Johns Hopkins University

EN.685.621.81.SP20 Algorithms for Data Science

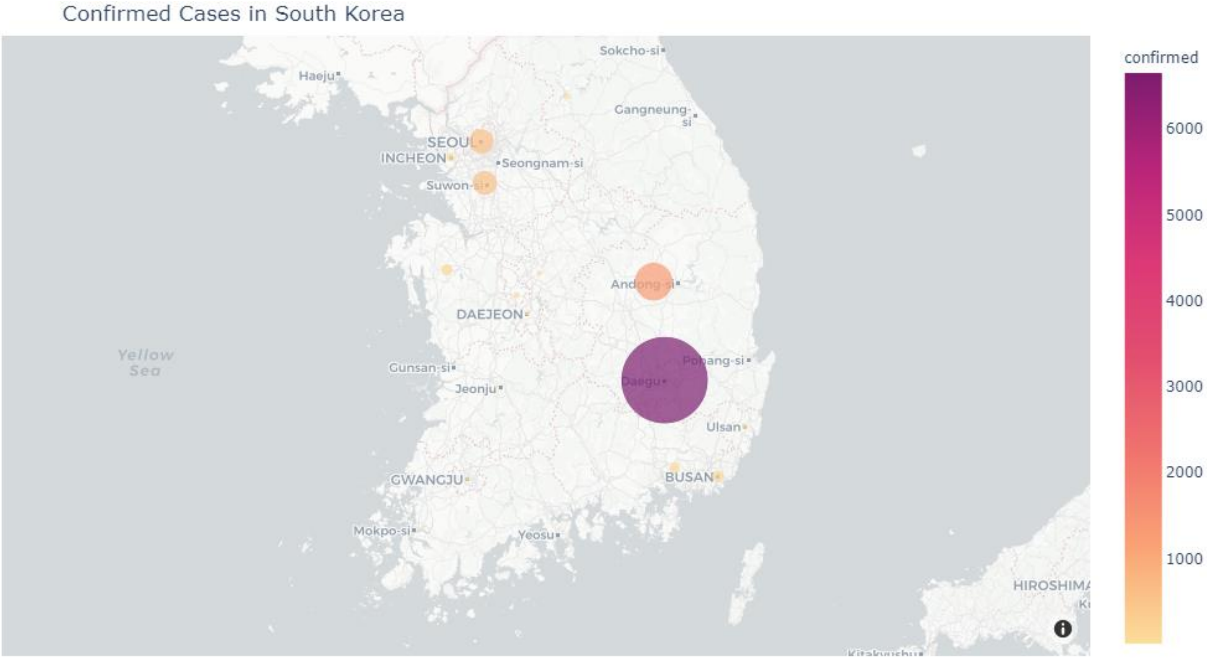
Alternate Programming Assignment 2

Author: Ivan Sheng

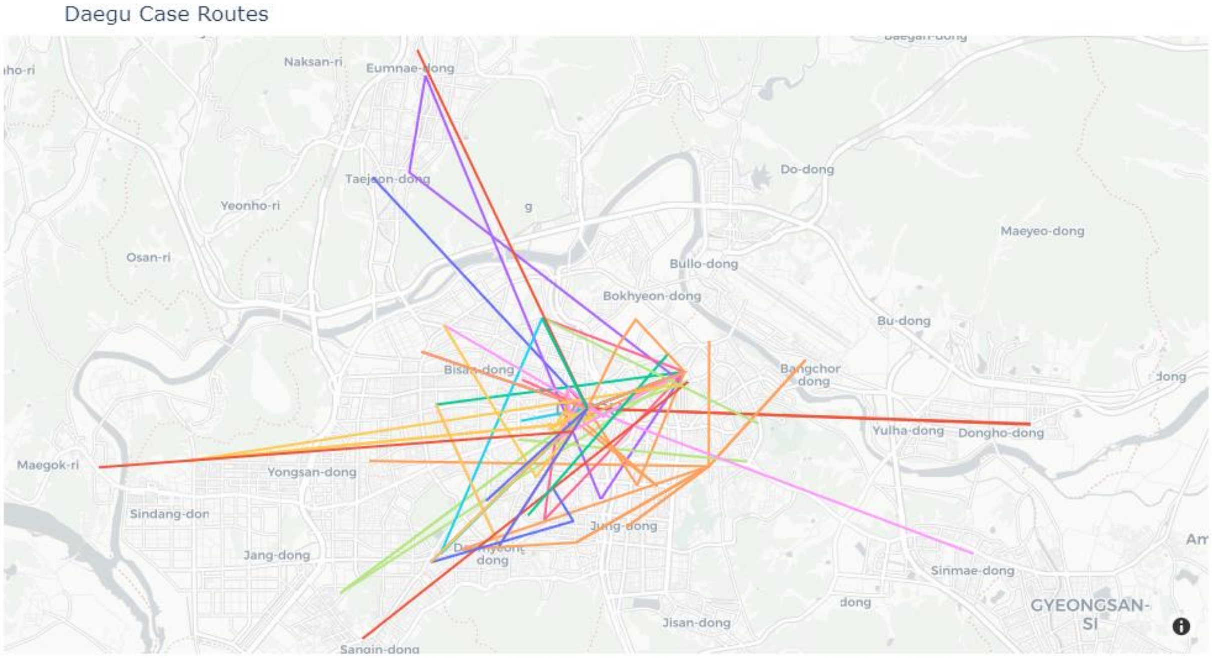
Telephone: 718-839-0967

Date of Submission: 5/12/2020

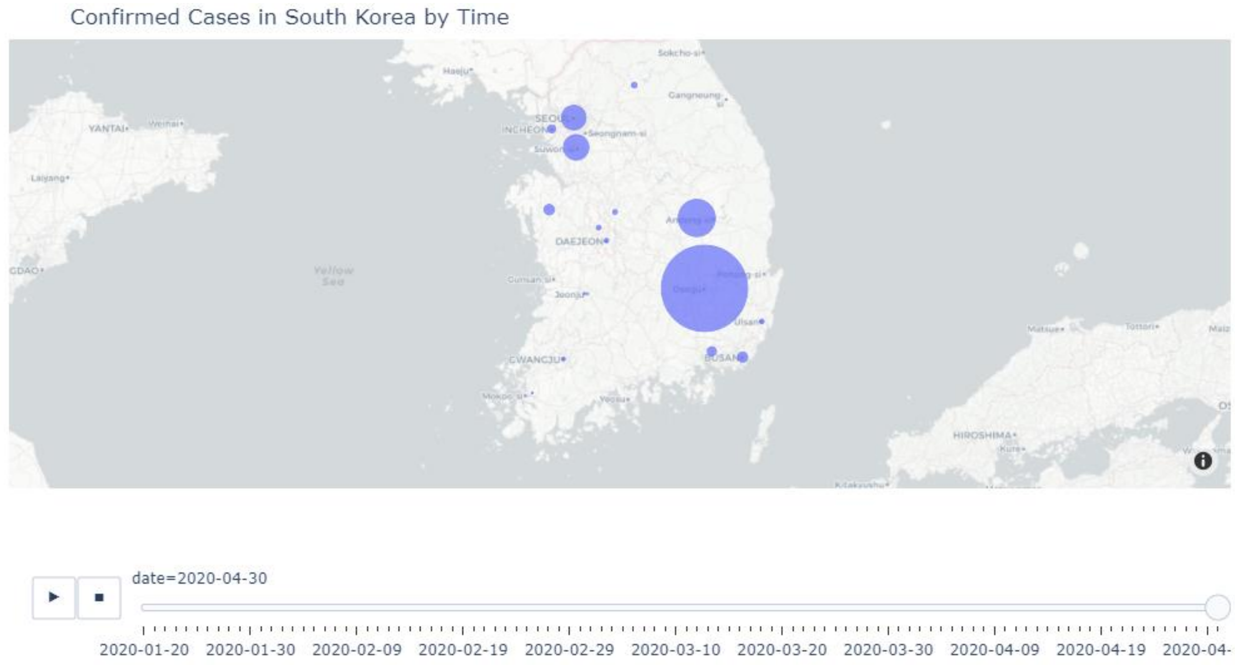
Problem 1-1



Problem 1-2



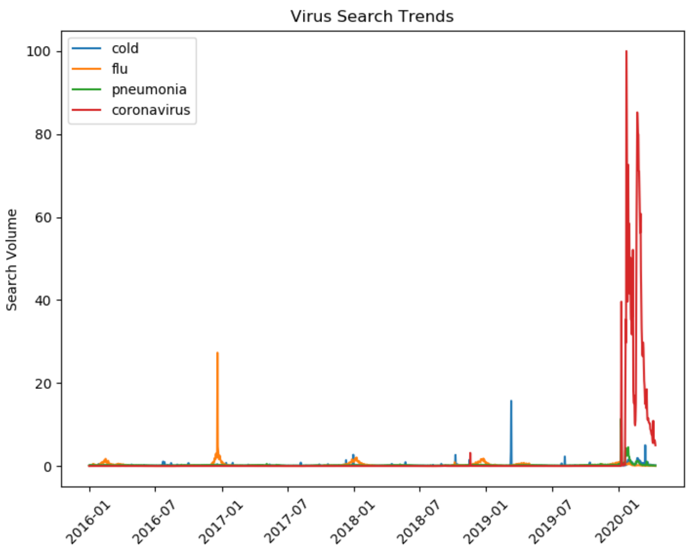
For performance reasons, the number of routes displayed were limited. The Daegu region was specifically selected because of the Shincheonji Church of Jesus incident.



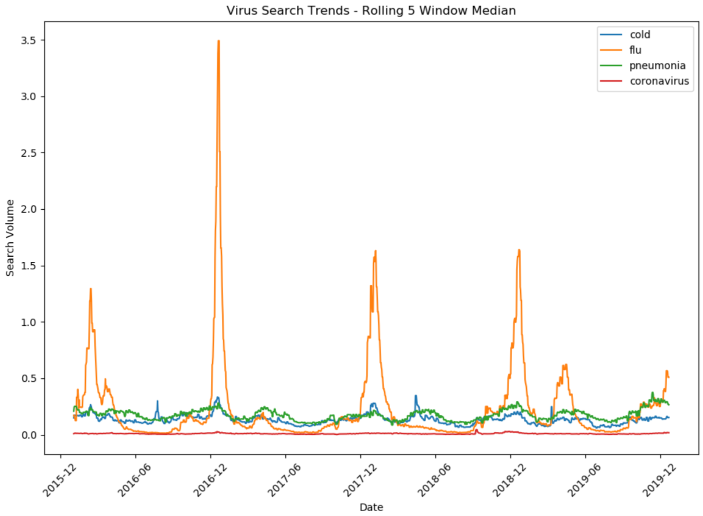
Problem 1-3

Looking at the raw search data, it's difficult to notice any details because Coronavirus searches in South Korea have drastically increased in volume compared to historical search volume of the cold, flu, and pneumonia. However, three things to note are the spikes in flu searches near the end of 2016, cold searches around the first quarter of 2019, and pneumonia searches mimicking Coronavirus searches in 2020.

- The spike in flu searches can be attributed to the bird flu outbreak, H5N8, that affected South Korea.
- The spike in cold searches in 2019 could possibly be attributed to the rising tensions and mentions of a Cold War between the US, North Korea, and South Korea during March.
- Because pneumonia is telling symptom of coronavirus, its search trends will mimic coronavirus searches.

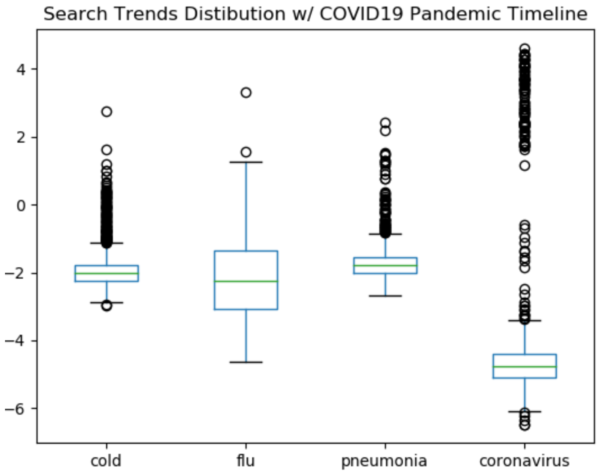


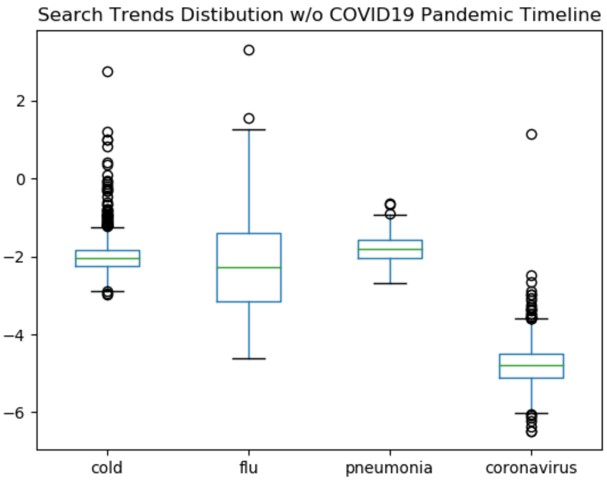
To analyze the data in a view that would allow trends to show, the axes must be set to cut off the Coronavirus pandemic, and the data itself should be smoothed out by applying a rolling n-window median.



After these techniques were applied, it can be observed that there is an obvious seasonal pattern for flu searches that occurs at the end of the year, with the exception of 2019 where there was an increase in searches within the first half of the year. As for cold and pneumonia searches, they peak more often than flu searches – approximately every half year – but their second peaks line up with flu search peaks in December.

Coronavirus has a standard deviation that's almost 5 times its mean. The mean is also significantly higher than cold, flu, and pneumonia despite its quartile limits being less than the other three. This is probably due to the fact that its max reaches 100, while the other three don't even come close. After looking at both the first order test statistics along with the line plots, a box and whisker plot that includes the Coronavirus timeframe would include significant outliers, so two box plots will be created: with and without the pandemic timeline.



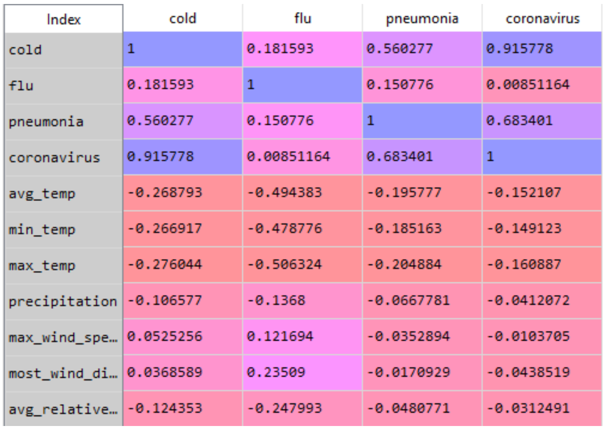


The number of outliers above the max whisker for coronavirus searches is significant, which is expected from looking at the 2020 search data. Because a relationship between pneumonia and coronavirus searches was observed during the 2020 timeframe, it makes sense that almost all of the outliers for pneumonia disappeared when that timeframe was cut off. Unfortunately, there still remains a significant number of cold search outliers.

Problem 1-4

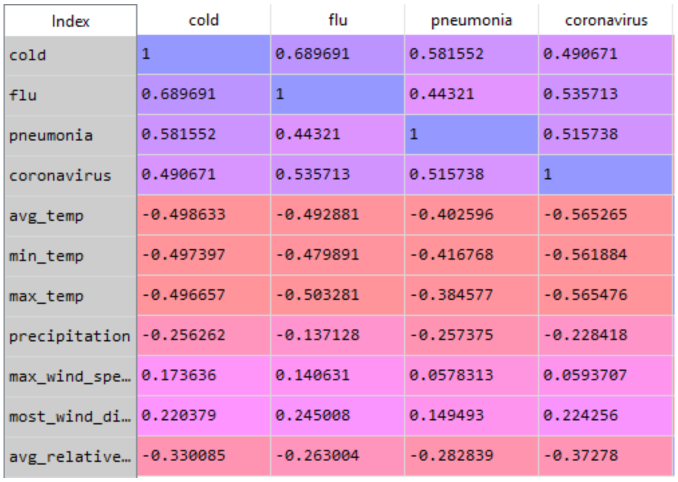
Seasonality was identified in the previous analysis, but are there any correlations that would show a relationship between virus search trends and weather patterns?

As mentioned previously, a 5-window rolling median had been applied to the search data to smooth it out – the same procedure will be applied to the weather data as well to filter and smooth out the noise seen in sporadic daily data.



Index	cold	flu	pneumonia	coronavirus
cold	1	0.181593	0.560277	0.915778
flu	0.181593	1	0.150776	0.00851164
pneumonia	0.560277	0.150776	1	0.683401
coronavirus	0.915778	0.00851164	0.683401	1
avg_temp	-0.268793	-0.494383	-0.195777	-0.152107
min_temp	-0.266917	-0.478776	-0.185163	-0.149123
max_temp	-0.276044	-0.506324	-0.204884	-0.160887
precipitation	-0.106577	-0.1368	-0.0667781	-0.0412072
max_wind_spe…	0.0525256	0.121694	-0.0352894	-0.0103705
most_wind_di…	0.0368589	0.23509	-0.0170929	-0.0438519
avg_relative…	-0.124353	-0.247993	-0.0480771	-0.0312491

At first, there isn’t anything too strong between weather data and search trends. Generally, there is a negative relationship between temperature data and flu searches: as temperature decreases, flu searches increase, however, nothing else is particularly notable. It’s worth mentioning that strong correlations are observed between Coronavirus, the common cold, and pneumonia searches because when looking at the search trend graphs, these two viruses were also searched in conjunction with the rise of Coronavirus searches.



Index	cold	flu	pneumonia	coronavirus
cold	1	0.689691	0.581552	0.490671
flu	0.689691	1	0.44321	0.535713
pneumonia	0.581552	0.44321	1	0.515738
coronavirus	0.490671	0.535713	0.515738	1
avg_temp	-0.498633	-0.492881	-0.402596	-0.565265
min_temp	-0.497397	-0.479891	-0.416768	-0.561884
max_temp	-0.496657	-0.503281	-0.384577	-0.565476
precipitation	-0.256262	-0.137128	-0.257375	-0.228418
max_wind_spe…	0.173636	0.140631	0.0578313	0.0593707
most_wind_di…	0.220379	0.245008	0.149493	0.224256
avg_relative…	-0.330085	-0.263004	-0.282839	-0.37278

When looking at the data before the Coronavirus pandemic began, there are negative correlations similar to the flu: as temperatures decrease, searches for all four viruses will increase. There is also not as strong of a correlation between Coronavirus, the common cold, and pneumonia.

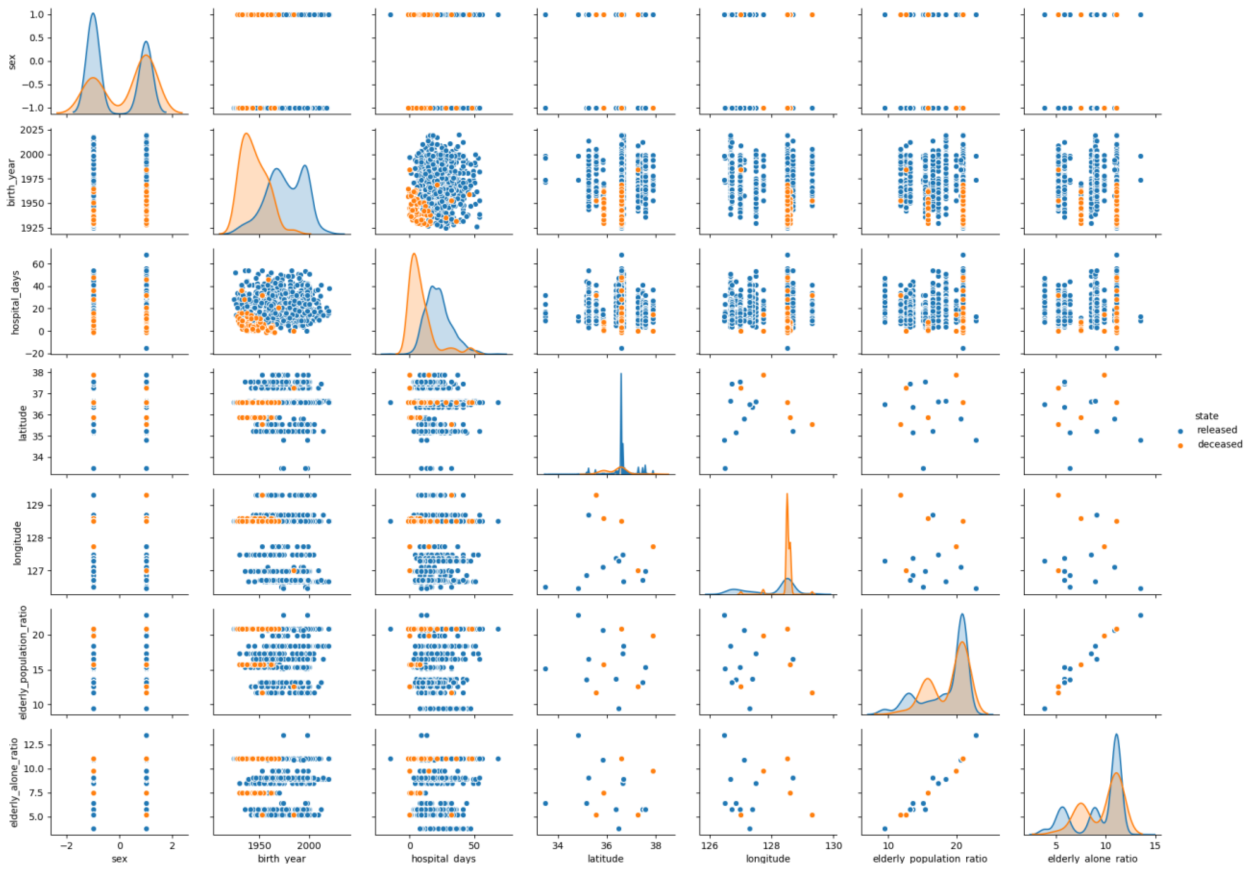
Problem 1-5

When filtering the patientinfo dataset for only those who have passed due to from the, there were only 63 patients who had both confirmation date and date of passing. Of those 63 patients:

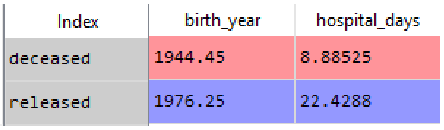
- 39 were from the province of Gyeongsangbuk-do, which has the second highest average elderly population ratio and elderly alone ratio.
- 20 were from Daegu, which was the first Coronavirus epicenter outside of China due to Shincheonji Church of Jesus followers who purposefully spread the virus.

Looking at the pairplots of the patientinfo dataset filtered to patients who had both a date of confirmation and date of release or passing, it seems that elderly population and alone ratios are actually poor class separators for released and deceased. A significant majority of deceased patients fall within very similar longitudes and latitudes – this was previously referenced as Daegu (35.872, 128.602) and Gyeongsangbuk-do (36.576,128.506).

Both birth year and days since confirmation seem to be good separators – it seems that those who have passed weren’t diagnosed for long and were older in age.



Diving a bit further into birth year and days since confirmation, the average number of days since confirming a case before passing is 8.88 days with an average birth year of 1944 (76 years old) while a release is 22.42 days with an average birth year of 1976 (54 years old). This raises concerns because bed occupancy would take a bit more than two-thirds of a month, which could overload hospitals if nothing was done to slow down the rate of spread.



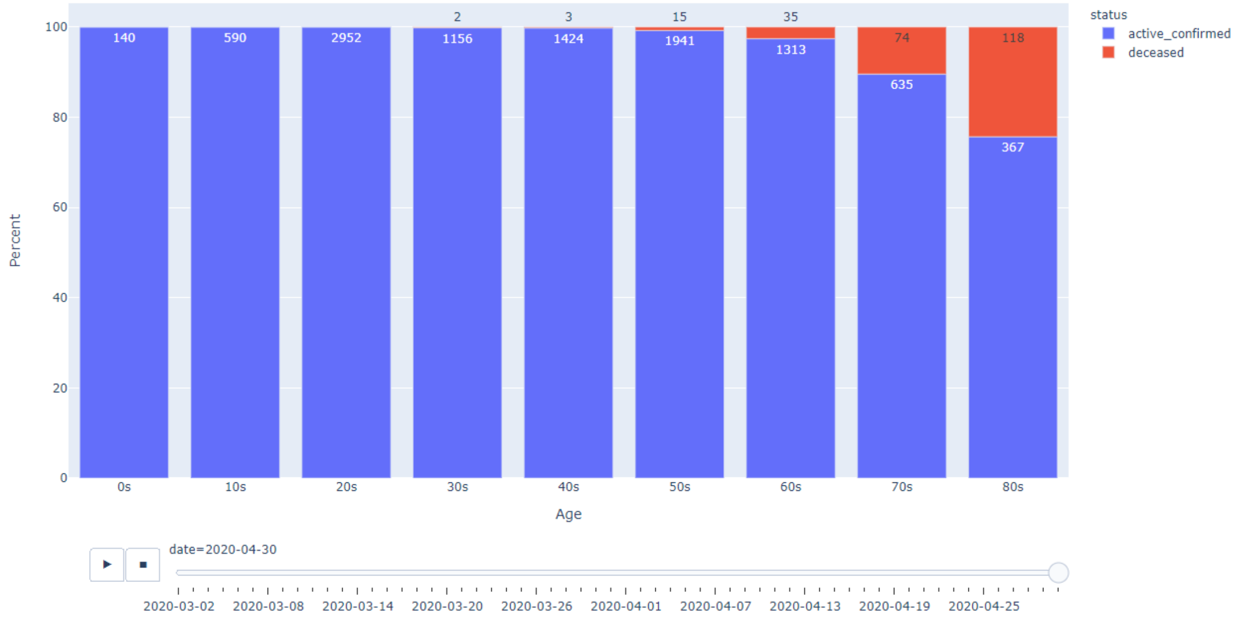
Index	birth_year	hospital_days
deceased	1944.45	8.88525
released	1976.25	22.4288

Graphing out the number of confirmed cases and fatalities in aggregate and by time, the data supports the observation that the elderly are more susceptible to succumbing to COVID-19. A couple things pop out:

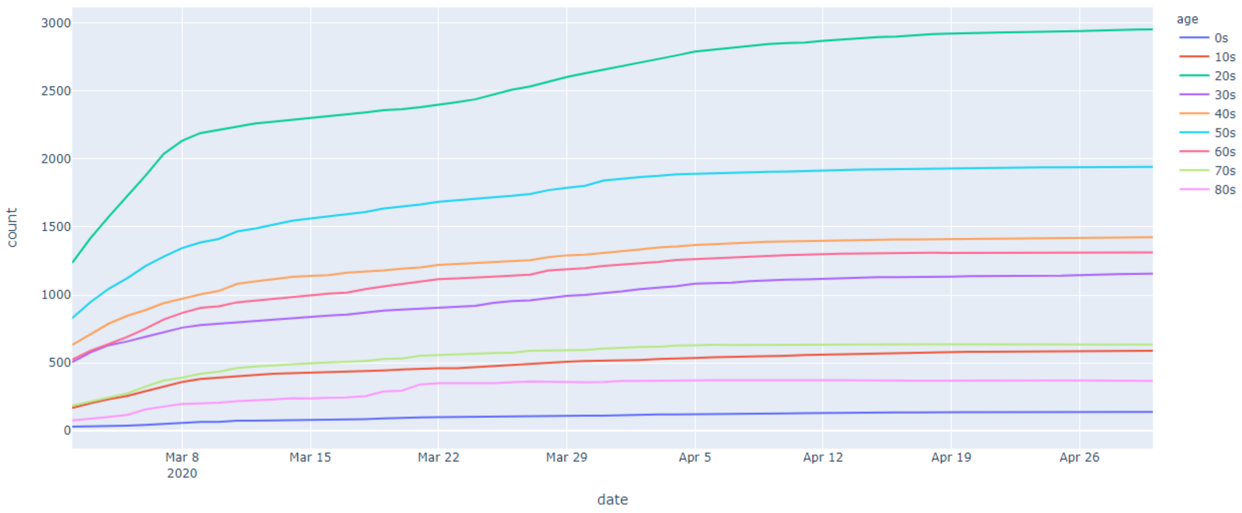
- In comparison to global death rates from Statista as of February 11th 2020, South Korea has a significantly higher mortality rate of their 80+ population [1].
  - The global rate for 80+ is 14.8% versus South Korea's 24.3%
- South Korea has a high number of confirmed cases for the age group of 20-29 years old.
  - In comparison, Canada's highest distribution of cases by age group is between 50-59 years old and 80+ years old [2].
- While there have been more confirmed cases for females, the number of fatalities is less than those of males who have approximately 2,000 less cases.

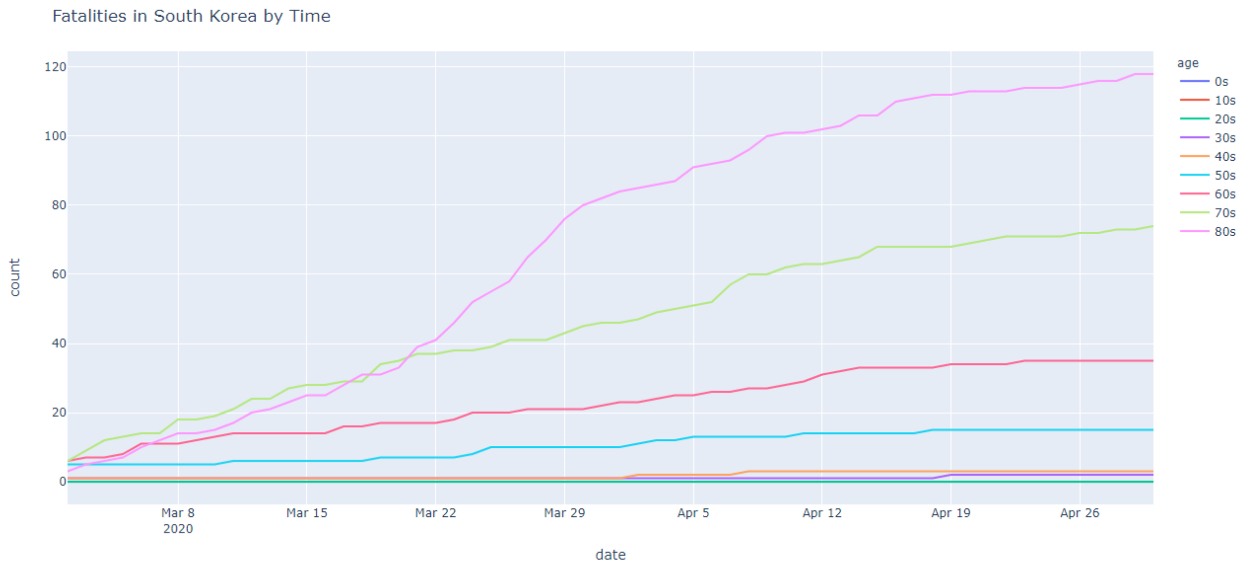
- ○ Regardless, the looking at the data by gender doesn't show whether or not one gender is more susceptible than the other since the percent deceased is already so small.

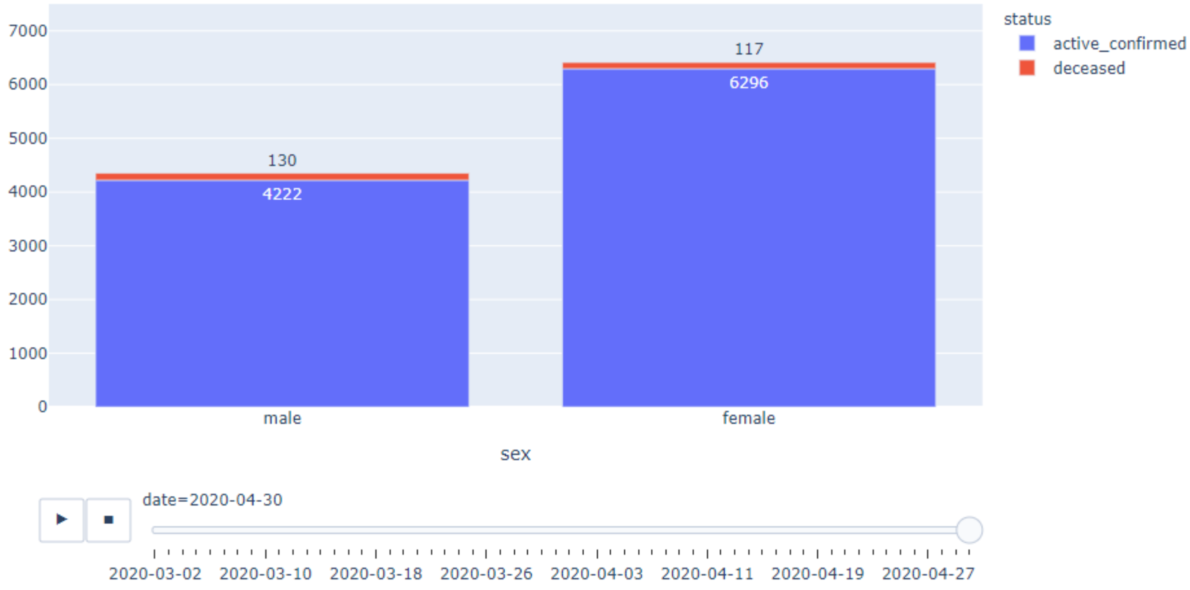
Active Confirmed Cases vs Fatalities in South Korea



Actived Confirmed Cases in South Korea by Time







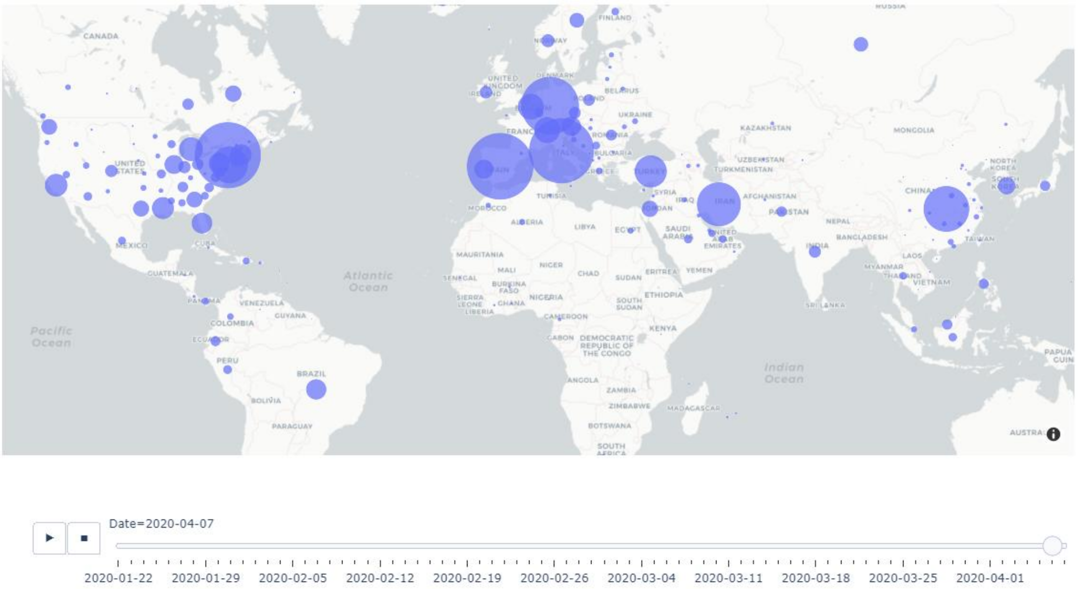
Runtime Analysis of 1-5 Algorithms

- The most time-consuming line of the algorithm used to prepare the data for visualization would be joining different dataframes. While there's no documentation on the time complexity of the pandas join() function, under the assumption that the join function compares every record in one dataframe to another the complexity should be nm.
- Fillna() and map() should take a time complexity of n
- Groupby() would take nlogn if sorting was necessary.

Since the bulk of the data preparation was performed at the beginning of problem 1-5; the runtime of that section would appear to be T(n) = nm +nlogn + bn + c, which would result in an a big-O of O(nm), since nm would be our largest growth factor.

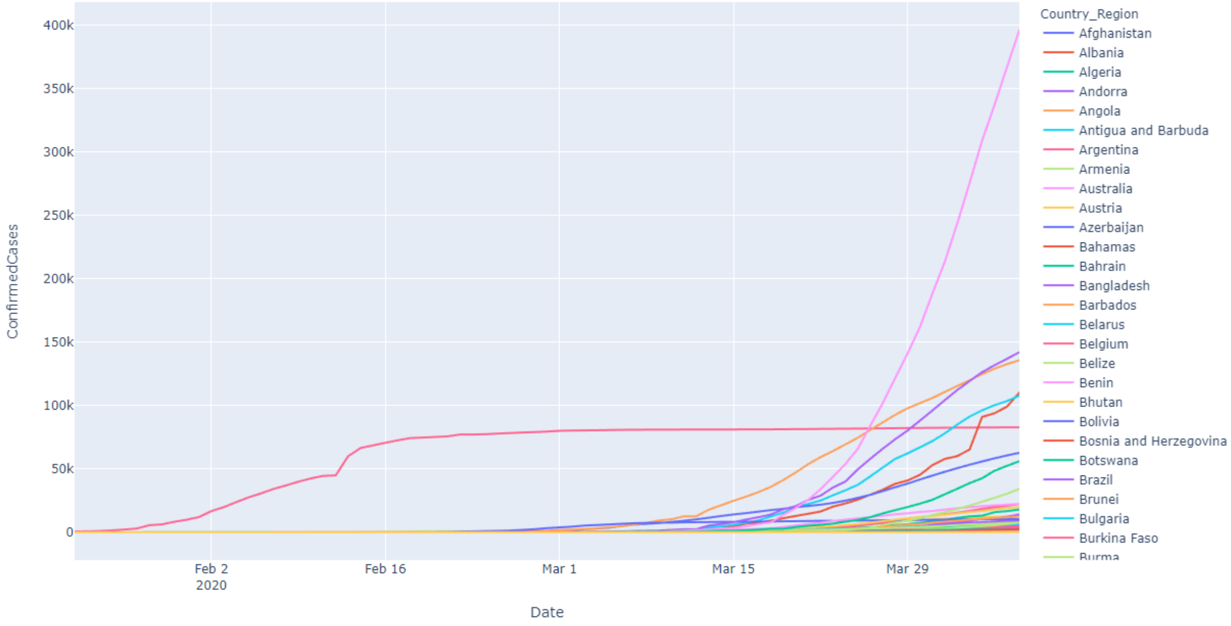
Problem 2-1

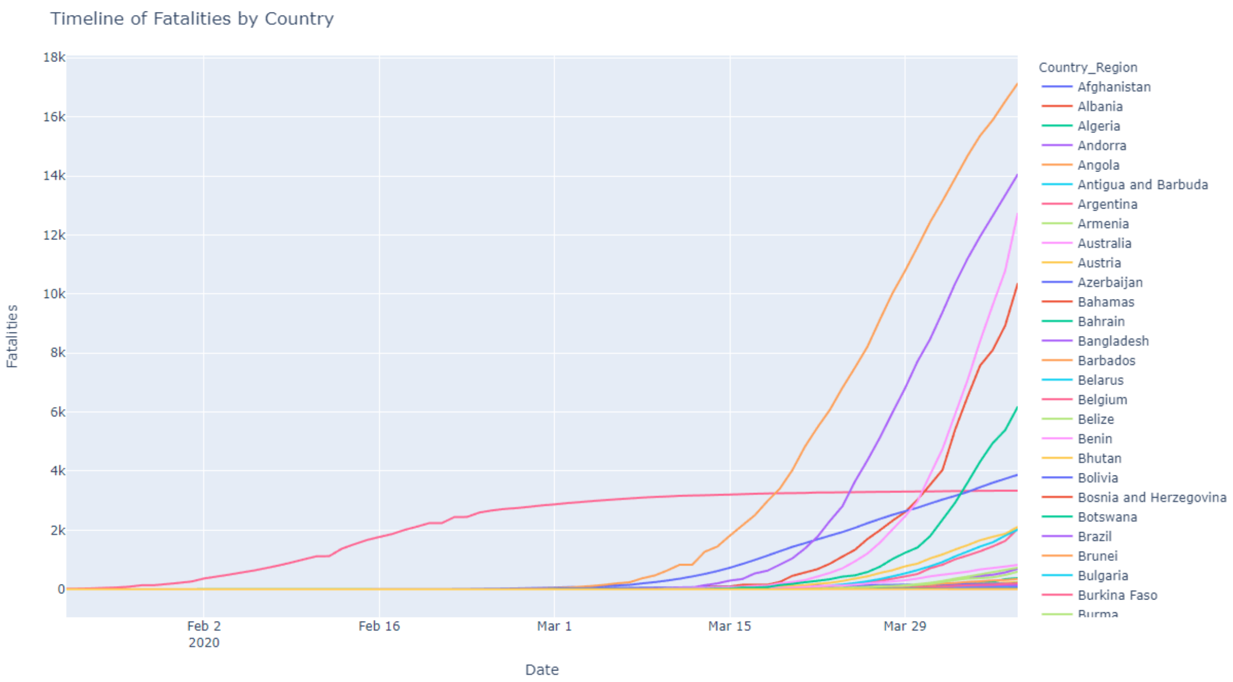
COVID19 GLobal Spread



Problem 2-2

Timeline of Confirmed Cases by Country

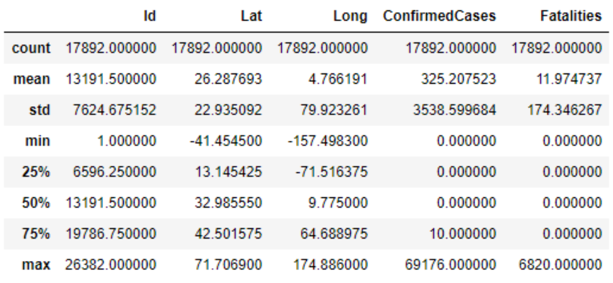




Problem 2-3

Looking at the timeline of confirmed cases by country, every country except the US and France appear to be experiencing leveling off in daily confirmed cases – these countries' growths are beginning to take a logistic shape instead of exponential. With the addition of the United Kingdom, similar trends can be seen when looking at daily fatalities.

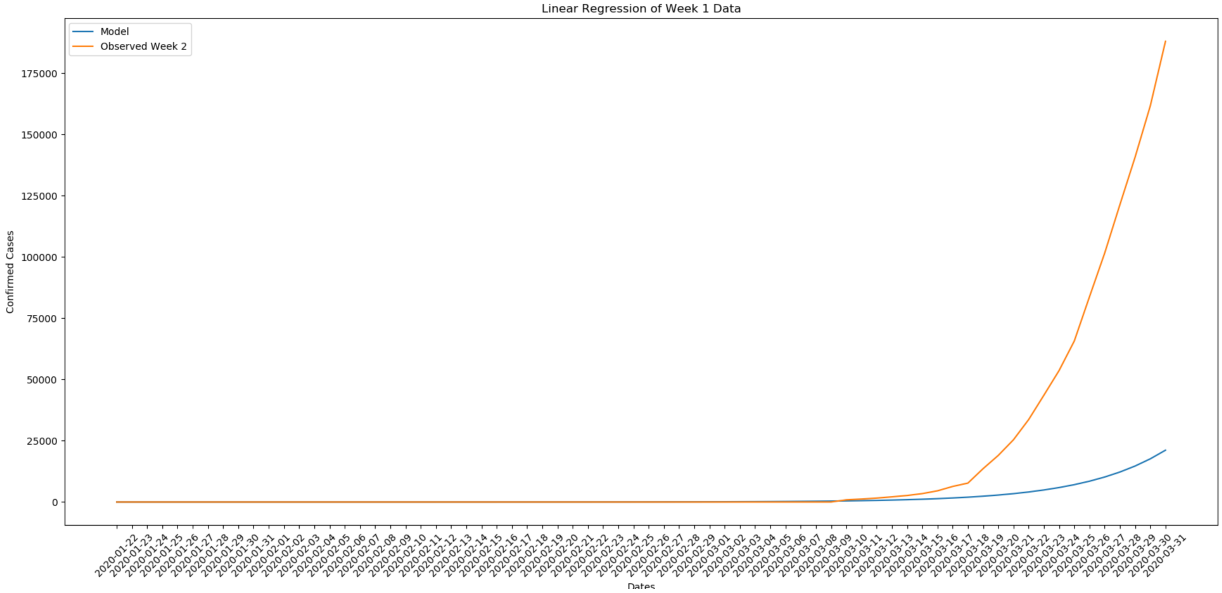
Analyzing the first order statistics for week 1 training data, most of the confirmed cases don't appear until the 3rd quartile because the data for week 1 begins tracking since late January 2020, when the large majority of documented confirmed cases were in China. Since most of the data will fall in the last quartile and each country was infected at different time periods, it'd be best to look at countries isolated since virus spread can be modeled as exponential growth in the early stages, and logistic later on. We can expect the US, UK, and France to follow exponential growth, and the rest of the world to follow logistic growth based on the line graphs.



	Id	Lat	Long	ConfirmedCases	Fatalities
count	17892.000000	17892.000000	17892.000000	17892.000000	17892.000000
mean	13191.500000	26.287693	4.766191	325.207523	11.974737
std	7624.675152	22.935092	79.923261	3538.599684	174.346267
min	1.000000	-41.454500	-157.498300	0.000000	0.000000
25%	6596.250000	13.145425	-71.516375	0.000000	0.000000
50%	13191.500000	32.985550	9.775000	0.000000	0.000000
75%	19786.750000	42.501575	64.688975	10.000000	0.000000
max	26382.000000	71.706900	174.886000	69176.000000	6820.000000

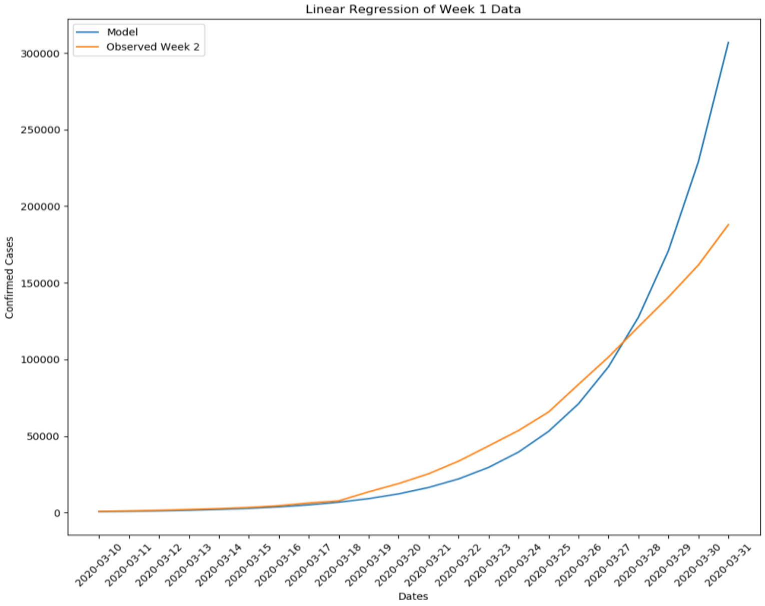
Problem 2-4

A linear regression was performed on week 1 US data after log transforming the y-axis: confirmed cases of COVID19. Initially this regression was performed on data for all available dates even if there were no recorded cases yet. This resulted in the following model:



Summary Statistics	Value
R^2	0.819
P-Value	< 0.001
RMSE	9372.59

Despite the high R^2 value, the linear regression seems to be a poor predictor for COVID19's spread in the United States; this could be attributed to the significant amount of time that the data remained at zero. In an attempt to improve the resulting linear regression, the data can be cut to only preserve data starting from when confirmed COVID19 cases began appearing. Looking at the regression results, there are improvements across the board with a higher coefficient of determination, as well as an RSME that's almost half the pervious iteration's RSME.

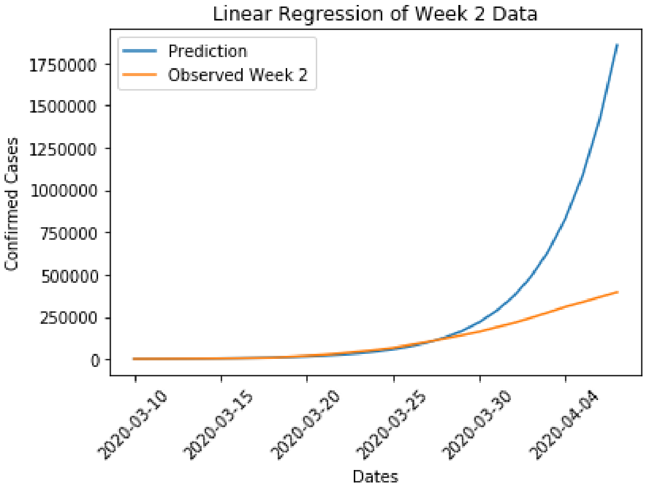


Summary Statistics	Value
R^2	0.900
P-Value	< 0.001
RMSE	4672.89

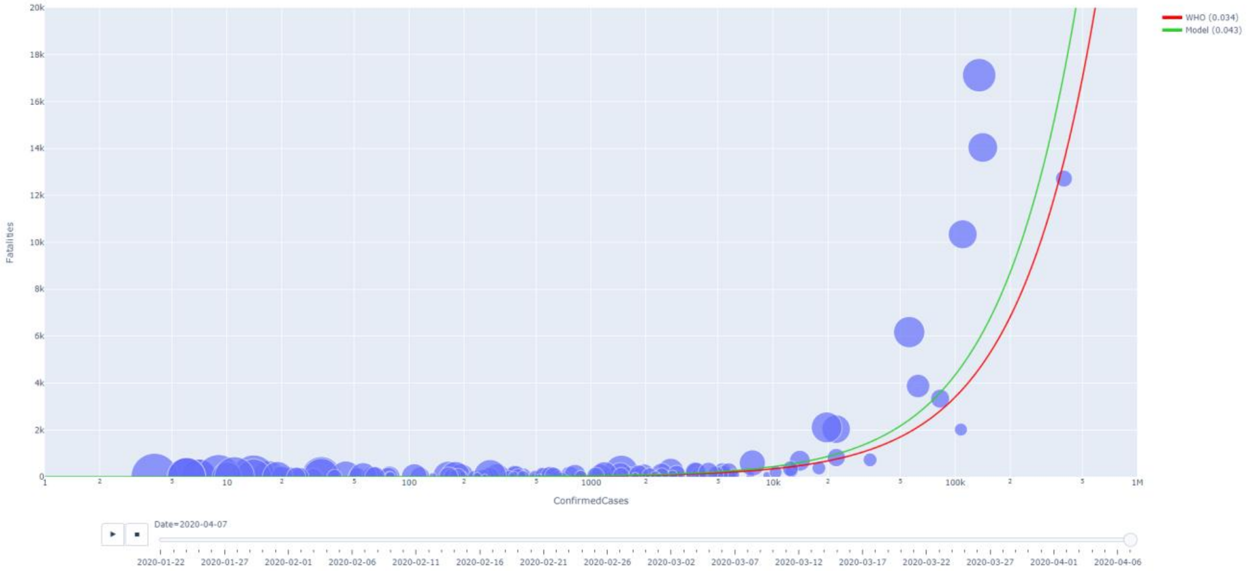
Based on our improved linear regression, the actual rate of confirmed COVID19 cases appears to be less than what the model is predicting by March 31st. The decreased confirmation rate also begins shortly after quarantine measures were taken on March 19th, however week 3 should be observed to confirm that the number of confirmed cases is continuing to flatten.

Problem 2-5

Comparing the predictions for week 3 with the improved regression versus the actual observed data for week 3, social distancing policies seem effective since the actual spread of Coronavirus is no longer following exponential growth.



According to WHO Director-General, it was reported on March 3rd that, “globally, about 3.4% of reported COVID-19 cases have died” [3]. Graphing out the log of confirmed cases versus fatalities, it can be observed that while the majority of countries fall fairly close to that 3.4% mortality line, by April, a better fit that was determined by a linear regression of confirmed cases and fatalities, would be 4.3% mortality. Interestingly, an article published by the New York Times on April 17th also cited a 4.3% mortality rate [4].



Around March 16th is the timeframe where the global data no longer follows a 3.4% mortality rate. Looking at the animation, it's observed that Italy, Spain, France, and the United Kingdom began to report higher mortality rates.

Runtime Analysis of Data Generation:

The algorithm created to forecast Coronavirus cases utilized linear regressions. According to slide 59 on the lecture slides found from KTH Royal Institute of Technology in Stockholm, Sweden the time complexity of linear regression would be O(d^3+nd^2), where d is the number of features and n is the size of the sample. This would have a runtime of T(n)= d^3+nd^2[5].

Problem 2-6

Between SVM and LDA, SVM created a more accurate model to classify confirmed cases with the features being latitude and longitude for week 1 and week 2 data. Date wasn't an included feature because the resulting scores for both classification models reported scores of over 95%. When scoring on week 5 data, the score was around 52%, which is a telling sign that the model was overfitted when the date feature was included.

However, using only latitude and longitude as the features resulted in models that only scored 60% for SVM and LDA. This is due to the fact that the model is now only broadly classifying based on location only instead of location and time.

The final iteration of the X variables were latitude, longitude, and the number of confirmed cases. For the week 1 data, this resulted in a score of 90% using SVM and 60% using LDA when tested on future dates. Week 2 also resulted in a high score of 90% for SVM, but an even poorer 59% for LDA.

Week 1 SVM Score: 0.904	
58095	772
8687	31075

Week 1 LDA Score: 0.602	
58012	855
38391	1371

Week 2 SVM Score: 0.901	
56015	995
8726	32893

Week 2 LDA Score: 0.593	
55844	1166
38986	2633

In regards to classifying fatalities, the probability of confirmed cases as well as dates, was added as a feature, but in the case of both SVM and LDA, there were negligible differences. Favorable results came from adding the number of fatalities as a feature - similar to confirmed cases classification - both SVM and LDA models resulted in a score of approximately 90%. Week 2 data resulted in stronger scores when testing with week 5 data: SVM scored 93%, while LDA scored 88%.

Week 1 SVM Score: 0.897	
79521	8732
1455	8921

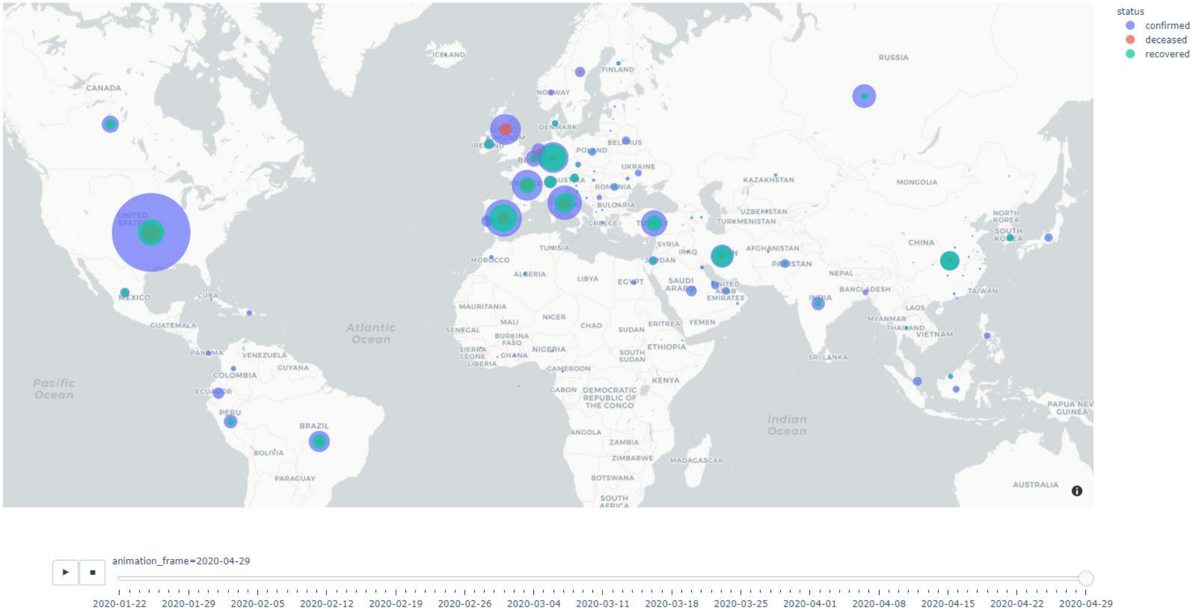
Week 1 LDA Score: 0.896	
88251	2
10299	77

Week 2 SVM Score: 0.929	
80741	5955
1025	10908

Week 2 LDA Score: 0.879	
86162	534
11361	572

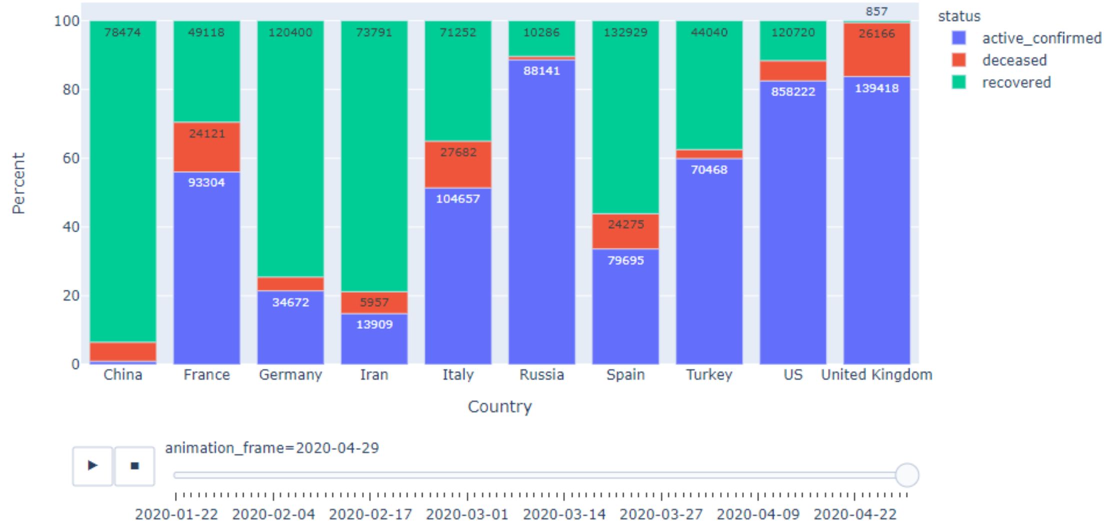
The weakness seen throughout testing are the false negative results when looking at the various confusion matrices – especially for LDA, which incorrectly classified *no fatalities* significantly more often than it did correctly. The better model for classification of confirmed cases and fatalities is SVM.

Problem 3-1

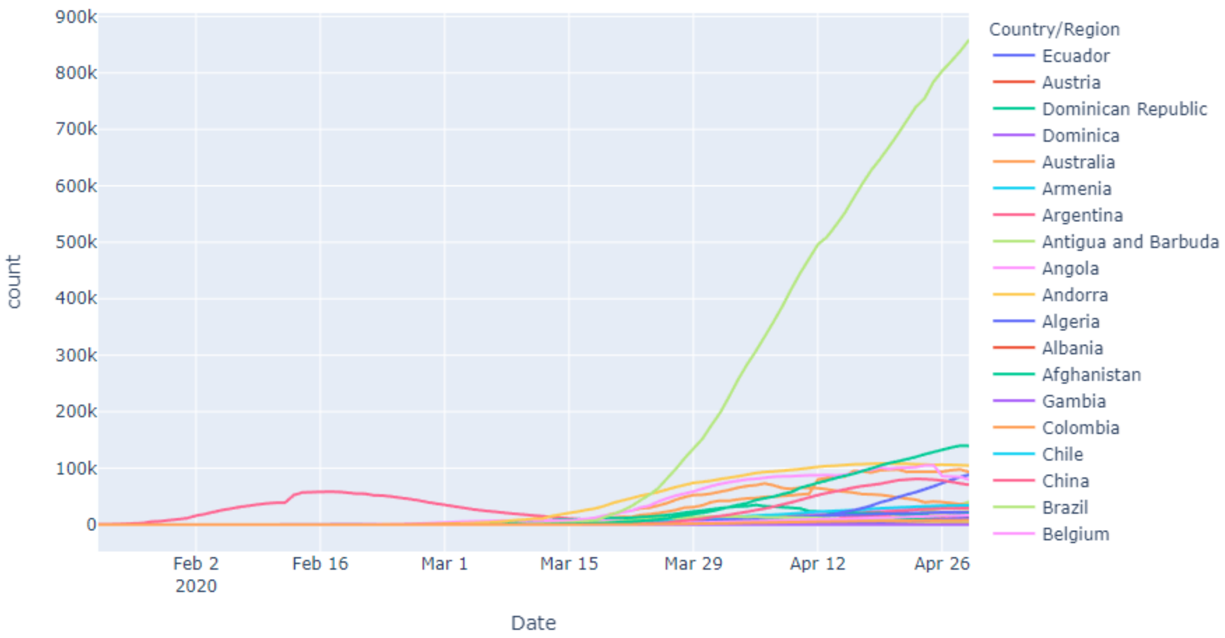


Problem 3-2

Case Status of Top 10 Infected Countries by Time

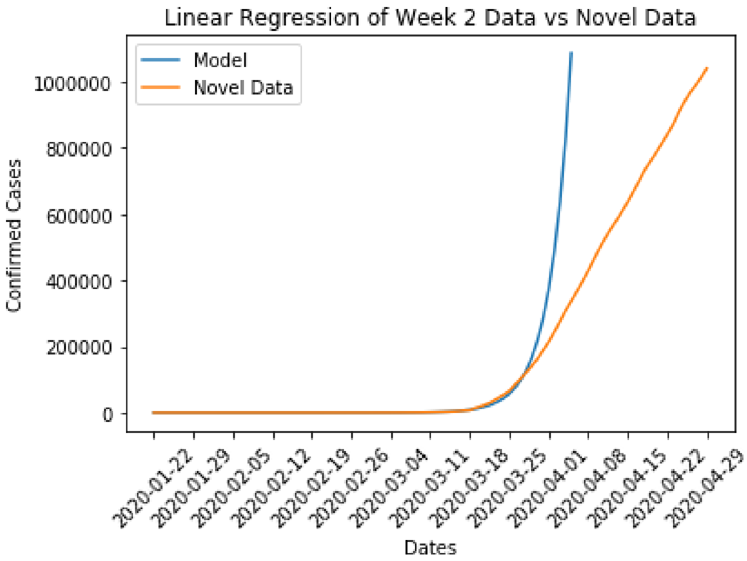


Active Cases Over Time by Country



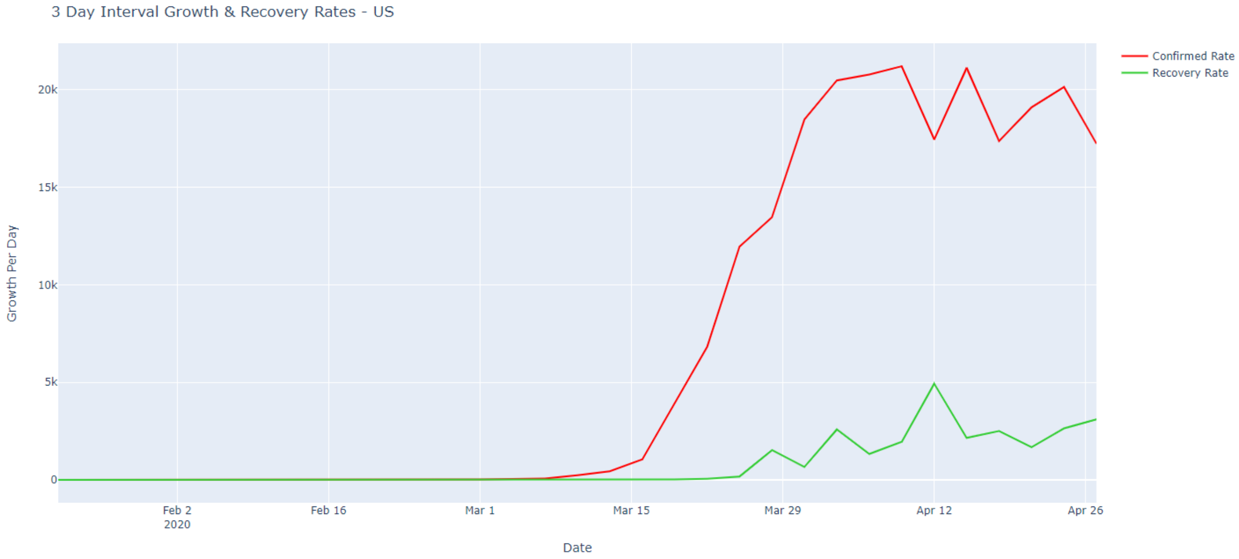
Problem 3-3

In comparison to the linear regression prediction generated from problem 2, week 2 data, the actual growth of confirmed Coronavirus cases within the United States is significantly slower - possibly because of the social distancing policies implemented to flatten the curve.

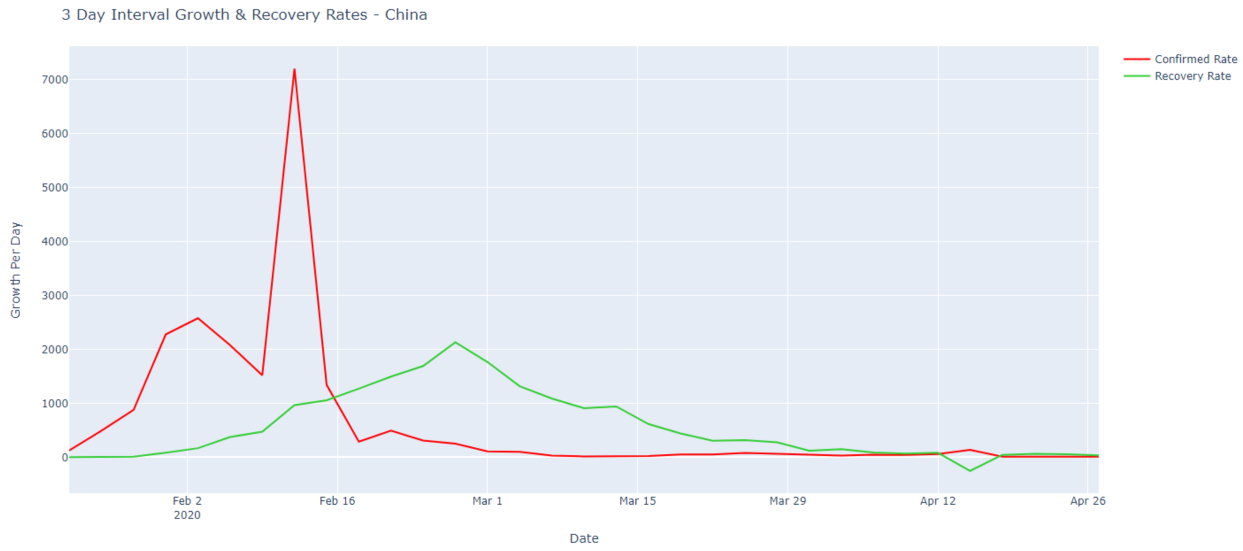


Problem 3-4 & 3-5

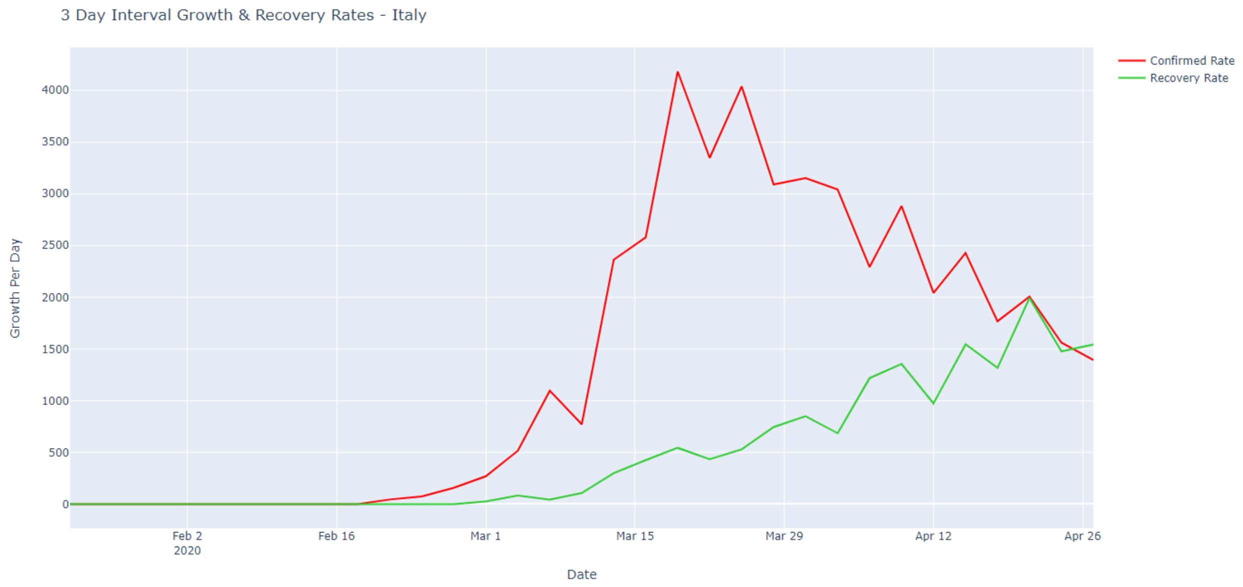
Filtering the time_series* data down to the US – the growth rate of confirmed cases is stagnating, albeit, still high. However, the recovery rate is steadily increasing.



Looking at a country that is on the road to recovery, the recovery rate should overtake the confirmed rate. From a previous analysis, it was seen that in most countries, the number of confirmed cases per day was beginning to stabilize. Looking at China, in particular, the recovery rate crossed the confirmed rate of increase around February 15th and is now on its way towards zero since there are only a few hundred cases remaining.



Similarly, in Italy, the recovery rate has very recently crossed the confirmed growth rate with an experienced downward trend ever since March 19th.



Runtime Analysis for rate calculation:

The function used to calculate the rate of change for n-days was the same for both growth rate and recovery rate. The runtime of the rate calculation would depend on the largest growth factor being the while-loop used and the efficiency of the built-in list insert function. After doing some research, according to the Python Wiki page [6], appending to a list is only O(1) on average; this would conclude our total runtime to be $T(n) = n^2+b$, or a $O(n) = n^2$.

References:

1. https://www.statista.com/statistics/1105088/south-korea-coronavirus-mortality-rate-by-age/
2. https://www.statista.com/statistics/1107149/covid19-cases-age-distribution-canada/
3. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---3-march-2020
4. https://www.nytimes.com/2020/04/17/us/coronavirus-death-rate.html
5. https://www.kth.se/social/files/5a040fe156be5be5f93667e9/ID2223-02-ml-pipelines-linear-regression.pdf
6. https://wiki.python.org/moin/TimeComplexity