Targeting Covid Remediation

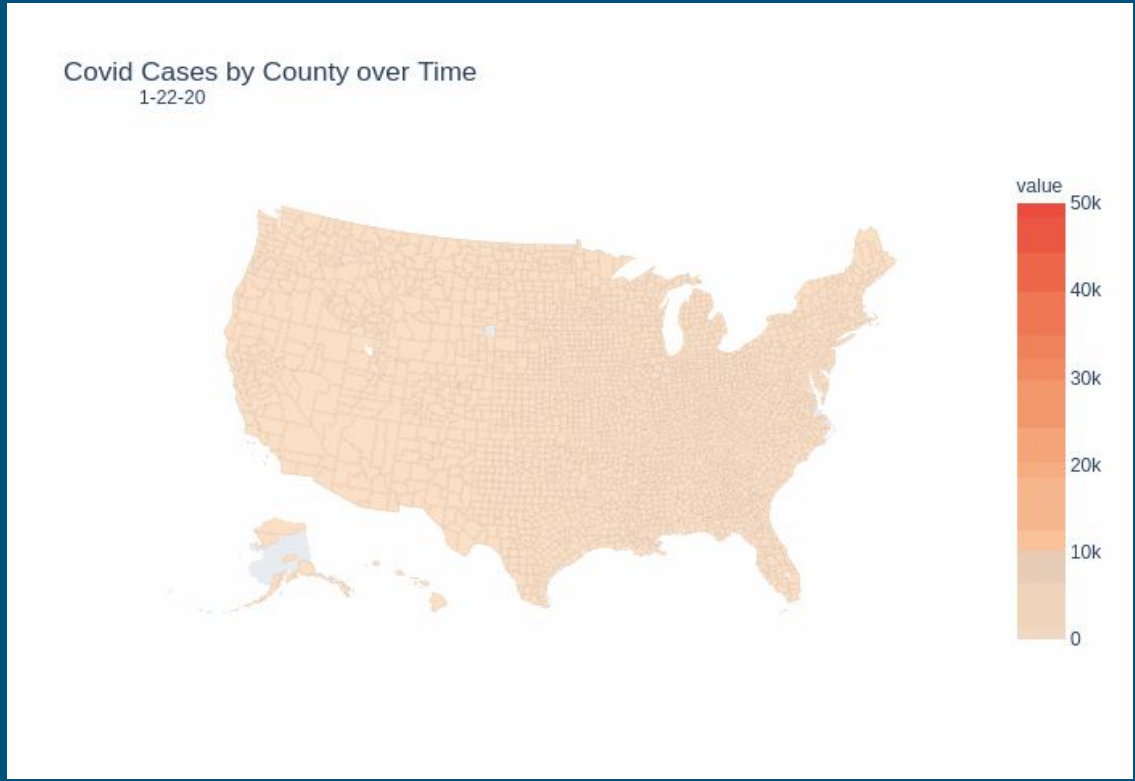
Amanda, Jennifer, Sam
DSIR-Nightingale (Project 5)

Problem Statement

- We are a team of data scientists who have been hired by a private national aid organization to identify and assist counties that are being hit hard by COVID-19. Though the CDC's dataset is comprehensive, there is a high degree of uncertainty when forecasting the spread of the virus. By conducting an independent audit of this data, we aim to improve this assessment model and provide geovisualization tools, so resources can be directed to areas of need more quickly than has previously been possible.

Background

- As we all learned from COVID-19, pandemics spread quickly and can be difficult to track.
- In addition to tracking where the virus is growing or falling, can we detect any patterns, whether it be higher vaccination rates or stricter social distancing regulations?



How we targeted our efforts

Three-prong approach:

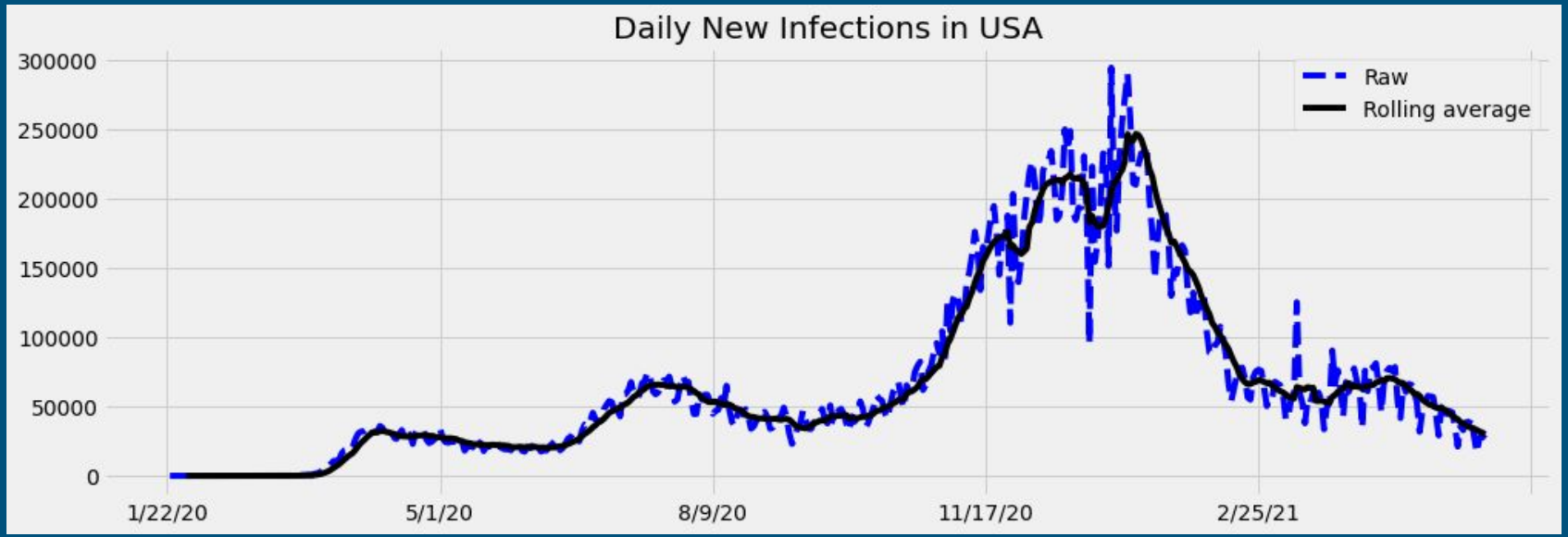
Context

Verification

Impact

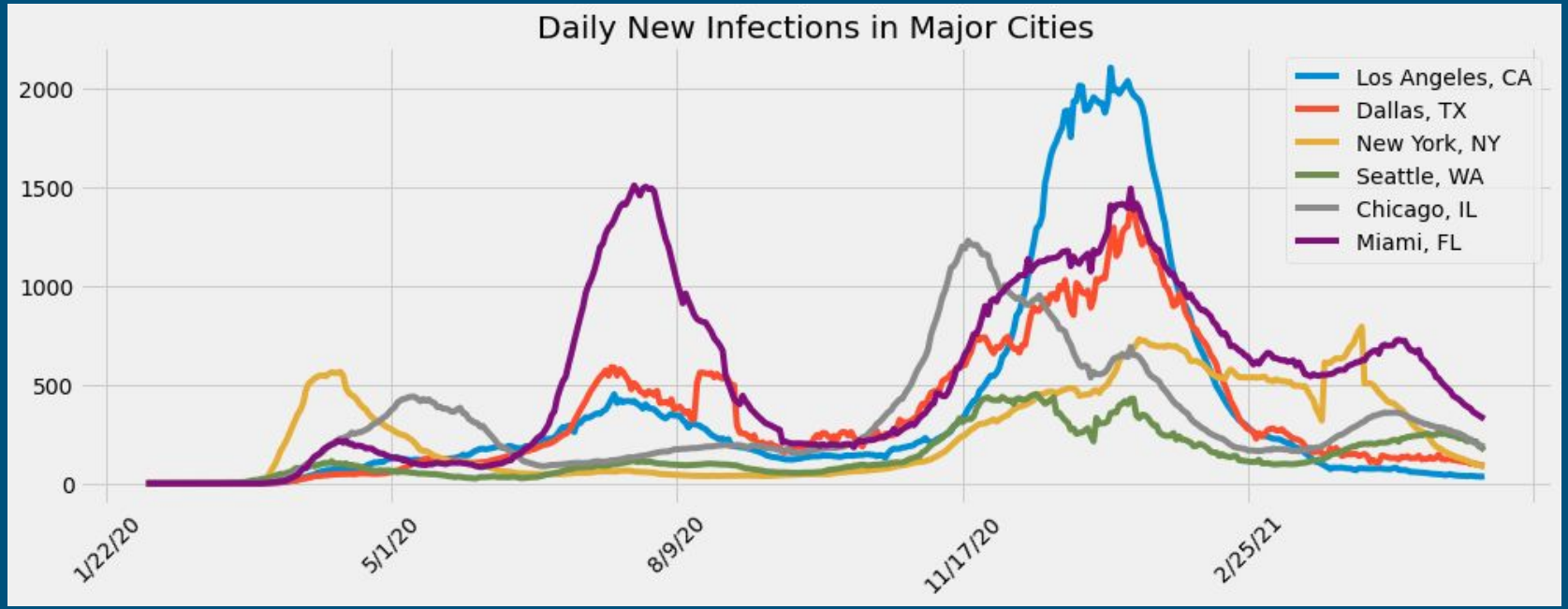
Source: Flaticon

First approach: Assessment Tools



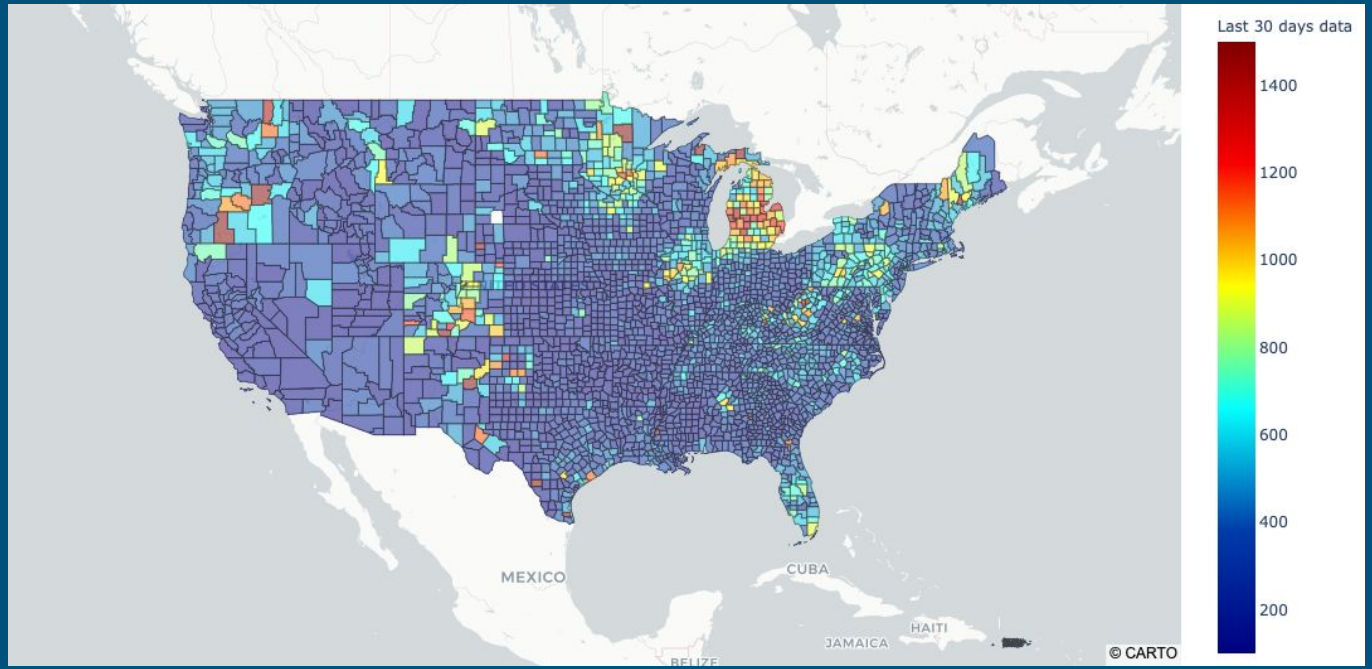
Top-level trends help indicate the country's overall success in combating the virus but provide little in the way of granularity.

First approach: Assessment Tools



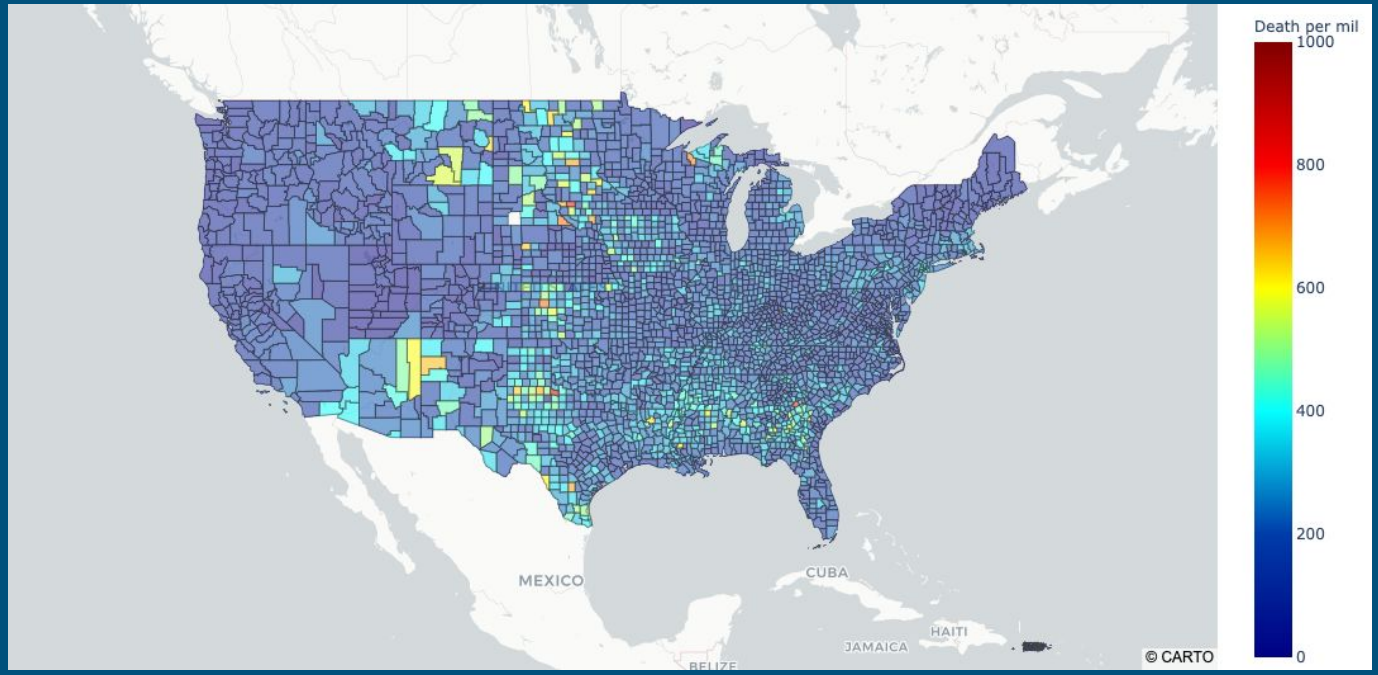
Zooming into different cities in can provide insights into where outbreaks are occurring as well as which policies may or may not be working to control the spread.

First approach: Assessment Tools



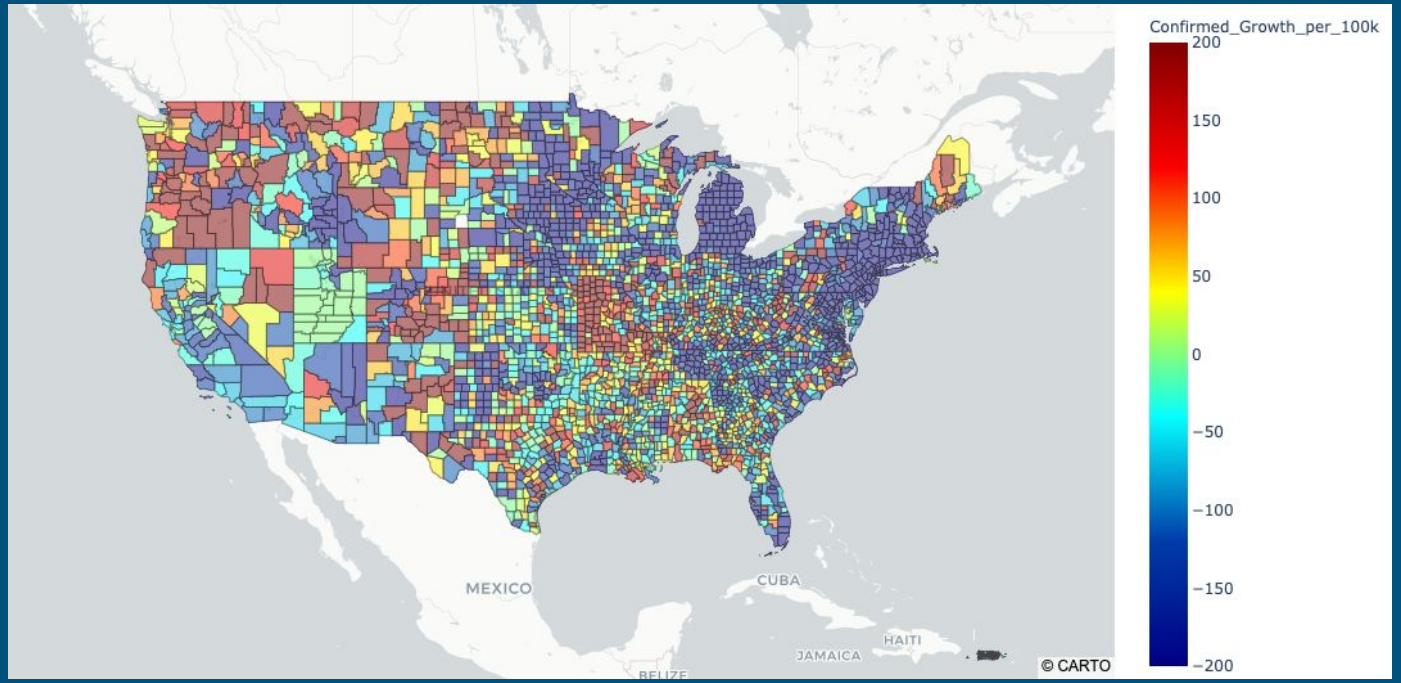
Confirmed cases have declined across the country with the rollout of vaccines, but there are still several pockets of hotspots.

First approach: Assessment Tools



Per 1 million residents, deaths have been higher in rural counties, particularly in the south and midwest.

First approach: Assessment Tools



As of a month ago, growth in cases has declined in urban areas while increasing in rural counties, particularly in the midwest, and pacific northwest.

Second approach: CDC Verification

Verifying government data

- Official reports (CDC)
  - Almost daily Excel worksheets by County (3k)
  - Data publish began in mid-December 2020
  - Examination Range: 1/02/2021 - 4/30/2021
  - Features include:
    - Changes in % cases, % deaths, %vaccinations
    - Population vulnerability scores
    - Hospital ICU/ventilator usage

Second approach: CDC Verification

Verification Target: Observed County Trajectories

	Value Counts
Likely Decreasing	~10.9%
Uncertain	~89.0%
Likely Increasing	~.1%

As we can see, the CDC had a high degree of uncertainty. We were tasked to check these results to see if we could have a better predictor.

Feature engineering was done to diff day-to-day feature values to capture more of the trajectory at row level

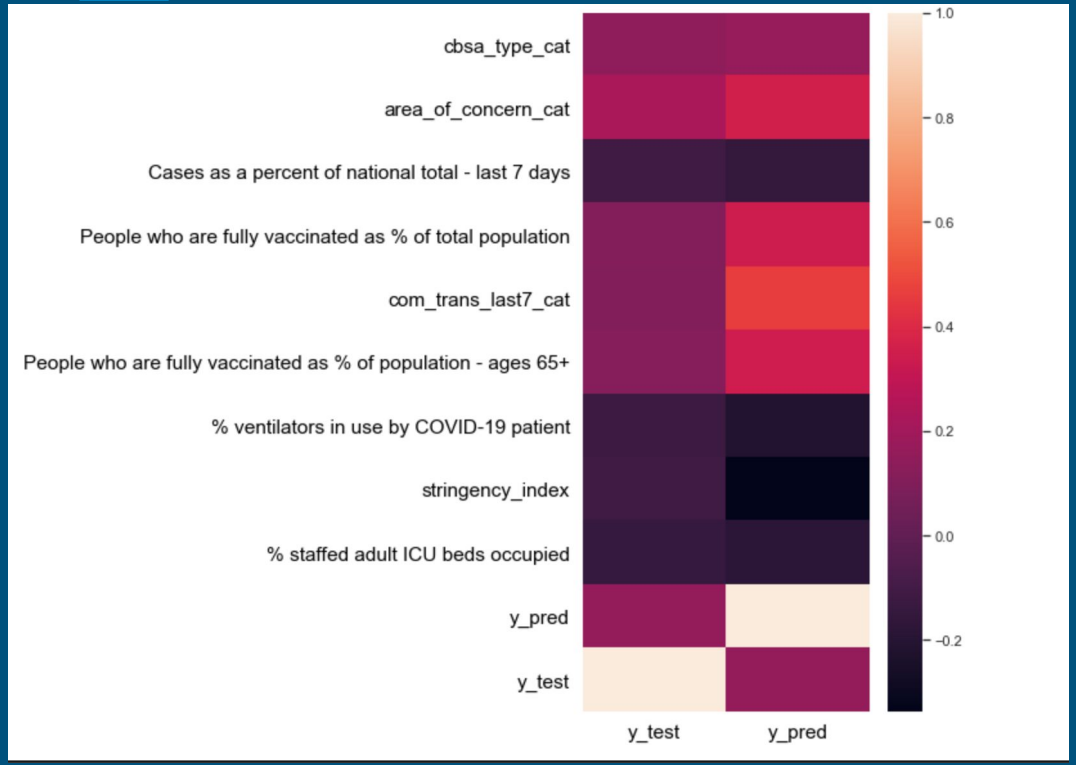
Second approach: CDC Verification

Imbalanced Data

Limitations

- Multinomial (multi-class) Logistic Regression
  - Could not use liblinear solver due to classes
  - Saga could not resolve
  - Accuracy score was .889
  - 99.8% predictions were the majority class compared to 88.9% (true)
  - “Likely increasing” category was not predicted
- SMOTE was used to balance out data.
  - the model accuracy went down to .541
- class balancing, accuracy was at .470

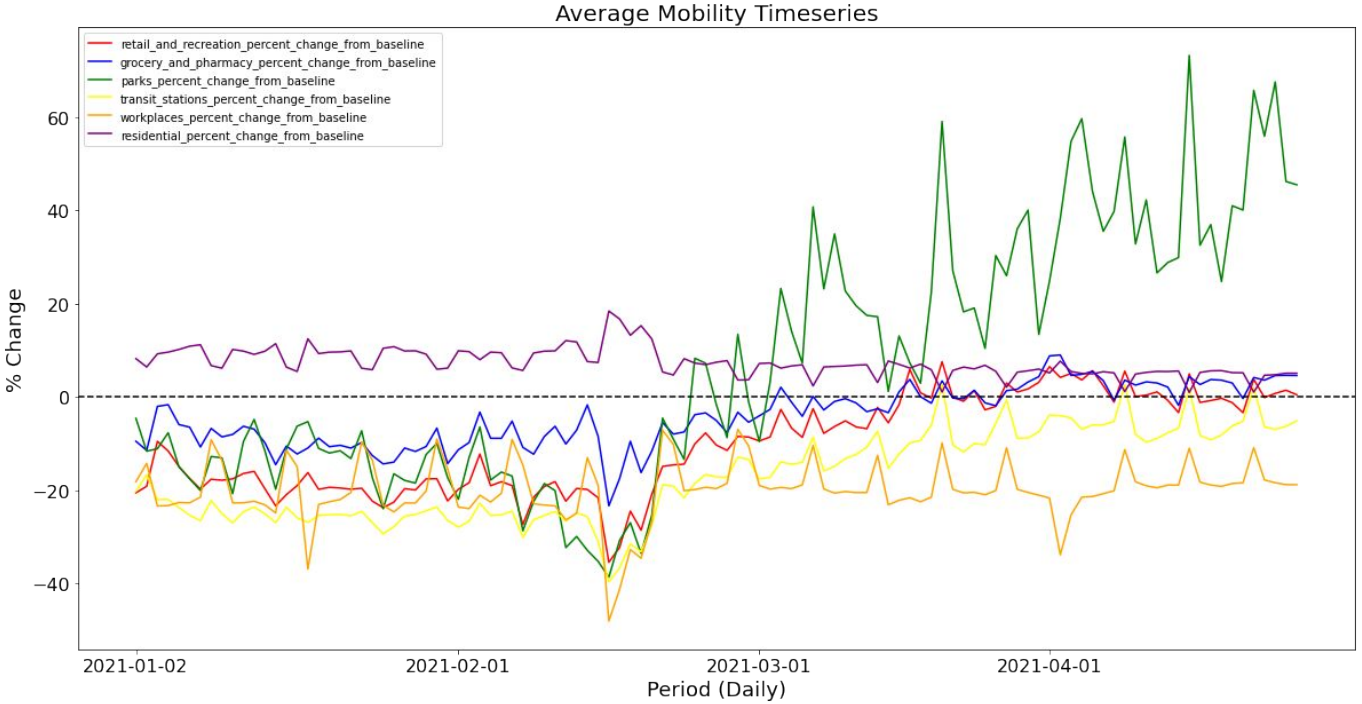
Second approach: CDC Verification



CDC County Correlations

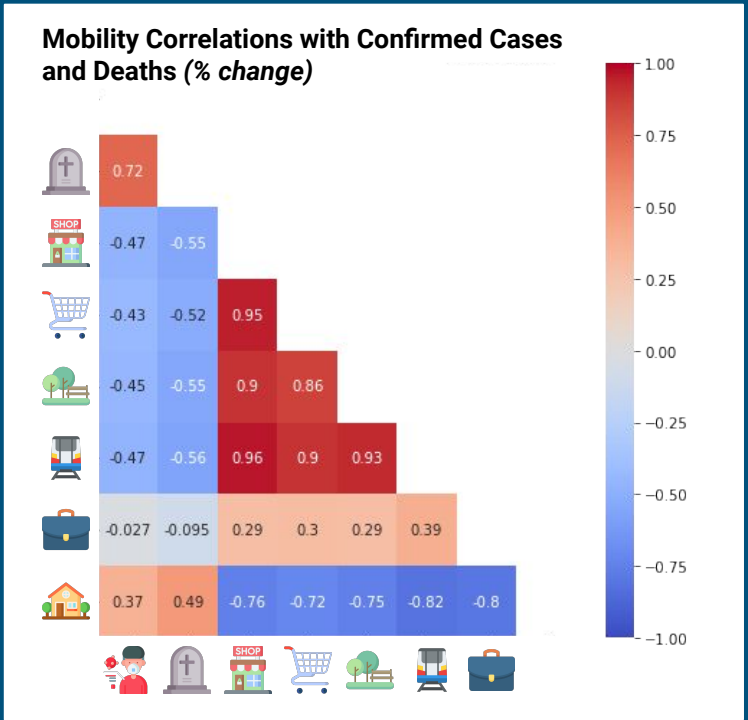
Since the target was the case trajectories, it would seem that this would be reflected in features such as % case/death changes. Instead we can see that this was more affected by area of concern, vaccinations, ICU bed availability

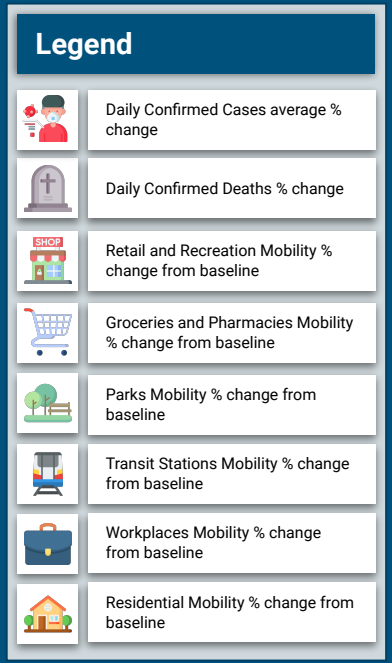
Third approach: Social Impact - Mobility



Looking at the graph, we can see that aside from mobility within residential areas, most locations' % change in mobility vs baseline (baseline being pre-covid numbers) is negative. Average mobility for these areas have decreased by around -6.2%. However, with vaccine roll outs in 2021, an upward trend in mobility can be observed - especially for areas like parks, followed by retail and recreation, and groceries and pharmacies

Third approach: Social Impact - Mobility

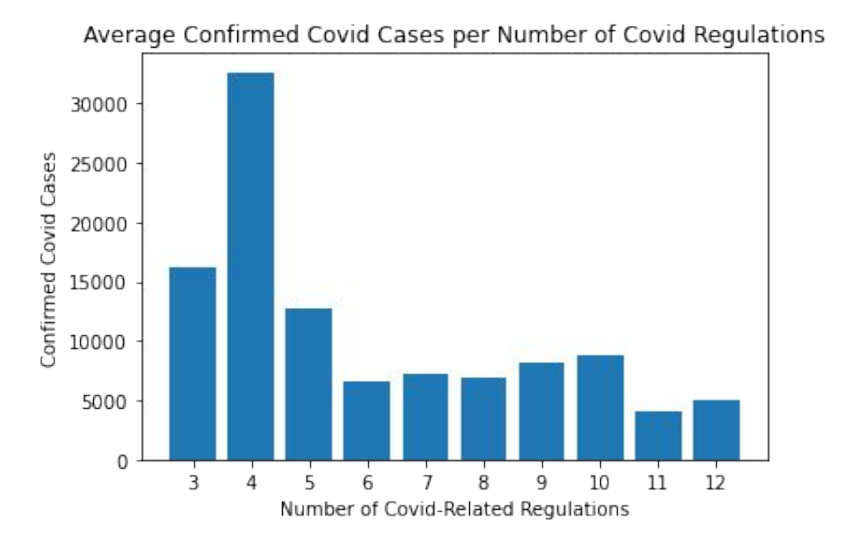


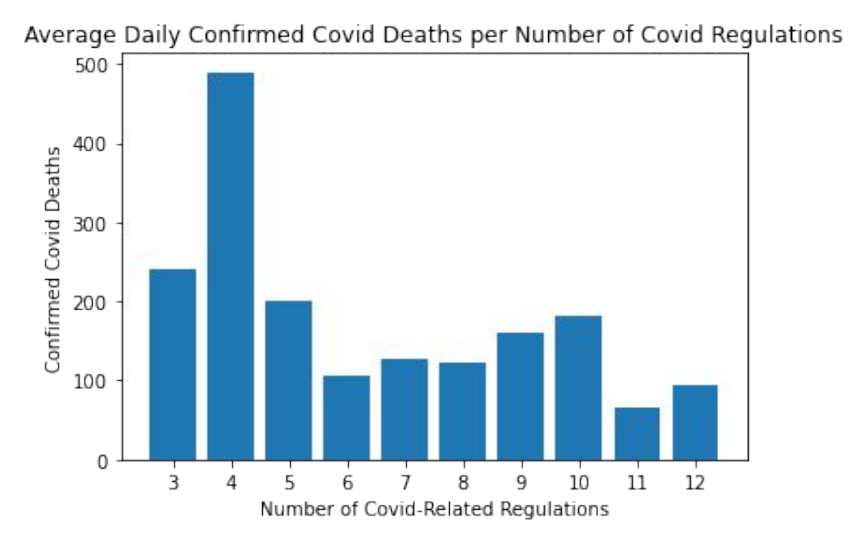


- We can see a strong negative correlation between mobility in public areas and the numbers of covid cases and deaths.
- What this means is that as cases rise, people are more likely to stay close to home.
- The reverse is also true - as the virus slows, we're seeing the public embrace a "new normal," returning to the public areas classified in our four categories.

Source: Flaticon

Third approach: Social Impact - Regulation





More regulations result in less confirmed cases and relatively less deaths. Although, **correlation does not always equal causation!**

Third approach: Social Impact - Unemployment

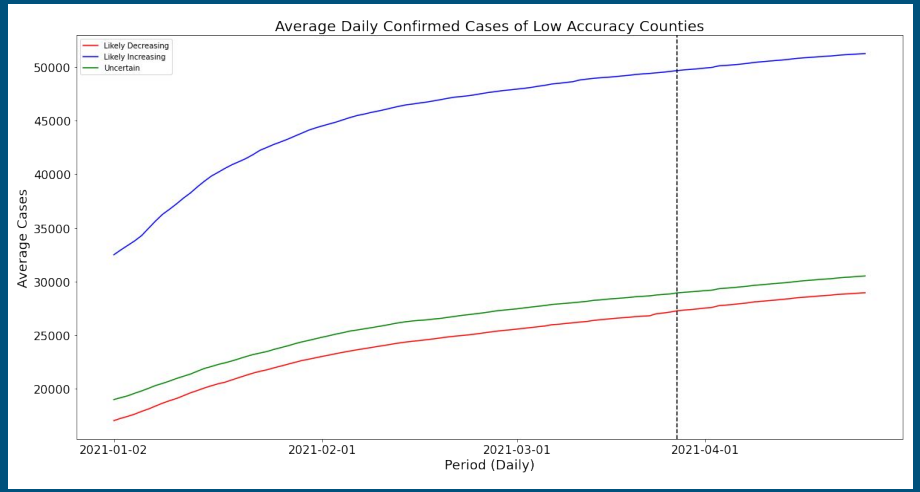
	Avg. Unemployment Rates	Avg. Daily Confirmed Cases	Avg. Daily Deaths
10 Counties with Lowest Unemployment Rates	1.4%	~73	~2
10 Counties with Highest Unemployment Rates	17.5%	~16.2K	~645

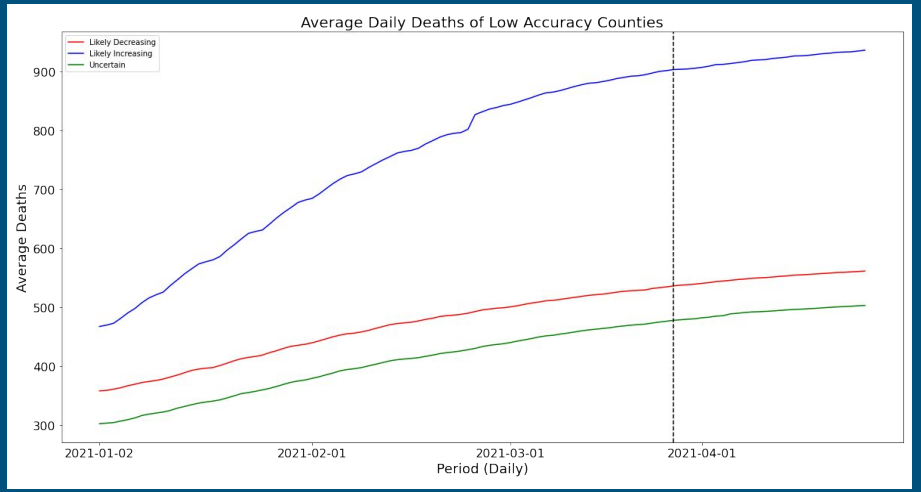
Looking at the numbers, counties with higher average daily confirmed cases and deaths have higher unemployment rates which may be a result of their economy being hit harder by the pandemic.

Modeling Summary

- Logistic Regression
- Random Forest
- Neural Networks
- Recurrent Neural Network

Modeling Summary: RNN





- 85% Accuracy - worse than Baseline
- Majority Class of Inaccurate Prediction: Likely Decreasing (~45%)

Key Takeaways

- Regulation/Restrictions correlated to decrease in Covid cases and deaths
- Rural areas continue to experience increases in Covid
- Government uncertainty is linked to data delays
- John Hopkins academic data provided more certainty and greater re-use
  - More narrow focus, but consistent reporting
- Government data is not geared towards public re-use
  - Format not accessible
  - Bias behind decisions not clear

Next Steps and Recommendations

- Live dashboards which could help client
- Timely information improvement
- Centralized, federally controlled data warehouse + tool
- Public information should be accessible and easily digestible