Data Science

Stage 3 Report

Utilize machine learning and statistical models to predict the trend of COVID-19 cases / deaths

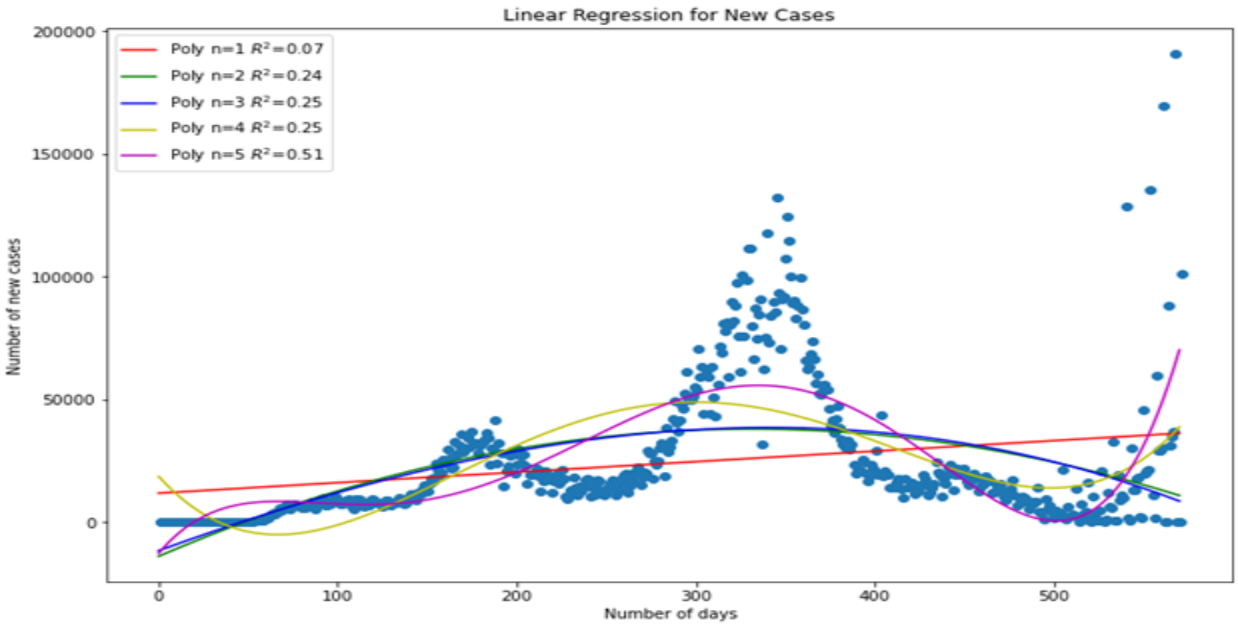
The main agenda of this stage is to fit a linear or non-linear regression model for new covid cases and death cases dataset and predict the future infection count. We first calculate the x i.e number of days from where the first covid cases occurred upto recent date the dataset has. Y is the number of new cases registered for that day. Similarly, death cases dataset is built for first death case occurred upto recent date and corresponding number of death cases. Two different models are generated, one for new cases and other for death cases. We calculate the Rsquared and RMSE values for each model and find the best fit model.

Team Task:

Develop Linear and Non-Linear (polynomial) regression models for predicting cases and deaths in US

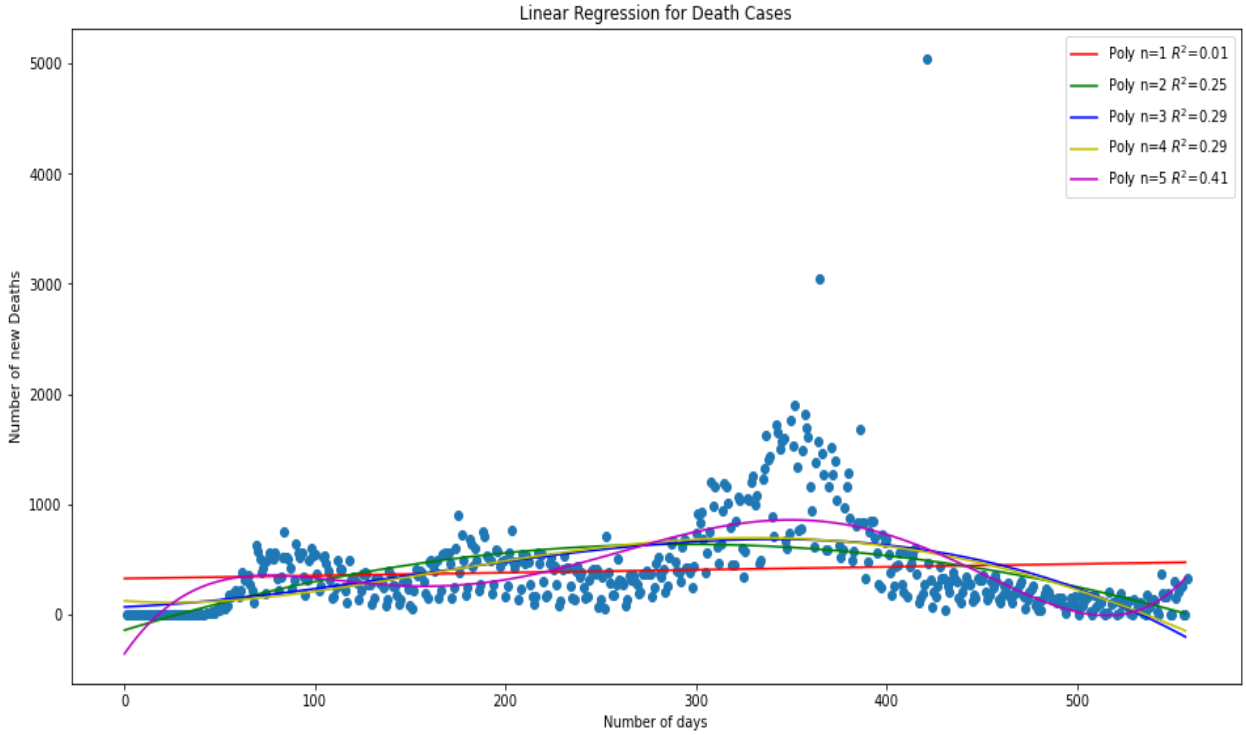
Here we developed a linear and non-linear regression model for the United states and results are shown below.

US everyday new cases:



The linear model has an insignificant R Squared value, with the increase in degree, there is a big increase when a non-linear model is fit for n=2 and it remains almost the same utp 4th degree and at n=5 the model's R Square is increased to 0.51. We will choose the model n=5 because there is an increase of more than 10% from the earlier model.

US everyday death cases:



The trend lines for the US determine an increase in the number of cases, but the death cases remain almost the same. The future prediction determines an increase of cases and no significant change with death cases

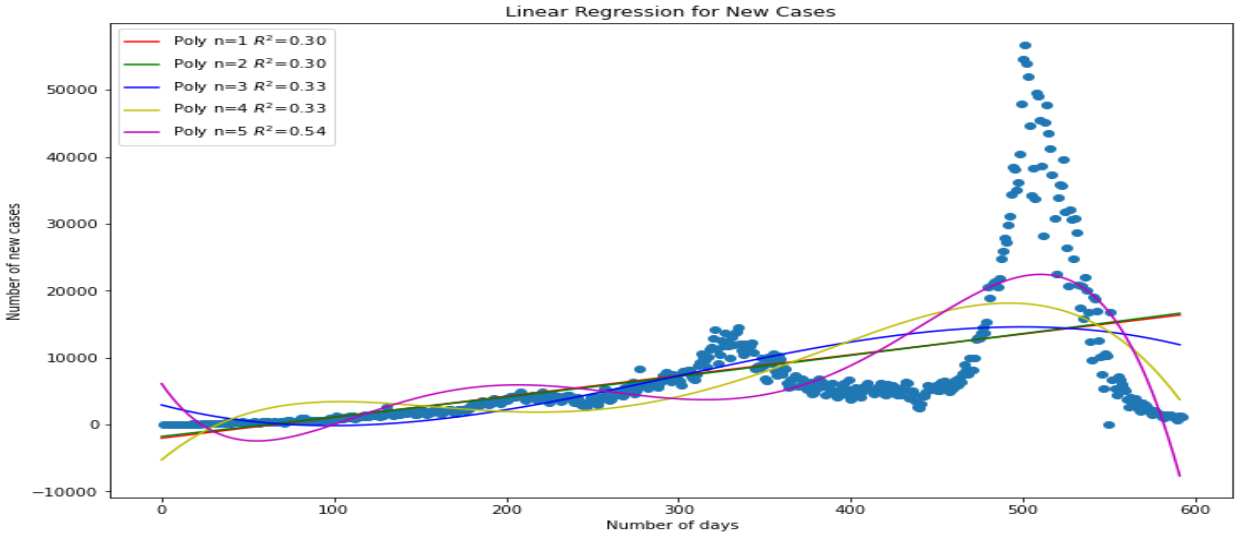
Implementing Linear/Non-Linear Regression models for other 5 countries

For this task, we chose 5 states

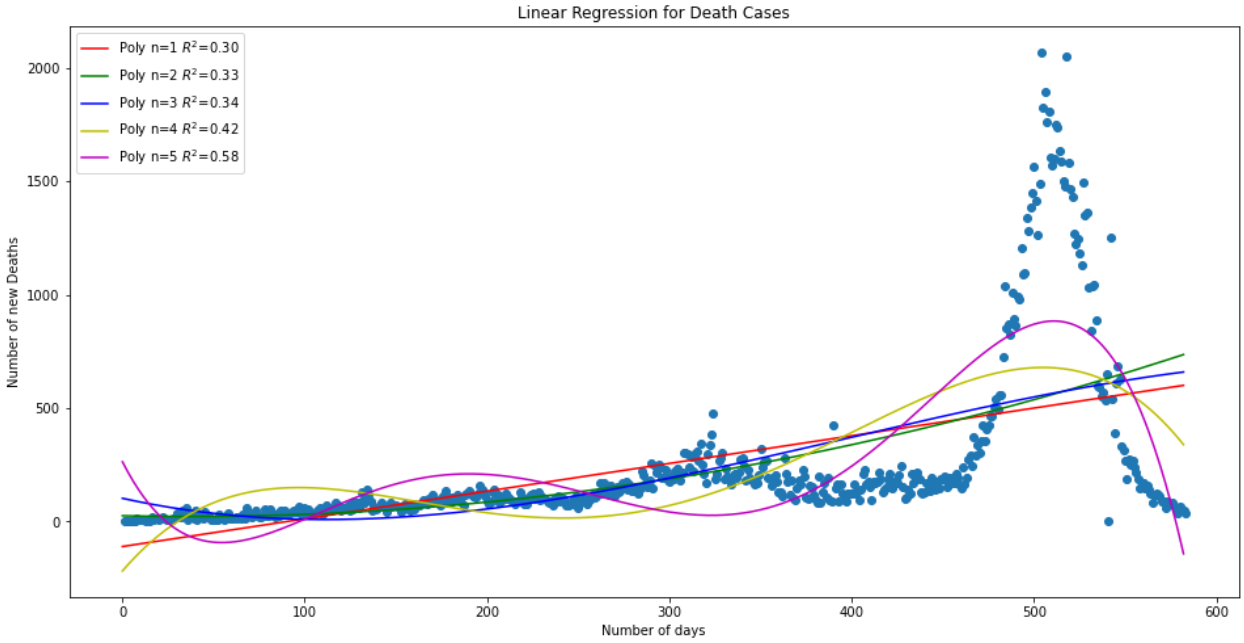
1. Indonesia
2. Brazil
3. Nigeria
4. Russia
5. Japan

Indonesia - everyday new cases -

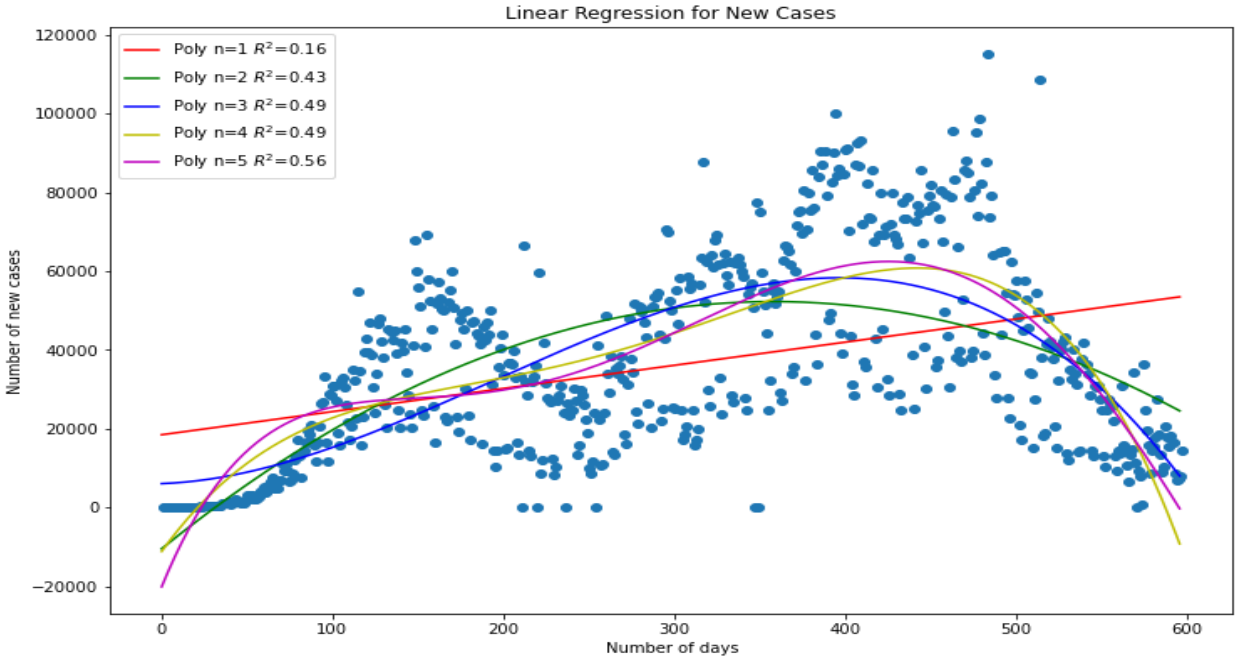
From the scatter plot we can observe that cases have decreased from peak and continuing a flat line pattern meaning a similar number of cases are occurring.



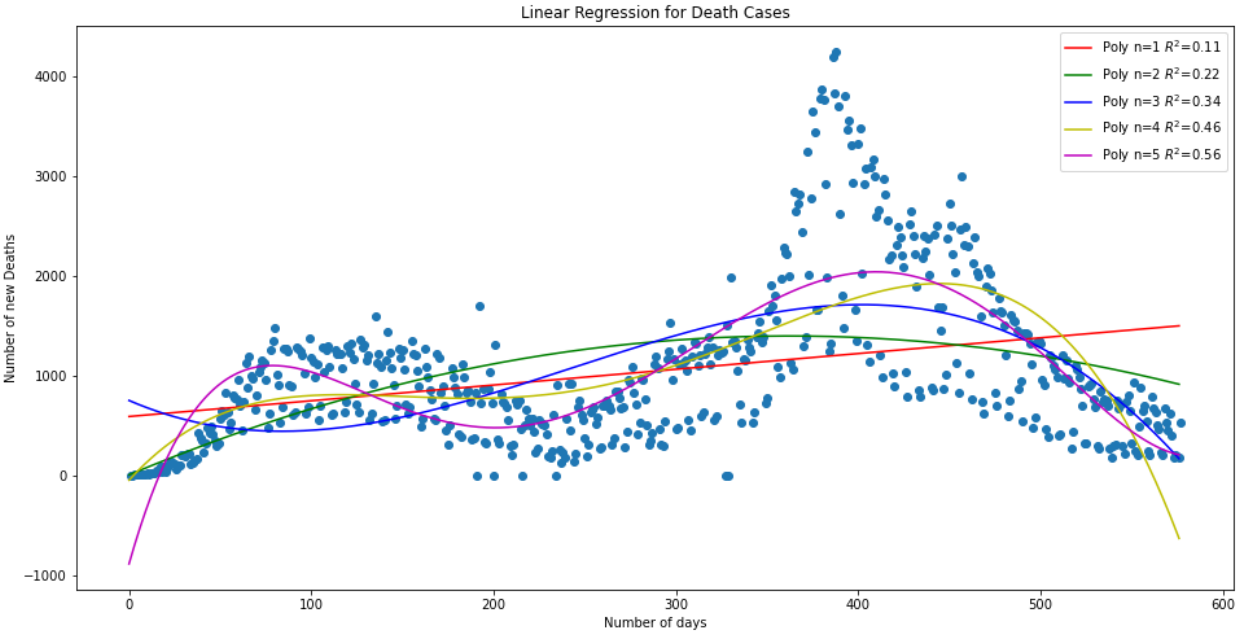
Indonesia - everyday death cases - death cases seems decreasing



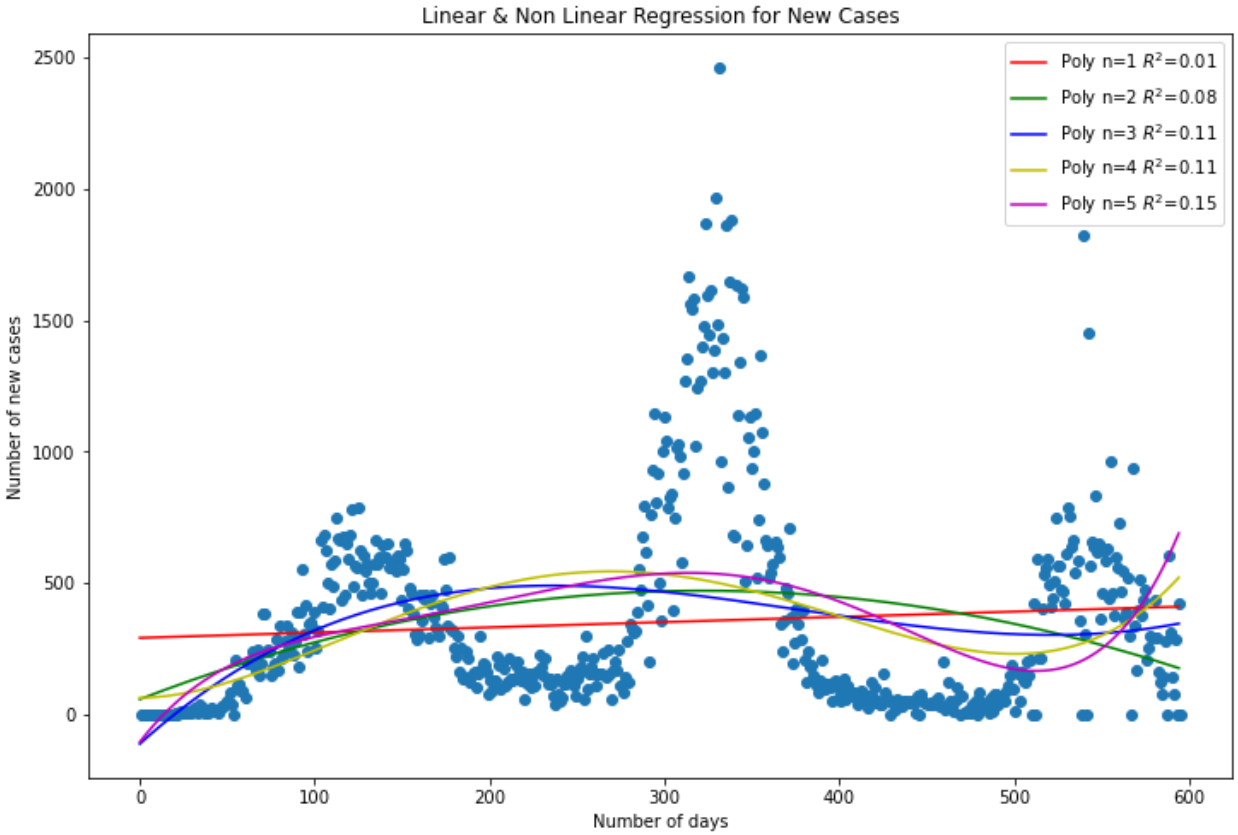
Brazil - everyday cases



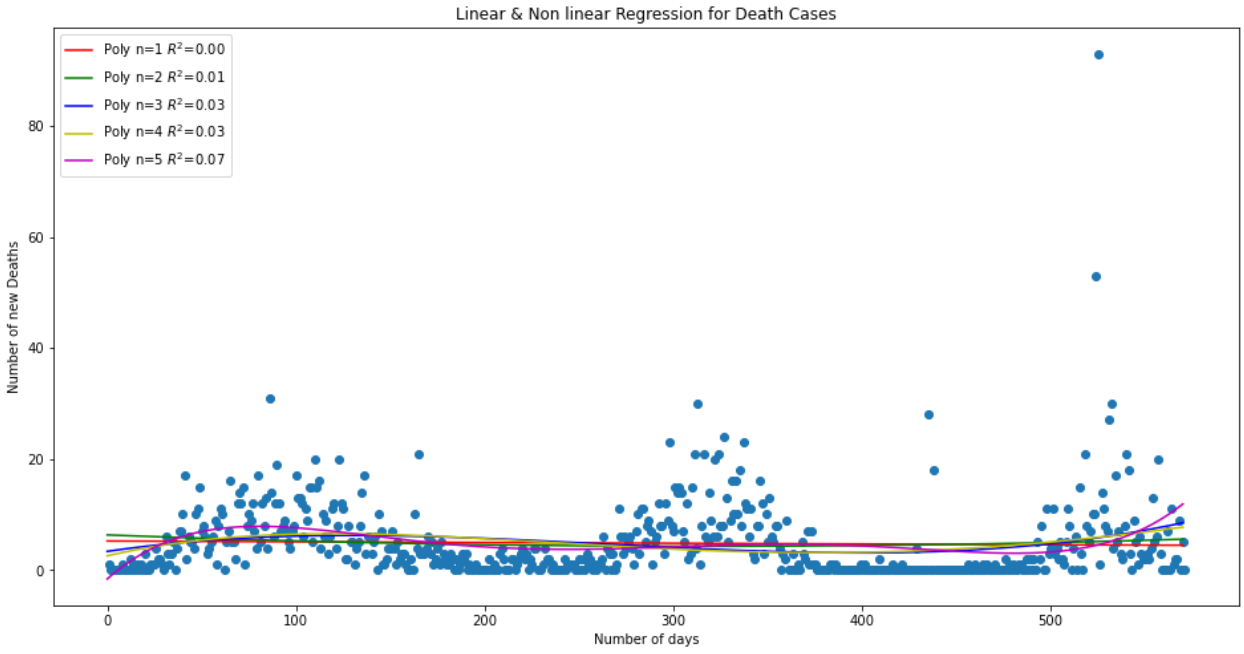
Brazil - death cases modelling



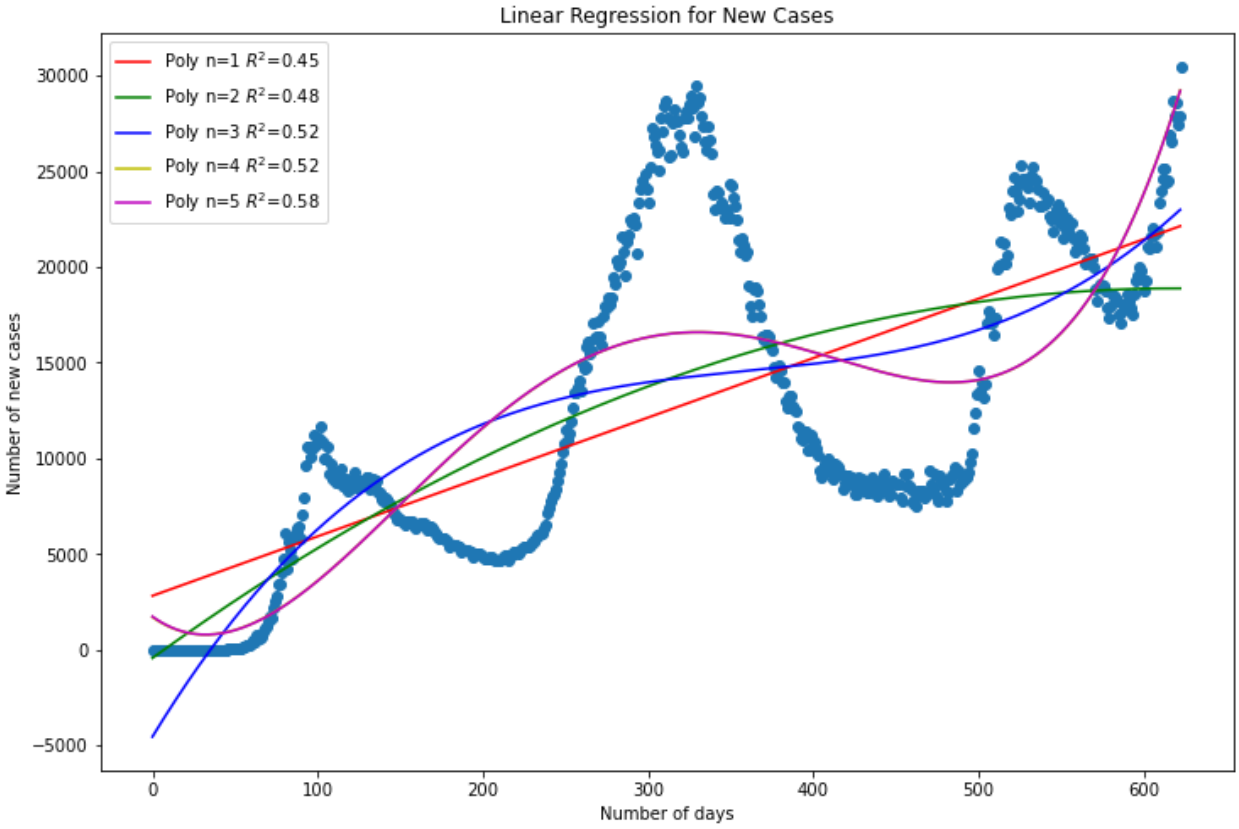
Nigeria - everyday new cases:



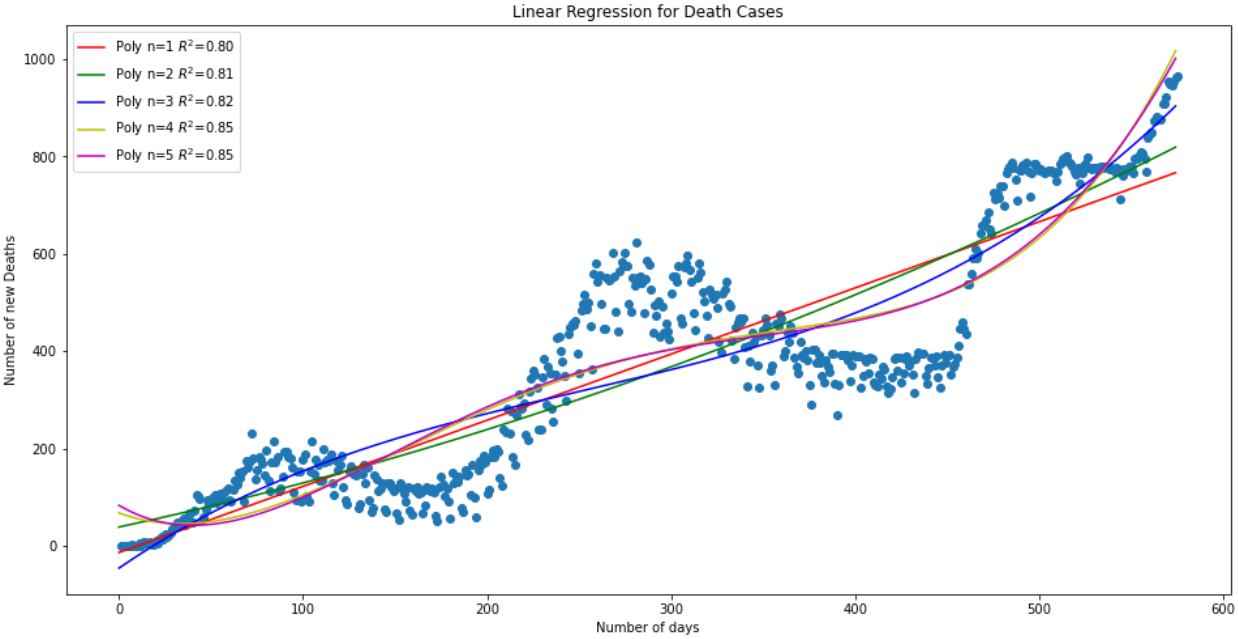
Nigeria - new death cases



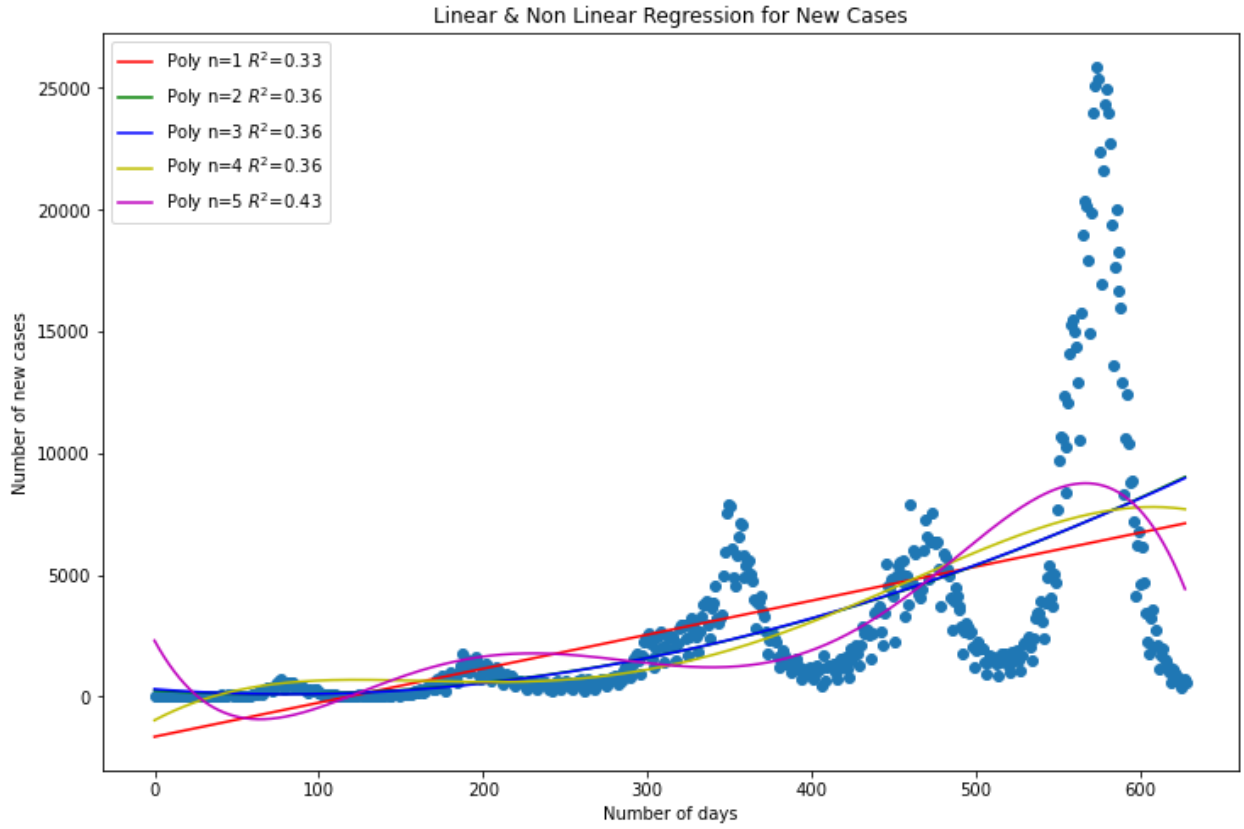
Russia - everyday new cases:



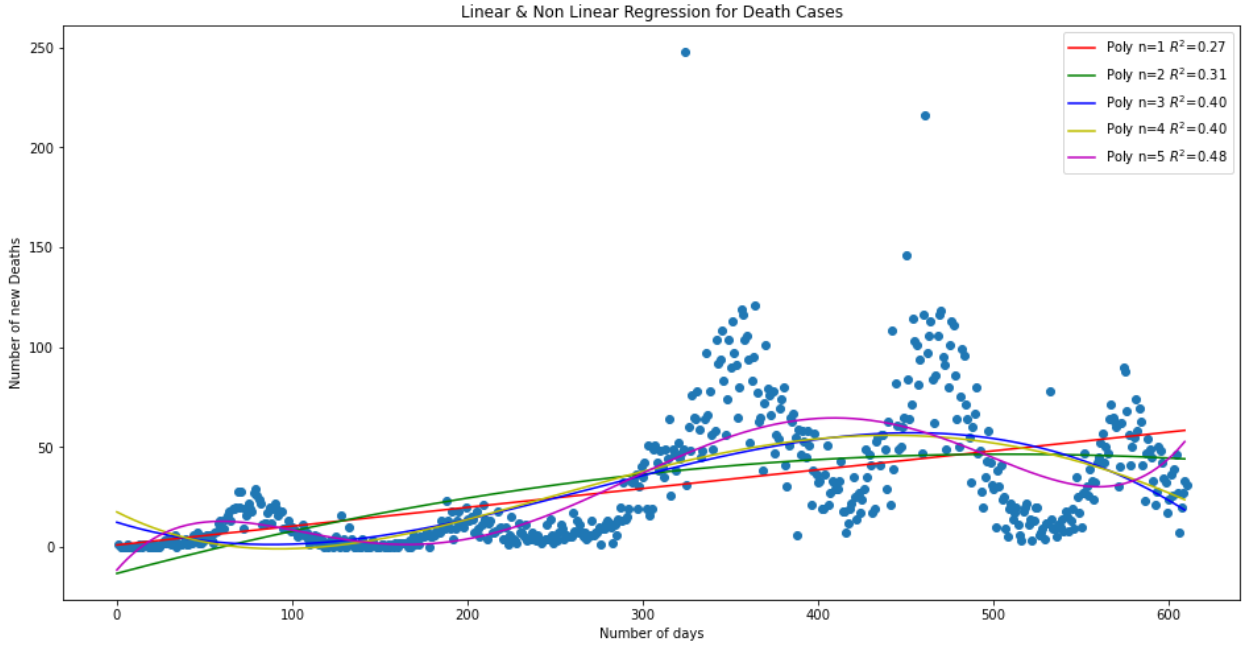
Russia - new death cases



Japan- everyday new cases:



Japan - new death cases:



Describe the trends of US compared to other countries:

When compared to other countries, the US is experiencing the most number of cases per day. But Russia is facing a higher number of death cases compared to all other states. Other than Russia and the US the other countries had peak cases at some point and the cases were reduced. Russia's curve looks like it is trending upwards in cases and deaths. The cases and deaths for the country Nigeria is very low(maximum 80 deaths per day) when compared to other countries. In total, after the US, Brazil had the highest cases per day. Indonesia's trend is almost flat for a year and then started increasing upto 55,000 cases per day and then suddenly dropped. Brazil's new cases and deaths are pretty much scattered for the whole time. Japan's highest case peak was in the last hundred days similarly it has highs and lows in its death rate as well.

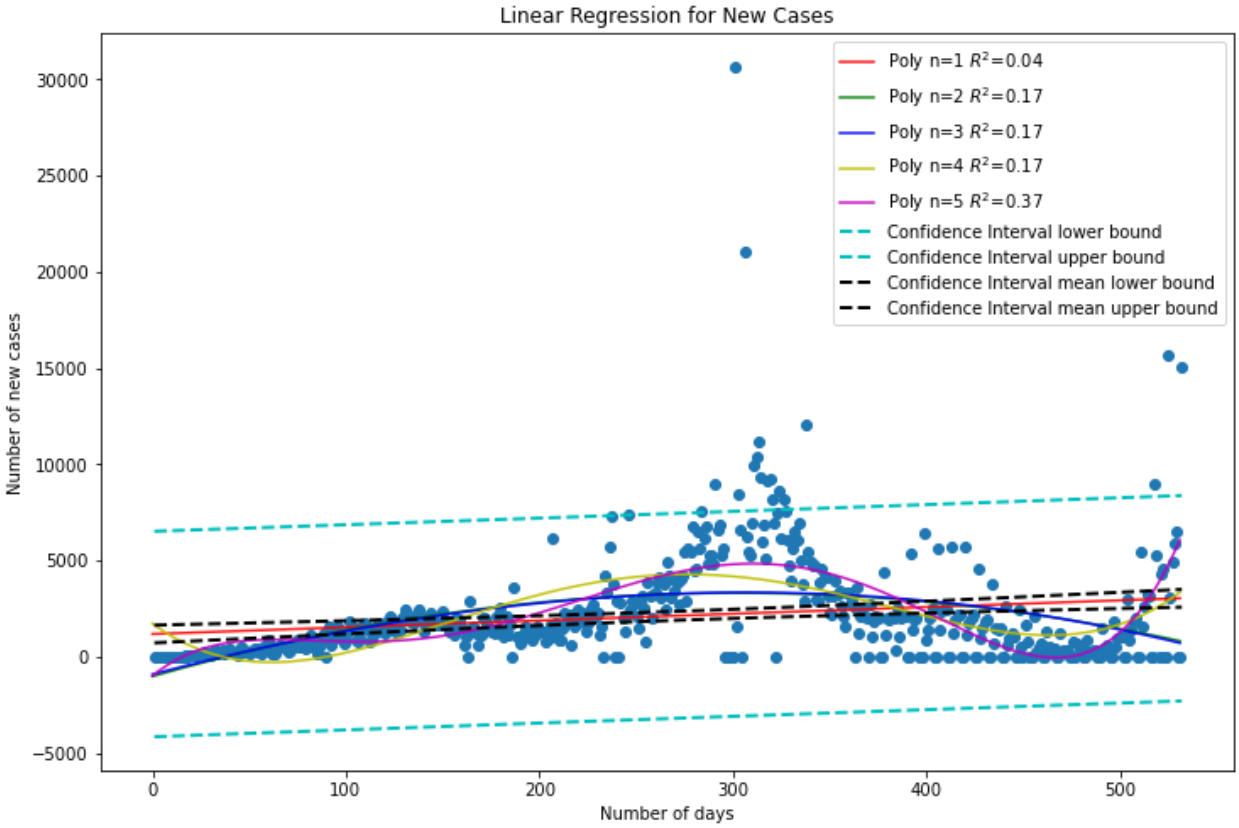
Individual Task

Deepa Jayanna

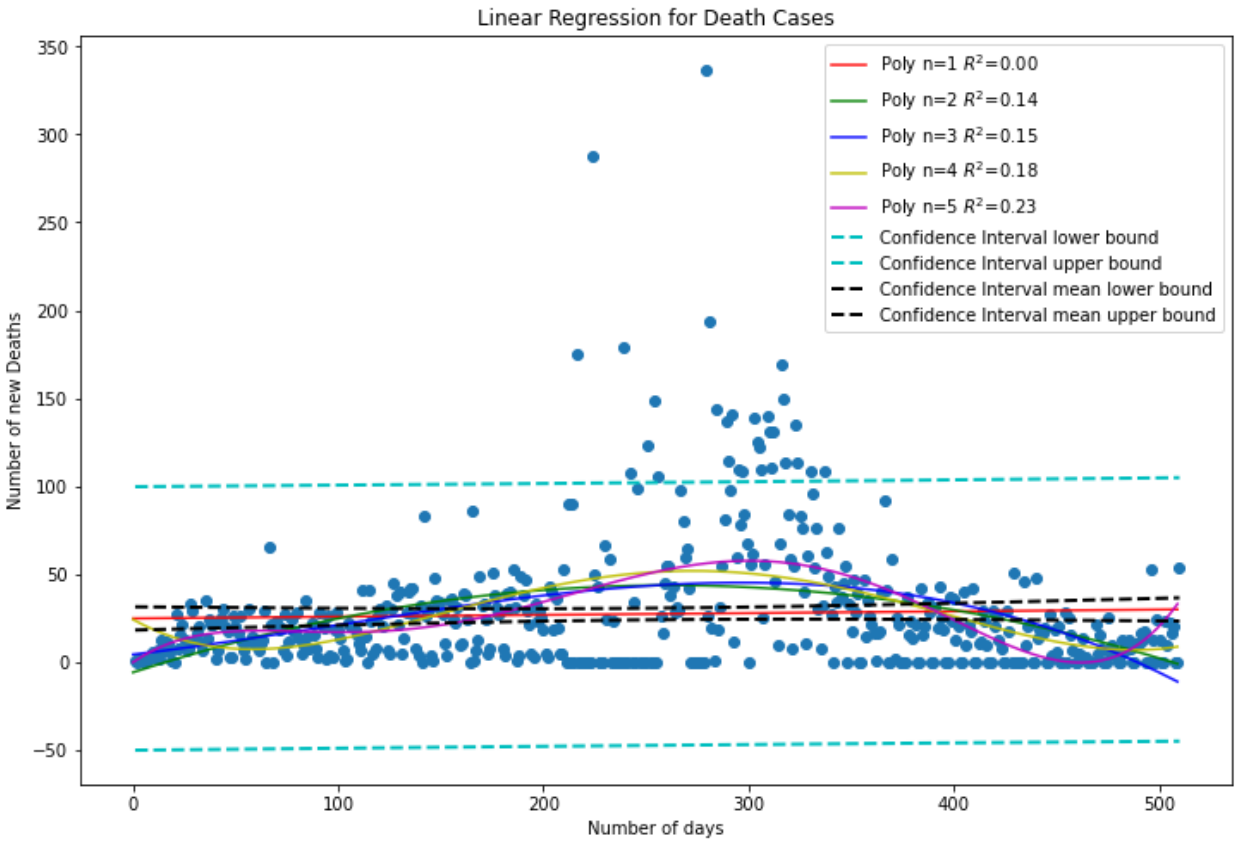
Utilize Linear and Nonlinear (polynomial) regression models to compare trends for a single state and its counties (top 5 with highest number of cases). Start your data from the first day of infections.

NC state everyday new cases:

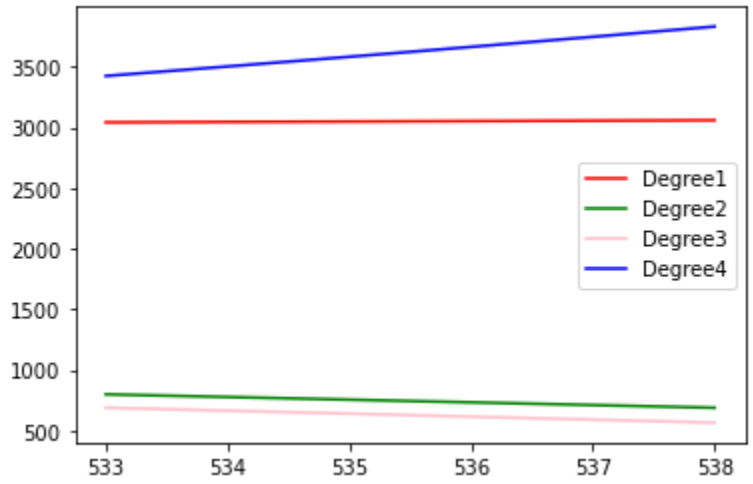
For this task, I am modeling linear regression and non-linear regression for North Carolina state. With n=5, Rsquared value stays at 0.37. Confidence intervals are also drawn



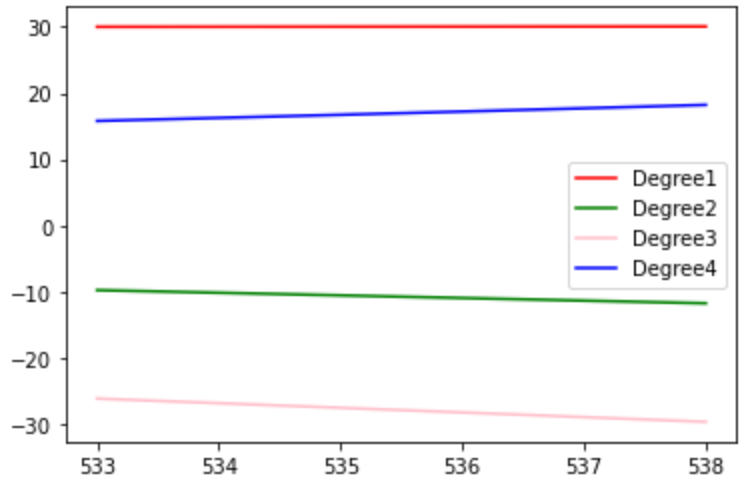
NC state new death cases:



NC future predictions: New cases

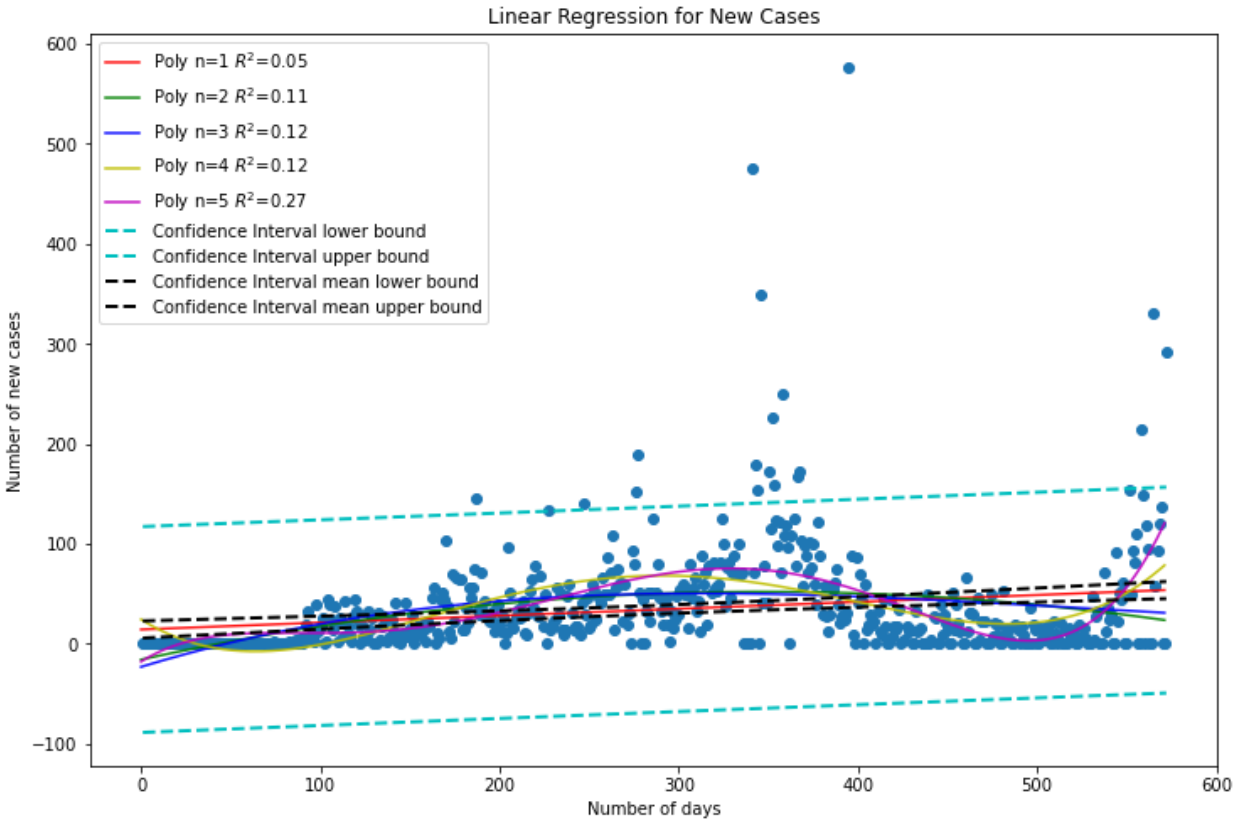


NC future predictions: Death cases

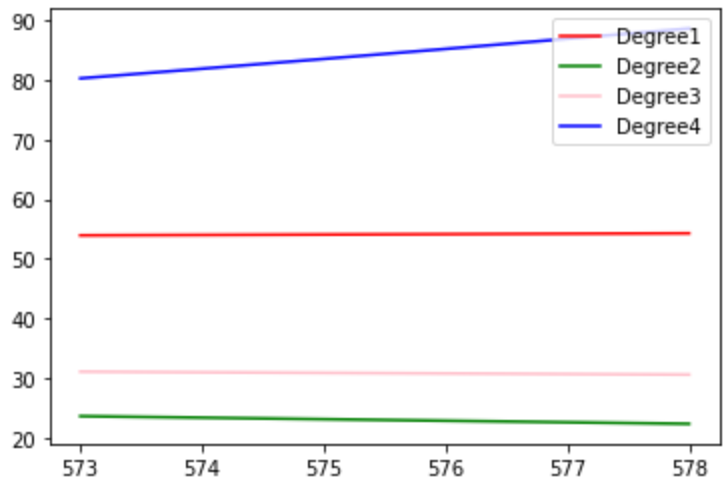


Comparison with counties which are highly impacted with new cases

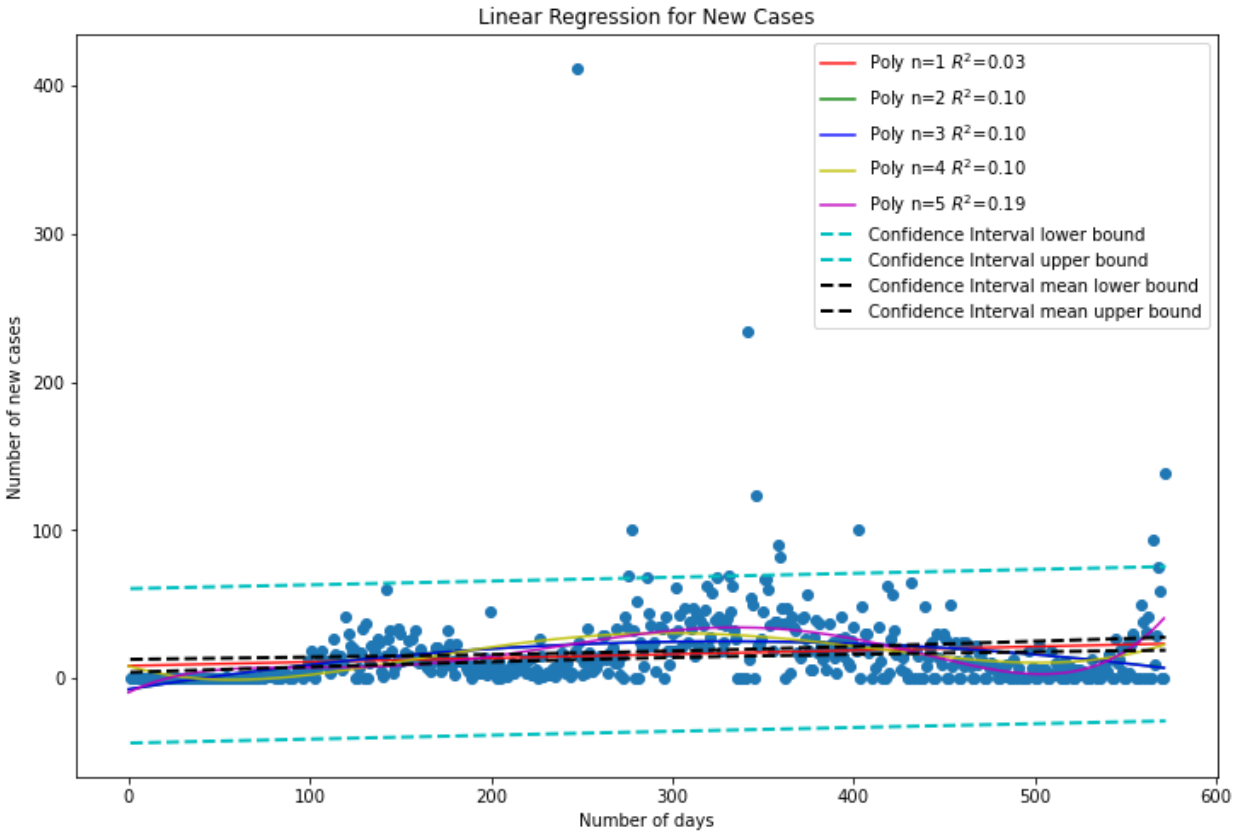
Robeson County:

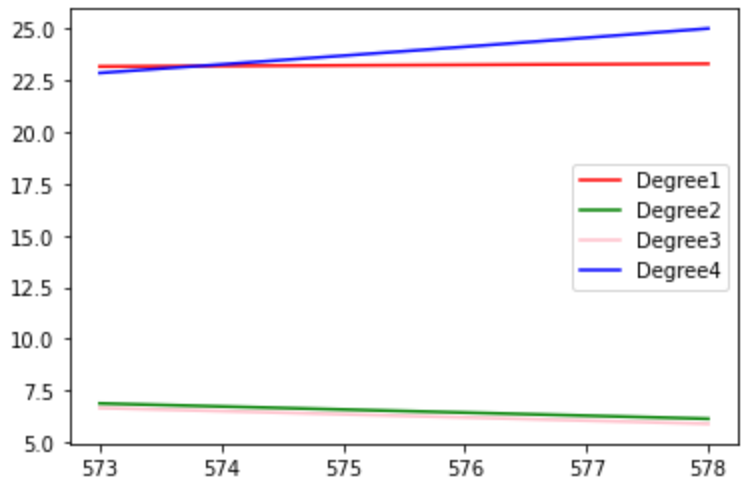


Future prediction of cases - Robeson county

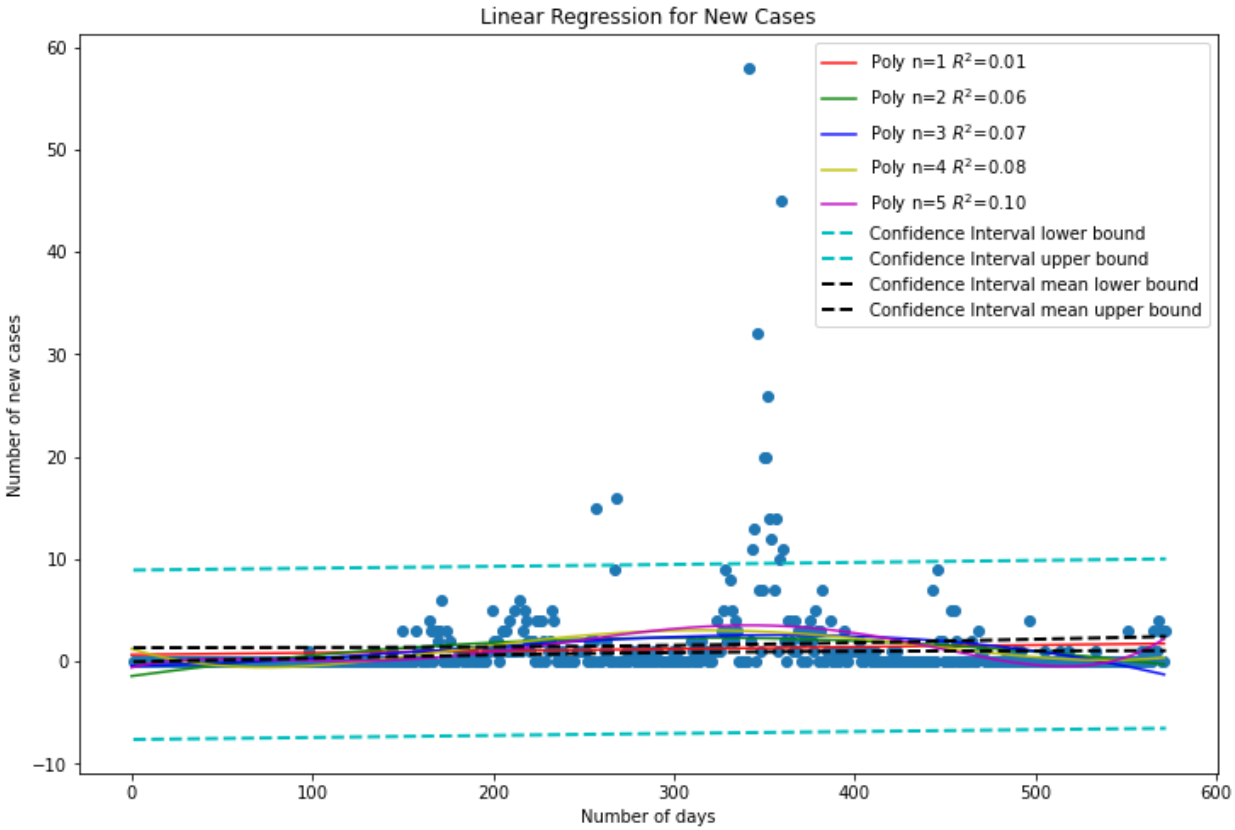


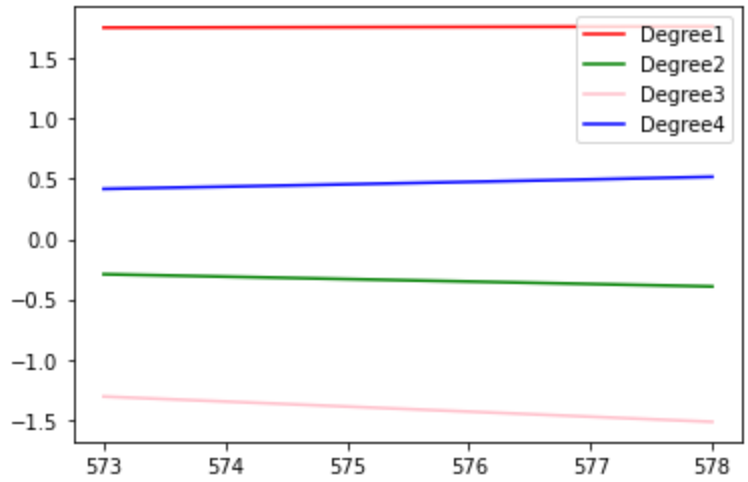
Sampson County:



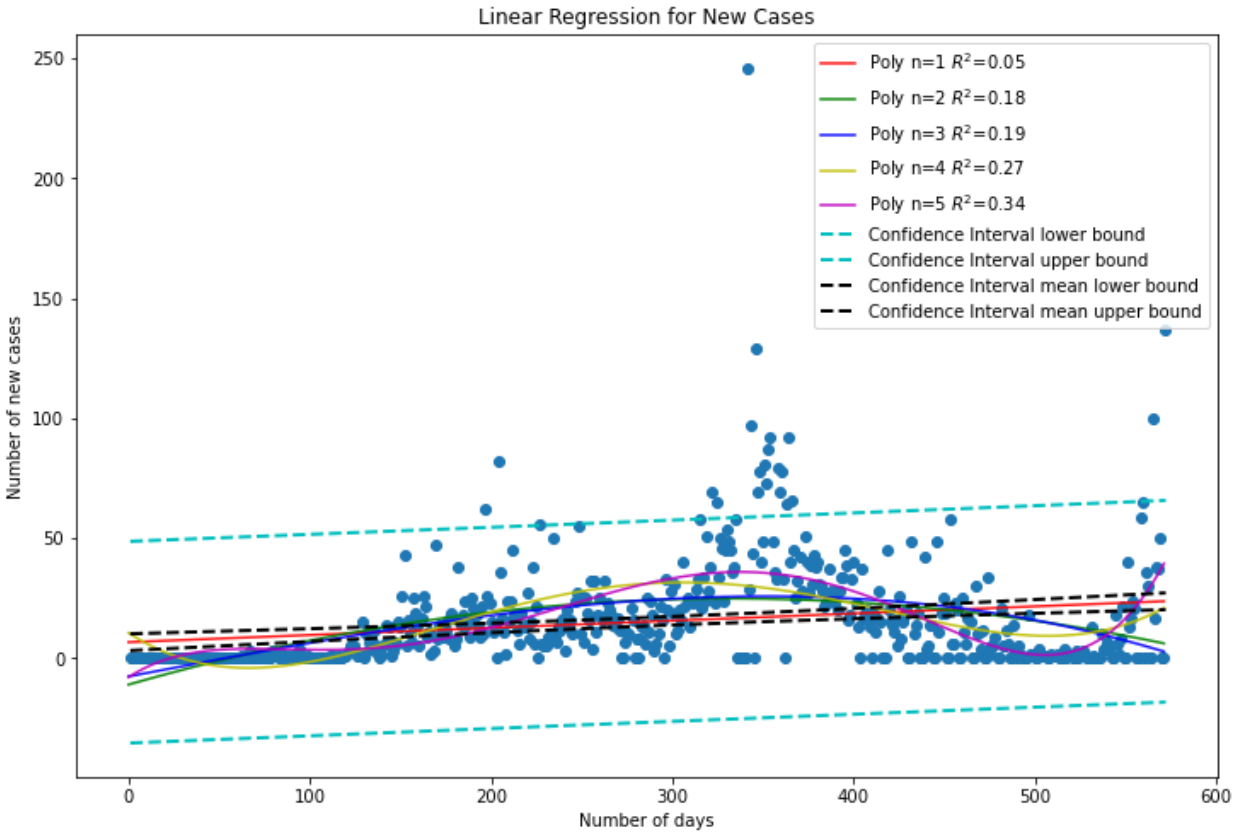

25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
573
574
575
576
577
578
Degree1
Degree2
Degree3
Degree4

Hyde County


Linear Regression for New Cases
Poly n=1 R^2=0.01
Poly n=2 R^2=0.06
Poly n=3 R^2=0.07
Poly n=4 R^2=0.08
Poly n=5 R^2=0.10
Confidence Interval lower bound
Confidence Interval upper bound
Confidence Interval mean lower bound
Confidence Interval mean upper bound
Number of new cases
Number of days
60
50
40
30
20
10
0
−10
0
100
200
300
400
500
600



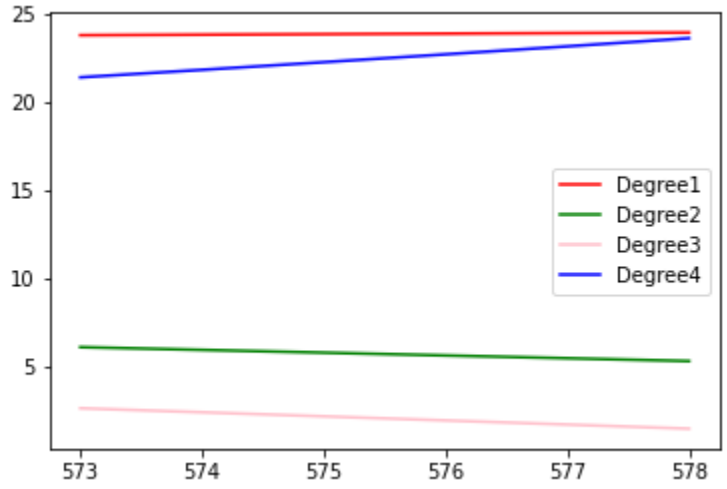
Stanly County



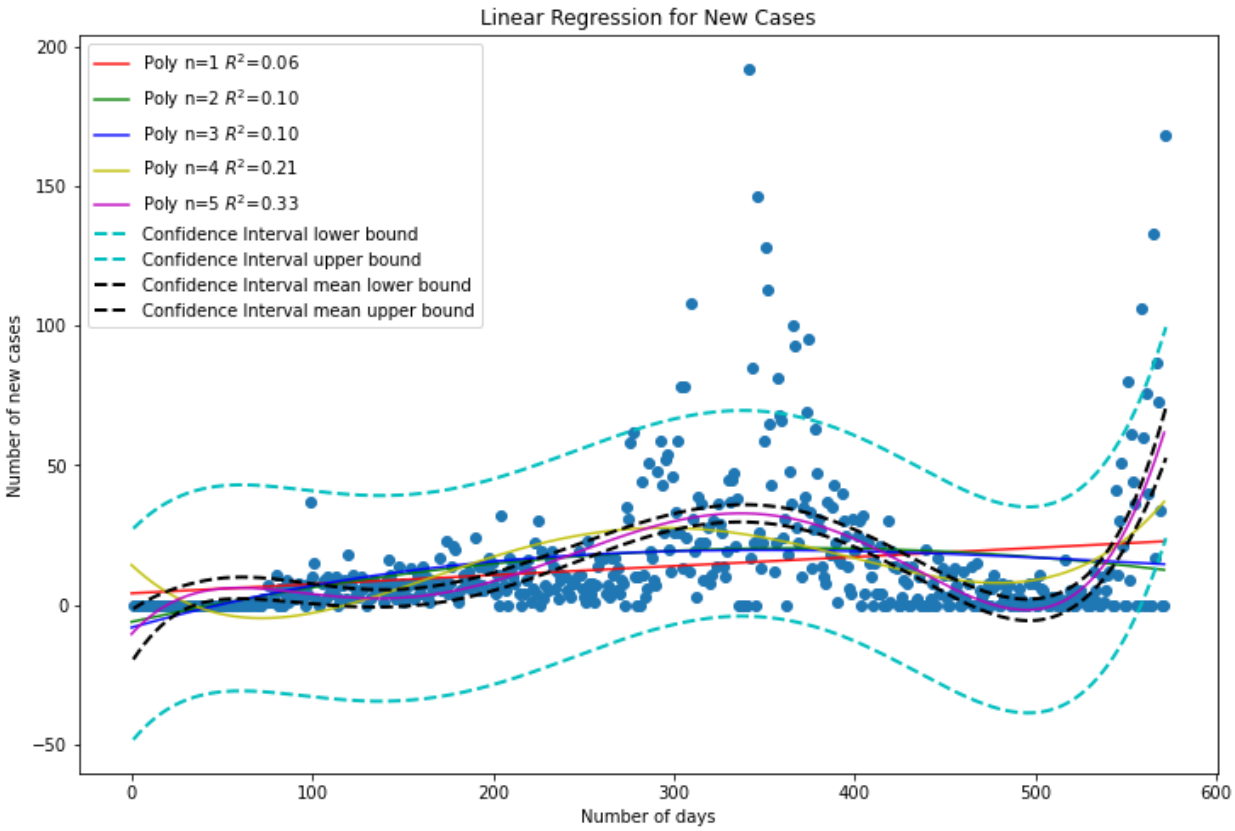
Top 5 infected counties Confidence intervals for Linear model - Death Cases

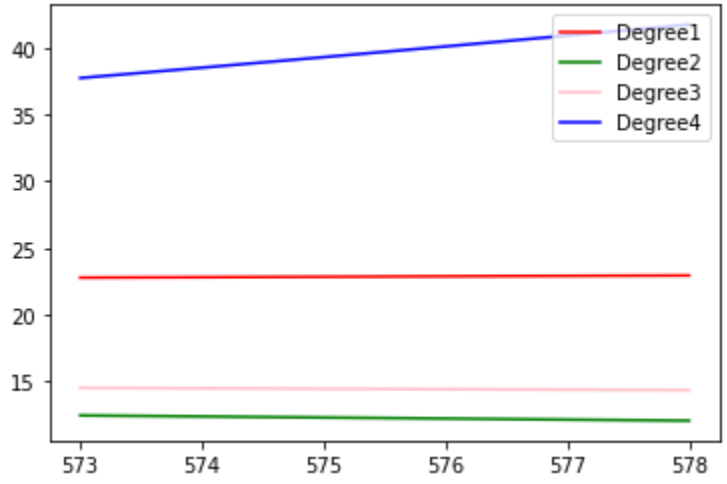
In [92]:
```
print("Montgomery County  Confidence interval",mo_model1.conf_int())
print('-----------------------------------------------------------')
print("Rutherford County Confidence interval",rf_model1.conf_int())
print('-----------------------------------------------------------')
print("Northampton County Confidence interval",nh_model1.conf_int())
print('-----------------------------------------------------------')
print("Jones County Confidence interval",jo_model1.conf_int())
print('-----------------------------------------------------------')
print("Columbus County Confidence interval",co_model1.conf_int())
```

```
Montgomery County  Confidence interval                    0         1
Intercept    0.032289  0.211989
num_of_days -0.000106  0.000437
-----------------------------------------------------------
Rutherford County Confidence interval                    0         1
Intercept   -0.051762  0.429568
num_of_days -0.000014  0.001441
-----------------------------------------------------------
Northampton County Confidence interval                    0         1
Intercept   -0.020421  0.136694
num_of_days -0.000068  0.000407
-----------------------------------------------------------
Jones County Confidence interval                    0         1
Intercept   -0.039316  0.093852
num_of_days -0.000120  0.000283
-----------------------------------------------------------
Columbus County Confidence interval                    0         1
Intercept    0.481387  7.778775
num_of_days  0.021518  0.043586
```



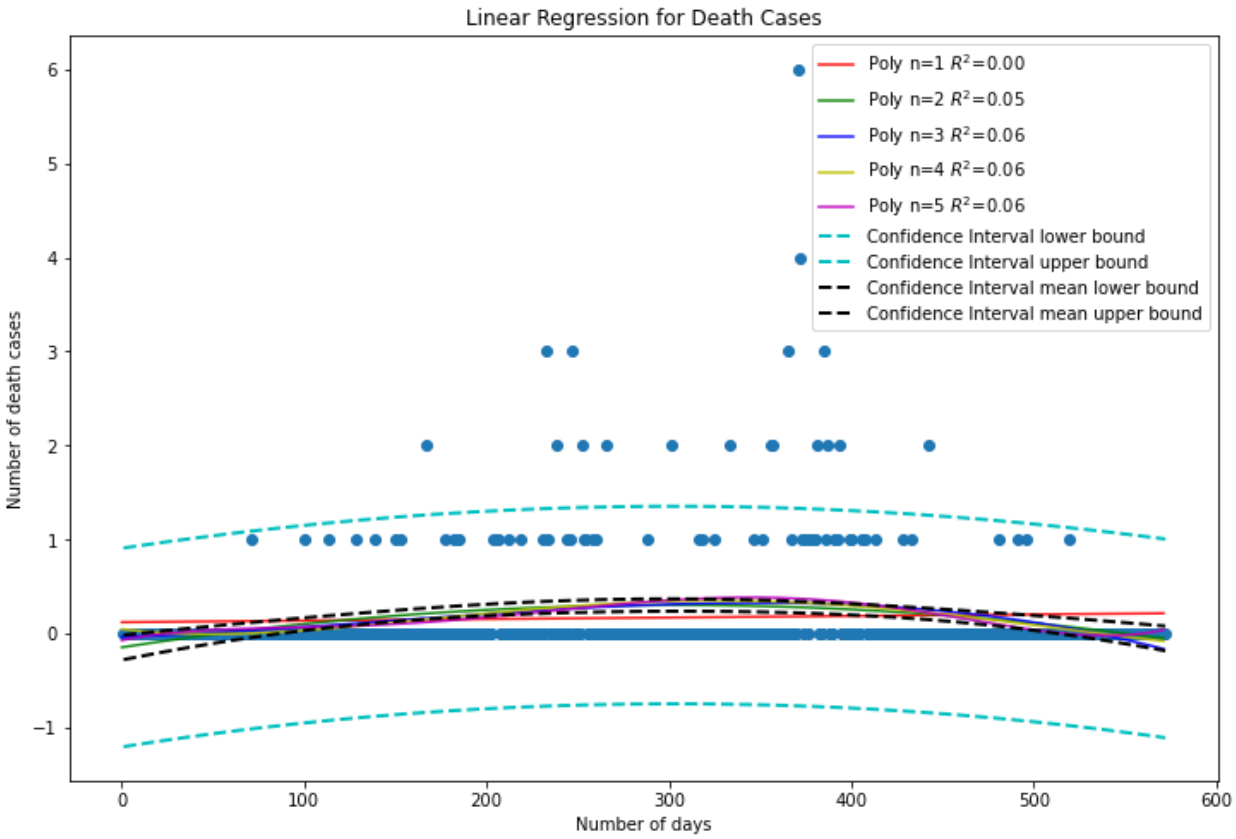
Columbus county

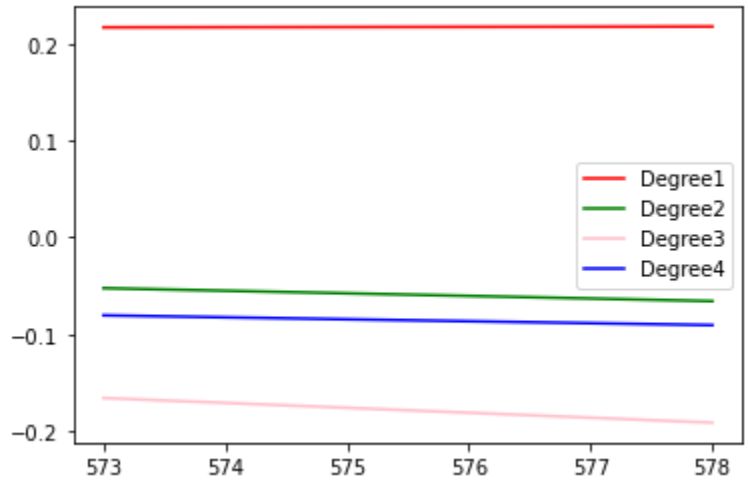




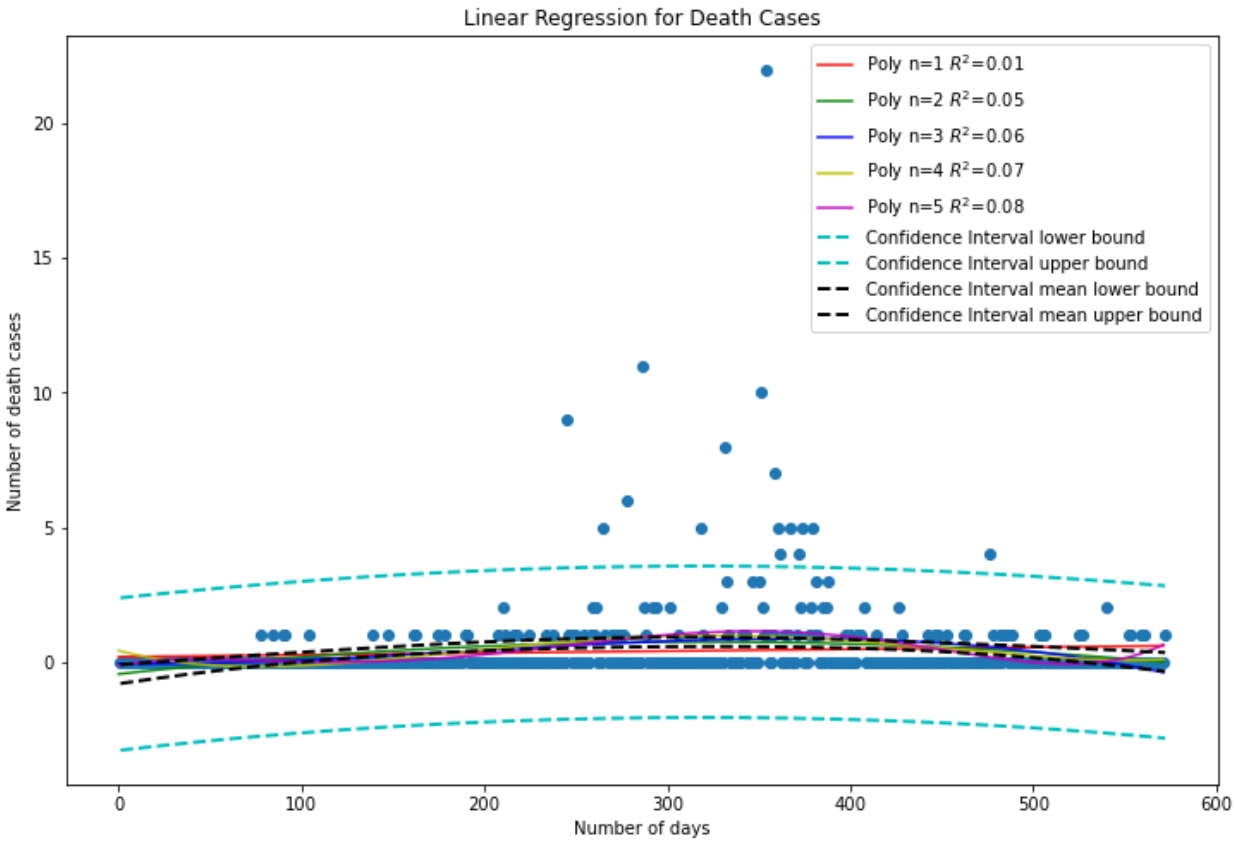
Modelling and Predictions for Top 5 Covid Death Counties in NC

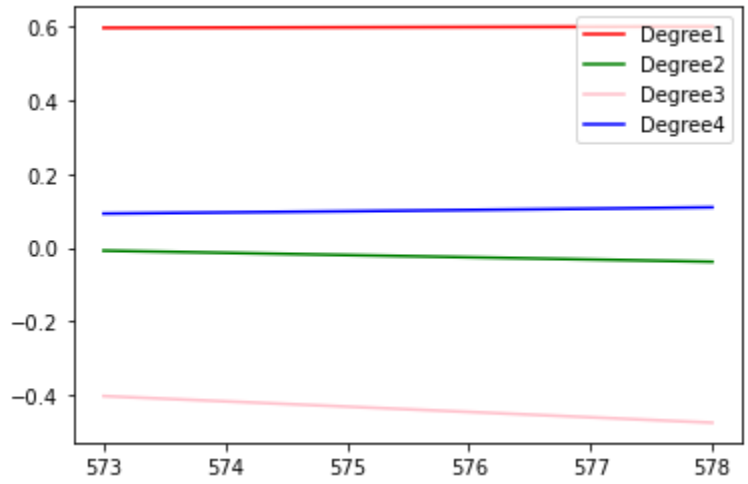
Montgomery County:



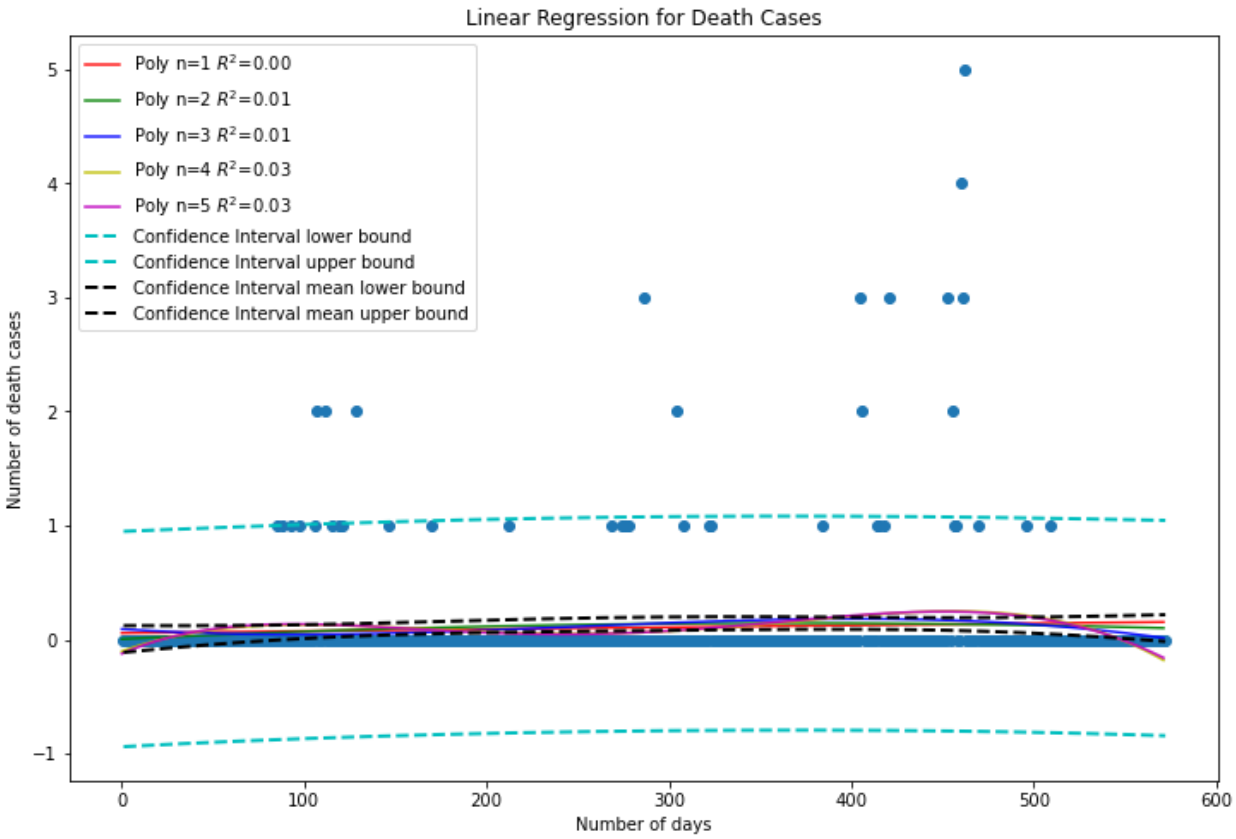


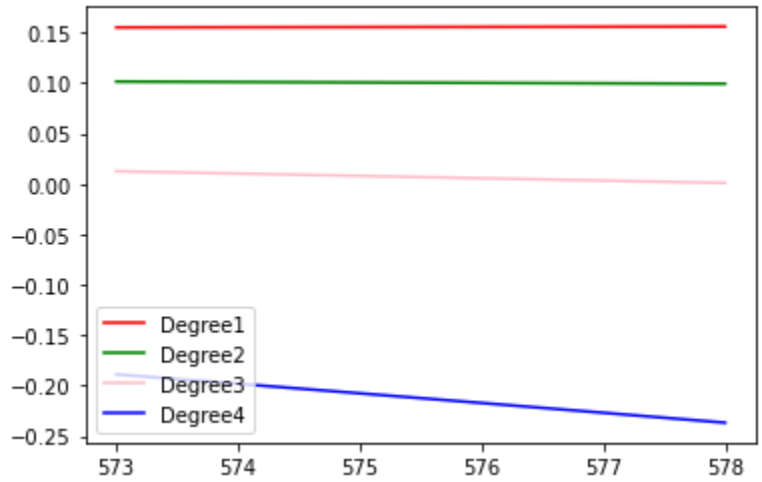
Rutherford county:



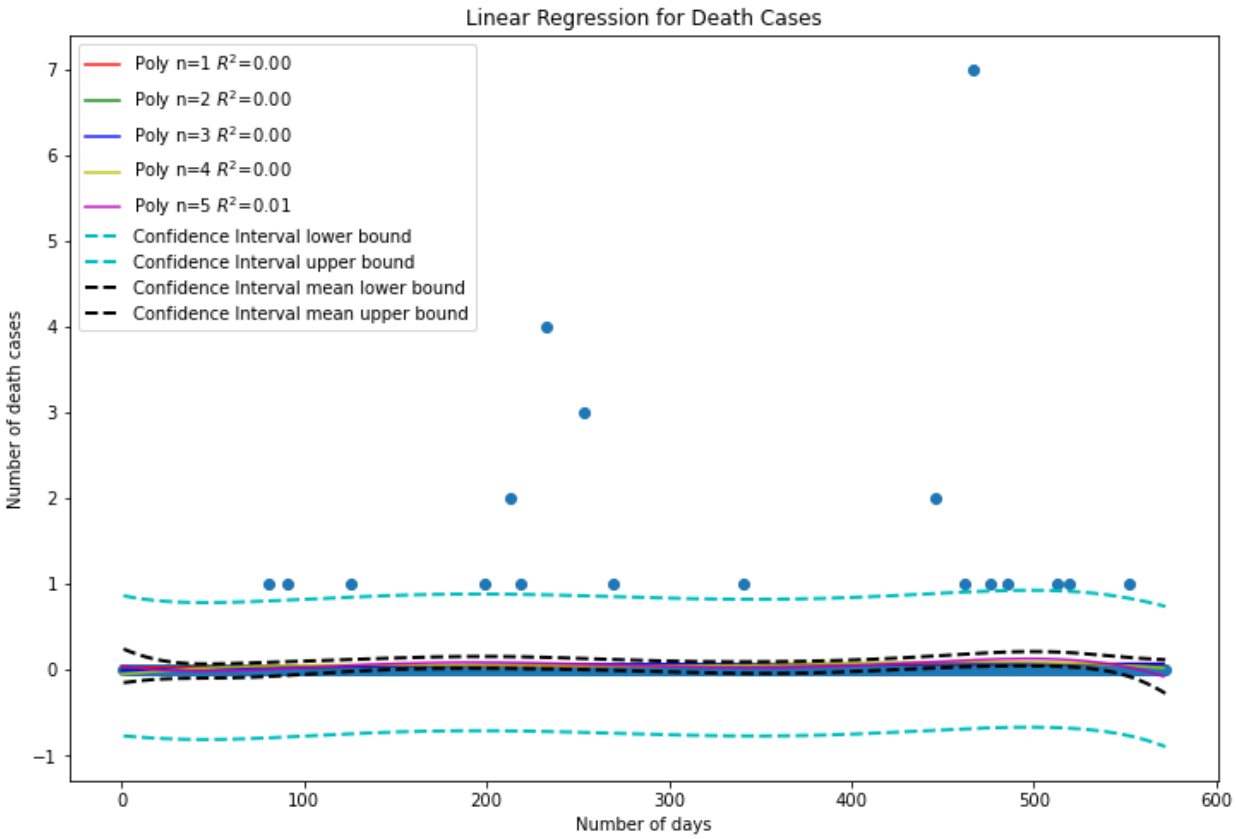


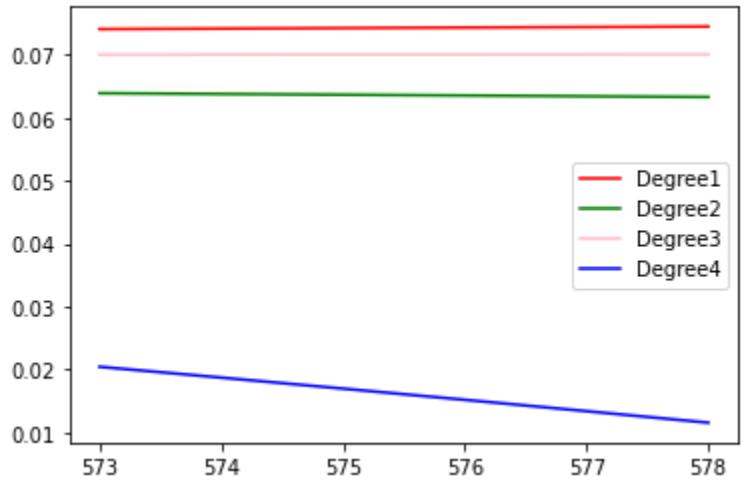
Northampton county



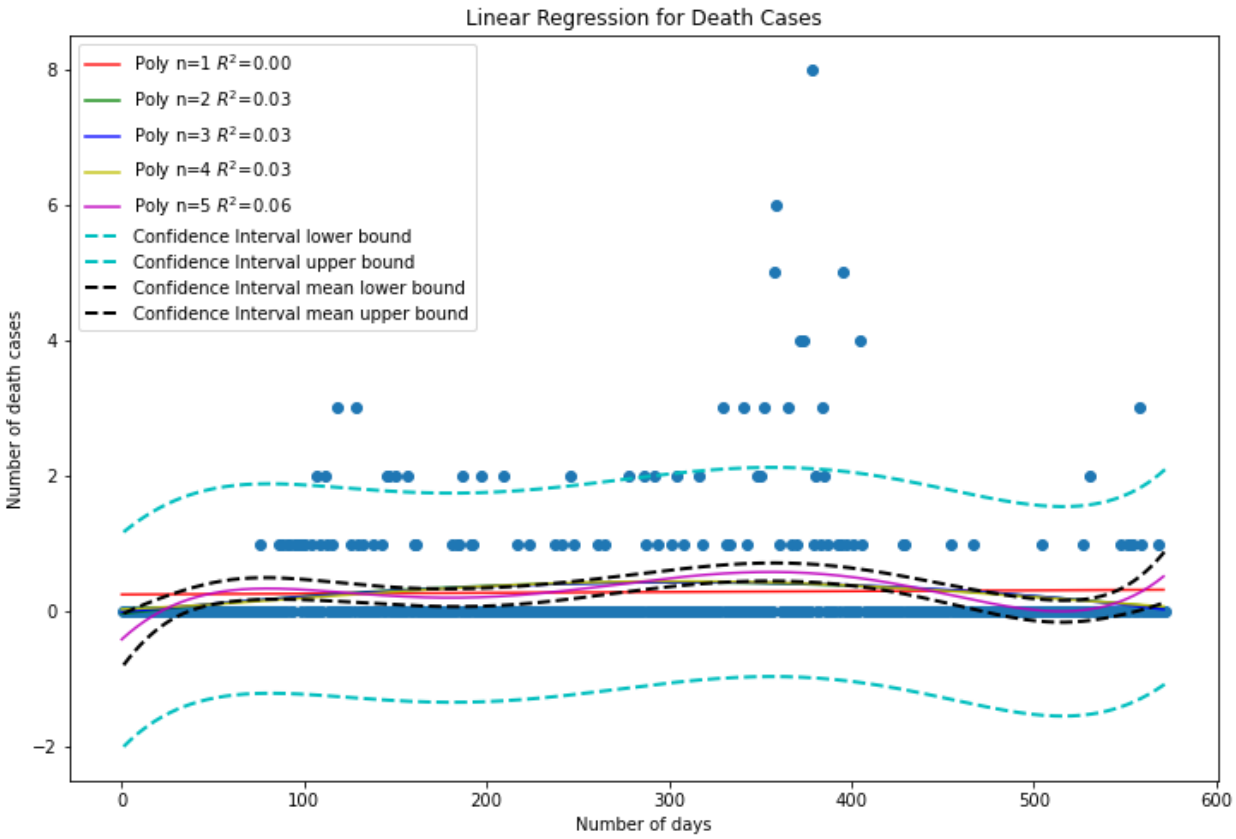


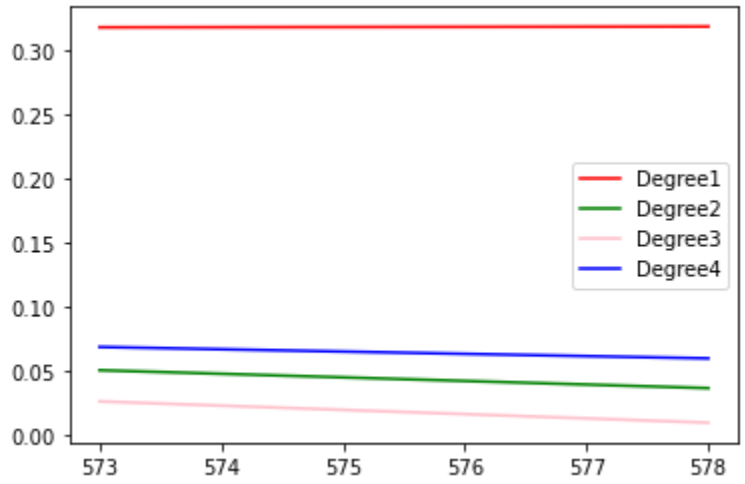
Jones county





Columbus county

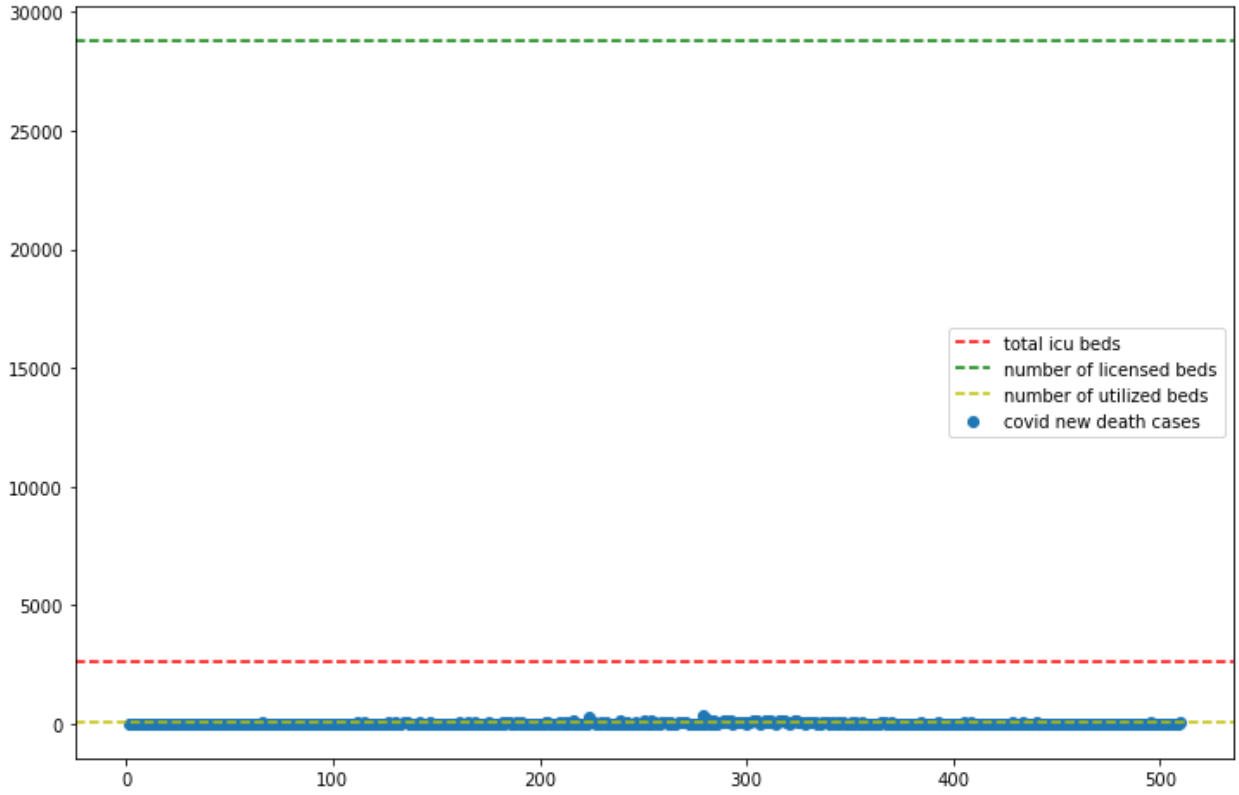


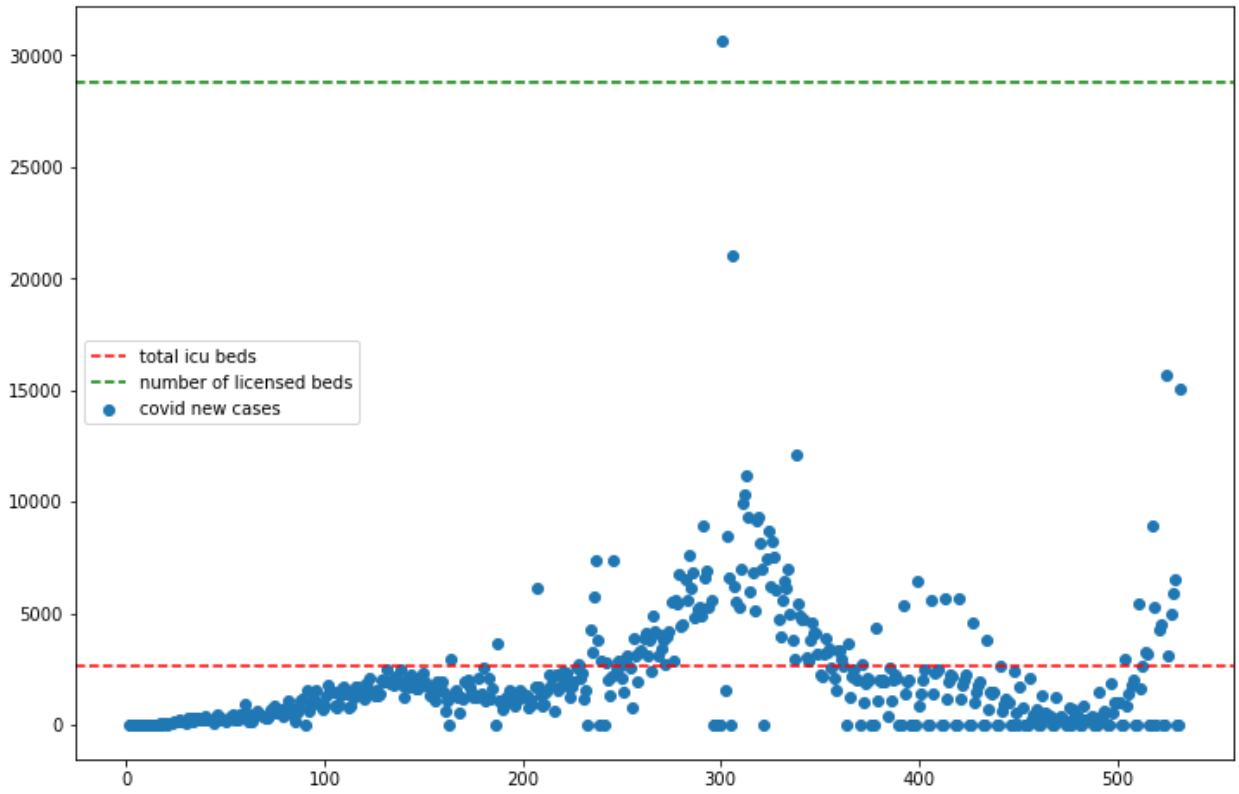


Confidence intervals

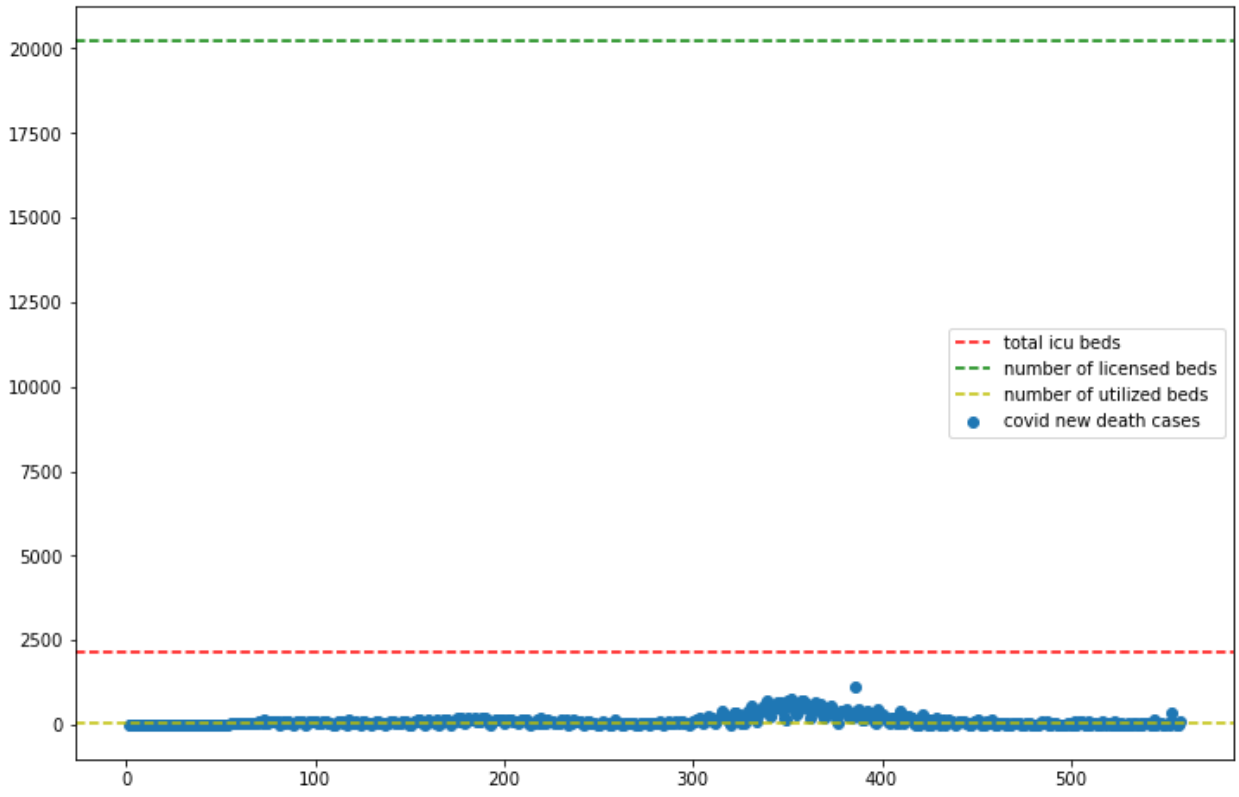
```
Robeson County Confidence interval                    0          1
Intercept    5.817222  22.991640
num_of_days  0.043001   0.094939
-------------------------------------------------------
Sampson county Confidence interval                    0          1
Intercept    3.876744  12.577275
num_of_days  0.012901   0.039213
-------------------------------------------------------
Hyde County Confidence interval                     0         1
Intercept   -0.030564  1.349707
num_of_days -0.000179  0.003995
-------------------------------------------------------
Stanly County Confidence interval                     0          1
Intercept    3.182699  10.194997
num_of_days  0.019236   0.040442
-------------------------------------------------------
Columbus County Confidence interval                     0         1
Intercept    0.481387  7.778775
num_of_days  0.021518  0.043586
```

Hospital bed - Utilize the hospital data to calculate the point of no return for a state. Use percentage occupancy / utilization to see which states are close and what their trend looks like


30000
25000
20000
15000
10000
5000
0
0
100
200
300
400
500
total icu beds
number of licensed beds
number of utilized beds
covid new death cases


30000
25000
20000
15000
10000
5000
0
0
100
200
300
400
500
total icu beds
number of licensed beds
covid new cases

Calculating California beds and cases to compare with North Carolina



Hypothesis Testing:

Hypothesis Testing from Stage 2

Does the total population of Male affect the number of covid cases?

Null Hypothesis:- There is no change in the covid cases wrt total male population

Alternative Hypothesis:- There is a change in the covid cases wrt total male population

Answer :

The pvalue is 0.00023. If we are using 95% confidence intervals, our null hypothesis can be rejected. Meaning, the male population impact the total number of covid cases . since the p-value is less than the corresponding significance level of 5%, we can reject the null hypothesis

Hypothesis Testing2 - Are the new cases and age group people between 45 to 54 years remaining the same?

Null Hypothesis - There is no difference between number of newcases and total population between ages of 45 to 54 years

Alternate Hypothesis - There is a change between number of newcases and total population between ages of 45 to 54 years

Answer:

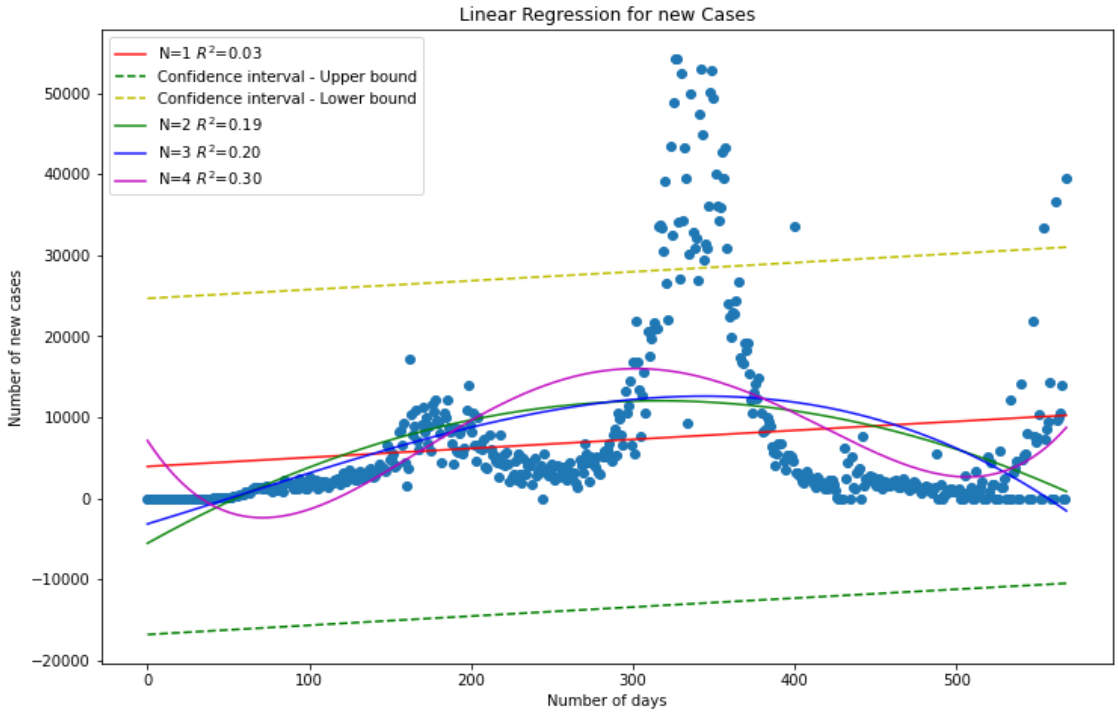
The p-value is 0.408. With the confidence interval being 95%, the p-value is 40.8% which is way more than 5% threshold. Hence we fail to reject the null hypothesis. These results tell us that people over 45 years are more prone to get infected. Also, there is a 40% chance we'd see sample data this far apart if the two groups tested are actually identical.

Poojitha Kalidindi

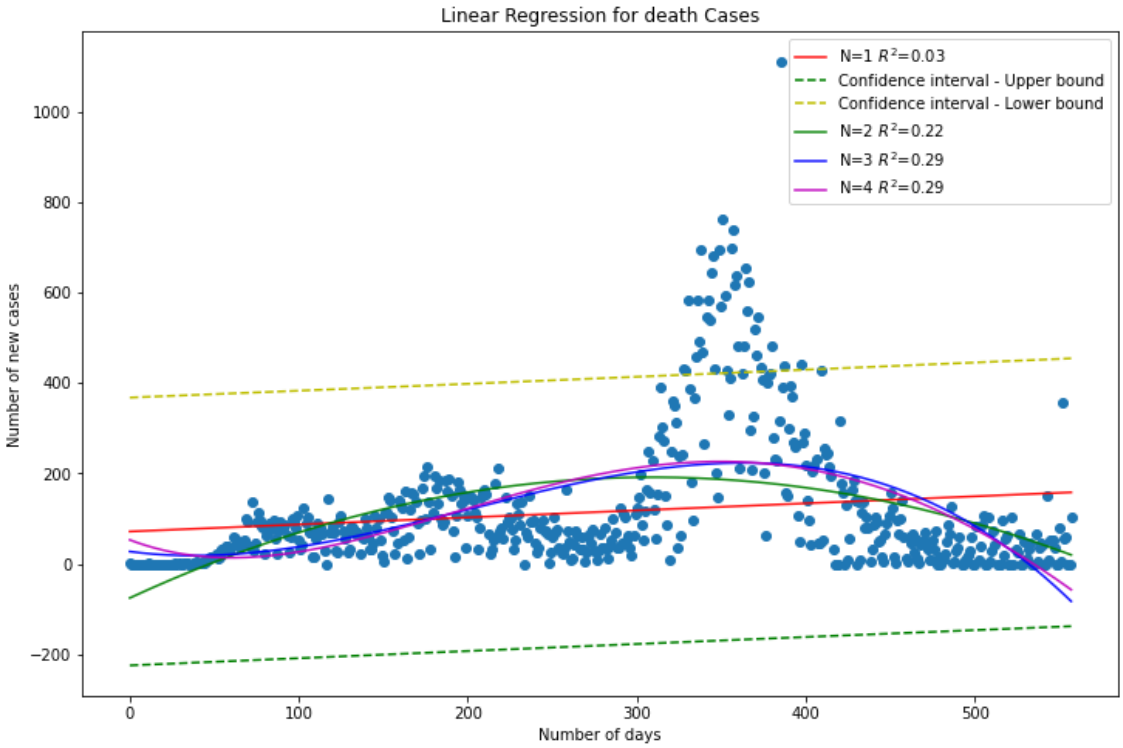
Stage - 3 Member Task

To implement Linear and non-linear regression, I have selected California state and its counties with highest covid cases. Confidence intervals for linear regression are also plotted on the same graph.

California State New Cases:

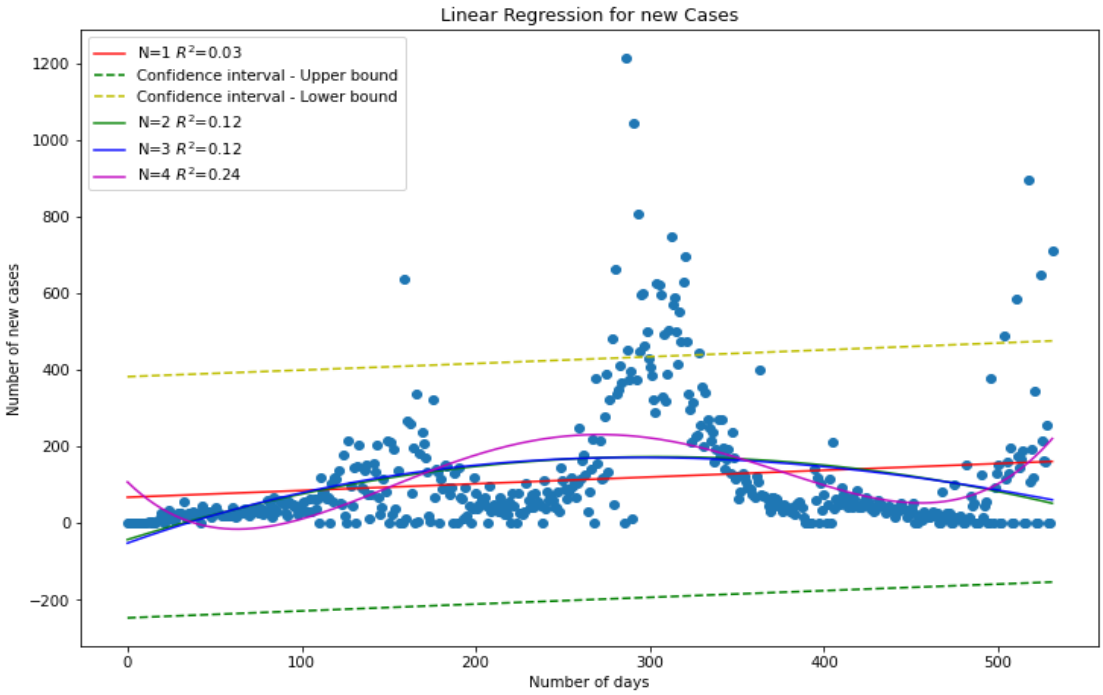


California State New Deaths:

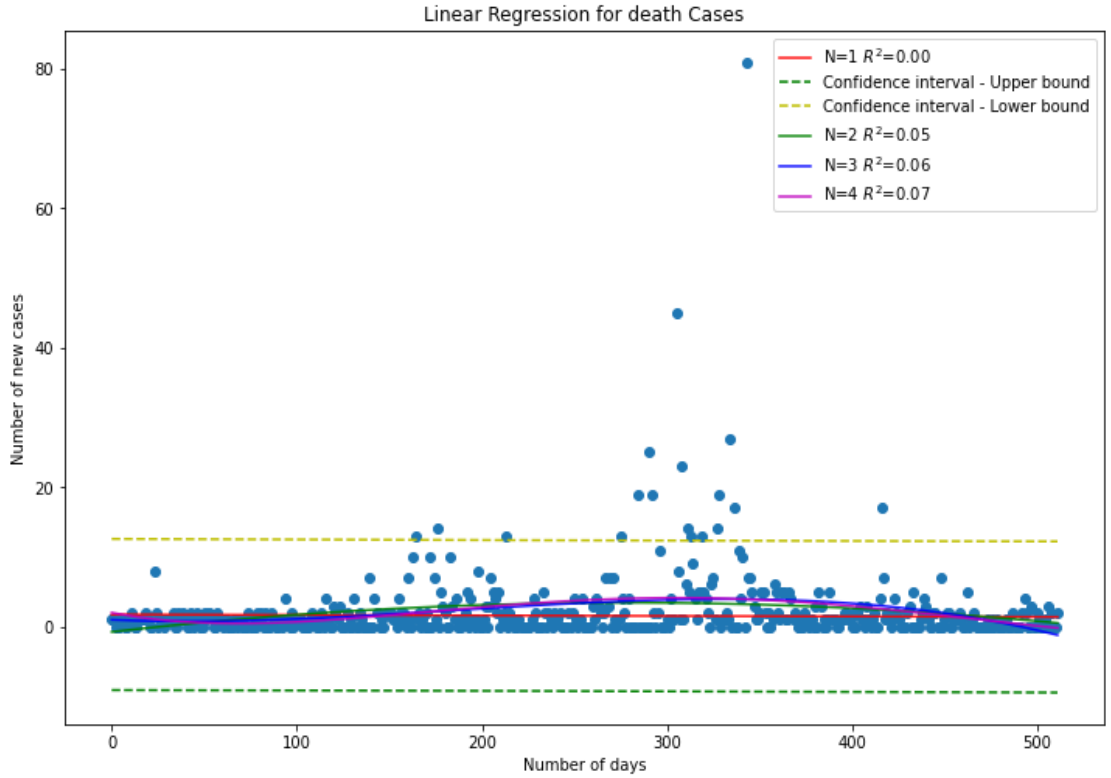


Counties:

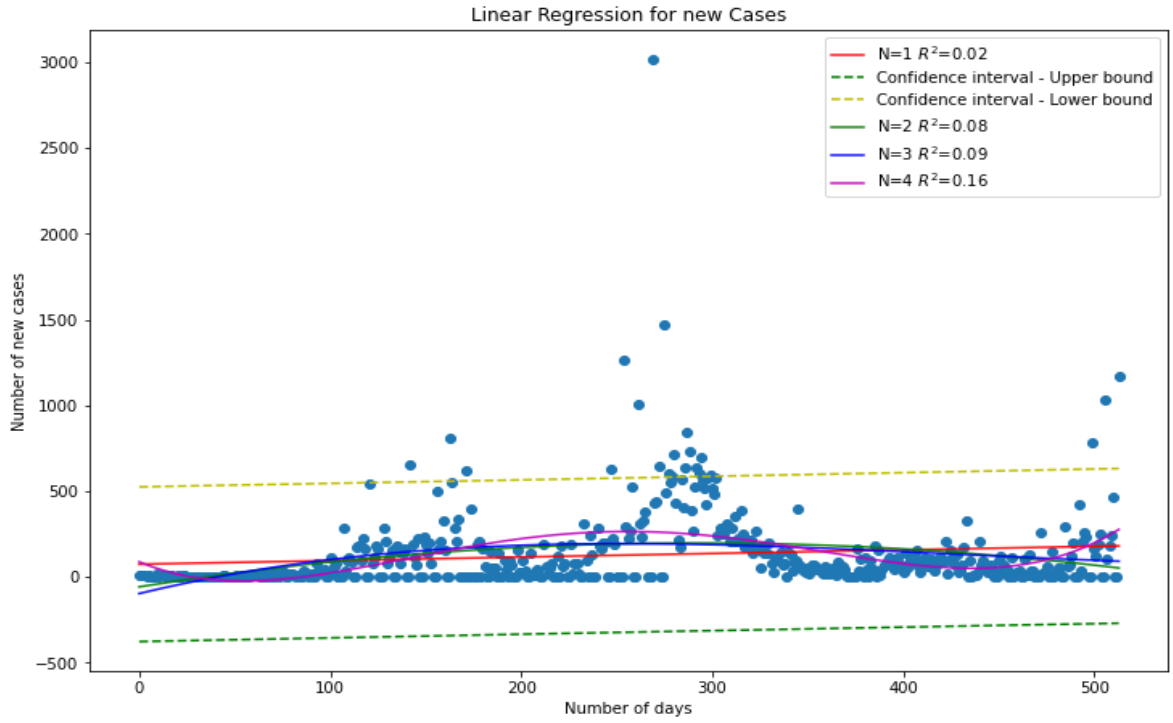
Alameda County New Cases:



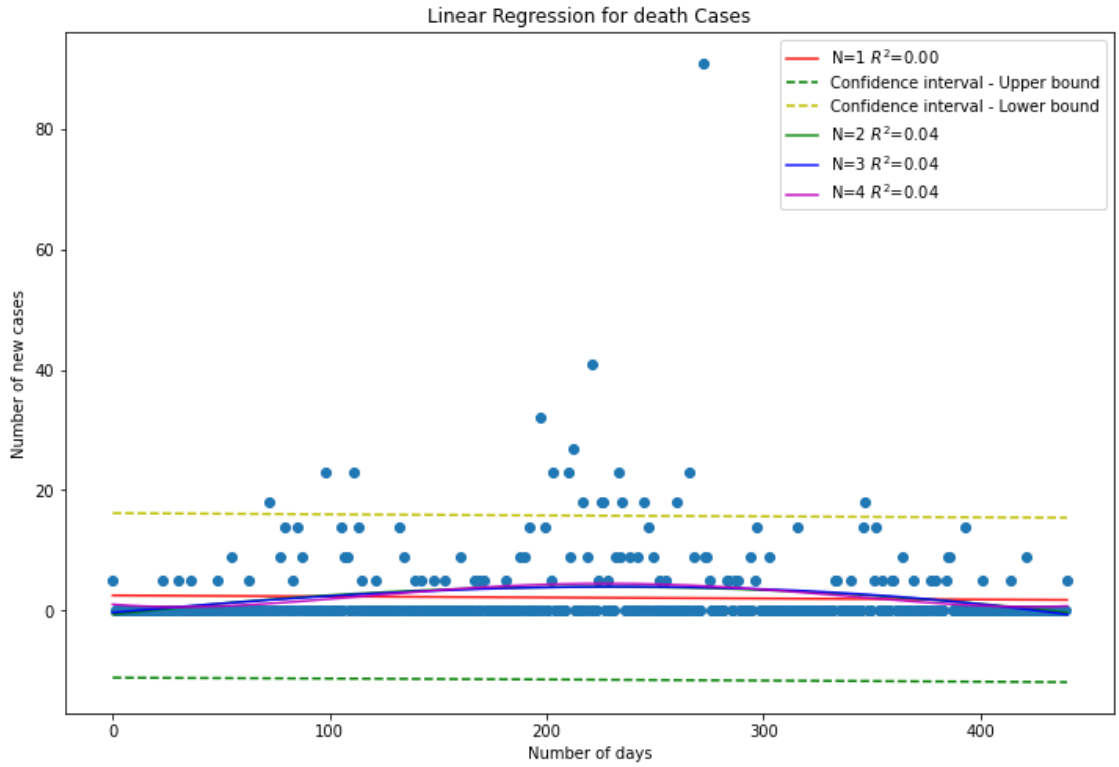
Alameda County New Deaths:



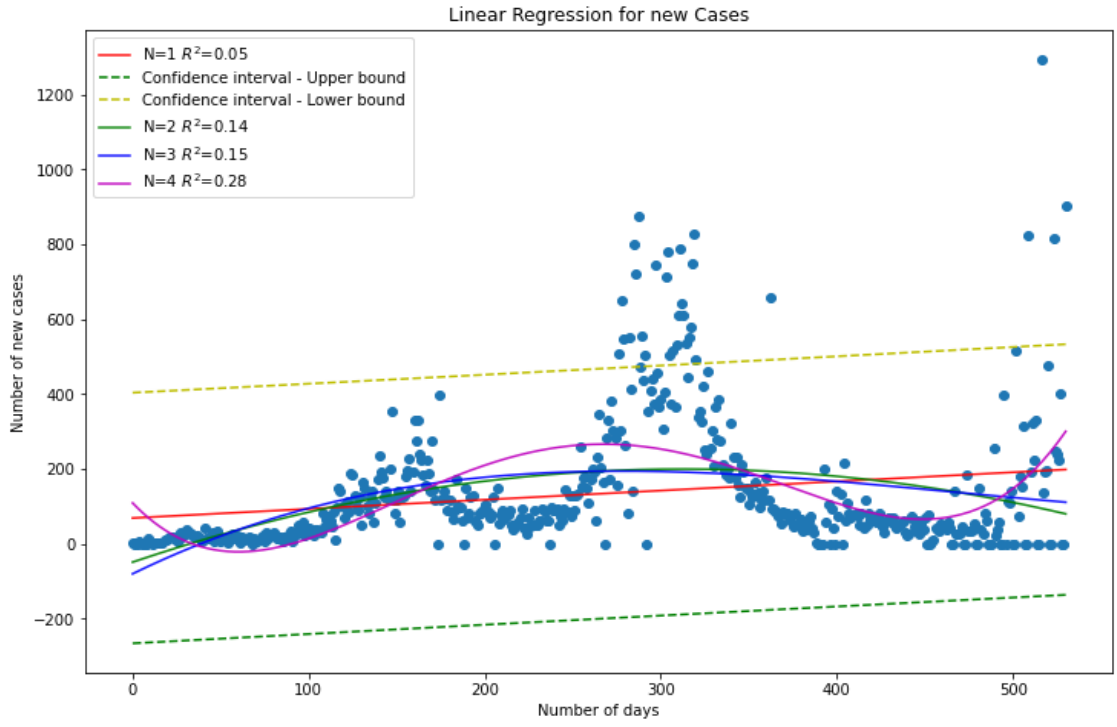
Butte County New Cases:



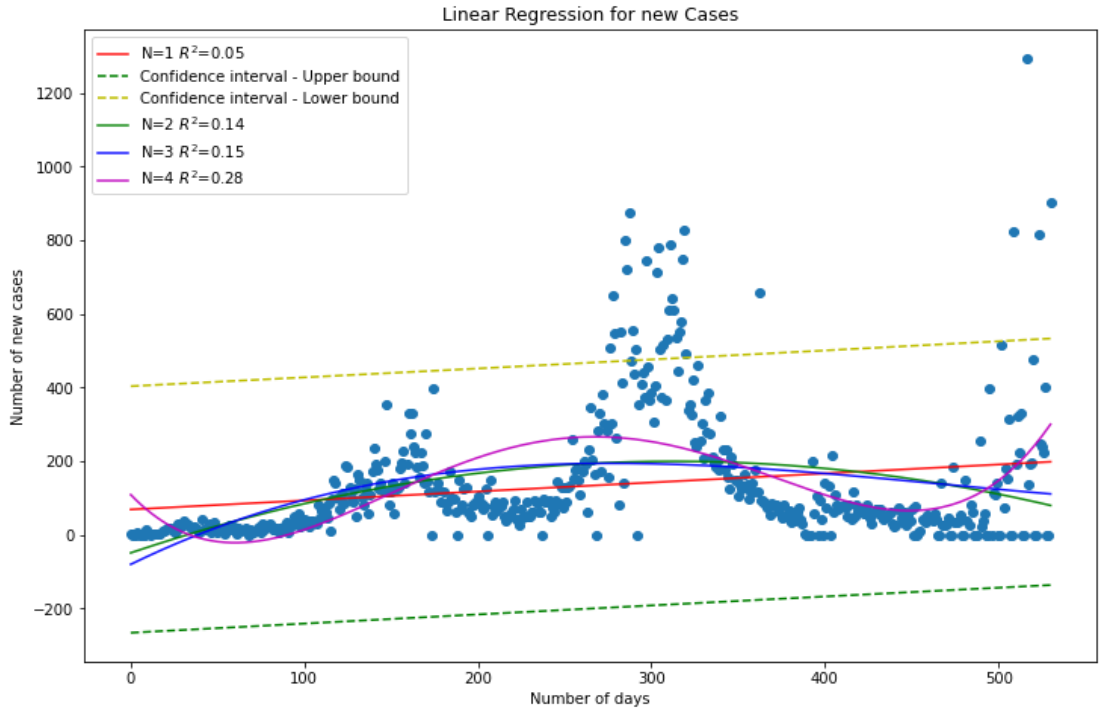
Butte County New Deaths:



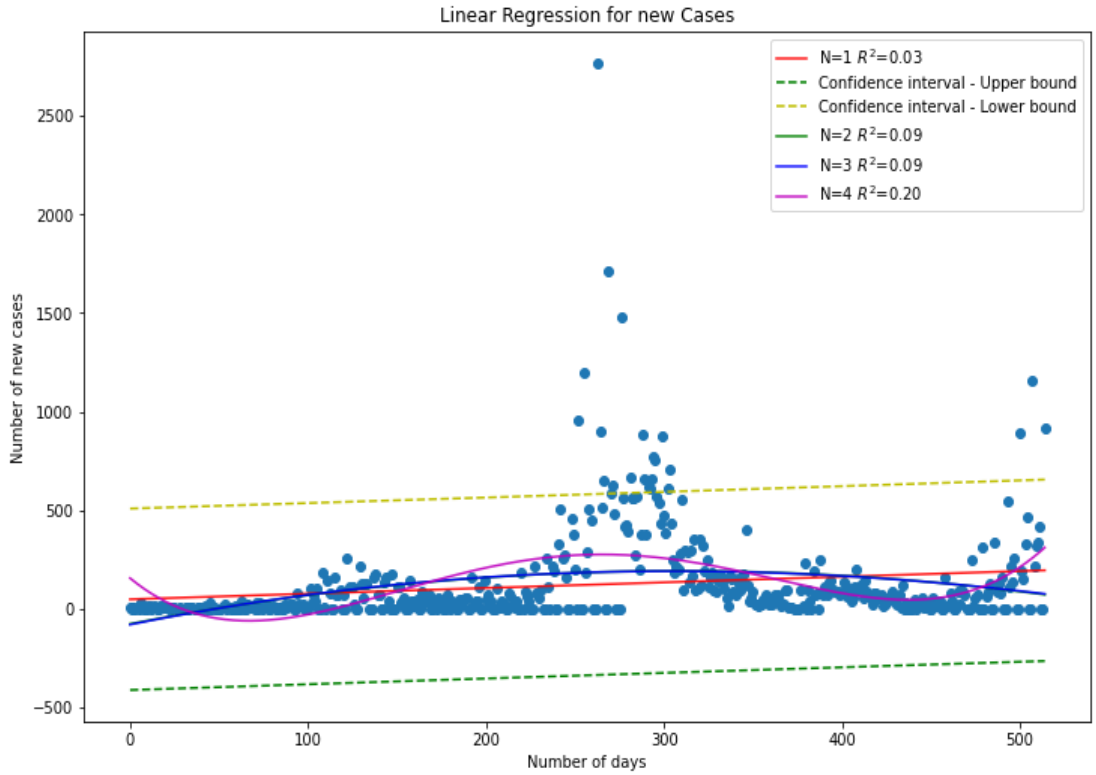
Contra Costa County New Cases:



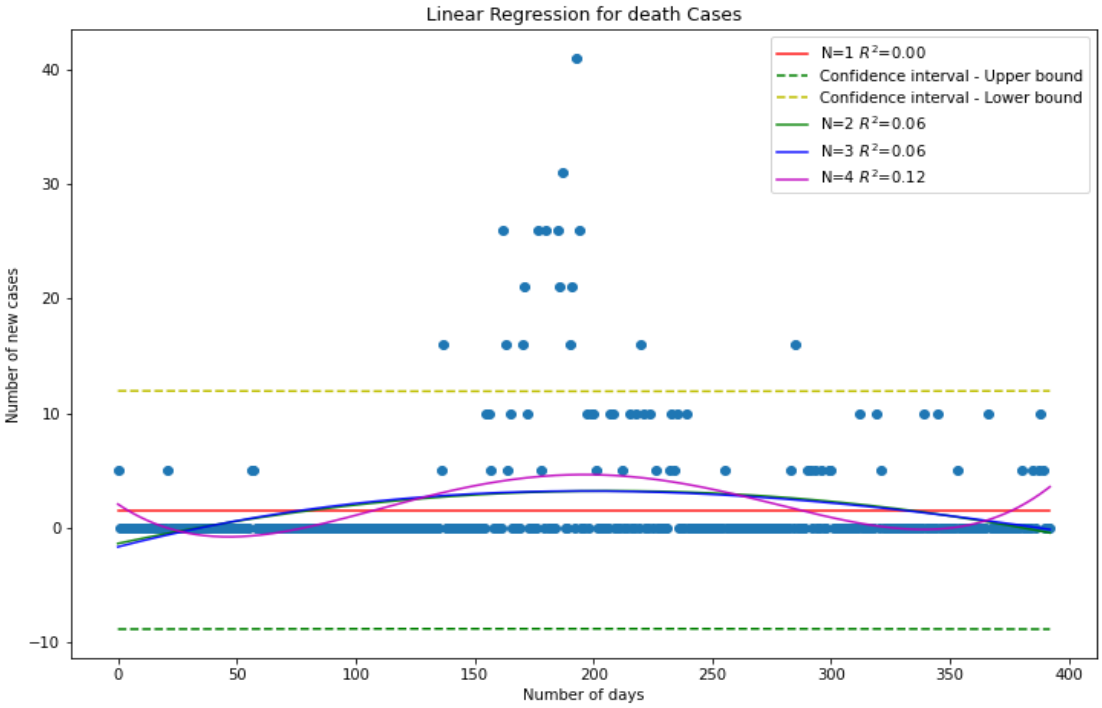
Contra Costa County New Deaths:



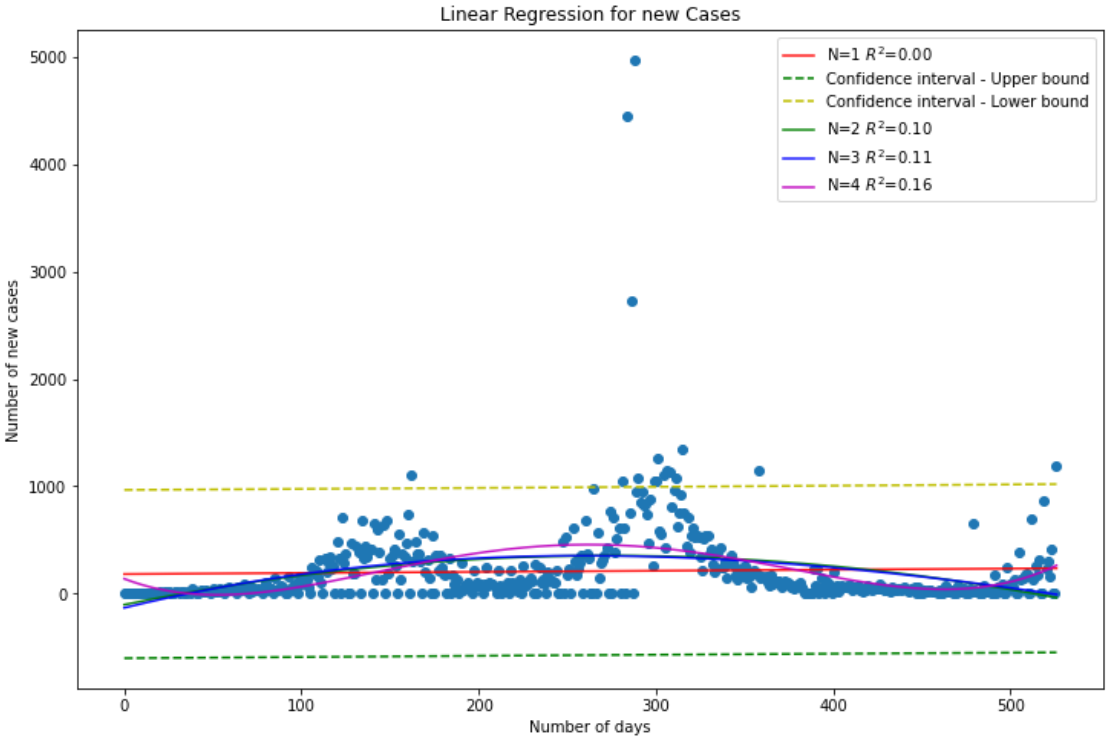
El Dorado County New Cases:



El Dorado County New Deaths:



Fresno County New Cases:



Fresno County New Deaths:

Hypothesis Testing is performed for the hypotheses that are made in stage 2:

Hypothesis 1 - States with high voting population have higher cases

Florida and Texas were picked for the test - Florida has 12% higher voting than Texas (both have large populations)

Null Hypothesis - The mean cases in Florida are lower than or the same as mean cases in Texas

Alternate Hypothesis - The mean cases in Florida are higher than cases in Florida

Performed a 1-tail test using an alpha value of 0.2 (Picked a higher value as the results do not need absolute certainty)

P Value = 0.86, t-statistic > 0

P Value/2 > alpha value - This means that we cannot reject null Hypothesis.

Hypothesis 2 - States with high dem population have lower cases

Massachusetts and Tennessee were picked for the test - Massachusetts has 2x democratic voter percentage compared to Tennessee

Null Hypothesis - The mean cases in Massachusetts are greater than or same as the mean cases in Tennessee.

Alternate Hypothesis - The mean cases in Massachusetts are lower than cases in Tennessee

Performed a 1-tail test using an alpha value of 0.2 (Picked a higher value as the results do not need absolute certainty)

P Value = 0.003, t-statistic < 0

P Value/2 < alpha value and t-statistic < 0. This means we can reject the null hypothesis also say that mean cases are lower in MA than TN.

Hypothesis 3 - States with high republican population have higher cases

KY and MD were picked for the test - KY has 2x democratic voter percentage compared to MD

Null Hypothesis - The mean cases in KY are lower than or same as the mean cases in MD.

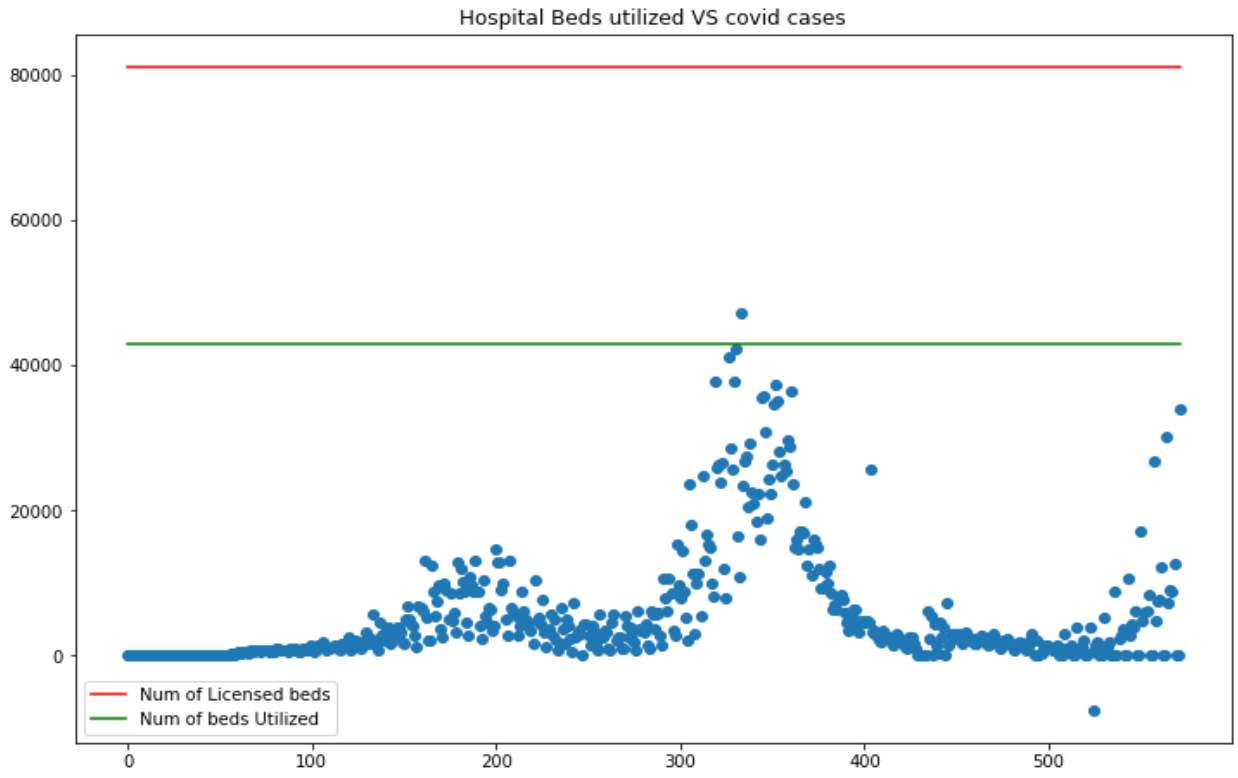
Alternate Hypothesis - The mean cases in KY are higher than cases in MD

Performed a 1-tail test using an alpha value of 0.2 (Picked a higher value as the results do not need absolute certainty)

P Value = 9.66e^-0.8, t-statistic > 0

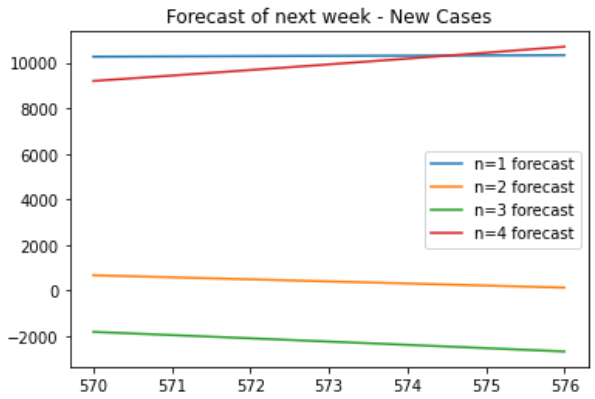
P Value/2 < alpha value and t-statistic > 0. This means we can reject the null hypothesis and also say that mean cases are greater in KY than MD.

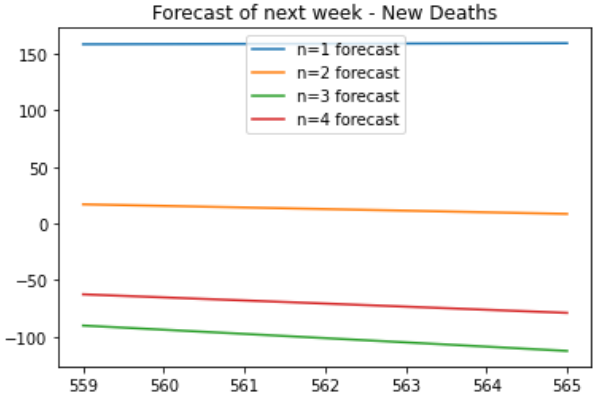
Hospital Bed Occupancy vs No of Covid Deaths:



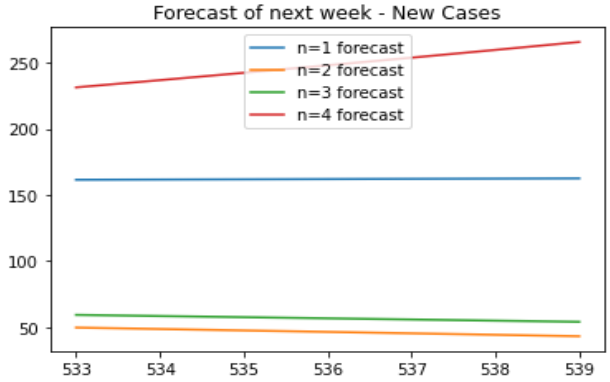
Forecast of prediction path for the next one week

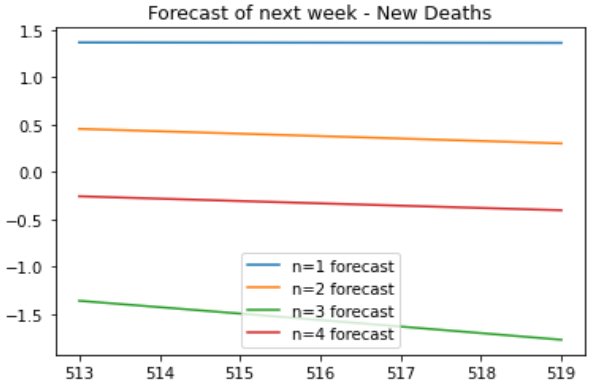
California State New cases and Deaths:



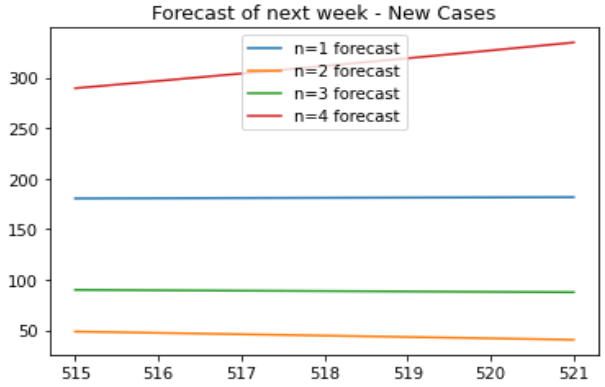


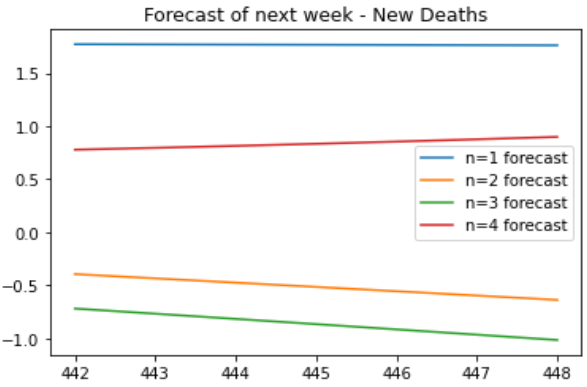
Alameda County New cases and deaths:



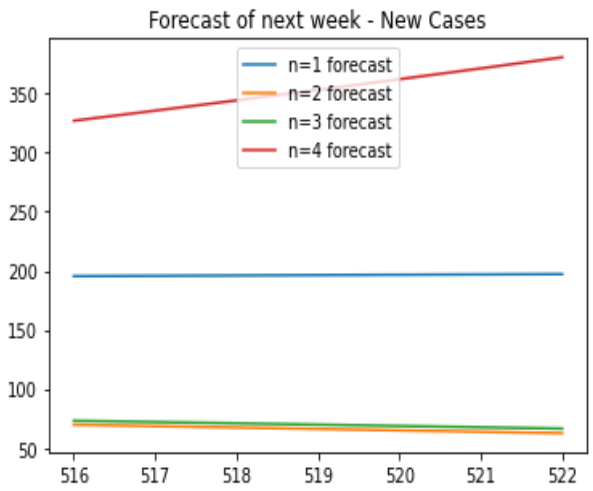


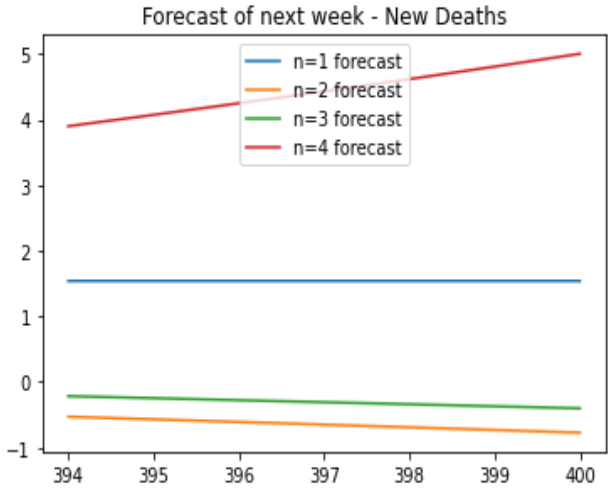
Butte County New cases and deaths:



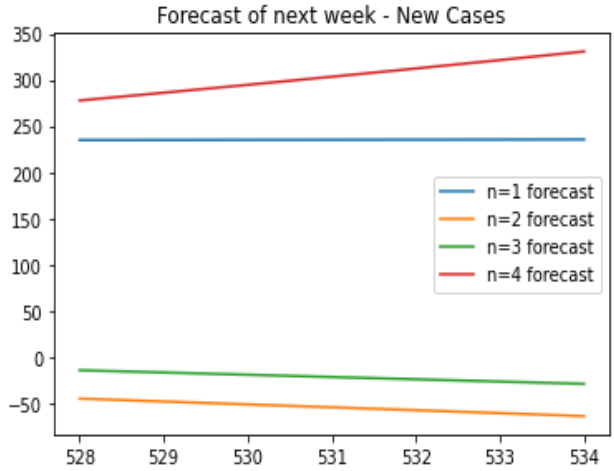


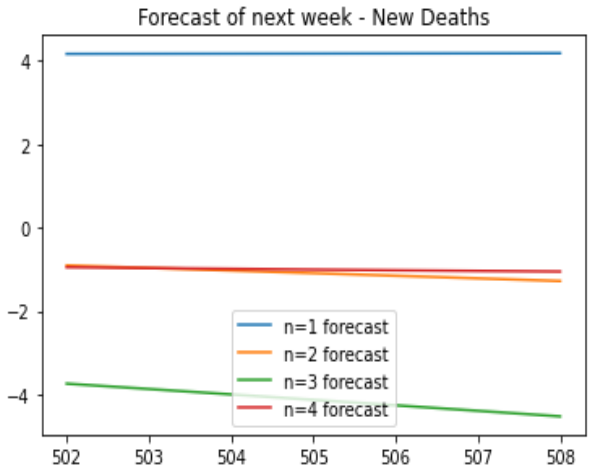
El Dorado County New cases and deaths:



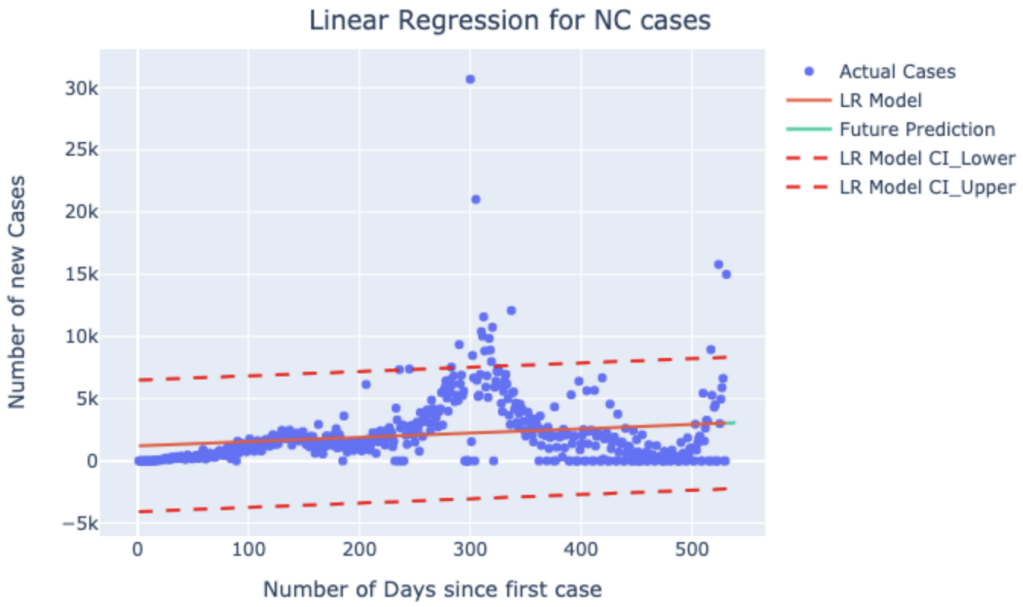


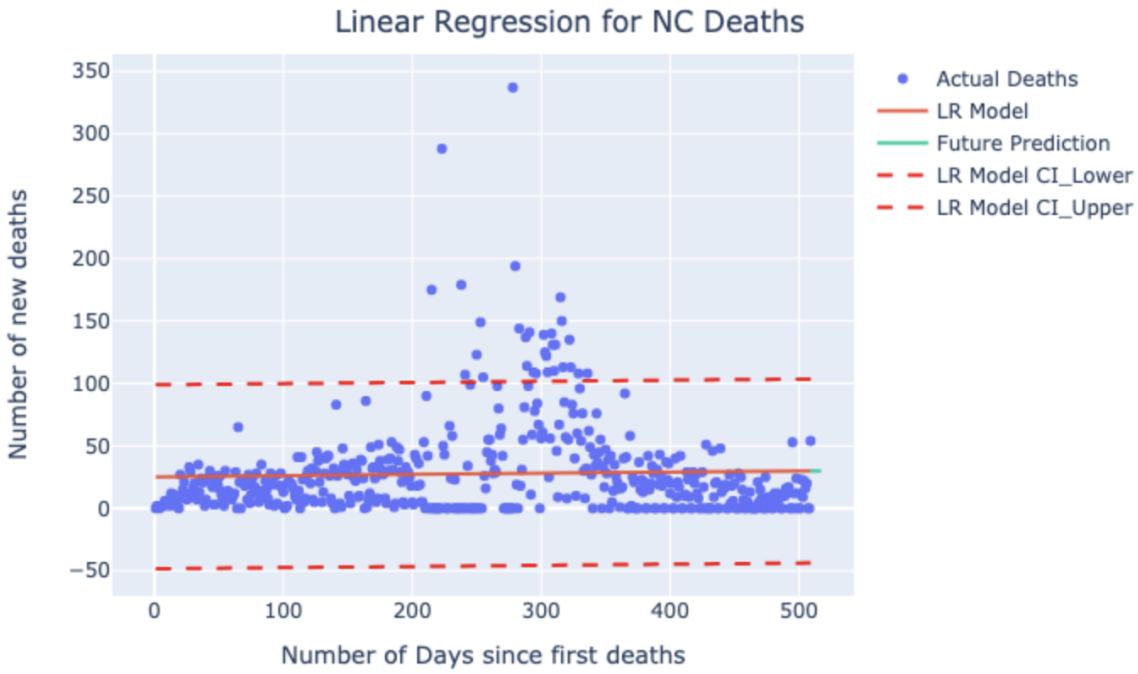
Fresno County New cases and deaths:

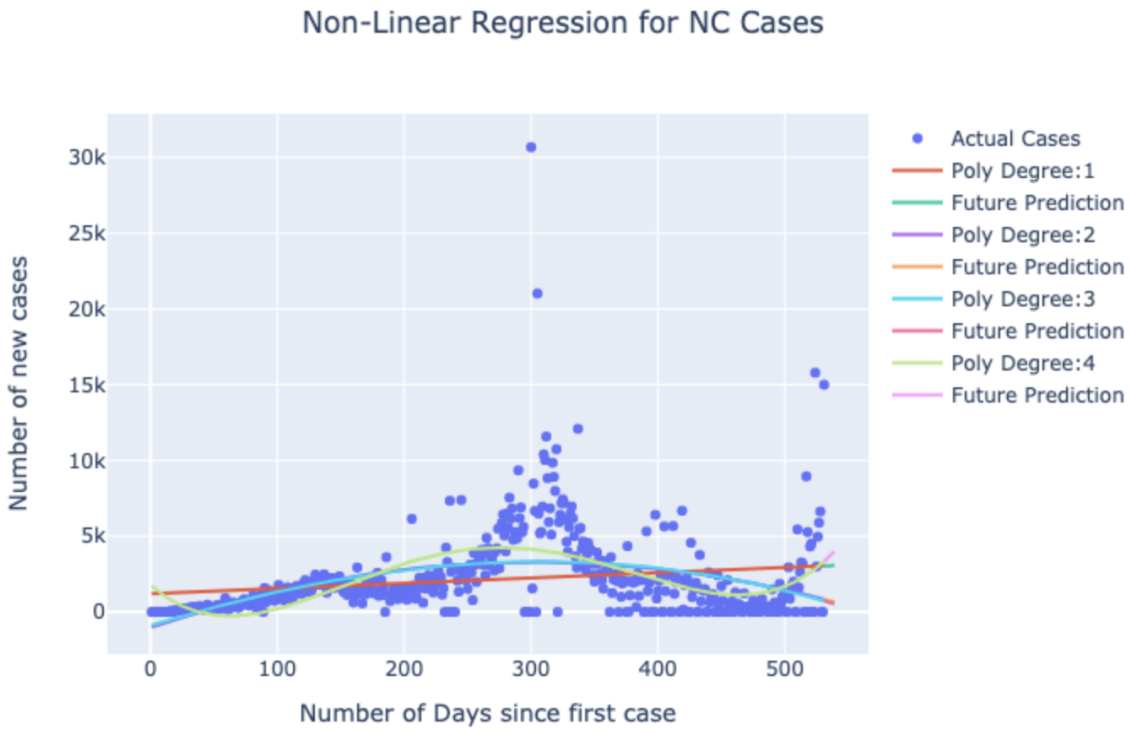




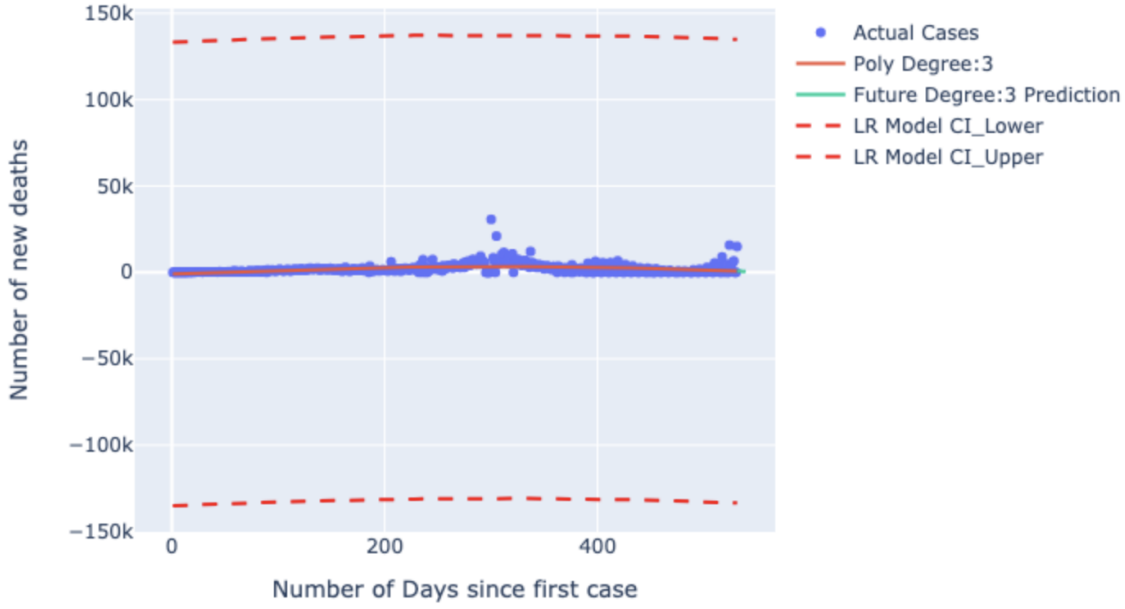
SHIPRA SHANU
Stage-3 Member Task



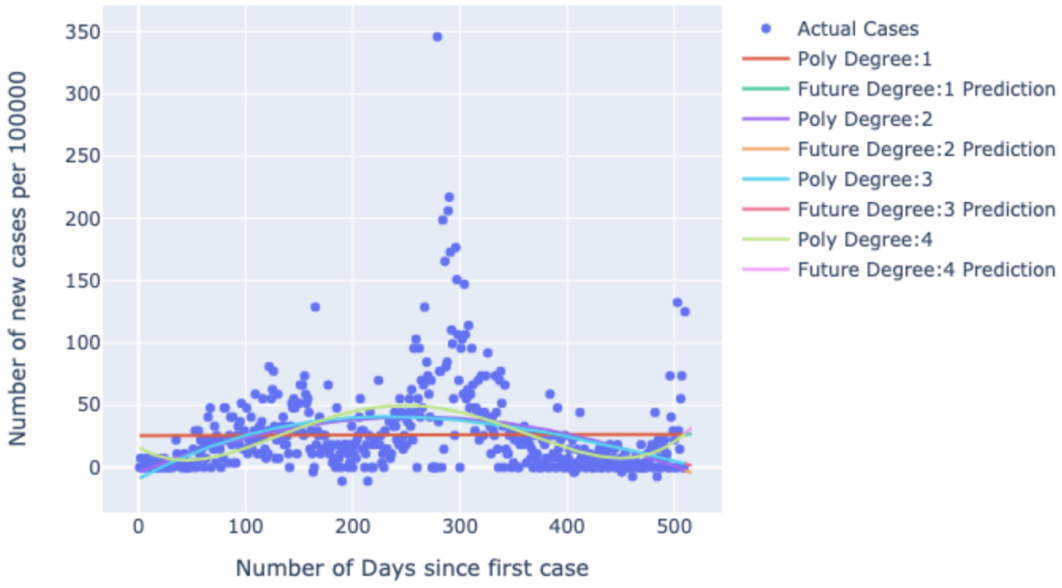




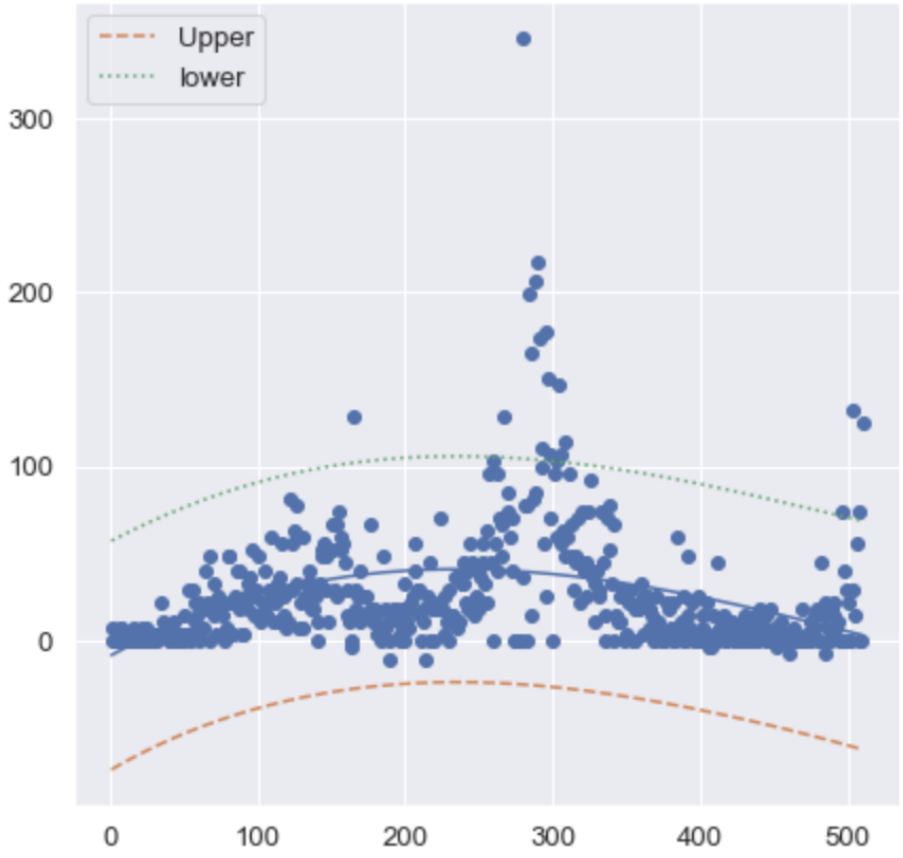
Non- Linear Regression for NC Deaths



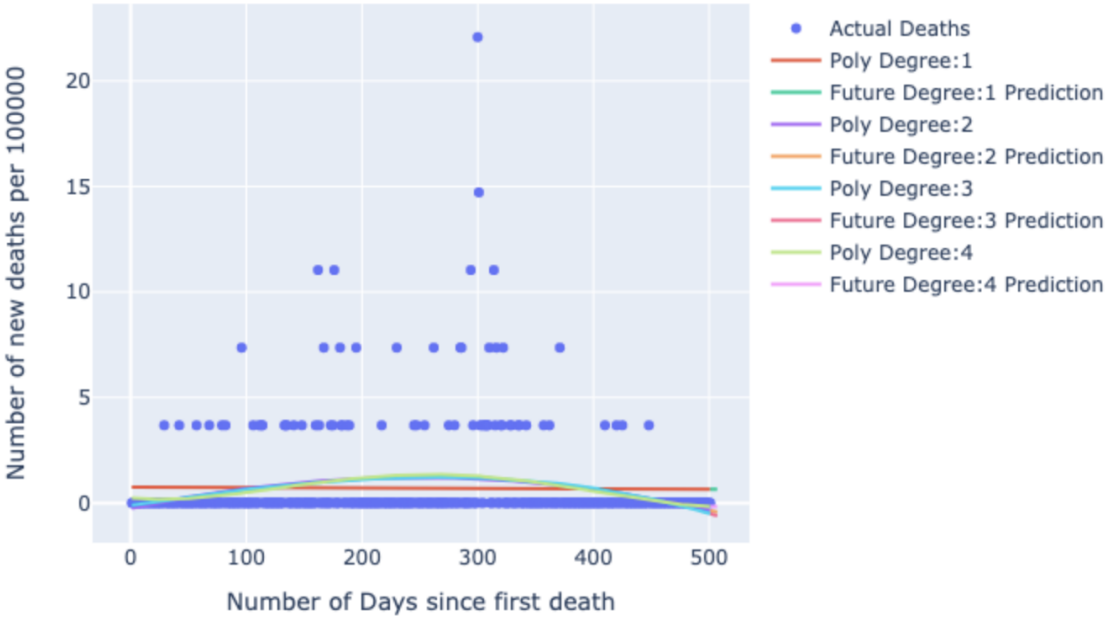
Non-Linear Regression for NC montgomery_county Cases

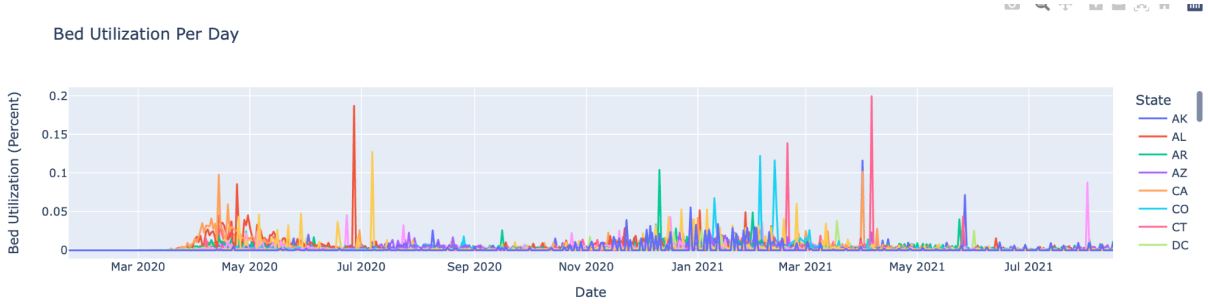


Confidence interval of Motgomery_county



Non-Linear Regression for montgomery_county Deaths

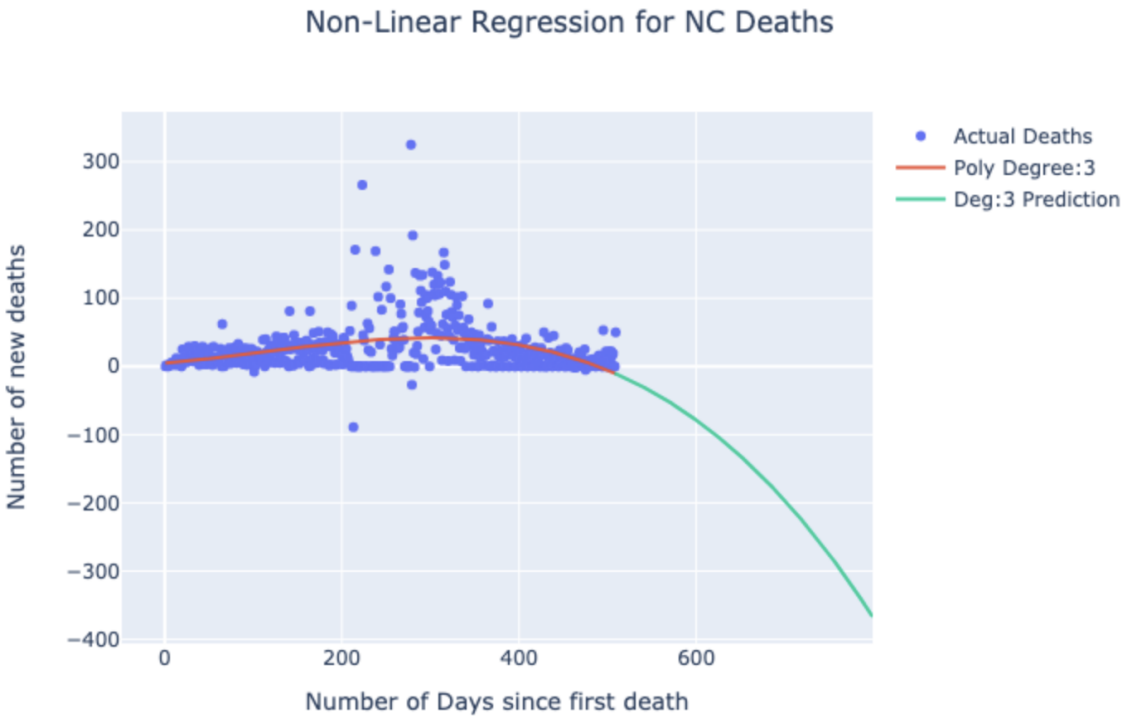




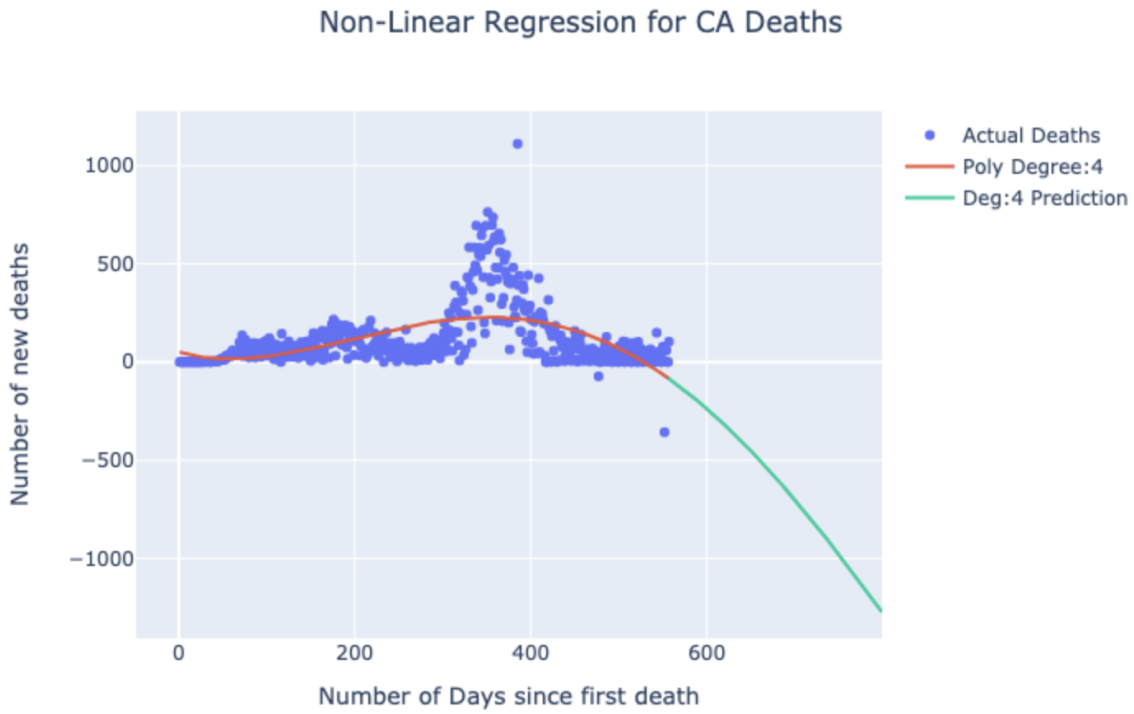

```
#NC Prediction
state="NC"
fig_name = nonLinearModelStateDeaths("NC",[3],True)

Image(filename=fig_name)
```

```
MAE for degree = 3 is  21.000538994984243
MSE for degree = 3 is  1129.2445210997323
RMSE for degree = 3 is  33.60423367820984
R^2 for degree = 3 is  0.1399553120424919
---------------------------------------------------------
Maximum number of ICU beds in NC  is  2657
```



```
MAE for degree = 4 is  87.84527482942609
MSE for degree = 4 is  16102.717061326312
RMSE for degree = 4 is  126.8964816743408
R^2 for degree = 4 is  0.30314221927255425
---------------------------------------------------------
Maximum number of ICU beds in CA  is  8719
```



Hypothesis
* Are married couple a factor for higher covid cases
* Does being married influence increase in no of deaths

[10]:

	State_x	Num of Cases	Num of Deaths	Married-couple family	Married-couple family with their children under 18 years	Cohabiting couple household	Cohabiting couple household with their own children under 18 years	Male householder, no spouse/partner
0	GA	187605358	3491823	782437230	320478900	87301134	26771133	305164902
1	IL	319578048	6280842	1088812308	432028821	143126232	44206377	446156136
2	MI	194801262	5320521	874378518	319931112	128735910	39974199	372191577
3	NJ	262468281	9403883	955456299	390041673	101280042	33694119	303185760
4	OH	234994530	4249976	999431757	357986907	169538094	56237658	447893472
5	PA	268407676	7232056	1247655930	434440005	188555964	57974994	499831911

[11]:
```
stats.ttest_ind(a=analysis_data_grp['Married-couple family'], b= analysis_data_grp['Num of Cases'],equal_var=False)
```

[11]:
```
Ttest_indResult(statistic=10.696548065945358, pvalue=4.3356645396679225e-05)
```

In this case, the p-value is greater than our significance level α (equal to 1-conf.level or 0.05) so, we would fail to reject the null hypothesis.

[14]:
```
stats.ttest_ind(a=analysis_data_grp['Num of Cases'], b= analysis_data_grp['Married-couple family with their children under 18 years'],equal_var=False)
```

[14]:
```
Ttest_indResult(statistic=-4.487803929856863, pvalue=0.0011690581987126702)
```

In this case, the p-value is lower than our significance level α (equal to 1-conf.level or 0.05) so, we should reject the null hypothesis.

[15]:
```
stats.ttest_ind(a=analysis_data_grp['Married-couple family'], b= analysis_data_grp['Num of Deaths'],equal_var=False)
```

[15]:
```
Ttest_indResult(statistic=14.74943936676968, pvalue=2.583178880046193e-05)
```

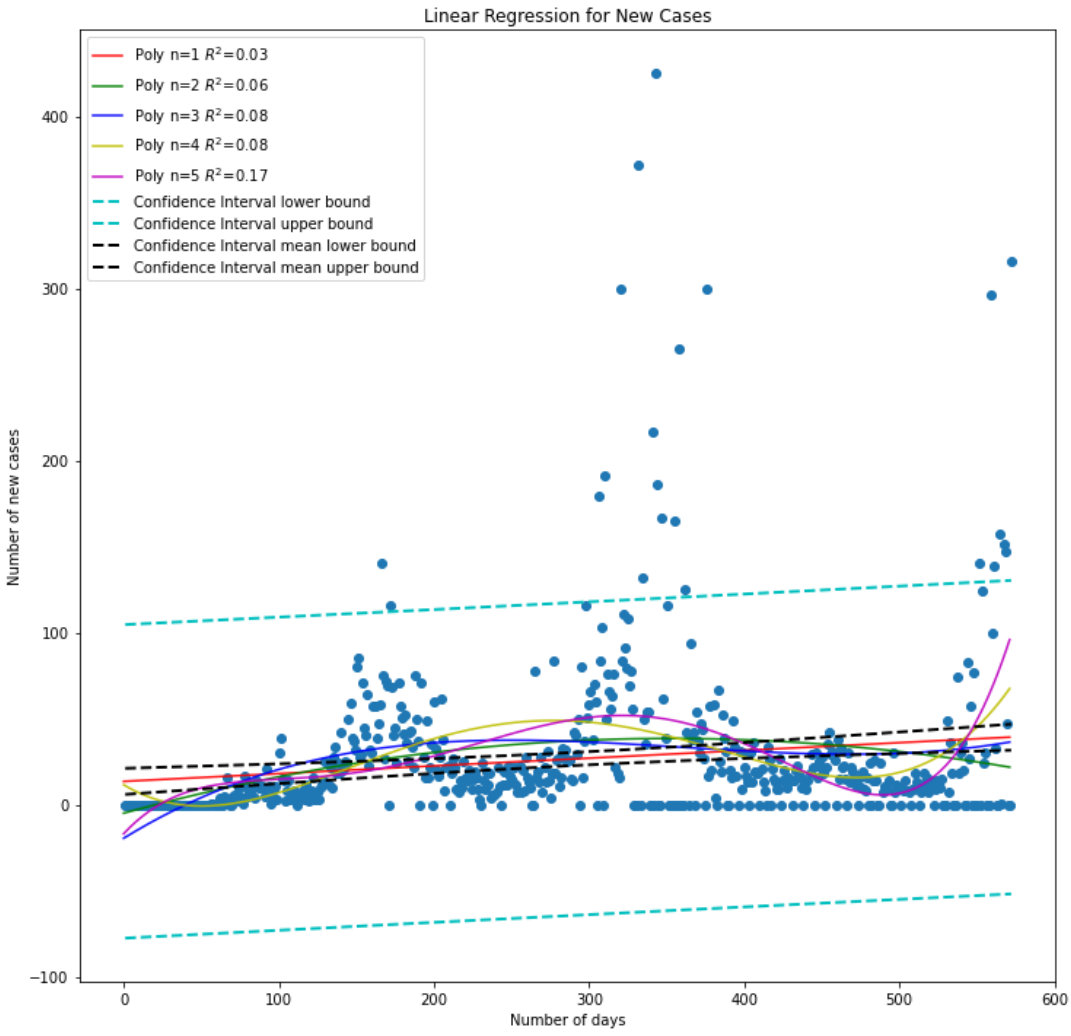
[ ]:
```
In this case, the p-value is greater than our significance level α (equal to 1-conf.level or 0.05) so,
we would fail to reject the null hypothesis.
```

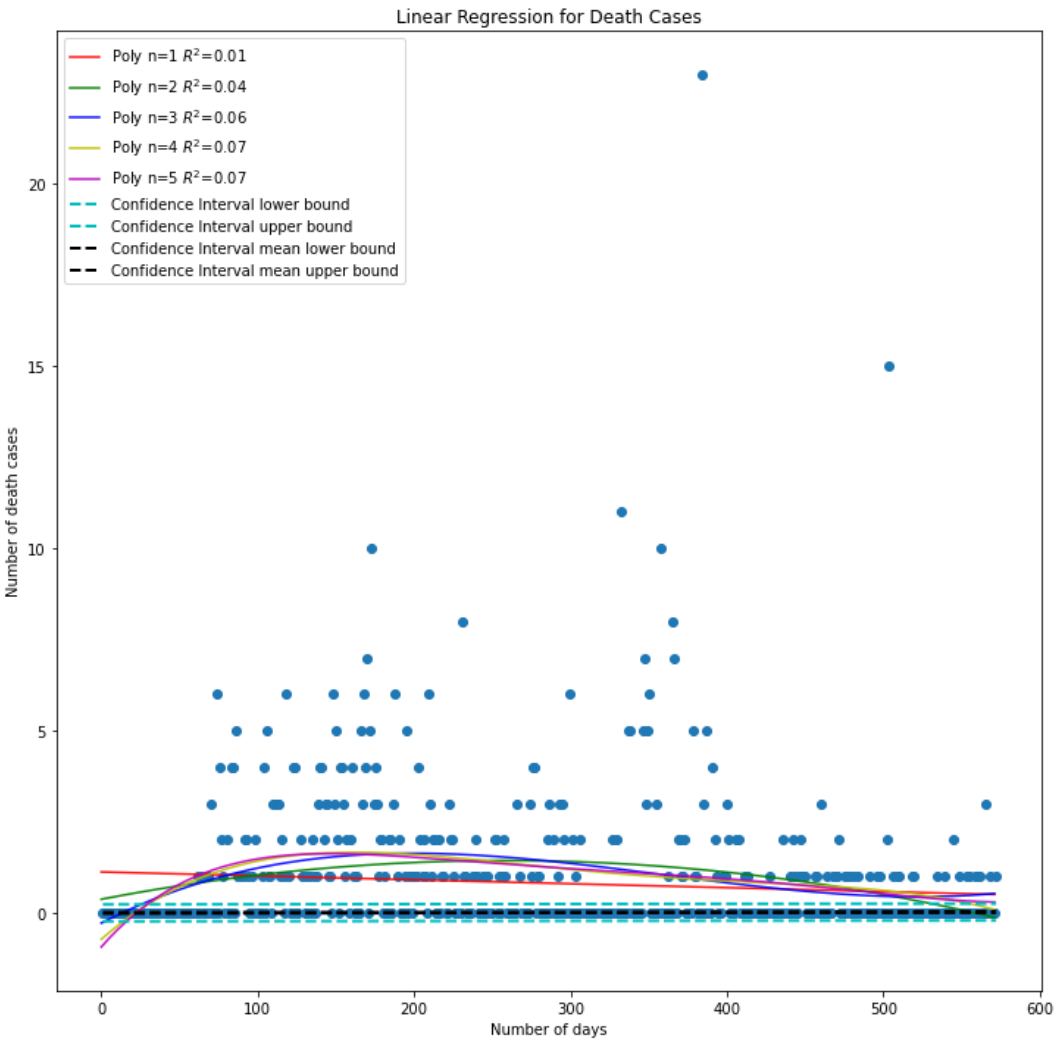
Heng's member task

I did WA state for all my graphs

Here's one example of a county graph for deaths and cases

This is Yakama County in WA state for deaths and cases


Linear Regression for New Cases
Poly n=1 R^2=0.03
Poly n=2 R^2=0.06
Poly n=3 R^2=0.08
Poly n=4 R^2=0.08
Poly n=5 R^2=0.17
Confidence Interval lower bound
Confidence Interval upper bound
Confidence Interval mean lower bound
Confidence Interval mean upper bound
Number of new cases
Number of days
400
300
200
100
0
−100
0
100
200
300
400
500
600


Linear Regression for Death Cases
Poly n=1 R^2=0.01
Poly n=2 R^2=0.04
Poly n=3 R^2=0.06
Poly n=4 R^2=0.07
Poly n=5 R^2=0.07
Confidence Interval lower bound
Confidence Interval upper bound
Confidence Interval mean lower bound
Confidence Interval mean upper bound
Number of death cases
Number of days
0
5
10
15
20
0
100
200
300
400
500
600