# **Project Summary on COVID-19 cases**

## **Data:**

The visualizations have been made based on the data that has been taken from four different sources as shown below. The FIPS code or county code is taken as the common key to compare different data elements.

## **Data analysis:**

The initial data analysis has been completed by deleting invalid information from each dataset. Duplicate values have been removed from the datasets and county and state names have been standardized. Hospitals where data for number of beds is unavailable or hospitals with closed statuses have been dropped from the dataset. Numerical values with floating zeros and leading zeros have been stripped off for comparison purposes. State names and state codes have been split into multiple columns. Values have been grouped to find the total number of cases by county. Recurring COVID counts have been dropped and latest data with complete statistics has been taken for data analysis.

## **Summary:**

Completed datasets have been merged using FIPS or County code and stored in a SQL lite database. A new table has been created to store the data. The following columns have been kept in the final dataset.

1. State code
2. Population Density
3. No of Beds
4. FIPS (county) code
5. No of COVID-19 cases
6. No of COVID-19 deaths
7. County name
8. State Name
9. Date of counts taken

## **Visualizations**

* Initial summary statistics indicate that the highest number of COVID deaths occurred in one county is 67551. The average number of deaths is found to be around 30 per county.
* A bubble chart has been developed to see the number of cases across different states. New York is the highest with 160000+ cases as of 05/23/2020. But, there are many counties with just one COVID case.
* When all of the COVID cases are plotted on a scatter chart, two outliers have been identified with case counts of 65000+ and 42000+. Many counties across US are in the range of 100 - 1000 cases
* A bar chart is plotted for the top 25 counties with the highest number of COVID cases. Although New York State tops with the highest number of cases, Cook County, Illinois tops in the county list. This is found to be the same outlier that is observed in the scatter plot. The number of deaths also the highest in Cook County with 3114 deaths. Compared to the number of cases, the number of deaths is very low across all counties.
* Population density is measured as the number of people per square mile. A line chart is plotted between the numbers of cases in each county with the population density. But, there is no correlation observed. However, one interesting observation is that Philadelphia, Suffolk (MA), and Hudson have a higher population density than Cook County but the number of COVID cases are far greater in Cook.
* A bubble chart is plotted with the size of the bubble as the number of beds available in county hospitals. Cook County has the highest positive cases, but only 16865 available beds in county hospitals. Harris County on the other hand has more than one bed per every COVID case. Also, Nassau, Suffolk, and West Chester counties in New York have very little number of beds compared to the total number of cases.
* The tree map indicates that there is no correlation between population density and the number of COVID cases.
* The histogram is left skewed showing that the majority of counties have less than 5000 cases.

## **Conclusion**

The number of cases are the highest in New York State followed by New Jersey but when you look at counties, Cook County has the highest number of positive COVID-19 cases. Many counties in New York have lower number of beds than the actual number of cases. The population density is not at all correlated to the number of cases but the number of deaths are correlated to the number of positive cases. Surprisingly some counties with the highest population density have the lowest number of positive cases.

**Sources:**

https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset

<https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html#IPE>

<https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals>

https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html#IPE