STAGE-4

MEMBER TASK:

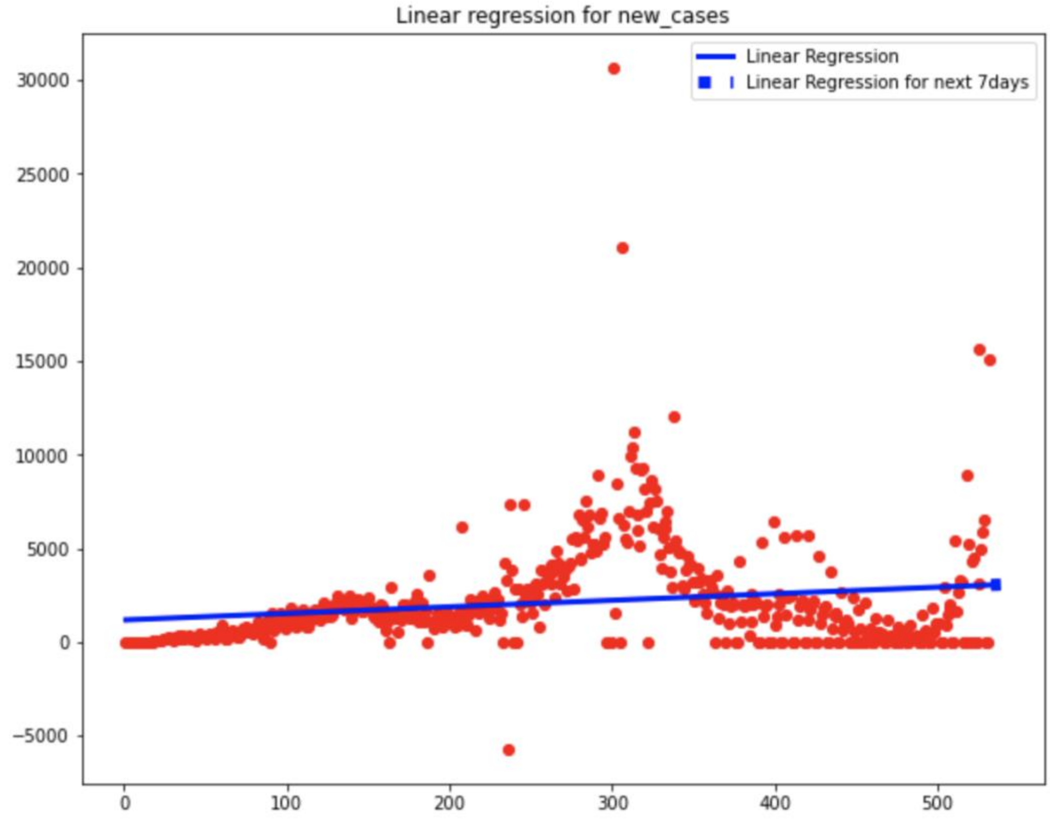
1) Utilize Linear and Non-Linear (polynomial) regression models to compare trends for a single state and its counties (top 5 with highest number of cases). Start your data from the first day of infections.

   The state that I chose to compare trends is North Carolina.

   I read the long_large_covid dataset which we generated in the stage-1. From that I separated the data of North Carolina state into a separate dataframe. Then I removed the rows whose countyFIPS is zero. Then I calculated the new cases and new deaths for each day by taking the difference of the rows. Then I grouped the whole new cases and new deaths by date and stored them into a separate dataframe.

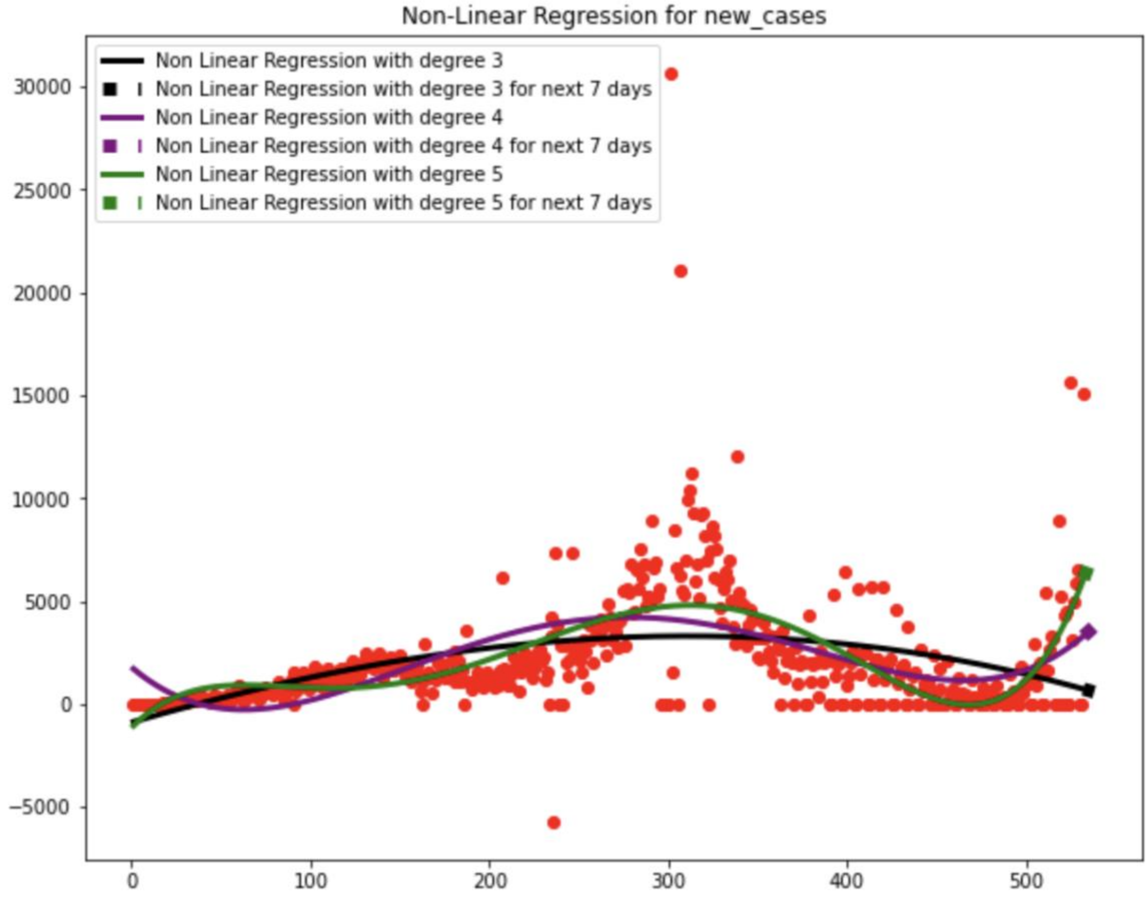
   In order to find the start date of the infection, I took the column of the new cases into a list and on that list, I wrote a for loop which gives the index of the row where the first case is registered. I wrote another for loop on the list of new cases appending count from the index of first new case registered. Then I stored the count into a particular column of the respective dataframe. I dropped the rows of the dataframe until the start day of the infection. So, the final dataframe contains the information from the start of the new cases. I repeated the same process for finding the start day of the deaths.

The trend of new cases across North Carolina state and prediction forecast of one week ahead using Linear regression model is as follows:



root mean square error for Linear Regression is: 2702.3695632518647

The trends of new cases across North Carolina state and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):

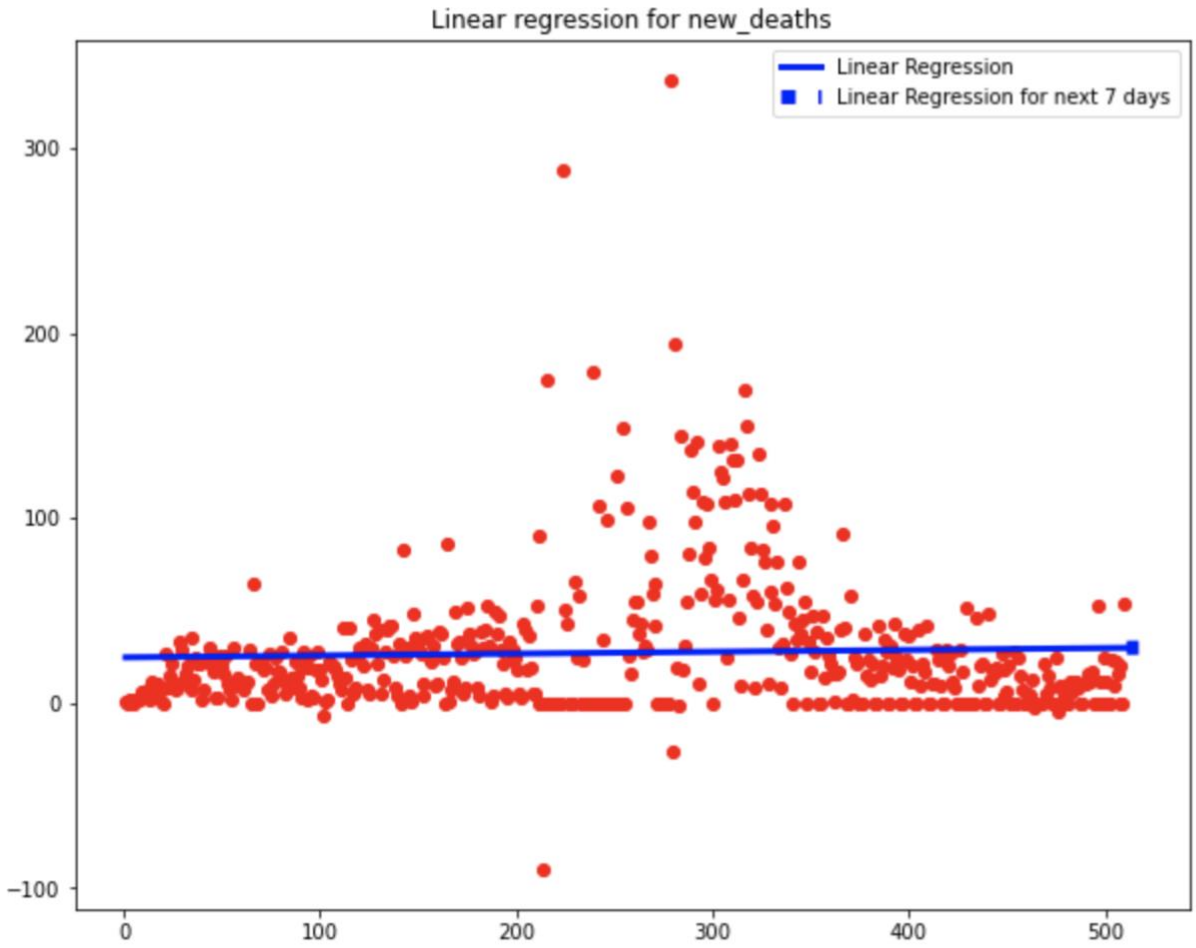


root mean square error for Non-Linear Regression with degree 3 is: 2517.740125969658

root mean square error for Non-Linear Regression with degree 4 is: 2361.255497995644

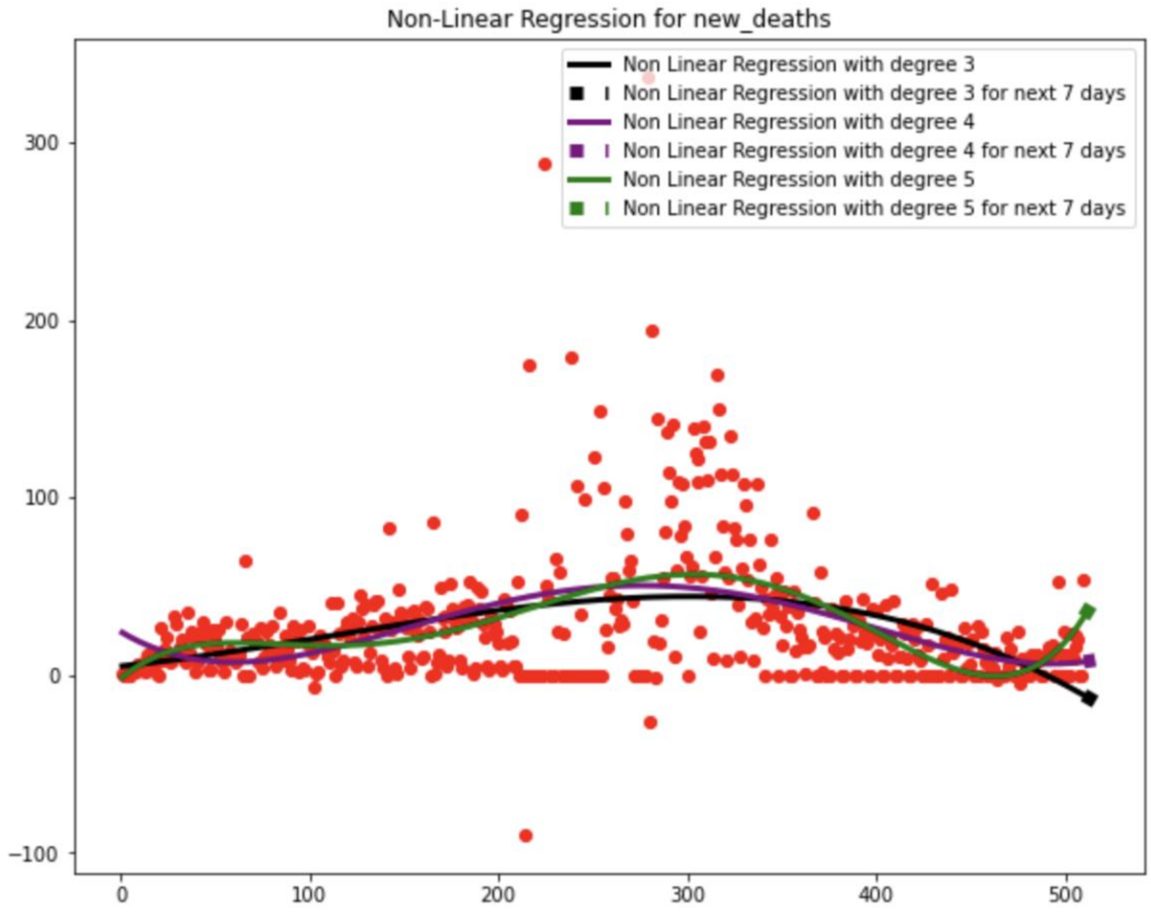
root mean square error for Non-Linear Regression with degree 5 is: 2200.131575367828

The trend of new deaths across North Carolina state and prediction forecast of one week ahead using Linear regression model is as follows:



root mean square error for Linear Regression is: 37.93658866059936

The trends of new deaths across North Carolina state and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):



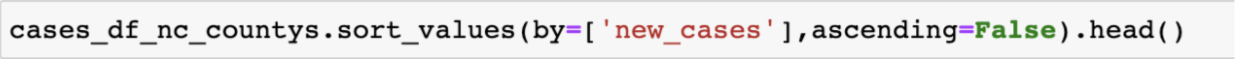
root mean square error for Non-Linear Regression with degree 3 is: 35.1556949750982

root mean square error for Non-Linear Regression with degree 4 is: 34.56947528380196

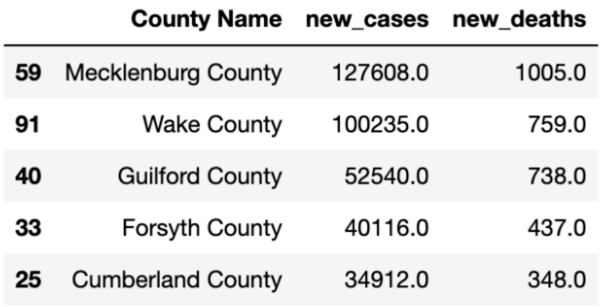
root mean square error for Non-Linear Regression with degree 5 is: 33.67250052860087

We can observe that the root mean square error for linear regression is more compared to the root mean square error of non-linear (Polynomial) regression. Even in the polynomial regressio n as we increase the value of 'n' the root mean square error is getting reduced.

2) Identify which counties are most at risk. Model for top 5 counties with cases within a state and describe their trends.
In order to find the counties that are more at risk in NC state. I grouped the dataframe by countyname and found the new cases and new deaths across each county. After that I sorted the dataframe according to the new cases by giving ascending is false and took the top 5 counties in the dataframe which are more at risk. Top 5 counties are as follows:

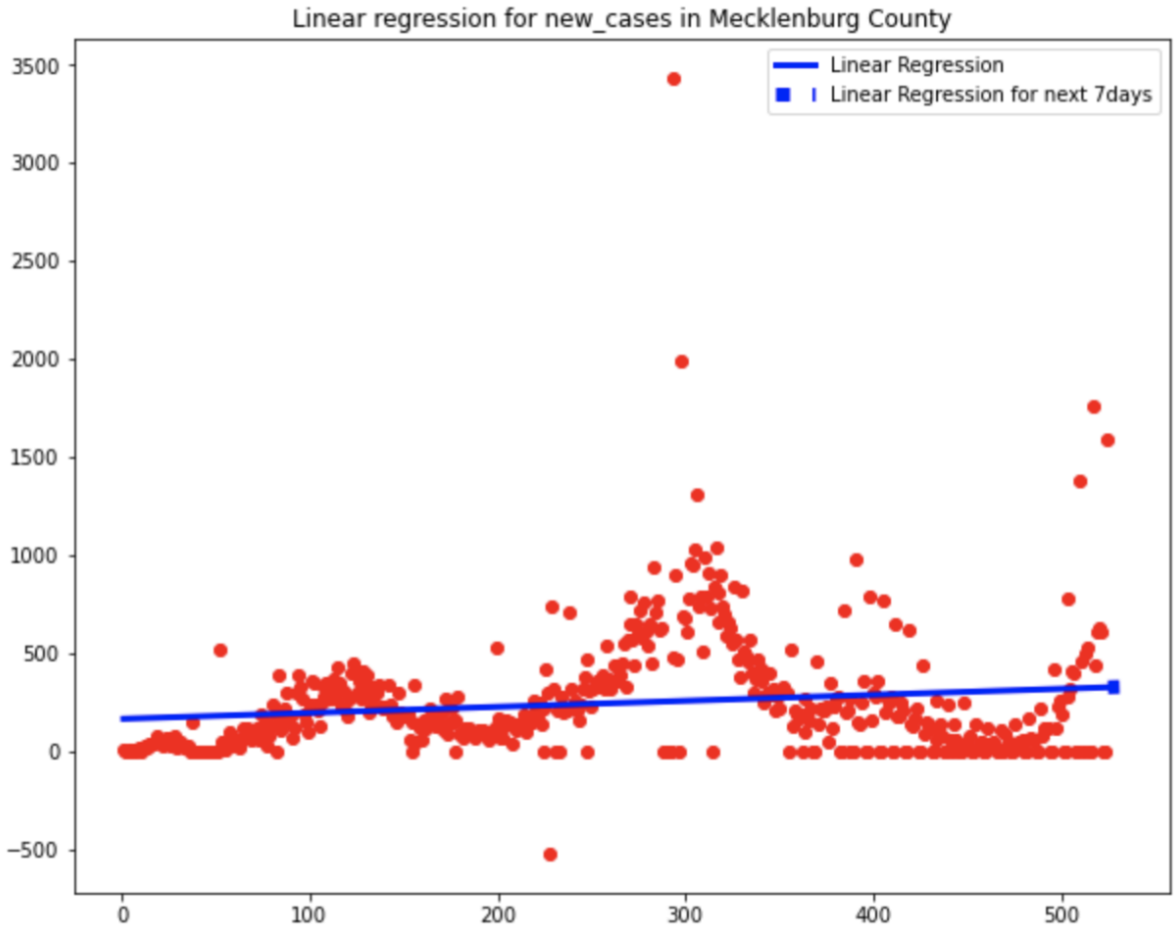


```
cases_df_nc_countys.sort_values(by=['new_cases'],ascending=False).head()
```



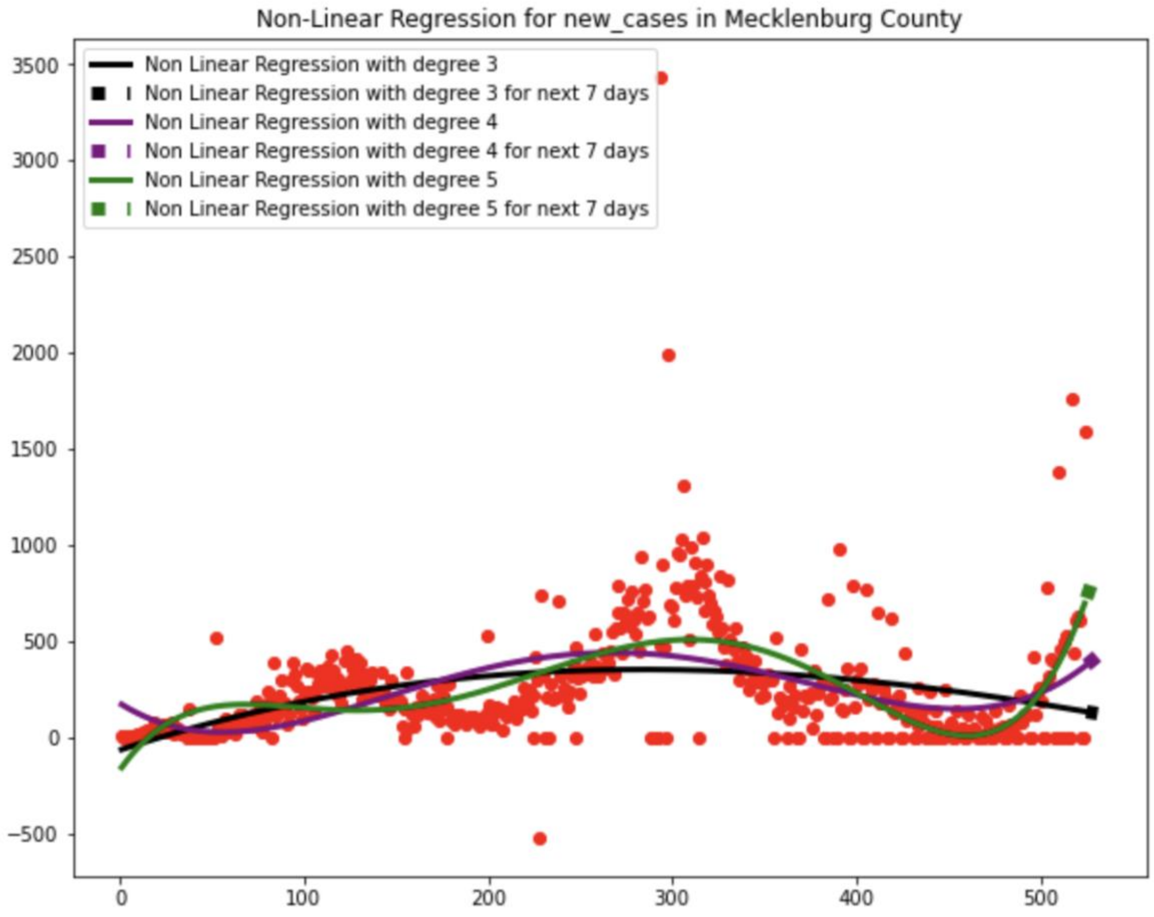
	County Name	new_cases	new_deaths
59	Mecklenburg County	127608.0	1005.0
91	Wake County	100235.0	759.0
40	Guilford County	52540.0	738.0
33	Forsyth County	40116.0	437.0
25	Cumberland County	34912.0	348.0

The trend of new cases across Mecklenburg County and prediction forecast of one week ahead using Linear regression model is as follows:



root mean square error for Linear Regression is: 296.9387602585522

The trends of new cases across Mecklenburg County and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):

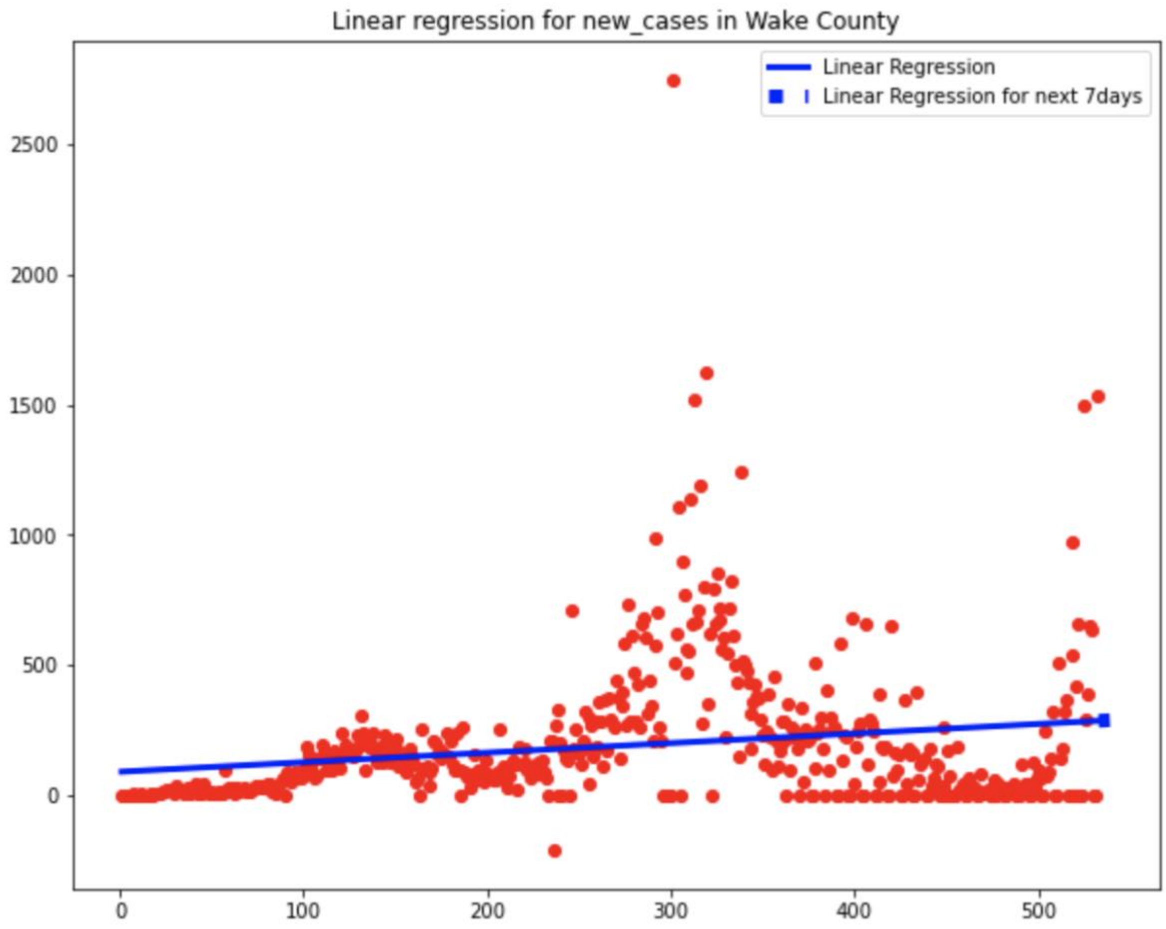


root mean square error for Non-Linear Regression with degree 3 is: 281.51278655887876

root mean square error for Non-Linear Regression with degree 4 is: 270.056320065757

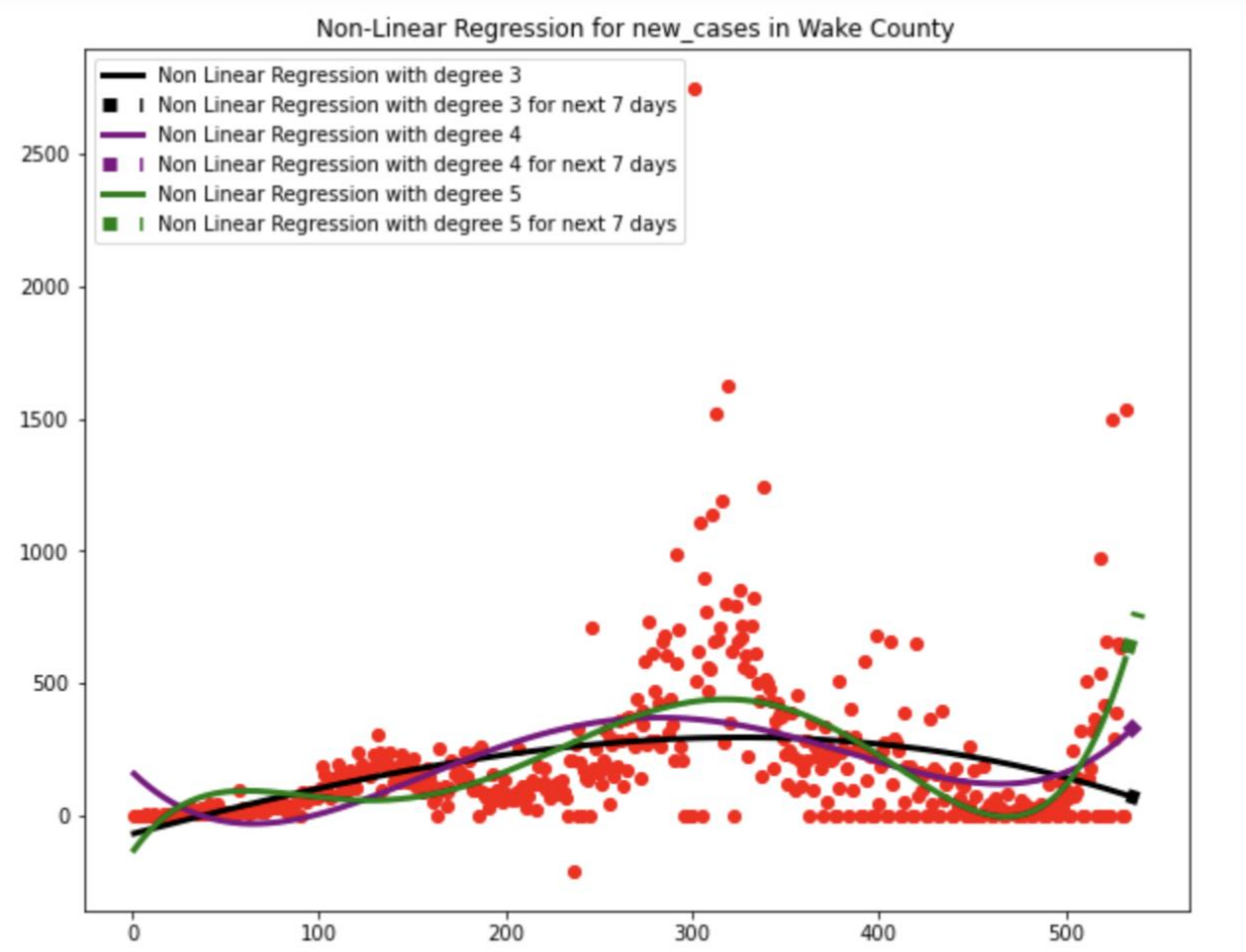
root mean square error for Non-Linear Regression with degree 5 is: 250.16458695168623

The trend of new cases across Wake County and prediction forecast of one week ahead using Linear regression model is as follows:



root mean square error for Linear Regression is: 259.3305660883798

The trends of new cases across Wake County and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):



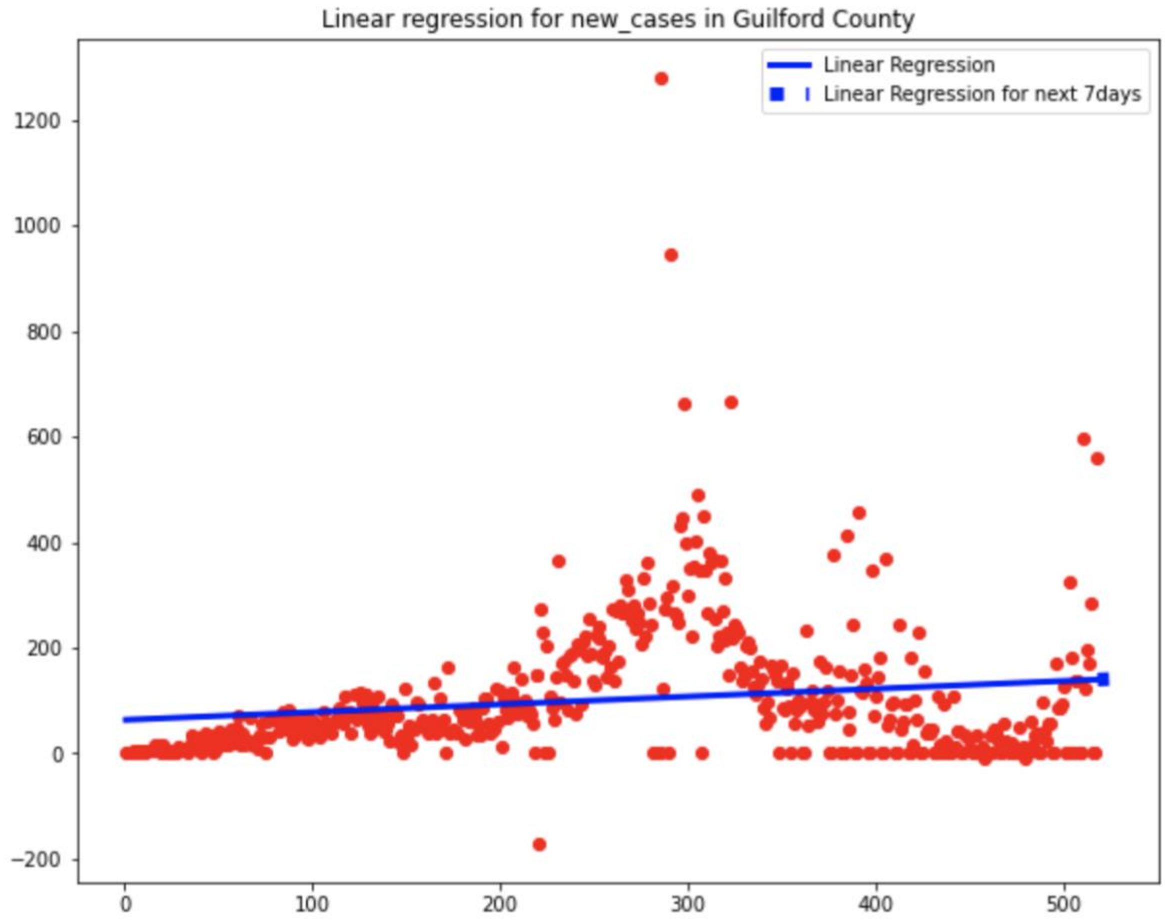
root mean square error for Non-Linear Regression with degree 3 is: 245.8293495821034

root mean square error for Non-Linear Regression with degree 4 is: 233.57456795981975

root mean square error for Non-Linear Regression with degree 5 is: 215.87631101195313

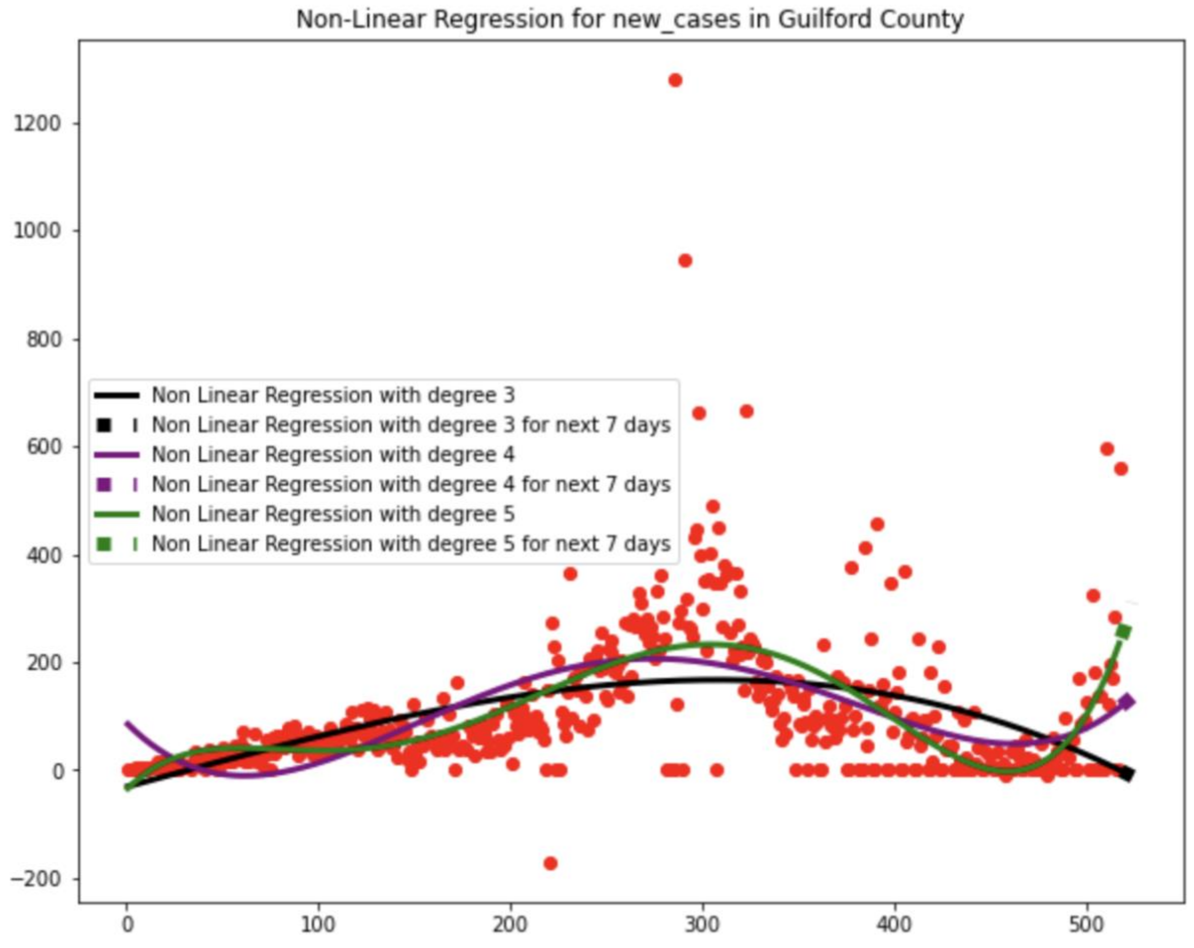
The trend of new cases across Guilford County and prediction forecast of one week ahead using Linear regression model is as follows:

-



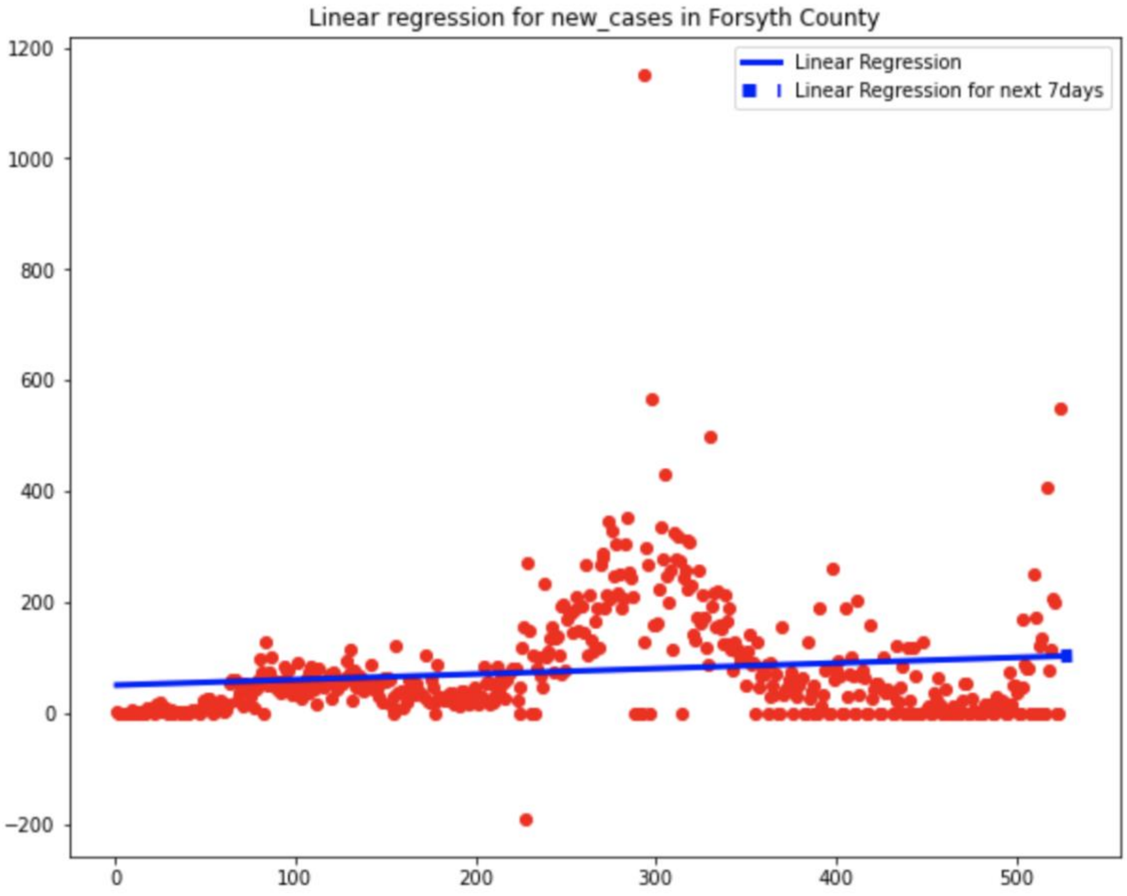
root mean square error for Linear Regression is: 125.63593727198354

The trends of new cases across Guilford County and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):



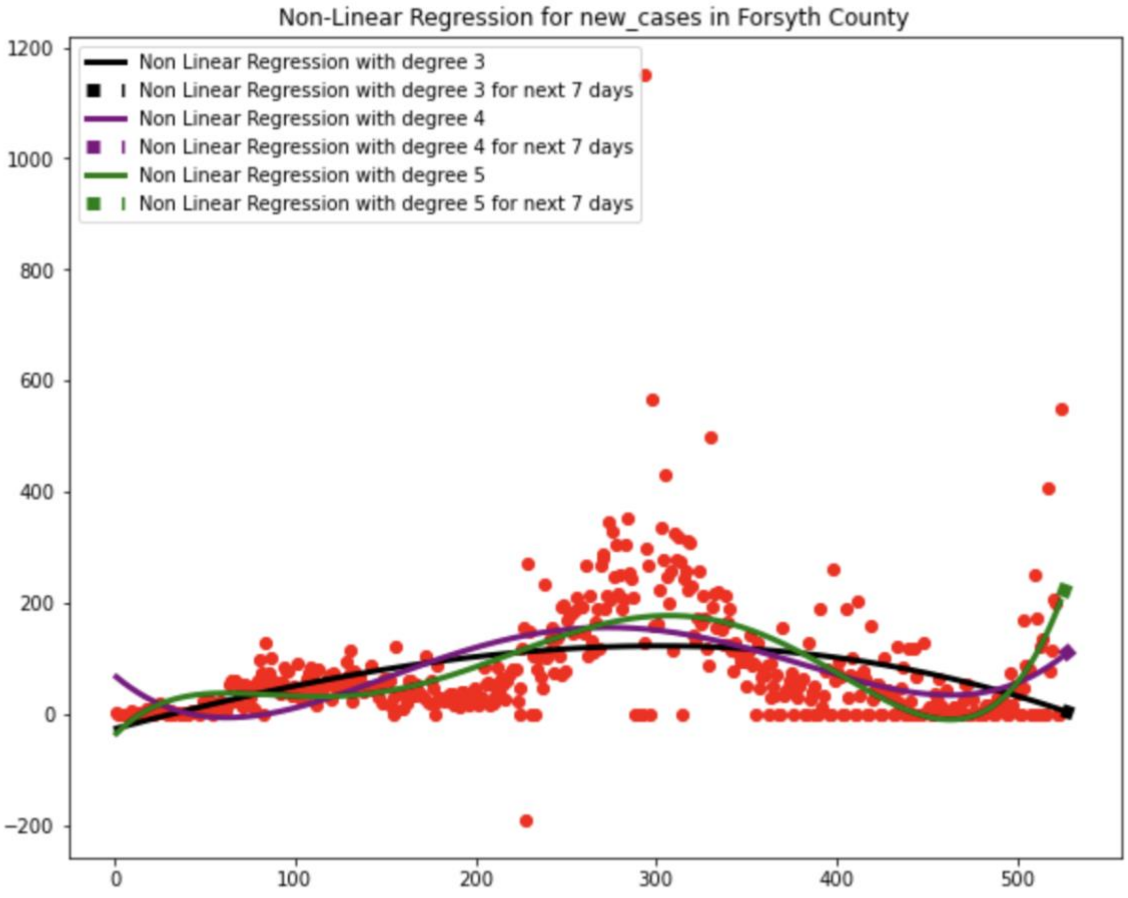
```
root mean square error for Non-Linear Regression with degree 3 is: 113.85701076554291

root mean square error for Non-Linear Regression with degree 4 is: 106.88344634923264

root mean square error for Non-Linear Regression with degree 5 is: 100.30089474482764
```

The trend of new cases across Forsyth County and prediction forecast of one week ahead using Linear regression model is as follows:



root mean square error for Linear Regression is: 99.4532093658823

The trends of new cases across Forsyth County and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):

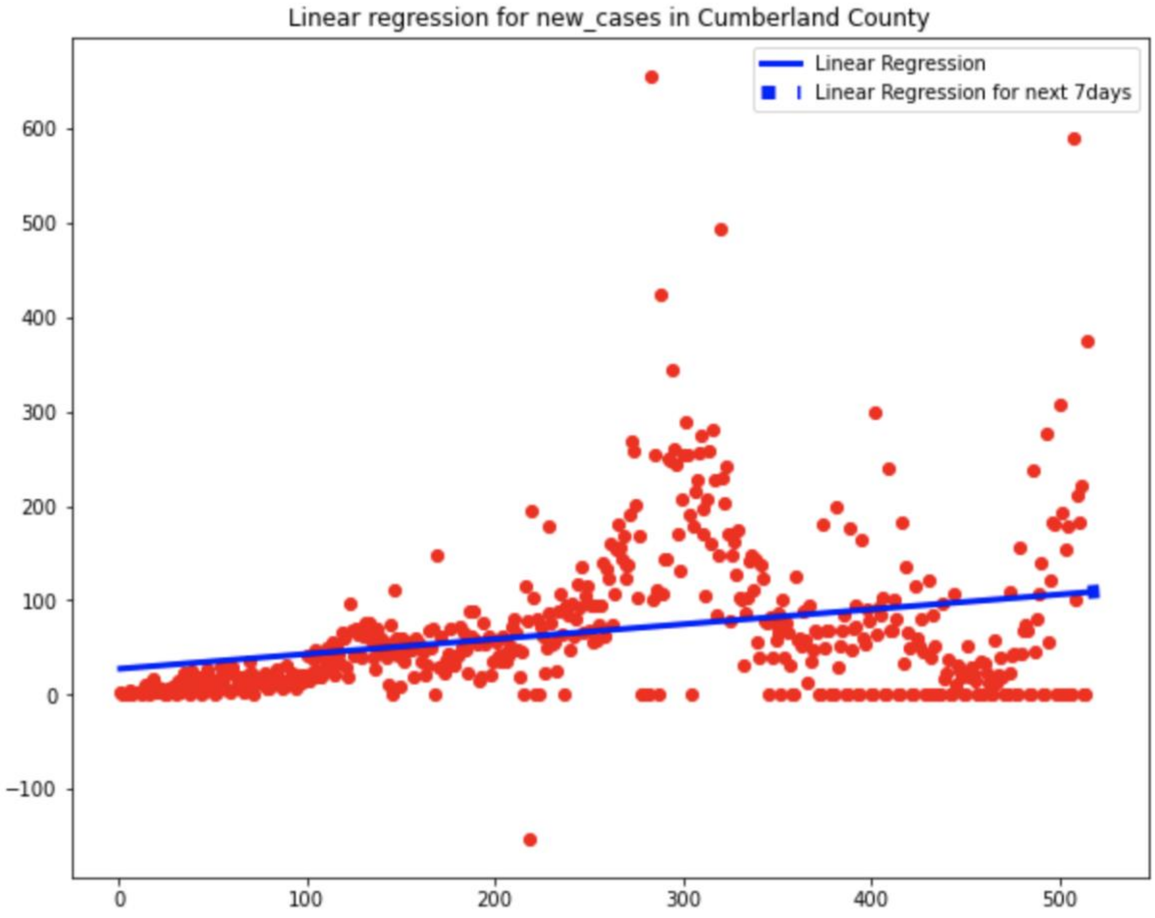


root mean square error for Non-Linear Regression with degree 3 is: 91.6406316573031

root mean square error for Non-Linear Regression with degree 4 is: 86.05734602342757

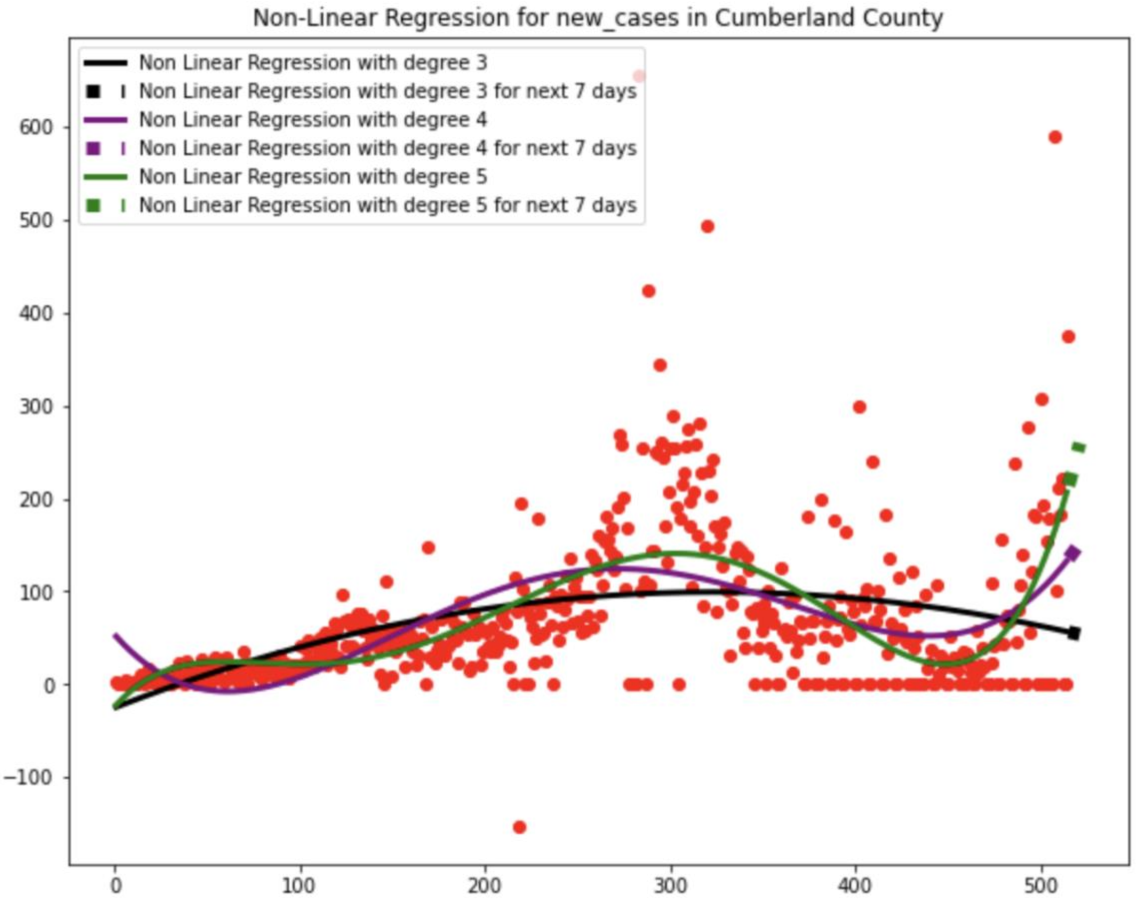
root mean square error for Non-Linear Regression with degree 5 is: 80.0285163912363

The trend of new cases across Cumberland County and prediction forecast of one week ahead using Linear regression model is as follows:



root mean square error for Linear Regression is: 78.24639341789913

The trends of new cases across Cumberland County and prediction forecast of one week ahead using Non-Linear (Polynomial) regression model is as follows (I started taking degree of polynomial from 3):



root mean square error for Non-Linear Regression with degree 3 is: 74.68277712002978

root mean square error for Non-Linear Regression with degree 4 is: 70.05887344684376

root mean square error for Non-Linear Regression with degree 5 is: 66.19239397740758

3) Perform hypothesis tests on questions identified in Stage II

i) Alternative Hypothesis: Covid cases are higher where there are a greater number of males?
Null Hypothesis: Covid cases are normal where there are greater number of males.

After applying two tail-two sample t-test the obtained statistic value and pvalue are

```
Ttest_indResult(statistic=4.738051025352931, pvalue=7.147880361398873e-06)
```

From the above result we can see that p-value is less than the threshold value i.e., 0.05 so we can reject the null hypothesis.

Now apply one-tail-two sample t-test to check whether it lies on lower side or greater side.

From the above result we can see that p/2 < 0.05 and t-statistic value >0. So, it lies on the greater side.

ii) Alternative Hypothesis: Covid cases are higher where there are a greater number of white population?

Null Hypothesis: Covid cases are less where there are a greater number of white population?

After applying two tail-two sample t-test the obtained statistic value and pvalue are

```
Ttest_indResult(statistic=5.703940213322775, pvalue=1.1942291564933215e-07)
```

From the above result we can see that p-value is less than the threshold value i.e., 0.05 so we can reject the null hypothesis.

Now apply one-tail-two sample t-test to check whether it lies on lower side or greater side.

From the above result we can see that p/2 < 0.05 and t-statistic value >0. So, it lies on the greater side.

iii) Alternative Hypothesis: Covid cases spread is higher where there are more population?

Null Hypothesis: Covid cases spread is less where there are more population?

After applying two tail-two sample t-test the obtained statistic value and pvalue are

```
Ttest_indResult(statistic=5.536557789493981, pvalue=2.492792004182535e-07)
```

From the above result we can see that p-value is less than the threshold value i.e., 0.05 so we can reject the null hypothesis.

Now apply one-tail-two sample t-test to check whether it lies on lower side or greater side.

From the above result we can see that p/2 < 0.05 and t-statistic value >0. So, it lies on the greater side.