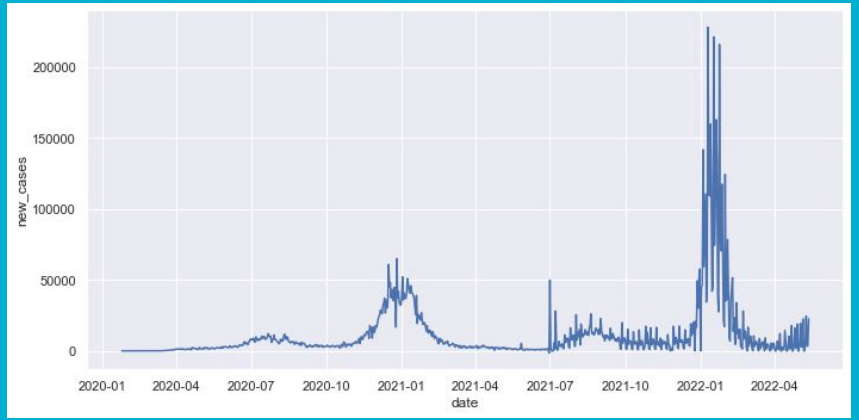
Covid-19 New Case Prediction in CA



Springboard Capstone#2 Presentation

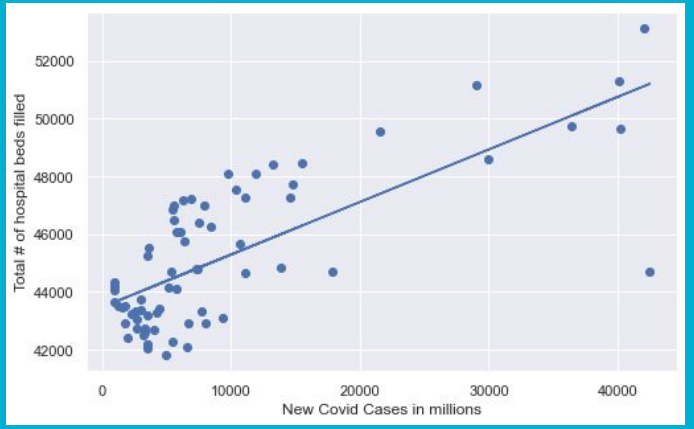
Problem Statement

- Covid-19 has affected society in many ways, but one that I have a personal connection to is the impact it has had on hospitals and hospital staff.
- Many hospitals have at some point during the pandemic had to put a freeze on all non essential operations
- Hospital staff burnout is at an all time high. Dozens of academic studies can be found with a quick google search confirming this.
- The goal of this project is to use time series to predict future covid cases in the state of CA, in order to give hospitals a chance to prepare for changes in patient demand due to covid.

Justification

The value of this project of this project for hospitals is dependent on covid infections actually having an impact hospital occupancy. To confirm this I ran a regression of the impact of new covid cases on total hospital occupancy in the state of California.

- The results were significant at 99% confidence
- Coefficients indicate that on average for roughly every 5.5 new covid cases an additional hospital bed is filled



Data

- New York Times Covid-19 Infection Data: https://github.com/nytimes/covid-19-data

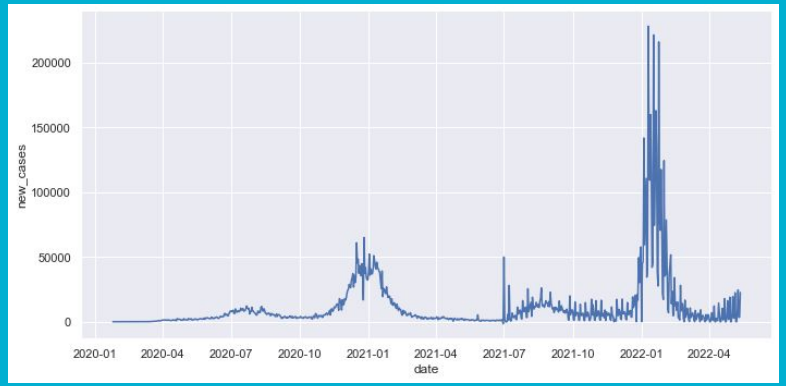
- HealthData.gov Hospital Data: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u

Data Wrangling

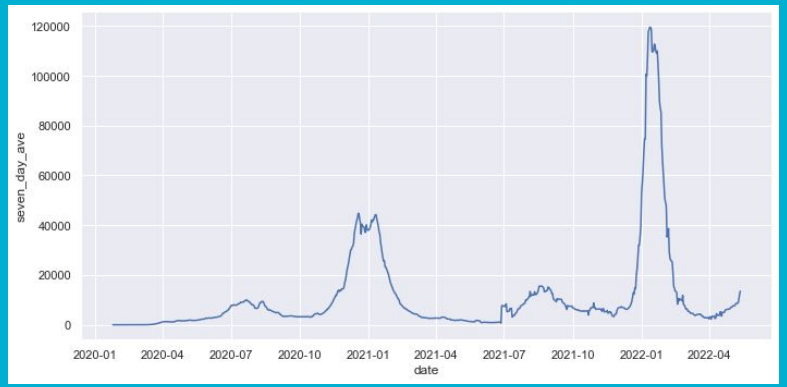
Process:

- Remove missing values
- Create features for EDA and modeling
- Group and aggregate county data at state level
- Resample covid data for comparison with hospital data
- Smooth covid data for modeling

Raw Covid Data

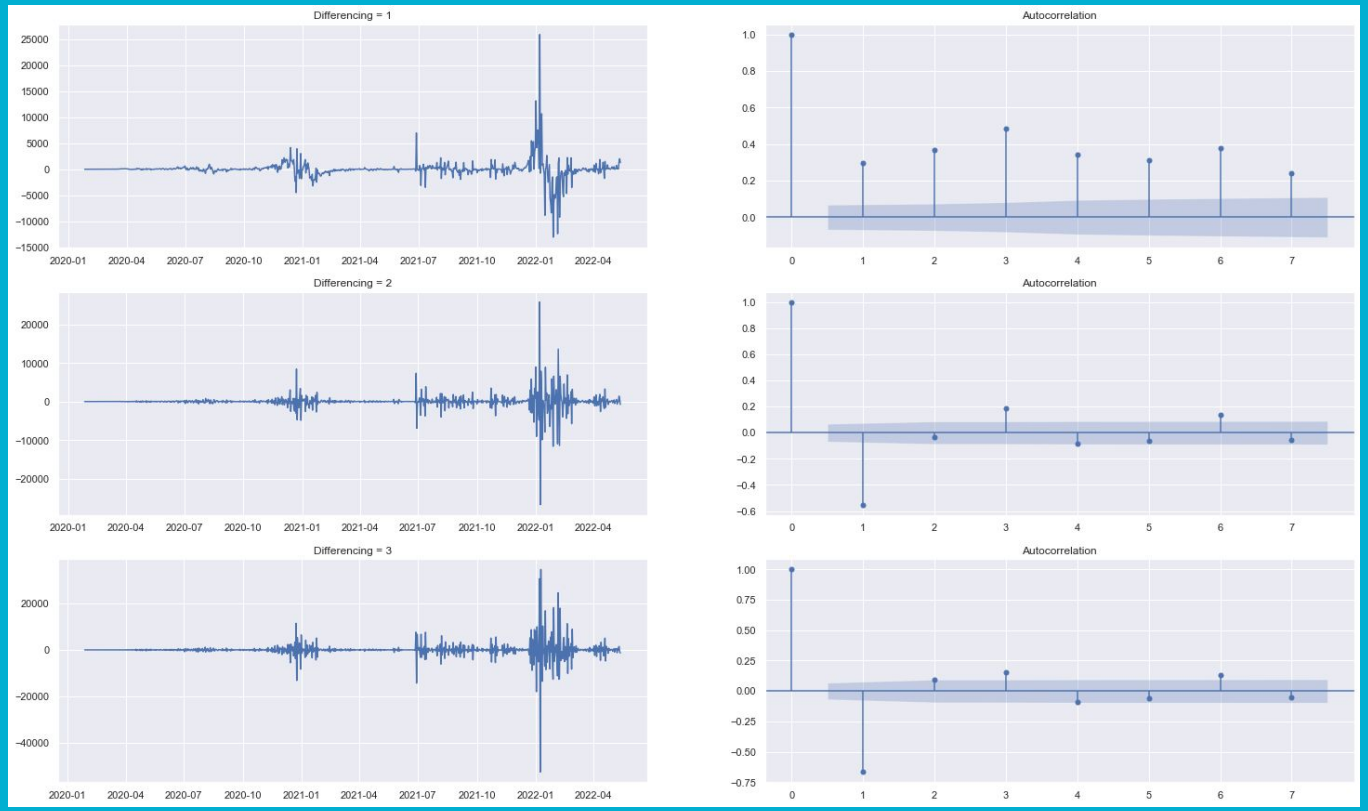


Smoothed Covid Data



EDA:

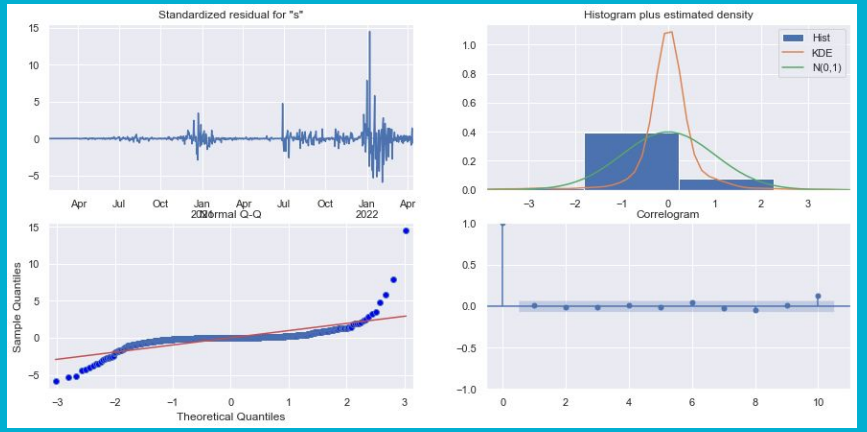
- Primary focus was to make data stationary for ARIMA model
- Performed Dicky Fuller test and examined plots of differenced data and ACF



ARIMA Model

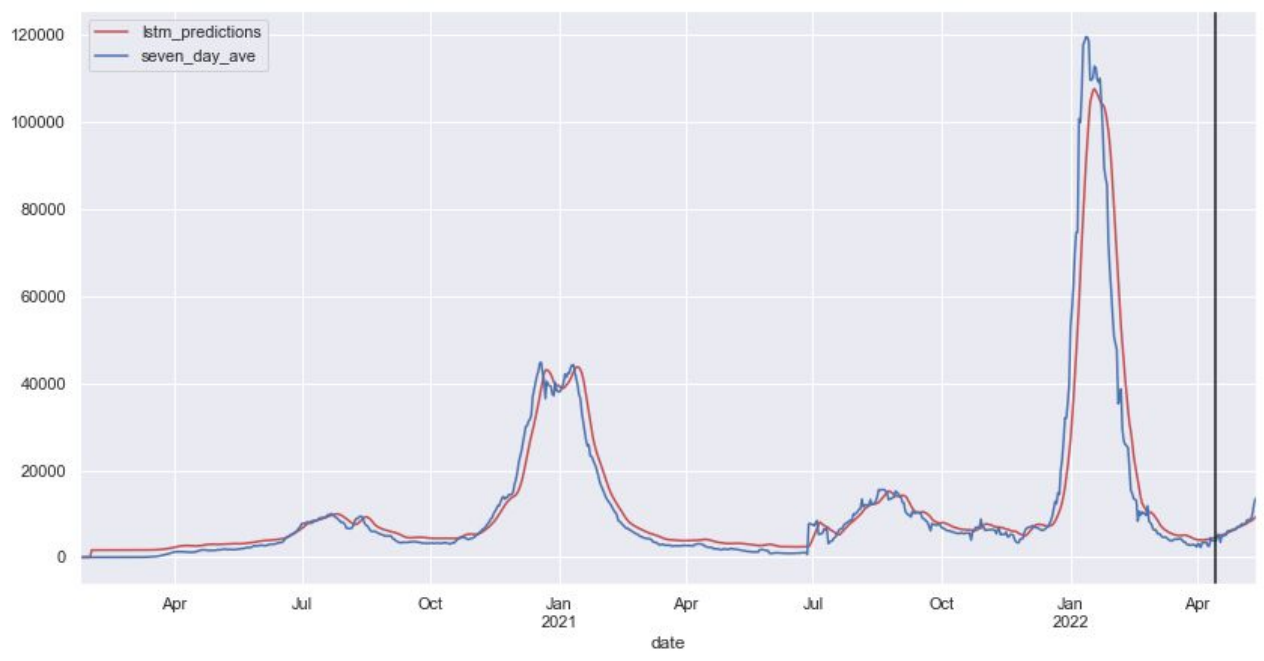
- Combination of AR and MA models
- Initially selected AR and MA terms using ACF and PACF visualizations
- Second attempt was with an auto ARIMA grid search
- Both models suffered from over fitting, and did not have normally distributed residuals.

$$Y_t = (\beta_1 + \beta_2) + (\Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p}) + (\omega_1 \varepsilon_{t-1} + \ldots + \omega_q \varepsilon_{t-q} + \varepsilon_t)$$



LSTM Model

- Final model was an LSTM recurrent neural network.
- Defined a function that takes input data and a list of parameters and uses them to create and score a model.
- Ran this function through a loop to perform hyper parameter tuning.



Model Evaluation

The best model by far was the LSTM model, it scores higher than either ARIMA model in all evaluation metrics related to the testing data. Most importantly, it is capable of reliably predicting covid infections up to one month in the future.

The LSTM model does a much better job generalizing the data, unlike the ARIMA models it did not overfit to the training set.

Model	MAPE	MRSE	R-Squared
ARIMA1	0.2964	3135.65	-0.9433
ARIMA2	0.3673	2885.83	-0.646
LSTM	0.0885	1126.78	0.749

Future Work and Other Applications

- Directly linking Covid Data to hospitalizations

- Applying findings to other industries

  1.) Retail

  2.) Transportation

  3.) Home entertainment

Questions?