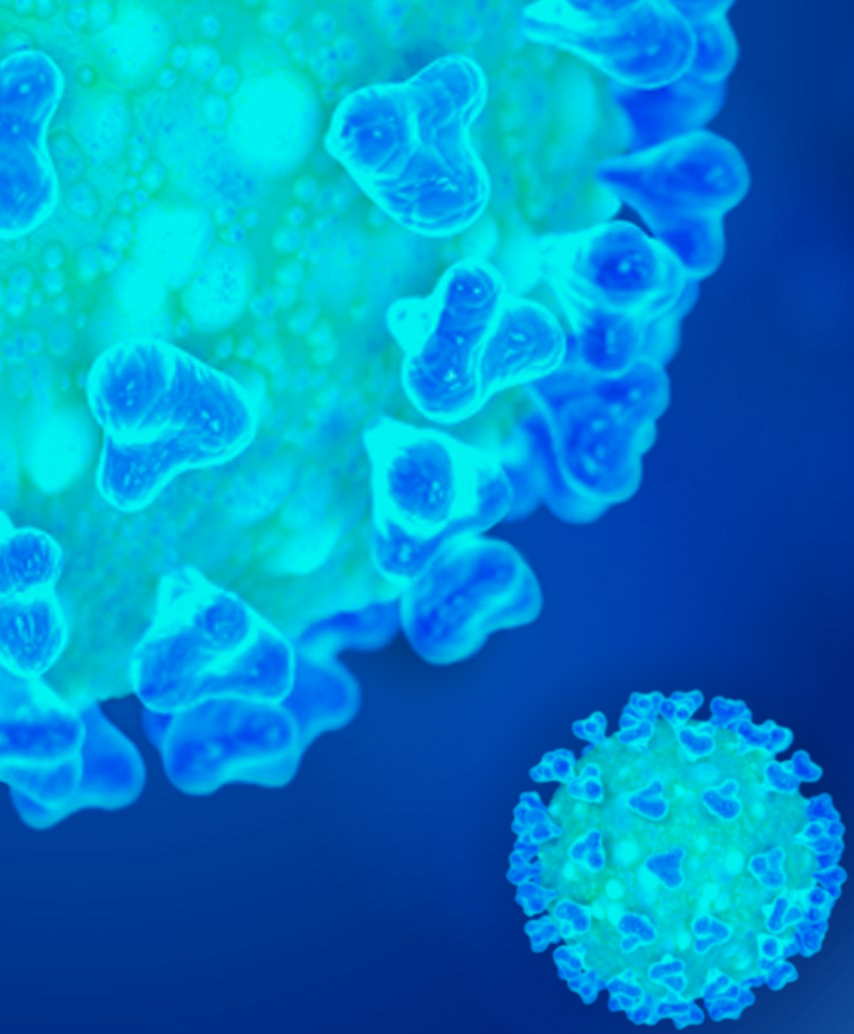


The ‘VID

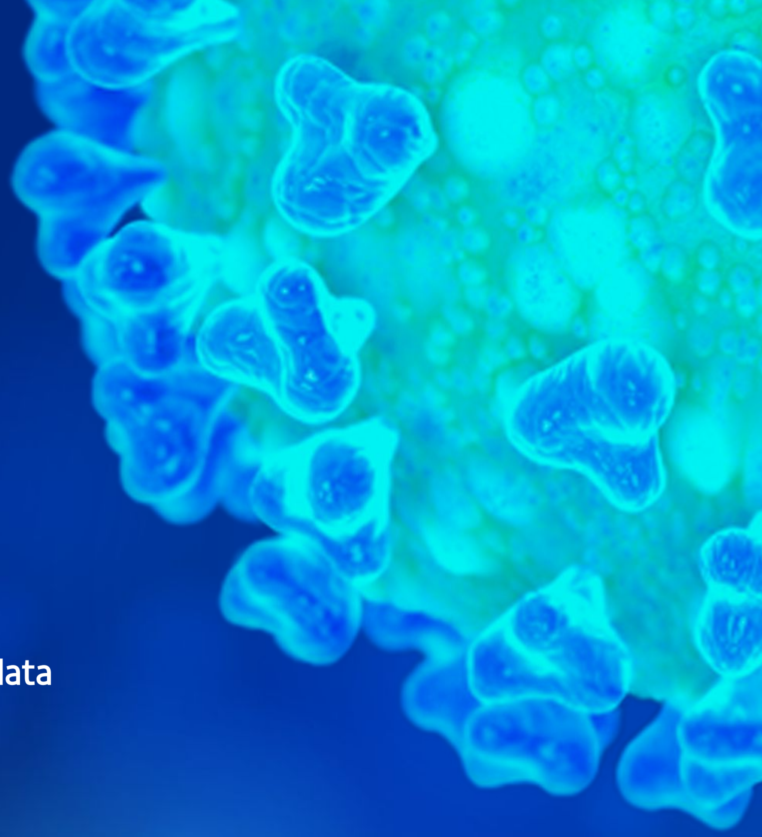
Anti-Rona Task Force

Clay Beabout, Brandon Pramann,
Benjamin Wyss, Jon Volden,
Bailey Srimoungchanh

Why COVID?

The COVID-19 epidemic continues to have an unprecedented negative impact on our economy and communities.

For this presentation, we will be focusing on COVID-19 related data sets to analyze and diagnose total cases, total deaths, survival rates, state success stories and anomalies in infection numbers.



Problems

We identify three main data science problems that exist in U.S. COVID-19 related data sets:

1. Can we predict future U.S. COVID-19 trends based on historical data? Is time-series forecasting alone enough to predict infections and deaths?

2. Is the overall rate of infection related to state and county responses? How can we compare U.S. counties based on their COVID-19 response performance?

3. Can we detect and explain anomalies in U.S. COVID-19 infections? Are there new insights about COVID-19 in the U.S. that can be determined by exploring these anomalies?

Problems

We identify three main data science problems that exist in U.S. COVID-19 related data sets:

1. Can we predict future U.S. COVID-19 trends based on historical data? Is time-series forecasting alone enough to predict infections and deaths? Time Series

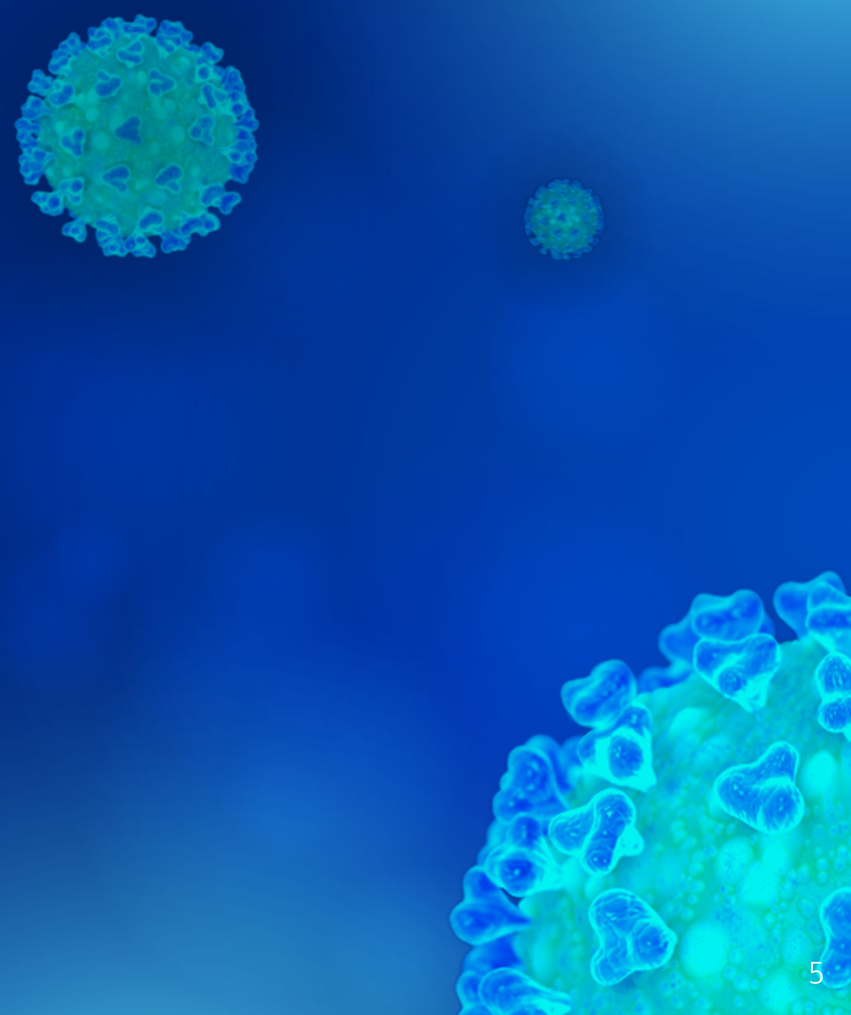
2. Is the overall rate of infection related to state and county responses? How can we compare U.S. counties based on their COVID-19 response performance? Clustering

3. Can we detect and explain anomalies in U.S. COVID-19 infections? Are there new insights about COVID-19 in the U.S. that can be determined by exploring these anomalies? Anomaly Detection

Sec. I

Time Series Forecasting

Predict Future U.S. COVID-19 Trends

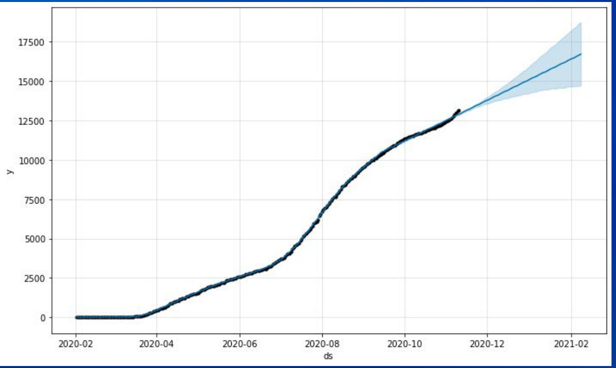


Time Series Forecasting

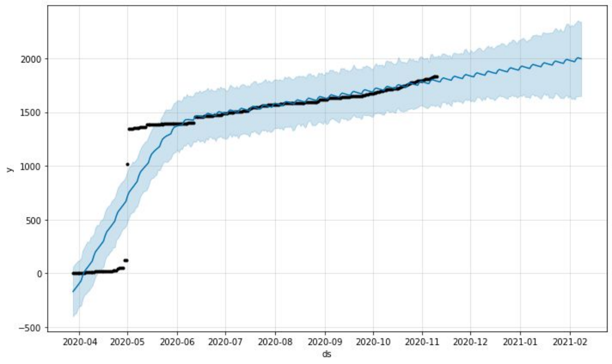
Can we predict future U.S. COVID-19 trends based on historical data? Is time-series forecasting alone enough to predict infections and deaths?

Using Facebook Prophet we looked at record data to project the rates per county 90 days into the future. (With and without feature engineering.)

San Francisco County, CA



Trousdale County, TN



Resources/Tools

Facebook Prophet

- facebook.github.io/prophet
- Time Forecasting using an additive model

NY Times

- Data contains a running total of cases accumulated per day
- ~700k lines of data
- Updated multiple times per day
- github.com/nytimes/covid-19-data

```
date,county,state,fips,cases,deaths
2020-01-21,Snohomish,Washington,53061,1,0
2020-01-22,Snohomish,Washington,53061,1,0
2020-01-23,Snohomish,Washington,53061,1,0
2020-01-24,Cook,Illinois,17031,1,0
2020-01-24,Snohomish,Washington,53061,1,0
2020-01-25,Orange,California,06059,1,0
2020-01-25,Cook,Illinois,17031,1,0
2020-01-25,Snohomish,Washington,53061,1,0
2020-01-26,Maricopa,Arizona,04013,1,0
2020-01-26,Los Angeles,California,06037,1,0
```

US Census

- Data from 2019 to calculate rates per 100,000
- Adjacent County Data
- www.census.gov

InformationIsBeautiful

- Deadliness/Contagiousness chart
- https://www.informationisbeautiful.net/visualizations/the-microbescope-infectious-diseases-in-context/

Center for Disease Control and Prevention

- https://www.cdc.gov/flu/weekly/fluviewinteractive.htm

Method

Standard data science libraries + Prophet

```python
import pandas as pd
import numpy as np
import plotly.express as px
from fbprophet import Prophet
```

Merged and preprocessed data into cases per 100,000

```python
df = df.merge(popdf, how='left', left_on='fips', right_on='FIPS')
...
df['rate'] = df['deaths'] / df['ratio']
```

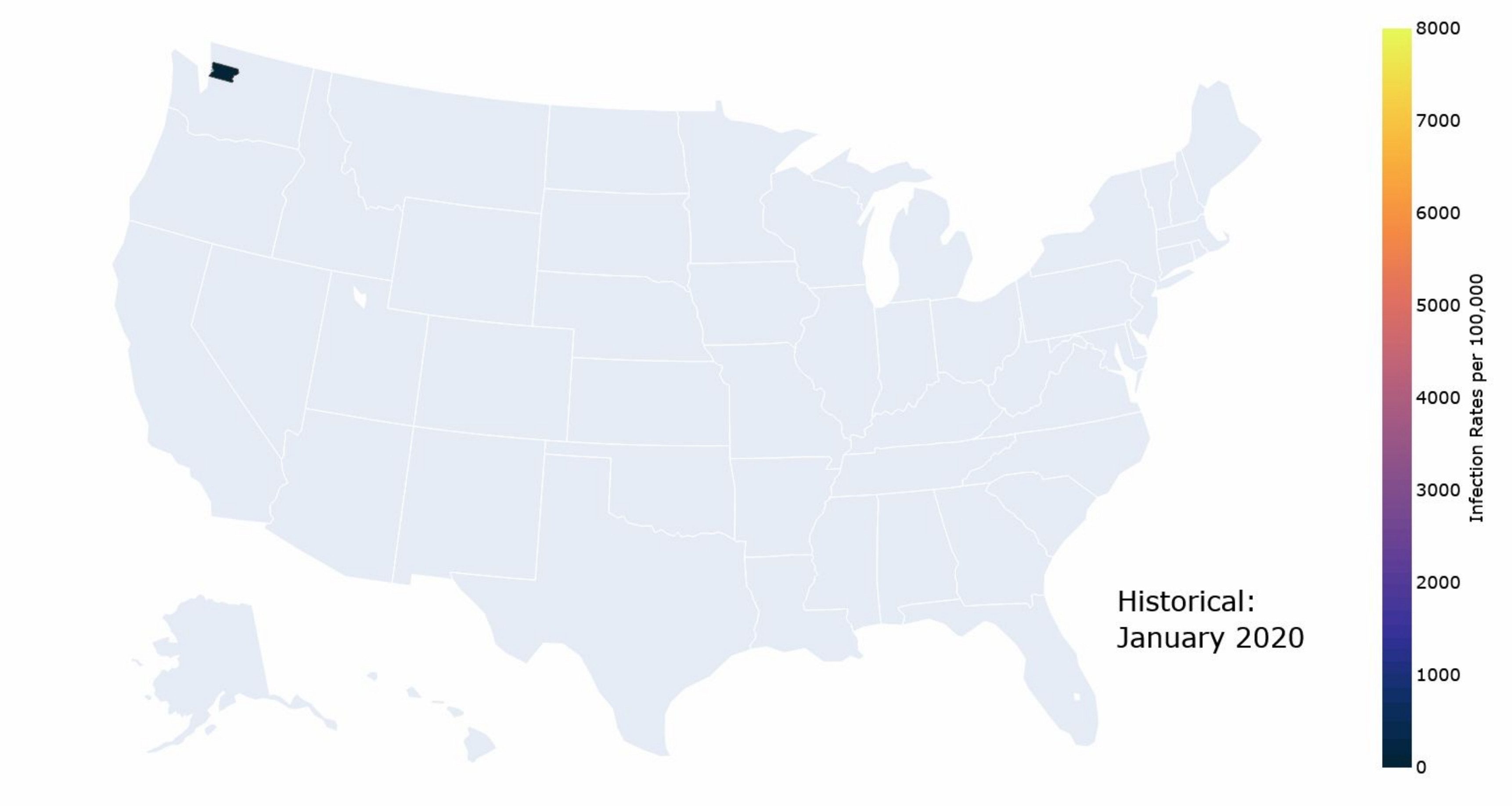
Concurrently ran each county through Prophet

```python
m = Prophet()
m.fit(data)
result = m.predict(future)
```

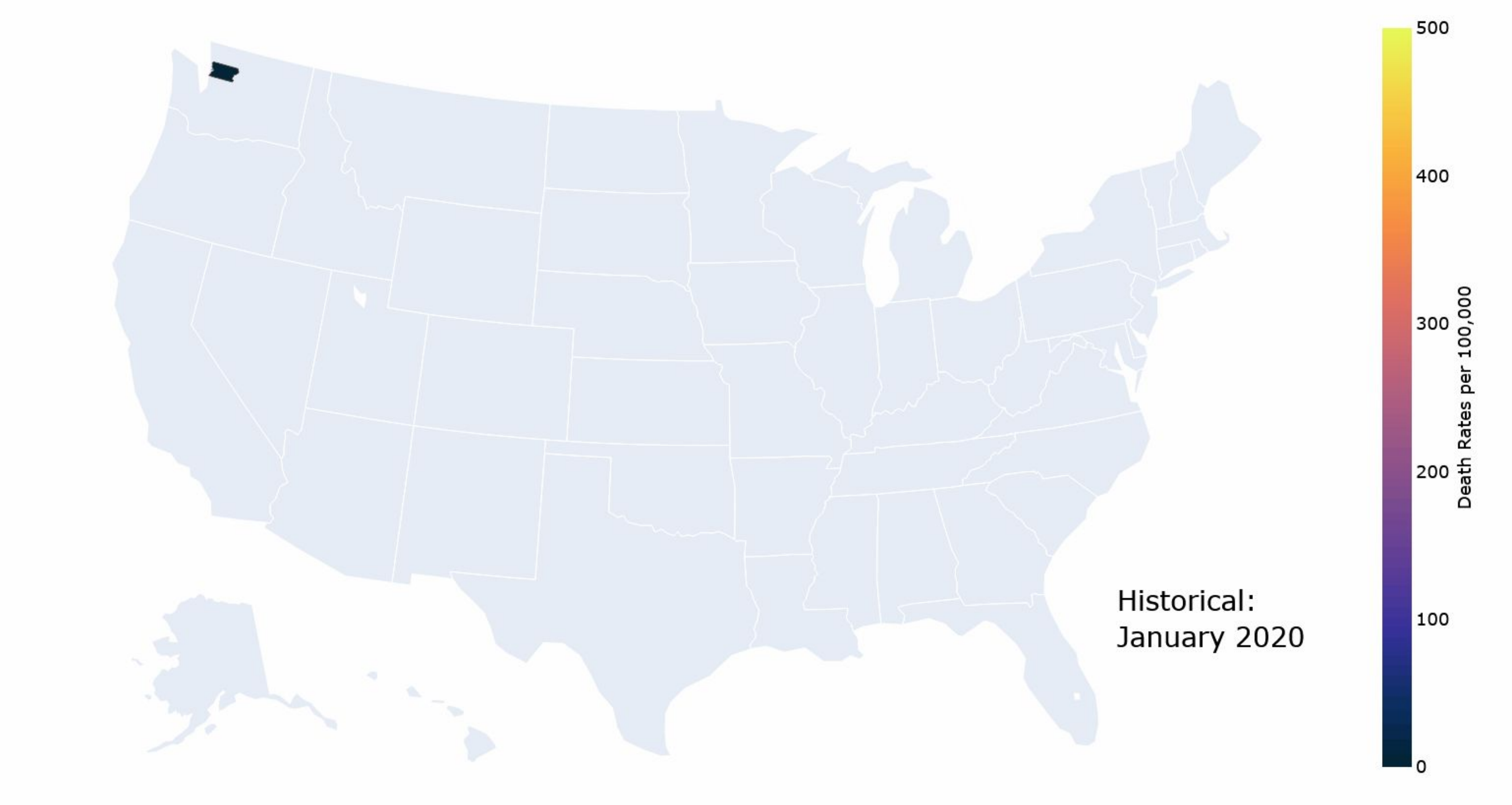
Plotted counties historical and projected rates

- Data had to be condensed into weekly
- Plotted using choropleth maps and county data provided by plotly

U.S. COVID-19 Cases per 100,000



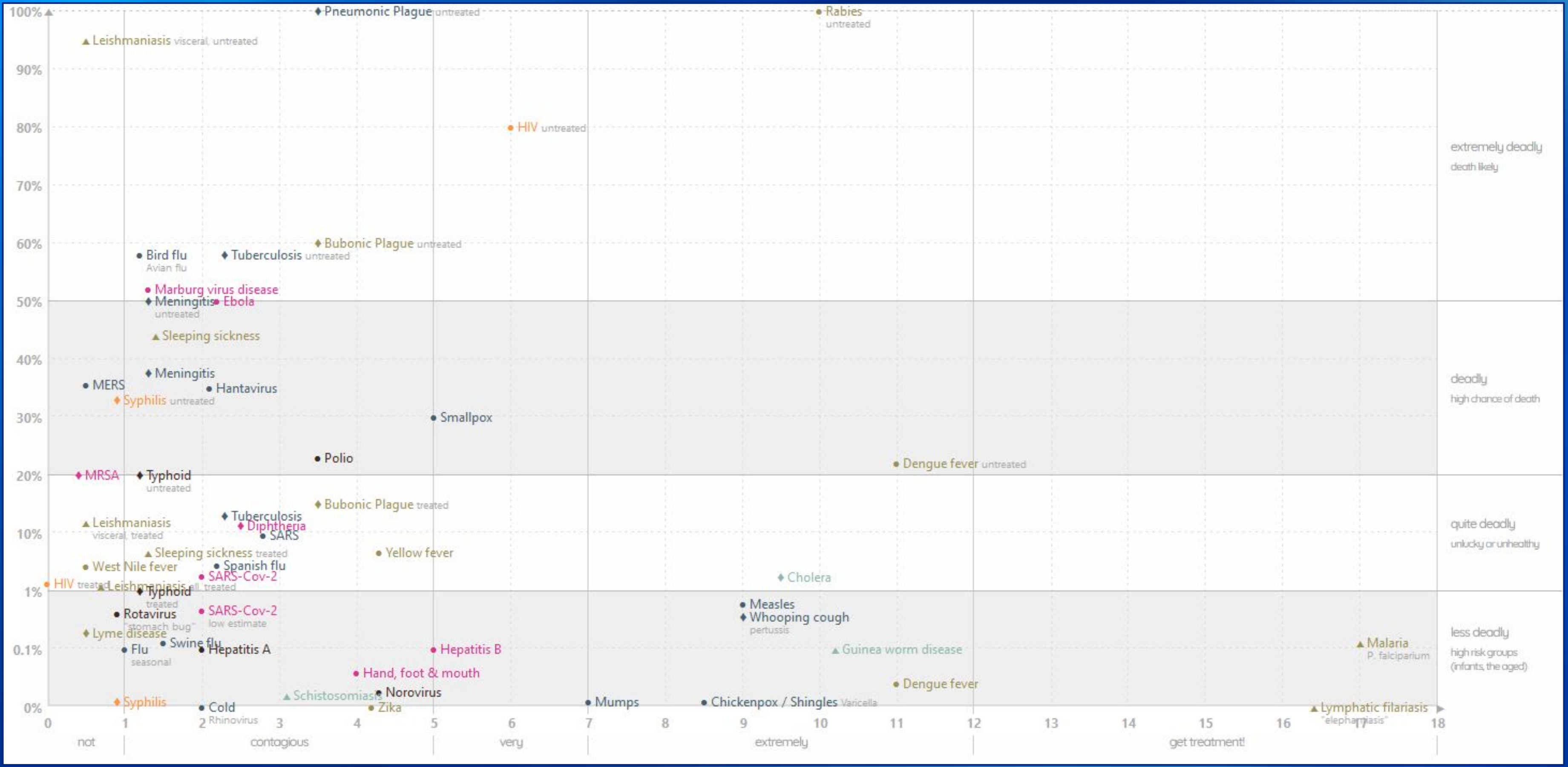
U.S. COVID-19 Deaths per 100,000



Further Feature Engineering

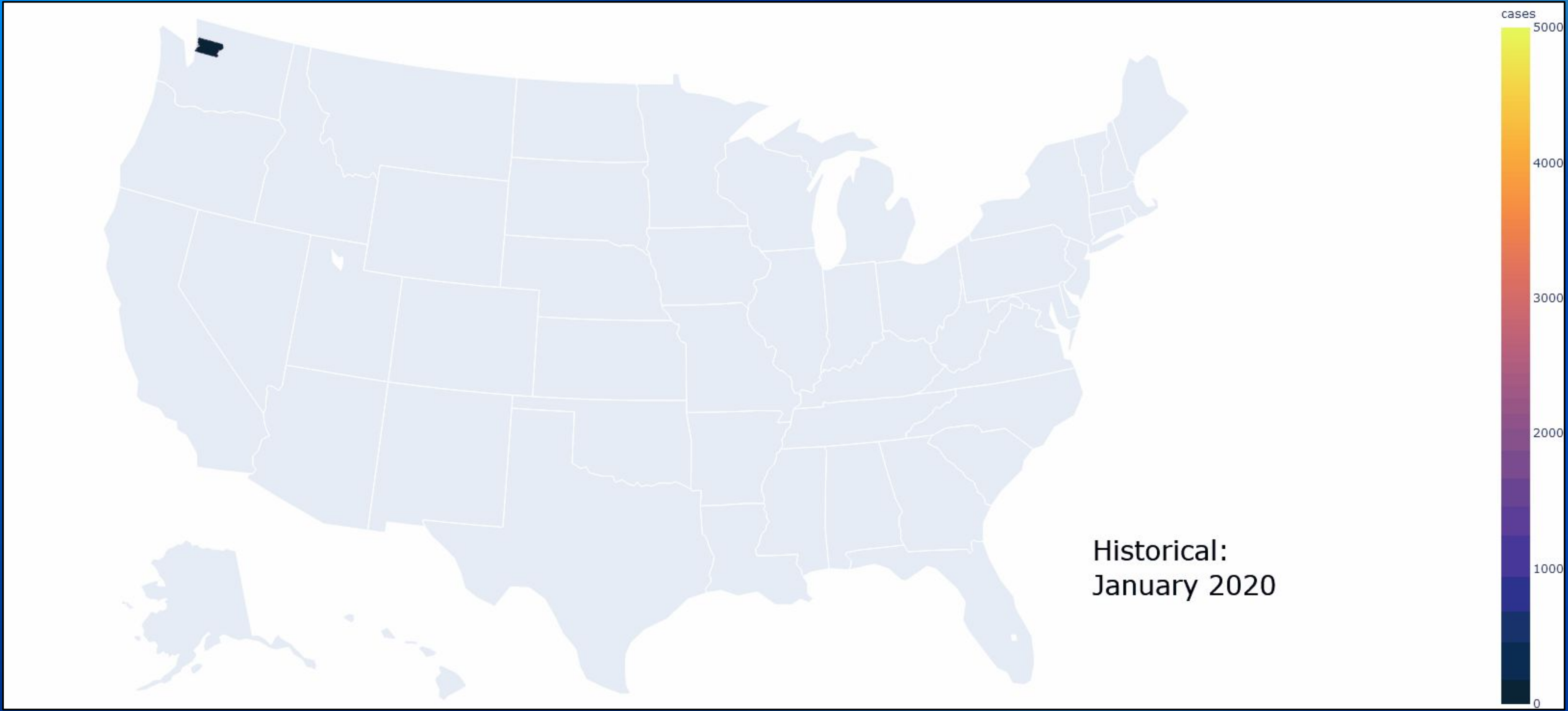
- **Adjacency**
  - https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/county-adjacency-record-layout.html
  - **Training Data**
    - Using provided census data about adjacent counties the sum of cases in and around a county can be generated when combined with the NYTimes dataset
  - **Predicting Data**
    - Prophet was utilized to predict the sum of adjacent cases
- **Virus Projections**
  - https://www.cdc.gov/flu/weekly/fluviewinteractive.htm
  - **Training Data**
    - Using CDC data from 2018-2019 on the amount of influenza-like illness cases weekly with linear interpolation to fill in daily data. Shift the dates to line up with the NYTimes data
    - Normalize ILI cases for each state, then average over all states for each day to be used as a feature
  - **Predicting Data**
    - "Future" ILI normalized values are already known

Why the flu?



100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
1%
0.1%
0%
Pneumonic Plague untreated
Rabies untreated
Leishmaniasis visceral, untreated
HIV untreated
extremely deadly
death likely
Bubonic Plague untreated
Bird flu
Avian flu
Tuberculosis untreated
Marburg virus disease
Meningitis
untreated
Ebola
Sleeping sickness
Meningitis
MERS
Hantavirus
Syphilis untreated
deadly
high chance of death
Smallpox
Polio
Dengue fever untreated
MRSA
Typhoid
untreated
Bubonic Plague treated
Leishmaniasis
visceral, treated
Tuberculosis
Diphtheria
SARS
quite deadly
unlucky or unhealthy
Sleeping sickness treated
Yellow fever
West Nile fever
Spanish flu
SARS-Cov-2
Cholera
HIV treated
Leishmaniasis all, treated
Typhoid
treated
Rotavirus
"stomach bug"
SARS-Cov-2
low estimate
Measles
Whooping cough
pertussis
less deadly
high risk groups
(infants, the aged)
Lyme disease
Flu
seasonal
Swine flu
Hepatitis A
Hepatitis B
Guinea worm disease
Malaria
P. falciparium
Hand, foot & mouth
Dengue fever
Syphilis
Cold
Rhinovirus
Schistosomiasis
Norovirus
Zika
Mumps
Chickenpox / Shingles Varicella
Lymphatic filariasis
"elephantiasis"
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
not
contagious
very
extremely
get treatment!

U.S. COVID-19 Projected Cases with Adjacency



Sec. II

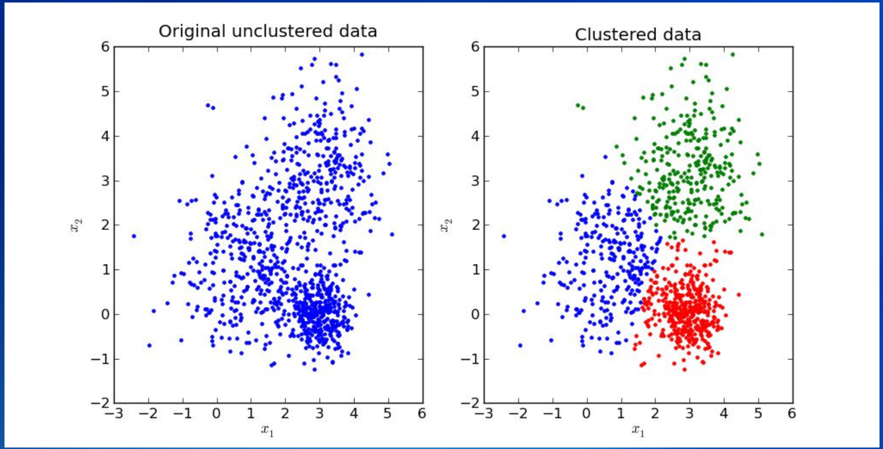
Clustering

Compare and Contrast County-by-County COVID-19 Response

Clustering

- Is the overall rate of infection related to state and county responses? How can we compare U.S. counties based on their COVID-19 response performance?

- To answer these questions, we construct a clustering model that groups U.S. counties together based on COVID-19 response performance metrics. We investigate if these metrics are correlated to overall infection rates.



New York Times Live COVID-19 Infections Dataset

https://github.com/nytimes/covid-19-data

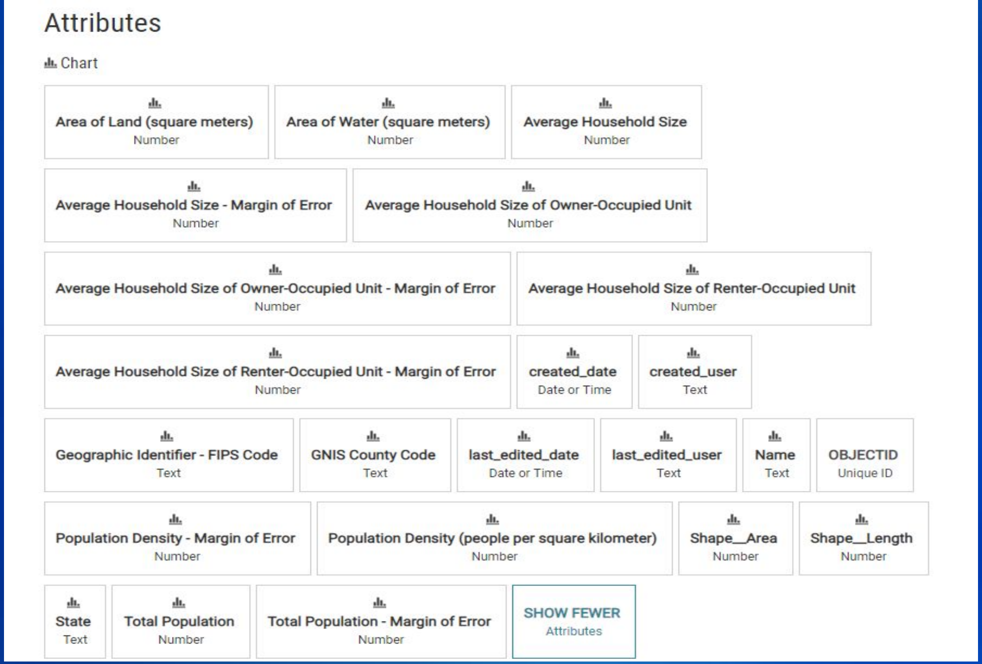
County-By-County Case Data. This information is necessary to compare counties based on response performance and COVID-19 infections per capita.

county	state	fips	cases	deaths	confirmed_cases	confirmed_deaths	probable_cases	probable_deaths
Autauga	Alabama	01001	2456	36	2199	33	257	3
Baldwin	Alabama	01003	7646	84	6397	80	1249	4
Barbour	Alabama	01005	1128	9	766	9	362	0
Bibb	Alabama	01007	986	17	887	13	99	4
Blount	Alabama	01009	2549	34	1949	33	600	1
Bullock	Alabama	01011	677	19	631	15	46	4
Butler	Alabama	01013	1087	41	1031	40	56	1
Calhoun	Alabama	01015	5608	77	4738	68	870	9
Chambers	Alabama	01017	1570	48	1010	41	560	7
Cherokee	Alabama	01019	908	15	650	14	258	1
Chilton	Alabama	01021	2078	36	1869	29	209	7
Choctaw	Alabama	01023	411	12	378	12	33	0
Clarke	Alabama	01025	1503	18	1225	16	278	2
Clay	Alabama	01027	850	13	727	13	123	0
Cleburne	Alabama	01029	669	11	623	11	46	0
Coffee	Alabama	01031	2170	14	1609	7	561	7

U.S. Census County Populations Dataset

https://covid19.census.gov/datasets/21843f238cbb46b08615fc53e19e0daf?geometry=136.810%2C28.795%2C-136.179%2C67.148

Attributes as Features. Many of these attributes, such as total population, population density, and average household size should intuitively be correlated with COVID-19 infections

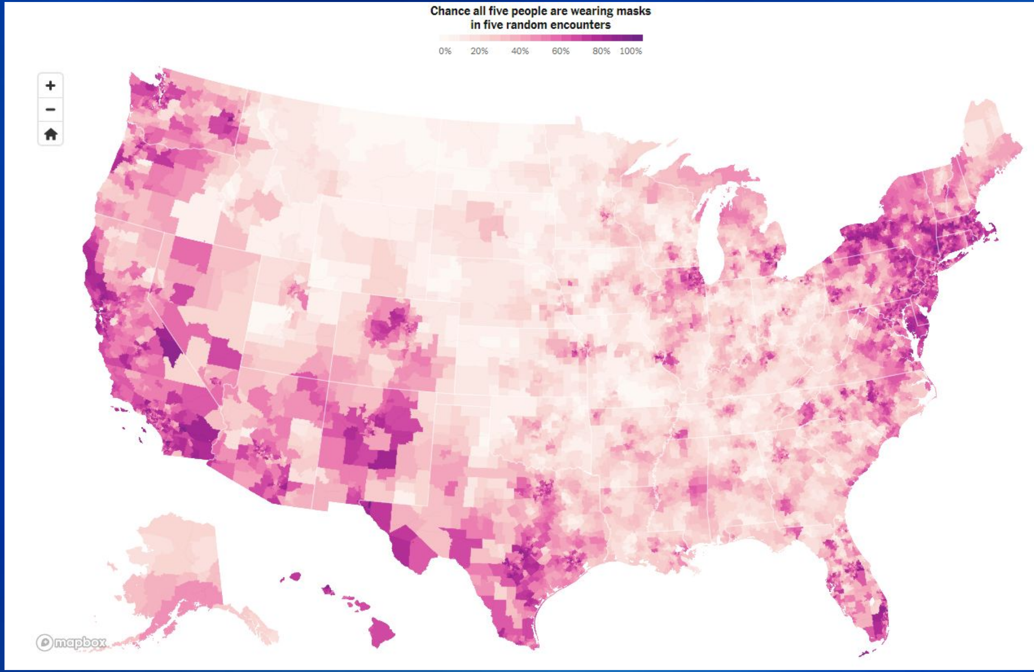


The New York Times COVID-19 Mask Use Dataset

https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html

How Does Mask Use Correspond to Infection Rates?

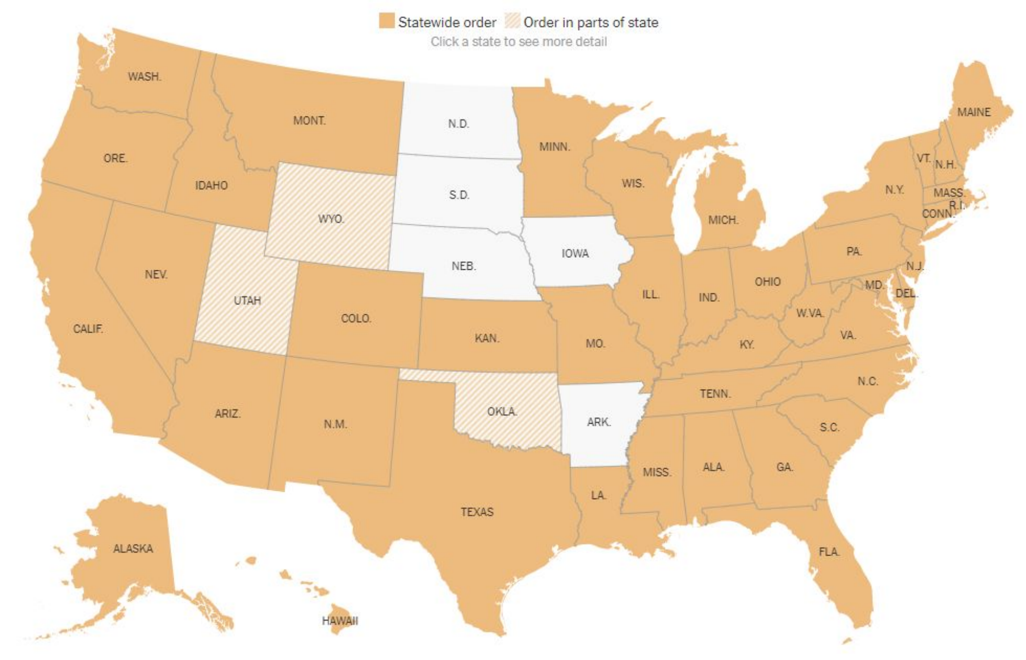
Mention Back to Time Series. Counties with high mask use seemed to have high infection rates early on, but our forecasting model predicts lower infection rates for these counties in the future.



U.S. County Lockdown Dates Dataset

https://www.kaggle.com/lin0li/us-lockdown-dates-dataset

How Do Lockdown Dates Correspond to Infection Rates?



Tools

- Pandas dataframes
- Matplotlib visualizations
- Scikit-learn clustering algorithms
    - DBSCAN, MeanShift, AgglomerativeClustering, OPTICS
- Plotly express chloroplex maps

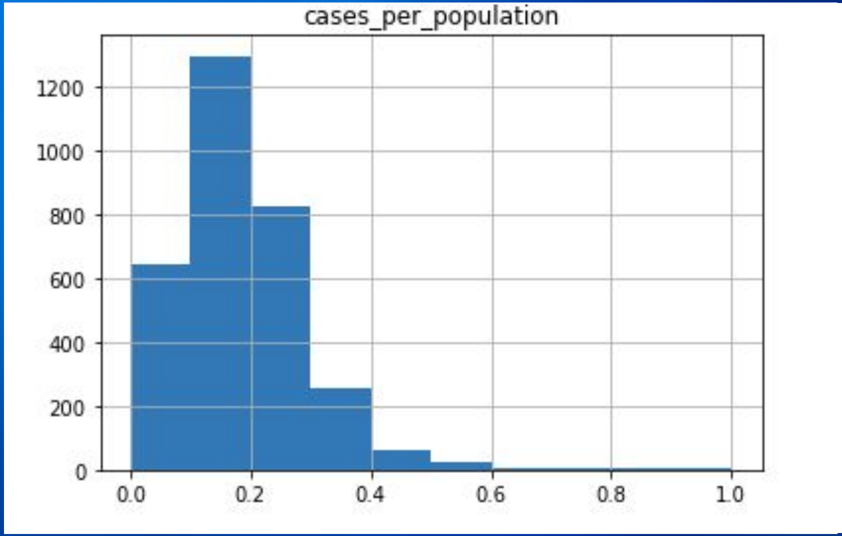
Feature Engineering

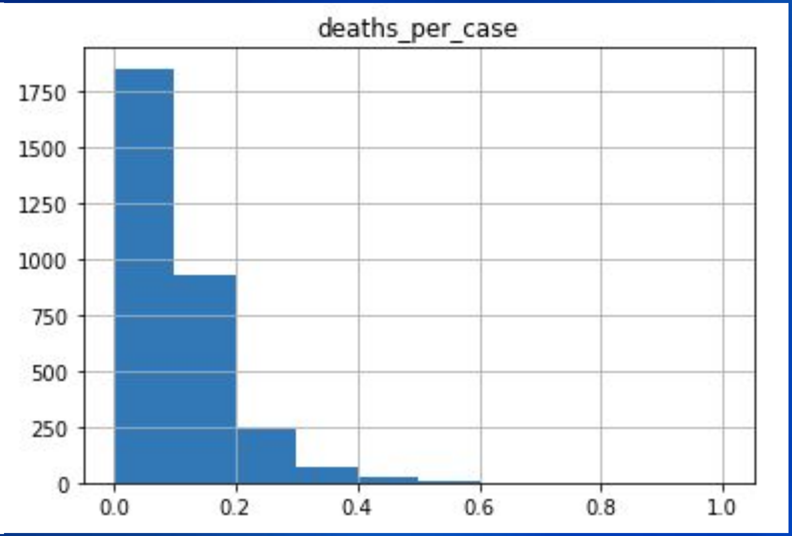
- COVID-19 Infections per Capita
  - ```
    df['cases_per_population'] = df['cases'] / df['population']
    ```
- COVID-19 Deaths per Case
  - ```
    df['deaths_per_population'] = df['deaths'] / df['cases']
    ```
- Population Density
- Average Household Size
- Mask Use Score
  - ```
    df['mask_use'] = df['mask_always'] * 1 + df['mask_frequently'] * 0.75 +
    df['mask_sometimes'] * 0.5 + df['mask_rarely'] * 0.25 + df['mask_never'] * 0
    ```
- Lockdown Score
  - ```
    df['lockdown_score'] = df.lockdown.dt.dayofyear
    ```
  - ```
    df['lockdown_score'] = (df.lockdown_score - df.lockdown_score.min())
                         / (df.lockdown_score.max() - df.lockdown_score.min())
    ```
  - ```
    df['lockdown_score'] = 1 - df.lockdown_score
    ```

- All input features are then standardized to range between 0 and 1
  - ```
    df[column] = (df[column] - df[column].min()) / (df[column].max() - df[column].min())
    ```

- Except for the lockdown score of counties that did not go into lockdown; they receive a lockdown score of -1.
  - ```
    df['lockdown_score'] = df.lockdown_score.fillna(-1)
    ```

Input Feature Histograms

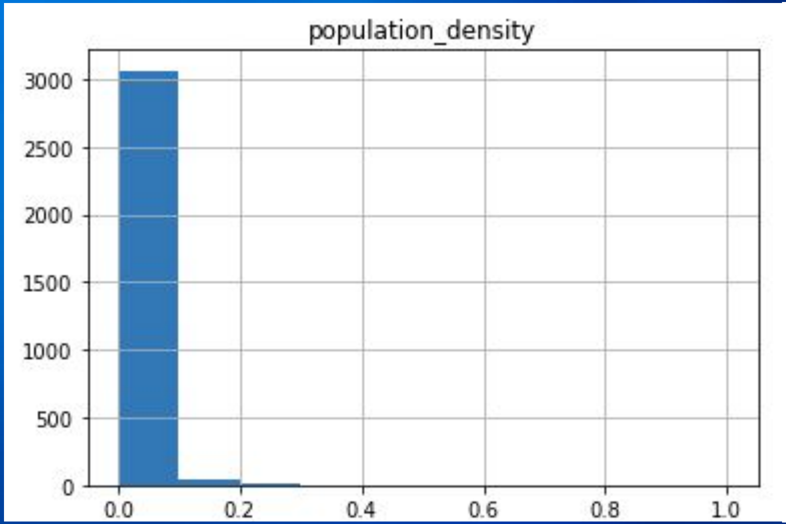


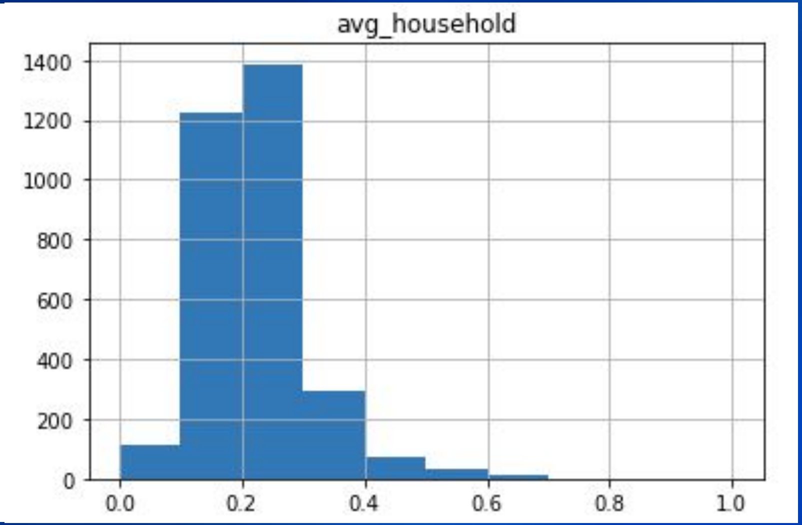


Mean: 0.184398

Mean: 0.102980

Input Feature Histograms Contd.

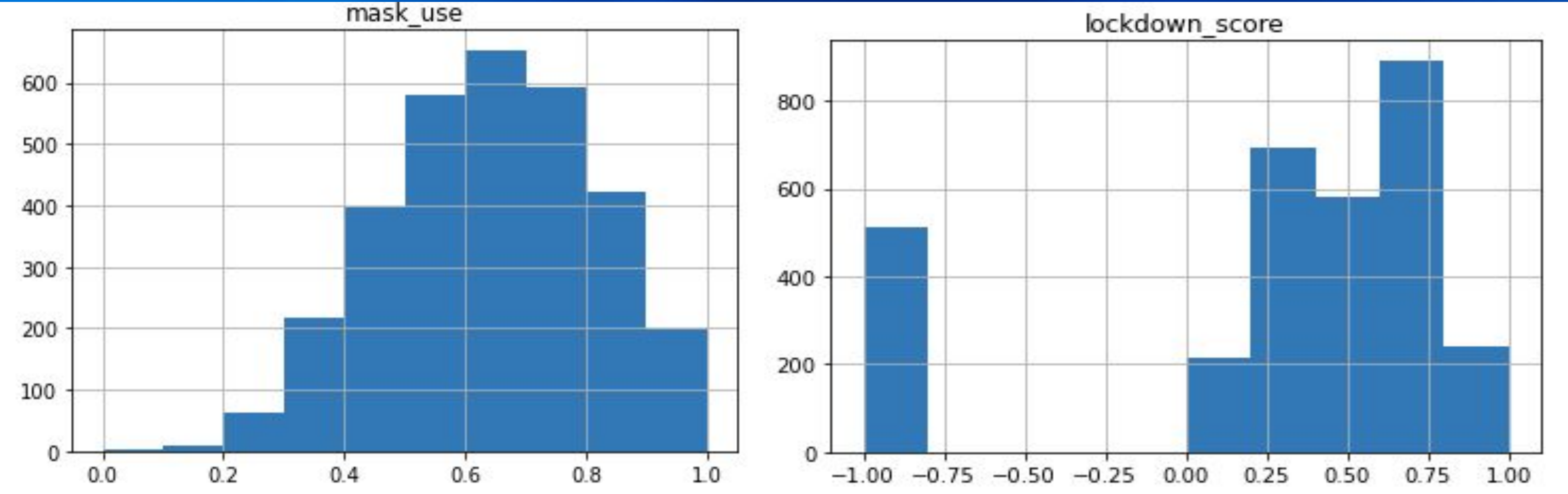




Mean: 0.011643

Mean: 0.222057

Input Feature Histograms Contd.



Mean: 0.640252

Mean: 0.254943

Input Feature Sets

- All Input Features
  - Cases per capita
  - deaths per case
  - population density
  - average household size
  - mask use score
  - lockdown score
- Reduced Input Features
  - Cases per capita
  - population density
  - mask use score
  - lockdown score
- Reduced and Infectionless Features
  - Population density
  - mask use score
  - lockdown score

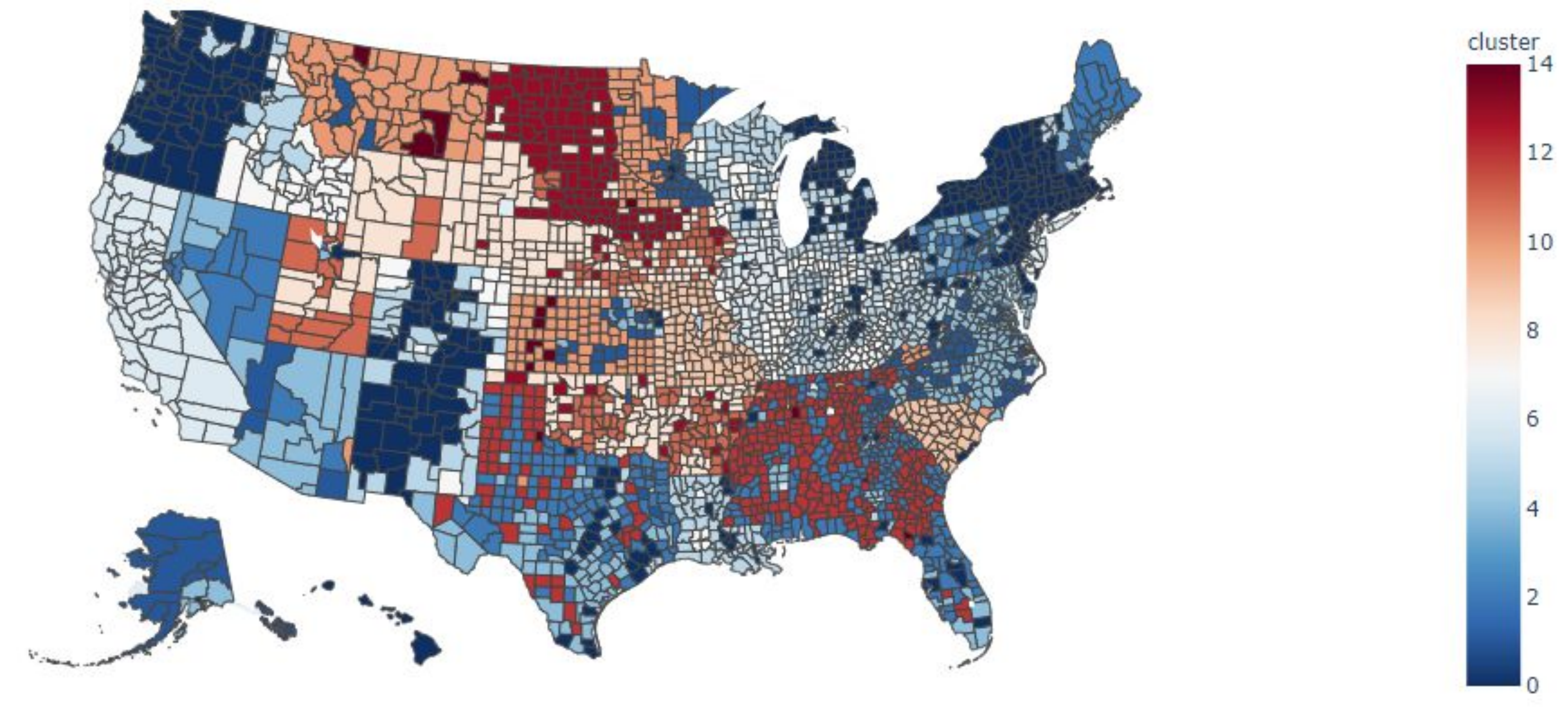
Clustering Model Performance

- The feature set containing all input features negatively impacted clustering models by adding greater variance to cluster sizes and feature values.

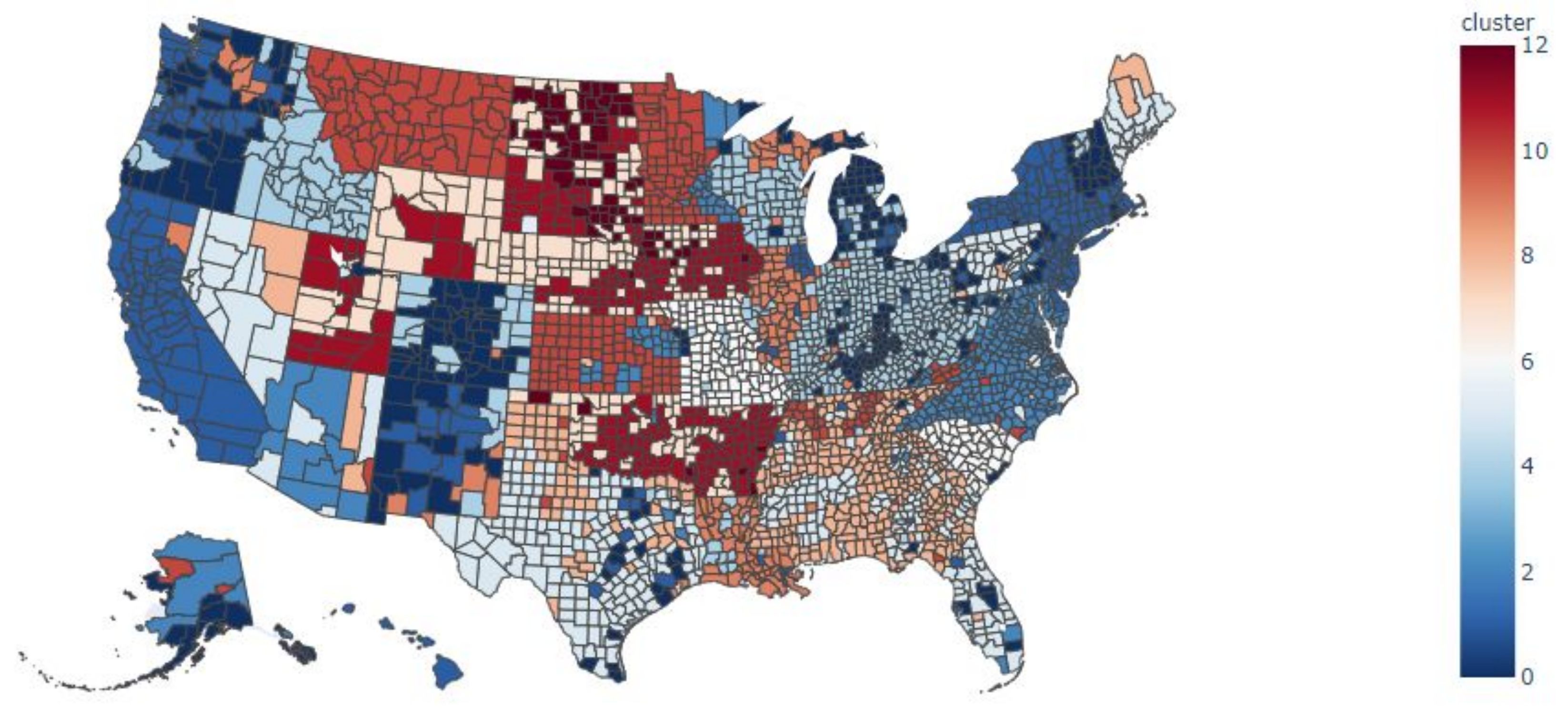
- OPTICS clustering algorithm failed to cluster almost every county
- DBSCAN and MeanShift models generated a few superclusters that were too varied to perform detailed analysis

- AgglomerativeClustering model performed very well, generating ~10 similarly sized clusters with discernable similarities in feature values.

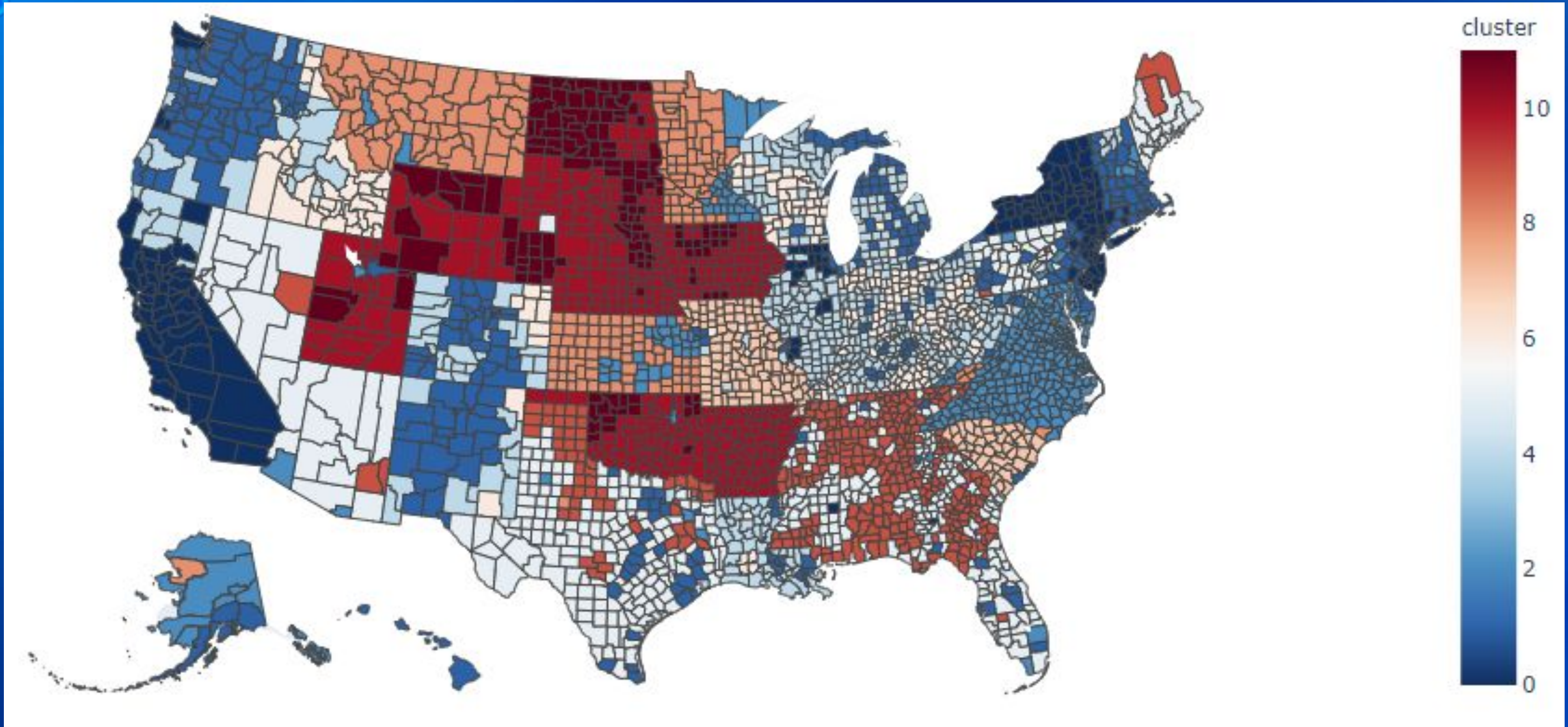
Clustering Model Visualization - Yesterday Reduced Input Feature Set



Clustering Model Visualization - Today
Reduced Input Feature Set



Clustering Model Visualization
Reduced and Infectionless Input Feature Set



Results

- County-by-county COVID-19 response varies with county political affiliation
- COVID-19 response metrics are strongly correlated with decreased infection rates
  - **Mask use score and COVID-19 cases per capita** correlation coefficient = -0.361809
  - **Lockdown score and COVID-19 cases per capita** correlation coefficient = -0.351241
  - **Population density and COVID-19 cases per capita** correlation coefficient = -0.050637
  - **Cluster and COVID-19 cases per capita** correlation coefficient = 0.435473
  - **Cluster and population density** correlation coefficient = -0.212152
  - **Cluster and mask use score** correlation coefficient = -0.744674
  - **Cluster and lockdown score** correlation coefficient = -0.765620
- We would expect population density to be positively correlated with COVID-19 cases per capita, but this is not the case as population density is also positively correlated with COVID-19 response measures, like mask usage and lockdowns. **Counties at higher risk are taking more precautions, and the data shows them to be effective.**

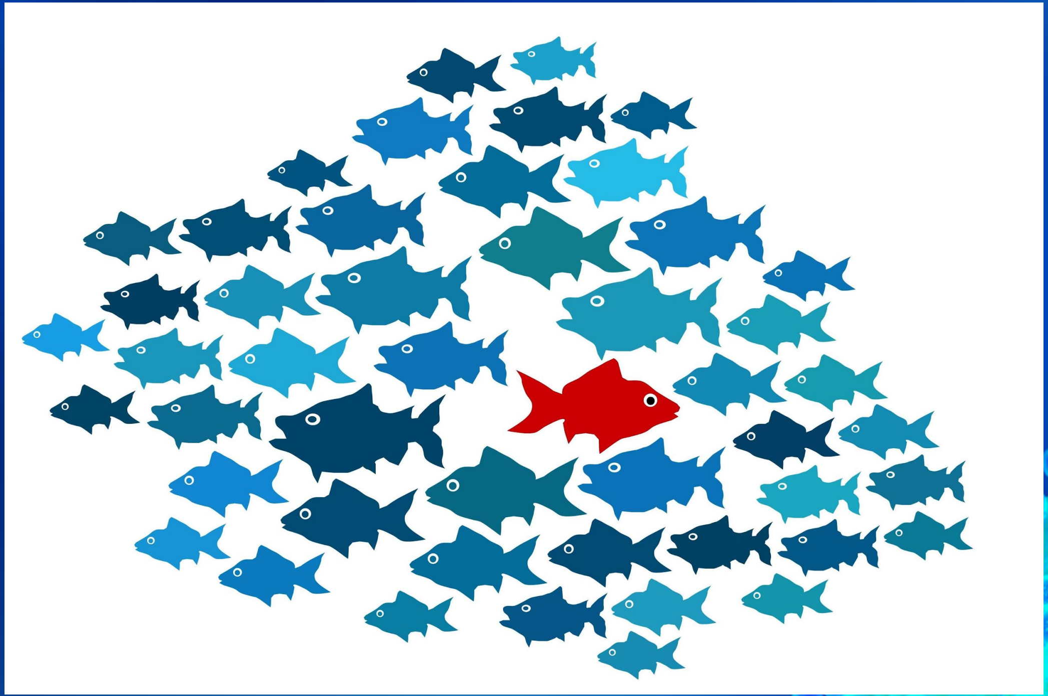
Sec. III

Anomaly Detection

Detect Abnormalities in U.S. COVID-19 Data

Anomaly Detection

Can we detect and explain anomalies in U.S. COVID-19 infections and deaths? Are there new insights about COVID-19 in the U.S. that can be determined by exploring these anomalies?

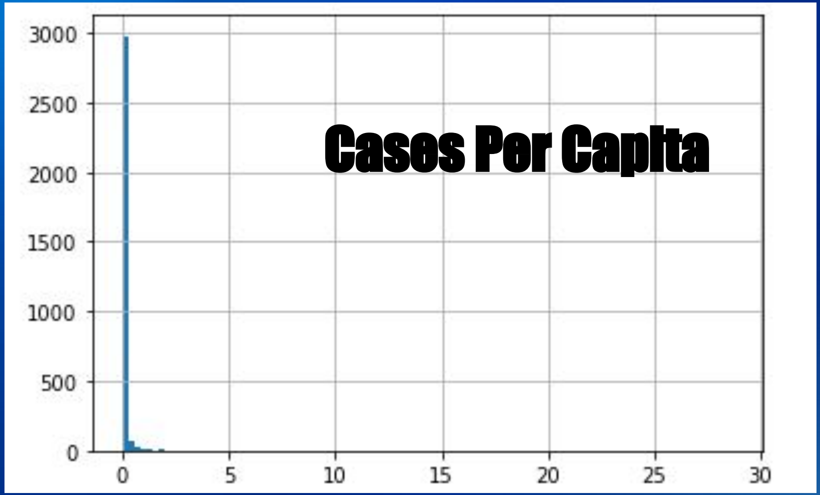


Tie in

- Find anomalies in data to see if we can find any counties that did a good/bad job at preventing COVID spread
- Sanity check data to make sure that there isn't any obvious mistakes

Exploratory Data Analysis

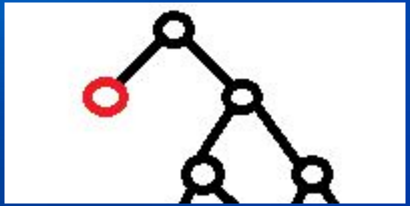
- Created histograms of cases and deaths data per county per capita to check to see if data matches any known models/pdfs
- Discussed interdependence of cases and deaths
- Decided Classical Methods for anomaly detection was out of play

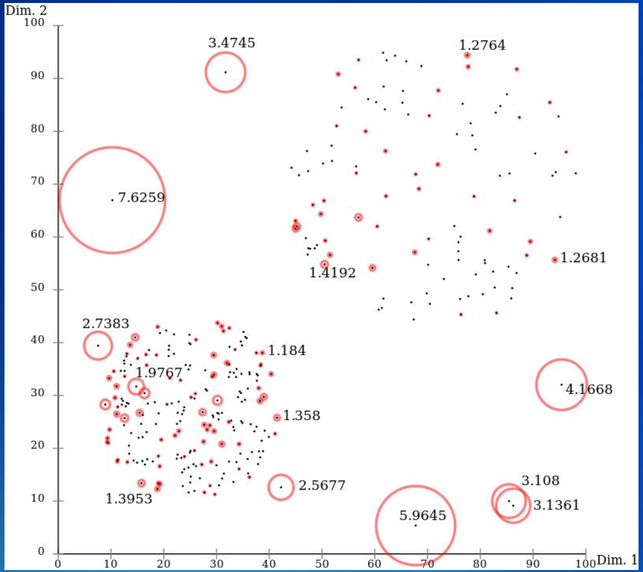


Methods

- Isolation Forests and Local Outlier Factor to find anomalies on both the cases and deaths per county per capita
- Isolation Forests worked well in both cases and deaths
- LOF worked poorly for deaths because of low amounts of deaths

Isolation Forest

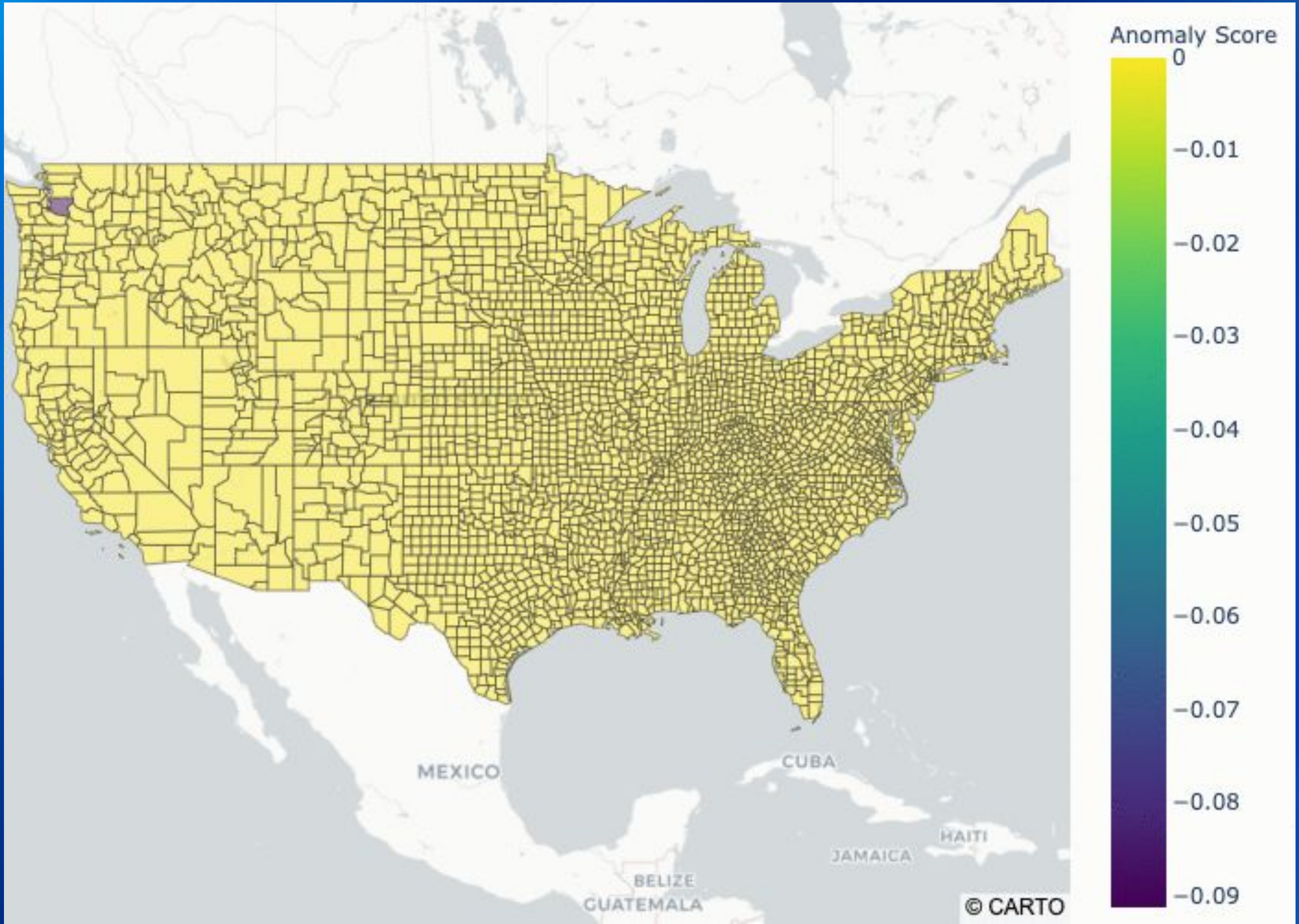




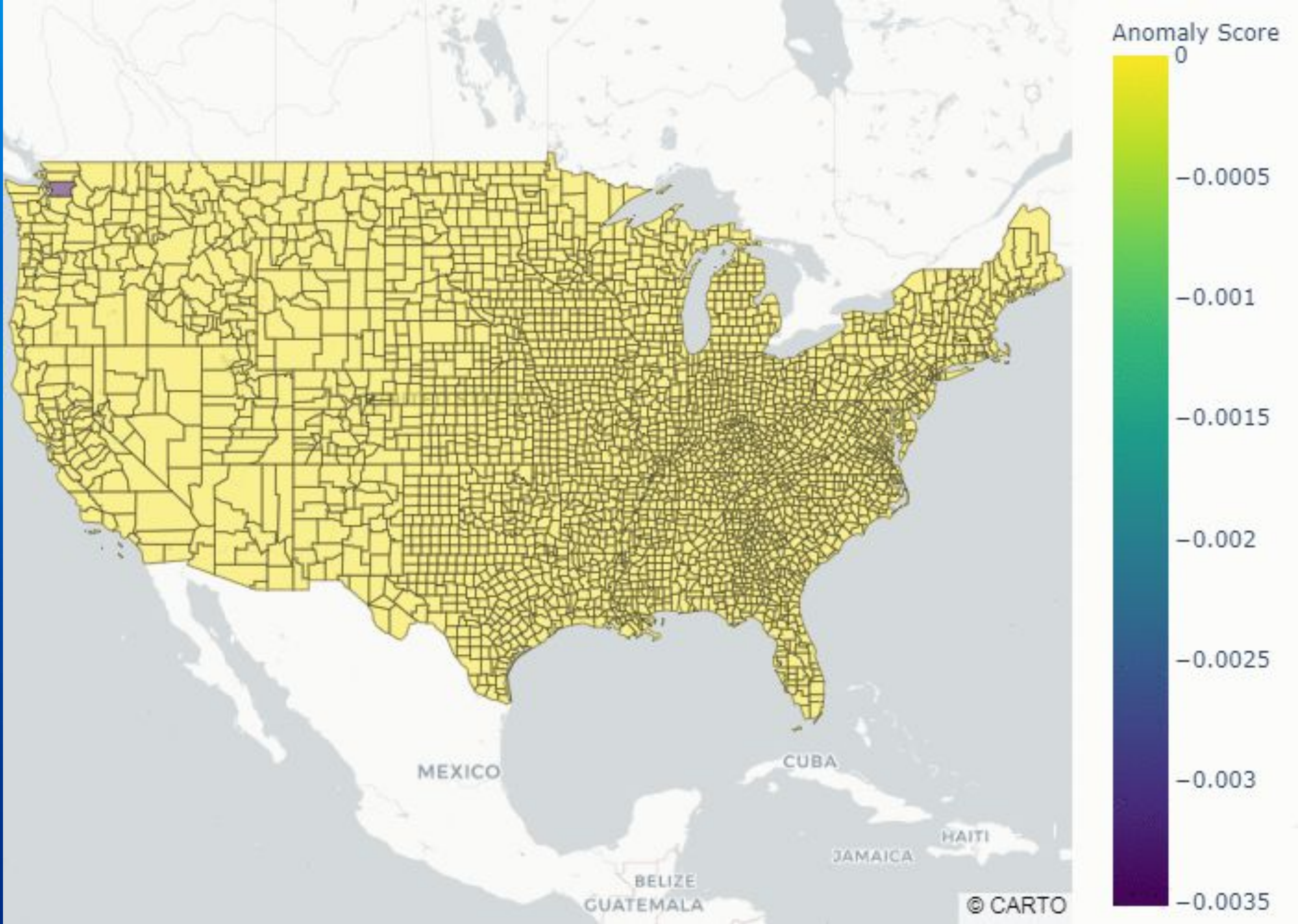
Tools

- Pandas for using dataframes
- Matplotlib.pyplot for creating histograms
- sklearn.ensemble for IsolationForest
- sklearn.neighbors for LocalOutlierFactor
- Plotly.express for animating anomaly data on US county map

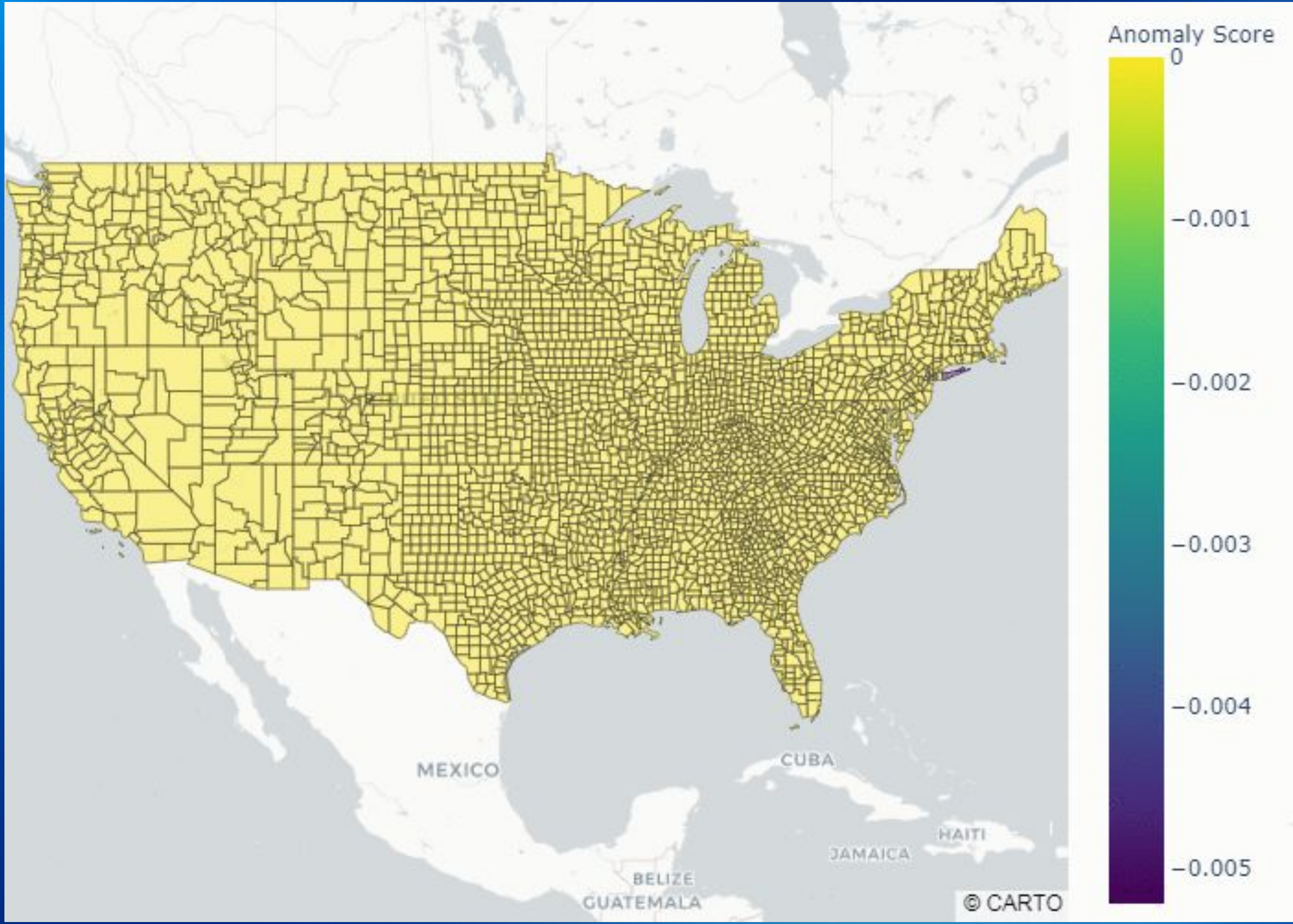
US COVID Death (Isolation Forest) Focal Points



US COVID Cases (Isolation Forest) Focal Points



US COVID Cases (Local Outlier Factor) Focal Points



Highest Anomaly Counties

County, State - Max Cases Per Capita

Suffolk County, New York - 3.371141

Suffolk County, Massachusetts - 3.042953

Westchester County, New York - 3.022067

Cook County, Illinois - 3.059203

Los Angeles County, California - 3.045838

Prince George's County, Maryland - 3.140716

Sussex County, Delaware - 2.914094

Miami-Dade County, Florida - 3.078938

Providence County, Rhode Island - 3.05906

Maricopa County, Arizona - 3.075498

San Bernardino County, California - 3.025381

San Diego County, California - 3.019064

Broward County, Florida - 3.038232

Milwaukee County, Wisconsin - 3.02498

Dallas County, Texas - 3.041544

Harris County, Texas - 3.018299

Orange County, California - 3.041009

Riverside County, California - 3.030819

Summit County, Ohio - 3.02349

Mecklenburg County, North Carolina - 3.018874

Santa Clara County, California - 3.035557

Sumner County, Tennessee - 3.041611

Bexar County, Texas - 3.04314

St. Louis County, Missouri - 3.028151

Prince William County, Virginia - 2.913647

Tulsa County, Oklahoma - 2.913457

Orange County, Florida - 3.023089

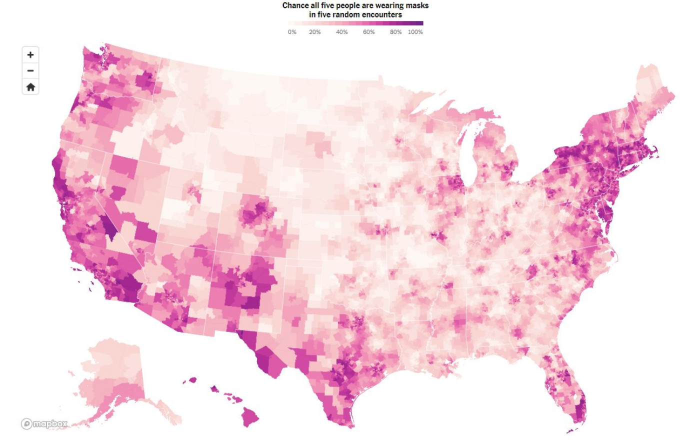
Marion County, Indiana - 3.022992

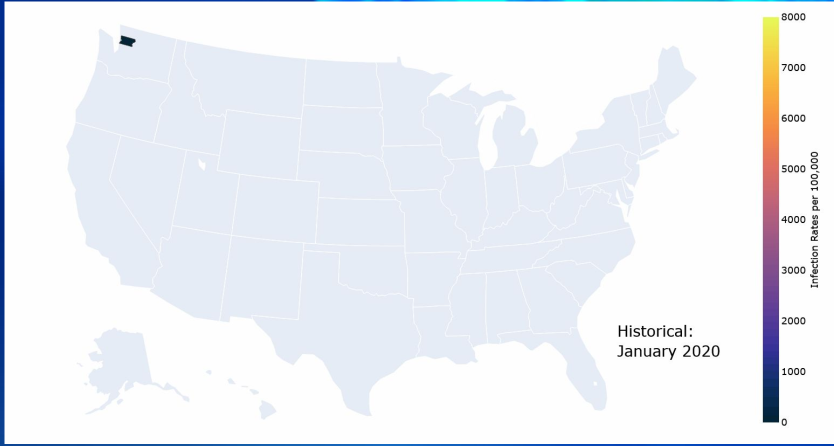
Hennepin County, Minnesota - 3.052729

Results

- The counties with highest chance of anomalies tended to be densely populated counties
- Counties hit early seemed to have a high chance of an anomaly even after they recovered from the initial hit

Expectations vs. Reality





What did our end results yield vs. what did we expect?

Thank You

Questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik