Covid-19-Time-Series-Modeling

Public

main 1 Branch 0 Tags Go to file t Go to file Add file Code

Bella3s spelling er... 6a0c864 · 1 minute ago 26 Commits

images	add files	3 days ago
.gitignore	Initial commit	3 weeks ago
README.md	updated recommendat...	11 hours ago
index.ipynb	spelling error	1 minute ago
owid-covid-data...	add downloaded data	3 weeks ago
world_country_a...	add files	3 days ago

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

README

Covid-19-Time-Series-Modeling

Abstract

This project creates a model that forecasts the number of deaths by Covid-19 given the historical data. The Covid-19 pandemic was detrimental in the number of people who perished from this disease. The idea behind this project is that if we can create a predictive model that will forecast the amount of harm, then public health officials can be better informed, advised, and empowered to mitigate future deaths. Specifically, this project focuses on a continent-wide evaluation, looking at Asia.

This project goes through a model iteration process, starting with a naive time series model, and iterating through ARIMA, multivariate ARIMA, Linear Regression, Facebook's Prophet, and multivariate prophet models. The project finalizes on the multivariate prophet model as the final model based on the evaluation metric, root mean squared error, and evaluates this model on the holdout test set. Lastly, the project offers a few recommendations to the proposed business and next steps for future projects.

Repository Structure

No packages published
Publish your first package

Languages

Jupyter Notebook 100.0%

```
├── images
    (https://github.com/Bella3s/Covid-19-Time-Series-Modeling/tree/main/images)

├── pdfs
    (https://github.com/Bella3s/Covid-19-Time-Series-Modeling/tree/main/pdfs)

├── README.md
    (https://github.com/Bella3s/Covid-19-Time-Series-Modeling/blob/main/README.md)

├── index.ipynb
    (https://github.com/Bella3s/Covid-19-Time-Series-
Modeling/blob/main/index.ipynb)

├── owid-covid-data.csv
    (https://github.com/Bella3s/Covid-19-Time-Series-Modeling/blob/main/owid-
covid-data.csv)

├── world_country_and_usa_states_latitude_and_longitude_values.csv
    (https://github.com/Bella3s/Covid-19-Time-Series-
Modeling/blob/main/world_country_and_usa_states_latitude_and_longitude_values.csv)
```

The non-technical presentation can be found under [pdfs/presentation]

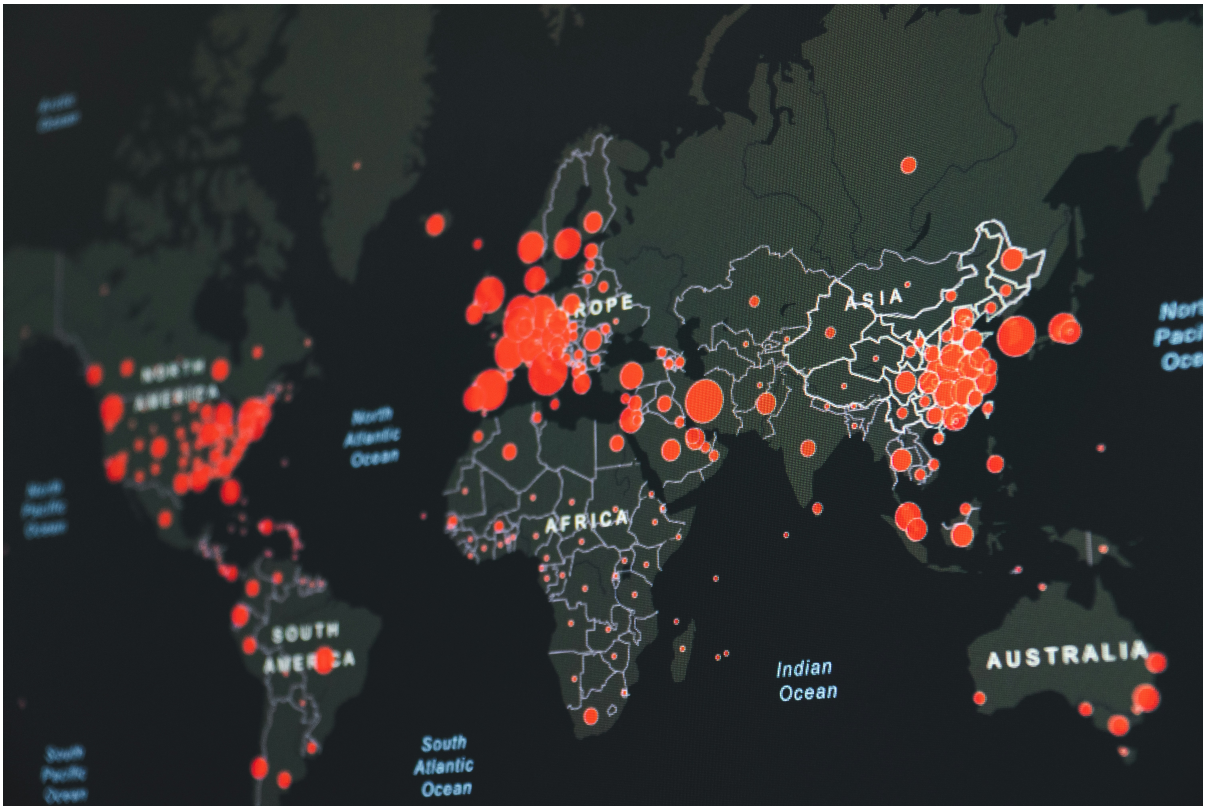
Reproducing via Google CoLabs

If reproducing this project via Google CoLabs, there are certain extra steps that need to be taken to run the code properly. Firstly, the pmdarima and prophet libraries need to be installed. Secondly, the environment needs to be set up to download the datasets from Kaggle. **Note that this step requires you to have a Kaggle username and API key.** Lastly, the datasets need to be downloaded and unzipped prior to use. This code is in the notebook to accomplish this that just needs to be uncommented, or see below for the snippet of just the extra code that needs to be run (all the imports from the notebook will need to be run as well for the project to run properly).

```
# For Google CoLab might need to install pmdarima and Prophet
!pip install pmdarima
!pip install Prophet

# Prep Google CoLab environment to download data from Kaggle
!mkdir ~/.kaggle
!touch ~/.kaggle/kaggle.json

username = ''  ## Your Kaggle username
api_key = ''  ## Your Kaggle API key

api_token = {"username": username,
             "key": api_key}

with open('/root/.kaggle/kaggle.json', 'w') as file:
    json.dump(api_token, file)

!chmod 600 ~/.kaggle/kaggle.json

# Download the dataset from Kaggle
!kaggle datasets download -d caesarmario/our-world-in-data-covid19-dataset
!kaggle datasets download -d paultimothymooney/latitude-and-longitude-for-every-
```

```
country-and-state

# Unzip the downloaded data
shutil.unpack_archive('our-world-in-data-covid19-dataset.zip', '/content')
shutil.unpack_archive('latitude-and-longitude-for-every-country-and-state.zip',
'/content')
```

The Business + Business Problem



The Covid-19 pandemic needs little introduction as it left no corner of our world untouched. It devastated lives, and economies -- day-to-day life was dramatically altered for about 2-3 years. The specific business for this project is one akin to the World Health Organization (WHO) that 'treats data as a public good'(WHO Principles, but also wants to leverage this data to make recommendations to public health officials. The idea for this project is to create a model that will forecast how much worse the pandemic will be in order to advise public health officials and policy makers on how to best handle the situation such that damages are minimized.

The Data Source

The Data from this project is sourced directly from Kaggle, a dataset called Our World in Data - COVID-19. The company, Our World in Data combined their own data along with data from the John Hopkins University and the WHO. Further details on their data and sources can be found on the GitHub page here.

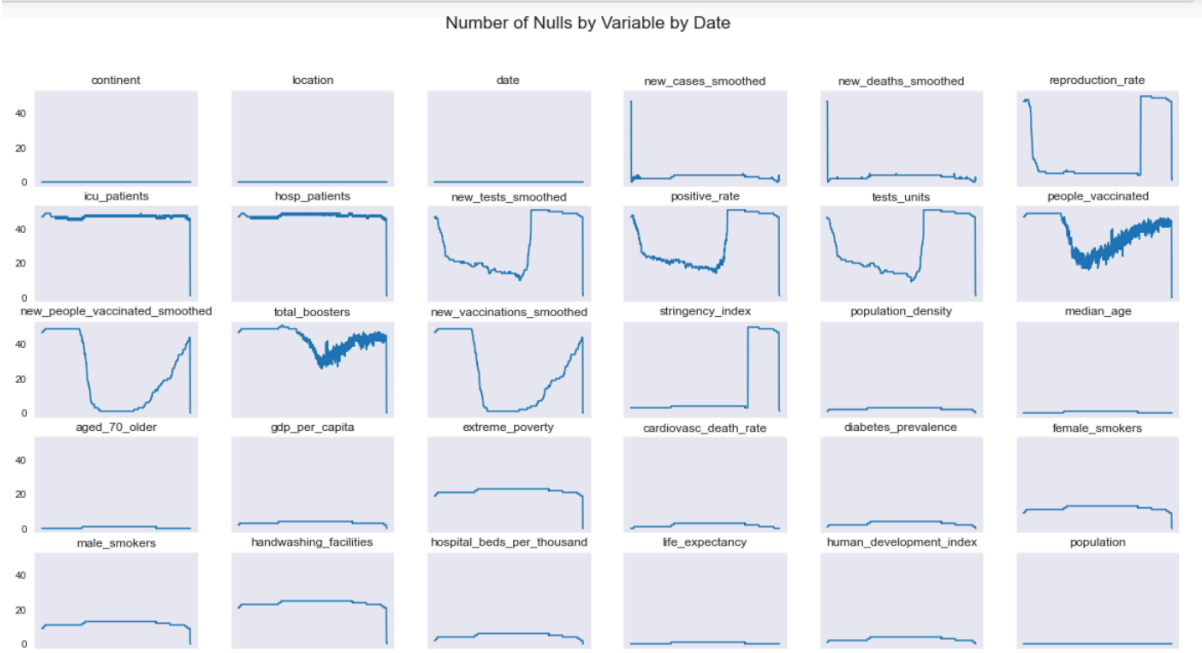
For visual purposes, the project also utilizes a dataset of latitude and longitude data of countries from Kaggle, originally sourced from google. See that data set here.

Data Exploration

In the initial data exploration, many of the variables are dropped as they are redundant for our purposes. The project looks at a visual of where the data is sourced from which leads to the decision to group the data by continent. This project focuses solely on the model created from the data in Asia.

Data Cleaning

There are numerous missing values in the complete dataset, as well as the Asia subset. The project looks at each of these and handles the missing values for each variable. This is done prior to the data being aggregated to one entry per day for the whole continent, but also while keeping mind to separate the train, validate, and test groups to avoid data leakage.

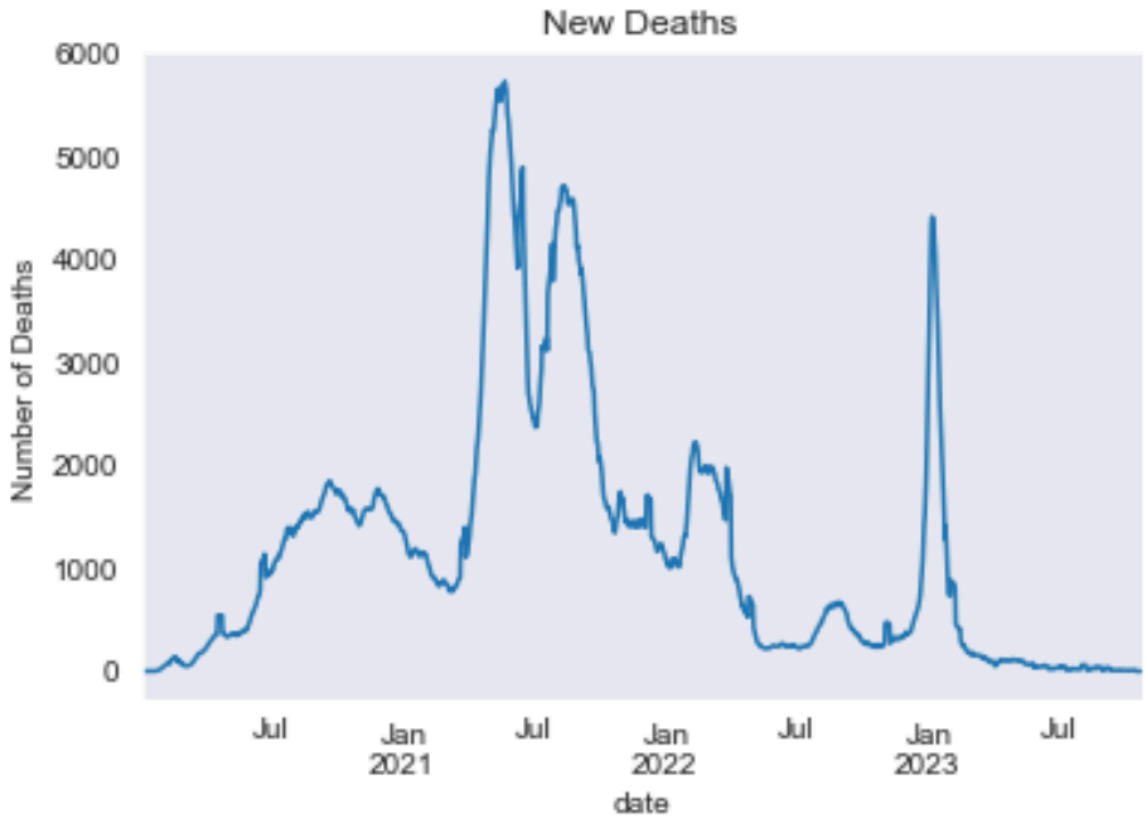


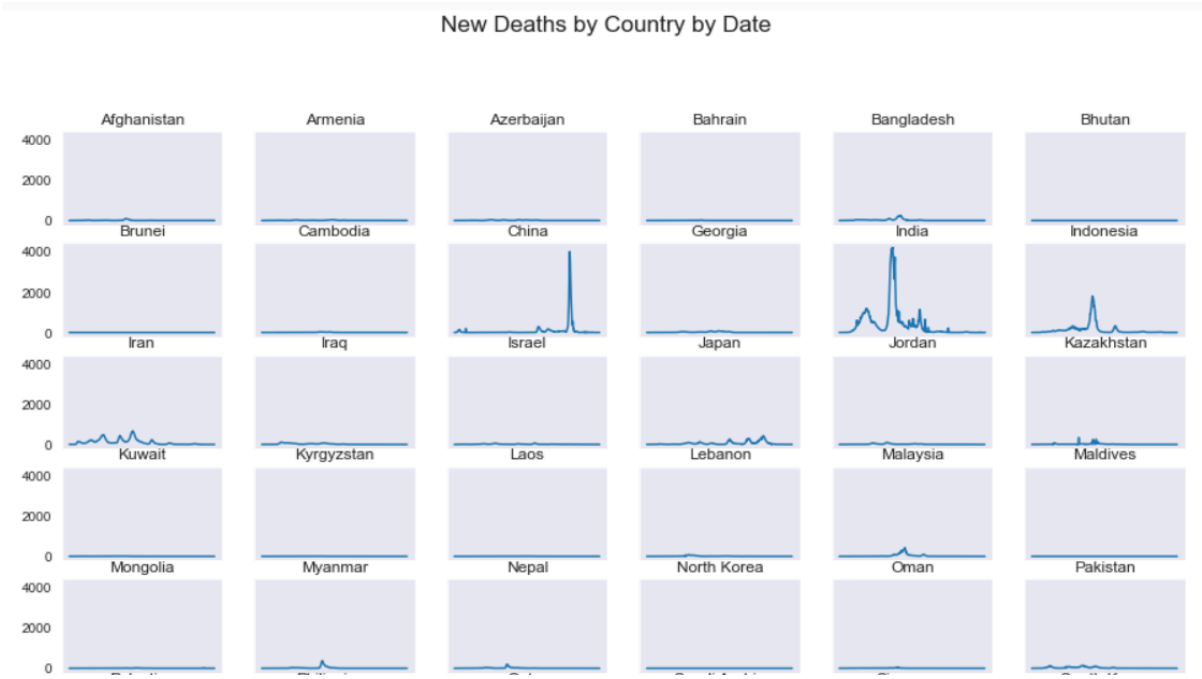
Aggregation + Final Exploration

After the data is aggregated (see table below), the project displays some visuals of the final dataset for a more thorough understanding of the data being used.

Variable	How to aggregate
`continent`	drop
`location`	drop
`date`	unique value per day
`new_cases_smoothed`	sum per day
`new_deaths_smoothed`	sum per day
`people_vaccinated`	sum per day (cumulative value)
`new_vaccinations_smoothed`	sum per day
`stringency_index`	average per day

Below is the target variable, the number of deaths per day, shown as a total aggregated value for the continent as well as by country.





Model Iteration

The models explored in this project include:

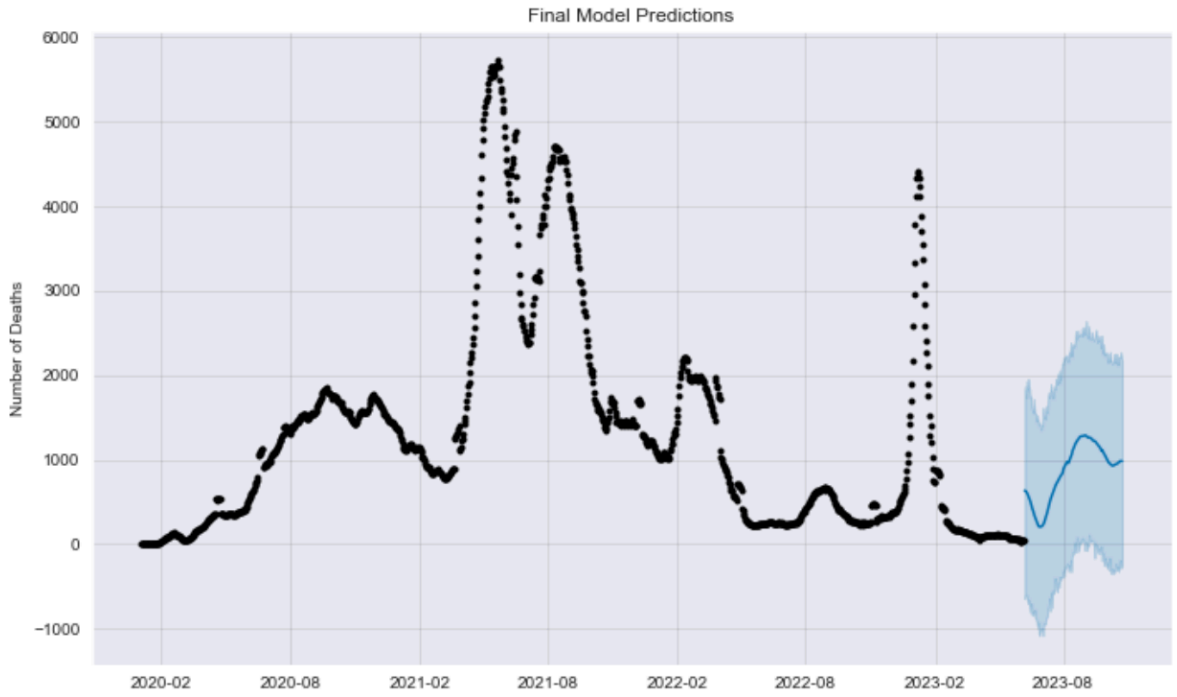
1. Naive Model
    - Our baseline model. The naive method uses 'today's' value as the prediction for 'tomorrow.'
2. ARIMA
    - Utilizing auto_arima to retrieve optimal p,d,q values
3. Multivariate ARIMA
    - Again utilizing auto_arima, but also including the other variables in our dataset, not just the target variable.
4. Linear Regression
5. Prophet
6. Multivariate Prophet

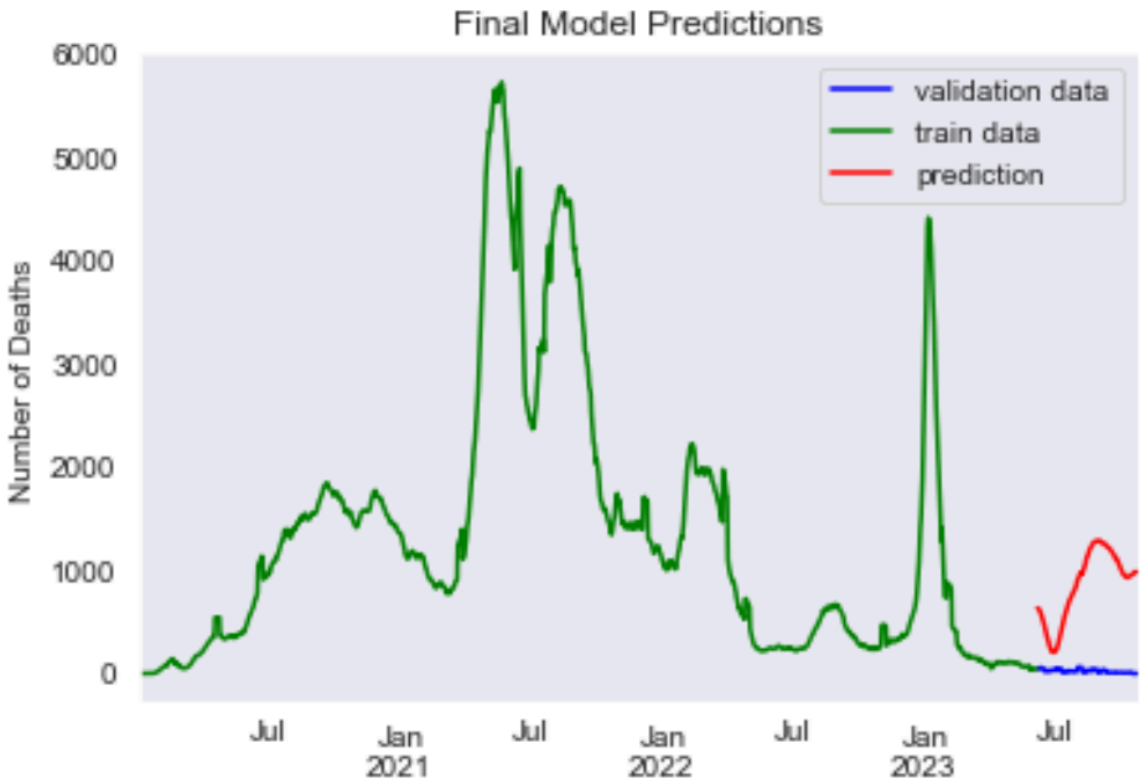
All Models were compared by their root mean squared error and mean absolute error from the predictions on the validation set.

| Model | RMSE | MAE |:--|:--| | Naive | 2,343 | 2,305 | | ARIMA | 1,944 | 1,897 | | Multivariate ARIMA | 968 | 931 | | Linear Regression | 809 | 731 | | Prophet | 2,164 | 1,972 | | Multivariate Prophet | 904 | 760 |

The Multivariate Prophet model was chosen as the final model even though the Linear Regression model had a lower RMSE because the Linear Regression model was predicting negative numbers for the number of deaths, the scores for these two models were quite close, and the prophet module is more suited for forecasting and analyzing trends in time series data.

Final Model Evaluation





Our model performed better on the validation data than it did on the test data. In part, this is because the pandemic had more or less ended by this time (mid 2023), which means the actual values for our data lie very close to zero, whereas our model is predicting the deaths to increase once again (similar to the actuals of July in 2020).

With an ending root mean squared error of 904 and mean absolute error of only 836, I would say this is a fairly good model, taking into account that this estimation is for the whole continent of Asia rather than one country alone.

Conclusion + Recommendations

Our model performed better on the validation data than it did on the test data. In part, this is because the pandemic had more or less ended by this time (mid 2023), which means the actual values for our data lie very close to zero, whereas our model is predicting the deaths to increase once again (similar to the actuals of July in 2020).

With an ending root mean squared error of 817,900 I would say this is a fairly good model, taking into account that this estimation is for the whole continent of Asia rather than one country alone.

1. **Utilize the model to advise public health officials.** Once we have forecasting the amount of harm predicted, we can then advise public health officials on what actions to take concerning public health measures such as mask mandates, social distancing, and stay-at-home orders (either to make more strict or lessen).

2. **Utilize the model to aid in resource planning.** From our forecasts of when spikes will occur, we can make recommendations to vendors and hospitals concerning resource planning.

3. **Investigate why deaths tend to be reported higher on Fridays.** From our final model, we can see a weekly trend where reported number of deaths is generally higher on Fridays by about 4 people, and lower earlier in the week by about 3 people. While a slight trend, it would be interesting to investigate why this is; if it is anything in the hospitals systems or public health trends that can be addressed in such a way that number of deaths lessens.

Next Steps

There is always more to do and try! Below are some ideas for expanding on this project:

- Continue to make data more accurate and complete
    - The missing, purposefully mis-reported, and unusable data all made our model less accurate than it could have been. Looking for ways in the future to gain accurate,