Semester Project – Final report:

Vacationing during Covid-19

Casey Lewis / Amanda Schneider

Department of Computer Science, Indiana University – Southeast

SU21: INTRODUCTION TO DATA SCIENCE: 3234

Professor Suranga Hettiarachchi

August 6th, 2021

Task Definition:

The goal of this project in simplest terms has always been to incorporate live data surrounding the Covid-19 pandemic and predict upon those values using Neural Networks to generate an accurate estimate for where the safest places might be to travel to within the United States and its territories. This can prove to be a life-saving endeavor given enough time to train the Neural Network based upon this data given the increase in infection rates caused by the spread of the delta variant of Covid-19. We will do that by forecasting the percentage of the population of each state that is vaccinated on each date, on dates by state in the future.

The Data:

We have from 5 different sources incorporated into this project.

From *The New York Times*, we have leveraged their master rolling-average Covid-19 dataset for states/territories within the United States purview. In this dataset, we have timeseries data ranging from 1/12/2021 to 7/06/2021 that details the vaccination rates by state.

	date	total_vaccinations	people_fully_vaccinated	daily_vaccinations	state	vaxxed_per_hundred
175	2021-07-06	3392366.0	1617584.0	6066.0	Alabama	32.99
351	2021-07-06	678029.0	322988.0	3219.0	Alaska	44.15
527	2021-07-06	47310.0	21583.0	172.0	American Samoa	38.76
703	2021-07-06	6826215.0	3172380.0	48692.0	Arizona	43.58
879	2021-07-06	2261649.0	1043217.0	5162.0	Arkansas	34.57

	date	state	cases_avg	cases_avg_per_100k	deaths_avg
26964	2021-07-06	Northern Mariana Islands	0.00	0.00	0.00
26965	2021-07-06	Guam	6.57	3.90	0.00
26966	2021-07-06	Puerto Rico	55.00	1.62	0.71
26967	2021-07-06	Virgin Islands	12.00	11.30	0.00
26968	2021-07-06	Wyoming	63.86	11.03	0.57

Unfortunately, because the Covid-19 project here stopped tracking certain data figures, like confirmed cases, we were only able to extrapolate a certain amount of data from John Hopkins directly.

	Confirmed	Deaths	Recovered	Active	Incident_Rate	Case_Fatality_Ratio	state
0	551298	11358	NaN	NaN	11243.671206	2.060229	Alabama
1	71384	377	NaN	NaN	9757.977978	0.528130	Alaska
2	0	0	NaN	NaN	0.000000	NaN	American Samoa
3	897010	17979	NaN	NaN	12323.737824	2.004325	Arizona
4	351825	5920	NaN	NaN	11658.311806	1.682655	Arkansas

In combination with John Hopkins, Sglavoie moderates a popular Covid-19 dataset on GitHub that is utilizing a variety of other reputable sources to pull a complete confirmed-active metric in timeseries/state.

7]:

	date	state	Active
0	2021-07-06	Alabama	81328
1	2021-07-06	Alaska	31286
2	2021-07-06	American Samoa	0
3	2021-07-06	Arizona	562308
4	2021-07-06	Arkansas	41703

.

We were able to source state population data from the US Census Bureau.

	state	population_est
0	Alabama	4903185
1	Alaska	731545
2	Arizona	7278717
3	Arkansas	3017804
4	California	39512223

Finally, we aggregated all of this data, as well as created some formulas of our own to get a percentage of the population that is either vaccinated or infected.

	date	state	people_fully_vaccinated	Confirmed	cases_avg	deaths_avg	Deaths	Incident_Rate	Case_Fatality_Ratio	daily_vaccinations
0	2021-01-12	Alabama	7270.0	404000.0	4036.43	76.83	5347.0	8239.54	1.32	13402.0
1	2021-01-12	Alaska	5400.0	50413.0	277.29	1.00	224.0	6888.71	0.44	13402.0
2	2021-01-12	Arizona	8343.0	627541.0	9580.43	169.00	10147.0	8621.59	1.62	13402.0
3	2021-01-12	Arkansas	8.0	256344.0	2952.14	40.71	4081.0	8494.39	1.59	13402.0
4	2021-01-12	California	100089.0	2784716.0	43491.86	517.00	30519.0	6980.17	1.11	13402.0
...	...	...	...	...	...	...	...	...	...	...
9323	2021-07-06	Virginia	4478317.0	681194.0	180.14	3.14	11431.0	7980.70	1.68	16147.0
9324	2021-07-06	Washington	4206349.0	452483.0	331.86	6.86	5939.0	5942.08	1.31	19116.0
9325	2021-07-06	West Virginia	671593.0	164149.0	41.00	1.67	2899.0	9159.35	1.77	1529.0
9326	2021-07-06	Wisconsin	2930845.0	678008.0	73.86	2.57	8144.0	11644.75	1.20	11972.0
9327	2021-07-06	Wyoming	204636.0	62445.0	63.86	0.57	747.0	10789.47	1.20	1350.0

9328 rows × 16 columns

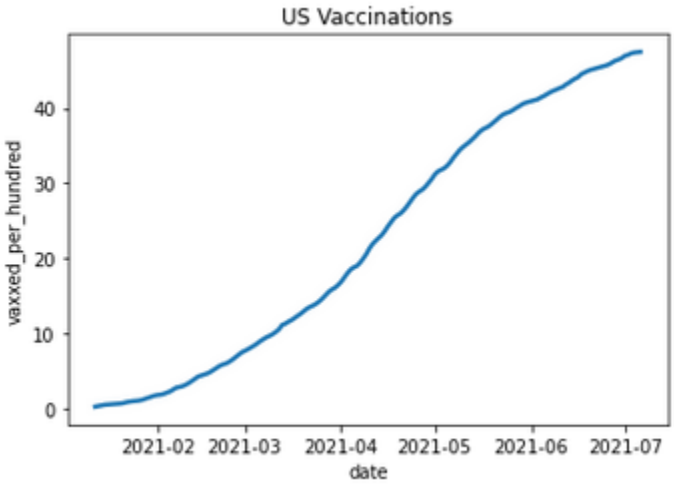
total_vaccinations	Recovered	active	total_state_pop	%_pop_vaxxed	%_pop_infected
78134.0	398653.0	59867	4903185.0	0.15	1.22
35838.0	50189.0	24214	731545.0	0.74	3.31
141355.0	617394.0	0	7278717.0	0.11	0.00
40879.0	252263.0	393732	3017804.0	0.00	13.05
816301.0	2754197.0	28788	39512223.0	0.25	0.07
...	...	...	...	...	...
9328521.0	669763.0	18069	8535519.0	52.47	0.21
8676836.0	446544.0	71755	7614893.0	55.24	0.94
1426490.0	161250.0	259885	1792147.0	37.47	14.50
5982183.0	669864.0	36407	5822434.0	50.34	0.63
431146.0	61698.0	104996	578759.0	35.36	18.14

Why we chose the data we chose:

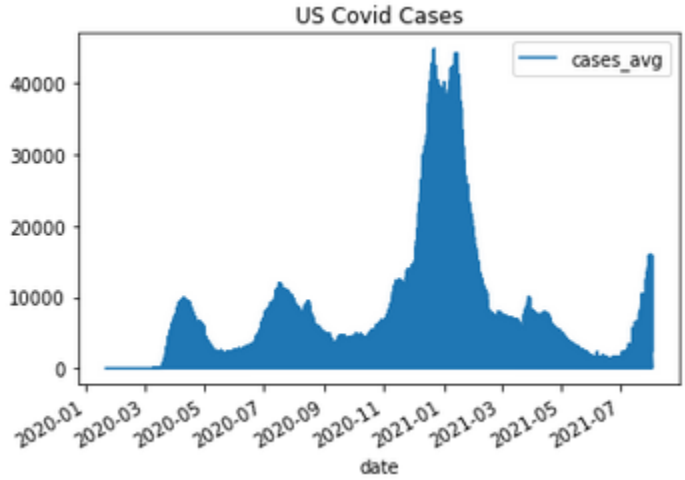
The first thought process behind getting a neural network to predict relative safety was that any data we were going to predict off of was going to have to be timeseries and also based off of state. As you can see in our dataset, it is essentially as if the date is our primary key and state is our secondary or foreign key. We ended up choosing column values from the various datasets because we felt these fields were the best to be able to come up with the best predictions, for example, in our case, we are going to be predicting the percentage of the population vaccinated to estimate which state may be at least relatively the safest to travel to. The rest of the data seemed like the most complementary data we can train the neural network with to make the most accurate prediction possible. This data may even lend itself to make additional predictions further on down the road after some refinement is done, and there is more data to compound onto what we already have. With the rise of the aforementioned delta variant, we could add data to this to see areas where the delta variant is the most widespread, i.e., warning a user that the particular state they are wishing to travel to may not be the safest by predicting which states will have the highest vaccination rates based upon the data provided to the LSTM (Long/Short Term Memory RNN).

Data Visualization:

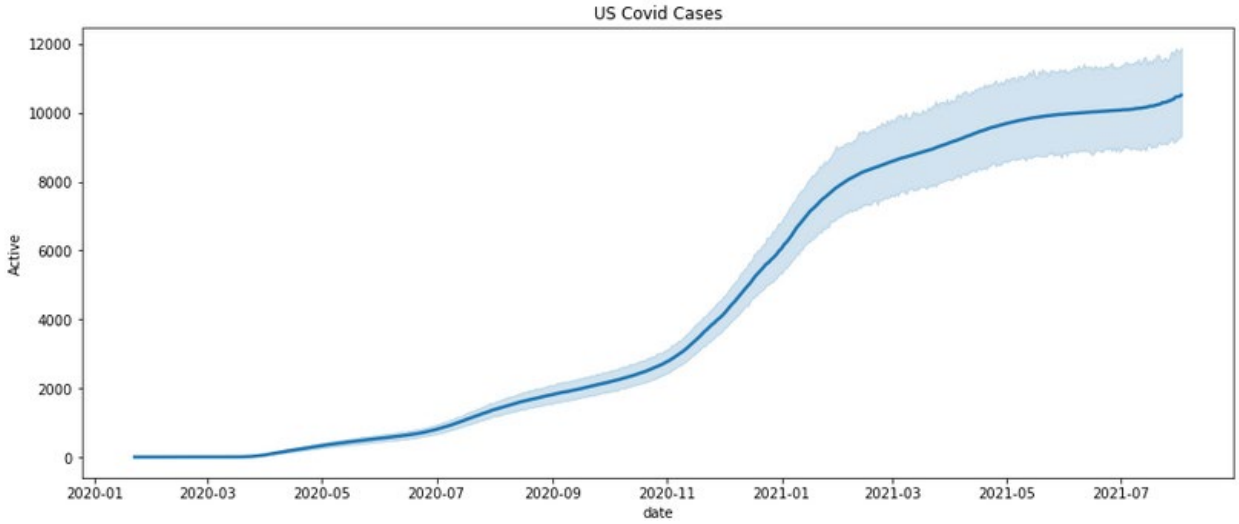
The US State Vaccination data set shows us the sharp rise in vaccinations across the US between January and July.



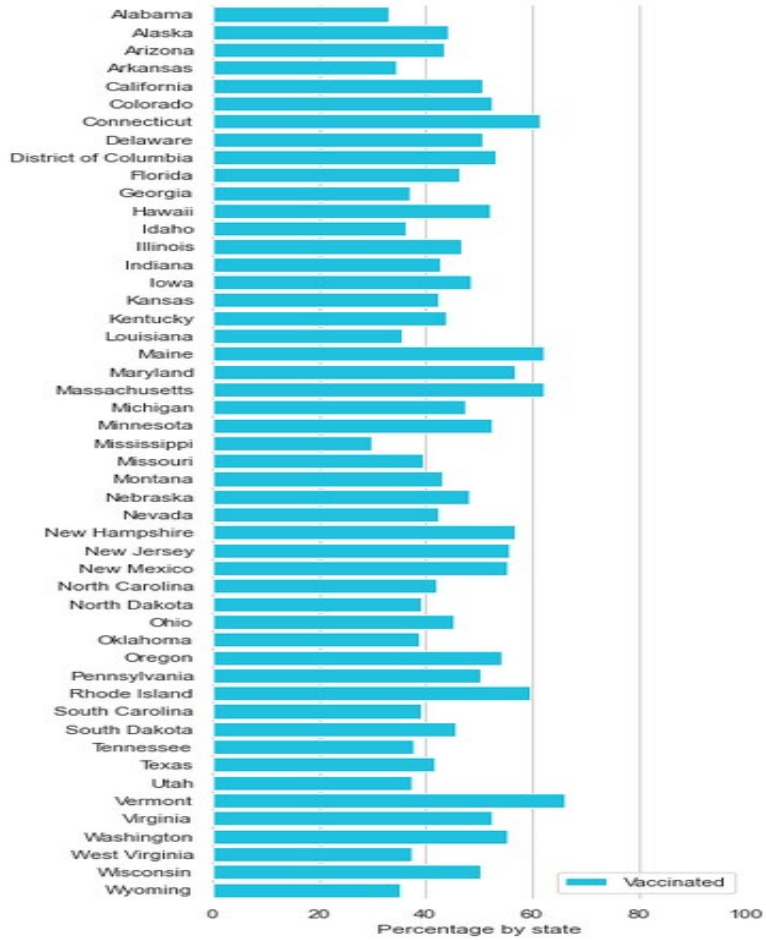
While the New York Times rolling average dataset reveals alarming spikes in covid cases across the country between January 2020 and July this year, while also revealing the rise in cases due to the spread of the delta variant.



The data sourced from John Hopkins reveals a sharp rise in Covid-19 cases from the onset of the beginning of last year to now.

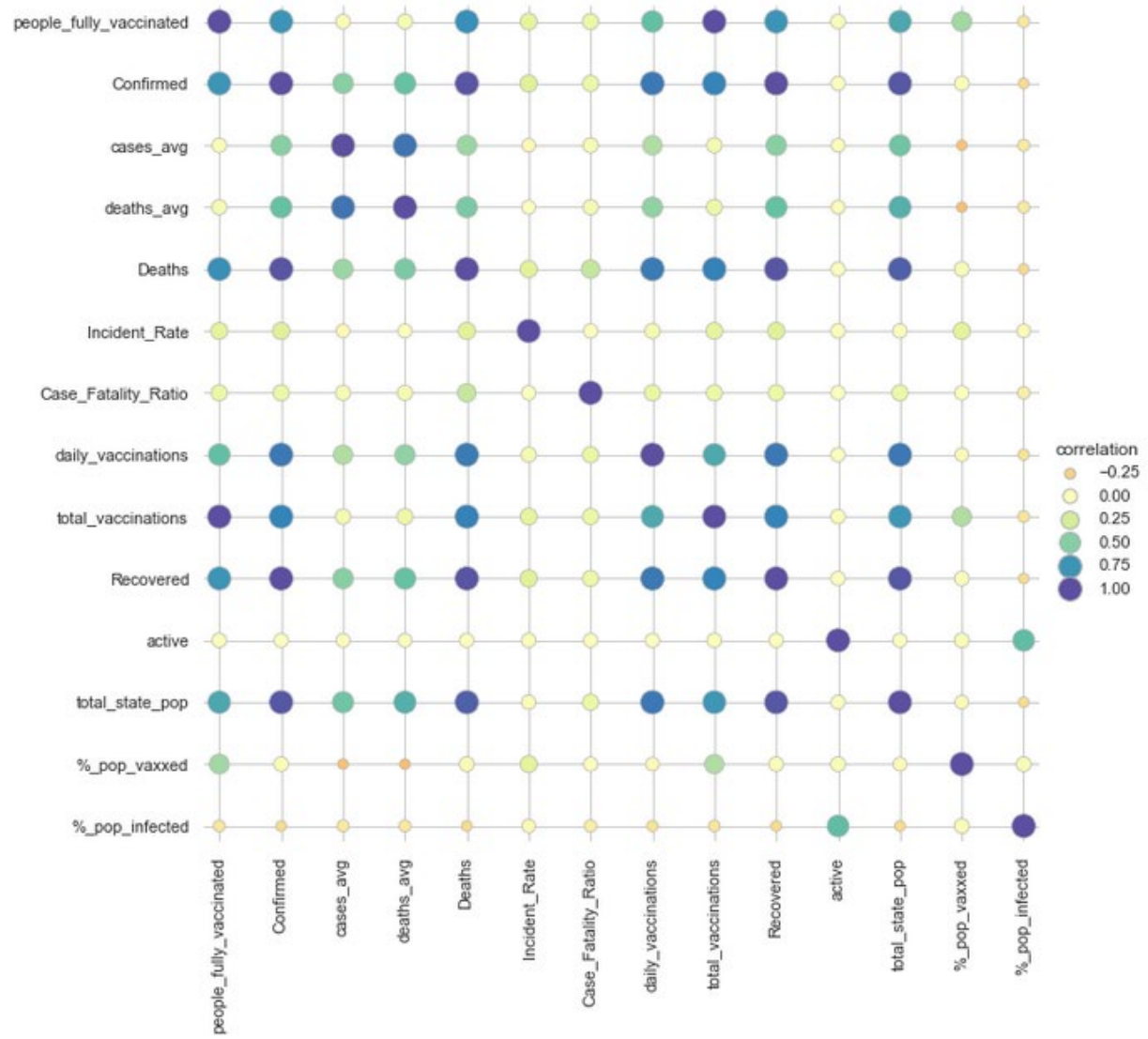


The combined data from John Hopkins and The New York Times reveals that the Northeastern states appear to be doing the strongest in terms of population vaccination percentages.



The final visualization done for this project was to show how strongly correlated the data was from the whole scope of the project. You can clearly see from the numbers how the data lines up

with the other features which will make for excellent training data for the neural network. This helps explain how our Y value for the neural network will be predicted because we can see the trend of correlation between the data clearly to the percentage of the population vaccinated. However, this will also help us make other predictions as well.



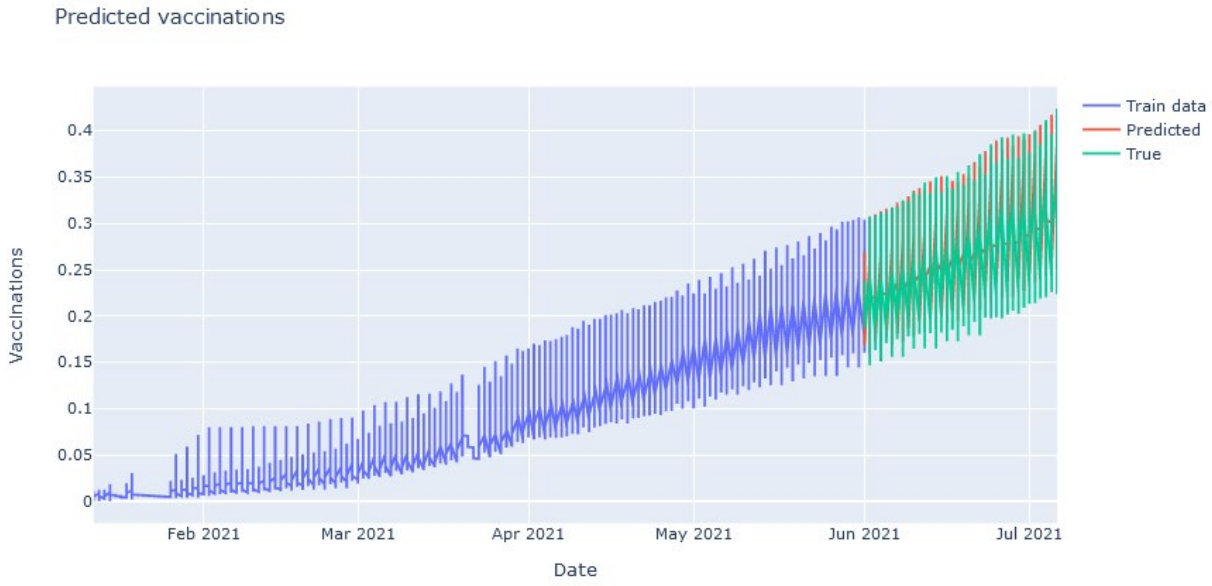
The Approach:

Prior to our attempts of actually producing a recursive neural network to model our timeseries data we had to figure out the best features to try to predict/forecast future vaccination rates, that may indicate potentially safe vacation states, for these RNN's the variables we have chosen are our dates, states, the percentage of the population that is currently vaccinated, daily vaccinations, people fully vaccinated, people fully vaccinated per 100, and the total state populations. We thought these features would work best for predicting the percentage population of the states vaccinated on dates in the future. One interesting thing to note about our final dataset, as shown above, is that our time series is stacked, for example, we have 176 total dates in the frame ranging from 1/12/2021 to 07/06/2021 with every single of the 176 dates having 53 instances per day.

The approach we decided to take to solve the problem of predicting vaccination rates by date/state really boiled down to only one clear cut option. LSTM (Long/Short term memory RNN's) are specifically equipped to deal with data in sequence like our dataset. There are two main forms of LSTM, univariat and multivariat lstms. Univariate being where we provide the datetime feature and then the Y variable and predict. In our approach we have used both approaches so we can say with relative efficcacy that this model is a valid option for us. Alternatively there is multivariat LSTM, where we feed our keras model with multiple features and predict upon our Y variable. In the instance of our model, we have chosen seemingly correlated data features that we thought would go along best for predicting the percentage of the state that is vaccinated by date. As an aside with the multivariat approach, it appears that using stacked LSTM, (multi-layer lstm), while including a dropout and dense layer were one of the most efficcient uses of the model we were able to find from our research. We tried all approaches for sequential time series analysis that we were able to find, as well as the previous standard linear ANNs.

The Results:

The first run using our keras LSTM was using the data by date/state and using the feature people fully vaccinated per hundred and our Y predictor as the percentage of the state vaccinated. We graphed our results as: Tthth



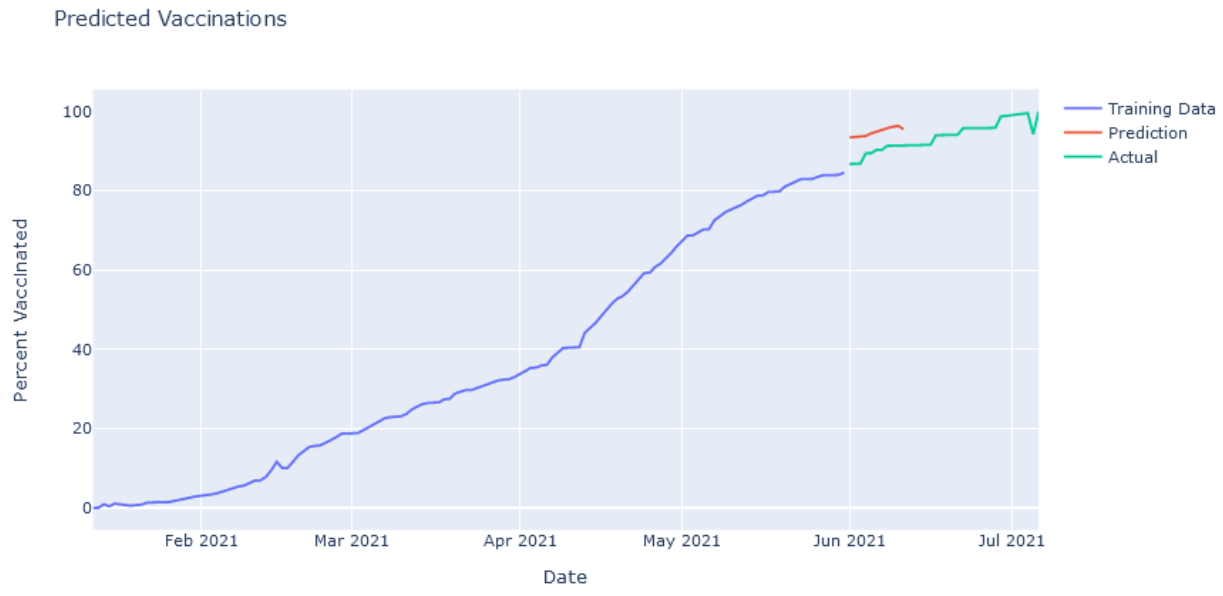
This data reveals a trend of elevating vaccinations as the dates move further in the year as one would expect, the jumps in the scatter plot line graph are due to the states on the individual dates, which made the expected output of the model startling. While the data on a graph reveals the data we would expect, the prediction was way off.

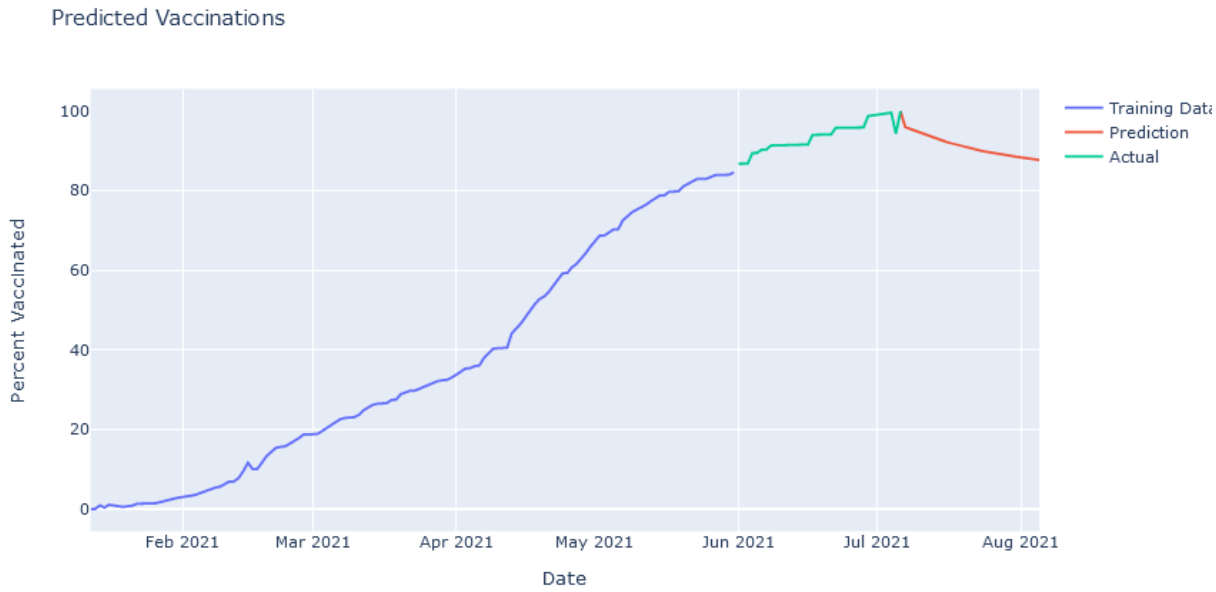
:

		Predicted_vaxxed
state	**date**	
Alabama	**2021-07-06**	35.360000
	2021-07-07	-296.822174
	2021-07-08	-262.870843
	2021-07-09	-126.019917
	2021-07-10	-67.154466
	2021-07-11	-163.587928
	2021-07-12	-38.359822
	2021-07-13	-81.671899
	2021-07-14	15.333788

Throughout all test runs of our lstms we experienced similar experiences where our output predicted frame was off, while the graph reflected the data as expected.

Predicting the percentage of the state vaccinated by date appeared to work better when we selected single states to predict on.





One thing to note, is a surprise why our predicted vaccinations always appeared to be starting with the highest value first and predicting downward.

		Predicted_vaxxed
state	**date**	
NaN	**2021-07-06**	36.980000
	2021-07-07	35.512228
	2021-07-08	35.336487
	2021-07-09	35.173998
	2021-07-10	35.011185
	2021-07-11	34.850540
	2021-07-12	34.693743
	2021-07-13	34.542024
	2021-07-14	34.395999
	2021-07-15	34.255865
Georgia	**2021-07-16**	34.121911
NaN	**2021-07-17**	33.994264
	2021-07-18	33.872734

The multivariate model appeared to look a bit better than the first many passes we took at it using the univariate model. Before scaling the data using this model, we passed a frame to the standard scaler that was using a stacked index.

		people_fully_vaccinated	daily_vaccinations	total_vaccinations	total_state_pop	%_pop_vaxxed
date	state					
2021-01-12	0.0	7270.0	13415.0	78134.0	4903185.0	0.15
	1.0	5400.0	13415.0	35838.0	731545.0	0.74
	2.0	8343.0	13415.0	141355.0	7278717.0	0.11
	3.0	8.0	13415.0	40879.0	3017804.0	0.00
	4.0	100089.0	13415.0	816301.0	39512223.0	0.25
...	...	...	...	...	...	...
2021-07-06	49.0	4478317.0	16147.0	9328521.0	8535519.0	52.47
	50.0	4206349.0	19116.0	8676836.0	7614893.0	55.24
	51.0	671593.0	1529.0	1426490.0	1792147.0	37.47
	52.0	2930845.0	11972.0	5982183.0	5822434.0	50.34
	53.0	204636.0	1350.0	431146.0	578759.0	35.36

9504 rows × 5 columns

For this model, the examples we found shown using a slice of data to predict from using past data was an effective method to predict a number of days in the future based on that data.

Using some data manipulation on the predicted frame for the multivariate model we were able to get the set looking semi decent, but we know it is still wrong.

	Date	State	Pred_vaxxed
0	2021-07-06	0.0	27.786766
1	2021-07-07	1.0	26.013676
2	2021-07-08	2.0	29.224445
3	2021-07-09	3.0	28.041592
4	2021-07-10	4.0	28.208984
...	...	...	...
3175	2021-08-30	53.0	22.241072
3176	2021-08-31	53.0	22.241072
3177	2021-09-01	53.0	22.241072
3178	2021-09-02	53.0	22.241072
3179	2021-09-03	53.0	22.241072

3180 rows × 3 columns

Final thoughts:

After weeks of mind wracking work trying to figure out why our model wasn't forecasting our data correctly, the one estimation we can make is that at least the way we were feeding the lstm

wasn't correct. We had tried everything we knew to try, but we never gave up, we kept going, and going, and going, and going. Our results seemed to improve after every attempt, but never quite ended up where we wanted. We still fully believe that the idea behind predicting future state vaccinations is an important one for a variety of reasons even beyond just simply "where might be a safe location to vacation?" We know it's absolutely possible, but it just seems like it is beyond our abilities at this time.