Data Mining
Professor Lv
CSCI 4502
Artem Nekrasov, Vinayak Sharma, Tristan Thomas, Harsh Deshpande

Disease Outbreak and its relation to Geopolitical Quality of Life Attributes.

Abstract:

This paper aims to explore a field sought over by many in the 21st century: pandemic datasets. The main datasets that were explored were a historic pandemic H1N1 dataset, a current pandemic Covid-19 dataset along with a quality of life mapping dataset provided by MoveHub. After acquiring the data, it was aggregated by country and then submitted to an exhaustive data mining search. Via means of SLR, MLR, K-Means Clustering, and 10 Fold Cross Validation we were able to look for trends. A hypothesis for the results of our data prior to data mining was that better quality of life attributes for a country would imply less susceptibility to a pandemic.Through our journey we were able to learn having a better Quality of Life index for a country seemed to make no implication that the country would fare better against the pandemic (as measured by Covid cases, Covid deaths, Covid deaths per 100 recovered, Covid deaths per 100 cases against HealthCare rating, MoveHub rating, Quality of life rating). Our hypothesis was disproved; from investigating H1N1 and Covid against Quality of Life metrics we can conclude that there seems to be no effect that the quality of life a country provides for its residents would help battle susceptibility to a pandemic. We must be careful against diseases, since they will wreak havoc irregardless of how well the citizens of a country are (health or wealth). No feature that a country has to provide for its citizens can stop the spread of a contagion. Furthermore, since countries such as India have reported a fraction of the coronavirus cases as America, it raises the validity of the coronavirus reporting into question.

Introduction:

Disease outbreak team was constructed out of an aspiration to create a project that can be related to real live people and a current situation in the world. We decided to dive into Covid-19 Data and also decided to use a historic pandemic dataset (H1N1). H1N1 would help us find commonality between these two diseases. Our main goal was to find any possible trends, factors and correlations about Covid-19 and Geopolitical Quality of Life features and make a prediction or take away some useful and helpful information that can be used in the future to predict new possible outbreaks and its scale. We want to prevent or at least warn people of possible factors that will lead to the next pandemic. From the output of our project people will be able to alert geographic regions, or groups of people about to be cautious about a future outbreak; authorities will be able to contain viruses that may emerge in the future a lot faster since we can identify some areas (and features of areas) that may arise with a pandemic outbreak.

After looking over many different datasets, we picked some Covid-19 dataset and H1N1 dataset that had only the important information that mattered to us (countries, number of cases, number of deaths, deaths per 100 recovered and deaths per 100 cases). After that we needed some data that would help us to find trends, factors, correlations, basically datasets that can be used with the Covid/H1N1 datasets together and can produce meaningful information. We decided to use MoveHub aggregates features from www.numbeo.com, the CIA World Factbook, the WHO, census data from several governments, and their own vast database of real international moves. Since move hub aggregates data from multiple reputable sources such as the CIA and the WHO and then normalizes the data from 0-100 ranking scale, it would be the best fit to look for patternы. We were able to use data mining skills that were taught in the class and have some very interesting outcomes.

.

Related Work:

Due to the nature of this project, most past studies referenced are epidemiological. One such interesting study was conducted by Cambridge University to study the spread of Dengue in the Guangdong Province of China. Over a course of six years, the researchers monitored and took blood samples from people all across the province to trace whether they had the disease and one of the 4 viruses that causes Dengue (Fan et al). While this is a very interesting study, there were some predictable factors to the spread of Dengue. The first is that Dengue is only transmitted by mosquitoes and this region has had a history of it, with there being a massive outbreak of it in 1978 (Fan et al). Our project will try to examine other possibly more unpredictable causes for spreading.

The second study that was referenced was done by the European Union. The study reported on the spread of the HLTV-1 virus (Human T-cell Lymphoma/Leukemia Virus) all across the world. This disease is fairly common around the world and is prevalent on nearly all continents. This virus spreads through liquids (similar to STDs), and a common way it gets spread is through blood donations, so the researchers tried to look at that. However, they did realize that there are many ways this virus could spread, so blood donations were not necessarily the most reliable data source. The study also presented a world map, showing that it is prevalent in South America, West Africa, and in Romania. The study also does specify however that there are no strong correlations to why the hotspots occur where they are, especially in the case of Romania.

In this paper, we will specifically be looking at the spread of COVID-19 during the 2020 Coronavirus pandemic as well as the spread of H1N1 during the swine flu pandemic of 2009 as historical reference. These viruses were more interesting to study as the transmission and spread of these viruses are fairly unpredictable, especially when compared to viruses like Dengue and Zika which have limited means of transmission.

Methodology:

Main Tasks

Our primary goal for this project was to establish features or trends that were shared amongst nations that were disproportionately affected by pandemics. This meant congregating a multitude of datasets corresponding to how countries were affected by certain pandemics as well as geopolitical factors indicative of the standards of living in these countries. This overarching task was then broken down into modular sub-tasks.

Aside from cleaning and preprocessing the data, we wanted to isolate each geopolitical feature against each of our pandemic statistics. For example, observing trends in National Crime Rating against National Covid cases, National Purchasing Power against National H1N1 cases, and so forth for all of our features. This exhaustive approach was then built up into multilinear regressions and higher dimensional clustering to achieve our primary goal of establishing features and trends that highly affected countries had in common.

Analytical Thinking

Our first approach was to identify which individual features from our geopolitical datasets had significant trends with respect to the pandemic statistics. This led us to perform a 10-fold cross-validation with a Root Mean Squared Error (RMSE) as our valuation metric. For the sake of exhaustiveness, a multitude of regressors were used in this analysis - Random Forest, Ada Boost, Gradient Boost, Bagging, Linear, and a Voting Regressor. While none of the individual features resulted in a particularly low RMSE, comparing relative to other features yielded some interesting trends. This preliminary analysis gave us insight into how the data could be further mined.

After identifying geopolitical features that seemingly had an impact on various pandemic statistics, we ran several rounds of k-means clustering. We used the elbow method to determine the optimal “k” number of clusters for each feature set. Certain features that seemed promising from the cross-validation proved insignificant when clustering was performed but other interesting trends arose which will be discussed further in the discussion section of this report. We also attempted to run multiple Density-based spatial clustering of applications with noise (DBSCAN). Due to the nature of our data, this proved inconclusive as over 80% of data points returned a "noisy" data and most clusters held the minimum number of data points. This can be explained by the "curse of dimensionality" that we observed when attempting to cluster in a higher dimensional space.

Data Understanding/Preprocessing

A significant portion of this project came down to understanding our various datasets and coming up with creative ways to both aggregate and concatenate the data in meaningful ways. Initially, all of our features corresponding to geopolitical factors were reported by cities and all of our pandemic data was reported at a national level. Our first task was to programmatically congregate all of the cities into their respective countries so the two sets of data could be concatenated. This was accomplished by calculating a weighted average value for each feature, including all cities present in a given country. It was in doing this that we first observed the significant variance in how countries reported certain data. This led to the presence of many outliers in the form of countries with null values. After filtering these out, we were able to begin analyzing distributions and visualizing the shape that our data held.

One matter that was recurring throughout our analysis and evaluation was the presence of outliers throughout our data set. As noted above with null data values, there were outliers present on both tails of our distributions. As these right-tailed outliers were confirmed to be accurately reported data, we felt that it should be necessary to maintain them in our datasets. However, most of our analytical techniques, especially k-means clustering and linear regression, are particularly sensitive to the presence of outliers. This elicited unintuitive results when interpreting both regression models and clusterings. We ultimately split our data into two sets - one with outliers and one where outliers had been removed. We did this to observe whether different trends arose when comparing the results of these two different sets. These findings are also discussed further in the discussion section of this report.

Evaluation:

The evaluation of our results comes from many parts. We look for clusters with means that indicate a trend, slr/mlr slope for features to indicate a trend combined with the relevant p-values, r-values, t-values, std. error to check for statistical significance that the slopes may provide, along with a 10 fold cross validation. This would thoroughly allow us to see if patterns exist, trends exist, and if trends exist with what statistical significance. Our evaluation comes from the methods of exploration themselves, and require no separate testing suite. We used H1N1 data to compare with the current covid data, since H1N1 is a relatively recent historic pandemic it has had a few years for data to be collected. If the H1N1 dataset varies wildly in light of the patterns found from the Covid dataset, then we will be enlightened about the infrastructural differences that countries have when it comes to their response to pandemics.

The first result set that set the tone for the datasets in subject were the results found from the K-Means clustering. This method of analysis within our datasets would help us determine if there were clusters of data signifying importance for our Quality of Life aggregate attributes. Having clusters of data would imply that patterns do exist, and in which format. This would then lead the direction of future techniques to be implemented. From applying a K-Means clustering using features ([Covid cases, Covid deaths, Covid deaths per 100 recovered, Covid deaths per 100 cases] , [QOL,Healthcare, Movehub Rating, Pollution]) it was clear that the data had a rather even spread across all features. This immediately implied that there were no splits in the data that would lead to pattern findings through clusters.

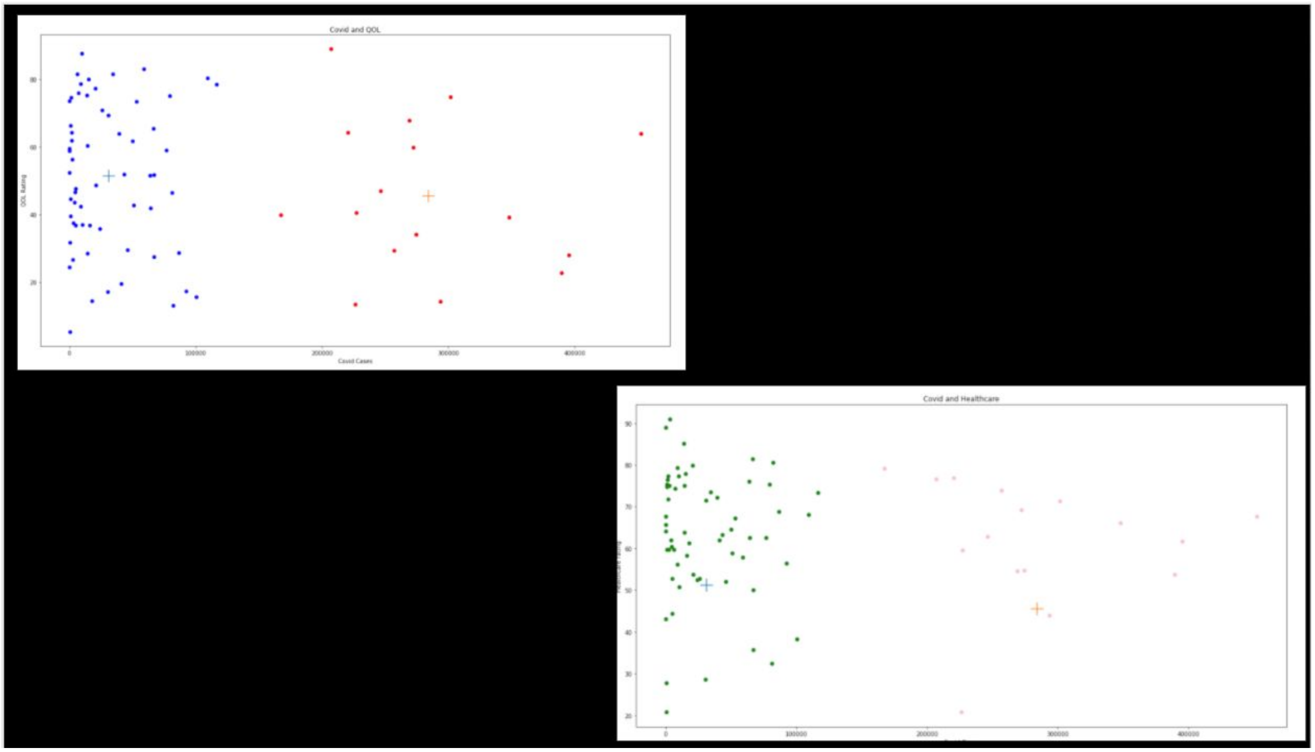


Figure 1- Clustering of Covid Cases (total) vs QOL and Healthcare. No pattern emerged, it seems

a country with any amount of Covid Cases could have any QOL rating and any Healthcare rating.

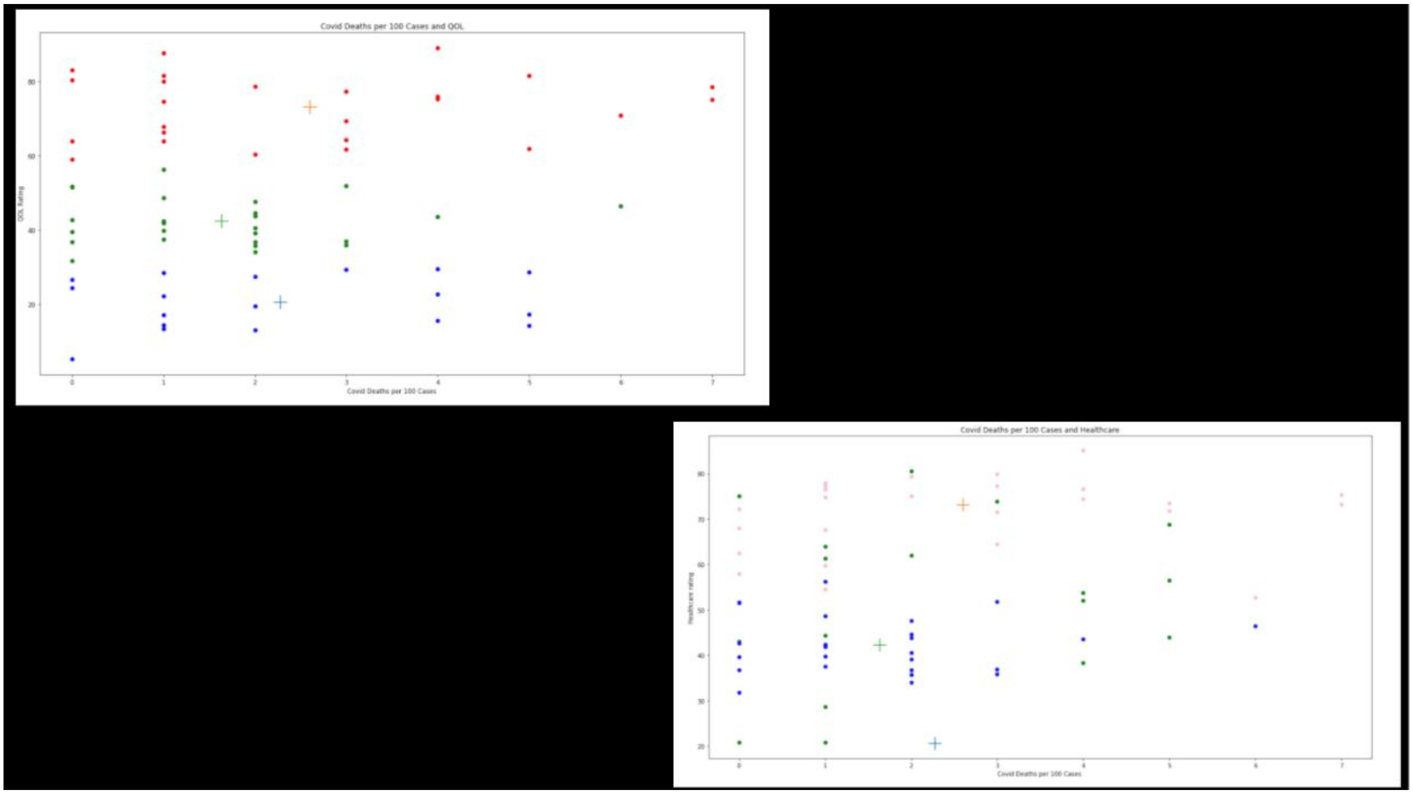


Figure 2 - Clustering of Covid deaths per 100 Cases vs QOL and Healthcare. No pattern emerged, it seems a country with any amount of Covid Death could have any QOL rating and any Healthcare rating.

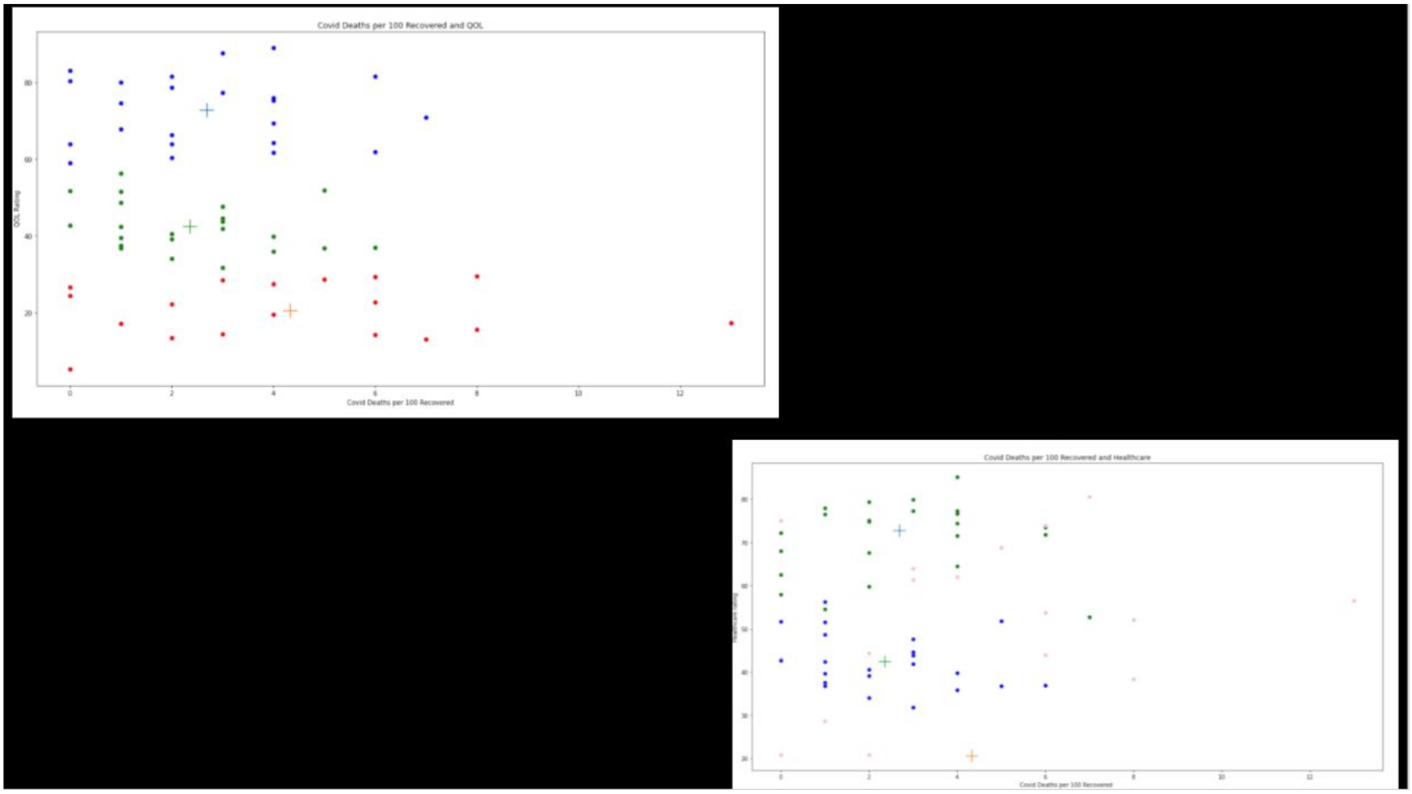


Figure 3 - Clustering of Covid Deaths per 100 Recovered vs QOL and Healthcare. No pattern emerged, it seems a country with any amount of Covid Deaths per 100 recovered could have any QOL rating and any Healthcare rating.

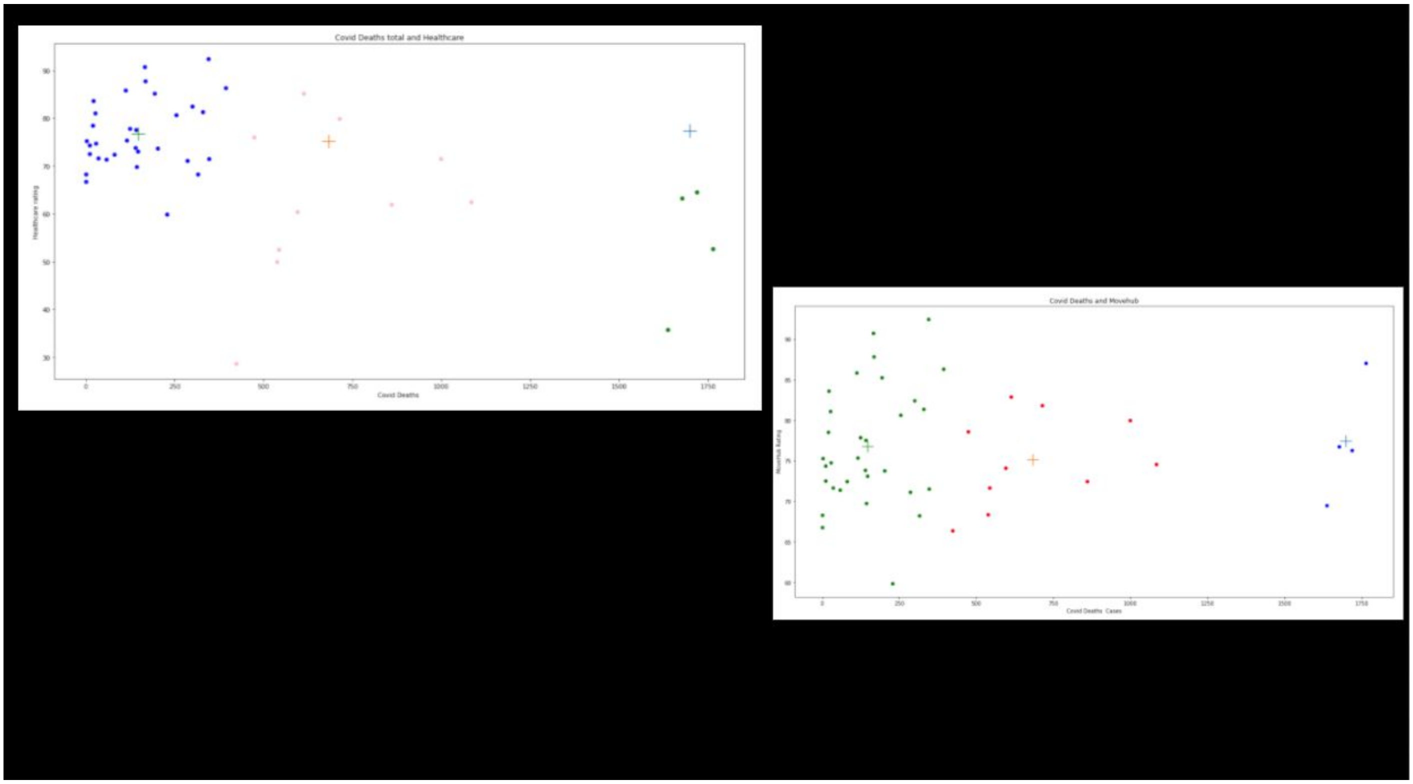
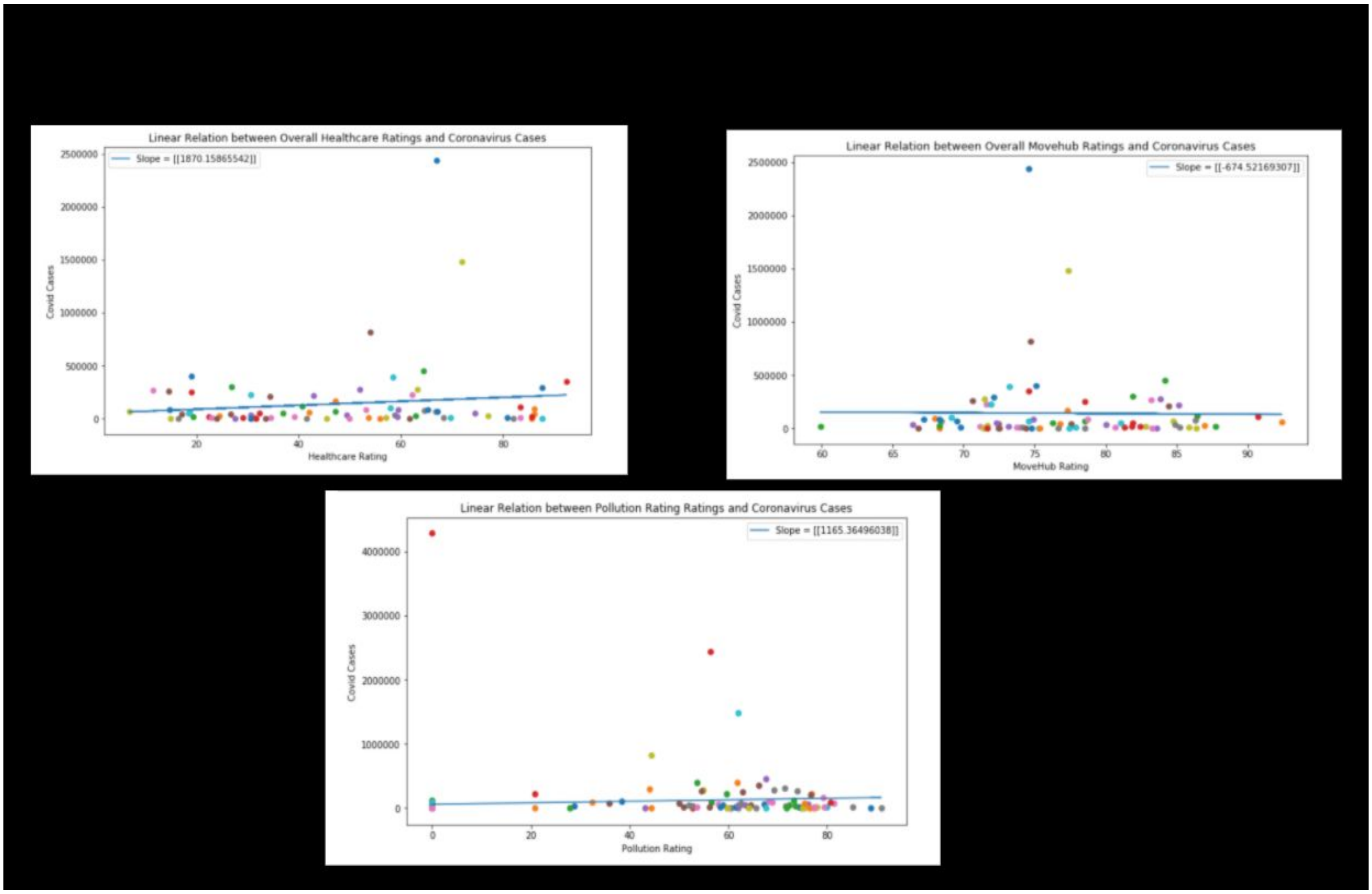


Figure 4 - Covid Deaths (total) vs Healthcare and Movehub rating. Clustering of Covid Deaths per 100 Recovered vs QOL and Healthcare. No pattern emerged, it seems a country with any amount of Covid Deaths could have any Healthcare rating and any MoveHub rating.

Looking at figures 1-4 it is evident that there are no patterns that can be established by clustering. All of the means in the figures above are placed at indexes that counter the argument of a visual cluster making sense. Further, for similar amounts of covid cases, they can range from the bottom of a ranking to the top of a ranking. It implies that having a certain number of covid cases does not bucket countries with similar rankings (better/well off countries do not fare better than worse off countries according to this clustering algorithm it would not be evident to prove the inverse).

After looking for meaning with a k-means clustering algorithm, we decided to look for trends instead of clustering. We initialized a search using a SLR toolbox. We used the best metric for a pandemics contagious capabilities within a country: total covid cases. Since from our clustering analysis we were unable to establish that there may be patterns within the other metrics of the Covid dataset, we wanted to focus our search on the best reported metrics, total covid cases and deaths. Even though deaths are less accurately reported due their natural hindrance to be tracked with ease.

Searching with SLR we see that there are many outliers that impact the slope for the trendlines that we were finding. K-means cannot work with outliers since it is very sensitive, this was our first opportunity to see how wildly the outliers would impact the trend lines that would be produced. (Figure 5).



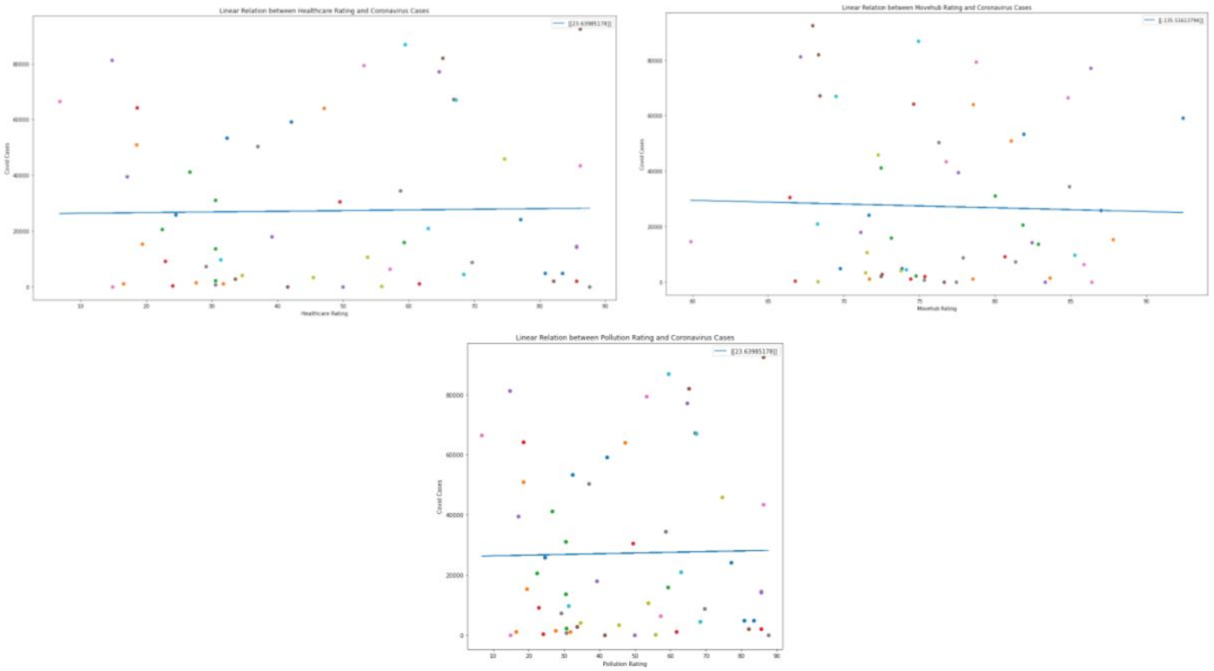
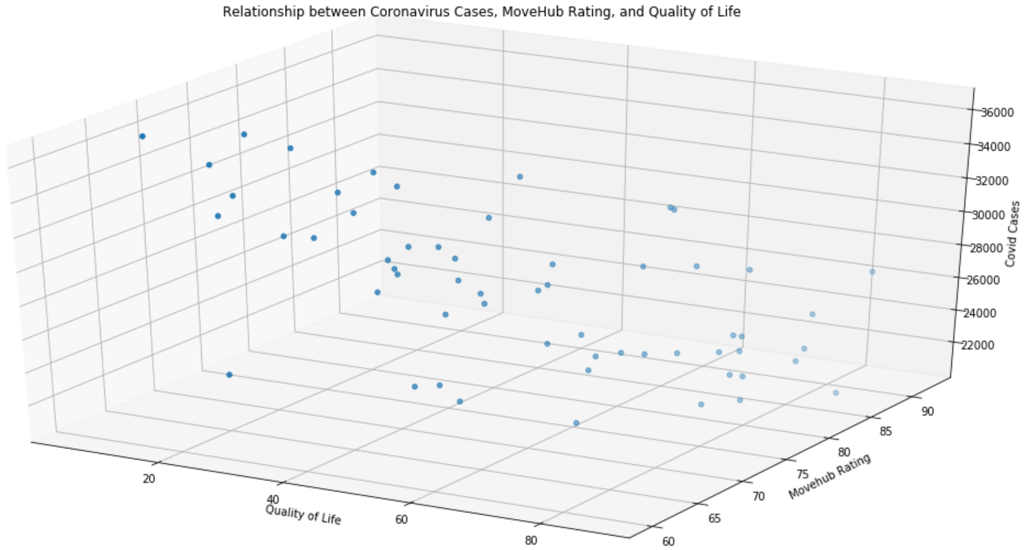
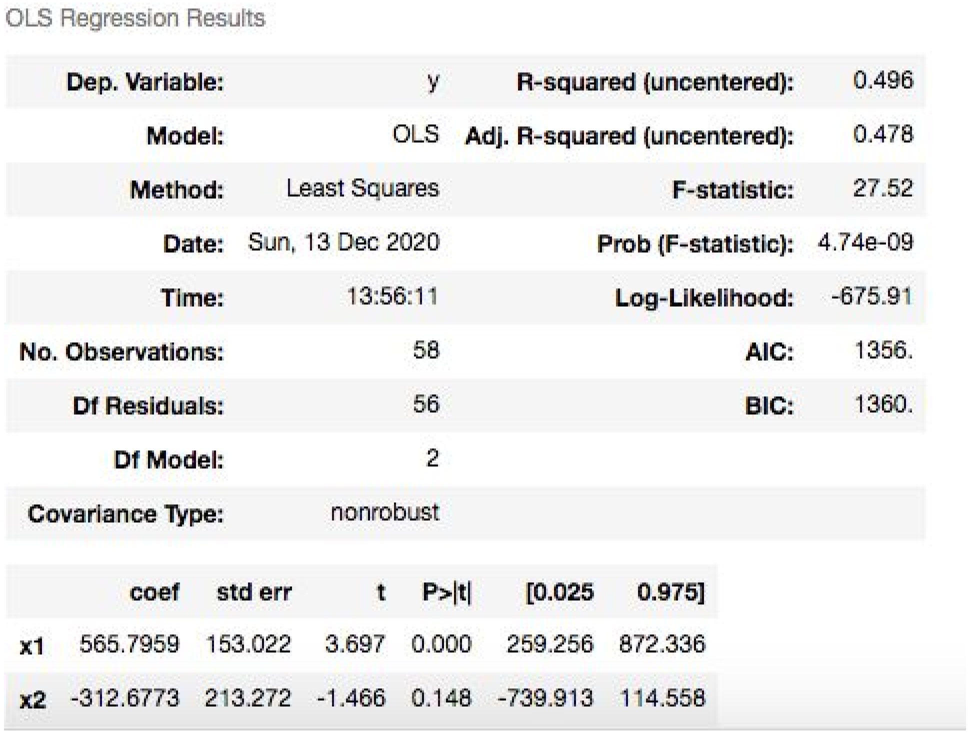


Fig 6

With figure 6 above, the trendlines end up not being as suggestive once the outliers are removed. These three graphs have R^2 values of approximately 3.8 * 10^(-4), 1.1 * 10^(-5) and 3.8 * 10^(-4) respectively. The R^2 values, being so close to zero, shows that there is little to no correlation between the factors listed above and COVID-19 cases and deaths.



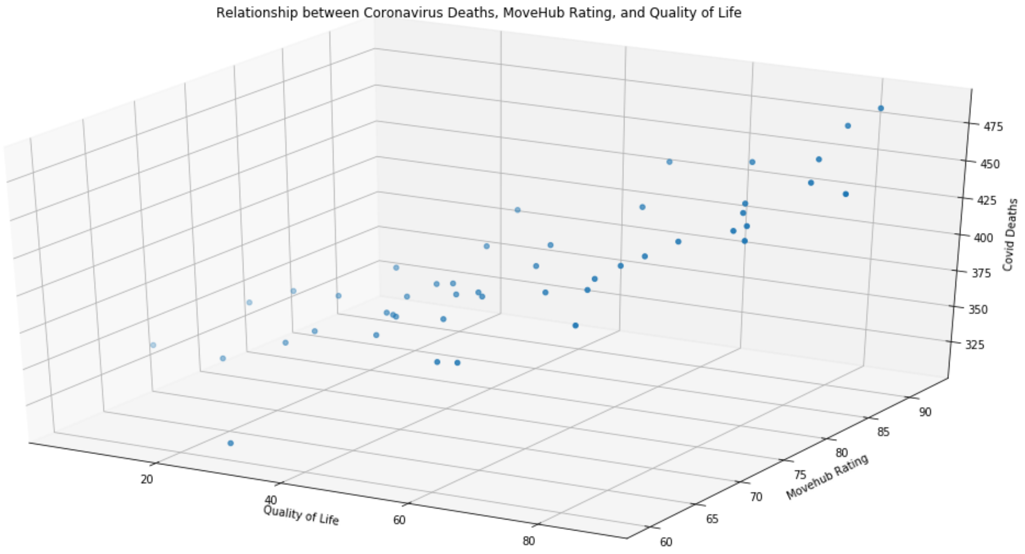


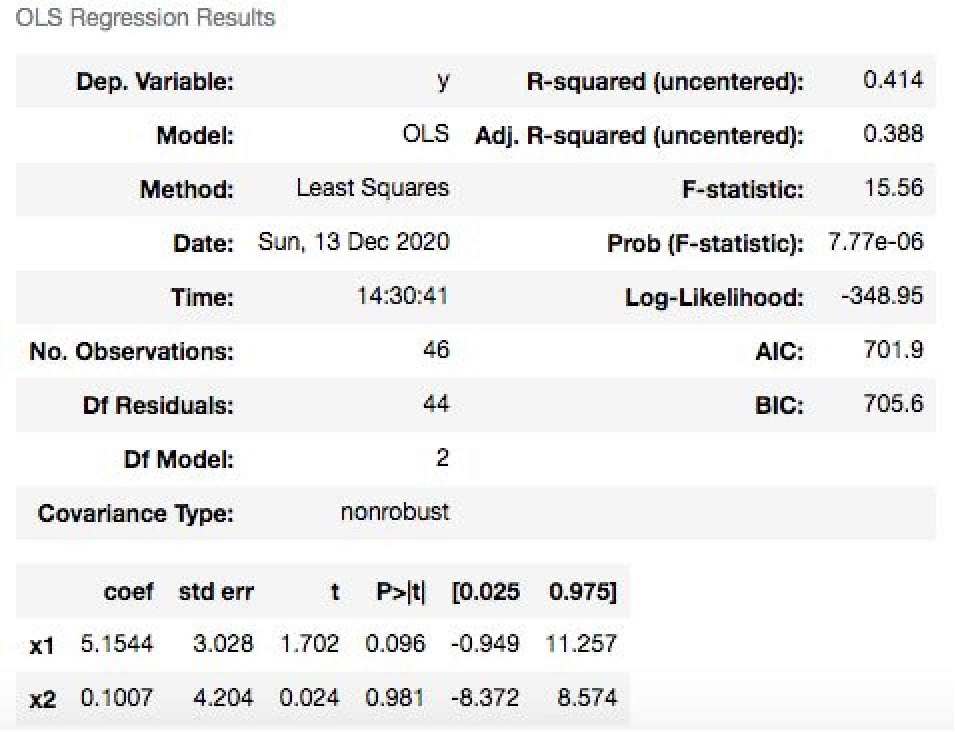
OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.496
Model:	OLS	**Adj. R-squared (uncentered):**	0.478
Method:	Least Squares	**F-statistic:**	27.52
Date:	Sun, 13 Dec 2020	**Prob (F-statistic):**	4.74e-09
Time:	13:56:11	**Log-Likelihood:**	-675.91
No. Observations:	58	**AIC:**	1356.
Df Residuals:	56	**BIC:**	1360.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
x1	565.7959	153.022	3.697	0.000	259.256	872.336
x2	-312.6773	213.272	-1.466	0.148	-739.913	114.558

Fig 7





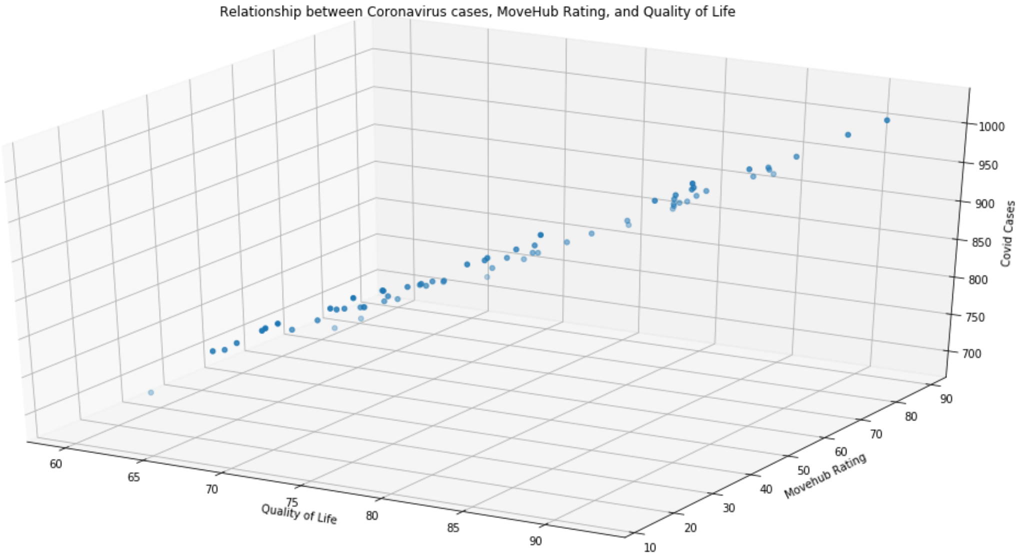
OLS Regression Results

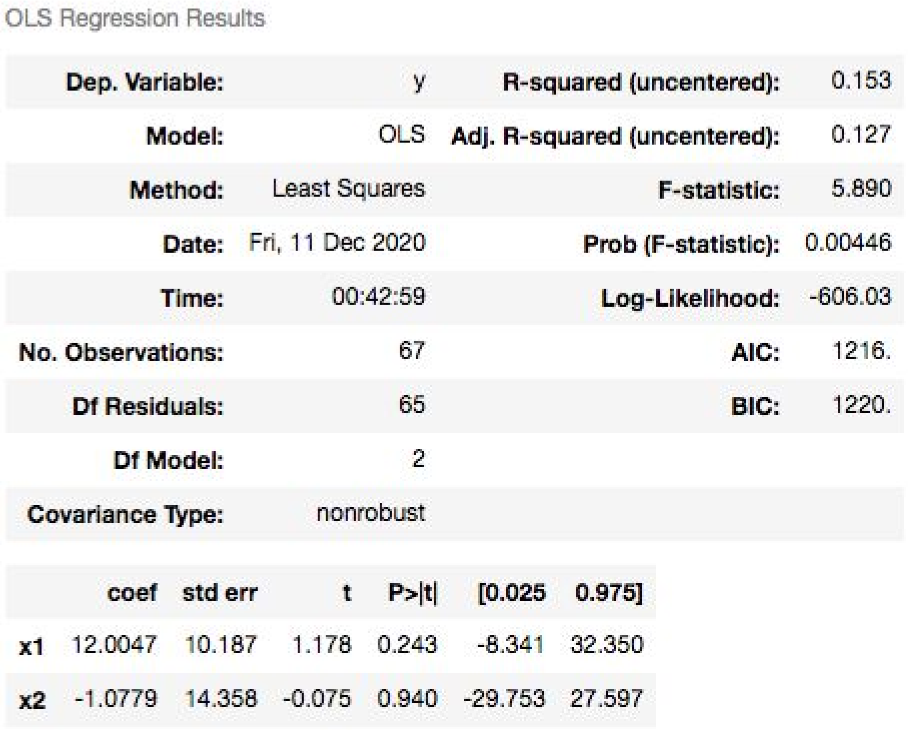
Dep. Variable:	y	R-squared (uncentered):	0.414
Model:	OLS	Adj. R-squared (uncentered):	0.388
Method:	Least Squares	F-statistic:	15.56
Date:	Sun, 13 Dec 2020	Prob (F-statistic):	7.77e-06
Time:	14:30:41	Log-Likelihood:	-348.95
No. Observations:	46	AIC:	701.9
Df Residuals:	44	BIC:	705.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	5.1544	3.028	1.702	0.096	-0.949	11.257
x2	0.1007	4.204	0.024	0.981	-8.372	8.574

Fig 8

With Figure 7 and 8, we tried to see if there was any correlation between COVID cases/deaths and a combination between any of the two factors. We decided to look at the Quality of Life and MoveHub ratings, as these are essentially two aggregate ratings. A Multi-Linear Regression was done with these two factors and both COVID cases and deaths. These seemed to be more promising as there was more correlation here. The R^2 values for both of these graphs were about 0.5, so there was some correlation going on. However, these trends do not prove to be completely reliable as the standard error for each of the slopes of both of the MLR were still relatively large.




OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.153
Model:	OLS	**Adj. R-squared (uncentered):**	0.127
Method:	Least Squares	**F-statistic:**	5.890
Date:	Fri, 11 Dec 2020	**Prob (F-statistic):**	0.00446
Time:	00:42:59	**Log-Likelihood:**	-606.03
No. Observations:	67	**AIC:**	1216.
Df Residuals:	65	**BIC:**	1220.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
x1	12.0047	10.187	1.178	0.243	-8.341	32.350
x2	-1.0779	14.358	-0.075	0.940	-29.753	27.597

Figure 9

For historical reference, however, Figure 9 shows that there is not much of a correlation between these same two factors when compared to the spread of the Swine flu. In this case, there is a much lower R^2 value of 0.15 and the slopes still have a fairly large standard error. This means that looking at the quality of life rating and the MoveHub rating as reliable factors for predicting the spread and severity of any virus/disease is very inaccurate, and only seems to be somewhat applicable to COVID-19.

Discussion:

What we learned from the data was that countries were reporting at different accuracies. It just does not make sense that a communicable airborne disease would have higher transmission rates in a country like the United States than it would in a country like India. This is because India has many times the population of the United States yet only a fraction of the land, which would imply that an airborne disease would transmit to more people in India than it would in the United States. Further from the SLR, before we remove outliers and look at what the data tells us, it shows that with outliers (such as the US), having better healthcare would actually imply more Covid Cases. This can demonstrate an instance of misreporting, because obviously countries with better healthcare can do more tracing and keep better records than countries with worse healthcare. After we removed the outliers(a lot of the outliers had high healthcare ratings) it evened out the playing field and the distribution showed no trends. We learned that pandemic datasets are hard to analyze because even though some countries may be reporting high values of cases and would be classified as outliers when compared to an interquartile range, it is hard to say they are outliers in any other sense. These countries may be the only countries with the facilities to report accurately.

What we also learned was that after removing the countries with high number of covid cases, it seems as though via means of investigation using K-Means Clustering, SLR, MLR, and 10-Fold Analysis that for the most part all of the data was distriubuted in a fashion that indicated the quality of life a country provides for its residents would help battle susceptibility to a pandemic. We must be careful against diseases, since they will wreak havoc irregardless of how well the citizens of a country are (health or wealth).

What did work well was how large the datasets were. We were given almost over a dozen features, and thousands of data points to analyze. Working with a dataset this large means that there was always something to be learned from every feature or technique we encountered. We were genuinely thrilled throughout the course of this project. For the future we will try to look at a more historical sense of the data, since H1N1 is relatively recent (even though it can be classified as a historic dataset), working with datasets solely off a historical basis we would be able to tell what Covid will turn into. It is very young right now and will be a decade before we see the real impact.

Conclusion:

This paper aims to explore a field sought over by many in the 21st century: pandemic datasets. The main datasets that were explored were a historic pandemic H1N1 dataset, a current pandemic Covid-19 dataset along with a quality of life mapping dataset provided by MoveHub. Since move hub aggregates data from multiple reputable sources such as the CIA and the WHO and then normalizes the data from 0-100 ranking scale, it would be the best fit to look for patterns. Since many features such as pollution and healthcare are measured in intricate manners such as ppm for pollution and hospitals per capita; the MoveHub data allows us to look at countries from a relative scale. After acquiring the data, it was aggregated by country and then submitted to an exhaustive data mining search. Via means of SLR, MLR, K-Means Clustering, and 10 Fold Cross Validation we were able to look for trends. A hypothesis for the results of our data prior to data mining was that better quality of life attributes for a country would imply less susceptibility to a pandemic. Immediately it was evident that the Covid data had a handful of extreme outliers (such as the United States, Turkey, India). These several countries that would be “extreme outliers” had Covid-19 stats many times larger than those of all the countries in the world. We started off with clustering and initially saw almost no real clusters. It seemed as though the Covid-19 data was evenly scattered. Having a better Quality of Life index for a country seemed to make no implication that the country would fare better against the pandemic (as measured by Covid cases, Covid deaths, Covid deaths per 100 recovered, Covid deaths per 100 cases against HealthCare rating, MoveHub rating, Quality of life rating). Our hypothesis was disproved again through SLR, where we analyzed individual quality of life features against Covid Cases (since covid cases seemed to be the most likely to be accurately reported globally). SLR showed many slopes of close to 0, meaning that the Covid Cases were distributed throughout Quality of Life attributes with no skew towards lower or higher QOL. This theme was re-iterated through an MLR investigation.During a Ten Fold Cross Validation we were able to see which features had the most impact.Even though we had low correlation and statistical evidence for causation we were able to see that the MoveHub aggregate feature was the most powerful. After a dive within the covid dataset, we decided to look into the H1N1 dataset to again see a similar distribution of data points, where no trends were visible from SLR and MLR. From investigating H1N1 and Covid against Quality of Life metrics we can conclude that there seems to be no effect that the quality of life a country provides for its residents would help battle susceptibility to a pandemic. We must be careful against diseases, since they will wreak havoc irregardless of how well the citizens of a country are (health or wealth). No feature that a country has to provide for its citizens can stop the spread of a contagion. Furthermore, since countries such as India have reported a fraction of the coronavirus cases as America, it raises the validity of the coronavirus reporting into question.

Appendix:

Honor pledge: "On my honor, as a University of Colorado Boulder student I have neither given nor received unauthorized assistance."

Individual Contributions are as follows, Tristan: DB-Scan, K-Means Clustering. Vinayak: K-Means Clustering, SLR, Dataproccessing. Harsh: MLR, Statistical Significance, Data Aggregation, Data Processing.