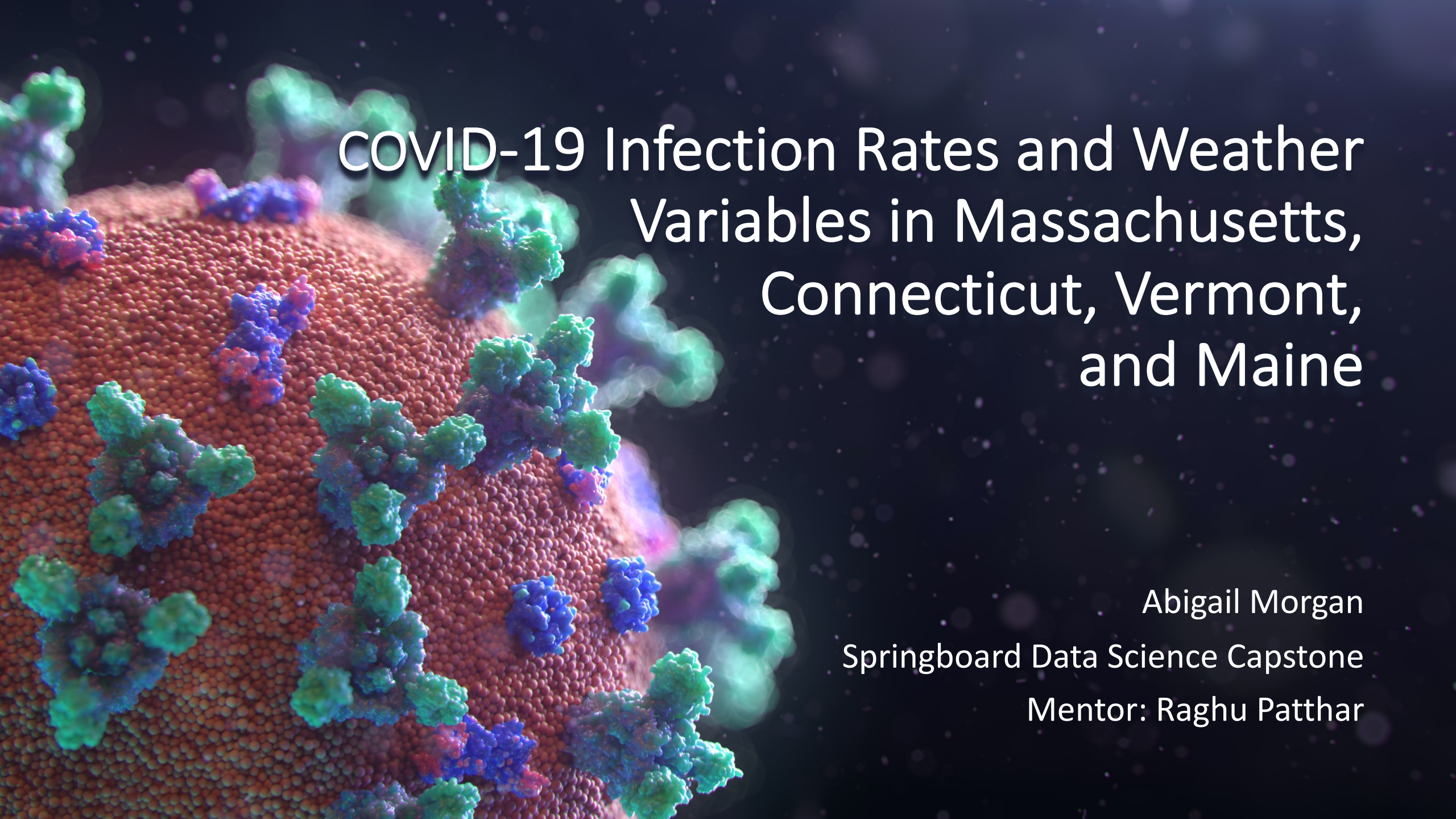


COVID-19 Infection Rates and Weather Variables in Massachusetts, Connecticut, Vermont, and Maine

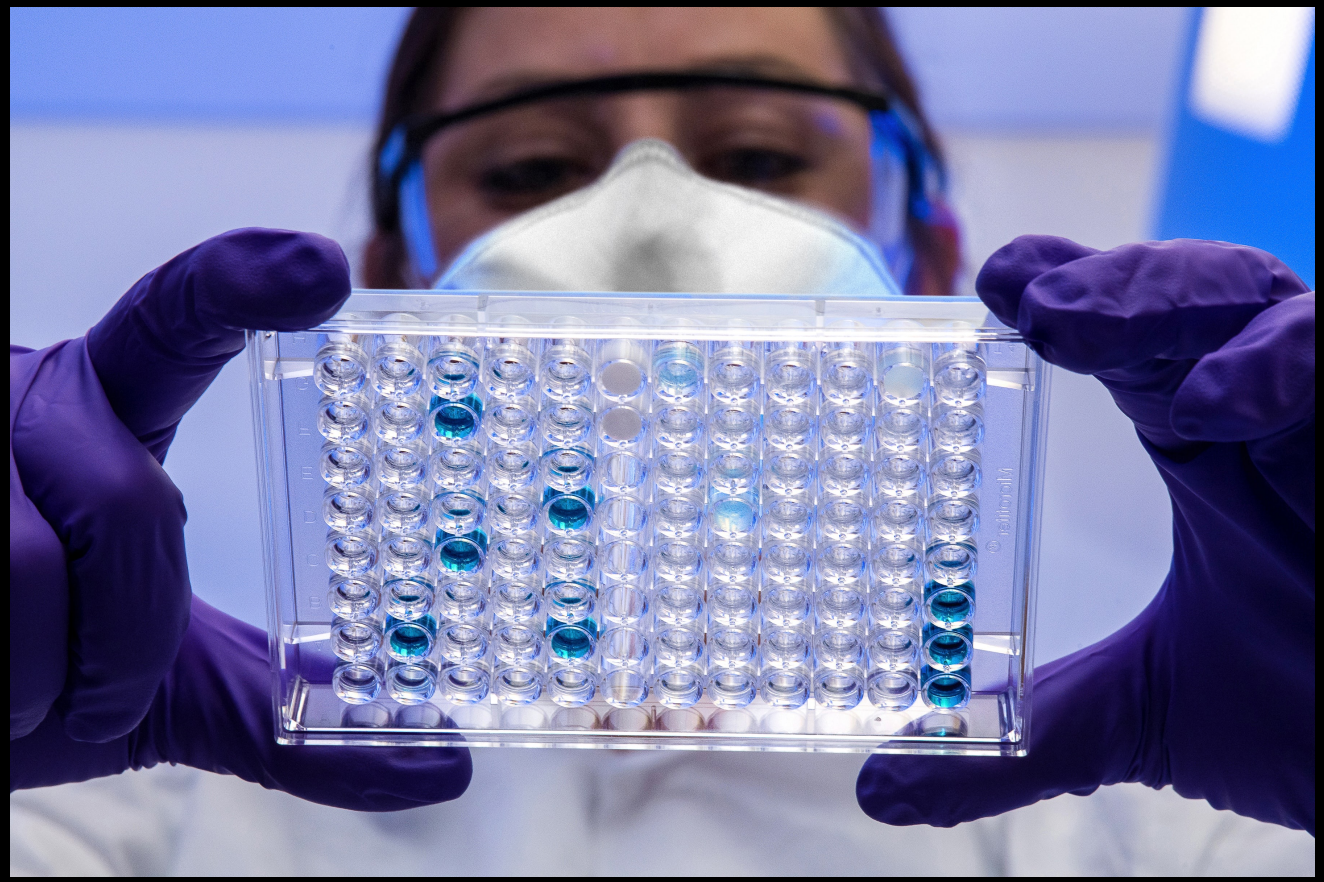
Abigail Morgan

Springboard Data Science Capstone

Mentor: Raghu Patthar

Epidemiological effects of weather variables

- Transmission dynamics
- Host susceptibility
- Virus survival



Behavioral effects of weather variables

- Social distancing

- Mobility levels

- Frequency and location of social gatherings

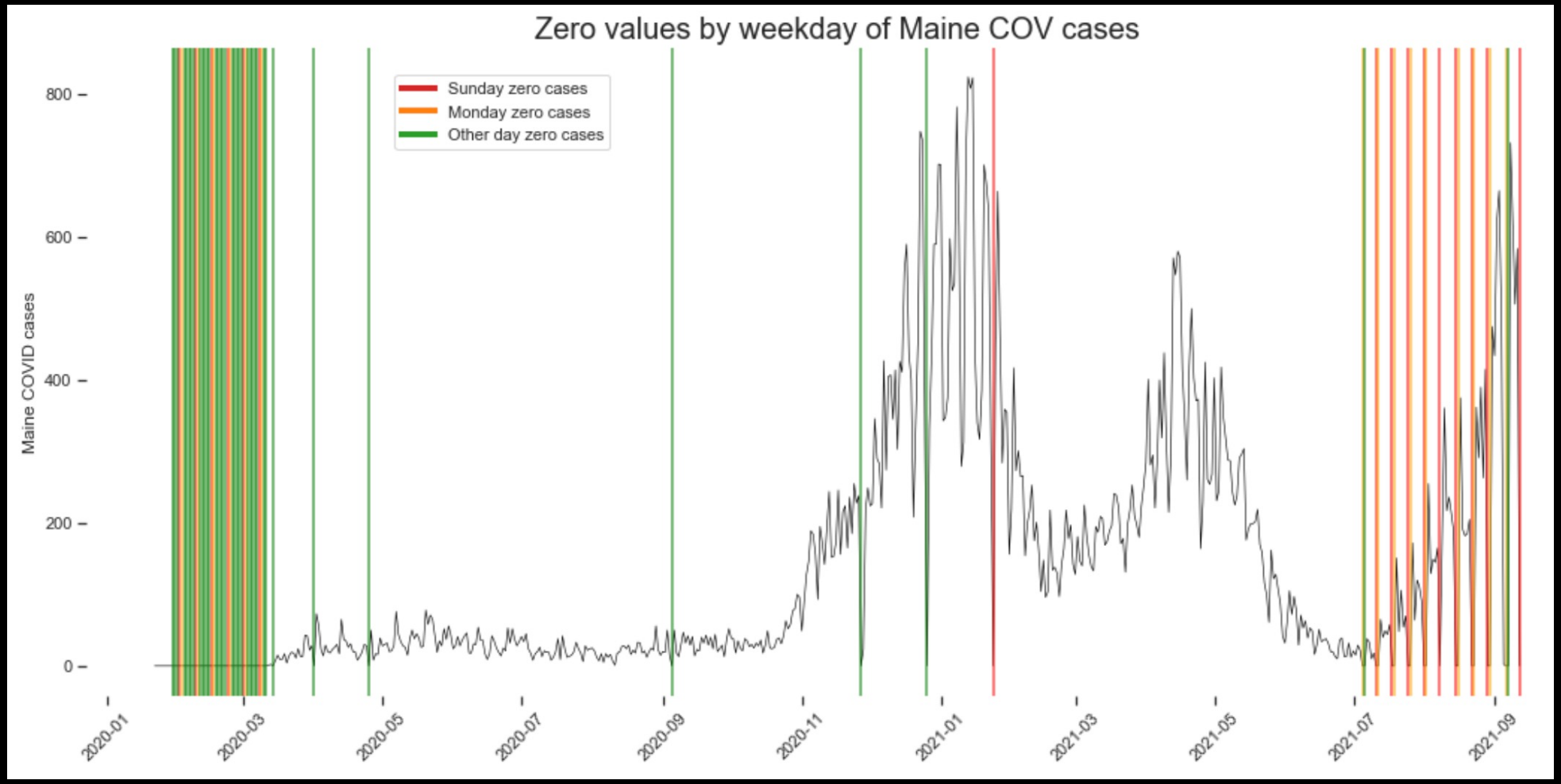
Guiding questions:

- 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?
- What are the capabilities and limitations of a model that uses these weather variables to predict COVID-19 infection rates?
- What other interpretations or explanations may exist to explain the results of this model?

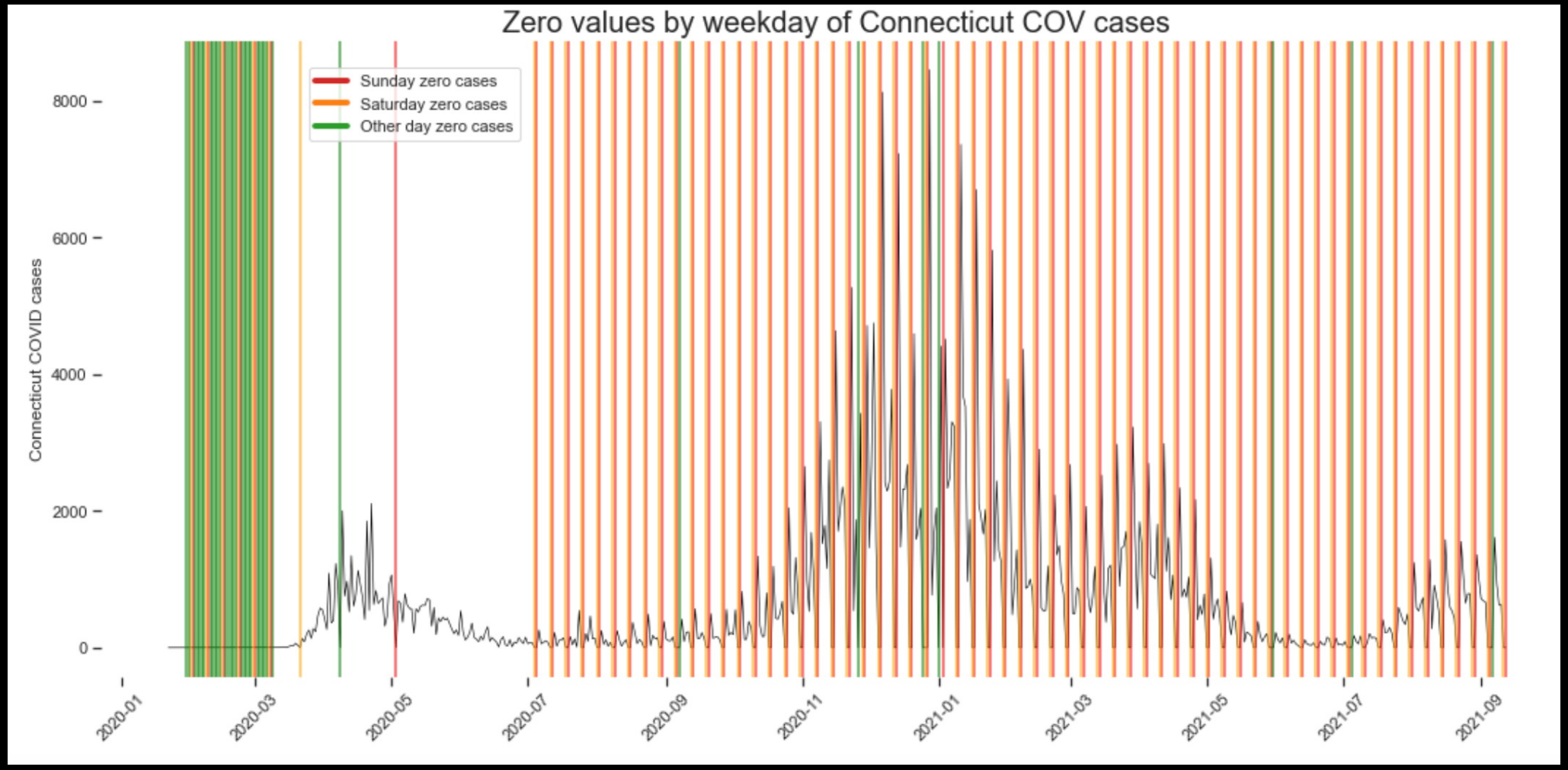
Data Sources

- NOAA's National Centers for Environmental Information Database
  - TAVG: Daily Average Temperature
    - TMIN: Daily Minimum Temperature
    - TMAX: Daily Maximum Temperature
  - PRCP: Daily Total Precipitation
- COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
  - Daily confirmed cases
- 03/01/2020 – 09/12/2021

Case Reporting Schedules: Maine



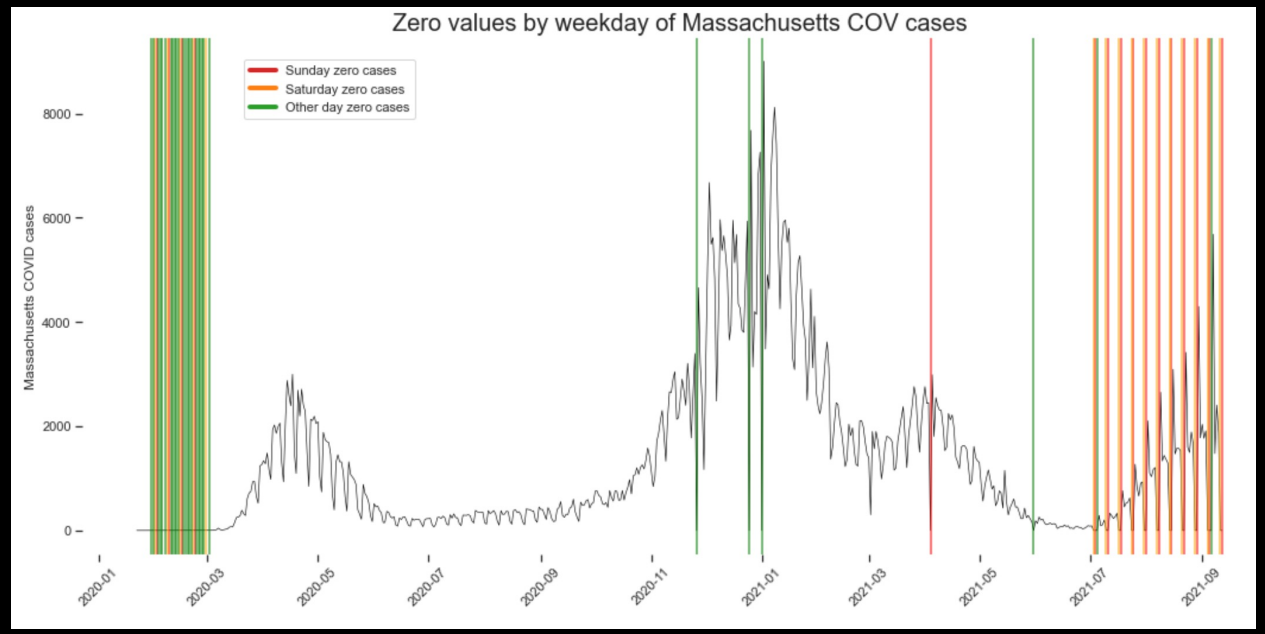
Case Reporting Schedules: Connecticut

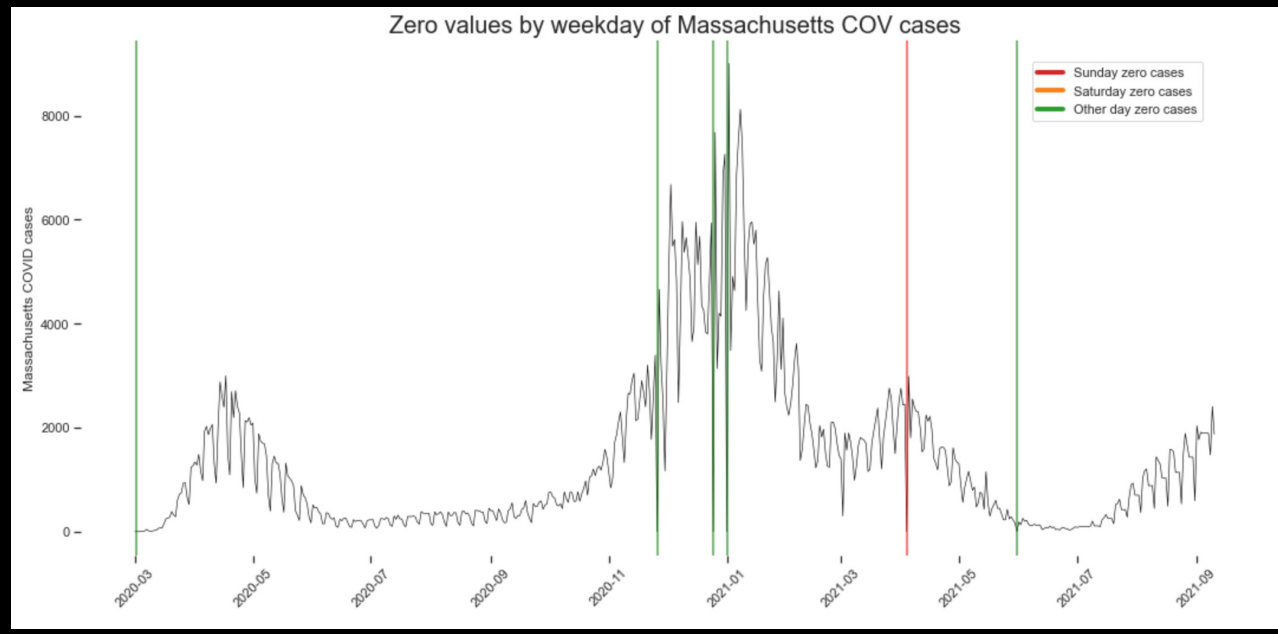


Case Reporting Schedules:

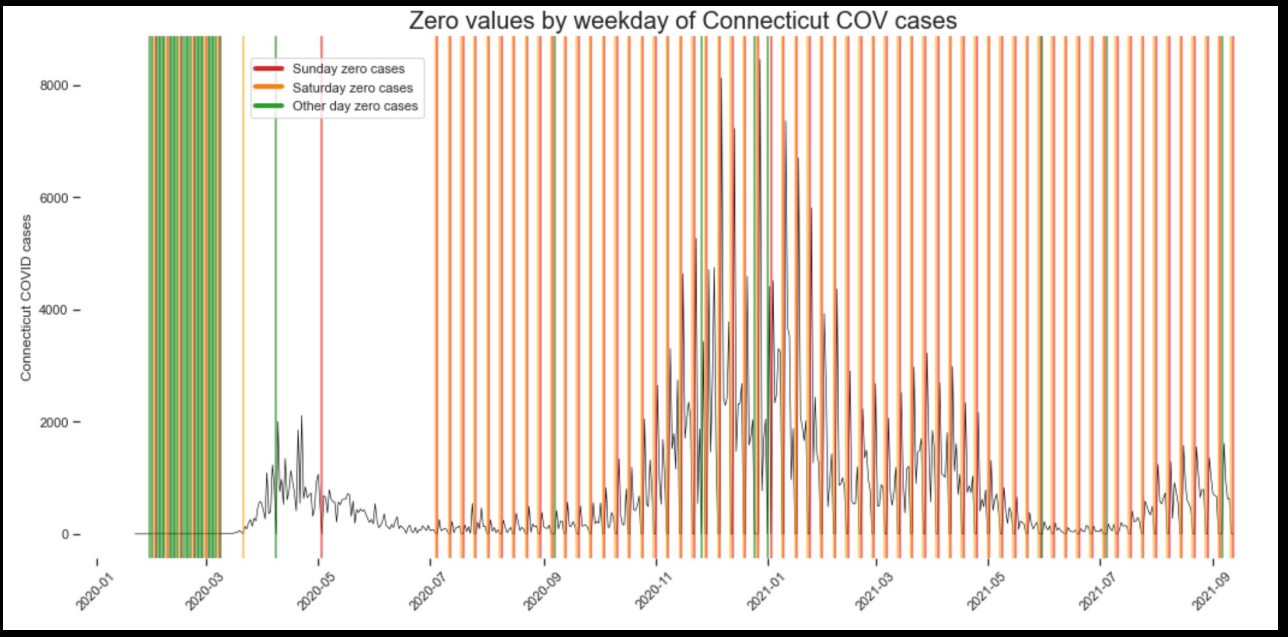
- Maine: Tuesday- Saturday from 2020-07-01 onwards (and daily before this date)
- Massachusetts: Monday- Friday from 2021-07-01 (and daily before this date)
- Vermont: Daily 7 days a week except for 2021-06-01 until 2021-08-23 when it temporarily shifted to a Monday- Friday reporting schedule
- Connecticut: Monday- Friday from 2020-07-01 onwards (and daily for only a very short period before this date)

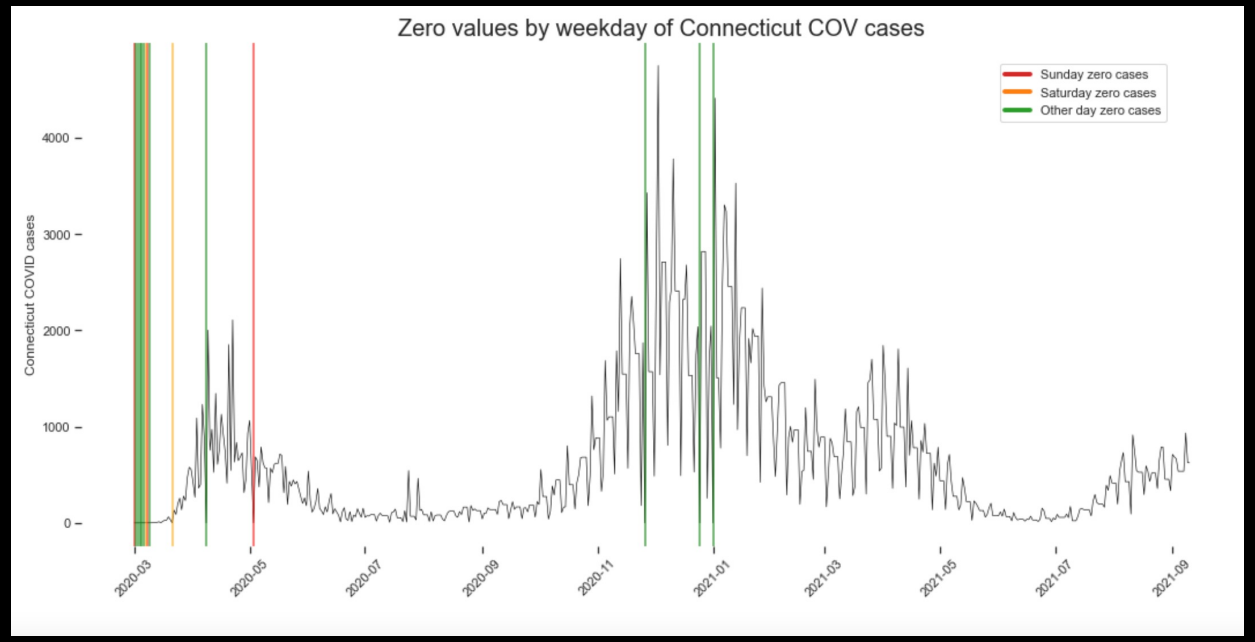
MA Missing values: Before and After



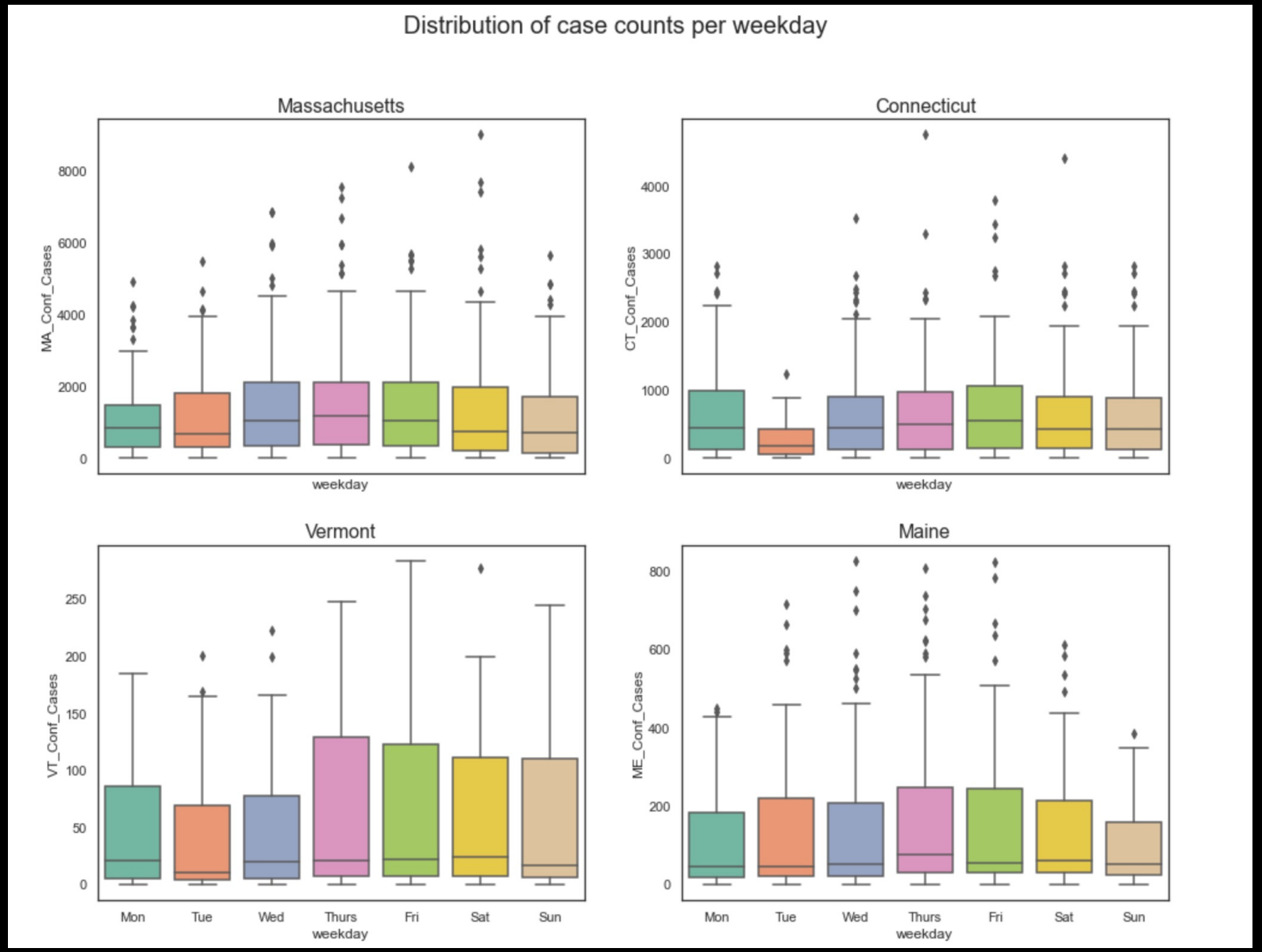


CT Missing Values: Before and After

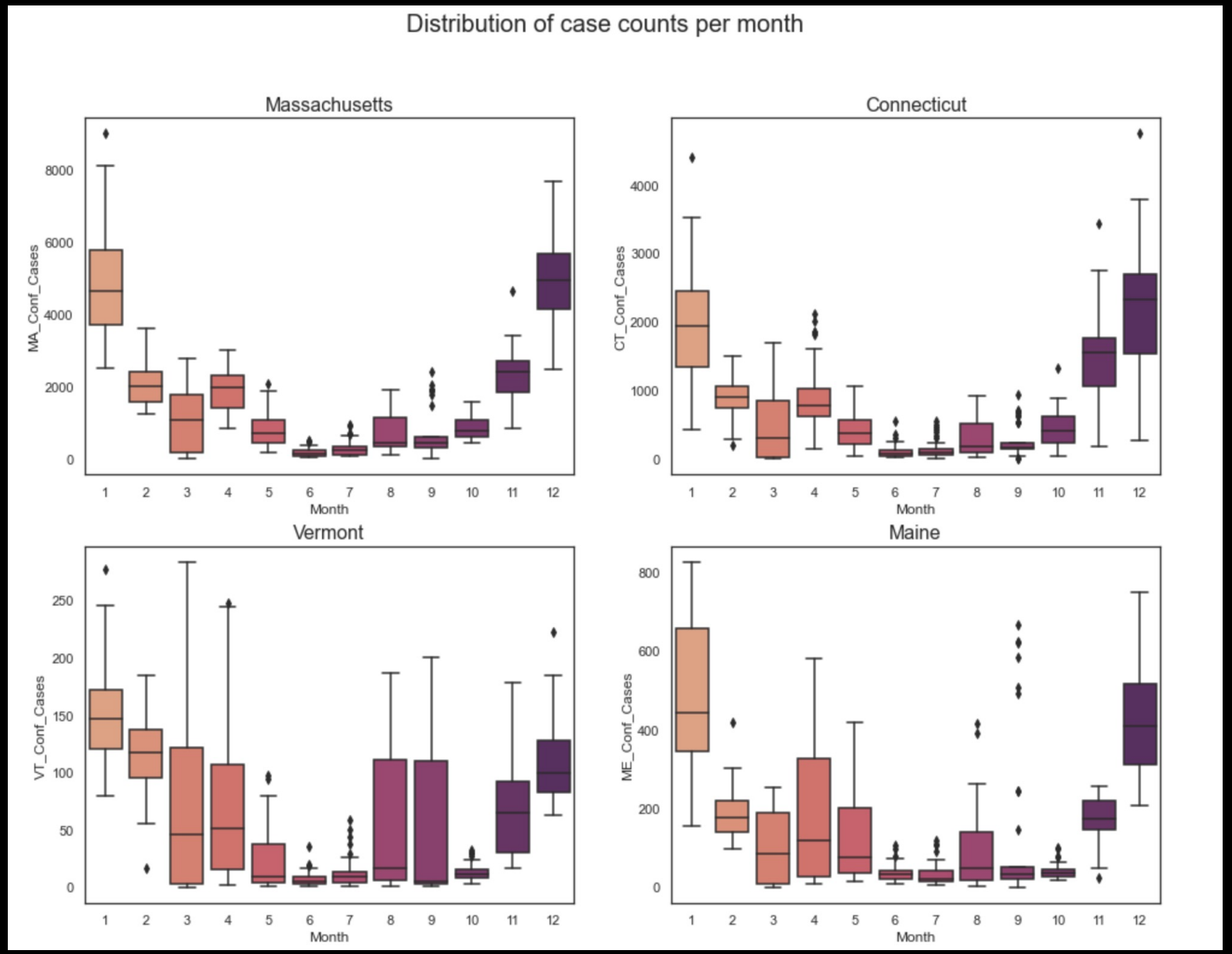




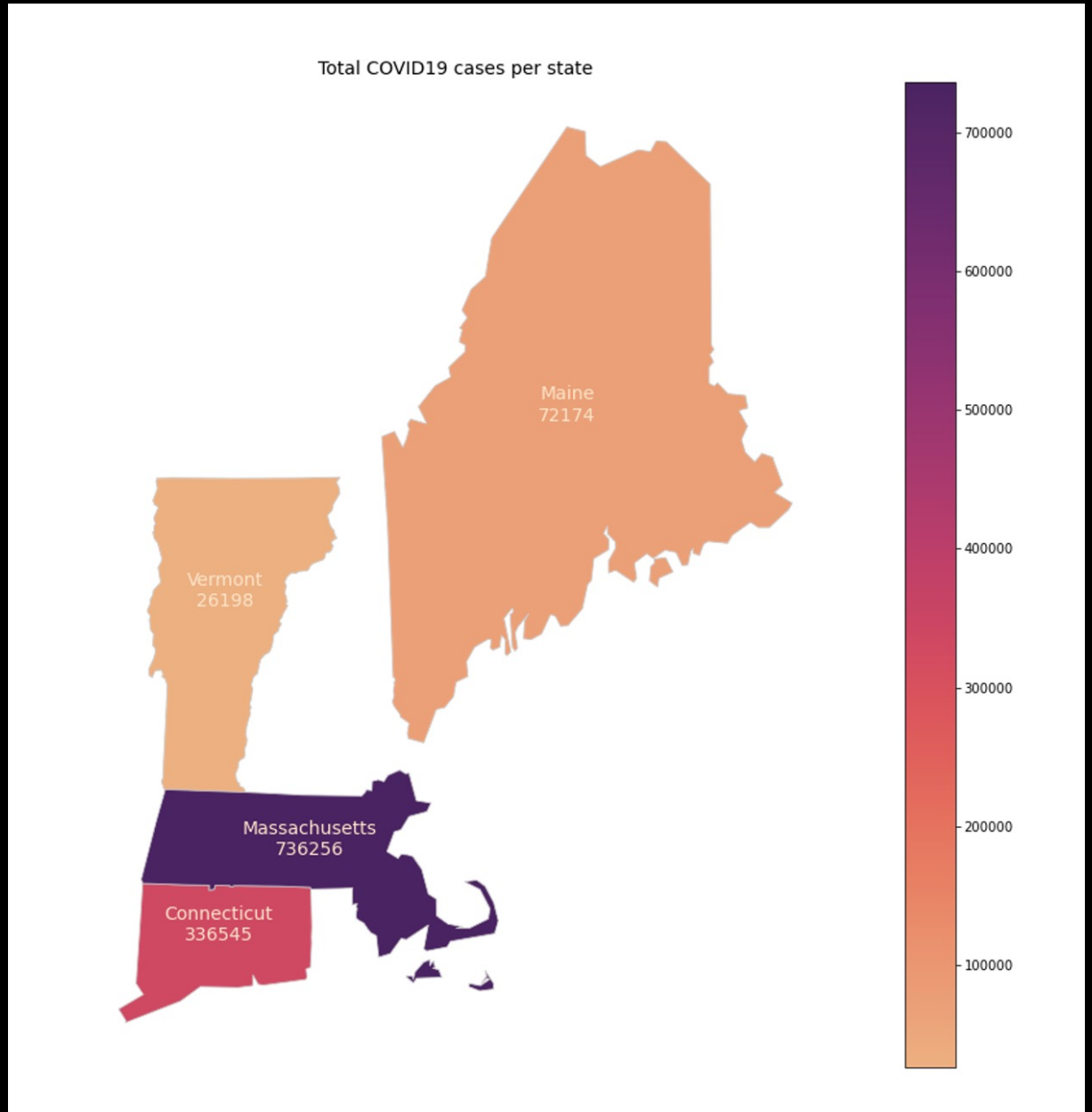
Distribution of case counts per weekday



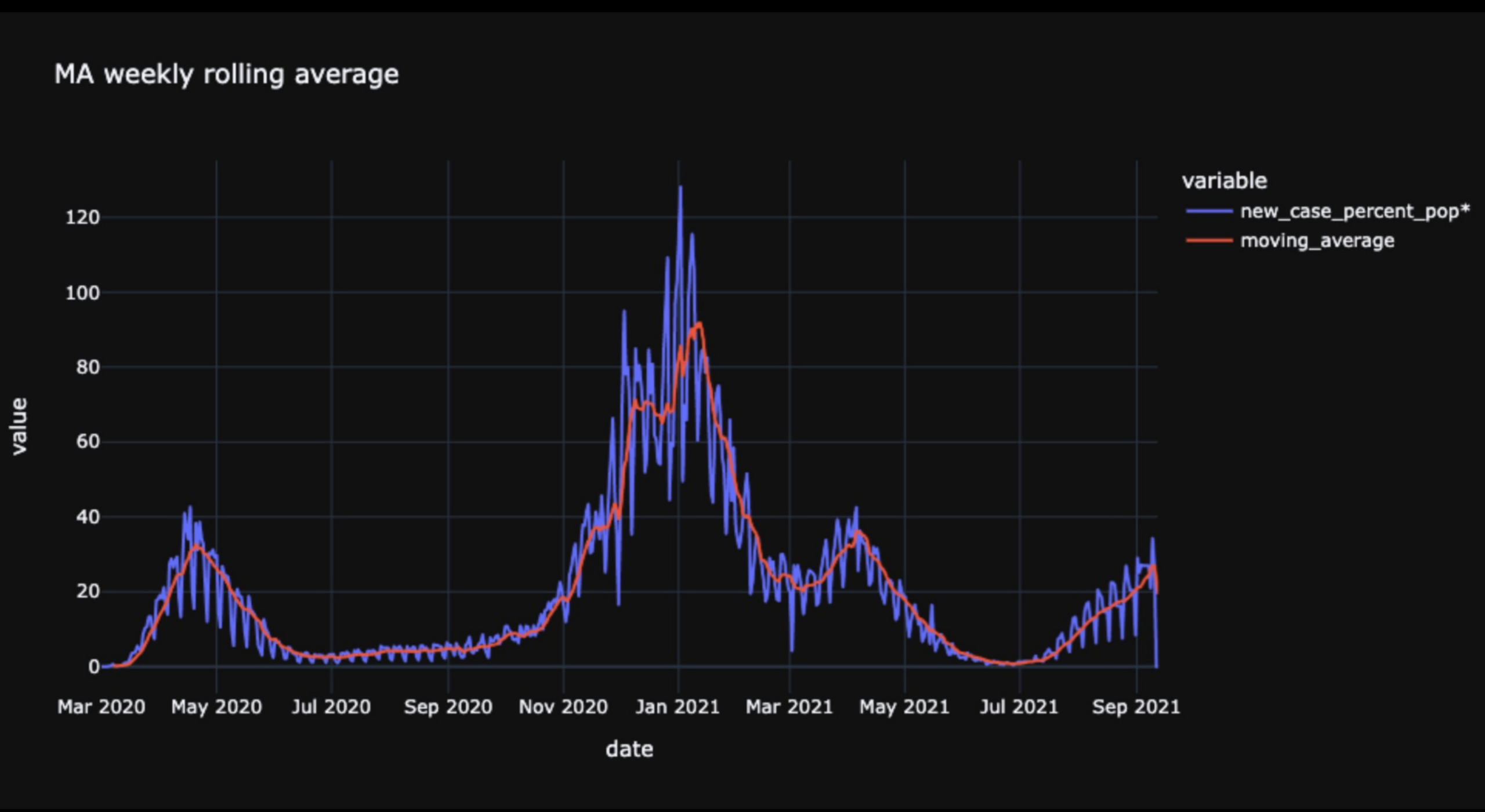
Distribution of case counts per month

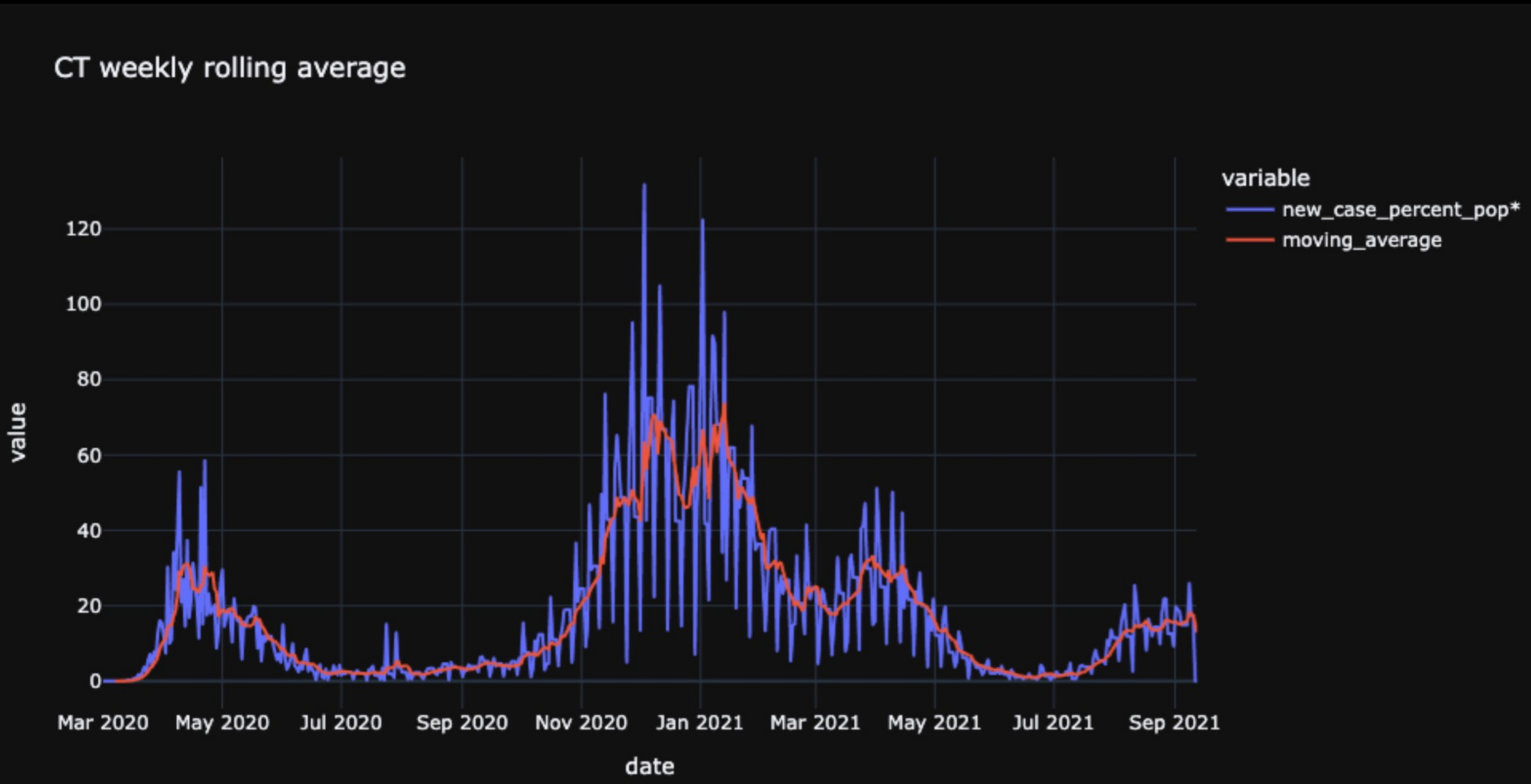


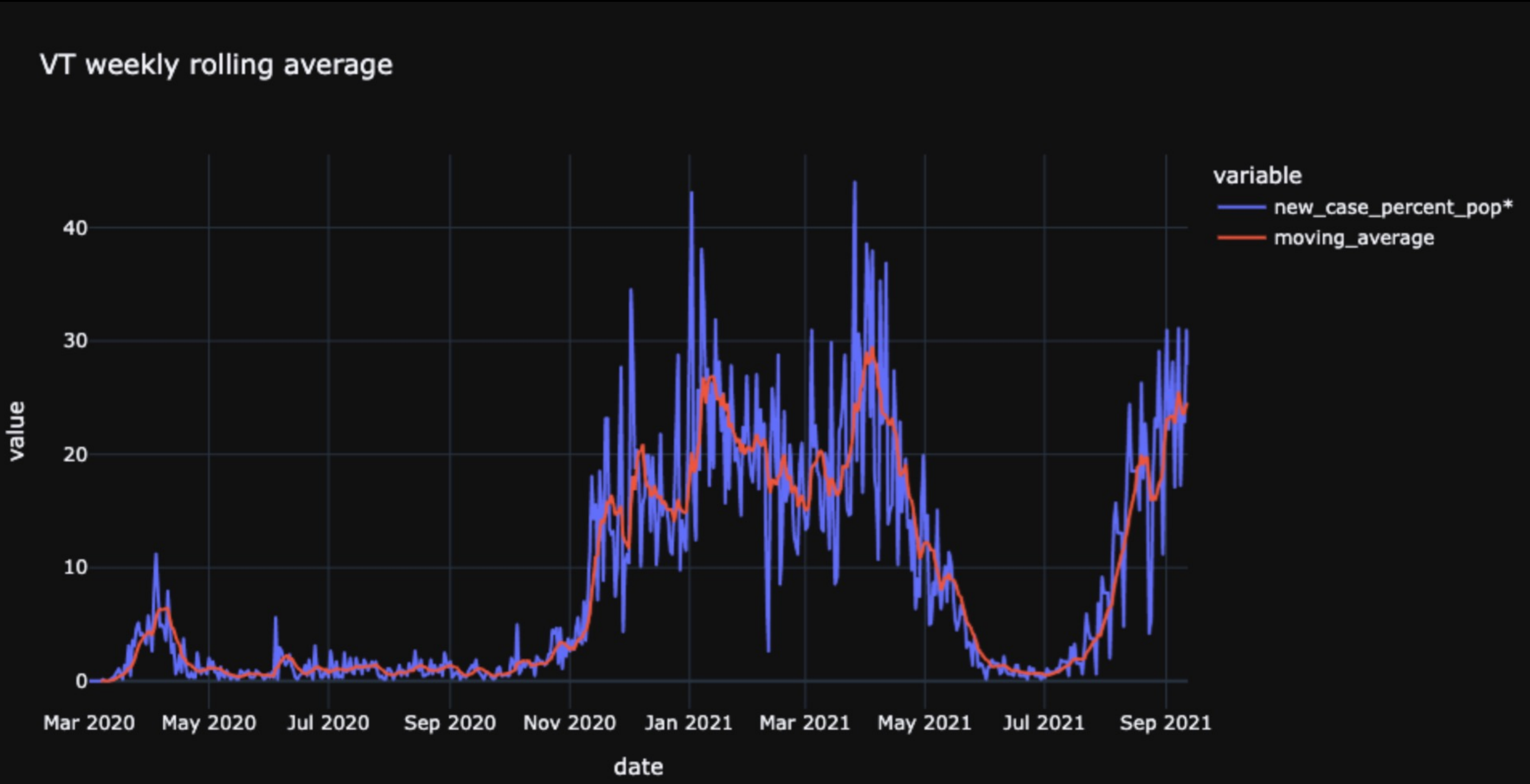
Distribution of case counts per state

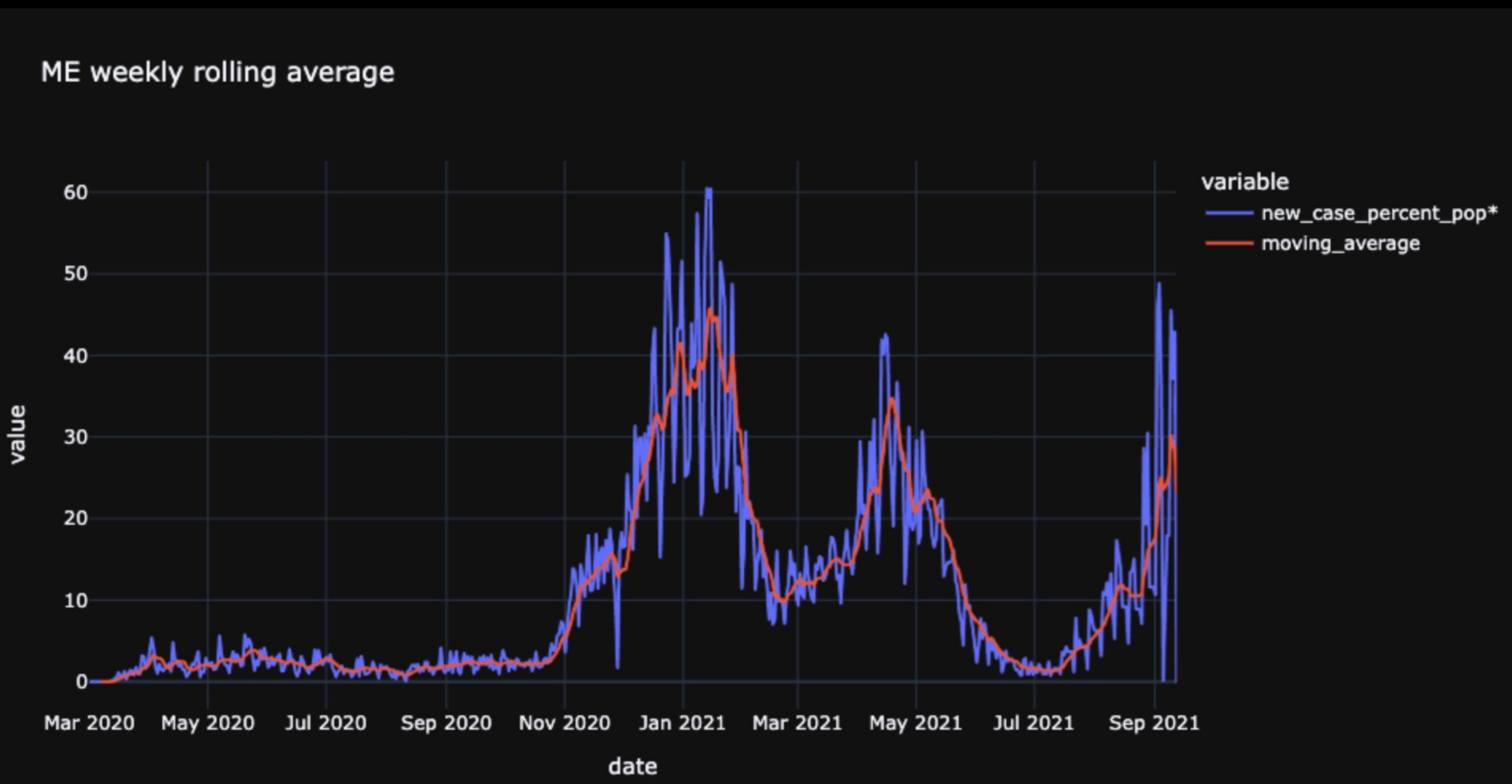


Visualizing patterns over time


MA weekly rolling average
variable
new_case_percent_pop*
moving_average
120
100
80
60
40
20
0
value
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
Sep 2021
date

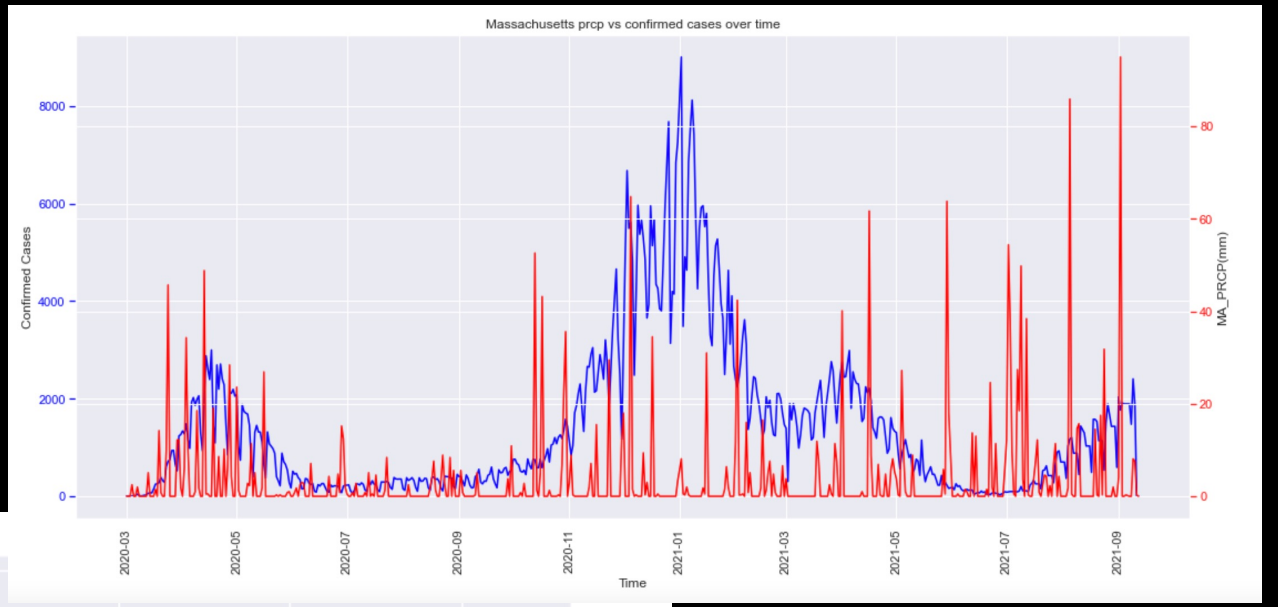

CT weekly rolling average
variable
new_case_percent_pop*
moving_average
value
120
100
80
60
40
20
0
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
Sep 2021
date

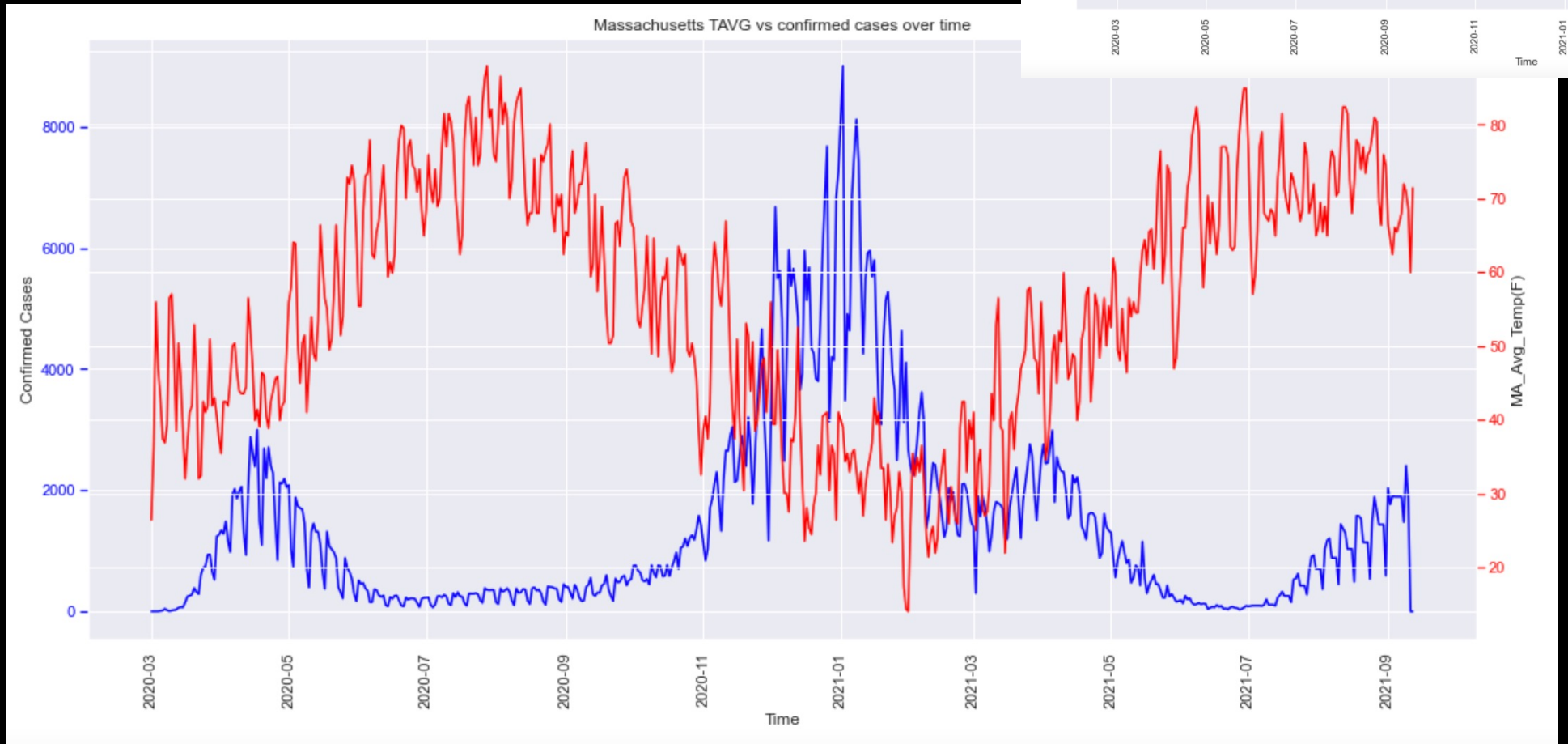

VT weekly rolling average
variable
new_case_percent_pop*
moving_average
40
30
20
10
0
value
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
Sep 2021
date


ME weekly rolling average
variable
new_case_percent_pop*
moving_average
value
60
50
40
30
20
10
0
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
Sep 2021
date

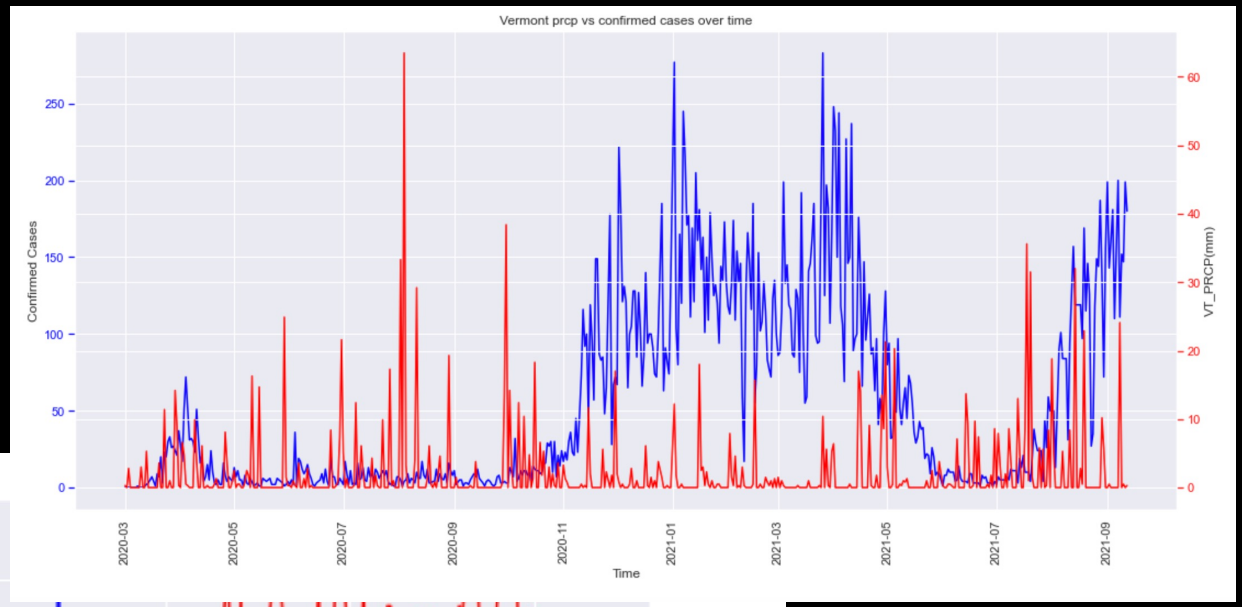
Visualizing relationships between features

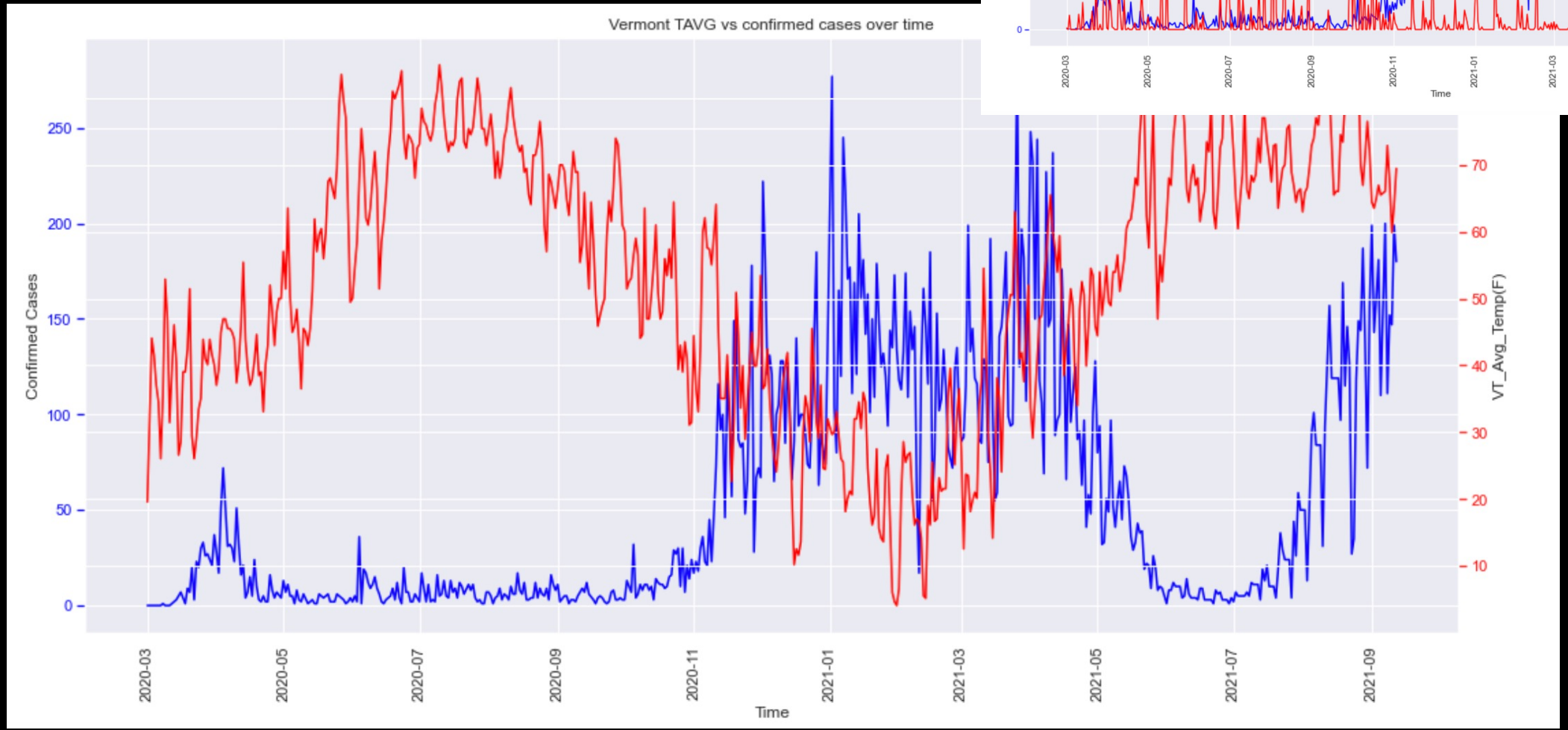
Massachusetts



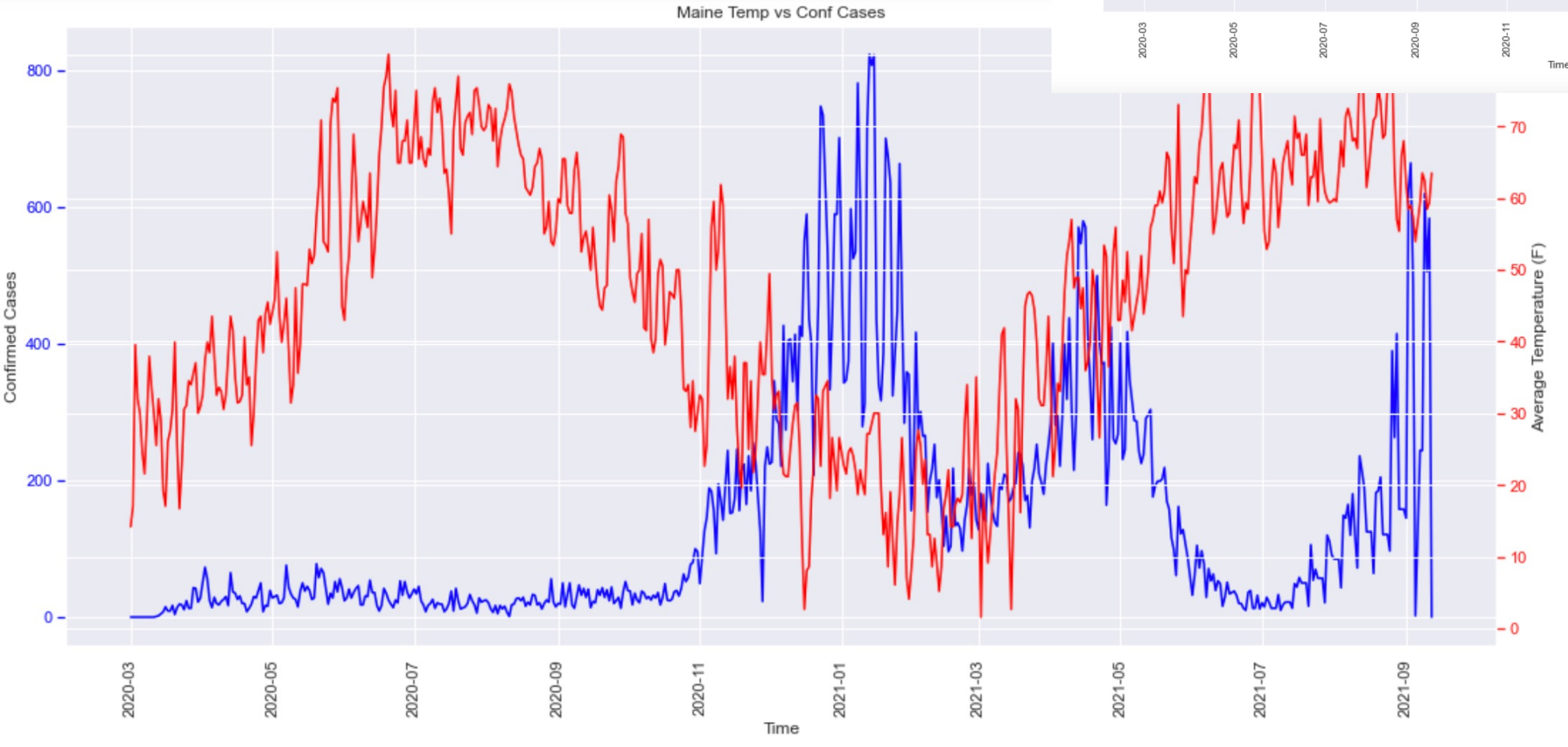


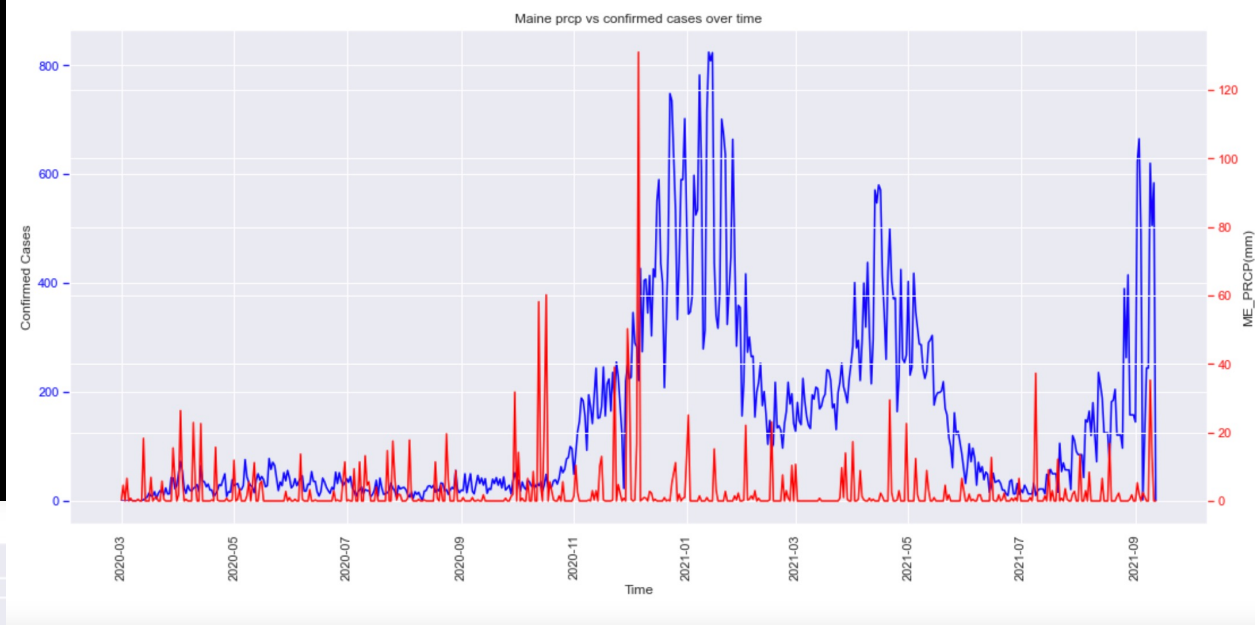
Vermont


Vermont prcp vs confirmed cases over time
Confirmed Cases
VT_PRCP(mm)
Time

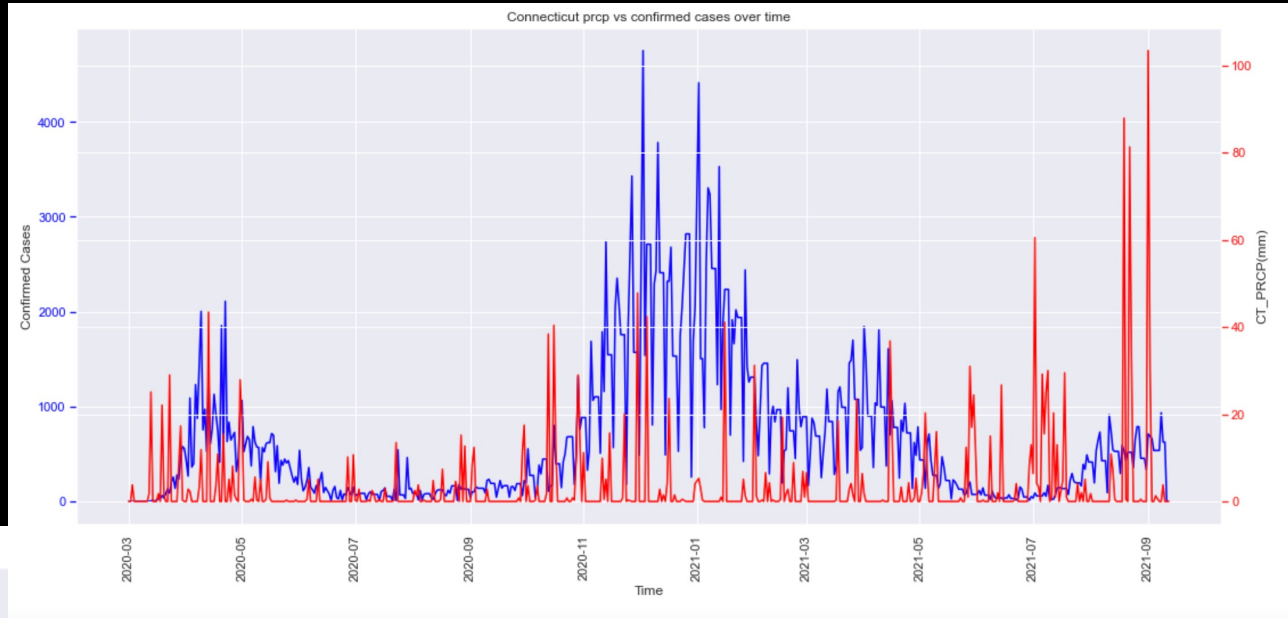

Vermont TAVG vs confirmed cases over time
Confirmed Cases
VT_Avg_Temp(F)
Time

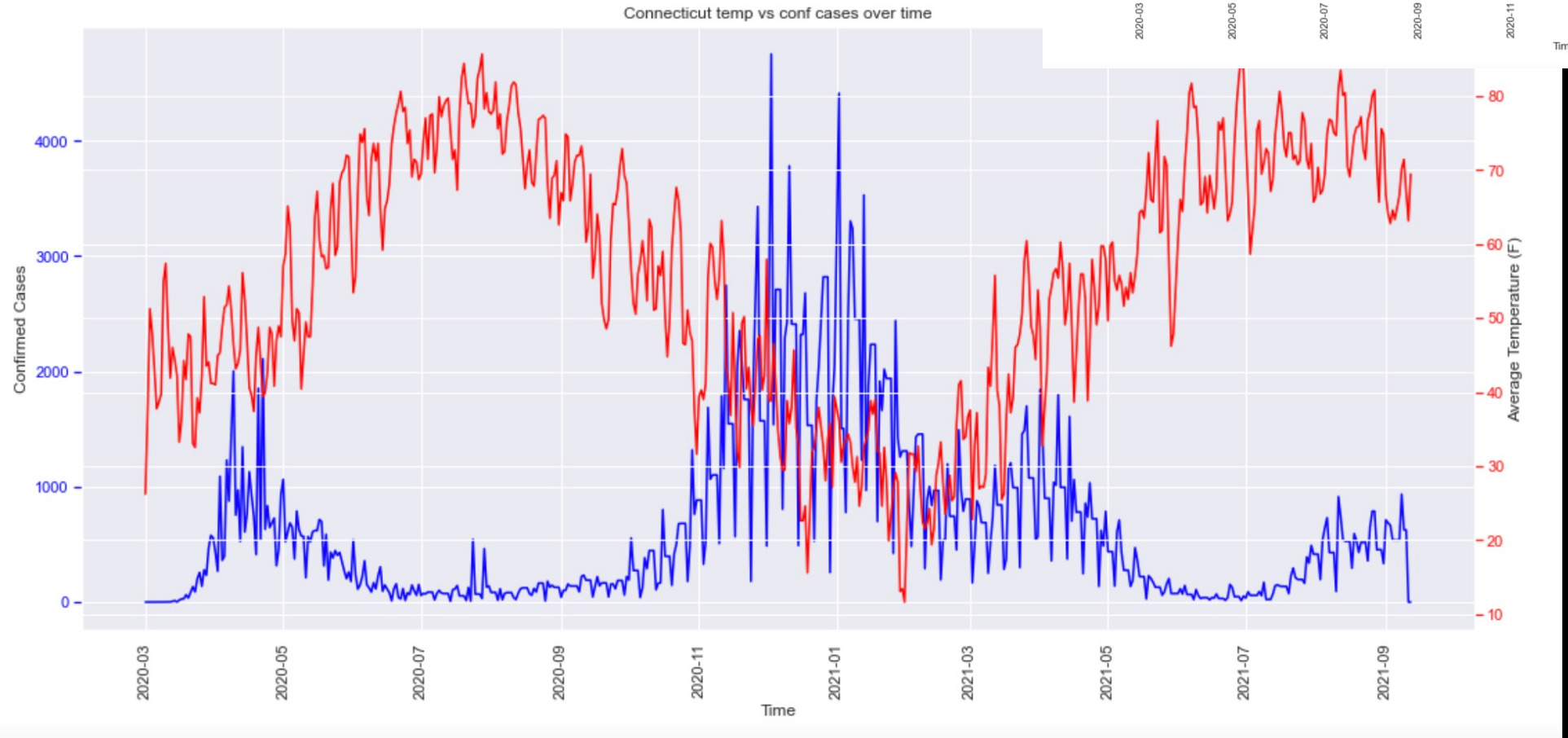
Maine


Maine Temp vs Conf Cases
Confirmed Cases
Average Temperature (F)
Time
0
200
400
600
800
0
10
20
30
40
50
60
70
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09


Maine prcp vs confirmed cases over time
Confirmed Cases
ME_PRCP(mm)
Time
0
200
400
600
800
0
20
40
60
80
100
120
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09

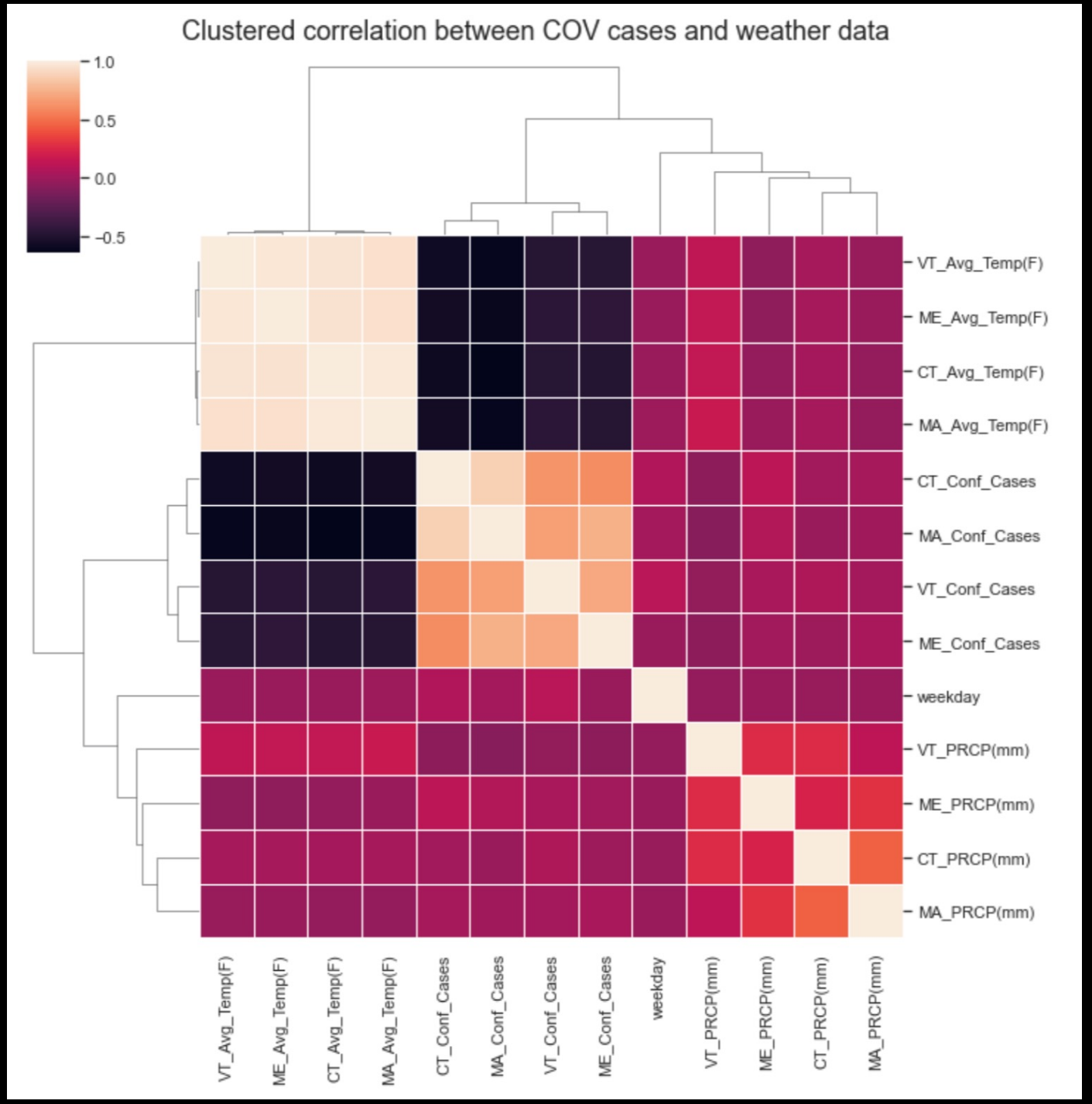
Connecticut





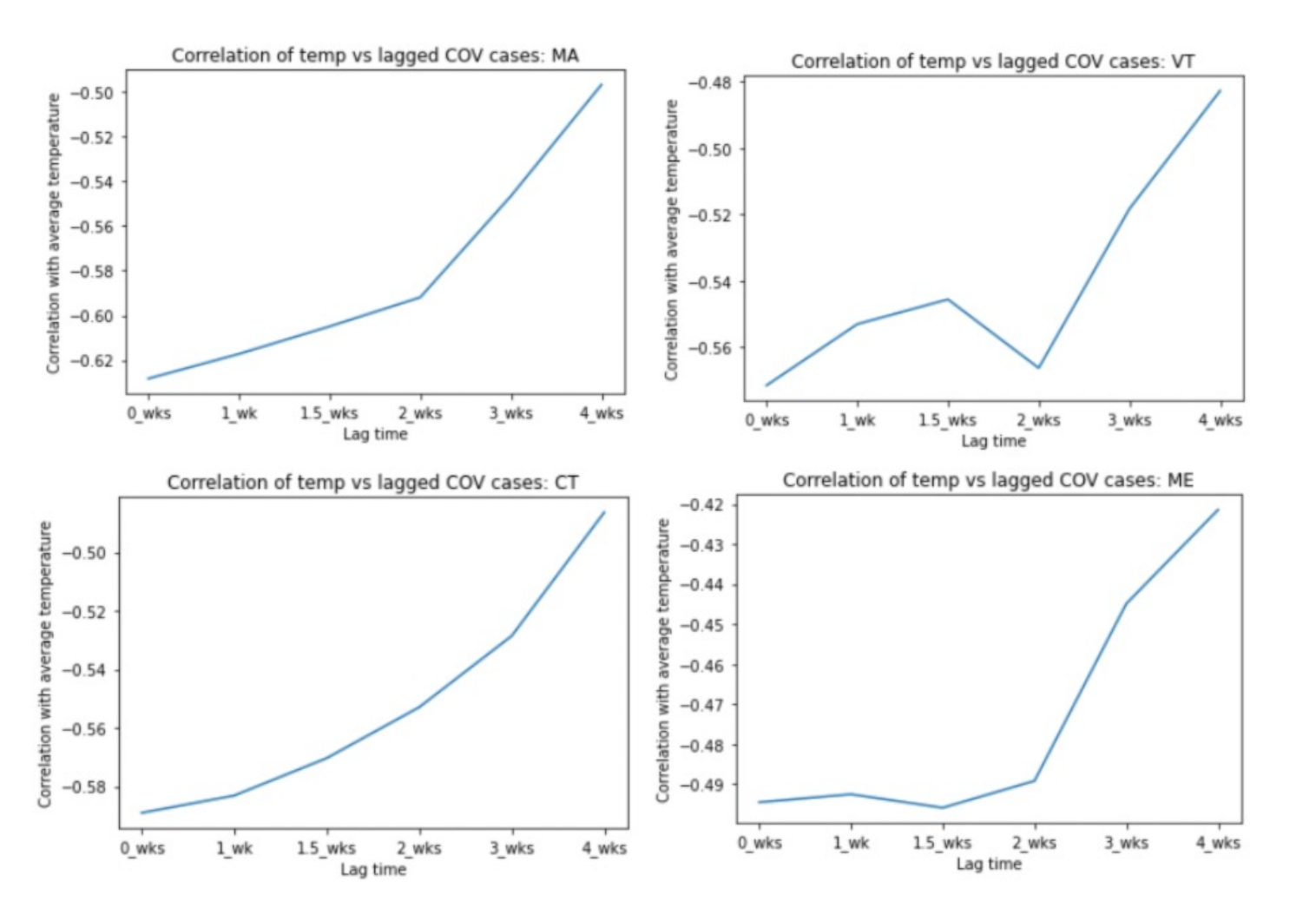
Feature correlation

- Average temperatures and confirmed case counts: near-black squares
- Massachusetts values are grouped most closely with Connecticut values = geographical grouping
- Vermont values grouped most closely with Maine values = geographical grouping



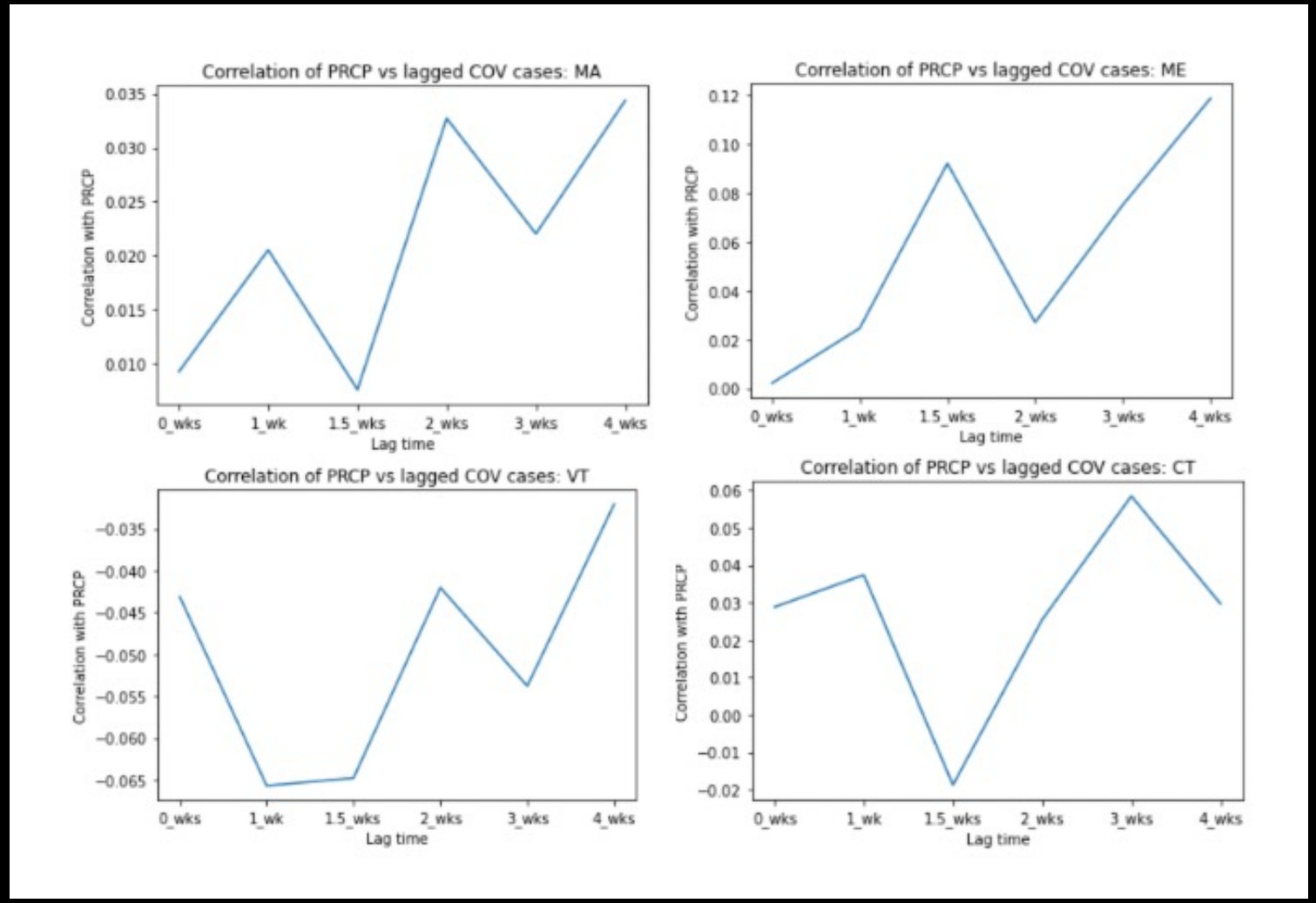
Feature Engineering

Lagged weather variables: TAVG



- Possibly a correlation between some states and a lagged value of TAVG
- At least 50% of the data do not indicate a correlation with lagged values
- Lagged TAVG will not be considered in this project

Lagged weather variables: PRCP



- All states demonstrate a higher correlation between lagged PRCP values and COVID-19 cases
- None agree on a optimal lagged value
- Maximum correlation improvement with lagged PRCP: 0.1
- There is no clear pattern in the trends, suggesting these results may very likely be random

One-hot-encoding categorical variables

- ‘state_id’ must be one-hot-encoded
- ‘Month’ and ‘day_of_week’ could be represented categorically or numerically

  - With only ‘state_id’ one-hot-encoded

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	4.0767	51.4359	7.1543	0.8280	0.5081	1.4200	0.3667
rf	Random Forest Regressor	3.9441	52.8238	7.2373	0.8233	0.4268	1.0654	0.1200
et	Extra Trees Regressor	3.8059	54.6242	7.3398	0.8205	0.3953	0.7701	0.1067
lightgbm	Light Gradient Boosting Machine	4.2811	55.2234	7.4053	0.8149	0.5077	1.5072	0.2067

  - With all three columns one-hot-encoded

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	4.0272	49.6088	7.0267	0.8340	0.5039	1.4677	0.4667
et	Extra Trees Regressor	3.7804	53.9427	7.2986	0.8225	0.3934	0.7279	0.1067
rf	Random Forest Regressor	3.9468	53.1076	7.2567	0.8225	0.4255	1.0604	0.1200
lightgbm	Light Gradient Boosting Machine	4.2811	55.2234	7.4053	0.8149	0.5077	1.5072	0.2800

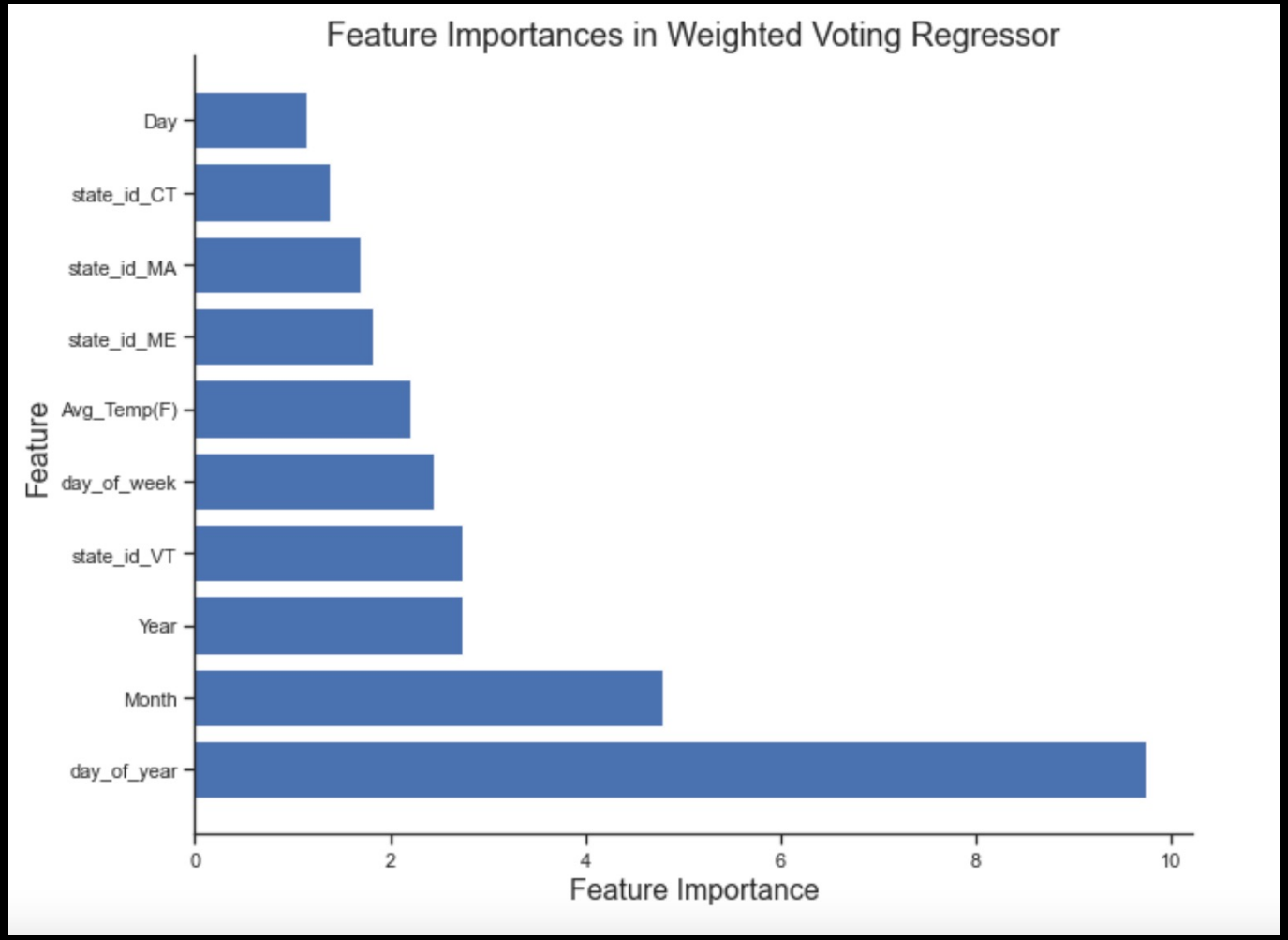
Weights and results

After tuning the hyperparameters of a Random Forest Regressor, Extra Trees Regressor, and CatBoost Regressor, feeding them into a Voting Regressor and determining the optimal weights of each estimator, the final model metrics were:

	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

```
Final VotingRegressor R2:  0.8628204220836 06
Final VotingRegressor MSE:  38.99180635418692
Final VotingRegressor RMSE:  6.24434194725008
```

Feature Importance



Thank you!