IS597: COVID-19 Project

Amrutha Kumaran, Erick Li,
Kerstin Wolf, Sai Jayanth

Research Questions

1. What is the relationship between COVID-19 prevalence and political parties chosen in the 2020 presidential election? How did people on Twitter portray the COVID-19 pandemic? Did the sentiment of people have a relationship with the U.S. 2021 political outcome in battle states?
2. How does the COVID-19 daily case numbers affect bar visits, and do bar visits contribute to COVID-19 case surges?
3. How did the pandemic impact employment, hours, earnings, etc. in 2020?
4. Are county obesity rates in Illinois associated with the number of COVID-19 deaths (or hospital utilization) in those counties?
5. 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

Covid-19 vs 2020 Political Outcome

Dataset

- Kaggle: COVID-19 in USA
- MIT Election Data Science Lab
- Kaggle: Covid 19 Tweets

MIT ELECTION DATA
+ SCIENCE LAB

kaggle

Covid-19 vs 2020 Political Outcome: Dataset Use

- Covid19 Tweets
  - Sentimental Analysis of people towards Covid-19
  - Focus on tweets from battle states
- MIT Election Data Science Lab
  - Analyze total candidate votes by party in the 10 battle states for 2020 year
    - States: Pennsylvania, Texas, Wisconsin, Iowa, Michigan, Minnesota, Nevada, New Hampshire, North Carolina, and Ohio
  - Battle states were the key in the 2020 Presidential Election Outcome
    - Easier to analyze the data by focusing on certain states and compare with the covid affected rate in those states
- Covid-19 in USA
  - State level data
  - Analyze Covid positive and death rate in the battle states

Covid-19 vs 2020 Political Outcome

Data Cleaning: Covid19 Tweets

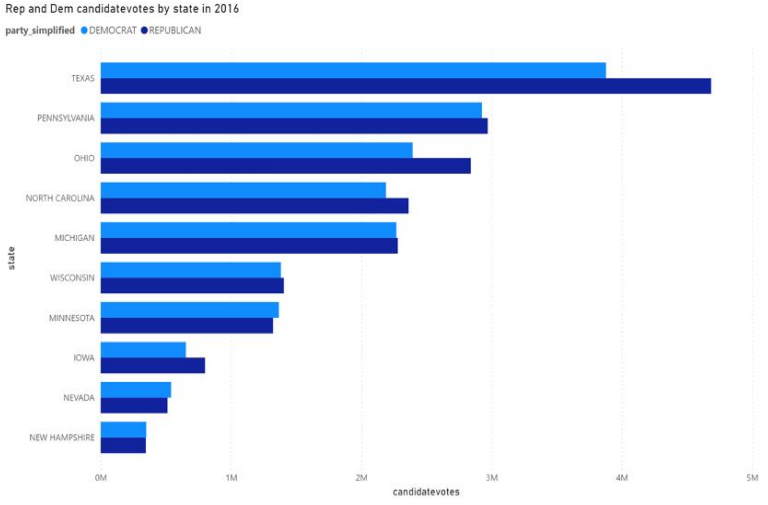
- Tool: OpenRefine 3.4.1
- Columns
    - 13 Columns & approx. 50,000 rows
    - 4 Columns (Date, user_location, text_description, and hashtags)

Data Cleaning

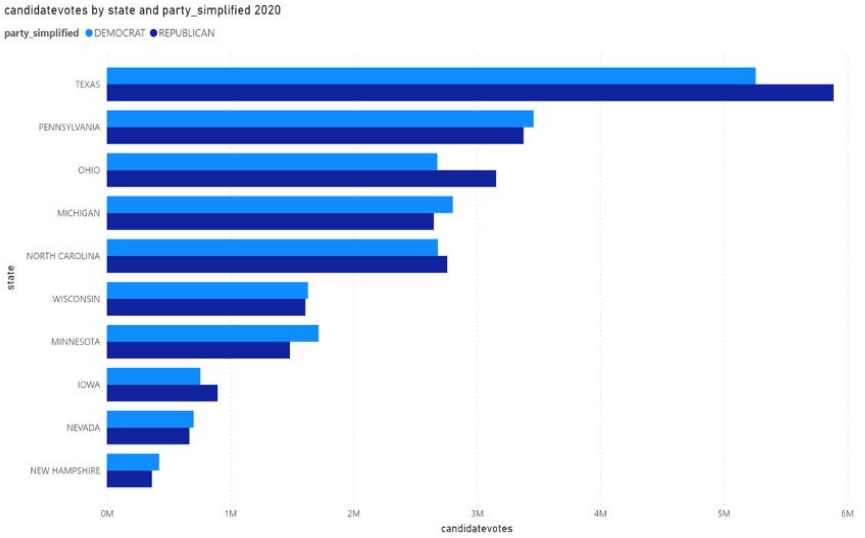
- Date
  - Change date format to ISO standard
- Hashtag and Text_description
  - remove URL links
  - Remove special characters
  - Merge similar data

- Data log
  - Transform
  - General Refine Expression Language (GREL)
    - To remove url links: value.replace(/(http:\/\/www\.|https:\/\/www\.|http:\/\/|https:\/\/)?[a-z0-9]+([\-\.]{1}[a-z0-9]+)*\.[a-z]{2,5}(:[0-9]{1,5})?(\/.*)?/, ”)
    - Clustering
      - Fingerprint
      - Metaphone3

Visualization: MIT Election Dataset

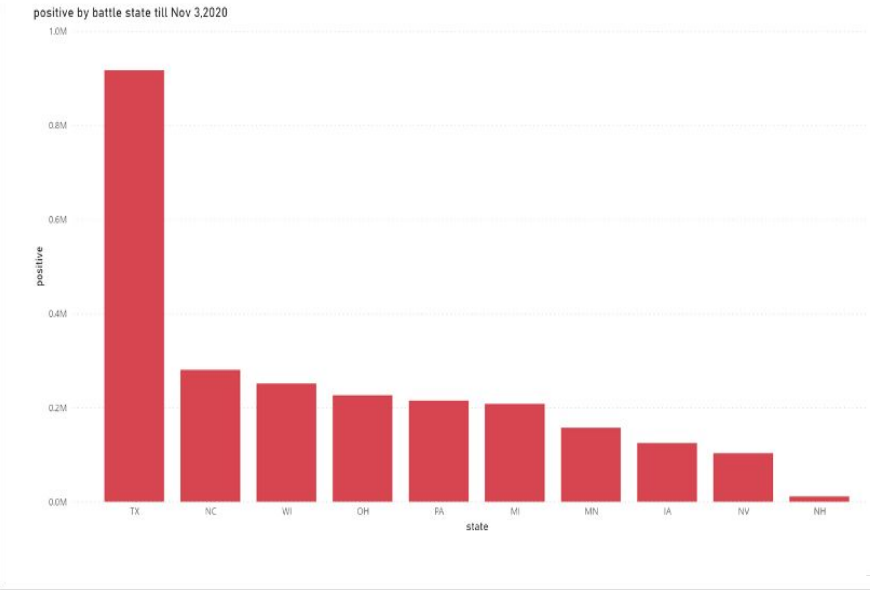


Candidate Votes by Battle States and Party in 2016

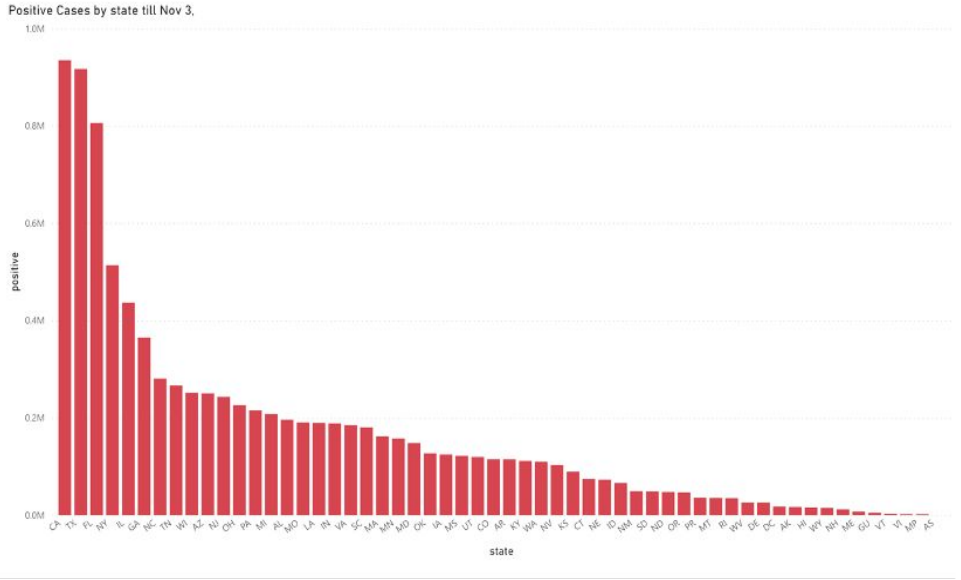


Candidate Votes by Battle States and Party in 2020

Visualization: Covid-19 in USA Dataset

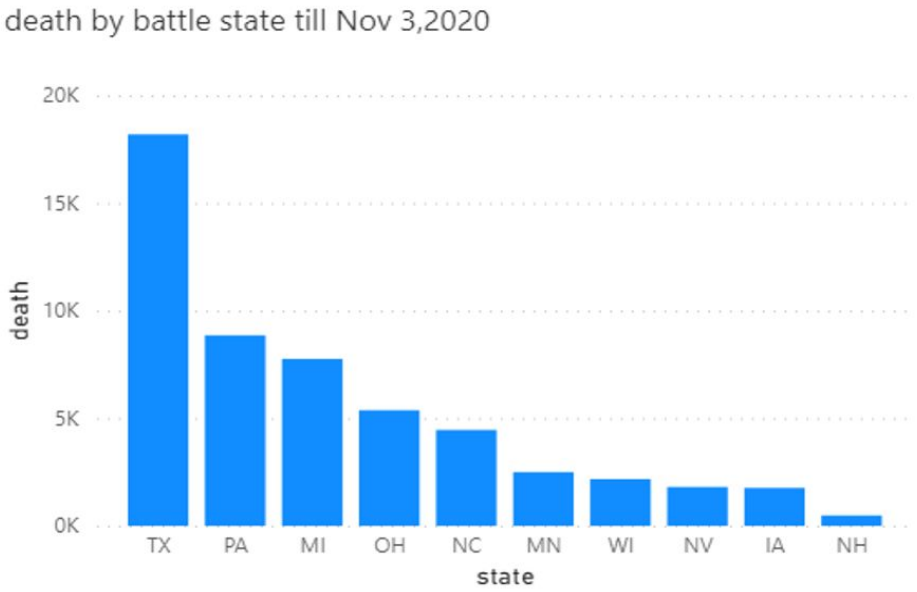


Positive Rate by Battle States Until Nov. 3, 2020

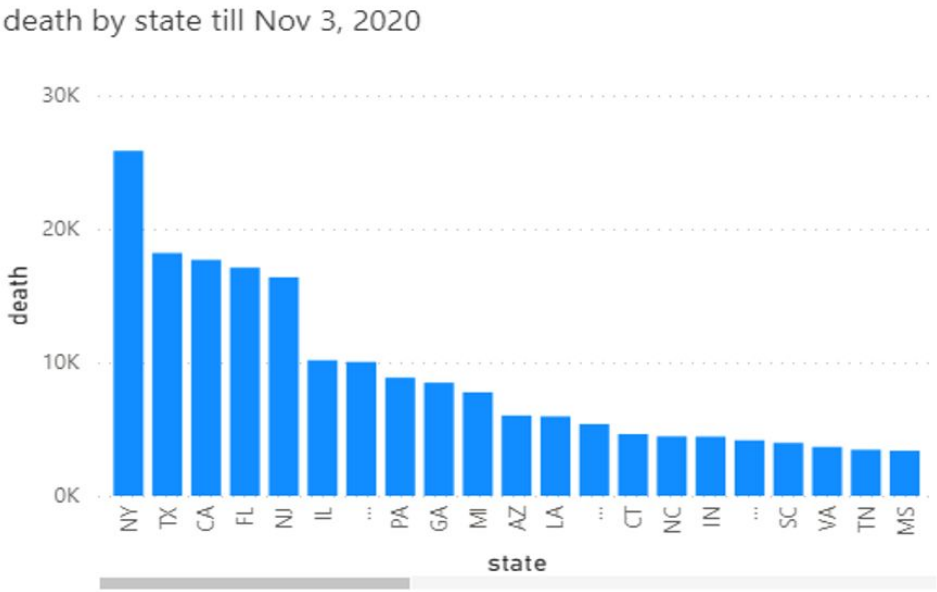


Positive Rate by States Until Nov. 3, 2020

Visualization: Covid-19 in USA Dataset



Death Rate by Battle States until Nov. 3, 2020



Death Rate by States until Nov. 3, 2020

Next Steps

- Data Cleaning with Twitter Dataset
  - Power Query/LDA (Latent Dirichlet Allocation-classify text in a document to a particular topic)
  - Visualization
- Rank the states by Covid positive and death rate
- Combine all results

Research Question #2

Illinois Daily Cases vs. Bar Visits

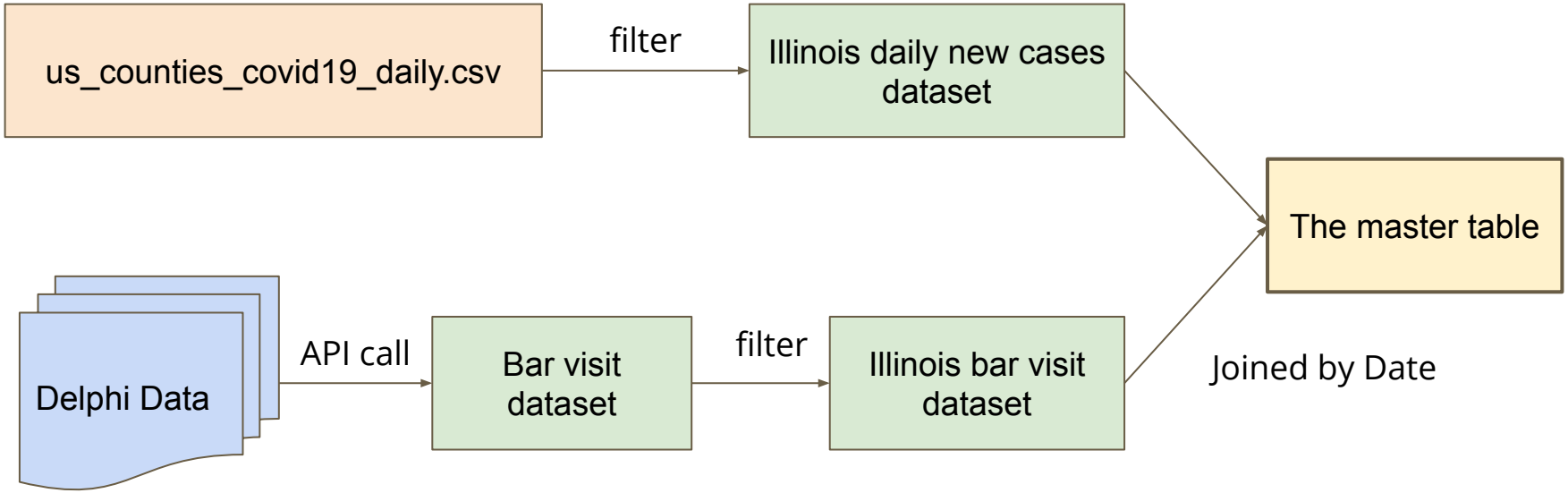
Introduction

- Study whether the number of daily new cases is associated with the number of bar visits in Illinois
- Use time series and data science models to find out
    - Whether the changes of daily case number contribute to the fluctuation of bar visits
    - Whether the surge of bar visits in the middle of the pandemic linked to the increases of the daily new cases

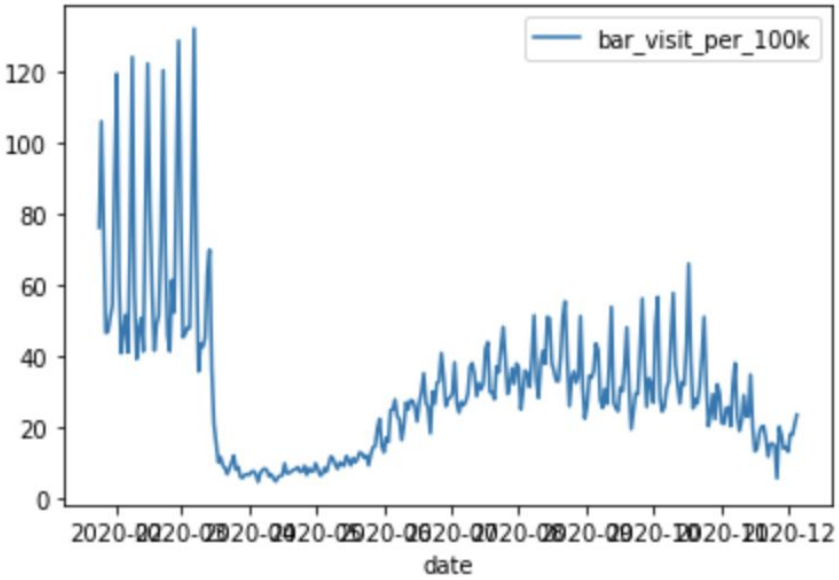
Data

- Data
  - 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

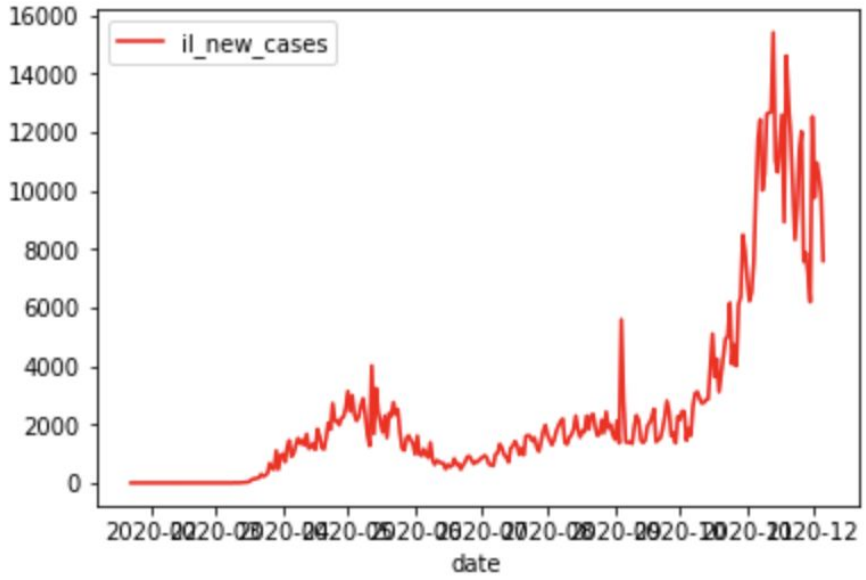
Data Preprocessing



Data Visualization



Bar Visits per 100k Population in IL



IL Daily Confirmed Case Number

Next Steps

- Use moving average instead of daily number
  - Bar visits fluctuate greatly between weekdays and weekends
  - Reduce the impact of the data errors
- Add the national daily confirmed case number into consideration and comparison
- Apply data science models to the dataset and see if there is a strong correlation between the bar visits and IL/national case numbers
  - Does the surge of the local/national number discourage people from visiting bars?
  - Does the increases of bar visiting somehow cause a later surge?

Research Question #3

COVID-19 and Unemployment

COVID-19 and Unemployment Datasets

- Centers for Disease Control and Prevention (CDC)
  - United States COVID-19 Cases and Death by State over Time
  - Provisional COVID-19 Death Counts by Week Ending Date and State (data provided by the NCHS)
- U.S. Bureau of Labor Statistics (BLS)
  - Illinois Unemployment Rate 2019-2020
  - State Unemployment Rate 2019-2020

CDC's Cases and Deaths by State Data

25,500 rows and 15 columns; 8 columns contain null values

Submission Date	State	Total Cases	Confirmed Cases	Probable Cases
New Cases	New Probable Cases	Total Deaths	Confirmed Deaths	Probable Deaths
New Deaths	New Probable Deaths	Record Creation Date/Time	Consent Cases	Consent Deaths

CDC's Provisional Death Count Data

4,428 rows and 17 columns; 12 columns contain null values

Date of Analysis	Start Date	End Date	Group	Year	Month
MMWR Week	Week Ending Date	State	COVID-19 Deaths	Total Deaths	Percent of Expected Deaths
Pneumonia Deaths	Pneumonia and COVID-19 deaths	Influenza Deaths	Pneumonia, Influenza, or COVID-19 Deaths	Footnote	

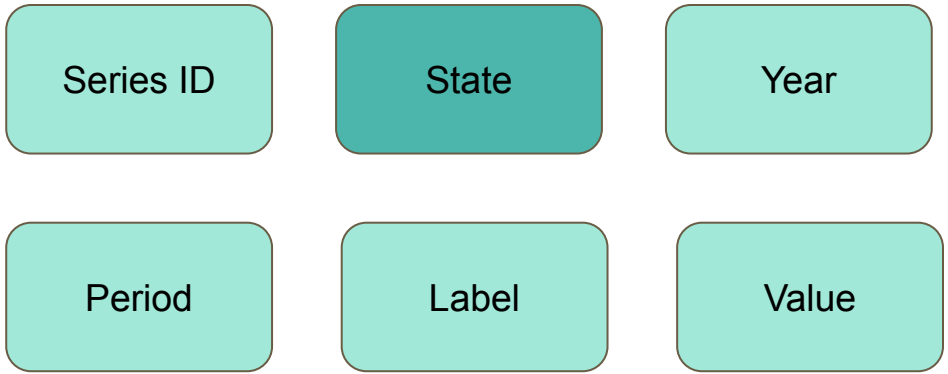
U.S. Bureau of Labor Statistics Data

Illinois Unemployment Rate 2019-2020

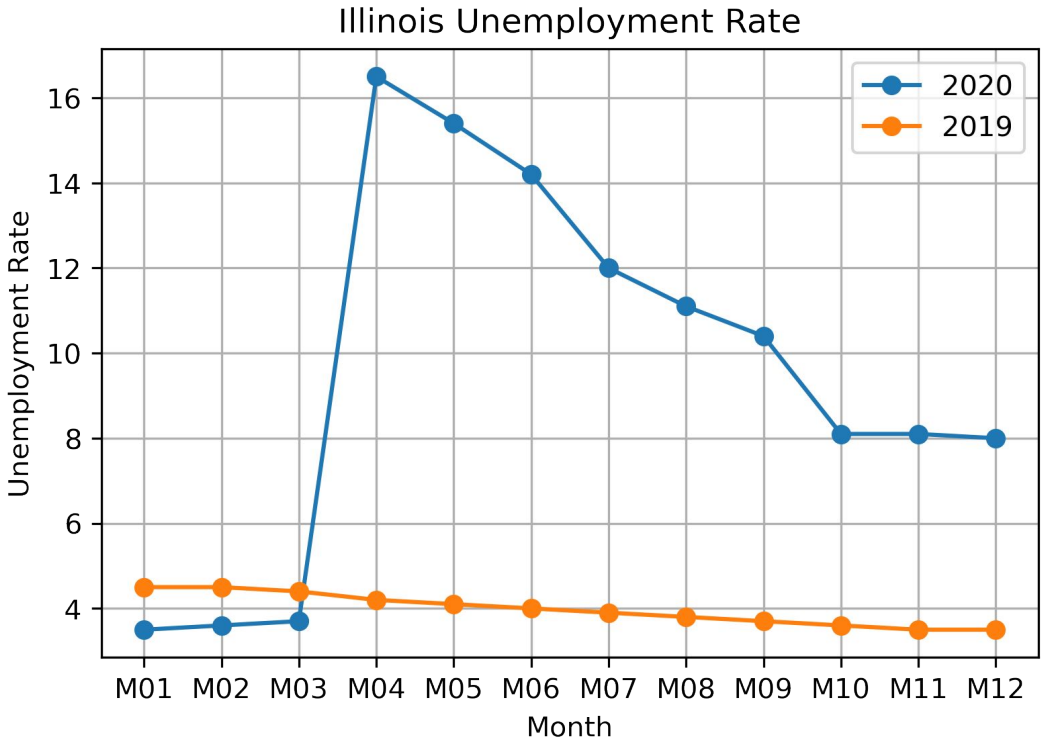
24 rows and 6 columns; seasonally adjusted

State Unemployment Rate 2019-2020

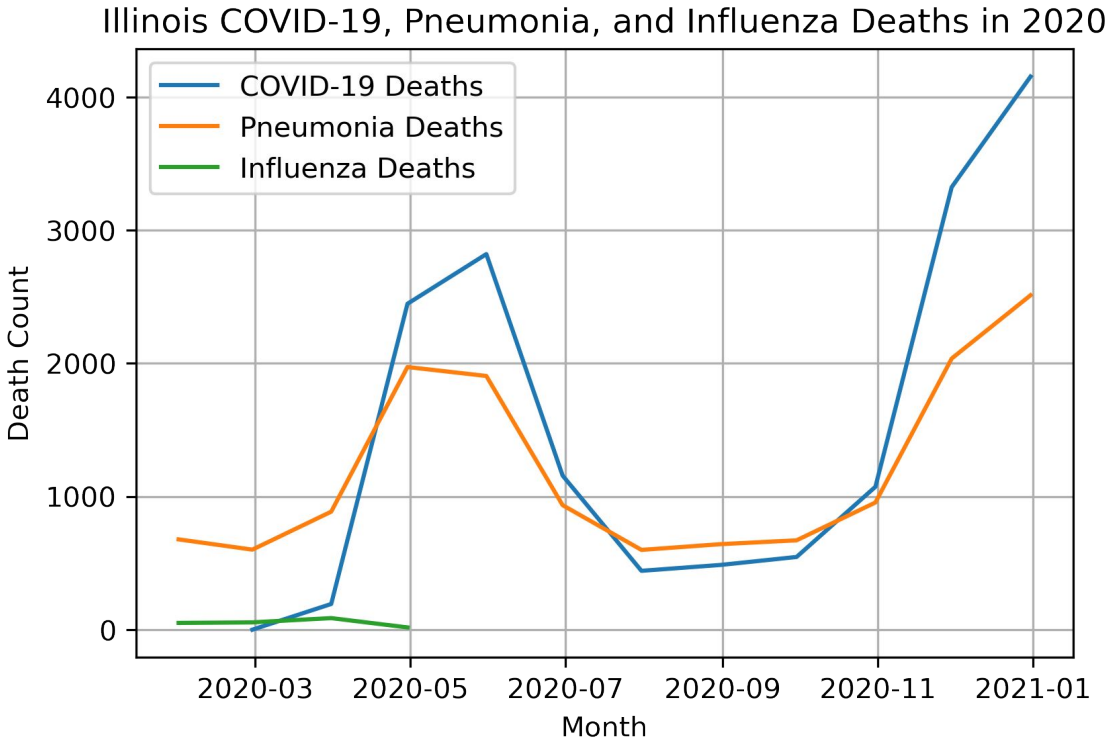
1,248 rows and 6 columns; seasonally adjusted



Visualization: Unemployment Comparison



Visualization: COVID-19 Deaths



Next Steps

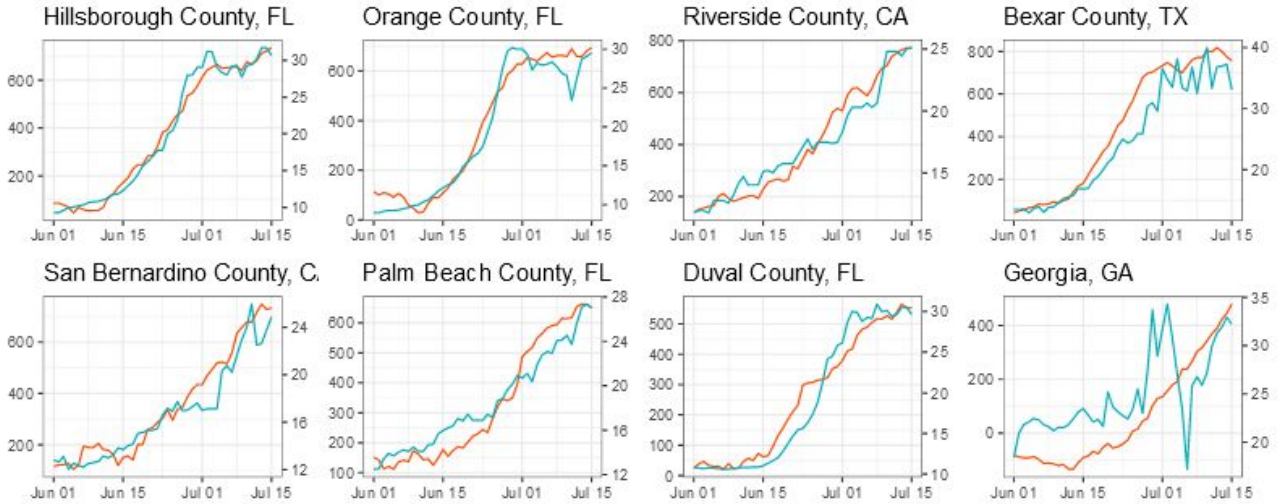
- Continue evaluating datasets
- Add more datasets exploring remote working, hours, and earnings
- Explore how a rise in cases may have affected the rapid rise in unemployment
- Compare pneumonia deaths in the winter of 2017-2018 to the winter of 2018-2019

Research Question #4

COVID-like illness in the Community

COVID-like Illness in Community

- 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**



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/

Data

- 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

Next Steps

- Focus on new research question (obesity and COVID-19 severity)
- **Only if there is time:**
  - Calculate Spearman correlation between CLI-in-community reported by Illinois and COVID cases reported in Illinois
    - Visualize relationship between variables
  - Enable CDC API to compare strength of correlation when using CDC's COVID data (different from John Hopkins' numbers

Research Question #5

Obesity and COVID-19 severity

Obesity and COVID-19 Severity (Just Started)

- Pre-existing conditions and traits have an impact on the severity of COVID-19
- Data sources:
  - Illinois county obesity rates from 2015 CDC BRFSS Survey
  - County COVID deaths from Kaggle dataset “COVID-19 in USA”
  - % ICU utilization by county is available through CDC

Next steps:

- Data cleaning and reorganization to focus solely on Illinois counties

Future goal:

- Comparing state obesity rates with state COVID-19 deaths and % ICU utilization

Questions?