ETL PROJECT II

COVID Cases in Prisons in the US

Team

Kudzanai Mukahiwa | Noshaad Ahmed | Diana Fernández

Description

We have chosen to make an ETL on Covid in US prison data. What happens in prisons is often looked passed as prisoners are looked down upon in society. These individuals are vulnerable and in fact are at a higher risk than the general public.

Our aim is to provide a good source of data to analyze and give a better understanding of how Covid impacted prisons. These have been achieved with official data from different sources from the United States by State.

EXTRACT

We started the ETL process by extracting data from:

1. **The Marshall Project & The Associated Press**

The Marshall Project, the nonprofit investigative newsroom dedicated to the U.S. criminal justice system, has partnered with The Associated Press to compile data on the prevalence of COVID-19 infection in prisons across that country.

Data collected by The Marshall Project and AP shows that hundreds of thousands of prisoners, workers, correctional officers and staff have caught the illness as prisons became the center of some of the country's largest outbreaks. And thousands of people, most of them incarcerated, have died.

Link to public source available for downloading:
https://data.world/associatedpress/marshall-project-covid-cases-in-prisons

2. The New York Times

From March 2020 until the end of March 2021, The New York Times has also collected data about coronavirus infections, deaths and testing for state and federal prisons; immigration detention centers; juvenile detention facilities; local, regional and reservation jails; and those in the custody of the U.S. Marshals Service.

Link to public repository: https://github.com/nytimes/covid-19-data/tree/master/prisons

3. Worldometers

Worldometer.info is an international website that publishes world statistics data that is sourced from reliable sources worldover.

For this project, we have scraped from their website US national Covid statistics by state. That will provide the possibility of comparison, for example the likelihood of a person in prison dying from Covid-19 compared to those that are not in prison.

Link to source: https://www.worldometers.info/coronavirus/country/us/

4. Social Security

Social Security is a United States government department that is responsible for providing financial security to US citizens. We have obtained from their website a table with state names and two letter state abbreviations of every state in the US. This allows us to join N°1, N°2 and N°3 datasets.

Link to source: https://www.ssa.gov/international/coc-docs/states.html

In the Resources folder there are all those files downloaded and available for analysis.

TRANSFORM

For the transform stage we have use Pandas and Datetime for dates formatting:

```
```
import pandas as pd
import datetime as dt
```
```

No columns were dropped on the dataframe that was extracted from the social security website. This dataframe had two columns: state and state abbreviations.

The most significant data cleaning was removing columns that were not needed on the table that was scrapped from the worldometers website. To do this the table extracted was transformed into a data frame (state_covid_data). Thereafter the columns ['#', 'NewCases', 'NewDeaths','Source', 'Projections', 'TotalRecovered', 'ActiveCases', 'Deaths/1M pop'] were dropped from the state_covid_data dataframe.

The CSV files from the NYT and Marshall project were also imported and transformed into a pandas dataframe using the pd.read_csv function.

To make the dataframes compatible with each other after loading the data in the dataframes into the SQL database, it was decided to ensure that each dataframe would have a state and abbreviations (of the state) columns on it that are compatible. To do this, the name of the state written on each dataframe was changed to ensure that it was all lower case using the pandas str function. For example ALABAMA or Alabama was changed into alabama. This allowed us to use “state” columns as primary / foreign keys in SQL tables.

Also some of the column names on all the dataframes were changed to ensure that they are written in a way that is compatible with the SQL language. Examples of these can be found in the final code for the project.

The data type for columns with dates or time were also transformed into a timestamp using the to_datetime function from the datetime module.

LOAD

The final function in the ETL process required us to Load the data we cleaned in the Transform phase.

First we created the tables in PostgreSQL. In LOAD.md document, there are the queries to CREATE TABLES in an SQL Database and the queries to generate the ER Diagram with all the tables.

We connected to our local database in PostgreSQL using SQLAlchemy:

```
```
from sqlalchemy import create_engine

engine =
create_engine(f"postgresql://postgres:{pg_key}@localhost/covid-in-us-pr
isons")
```
```

```
con=engine.connect()
```

The 8 tables we have Loaded are:

- state_abbreviations_df
- state_covid_datadf_left
- nyt_df1
- nyt_df2
- ap_data_df1_cases
- ap_data_df2_rates
- ap_data_df3_prisonpop
- ap_data_df4_staffpop