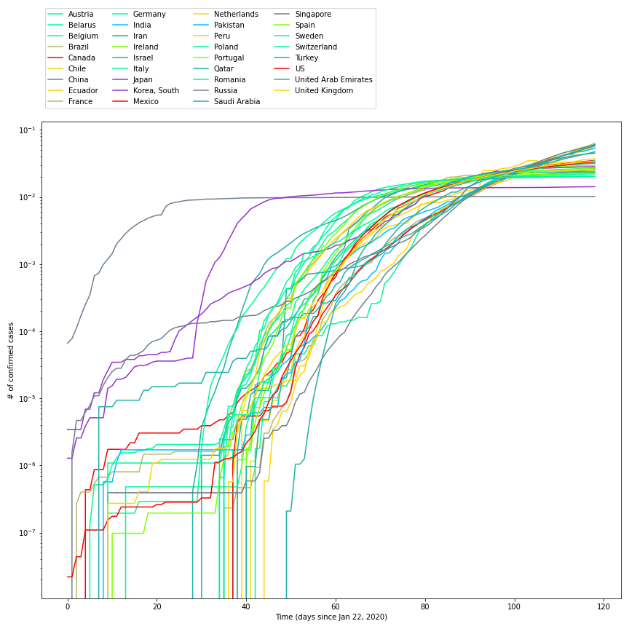
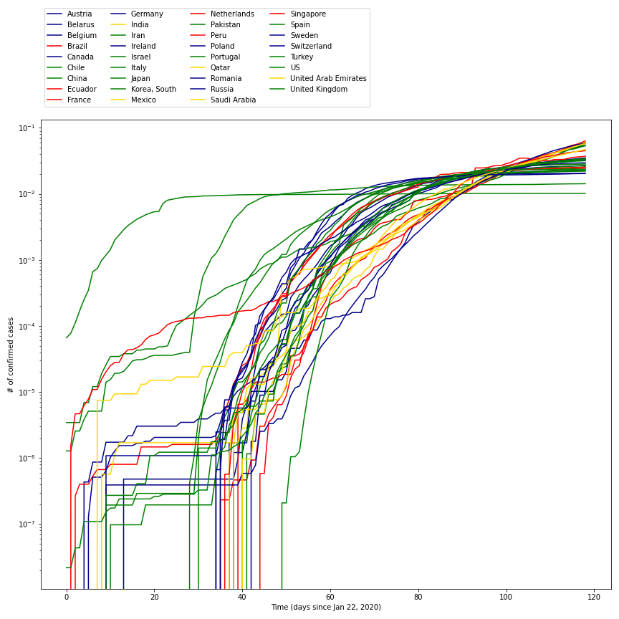
**Final Project Option 1**

**Classification**

For the final project, I start by performing some visualizations on the global confirmed cases data. I want to see that if there are some trends that I can find by grouping countries based on different traits.

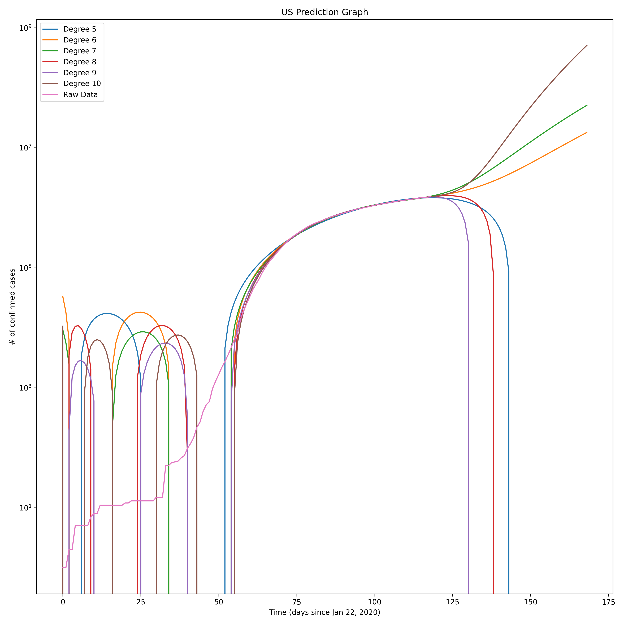
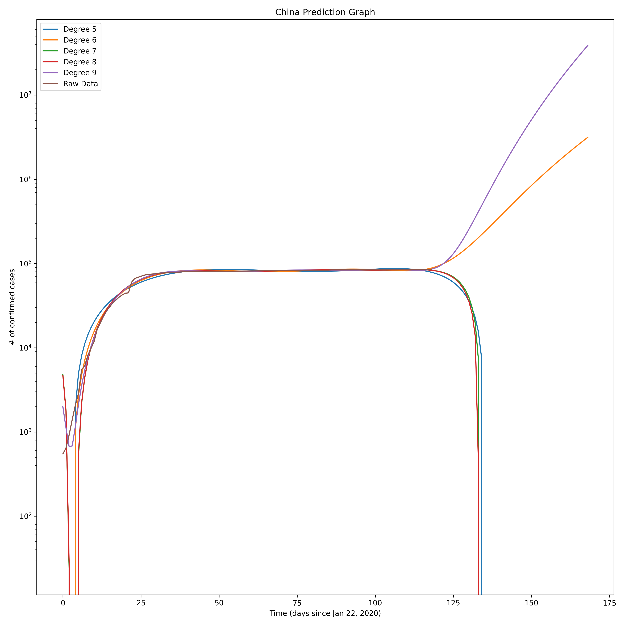


The plot on the left is made by grouping countries based on their latitude, with brighter colors representing countries that are closer to the equator. I want to test if countries with warmer climate have lower growth rate for confirmed cases. Since the data is normalized, we the rate of change of the cases is directly reflected by the slope of the plots. However, the plot shows no correlation between latitude and growth rate as trajectories marked by bright colors show no distinctive pattern compared with others.

The plot on the right is made by grouping countries based on their longitude. Countries with similar longitudes have similar color. The assumption is that similar longitude indicates that the countries are on the same continent, which suggests that they have similar culture and political system. The result shows that counties in Asia like China, Korea, Japan, and Singapore, have lower growth rate. It could suggest that the emphasis on individual freedom in Western countries hinders their efforts in combating the pandemic.

**Prediction**

Then, I tried to perform predictions on the data using np.polyfit(). To find the optimal degree, I plotted the fit result using degree five to degree ten. I also used the curve for US and China to investigate the performance of this approach on different trajectory. The plots are attached below.



The fitting if performed using data from available 119 days. Then, based on the polynomial we obtain, we performed predictions for additional fifty days. Immediately, we can see a drawback for this approach. Because we are dealing with total confirmed cases, the predictions that becomes negative are clearly invalid. Furthermore, we noticed that the predicted trends are unrealistic. Based on the original data, the increase of confirmed cases has slowed down in both US and China. However, the polynomials we obtained both show drastic increase when they enter the prediction range. This is likely caused by the fact that polynomials tend to diverge when the absolute value of independent variable goes to infinity, which means it is not a great choice for making predictions. We also performed predictions using higher degree polynomials. The result is shown in the graph below. Even at degree 50, the polynomial is not capable of producing results with great value. Therefore, we believe that we need hand-crafted mathematic models to explain the behavior of virus spread.

