Stage-III (Basic Machine Learning)

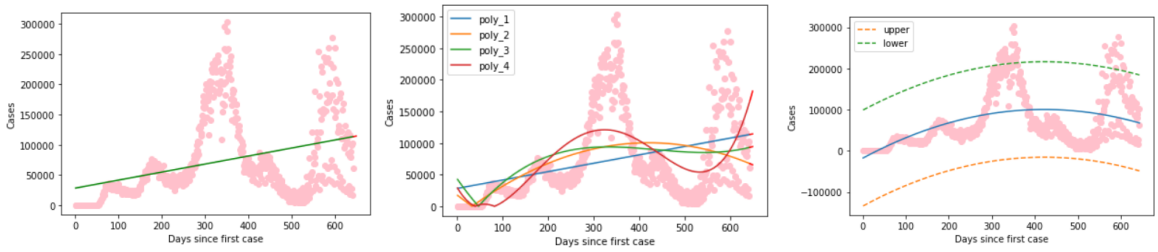
Team Task:

Regression Trends Description

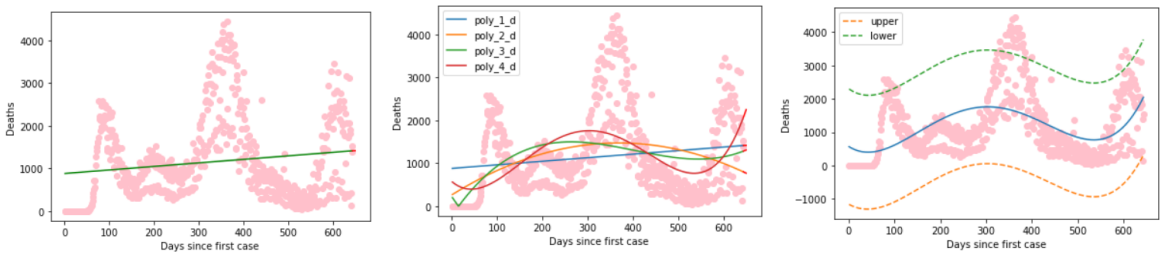
For the regression line trends compared to other countries, we are mostly seeing increases in cases. For the US, predicted new cases are around 114000 while predicted new deaths are around 1400. Indonesia's new cases are predicted to be around 15000 while deaths are predicted to be around 600. Pakistan's predicted new cases are around 3000 while predicted deaths are around 70 which is very low compared to most of the other countries. Brazil's predicted cases are around 51000 while its deaths are around 1400. Bangladesh's predicted new cases are around 4600 while it's predicted new deaths are around 93. Nigeria comes in last place for predicted new cases and deaths with a prediction of 400 new predicted cases and around 6 for new deaths.

When it comes to predicted new cases, Nigeria has the smallest prediction of around 400 while the United States has the highest predicted cases of around 114000 with Brazil in second, Indonesia in third, Bangladesh in fourth, and Pakistan with the fifth highest new cases. For predicted new deaths, the US and Brazil have a similar amount of deaths both being around 1400, then Indonesia would be second highest, Bangladesh is third highest, Pakistan fourth highest and Nigeria in last once again.

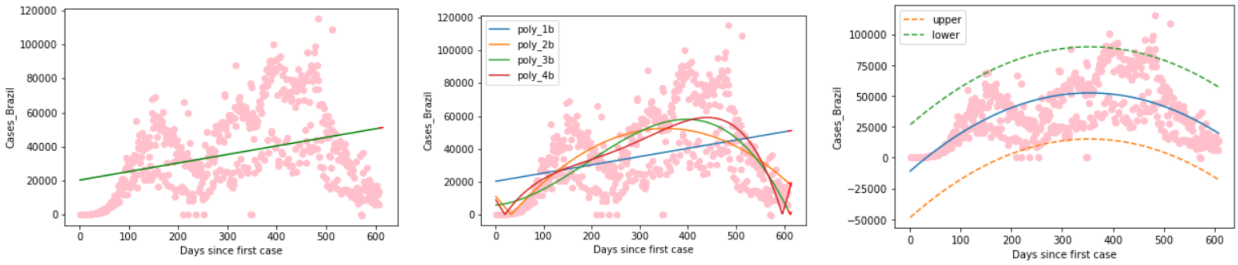
Linear-model, Non-linear model and confidence-interval for US cases with trend lines:



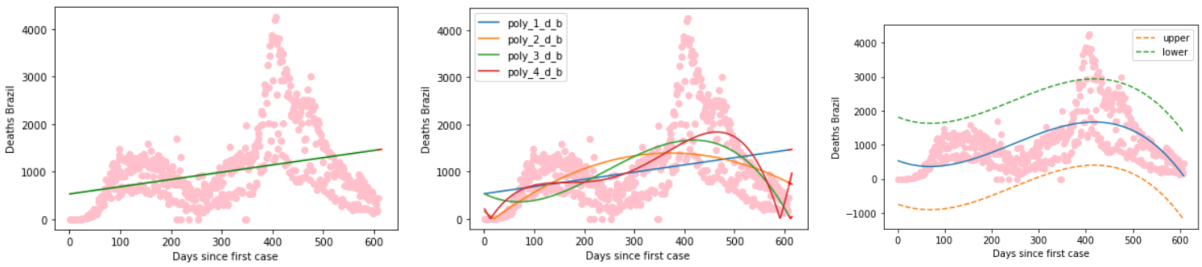
Linear-model, Non-linear model and confidence-interval for US deaths with trend lines:



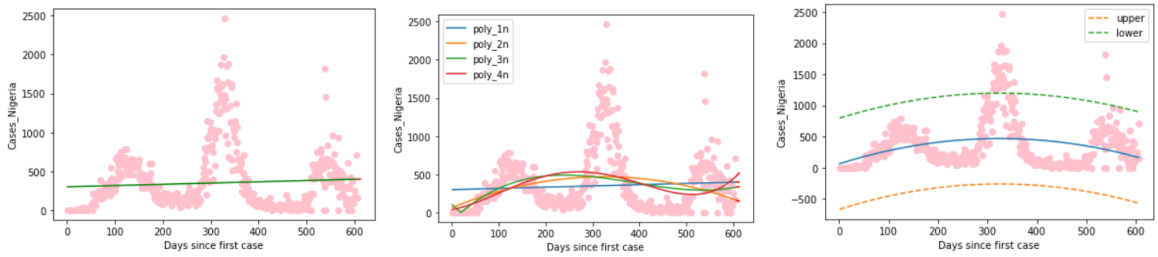
Linear-model, Non-linear model and confidence-interval for Brazil cases with trend lines:



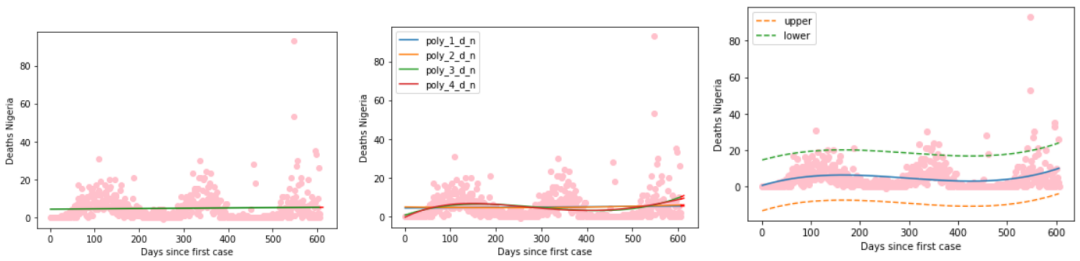
Linear-model, Non-linear model and confidence-interval for Brazil deaths with trend lines:



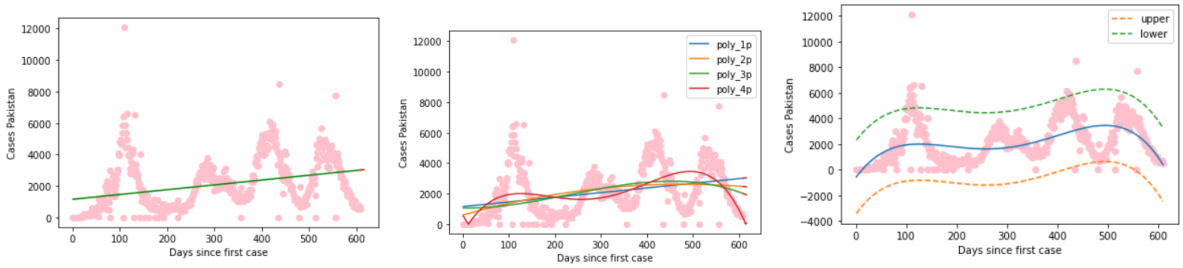
Linear-model, Non-linear model and confidence-interval for Nigeria cases with trend lines:



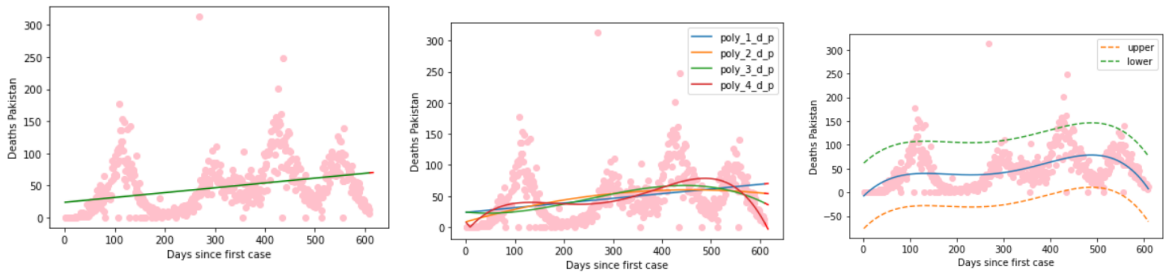
Linear-model, Non-linear model and confidence-interval for Nigeria deaths with trend lines:



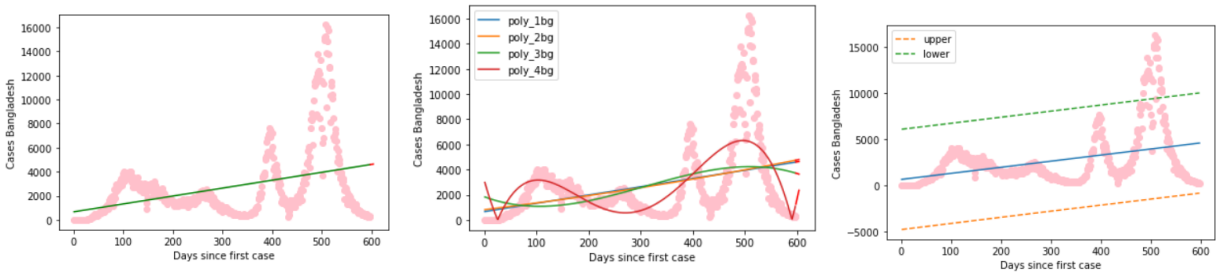
Linear-model, Non-linear model and confidence-interval for Pakistan cases with trend lines:



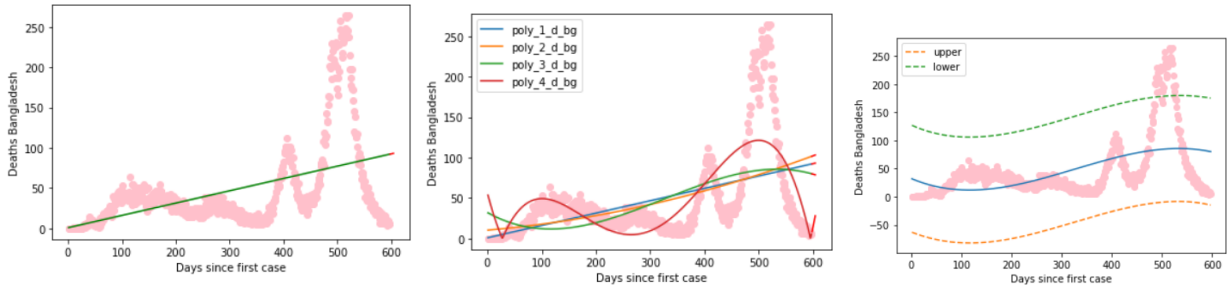
Linear-model, Non-linear model and confidence-interval for Pakistan deaths with trend lines:



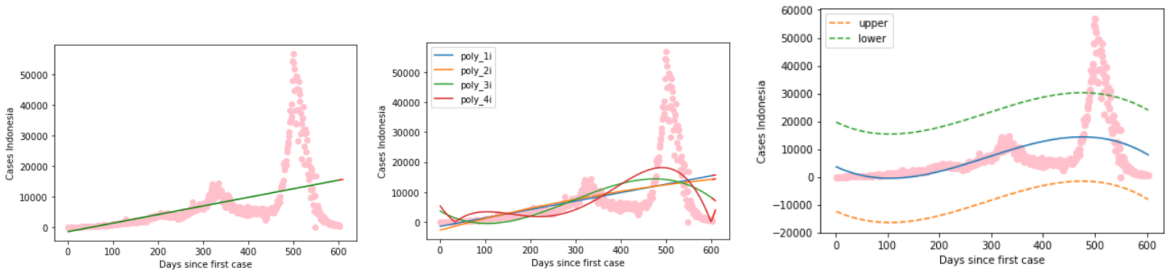
Linear-model, Non-linear model and confidence-interval for Bangladesh cases with trend lines:



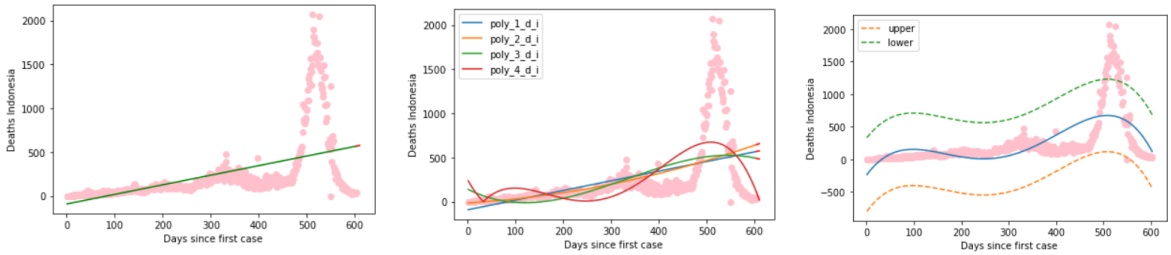
Linear-model, Non-linear model and confidence-interval for Bangladesh deaths with trend lines:



Linear-model, Non-linear model and confidence-interval for Indonesia cases with trend lines:



Linear-model, Non-linear model and confidence-interval for Bangladesh deaths with trend lines:

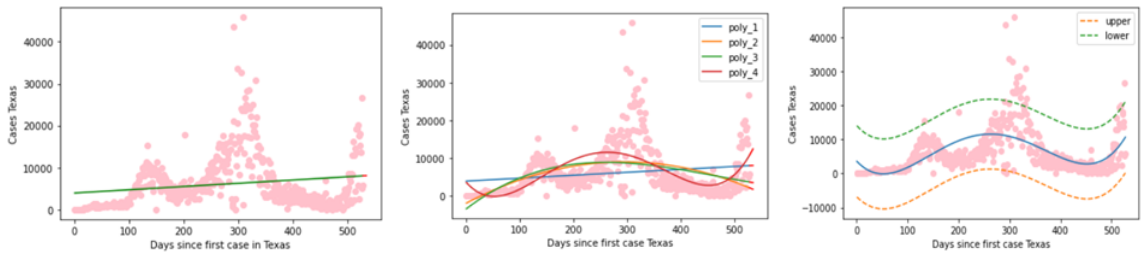


Member Task:

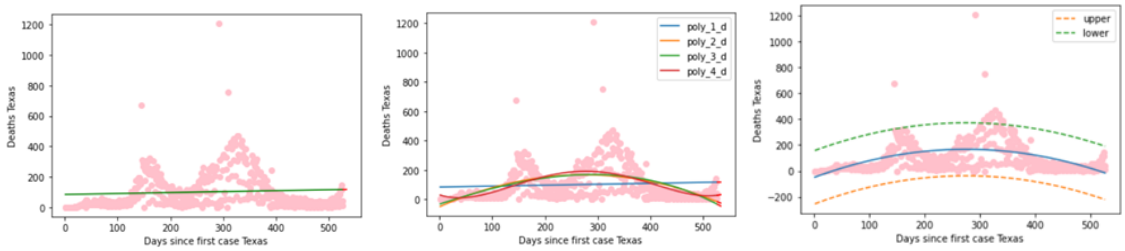
Aka'sh:
Task-1: Analyzed Pennsylvania
Task-2:

Chandan:
I have taken Texas as my state of choice and calculated the daily new cases and deaths similar to stage 2. I have statsmodelsformula.api for generation of linear regression models and non-linear regression models.
After performing required filters like to select data from the first day of cases in the states and removed the first row which is the start date cases.
Calculated number of days since first case was seen.
Trained the model providing new cases and number of days since the first case and predicted the next week's cases and deaths respectively and plotted their trend lines as below for both linear and non-linear regression .

Linear-model, Non-linear model and confidence-interval for Texas cases with trend lines:



Linear-model, Non-linear model and confidence-interval for Texas deaths with trend lines:



Now for the state of texas using hospital available bed data I have checked the point of no return.After training the model with available data and predicted number of deaths for next week the values by linear model are as below.

```
pred_next_deaths
```

```
0    117.185432
1    117.247955
2    117.310478
3    117.373001
4    117.435524
5    117.498047
6    117.560571
```

The number of total staffed adult ICU beds is 6857.So I assume we won't reach the point of no return in no time soon.

Similarly, we have done regression models both linear and non-linear. Plotted the trend lines and confidence intervals.

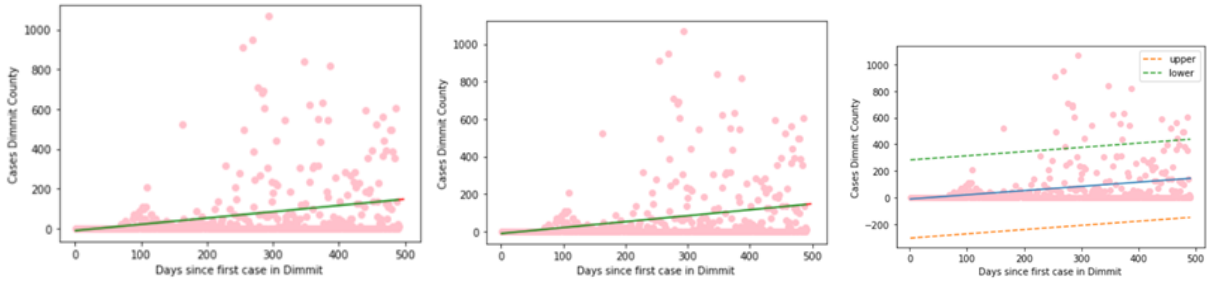
I have done this for 5 counties in the state of Texas.

1. Dimmit County
2. Crockett County
3. Concho County
4. Hansford County
5. Floyd County

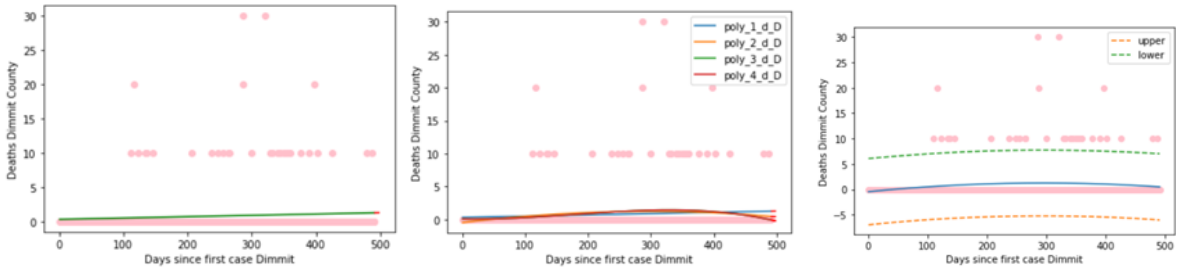
These counties have less populations but the number of cases in these counties is high. Normalized the cases and deaths to 100000.

Dimmit County

Linear-model, Non-linear model and confidence-interval for Dimmit County cases with trend lines:

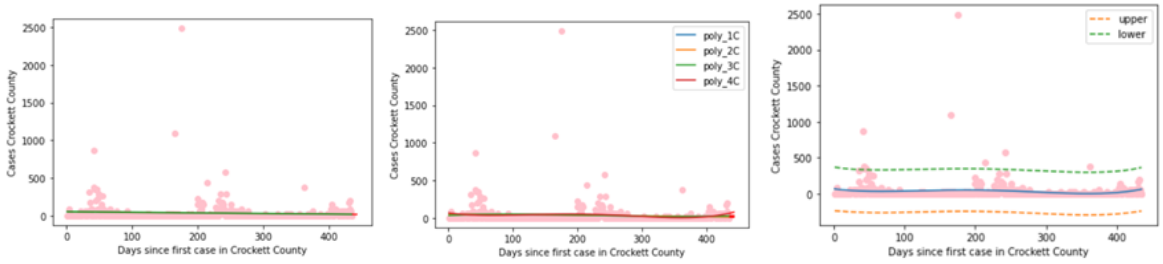


Linear-model, Non-linear model and confidence-interval for Dimmit County deaths with trend lines:

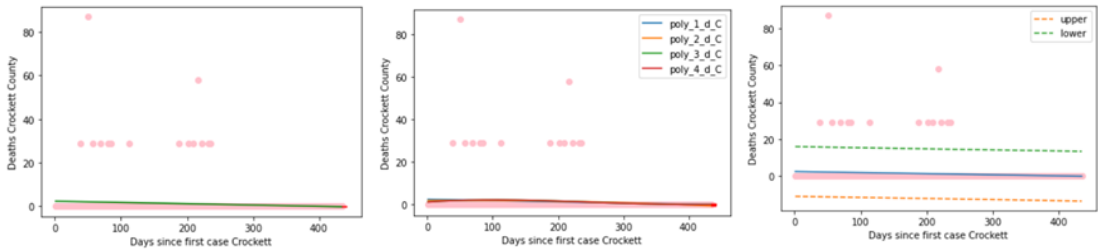


Crockett County

Linear-model, Non-linear model and confidence-interval for Crockett County cases with trend lines:

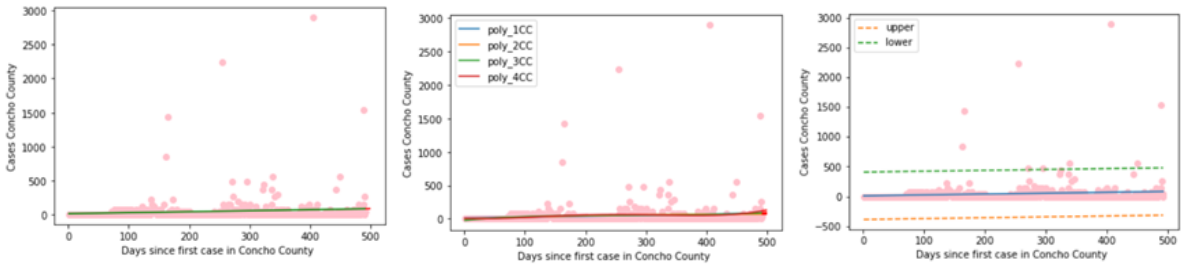


Linear-model, Non-linear model and confidence-interval for Crockett County deaths with trend lines:

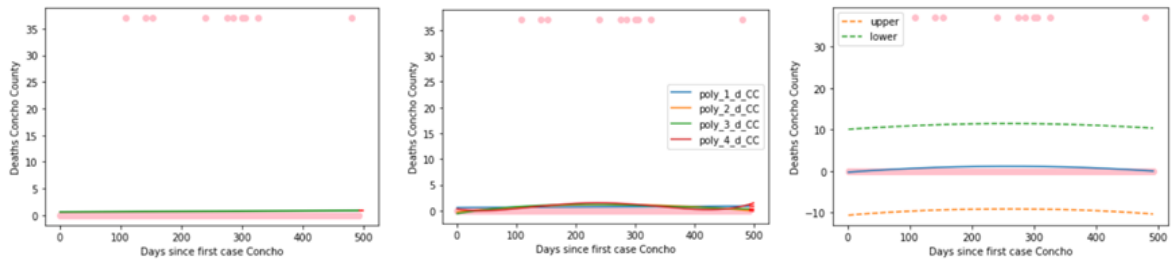


Concho County

Linear-model, Non-linear model and confidence-interval for Concho County cases with trend lines:

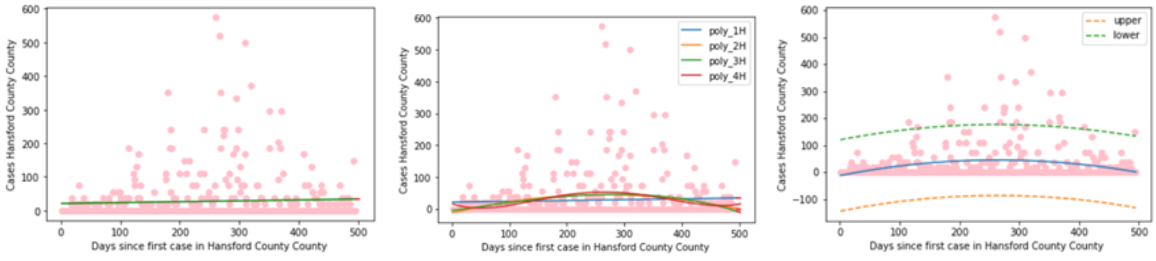


Linear-model, Non-linear model and confidence-interval for Concho County deaths with trend lines:

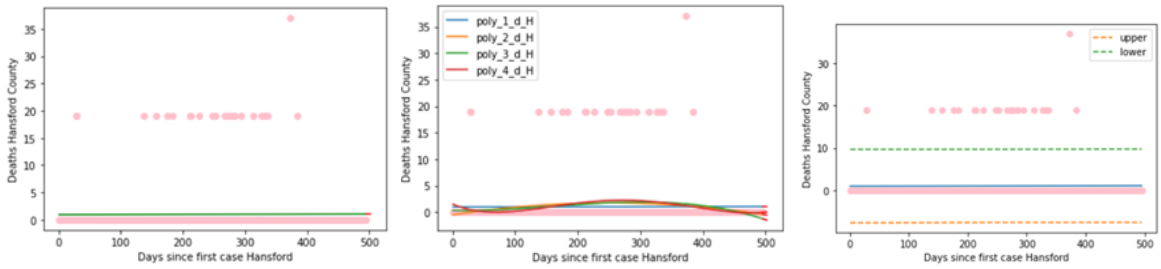


Hansford County

Linear-model, Non-linear model and confidence-interval for Hansford County cases with trend lines:

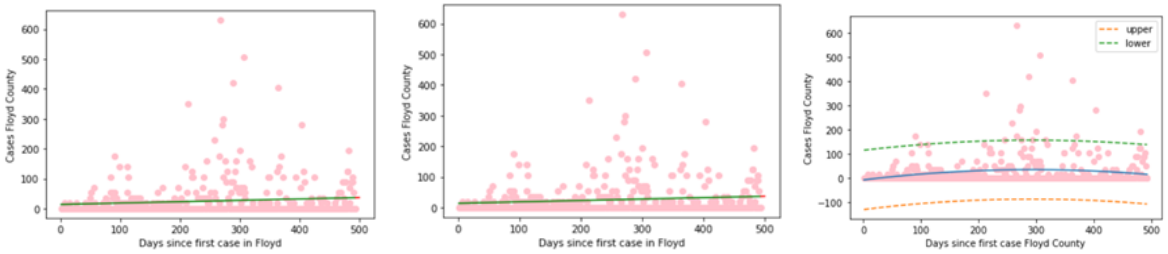


Linear-model, Non-linear model and confidence-interval for Floyd County deaths with trend lines:

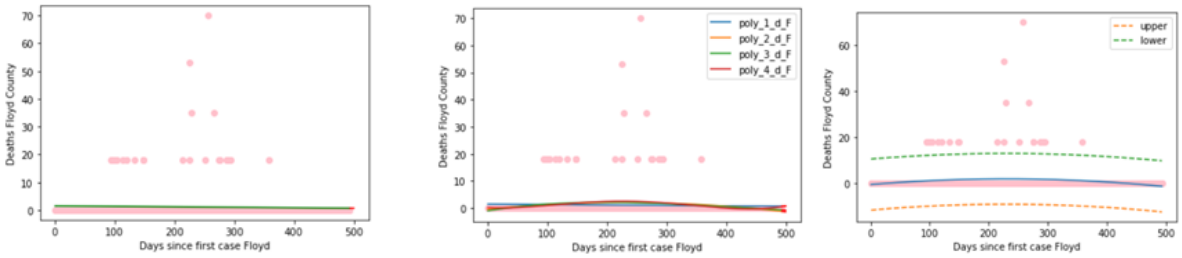


Floyd County

Linear-model, Non-linear model and confidence-interval for Floyd County cases with trend lines:



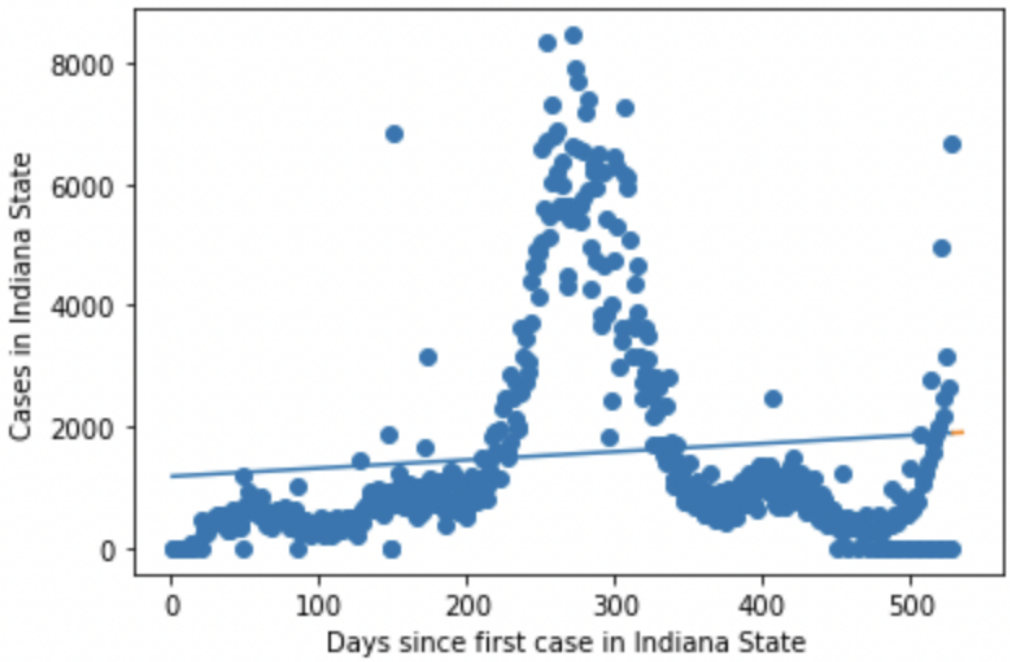
Linear-model, Non-linear model and confidence-interval for Floyd County deaths with trend lines:



Chandra Shekhar:

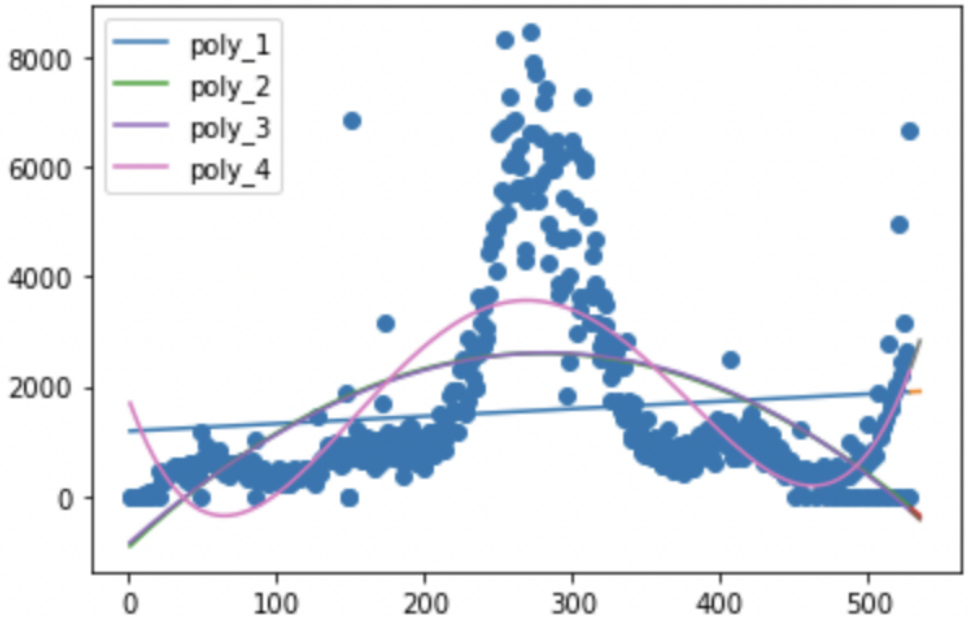
Task-1:

I have selected Indiana State as my preferred state of choice to predict the cases and deaths of a state. I have used linear modelling to fit the prediction model of the available cases and deaths. Below is the plot of cases per day from the first cases reported in the state and the linear regression model.

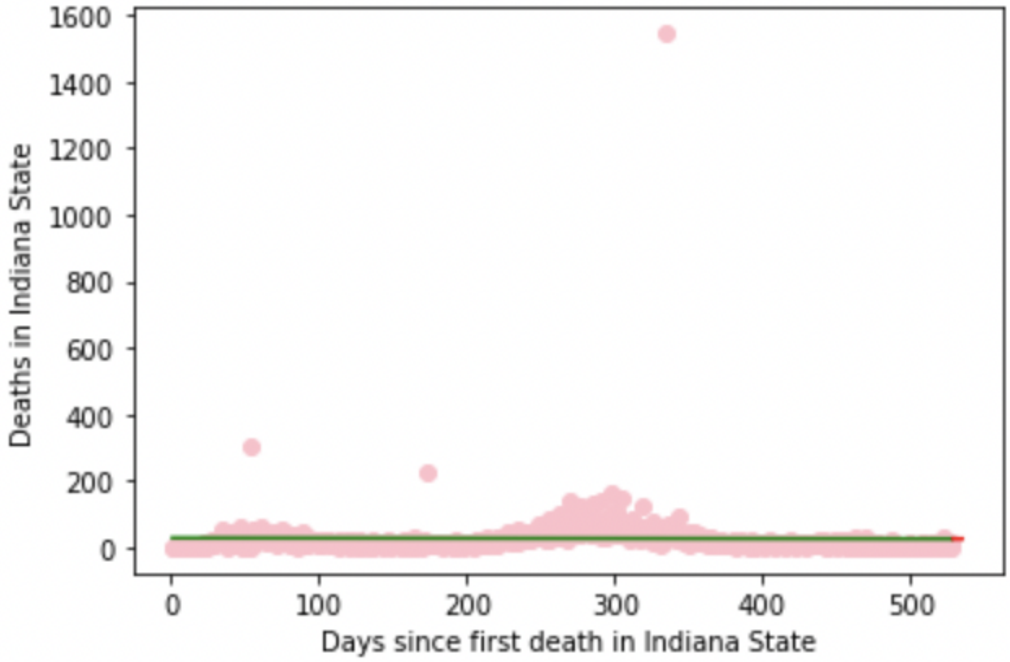


The root mean square error values for polynomial regression are:
RMSE of ploy1 is: 1798.553419382575
RMSE of ploy2 is: 1532.7366842730946
RMSE of ploy3 is: 1532.5479489199333
RMSE of ploy4 is: 1267.6260888954878

The prediction of cases of next week of Indiana State.

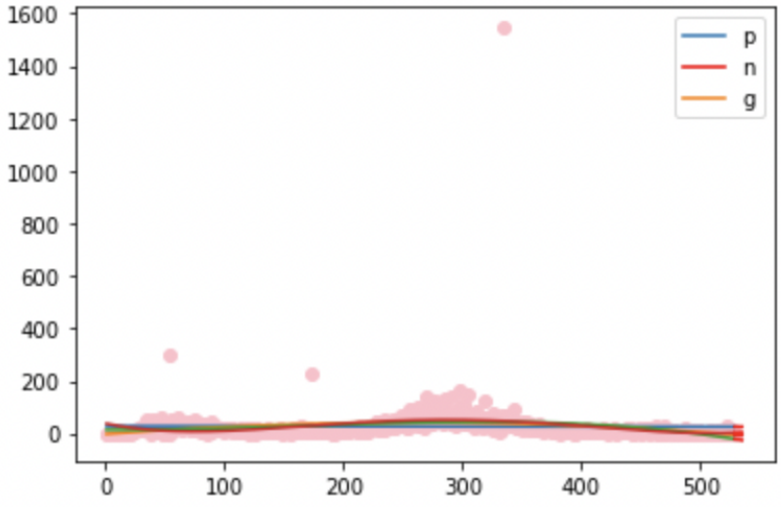


Below is the plot of deaths per day from the first cases reported in the state and the linear regression model.



The root mean square error values for polynomial regression are:
RMSE of ploy1 is: 1528.0319390717375
RMSE of ploy2 is: 1787.4075282858523
RMSE of ploy3 is: 1787.4867439705563
RMSE of ploy4 is: 1980.7582226374248

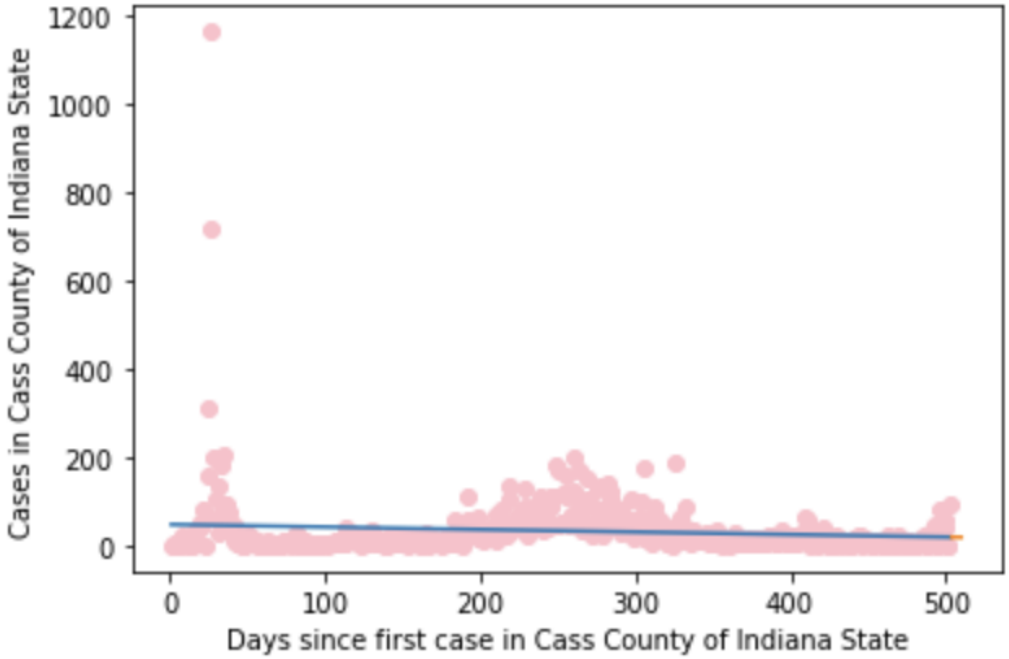
The prediction of deaths of next week of Indiana State.



We calculated the top 5 infected counties in Stage 2. So, using the same counties here to fit a linear model and predict future cases and deaths.

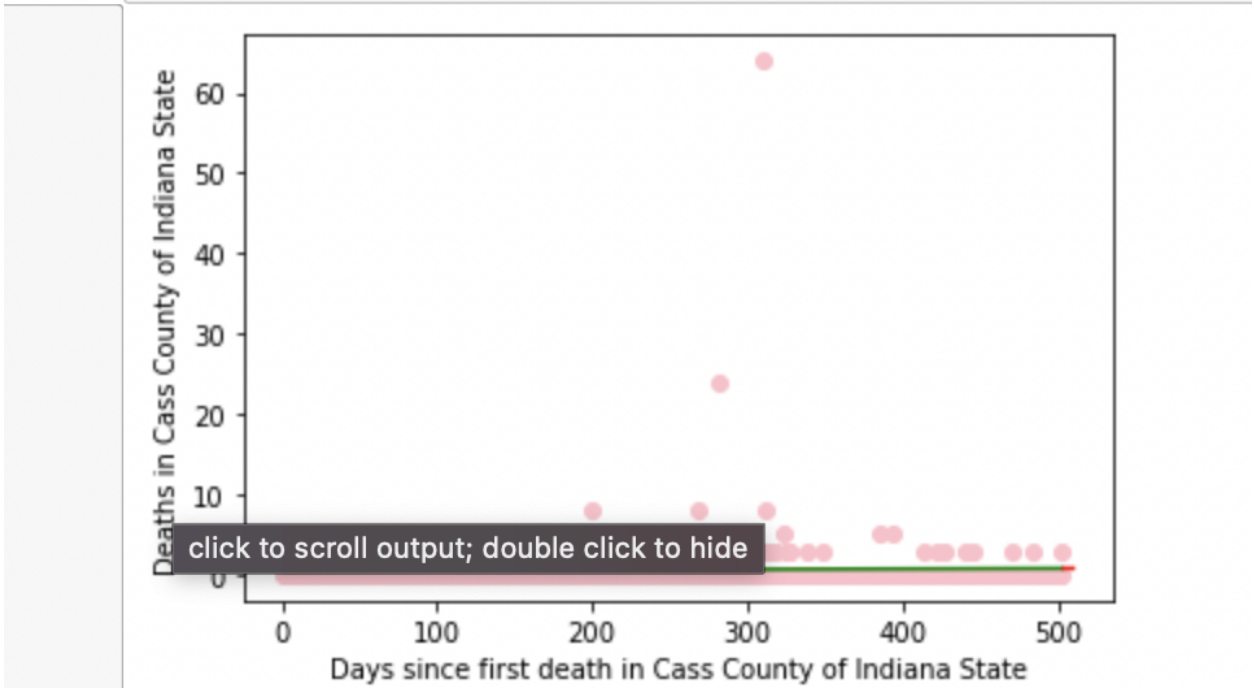
18017 - Cass County,
18037 - Dubois County,
18045 - Fountain County,
18023 - Clinton County and
18039 - Elkhart County

For Cass county
The plot of cases per day from the first day of cases in Indiana State and its liner regression is.



Root Mean square error of linear regression of cases in Cass county
RMSE of ploy1 is: 70.80798585272849
RMSE of ploy2 is: 70.66944566124013
RMSE of ploy3 is: 69.99891067570582
RMSE of ploy4 is: 67.5357018500736

The plot of deaths per day from the first day of cases in Indiana State and its liner regression is.



Root Mean Square Error of Cass county Deaths Linear regression model.
RMSE of ploy1 is: 33.90619864603722
RMSE of ploy2 is: 34.14277899819973
RMSE of ploy3 is: 35.44399656437719
RMSE of ploy4 is: 39.7915000132957

Reading the hospital dataset to find total available beds in Indiana State's hospitals. Then using this data to find the point of no return.
Total beds in Indiana State's hospitals 19419
In Indiana state there are 6107 hospital beds available where as the prediction for cases in next month is close to 2000. This states that the the Indiana states is not reaching the point of return in next month.

Sytiva:
Task-1:

The state I chose was New York once again. I found that for the state, the best fitting regression lines were fifth order polynomial regression lines for both new cases and new deaths. For new cases, $R^2 = 0.72$ and for new deaths, $R^2 = 0.38$. The RMSEs for new cases and new deaths were 5348.60 and 289.10 respectively. The counties modeled to be top 5 with cases were found to be Rockland, Richmond, Westchester, Nassau, and Suffolk. Rockland currently has the highest amount of cases while Westchester has the lowest among the top 5 counties. Every county's cases increase in the beginning of the pandemic from 0 to between

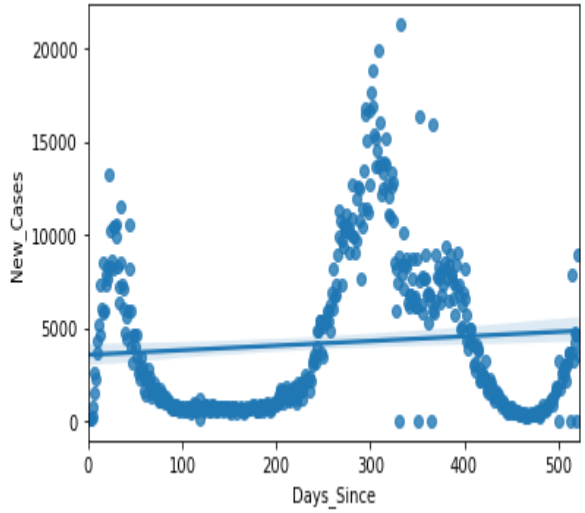
either 200 or 400 cases, stay mostly constant until between october 2020 and january 2021 where they increase steadily until becoming constant again at between 1300 and 1500 cases.

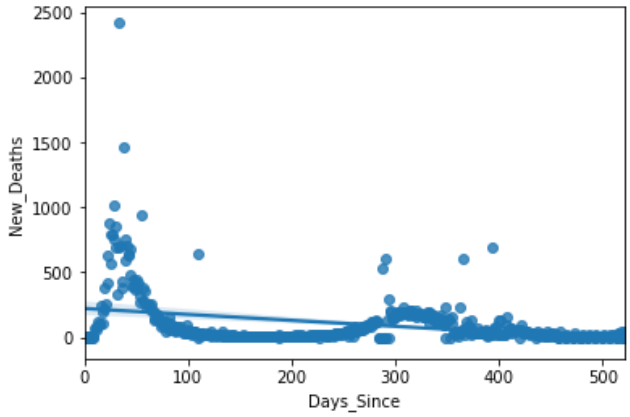
When finding regression lines for these counties, a linear regression line was best to use for each of them. The $R^2 = 0.91$ for Nassau, Suffolk, and Richmond while $R^2 = 0.94$ for Rockland and $R^2 = 0.92$ for Westchester. The capacity for New York's hospitals was 4767 total ICU beds and the trendline found from the hospital data for this state showed that New York has not approached its point of no return at all. Using weekly occupied ICU beds from the data, the trendline shows that this number only peaks at a little under 4000 beds then starts declining and has not reached that point since.

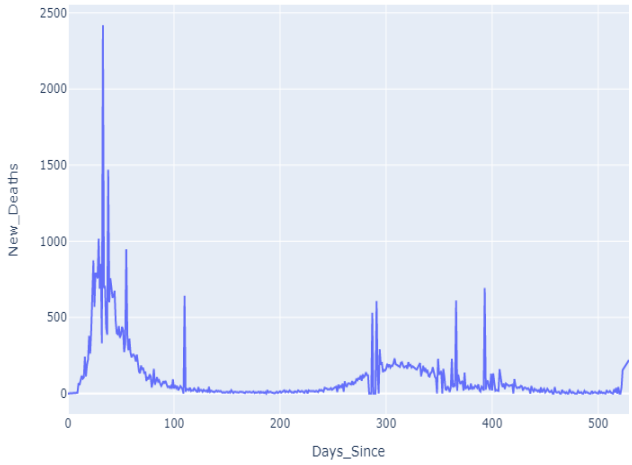
My hypotheses for my enrichment variables from stage 2 were that the covid cases and deaths had no relation to any of them. However, after performing chi-square tests of independence for all three of the enrichment variables, it was found that there is a relationship between the variables and the covid cases and deaths for the state of NY.

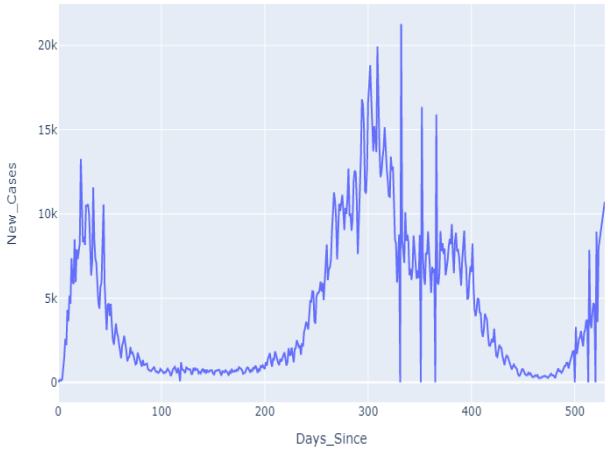
Task-2:

For Task 2, the confidence intervals, trendlines and 1 week forecasts for NY's new cases and deaths were graphed. The hospital datas confidence interval, trendline and forecast was graphed as well.

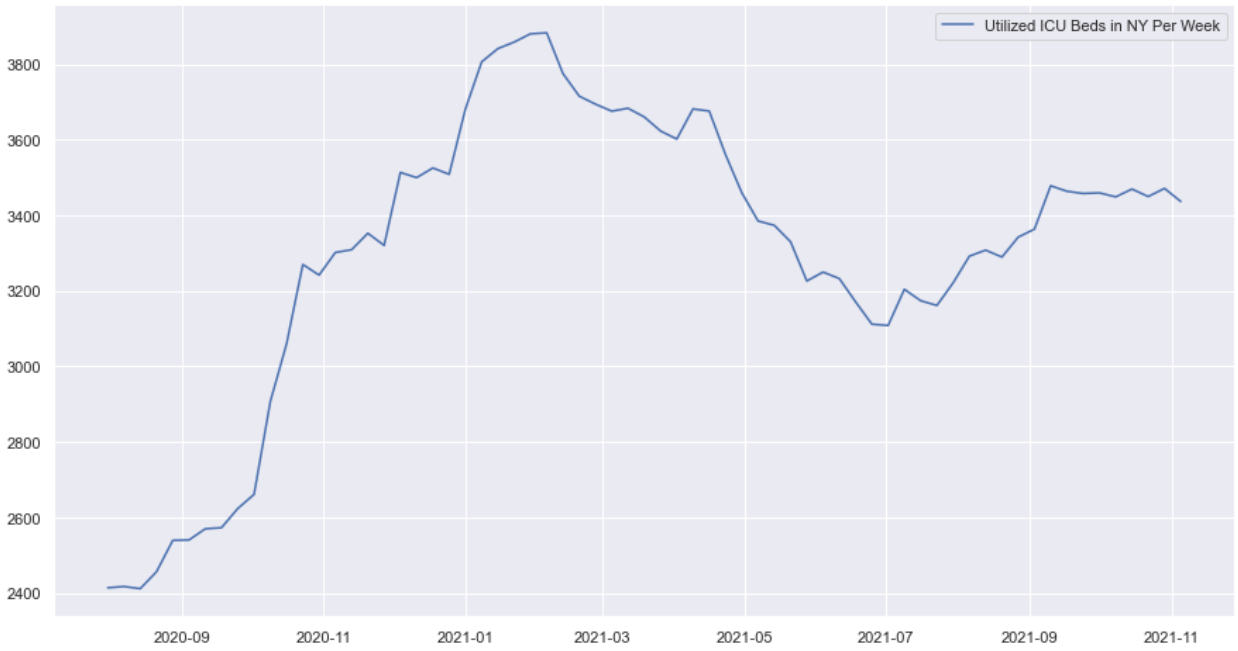


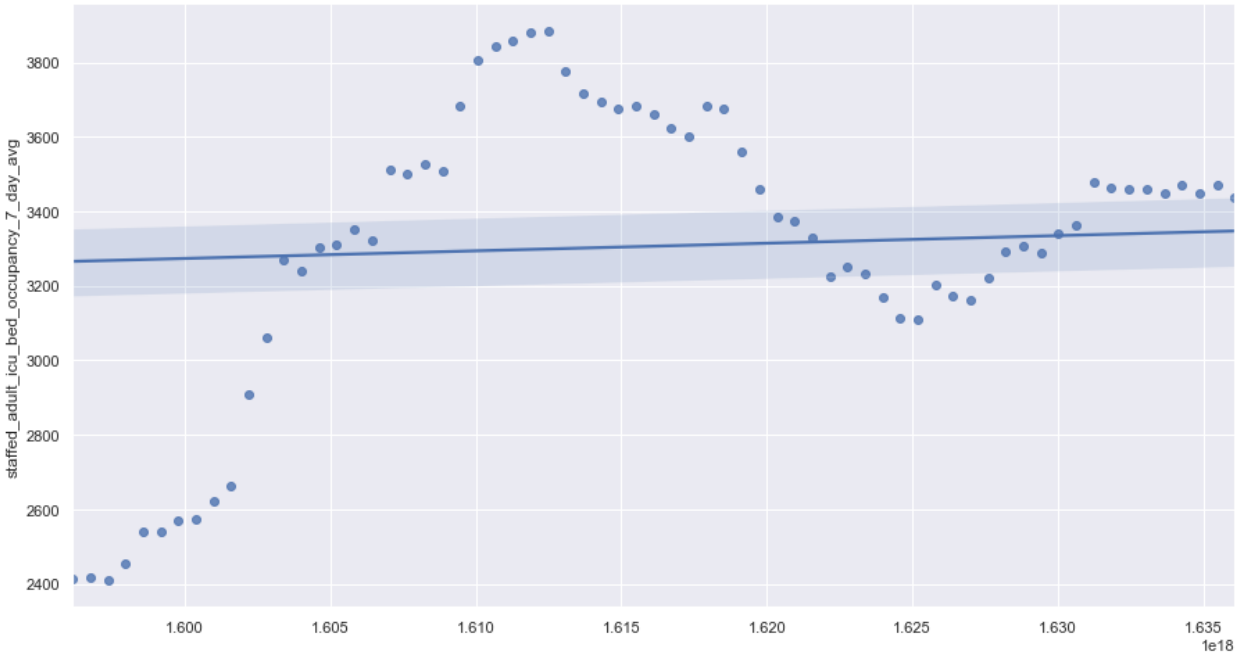


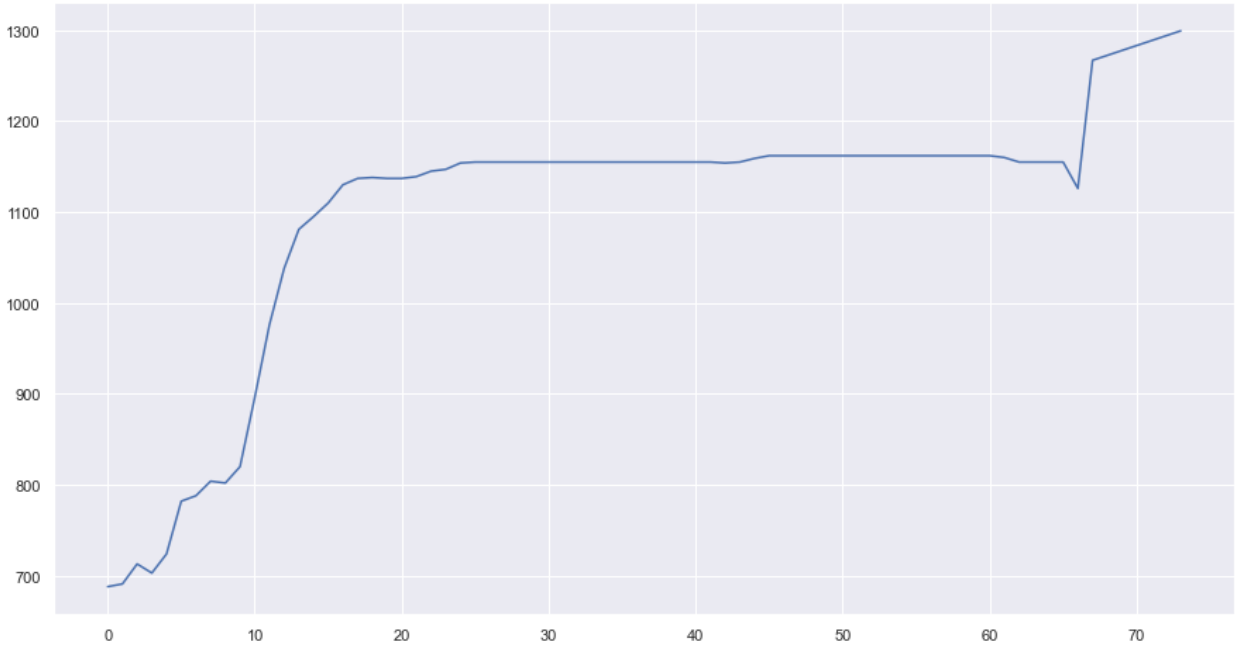




Hospital Data (Order: Trendline, Confidence interval, Forecast)


Utilized ICU Beds in NY Per Week
3800
3600
3400
3200
3000
2800
2600
2400
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11


staffed_adult_icu_bed_occupancy_7_day_avg
3800
3600
3400
3200
3000
2800
2600
2400
1.600
1.605
1.610
1.615
1.620
1.625
1.630
1.635
1e18


1300
1200
1100
1000
900
800
700
0
10
20
30
40
50
60
70