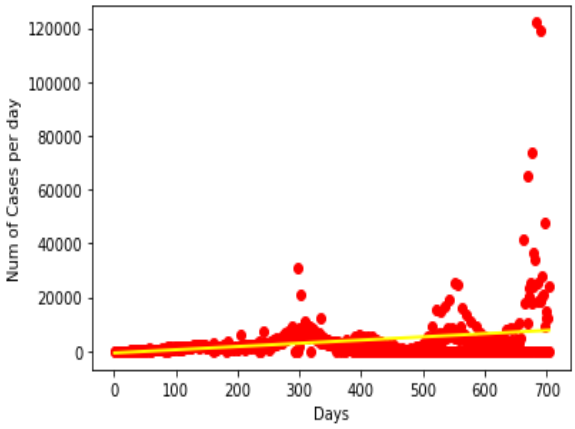
Stage IV

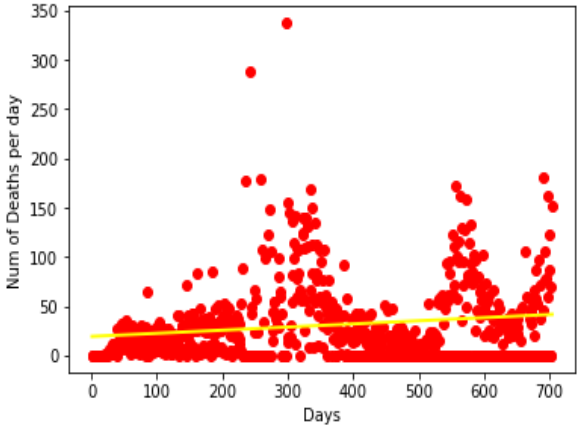
Task1: Utilize Linear and Non-Linear (polynomial) regression models to compare trends for a single state and its counties (top 5 with highest number of cases). Start your data from the first day of infections.

For this task I have selected North Carolina as my state of choice.

- To calculate the first occurrence of infections, I have used the method of idmax() to find the first day when cases start to report. Once I have found that index, I trim the above data.
- I grouped the data by date to get the total cases and deaths daily count.
- I have used the method of linear regression of statsmodel.formula.api to build the linear regression model on the given data. I have applied the same method to plot the regression model for cases/deaths.

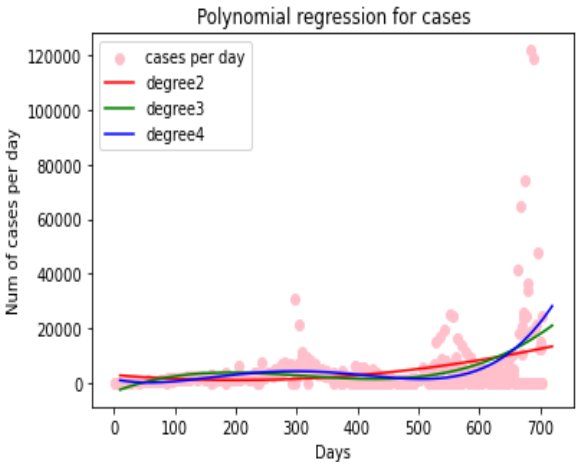


Linear regression for Cases

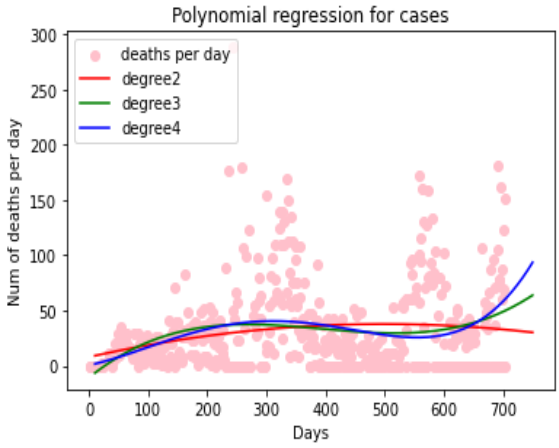


Linear regression for deaths

- For non-linear regression I have used the same method that we have used in the team task. I have mentioned it in the notebook also.
- I have shown the non-linear regression for degree 2, 3 and 4.



Nonlinear regression for cases



Nonlinear regression for deaths

- I have selected top 5 counties in NC as per cases by grouping the data by county and sorting in descending by cases. Then I followed the same steps to plot linear and nonlinear regression for counties.
- I plotted the trendline for cases and deaths for state as well as its counties using the plotly plots with trendline = 'ols' function.

Task2: 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.

- For this task, I have made use of the super hospital enrichment dataset generated by one of our team members in stage 1 of the project.
- To give a brief idea about the point of no return: if the number of bed utilized during the covid exceeds total beds availability, then it we can say that there is a point of no return.
- I have plotted the line plots for total utilization and total available. Those lines do not cross each other and from that I came to conclusion that NC do not have point of no return.

Task3: Perform hypothesis tests on questions identified in Stage III.

- To perform hypothesis testing, I have used census demographic dataset. I generated null and alternative hypothesis for testing.

- I calculated stats.ttest_ind() for population, age range with confirmed cases and check for statistic and p value score. I have shown criteria:
    - Greater: p/2 < α and t-statistic > 0
    - Less: p/2 < α and t-statistic < 0
- Based on that I proved that if my null hypothesis values are true or if alternative hypothesis are true.

For each of the scenario I have plotted confidence interval, trendline and prediction also.