ETL Project

Covid-19 Cases/ Deaths Data vs Masks Usage in the U.S. Analysis

Members:

Kefan Li
Himani Manglik
Akibo Watson

ReadMe:

Covid19 has been the central topic around the world for quite some time since late February. The numbers are still increasing drastically in many places in the states. We found that New York State Governor Andrew Cuomos was the first official governor to urge citizens to use masks in the public area and provide a path of accessing free masks at different locations across New York state. With the effort of social distancing and the enforcement of masks usage, New York state has reached a flat increase in death cases just a day ago on Aug 4th, 2020 which sets a tremendous precedent on pandemic control when compared to New York state's early status during mid March.

Meanwhile, western and southern states have been experiencing a skyrocket in both confirmed cases and deaths numbers over a short span of time. As some of the states are still hesitating to put on a mask enforcement to the citizens in public, we would like to analyze a set of live data released by New York Times to provide some insights of Covid19 statistics vs masks usage.

References :

- https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html
- https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset?select=us-counties.csv
- https://github.com/nytimes/covid-19-data

Outline:

Data Research/Collecting -

With the concept of comparing Covid19 cases vs masks usage in mind, we have used kaggle, google and data.world to search the related topics csv, database and articles etcs. Out of all the available options, we luckily found a few consolidated csv files that display the data we needed which all generated from New York Times by above github link.

Data Selection -

We studied through all the provided csv files and selected the most relevant dataset to use which is Covid 19 cases by county under 'live' folder and masks usage under 'mask' folder. The Covid19 csv file is reflecting an accumulative data until Aug 1st, 2020.

Data Cleaning - Pandas/Jupyter

The 1st step of data cleaning we did was importing the csv files into Jupyter notebook and set each file as a corresponding dataframe. Under the Analysis.ipynb file, we conducted codes from dropping Not Available rows, converting float numbers to integers, setting new index columns and renaming column names for 'county' dataset. For 'mask' dataset, we renamed column names to lower cases in order to match 'county' dataset columns, converted float numbers to percentages and reseted the index column.

Lastly, we created the dataset connection to postgres engine with the existing two table names and loaded the dataframe to the database.

Creating Database Tables - PostgreSQ

We created the schema code in PostgreSQL. We first created a database called covidmask_db. We created two tables, one called 'county' and another called 'mask' to match the names given in Pandas/Jupyter. We made sure to set the primary keys to CountyFP (or the county FIPS code). Somewhat unusually, we made our survey words of "Never," "Rarely," "Sometimes," etc. into varchars because used string formatting on them during the data cleaning process. Finally, we ran the tables on PostgreSQL. After the data had been loaded from Pandas, we were able to see the transformed output data in PostgreSQL as well.

Querying the Data- PostgreSQL

Again, we used PostgreSQL to query the data. We ran 17 different queries to find any trends between COVID19 confirmed cases/deaths and the reports of mask usage. In the end, our main takeaways were that counties with higher density populations were more likely to have more COVID cases and deaths, despite having a high percentage of people stating they 'Always' use masks. Another main finding was that midwestern counties and states were more likely to have higher numbers of COVID-related cases and deaths.

Final Analysis:

We conducted our final analysis on PostgreSQL itself by querying the relevant data. We first joined the tables for '*county'* and '*mask.'* We were able to join the two tables by the FIPS county codes that were the primary keys. Some findings from our queries are:

1. Despite some counties reporting that they "Always" wear masks, they may still be at greater risk of confirmed deaths and cases of COVID19, perhaps because those counties have higher density populations. For example, Cook, Illinois (where Chicago is located) had confirmed deaths of 4,886 despite reporting that 72% of the population always wear their masks. Smaller counties across the country report they wear masks at a much lower percentage, but still their case and death count is lower. Further analysis should look at how population density related to mask reporting and the number of COVID cases.
2. Suffolk county, New York had the highest number of confirmed cases in NY, while Nassau county, New York had the highest number of confirmed deaths in NY.
3. California, Texas, and New York had the highest number of confirmed cases, respectively.
4. New Jersey, New York, and California had the highest number of confirmed deaths, respectively.
5. Inyo, California, Yates, New York, and El Paso, Texas had the greatest percentage of people saying they 'Always' wear their masks. Despite this these states still had a high number of confirmed cases and deaths, suggesting that there are other variables affecting the spread of COVID19 besides just mask usage.
6. Millard, Utah, Wright, Missouri, and Cass, Iowa had the highest percentage of people saying they 'Never' wear their masks.
7. Delaware, Rhode Island, and Connecticut, respectively, had the highest percentage of people saying that they on average 'Always' wear their masks.
8. North Dakota, Utah, and Montana, respectively, had the highest percentage of people saying that they on average 'Never' wear their masks. There seems to be a trend where Midwestern states and counties are reporting less usage of masks.