Rickord, Jake JJ
2021-7-14
Data Mining
Final Project Case Study

Isolation. Monotony. Sickness and Death. Our previous year can be summed up in those 3 succinct descriptions for the majority of us. With the rapid spread and malevolence of COVID-19, we've likely suffered from one of those elements shown above. As we peak into the horizon and continue to meld back into a world that looks akin to the one we left some short 12-15 months ago, I feel it's important to do a post-mortem on our efforts to stop the virus. Through this analysis, perhaps we can pull actionable insights as we move forward and face future challenges such as the rising threat of the Delta variant, as well as other crucibles we may encounter in the coming years. In our disposal, we have access to a list of all 50 states with data spanning death totals, recovered totals, testing done, state population, and other indicators we can use to investigate the effect that state restrictions have had on stemming the spread.

With how divisive the political opinions have been on either side of the aisle have been regarding the virus, our goal from this analysis is not to be a breakdown of different factors such as false positive testing nor an investigation into whether deaths were solely from covid or not, rather, we aim to simply show that based on statistical values we have from the dataset available, in conjunction with the general restrictions employed by each state, whether we can say that the data supports that restrictions aided in limiting the number of cases (with respect to population) of the coronavirus.

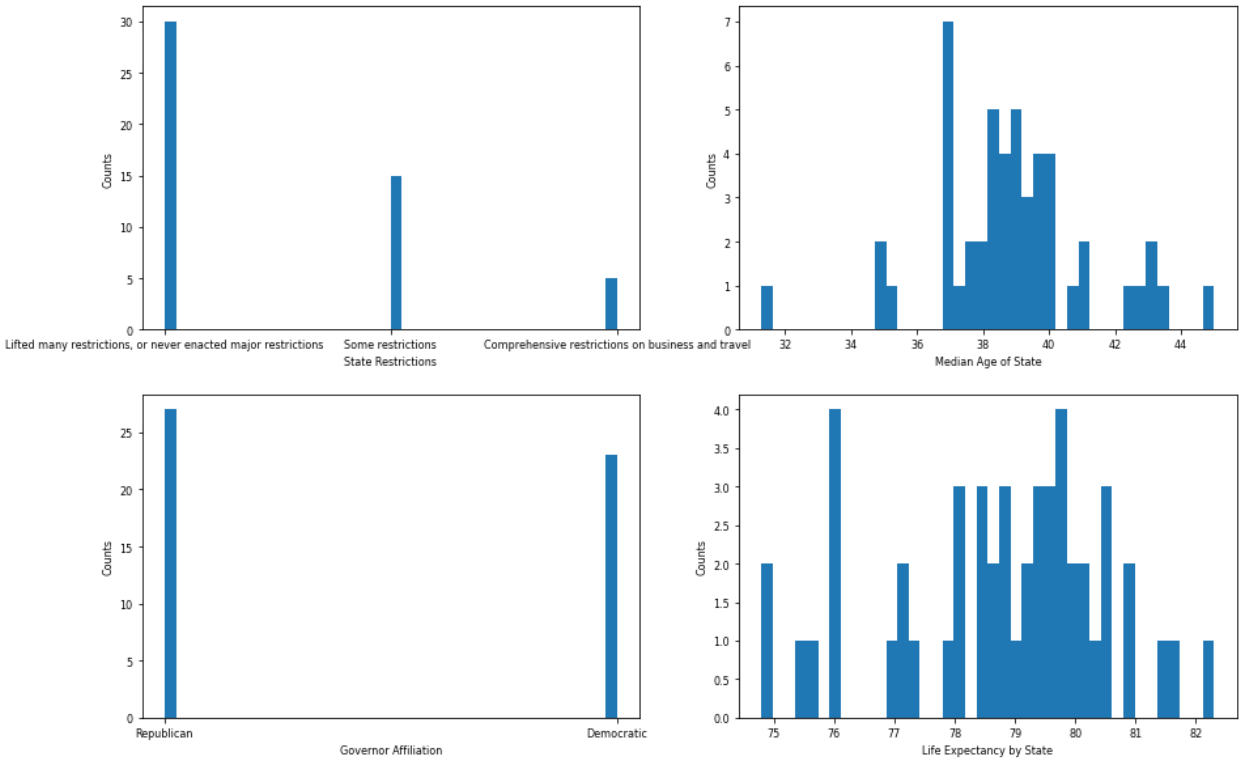
To begin this process, we needed to make the first leap, which was completed via Milestone 1. We needed to find relevant datasets we could clean and incorporate into our project to base our research, data manipulation, modeling, and results off of. With the recency of COVID's data, I found it difficult to find only one or two sources that would compound to contain all the relevant information I would have liked to compare as it pertained to the analysis I was planning on performing. Instead, I was able to identify, read in, and coagulate several different originators of information that would be pertinent to our project.

The key player in starting our base of data was a source found in Kaggle. With a simple download and pandas read call we were able to bring that dataset into the fold. We now had access to statistics for each state regarding Covid total cases, deaths, recovered, active cases, and often more importantly Total Cases/1mil, death/1mil, total tests, tests/1mil, and overall population of the state. We were able to backfill a few pieces of missing data including percent recovered for a few states that did not contain that data previously. Previously we leveraged the mean for the fill, however, per Professor Werner's suggestion, we modified our calculation mechanism to the median in order to have this percentage swayed less by potential outliers (there were a few noticed in our data surrounding active case number representation).

For those lacking Active case numbers, we leveraged some simple arithmetic to fill in additional missing details. From here, we added in what many would consider the columns of essential

pieces to our case study: State restriction level, median age of state, life expectancy of each state, and governor affiliations. We would treat these features as the primary independent variables we would test on, the most important of which being the state restrictions, the correlation of which could lend credence for or against their use in stemming the spread of COVID-19 in the U.S. and elsewhere.

Once we'd appended these additional columns to our pre-existing dataframe from various JSON, HTML, and csv sources (mostly pulled from wikipedia and other webscraped or downloaded web charts), we could then begin the second half of Milestone 1: using this newly established dataset with graphics in order to perform introductory analysis visually. We compiled graphics that ran multiple different comparisons to start the conversation of how these variables interacted with each other. The first of which was a simple histogram run for all 4 relevant independent variables:

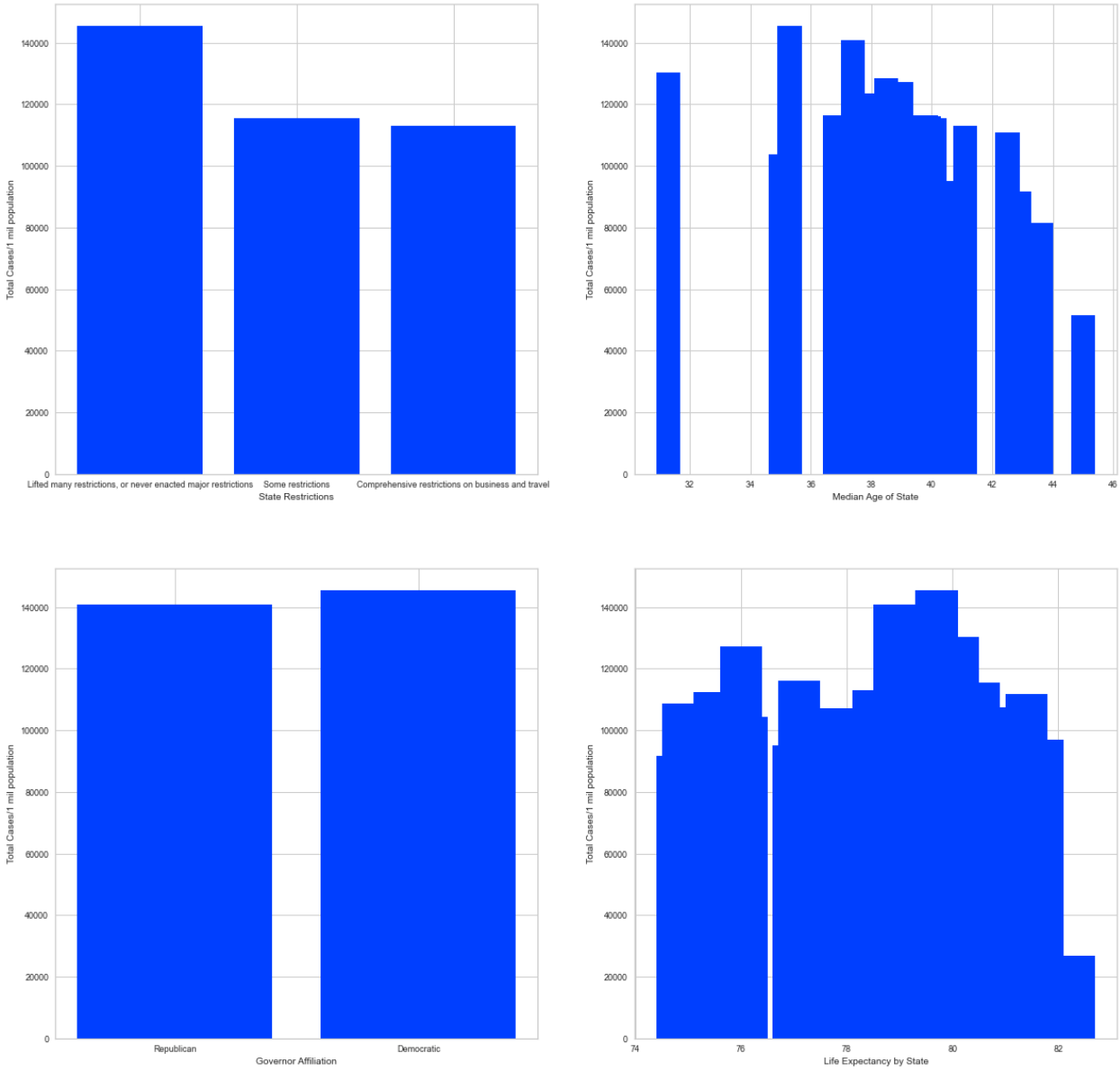


From this graph of features, we can see a few observations:

1.Looks like we can see a good amount of comparisons from no restrictions against some restrictions, though comprehensive restrictions has a much smaller sample size.
2.Median age looks well distributed (though looks like there's possibly an outlier for a state short of 32. We'll look at that
3. Republican vs democrat looks well distributed

4. Life Expectancy also looks well distributed

From there, we jumped into a direct visualization of our question: Visual 2 depicts our features charted against their case number/1 mil population:

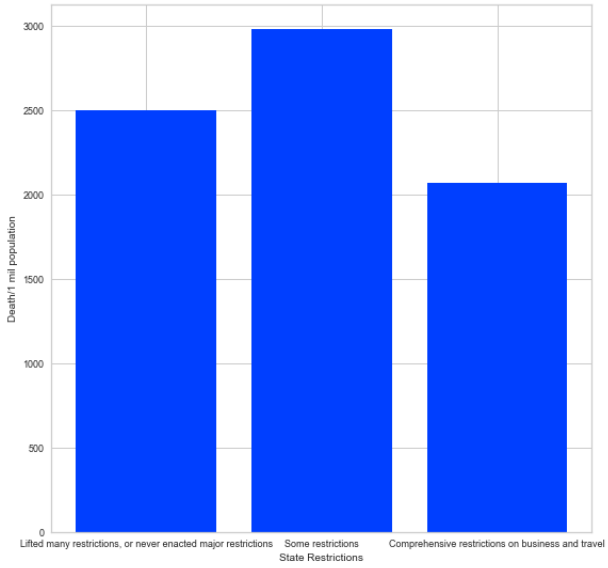


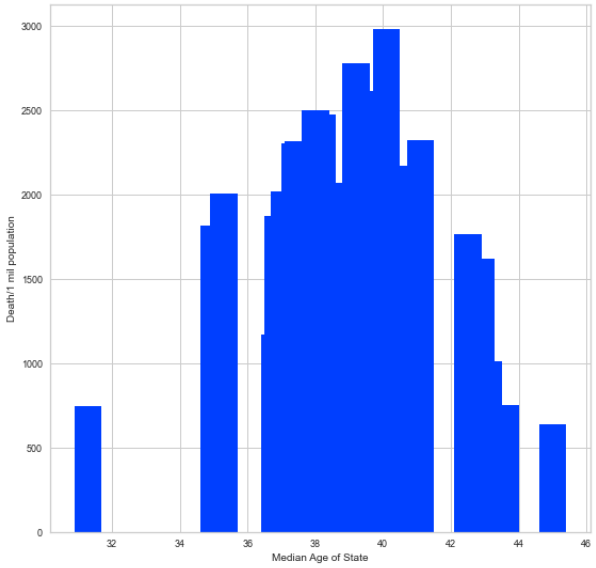
Examining these graphs for items that stand out in comparison to our histograms, we find that most are in-line with those initial graphs. One interesting point may be to see that while no restrictions looks to have had more cases, some restrictions vs heavy restrictions seems to not have had much effect on lowering their respective cases. Another piece of interest is the political affiliations, which seem to show that regardless of governor political affiliation (which one might have tied to restriction level), seemed to have little effect. In fact, it looks like despite having less

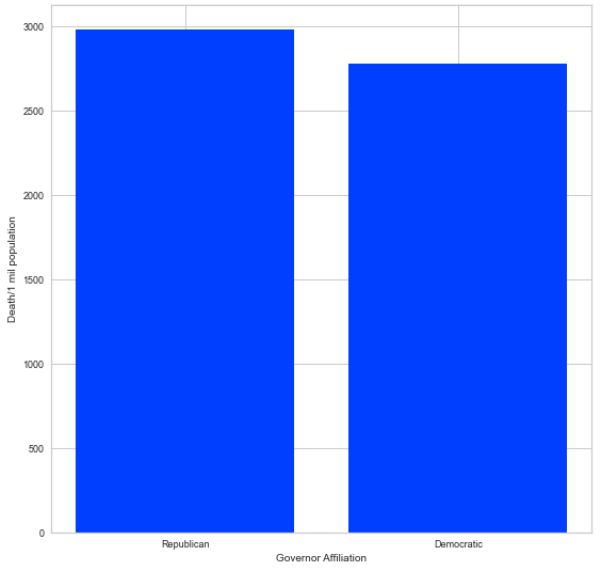
democratic governors, there might actually be more cases per population in those areas. This might be offset by testing levels, but still a highlightable point.

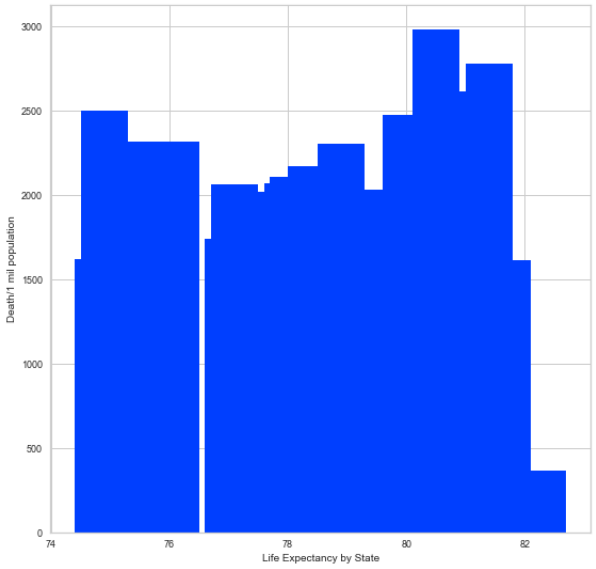
Other than those more obvious ones, the life expectancy and median state age seemed to have little disparity between their total case charts and their distributions. One might have expected that if median age was higher, than that would mean more elderly population would be affected due to slowed immune systems possibly, however if anything, it looks like those skewed towards younger generations were more heavily infected.

Mentioning this difference in immune systems also led us to discussions of death, which provided this graphic of features vs death counts, which, while not directly answering our original question on spreading the virus, is still an important aspect to be understood:



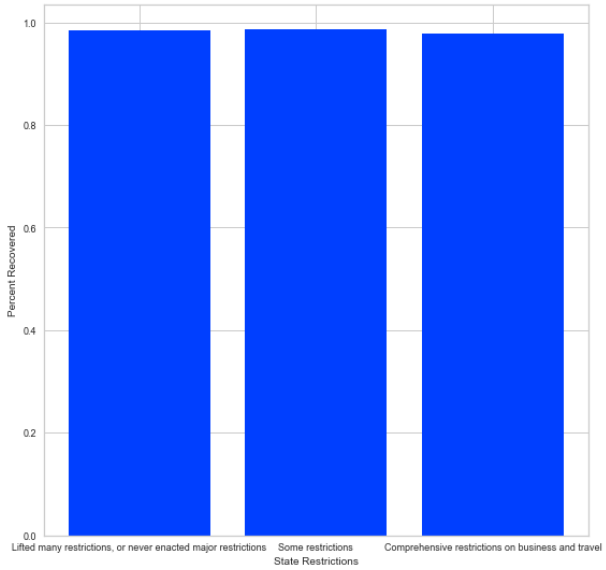


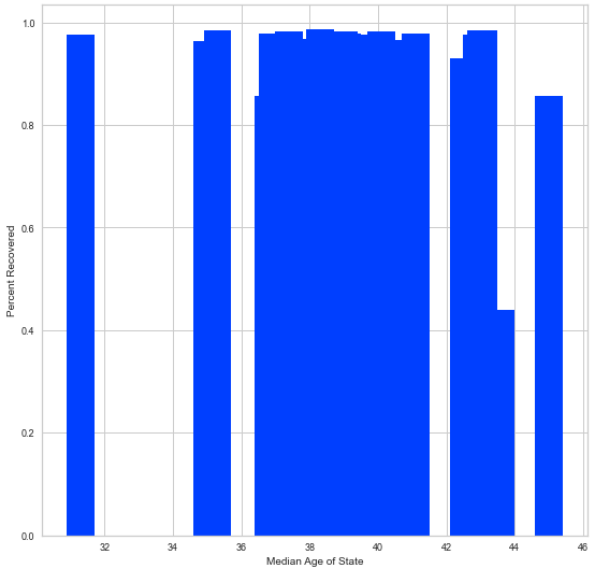


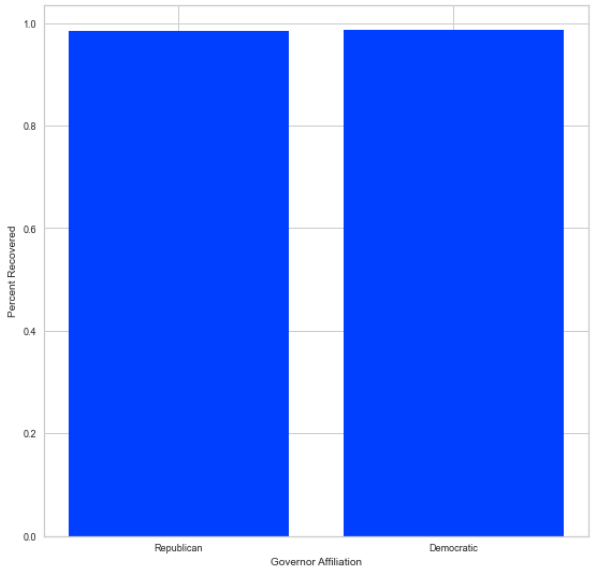


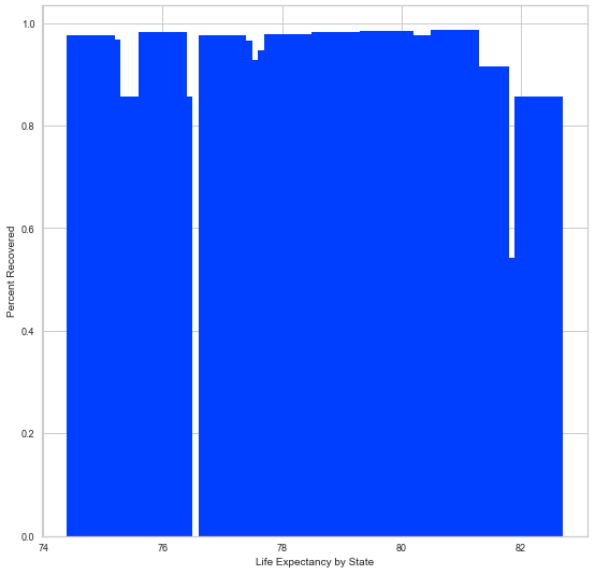
Lots of interesting insights here. Let's go from left to right here. Despite being less common than states with no restrictions, states with some restrictions had a higher death/1 million pop. Looks like peak for deaths/1 mill did indeed move further towards older age states on med age chart, however it didn't move as far as I would've thought. For partisanship, it looks like while democratic governors may have had more cases, republican governors dealt with more deaths/1 mill. Lastly, the life expectancy chart seems to have little to do again here, and visually seems like the least impactful of the feature vectors.

Tied somewhat directly to this, we capped our graphical analysis off with a visualization of recovery rate vs features:





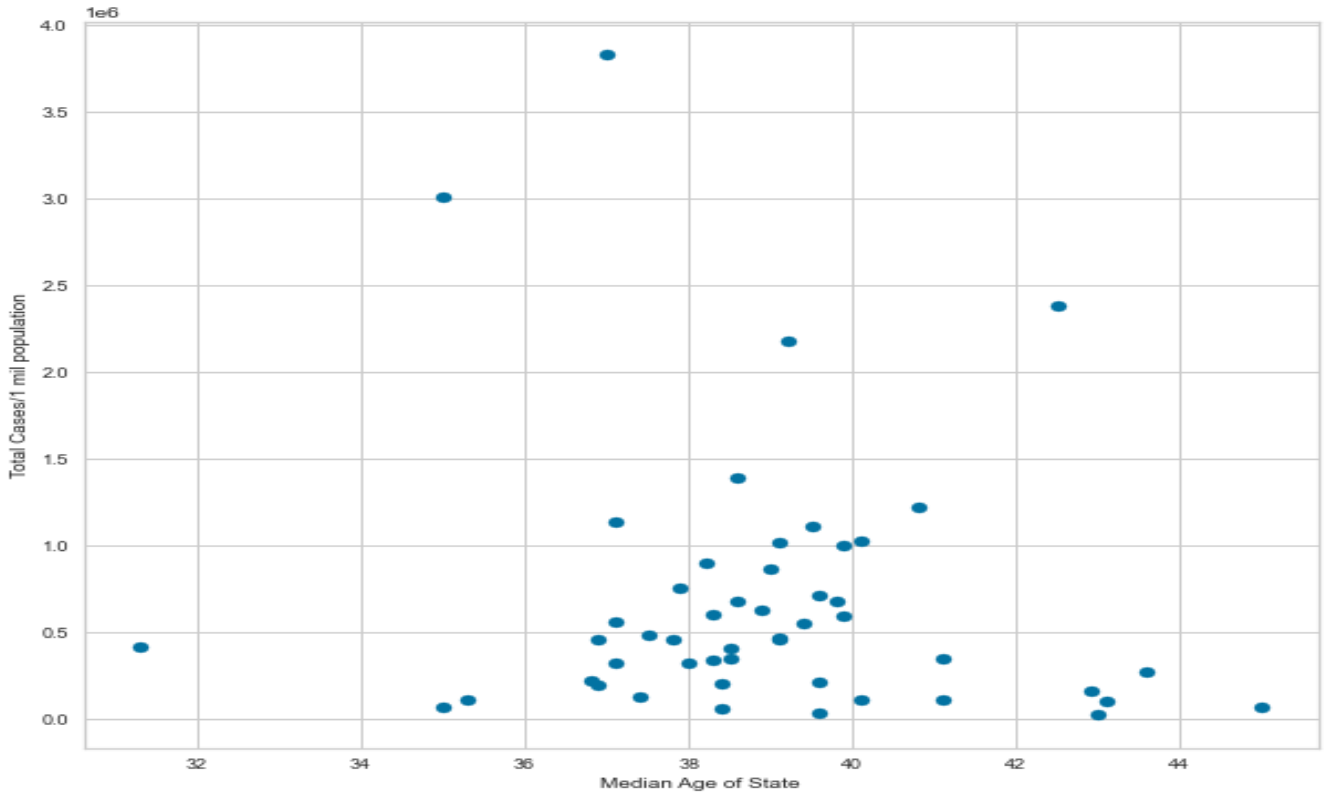


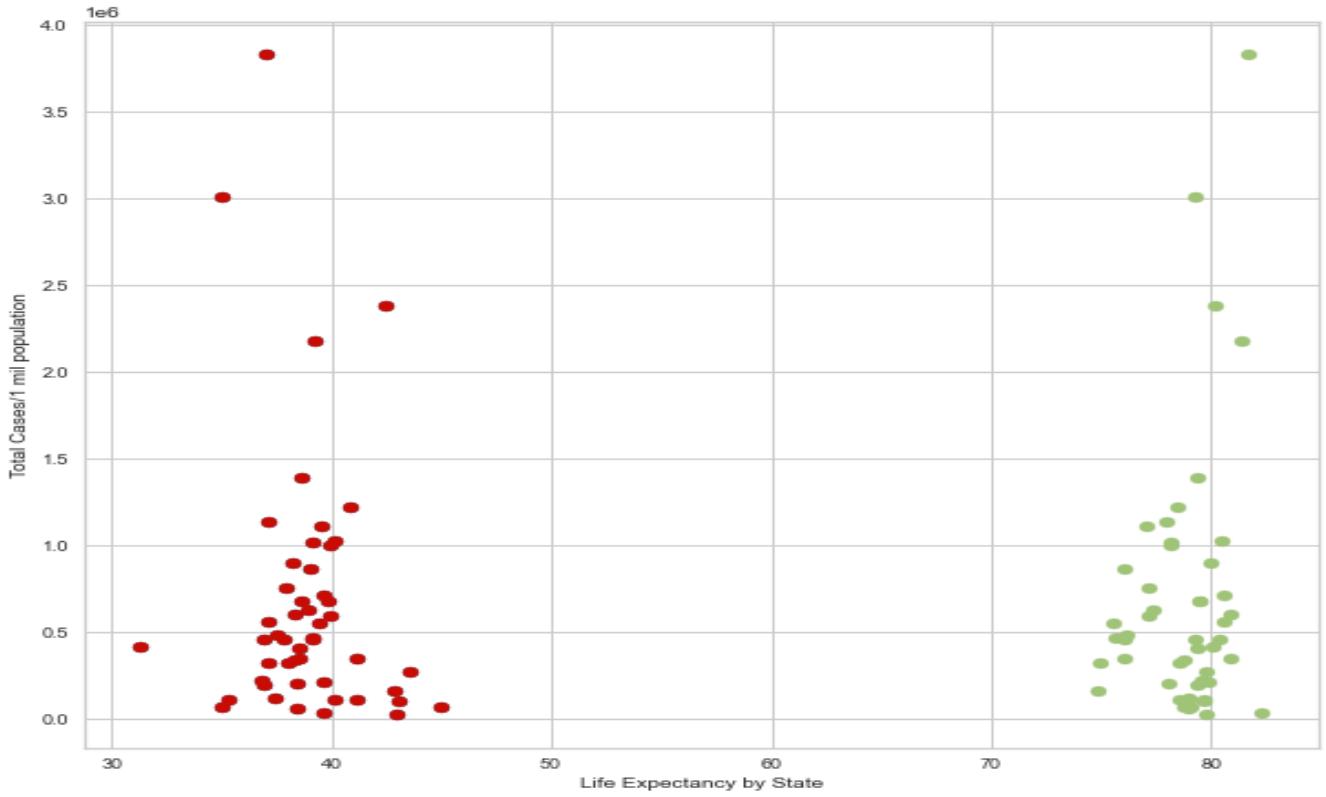


Good news here is that on a percentage basis, it looks like the percent of those infected were all near a similar, high, recovered rate with exception of those few oddballs we saw earlier that listed decent numbers of active cases instead of recovered.

From our graphs, we determined that the least influential it would seem would be the life expectancy. Usually low life expectancy is tied to harsh conditions, lower economic standing, and/or poor access to healthcare. However, it seems to be fairly on point with it's own histographic distribution. On the other hand, restriction policies did seem to have the highest impact on the amount of cases/population when compared to no restrictions, though it would seem that some restrictions versus high restrictions seem to have little difference in their impact. Thus concluded our endeavors in Milestone 1.

As we delved into Milestone 2, our focus shifted from graphics to feature reduction in anticipation of model building. With that in mind, we searched to find whether any variables truly seemed to have little to no effect on case amounts with regards to population. To that end, we capitalized on our graphical analysis from Milestone 1 to understand that some of the most randomized data seemed to be life expectancy, and median age of state. To verify these in alternative statistical and visual ways, we leveraged the Pearson Ranking and populated a graphic to see the covariance of these elements together, which yielded truthfully little to no actionable intel. However, checking our correlations of each against our total cases with respect to population, we were able to pull the following charts:


1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total Cases/1 mil population
32
34
36
38
40
42
44
Median Age of State


1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total Cases/1 mil population
30
40
50
60
70
80
Life Expectancy by State

While neither piece seems to have that great a connection to the dependent variable, life expectancy seems incredulously offset. As such, we will eliminate this piece as we move into the modeling phase. One of the final and most telling elements of our dimensional reduction endeavors was taking our remaining categorical variables and checking their Pearson correlation coefficients versus cases with respect to population, which resulted in the following outputs:

-0.5139176885705462 for state restrictions
-0.0811441424424006 for governor affiliation

As a result, we have some solid statistical backing that state restrictions are the strongest of our indicators given to lower COVID case numbers. This concluded our work in Milestone 2.

Stepping into our final milestones we examined our dataframe through the lens of a modeling scenario to get further insight on whether our correlation we found in Milestone 2 held any weight. Given our dependent variable being a prediction for a quantitative amount, we felt the most logical method of moving forward with modeling was to idealize it as a linear regression problem. As such, we compared our model with the sole affecting feature being the state restrictions, against each of the other independent variables. From our results, we determined that while it did seem adding in the second and third variables improve the model slightly, the major driving factor for case numbers with respect to population was the level of state restrictions.

Having completed each of these milestones, we can now compile our findings here. Through each of our steps in the process, we seemed to be able to recursively identify that certain features were less telling, starting with life expectancy and median state age's graphical piece being all over the board with little regard to case numbers, to correlation numbers showing insignificant relations to governor affiliation, and finally driving the point home with modeling that state restrictions, in and of themselves, were likely the best and most telling indicator of the percentage of COVID cases prevailing in the States.

Overall, recommendations on this would likely be a further understanding of this schism. Taking this modeling and analysis further we could break up the state restrictions away from a generic label and investigate if stay at home orders, mask mandates, or vaccinations each had a more telling effect on these case numbers. We could also take snippets in time for each state to see how they individually had case percentages ebb and flow over time and see if their implementation of mandate policies slowed the spread.

At the end of the day, this is a mission critical analysis to be done on a grander scale that will likely have similar case studies in larger environments examining hundreds of more variables to determine the best outcomes, and businesses and governments alike don't need to even be sold on this. COVID has held a stranglehold on the world for over a year and a half now, with no signs of stopping, so as things progress, understanding the best method of tackling and lowering COVID numbers will continue to be researched to help us in current times and in the

case that any other virus is to come even after we're through the newest Greek alphabet soup of variants. As for now, it looks like generically restricting states does seem to stem the flow of the virus, however the ethical debate of whether they should be implemented is not within the scope of our scientific engagement here.