

# IS 597 COVID-19 Project

Amrutha Kumaran, Erick Li, Kerstin Wolf, Sai Jayanth

# Research Questions

1. After vaccination in United States began how did the infection trend move? In which states the vaccination program is more advanced?
2. How does the Illinois bar visits statistically related to IL COVID-19 daily case numbers?
3. How did the pandemic impact employment, hours, earnings, etc. in 2020?
4. Are rates of “COVID-like illness” in the community predictive of COVID-19 cases by county in Illinois recorded by the CDC?

# Research Question 1

# Research Question 1

## *Covid Vaccination Progress*

Covid-19 vaccination will be an important tool that could help stop pandemic. Based on the evidence it is states that fully vaccinated people are less likely to get infected and less likely to spread the virus that causes Covid-19 to others. This would help in reducing number of hospitalization and death. Potentially leading to normal lives.

- Vaccination started December 2020.
  - Is vaccination effective vs daily new cases?
  - In which states vaccination program is more advanced?

# Datasets

- [United States COVID-19 Cases and Deaths by State over Time](#): Data.Gov or CDC
- [US State Vaccination](#): Our World in Data

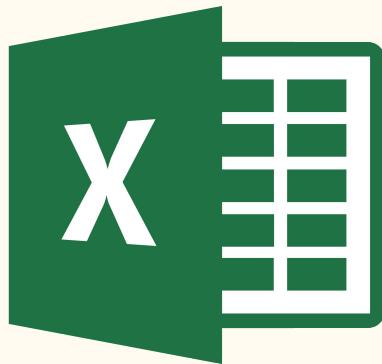


# Data PreProcessing

- Data Cleaning
  - NA
  - Rename state name in code format to full name.
- Data Blending
  - Both Datasets were merged by state name and date using Inner Join

# Tools

- Tableau
- Excel



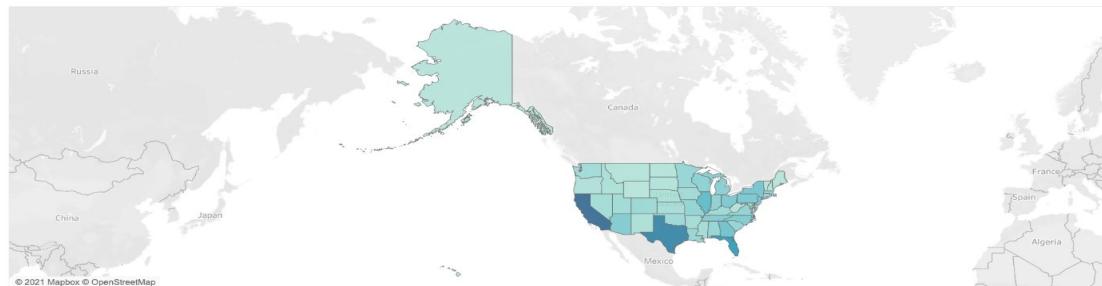
# Preliminary Results - Infection Trend

Tot Cases

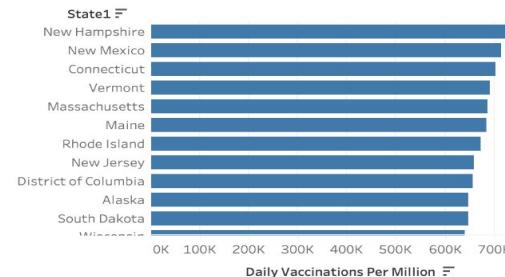
2,156,456

528,720,690

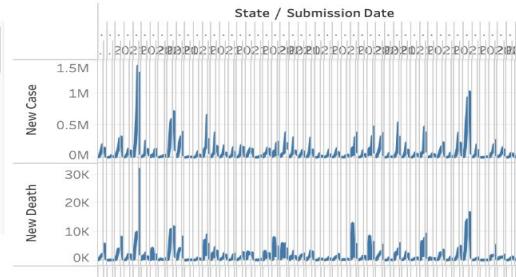
Geomap by Total Cases



DailyVaccination/Million



New Cases

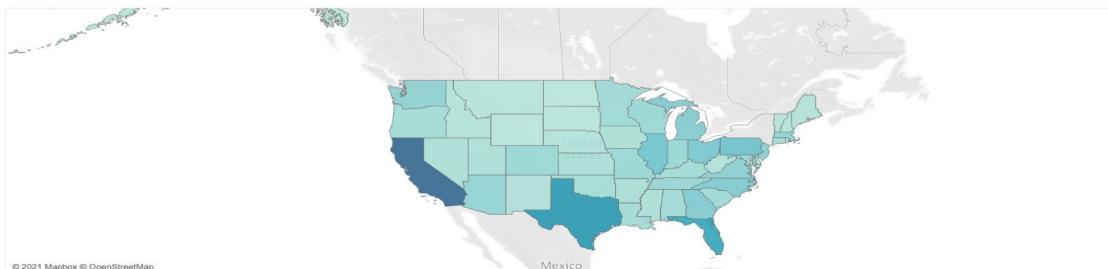


# Preliminary Results - Top Performing Vaccination

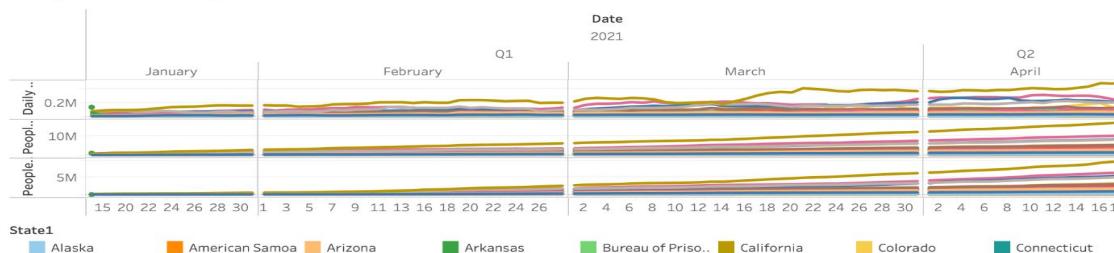
Total Distributed  
468,615

32,618,010

Total Distributed Vaccines



DailyVaccinationByState



# Research Question 2

# Datasets

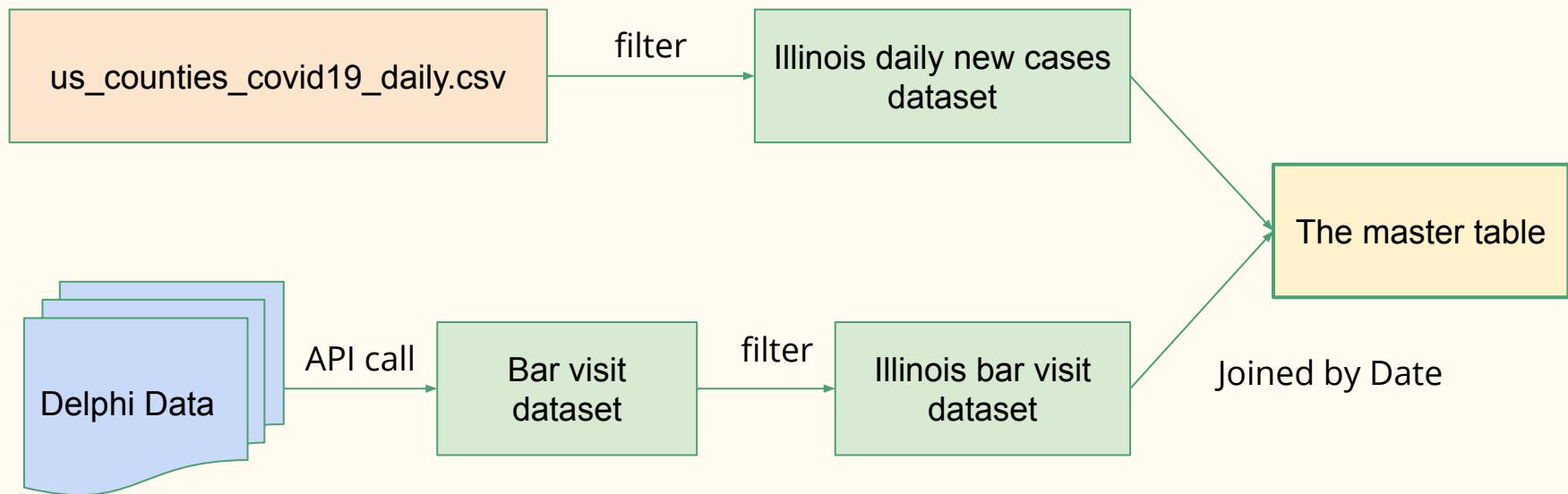
- Daily Case Number
  - From a Kaggle Dataset called “COVID-19 in USA”
  - <https://www.kaggle.com/sudalairajkumar/covid19-in-usa>
  - Contains state level data including number of new cases and deaths
  - Data range from Jan. 2020 to Dec. 2020
- Bar Visit Data
  - From the Carnegie Mellon Delphi API SafeGraph
  - <https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/safegraph.html#safegraph-social-distancing-metrics>
  - Data collected by SafeGraph via anonymous mobile phone locations
  - Contains daily number of bar visits per 100,000 population on the state level

# Data Preprocessing

- Data cleaning
  - NA
  - Bad format
- Data merging
  - Merge by date
  - Outer join
- Tools
  - PySpark pipeline
  - Python

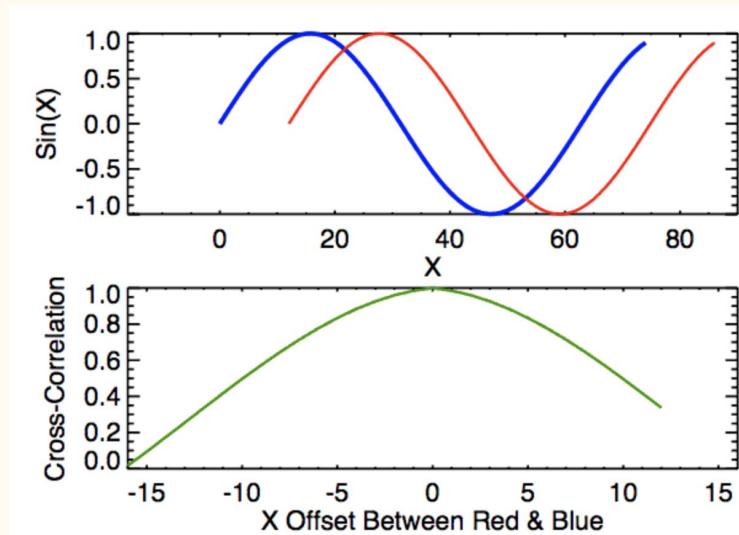


# Data Pipeline



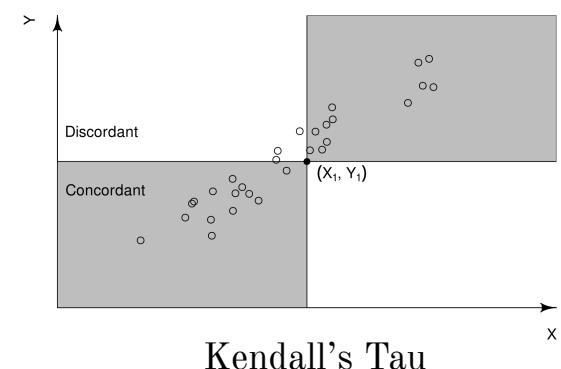
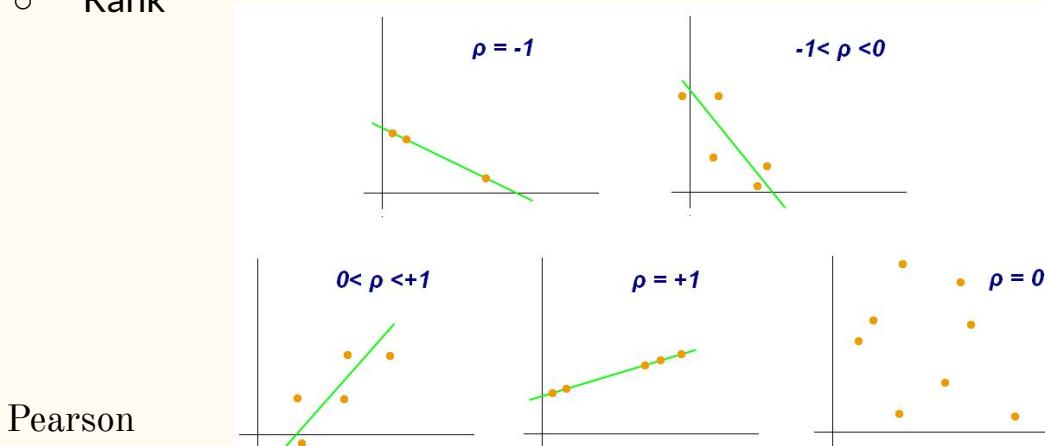
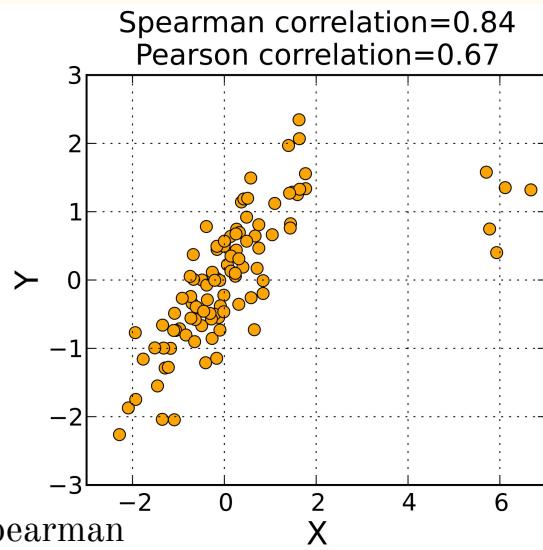
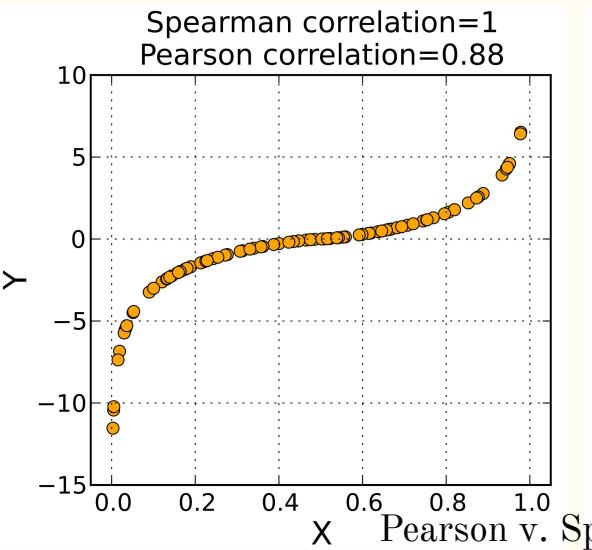
# Time Lagged Cross Correlation (TLCC)

- Show how the two time-series data are statistically correlated with each other with different lags
- Discover delayed impacts
- Measures
  - Pearson
  - Spearman
  - Kendall's Tau

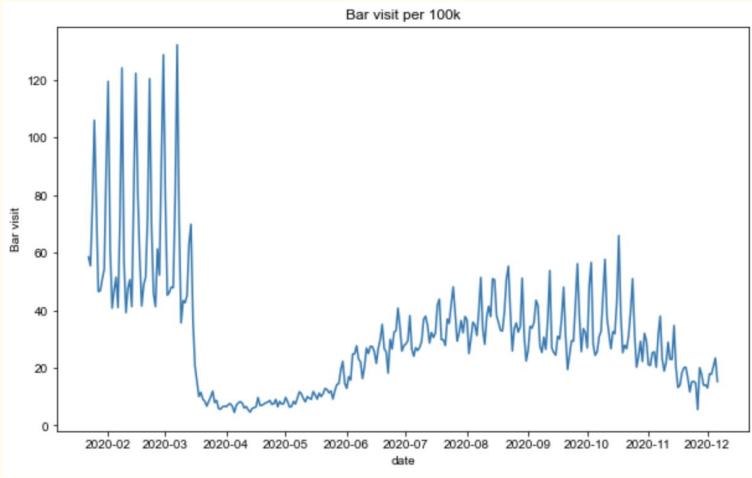


# Metrics

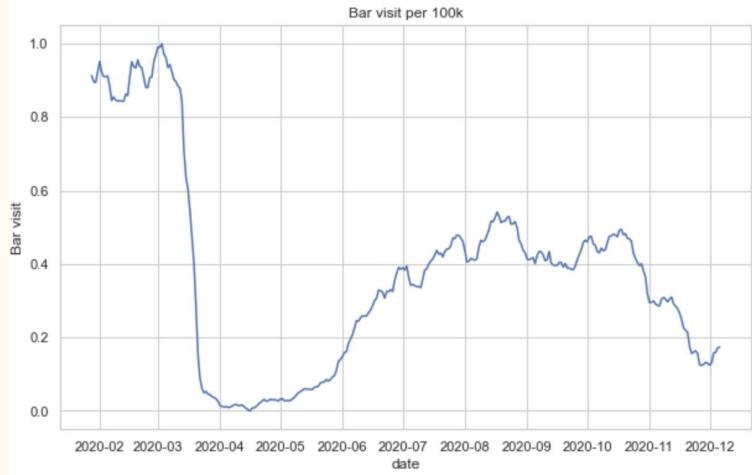
- Pearson Correlation
  - Linear
- Spearman Correlation
  - Rank
- Kendall's Tau
  - Rank



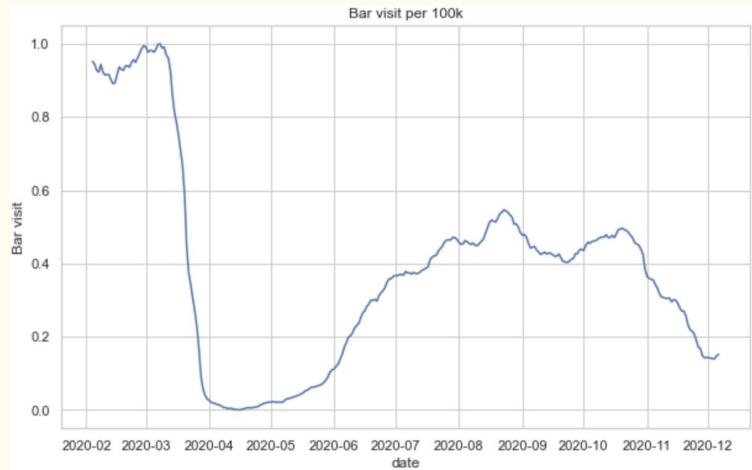
# Moving Average



Without MA



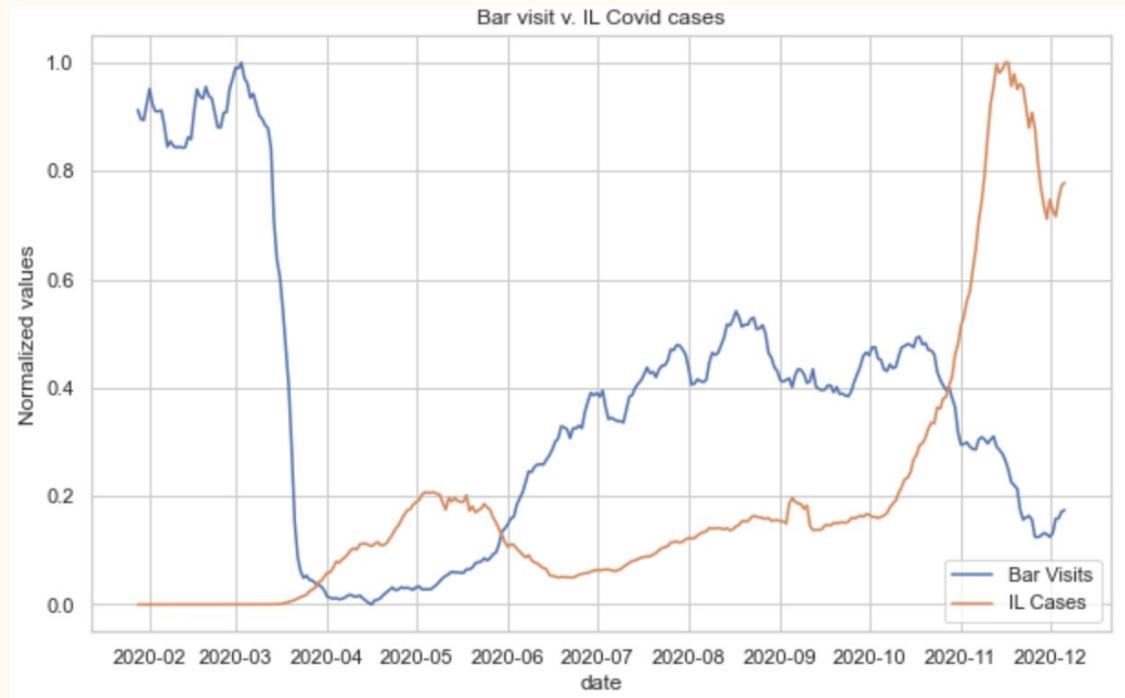
7d



14d

# Normalization

- Max-min normalization
  - Max: 1
  - Min: 0
- Good for comparison between two variables with different scales



# Cases

- Bar visits to cases, 7-day moving average
- Bar visits to cases, 14-day moving average
- Cases to bar visits, 14-day moving average

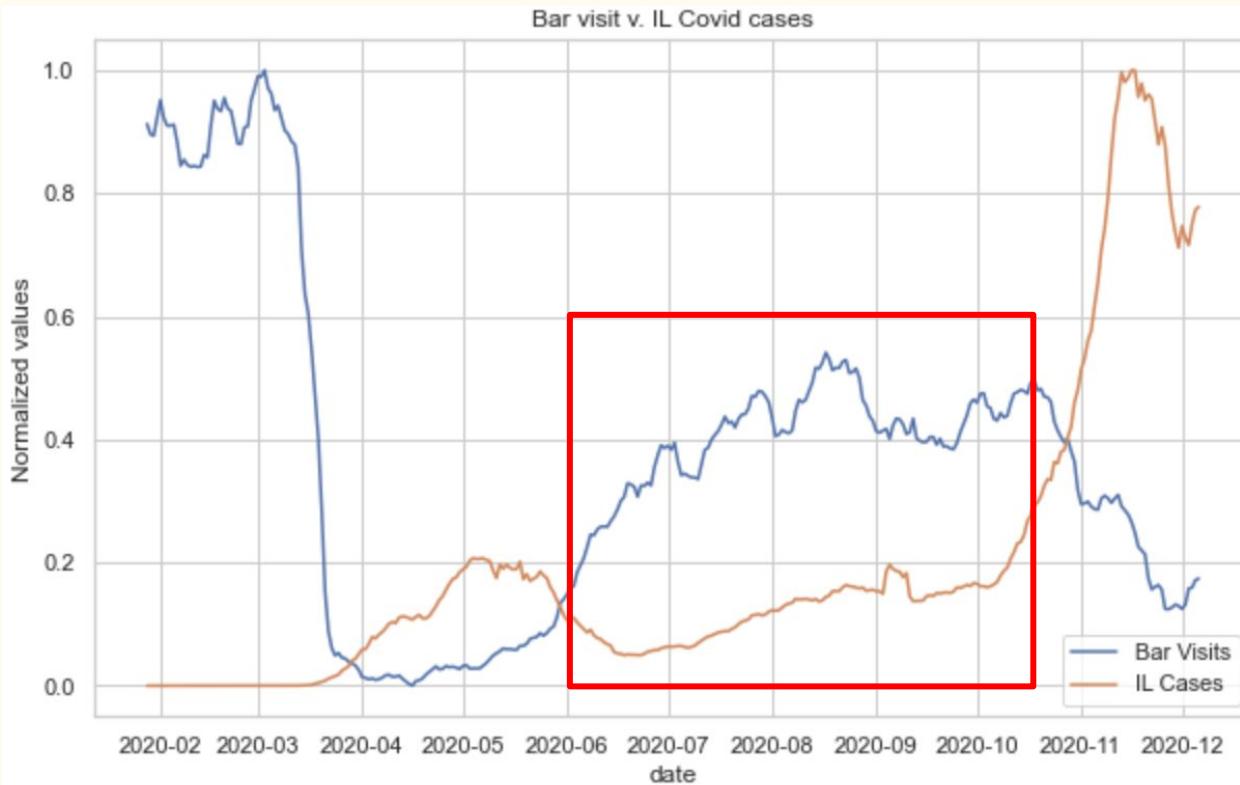
# Bar Visit to IL Covid Cases

- Questions to be answered
  - How is bar visit related to IL Covid Cases?
  - Was there an increase of bar visits followed by a case number surge in Illinois?
- Focus on the time range where
  - Bar visit was increasing, and
  - IL case number was growing

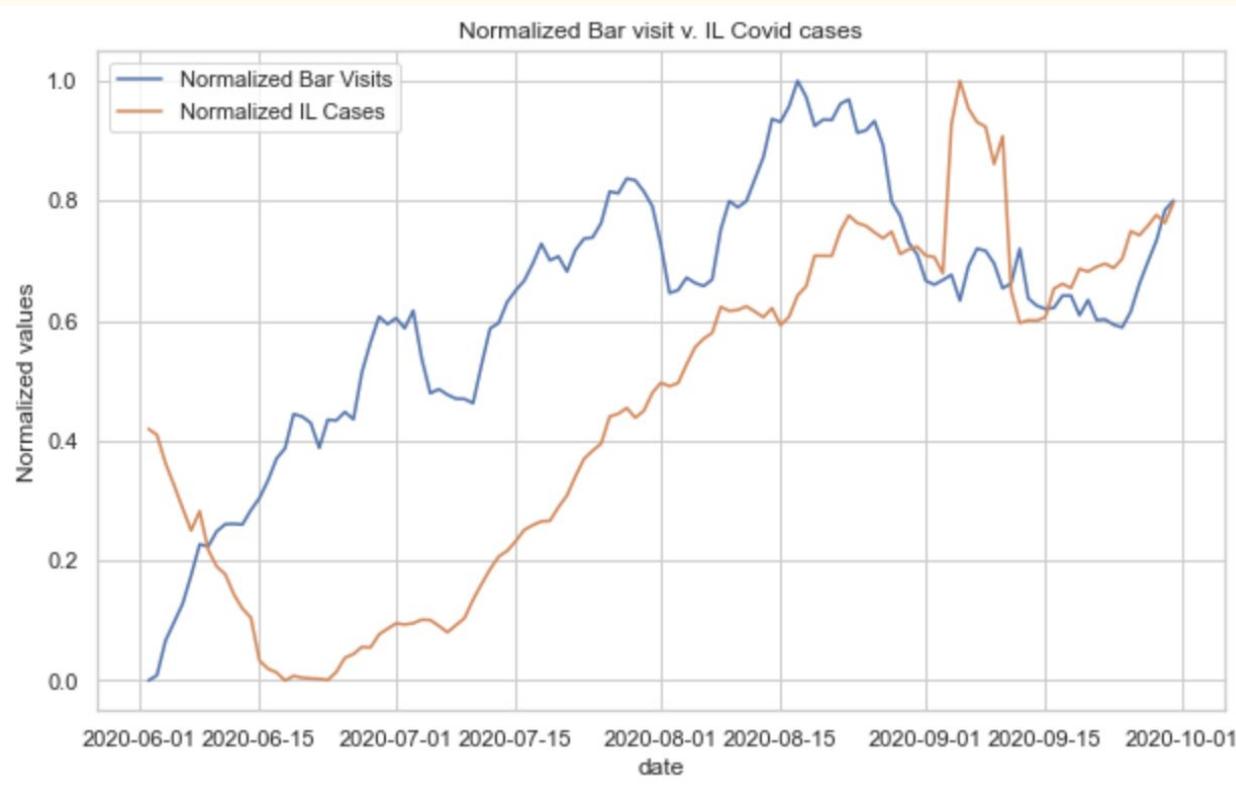
# IL Covid Cases to Bar Visits

- Questions to be answered
  - How did IL Covid Cases affect bar visits?
  - Was there a drop of bar visits after a case number surge in Illinois?
- Focus on the time range where
  - Bar visit was decreasing, while
  - IL case number was growing

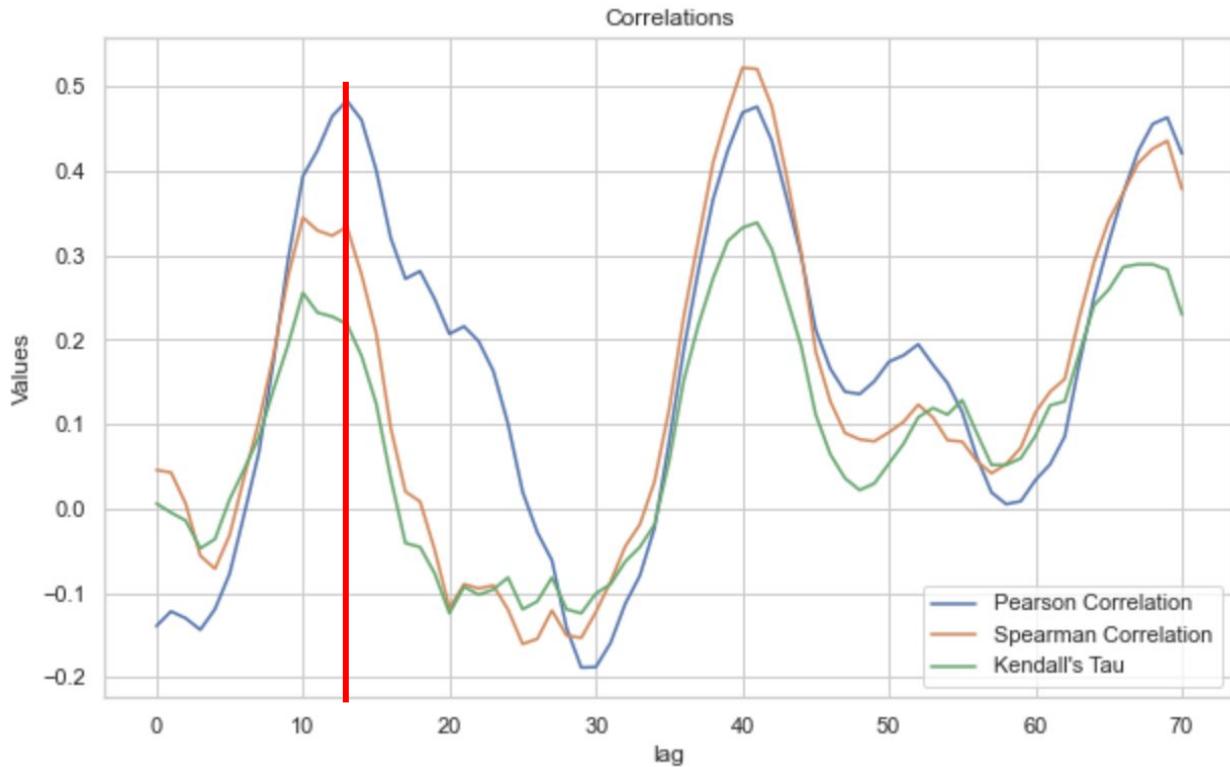
# Bar Visit to IL Covid Cases - 7d MA



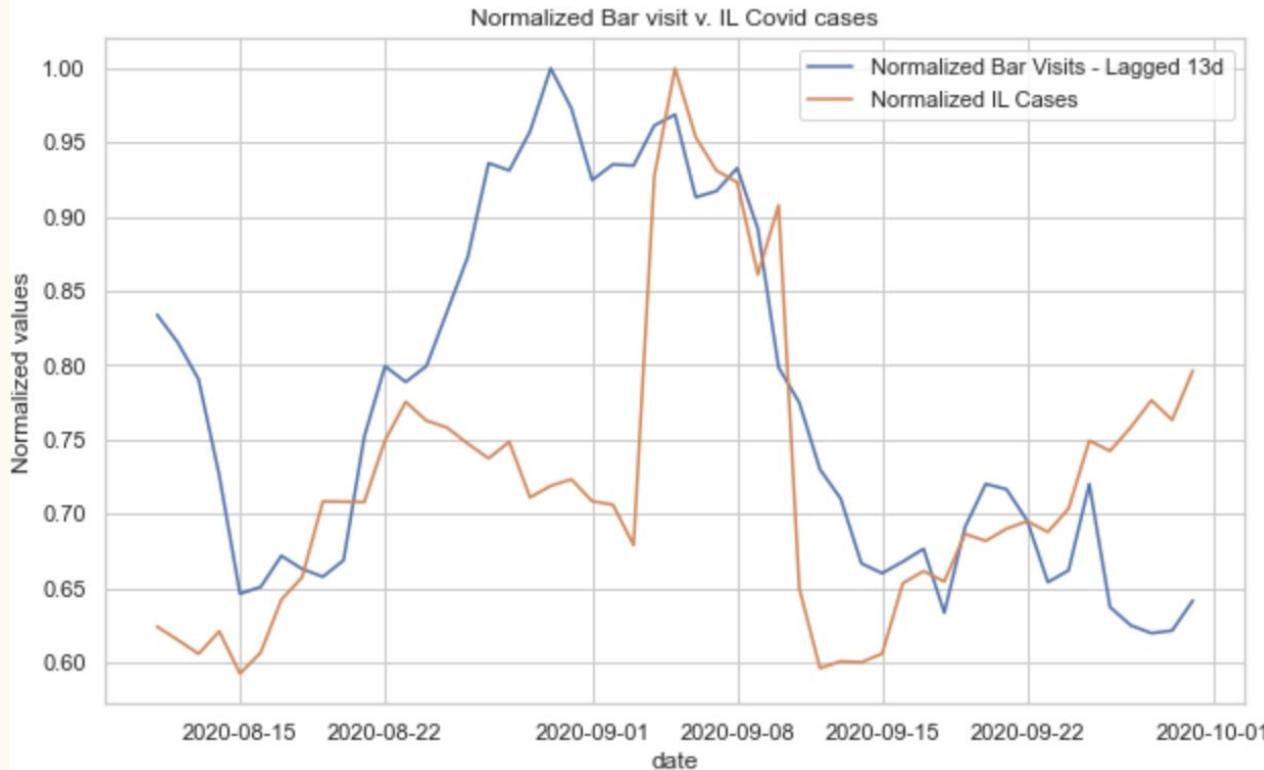
# Bar Visit to IL Covid Cases (June-Mid Oct) - 7d MA



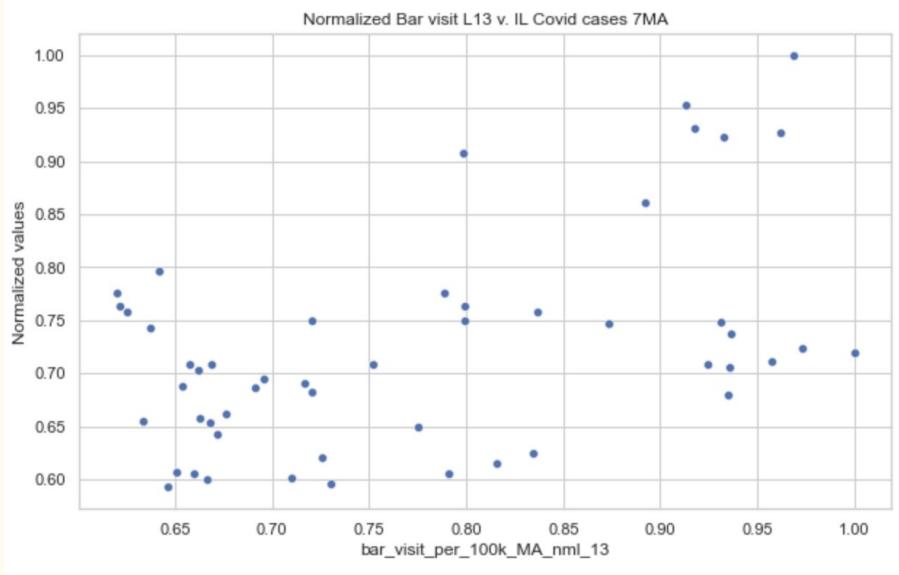
# TCLL - 7d MA



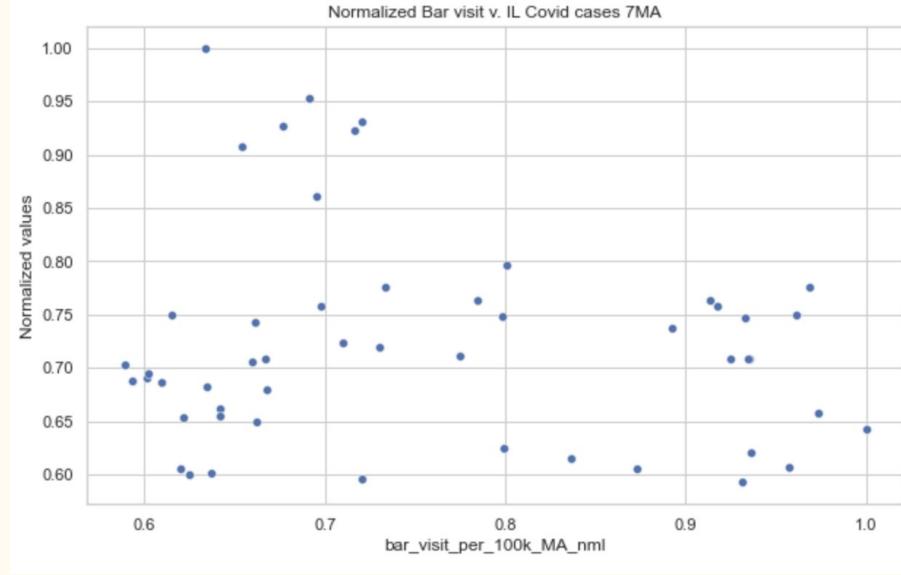
# Lag - 13 days - 7d MA



# Scatterplots - 7d MA

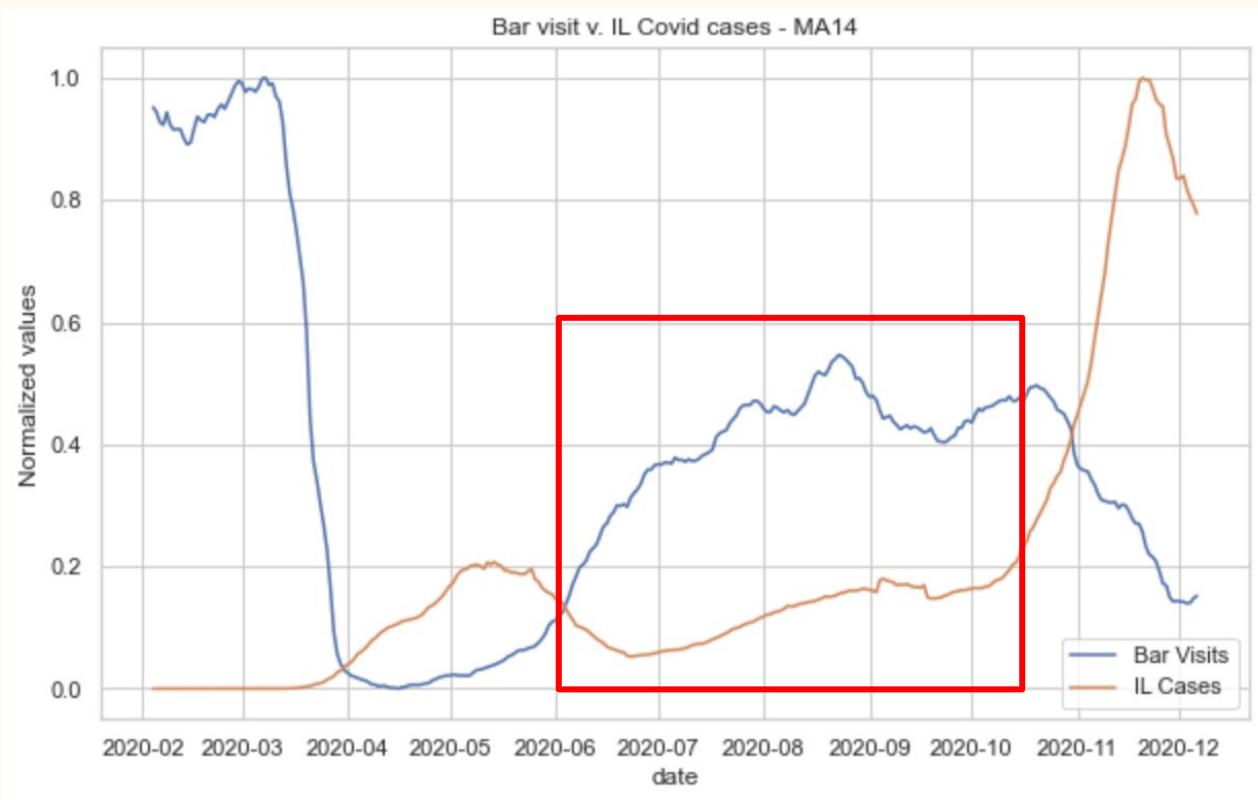


13-day lag

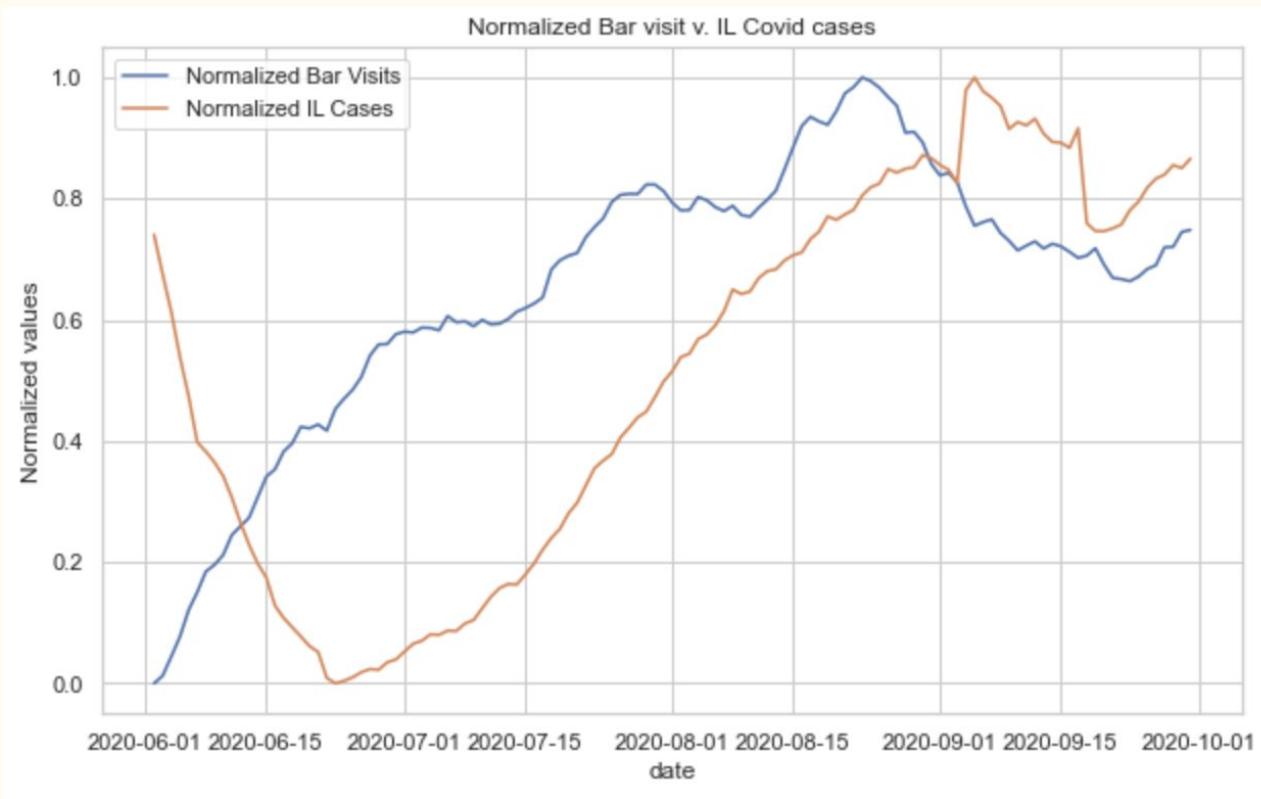


No lags

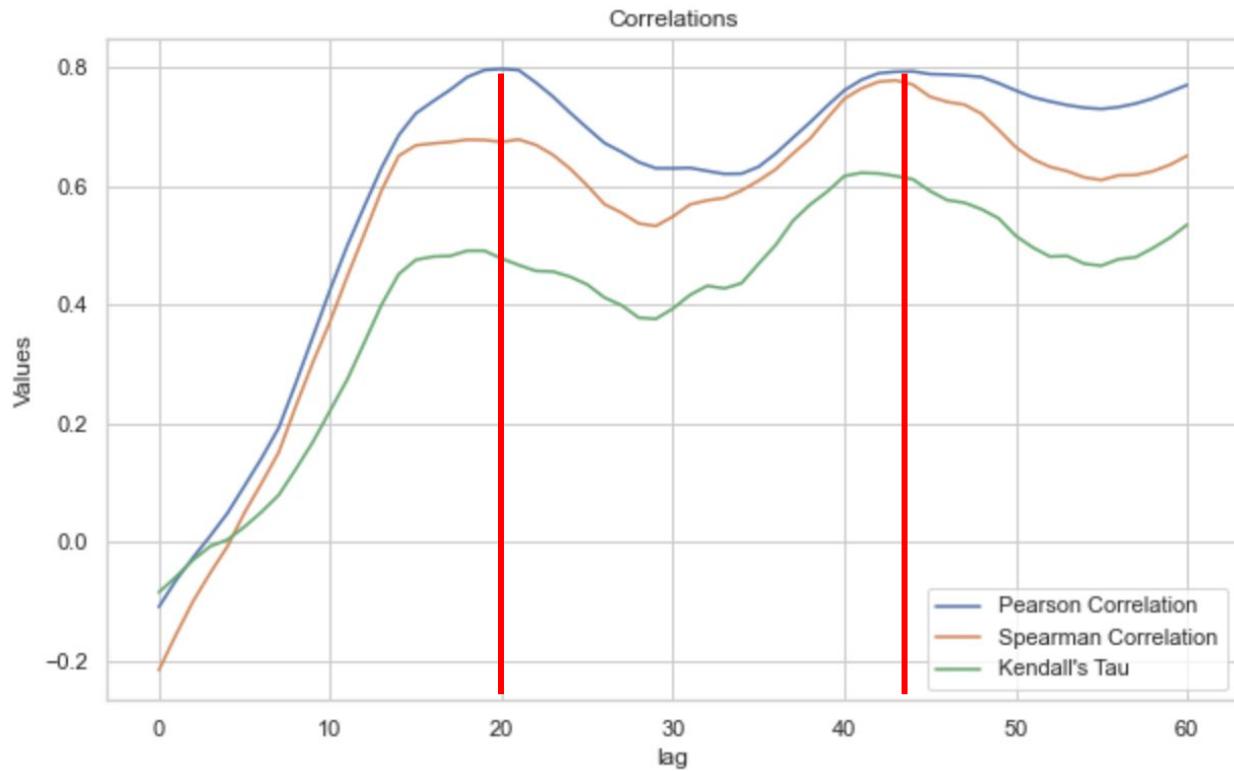
# Bar Visit to IL Covid Cases - 14d MA



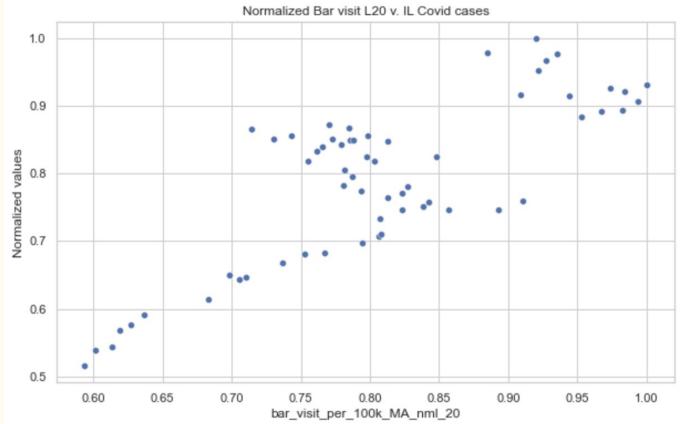
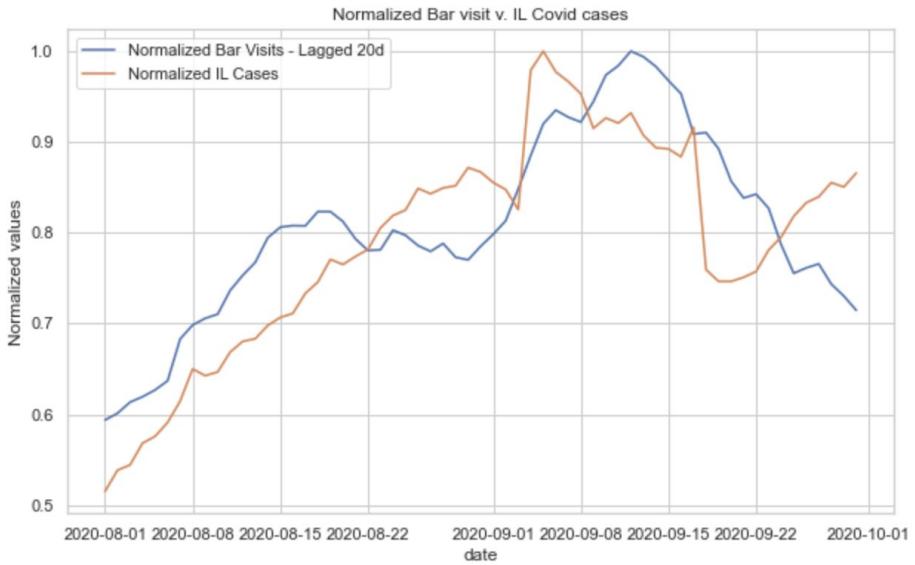
# Bar Visit to IL Covid Cases (June-Mid Oct) - 14d MA



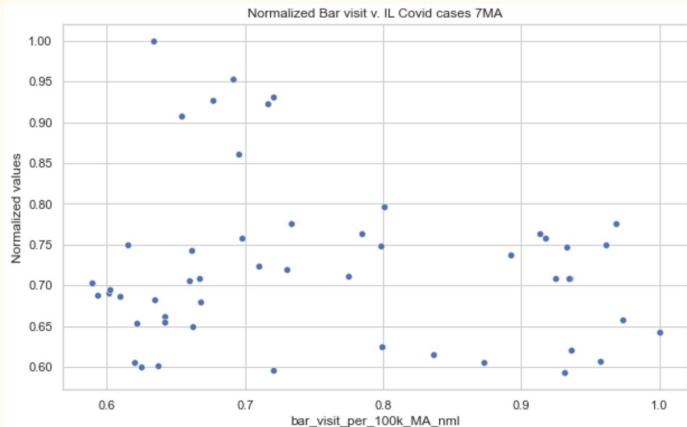
# TCLL - 14d MA



# Lag - 20 days - 14d MA

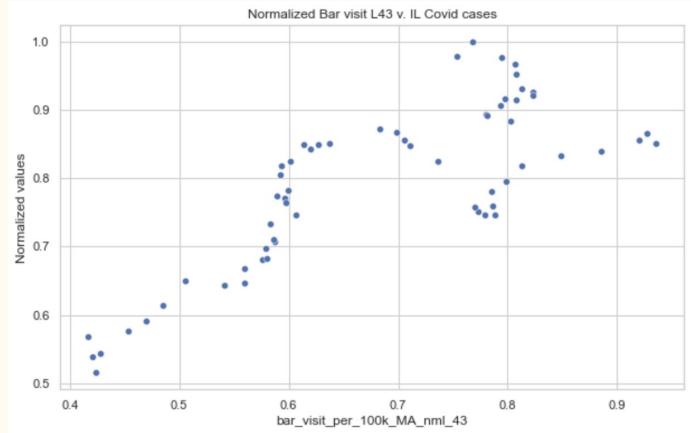
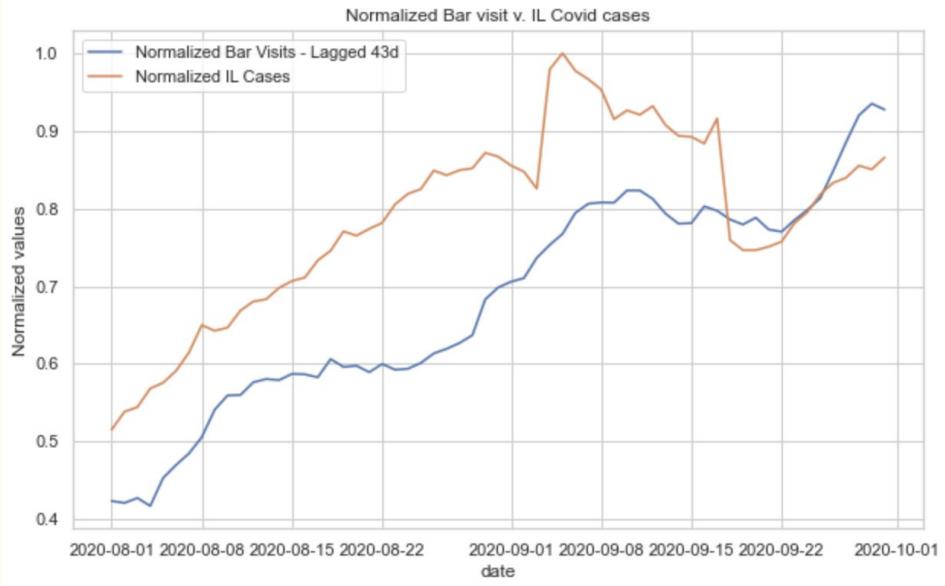


With a lag

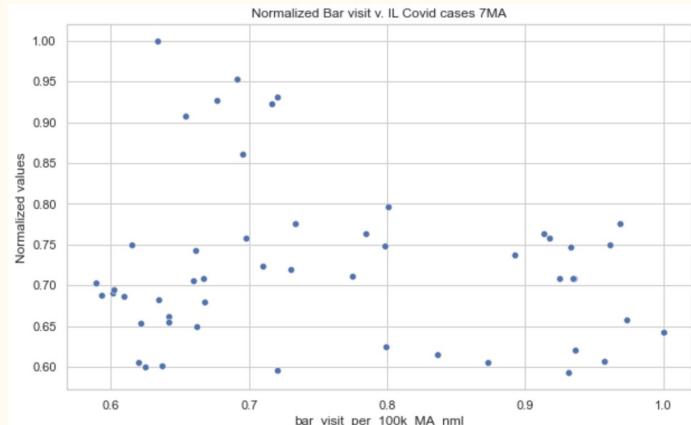


No lags

# Lag - 43 days - 14d MA

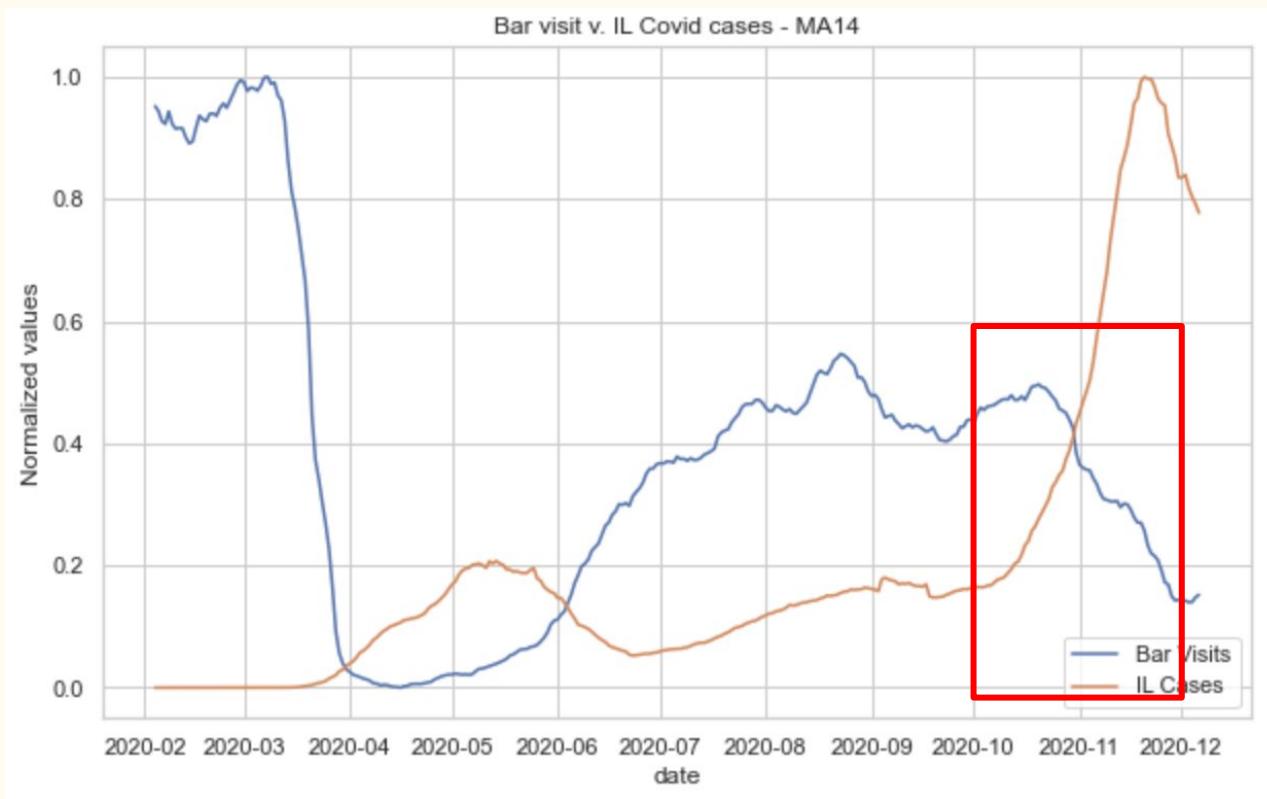


With a lag

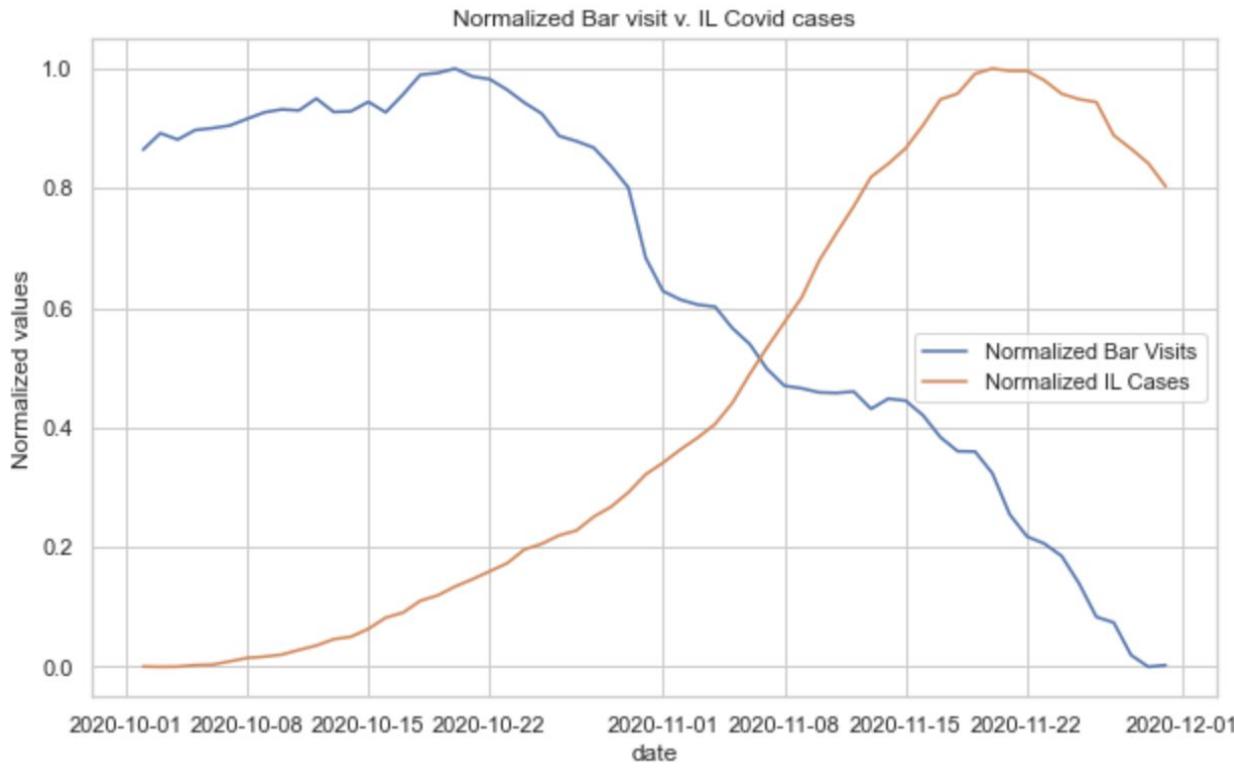


No lags

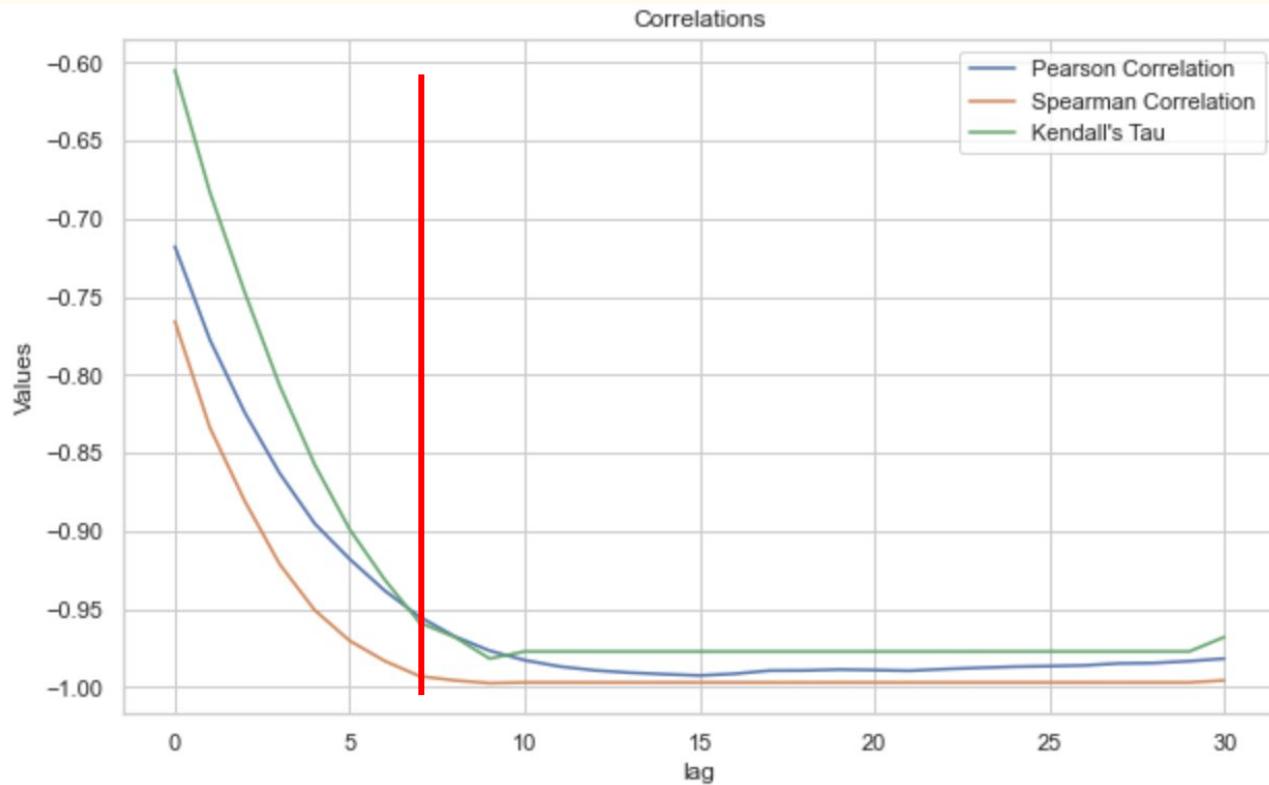
# IL Covid Cases to Bar Visits - 14d MA



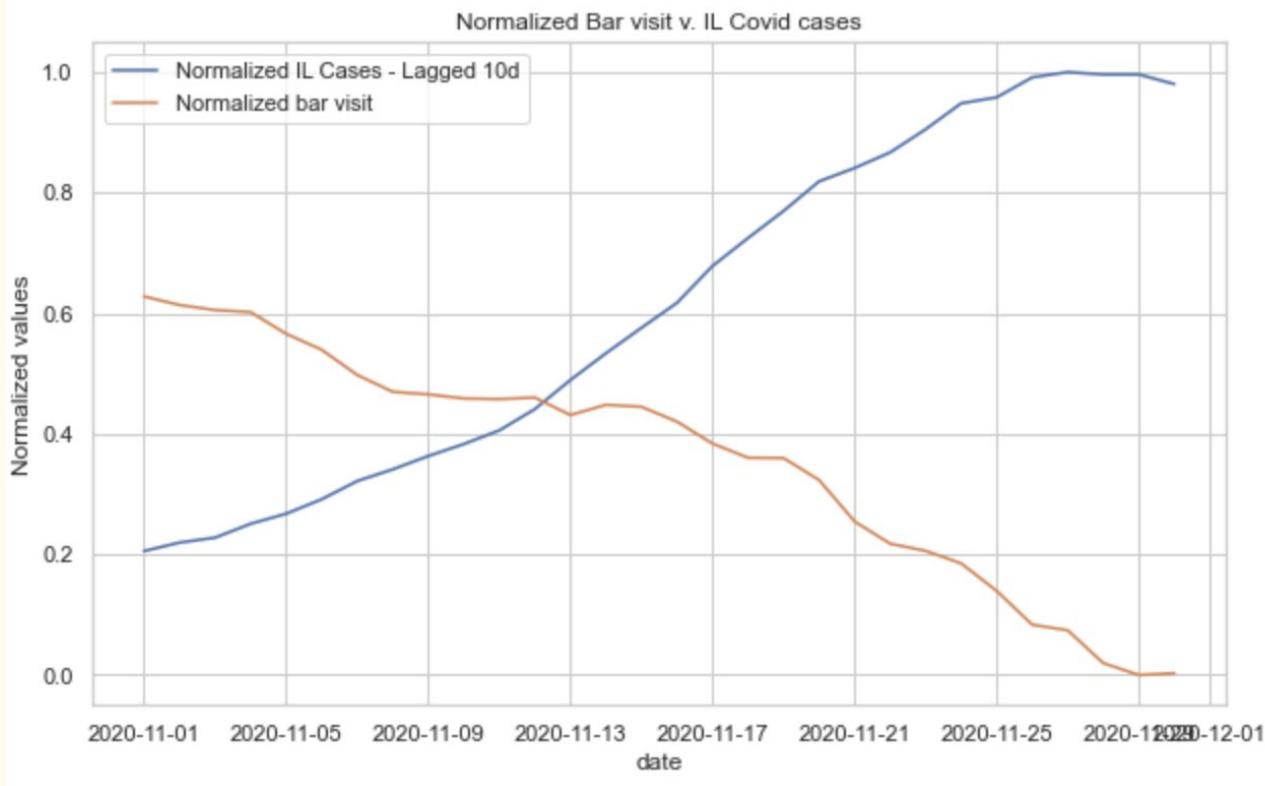
# IL Covid Cases to Bar Visits (Oct - Dec) - 14d MA



# TCLL - 14d MA



# Lag - 7 days - 14d MA



# Summary

- High correlations between the IL bar visits and IL Covid daily case number
  - 7-day moving average: 13 days
  - 14-day moving average: 20 days
- High correlations between the IL Covid daily case number and IL bar visits
  - 14-day moving average: 7 days
- 14-day MA is smoother than 7-day MA
- People visiting bar are more quickly affected by the Covid surge
  - Local government's regulation
  - People can control whether they want to visit a bar

# Research Question 3

# Datasets

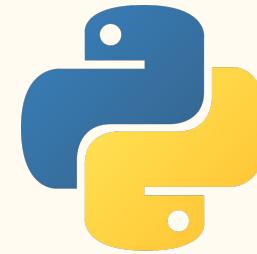
- Centers for Disease Control and Prevention (CDC)
  - [United States COVID-19 Cases and Deaths by State over Time](#)
  - [Provisional COVID-19 Death Counts by Week Ending Date and State](#) (data provided by the NCHS)
  - [Illinois Pneumonia Deaths in 2019](#)
- U.S. Bureau of Labor Statistics (BLS)
  - [Illinois Unemployment Rate 2019-2020](#)
  - [State Unemployment Rate 2019-2020](#)
  - [Weekly Hours of US Employees in the Private Sector in 2019-2020](#)

# Datasets (Continued)

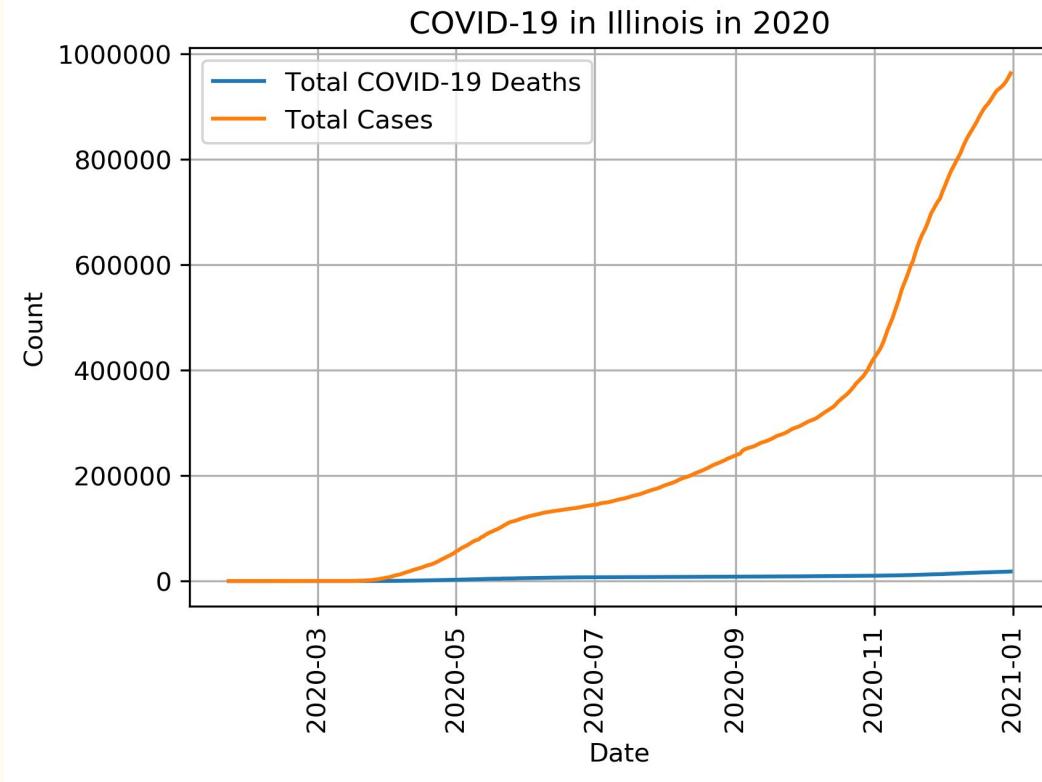
- Carnegie Mellon Delphi Epidata API (covidcast)
  - [SafeGraph](#)
    - 7-Day Average of the Fraction of People Who Did Not Leave Home
    - 7-Day Average of the Fraction of People Working Full Time Outside the Home
    - 7-Day Average of the Fraction of People Working Part Time Outside the Home

# Tools for Data Visualization

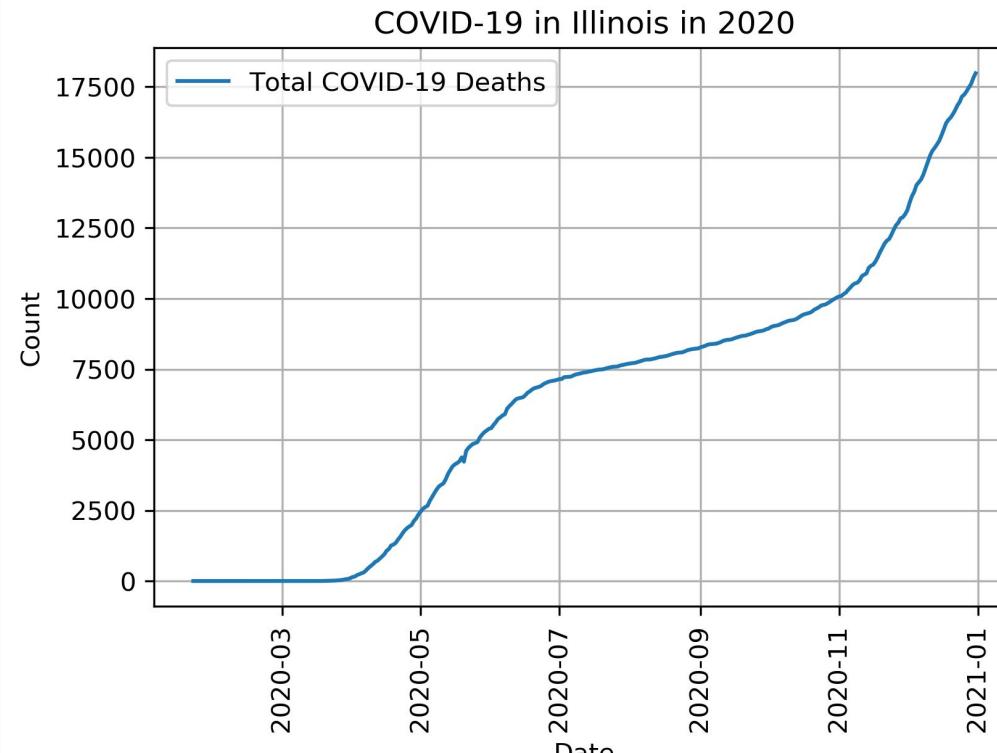
- Python
  - Matplotlib
- Tableau



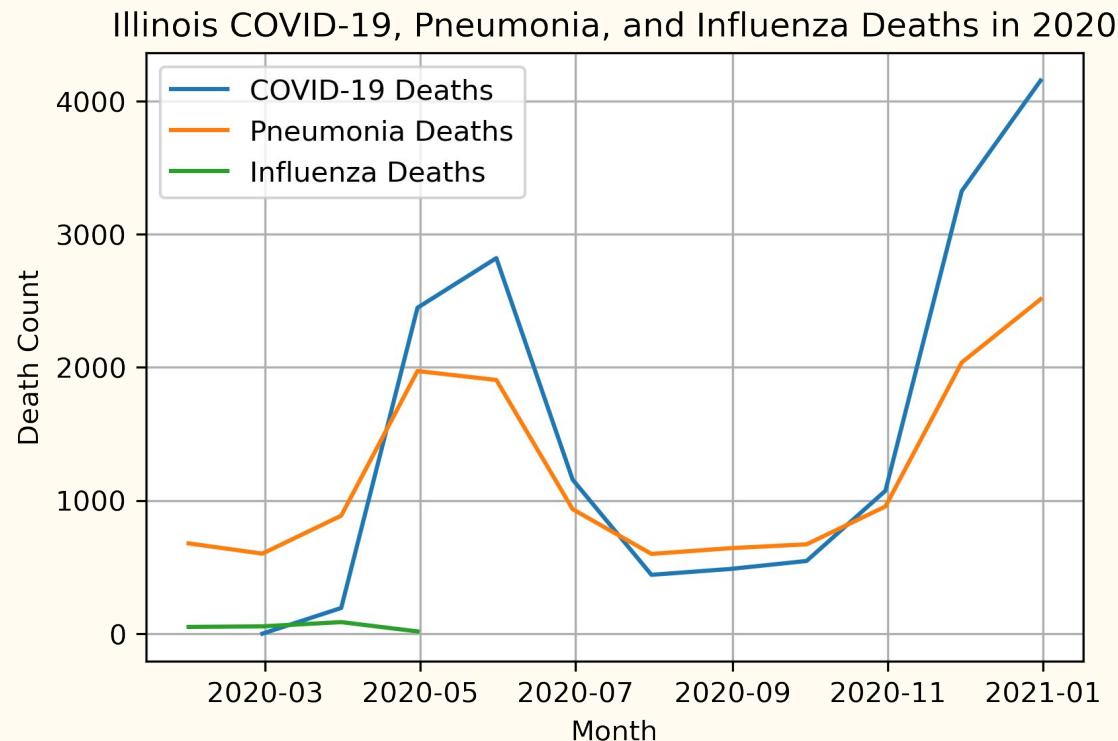
# Illinois COVID-19 Cases and Deaths in 2020



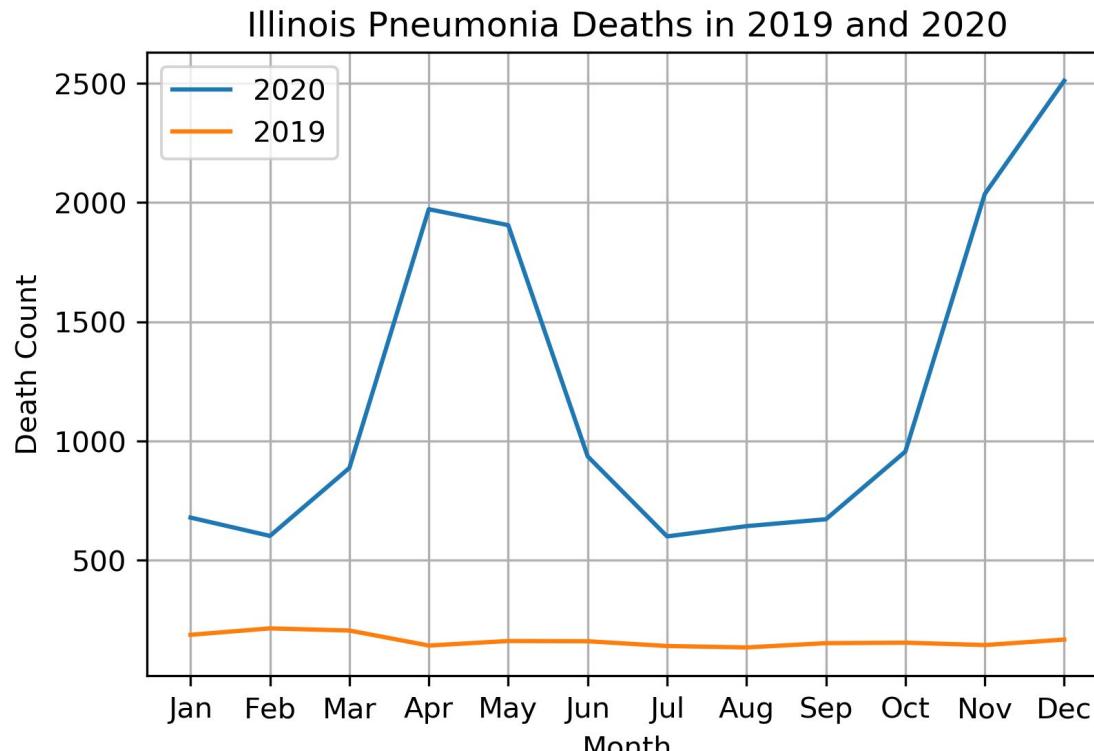
# Illinois COVID-19 Deaths in 2020



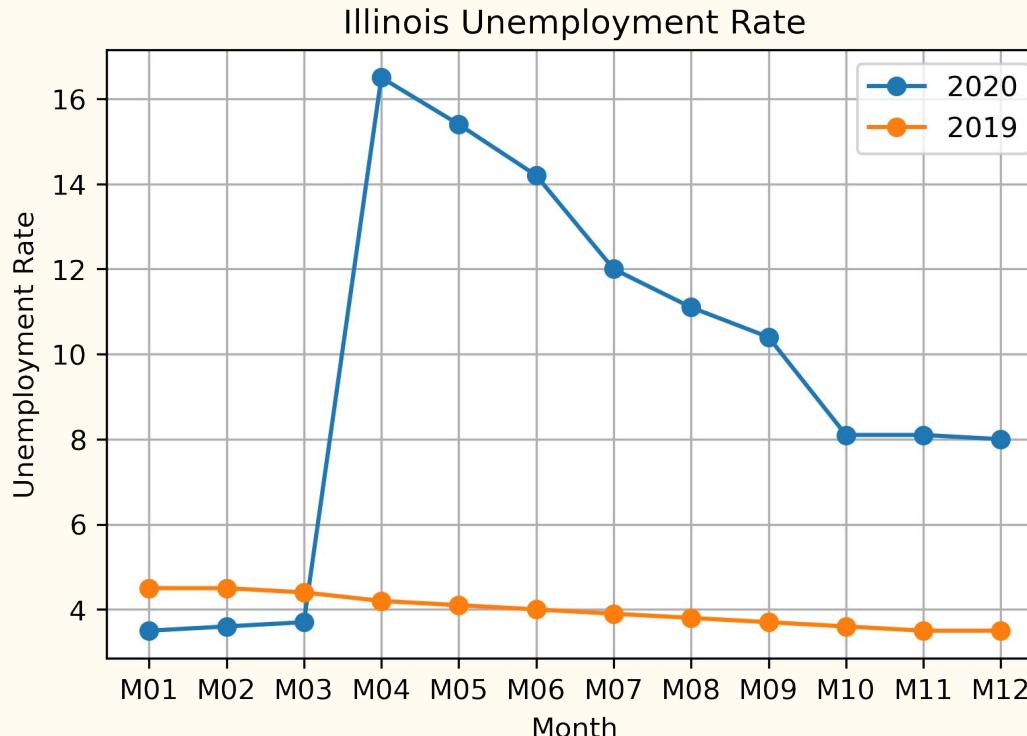
# Illinois COVID-19, Pneumonia, and Influenza Deaths in 2020



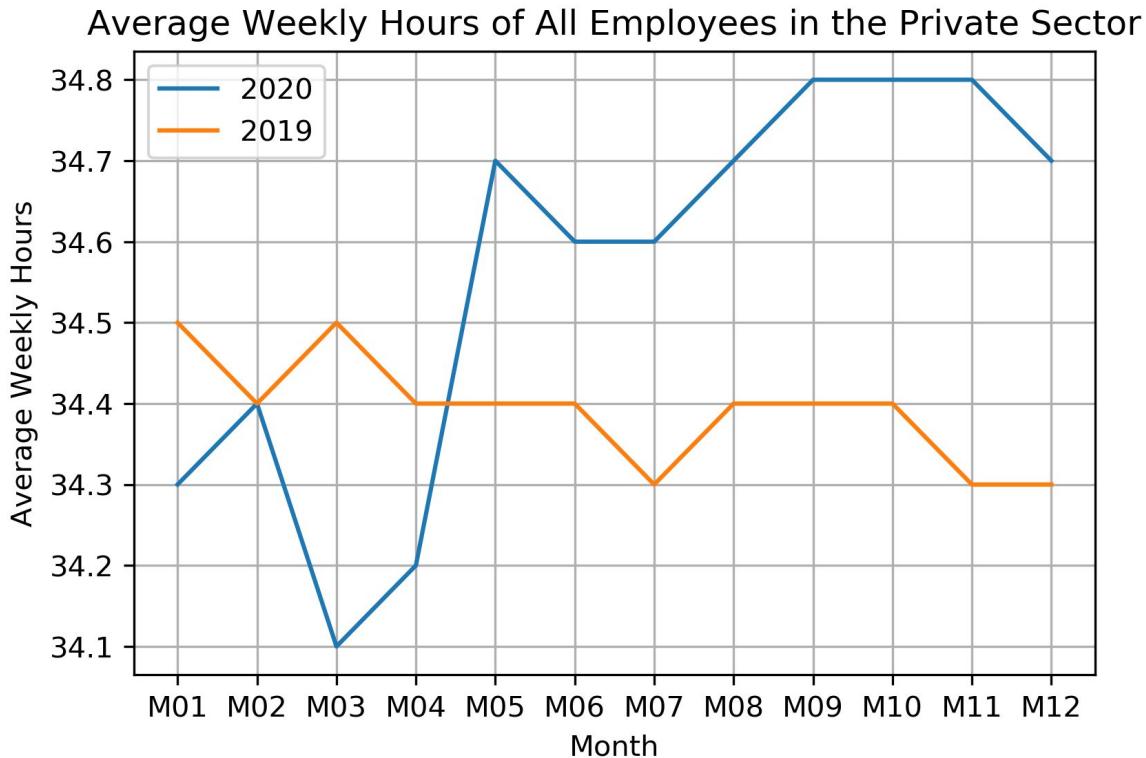
# Illinois Pneumonia Deaths in 2019 and 2020



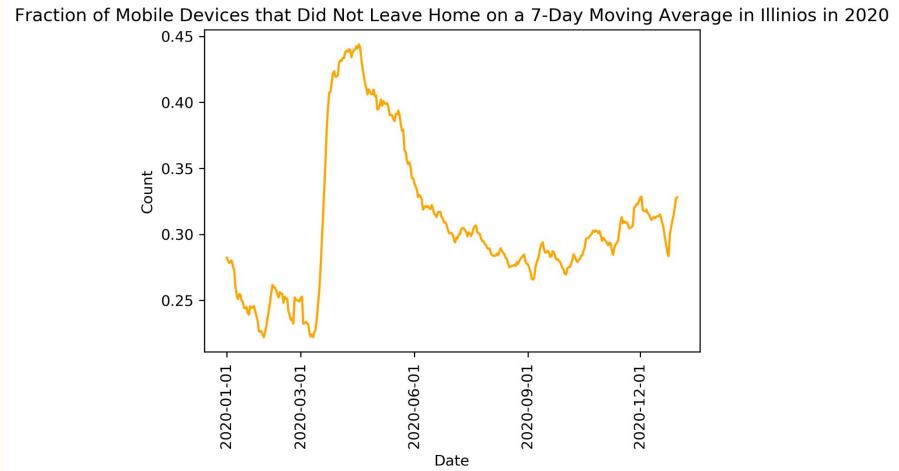
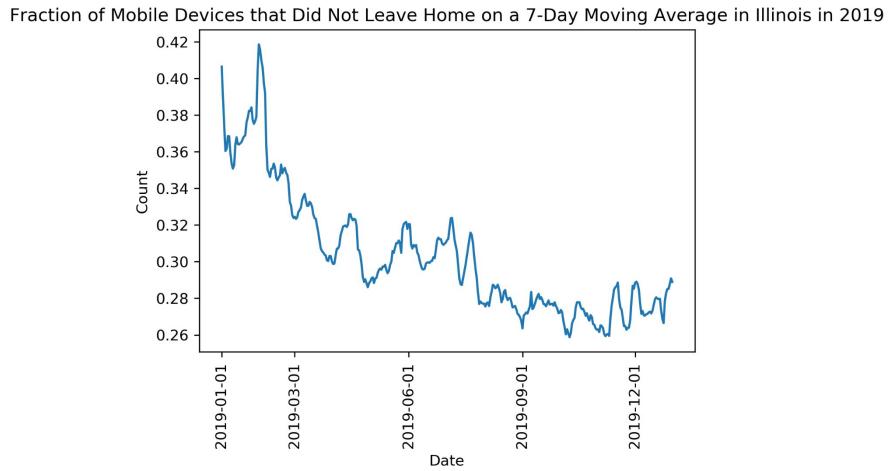
# Illinois Unemployment Rate in 2019 and 2020



# National Average Weekly Work Hours

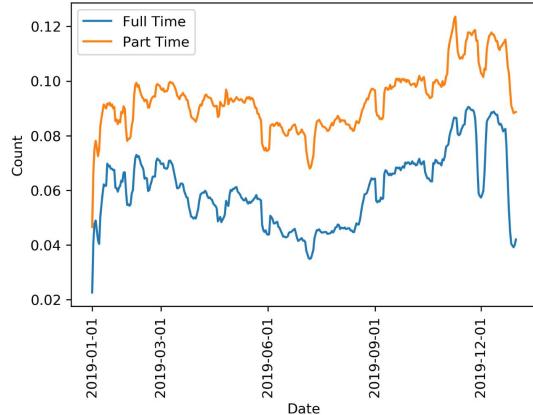


# Leaving the House in 2019 and 2020

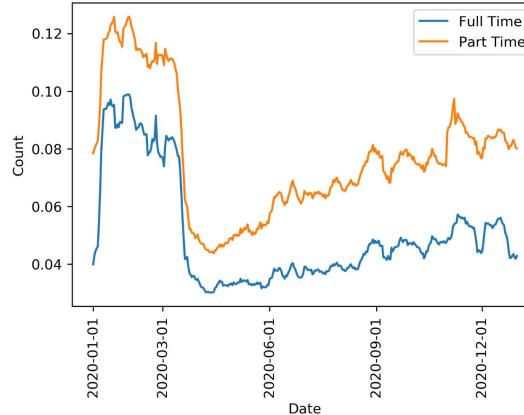


# Full Time and Part Time in 2019 and 2020

People Working Possible Full Time and Part Time Jobs Requiring Leaving the House in Illinois in 2019



People Working Possible Full Time and Part Time Jobs Requiring Leaving the House in Illinois in 2020



# Results

- Illinois cases and deaths began rising suddenly in April 2020.
- A number of pneumonia deaths in 2020 may actually be COVID-19 deaths or COVID-19 related deaths.
- COVID-19 was likely moving through Illinois before cases were being reported here.
- Unemployment rate spikes suddenly in March going into April 2020
- The national average weekly work hours drops significantly in February 2020 before rebounding and increasing significantly through March and April and topping out at 34.8 hours in September.
- There's a spike in people in Illinois staying home in March, but this somewhat quickly decreases.
- Around March and April in 2020, there is decrease in people leaving their houses for long stretches of time possibly due to work.
- Those potentially working part-time jobs return to work at a faster rate than those potentially working full time jobs.

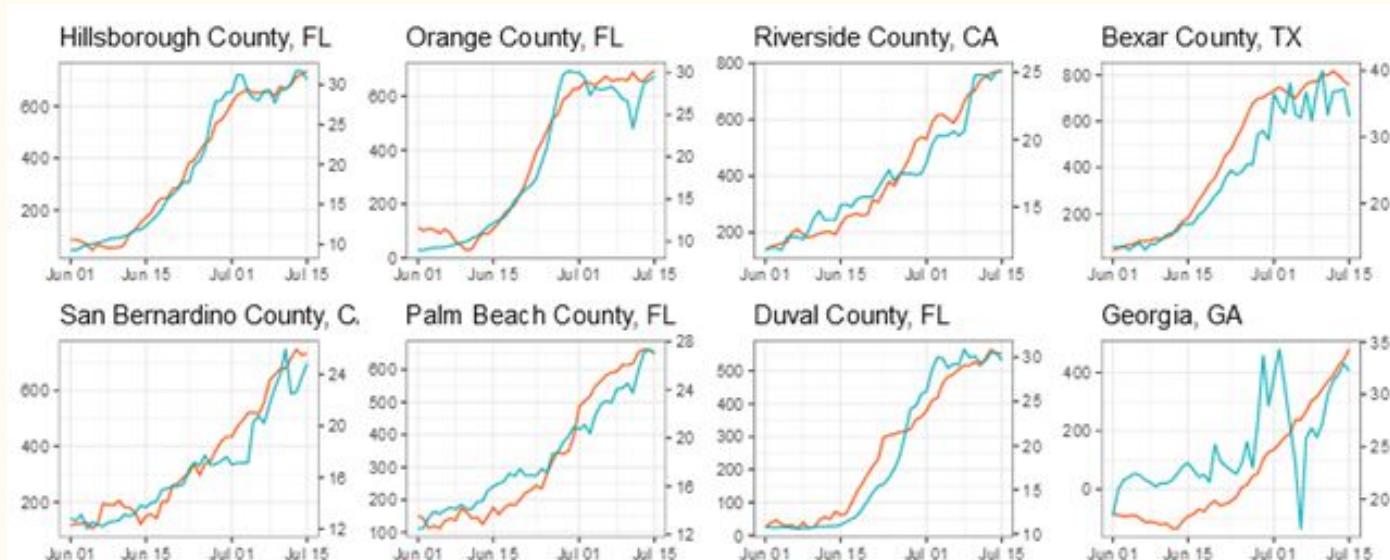
# Research Question 4

# Research Question 4

- How are survey responses of COVID-like illness in the community related to COVID-19 cases?

# COVID-like Illness in the Community (CLI)

- Defined as the percentage of FB survey respondents that are reporting symptoms of fever AND at least one of the following: along with cough, shortness of breath, or difficulty breathing **in their community**



# Data Sources

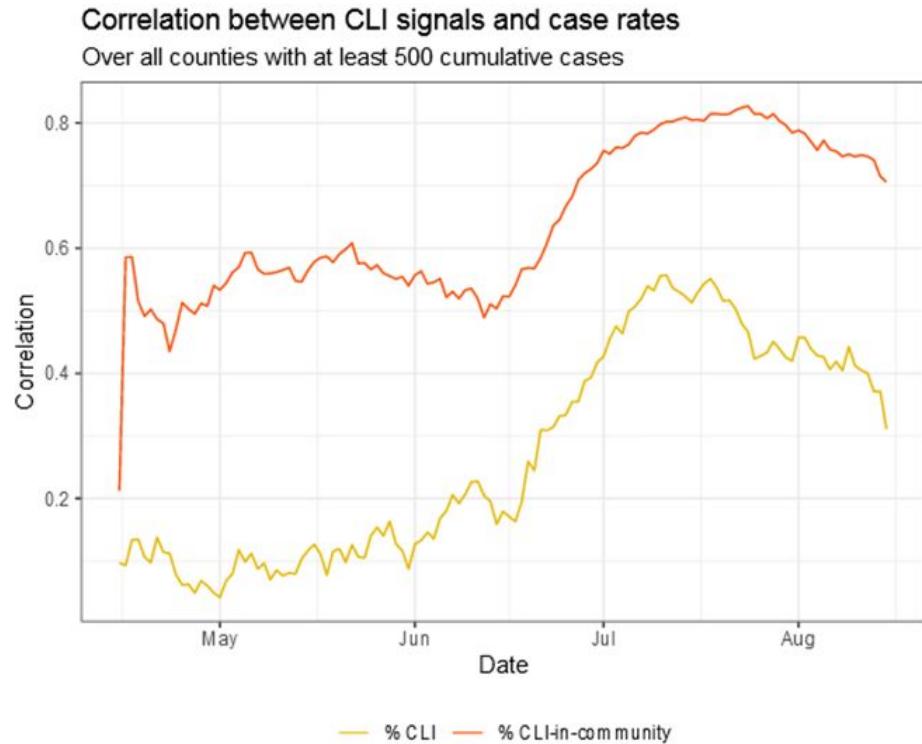
- John Hopkins' university case reports from local/state government
- Carnegie-Mellon University x Facebook survey questions

	geo_value	signal	time_value	issue	lag	value	stderr	sample_size	geo_type	data_source
0	01000	smoothed_wnohh_cmnty_cli	2020-05-01	2020-09-03	125	19.468615	0.976401	1644.5362	county	fb-survey
1	01001	smoothed_wnohh_cmnty_cli	2020-05-01	2020-09-03	125	7.445426	2.523874	108.1813	county	fb-survey
2	01003	smoothed_wnohh_cmnty_cli	2020-05-01	2020-09-03	125	12.467176	1.398068	558.3194	county	fb-survey
3	01015	smoothed_wnohh_cmnty_cli	2020-05-01	2020-09-03	125	12.263017	3.047784	115.8275	county	fb-survey
4	01031	smoothed_wnohh_cmnty_cli	2020-05-01	2020-09-03	125	21.194579	3.892180	110.2542	county	fb-survey
...	...	...	...	...	...	...	...	...	...	...
909	55141	smoothed_wnohh_cmnty_cli	2020-05-07	2020-09-03	119	9.216453	2.094310	190.7608	county	fb-survey
910	56000	smoothed_wnohh_cmnty_cli	2020-05-07	2020-09-03	119	13.702421	1.426436	581.1541	county	fb-survey
911	56001	smoothed_wnohh_cmnty_cli	2020-05-07	2020-09-03	119	21.755616	4.112058	100.6714	county	fb-survey
912	56021	smoothed_wnohh_cmnty_cli	2020-05-07	2020-09-03	119	16.756208	2.264684	271.9646	county	fb-survey
913	56025	smoothed_wnohh_cmnty_cli	2020-05-07	2020-09-03	119	11.344546	2.100875	227.8725	county	fb-survey

6828 rows × 10 columns

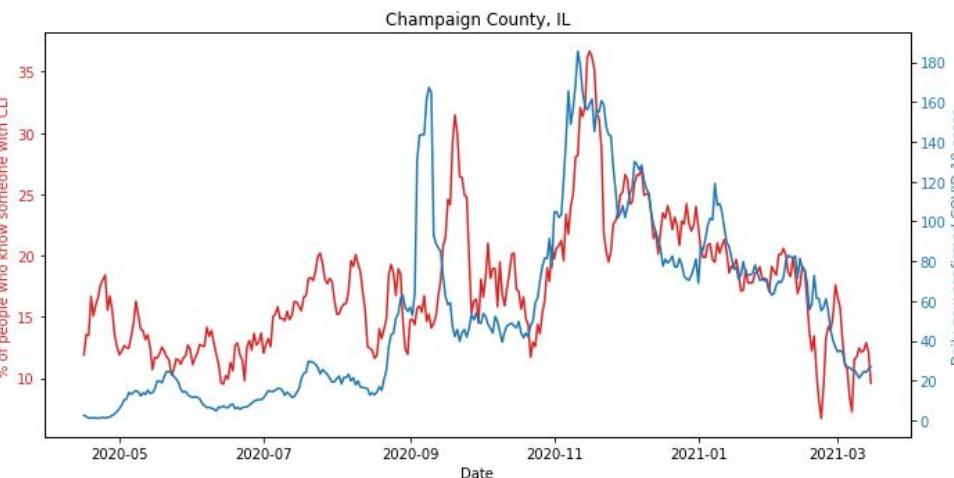
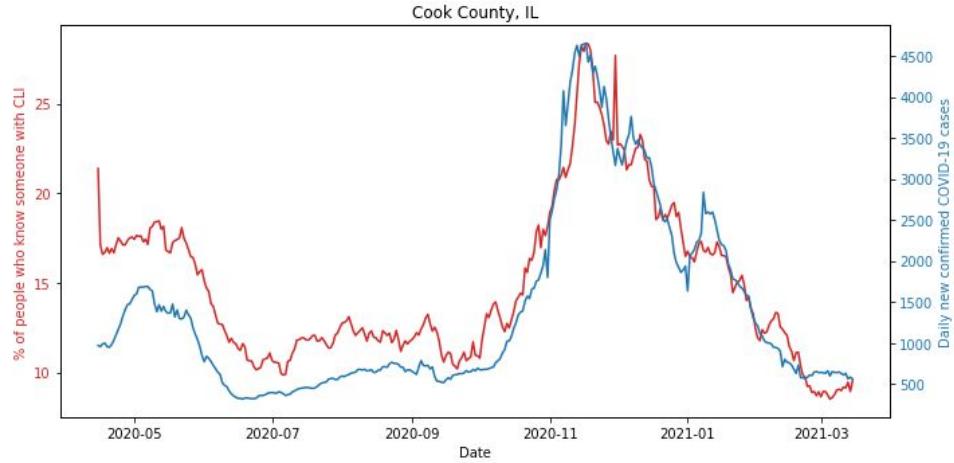
# Methods

- Spearman's correlation
  - Based on the assigned rank of the data
  - Assumes the data is non-parametric and continuously increasing or decreasing
- Pearson's correlation
  - Assumes the data is normally distributed



# Preliminary Results: Survey responses vs. daily new cases

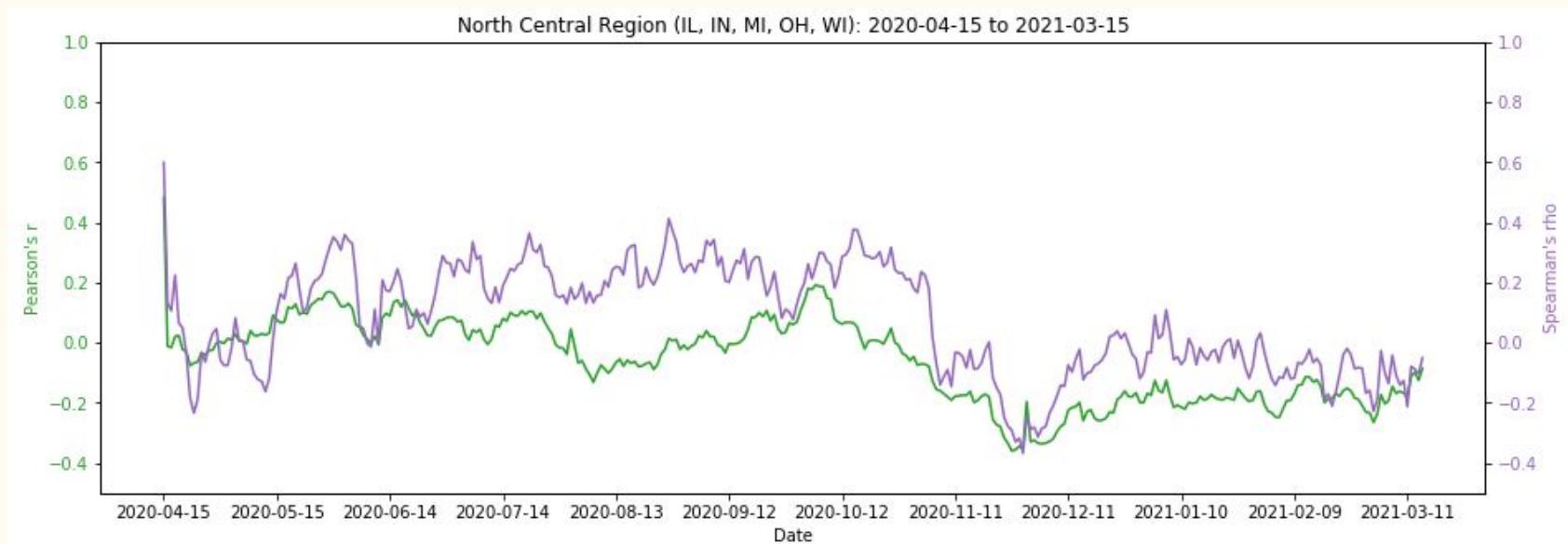
- Summer wave did not occur
- Compare the November wave in Cook County
- Very slight anticipation of the surge by survey responses



# Preliminary Results:

## Survey responses vs. daily new cases

- Weak positive correlation around period of largest surge (early/mid-October)
- Requires greater geographic range to ascertain relationship



# Questions?