [7]:
```
connection.close()
```

Data Analysis using SQLite3

In [8]:
```
# Open SQLite3 database.
# - define connection and cursor.
connection = sqlite3.connect('covid_large_dataset.db')
cursor = connection.cursor()
```

In [9]:
```
# Query data from sqlite3 database for dates, states, and count totals per state/da
command1 =  """
            SELECT DISTINCT
                case_month,
                res_state,
                count(res_state) AS state_total
            FROM
                covid_data
            WHERE
                case_month IS NOT NULL AND res_state IS NOT NULL
            GROUP BY
                case_month, res_state
            ORDER BY
                case_month, res_state;
            """
```

In [10]:
```
# Import filter data from SQL to DataFrame.
# Execute command and read into DataFrame.
df = pd.read_sql(sql=command1, con=connection)
```

In [11]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1485 entries, 0 to 1484
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   case_month   1485 non-null   object
 1   res_state    1485 non-null   object
 2   state_total  1485 non-null   int64 
dtypes: int64(1), object(2)
memory usage: 34.9+ KB
```

In [12]:
```
df.head()
```

Out[12]:

	case_month	res_state	state_total
0	2020-01	AL	103
1	2020-01	AR	17
2	2020-01	AZ	63
3	2020-01	CA	389
4	2020-01	CO	85

In [13]:
```
# Create a list of sorted dates and a list of sorted states
# for use as index and columns of new table.
dates = df['case_month'].value_counts()
states = df['res_state'].value_counts()

dates.sort_index(inplace=True)
states.sort_index(inplace=True)

dates_list = list(dates.index)
states_list = list(states.index)
```

In [14]:
```
# Create the new table.
df2 = pd.DataFrame(index=states_list, columns=dates_list)

# Assign values to each date, state pair in new table.
for index, row in df.iterrows():
    state = row["res_state"]
    date = row["case_month"]
    total = row["state_total"]
    df2[date][state] = total
```

In [15]:
```
df2.head(10)
```

Out[15]:

	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-1
AK	NaN	NaN	239	137	146	614	2293	2104	3068	936
AL	103	59	2677	6208	11427	21971	49743	37556	27988	3794
AR	17	18	1156	2742	5354	16458	21307	18628	21805	2896
AZ	63	57	2679	7243	16190	71744	86892	22845	16826	3341
CA	389	489	19293	46525	71232	174261	282735	151891	101378	12494
CO	85	98	6793	12128	9287	7091	14650	9801	14372	4331
CT	17	36	3766	9257	3629	29767	4007	3505	4564	1634
DC	NaN	36	1232	3305	2175	566	4361	712	633	79
DE	NaN	NaN	300	4242	4384	1929	3134	2397	2921	427
FL	132	290	16081	20817	21705	81505	135395	149549	111662	12942

10 rows × 29 columns

In [ ]:
```

```