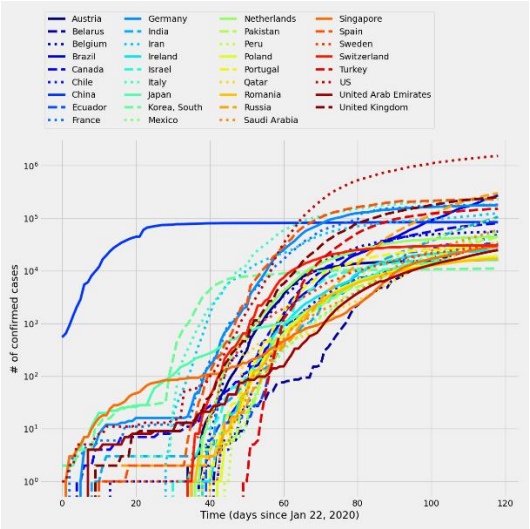
Final Project Option 1

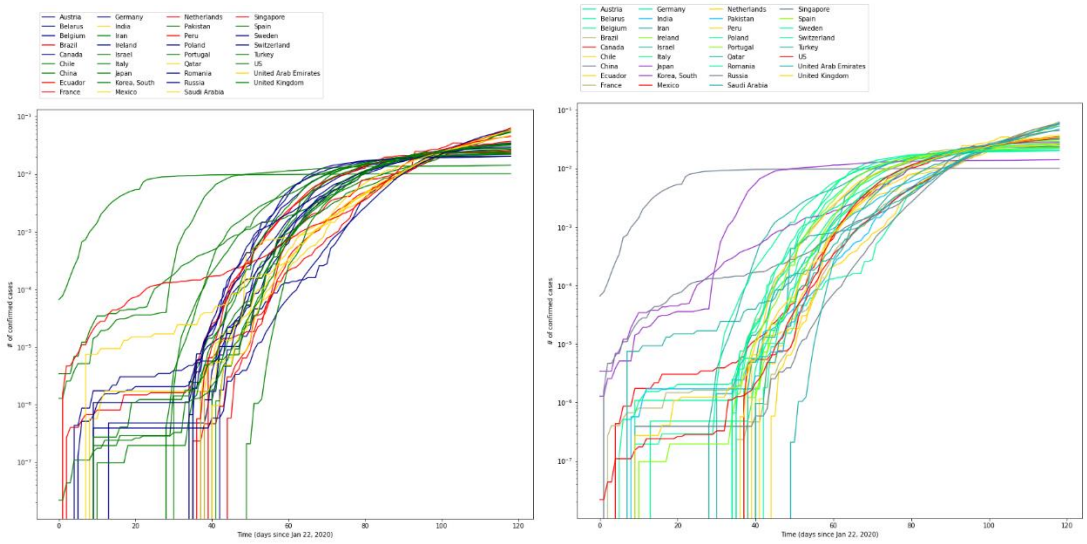
Introduction



First, we can examine the data by directly plotting them. From this plot, we can see that there are roughly two patterns of growth. Countries like China and France have a rate of increase that is high from the beginning and gradually slows down, whereas UK, US, and Japan first slowly increase, and the growth rate then suddenly increases and stabilizes towards the end. The second pattern is likely caused by the improvement in testing capacity and awareness. However, clustering gives us some different results, which might be caused by the delay in growth in different countries. Therefore, we focus primarily on visualization results. Recognizing the growth pattern, we can do more investigations.

Classification

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.

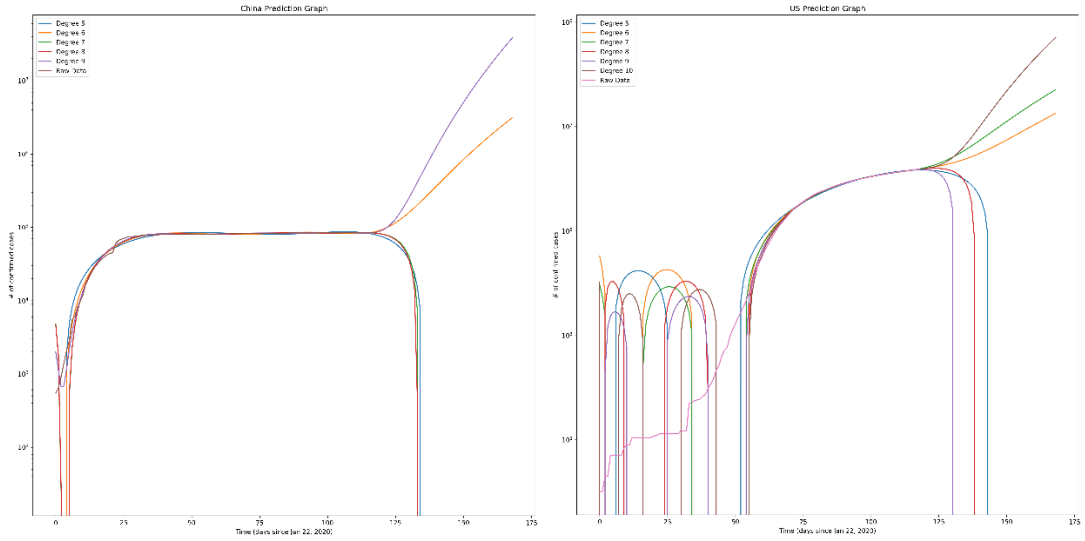


The plot on the left is made by grouping countries based on their latitude, with brighter colors representing countries that are closer to the equator. I want to test if countries with warmer climates have lower growth rates for confirmed cases. Since the data is normalized, the rate of change of the 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 colors. The assumption is that similar longitude indicates that the countries are on the same continent, which suggests that they have similar culture and political systems. The result shows that counties in Asia like China, Korea, Japan, and Singapore, have lower growth rates. It could suggest that the emphasis on individual freedom in Western countries hinders their efforts in combating the pandemic. This conclusion is also backed by the clustering result, which shows resemblance in the respective groups.

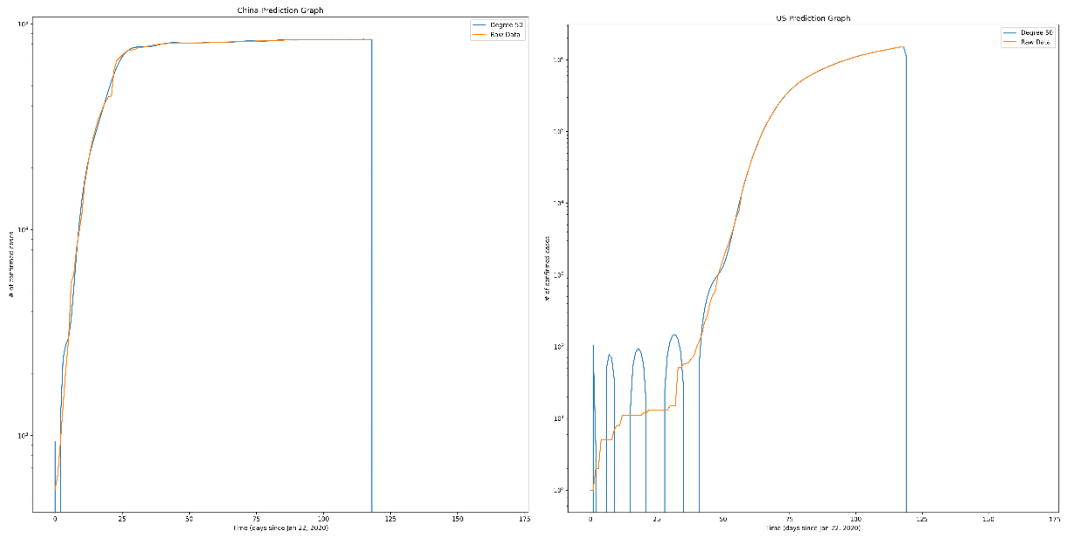
Prediction

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 the US and China to investigate the performance of this approach on different trajectories. The plots are attached below.



The fitting is performed using data from the available 119 days. Then, based on the polynomial we obtain, we performed predictions for additional fifty days. Immediately, we can see a drawback to this approach. Because we are dealing with total confirmed cases, the predictions that decrease are 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 increases when they enter the prediction range. This is likely caused by the fact that polynomials tend to diverge when the absolute value of the 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.



Link to Code: https://github.com/JerryXRQ/Final-Project
Under exp folder