Analyzing Relationships among Social Vulnerability Index factors, Crime Cases and Covid Cases

Md Rakibul Hasan Talukder

Mridul Banik

Nick Kowalczyk

Team 15

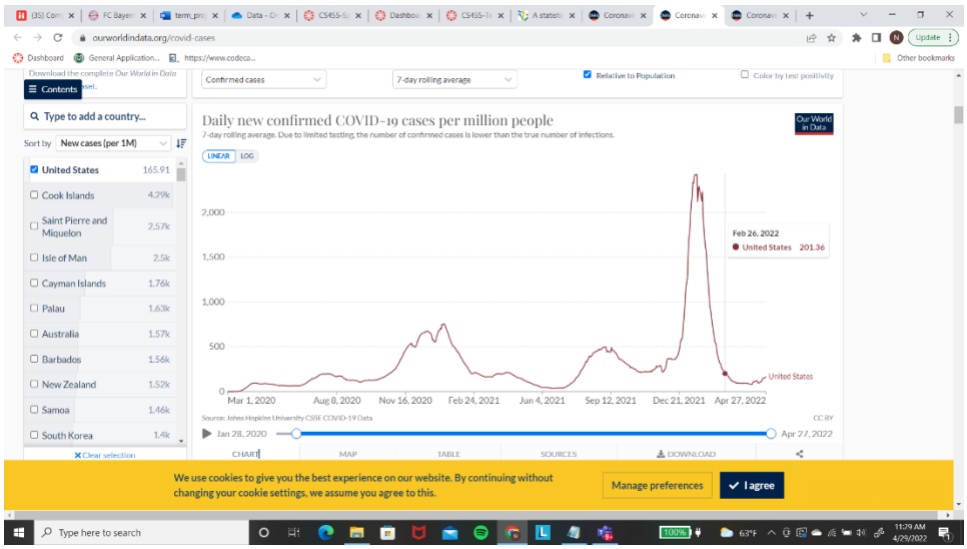
1 Introduction

The corona virus pandemic that was reported in late 2019, has impacted every aspect of our lives in the past two years. In this paper we will be looking at and comparing two datasets with the Covid 19 pandemic datasets for the years 2020 and 2021. The datasets are coming from the United states and can be found on [5] as well as crime data found on [6]. The datasets come from a variety of states dispersed across the U.S, including Colorado, Texas, Georgia, Kansas, Kentucky, Virginia, Iowa, and Missouri. We chose these states as they have a variety of different geographies, cultures, populations, and economic output. The datasets we will be comparing this against are the Social Vulnerability Index (SVI) which is calculated from four different factors, as well as the total crime cases. This analysis will be looking at data on a county-by-county basis.

Other studies have found, [1] “robbery, aggravated assault, and simple assault declined post-pandemic,”, but different trends have been spotted following, [1] “homicides and shootings”, which can be attributed to gang and drug market crimes. One possible answer for this follows that people had more free time at home, and some may have turned to alternative forms of recreation including increased drug use. Thus, many gangs and drug markets seized the opportunity to expand their respective markets.

Our analysis will be focused on discovering whether there is a correlation between SVI, crime and the Covid pandemic. We believe there is a connection between crime and SVI with Covid case and death totals. This is because economically poorer or rural areas generally have less access to healthcare and were hit the hardest by the lockdowns that swept the nation.

Because of the strong covid reporting done by the U.S. government, we have an accurate representation of total cases and death totals. The below graphs come from [2] ourworlddata.org/pandemic. These graphs portray that although many people contracted the Covid-19 virus, many of them did not die. This can be seen in graph one, where there is a large spike in case totals, and one can only assume that this affected people ranging from poor to high economic status across the U.S.



The death total graph shows an interesting trend, and we hypothesize this is because some areas had better access to healthcare than others when they were affected.

In the analysis below we will be looking at the correlation between SVI, crime against the COVid-19 dataset, and supplementing our findings with other calculations including the worst and least impacted counties by state.

2 Problem characterization

The problem we have chosen to characterize, analyze, and present is relevant, scalable, and personal. We are interested in determining whether crime, socioeconomic status, household composition, minority status, and housing influenced the COVID-19 pandemic. We all have a personal stake in this, as we constantly relived the “two more weeks” lockdowns, and all of us have had, or know someone that had COVID. This problem is important to us because not only is it very recent and impactful, but various aspects of our daily lives have

been affected. Our goals include finding whether your standard of living and safety in your county influenced pandemic response from government officials, and survivability.

The datasets we utilized to get our findings included the social vulnerability index(SVI), which looks at a variety of factors including socioeconomic status, household composition, minority status, and housing, the total crime cases, as well as total case counts and death totals by county in 10 different states spread out around the country.

Our approach was to determine whether there is a correlation between the four SVI factors, crime totals, and covid cases and death totals for the years 2020 and 2021. Some factors that make this analysis difficult included how recent the covid pandemic is, which made it difficult to find other documentation or similar analysis to compare our findings. Because of the global nature of COVID 19, and no other pandemic has reached this scale in the modern age, it has been difficult to analyze our findings. Yet, this data will provide valuable information about preparing for future pandemics, and how to protect vulnerable communities and groups of people.

Another issue that makes this difficult is collecting accurate and reliable data. It is hard to say how many covid deaths were linked to covid, and no other types of traumas, so these statistics may be skewed. Additionally, many poorer comminutes had people never visit the hospital after falling ill, and many had difficulties accessing self-reporting resources. Many of these poorer communities were hit the hardest by the pandemic and it is difficult to tell the true scale. Additionally, many people reported feeling no symptoms, and failed to report that they may have had covid. This could also skew the total case counts.

We were interested in discovering how different governments and local areas dealt with the pandemic. We did this by sampling a variety of states ranging across the U.S to get a better picture of how-to pandemic was treated. We treated all the data equally and provided an unbiased analysis of the correlation between crime, SVI, and covid.

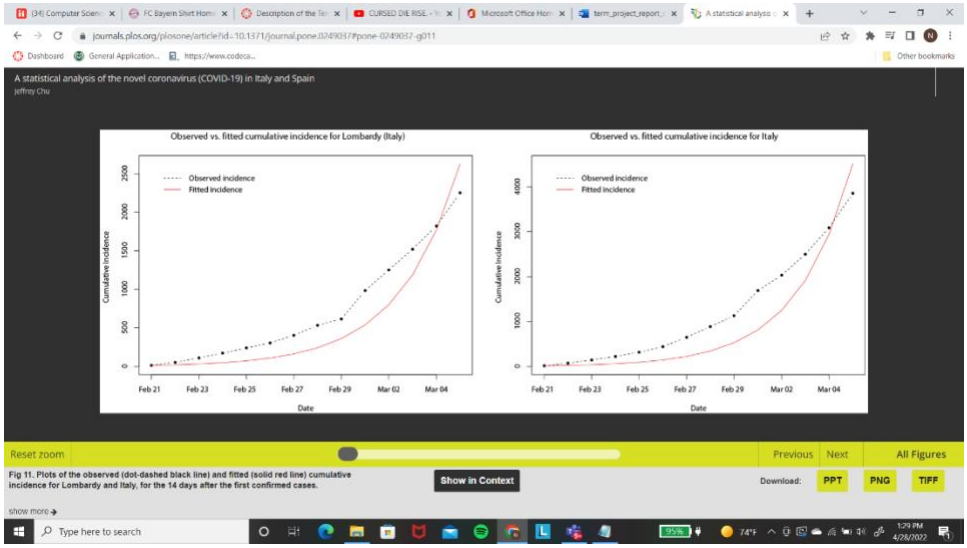
3 Dominant Approaches to solve the problem

This study [3] utilizes the Susceptible-Infectious-Recovered model and the log-linear regression models to generate a machine learning algorithm to fit the sample data and create predictions going forward. This study looked at the total cases in Spain and Italy for 2020 and analyzed covid in all regions of the two countries. The susceptible infectious recovered model is split into three groups, and this model has been used previously for many other pandemics. The team utilized the equation [3] "$S(t) + I(t) + R(t) = N$"

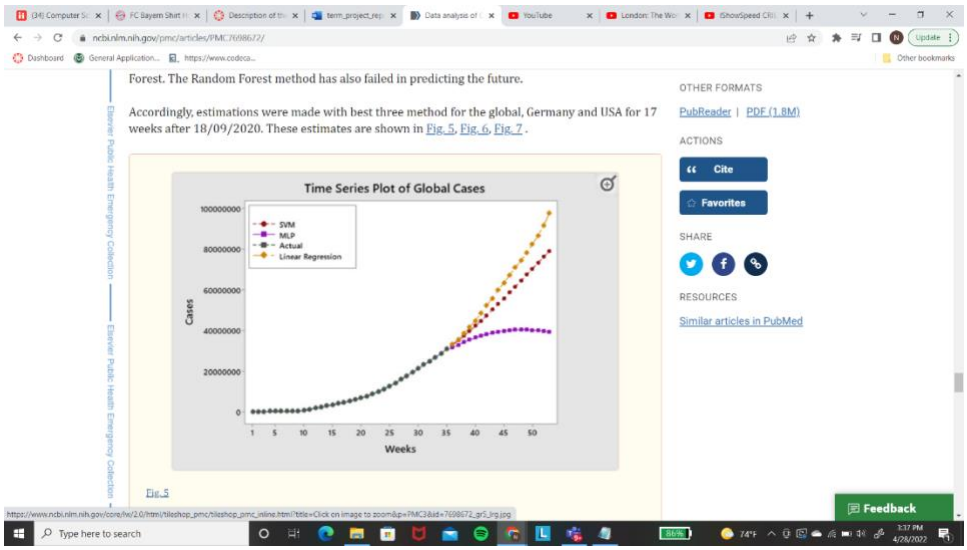
$$RMSE = \sqrt{\frac{\sum_{i=2}^{n} \left(y\left(i\right) - \hat{y}\left(i\right)\right)^2}{n-1}}$$

to solve the minimization problem. These models were found to be effective at predicting future spread of the covid. When comparing the accuracy of the predicted vs sample against our own analysis of correlation it can be seen below [3], that their models lined up well with the real-world data. In our case, many correlations we thought were accurate in impacting covid were found to have little or negative correlation when comparing each state against each other. Therefore, our analysis was not as accurate and robust as this paper's findings,

where they accurately trained a model to fit the data, and predict the future. On the other hand, it was seen that when looking at the states individually they followed a strong trend.

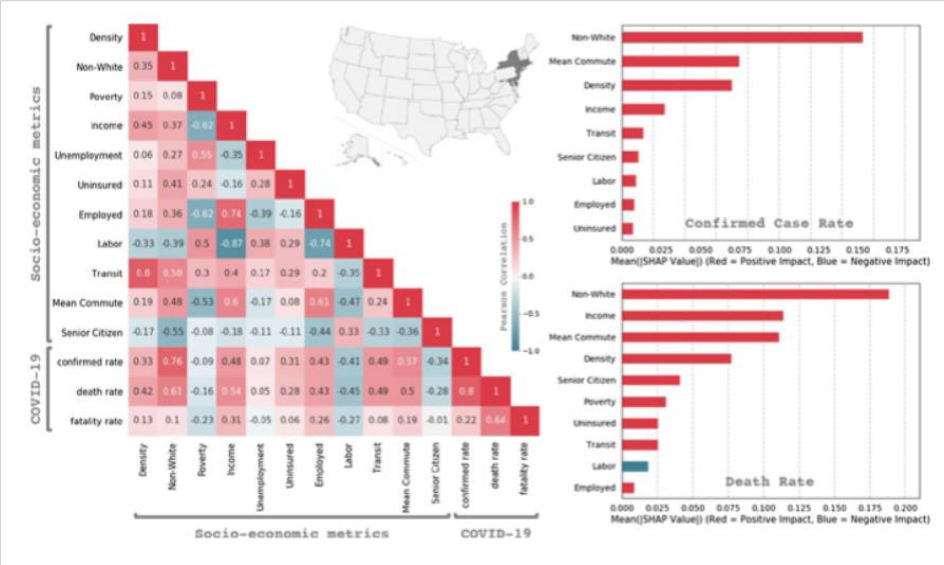


[4] This paper analyzed a variety of machine learning approaches in order to estimate the Covid – 19 pandemics in 2020. They utilized models like linear regression, multi-layer perceptron's, and SVM compared against RMSE. In particular it was found the SVM was most accurate, and they used the equation here [4]. Our approach focused on a county-by-county basis through a few states. When comparing the scalability of our approach, it can be seen that this study was more successful in scaling their analysis to the global pandemic. This study was successful in predicting future covid trends on a global scale with a variety of methods, while our analysis was primarily based on the correlation between factors.



Finally, we will be comparing the socio-economic status of counties against covid data found. [7] This study compared a variety of factors(seen below) against various covid data. This study found that minority groups

had the highest correlation in both case and death totals. When compared against our findings, there is some discrepancy. Although we looked at a variety factors in the SVI, we did not notice a dramatic correlation. This is possibly attributable to other factors in the SVI canceling out some of the correlations, or the larger dataset from this study allowed for the correlation to be better seen.



4 Methodology

4.1 Data sets

For the term project we have selected three datasets. Two of them were from https://urban-sustain.org website and another dataset were collected from https://crime-data-explorer.fr.cloud.gov/pages/home

This dataset of SVI index contains Social Vulnerability Index. In attribute lists the dataset had overall SVI index, and four other factors. Dataset of covid cases contains covid case and covid death cases by county of two years. The dataset of crime contains information of total crime of each county of a state.

4.2 Data Collection, Data formatting and Conversion process

For SVI data we collected data in json format and used python to covert json file to make a csv file. We only took overall SVI, Socioeconomic Status, Household Composition & Disability, Minority Status & Language, and Housing & Transportation in our SVI csv file.

The crime data was downloaded in xls format and was converted into csv file.

To begin the data processing for the covid dataset, we created a directory and staged all of the covid_data json files by state. Then, we created a simple parsing script that imported the json files, opened an output file, and looped over the opened json files on by one to perform some data processing. We organized the covid data in the years 2020 and 2021, and outputted the state, county, total cases, deathcount, and date. Because of the nature of our script, we were forced to parse this output csv in order to create a final covid csv file that contained the covid totals for both years on a single line. This allows us to have all of the covid totals in a single row, and have the data split up in columns for analysis with spark. This was done by opening the csv file, looping over it line by line, and checking whether our key (state, county) was present in the dictionary, and appending the value. If the key doesn't exist, we would add it with its associated value. Finally, we opened a final output file that wrote the dictionary to it.

4.3 Merging and Conversion Data

When we have all our individual dataset got ready, we have merged them into one csv file. We have used python panda's library to build the combined csv file. We then had separate datasets for all the states. As we had collected dataset from 7 states, we end up having 7 different combined files with respected to state.
Additionally, we also wanted to have a version of dataset where we will explore counties across states. Therefore, we concatenated all the separate state wise csv files into one csv file. We had two version of data which enabled us to explore the data in two categories.

4.4 Processing Data

First, we filled up our empty cells with average values and continue our processing of next stage.

While studying the data we had, we found the pattern of each column is different. For example, svi indexes stays into 0 to 1, whereas crime cases or covid cases had a lot of variation which may oscillates into 0 to few thousands. Therefore, we decided to apply normalization techniques into our datapoints which will enable us to have the data into a certain range.

We have applied two different preprocessing in term of normalization. We used *sklearn* library to transform our data. The first way of preprocessing was *Standardscaler* approach. Standardization is a common requirement for many machine learning estimators: algorithms might behave badly if the individual features do not more or less look like standard normally distributed. For instance, many elements used in the objective function of a learning algorithm assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. That's why we found this *Standardscaler* techniques will be appropriate to apply.

This *Standardscaler* techniques also transform some values into negative values. Therefore, we had applied another form of normalization using *Sklearn* Normalizer. Normalizer provides us the Normalization, which does scaling individual samples to have unit norm. It does not transform the positive values into negative values. Thus, we applied two different forms of scaling in our dataset.

Since transformation with normalization will change the values of every column, we keep both unchanged and transformed version of covid datapoints in our dataset. That made the final version of our dataset where we performed analysis in *Apache Spark* with Scala programming language and executed in distributed fashion.

4.5 Problems We are Solving

The primary idea of our term project was to find out how our data's are corelated with each other. We have calculated correlation of multiple columns. We calculated how different SVI indexes are correlated with total crimes of state. To do that, we found correlation of Socia-Economic factor, house composition disability factor, minority status factor, house transportation factor with total crime cases.

We also calculated correlation of covid cases and covid death cases with SVI index of two different years.
To calculate, the correlation we first read our dataset as dataframe in scala programming language. Then we selected our desired column of which we will find the correlation and assign that into a new dataframe. This new dataframe was casted to double as we need to perform operation of correlation on this. We used corr() function of dataframe which is basically perform Pearson correlation coefficient between those two columns. Pearson correlation coefficient gives value within -1 to 1. The higher value represents strong correlation. After getting the correlation we saved the correlation value with factor names in txt file using saveAsText method of Scala.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Fig. : Pearson Correlation Equation

Apart from finding correlation, we found the highest and lowest value of covid cases and covid death cases of two years and return the corresponding county respective to it's state. We were also able to find the highest and lowest county name across multiple state. For calculating highest and lowest value, we first read the csv file as rdd. Then we drop the first row from the rdd. We took the coumn values by split lines with comma using map function of rdd. Then we ordered our rdd in both ascending and descending and assign the first value in a rdd. That rdd value, we saved in text file as our desired output.

5 Experimental Benchmarks

The main objective for our analysis was to find out the relation among different attributes for SVI, Crime and Covid datasets. For this purpose, we have chosen Pearson Correlation as experiment benchmark. For our experiments, the correlation value ranged between 0.51 and -0.14. Positive correlation represents higher dependency between two attributes, whereas negative correlation provides lower dependency relation. We tried to find the correlation between two attributes (combination of attributes are given in the following table) for each state using the attribute values of counties of corresponding state.

Suppose, For Colorado state, we had data for 50 counties. We have assumed, the correlation between one of the four SVI factors (Eg. Socio_eco) and total crime cases can be derived considering only the attribute values of 50 counties. Correlation of attributes for each state are considered to be dependent on the counties they govern. Counties from outside of a state does not have an impact on the correlation on that specific state. This assumption seems reasonable to find actual relationship with the help of correlation.

Moreover, For each state, we have calculated maximum and minimum for both covid case and death case of 2020 and 2021. This helped us to understand the range of cases per state which can be related with other factors (SVI and Crime).

Our main target was to find any kind of relation among these three aspects Crime, Social Vulnerability and Pandemic situation. We have not only calculated correlation for all three combinations of three factors (SVI, SVI and Covid), but also divided each factor into other subfactor to find detail relationship. SVI is a composite function of four other factors. Again, covid cases are divided into affected and death case for two years.

All these combinations support the benchmark to do a reasonable and rational analysis.

Total Crime Cases	Socioeconomic status
	Household composition & disability
	Minority status & language
	Housing type & transportation
SVI Index	Covid case 2020
	Covid death case 2020
	Covid case 2021
	Covid death case 2021
Total Crime Cases	Covid case 2020
	Covid death case 2020
	Covid case 2021
	Covid death case 2021

Table: Pair combinations of attributes for correlation

6 Insights Gleaned

Initially, we thought that there will be strong correlation of all factors of SVI with crime cases and covid cases. However, according to our experiments we found that there's a strong correlation between some of the factors and some of factors do not have any correlation.

From the graph ***Normalizer : Correlation : SVI Factors - Total Crime*** we can understand that we found strong correlation between Socio-Economic factor of SVI and Total crime for 7 states. On the other hand, house composition disability factor is not strongly correlated with total crime cases. Minor statistics and house transport are somewhat neutrally correlated.

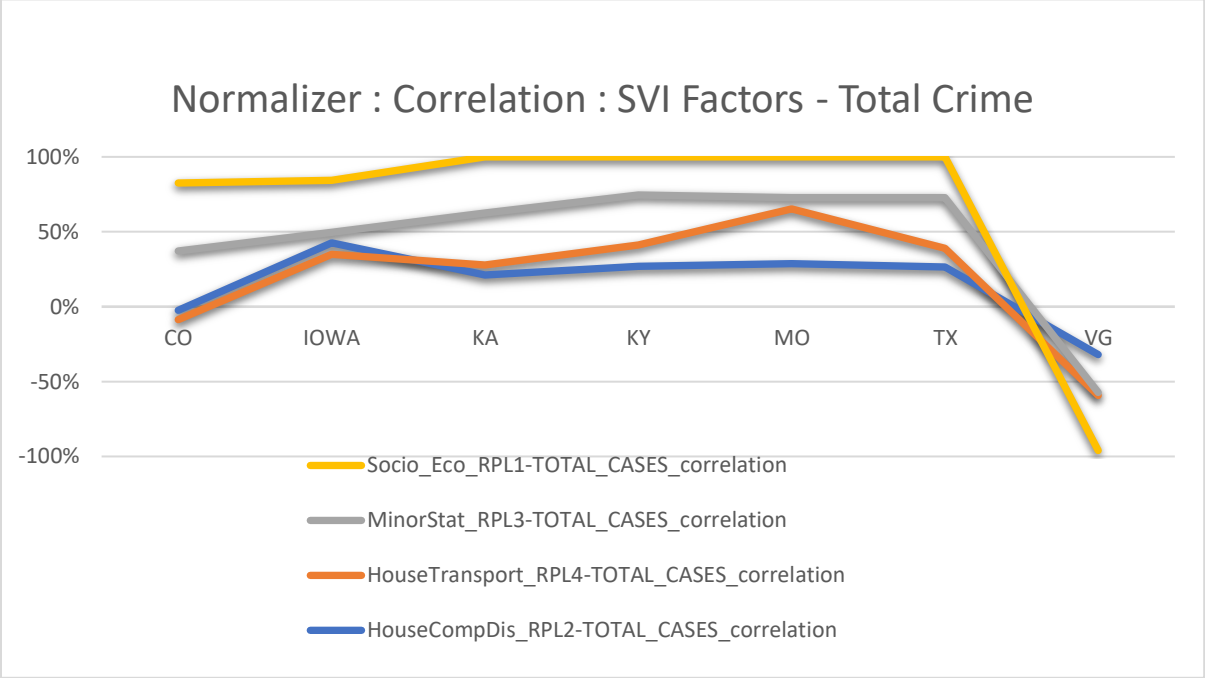


Fig : Normalizer : Correlation : SVI Factors - Total Crime

The next graph ***Normalizer : SVI - Total Crime - Covid 2021*** compares, relation of SVI with covid 2021 and relation of total crimes with covid situation of 2021. From the graph we can see, all these lines are following the same trend. All the states have correlation points very close which was surprising to us. Therefore, we can say that there was correlation with total crime, covid situation and SVI index across the state.

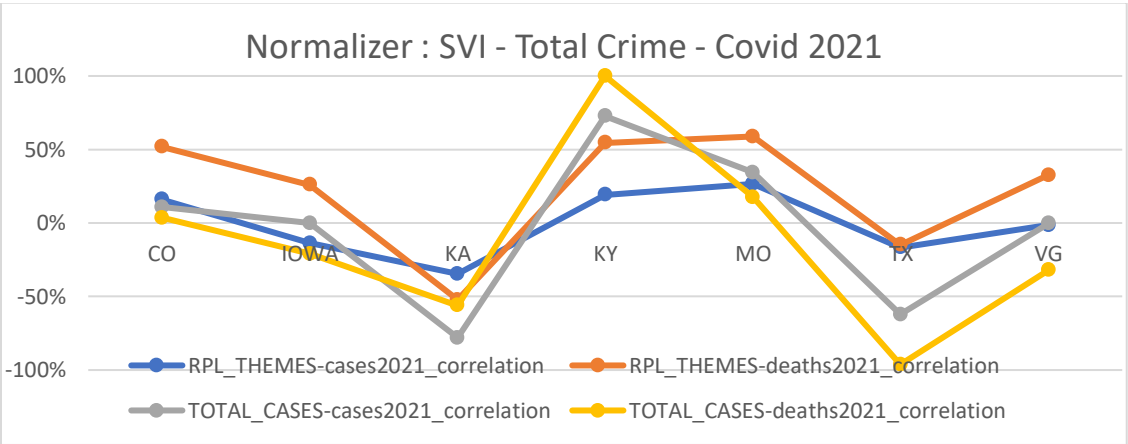


Fig : Normalizer : SVI - Total Crime - Covid 2021

The next pair of graph showing us how same data makes a different correlation trend with a different normalization technique. Even though, these two pair of graph shows a different trend and correlation values

are different than previous technique, we can see all these four graphs follow similar trend across the state. For example, 4 SVI factors had higher correlation with factor of crime and covid for Texas state compare to other states in standard scaler technique (*Fig. : Standard Scaler : Correlation : SVI Factors - Total Crime*). Each graph shows that all correlation points are close to each other across the state which is an interesting finding. Another interesting insight is, scaling plays an important role finding out correlation. Therefore, to get which scaling technique works best require intensive research and experiments.

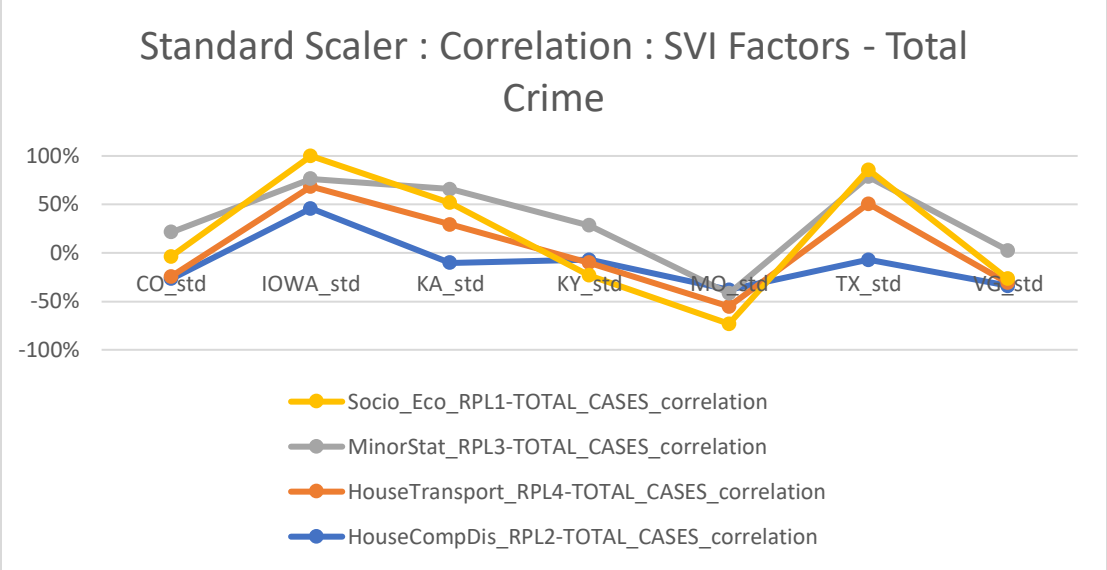


Fig. : Standard Scaler : Correlation : SVI Factors - Total Crime

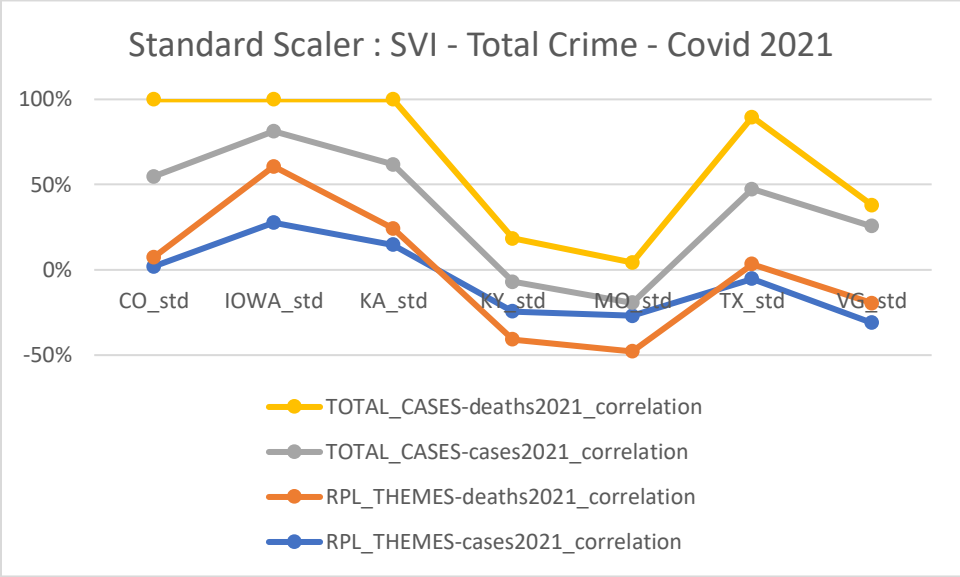


Fig. Standard Scaler : SVI - Total Crime - Covid 2021

State	County Name, Max covid 2020	Min covid 2020	Max death 2020	Min death 2020
IOWA	Polk, 47978	Fayette, 0	Polk, 500	Fayette, 0
KANSAS	Sedgwick, 49884	Stafford, 0	Johnson, 655	Hodgeman, 0
Kentucky	Jefferson, 200826	Greenup, 0	Jefferson,1938	Greenup, 0
Missouri	St. Charles, 31548	Henry, 0	St. Charles, 344	Franklin, 0
Texas	Dallas, 54402	Washington, 0	Dallas, 6040	Washington, 0
Virginia	Stafford, 28044	Gloucester, 0	Shenandoah, 195	Franklin, 0
Colorado	El Paso, 47777	Hinsdale, 16	El Paso, 705	San Miguel, 0

Table1: State wise max-min covid cases 2020 with county name

This table 1 shows, which county had the highest and lowest covid cases and death cases of 2021. In most cases, the counties that had lowest covid cases had lowest death rate with some exceptions. The county that highest death cases across the state was Dallas of Texas which also had the highest covid cases of these seven states in 2020. The table 2 shows which county had best and worst covid situation in 2021. In 2021, The highest death cases happened in Sedgwick of Kansas state with most covid cases.

State	Max covid 2021	Min covid 2021	Max death 2021	Min death 2021
IOWA	Polk, 112416	Fremont, 531	Polk, 948	Worth, 4
KANSAS	Sedgwick, 138916	Lane, 285	Sedgwick, 11175	Greeley, 3
Kentucky	Fayette, 85597	Ballard, 308	Jefferson,690	Ballard, 3
Missouri	St. Charles, 90294	Carter, 1424	Jackson, 395	Carter, 19
Texas	Tarrant, 522861	Roberts, 65	Tarrant, 5277	Loving, 0
Virginia	Loudoun, 62253	Highland, 81	Loudoun, 318	Highland,0
Colorado	El Paso, 170043	Hinsdale, 117	El Paso, 1511	Hinsdale, 0

Table2: State wise max-min covid cases 2021 with county name

Though in 2020, many counties had 0 covid cases with 0 death cases, We saw the scenario went worse in 2021. In 2021, the county that had minimum cases (65) of these seven states was Roberts from Texas. There were also few counties which had 0 death cases in 2021.

6 How the problem space will look like in the future

Looking forward we can see that the Covid pandemic disrupted the global economy in unforeseen ways, as well as how overwhelmed healthcare and difficult it can be to create a cure with little to no knowledge and the general mystery surrounding similar diseases. In the future global, state, and local governments will be looking into ways to minimize the loss of life and impacts on communities. We have already seen a large shift towards a remote economy, learning, and social life.

Similarly, the massive amount of data that was collected by the pandemic will be analyzed for decades to come. The force driving this change will be the desire to create a more effective pandemic response team. Technological advancements in machine learning have allowed us to predict cases and death totals. This machine learning data can be used to train more models, or AI to predict pandemics before they happen. This may include utilizing correlation data, death totals, and SVI to predict vulnerable communities. In the future we can expect to see AI capable of classifying possible outbreaks, creating contingency plans, and providing advice about how to move forward. These AI models may be the new way that pandemic response teams respond to pandemics.

Many hospitals and services around the world have created more robust data collection and archives for the future if another pandemic arrives. This data will be utilized to create contingency plans in case another pandemic occurs. There will likely be new policies and differences in how business is conducted based off the data collected from the pandemic. but it is still hard to say exactly what will happen going forward. We can certainly expect that data collection will become even more important in the future, and this will require larger datacenters, stronger computational power, and unique analysis of how the pandemic became a global phenomenon.

7 Conclusions

In our work, we have performed a set of analysis to understand relationship among three datasets. The reason behind choosing these three datasets is we assumed that there might be some correlations between social vulnerability factors and crime cases. Later we included covid affected cases and death cases to see if there is any effect of social vulnerability or crime cases on any pandemic situation.

Our analysis follows to support our assumptions. Though correlation values are not that much high for each pair of attributes we have considered, there is a similarity of all the correlations for each state and this similarity is consistent across the state too.

The first assertion is for each state all the correlations (SVI with Crime, Crime with Covid or SVI with Covid) behave same. If SVI with Crim correlation is higher for one state, the other two correlations follow the same. And if there is negative correlation (no correlation) for one state, that means that state does not have any dependencies among these three factors: Social vulnerability, Crime and Pandemic.

We also assert this behavior of similarity of different correlation types is consistent across different state. All the states data we have performed experiments on validates the first assertion which results in our second assertion. In respect of social and overall governance aspect, we can assert that all the states act like a bubble where all the dependency relations stay there and does not affect other state much.

According to the methodology of our experiments, we can also assert that the counties under one state defines the relationship among social, economic, good governance, capability of handling emergency situation like

pandemic. As a result, we can come to another conclusion that localization characteristics of counties under the same state represents a clustering nature of different socio-economic aspects like social vulnerability, crime and pandemic.

8 Bibliography

[1] N. Johnson, C. Roman, "Community correlates of change: A mixed effects assessment of shooting of shooting dynamics during COVID-19", *PLOS ONE,* Feb. 2022, Accessed: Apr. 29, 2022. [Online]. Available: HTTPS://WEB-P-EBSCOHOST-COM.EZPROXY2.LIBRARY.COLOSTATE.EDU/EHOST/PDFVIEWER/PDFVIEWER?VID=3&SID=99D2AFAB-AEF4-498A-BCF6-9AA38DB0DC39%40REDIS

[2] H. Ritchie *et al.*, "Coronavirus Pandemic (COVID-19)," *Our World in Data*, Mar. 2020, Accessed: Apr. 29, 2022. [Online]. Available: https://ourworldindata.org/coronavirus

[3] J. Chu, "A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain," *PLOS ONE*, vol. 16, no. 3, p. e0249037, Mar. 2021, doi: 10.1371/journal.pone.0249037.

[4] S. Ballı, "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods," *Chaos Solitons Fractals*, vol. 142, p. 110512, Jan. 2021, doi: 10.1016/j.chaos.2020.110512.

[5] "SUSTAIN - Catalyzing Urban Sustainability Research." https://urban-sustain.org/ (accessed Apr. 29, 2022).

[6]"CDE :: Home." https://crime-data-explorer.fr.cloud.gov/pages/home (accessed Apr. 29, 2022).

[7]A. Paul, P. Englert, and M. Varga, "Socio-economic disparities and COVID-19 in the USA," *J. Phys. Complex.*, vol. 2, no. 3, p. 035017, Jul. 2021, doi: 10.1088/2632-072X/ac0fc7.

[8]"Pearson's Correlation Coefficient - A Beginners Guide," *Analytics Vidhya*, Jan. 06, 2021. https://www.analyticsvidhya.com/blog/2021/01/beginners-guide-to-pearsons-correlation-coefficient/ (accessed Apr. 29, 2022).