STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Informatics for Engineering Management

EM 624 Fall 2022 - Final Project

COVID-19 Data Analysis

Author
Shashank Khanna

Date: December 3, 2022

Table of Contents

Abstract

The data in the file “COVID-19_US-Counties.csv” provided by Professor Carlo Lipizzi on the Canvas platform presents the distribution of the COVID 19 disease in the United States. It gives us the number of cases and deaths in each state as Ill as the cases and deaths in each county of each state. This was the initial file which was used to work on the propogation of the COVID-19 virus in the US. I then move on to other datasets provided by Prof. Lipizzi to find the correlation betweenthem and create tables to help us visualize the data in great detail.

In addition to the 4 files provided by Prof, I also extracted data from the internet which is about the pollution data for each state in the year 2019 as Ill as the Housing Units present in the states for the year 2019.

I shall focus on the spread of COVID-19 in the United States and present my analysis on the same. My analysis is completely using pandas to generate Matplot graphs as Ill as Bokeh plots for data visulization. Using these graphs I shall present my interpretation.

In my interpretation I shall answer the following questions to the best of my ability :

- Which states have the highest mortality rate ?
- How is the population and housing units related ?
- What is the relation between mortality and air pollution of each state
- What are some of the major factors that play a part in the spread of the COVID-19 virus?

Introduction

The coronavirus illness of 2019, also known as COVID-19, has spread to practically every nation and territory in the globe, sickening millions of people and wreaking havoc on the international economy. There are a total of 646 million COVID-19 instances globally as of November 28, 2022. Additionally, 6.6 million people died from COVID-19.

The majority of confirmed cases and fatalities occur in the United States. State governments have come under scrutiny for implementing regulations that Ire not harsh enough and removing limitations too soon. The U.S. government's overall approach to the epidemic has also drawn criticism.

The nation's vaccination program has been successful, and the United States is one of the nations with the largest number of immunizations delivered globally. But even then due to high air pollution and dense populations, the virus has been spreading at a rapid rate.

To prevent a resurgence of new cases, experts still continue to caution against complacency and emphasize the significance of adhering to regulations and keeping watchful. This is especially crucial in light of the rise in instances brought on by novel COVID-19 mutations.

According to a study conducted in March 2020, just 12% of respondents in the United States believed that far greater than 10,000 people will pass away in the country during the following year due to COVID-19. The White House's coronavirus task committee predicted between100,000 and 200,000 Americans may perish at the end of that month; the actual death toll has been much higher. When examining the proportion of COVID-19 mortality in the U.S. by age, it is evident that the aged and those with pre-existing health problems are more susceptible to the disease. California, Texas, and New Jersey have now reported the highest number of cases in the United States.

Dataset Description

Prof. Lipizzi provided the following datasets :

- COVID-19_US-counties.csv

This dataset contains a list of cases and deaths in each county by date as Ill as the associated state. It contains the date on which this data was extracted for COVID-19 analysis. It also contains the county codes in the ‘fips’column of the table.

- CO-EST2019-ANNHU.xlsx

The estimates of housing units present in the counties, states is present in this dataset. The dataset consists of the following columns : Geographic Area, Housing Estimates from year 2010 to year 2019. The geographic area contains the county and state information together.

- GDP_Counties.xlsx

This excel file contains the estimates of GDP for each county as Ill as each state in the same column. The remaining columns consist of the GDP of each place from the year 2015 to year 2018, all in individual columns. Columns after that contain the percentage change with respect to the previous year from 2016 to 2018 which means that the percentage change 2016 column gives the data of change in GDP for that state/ county over the previous year which in this case is 2015.

- cc-est2019-alldata.csv

Consensus Data along with the racial distribution is present in this file. The data set consists of state code and county code data and then the state name and county name. The YEAR column has codes numbering from 0 to 12 which identify from year 2010 to year 2019. The AGEGRP column has codes similar codes but this time from 0 to 18 which are identified from 0 years to 85 or more years. Example : code 5 in AGEGRP = age 20 to 24. The next 3 columns identify the total population, the populations of female and population of males in that county/state. The remaining columns are just a racial distribution of the total population given to us in the previous columns.

- cc-est2019-alldata.pdf

This PDF doc provided to us from Canvas is just a dataset description of the c-est2019-alldata.csv file. It tells us in detail what each and every column stands for as Ill as all the codes in the AGEGRP and YEAR columns. Refer to this in-order to get an in-depth understanding to the covid-19_alldata excel file.

In addition to the above Canvas provided files, I downloaded the following datasets from the internet :

- 2019-Annual.csv

This pollution dataset was extracted online to bridge the gap between the states which have high cases or deaths due to COVID-19 and the air pollution in those states.

- NST-EST2019-ANNHU.xlsx

This dataset was downloaded from the net to find a relation betweenthe covid cases, deaths and the housing units. It is a simple dataset that contains the name of the states in one column and the estimated total housing units in the other column.

Dataset Preparation

For the dataset to be usable I need to perform some cleaning and preprocessing. I have done the pre-processing of each file as given below :

- COVID-19_US-counties.csv

In the COVID-19_US-counties.csv file I 1st need to read the file using the read_csv function and then convert the file into a dataframe using the Dataframe function. Then I removed the ‘fips’, ‘date’ and ‘county’ columns from the dataframe using the drop() function as they are not useful to me. This is because my analysis is focused on the states rather than counties. After dropping the tables I use the pivot_table function and the aggregate function np.sum to obtain the sum of cases and deaths as per states. I can then print the pivotted table; this table is going to be used for further analysis.

- CO-EST2019-ANNHU.xlsx

In the CO-EST2019-ANNHU.xlsx file I 1st need to read the file using the read_excel function, also remove the headers as I shall add my own custom column names. Then convert the file into a dataframe using the Dataframe function. After this I removed the headers by dropping the 1st 5 rows of the table using the drop function. Removing the 2nd and 3rd column as Ill in the same way as this data is not useful for analysis. Renaming the remaining columns using the rename function of pandas. Then dropping the last 6 rows of the table as they do not contain any useful data. Finally sorting the table as per ‘FY2019’ column and the dataframe is ready.

- GDP_Counties.xlsx

For the GDP_Counties.xlsx file I read the file and remove the headers to add custom column names. Then convert the file into a dataframe using the Dataframe function. And then removing the 1st 6 rows as they are the headers of the file. Renaming as per requirement and then dropping all columns after the “2018 GDP” column as they are not required for analysis. Finally setting the index of the table as “state” and then resetting the index using the reset_index function and the file is ready for futher analysis.

- cc-est2019-alldata.csv

After reading the cc-est2019-alldata.csv file and converting into a dataframe. I can remove all columns that are not required for processing like "SUMLEV", "STATE", "COUNTY", "YEAR", "AGEGRP" , "CTYNAME". Sorting the table in descending order by the “TOT_POP” column and then creating a pivot table to obtain the sum of population for each state. The pivot function disregards all other columns. Now the file is ready to be processed.

- cc-est2019-alldata.pdf

No preparation needed as this is just a dataset description file.

- 2019-Annual.csv

Reading the 2019-Annual.csv file and converting into a dataframe using pandas function. I then go ahead and drop any columns I do not need for analysis - the columns dropped are 'Edition', 'Report Type', 'Score', 'Lower CI', 'Upper CI', 'Source' and 'Source Year'. Then I select all the values from the table which contain “Measure Name” as “Air Pollution” and then slice the dataframe to make a table which contains the data that is useful to us. I can then rename the “State” column to “state” so that it can be used to merge the tables together and finally sort the table by “value” column.

- NST-EST2019-ANNHU.xlsx

1st read the NST-EST2019-ANNHU.xlsx file and drop all headers and convert into a dataframe using pandas. After this we can remove the 1st 8 rows from the housing units table as it is not useful. Then rename columns to custom headers and we can reset the index for the table. After that I sort the file using the “FY2019” column. The file is now ready for further processing.

Methodology

My main approach to the analysis of the COVID-19 Datasets is using Pandas. Once all the files are cleaned and pre-processed. I can go ahead and use them in my analysis.

To summarize; The main factors that I have chosen in this analysis are :

- Mortality Rate
- Average Number of people per household
- Air Pollution

I have chosen a step by step approach which is given below :

1. Read, pre-process and clean all the files as mentioned in the Data Preparation section of the report and print the outputs in Python as Tables.
2. Once I get all the output tables I can then go ahead plot graphs using these tables to make my initial analysis. You can find the graphs in the Result section of the report.
3. Along with the graphs I also calculated the mortality rate for each state using the formula:

   (Number of Deaths / Number of Cases) * 100 %

4. After the initial analysis I can start merging the files to obtain one master table which can act as my main analysis table. The final merged table (master table) is attached below.

Following table contains the Merged Final Table :

	state	cases	deaths	mortality	Female	Male	Total	Measure Name	Rank	Value	FY2019
0	California	77374476	1603639	2.072568	463547576	457833174	921380750	Air Pollution	50	12.8	159861376
1	Texas	67974708	1306932	1.922674	324336574	319688662	644025236	Air Pollution	40	8.3	180533648
2	Florida	65213390	1265202	1.940095	244000222	233404154	477404376	Air Pollution	25	7.4	154778912
3	New Jersey	33009985	2549308	7.722839	108793786	103546994	212340780	Air Pollution	47	9.2	38268992
4	Illinois	32987839	1255840	3.806979	156701378	151069528	307770906	Air Pollution	48	9.3	86209056
5	Georgia	27949474	696950	2.493607	123895158	117807046	241702204	Air Pollution	40	8.3	70054256
6	Arizona	22271168	545236	2.448170	81313728	80353816	161667544	Air Pollution	49	9.7	49215696
7	Massachusetts	20355693	1426483	7.007784	83152802	78181164	161333966	Air Pollution	40	8.3	26859712
8	Pennsylvania	20251638	1194116	5.896392	156602048	149701182	306303230	Air Pollution	36	8.1	71722048
9	North Carolina	19349573	346798	1.792277	122260012	116046416	238306428	Air Pollution	22	7.2	75967088
10	Louisiana	17975954	708691	3.942439	56589874	54166796	110756670	Air Pollution	33	7.9	33436432

5. It contains the cases, deaths, mortality rate, female population, male population, total population, the air pollution value along with the rank of air quality with 50

being the more polluted and 1 being the least, and the housing units estimate for Financial Year 2019 (FY2019).

6. From the master table I can find the state with the most cases, as the final merged table is sorted as per the COVID-19 cases in descending order.
7. I, then find out the relation between the deaths and total population of that particular state along with the relation between housing units and air pollution with respect to the mortality rate.
8. Also I can find a relation between the number of cases to pollution, population.
9. After I plot the graphs on Bokeh Plots using the merge table data, I can analysis the plots and give the relationships between the factors.
10. Along with this I also calculated the average number of people residing in a household which will help us in identifying if density of people in a house also plays a factor in the number of COVID-19 cases/deaths.
11. The data of average people living in a household was calculated using the formula:

(Total Population of State / Number of Housing Units in State)

Results

Finally we have reached the results section, which involves data visualization and representation along with our analysis of the COVID-19 datasets provided.

The first table that is generated using Pandas in Python (with the use of the COVID-19_US-counties.csv file) is of the number of cases and deaths caused by COVID-19 in each state. This data is then put into variables as lists and a multi-bar graph is plotted against it. As seen in Figure 1, that California is the state with the maximum cases of COVID-19 as of 2019 with Texas, Florida and New Jersey following.

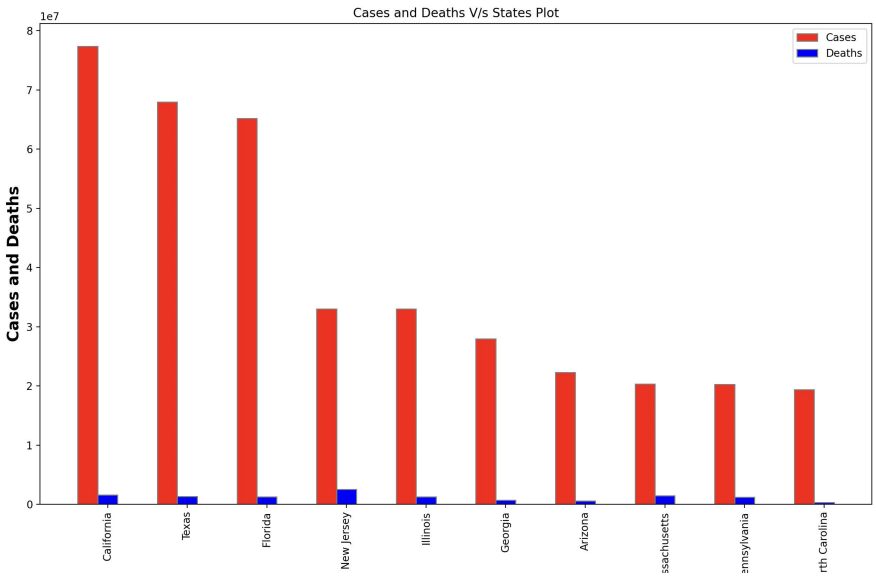

Figure 1: COVID-19 Cases and Deaths V/s States

Even though the number of cases are maximum in California, it can be seen that the number of deaths are very less with respect to other states. Especially when looking at New Jersey and Massachusetts. The number of COVID-19 deaths to cases ratio is quite high. I can safely assume that the Mortality Rate of New Jersey and Massachusetts will be far greater than California, Texas or Florida. We shall look into the mortality rate analysis a bit further.

The next table was generated from the CO-EST2019-ANNHU.xlsx dataset, which contains the housing unit estimations from year 2010 to 2019. But we shall be looking at only the 2019 data as all data from other datasets belongs from the year 2019. The Figure 2 shows the bar graph of the dataset and gives us the data of housing units per county.

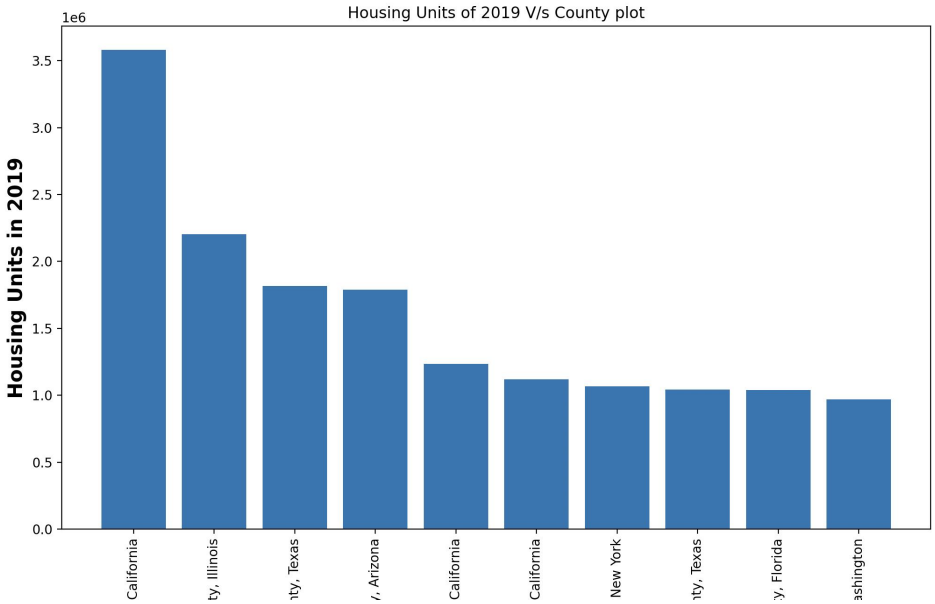

Figure 2: Housing Units in 2019 V/s Counties

I ended up downloading a new dataset which would give me the housing units per state. But in Figure 2 we can see that the Housing Units for Counties in California are most hence, we can assume that housing units will be maximum for the state of California. The same will be analyzed in later stages of the report.

Next we move on to the GDP_Counties.xlsx file which gives us data about the GDP from year 2015 to 2018 for each state. This data, once it has been processed into a table, we can go ahead and use the table to make a line graph which can be seen in Figure 3. The line graph shows the GDP for all four years for each state. This gives a good visual about how the GDP changes for each state. Also of-course California has the highest GDP as it is booming renewable energy industry, job creation, falling unemployment rates, and growing market values for companies as factors that are driving California's economic growth. The next being Texas which is a gold mine for any kind of manufacturing, especially automobiles.

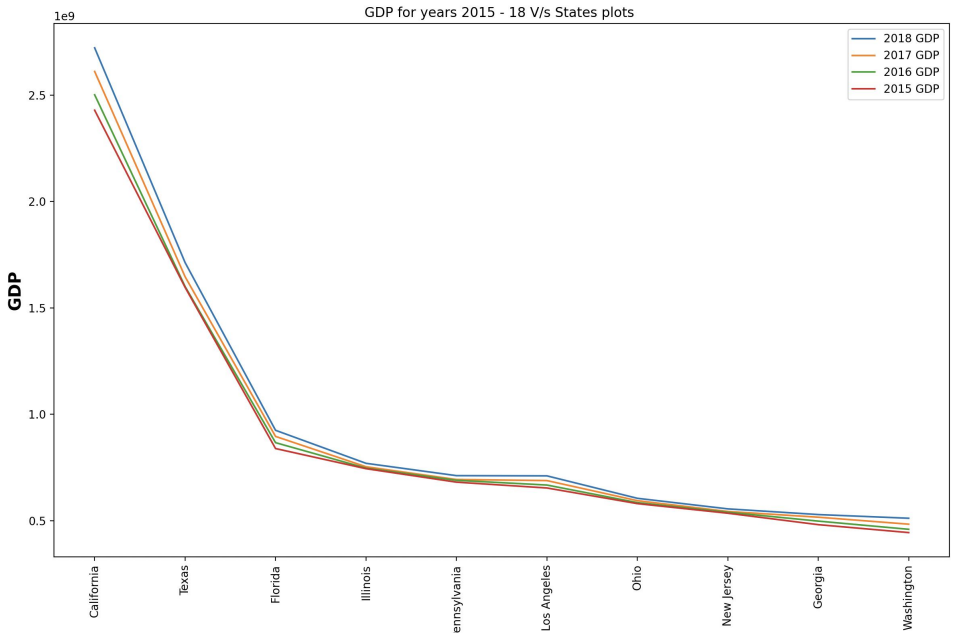

Figure 3: GDP from 2015 to 2018 V/s States

The GDP in the next few years would be greatly less than that of its preceding years due to the COVID-19 virus. As a great amount of people lost their jobs and hence the unemployment rate sky rocketed. But still the change GDP over all stayed the same hence even now California has the highest GDP of all the US states.

My analysis now moves on to the census dataset which is being taken from the file - cc-est2019-alldata.csv. Now most of the data in the file is not useful to me for the analysis that I wish to provide but it does give great details about the population of each state. Going into the racial as well as age wise distribution. The population data is quite important as COVID-19 was a virus that could be transmitted from one individual to another hence, it was seen that more populated places were impacted the most from the virus. Figure 4 shows the bar graph that is obtained using the Total Population data from the dataset for each state. We can see clearly that California is the state with the highest population, followed by Texas and Florida, while New Jersey and Massachusetts seem to be having much less population than them, so would it be safe to assume that the impact of COVID-19 was mostly in California, Texas and much less in New Jersey. Not at all. The population density also plays a huge part in the propagation of the virus.

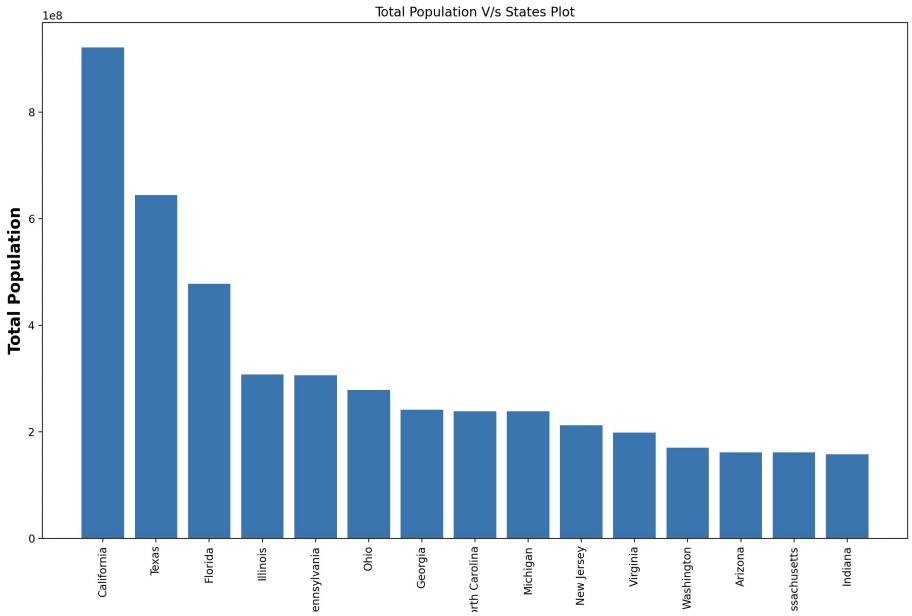

Figure 4: Total Population V/s States

As per the data extracted from the net regarding the Total Area for each state, which can be found on the website : https://www.census.gov/geographies/reference-files/2010/geo/state-area.html, we can clearly see that California is around 160,000 sq.miles while New Jersey and Massachusetts are both under 10,000 sq.miles of total area. Hence, we can conclude that New Jersey and Massachusetts both have a very high population density. Which can cause the COVID-19 virus to spread rapidly among the people. As a precaution it is advisable to stand 6 feet away from everyone to not be exposed to the virus, but with a population density this high, its impossible and hence we can say that this is a factor that is directly proportional to the high mortality rates. As even if people of California might have more cases. They have much more area to isolate and recover, while the same is not possible in Massachusetts or New Jersey.

Next we finally reach the mortality rate analysis. The mortality rate was calculated using the formula which I specified in the Methodology section of my report. We cam see in Figure 5 that As we had suspected, the mortality rate came the highest for New Jersey and Massachusetts.

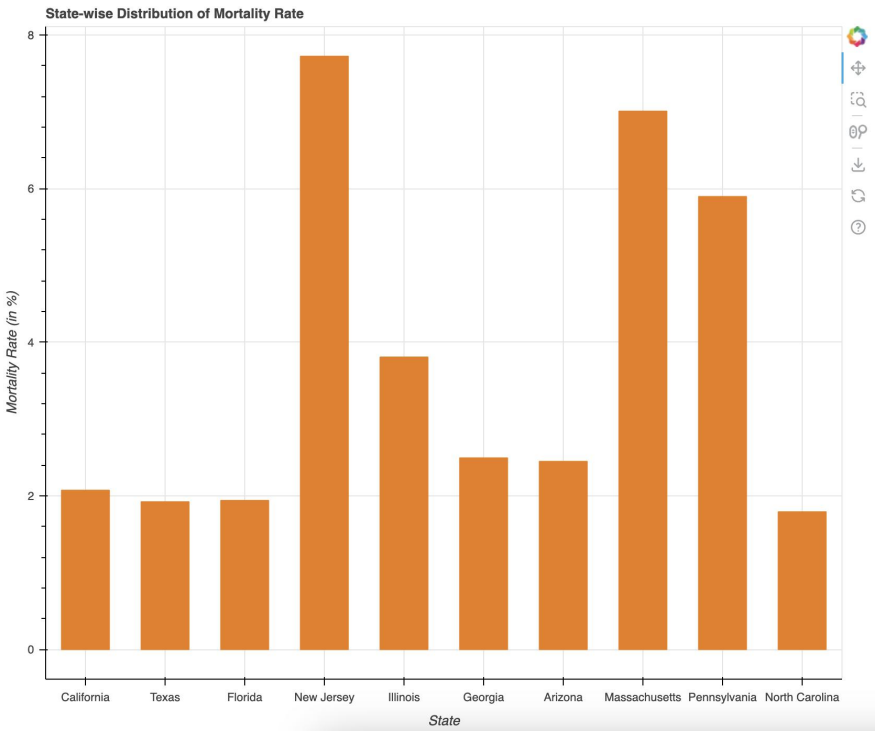

Figure 5: COVID-19 Mortality Rate V/s States

As discussed earlier that the mortality rate i.e. the death rate of the COVID-19 virus would be highest in New Jersey and Massachusetts due to the following reasons as discussed till now :

- High number of cases
- High population density
- Not as well developed as some other states like California in the medical services

These reasons have been some of the major factors leading to the high death rate due to the coronavirus with a few more incoming next.

We shall now continue our analysis of the Housing Unit estimations but this time with an all new dataset which we found on the internet. The dataset is NST-EST2019-ANNHU.xlsx file. This contains the housing units in 2019 for each state rather than county. Which helps us to determine its relationship with the total population data as well. I went ahead and calculated the Average number of people per household in each state using the formula as mentioned in the methodology section of the report. This data is further used to plot the horizontal bar graph as seen in Figure 6.

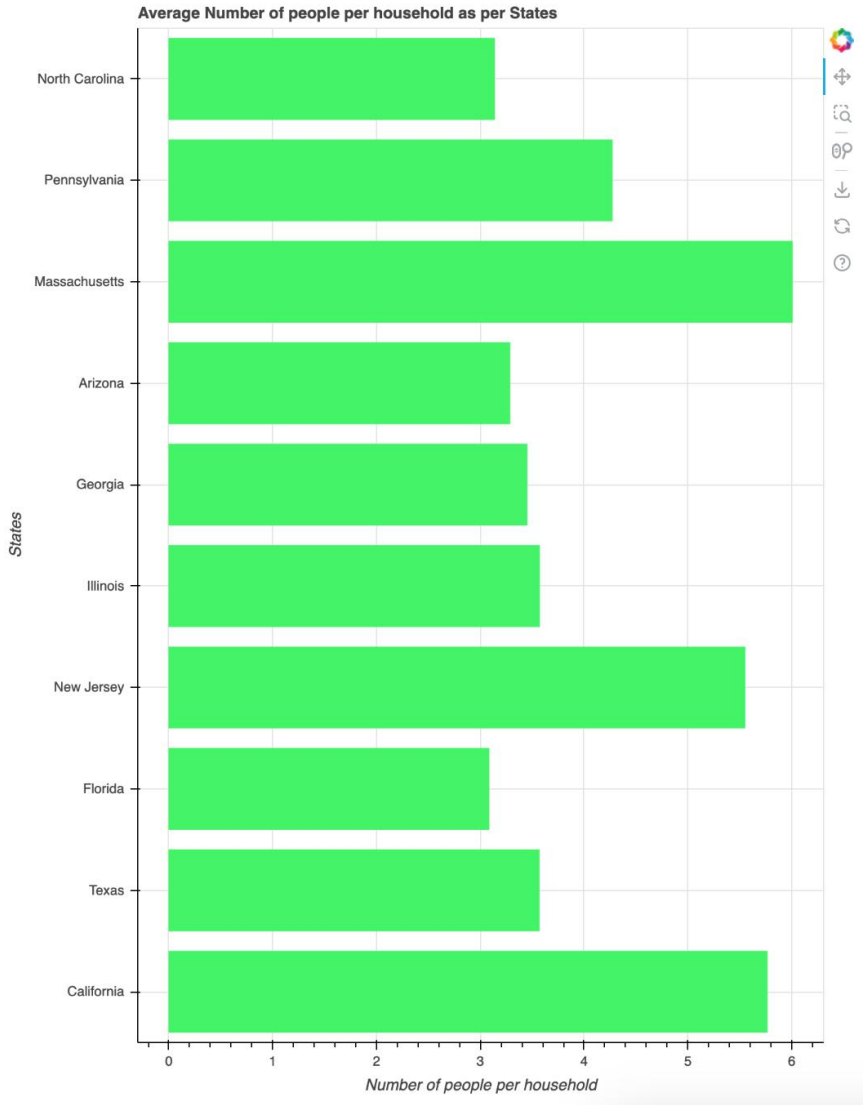

Figure 6: Number of people per household V/s States

As predicted, the number of people per household is in Massachusetts and New Jersey is quite high, hence the highest mortality rate also is in both of those states. But the most surprising discovery is that the number of people per household in California is the highest, which means that even though California has a much larger total area, it is overwhelmed by the much higher population. But the mortality rate is lower as said before due to the well connected medical services available. As well as its political affiliations, which make it one of the first states to reach the vaccine for the COVID-19 virus. While New Jersey and Texas do not have such great advancements in the medical sector, as well as having such a high population density per household is enough for the virus to spread and re-spread until it eventually leads in death before being treated.

The last factor that is included in my analysis comes from the 2019-Annual.csv dataset. This dataset includes the pollution data for each state, the rank of each state as per the air pollution with which increase with increasing pollution and the value of the pollution as per the air quality index. This value is what we used in our plotting of the pollution vs states graph as seen in Figure 7. We can see that the air pollution is maximum for California owing to its massive population and its overwhelming industry sector. The next states to have bad air quality is Massachusetts, Illinois and New Jersey. But as Illinois has low population density and a high total area, it has kept the covonavirus spread to a minimum.

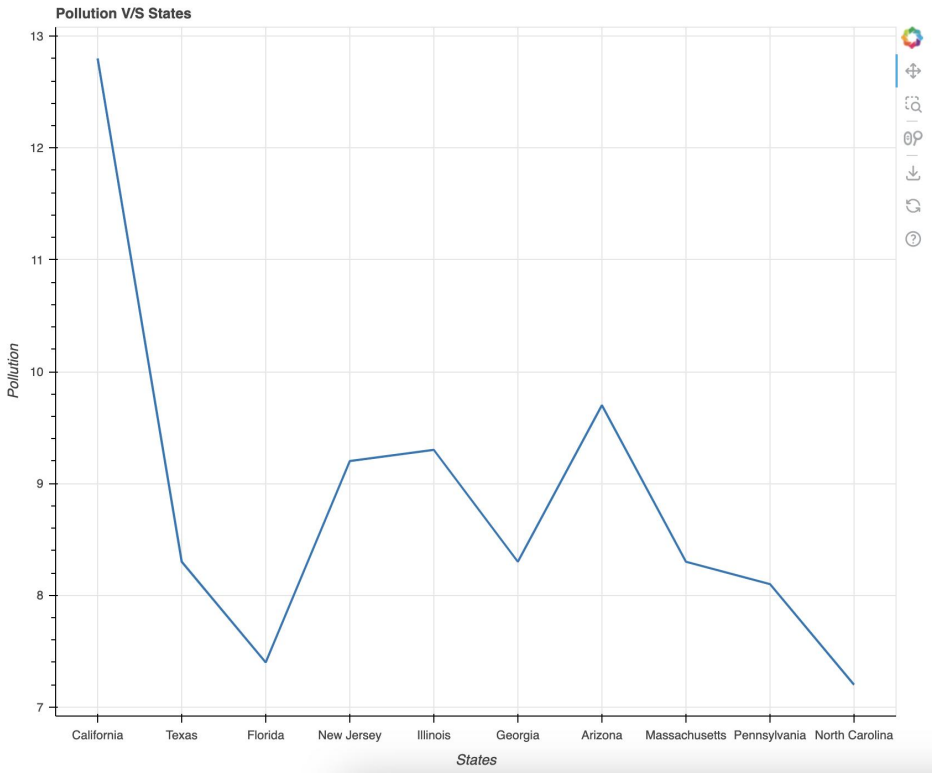

Figure 7: Air Pollution V/s States

Unfortunately New Jersey and Massachusetts having the worst air quality which has been a major factor in the spread of COVID-19 as it is an air borne virus has led to those states having a much higher mortality rate than other states.

Conclusion

In conclusion I would like to highlight all the major factors that have made the COVID-19 virus to propagate the highest in New Jersey and Massachusetts. The factors are :

- High number of cases
- High population density
- Not as well developed as some other states like California in the medical service
- Higher number of people per household
- Lower Total Area of the state
- Bad air quality or Higher air pollution

It is the combination of all these factors which lead to such a rapid spread of coronavirus in the state of New Jersey and Massachusetts. Other states like California and Illinois also do have many factors which could have made the virus spread faster, but it was due to their great medical infrastructure and large total area which has helped them recover and stop the virus from spreading further. Also my analysis has answered all the questions in the Abstract section of my report.

References

Datasets provided by Prof. Lipizzi - Stevens Canvas website :

https://sit.instructure.com/courses/59482/modules

Datasets provided by External websites :

https://www.census.gov/data/datasets/time-series/demo/popest/2010s-total-housing-units.html

https://www.americashealthrankings.org/explore/annual/measure/air/state/ALL?edition-year=2019

Other websites referred :

https://www.statista.com/topics/6084/coronavirus-covid-19-in-the-us/#topicHeader__wrapper

https://www.census.gov/geographies/reference-files/2010/geo/state-area.html