Abigail Morgan

Capstone 2: Predictive power of weather variables on COVID-19 infection rates in Massachusetts, Connecticut, Vermont, and Maine

I. INTRODUCTION

Weather and climate are believed to play a significant role in the transmission of infectious diseases, and their relationships with COVID-19 have become a focal point of research since the onset of the pandemic. From an epidemiological perspective, weather variables can affect virus spread by influencing transmission dynamics, host susceptibility, and virus survival rates. From a behavioral perspective, weather often has big implications on social distancing, mobility levels, and frequency and location of social gatherings. And while there is no evidence directly linking COVID-19 with climate change, it's safe to say that our world has been fundamentally affected by climate change in ways that have direct consequences on our health and risk of infection. Indeed, many of the root causes of climate change (air pollution, deforestation, wildlife displacement, global warming) increase the risk for pandemics in general and recent decades have witnessed an accelerated emergence of infectious diseases. Further examination of the relationships between weather variables and infectious diseases (specifically COVID-19) could lead to a better understanding of the roles they may play in the seasonality of transmission and outbreak development, as well as help to establish early warning signs and inform outbreak response.

The goal of this project is to explore the relationship between COVID-19 infection rates and weather variables (specifically temperature and precipitation) and determine to what extent these weather variables may be useful in predicting virus transmission rates. For this project, the scope of the investigation has been limited to four states in the New England region of the United States: Massachusetts, Connecticut, Vermont, and Maine. These states represent areas of varying population densities, temperature ranges, and political views (often linked to mask usage, social distancing, and other precautionary behaviors). The guiding questions of this project are:

1. What correlation, if any, can be demonstrated between COVID-19 infection rates and weather variables (specifically temperature and precipitation) in Massachusetts, Connecticut, Maine, and Vermont?
2. What are the capabilities and limitations of a model that uses these weather variables to predict COVID-19 infection rates?
3. What other interpretations or explanations may exist to explain the results of this model?

II. DATA WRANGLING

Weather data was gathered from NOAA's National Centers for Environmental Information Database. Massachusetts used Middlesex County weather stations (primarily a station in Hingham, MA), Connecticut used Hartford County weather stations (primarily Hartford International Airport), Vermont used Essex County weather stations (primarily Burlington International Airport), and Maine used Piscataquis County

weather stations (primarily the Greenville Maine Forestry Service). There were very few missing values in the weather data, but missing values were replaced with data from the closest available weather station when possible. Only Hartford International Airport reported daily average temperature (TAVG) data. Massachusetts, Vermont, and Maine, reported minimum (TMIN) and maximum (TMAX) daily temperatures and these were averaged and used as TAVG. The temperature data was reported in tenths of a degree Celsius and so was converted into degrees Fahrenheit for this project. Precipitation was reported in tenths of a millimeter, and was converted to millimeters.

All COVID-19 data were obtained from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University." The confirmed case counts were aggregated for each relevant state and the first differences of these aggregated case counts were taken in order to calculate daily new cases, rather than total cumulative cases. A total of nine negative daily COVID-19 case counts were detected, which posed an obvious problem as negative new cases are not possible. After examining the COVID-19 rates surrounding the dates of these erroneous counts, they were replaced with the averages of the case counts on the dates preceding and following them. Other basic data cleaning was also performed to ensure all features were of the correct data type and the weather and COVID-19 datasets were merged correctly.

The main challenge in wrangling this data arose out of the variability in COVID-19 reporting schedules, which did not only vary among states, but also across time. In fact, it wasn't until visualizing the data that the reporting schedules became recognizable, and so the process of addressing this issue will be discussed more in depth in the exploratory data analysis section below.

III. EXPLORATORY DATA ANALYSIS

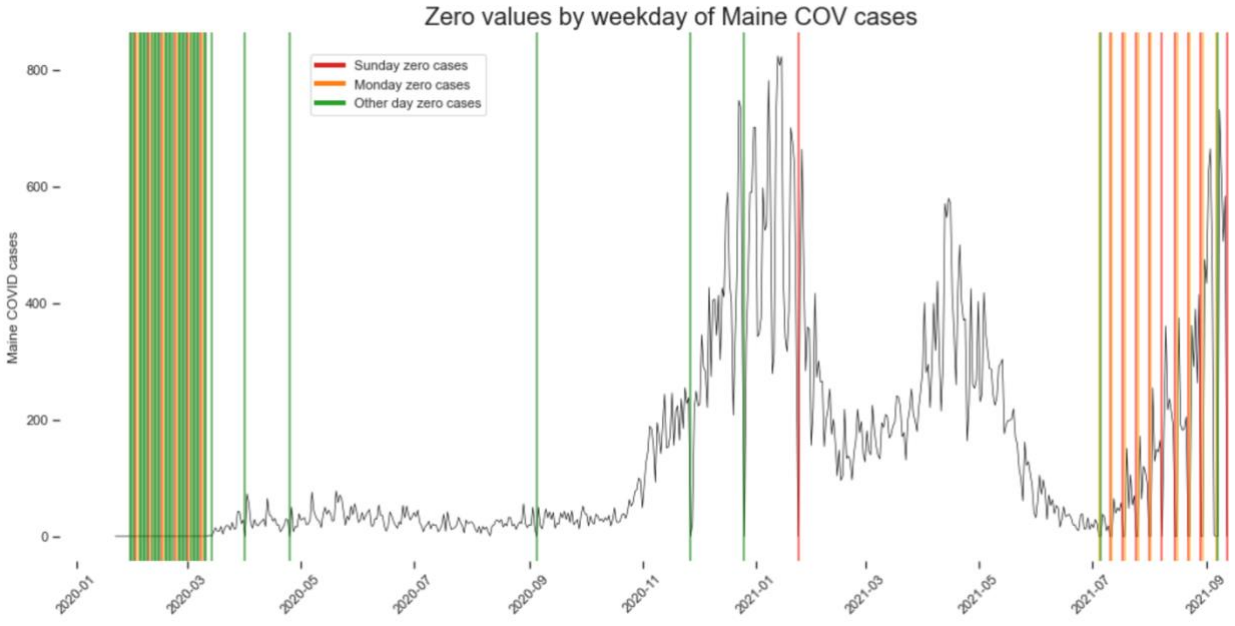
Missing case values

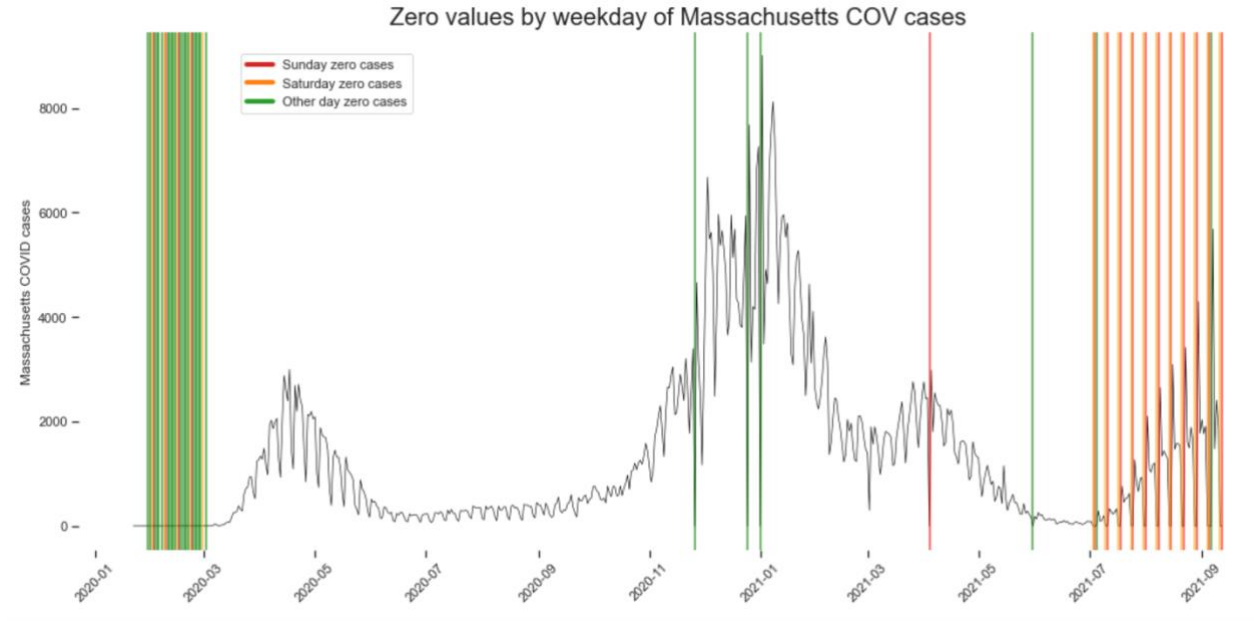
We may reasonably expect a certain proportion of zero values in a given dataset, but the clear pattern of zero values in this dataset immediately raised suspicion; for every five days of confirmed cases values, there were two days of zero values. For Connecticut, Massachusetts, and Vermont, a quick calendar check confirmed that these zero values all corresponded to weekend (Saturday- Sunday) dates (and a few Monday bank holidays). For Maine, these zero values all corresponded to Sunday and Monday dates (and a few Tuesday bank holidays). To further complicate matters, however, not *all* weekend values were missing for Massachusetts, Connecticut, and Vermont (or, Sunday-Monday values for Maine), but only certain periods of these values. These periods, too, seemed to differ among states (in both duration and location).

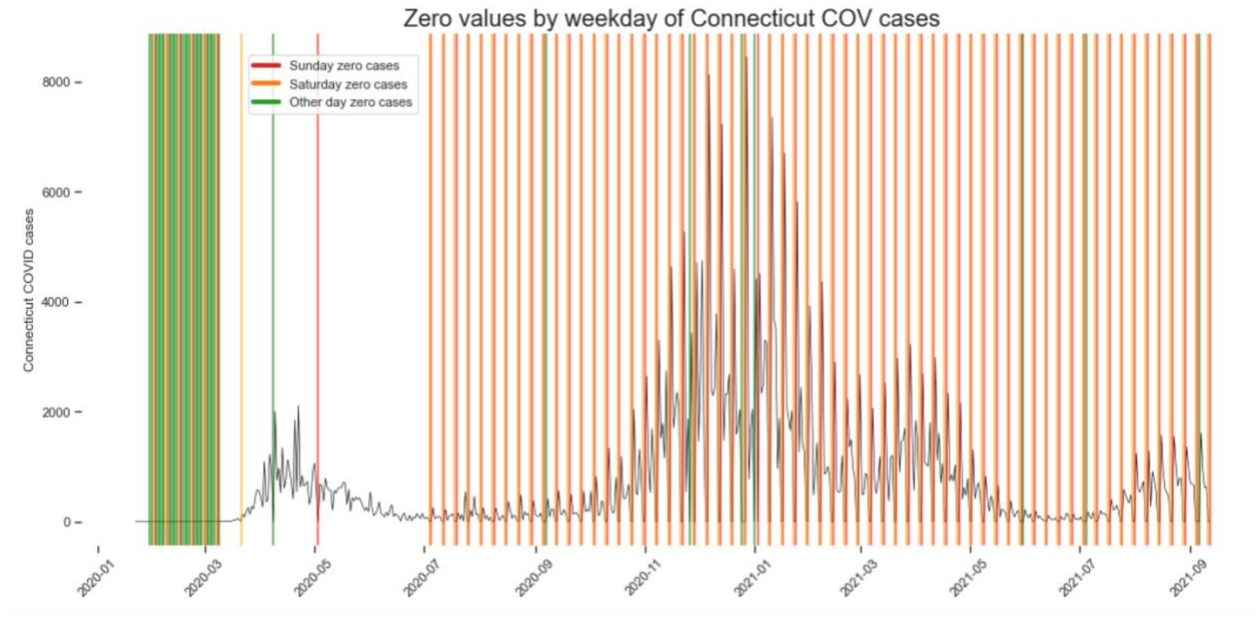
A column indicating the day of the week was first added to the dataset and all zero values were then plotted by state, as shown below. In the plots for Massachusetts, Connecticut, and Vermont, there is an orange vertical line for every zero value that falls on a Saturday, a red vertical line for every zero value that falls on a Sunday, and a green vertical line for every zero value that falls on any day of the week other than Saturday or Sunday. Because Maine seems to follow a different reporting schedule, its plot contains red vertical lines for every zero count that falls on a *Sunday*, an orange

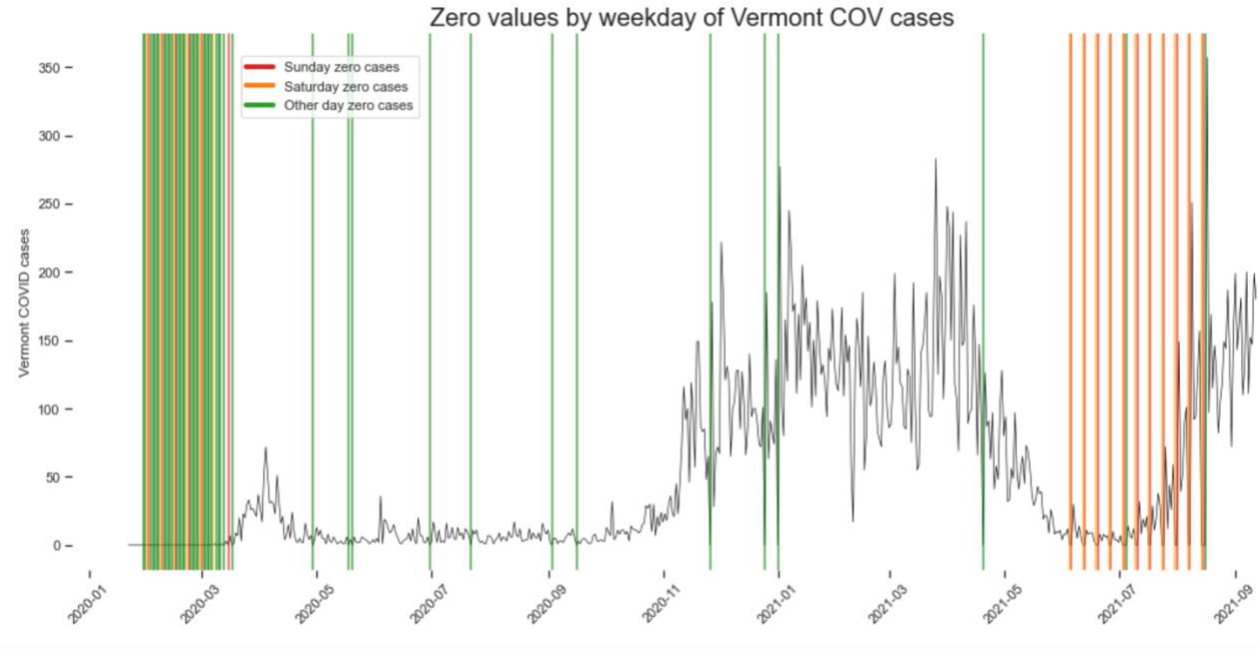
vertical line for every zero value that falls on a *Monday*, and a green vertical line for every zero value that falls on any day of the week other than Sunday or Monday.

Plotting the orange and red lines helps to determine the duration and placement of the abbreviated reporting schedules hypothesized above. Plotting the green lines double checks for any unusual cases of zero values that do not fit within these hypothesized reporting schedules.







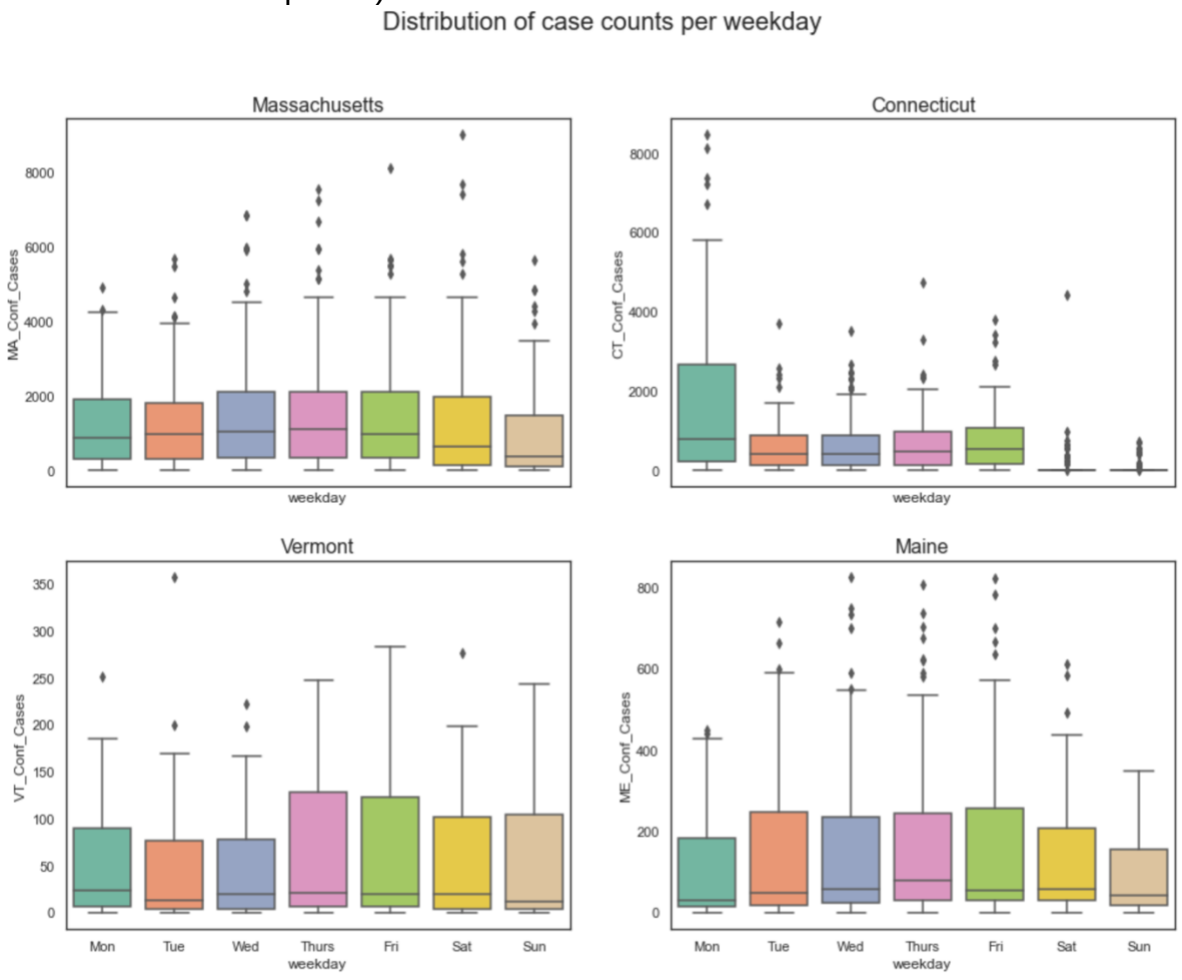


It becomes clear that all data prior to March, 2020 can be cut, as there is only one non-zero value count (from Massachusetts) prior to March, and the repetitive zero values from this period can only serve to skew any analysis of the data. Examining each state's plot, along with each state's government webpage (see here: Maine, Massachusetts, Vermont, Connecticut), the following reporting schedules can be concluded (excepting the occasional bank holiday):

- Maine reported cases Tuesday- Saturday from 2020-07-01 onwards (and reported cases daily before this date).

- Massachusetts reported cases Monday- Friday from 2021-07-01 onwards (and reported cases daily before this date).
- Connecticut reported cases Monday- Friday from 2020-07-01 onwards (and reported cases daily for only a very short period before this date).
- Vermont reported cases daily except for a period between 2021-06-01 until 2021-08-23 when it temporarily shifted to a Monday- Friday reporting schedule (before reverting back to daily reporting due to a surge).

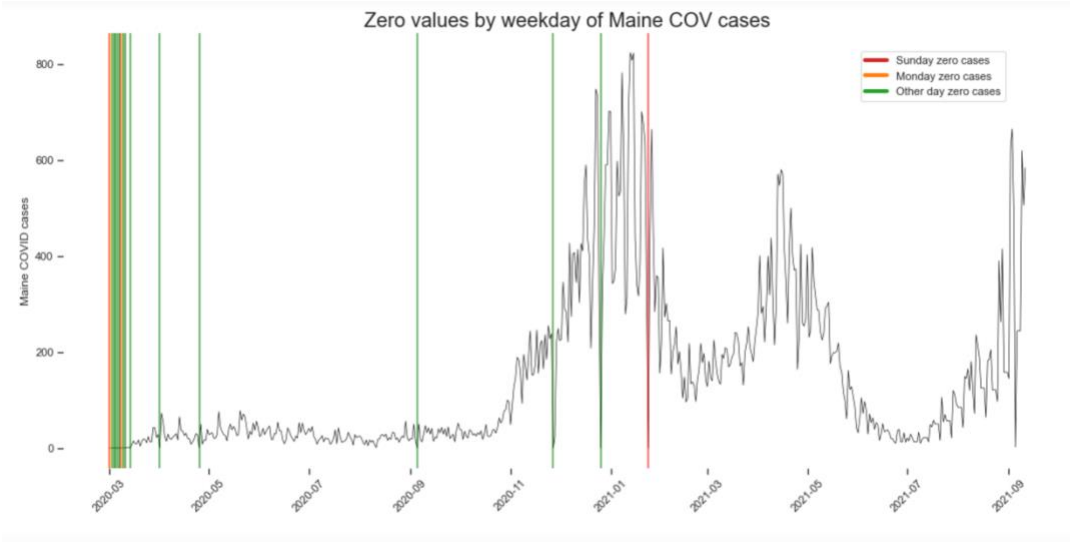
The above hyperlinked resources also describe that each state reports a cumulative total of new cases on the day following a two-day break in reporting. For Massachusetts, Connecticut, and Vermont this would be Monday; for Maine, this would be Tuesday. The resources also indicate that if the day following a two-day break in reporting (either a Monday or a Tuesday) is a bank holiday, then a cumulative case count for the *three* prior days will be reported on the day following the bank holiday (for Massachusetts, Connecticut, and Vermont, this would be Tuesday, and for Maine, this would be Wednesday). Because each state reports a *cumulative total* the day following a break in reporting, this cumulative count can be divided by three and each of the three days' values can be replaced with one third of the cumulative value for those days (so long as the date of the cumulative total does not also correspond to a bank holiday when cases were not reported).

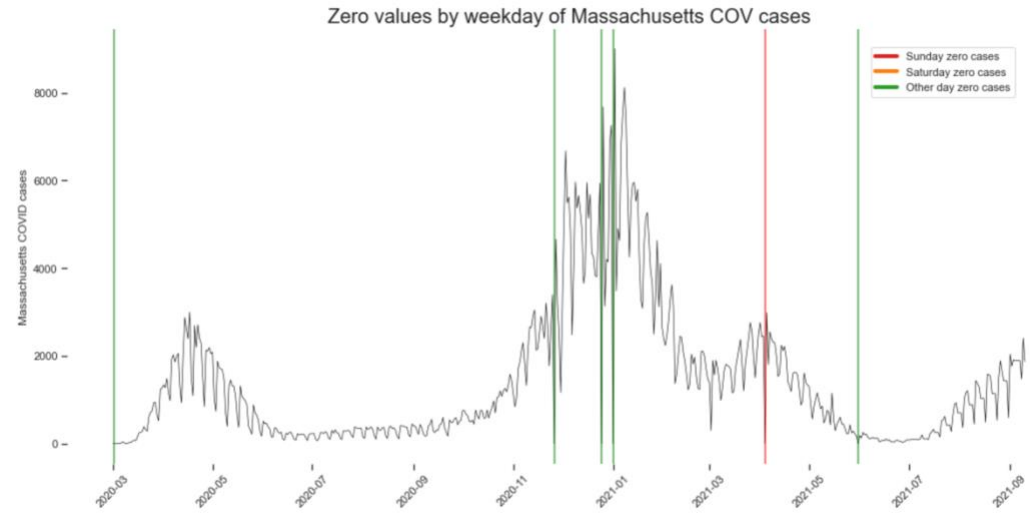


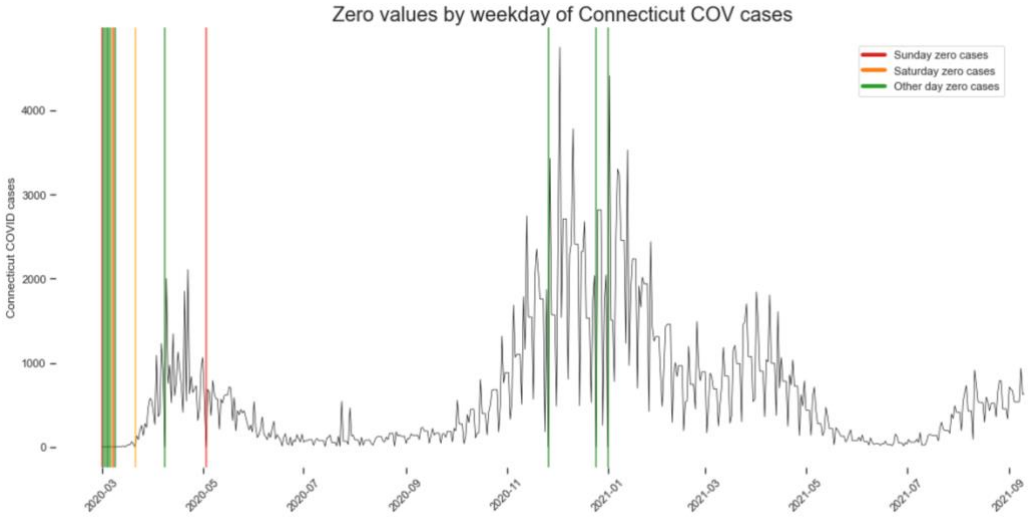
This boxplot illustrates the distribution of cases per weekday per state before addressing any of the missing weekend values. As might be expected, Connecticut's case counts are the most inconsistent by weekday. Specifically, the mean, median, and mode of case count values for Connecticut on Saturdays and Sundays appear to be close to zero. This makes sense, as Connecticut was the only state that didn't report weekend case counts for almost the entirety of the pandemic. Massachusetts, Vermont, and Maine each reported case counts daily for the vast majority of the pandemic, with only shorter temporary periods of partial-week reporting schedules. Maine clearly has the lowest number of case counts on Sundays and Mondays, which are precisely the days Maine didn't report cases towards the later stages of the pandemic. Once the various reporting schedules have been addressed, the distribution of cases counts per weekday per state will be replotted.

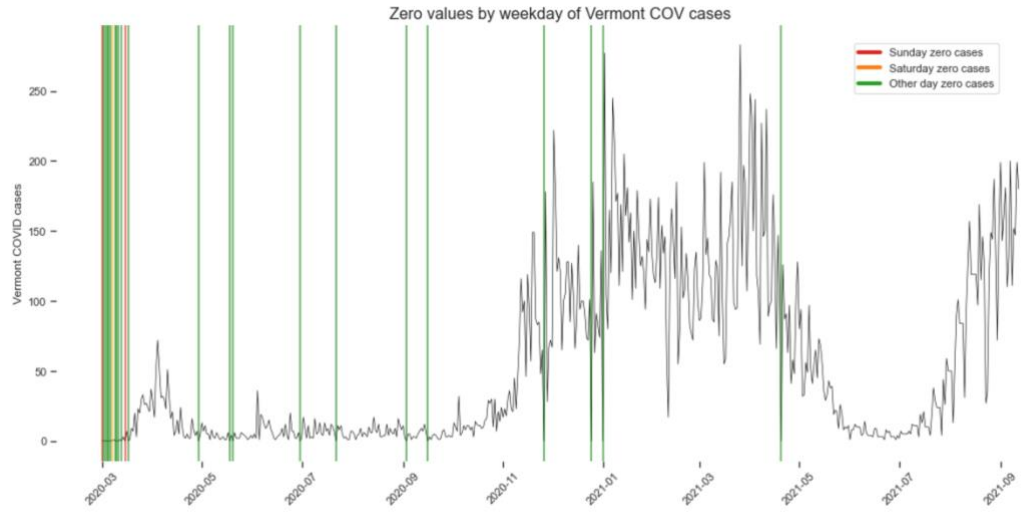
In order to address the missing case counts, each state will have to be examined individually, and the missing values addressed according to that state's unique reporting schedule. For Maine, this means checking each confirmed case count occurring on or after 2021-07-01, on a Tuesday, that did not have a zero-case count (and so must not be a bank holiday on which cases were not reported). A column is created, `ME_third`, representing one third of that value and the original cumulative Tuesday value is replaced with one-third of itself. Three more columns are added, representing lag-1, lag-2, and lag-3 of 'ME_third.' Monday zero values are then replaced with lag-1 of 'ME_third', Sunday zero values with lag-2 of 'ME_third', and so on. Once this is complete, the 'ME_third' column, as well as the lag columns, can be removed and the process is repeated on the case counts for Connecticut, Massachusetts, and Vermont, according to their respective reporting schedules.

In order to address instances of long weekends, where the day following a two-day break in reporting falls on a bank holiday and zero cases are reported, a similar process can be applied by taking a third of the case count following the bank holiday and replacing missing values appropriately. Finally, a list is made of all other instances of bank holidays that fall mid-week, where zero case counts are reported in at least half of the states, and these dates (five, in total) are deleted from the dataset. The results are four much cleaner and complete time series plots, as shown.

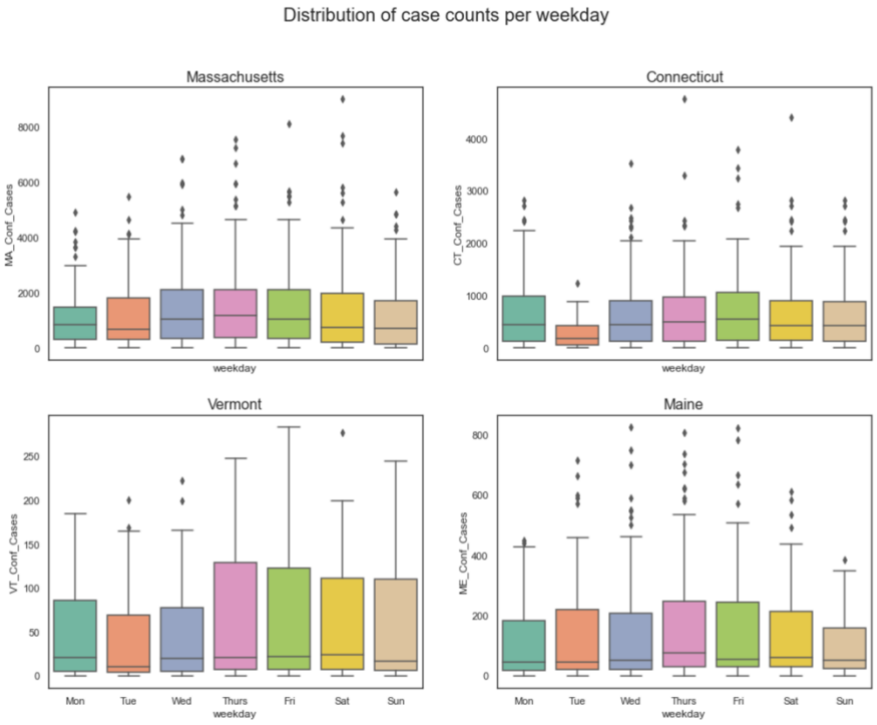



Zero values by weekday of Massachusetts COV cases
Sunday zero cases
Saturday zero cases
Other day zero cases
Massachusetts COVID cases
8000
6000
4000
2000
0
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09


Zero values by weekday of Connecticut COV cases
Sunday zero cases
Saturday zero cases
Other day zero cases
Connecticut COVID cases
4000
3000
2000
1000
0
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09

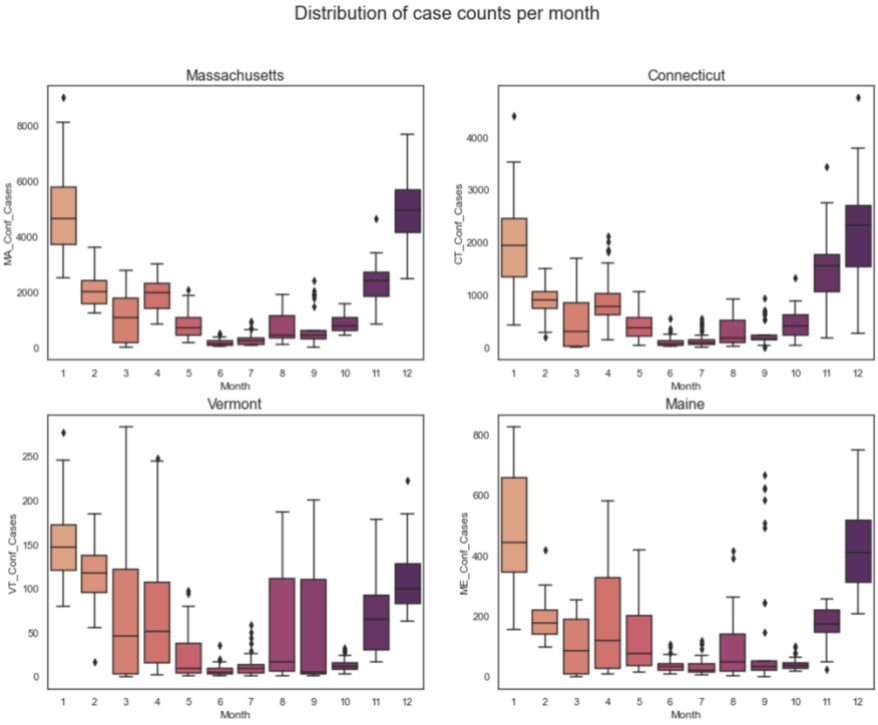

Zero values by weekday of Vermont COV cases
Sunday zero cases
Saturday zero cases
Other day zero cases
Vermont COVID cases
250
200
150
100
50
0
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09

Distribution of case values



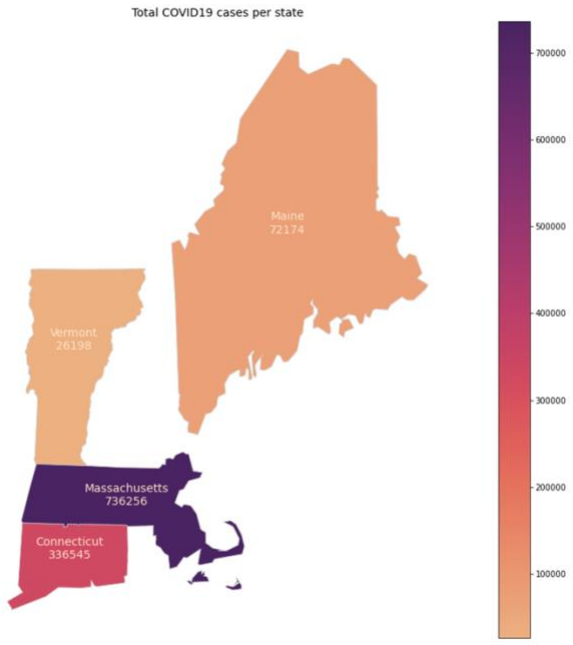
Replotting the data after addressing the missing values reveals a few remaining zero values per state. Overall, however, there is a vast improvement in both cleanliness and completeness of the dataset. The boxplots of reported cases per weekday per state are also replotted, and support the notion of a more balanced distribution of case counts per weekday.

A boxplot of the distribution of case counts per *month* per state reveals that more cases are reported at the beginning and at the end of the year, with less cases generally reported from about May to September.

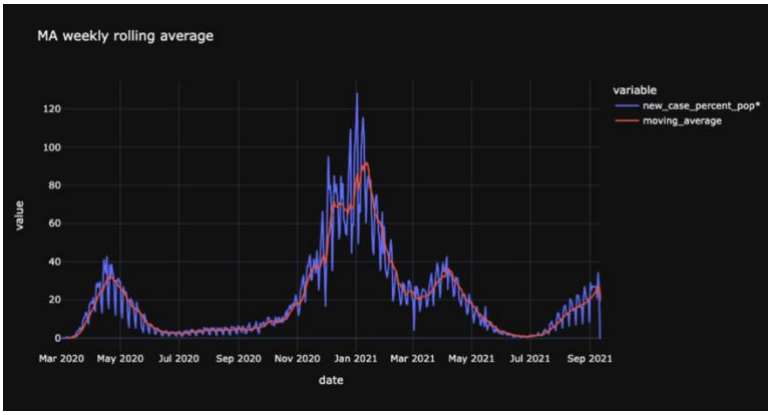


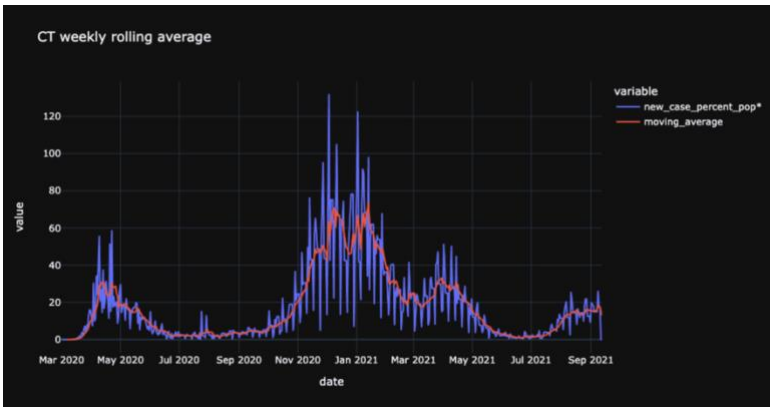
This could be due to weather changes from winter to summer months (particularly temperature), but this will need to be further investigated before making any conclusions. Regardless, each state demonstrates a very similar U-shaped distribution of case counts per month.

The distribution of total COVID-19 cases by state was plotted using geopandas, with color representing total number of COVID-19 cases from 2020-03-01 until 09-12-2021. In deep purple, Massachusetts clearly has the largest total cases, with 736,256. Connecticut has the second highest total case count with 336,545. Maine and Vermont have the lowest total case counts, with 72,174 cases and 26,198 cases respectively. There is also an interactive map illustrating total COVID19 cases by state over time. The interactive element of this map can't be rendered within GitHub itself, but it is functional within the downloaded or cloned version of the .ipynb notebook.

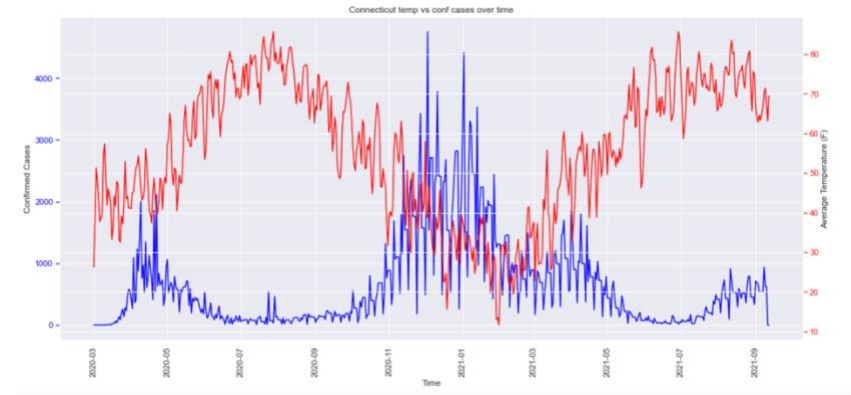


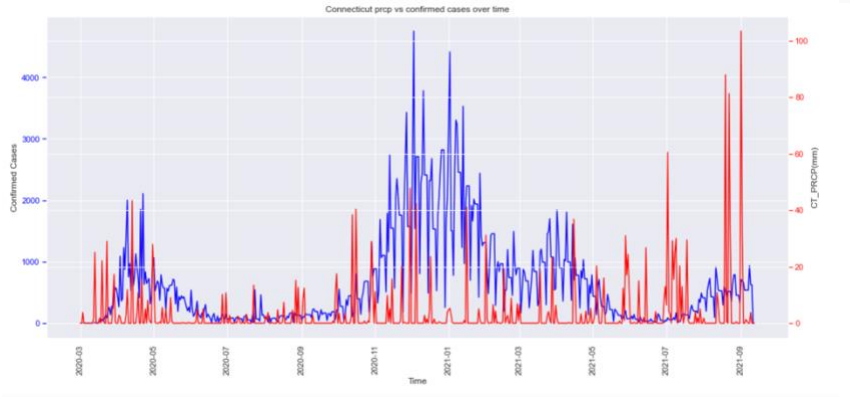
Lastly COVID-19 cases were plotted per state as a weekly rolling average. The 7-day moving average is plotted in red and the actual daily case counts are plotted in blue. The weekly rolling average plots are much cleaner than the daily case count plots, and more clearly show overall trends. Massachusetts, Vermont, and Connecticut each show a small spike in cases in the Spring of 2020. All states experienced their greatest surge (in both number of cases and length of time) in the winter of 2020-2021. All four states also seem to be witnessing the beginning of a third spike in cases towards the end of the plot in September 2021. Interestingly, each state also seems to demonstrate at small dip towards the end of their largest surge, following by a small spike. One possible explanation for this could be the emergence of the vaccine to the greater population (the dip),





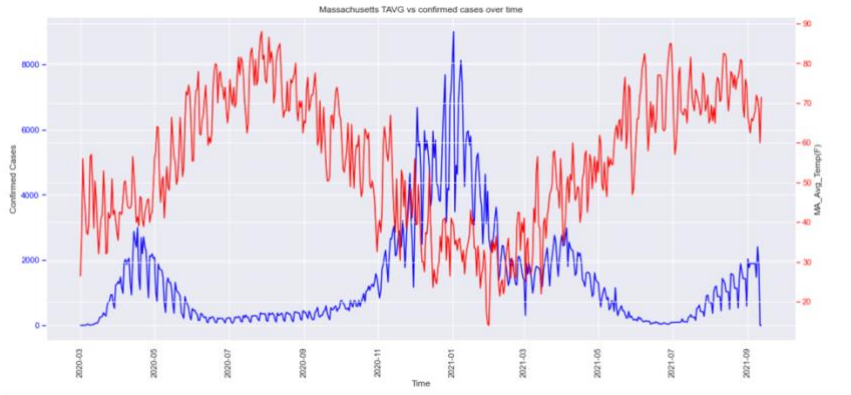
followed by the arrival of the Delta variant (the second spike), both of which occurred in the spring of 2021.

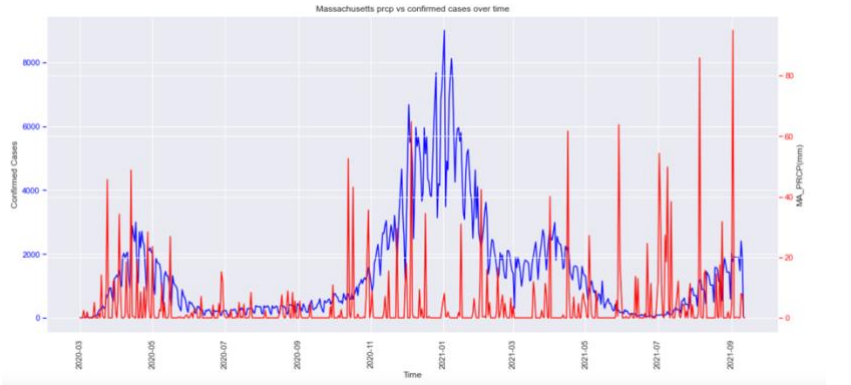




Visualizing relationships between features

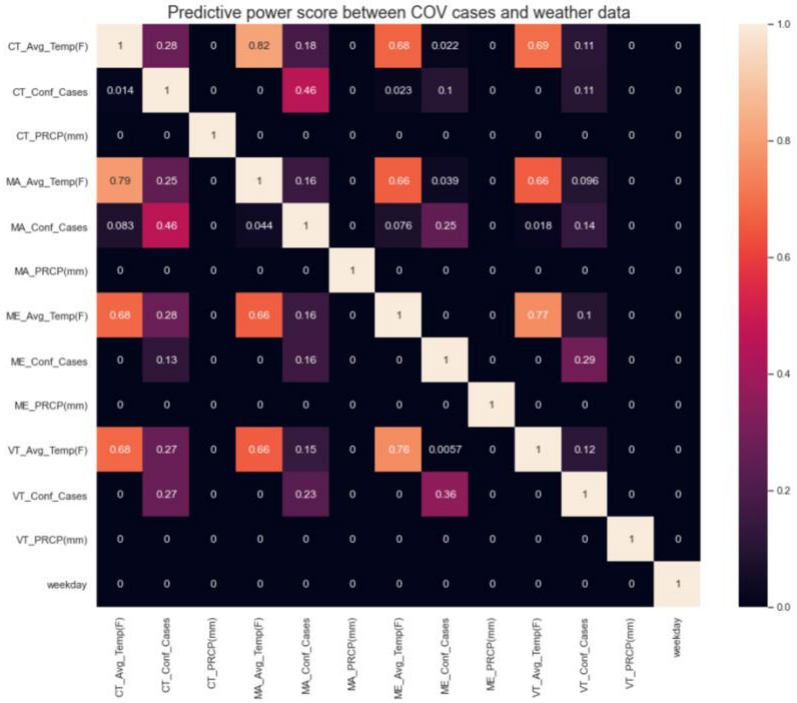
As a first step in visualizing the relationships between features, the data is plotted by state and by weather variable as time series (line) plots with twin axes to compare each state's weather variables against daily confirmed cases. In other words, two time series plots with twin axes are plotted per state: the first, of daily average temperature (TAVG) and COVID-19 cases over time, and the second, as daily precipitation (PRCP) and COVID-19 cases over time. Confirmed COVID-19 cases are all plotted in blue and weather variables are all plotted in red.





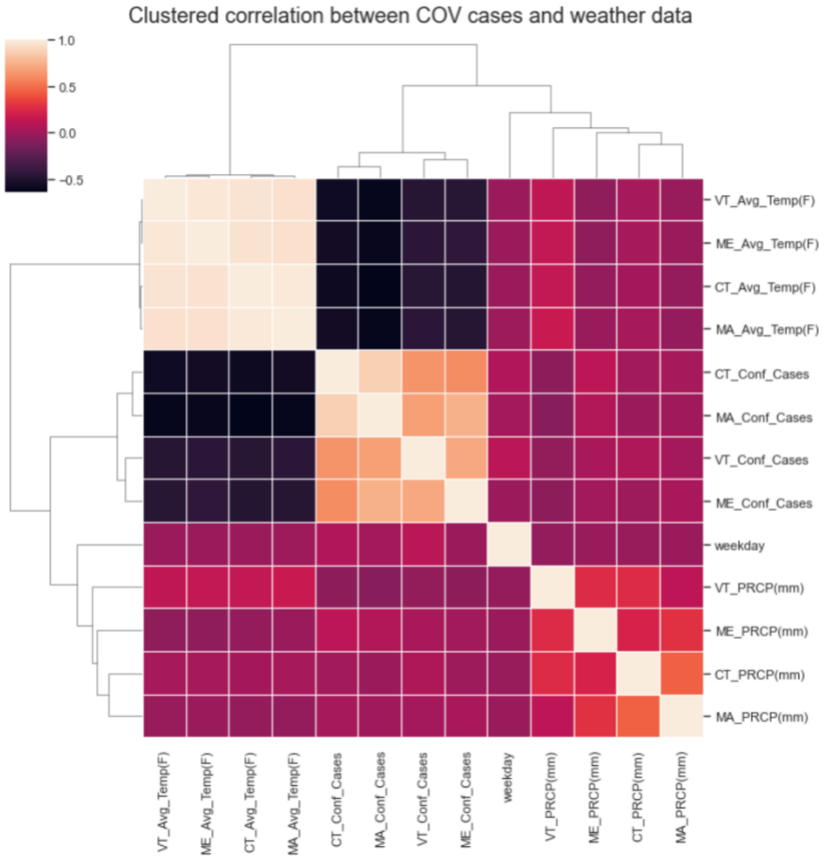
Unsurprisingly, there is a very clear seasonal trend in TAVG for each of the four states. This could be expected, especially in a region like New England with four very distinct seasons. However, there also seems to be a clear seasonal trend in COVID-19, and most immediately striking is the strong negative correlation between TAVG and daily COVID-19 cases. Each of the four states seems to demonstrate this negative correlation, though the plots for Massachusetts and Connecticut are perhaps the most striking.

The correlation between COVID-19 cases and daily precipitation rates is far less clear (if there is one at all). There may be some signs of a (positive) correlation between precipitation rates and a *lagged* COVID-19 rate, but this requires further exploration before drawing any conclusions. Additionally, there is a known correlation between temperature and precipitation, so care will need to be taken not to overfit a model with these two correlated variables (should both be used in the final model).



It may be easier to determine correlation with some numerical values, so a predictive power heatmap is plotted next. The temperature variables do seem to carry some predictive power, but not much. The precipitation variables have such little predictive power that they register as zero (if there is any predictive power to these variables at all).

Lastly, a clustered correlation heatmap is plotted. The clustermap is much easier to interpret than an unclustered heatmap, and visually points the audience immediately to the correlation between average temperatures and confirmed case counts, which appear as two dark near-black squares in the upper left-hand side. The clustermap has also grouped Massachusetts values most closely with Connecticut values for each observation, as well as Vermont values most closely with Maine values. This is not surprising, as this is also the geographical grouping of the states.



IV. PREPROCESSING

To time series or not to time series

One of the first questions that arose when beginning the preprocessing stage of this project was whether or not to analyze the dataset as a timeseries dataset. Clearly, the dataset fits the definition of a timeseries dataset. That is, it is a sequence of datapoints indexed chronologically with time stamps. However, as is illustrated in the twin axes plots of COVID-19 cases against temperature over time, the data only spans about 1.5 seasonal cycles. It would likely be very difficult (if not impossible) for a model to learn the patterns of the dataset with only 1.5 seasonal cycles without drastically overfitting.

Keeping this in mind, the data was stationarized by taking first differences until the Augmented Dickey-Fuller tests returned p-values of under 0.05 for each variable. The steps of this process can be found in the accompanying notebooks contained in the "Experiments" folder of the project repository. Pycaret was used to identify top performing models for the dataset and returned the resulting R2s. Unfortunately, the top performing model only achieved an R2 of around 0.2. For this reason, it was concluded that there was not enough data to analyze the dataset using timeseries methods, and it was decided that it would not be treated as such.

IV. Feature Engineering

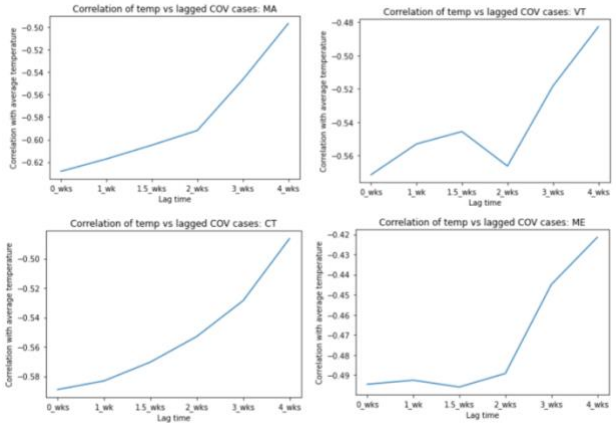
Controlling for population density

Because some of the states are very densely populated (Massachusetts, Connecticut), while others (Maine, Vermont) are much more sparsely populated, confirmed COVID-19 cases as a *percentage of overall population* were calculated for each state, in place of total cases. Total population data was retrieved from the U.S. Census Bureau's April 1, 2020 census. Each state's daily confirmed cases value was divided by its total population (as of April 1, 2020), and then multiplied by 100,000 for ten thousandths of a percent of total population.

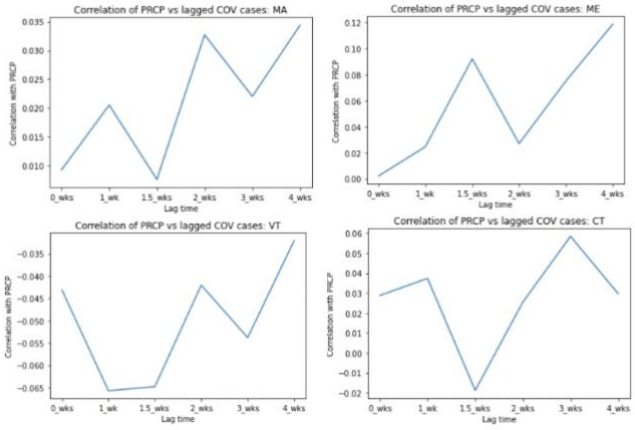
Lagged weather variable values

There is a period of incubation between the moment of COVID-19 infection and the manifestation of symptoms, and so it might be reasonable to expect a higher correlation between COVID-19 cases and some lagged value of the weather variables. If cases get tested *because* they've manifested symptoms, one might expect weather variables with a lag of about 1.5- 2 weeks to demonstrate a higher correlation than variables with no lag. On the other hand, if cases are getting tested regularly (for work, medical reasons, etc.), there might not be much of a lag, if any at all, with weather variables.

Correlation coefficients of COVID-19 cases were determined for each weather variable separately (TAVG, PRCP), with lags of 0 weeks, 1.5 weeks, 2 weeks, 3 weeks, and 4 weeks. These correlations were then plotted with line plots, as shown. In the lagged TAVG plots, Maine seems to have the strongest correlation between COVID-19 cases and a 1.5-week lag of temperature. Vermont seems to have the strongest correlation between COVID-19 cases and current temperatures, as well as those lagged by two weeks. However, Massachusetts and Connecticut both very clearly demonstrate the strongest correlation between cases and lag 0 (or, current) temperatures. While it seems that for some states there may well be a stronger correlation between COVID cases and lagged temperature values, since 50- 75% of the data don't show any increased correlation using temperature lags (and the remaining 25- 50% of the data disagree on an optimal lagged value), this project will stick with the original lag 0 (current) temperature cases.



Next, correlation was calculated between COVID-19 cases and lagged values of *precipitation* total for each of the four states. Massachusetts showed the highest correlation with lag 4 weeks of PRCP, but there was no real pattern in the data suggesting this correlation was anything more than random. Connecticut's case counts also showed no clear pattern with lagged values of PRCP, though it did show a small spike in correlation at around lag 3 weeks of PRCP. Vermont showed elevated correlation between COVID-19 cases and lag 2 weeks and 4 weeks of PRCP, again with no apparent pattern and Maine showed an elevation in correlation at lag 4 weeks PRCP. Given the spiky, pattern-less nature of the plots, however, and the fact that the correlations only varied about 0.01-0.03, it was also decided not to use a lagged value of PRCP, for fear of overfitting.



All lagged values of weather variables are therefore removed and the state-specific data is prepped for a merge back into one single DataFrame. A new column is placed in each state's separate DataFrame, 'state_id'. With this new identifier column, state-specific identifiers will no longer be necessary in the column names themselves, and each state's column names are updated to mirror each other. The PRCP columns are dropped at this point, as they have demonstrated little, if any, correlation to

confirmed COIVD-19 cases. The state DataFrames are then concatenated vertically in preparation for use with machine learning models.

One-hot-encoding categorical variables

Many Machine Learning models struggle with categorical data, and this dataset contains at least one categorical data column (the newly created ‘state_id’ column). Other columns, like ‘Month’ or ‘day_of_week’ could be considered categorical data, but can also be represented numerically without using a technique like one-hot-encoding. Therefore, the ‘state_id’ column is one-hot-encoded, and a second version of the dataset is also created, which one-hot-encodes the ‘state_id’, ‘Month’ and ‘day_of_week’ columns. Both of these datasets will be run through pycaret to determine optimal base estimators to use in a final Voting Regressor model.

Determining optimal estimators with pycaret

Without any hyperparameter tuning, pycaret determined that a CatBoost Regressor produced the best model metrics for Massachusetts, Maine, and Connecticut, and a Random Forest Regressor produced the best model metrics for Vermont (using R2 as an evaluation metric). The top three performing models on the entire dataset (without any one-hot-encoding) were: CatBoost Regressor, Random Forest Regressor, and Gradient Boosting Regressor (in that order).

Interestingly, there was a somewhat significant improvement in R2 when the ‘state_id’ column was one-hot-encoded (from a maximum of 0.7989 to a maximum of 0.8280). With the ‘state_id’ column one-hot-encoded, the Extra Trees Regressor now also outperforms the Gradient Boosting Regressor. However, when pycaret was run on the dataset containing all three one-hot-encoded variables: ‘state_id’, ‘Month’, and ‘day_of_week,’ it produced the best evaluation metrics of them all (with a maximum R2 of 0.8340). The dataset containing all three variables one-hot-encoded is therefore saved for use with a CatBoost Regressor, Extra Trees Regressor, and Random Forest Regressor, which will, in turn, be fed into a Voting Regressor.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	4.0272	49.6088	7.0267	0.8340	0.5039	1.4677	0.4967
et	Extra Trees Regressor	3.7804	53.9427	7.2986	0.8225	0.3934	0.7279	0.1500
rf	Random Forest Regressor	3.9468	53.1076	7.2567	0.8225	0.4255	1.0604	0.1467
lightgbm	Light Gradient Boosting Machine	4.2811	55.2234	7.4053	0.8149	0.5077	1.5072	0.1833
xgboost	Extreme Gradient Boosting	4.5464	63.4020	7.9425	0.7885	0.5245	1.3973	0.2133

V. MODELING

Before training any models on the data, it is split into training and testing sets to keep the data consistent and separated. Because the data has a much smaller proportion of high case counts, it is separated into bins (without sorting) and the bins are stratified in the train test split. This keeps the X, y pairs together, but retains some balance in the split. Next, a Dummy Regressor is fit to the data, using the mean to

predict the target variable y (COVID-19 cases). Unsurprisingly, this Dummy Regressor performs very poorly, scoring an R2 of about -0.00086.

Tuning the top performing models for the ensemble model

In the preprocessing stage, pycaret determined that the top performing models for the dataset were a CatBoost Regressor, Random Forest Regressor, and Extra Trees Regressor. The Random Forest Regressor's and Extra Trees Regressor's hyperparameters are tuned first with Coarse to Fine hyperparameter tuning (using Randomized Search CV and then Grid Search CV), as well as with Bayesian Optimization using Hyperopt. All hyperparameter tuning methods utilized five-fold cross-validation as well. In order to use Bayesian Optimization to determine optimal hyperparameter values, a 'space' of various hyperparameters, and then a function ('hyperparameter_tuning') which returns negative R2, are defined and then passed into a minimizing function. Once optimal hyperparameter values have been determined using each method, their resulting R2s are compared and the top performed hyperparameter values are chosen for each of the two decision-tree-based estimator models (Random Forest Regressor and Extra Trees Regressor).

Finally, the CatBoost Regressor is instantiated. It is first run with, and later without, the Pool feature to determine which method performed better. The Pool method performed slightly better on the dataset, and so it is used while conducting Coarse to Fine hyperparameter tuning (using Randomized Search CV and Grid Search CV with five-fold cross-validation, as was used in the previous two estimators).

Once the optimal hyperparameters had been determined for all three estimator models, a Voting Regressor was instantiated with each of three optimal estimators determined in the previous modeling steps. This was run on the data and the evaluation metrics were calculated. Finally, in order to determine the best weights for each of the estimators, a nested for loop was created to test all combinations of 10 percent intervals between 0.1 and 1.0 that added up to 1. The top performing weights were determined to be 0.1 for the Random Forest Regressor, 0.6 for the Extra Trees Regressor, and 0.3 for the CatBoost Regressor. The final Voting Regressor was then re-instantiated with the optimal weights and run on the entire training and testing datasets for a final R2 of about 0.863 and a final RMSE of about 6.244.

	Weight1	Weight2	Weight3	Test Score	sum_weights
47	0.1	0.6	0.3	0.862820	1.0
39	0.1	0.5	0.4	0.862796	1.0
119	0.2	0.5	0.3	0.862615	1.0
111	0.2	0.4	0.4	0.862521	1.0
55	0.1	0.7	0.2	0.862300	1.0

Lastly, the features were plotted in order of importance (contribution). Because Voting Regressors do not currently support the feature importance method, a function was created to calculate the feature importances for each individual contributing estimator model, and these feature importances were weighted in the same way that the final voting regressor was weighted. The feature found to most affect total number of COVID-19 cases was day_of_year, but surprisingly, TAVG scored next to last. One possible interpretation of this finding may be that average temperature could have less of a direct influence on COVID-19 cases than originally hypothesized. Instead, there

may be a stronger correlation between the indirect effect that temperature has on behavioral patterns (like meeting indoors vs outdoors) and other events (like school attendance) that in turn happen to correspond to temperature.