COVID-19 Trend Prediction and Analysis

CSE163 Project

Authors: Anny Kong
Forrest Jiang
Zealer Xiao

Summary of research questions AND RESULTS:

Q1: What's the future trend of COVID-19 in each country?
A1: Our result shows that if every country does not enforce more health guidelines, the number will continue to grow according to the polynomial model, with 83% accuracy.
Q2: By data visualization, what can we say about how different regions (countries) may affect the number of confirmed, deaths, and recovered cases?
A2: The month is a large factor in the relationship with the current Covid-19 condition of different regions. From February to August, Covid-19 has spread all over the world. Though for those neighboring affected countries/regions, they are more likely to be affected, there isn't evidence for a causal relationship between longitude/latitude and the number of cases.
Q3: How do the infection rate and fatality rate of COVID-19 relate to regions' income levels?
A3:

Motivation and background:

As the health crisis triggered by the pandemic COVID-19 keeps the world in its grasp, most people are forced to work from home as well as to practice social distancing. Our project could potentially foretell how future pandemics are going. We will visualize the virus spread with respect to different regions/countries and analyze how the virus spread is correlated to different geolocations, and see how the projection of infected cases will be like in the near future.

Our modeling and forecasting could potentially help evaluate the course of the pandemic. It could also provide insightful ideas to the governments on how different implementations of public health measures may impact the spread. In a narrow sense, investigating the relationship between local income level and rate of infection and fatality could help us specifically understand the situation in our community. The information could be used to make more adaptive policies since current guidelines from the federal government are drafted on the national scale while COVID-19 did not hit every state at the same time. With real-time geographical visualizations of the virus spread, we could efficiently avoid visitors traveling to those regions/countries to help reduce COVID-19 spread. Future predictions of virus spread may help the government or associated organizations pay more attention to specific regions and take effective measures.

Dataset source:

Income by County in the United States
The dataset contains the average income of each county in the United States from 2016 to 2018 and percent change from the preceding year.

About Natural Earth Vector
The same dataset introduced in the lecture contains Geospatial information about countries around the world. It was built through a collaboration of many volunteers and is supported by NACIS.

COVID-19 Globe Daily Case Report
The dataset that we are using is the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The data contains daily cases reported in UTC. Each row is a single case that contains information about each case such as the time (Last Update), the place (Province_State, Country_Region, Lat, Long_), details of the case (Confirmed, Deaths, Recovered, Active, Incidence Rate, Case-Fatality Ratio (%)).
Specifically, here is a list of field description:

- **Province_State**: Province, state, or dependency name.
- **Country_Region**: Country, region, or sovereignty name.
- **Last Update**: MM/DD/YYYY HH:mm:ss (24-hour format, in UTC).
- **Lat** and **Long_**: Dot locations on the dashboard. All points (except for Australia) shown on the map are based on geographic centroids and are not representative of a specific address, building, or any location at a spatial scale finer than a province/state. Australian dots are located at the centroid of the largest city in each state.
- **Confirmed**: Counts include confirmed and probable (where reported).
- **Deaths**: Counts include confirmed and probable (where reported).
- **Recovered**: Recovered cases are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number. US state-level recovered cases are from COVID Tracking Project.
- **Active**: Active cases = total cases - total recovered - total deaths.
- **Incidence_Rate**: Incidence Rate = cases per 100,000 persons.
- **Case-Fatality** Ratio (%): Case-Fatality Ratio (%) = Number recorded deaths / Number cases.

Methodology (algorithm or analysis):

Q1: What's the future trend of COVID-19 in each country?
M1: After we gathered all the data from all datasets, we need to join, filter data, and clean missing entries. Join data: join data by different days. Filter data: filter out less important data columns. Missing data: analyze and fix missing data. Then, our objective is to predict the spread of this virus. Our first attempt will be using linear regression for one country, this model is

straightforward, and only elemental features will be considered are Country/Region, date information. Then we start performing the linear regression, and visualize the difference between predicted and actual trends, and log the accuracy of the model. Then we will try the ridge regression fit, and we repeat the above steps. Afterward, we will perform a polynomial regression. Finally, we will perform some hyper-parameters tuning for optimal accuracy of the model above that has the best accuracy.

Q2: By data visualization, what can we say about how different regions (countries) may affect the number of confirmed, deaths, and recovered cases?

M2: For research question 2, we specifically follow the steps below.

1) Use geographical data to visualize the number of confirmed/deaths/recovered cases using a heat map with geopandas. If the time comes in as an interest component, we may include a heat map for different days/months as well.
2) Plot the number of confirmed/death/recovered cases with respect to different regions in a single month with Seaborn and Matplotlib. And possibly adding plots for more months to see whether a specific region is more influenced in a specific month. We may analyze related events around that time and see how three different cases change differently.
3) Plot the number of confirmed/death/recovered cases with respect to longitude and latitude for exploring the correlation between geolocation and the number of confirmed/deaths/recovered cases in one month.

Q3: How do the infection rate and fatality rate of COVID-19 relate to regions' income levels?

M3: Dataset Filtering: remove data columns we are not interested in and any missing entries in the study subject

Data Join: merging the COVID-19 dataset with Income by County dataset

Visualization: set thresholds that considered to be high in income level, infection rate, and fatality rate respectively; plot two maps of infection rate and fatality rate by county, highlights the edge of the county has a high-income level

Analyzing the plot to see how income level relates to the infection rate and fatality rate.

Results:

Q1: What's the future trend of COVID-19 in each country?

The trend of the spread in each country

After we cleaned and merged all the data, we first converted the time series into days for better machine learning. Our first guess was using a simple linear regression model. However, we only had about 70.6% accuracy. (all the models were using 70% training, 30% testing). To understand why, we plotted the prediction and actual trend of two most representative countries, India and the US. (Figure 1.1)

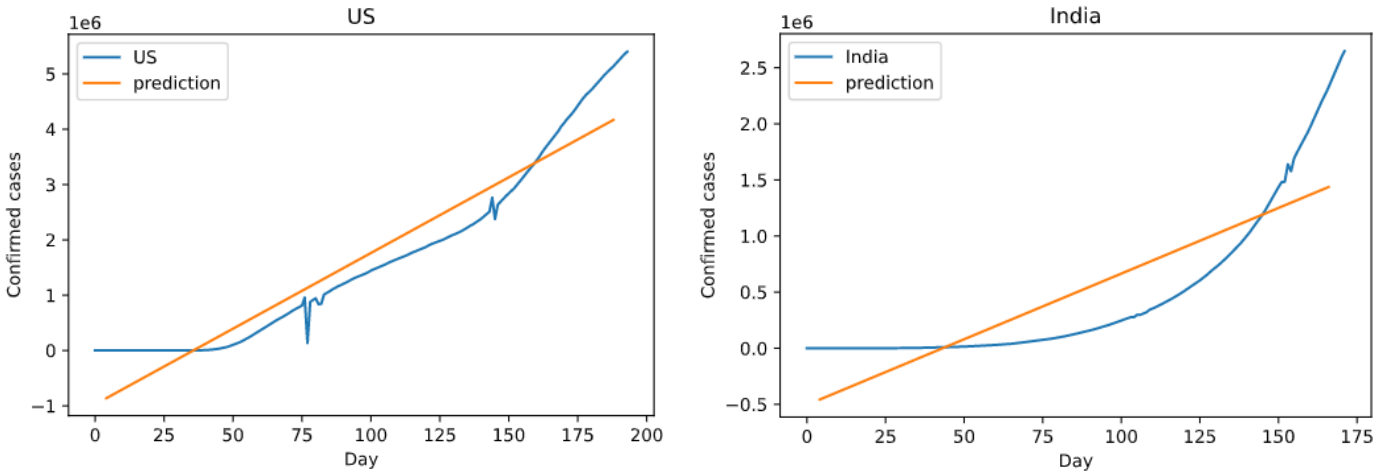

Figure 1.1 Actual Trends and Predictions for India and US using Linear Regression

It is clear that the actual trend is a curve and our model is trying to fit it using a line. Our next idea was to use regularization to improve the accuracy of our model. Then we performed ridge regression. However, the accuracy was still about 69.8%. (Figure 1.2)

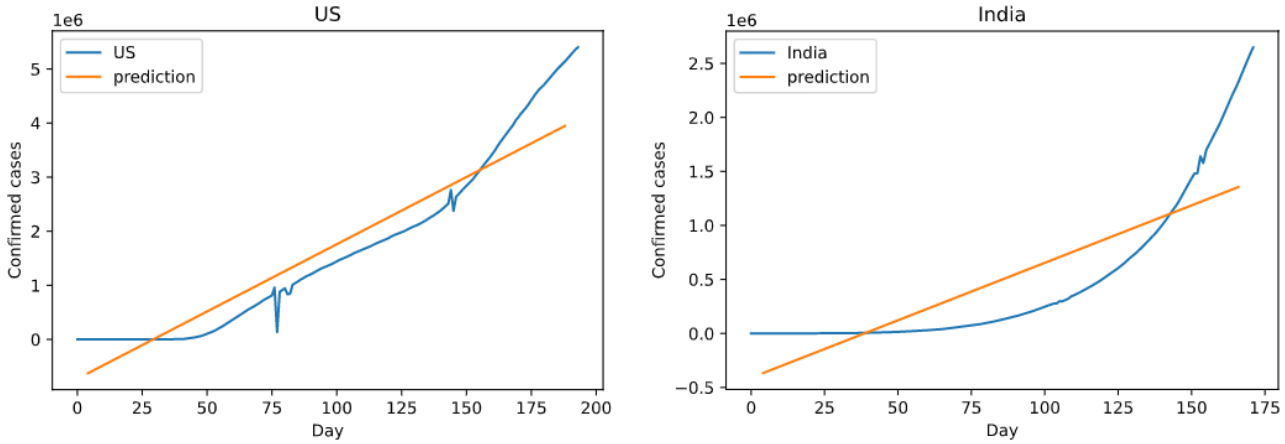

Figure 1.2 Actual Trends and Predictions for India and US using Ridge Regression

It was hard to tell the difference just by eyeballing, and the result was about the same. Afterward, since the trends were curves, why not use polynomial regression. After using polynomial regression, the accuracy improved significantly. It was about 83.5%, and that was a huge improvement. For the polynomial regression model, we used degree 5. (Figure 1.3)

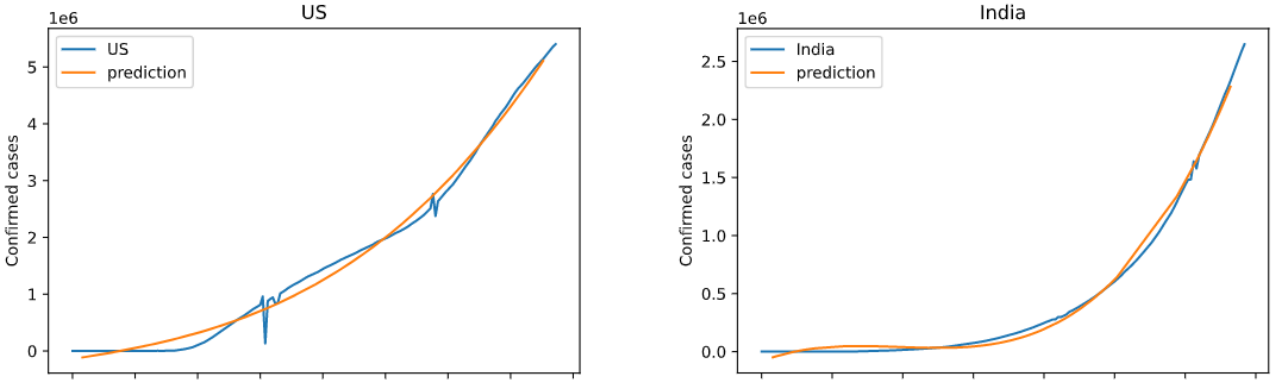

Figure 1.3 Actual Trends and Predictions for India and US using polynomial regression

Needless to say, the model fit really well, almost overlapped the original curves.
To achieve the best accuracy for this model, we used a grid search trying out different values of degrees.

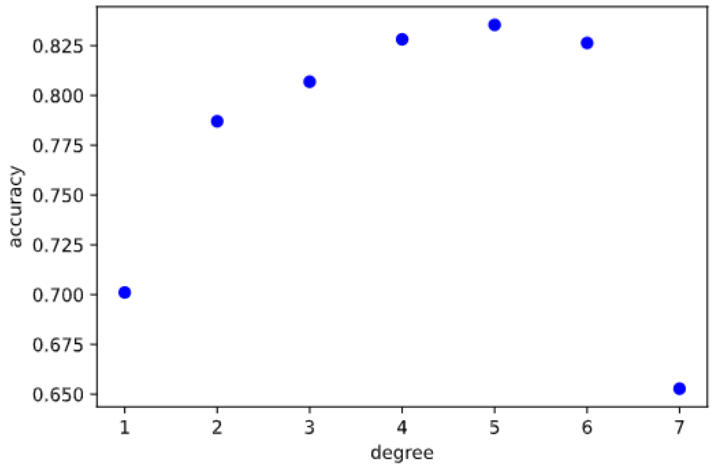

Figure1.4 Degree VS Accuracy for Polynomial Regression

As you can see(Figure 1.4), the accuracy peaked at degree 5, and going down abruptly afterward, our guess was that higher degrees tended to overfit the model.
Finally, our result shows that if every country does not enforce more health guidelines, the number will continue to grow according to the polynomial model, with 83% accuracy. (Figure 1.4)

Then let's put this polynomial model into use. We are going to make predictions for "India" and "US" for days between 200 and 350, after the onset of the spread.

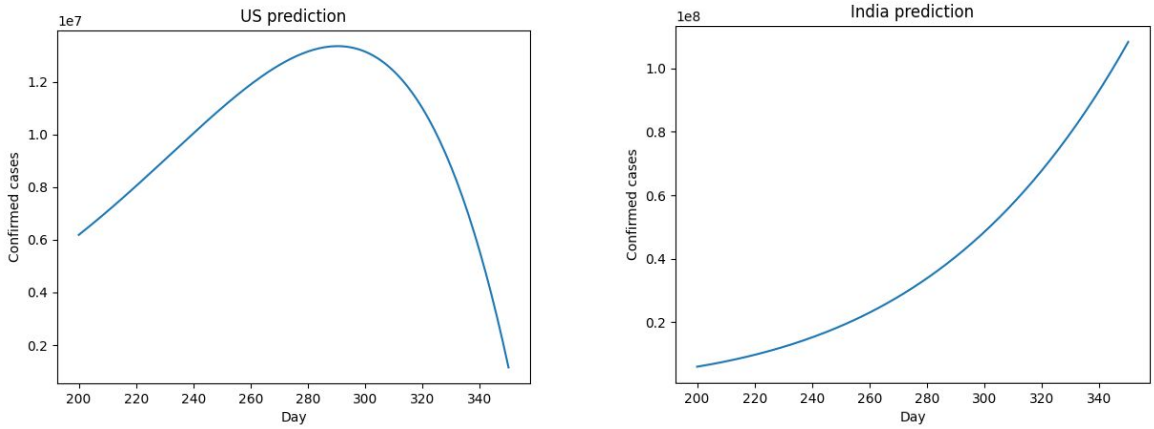

Figure 1.5 Prediction for US and India Days between 200 and 350

According to our model cases start to decline for the US and are still climbing for India under the same health guidelines and policies. (Figure 1.5)

Q2: By data visualization, what can we say about how different regions (countries) may affect the number of confirmed, deaths, and recovered cases?

Global Geographical visualization of Covid-19

We notice that the location of concentrated Covid-19 cases has changed over the month from February to August. The confirmed cases started to occur in March as demonstrated by Figure 2.1 and Table 2.1. The color legend on the right, as well as the size of the dots on the figure, gives a measure of severity. The top ten affected regions in March are mainly China, the US, Vietnam, Kuwait, United Arab Emirates, Australia, and Canada areas.

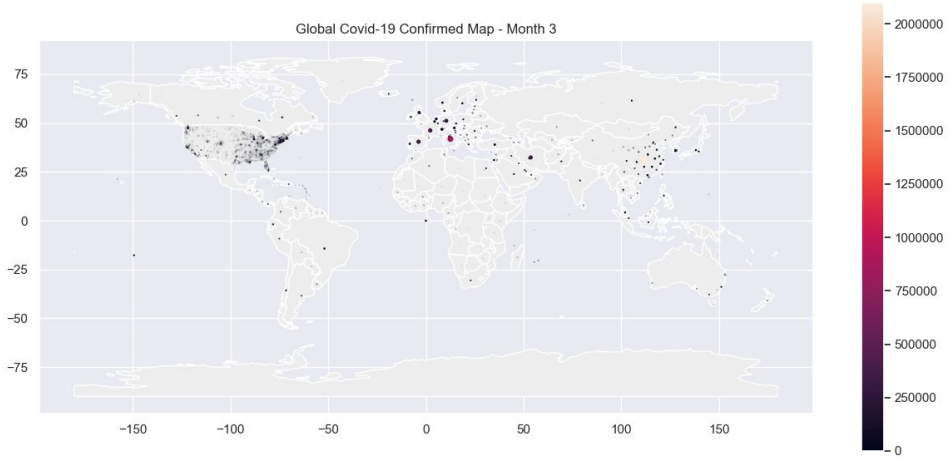

Figure 2.1 Global Covid-19 Confirmed Map in March.

Global Covid-19 Confirmed Cases by Country/Region - Month 2

	Country_Region	Confirmed
20	Others	705.0
7	China	703.0
13	Mainland China	409.0
26	US	129.0
28	Vietnam	96.0
10	Kuwait	45.0
27	United Arab Emirates	42.0
2	Australia	34.0
6	Canada	28.0
12	Macau	20.0

Table 2.1 Global Covid-19 Confirmed Cases in March.

The confirmed cases have increased a lot in July as demonstrated by Figure 2.2 and Table 2.2. The top ten affected regions in July are mainly the US, Brazil, India, Russia, Peru, Mexico, Chile, South Africa, United Kingdom, and Iran areas. We may also notice the number of dots with lighter color has largely increased - spread all over the world. According to the table, the number of top 1 region's confirmed cases has increased from 705 to over 108,398k, which demonstrates the extremely fast spread speed of Covid-19 as well as the severity of the current situation.

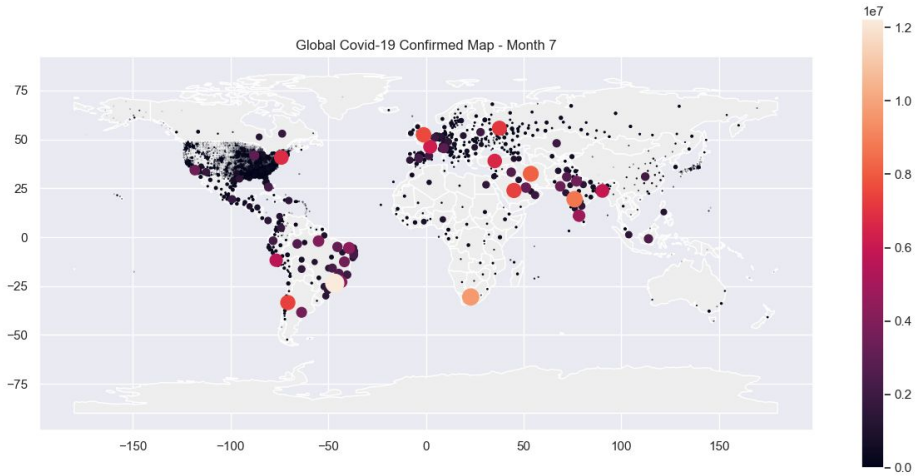

Figure 2.2 Global Covid-19 Confirmed Map in July.

Global Covid-19 Confirmed Cases by Country/Region - Month 7

	Country_Region	Confirmed
172	US	108398676.0
23	Brazil	61381421.0
78	India	31601438.0
138	Russia	23041509.0
132	Peru	10507773.0
111	Mexico	9909190.0
35	Chile	9906516.0
154	South Africa	9715943.0
176	United Kingdom	8924266.0
80	Iran	8202806.0

Table 2.2 Global Covid-19 Confirmed Cases in July.

Then, we tried to explore if there are any correlations between the number of confirmed Covid-19 cases and longitude/latitude. Based on the figures 2.3-2.6, there isn't clear evidence for a causal relationship between the number of confirmed cases and geolocation as the concentrated Covid-19 regions appear pretty random. However, we did see the month as an important factor impacting the number of cases.

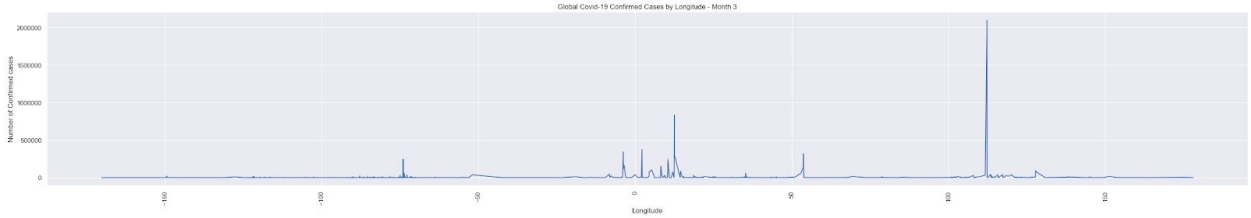

Figure 2.3 Global Covid-19 Confirmed Cases v.s. Longitude in March.

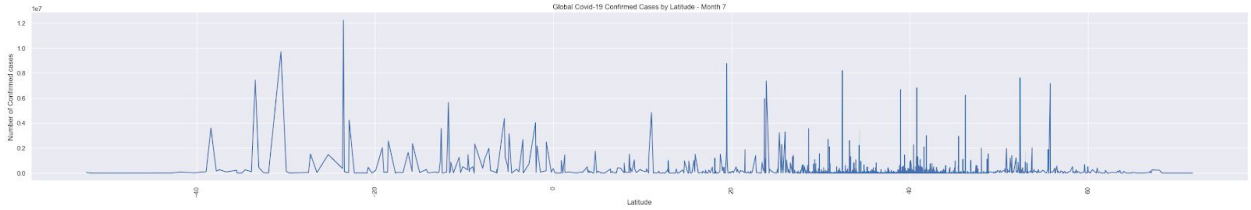

Figure 2.4 Global Covid-19 Confirmed Cases v.s. Latitude in March.

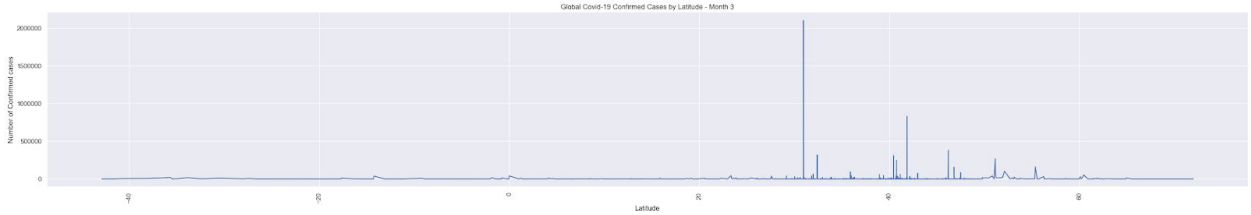

Figure 2.5 Global Covid-19 Confirmed Cases v.s. Longitude in July.

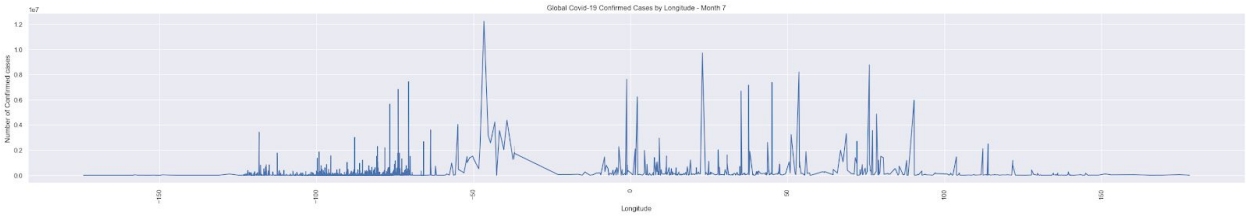

Figure 2.6 Global Covid-19 Confirmed Cases v.s. Latitude in July.

We observed some relationships between population and confirmed cases based on Figure 2.7. The US, China where countries with larger populations are witnessed in top ten lists of Covid-19 confirmed cases throughout months. It also makes sense. Since the larger the population, the more likely the Covid-19 can be spread among people. Going further in this direction of future research, it might be an interesting area to explore.

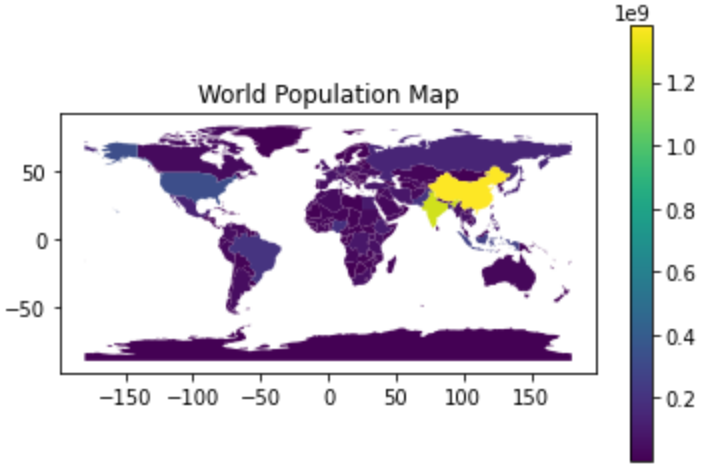

Figure 2.7 World Population Map.

For more global Covid-19 confirmed/deaths/recovered results, check out our GitHub repo or Research Question 2 results gallery webpage for more information.

Q3: How do the infection rate and fatality rate of COVID-19 relate to regions' income levels?

Challenge Goals:

Multiple Datasets - We have seen a lot of datasets with different dates and it is constantly updating. It requires us to extract important portions of data and combine related ones. Evaluation: We combined daily reports ranging from 2020-01-22 to 2020-08-15 into a big dataset as well as resolved naming conflicts of different columns. We also did preprocessing to extract useful and non-null rows for data analysis

Machine Learning - We will apply machine learning concepts learned in class to predict the future trends of Covid-19. We incorporated linear regression, ridge regression, and polynomial regression in our machine learning models.
Evaluation: As expected, we applied machine learning concepts learned in lecture. It comes in as a challenge to select the most appropriate ML model for prediction and apply appropriate data as x and y for training the model.

Other - This is a new challenge we met while implementing the second research question. As we found it could be really helpful to give a demo web page of how real-time visualization of confirmed/deaths/recovered cases on the global map can look like, we looked into creating a HTML/CSS/JS website with tutorials provided by w3schools.com.

Work Plan Evaluation:

Work Plan From Part I:
We will be using Git as our Version Control System for code collaboration.
Github Repo Link: https://github.com/AnnyKong/cse163-20su-final

Design (~5 hours each)
- Design data structures or specific functions for Q1
- Design data structures or specific functions for Q2
- Design data structures or specific functions for Q3

Implementation (~20 hours each)
- Implement Q1
- Implement Q2
- Implement Q3

Testing (~10 hours each)
- Implement tests for Q1
- Implement tests for Q2
- Implement tests for Q3

Modified Work Plan From Part I:
We will be using Git as our Version Control System for code collaboration.
Github Repo Link: https://github.com/AnnyKong/cse163-20su-final

Design (~5 hours each)
- Try to model pandemic trends for each country using Linear, ridge, and polynomial regressions. (1 hour) - Zealer
- Preprocess data and design plot functions for Q2 (~5 hours) - Anny
- Design data structures or specific functions for Q3 - Forrest

Implementation (~20 hours each)
- Split data for each country and create a model for each of them, then compare the accuracy between models mathematically and visually. (15 hours) - Zealer

- Implement functions, create global Covid-19 map, barplot, and line plots, and make a website to visualize results for Q2 (~24 hours) - Anny
- Implement Q3 - Forrest

Testing (~10 hours each)

- Plot the actual trends of the pandemic along with the predictions to test accuracy. (5 hours) - Zealer
- Tests for Q2 with sample data of a specific month, and use Colab for intermediate outputs/tables/plots to ensure correctness (~5 hours) - Anny
- Implement tests for Q3 - Forrest

Evaluation:

The time estimates of the work plan from part 1 was pretty accurate. For machine learning, we followed exactly the steps listed by the work plan. The estimation of time was a little bit optimistic. We spent a lot of time debugging as well as looking up documentation.

As for RQ2, since we removed the ML component as we couldn't see any obvious correlation between geolocation and the number of cases, later we added in a new part to make a website for visualizing results. Throughout the process, we learned that changes are normal in code development, it is helpful to learn to accept and adapt changes.

Testing:

The testing for machine learning was using 70% training and 30% testing as well as using random states for each run and we ran our model for several times, the result was consistent. We also used plots accompanying our machine learning models by comparing the predictions and actual trends visually. Higher accuracy tends to have more overlaps with the original plots.

As for RQ2, we used Colab notebook as a playground for printing out intermediate outputs/tables/plots to ensure correctness. We also included a test file test_research-question-2.py and saved intermediate outputs to test_result/. By comparing table results and plot results, we may witness any errors if there is. Also I used the data for May as the sample data to ensure the correctness of plots.

Collaboration:

The main resources for machine learning and plotting used were sklearn, seaborn and matplotlib official documentation and *stackoverfow.com.* For building the HTML/CSS/JS webpage visualizing RQ2 results, *w3schools.com* was used as a reference.