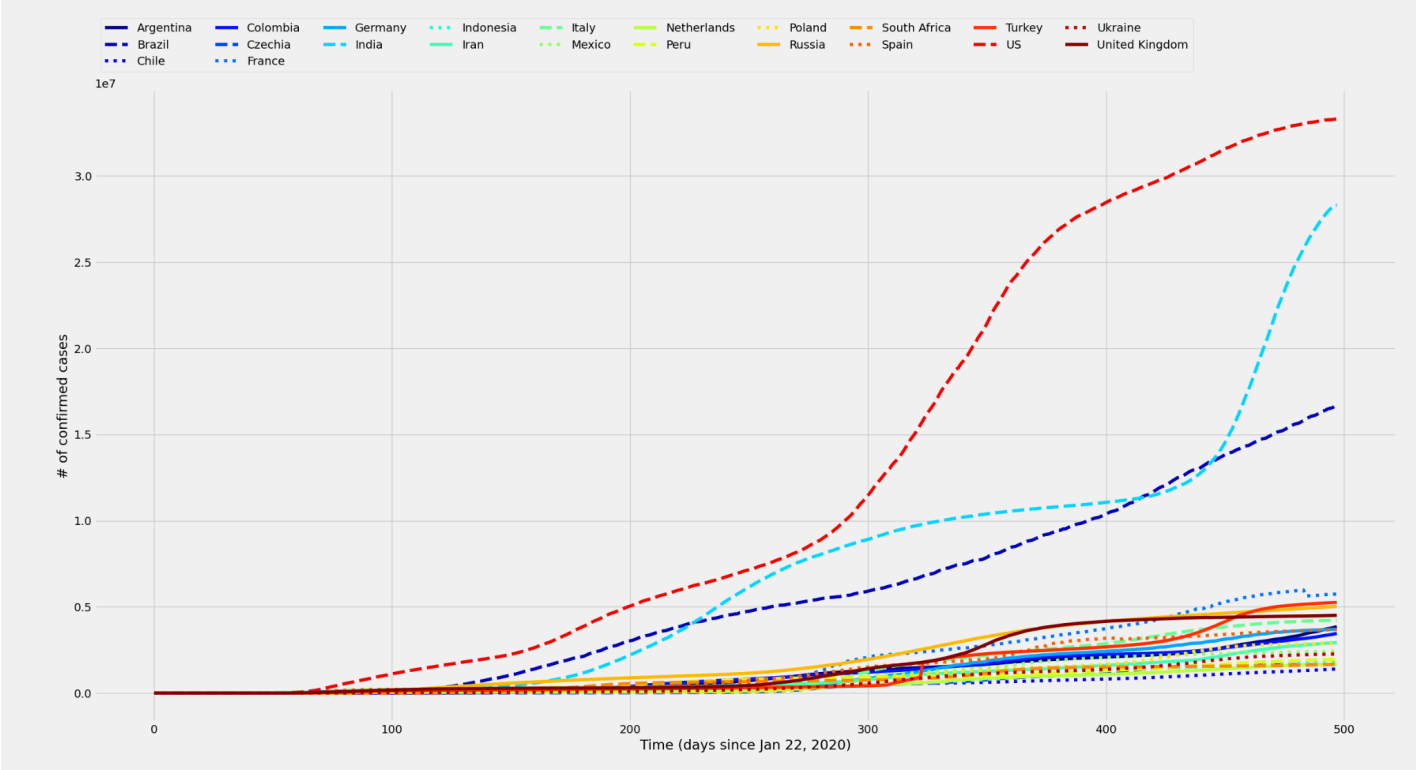
ML Final Assignment
Jack McCallie

For my final assignment in machine learning I decided to investigate a few countries and the effects of the hyper-parameters on their results. I was interested in finding the true closest neighbors to the US and seeing how many neighbors until the results seemed to stay consistent.

Here are the results for that investigation:

KNN raw (Manhattan):
Closest Neighbor to US: Denmark (1 neighbor)
Closest Neighbor to US: Canada(2-4 neighbors)
Closest Neighbor to US: China (5 neighbors)
Closest Neighbor to US: China (10 neighbors)



- I found it interesting that the closest neighbor of the US with k set to 1 is Denmark, especially considering the considerable differences in population and location. Also the final results for Denmark show that it is not a great prediction. This highlights the need for careful analysis of which K value to use.
- Also, China's final case number is very different from the United States, but China is what algorithm predicts as the closest when a decent amount of neighbors are used. The US and China had similar trajectories, but with such drastic differences in the final result

there must be significant differences in how the virus was handled at a certain point or how cases were reported. The graph above shows a visual of the real case number values over time to highlight the strong difference between the US and china. China is not even viewable on the graph due to not having enough cases.

The other experiment was in knn_dist_diff.py, where I decided to test the difference between altering the number of neighbors and number of bins to see if there are significant differences in outcomes.

Results:

KNN Diff Dist (Manhattan): 20 bins
Closest neighbor to us: Ethiopia (1 neighbor), population: 112 million
Closest neighbor to us: Angola (3 neighbors), population: 31 million
Closest neighbor to us: Angola (3-10 neighbors), population: 31 million
Closest neighbor to us: Canada (20 neighbors) population: 31 million
Closest neighbor to us: Canada (50 neighbors), population: 31 million

KNN Diff Dist (Manhattan): 5 Neighbors
Closest neighbor to us: Czechia (5 Bins) population: 10 million
Closest neighbor to us: Ethiopia (10 Bins), population: 112 million
Closest neighbor to us: Ethiopia (15 Bins), population: 112 million
Closest neighbor to us: Angola(20 Bins) Population: 31 million
Closest neighbor to us: Armenia(50 Bins) population 3 million

From the results of the experiment there are some significant differences. We see that the closest neighbor still changed after increasing the bins from 20 to 50, but did not change when increasing k from 20 to 50. This result makes sense as the decision boundary will be smoother for the greater k values, but when running the experiment using 50 neighbors it took significantly greater computation time showing a tradeoff with increasing the amount of neighbors. Computation time could become a serious problem if the data set and dimensionality were large enough that the optimal value of K was not feasible to use for predictions. However, in this case that is not a serious problem.