An in-depth look at Covid-19; A rebuttal

In a paper that was published in late 2020, as the height of the new novel SARs virus was only starting, a paper was published called, “How Political is the Spread of COVID-19 in the United States? An Analysis using Transportation and Weather Data.” it was suggested that covid-19 was being spread due to political affiliation. in the end, they ultimately decided that it was not the case for the spread, the format in which that analysis was completed seemed to be short-sided, and many corners were cut in the initial planning process for the analysis that was presented.

As a student new to this field, I was all too eager to take on the challenge of looking into the information that, in my opinion at least, was overlooked. Not to mention that with my medical background this analysis really had struck a chord with me. As I work in a place where I am face-to-face with this infection on a daily basis, and I get told people’s opinions about this virus has unsolicited advice just as often, I feel that with that experience I may have a slightly different outlook on this issue than the people who did the original no analysis of that I had read.

As I am well aware that I am doing this analysis paper for school, I know that I probably should write it with a more formal tone. But that would be writing my paper to only accommodate for one potential audience. As I feel that this issue deserves more attention than that, I’m going to be choosing to write it with a more informal tone. This may end up being to my detriment in the end, I feel that if more people can read this and not have it be so dry that their eyes glaze over a few paragraphs in, well then- that’s a win.

So, before I start off on explaining my method and process, let’s first take a look at the other article that got me thinking about this analysis to begin with. The article had taken our 50 states and broken them up into five clusters trying to keep populations similar to each other as well as maintaining the split along the party line. As this article was written in September 2020 the new polling for presidential election of 2020 had not transpired yet. The clusters that they chose to use was Florida to New York, Georgia, Indiana, Tennessee, Michigan, North Carolina, Ohio, Pennsylvania, to Illinois, Massachusetts, New Jersey, Virginia, and Washington in cluster two. Cluster three comprised of Texas and California. Cluster four was West Virginia and South Carolina to Connecticut, District of Columbia, Delaware, Hawaii, Maryland, Maine, New Hampshire, Rhode Island, and Vermont. And finally for cluster five we had Alaska, Alabama, Arizona, Arkansas, Iowa, Idaho, Kansas, Kentucky, Louisiana, Missouri, Mississippi, Montana, North Dakota, Nebraska, South Dakota, Utah, Wisconsin, and Wyoming, compared to Colorado, Minnesota, New Mexico, Nevada, and Oregon.

And where I understand that clustering makes it easier to do an analysis on data of this capacity it also overlooks the fact that a single state is not a single color of red or blue. An easy to see example of this is looking at the state of Illinois. Chicago is very much a blue city because his population there is larger than the rest of the state, the electoral college boats for that state go to the Democrats. However if you were to look at the county data, you would see that the rest of the state is red, and as Republican as they wish they could vote. So taking a state like that and assuming that everyone in that state is going to abide by the same mandates there set out by the government to stay in and quarantine, the reality is that we have to look at the data a lot deeper than by chunking it up into five sections based on red or blue.

The analysis in question also used temperature logs by states to show correlations of infection rates. To be exact a were looking at May June and July for the temperature rates. As I would hope would be common knowledge by most people at this point in their lives, we know that the further south you go the hotter it is. Which means on their chloropleth map they had generated into their analysis shows it being extremely warm from California over to North Carolina. They then used a Pearson correlation coefficients to look at the two sets of data and assume that the higher rates in those warm states were due to the warm weather. And that was all they did for the analysis. They didn’t choose to look deeper into the data to see if there was any other correlation and/or causation for the higher rates in those warm places.

So let’s take a look at the data and see if we can surmise a different reason other than the weather for Covid rates to be so high in some places and not as high in others. We will start with just the Covid infection and death rates for the states:

I would also like to point out that for this analysis, I will be including the US Virgin Islands and Puerto Rico. I chose not to include quantum as because it is so far away for easy travel and that I could not find all of the matching data that I used for the rest of this analysis, that it would not be viable information to include. I chose to include the US Virgin Islands and Puerto Rico as they are part of the United States, even if they are not states. In the United States will not give the land back to the people nor give them statehood, which means that the people who reside on that land should be the responsibility of the United States. So when we’re looking at infection rates for the United States it seems immoral to not include them.Personally, if I had an option not to include anywhere for this analysis it would probably either be Florida because they’re crazy, or New Jersey because that show was awful.

Thanks to the New York Times and their GitHub, I was able to render more accurate data as I was finishing up this analysis. In looking at all those numbers it seems like a mess. So let’s look at the state cases, and then state deaths:

c19 <- read.csv("-c19statedeathdata.csv")  
  
# Setup  
options(scipen=999)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

data("c19", package = "ggplot2")

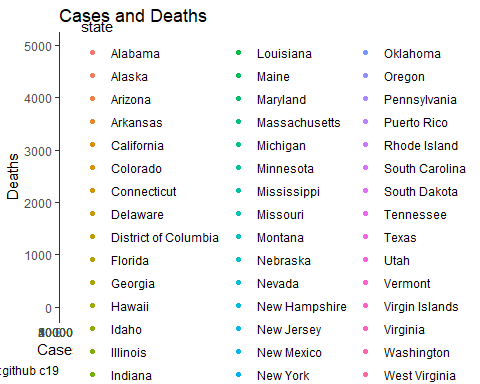
## Warning in data("c19", package = "ggplot2"): data set 'c19' not found

theme\_set(theme\_bw())  
  
# Add plot components --------------------------------  
gg <- ggplot(c19, aes(x=cases, y=deaths)) +   
 geom\_point(aes(col=state, size=population)) +   
 geom\_smooth(method="loess", se=F) + xlim(c(0, 500000)) + ylim(c(0, 5000)) +   
 labs(title="Cases and Deaths", y="Deaths", x="Cases", caption="Source: nyt:github c19")  
  
# Call plot ------------------------------------------  
plot(gg)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 33 rows containing non-finite values (stat\_smooth).

## Warning: Removed 33 rows containing missing values (geom\_point).



As we can see from this graph there are a few states that have quite a few more cases and deaths than others. Which leads us to the ultimate question, what causes those states to have more Covid cases and more deaths than the other states. There has been an assumption placed on the fact that the spread of Covid has been more political than not. But there have been many speculations as to why despite risks has spread so rapidly through not only United States, but through the world.

The data that I have collected for this particular analysis is quite extensive. Therefore I will only be examining a portion of it for this project. I will continue on with my analysis throughout the upcoming hopefully just months, to see if I can reach a conclusion that can be agreed upon, if for no other reason than is proven through the analysis which was conducted.

df <- read.csv("data.csv")  
# Load data  
data("df")

## Warning in data("df"): data set 'df' not found

my\_data <- df[, c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36)]  
# print the first 6 rows  
head(my\_data, 6)

## sc p19 pw pb pn pa pm ph pv pph  
## 1 48 20785296 22819758 3740468 318954 1507785 608913 11511364 866194 2.86  
## 2 13 7815734 6391688 3461279 63704 467166 233583 1051124 377926 2.71  
## 3 12 16763162 16602290 3629737 128866 644332 472510 5670122 829169 2.65  
## 4 6 29586826 28409288 2568294 829756 6124394 1580488 15567815 946070 2.96  
## 5 37 7895647 7404587 2328354 178297 335618 241225 1027832 391408 2.52  
## 6 36 14932565 13539678 3423826 213989 1750820 525246 3754537 437529 2.60  
## hse bde pd nhi clf pip pt la psm  
## 1 14186487 4156640 1347034 6039668 18615355 3943439 28995881 261962.0 110.69  
## 2 5461042 1676539 547993 1648849 6625271 1412117 10617423 57915.0 183.33  
## 3 10793926 3151826 1054860 3507003 12521520 2727672 21477737 53999.0 397.74  
## 4 19129080 6369983 1569092 3528124 24932212 4662442 39512223 156147.0 253.05  
## 5 5370339 1637953 583732 1408475 6429195 1426379 10488084 53664.0 195.44  
## 6 10065502 3613515 884367 1192485 12255743 2528962 19453561 54341.6 357.99  
## vap cf tmp tfp a019 a019f a019m a2064 a2064f  
## 1 20785296 551 14402702 14593179 8210585 4020641 4189944 17051067 8507476  
## 2 7815734 288 5160424 5456999 2801689 1375698 1425991 6298780 3225998  
## 3 16763162 279 10497056 10980681 4714575 2306134 2408441 12265825 6206281  
## 4 29586826 264 19079447 19287602 9629531 4707294 4922237 23074886 11441150  
## 5 7895647 175 5100264 5387820 2592437 1271077 1321360 6144553 3132385  
## 6 14932565 174 9447846 10005715 4520996 2211837 2309159 11636419 5920703  
## a2064m a65 a65m a65f ccc ccd tv nv  
## 1 8543591 3734229 1669167 2065062 2969114 52019 11323985 9461311  
## 2 3072782 1516954 661651 855303 1103192 20421 4998566 2817168  
## 3 6059544 4497337 2029071 2468266 2329859 36972 11067456 5695706  
## 4 11633736 5662632 2523474 3139158 3800049 63647 17395590 12191236  
## 5 3012168 1751094 766736 984358 1008036 13268 5524813 2370834  
## 6 5715716 3296146 1422971 1873175 2106056 52927 7170577 7761988

I was having an issue with my data and showing either an exceptionally high correlation or extremely low correlation when looking at a correlation plot in R.

I then realized that the correlation plot was looking specifically at numbers, and nothing else. Which, thinking about it makes sense as a correlation plot only can look at the data that it is given.

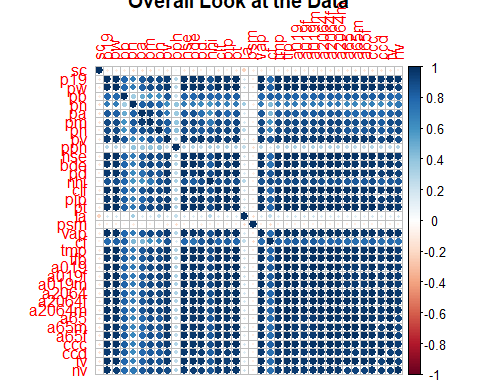
So I reran the numbers that I had changed into percentages and put them back into numerical digits. I then put that data back into a master sheet so that way I would have all the data quickly at hand. When I run a correlation plot against that data it gives me pretty much the same results, except for now pretty much everything is highly correlated. This also is not going to be very helpful as looking at the data, some of the values that it says are not correlated, in my opinion are highly correlated. This means the data will need to be gone through more thoroughly and more in depth to be able to reach an accurate conclusion.

corr\_mat <- cor(df)  
  
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.5

## corrplot 0.89 loaded

corrplot(corr\_mat,title = "Overall Look at the Data")



As the names for all of my data fields ended up being a little large I had to abbreviate most of them for which I kept on a separate sheet and I will include at the bottom of this analysis. However for now if you look at the columns listed LA, and PSM you will notice that there is a either very faint none or very minimal negative correlation. These two fields indicate the land area for the given state as well as the population per square mile. As some states have quite a bit larger areas and metropolitan areas tend to have quite a few more people than rule areas, I believe that these two fields actually have quite a bit more to do with the data at hand and is shown by the correlation plot.

If you happen to live in an area that has a average population of two people per square mile, is far less likely for you to contract virus than if you were living in an area that has a population of one person per every 250 miles as we see in Yucatat County, Alaska- versus living in an area such as New York County, New York which boasts a staggering population of 48,299 people per square mile. Which, is that particular nuance is not able to be seen by looking at just the state data, I will go into that particular portion of my analysis at a later date and time.

library(ggplot2) # visualizing data  
library(gridExtra) # combining multiple plots

## Warning: package 'gridExtra' was built under R version 4.0.5

library(corrgram) # visualizing data

## Warning: package 'corrgram' was built under R version 4.0.5

library(corrplot) # visualizing data   
library(Hmisc) # produces correlation matrices with p-values

## Warning: package 'Hmisc' was built under R version 4.0.5

## Loading required package: lattice

##   
## Attaching package: 'lattice'

## The following object is masked from 'package:corrgram':  
##   
## panel.fill

## Loading required package: survival

## Loading required package: Formula

## Warning: package 'Formula' was built under R version 4.0.3

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(ppcor) # assesses partial correlations

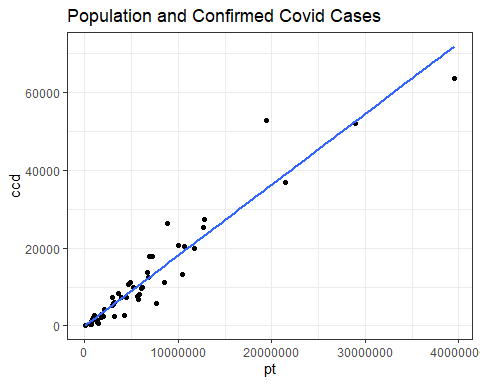
## Warning: package 'ppcor' was built under R version 4.0.5

## Loading required package: MASS

library(gridExtra)  
library(grid)

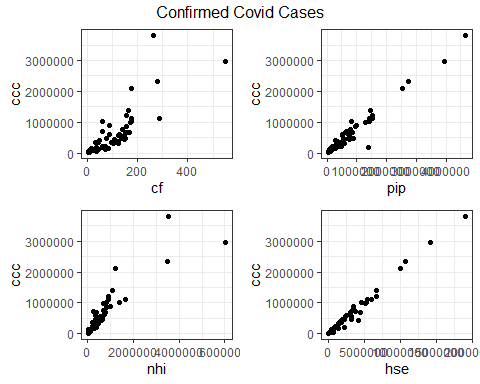
qplot(x = pt, y = ccd, data = df) +  
 geom\_smooth(method = "lm", se = FALSE) +  
 ggtitle("Population and Confirmed Covid Cases")

## `geom\_smooth()` using formula 'y ~ x'



As we can see from this linear regression model the confirmed Covid deaths are directly correlated with the total population of the states. This is not really surprising as states with higher populations are going to end up with more Covid cases and ultimately Covid deaths.

library(gridExtra)  
library(grid)  
  
p1 <- qplot(x = cf, y = ccc, data = df)  
p2 <- qplot(x = pip, y = ccc, data = df)  
p3 <- qplot(x = nhi, y = ccc, data = df)  
p4 <- qplot(x = hse, y = ccc, data = df)  
  
grid.arrange(p1, p2, p3, p4, ncol = 2,   
 top = textGrob("Confirmed Covid Cases"))

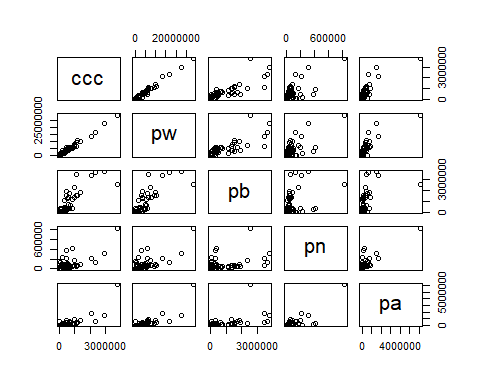


Another linear regression model that I had used on this data was taking a look at the confirmed cases, and looking at any correlation between the correctional facilities in each state, as seen in the top left graph, the confirmed covert cases in any relation to the total population in poverty as seen in the top right graph, confirmed covert cases with any relation to the population with no health insurance, which is seen in the bottom left graph, and finally the total of it cases in regards to the population that has only high school education as seen in the bottom right graph. As expected, once again all this data shows correlation. I feel that the ultimate question behind the nuances of this data set though, is if this is correlation or if there is any possible causation.

I opted to look at a few regression models just looking at the confirmed cases. I broke up the remaining variables into groups as evenly as possible.

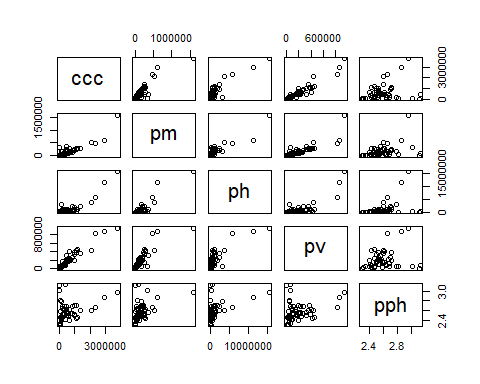
In this model we have PW, PB, PN and PA. These codes are looking at are confirmed Covid cases against our population of Anglo-Saxon descent, R population of African descent , are population of either Native American, native Hawaiian, Pacific Islander or other descent, and are population of Asian descent.

pairs(df[, c(33, 3:6)])



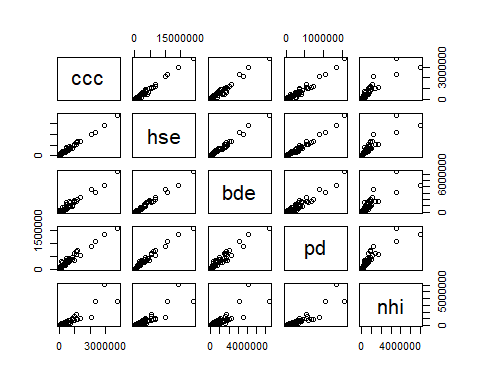
As we can see him this matrix the commonality of each dissent of the population does not show a high correlation with the Covid cases. However this will be looked at in a lot more detail when I go through each state individually, and then compare them to the Covid cases. I expect to find the population origin of familial dissent to be only strongly correlated with the region that is the highest in each given area. For example, I expect to find a higher Hawaiian population in Hawaii that has contracted the Covid virus, then I expect to find in a location such as South Dakota. However because this is going only by state, and overall, I do not find the results from any of these graphs to be strongly to a link of one result or another.

pairs(df[, c(33, 7:10)])



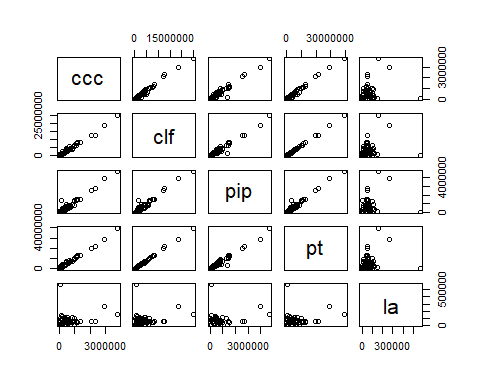
In the above matrix we have again the confirmed Covid cases, and we are looking at codes PM, PH, PV, and PPH. This is looking at our population of mixed descent, our population of Hispanic or Latino descent overall population of veterans, and lastly the average persons per household.

pairs(df[, c(33, 11:14)])



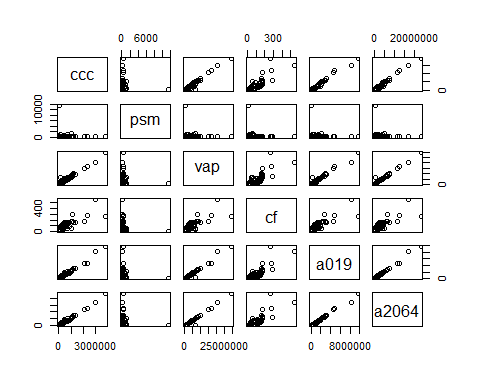
This particular matrix model has a much more defined positive correlation for all the values listed than the one above it does. However I do believe that the PPH which is the person per household on average does have a strong correlation to this virus in the spread of it, this will also have to be looked at at a later time. This matrix shows codes HSE, BDE, PD and NHI. These codes are looking at the population with a high school education, our population with a bachelors degree or higher education, our population that has a disability that is within the working age force, and lastly our population without health insurance. All once again looking if there’s any correlation with the overall confirmed Covid cases.

pairs(df[, c(33, 15:18)])



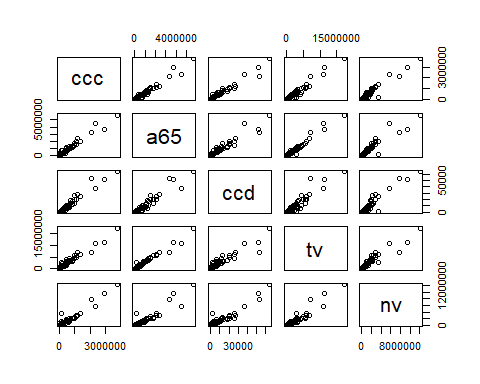
This matrix shows us are confirmed Covid cases against the codes for CLF, PIP, PT, and LA. This is our population age 16 and up in our civilian labor force, or total population in poverty, the population total, and the land area per square mile of states. Once again I know that this correlation plot and regressive model shows that there is no correlation between our land area and are confirmed Covid cases, but again I feel that the lack of correlation is more telling of the fact that there is a correlation.

pairs(df[, c(33,19,20,21,24,27)])



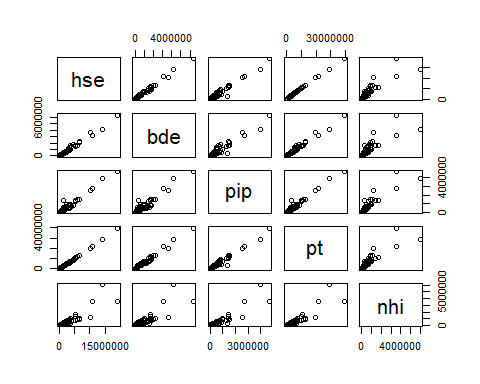
This matrix, we have our total Covid cases against codes PSM, VAP, CF, A019, and A2064. This again is looking at are confirmed Covid cases against our population per square mile, our voting age population, our total correctional facilities, our population between the age of 0 and 19, and our population between the age 20 and 64.

pairs(df[, c(33,30,34,35,36)])



And finally on this plot, for the rest of the major contributing variables, we have confirmed Covid cases against code A65, CCD, TB and and NV. These codes are representative of the variables for our age population of 65 and up are confirmed Covid deaths, our total votes counted, and are persons who are of legal age to vote and did not choose to participate in that process nor are registered to vote.

pairs(df[, c(11,12,16,17,14)])



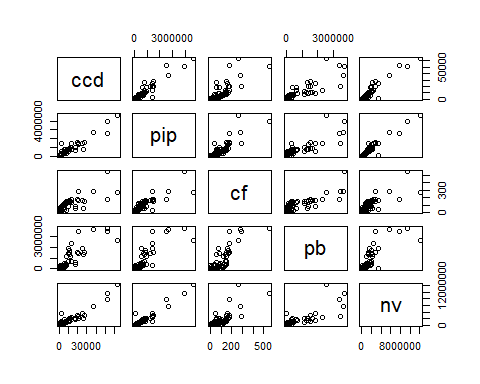
Here is the beginning of some of the correlation matrices I want to look at in more depth later. This is looking at the population that has a high school education, the population with a bachelors degree or higher as well as our population in poverty, the overall population, as well as our population without health insurance. As we can see from the model, except with a a few outliers, the data is pretty strongly correlated. And this is where the data starts to get quite interesting. Goal for our country is to have our persons in poverty percentage be under 10%,

Then only 11 states have achieved this. This would be Nebraska, Virginia, Washington, Massachusetts, Colorado, Hawaii, New Jersey, Maryland, Minnesota, Utah, and New Hampshire. If we want to look at the other end of the spectrum, our top 10 states for the persons in poverty is going to be Oklahoma, with a staggering percentage of 15.2% of their population in poverty. Alabama with 15.5% of their population in poverty, West Virginia with 16% of their population in poverty Arkansas 16.2% Kentucky 16.3% New Mexico 18.2% Louisiana at 19% Mississippi at 19.6% of the population in poverty and our top two winners, if we call them that is Puerto Rico at 43.5% and Virgin Island’s with 44.23% of the population in poverty.

That information alone is quite staggering. However because there are many states that have higher populations those are not necessarily the states that have the highest amount of people in poverty. For that we would have to look at Pennsylvania, Ohio, Arkansas, Illinois, New York, North Carolina, Georgia, Florida, California, and finally Texas. Interestingly enough all the states that are listed with the highest poverty as well as the states that contain the highest amount of people in poverty are all listed in the same data frame that includes the highest amount of correctional facilities in the United States, as well as they are also included on the original analysis paper that I had read. Their analysis stated that the spread of Covid was due to the heat, and the fact that the people were traveling during the summertime.

However looking at Texas with their 551 correctional facilities and their 3.95 million people in poverty, not to mention there are 9.5 million adults that are not eligible to vote in their state, I would have to argue that there is another reason that Covid spread so much in Texas not to mention other states due to something much more than it being hot.

pairs(df[, c(34,16,21,4,36)])



As there is no denying the racial disparity in our country, here I was looking at the confirmed Covid deaths, our population in poverty are correctional facilities, are persons of African descent and our population people who are not eligible to vote, our elected to not register.

And where we see a high correlation with these factors, as is shown in our correlation plots at the beginning of this paper, all of this data seems to be highly correlated save the land area in population per square mile- however again, I feel that the lack of correlation of that information shows a higher correlation than not.

Here is the data from the top 10 states with the lowest correctional facilities.

low <- read.csv("datalow.csv")  
print(low)

## sc p19 pv hse bde pd nhi clf pip pt  
## 1 9 2735705 102746 1905442 741216 153698 250622 2360219 356528 3565287  
## 2 38 561285 26971 408332 120457 31342 61947 527346 80778 762062  
## 3 23 1063054 61989 717899 221830 91779 136153 845509 146519 1344212  
## 4 33 1068673 58503 756957 276289 71703 103745 920524 99258 1359711  
## 5 44 820818 34413 557719 185720 62109 51165 682228 114410 1059361  
## 6 15 1085694 60969 750267 243836 53123 71197 874955 131668 1415786  
## 7 10 744852 38026 499239 156761 47255 79151 606654 110035 973764  
## 8 50 489574 22564 337559 125909 38276 35125 409960 63646 623989  
## 9 78 79890 3247 40687 7811 11509 39941 82634 47062 106405  
## 10 11 556412 17918 424976 244786 40339 27758 490495 95276 705749  
## vap cf ccc ccd tv nv  
## 1 2735705 33 348319 8260 1824317 911388  
## 2 561285 32 110382 1548 361819 199466  
## 3 1063054 28 68449 843 816428 246626  
## 4 1068673 17 99040 1360 804430 264243  
## 5 820818 17 152187 2722 513627 307191  
## 6 1085694 13 35522 502 574469 511225  
## 7 744852 10 109258 1679 504010 240842  
## 8 489574 9 24314 256 367428 122146  
## 9 79890 3 4002 88 17561 62329  
## 10 556412 3 49151 1137 344356 212056

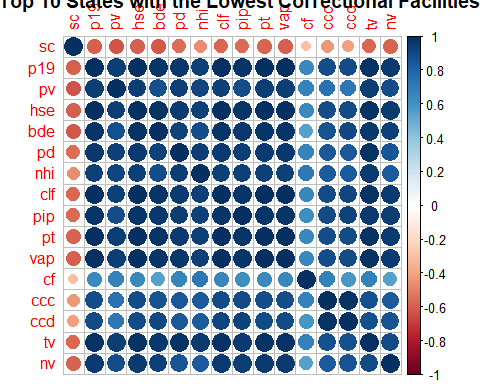
This is the information from the states with the highest amount of correctional facilities.

Incidentally enough are states that have the top 10 correctional facilities: Texas, Georgia, Florida, California, North Carolina, New York, Ohio, Michigan, Virginia, and Illinois. Also show up on our states with the highest amount of Covid cases and the highest amount of Covid deaths in the country. States are California, New York, Texas, Florida, Pennsylvania (which is state number 11 for the highest amount of correctional facilities), New Jersey, Illinois, Michigan, Georgia, and Ohio.

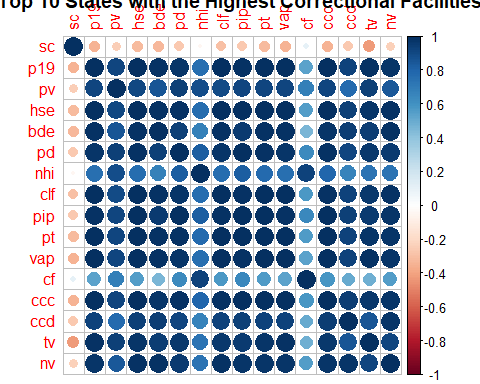
high <- read.csv("datahigh.csv")  
print(high)

## sc p19 pv hse bde pd nhi clf pip  
## 1 48 20785296 866194 14186487 4156640 1347034 6039668 18615355 3943439  
## 2 13 7815734 377926 5461042 1676539 547993 1648849 6625271 1412117  
## 3 12 16763162 829169 10793926 3151826 1054860 3507003 12521520 2727672  
## 4 6 29586826 946070 19129080 6369983 1569092 3528124 24932212 4662442  
## 5 37 7895647 391408 5370339 1637953 583732 1408475 6429195 1426379  
## 6 36 14932565 437529 10065502 3613515 884367 1192485 12255743 2528962  
## 7 39 8802227 421568 6087072 1692206 682346 915126 7375822 1531272  
## 8 26 7579167 329059 5261458 1504776 598817 691999 6121943 1298291  
## 9 51 6448093 408152 4544643 1736053 407134 796347 5479803 845016  
## 10 17 9526512 351714 6652635 2268548 531312 1093517 8249355 1457259  
## pt vap cf ccc ccd tv nv  
## 1 28995881 20785296 551 2969114 52019 11323985 9461311  
## 2 10617423 7815734 288 1103192 20421 4998566 2817168  
## 3 21477737 16763162 279 2329859 36972 11067456 5695706  
## 4 39512223 29586826 264 3800049 63647 17395590 12191236  
## 5 10488084 7895647 175 1008036 13268 5524813 2370834  
## 6 19453561 14932565 174 2106056 52927 7170577 7761988  
## 7 11689100 8802227 174 1106064 20021 5929361 2872866  
## 8 9986857 7579167 169 994668 20698 5544102 2035065  
## 9 8535519 6448093 166 677425 11270 4460524 1987569  
## 10 12671821 9526512 161 1390705 25430 6019642 3506870

corr\_mat <- cor(low)  
  
library(corrplot)  
corrplot(corr\_mat,title = "Top 10 States with the Lowest Correctional Facilities")

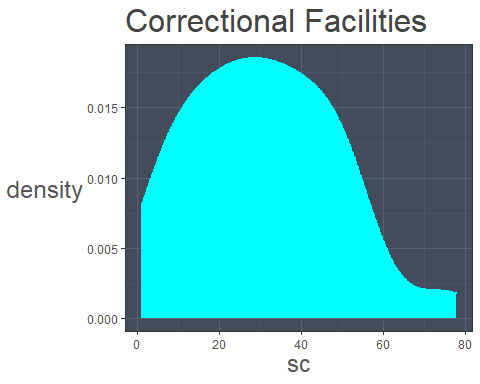


corr\_mat <- cor(high)  
  
library(corrplot)  
corrplot(corr\_mat,title = "Top 10 States with the Highest Correctional Facilities")



Although hard to determine, due to the high correlation of all my values together, you can see with the top 10 states with the highest correctional facilities you can see that there is a higher number of individuals who do not have health insurance. Which on the graph that shows the top 10 states with the lowest number of correctional facilities they have more people who are insured.

ggplot(data = df, aes(x = sc)) +  
 geom\_density(fill = 'cyan', color = 'cyan') +  
 labs(title = 'Correctional Facilities') +  
 theme(text = element\_text(family = '', color = "#444444")  
 ,panel.background = element\_rect(fill = '#444B5A')  
 ,panel.grid.minor = element\_line(color = '#4d5566')  
 ,panel.grid.major = element\_line(color = '#586174')  
 ,plot.title = element\_text(size = 24)  
 ,axis.title = element\_text(size = 18, color = '#555555')  
 ,axis.title.y = element\_text(vjust = .5, angle = 0)  
 ,axis.title.x = element\_text(hjust = .5)  
 )



Here is a kernel density plots where I tried to show where the states with the highest amount of correctional facilities are compared to other locations. As we can see from this density graph the states that generally have the highest amount of correctional facilities tend to reside between mistake codes 20 and 40. This is congruent with the information that is shown above, however due to the fact that the facilities vary greatly from state to state, this plot does not show the nuances of that information. Therefore I will not be using a density plot to show this information going forward, as the plot does not show the data accurately. At least not for my needs.

pca <- prcomp(df)  
summary(pca)

## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 16108888.3168 1298965.3489 1116798.25329 736610.04657  
## Proportion of Variance 0.9835 0.0064 0.00473 0.00206  
## Cumulative Proportion 0.9835 0.9899 0.99465 0.99671  
## PC5 PC6 PC7 PC8  
## Standard deviation 652360.93029 480613.78483 331522.41895 200001.70462  
## Proportion of Variance 0.00161 0.00088 0.00042 0.00015  
## Cumulative Proportion 0.99832 0.99920 0.99962 0.99977  
## PC9 PC10 PC11 PC12  
## Standard deviation 161703.8808 117479.24882 84911.28988 75938.08286  
## Proportion of Variance 0.0001 0.00005 0.00003 0.00002  
## Cumulative Proportion 0.9999 0.99992 0.99995 0.99997  
## PC13 PC14 PC15 PC16 PC17 PC18 PC19  
## Standard deviation 54898.51672 45264.83504 34200 31526 22720 18262 13417  
## Proportion of Variance 0.00001 0.00001 0 0 0 0 0  
## Cumulative Proportion 0.99998 0.99999 1 1 1 1 1  
## PC20 PC21 PC22 PC23 PC24 PC25 PC26 PC27 PC28 PC29  
## Standard deviation 10257 8650 5789 5040 1421 1317 888.3 323.8 18.04 11.14  
## Proportion of Variance 0 0 0 0 0 0 0.0 0.0 0.00 0.00  
## Cumulative Proportion 1 1 1 1 1 1 1.0 1.0 1.00 1.00  
## PC30 PC31 PC32 PC33  
## Standard deviation 0.07066 0.000000001592 0.000000001592 0.000000001592  
## Proportion of Variance 0.00000 0.000000000000 0.000000000000 0.000000000000  
## Cumulative Proportion 1.00000 1.000000000000 1.000000000000 1.000000000000  
## PC34 PC35 PC36  
## Standard deviation 0.000000001592 0.000000001592 0.000000001592  
## Proportion of Variance 0.000000000000 0.000000000000 0.000000000000  
## Cumulative Proportion 1.000000000000 1.000000000000 1.000000000000

library(devtools)

## Loading required package: usethis

## Warning: package 'usethis' was built under R version 4.0.5

library(plyr)

## Warning: package 'plyr' was built under R version 4.0.5

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:Hmisc':  
##   
## is.discrete, summarize

## The following object is masked from 'package:corrgram':  
##   
## baseball

library(ggbiplot)

## Loading required package: scales

## Warning: package 'scales' was built under R version 4.0.5

df.pca <- prcomp(df, center = TRUE,scale. = TRUE)  
summary(df.pca)

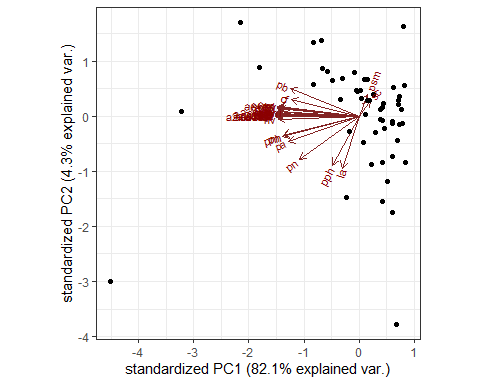
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 5.437 1.23877 1.10973 1.05538 0.88244 0.81039 0.57065  
## Proportion of Variance 0.821 0.04263 0.03421 0.03094 0.02163 0.01824 0.00905  
## Cumulative Proportion 0.821 0.86365 0.89786 0.92879 0.95043 0.96867 0.97771  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.44284 0.41458 0.35932 0.30773 0.26541 0.24383 0.15179  
## Proportion of Variance 0.00545 0.00477 0.00359 0.00263 0.00196 0.00165 0.00064  
## Cumulative Proportion 0.98316 0.98794 0.99152 0.99415 0.99611 0.99776 0.99840  
## PC15 PC16 PC17 PC18 PC19 PC20 PC21  
## Standard deviation 0.14357 0.12231 0.09386 0.08252 0.05284 0.04786 0.02758  
## Proportion of Variance 0.00057 0.00042 0.00024 0.00019 0.00008 0.00006 0.00002  
## Cumulative Proportion 0.99897 0.99939 0.99963 0.99982 0.99990 0.99996 0.99999  
## PC22 PC23 PC24 PC25 PC26 PC27  
## Standard deviation 0.01702 0.01191 0.00654 0.005374 0.004373 0.002269  
## Proportion of Variance 0.00001 0.00000 0.00000 0.000000 0.000000 0.000000  
## Cumulative Proportion 0.99999 1.00000 1.00000 1.000000 1.000000 1.000000  
## PC28 PC29 PC30 PC31  
## Standard deviation 0.00134 0.000905 0.0000592 0.0000000000000004963  
## Proportion of Variance 0.00000 0.000000 0.0000000 0.0000000000000000000  
## Cumulative Proportion 1.00000 1.000000 1.0000000 1.0000000000000000000  
## PC32 PC33  
## Standard deviation 0.0000000000000004963 0.0000000000000004963  
## Proportion of Variance 0.0000000000000000000 0.0000000000000000000  
## Cumulative Proportion 1.0000000000000000000 1.0000000000000000000  
## PC34 PC35  
## Standard deviation 0.0000000000000004963 0.0000000000000004963  
## Proportion of Variance 0.0000000000000000000 0.0000000000000000000  
## Cumulative Proportion 1.0000000000000000000 1.0000000000000000000  
## PC36  
## Standard deviation 0.0000000000000004963  
## Proportion of Variance 0.0000000000000000000  
## Cumulative Proportion 1.0000000000000000000

str(df.pca)

## List of 5  
## $ sdev : num [1:36] 5.437 1.239 1.11 1.055 0.882 ...  
## $ rotation: num [1:36, 1:36] 0.0245 -0.1822 -0.1828 -0.1501 -0.1321 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:36] "sc" "p19" "pw" "pb" ...  
## .. ..$ : chr [1:36] "PC1" "PC2" "PC3" "PC4" ...  
## $ center : Named num [1:36] 30.7 4813340 4767652.8 840240.3 93786.6 ...  
## ..- attr(\*, "names")= chr [1:36] "sc" "p19" "pw" "pb" ...  
## $ scale : Named num [1:36] 17.9 5420072.7 5415141.8 1033324.1 136885.9 ...  
## ..- attr(\*, "names")= chr [1:36] "sc" "p19" "pw" "pb" ...  
## $ x : num [1:53, 1:36] -17.51 -3.6 -11.73 -24.5 -3.17 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : NULL  
## .. ..$ : chr [1:36] "PC1" "PC2" "PC3" "PC4" ...  
## - attr(\*, "class")= chr "prcomp"

Principal component analysis is generally used for data sets that are large and wide, and in this overall plot you can see that a lot of my lines lay along zero. I think that I will definitely use this type of plot in my next analysis when I am going through each and every state county/FIPS code, an in-depth getting more in-depth information from this analysis. For now, because the data is so large and varies so much, range of creates and around each group, which would in-depth being each state would in-depth even messier looking graph.

ggbiplot(df.pca)



So for now we will leave this graph as is, and move on.

As we can see, from the amount of data and information that I have even in this analysis, it is far from being over. What will be coming in the future is a more in-depth analysis of each of the factors in each of the counties that is represented. I have attempted on multiple occasions to obtain demographic information from the prisons and jails of each of the states, however, the data they have provided to me seems inaccurate. For example I had one prison in Texas advise me of a listed prison capacity of over 2500 with a current inmate population of only 97. As a quick Google search would identify that America’s prisons are over packed and at capacity or above, I find it highly unlikely that a prison would have a population of sub 100. Another issue that will have to be addressed in one way or another is the ICE population of prisoners. The United States is supposed to have a record of all prisoners in this population, and there are some registries of the population. But as the New York Times as well as many other new sites has made us aware of the issue on the border where parents and children were separated and kept in sight of empty warehouses the idea that we will have an accurate representation of the population seems unlikely.

What I will eventually show in my in-depth analysis of the data that I have collected scrubbed and looked over for general correlations and research for causations is that the widespread cases of Covid and the eminent deaths caused all starts from a different type of systemic issue.

This issue, cannot be summed up quickly or easily although there have been many papers and analysis that tried. The spread of Covid, had nothing to do it the temperature in the given cities or states, it had nothing to do with the political leaning of the population of people in the states. Populations in poverty, populations without adequate education, populations in correctional facilities, and the populations that come out of those facilities, the persons who were not eligible to vote – not always because one chooses not to possibly and more likely, due to the fact that their voting rights were stripped, populations that are living in areas with thousands of people per square mile, and the population of workers all contribute to the spread of this disease.

As it is a standard practice before the courts nationwide to use a recidivism scale, people who are not of Anglo-Saxon dissent generally end up with a higher rate and end up in jail, or prison- sometimes before even their trial. As jails are overpopulated, social distancing is not an option. And this becomes a petri dish for the disease to spread. As one of the scales that is looked at when assessing someone’s flight risk or the fact that they need to be in jail is generally looked at there is any family history of family being in jail.

As the article about how spread of Covid was political or not using the transportation and weather data analysis first got me thinking about how each county was different and that although one state may be read the population inside of it is not always 100% that way. Where you can live in an extremely liberal area and have a neighbor next door who does not believe in wearing masks and believes that Covid is a government conspiracy, another article that was published by Pro-Publica in 2016, got me thinking about another potential reason for the spread of Covid.

This article entitled “Machine Bias” goes into depth about the process used to keep inmates in jail. Under speaking to multiple social workers, as well as people in the court system, it was apparent that this is actually the case for many people. If you have a family member who was in jail you are than at higher risk for ending up in jail yourself. If you end up with a felony for one reason or another you are no longer allowed to vote. If you go to jail or are convicted of a crime, then you must declare it on your job application and your housing applications. This label of an ex-convict or someone with a criminal record in some way or another will then follow you. It makes it very difficult to obtain a higher education, or a higher paying job. You’re not able to obtain higher paying job there is a good chance that you will end up in an impoverished state physically or figuratively.

When Covid first came to United States, and almost every job shutdown or insisted that their employees start working from home, this was only a luxury afforded to people who had a job that allowed it. If the only job you’re able to find is a fast food worker or a customer service job that works face-to-face with the general public because of either lack of education or due to the fact that you have a criminal record, then staying home and keeping yourself quarantined in safe is no longer a possibility.

My later analysis will show that the populations that have higher members of people who have been in the prison system in one way or another or have less than college education, are the ones that were infected far more than anyone else.

The death rates did affect the elderly and disabled populations more yes, however, there infection was driven from the fact that members of the public that they interacted with were not able to stay home because they needed to work to feed either themselves or their family.

States and areas that had higher education and less correctional facilities overall ended up with a smaller percentage of their population were either infected with Covid or died from the virus.

That was only one leg of the chair that causes virus to spread. Another leg came from the overall general misinformation of our population. The fact that college is so expensive and mostly unobtainable for many people keeps our population in a predicament that leads them to believe that what is written on Facebook is truth. They have not been taught how to research and to think critically when it comes to the information they are given and bad word-of-mouth spreads faster than any of information from the CDC ever could.

This gives us our second leg. However trying to determine if it is the lack of education that causes the desire for misinformation, is another analysis that would need to be done at a different point.

The third leg, I touched on briefly before. The amount of poverty that we have the United States. As wage disparity is pretty well-known at this point, not to mention the fact that most of the will this not among people, the poverty line for the United States although it looks as if it is somewhat reasonable at on average 13%, is not exactly accurate. People that are earmarked for being poverty generally are not the only ones. As the federal minimum wage is still 7.25 an hour, and state I live in it is 13.69, that wage may seem like a fair amount. Unfortunately the average studio apartment is about 1400, at a minimum wage job if no taxes were taken out of your check, it was still take 102.25 hours just to pay your rent. My husband, has a bachelors degree the average pay for his job is 15-20 an hour. So even with a higher education, many people still end up in what should be considered poverty however due to the federal average it is not.

Another portion of this Lake also leads to the fact that if childcare is needed, it is extremely expensive. Therefore if you are a single parent it makes it your and possible to be able to afford a place to live, afford food, electric, and childcare. To be able to have state assistance to meet any of these needs you have to make very little. It is good that these services are available, however the threshold to be able to get these services is very small which means that not everyone who desperately needs the services is able to get them.

The fourth leg to our pandemic chair goes right with our other three is expected. This is the amount of people per each state who do not have healthcare. And I’m going to expand the umbrella of healthcare a little bit wider, as many Americans may have healthcare but have extremely high deductibles which discourage them from going to the doctor, or do not have sick leave with her place of employment for one reason or another.

I don’t think I know a single person who can say that they have never in their life gone to work when they knew they were sick. Personally, I have been sent home with an active case of the flu before - and so have many of my coworkers. And not the flu as most people think where you don’t feel good and your stomach hurts and you’re throwing up, but an active case of influenza. This may be a good time to tell you that I work in a doctor’s office next to the hospital. I have even had the doctors that I work with come into work actively sick.

The major hospital that I work with has locations in 47 states and 113 countries across the world. This hospital offers all of their employees one of two packages of insurance. You can elect to pay 300-600 per paycheck for a very low deductible and many of your costs met or you can choose the second option which is only hundred dollars out of your check but comes with a 3000-6000 deductible before any of your costs are covered.

Many Americans who were forced – or at least how they believe is forced – also opted to have these low cost upfront high deductible plans because that is all that they can afford. With insurance plan like that even though you can say you have insurance it might as well be is if you don’t. Because if all your money goes towards basic living expenses, then how are you going to afford a $300 Doctor visit if you’re sick? The problem is, is that you can’t.

And I know this analysis was long, as well as my conclusion was also verbose. But this is an issue that is not simple and not easy to rectify. However, if we are ever going to be able to make sure our country can withstand another pandemic such as the one we faced in 2020, then we need to start fixing these issues from the ground up.

So, thank you for your time and, please standby for my next bit of analysis on this data.