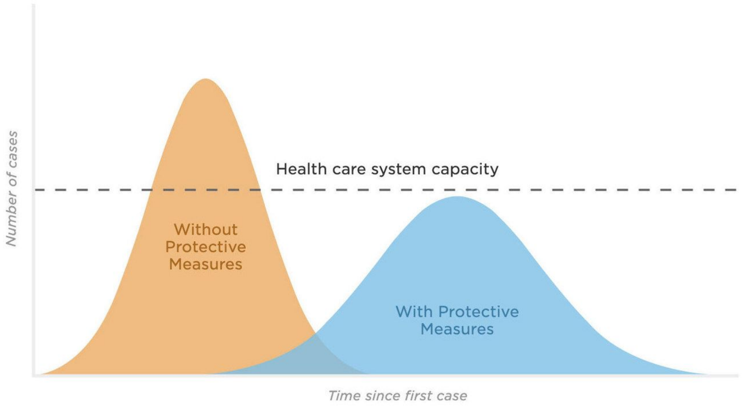
Daniel Kim
Professor Reuning-Scherer
Multivariate Statistics
1 May 2020

Multivariate Statistics Final Report

Introduction: The COVID-19 Pandemic

As SARS-CoV-2, more commonly referred to as COVID-19, proliferated to grip all corners of the country, governors scrambled to implement stay-at-home orders and quarantine directives as positive test cases skyrocketed. As of April 21, 2020, there are 824,698 confirmed cases in the United States and 45,297 deaths as a result of the virus (Centers for Disease Control and Protection 2020). Due to the rapid spread of the virus, hospitals all across the nation have seen the number of patients increase exponentially; in particular, hospitals in densely-populated urban areas such as Los Angeles and New York City are seeing their hospitals flood with sick patients, with the end of the pandemic nowhere near in sight.

As a result, many health experts and scientists pushed measures to stop the rampant spread of the disease; some of these include constant sterilization through hand soap or hand sanitizer and "social distancing," where people are encouraged to maintain six feet worth of distance between others in public. This is largely an effort to "flatten the curve" or reduce the peak number of infections so that the healthcare industry will not be overwhelmed.



Flattening the Curve Graphic, Source: NPR

Even with these efforts, however, health care systems across the country are struggling to meet the needs of all their patients; intensive care unit beds are going scarce, doctors are desperately needed, and the growth of the virus is steadily continuing. As a result, I chose to look at the abilities of each of the 50 states, Puerto Rico, and Washington D.C. to handle the load of COVID-19 in the near future.

Design and Primary Questions:

Specifically, I am interested in looking at the amount of current cases each state and territory has, along with other facts revolving around available Intensive Care Unit beds and projected growth of the virus in order to look at the measures each state would theoretically have to undergo to handle the amount of patients. Through this paper, I plan to delve into the following research topics: the extent to which markers such as available Intensive Care Unit beds and Projected Hospitalized Individuals offer redundant information, and how this can be conveyed more effectively using Principal Components Analysis; the amount of variation in ability to handle the projected growth of the coronavirus pandemic across states and territories in the United States through Cluster Analysis; and the presence of underlying factors that explain the markers of the hospital data through Factor Analysis.

Data:

My dataset, which I borrowed from globalepidemic.org, focuses on the hospital capacity per state and territory in the United State during the COVID-19 pandemic (Jha). This dataset is based on a predictive model of 306 U.S. Hospital markets across all 50 states and additional territories. This particular model was developed by an experienced group of health system researchers at Harvard Global Health Institute and the Harvard T.H. Chan School of Public Health.

Although the dataset contains more than 30 variables, I chose to focus on seven particular variables: Total Intensive Care Unit Beds, Projected Infected Individuals, Projected Hospitalized Individuals, Percentage of Total ICU Beds Needed in Six Months, Percentage of Total ICU Beds Needed in Twelve Months, and Percentage of Total ICU Beds Needed in Eighteen Months. The Total Intensive Care Unit Beds variable indicates the count of all ICU beds within an HRR that are set up and staffed. Total Available ICU beds is the number of unoccupied ICU beds on average. The Projected Infected Individuals variable states the amount of individuals over the age of 18 that are expected to get infected with COVID-19 over the course of the entire pandemic. The Projected Hospitalized Individuals is similar in that it states the amount of individuals over the age of 18 that are expected to get hospitalized due to COVID-19 over the course of the entire pandemic. Lastly, the Percentage of Total ICU Beds Needed, X months variables indicate how many ICU beds would need to be available to care for all patients requiring hospital care within X months. All of the variables are quantitative.

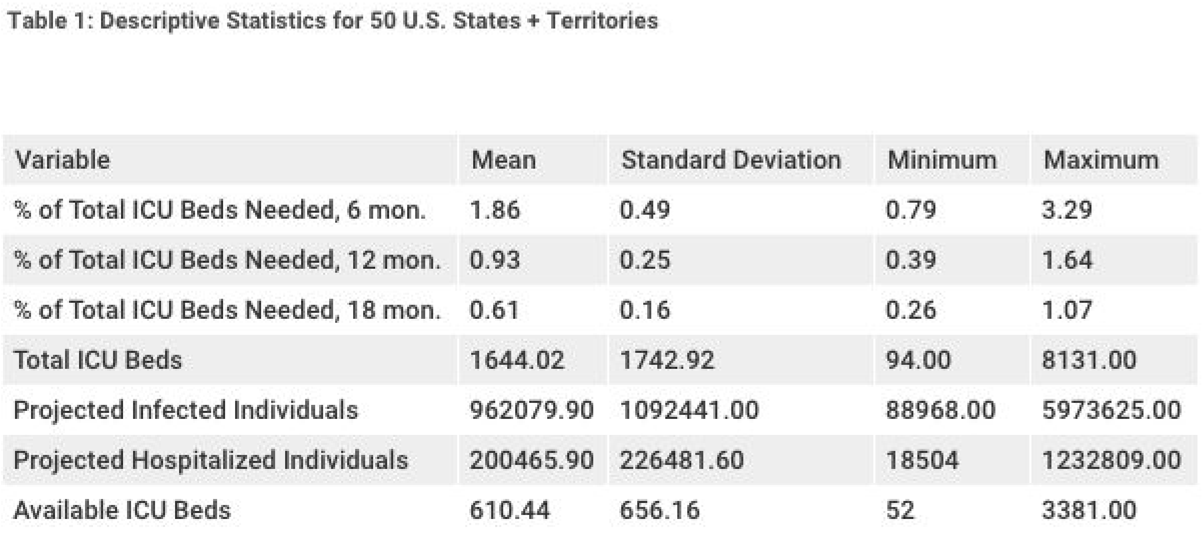
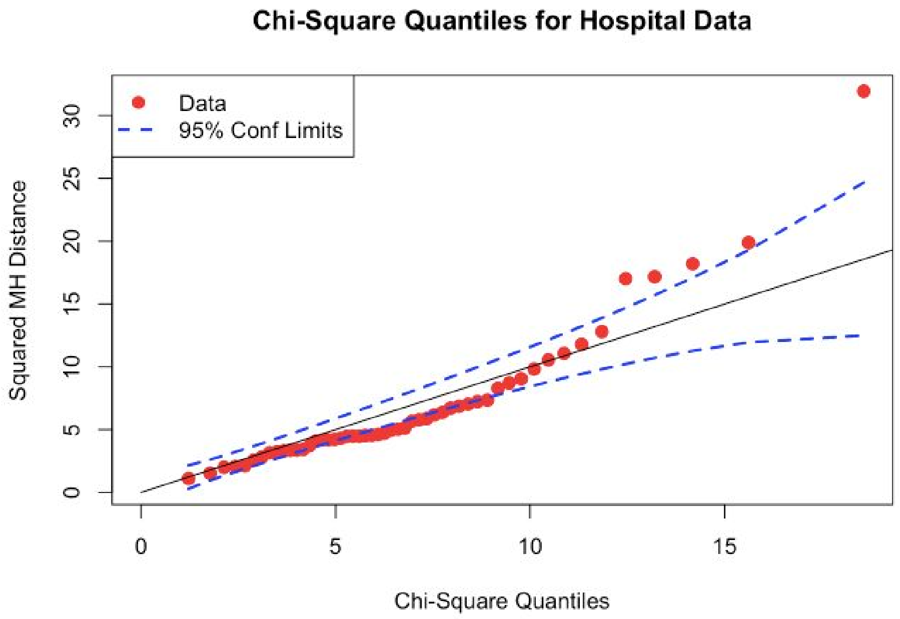


Table 1: Descriptive Statistics for 50 U.S. States + Territories

Variable	Mean	Standard Deviation	Minimum	Maximum
% of Total ICU Beds Needed, 6 mon.	1.86	0.49	0.79	3.29
% of Total ICU Beds Needed, 12 mon.	0.93	0.25	0.39	1.64
% of Total ICU Beds Needed, 18 mon.	0.61	0.16	0.26	1.07
Total ICU Beds	1644.02	1742.92	94.00	8131.00
Projected Infected Individuals	962079.90	1092441.00	88968.00	5973625.00
Projected Hospitalized Individuals	200465.90	226481.60	18504	1232809.00
Available ICU Beds	610.44	656.16	52	3381.00

I used histograms and normal quantile plots to examine all seven variables. I decided to log Total ICU Beds, Available ICU Beds, Projected Infected Individuals, and Projected Hospitalized Individuals due to the magnitude of those variables, their means, and standard deviations relative to the other variables. I examined the distributions for my variables again, and these distributions were much closer to normal after examining their respective normal quantile plots. However, univariate normal distributions do not guarantee a multivariate normal distribution, so I also used a chi-square quantile plot test (results are displayed in Figure 1 below)

Figure 1

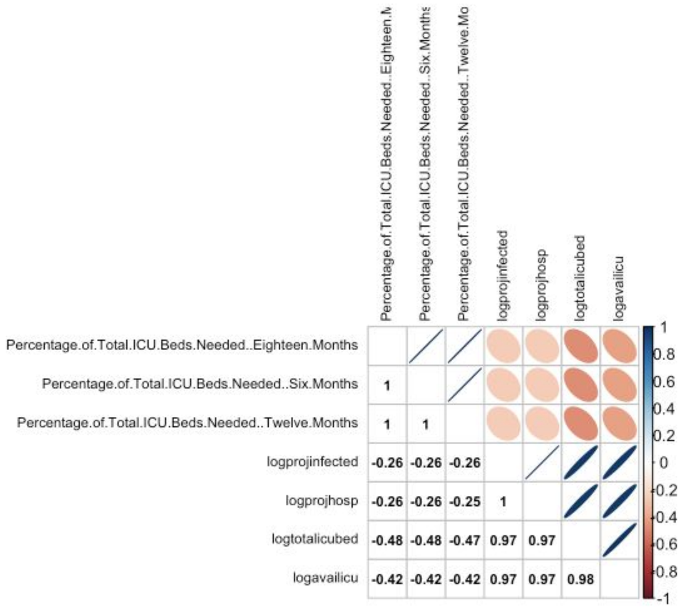


As evidenced by this Chi-Square Quantile plot, my transformed data is not perfectly normal, but it is close enough where I feel confident to move forward with my data analysis.

Descriptive Plots, Summary Statistics + Multivariate Analysis:

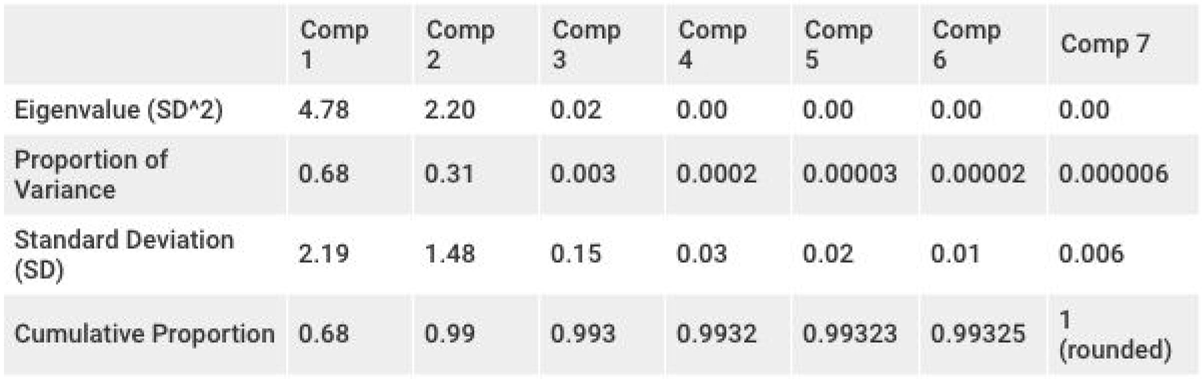
Section 1: Principal Components Analysis

Figure 2



While some of the variables did not have the strongest correlations, I still felt there was enough correlation between the seven variables to proceed with Principal Components Analysis. In particular, the percentage of Total Intensive Care Unit Beds Needed in Six, Twelve, and Eighteen months had strong correlations and the logarithms of Projected Infected and Hospitalized Individuals as well as Total ICU Beds and Available ICU Beds had strong correlations.

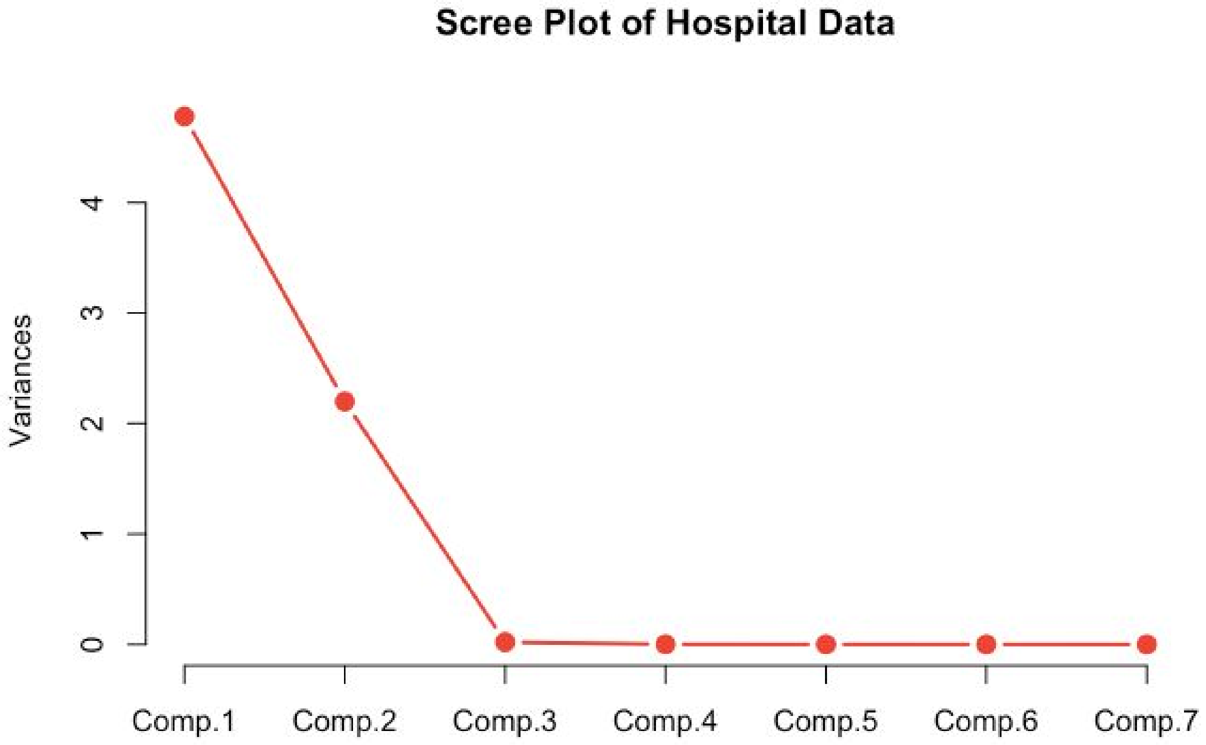
Table 2: Results from Principal Components Analysis



	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Eigenvalue (SD^2)	4.78	2.20	0.02	0.00	0.00	0.00	0.00
Proportion of Variance	0.68	0.31	0.003	0.0002	0.00003	0.00002	0.000006
Standard Deviation (SD)	2.19	1.48	0.15	0.03	0.02	0.01	0.006
Cumulative Proportion	0.68	0.99	0.993	0.9932	0.99323	0.99325	1 (rounded)

Table 2 shows that in order to explain 99% of the variance in my data, I would only need the first two principal components. Moreover, we can see that the first two eigenvalues are the only values greater than 1.

Figure 3



Now, I conduct Parallel Analysis, which is possible since my data has a multivariate normal distribution. This is to determine which components are statistically significant, using the Longman and Allen methods.

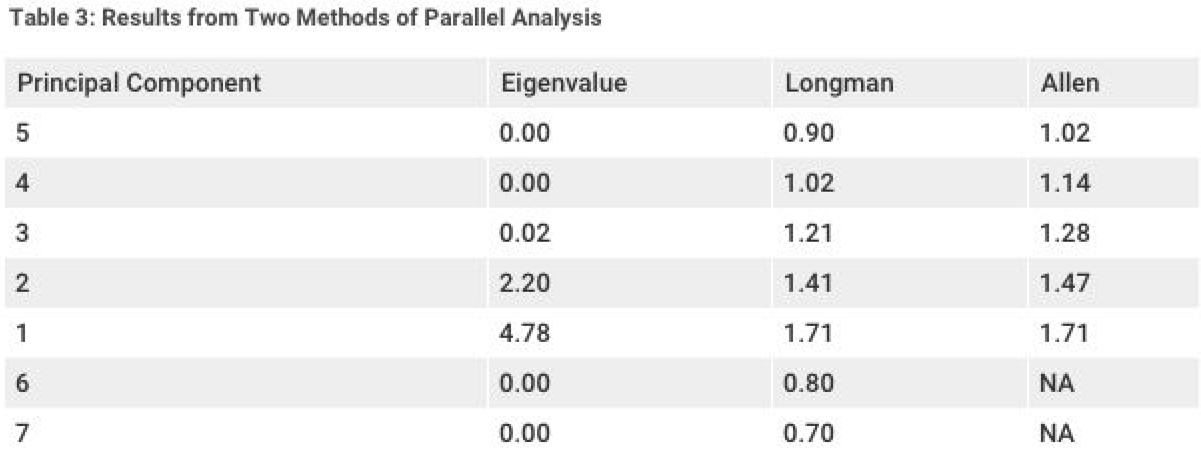


Table 3: Results from Two Methods of Parallel Analysis

Principal Component	Eigenvalue	Longman	Allen
5	0.00	0.90	1.02
4	0.00	1.02	1.14
3	0.02	1.21	1.28
2	2.20	1.41	1.47
1	4.78	1.71	1.71
6	0.00	0.80	NA
7	0.00	0.70	NA

Figure 4

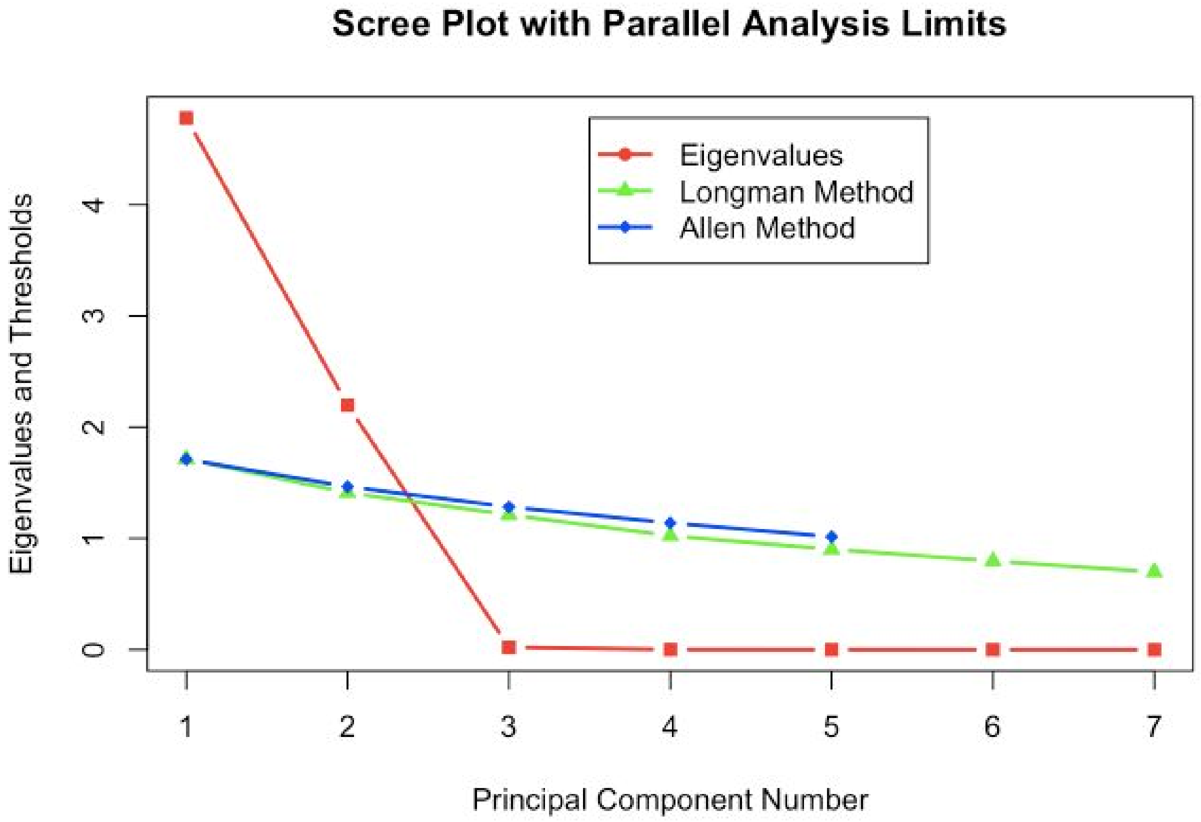


Table 3 and Figure 4 that the first two components are significant, since only their eigenvalues are greater than the thresholds. This reinforces what we already found earlier.

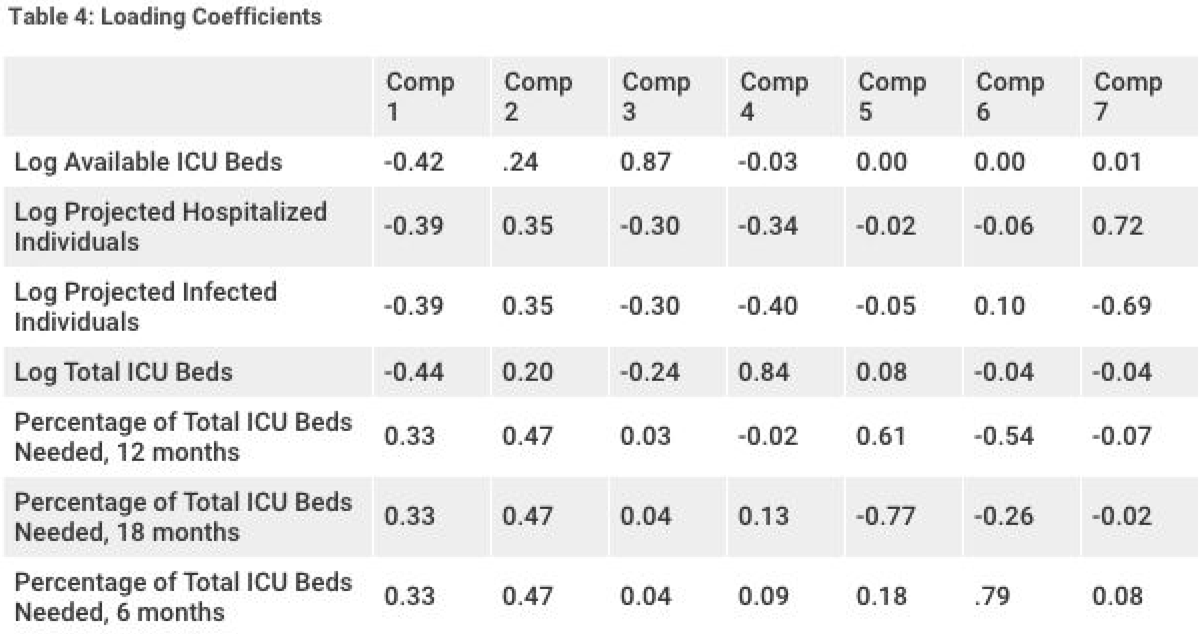
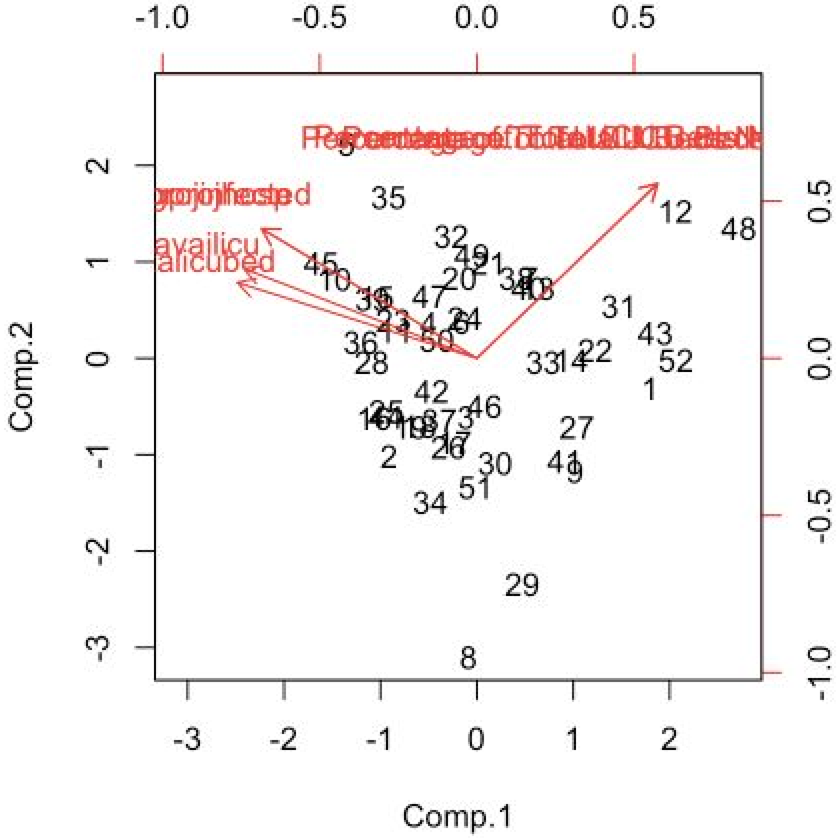


Table 4: Loading Coefficients

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Log Available ICU Beds	-0.42	.24	0.87	-0.03	0.00	0.00	0.01
Log Projected Hospitalized Individuals	-0.39	0.35	-0.30	-0.34	-0.02	-0.06	0.72
Log Projected Infected Individuals	-0.39	0.35	-0.30	-0.40	-0.05	0.10	-0.69
Log Total ICU Beds	-0.44	0.20	-0.24	0.84	0.08	-0.04	-0.04
Percentage of Total ICU Beds Needed, 12 months	0.33	0.47	0.03	-0.02	0.61	-0.54	-0.07
Percentage of Total ICU Beds Needed, 18 months	0.33	0.47	0.04	0.13	-0.77	-0.26	-0.02
Percentage of Total ICU Beds Needed, 6 months	0.33	0.47	0.04	0.09	0.18	.79	0.08

For the first two components, which are of primary focus, Comp 1 has positive correlations with the Percentage of Total ICU Beds needed in 6, 12, and 18 months but has negative correlations with the log variables of Available ICU Beds, Projected Hospitalized Individuals, Projected Infected Individuals, and Total ICU beds. The values with largest magnitude, which are Available ICU Beds and Total ICU beds, lead to the belief that Component one is picking up on the variation in incapability to take in sick patients for hospitals in terms of beds. Component two, which has positive correlations with all the variables, is picking up on the variation in the projected capacity of the hospitals based on health care system models.

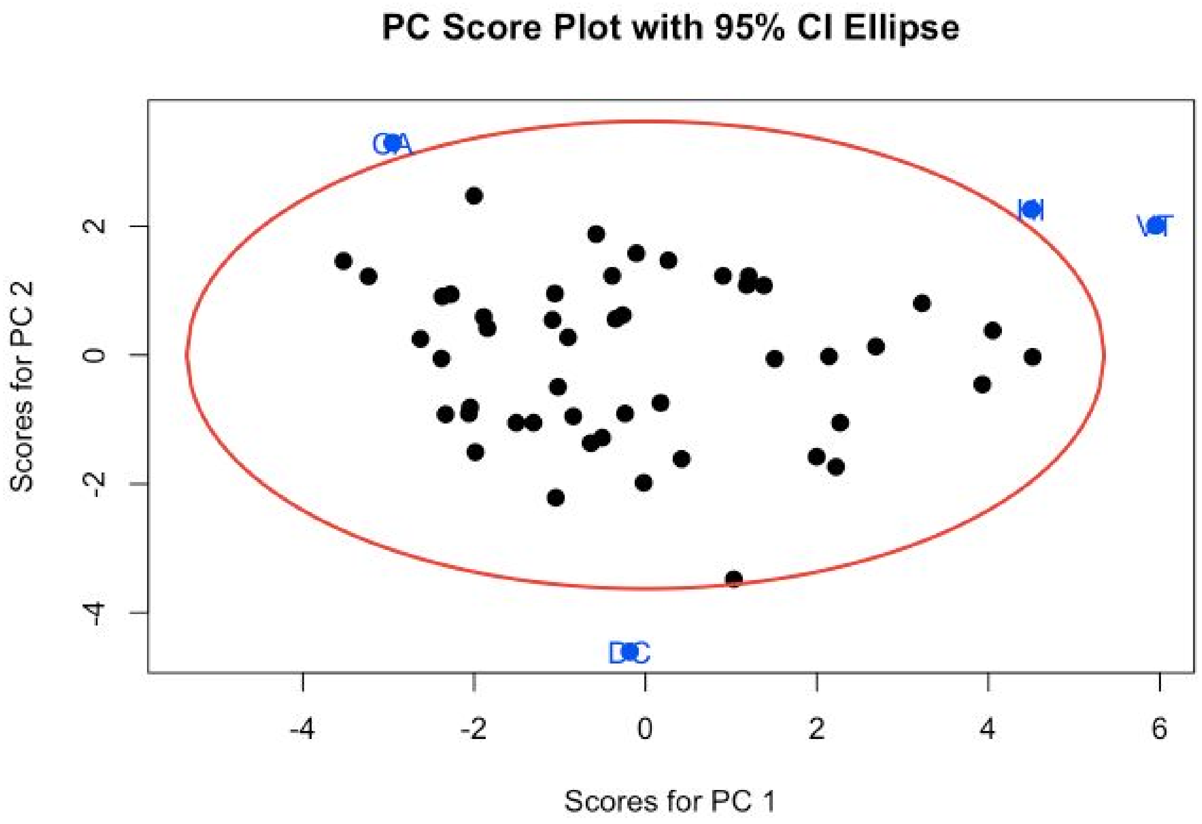
Figure 5



Looking at figure 5, which is the biplot of the first two principal components, the bottom and left-side axes are the principal component scores for the first and second components, respectively. The right-side and top axes are the loading coefficients for the principal components. There seems to be more spread in the direction of Comp2 than Comp1. The loose clustering indicates that the 50 states and additional territories tend to be different across the 7 variables. This indicates that although a state might have a lot of projected infected individuals, for example, there is variation on how many ICU beds they might need in the coming months.

Finally, I created a confidence ellipse plot, to test for outliers in Comp1 and Comp2.

Figure 6



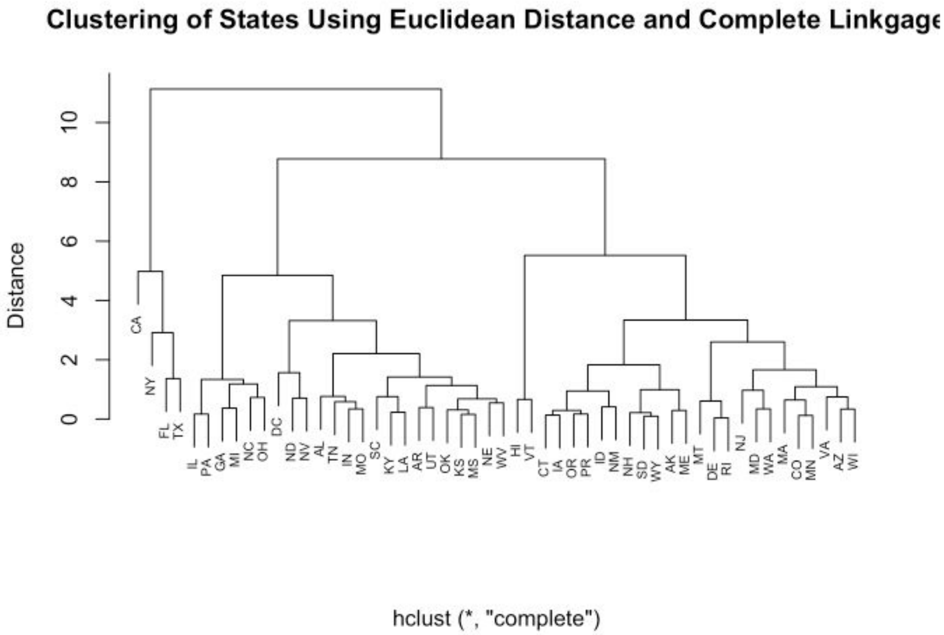
Looking at the outliers, most of the outliers come from the direction of the second principal component with the exception of Vermont. California, Washington D.C., and Hawaii all seem to be projected to be unable to handle the coronavirus load coming in the coming months, which is because this is what Component 2 was primarily loaded on. Vermont on the other hand might have an extreme shortage of ICU beds that are available compared to other states. Considering that Vermont is much less populated than the other outliers, I hypothesize that this might be due to the large elderly population, but this point could use more exploration.

Section 2: Cluster Analysis

For cluster analysis, I used the same 50 states plus additional territories and seven variables; however, instead of transforming my variables using logarithms, I instead standardized them. Standardizing the variables ensures that the differing ranges of the original variables won't be an issue when attempting to perform cluster analysis so that certain variables don't hold more weight than others.

I generated two different dendrograms grouping the 50 U.S. states along with the additional territories. This is with the goal of seeing how many groups there are in the United States in regards to preparedness of combatting the COVID-19 pandemic. This would give me a better idea to look for similarities between states within groups to see why certain states are more vulnerable than others which can be useful for future planning against pandemics.

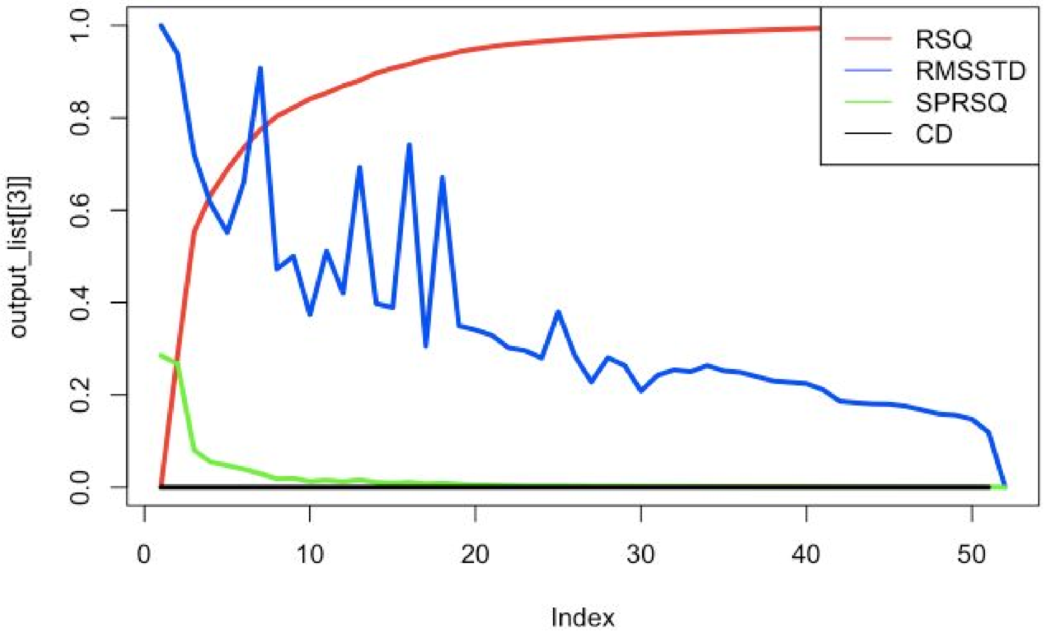
Figure 7:



The Euclidean Distance Metric gives equal weight to each of the seven different variables when used to determine the distance between states. The Complete Agglomeration method defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the clustering process, the two nearest clusters are merged into a new cluster. Looking at the produced dendrogram, there appears to be approximately three groups, with only one of those groups in a different major branch; I theorize that these states are experiencing a massive load of coronavirus cases and are struggling to take in all sick patients, in comparison to other states.

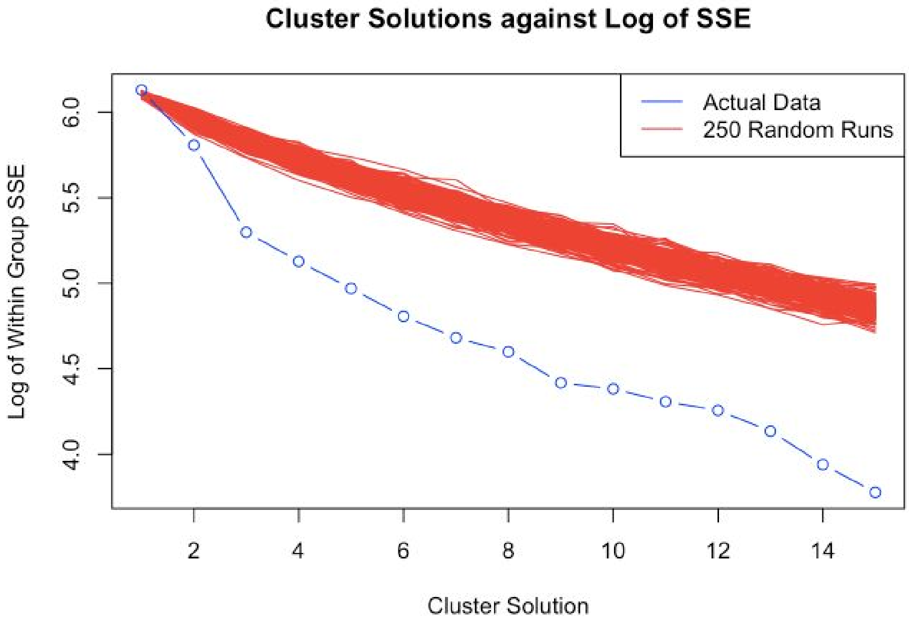
In order to more precisely determine the number of clusters, I looked at other metrics such as Root-Mean-Square Standard Deviation, R-Squared, Semi-partial R-Squared, and Cluster Distance. The Root-Mean-Square Standard Deviation is the average standard deviation for however many variables there are. R-squared is the total sum of squares between clusters divided by the total sum of squares total. Semi-partial R-Squared measures the relative change in within-clusters sum of squares and Cluster Distance measures the determined similarity between elements.

Figure 8:



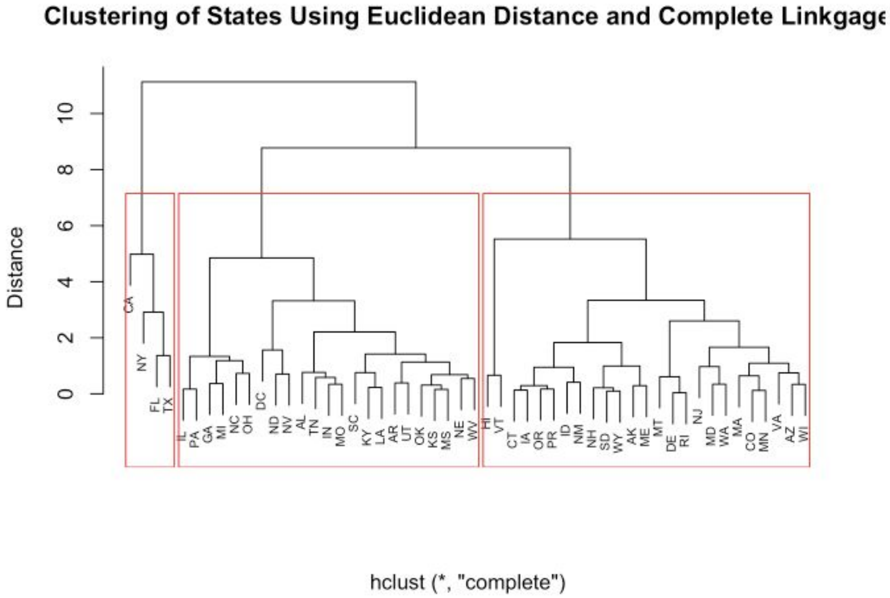
It looks like there are around 3 cluster groups. While the RMSSTD and CD lines offer little information, the points where the RSQ and SPRSQ curves start to level out is around three cluster groups.

Figure 9:



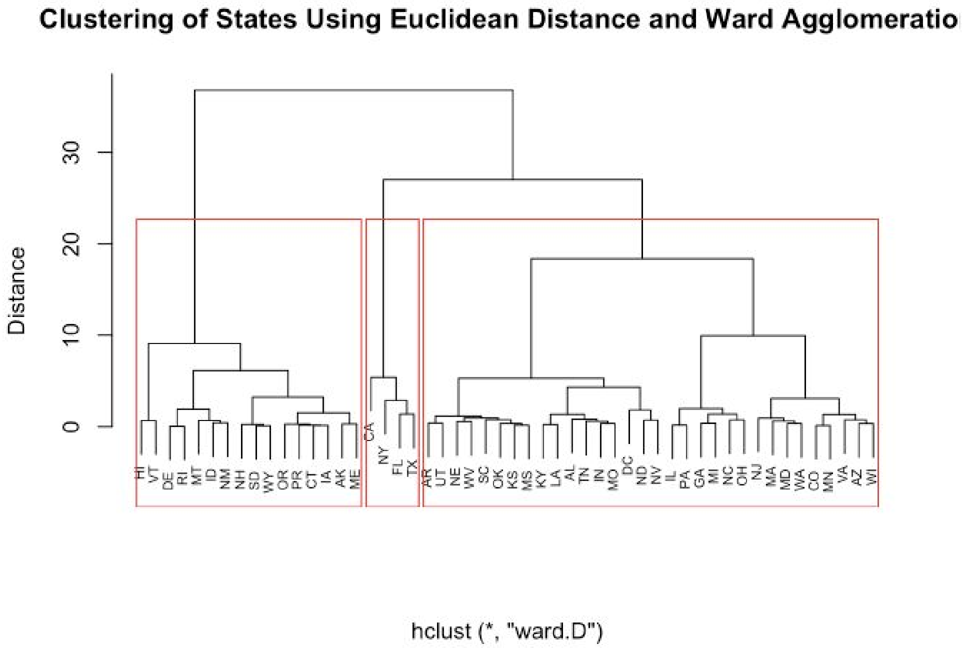
Moreover, looking at the Log of SSE plotted against the cluster groups for the actual data against 250 random runs, the point where the distance between the two stops changing is around 3 groups. Thus, we can produce dendrograms that cluster at around 3 groups.

Figure 10:



Once again looking at the dendrogram produced using Euclidean Distance and Complete Agglomeration, we can see that when we form three clusters, one in particular consists of California, New York, Texas, and Florida, which are all highly populated states struck hard by the coronavirus.

Figure 11:



Looking at this dendrogram using Euclidean Distance and Ward's Agglomerative Method, which minimizes distance between clusters, we can see that once again CA, NY, FL, and TX are in a cluster but the groups of other states are different. This cluster analysis shows us that these highly populated states with large urban areas are prone to pandemics, which could help for future planning.

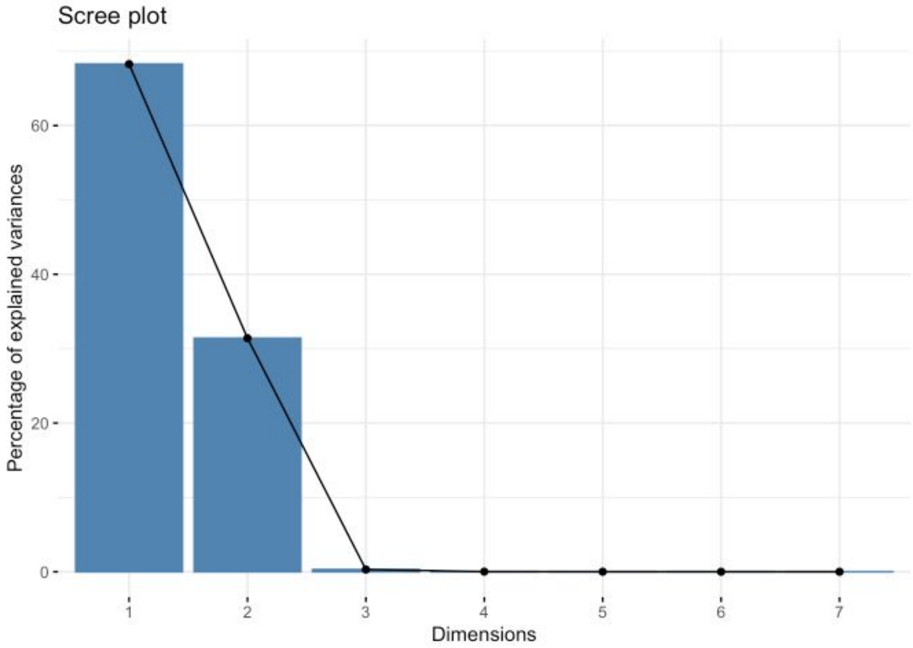
Section 3: Factor Analysis

When performing factor analysis, it is important to first identify whether there might exist relationships from the indicators. There appears to be relationships between the total amount of ICU Beds needed in 6, 12, and 18 months, which is self-explanatory. There also seems to be relationships between the Logarithms of Projected Infected Individuals, Projected Hospitalized Individuals, Total ICU Beds, and Available ICU Beds, which makes sense since states with higher populations with more projected ill citizens will tend to have more hospitals and ICU beds. This information is supplemented by the correlation matrix in Figure 2, generated during Principal Components Analysis.

Moreover, once calculating the KMO (Kaiser-Meyer-Olkin measure of adequacy) which measures how fit a dataset is for factor analysis, I achieved a value of 0.82, which is above the Meritorious cutoff created by Kaiser and Rice in 1974. This indicates that our dataset is indeed able to be manipulated for Factor Analysis.

After performing principal components on the dataset, I generated a scree plot to decide on the number of latent factors.

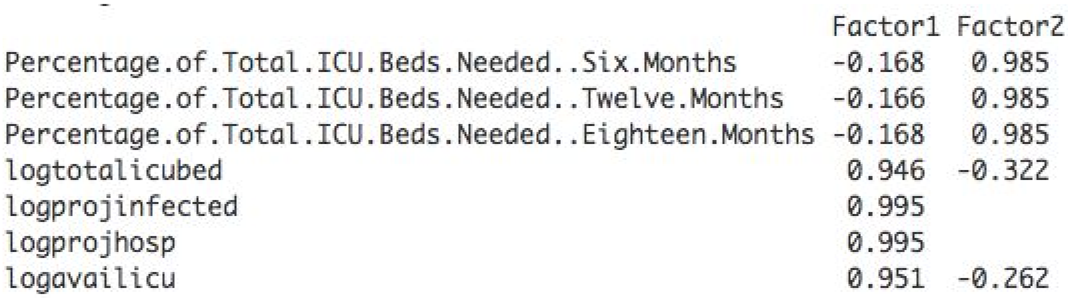
Figure 12:



The elbow appears at 3 dimensions indicating that there are probably two latent factors in our dataset.

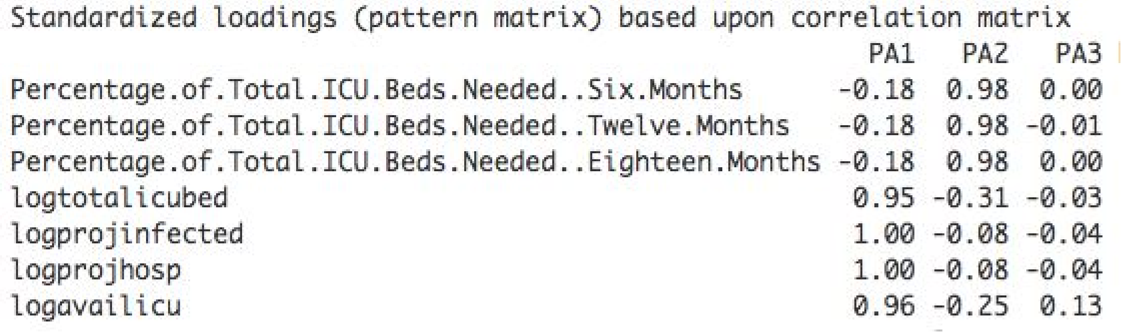
Next, I performed factor analysis using both Maximum Likelihood and Iterative PCA. Looking at the residual correlation matrices for both methods, the first two factors both had many indicators that were well determined.

Table 5: Loadings for Maximum Likelihood FA



```
                                          Factor1 Factor2
Percentage.of.Total.ICU.Beds.Needed..Six.Months      -0.168   0.985
Percentage.of.Total.ICU.Beds.Needed..Twelve.Months   -0.166   0.985
Percentage.of.Total.ICU.Beds.Needed..Eighteen.Months -0.168   0.985
logtotalicubed                                        0.946  -0.322
logprojinfected                                       0.995
logprojhosp                                           0.995
logavailicu                                           0.951  -0.262
```

Table 6: Loadings for Iterative PCA FA

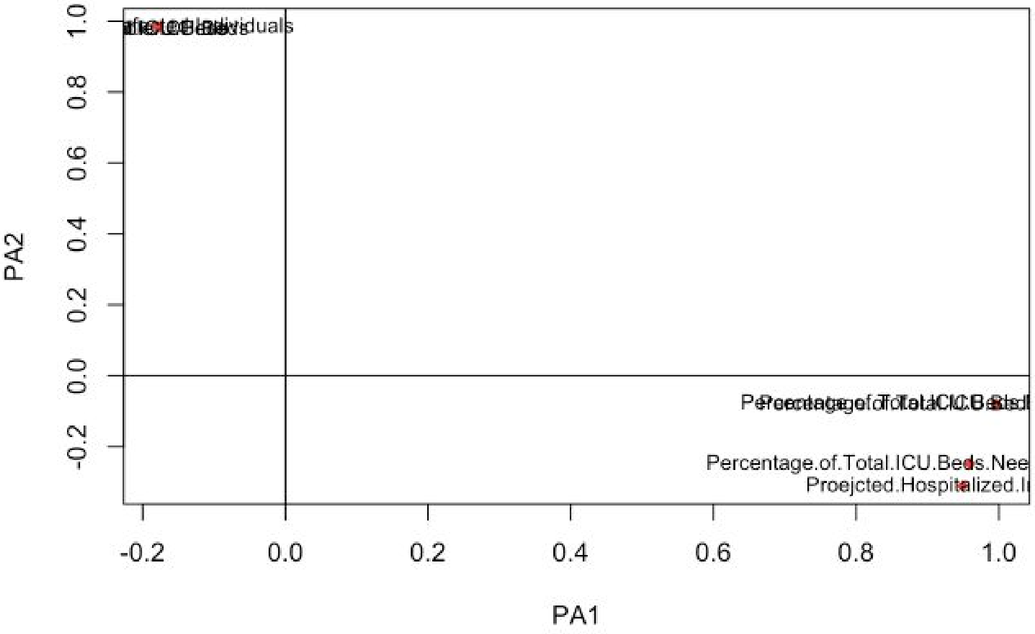


```
Standardized loadings (pattern matrix) based upon correlation matrix
                                                       PA1   PA2   PA3
Percentage.of.Total.ICU.Beds.Needed..Six.Months      -0.18  0.98  0.00
Percentage.of.Total.ICU.Beds.Needed..Twelve.Months   -0.18  0.98 -0.01
Percentage.of.Total.ICU.Beds.Needed..Eighteen.Months -0.18  0.98  0.00
logtotalicubed                                        0.95 -0.31 -0.03
logprojinfected                                       1.00 -0.08 -0.04
logprojhosp                                           1.00 -0.08 -0.04
logavailicu                                           0.96 -0.25  0.13
```

However, the root mean square residual is lower for the factor analysis using Iterative PCA, (6.657618e-05 vs. 0.000848924) so I proceeded using that method.

Now, I look at the loading plot of Factor Analysis using Iterative PCA with Varimax rotation, which is used to simplify the expression of a particular sub-space in terms of just a few major items each.

Figure 13:



Taking a look at the figure, Factor 1 seems to take in account states that are projected to have many infections and hospitalizations due to COVID-19 while Factor 2 seems to look at hospital capacity based on predicted models of the pandemic.

Conclusions and Discussion:

Based on our analysis conducted above, we examined the impact of the COVID-19 pandemic on the 50 U.S. states and the additional territories.

Using Principal Components Analysis, we were able to reduce the dimensionality of the dataset in the first and second most directions of maximum variability. While doing this, we were able to see which states were outliers compared to others; one notable observation was that Vermont's healthcare system is quickly projected to become overwhelmed, which might be due to the large eldery and retired population residing there. In cluster analysis, we were able to group states based on the projected severity of COVID-19 and their healthcare systems' abilities to combat the virus. Through cluster analysis, we observed that there were around three different situations that a state could be in. Large and densely-populated states such as NY and CA are struggling to contain the outbreak and hospitalize patients. Next, there are middle-of-the-pack states that are seeing cases of the virus, but not to the extent of larger states. Lastly, there are sparsely populated states such as Idaho that are not seeing rapid proliferation of the virus.

Lastly, using Factor analysis, we were able to see that the two prominent factors explaining our variables were the projected infections and hospitalizations due to COVID-19 and also hospital capacity based on predicted models of the pandemic.

Points for Further Analysis:

For further analysis, it might be helpful to look at the other variables of the COVID-19 state hospitals dataset and perform extensive analysis based on the addition of those variables. For instance, there is data on the number of adult citizens and citizens above the age of 65. Since studies show that elderly citizens are at a serious risk due to the fatality of COVID-19 (Arbaje). Moreover, it might be intuitive to compare the data of hospitals in the United States to that of other countries and try to conduct statistical techniques such as cluster analysis to see which countries the United States is similar to to see how the trajectory of the virus might take place in the United States. For example, based on if the United States projected COVID-19 growth and hospital capacity is more similar to a country like Italy that experienced a rapid proliferation of the coronavirus as opposed to South Korea which quickly controlled the spread, health professionals could take more extensive measures to prevent the same exponential growth of the virus.

Works Cited

Works Cited

Arbaje, Alicia. "Coronavirus and COVID-19: Caregiving for the Elderly." *Coronavirus and COVID-19: Caregiving for the Elderly | Johns Hopkins Medicine*, Johns Hopkins Medicine, www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-caregiving-for-the-elderly.

"Cases in U.S." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 29 Apr. 2020, www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

Jha, Ashish K. "US Hospital Capacity." *Pandemics Explained*, Harvard Global Health Institute, 17 Mar. 2020, globalepidemics.org/our-data/hospital-capacity/.