COVID-19 Dataset

By Declan Sheehan & Jack Stoetzel

Hypothesis

Judging from the values in the dataset, the number of COVID-19 cases is getting worse in every country across the globe. Despite it getting worse, some nations are better off than other nations due to lower death and case rates.

Methods

We used the following machine learning algorithms for our hypothesis:

- k Nearest Neighbor
- Linear Regression

For kNN, we used the data provided to create values that represent the Covid-19 “Status” for each country. Then we used kNN to predict the status of new data.

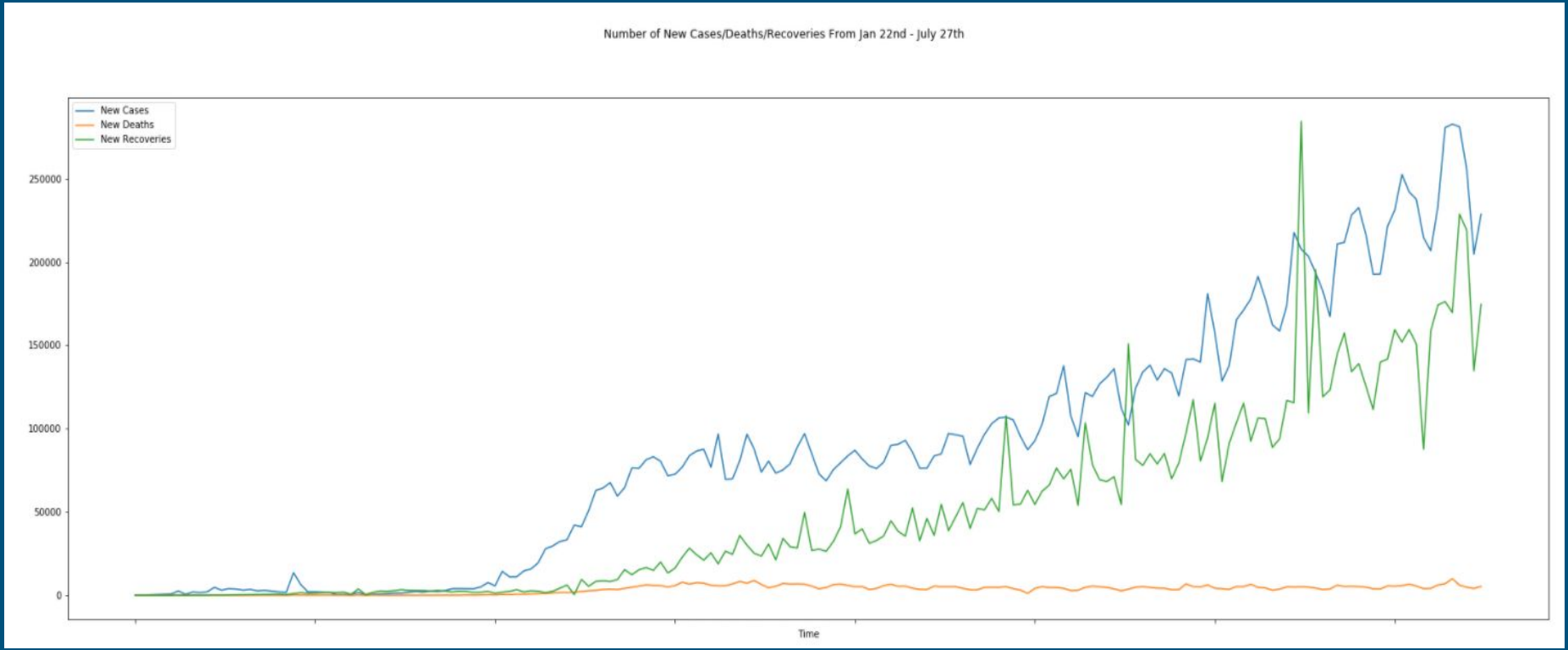
We graphed the cumulative cases for the world, and for different countries.

Our COVID-19 Dataset

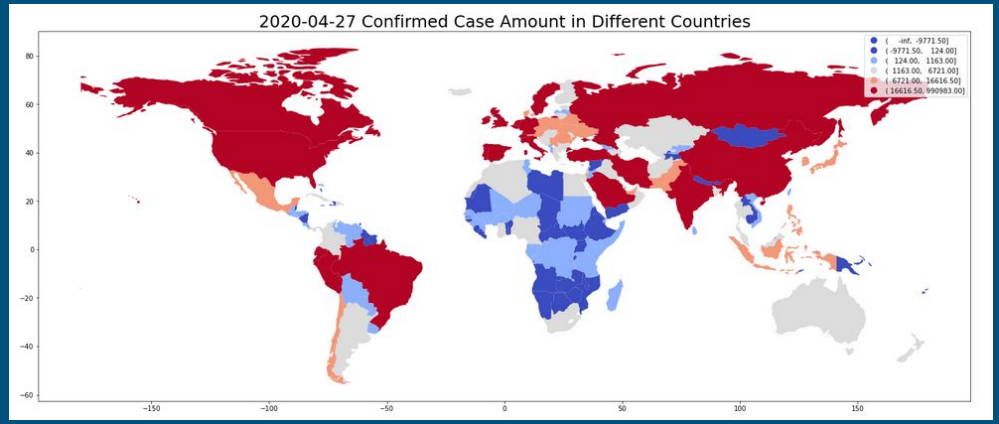
Our dataset for COVID-19 included the following from Jan 22nd, to Jul 27th:

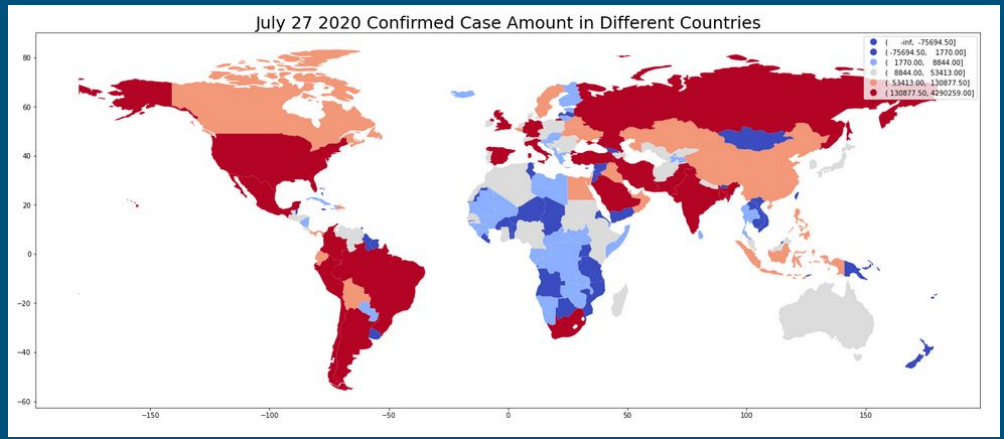
- Every day:
  - Country
    - Total Deaths, Total Cases, Total Recoveries
    - New Deaths, New Cases, New Recoveries
    - Active Cases
    - World Health Organization Region

New Cases/Deaths/Recoveries

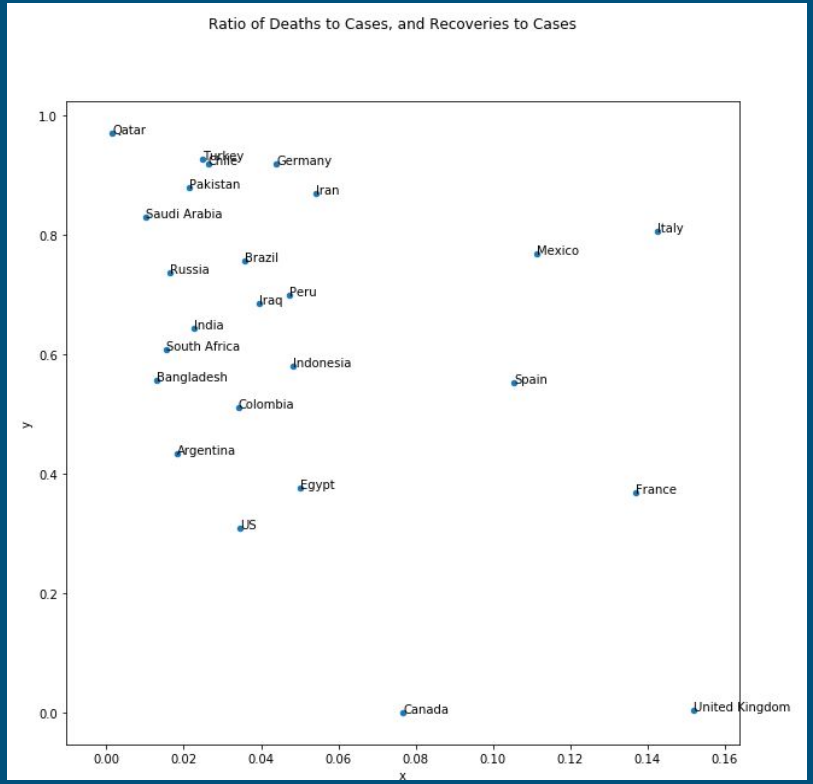


Map of Global Cases on July 27, 2020





Death Rates and Recovery Rates Visualized



k Nearest Neighbor Details

Features:

- The number of deaths proportional to the number of cases
- The number of recoveries proportional to the number of cases
- The number of cases proportional to the country's population.

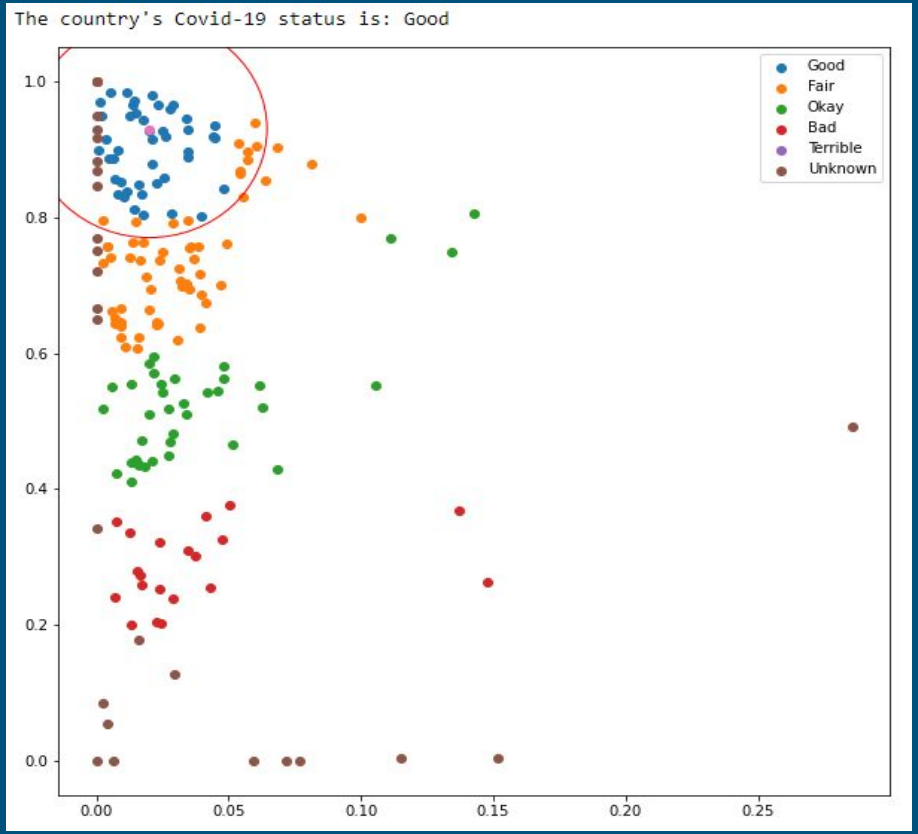
The "training" process of the kNN happens when we classify a datapoint based on its values. For example: Afghanistan is "Fair" with death rates=0.034, recovery rates=0.6948, and Cases/Population=0.01%.

The "testing" happens when we run the kNN on new values.

k Nearest Neighbor (2D)

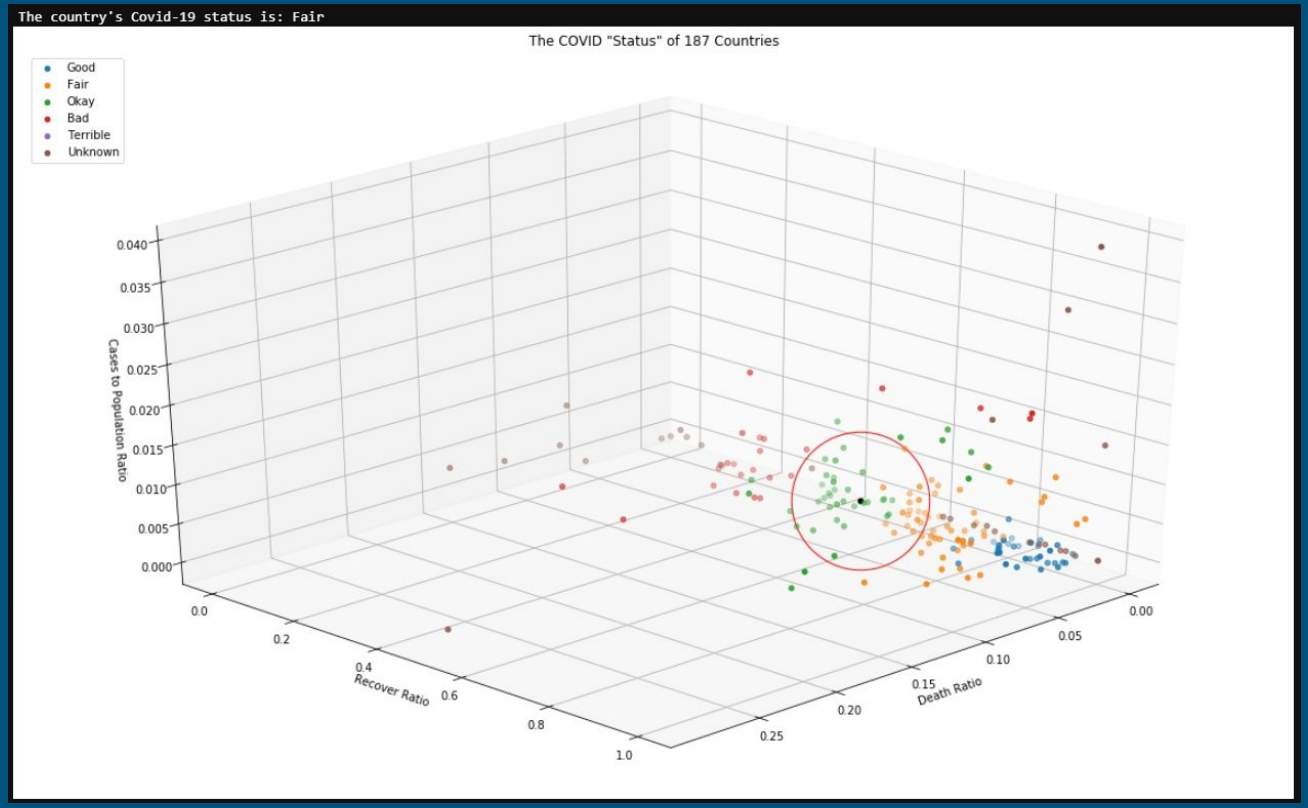
First we classified different countries based on **two** dimensions (death rates & recovery rates).

Our result is shown to the right.



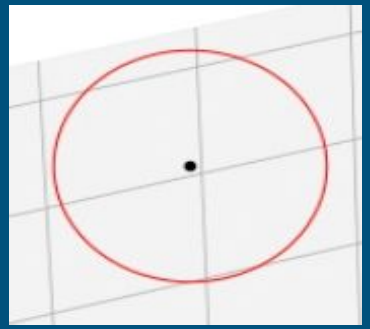
k Nearest Neighbor (3D)

We decided to add another dimension: Total cases / Country Pop

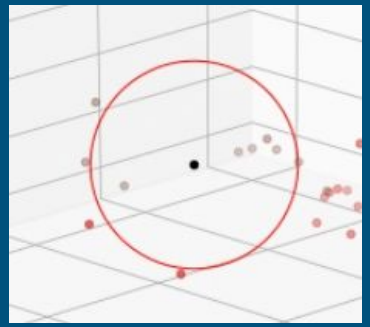


k Nearest Neighbors Accuracy

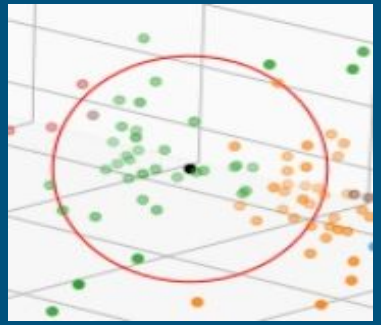
Terrible



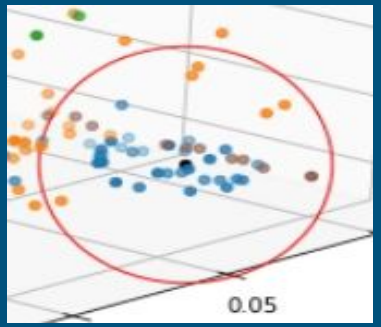
Bad



Fair



Good



Judging from the kNN prediction (k=11), the results are fairly accurate.

k Nearest Neighbor Next Steps

The next steps to validate kNN would to loop through a list of countries to see how the kNN classifies them. Then it could also be possible to compare our method of classifying to official (CDC) country classification to validate them.

There are also other variables we could place in each dimension to test which represents each country the best.

Linear Regression Details

Features:

- Projects future values for cumulative COVID-19 cases.

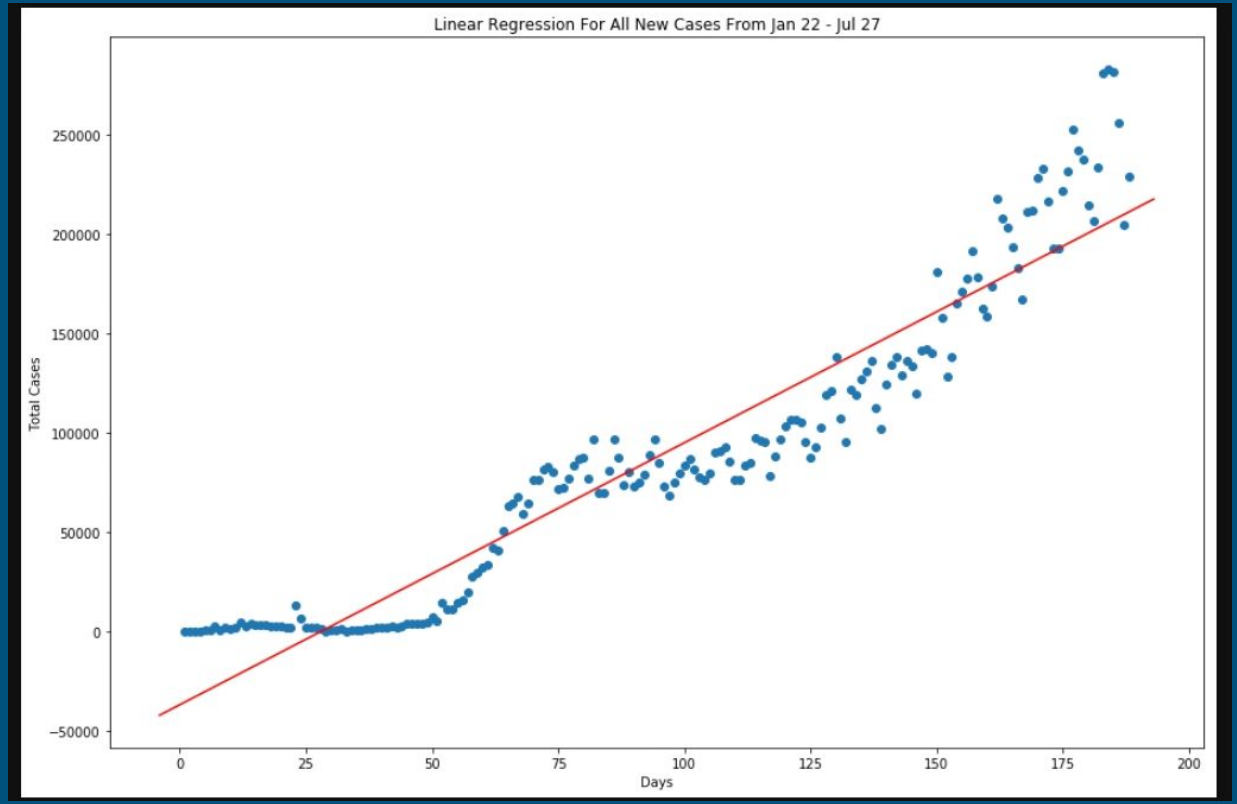
Here, the “training” process occurs when the current values are calculated using the linear regression formula.

Then the “testing” process occurs when we compare the slope of the linear regression to *actual* future values (to see if it is correct).

Linear Regression Visualized (Global)

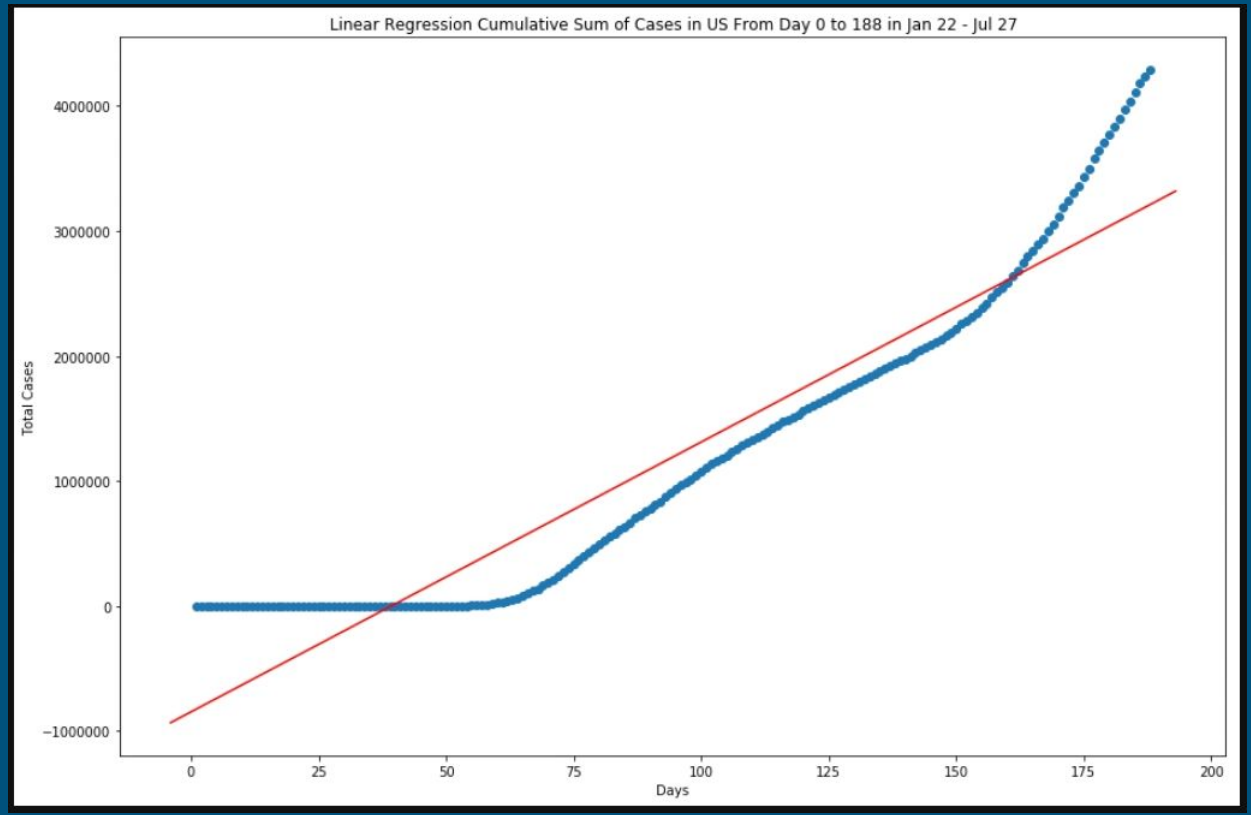
The linear regression for the cumulative COVID-19 cases globally

y=-36665.6976334+1316.7906763x

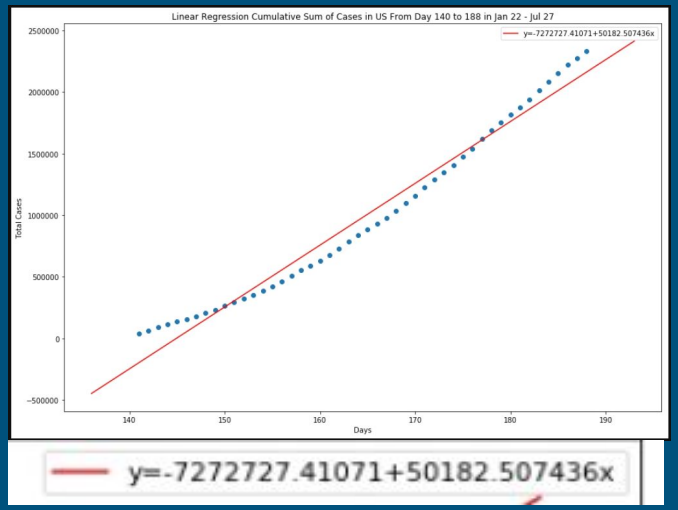


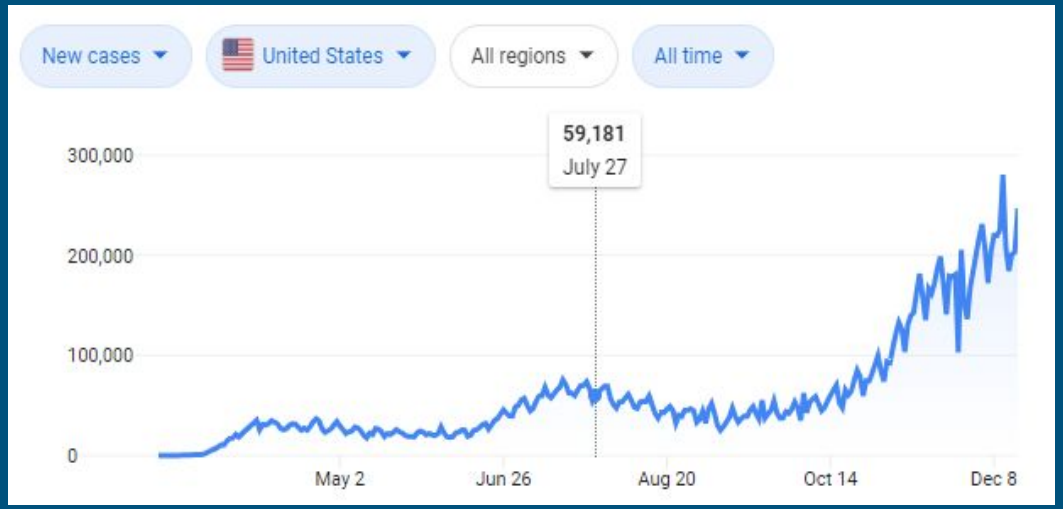
Linear Regression Visualized (USA)

The linear regression for the cumulative COVID-19 cases in the USA



Linear Regression Accuracy





When running the LR on the cumulative USA cases, the slope is 50,182.

Similarly, the number of *new cases* the next day is 59,181.

It is fair to say the Linear Regression is accurate (depending on the data sample).

Linear Regression Next Steps

The next steps for linear regression could be that we implement squared or cubed predictors to produce a graph with bends and curves. This would allow for a more accurate method of predicting future values in all intervals.

Another improvement to our data could be to add more up-to-date data.

Conclusion

In conclusion, our hypothesis holds true given our machine learning algorithms.

From a statistical standpoint, our algorithm's features *correlate* to real-life values. In addition, we do not believe our data to be statistically significant ($p<0.05$), so we would end up failing to reject the H_0.

Overall, the data does show that COVID-19 is getting worse, but some countries suffer less than others.

Fin

Questions?