MORTALITY RATES IN 2020

USING TIME-SERIES MODELS TO PUT STATE NUMBERS IN CONTEXT

Luken Weaver, General Assembly, DSI Immersive, 6/9/2020

The big picture

- The 2020 covid-19 outbreak has killed over 100,000 Americans, by official estimates.
- The ensuing lockdown has led to the loss of over 36.5 million jobs.
- Both of these presumably have large-scale ripple effects, making an accounting of the full consequences, their analysis, and the relative risk assessment of different policy approaches difficult.

THE HILL **The economic costs of reopening too soon**

REUTERS

Special Report: How the COVID-19 lockdown will take its own toll on health

The granular level

- Inconsistencies with how key metrics are tracked and reported, and the potential for blind spots toward other markers, further muddy the waters of what can be considered an accurate assessment of the situation.

7 KPLC News

New federal COVID-19 nursing home data fraught with inaccuracies, overcounts and undercounts

The Washington Post

Which deaths count toward the covid-19 death toll? It depends on the state.

The questions

- When controlling for past years in a given area, do we see a noticeable change in overall mortality rate in 2020, compared to what a ‘typical’ year would be?

- Can this change be reliably predicted from other features about the area? Are static features more or less predictive than policy response?

The proposal

- Use historical data to construct a time-series model that will predict death rates for 2020 for each state*
- If 2020 is in fact an aberrant year in terms of mortality, these forecasts will be inaccurate by design
- The *degree* by which the observed numbers differ from the model can be a parameter of interest in itself, representing the degree to which a state has been knocked 'off course' by the crisis
- Can this metric be predicted? Feed possible explanatory features into a supervised learning model and look for patterns.

** Comparing either smaller-scale (cities) or larger-scale (countries) areas would likely be a better fit for the second part of this project. Due to the constraints of available data, states were chosen as the unit of analysis as proof of concept.*

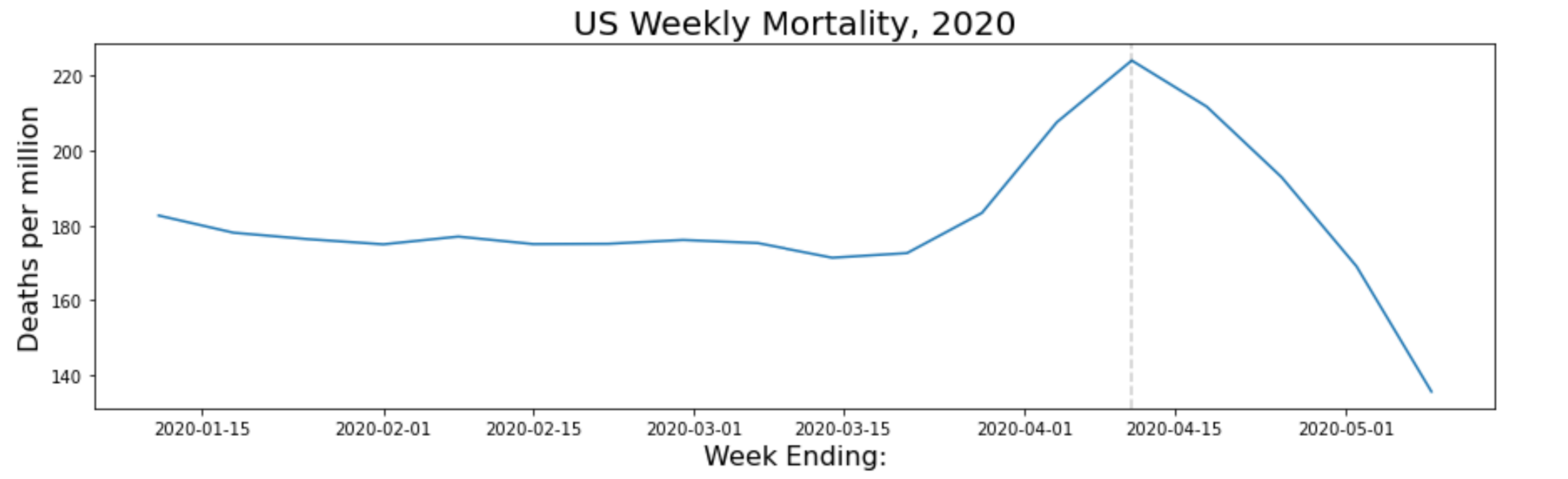
The data

- The Economist’s James Tozer and Martín González author and maintain a Covid-19 Excess Deaths tracker github. The historical weekly death rates they collated from the CDC will be the basis of part one of this analysis.
- For part two, select data was gathered about each state from the Department of Labor, Census Bureau, and independent websites listed in the citations section at the end of this presentation.

Part One

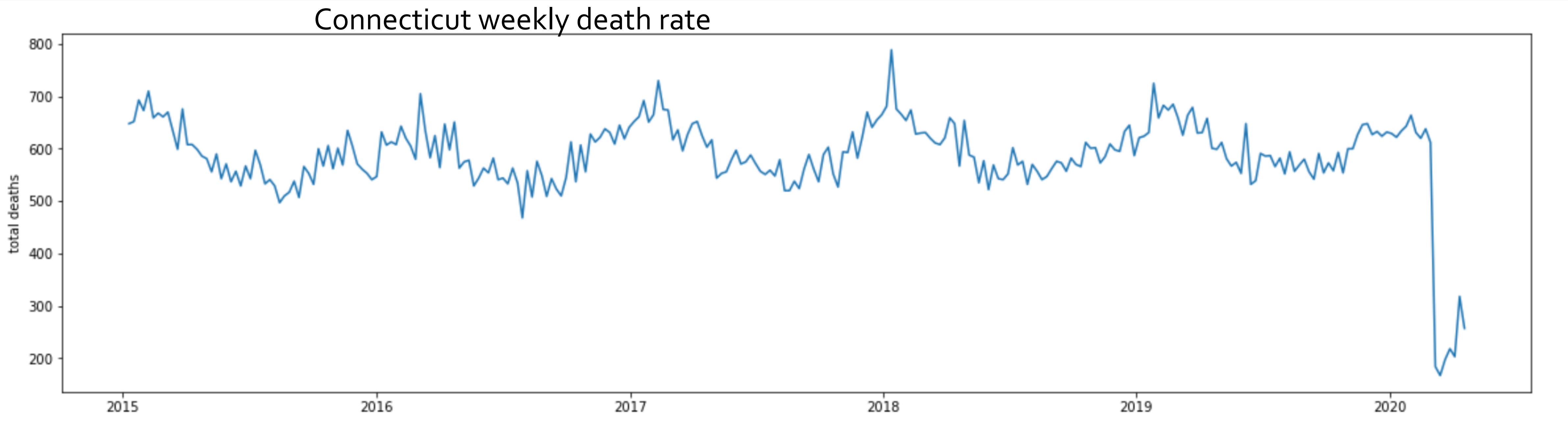
Time Series

Data problems



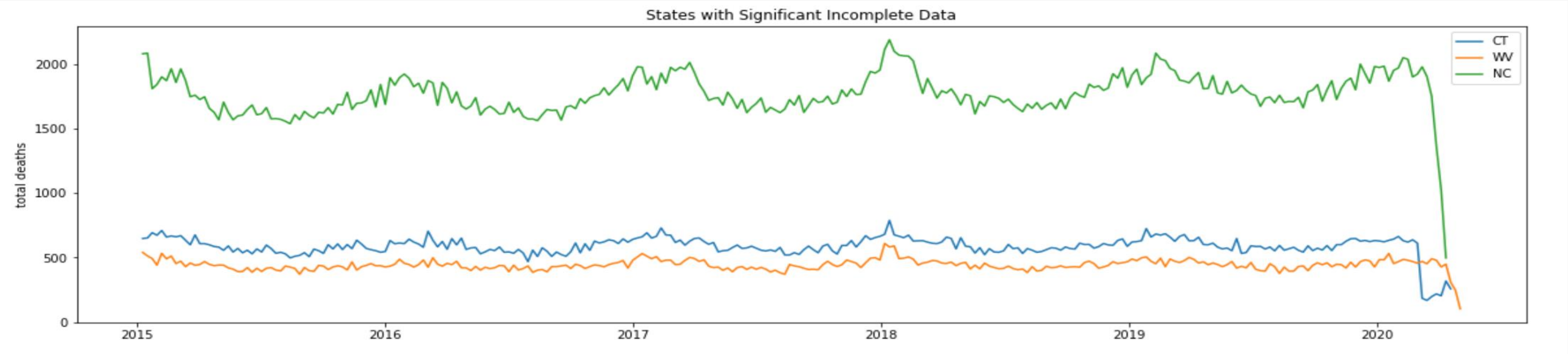
Death rates spike in the week ending April 11th. A come-down afterward is expected. However, we are seeing death rates drop significantly below averages for the year.

Not just missing values



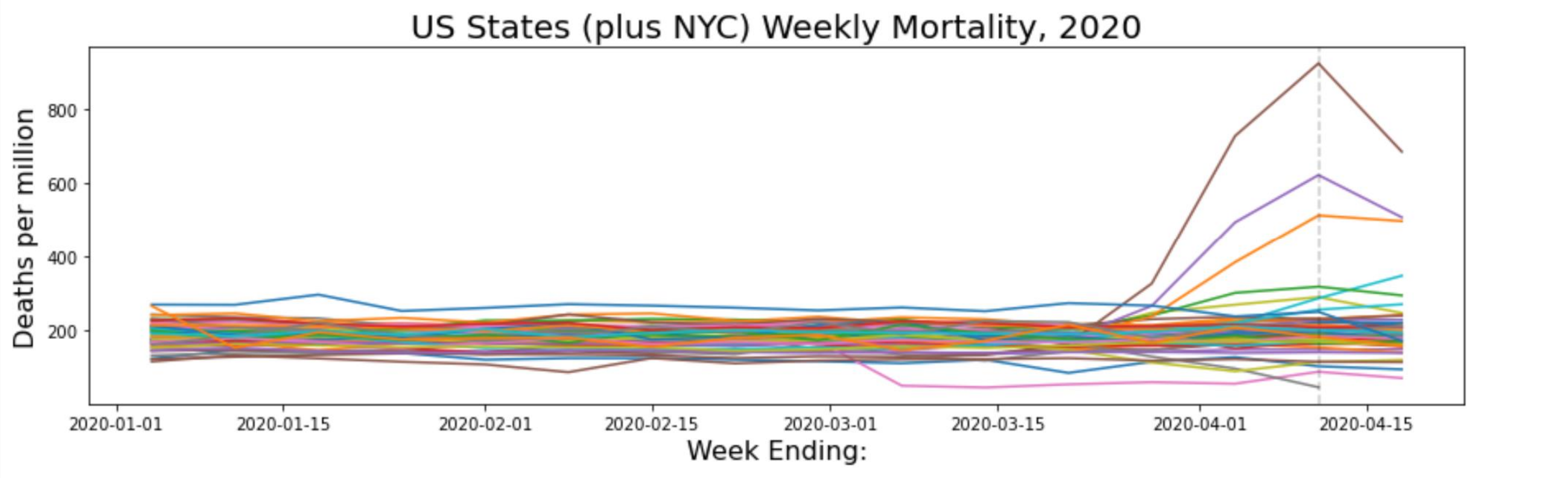
Connecticut leaps out as the most severe example. In addition to some missing values that had not yet been reported as of the data pull, the numbers that had been reported are significantly lower than average at precisely the time that one would expect them to be very high, from reports of covid-19's spread.

Not just Connecticut



And while not overly wide-spread, CT was not the only state that presented this kind of non-credible data set. States are highly inconsistent with their reporting, making more recent data shakier in its reliability than it might seem.

Cutting down the test data

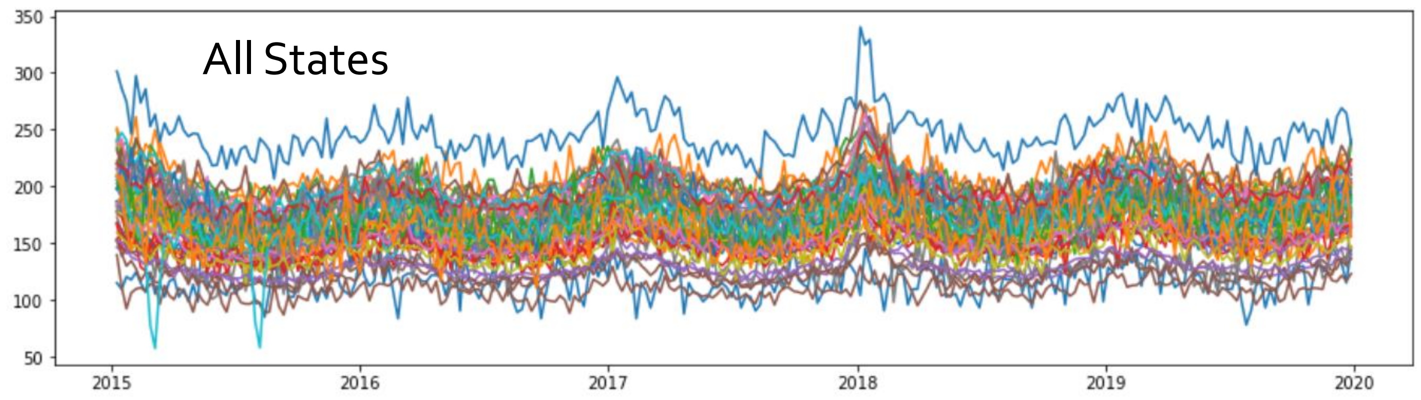


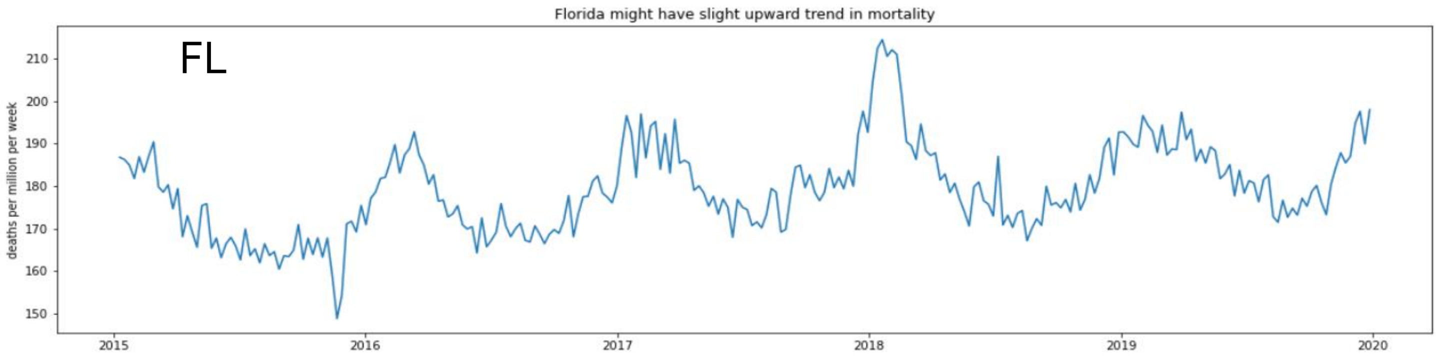
Ultimately, the decision was made to cut off the last three weeks of our dataset as unreliable in general, leaving the above as our 2020 data that our model's forecast will be compared to.

How do we handle training data?

Dickey-Fuller edge cases

	Test Statistic	p-value
AZ	-3.305209	0.014647
FL	-2.818836	0.055633
GA	-3.421682	0.010242
KS	-3.184181	0.020919
MN	-3.173855	0.021549
ND	-3.401563	0.010906
NM	-3.369659	0.012038
NV	-3.420847	0.010268
OH	-3.071547	0.028732
OK	-3.213938	0.019192
TN	-2.890752	0.046434
TX	-3.314551	0.014240





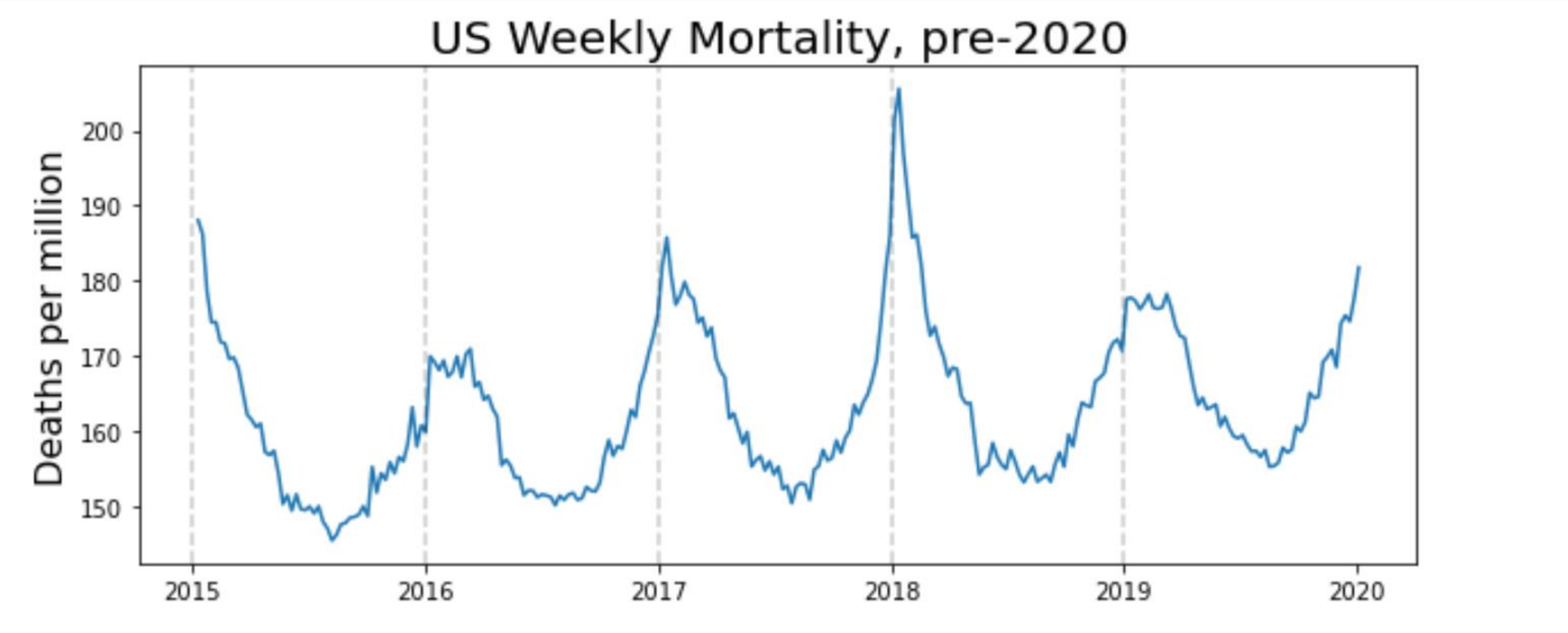
Most states, and the country overall, seem stationary, but some would fail a Dickey-Fuller test at sigma = 0.1

Florida fails even a more relaxed sigma = 0.5

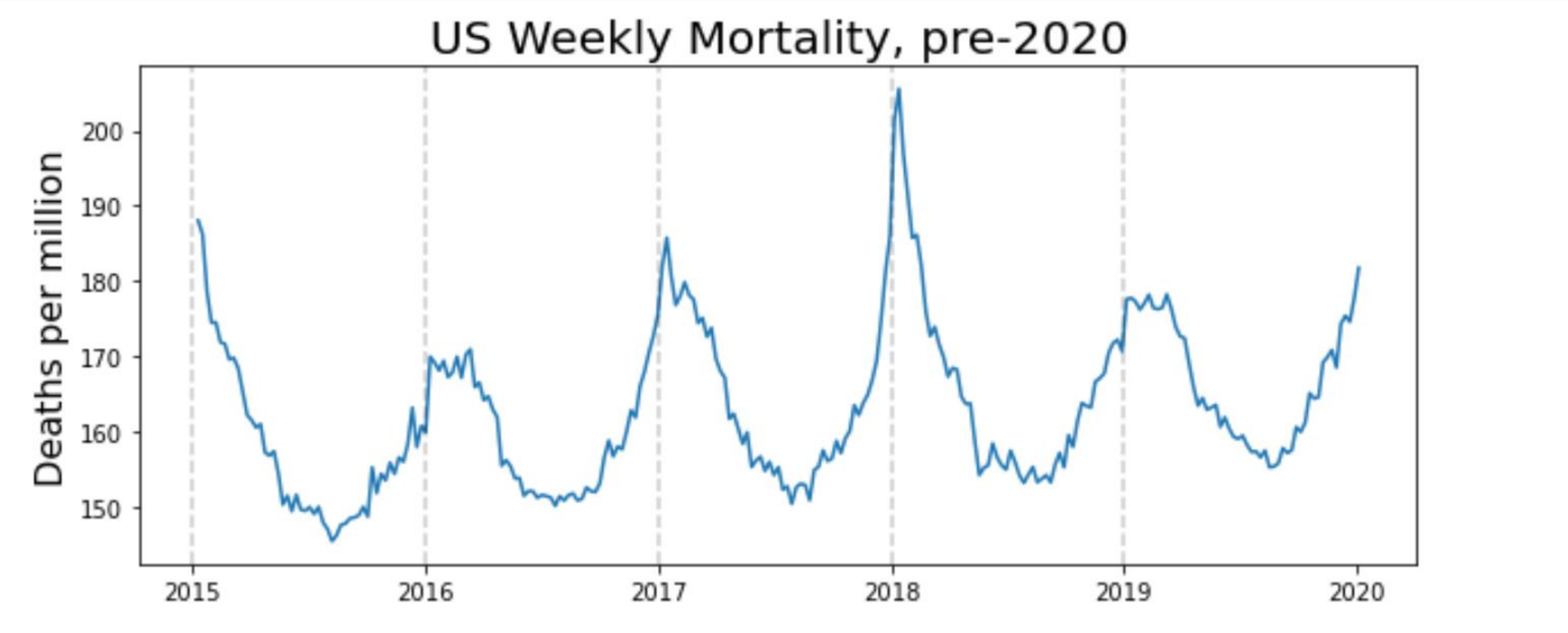
What do we train our model on?

- Using one model to forecast all states would miss any dynamics happening at the individual state-level, defeating the purpose of comparison. Each state needs its own model.
- However, there would be concerns over consistency if comparing the results of two differently designed models.
- Additionally, grid-searching to find the best parameters for each state would be computationally intensive.
- The decision was made to find tune a model to best predict trends at a national level, then use those parameters to fit on each state’s training data (i.e. death rate prior to 2020).

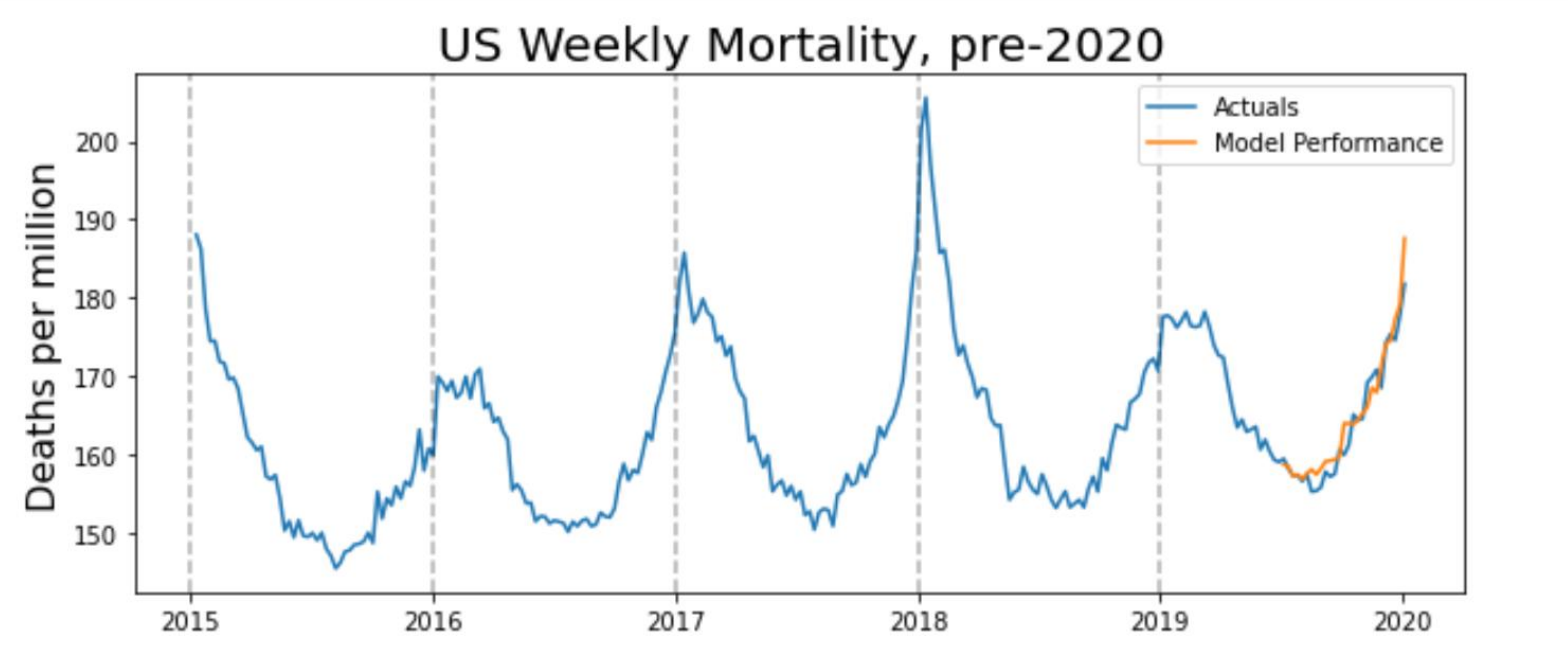
Our training data. Commence grid-search…



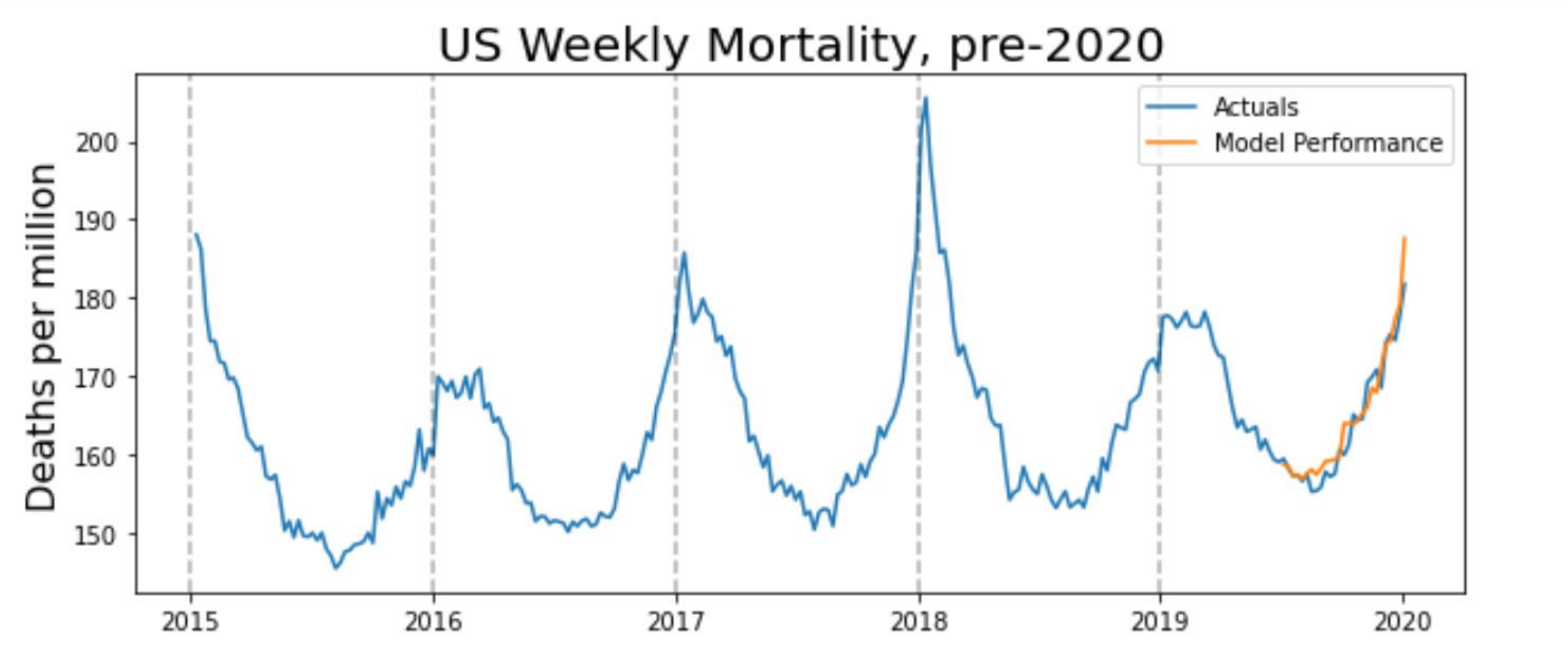
50+ hours later…



50+ hours later…



50+ hours later…



An Arima model order (3, 1, 1) with a seasonal component of order (3, 1, 0) received our best AIC score*

** some models did in fact score slightly better, but as they received convergence warnings and had features scored with concerningly high p-values for significance, they were discarded*

Converting forecasts into a parameter of interest

- A model with the parameters above was fit to each state's weekly death rate data for the year, and this model was then used to forecast the data for 2020 as far as the data that we do have was deemed reliable enough.
- The mean of each state's residuals was then calculated, deriving the target metric for each state, termed Death Rate Change (or DRC).

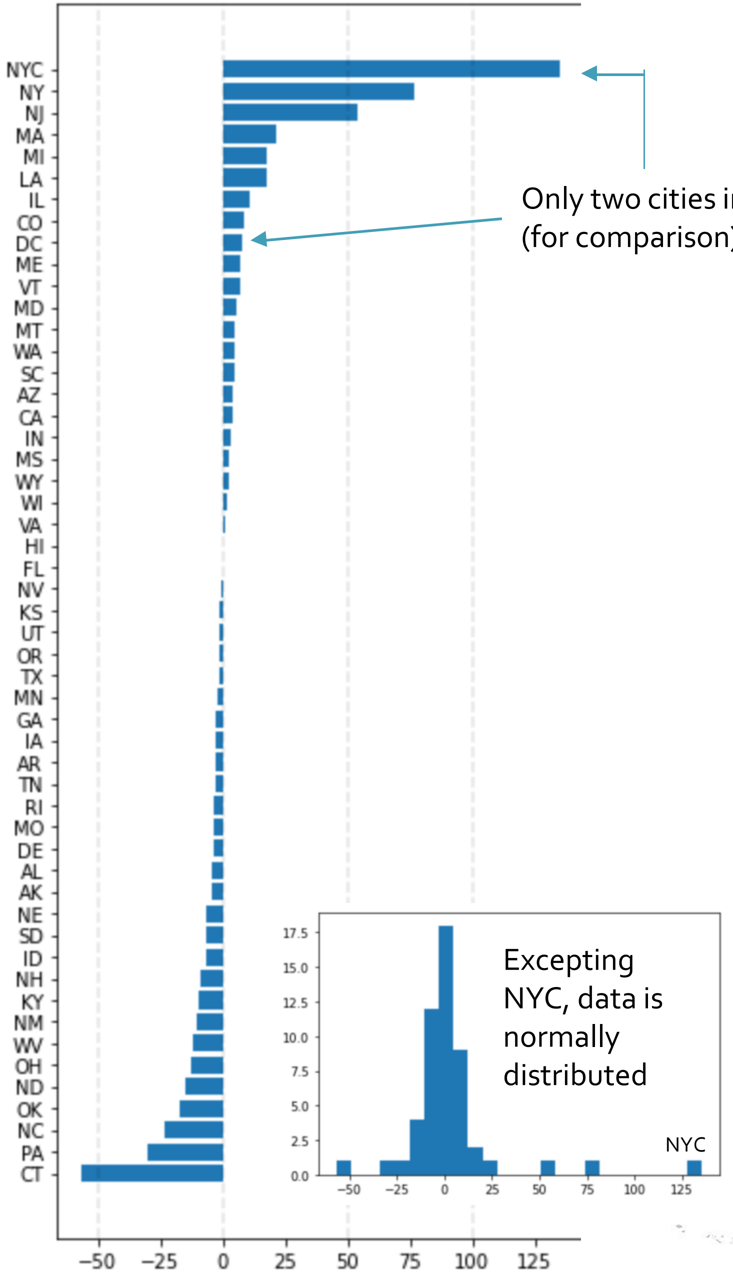
Results

Taking out the two cities (NYC and DC), exactly half of US states experienced fewer reported fatalities from all causes so far in 2020 than our model forecast.

death_rate_change_per_mil	state
135.308	NYC
76.4311	NY
54.0089	NJ
21.7192	MA
17.8868	MI
17.4065	LA
10.9861	IL
8.73387	CO
8.08606	DC
7.25544	ME
6.827	VT
5.43214	MD
5.04829	MT
4.7423	WA
4.63736	SC
3.89358	AZ
3.77528	CA
3.03703	IN
2.60094	MS
2.42269	WY
1.36567	WI
1.02514	VA
0.193375	HI
0.0977752	FL
-0.838527	NV
-1.12528	KS
-1.19768	UT
-1.39887	OR
-1.64559	TX
-2.13366	MN
-2.6343	GA
-3.00581	IA
-3.05901	AR
-3.10539	TN
-3.37742	RI
-3.45979	MO
-3.8229	DE
-4.13449	AL
-4.53104	AK
-6.3114	NE
-6.61537	SD
-6.74037	ID
-8.84692	NH
-9.82429	KY
-10.4693	NM
-12.1999	WV
-12.9971	OH
-15.3159	ND
-17.1047	OK
-23.7157	NC
-30.5055	PA
-56.8809	CT

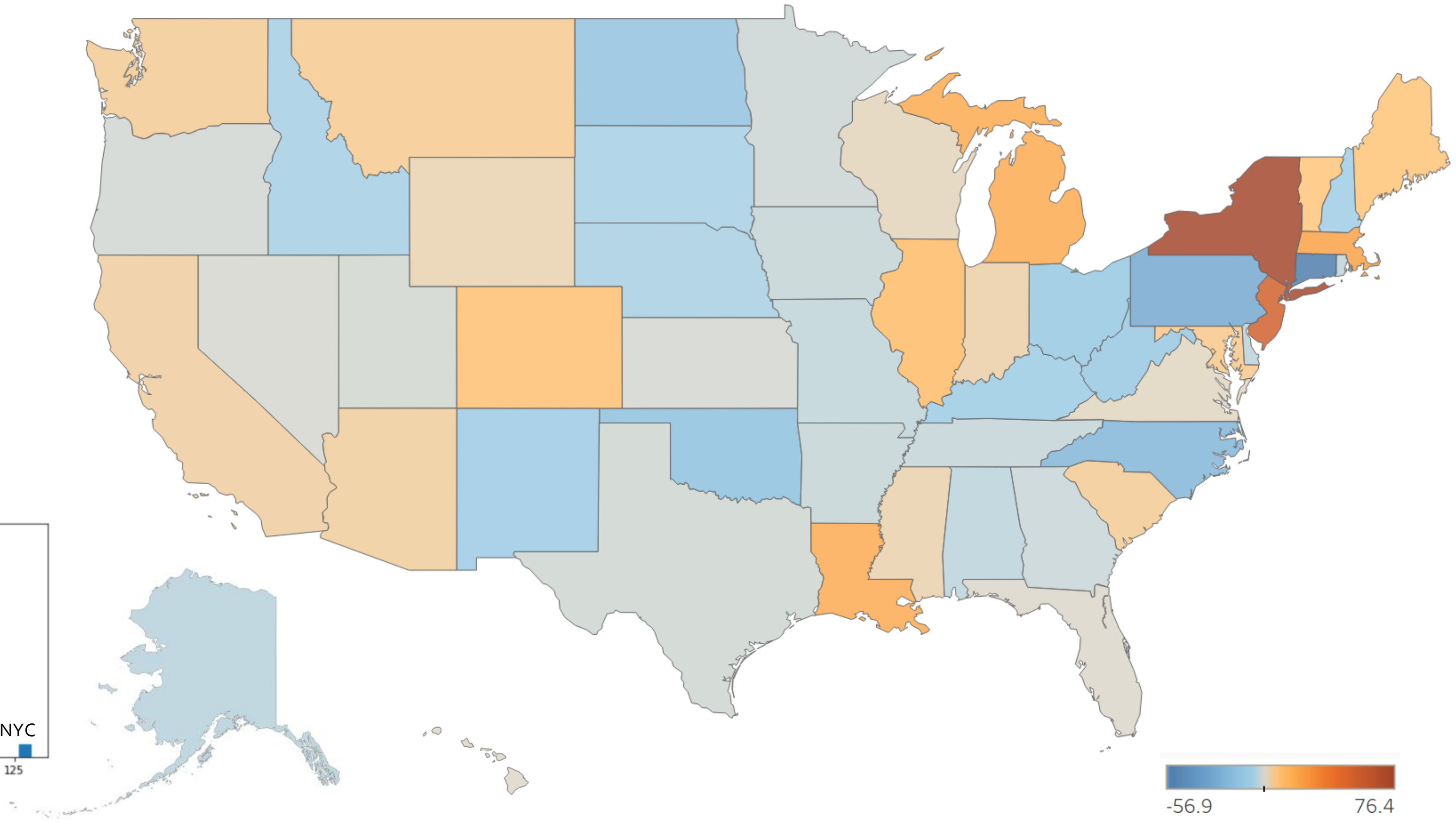
Results

Death Rate Change, per million population, US



Only two cities in our data set
(for comparison)

Excepting
NYC, data is
normally
distributed



Part Two

Regression Modeling

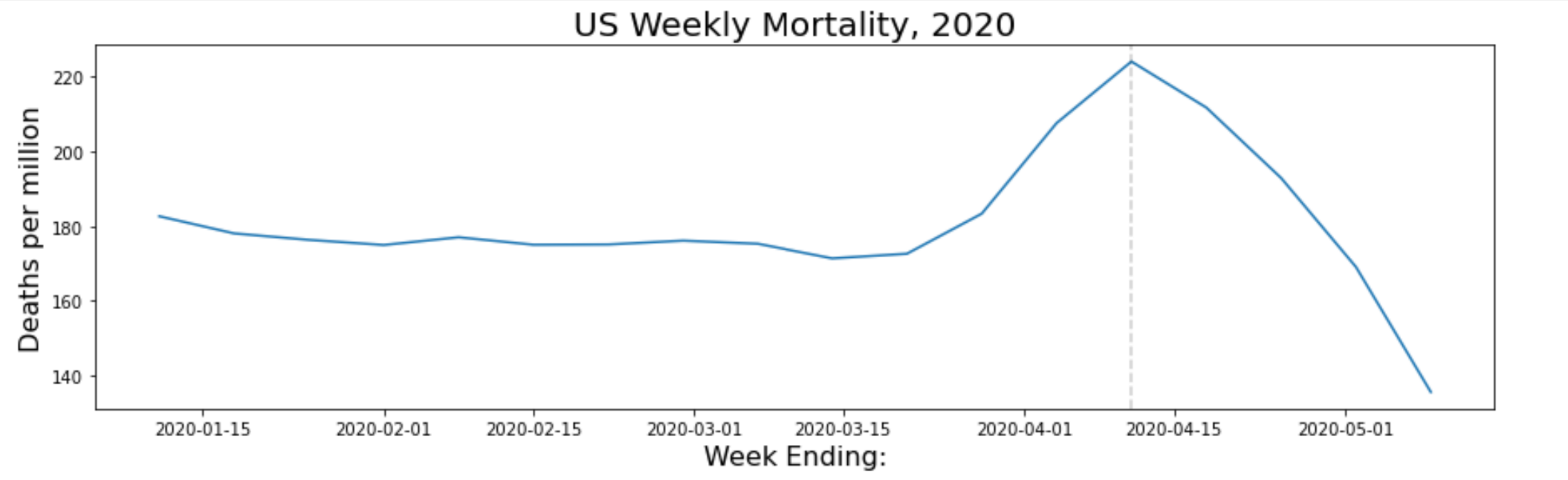
The data, revisited

Key metrics that might explain a state's resilience or vulnerability to a health crisis, or that have been heard explained as such, were gathered from government or journalistic sources cited at the end of this slideshow. Initial features:

- Median Household Income
- Median Age
- Population Density
- Latitude/Longitude
- % Population Un-Insured, under 65
- Political Party of Governor
- Proportion of time under stay-at-home order

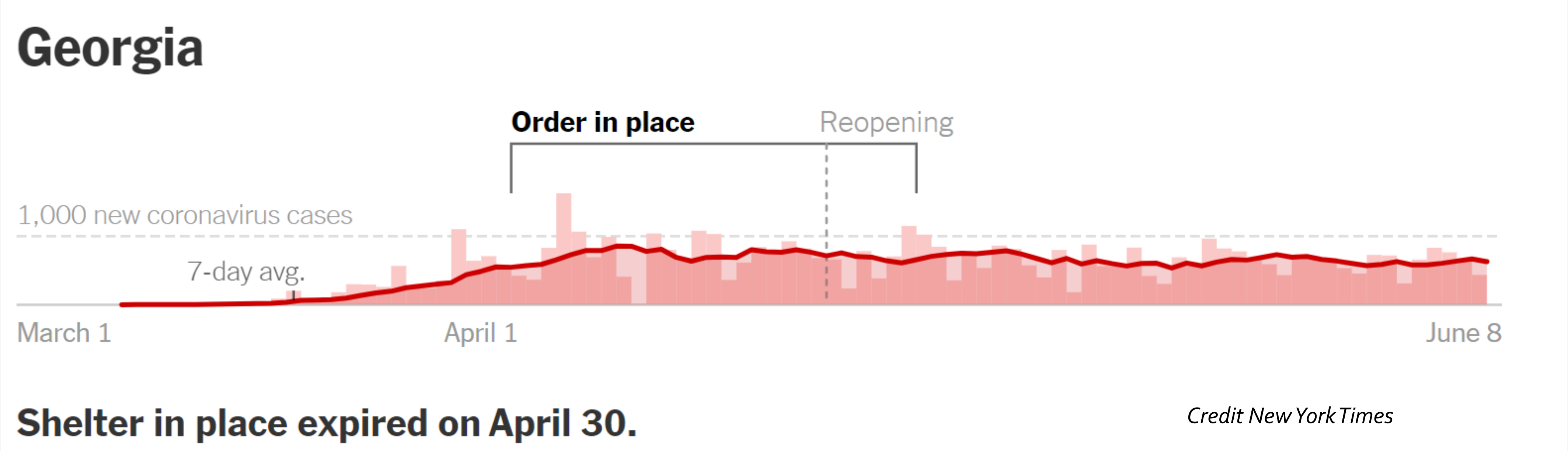
** NYC and DC are excluded from part two of this analysis for dissimilarity of information*

Data problems, part two



In part one, the decision was made to exclude the data after 4/18/2020 as incomplete...

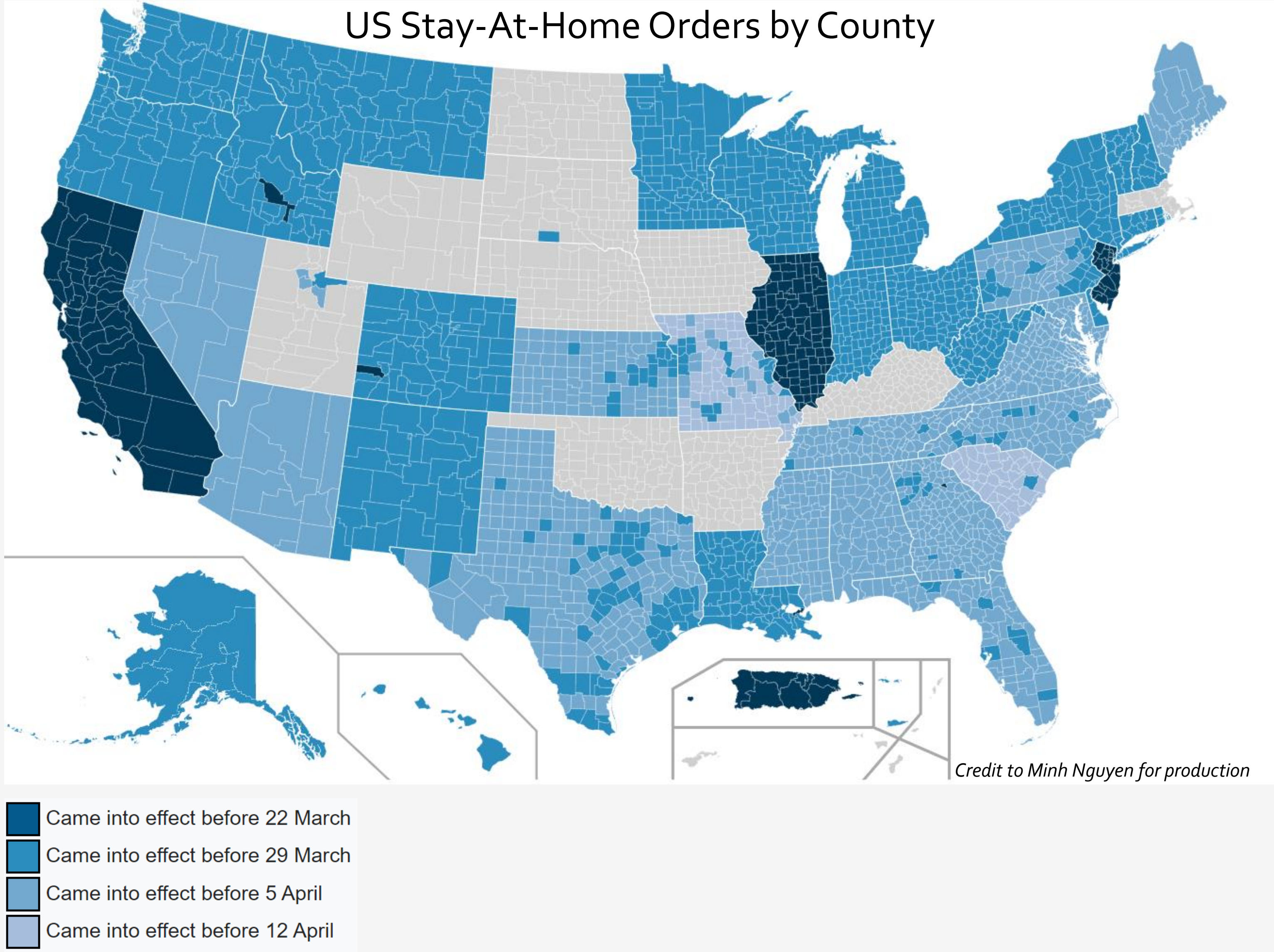
Data problems, part two



Credit New York Times

However, this means that our data does not include anything from the period after certain states re-opened. This difference in local policy will not be captured in our data.

Lockdowns

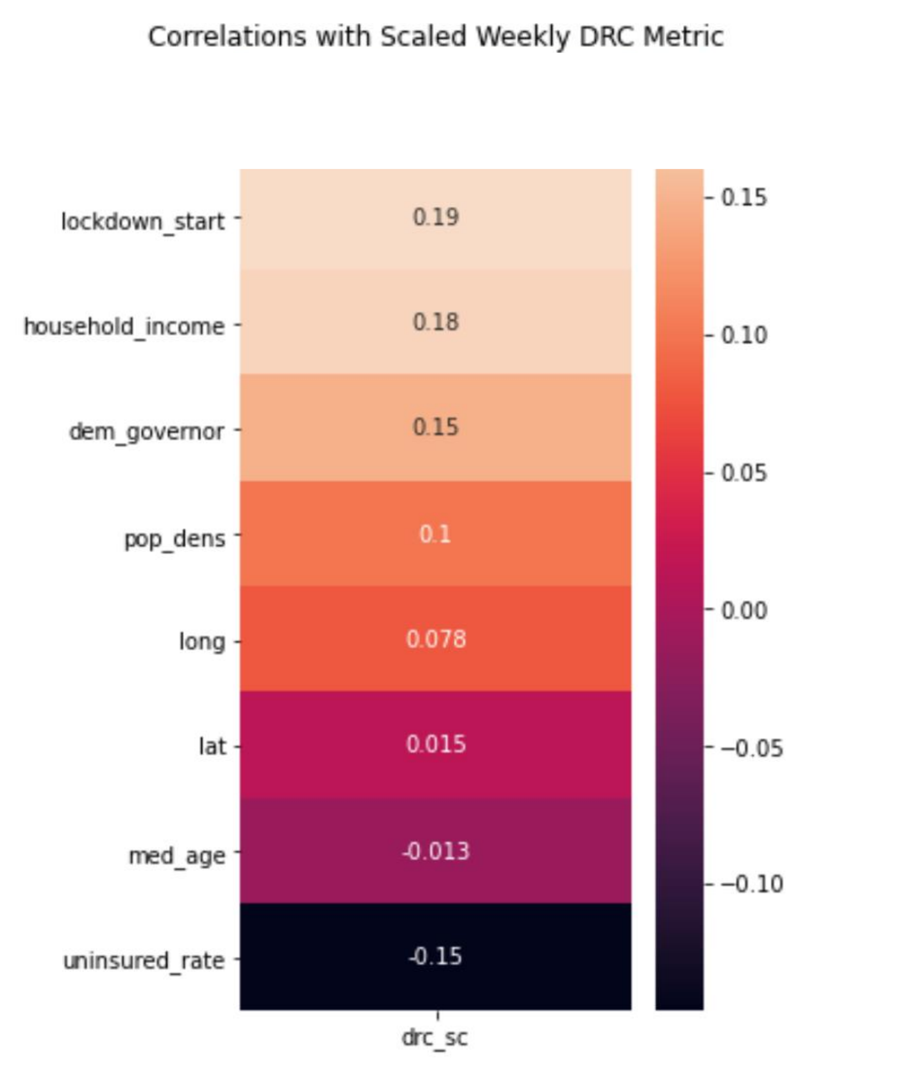


Due to the absence of any information regarding end-points of lockdowns, this feature was changed to an ordinal encoding of this information regarding stay-at-home order start-time.

Future versions of this model should use the originally proposed 'percent time under lockdown' metric, as the data permits.

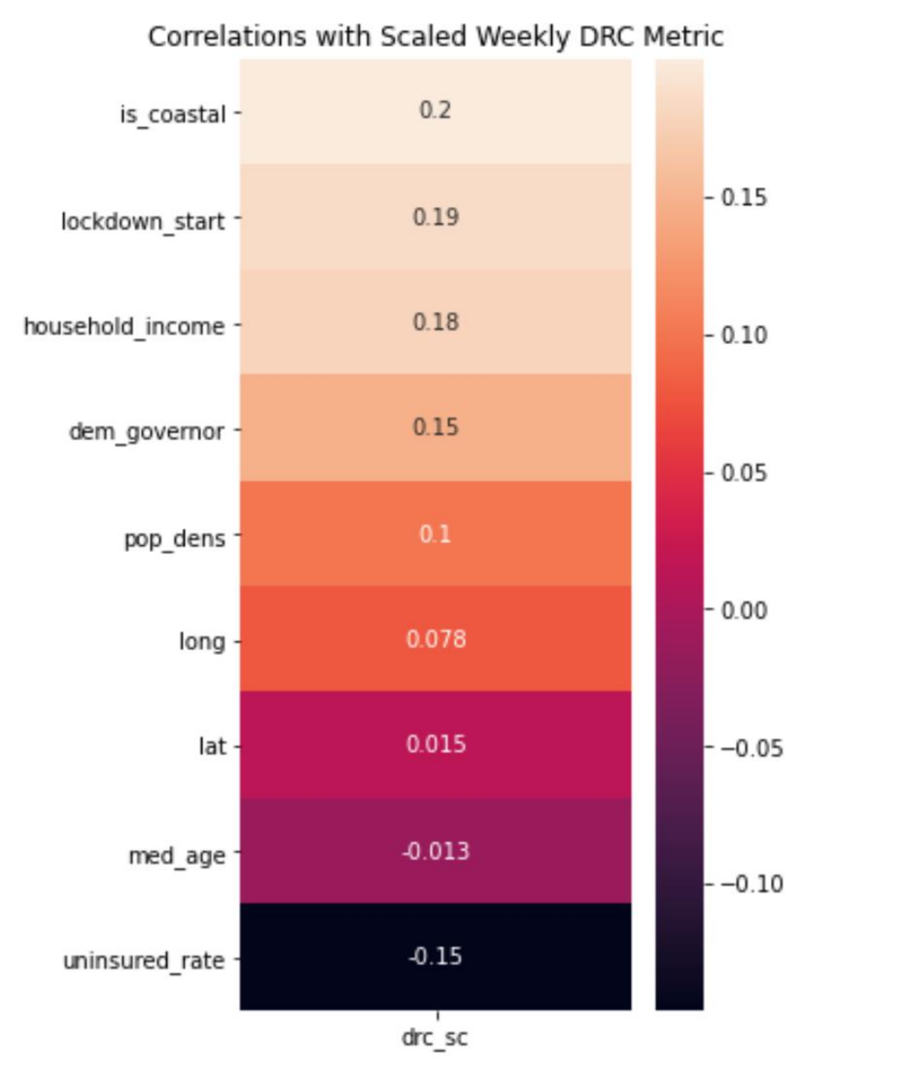
Underwhelming Correlations

The correlation numbers here were so low that a very simple, binary, ‘is_coastal’ feature was added…



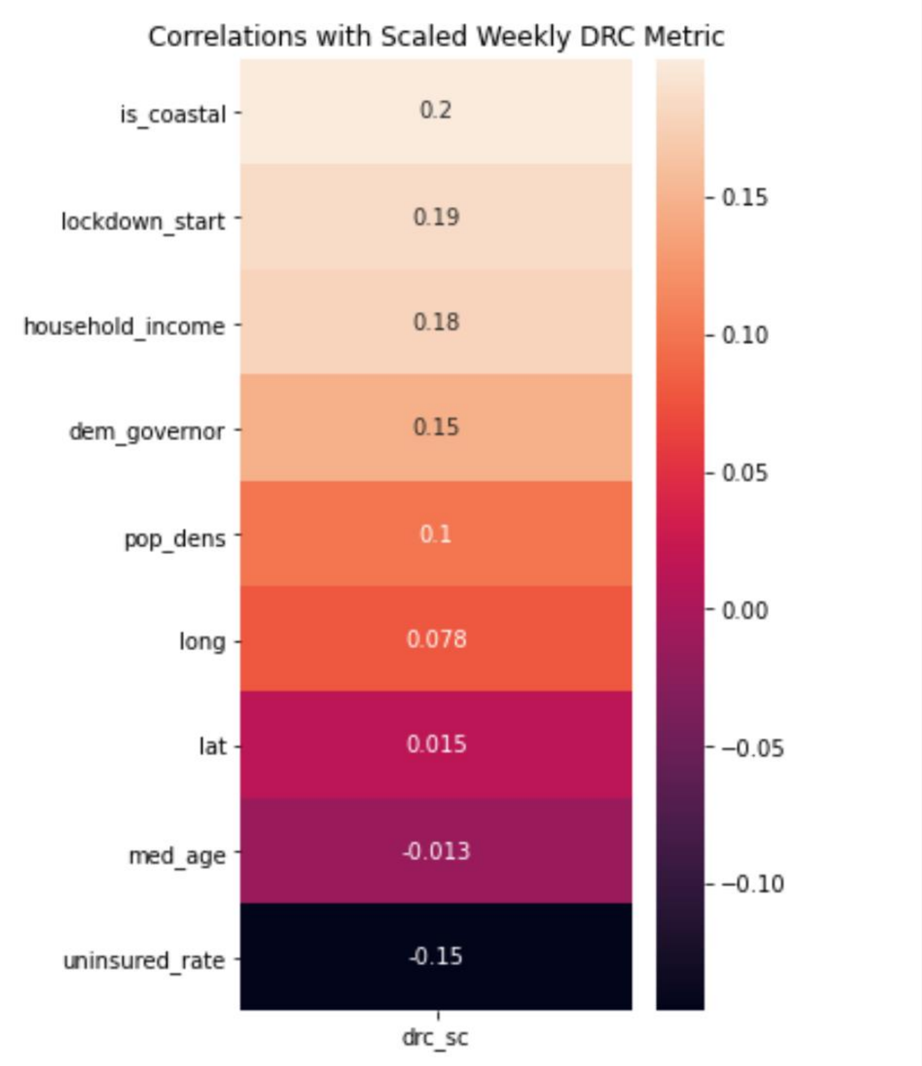
Underwhelming Correlations

... and immediately became our top feature.



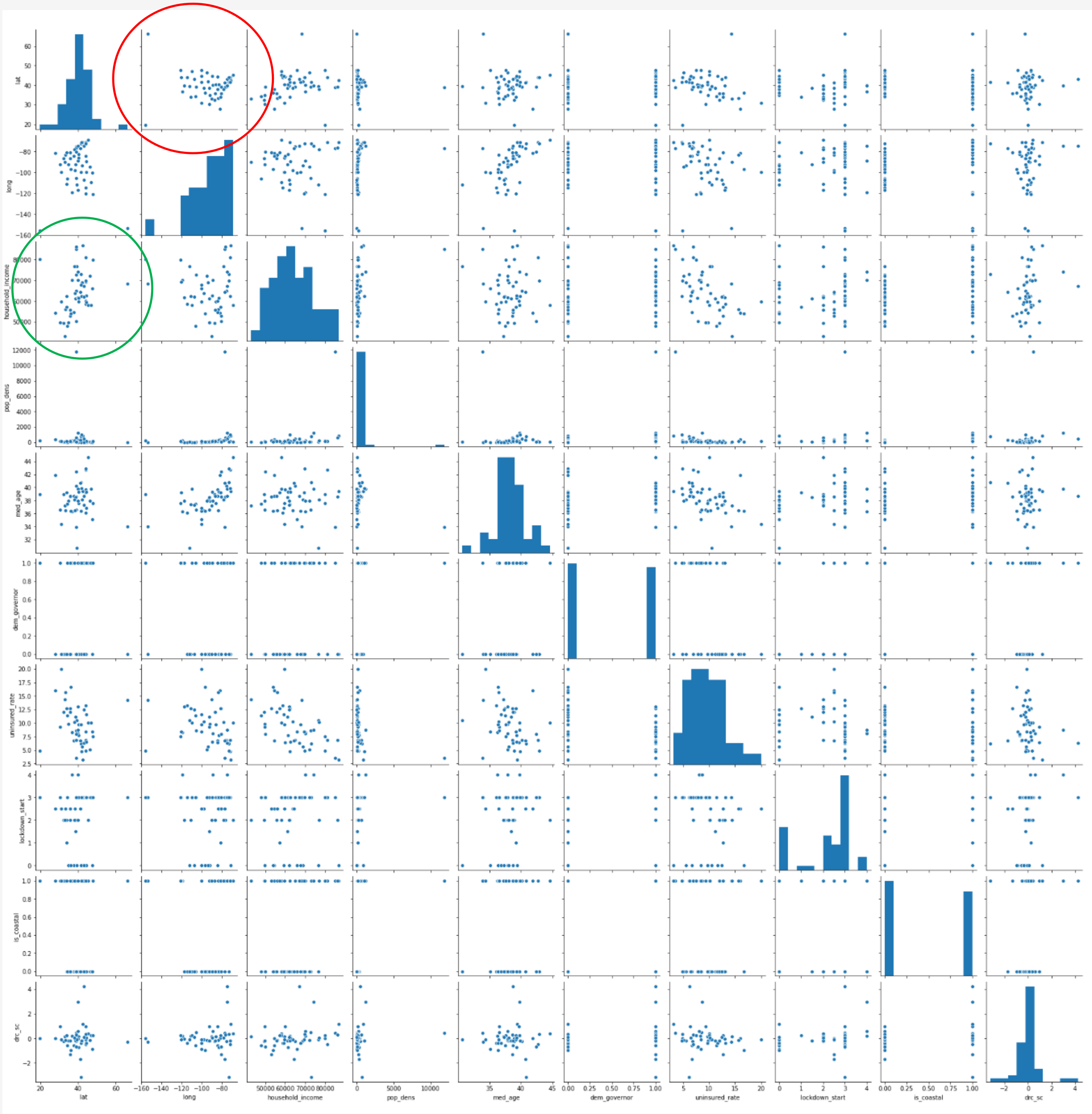
And Surprising

- Uninsured rate and median age negatively correlated with mortality
- Household income and early lockdowns positively correlated with a higher death rate



Not much to go on…

- This pairplot presented just to show the lack of linear relationships in our data
- Our parameter of interest is the bottom row and right column
- There is a stronger correlation between latitude and household income (circled in green) than anything to our scaled DRC metric
- There is a stronger correlation between latitude and longitude (red), and that is basically just an abstract picture of the United States



Confirmation

For the sake of completeness, a battery of simple models were run on the data, including:

- Linear regression
- Support vector regression
- Decision tree
- Bagged decision tree
- Random forest
- K Nearest Neighbors regression
- ADA boosting

None consistently outperformed baseline

Conclusions

Caveats

- There are more factors affecting mortality rates than just Covid-19.
- There are more objectives to policy and societies than just reducing mortality rates.
- The time-series model could just be over-estimating (a normal year's) death rates, hence the seemingly normal distribution in DRC (the relative comparison between states should still be relevant, even in this case).

- The concept of this project was always to get as large-scale a picture of what is happening as possible. The available data is not comprehensive or reliable enough yet to support this approach.
- At the end of the year, or even a couple years from now, after ripple effects have had time to play out, this will be a more representative picture of results.
- City and country comparisons remain very appealing levels of analysis for this approach.
- Another possibility would be to do the same from the other direction, trying to find inflection points in a localities economic prosperity.

Conclusions

- While the data is not settled enough yet, the fundamental idea in this approach seems to have promise
- Using a time-series model, rather than a rolling average, is a better baseline for any excess death analysis if there is the possibility of trend in the data.
- No considered feature, or combination thereof, was a strong predictor of having a higher mortality rate in 2020 than would be expected, covid-related or not.
- While it has dominated the consciousness and attention of America as a whole, the *lethal* consequences is largely only a true crisis in certain localities as of the time represented in this data set.
- This does not imply causation anywhere. It is still unclear whether lockdowns are good or bad, places are safe or unsafe from future infection, etc. At least as far as this analysis goes.

Thank You

Citations

- https://github.com/TheEconomist/covid-19-excess-deaths-tracker
- https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html
- https://www.latlong.net/category/states-236-14.html
- https://state.1keydata.com/
- https://dqydj.com/
- https://worldpopulationreview.com/states/state-densities/
- https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdowns#United_States
- https://www.census.gov/
- https://www.dol.gov/ui/data.pdf