Paul Tuck

IST 652

11/30/2021

**Homework 2**

Introduction and data

The purpose of this project is to compare the number of Covid 19 cases between the three most populous states in the U. S. (California, Texas, and New York) along with my home state (Ohio). Specifically, I wanted to observe how the number of covid cases and hospitalizations have changed over time between the four states. To explore this question, I obtained json data from the Covid Tracking Project API (<https://covidtracking.com/data/api>). A sample of the json data can be found here: <https://api.covidtracking.com/v1/states/ca/daily.json>. The json data contains many fields in a single dictionary for each date; however, the ones I was interested in were the ‘positive’, ‘positiveIncrease’ and ‘hospitalizedCrurrently’ fields. The former is the total number of confirmed cases in that state up to that date, and the latter is the number of hospitalizations for a specific date. The dates for the data ranged from the 2nd of March, 2020 until the 7th of March, 2021.

Data preparation and cleanup

After setting up the Mongo db server and a database for the covid data, I retrieved the data from the API for the four states. The data from the four states was then inserted into the database, after which the data was retrieved with the find() function. In order to create the data frame that would be used for the graphs, I first created a list for each variable that would be used for the graphs (‘date’, ‘state’, ‘positive’, ‘positiveIncrease’, and ‘hospitalizedCurrently’). After the list of dates was converted to a datetime format, these lists were then combined into a data frame. However, the resulting data frame had the rows as the data fields and the columns as the observations, so it needed to be transposed; after which the columns were renamed to ‘Date’, ‘State’, ‘Positive\_cases’, ‘Positive\_increase’, and ‘Hospital’.

After making sure that all the columns were correctly formatted, I explored how many null values were in the data with the .info function. I found that there was only one null value in the ‘Positive\_cases’ field (March 3rd, 2020 in Texas) and 124 null values in the ‘Hospital’ field. I figured that these null values were the result of no cases or hospitalizations for those days, so I replaced the null values with 0. The resulting data frame contained 1,478 values. I then created a new column ‘Hospitals\_per\_positive’ by dividing the values in the ‘Hospital’ column with the ‘Positive\_increase’ column.

Analysis Questions

*Question 1: How do the total number of covid cases and hospitalizations between the states compare over time?*

To answer this question, I constructed a line graph with date on the x-axis and the number of cases and hospitalizations on the y-axis. I then used these graphs to see if there are any interesting patterns that could warrant further investigation. It is expected that California would have the highest number of cases and hospitalizations overall, since it’s the most populous state in the U. S. I would also expect for the number of hospitalizations to be highest before in the early part of the pandemic, decline when the lockdowns are issued, and increase when regulations are relaxed with the introduction of the vaccines followed by a decline with more people becoming vaccinated.

*Question 2: How do the number of hospitalizations per new cases compare between states?*

After I divided the hospitalizations by the number of new cases, I graphed the results in the same time frame as the previous graphs. Like the previous graphs before, I plotted the date on the x-axis and hospitalizations per increase in covid on the y-axis. I expected most of the results to be less than one, since not all new covid cases would result in hospitalizations (as we will see, this wasn’t the case). I also expected this value to follow a similar pattern to the ones above; with an initial increase at the beginning of the pandemic, a decline when guidelines were in place, and an increase when they were relaxed before declining due to the vaccines.

Python Program

This program is split up into gathering the data, creating the data frame, dealing with null values, and graphing the data. This program used the libraries json, urllib.request, pandas, matplotlib.pyplot, and MongoClient from pymongo. Pymongo, datetime, and csv were also imported, but were unused. This program also requires a connection to the Mongodb client.

After connecting to the Mongo client and creating a covid database, the urls for each state’s json API was stored in a series of variables. These variables where then stored in a list, which was looped through to open the url, read the json files, and store them in the covid database, creating a collection for that database. This collection was then stored using the .find() function.

To create the data frame, I created empty lists for each of the fields that I was interested in (date, state, total positive cases, increase in positive cases, and hospitalizations). I then created a for loop that went through my database collection to fill those lists with the data I needed. At this point, the database collection was no longer needed, so the database was dropped to prevent multiple instances of the same data to populate the database. After converting the date list to datetime, I added all the lists to a pandas data frame. As discussed above, this data frame needed to be transposed, and the columns were renamed to ‘Date’, ‘State’, ‘Positive\_cases’, ‘Positive\_increase’, and ‘Hospital’.

To count the null values, I used the .info() function, which summarizes the columns of a data frame by counting the number of non-null values. The only fields which contained null values were the hospitalization and one value for positive cases. I assumed that these null values were the result of no cases for those days and states, so they were replaced with 0. I then created a new column called ‘Hospital\_per\_positive’ that measure the number of hospitalizations per new covid case by dividing the ‘Hospital’ column by the ‘Positive\_increase’ column. This new column was filled with a few null values as well (probably because of several instances of dividing by 0), so these were filled with 0. The resulting data frame had 1,478 entries, which was exported to a csv file.

Once the data frame was prepared, I created new data frames for each of the four states using the .loc function to find the subset of that data corresponding to that state. I then used plt.plot to graph the date on the x-axis and number of positive cases on the y-axis for each state. I made sure to label the axes, give the plot a title and legend. I then repeated these steps, changing the y-axis to hospitalizations and hospitalizations per new case.

Output and Result

*Question 1*

When comparing the total number of covid cases, California and Texas followed similar patterns: an initial increase around the 7th of July, followed by a plateau until towards December 2020-January 2021, with a sharp increase around this time before slowing down around February. Ohio followed a similar pattern, but not as pronounced. New York differed from the rest by having a much more dramatic rise earlier than the others (around April-May of 2020), followed by a much longer plateau until another increase at the end of 2020 (fig. 1).

Chart, line chart

Description automatically generated

Fig 1: Cumulative covid cases from the 3rd of March 2020 until the 7th of March 2021. The number of cases on the y axis are in the millions.

California and Texas also followed similar patterns with hospitalizations. There was an initial increase in April of 2020, followed by a spike around late July, early August of 2020. There was then a decline until the end of 2020, where both states started to rise again, peaking in late January-early February of 2021 before declining again. New York’s peak in hospitalizations occurred much earlier than any of the other states, around April of 2020, before declining and staying low throughout most of the year. This lasted until late 2020, early 2021 when New York followed a similar pattern to Texas and California. Ohio’s pattern was unique in that its summer peak was relatively small, and the final peak and decline happened sooner than the other states (fig. 2).

Chart, histogram

Description automatically generated

Fig. 2: Numbers of hospitalizations per day throughout the time of the datasets. Though all states had a final peak during late 2020-early 2021, they differ somewhat in their timing. New York was the only state with a large initial peak and no summer peak.

*Question 2*

The number of hospitalizations per new covid cases was more difficult to interpret, due to the jagged nature of the graph. However, it appears that all of the states follow a general pattern of the greatest number of hospitalizations occur during the first few months of the pandemic (April to May of 2020), followed by a decline. Only California seemed to have increased in hospitalizations per new covid cases from January to March of 2021. Also, most of the results appear to be above 1, indicating that there were more hospitalizations than new cases, as discussed below (fig. 3).

Chart, histogram

Description automatically generated

Fig. 3: Number of hospitalizations per new cases of covid. All the states seem to have the largest increase towards the beginning of the pandemic, followed by a decline. Only California seems to be increasing since January 2021.

Discussion and Caveats

My initial hypothesis regarding covid cases and hospitalizations was partially correct. There was an initial increase around the time the vaccine was beginning to be introduced, followed by a decline. However, the most substantial increase in covid cases and hospitalizations seemed to have occurred around the summer of 2020, rather than the beginning of 2020. Only New York followed the pattern I expected. One hypothesis to explain this is that New York was one of the first state to have a covid outbreak, after which it spread to the other states. Once the vaccine was introduced and restrictions were lessened, this would have led to an initial spread of covid until another decline as more people got vaccinated.

For my second question, it appears that all the states had the highest number of hospitalizations per new covid cases towards the beginning of the pandemic, without a second increase. There are a number of possible explanations for this pattern, such as the virus evolving to become less deadly, or treatments improving over time. The increase in California is harder to interpret and requires more data.

My initial assumption was that the number of hospitalizations would always be lower than the number of new cases, since new covid hospital visits would be the result of new cases. This doesn’t appear to be the case. It might be that there is a delay between someone getting covid and them requiring medical treatment. If several people test positive, it might be a few days before the symptoms become serious, resulting in new hospitalizations on a day when there are relatively few new cases.

One bit of caution when interpreting this data is that it does not take the population of the state into account. California has the largest population of all the states, so it will naturally have more covid cases than any of the other states. As mentioned on the Covid Tracking website, various states also have different means of measuring new cases and hospitalizations, which have undoubtedly affected these results. There are also other caveats that the metadata mentions that should be accounted for before coming to any large conclusion. Despite these complications, these results will hopefully offer insights and new avenues for future research.