Sai Jayanth (RQ #1)
Erick Li (RQ #2)
Kerstin Wolf (RQ #3)
Amrutha Kumaran (RQ #4)

Studying Attributes of the COVID-19 Pandemic in the United States

1. Introduction

The outbreak of Coronavirus disease (COVID-19) in the year 2020 had a major impact on people's health and lifestyle. The United States is one of the countries that has been hugely affected by the virus. There have been approximately 30,257,078 COVID cases recorded to date with a death toll of approximately 549,299. During that time, countries closed their borders, established lockdowns and curfews, encouraged social distancing and masking, and enforced the quarantining of those infected. Deaths soared as case rates rose and United States workers faced a situation that hasn't been seen in the U.S. in over a hundred years. We came up with four research questions that could help analyze various factors happening around COVID-19:

1. After vaccination in the United States began, how did the infection trend move? In which states is the vaccination program more advanced?
2. How did Illinois bar visits statistically relate to Illinois COVID-19 daily case numbers?
3. How did the pandemic impact the employment, hours, and working environment of U.S.—specifically Illinois—workers in 2020?
4. Are rates of "COVID-like illness" in the community collected from respondents to the Carnegie-Mellon Facebook Survey predictive of COVID-19 cases by county in Illinois recorded by John Hopkins University?

2. Methods

2.1. COVID-19 Vaccination Progress

This study has considered the COVID-19 cases and vaccinations particularly in the context of the United States country for data collection. We have used Tableau Desktop for the analysis of data as a software tool.

2.1.1. Data Source

For the proposed research, two datasets are collected from CDC and Our World in Data. The CDC dataset includes the information related to COVID-19 cases and death by state and date. The Our World in Data includes the information related to COVID-19 vaccination by state and date. Both of the datasets have an average sample size of approximately 7000 rows.

2.1.2. Data Manipulation

As part of the data analysis process, data cleaning was performed using Microsoft Excel. Data cleaning was performed to exclude all null values and to synchronize the state names for data blending purposes in the Tableau Desktop tool. Both the datasets were blended in the Tableau Desktop tool by state using the inner join method.

2.1.3. Exploratory Dashboard Analysis

Two different visualization plots are used to find the top-performing states in the vaccination program. Top-performing states are analyzed based on daily vaccination, people vaccinated, and people fully vaccinated. This depicts the total population being vaccinated every month. Results might vary based on the population to total vaccine distribution ratio, therefore, the dataset is also analyzed based on people fully vaccinated per hundred people. The dataset is filtered for analysis purposes by state. This is done using the Geo Map feature in the dashboard.

Three different visualization plots are used to perform data analysis on infection trends by state by comparing between 2020 and 2021 year. The COVID-19 vaccination program began in January 2021, therefore, the dataset trend is compared between the years 2020 and 2021. The dataset is filtered for analysis purposes by state. This is also done using the Geo Map feature in the dashboard.

2.2. Illinois COVID Daily Case Number v. Bar Visits

This study focuses on the relationship between the Illinois COVID-19 daily new cases and the bar visits in state. Through building the model, the goal is to discover whether they have connections with each other.

2.2.1. Data source

In this topic, the sources of the Illinois COVID-19 daily new cases and the Illinois bar visits data are from a Kaggle dataset called “COVID-19 in USA” and the Carnegie Mellon Delphi Project, respectively.

Kaggle Dataset - “COVID-19 in USA”

The data are stored in a CSV file downloaded directly from Kaggle. Based on the description of the data page, all the numbers are collected from the COVID-19 Tracking Project managed by the New York Times. The dataset contains the cumulative case number for each state by date. It requires us to convert the reporting method from total cases to daily new cases.

Carnegie Mellon Delphi

The bar visit data are collected by the location data from the project participants’ mobile phones. It used their location to determine whether the owner of the phone visited a bar. The dataset reports the daily number of bar visits per 100,000 population in the state. I used a Python API provided by Delphi to download the data and store them into a CSV file.

2.2.2. Data preprocessing

Although the datasets are not large, I used a big data processing tool called PySpark, which introduces parallel computation to boost up the speed. It makes the preprocessing pipeline flexible even facing different sizes of data.

The purpose of the preprocessing pipeline is to resolve any data corruption (NA, missing data), datetime format, and removing unnecessary data fields. More importantly, we want to merge the COVID and bar visits data appropriately. The Figure 1 shows the pipeline for preprocessing the data.

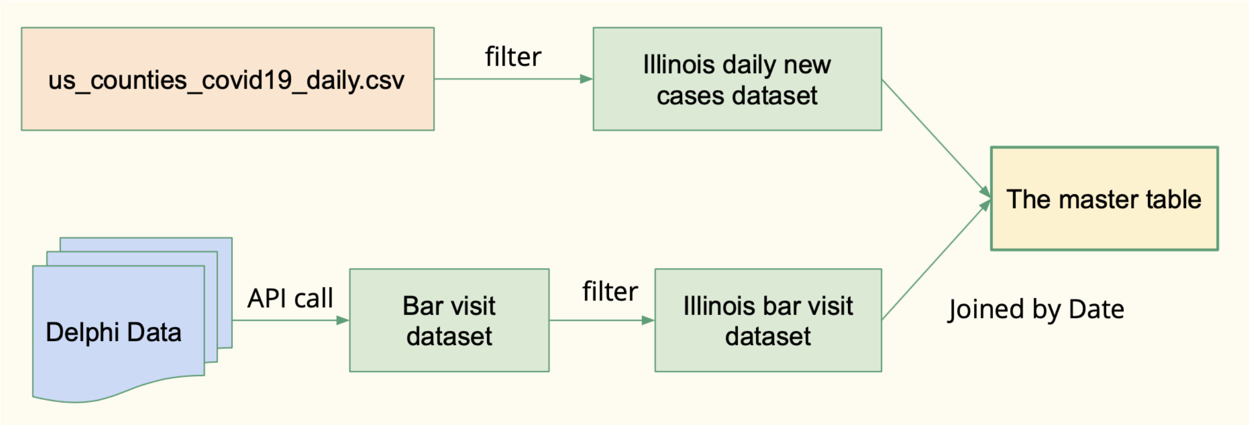

us_counties_covid19_daily.csv
filter
Illinois daily new cases dataset
The master table
Delphi Data
API call
Bar visit dataset
filter
Illinois bar visit dataset
Joined by Date

Figure 1: Data preprocessing pipeline

We applied filters on both datasets and only kept the columns containing date and the values reflecting the daily new cases and bar visit numbers. It follows by merging (or joining) the datasets into a master table by date. The master table will be used for the analysis.

2.2.3. Data Manipulation

We have done some extra steps, including calculating moving averages and normalization. The calculation of the moving averages can smooth the curve and reduce the influence of fluctuations and anomalies. Normalizing the data by max-min makes the datasets with different scales comparable since both data now share the same scale.

2.2.4. Modeling methods

Time lagged cross correlation (TLCC) is used for finding out the relationship between the COVID data and bar visit data. Hypothetically, there might be a lag between those two data. The increase of bar visits might not instantly contribute to a surge of the COVID cases, but it could have a lag of several days or even weeks. TLCC will calculate the correlations with different lags and find out the lag that best helps match those datasets.

Pearson correlation and Spearman correlation are used as the measures for deciding the best lag. Pearson correlation uses a linear approach to judge how close two series of data points are, while Spearman correlation uses a ranking approach. Two series of data can achieve a high Spearman correlation while low Pearson correlation when each set of data has the same ranking but their relationship is not linear, as shown in Figure 2. The ranking of the data points on X is the same as Y, while their relationship is not linear.

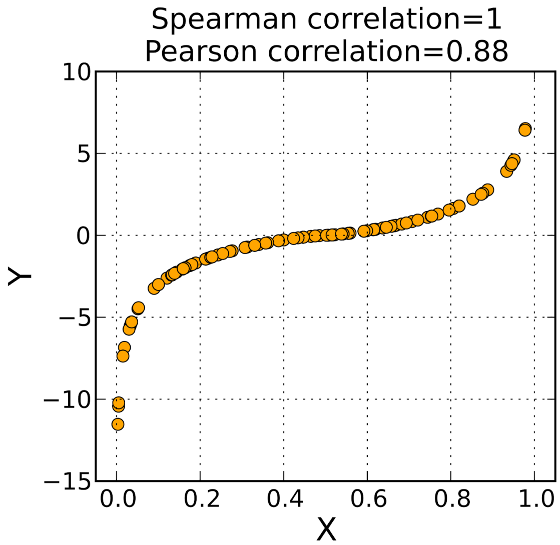


Figure 2: High Spearman correlation while low Pearson correlation

Considering both Spearman and Pearson correlation can help us discover more about the correlation between the Illinois COVID daily new cases and bar visits.

2.3. COVID-19 and Workers

This study focuses on the impact the COVID-19 pandemic has had on workers throughout 2020 in comparison to 2019. Using data visualizations to display the data, the impact can clearly be fully analyzed.

2.3.1. Data Sources

There were a number of data sources utilized for this study on COVID-19 and workers. All the datasets used on COVID-19 were from the Centers for Disease Control and Prevention (CDC). From the CDC, three datasets were pulled: “United States COVID-19 Cases and Deaths by State Over Time,” “Provisional COVID-19 Death Counts by Week Ending Date and State,” and a small dataset on Illinois pneumonia deaths in 2019. For datasets focusing on workers, the majority were pulled from the U.S. Bureau of Labor Statistics (BLS). These datasets included the Illinois unemployment rate from 2019 through 2020, the state unemployment rate from 2019 through 2020, and the weekly hours of U.S. employees in the private sector from 2019 through 2020. For tracking the general movements of people, three data signals from the SafeGraph data source through the Carnegie Mellon Delphi Epidata API, also called COVIDcast, were utilized.

These three datasets were the 7-day average of the fraction of people who didn't leave home, the fraction of people working full time away from home, and the fraction of people working part time away from home from 2019 through 2020.

2.3.2. Data Preprocessing

The couple of steps after selecting the appropriate datasets was to evaluate the data and clean and transform it into data that's usable for data visualizations. Evaluating the datasets led to some key observations that affected how the data was handled. The CDC datasets "United States COVID-19 Cases and Deaths by State Over Time" and "Provisional COVID-19 Death Counts by Week Ending Date and State" had the most surprises. For the dataset "United States COVID-19 Cases and Deaths by State Over Time," there were more than just the 50 states included. There were ten extra territories and metropolitan areas added into this dataset. As for "Provisional COVID-19 Death Counts by Week Ending Date and State," the documentation for the dataset and even the title stated that the data was weekly death counts by state. This was not the case. This dataset included national U.S. data besides just the states and also didn't just have counts by week increments. There were also month increments and year increments. Besides these two datasets, the rest didn't contain any big surprises.

2.3.3. Data Analysis

For this study, the primary focus was to use data visualizations to help analyze the data and uncover if there are any patterns or relationships that can be seen between COVID-19 cases and deaths and unemployment, hours, and working location of workers. Everything in this study was done through Python. All data cleaning, transformation, and exploration was done with the Pandas library in Python. The data visualizations in this study were also created in Python using Matplotlib.

2.4. Relationship between reported COVID-like-illness in community and COVID cases

Carnegie Mellon University's Delphi Epidata team originally collaborated with Google on a brief survey before continuing a longer partnership with Facebook to deploy a survey to a random sample of users each day (which has been in effect since March of 2020). The survey consists of questions about behavioral tendencies, mental health, social distancing, and the prevalence of ill people known to the respondent. The Delphi Group is specifically interested in conducting a COVID symptom survey for the purpose of potentially predicting surges in COVID cases before they occur.

There are five questions about the respondents' knowledge of ill people in their household or community. The Delphi Group used the CDC's guidelines for potential symptoms of COVID to create prerequisites for COVID-like-illness, or CLI. They defined CLI as the occurrence of fever and at least one of the following: sore throat, cough, shortness of breath, and difficulty breathing. The information concerning ill people in the community is covered by the first and fifth questions in the survey:

1. In the past 24 hours, have you or anyone in your household had any of the following (yes/no for each):

a. Fever (100 ºF or higher)

b. Sore throat

c. Cough

d. Shortness of breath

e. Difficulty breathing

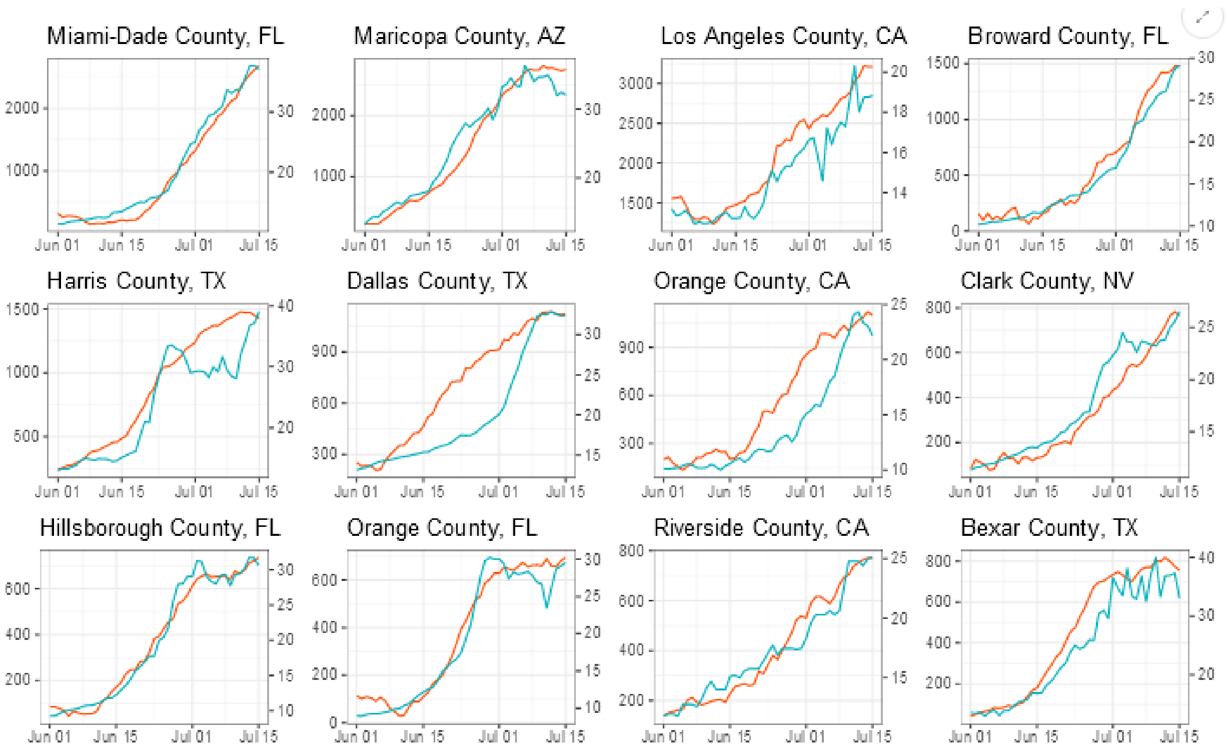
5. How many additional people in your local community that you know personally are sick (fever, along with at least one other symptom from the above list)?

2.4.1. Survey Bias

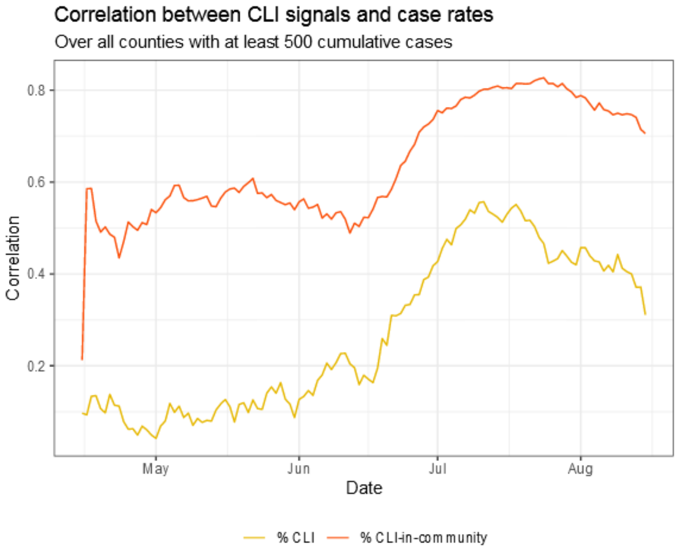
The Delphi Group identified some concerns about the legitimacy of their survey responses. Social media platforms are less frequented by underprivileged, rural, or low-income people. The sample of respondents is biased since it is a self-selected sample. Certain attributes of people who choose to respond to the survey (versus those who do not) could bias the responses. Responses need not be accurate or true. Additionally, the number of respondents per county varies from day to day.

2.4.2. Community Illness Anticipating COVID Cases

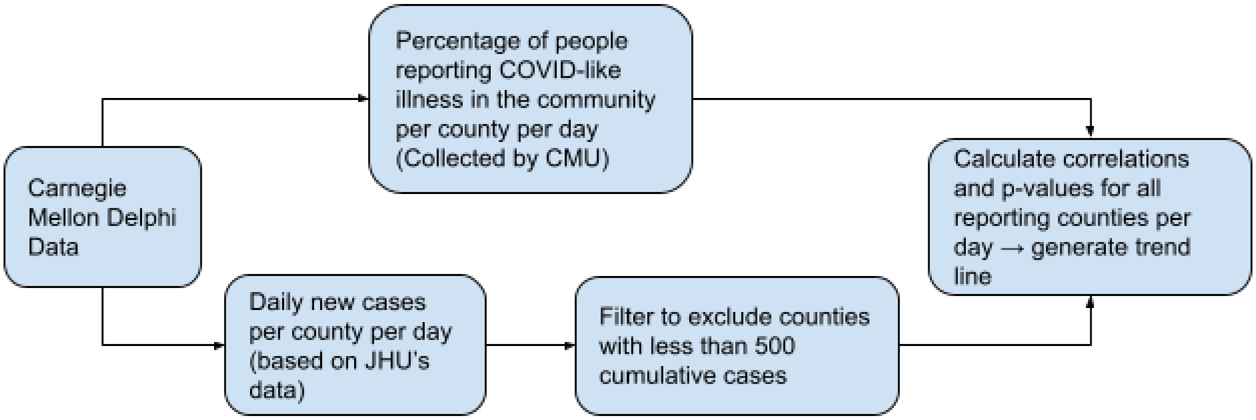
When they compared the trend lines of reported CLI-in-community and daily new cases in individual counties, they noticed an interesting phenomenon. In counties with large COVID surges during the summer of 2020, they saw that increases in reported CLI-in-community did in fact "anticipate" sharper increases in daily new cases. Anticipation was roughly defined as the presence of a steep slope in the CLI-in-community trend line occurring days before the slope of daily new cases also steepened. This feature is most visible around June 15 in Miami-Dade, Maricopa, Harris, and Dallas counties, shown below (Reinhart and Tibishari, 2020).



To better understand the qualitative appearance of anticipation, the Delphi Group also plotted the Spearman rank correlation between CLI-in-community and daily new cases for all U.S. counties with greater than 500 cumulative COVID cases over the spring and summer of 2020. They discovered that the correlation jumped to 0.8 after mid-June (Reinhart and Tibishari, 2020).



What follows is a parallel analysis of the same variables for the state of Illinois. The following investigation will attempt to determine the periods of case surges and what relationship (if any) CLI-in-community has to daily new COVID cases in those geographic areas.


Percentage of people reporting COVID-like illness in the community per county per day (Collected by CMU)
Carnegie Mellon Delphi Data
Calculate correlations and p-values for all reporting counties per day → generate trend line
Daily new cases per county per day (based on JHU's data)
Filter to exclude counties with less than 500 cumulative cases

2.4.3. Spearman’s and Pearson’s Correlations

Like the Delphi Group, we take reported CLI-in-community to be a product of daily new cases and rightfully assume that daily new cases are not a product of CLI-in-community. The Delphi Group calculated the Spearman’s rank correlation for CLI-in-community and daily new cases. Spearman’s correlation is typically used for ordinal, ratio, or interval data and not continuous data. They implied that they chose the Spearman correlation because it evaluates the monotonic relationship between two variables, regardless of whether their relationship is linear or not. The researchers may have chosen the Spearman correlation because they did not want to assume that an increase in daily new cases would correspond to a proportional increase in CLI-in-community.

Spearman’s correlation: $$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

This is counterintuitive to what we may naturally believe about the effect of daily new cases on CLI-in-community. Spearman’s correlation evaluates the strength of the monotonic relationship between the two variables and is used under the assumption that the two variables increase and decrease together, whereas Pearson’s correlation evaluates the strength of the linear relationship between two variables.

Pearson’s correlation: $$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Spearman’s correlation may be useful for merely discovering whether CLI-in-community can roughly signal oncoming surges in daily new cases, regardless of the mathematical relationship between the two variables. To more fully characterize the relationship between

CLI-in-community and daily new cases in Illinois and the North Central Region, both Spearman's and Pearson's correlations and their respective p-values were plotted on a per day basis from April 15, 2020 to March 15, 2021.

For Spearman's correlation:
H_o: There exists no monotonic association between CLI-in-community and daily new cases, or r =0.
H_a : There is a monotonic association between CLI-in-community and daily new cases, or r not equal to 0.

For Pearson's correlation:
H_o: There is no linear relationship between reported CLI-in-community and daily new cases, or r = 0.
H_a: There is a linear relationship between CLI-in-community and daily new cases, or r not equal to 0.
We set $\alpha = 0.05$ in order to ascertain whether there is less than a 5% chance that the relationship determined by the correlation coefficient occurred by chance given that the null hypothesis is true.

3. Results

3.1. COVID-19 Vaccination Progress

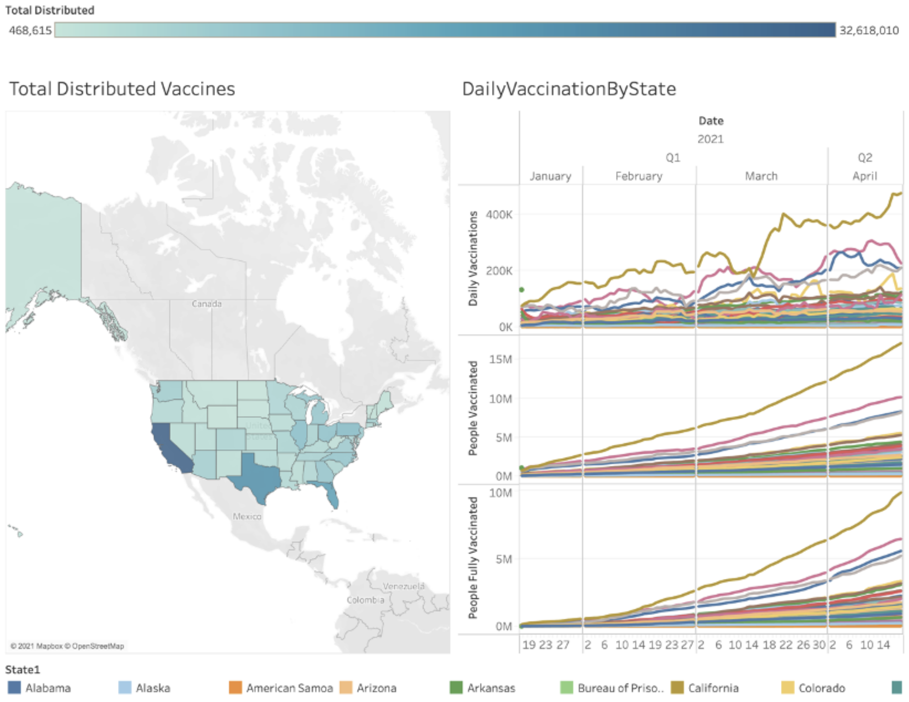


Figure 3: Top COVID-19 Vaccination Performing States Dashboard

Figure 3 shows the top performing states in the vaccination program. It lists the total distributed vaccines by states, people fully vaccinated per hundred in relation to the population of the state people, and vaccinations by states in relation to daily vaccination, people fully vaccinated, and people vaccinated.

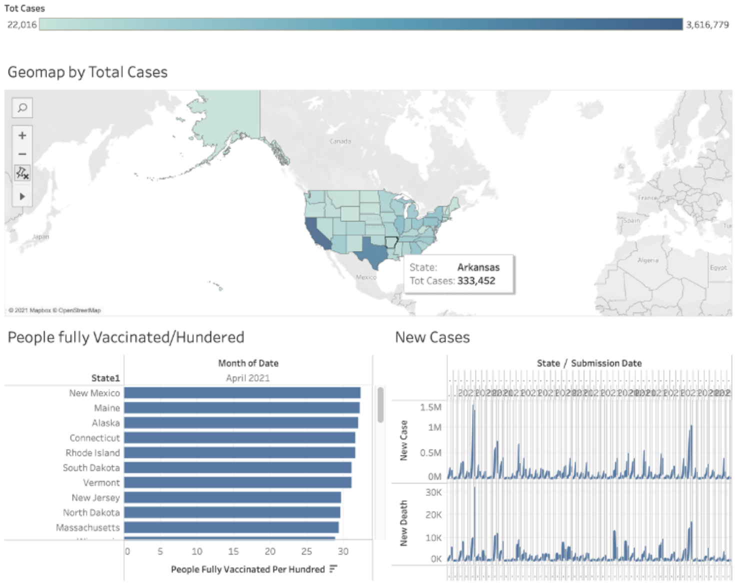


Figure 4: COVID-19 Infection Trend Dashboard

Figure 4 shows the infection trend in relationship to COVID-19 cases and deaths by year in each state. The total cases show the cumulative number of cases recorded from January 2020 to April 2021. The new cases chart shows the trend of COVID count between 2020 and 2021. The dashboard is designed in a way to show the data clearly when a state is selected on the Geomap. The Geomap acts as a filter of the dashboard.

3.2. Illinois COVID Daily Case Number v. Bar Visits

Figure 5 shows the normalized bar visits. It illustrates there was an increase in the number of people visiting bars from May 2021 to October 2021 after a jump in March due to the lockdown.

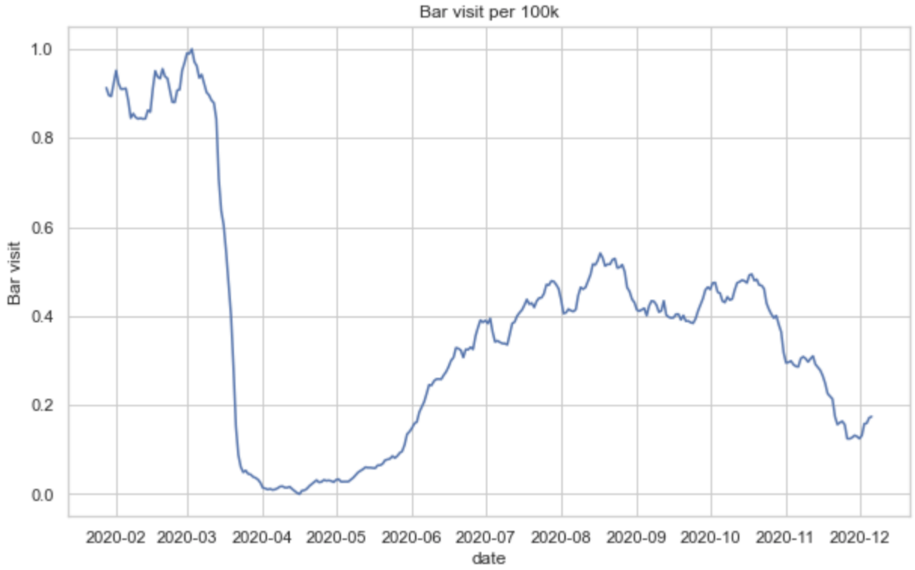


Figure 5. Normalized bar visit with 7-day moving average

Thus we will consider the time frame from May to October in 2021 and see how those two series of data are correlated with each other. Figure 6 shows the comparison between the Illinois COVID daily new cases and the bar visit data applied 7-day moving average and normalization.

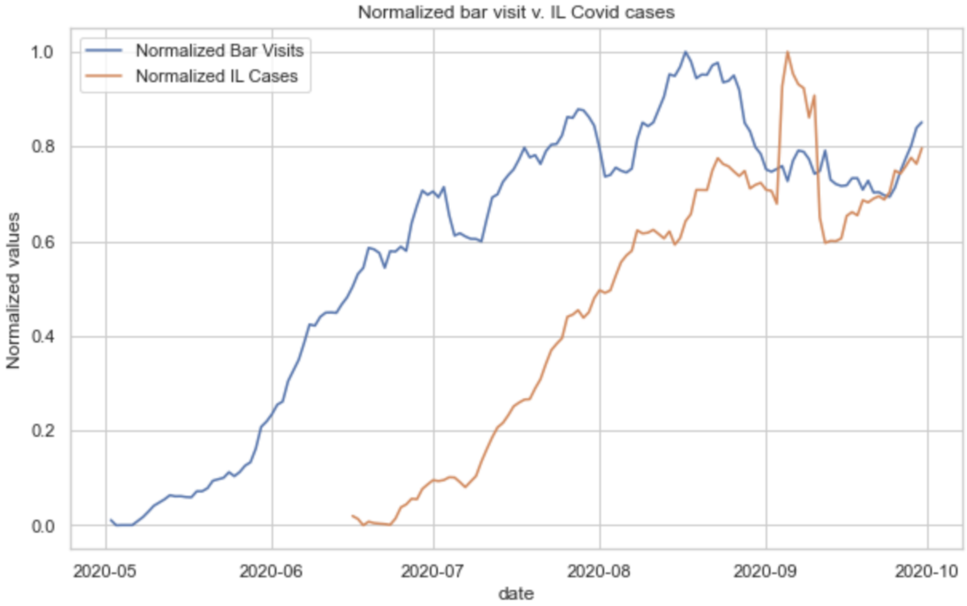


Figure 6: Normalized bar visits and Illinois COVID cases in Summer 2020

Applied to the TLCC, we get the correlation graph in Figure 7.

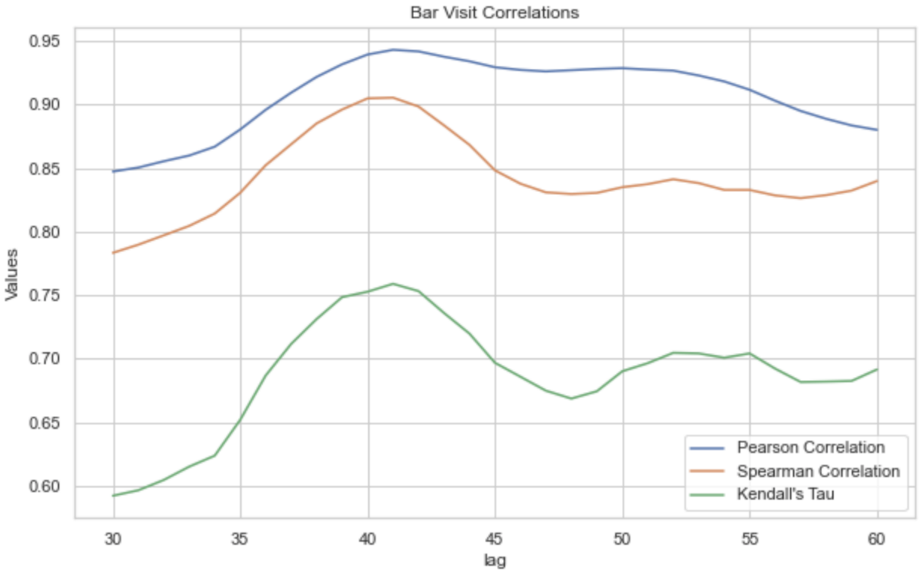


Figure 7: Correlations under different lags

From the figure, we can see that with a lag of 41 days of the bar visit data, the Pearson correlation is the largest, while the Spearman correlation is also high. In Figures 8 and 9, it shows the relationship between the Illinois COVID daily new cases and the bar visit data with a 41-day lag.

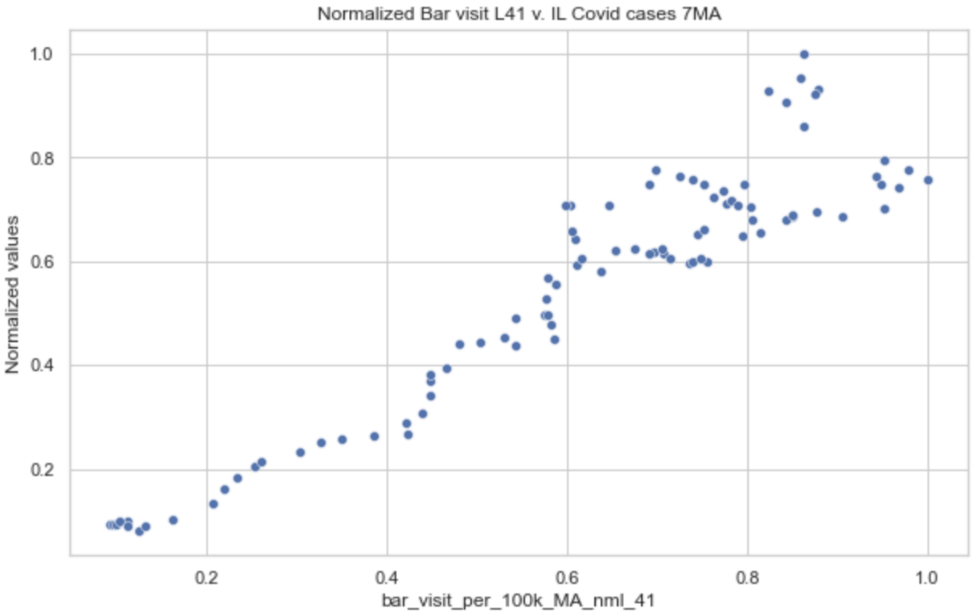


Figure 8: Scatterplot between normalized bar visits with 41-day lag and Illinois COVID cases

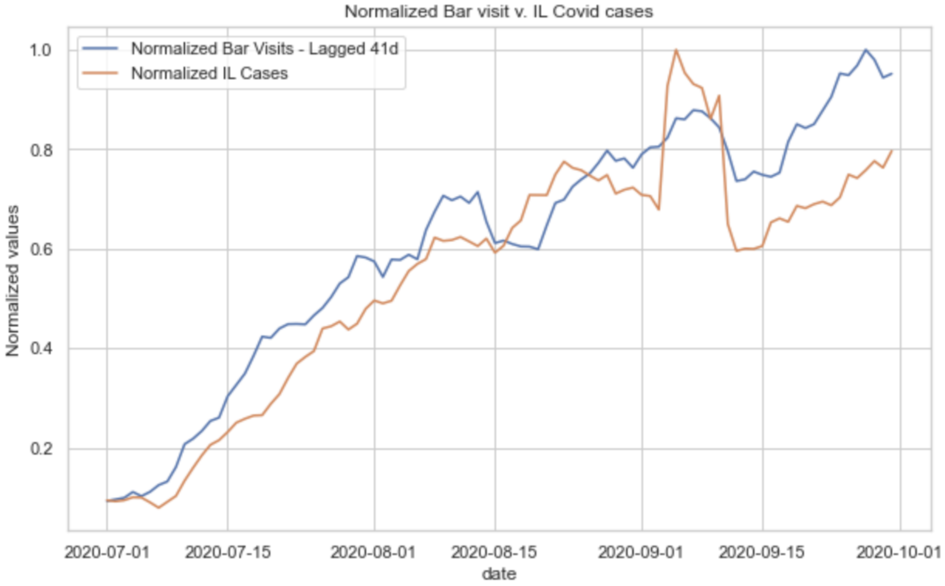


Figure 9: Line plot between normalized bar visits with 41-day lag and Illinois COVID cases

3.3. COVID-19 and Workers

The results of this study on COVID-19 and U.S. workers can be seen in the numerous visualizations shown below.

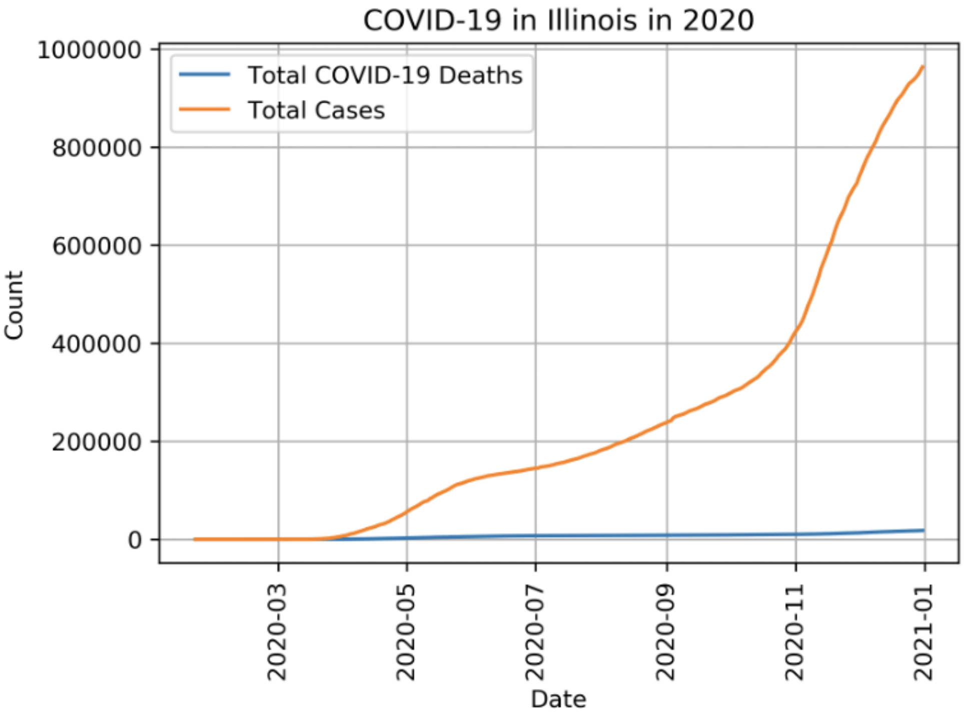


Figure 10: The total COVID-19 deaths and cases in Illinois in 2020.

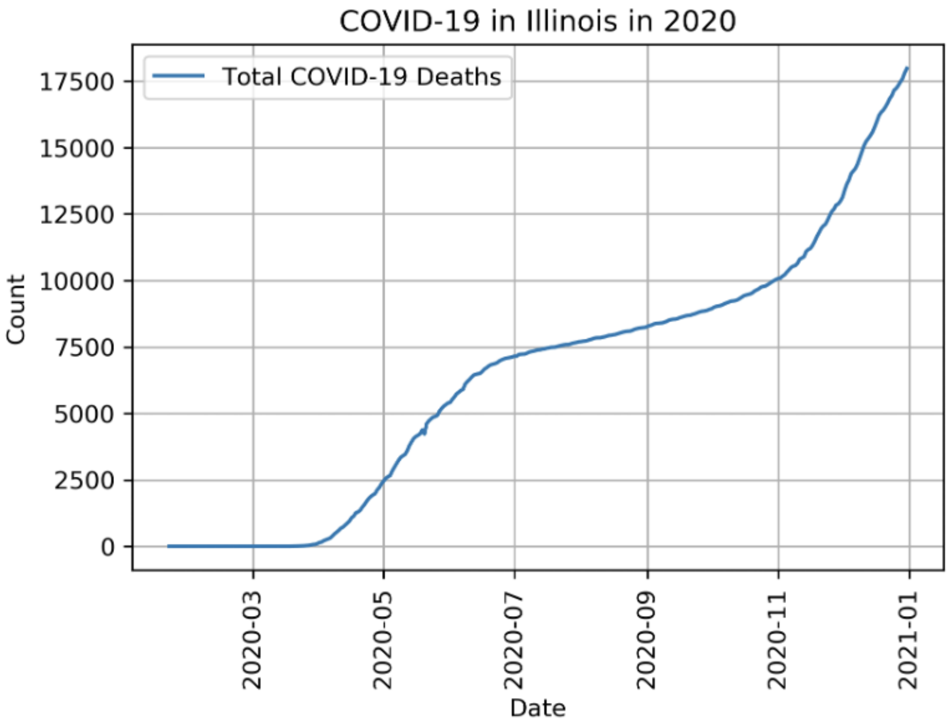


Figure 11: The total COVID-19 deaths in Illinois in 2020.

In order to properly evaluate the impact COVID-19 had on workers, it is important to first have an understanding of COVID-19 and the number of cases and deaths that occurred in Illinois in 2020 alone. Figure 10 displays the total cases and deaths in Illinois in 2020. It's clear that the case count began to noticeably rise in April 2020, and that there were just under one million cases in Illinois by the end of the year. The total deaths by COVID-19 is not clear in Figure 10 due to the scale. Figure 11 shows strictly the deaths by COVID-19 in Illinois. When graphed on its own, the death trend is more clear. Like with the cases shown in orange in Figure 10, the COVID-19 deaths began to noticeably increase in April 2020. By the end of the year, the number of deaths reached just over 17,500.

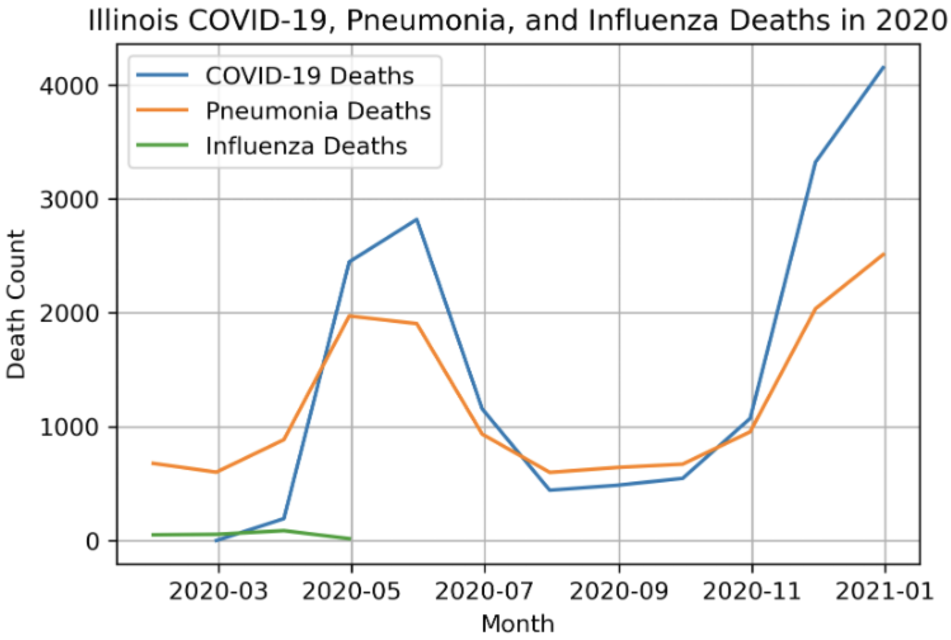


Figure 12: The monthly total of COVID-19, pneumonia, and influenza deaths in Illinois in 2020.

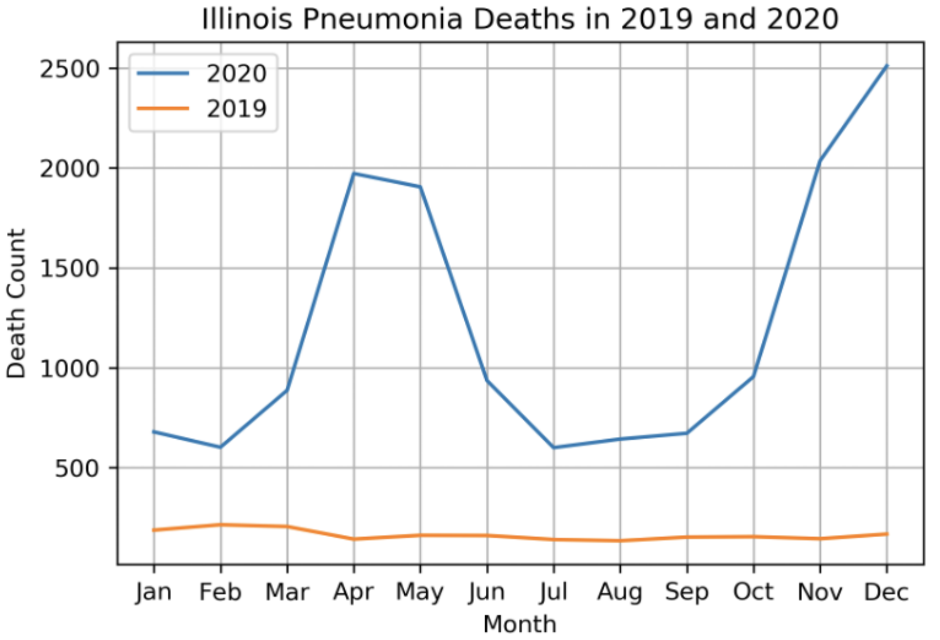


Figure 13: The total monthly pneumonia deaths in Illinois in 2019 and 2020.

While evaluating the data in the “Provisional COVID-19 Death Counts by Week Ending Date and State” dataset from the CDC, it became apparent that something interesting was happening with COVID-19 deaths and pneumonia deaths. Figure 12 shows the COVID-19, pneumonia, and influenza deaths that occured in Illinois in 2020. It is clear in the visualization that COVID-19 deaths started to be recorded in March 2020 for Illinois. It can be seen in Figure 12 that COVID-19 deaths and pneumonia deaths mirror each other with both having similar peaks and valleys. Influenza deaths on the other hand are extremely low and stopped being recorded after May 2020. It should be noted that while Figure 11 shows the cumulative death count as it increases over the year, Figure 12 and Figure 13 show the total death count for just each month.

In Figure 13, pneumonia deaths were evaluated a bit closer. Figure 13 shows pneumonia deaths in Illinois in 2019 and 2020. It is clear from the visualization that 2019 had a significantly lower record of pneumonia deaths than in 2020. At no point does the 2019 pneumonia deaths come even close to the 2020 pneumonia deaths.

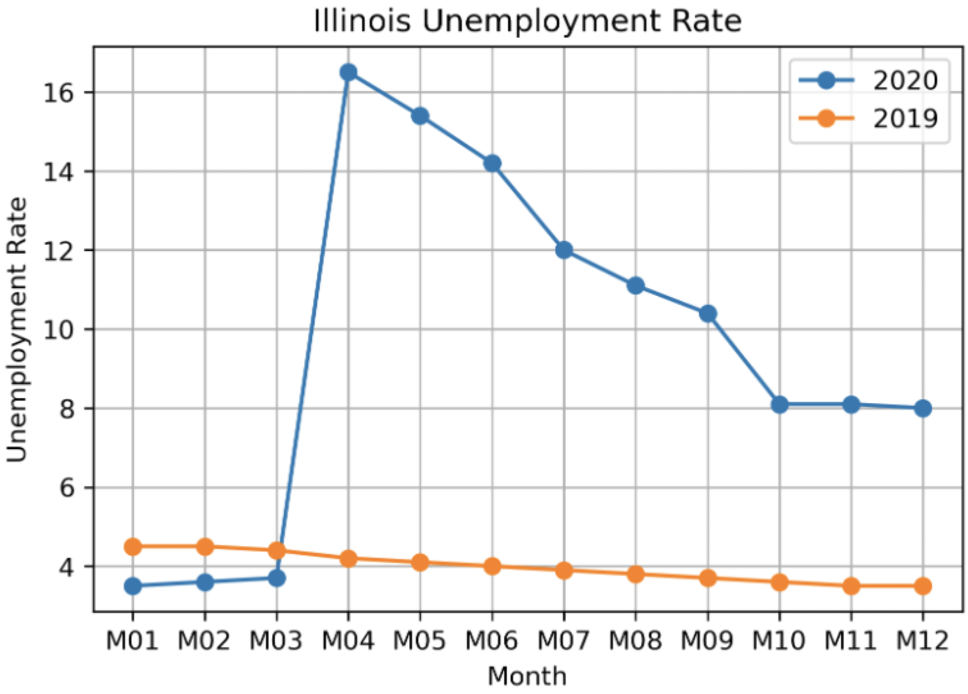


Figure 14: The Illinois unemployment rate in 2019 and 2020.

Figure 14 demonstrates the unemployment rate in Illinois in 2019 and 2020. The data in this graph is seasonally adjusted. In 2019, it can be seen that the unemployment rate was decreasing throughout the year. This changes in 2020 with a massive spike in March 2020 that reaches an unemployment rate of above 16. From that point, the unemployment rate decreases, but it appears to stagnate in October 2020 through December 2020 at an unemployment rate of around 8.

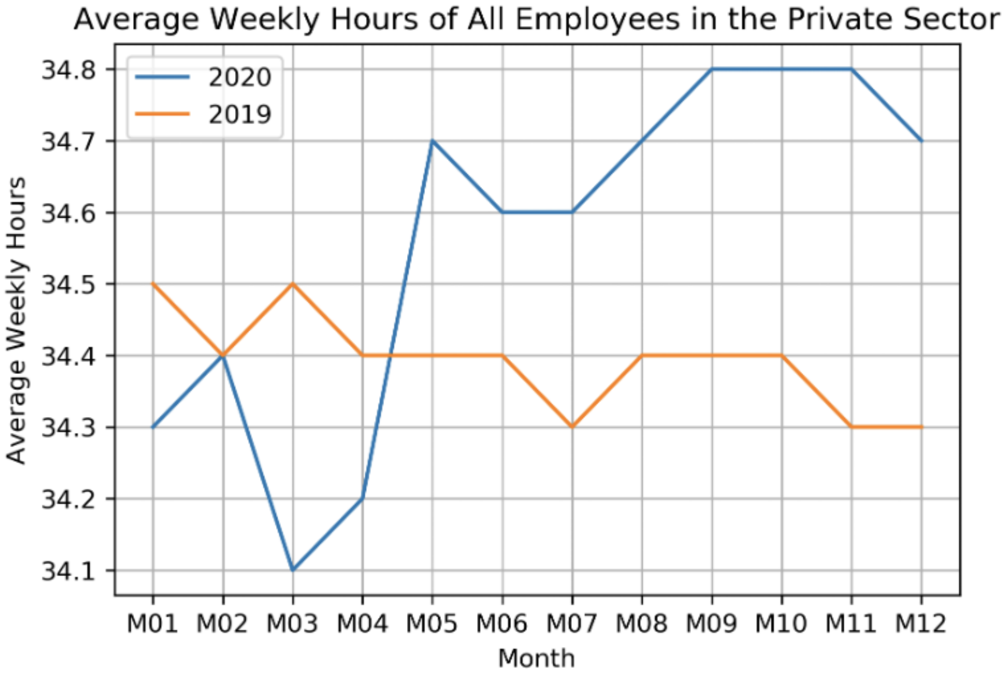


Figure 15: The average weekly hours of all U.S. employees in the private sector in 2019 and 2020.

The weekly hours employees work also changes in 2020. Figure 15 shows the average weekly hours of all U.S. employees in the private sector in 2019 and 2020. The data used in this graph is seasonally adjusted as well. In 2019, average weekly hours for U.S. employees was mostly at 34.4, but it fluctuated slightly between 34.3 and 34.5. It appears as though hours were slightly decreasing throughout 2019. This changes in 2020 when the average weekly hours drops suddenly starting in February 2020 and going into March 2020 hitting a low of 34.1. The trend then switches directions and hours increase starting in March 2020. The steepest slope occurs in April 2020, and the hours peak in October 2020 at 34.8. While the change occurring is just tenths of an hour, this data is still significant considering how stable the average weekly hours were in 2019 and how erratic the hours were in 2020. The 2020 hours show an unusual change that is not seen elsewhere.

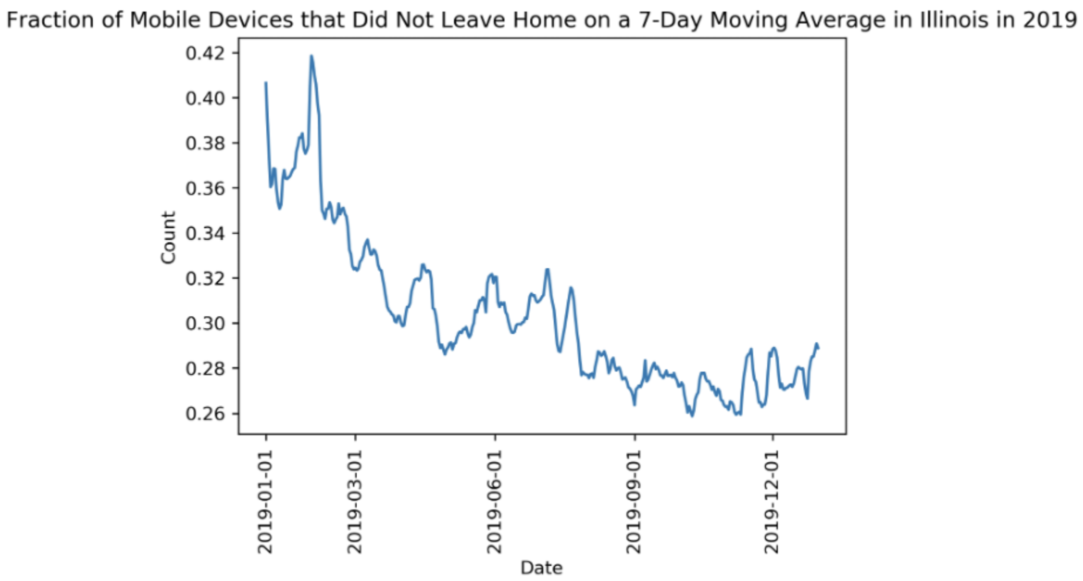


Figure 16: The 7-day moving average of the fraction of mobile devices that showed up on SafeGraph's panel of GPS pings that did not leave home in Illinois in 2019. The count is the mean over the census block groups (CBGs) in a state.

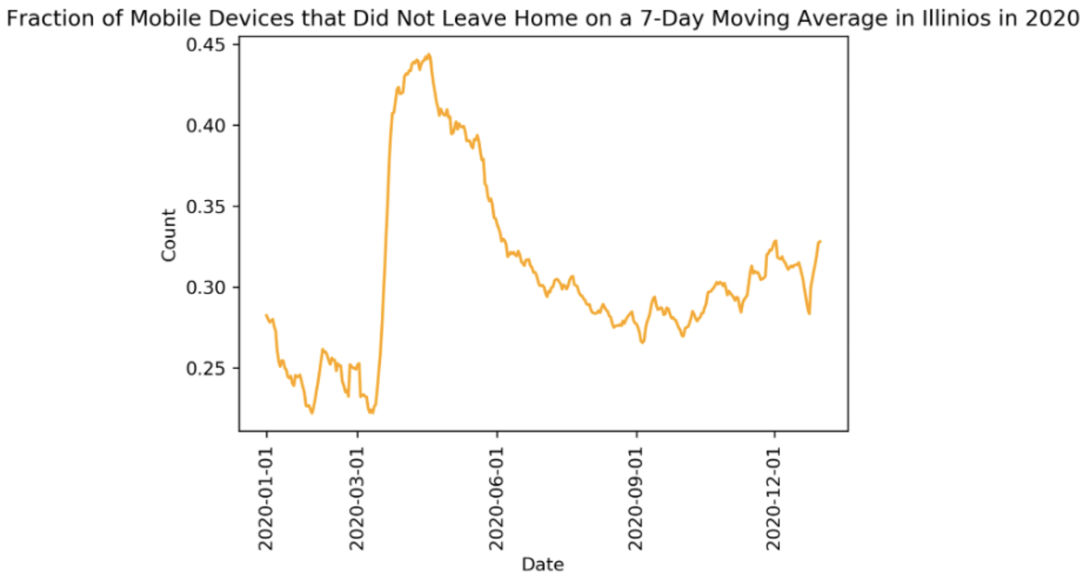


Figure 17:The 7-day moving average of the fraction of mobile devices that showed up on SafeGraph's panel of GPS pings that did not leave home in Illinois in 2020. The count is the mean over the census block groups (CBGs) in a state.

In Figures 16 and 17, the 7-day moving average of the fraction of people not leaving home in Illinois is shown for 2019 (Figure 16) and 2020 (Figure 17). It is important to note that technically it's the location of the mobile devices that are registering as not having left home, not the people. In a way, though, this is a proxy for people since so many don't leave the house without a mobile device such as a cellphone. As mentioned in the figure descriptions, these visualizations show the fraction of mobile devices that appear on SafeGraph's panel of GPS pings. Not every mobile device will show up on this panel, but the figures still give us an idea of people's movements. The home location for these mobile devices were determined by their common nighttime location over 6 weeks to a Geohash-7 granularity. The count in these visualizations is the mean over the census block groups (CBGs) in a state.

As can be seen in Figure 16, more mobile devices are registering as remaining home in the colder months of early 2019. As the year continues on, more and more devices are registering as having left home. Figure 17, showing data from 2020, has a very pattern from Figure 16 (2019). In the early months of 2020, a large fraction of the mobile devices appearing on SafeGraph's panel are shown leaving home. March is where a sudden increase in mobile devices staying home can be seen. From that point, the amount of mobile devices staying home begins to decrease as more are seen leaving. The slope changes once the colder months start creeping closer. From September on, an increasing slope of mobile devices are registering as staying home.

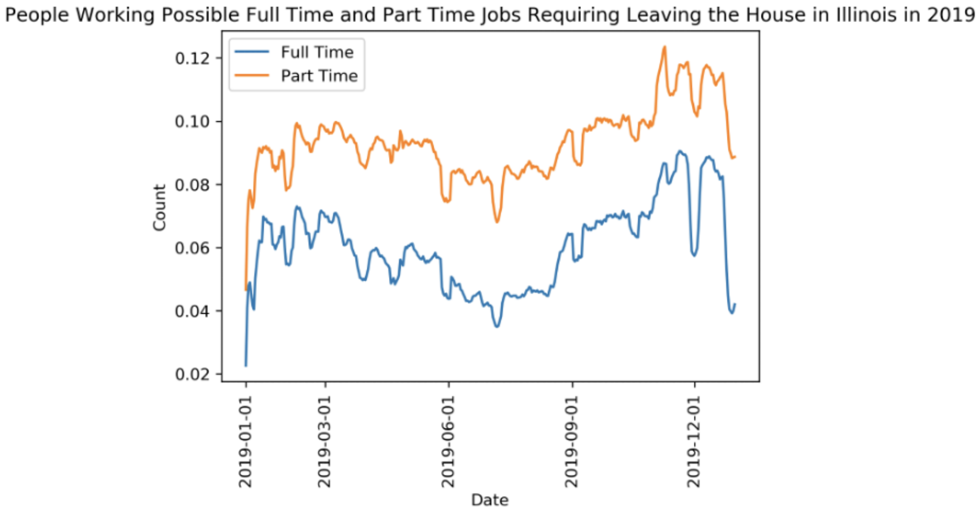


Figure 18: The 7-day moving average of the fraction of mobile devices that showed up on SafeGraph's panel of GPS pings that left home in Illinois in 2019 for three to six hours (part time) or over six hours (full time). The count is the mean over the census block groups (CBGs) in a state.

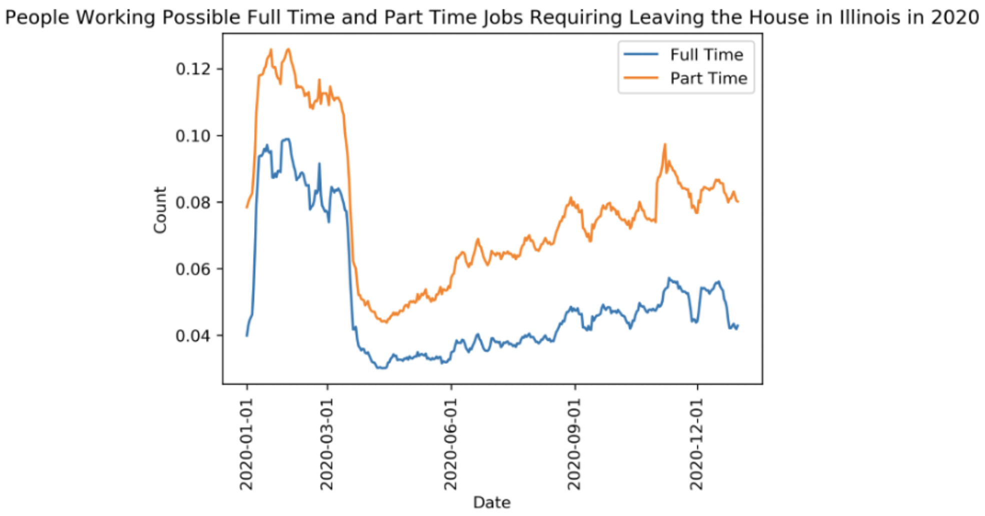
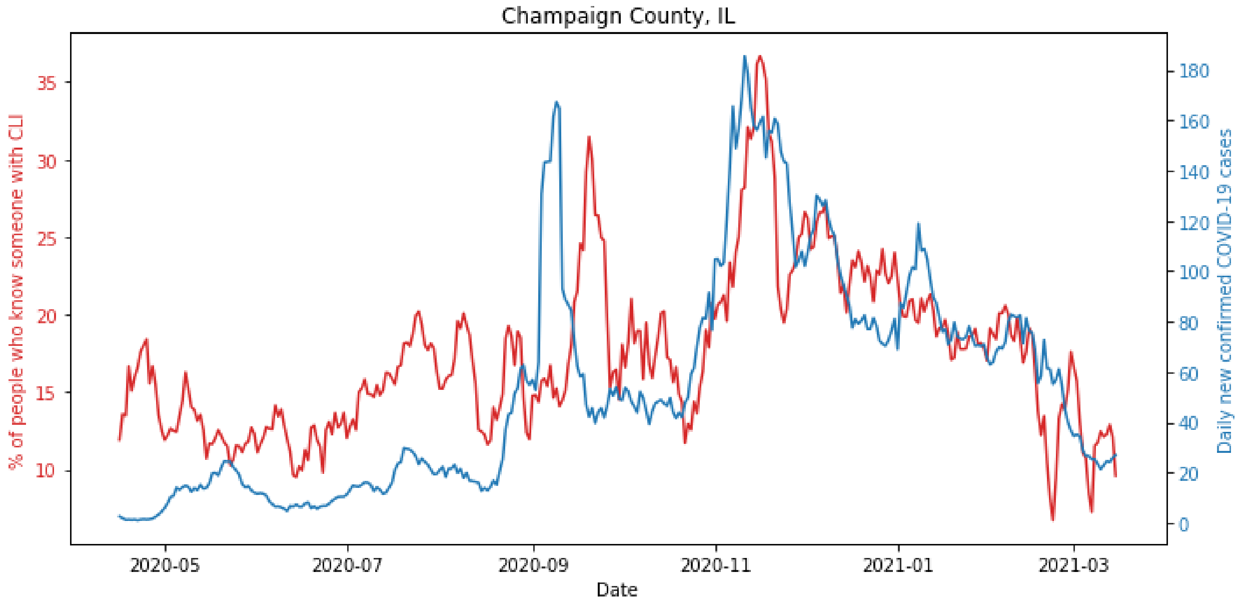


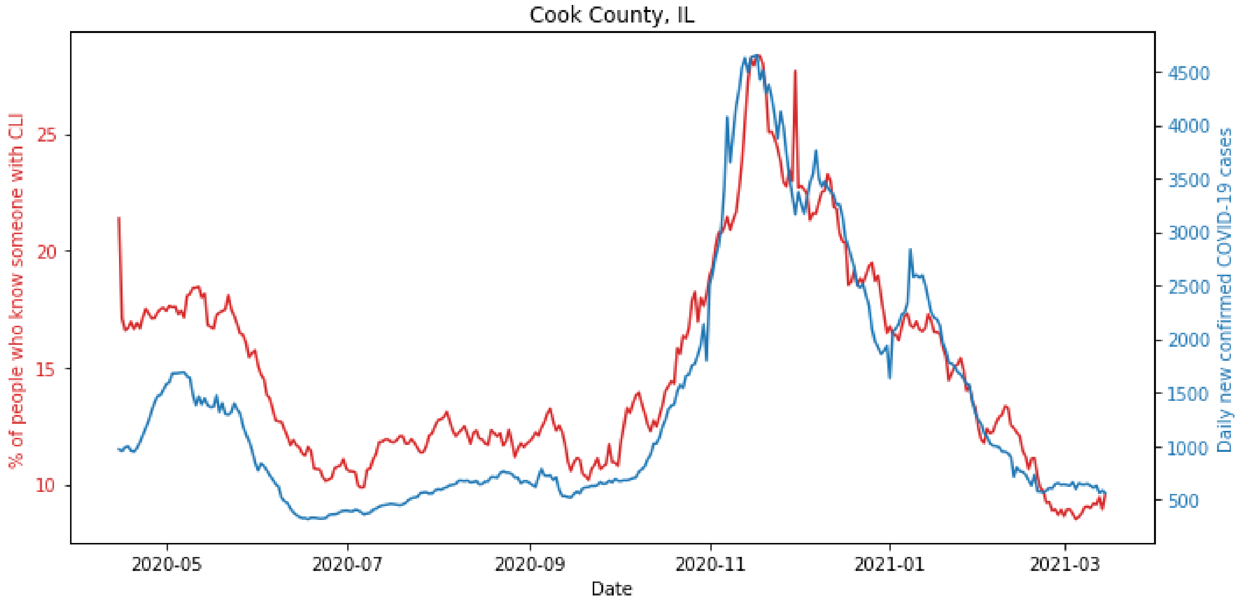
Figure 19: The 7-day moving average of the fraction of mobile devices that showed up on SafeGraph's panel of GPS pings that left home in Illinois in 2020 for three to six hours (part time) or over six hours (full time). The count is the mean over the census block groups (CBGs) in a state.

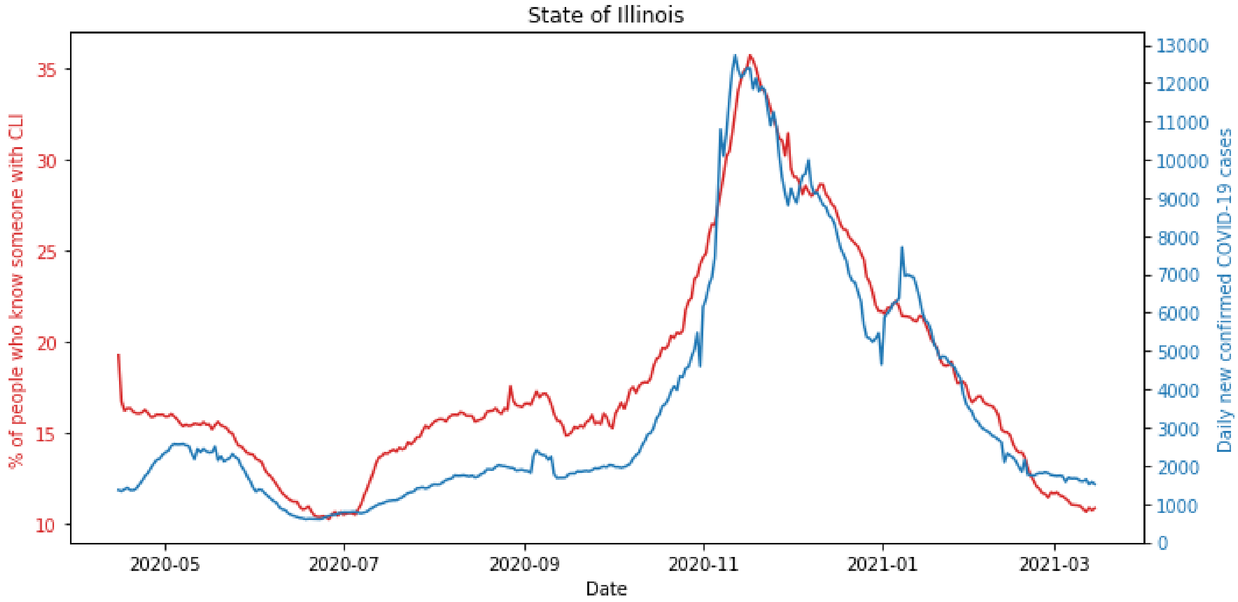
In Figures 18 and 19, the 7-day moving average of the fraction of people that left home in Illinois in 2019 (Figure 18) and 2020 (Figure 19) for three to six hours (possible part time job) or over six hours (possible full time job). Figures 18 and 19 are very similar to the previous two Figures 16 and 17. Like in Figures 16 and 17, these visualizations show the fraction of mobile devices that appear on SafeGraph's panel of GPS pings. Keep in mind that mobile devices include a wide variety of devices, not only cell phones. This is a contributing factor to the low fractions. Not every mobile device will also show up on this panel. Only the devices that ping on SafeGraph's GPS panel at the time are calculated. So, while the data isn't directly tracking people and there is some variability in the devices that ping the GPS panel, the figures still give us a basic idea of people's movements during this time period. Mobile devices are used in this case to track people's movements. So, if a mobile device leaves home and pings the GPS panel, then it is assumed that a person is moving the device and therefore leaving home as well. The home location for these mobile devices were determined by their common nighttime location over 6 weeks to a Geohash-7 granularity. The count in these visualizations is the mean over the census block groups (CBGs) in a state.

As can be seen in Figure 18, there are more mobile devices leaving the home for three to six hours, possibly indicating the owner works a part time job, than there are mobile devices leaving the home for six or more hours, possibly indicating the owner works a full time job. In 2019 (Figure 18), there is a slight dip in the middle of the year during the summer with less people leaving home for such long periods of time. There is then an increase in the fall to early winter with more people leaving home for three to six hours or six or more hours. In a way, the lines in Figure 18 almost make a slight "M" shape. Figure 19, showing 2020 data, is not like that at all. In Figure 19, there is a steep decrease in mobile devices leaving home for long periods of time right around March and April of 2020. After this point, there is a slow increase throughout the rest of the year in mobile devices being picked up by the GPS panel outside of their home. There is a slightly steeper increase in mobile devices being picked up outside the home for three to six hours than there is in mobile devices outside the home for six or more hours.

3.4. Relationship between reported COVID-like-illness in community and COVID cases

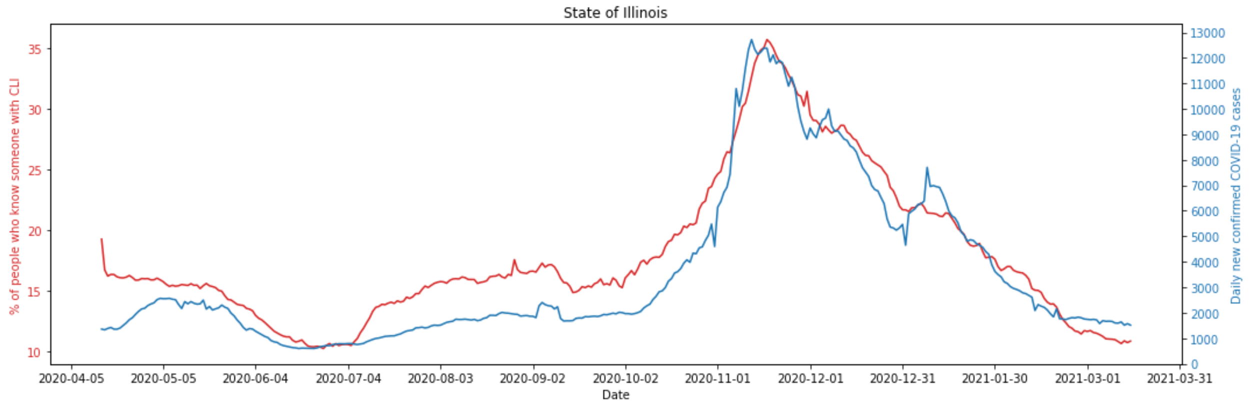




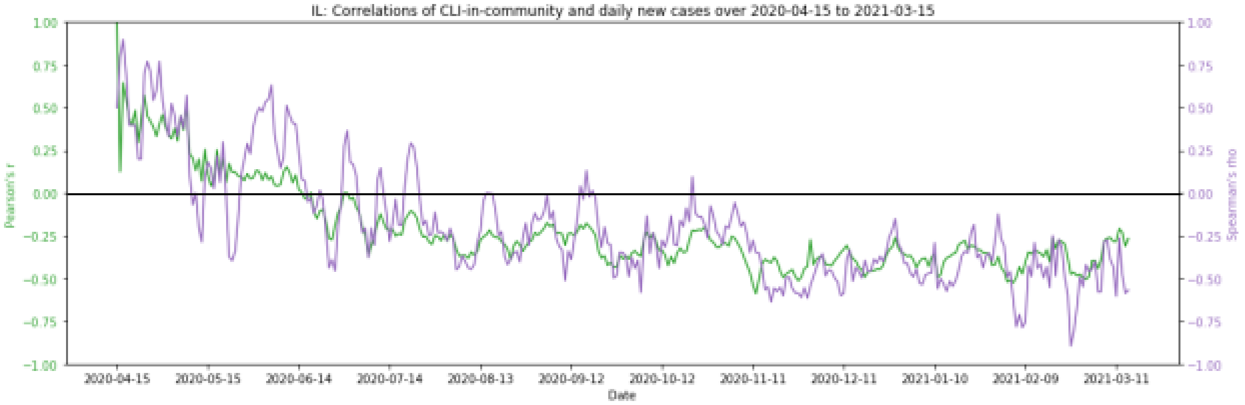


Comparing Champaign county to Illinois, we see that Champaign County's smaller surge in early September was not strongly reflected in the magnitude of the state region. Unlike some of the counties from the southern and western U.S. that were featured in the Delphi Group's analysis, it appears that Cook County did not experience a major surge in the middle of June of 2020. Instead, Cook County and Illinois experienced an earlier surge at the start of May, in which the daily new cases plateaued between 1000 and 1,500, followed by a very large surge in mid- to late November that peaked at 4,5000.

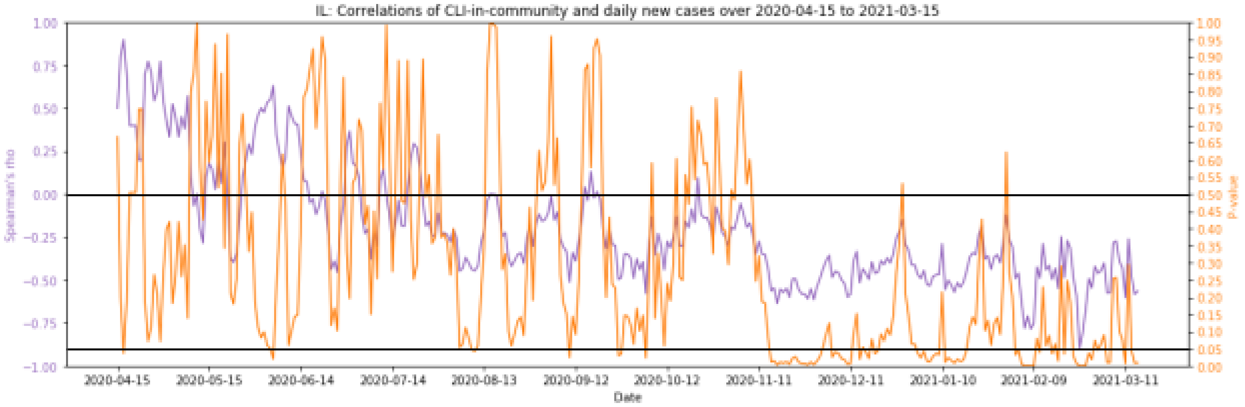
In the case of Champaign County, CLI-in-community fails to anticipate daily new cases and in fact lags behind daily new cases. For Cook County and Illinois, while CLI-in-community closely follows daily new COVID cases, it does not anticipate it. At the inflection point right around the beginning of October, the slope of CLI-in-community is slightly less steep than that of daily new cases.



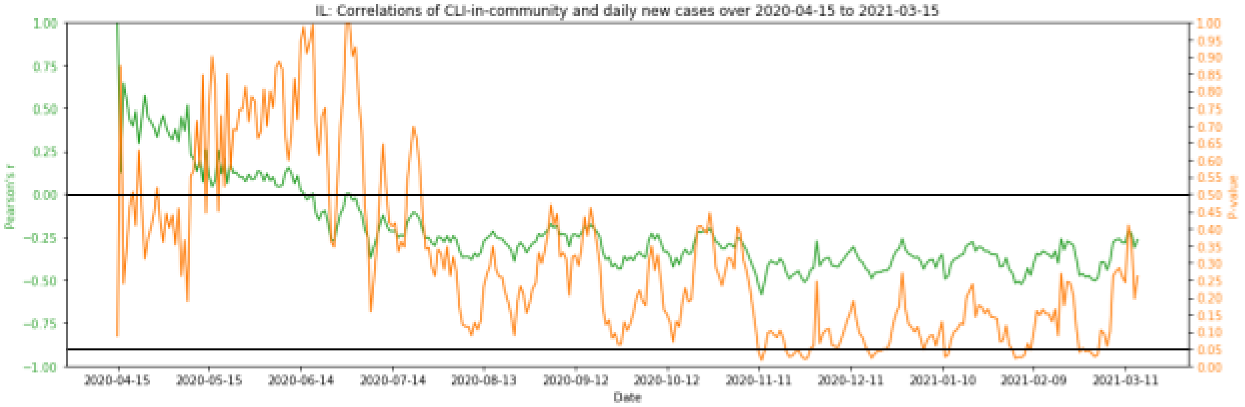
State of Illinois



IL: Correlations of CLI-in-community and daily new cases over 2020-04-15 to 2021-03-15



IL: Correlations of CLI-in-community and daily new cases over 2020-04-15 to 2021-03-15



IL: Correlations of CLI-in-community and daily new cases over 2020-04-15 to 2021-03-15

4. Discussion

4.1. COVID-19 Vaccination Progress

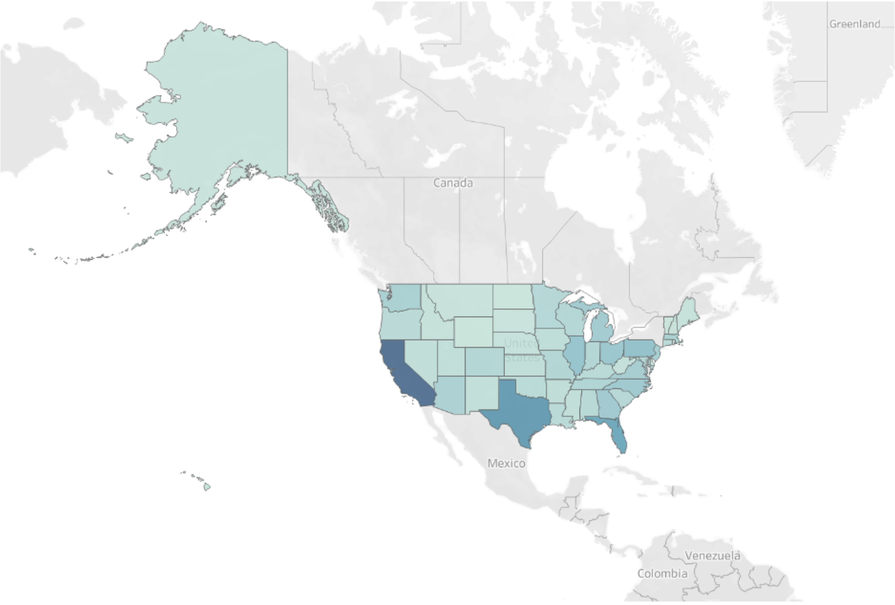


Figure 20: Total Vaccines distributed to each state.

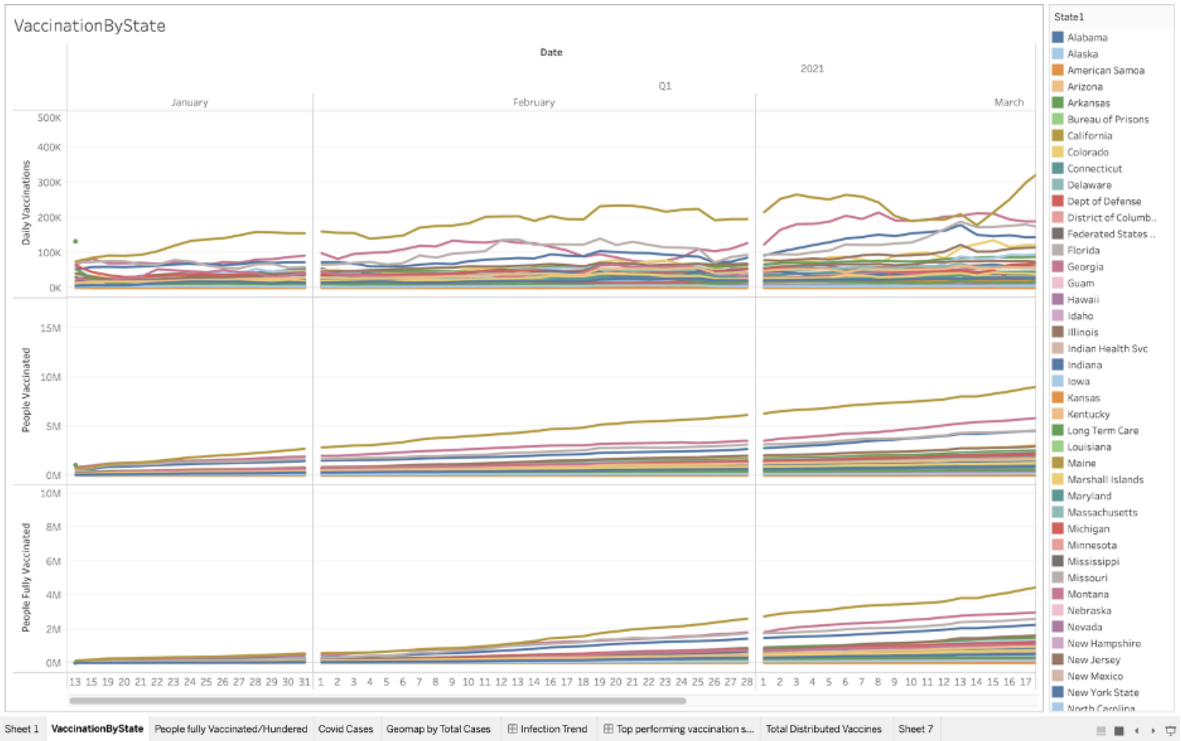


Figure 21: Vaccination by State.

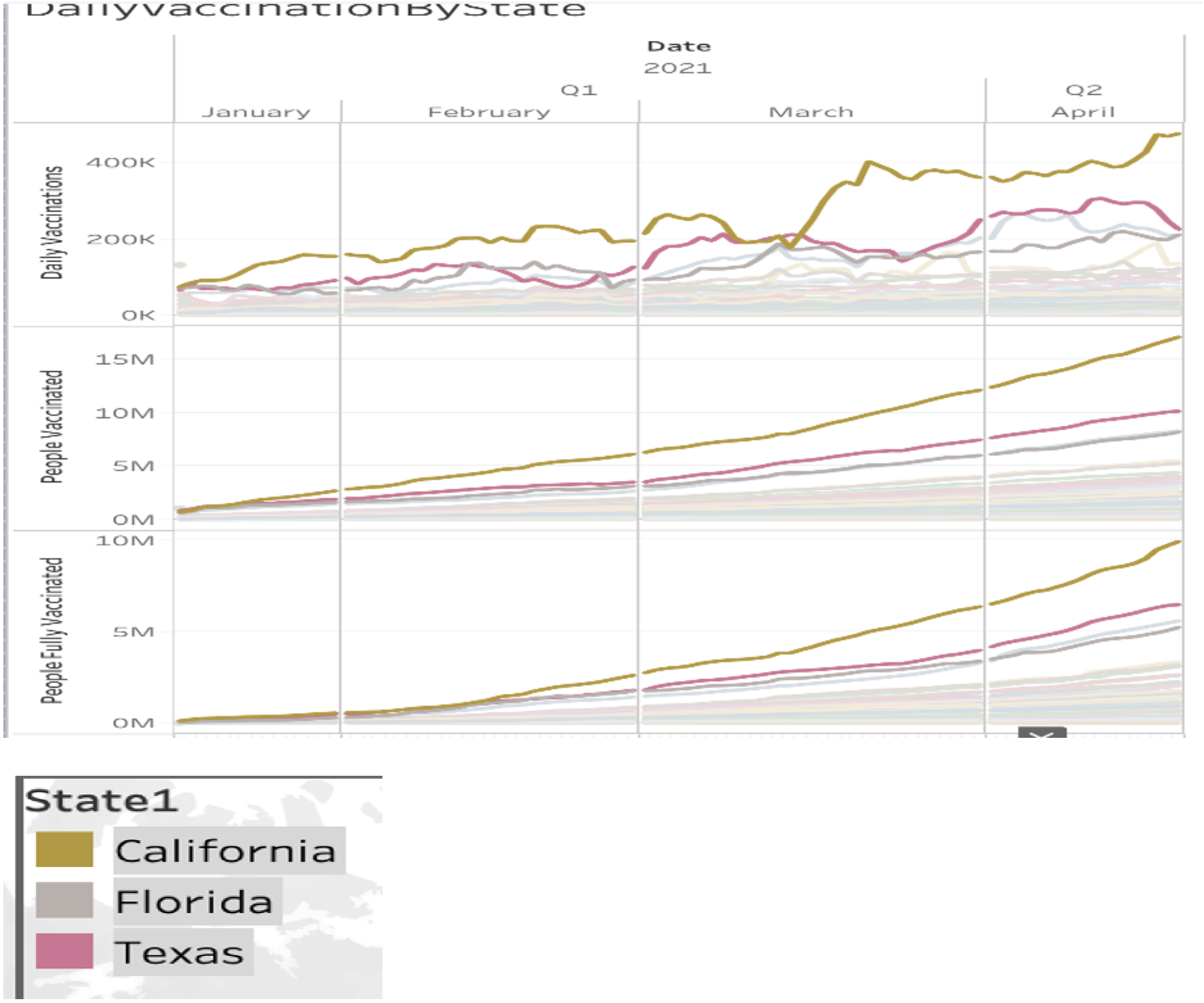
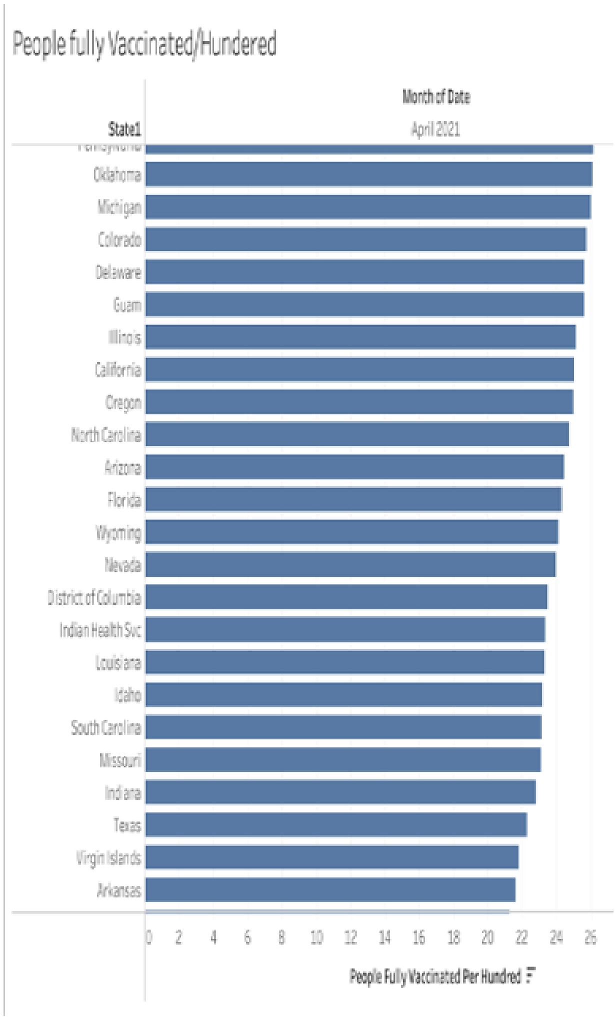


Figure 22: California, Texas, and Florida ranking in daily vaccination by state

Figure 20 from the Figure 3 dashboard shows the total vaccine distributed to each state in the United States. From the chart, we can see that the top three states with the highest number of vaccines distributed to date are California, Texas, and Florida. It is important to analyze if the three states are still standing in the top three positions in terms of the daily vaccination, people vaccinated, and people fully vaccinated because it would correlate to the total vaccines being distributed to the state. Figure 22 and Figure 21 show the top-performing states in the vaccination program; the results are that California, Texas, and Florida are the top-performing states. They have the highest value of vaccine distribution, daily vaccination, people vaccinated, and people fully vaccinated. Since those three states are some of the most highly populated states in the United States, it results in top-performing states. Also, in terms of people fully vaccinated per hundred people, New Mexico, Maine, and Alaska are ranking at the top. This could be because of the population density in those states. On average, 32 people out of 100 are fully vaccinated in New Mexico, Maine, and Alaska. But in California, Texas, and Florida on average 24 out of 100 people are vaccinated. This results that more ratio of people who are fully vaccinated in California, Texas, and Florida to population density in top-performing states.

The top three performing states might be performing well because of the number of populations in the state which results in a greater number of vaccines being distributed. Therefore, it is required to analyze based on fully vaccinated people with population ratio. Figure 23 shows the people fully vaccinated per hundred. The results show New Mexico, Maine, and Alaska are the three leading states in daily vaccination per hundred. On average, at least 32 people out of 100 in New Mexico, Maine, and Alaska are fully vaccinated. In California, Texas, and Florida on average least 27 out of 100 people are fully vaccinated. Even though New Mexico, Maine, and Alaska are the top three leading states in terms of people fully vaccinated per hundred, still California, Texas, and Florida are the best vaccine program performing states because of the population density.



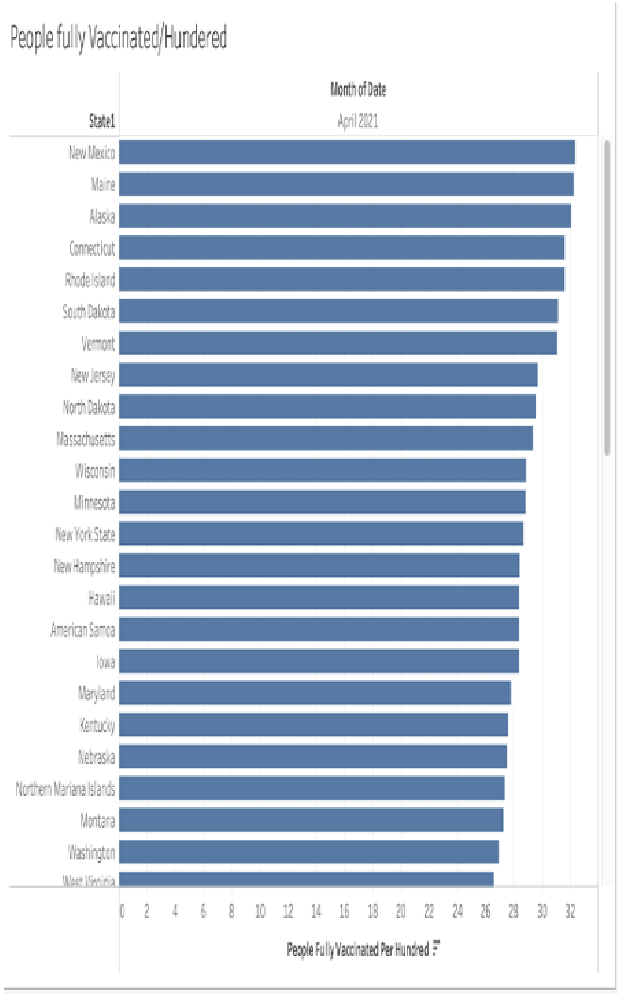


Figure 23: People Fully vaccinated per hundred by state.

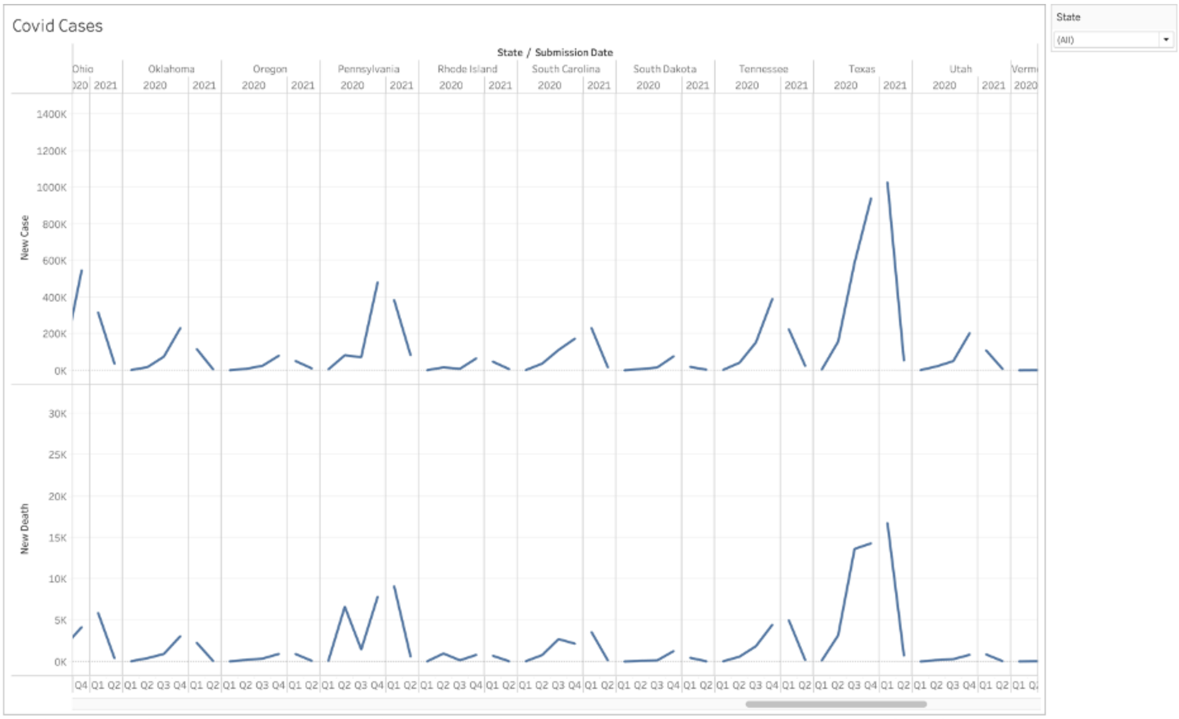


Figure 24: COVID Cases by State

Figure 24 from Figure 4 shows the infection trend in terms of COVID-19 cases and deaths declined gradually in relationship with an increase in COVID-19 vaccinations. There was a rise in COVID-19 cases and deaths by the end of 2020, but as the vaccination program began in January 2021, it resulted in a slow decline of COVID-19 cases and death curve in every state. It results that the COVID-19 vaccination is effective.

4.2. Illinois COVID Daily Case Number v. Bar Visits

Based on the TLCC analysis along with different correlation measures, we observe a 41-day lag using the 7-day moving average. One of the reasons for picking 7 days as the moving average is that the bar visits fluctuate on a weekly basis. More people visit bars on weekends while less people go there in the middle of weekdays, as shown in Figure 25.

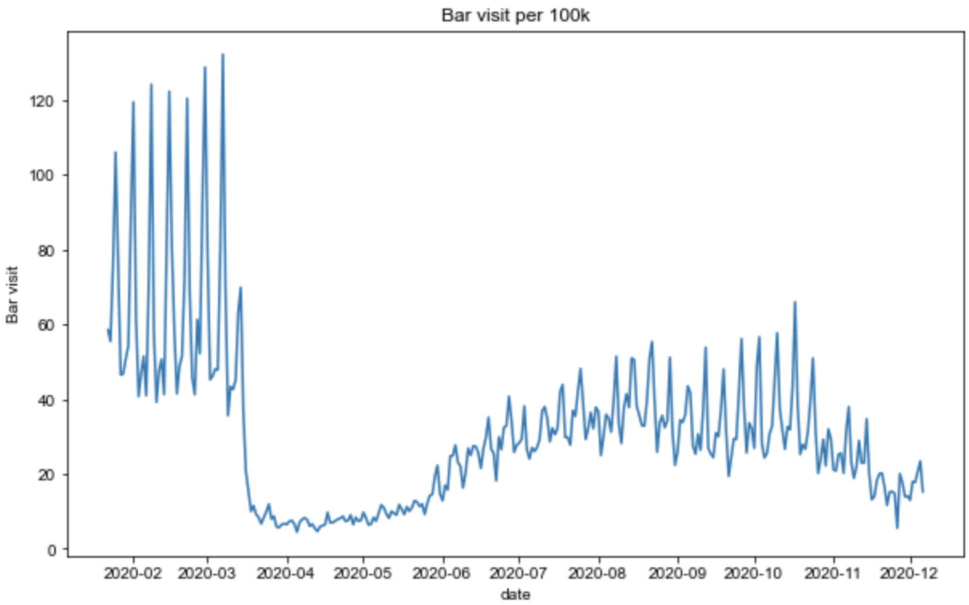


Figure 25: Bar visit plot without applying moving averages

I also built a model using 14 days as moving average rather than 7 days. The correlation result is shown in Figures 26.

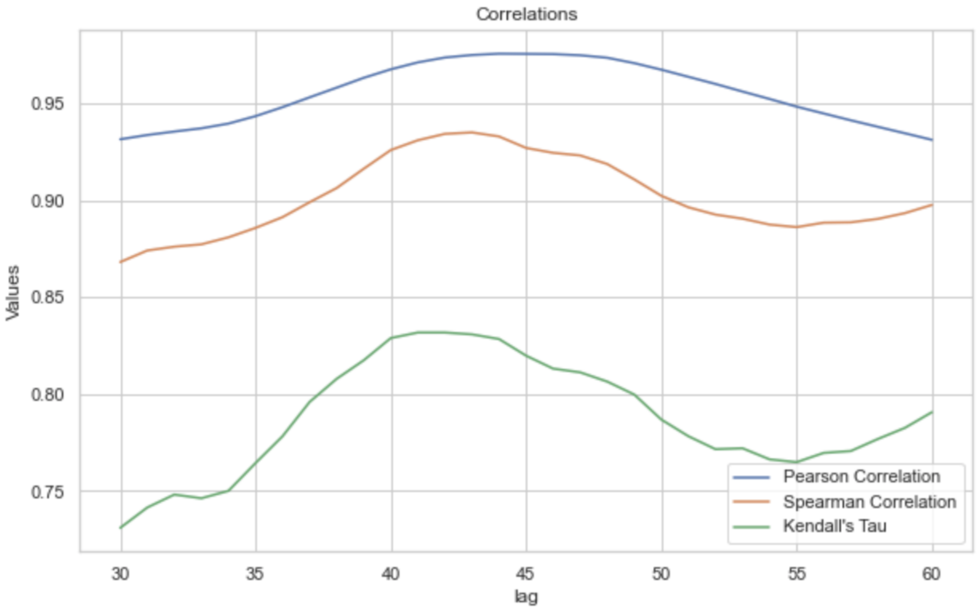


Figure 26: Correlations under different lags with 14-day moving average

The highest Pearson correlation occurs on the 44th day of lagging on the bar visits. The plots are shown in Figures 27 and 28.

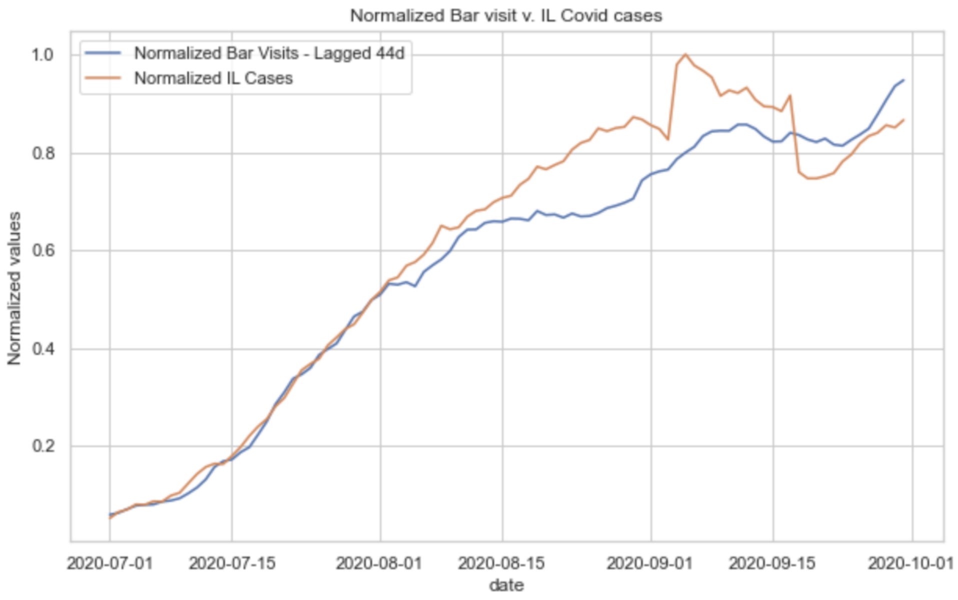


Figure 27: Line plot between normalized bar visits with 44-day lag and Illinois COVID cases (14-day moving average)

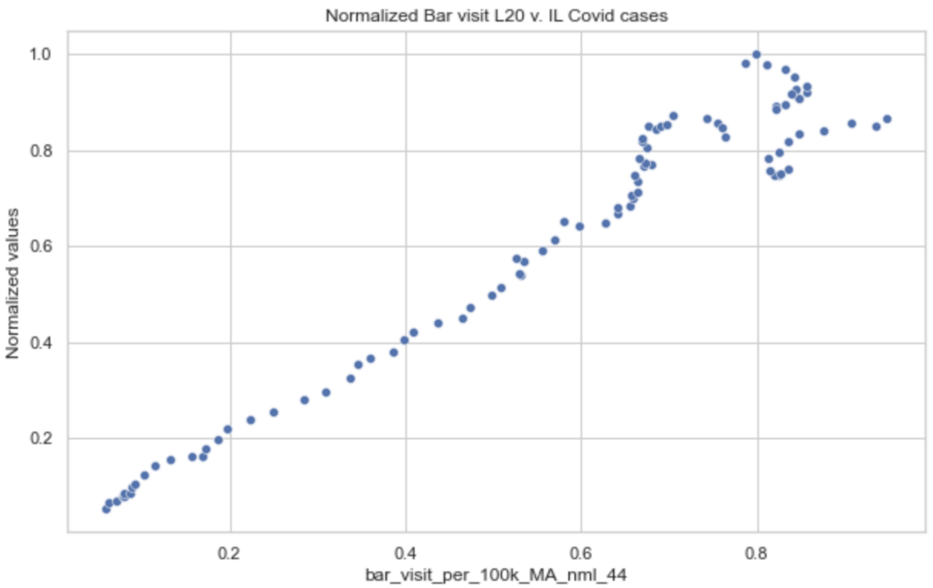


Figure 28: scatterplot between normalized bar visits with 44-day lag and Illinois COVID cases (14-day moving average)

The models using the 7-day moving average and 14-day moving average show a 3-day difference in the results (41st day and 44th day, respectively). The p-values of the correlation in both models are less than 0.01, from which we can reject the null hypothesis that the Illinois COVID case numbers are not correlated with the bar visits.

The results indicate that the surge of the Illinois COVID daily case number in summer 2020 had a lag around 42 days following the increase of in-state bar visits. The lag is more than one month, which can be explained that it takes some time for the virus to be widespread after the initial infection. Moreover, some infected people might be asymptomatic, which reduced the increase of the reported number. It is also possible that the increases in bar visits do not contribute to the surge of the Illinois cases in the summer, while some other factors contributed to the growth of numbers, such as the increase in restaurant visits or other untracked large gatherings.

The models I built return 0.95 as the Pearson correlation score, which still states that the pattern of the bar visits changes matches how the Illinois COVID daily new case numbers surged, with a 42-day lag.

4.3. COVID-19 and Workers

As can seen by the numerous visualizations in the results section for research question three, a lot of data was used and a lot of results were found while analyzing COVID-19 in 2020, primarily in Illinois, and how it impacted workers. In this section, we'll be analyzing the results in more depth as well looking at possible confounding factors and unknown information.

First, it's important to analyze COVID-19 cases and deaths on their own before comparing it with the trends seen throughout 2020 for workers. In Figure 10, the case count of COVID-19 in Illinois began noticeably increasing beginning in April 2020. When looking at the COVID-19 death count in Figure 11, it is also clear that April 2020 is the time that COVID-19 deaths began picking up as well, although Figure 12 does show that there were some COVID-19 deaths in March as well. The slope in Figure 11 levels off a little more between June 2020 through October 2020 due to a lower death rate during this time. This can be observed in the monthly total of COVID-19 deaths in Illinois in Figure 12 and Figure 13 where there is a big dip in COVID-19 deaths between June through October 2020. As mentioned in the results section, COVID-19 deaths and pneumonia deaths mirror each other very closely in Figure 12. There are a couple possibilities about why this is occurring. It could be that a lot of the pneumonia deaths are miscategorized COVID-19 deaths. As can be seen in Figure 13, there were not nearly as many pneumonia deaths in 2019 as there was in 2020. Plus, the trend of pneumonia deaths in 2019 was pretty stable throughout the year, which is very different from the trend of pneumonia deaths in 2020. There is a possibility that what was actually COVID-19 is being miscategorized as pneumonia and that there are actually a lot more COVID-19 deaths in Illinois than we ever realized. That is one possibility. The second possibility is that COVID-19 and pneumonia are being recorded as the primary and secondary causes of death for many of the people dying and that the numbers in this dataset are the cumulative count each month of both the primary and secondary causes. The documentation for the data used in Figure 12 didn't state whether only primary causes of death were recorded or if both primary and secondary causes were recorded, so there's no way to know for sure. One thing that can likely be determined from Figure 12, though, is that people were dying in Illinois from COVID-19 before March 2020. As can be seen in Figure 12, COVID-19 deaths really started being recorded in March, but pneumonia deaths were already much higher than typical. In January 2020, pneumonia deaths were already more than triple what they were for the month with the highest death count in 2019 (Figure 13). This is highly unusual in comparison to the year before. This means that there is a very real possibility that COVID-19 was killing Illinois residents well before March, possibly as early as January 2020.

When evaluating the data collected on workers, it is clear that the pandemic had a large impact on them. As seen in Figure 14, unemployment in Illinois skyrocketed from March to April 2020 hitting an unemployment rate of over 16. That's quadruple what the unemployment rate was at the same month the year before in 2019. Taking into context what was happening at the time, March 2020 was the month where Illinois entered its first lockdown due to growing concerns about the pandemic. Around this same time is when the average weekly hours of U.S. employees in the private sector hit its lowest (Figure 15). In February 2020, the average weekly hours of employees in the U.S. plummeted hitting its lowest point in March at 34.1 hours. From that point, the hours rebounded, easily surpassing the average hours in 2019, to hit its highest peak in September 2020 at 34.8 hours. A couple tenths of an hour likely doesn't seem like much, but this is the national average. Millions of employees' weekly hours make up this average meaning that there is a large base. Unless there are some dramatic changes in hours for a large portion of the population, the average isn't likely to change much. Also, when comparing the hours in 2019 to 2020 in Figure 15, the change is very clear. In 2019, the national average was slowly decreasing throughout the year, and even then, there still was only a change of two-tenths of an hour over that entire time. In 2020, the national weekly hours had a change of eight-tenths. That's four times as much as in 2019. A possibility for the decrease in hours seen in February 2020 could be due to the furloughs and businesses shortening their operational hours. As for the sudden increase in hours beginning in March 2020 and steeply climbing in April, two possibilities come to mind. The first possibility is that due to so many workers being furloughed and laid off, the remaining workers at various companies and businesses have to work even harder and longer to carry the added workload. The second possibility could be that a higher number of part time workers were fired during this time causing the national average to shift higher due to the lower data points no longer existing. March 2020 is when the unemployment rate rapidly rose (Figure 14), and March 2020 is also when the national average weekly hours began to increase after hitting its lowest point (Figure 15).

Lastly, the working environment changed for a number of workers throughout 2020. This can be seen when looking at Figure 16 and Figure 17. Both figures display the 7-day moving average of the fraction of mobile devices that did not leave home. As addressed extensively in the results section, these visualizations aren't all encompassing since the data is limited by the number of mobile devices that appear on SafeGraph's panel of GPS pings. This means that not every cell phone, tablet, or any other mobile device will show up on this panel. The data can be seen as a small sample of the population with the fraction being calculated from this small sample. The data is incomplete to a degree but this is some of the only data available to the public that tracks human movements during this time. Figure 16 shows a steady decrease in people staying home throughout the year in 2019. The 2018-2019 winter was a particularly brutal one in Illinois, so it makes sense that more people stayed home if possible during that time. As the weather warmed up, more people left home more often, and this continued into the very mild 2019-2020 winter. The pattern changes significantly for 2020 in Figure 17. The expectation due to the mild 2019-2020 winter would have been that the fraction of mobile devices not leaving home would be fairly low throughout the year and may increase slightly near the end. Instead, there was a large spike in the fraction of devices being read by the panel as staying home in March 2020 right when lockdown first occurred in Illinois and when unemployment rose rapidly. This point in the figure could potentially indicate an increase in workers working remotely or workers being fired or furloughed and spending more time at home. As for Figure 18 and Figure 10, these visualizations display the mobile devices that registered on the SafeGraph panel of GPS pings that left home for long stretches of time that could indicate part time or full time employment. Now, there is no way to determine for sure if the owners of these devices actually work a full time or part time job. The data is only indicating a possibility of a full time or part time job based on time alone. The fraction of people leaving the house in these visualizations is very low, so keep in mind that the data used has a number of unknown variables and is from a very small sample size. These visualizations are mostly for viewing general trends in movement. Comparing the trend of movement in 2019 (Figure 18) with 2020 (Figure 19) shows a somewhat similar trend to what we've seen in previously mentioned visualizations. Like with unemployment and hours, a change occurred right around March and April 2020. Figure 19 shows a sudden decrease in the fraction of mobile devices leaving the home for long stretches of time. While the change isn't the most significant considering the scale, there is still a change that occurred in time with a number of other changes that we know occurred at the same time.

4.4. Relationship between reported COVID-like-illness in community and COVID cases

The graph of the state of Illinois plots the daily COVID cases and reported CLI-in-community for each day between April 15, 2020 and March 15, 2020. As mentioned in the methods section, the correlation of CLI-in-community and COVID cases was calculated for all reporting counties on each particular day and plotted over time. The sample size and degrees of freedom varied from day to day since the number of reporting counties in the state of Illinois was not constant.

The comparison of the Spearman and Pearson correlations for the counties reporting in the state of Illinois reveals that they follow a similar trend even as the Spearman oscillates more than the Pearson. They are both mostly positive until mid-June, after which both trend lines cross over to mostly negative correlation values. In early June, the Spearman correlation reports a strong positive monotonic relationship (r = ~0.70) between the two variables and the Pearson correlation reports a weakly positive (r = ~0.12) linear relationship between the two variables.

Around the big surge in November, a weak to moderate negative monotonic relationship and weak to moderate negative linear relationship is observed at the county level. This seems unexpected because the two variables have sharp positive slopes at the state level in November. This shows that there is an association between CLI-in-community and daily COVID cases at the county level, but this association is not reflective of their trends at the state level (where all counties are summed). We take this to mean that CLI-in-community is decreasing while COVID cases are increasing and vice versa at the county level. This may be caused by the fact that CLI-in-community is lagging slightly behind COVID cases, but further work is needed to ascertain whether this is the case.

Since we set α = 0.05, any p-values less than 0.05 will allow us to reject the null hypothesis and accept the alternative hypothesis. The p-values vary greatly over time and they are more consistently low after mid-July. We take this to be evidence that the data is not very robust, since the probability that the correlation relationships were observed by chance vacillates so much. This may be caused by the fact that only about one fifth of Illinois counties reported CLI-in-community regularly. Further understanding could be gained by conducting the same tests at larger geographic ranges within the Midwest. Interestingly, the p-values appear to fall more consistently below 0.05 during and after the November surge. One point of note is that in early June, where the Spearman correlation jumps to r = ~0.70, the p-value is less than 0.05 while the p-value for the Pearson correlation (r = ~0.12) is extremely high, around 0.80. It would be interesting to conduct this comparison with a more robust data set or larger geographic region.

5. Conclusion

5.1. COVID-19 Vaccination Progress

The research presented in this paper sheds light on COVID-19 vaccination progress and results. We can conclude that California, Texas, and Florida are the top three performing states in terms of vaccination. COVID-19 vaccination is proving to be effective with decline in new COVID-19 cases and deaths. This analysis will help in predicting the future of the pandemic end.

5.2. Illinois COVID Daily Case Number v. Bar Visits

The models reflect a strong correlation between Illinois daily new COVID case numbers and the in-state daily number of bar visits with a 42-day lag. It only indicates that the surge of the COVID cases in summer 2020 matches the increase of the bar visits. 42-day lag is long, which makes the connection between COVID case numbers and bar visits weaker. Additionally, there might be other factors that caused the surge of the case numbers. In the future study, we might consider building a more completed multivariable model, including the restaurant visits and ratio of people working from home.

5.3. COVID-19 and Workers

The visualizations presented in this study shed light on the impact COVID-19 has had on workers. COVID-19 has affected people's health, their employment, their work hours, and even their work environment and location. Across the board, a significant change can be seen in every visualization right around March and April 2020. This is exactly when the case and death counts began to officially take off in Illinois and the first lockdown occurred. Interestingly, while the Illinois public became aware of the in-state cases and deaths at that time, it's very likely that COVID-19 had already been silently moving through the population starting back in January under the misclassification of pneumonia. The pandemic has greatly impacted how people work and has likely also affected how people view work. Future studies on this topic would expand to a national level and also analyze survey data on people's views on work and how they wish to continue work: whether in-office, at home, or with a flexible schedule.

5.4. Relationship between reported COVID-like-illness in community and COVID cases

The public health significance of being able to predict COVID case surges with reports of community illness is enormous. At the very least, such knowledge could help prepare governments, public health departments, business owners, and communities for upcoming surges. Ideally, this knowledge could guide purchasing decisions and resource allocation during a pandemic. The results of this investigation are inconclusive as to whether reports of community illness can adequately anticipate COVID case surges, but very promising. This work was limited by surveying the state of Illinois and its smaller daily reporting sample sizes in the Delphi Epidata Covidcast. Working with CLI-in-community responses is challenging because of the self-selection of sample sizes and certain inherent biases in the results, but may prove fruitful. Future research into larger geographic areas could give more significant measures and a greater understanding of the association between these variables.

6. References:

Centers for Disease Control and Prevention. (n.d.). *United States COVID-19 cases and deaths by state over time*. Centers for Disease Control and Prevention. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

CMU Delphi Research Group. (n.d.). *Safegraph*. Delphi Epidata API. https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/safegraph-inactive.html

National Center for Health Statistics. (2021). *Provisional COVID-19 death counts by week ending date and state*. Centers for Disease Control and Prevention. https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab

Owid. (2021, May 5). *owid/covid-19-data*. Ourworldindata. https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/us_state_vaccinations.csv

Owid. (n.d.). *owid/covid-19-data*. Owid vaccination. https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations

Reinhart, Alex and Tibishari, Ryan.. (2020, August 26). *COVID-19 Symptom Surveys through Facebook.* Carnegie Mellon DELPHI Group. https://delphi.cmu.edu/blog/2020/08/26/covid-19-symptom-surveys-through-facebook/

SafeGraph. (2021). *Social distancing metrics*. SafeGraph. https://docs.safegraph.com/docs/social-distancing-metrics

United States Department of Labor. *Average week hours of all employees, total private, seasonally adjusted*. BLS Beta Labs. https://beta.bls.gov/dataViewer/view/timeseries/CES0500000002

United States Department of Labor. *Local area unemployment statistics*. BLS Beta Labs. https://beta.bls.gov/dataViewer/view/613c964322424efea9e49069abe335e8

United States Department of Labor. *Unemployment rate: Illinois*. BLS Beta Labs. https://beta.bls.gov/dataViewer/view/timeseries/LASST170000000000003