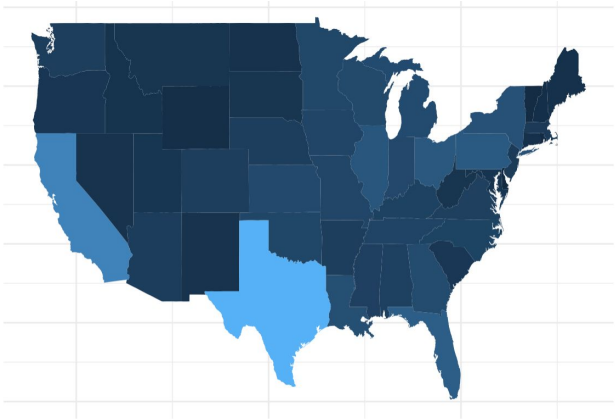
Patterns in COVID-19 data in the United States

Ishita Gupta, Praveena P, Rohith Krishna
Vijiyashree SB, Shravanth J



Introduction

We observe patterns in the United States Country-wise COVID-19 dataset. Its observations include the 50 states and the capitol hill - DC.

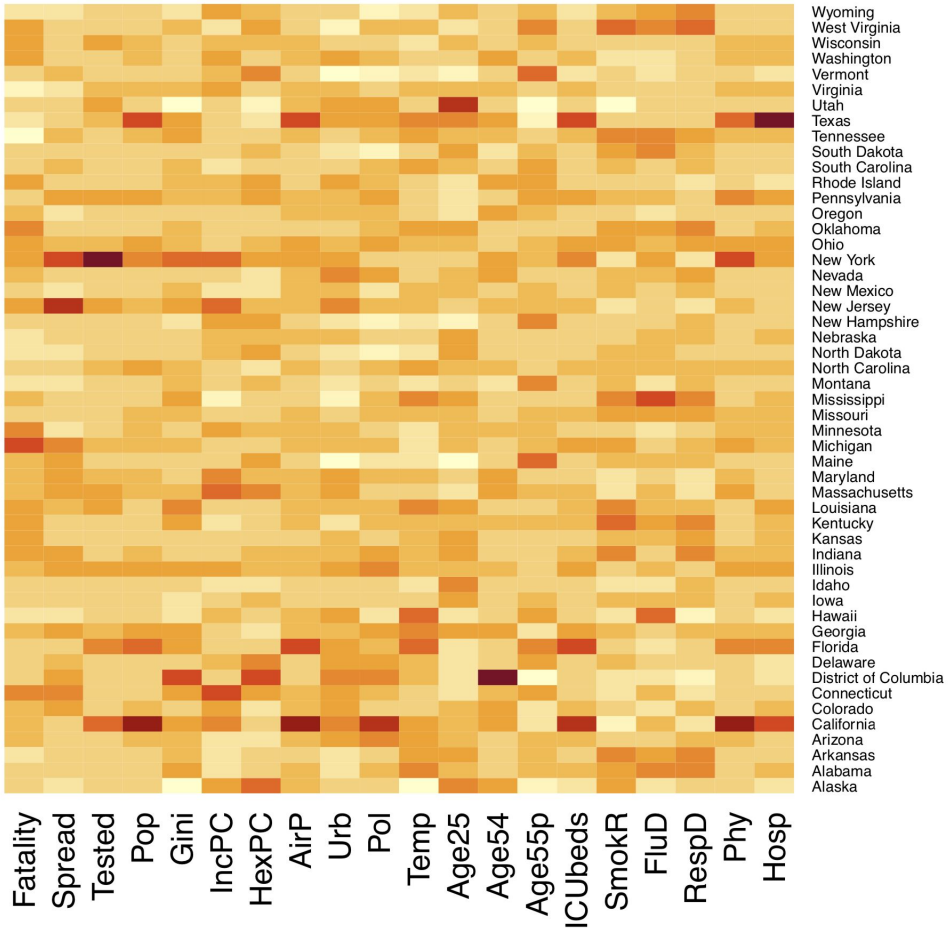
The number of features in the dataset are 20, which is exceedingly high dimensional for the given number of observations.

Thus it is pertinent to use methods of unsupervised learning such as dimensionality reduction and clustering. We make use of three major methods here. They are:

- Principal Component Analysis (PCA)
- Partitional methods such as K-means and PAM approach
- Hierarchical clustering method which takes a bottom-up agglomerative approach to clustering.

Fatality - deaths as proportion of number of persons infected
Spread - number of persons infected as a proportion of number of people tested
Tested - number of people tested
Pop - population estimates for the state
Gini - gini coefficient for income inequality
IncPC - per capita income as per 2018
HexPC - health expenditure per capita
AirP - number of medium and large airports in each state
Urb - urbanisation as a percentage of population
Pol - average exposure of the general public to particulate matter of 2.5 microns or less (PM2.5) measured in micrograms per cubic meter (3-year estimate)
Temp - average temperature of the state (2019)
Age25 - proportion of population aged between 0-25 years
Age54 - proportion of population aged between 25-54 years
Age55p - proportion of population aged over 55 years
ICUbeds - number of ICU beds in the state
SmokR - percentage of smokers in the population
FluD - influenza and pneumonia death rate per 100,000 people
RespD - Chronic lower respiratory disease rate per 100,000 people
Phy - Number of primary and specialty care active physicians
Hosp - Number of hospitals

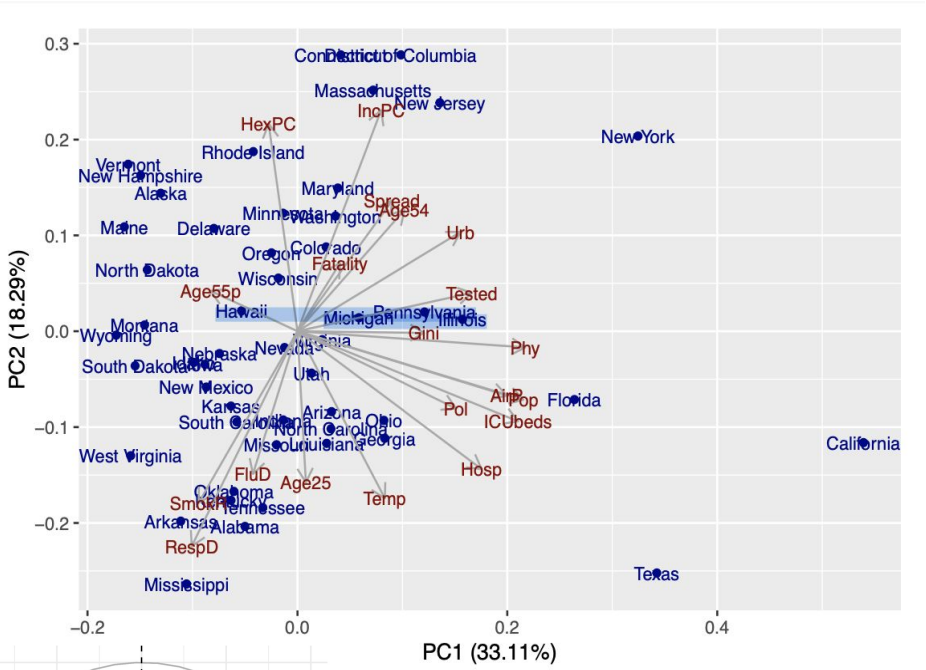
Heatmap of the dataset



- Michigan has the highest fatality, Tennessee the least.
- NY and NJ has the highest spread of COVID-19.
- NY and California are areas with high testing rates.
- CA, TX, FL → largest populations; followed by NY.
- High income disparity in NY and DC.
- HexPC is the highest in DC, Utah the least. Average health spending is ~$8300/person/year.
- Largest airports in CA, TX, FL.
- Maine and Vermont are largely rural and have the largest populations of people aged 55 and above.
- CA, DC, Illinois etc. highly polluted.
- Working age people disproportionately high in DC.
- CA, TX, FL have highest number of ICU beds and hospitals. NY closely follows.
- Largest numbers of smokers in West Virginia, Mississippi, Kentucky and Arkansas; The least in CA, Utah.
- CA, TX, FL and NY have the highest number of physicians.

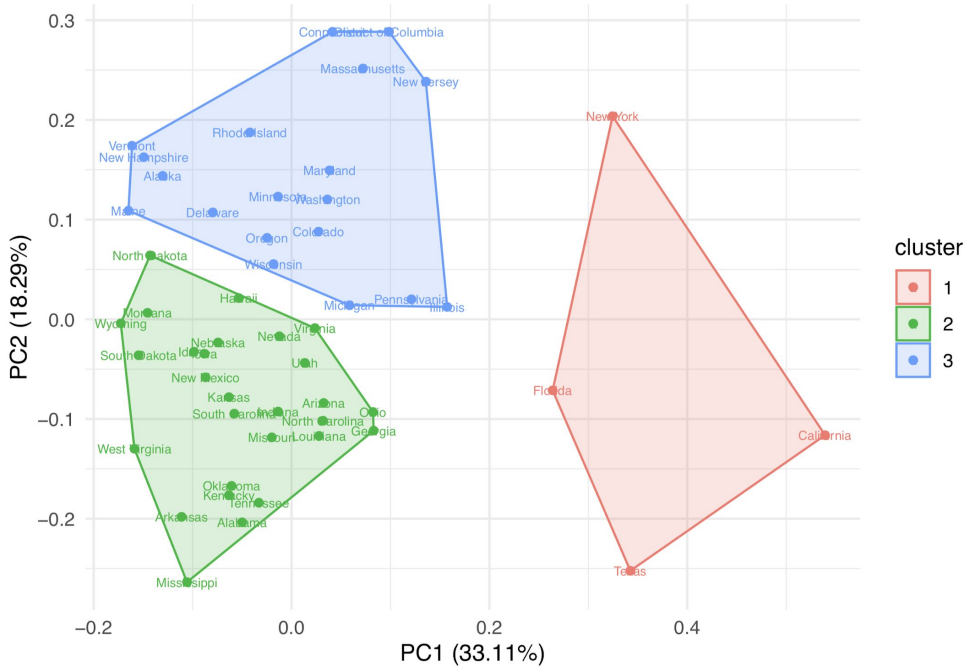
Principal Component Analysis

- The states that are close to each other in the PC Plane have similar values on all parameters
- States along certain feature vectors have high values of that particular feature. Eg. SmokR - Arkansas, Mississippi. Kentucky. This is confirmed by the numbers in our data.
- The first two principal components explain only 51% of the variation in data.
- This indicates that the number of features is exceedingly high dimensional for the given number of observations
- The variables explaining most of the PCs are ICU Beds, Population and Physicians.
- The states of Georgia and Ohio have very similar population sizes, number of physicians and ICU beds,
- In the PC plane, we see that these two states are close to each other.

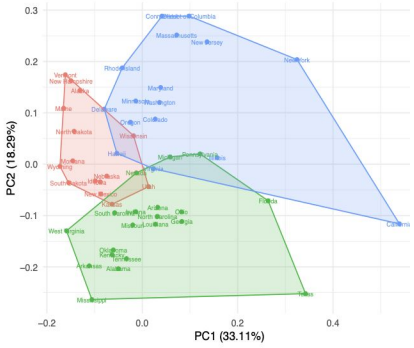


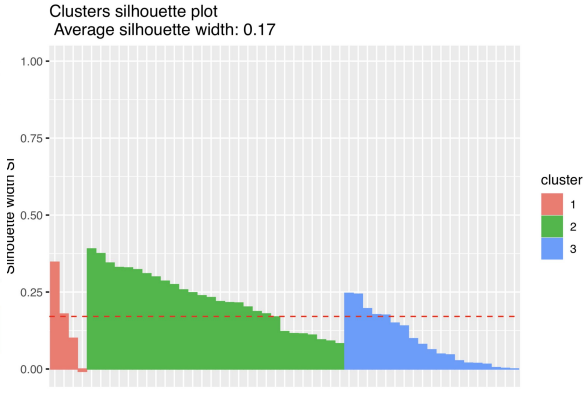
Partitioned Clustering K-means, PAM.

- The optimal value of k was found to be 3. This gave a Silhouette coefficient of 0.17.
- SC = 0.17 indicates no substantial structure in clustering.
- Try an alternative algorithm such as Partitioning Around Medoids (PAM)
- SC = 0.12 for PAM. → no structure.
- PAM results in overlapping clusters, negative SC for overlapped observations.
- Indicates a need to narrow down the feature space.



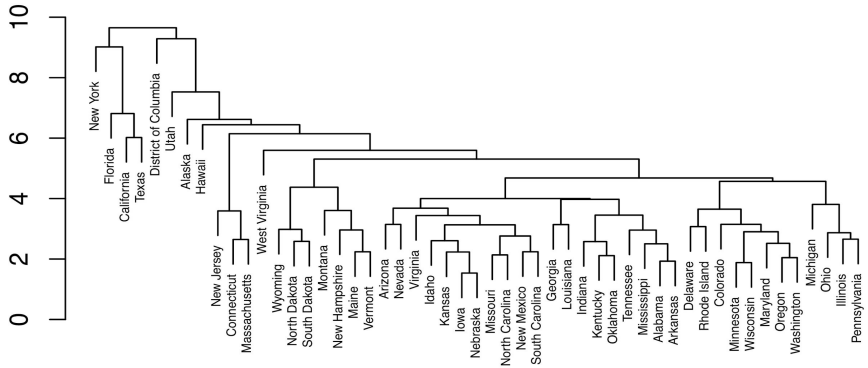
- K-means method sometimes allocates close points to different clusters in trying to minimize the within-cluster-variance
- Eg.North Dakota, Maine and Mississippi

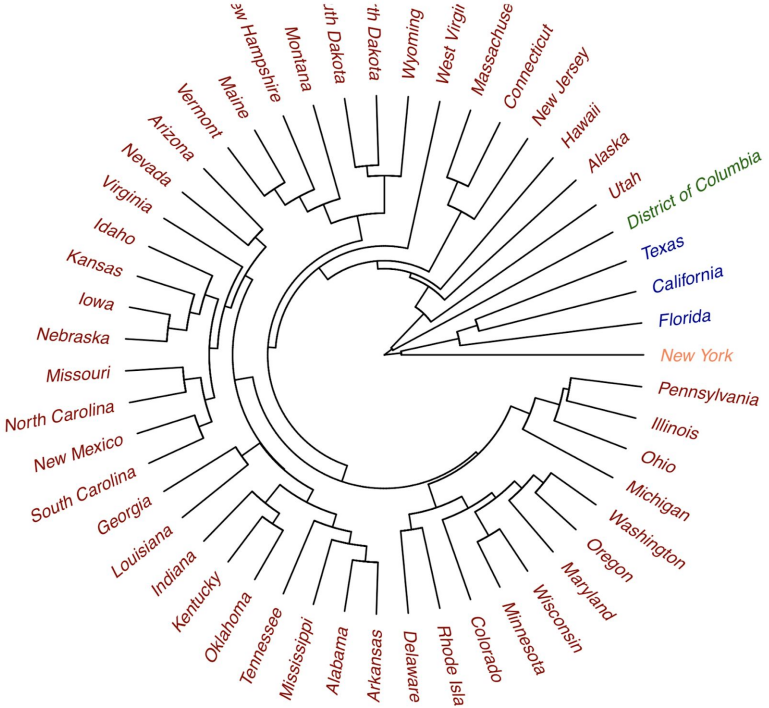




Hierarchical Clustering

- We use average linkage algorithm for hierarchical clustering.
- We get four major clusters with New York being a separate cluster.
- Clearly, the outliers in the PCA plot are forming separate clusters in Hierarchical Clustering.
- California, Florida and Texas are closer whereas New York and District of Columbia emerge as a complete outlier.





- The first cut off point for the cluster is at a height 9 which segregates New York, Florida, Texas and California and District of Columbia, Utah, Hawaii and others in a separate cluster. Next cutoff point is at 7. This goes on till it reaches the minimum cutoff of height 2.

Segmentation of data: Pandemic, Health, Economic

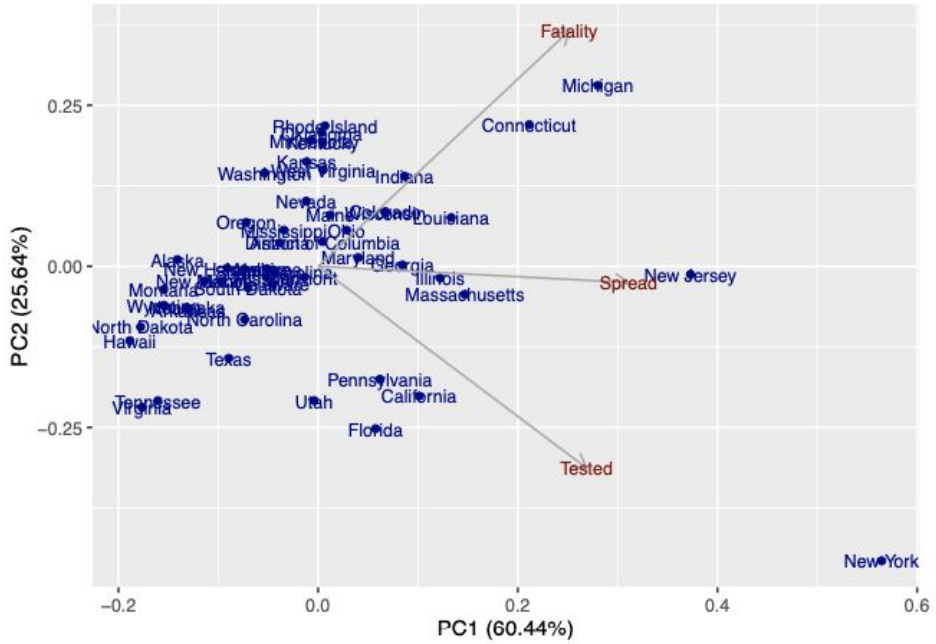
- For better feature analysis, we take up
  - Analysis on subsets of the data.
  - Variable selection.
- Under the segmentation, we choose:
  - Pandemic: Fatality, Spread, Tested
  - Health: ICUbeds, SmokR, FluD, RespD, Phy, Hosp.
  - Economic: Pop, Gini, IncPC, HexPC, AirP, Urb
- We then perform PCA, partitioning and hierarchical clustering on these data subsets, and interpret results.
- Further, we also perform variable selection using AIC selection method and perform clustering on the refined model.

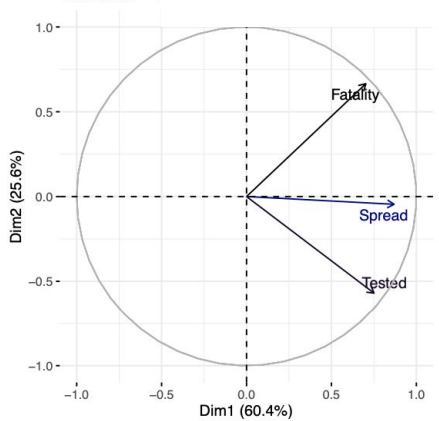
```r
#pandemic - 3 variables
dfp <- dselect(df,c(Fatality,Spread,Tested,))
#health  - 6 variables
dfh <- dselect(df,c(ICUbeds,SmokR,FluD,RespD,Phy,Hosp))
#economic - 6 variables
dfe <- dselect(df,c(Pop,Gini,IncPC,HexPC,AirP,Urb))
#climatic - 2 variables
dfc <- dselect(df,c(Pol,Temp))
#demographic - 3 variables
dfd <- dselect(df,c(Age25,Age54,Age55p))
# all predictors, with Fatality as response
dff <- dselect(df,-c(Spread,Tested))
# all predictors, with Spread as response
dfs <- dselect(df,-c(Fatality,Tested))
# all predictors, with Tested as response
dft <- dselect(df,-c(Fatality,Spread))
```

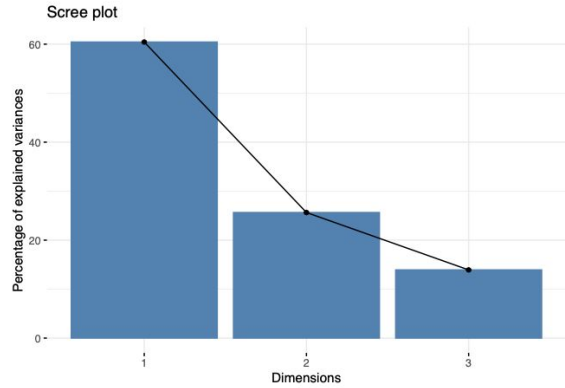
Pandemic - PCA

- The first two principal components explain about 86% of the variation in data.
- Each of the 3 variables contributes about 30-36% to the principal component.
- The states that are close to each other in the PC Plane have similar values in terms of spread, fatality and tested.
- As per the data,the states of Michigan and Connecticut have very similar values in all 3 variables. (shown below)
- In the PC plane, we see that these two states are close to each other.
- Michigan also has the highest fatality which is prevalent in our PC plot as well.

	Michigan	Connecticut
Fatality	0.075	0.06
Spread	0.285	0.301
Tested	107791	58213

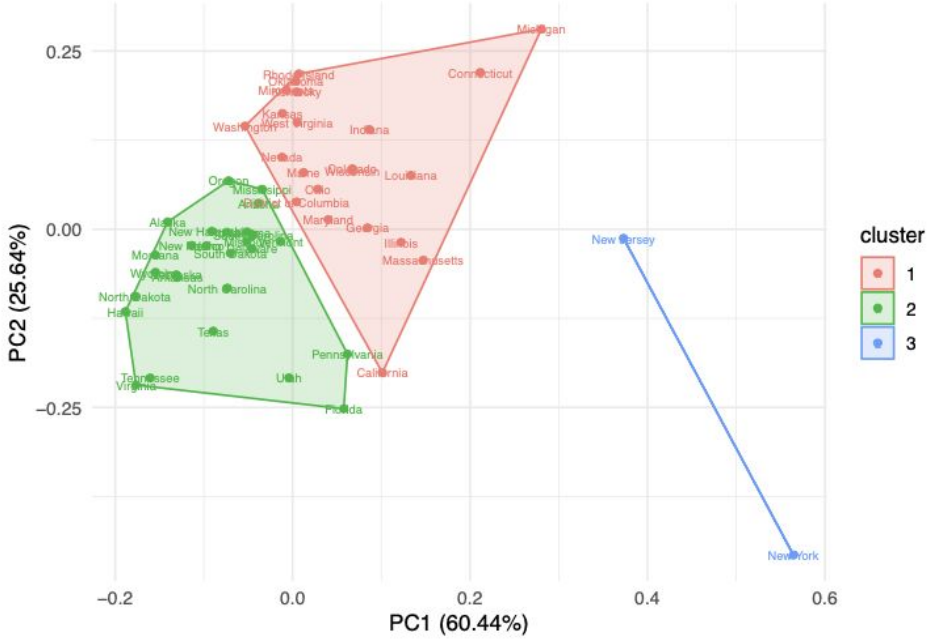


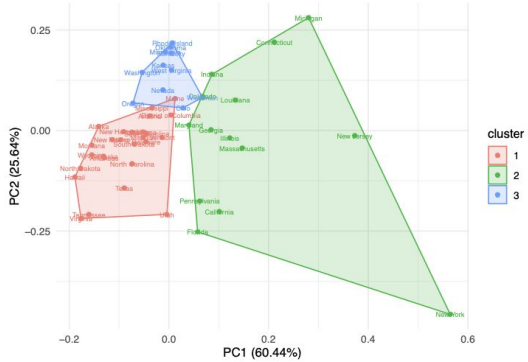


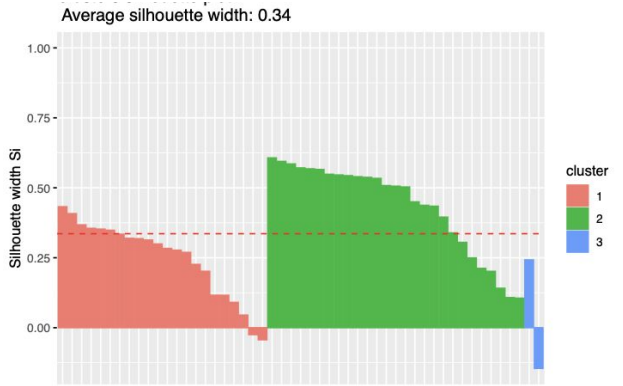


Partitioned Clustering Pandemic variables

- The optimal value of k was found to be 3. This gave a Silhouette coefficient of 0.34.
- SC = 0.34 indicates the presence of a weak structure.
- However, on trying alternate algorithms such as pam algorithm, fanny, etc, we observed that the silhouette value decreased further.
- Further PAM showed cluster overlap.
- Hence, we prefer K-means clustering for the given case.
- Although it has a low silhouette value, it turns out to be pretty high for real world data.
- Eg.North Carolina, Texas and Utah

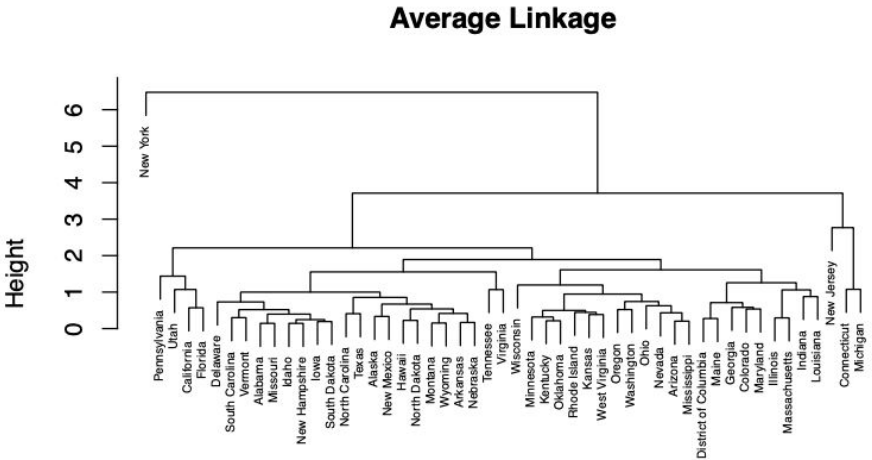


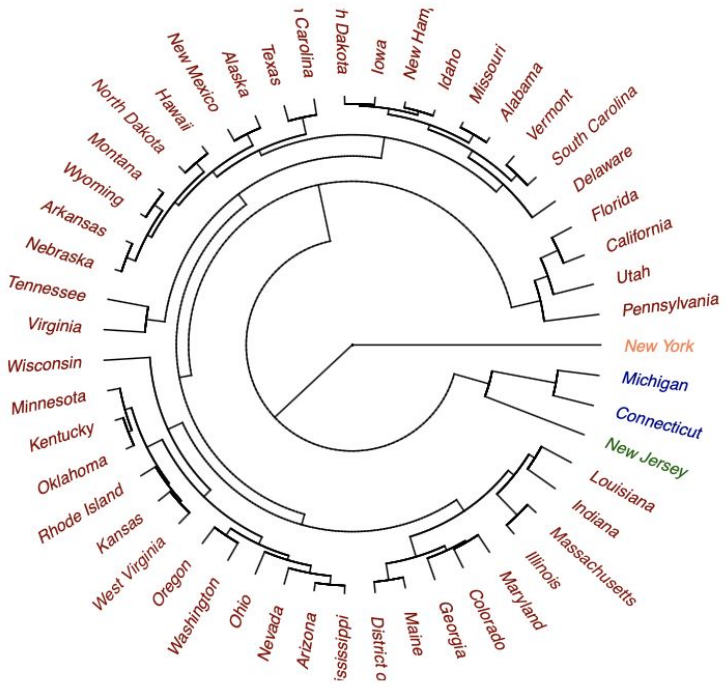




Hierarchical Clustering

- We use average linkage algorithm for hierarchical clustering.
- We get four major clusters.
- Clearly, the outliers in the PCA plot are forming separate clusters in Hierarchical Clustering.
- Michigan and Connecticut which are high on fatality rate form a separate cluster.

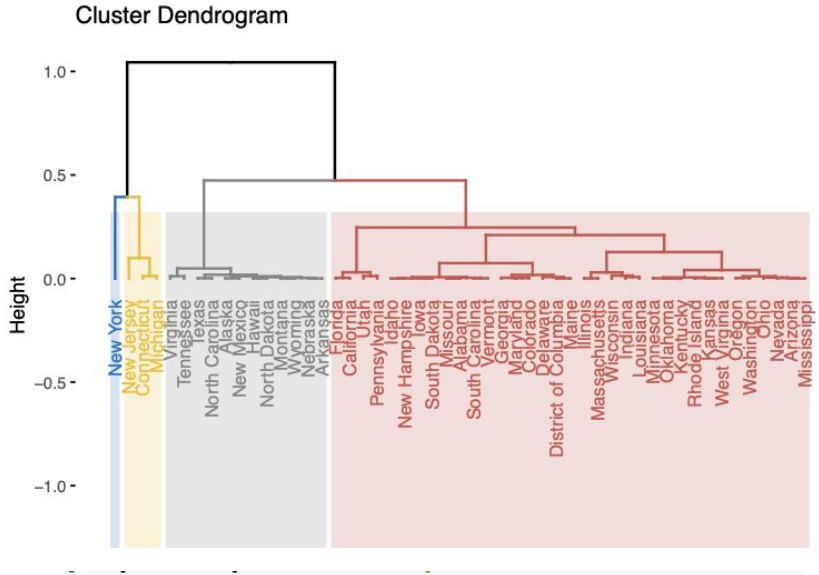


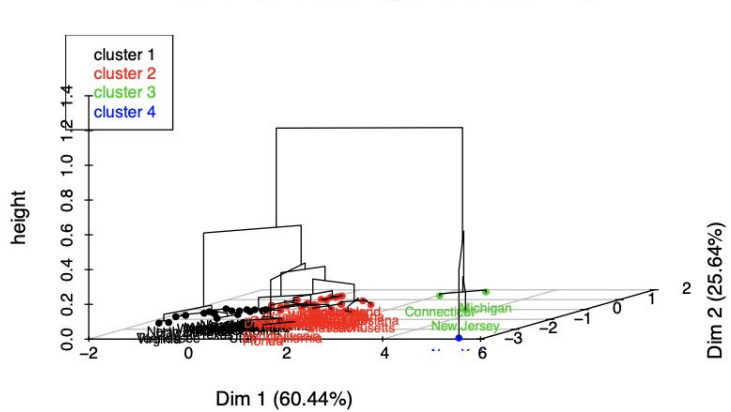


- New Jersey, which has the highest spread in the data, forms a different cluster. This points to the urgency of the situation in the state.
- New York, which has the highest number of tested persons also forms a different cluster.

Clustering Validation using HCPC analysis

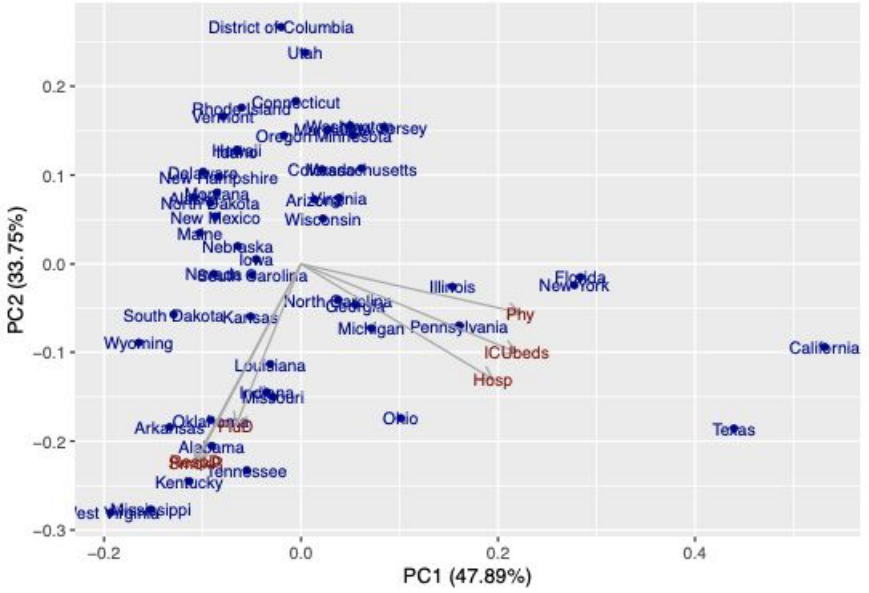
- We perform the validation of the above PCA and clustering analysis by the HCPC method here.
- The HCPC (Hierarchical Clustering on Principal Components) approach allows us to combine the three standard methods used in multivariate data analysis - PCA, Partitioning and Hierarchical clustering.
- From the dendrogram and from the 3D factor map, 4 groups is the optimal number of clusters.
- This is the number set as the cuttree level of the hierarchical method.



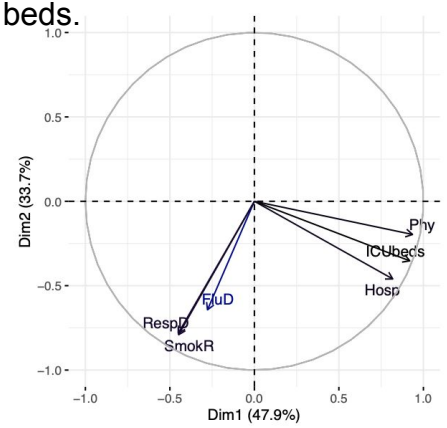


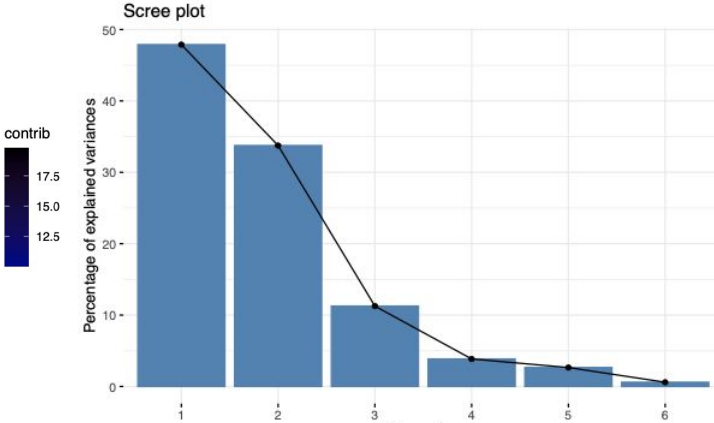
Analysis of Health Infrastructure - PCA

- The first two principal components explain about 82% of the variation in data.
- Flu deaths explains the least of the principal components.
- The states that are close to each other in the PC Plane have similar values in terms of Hospitals, Number of Physicians, etc.
- Clearly, the data validates our PCA plot as Oklahoma is much more higher on negative parameters such as smoking rate,flu death rate and respiratory disease death rate whereas Pennsylvania is more dominated by positive parameters such as physicians, hospitals and number of ICU beds.



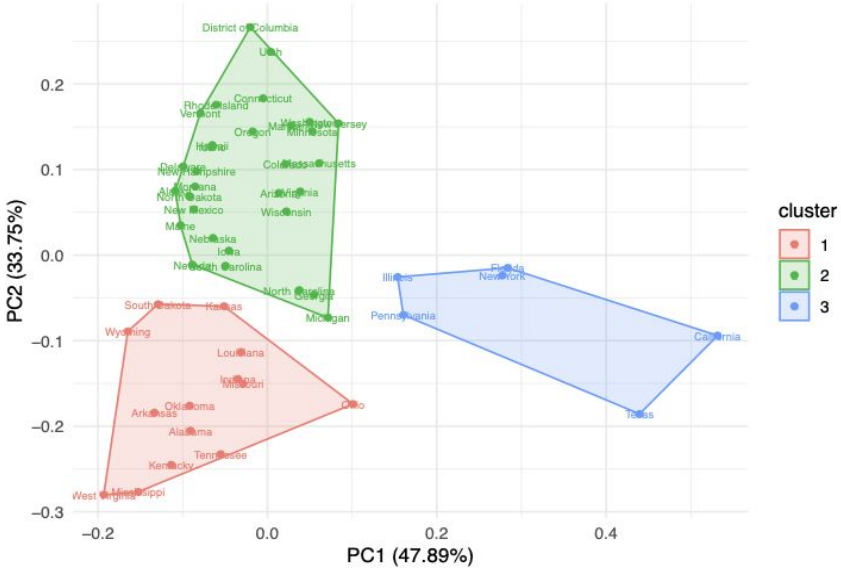
	Oklahoma	Pennsylvania
ICUbeds	1064	3169
SmokR	20.1	18.7
FluD	17.8	15.5
RespD	63.5	35.1
Phy	9472	51069
Hosp	125	199

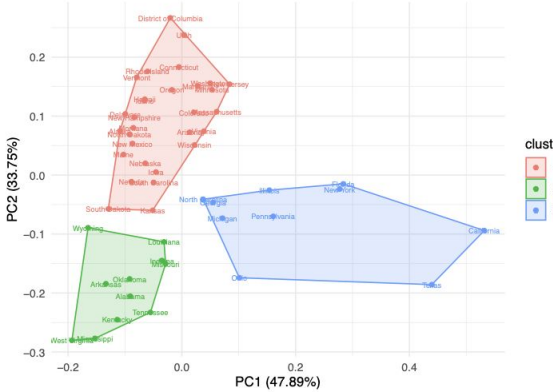


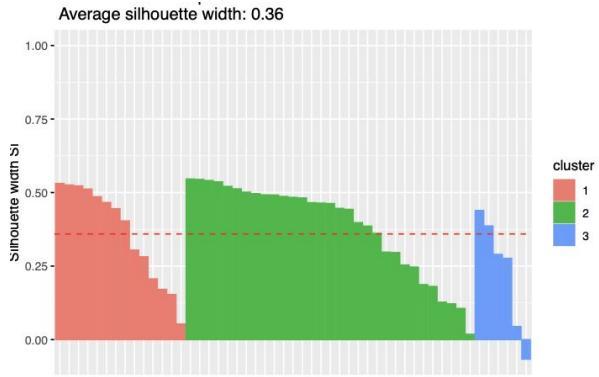


Partitioned Clustering- Health Infrastructure

- The optimal value of k was found to be 3. This gave a Silhouette coefficient of 0.36
- SC = 0.36 indicates the presence of a weak structure.
- Alternate algorithms such as pam algorithm, we observed a similar silhouette value of 0.35.
- Although it has a low silhouette value, it turns out to be pretty high for real world data.
- Eg.Oklahoma, Mississippi,Louisiana

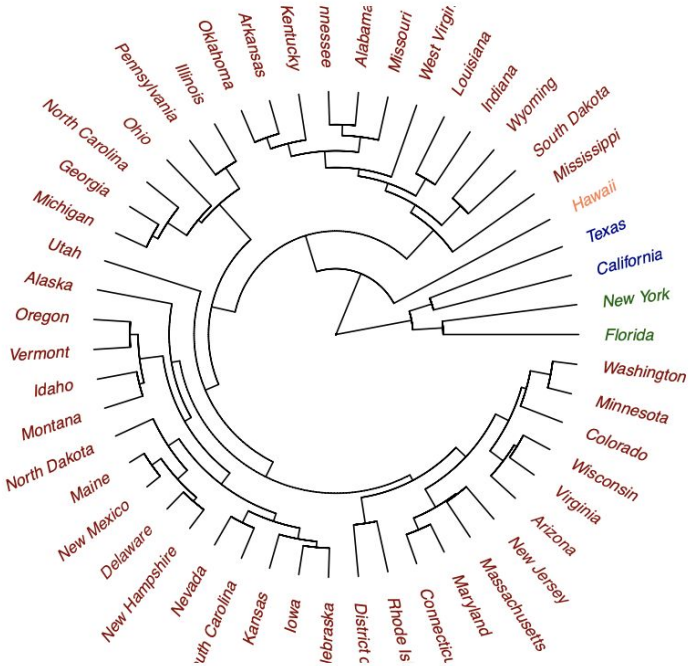


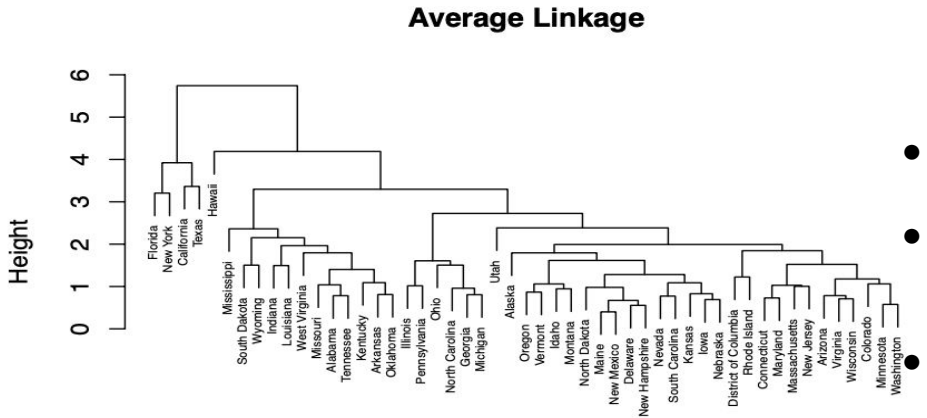




Hierarchical Clustering

- We use average linkage algorithm for hierarchical clustering.
- We get four major clusters.
- Clearly, the outliers in the PCA plot are forming separate clusters in Hierarchical Clustering.
- Texas and California are high on health infrastructure factors such as number of physicians, hospitals and ICU beds in the state.

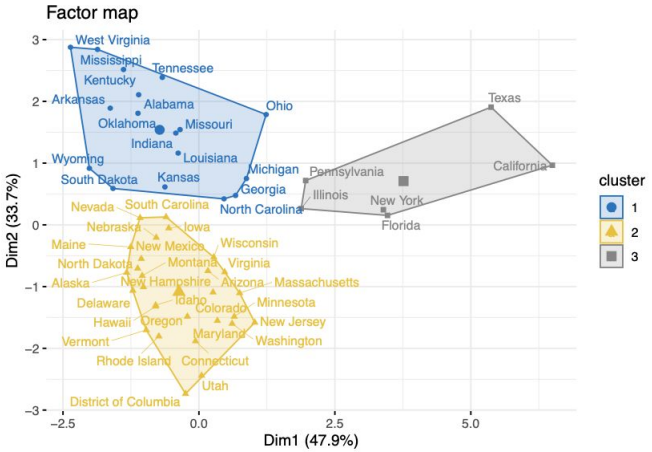


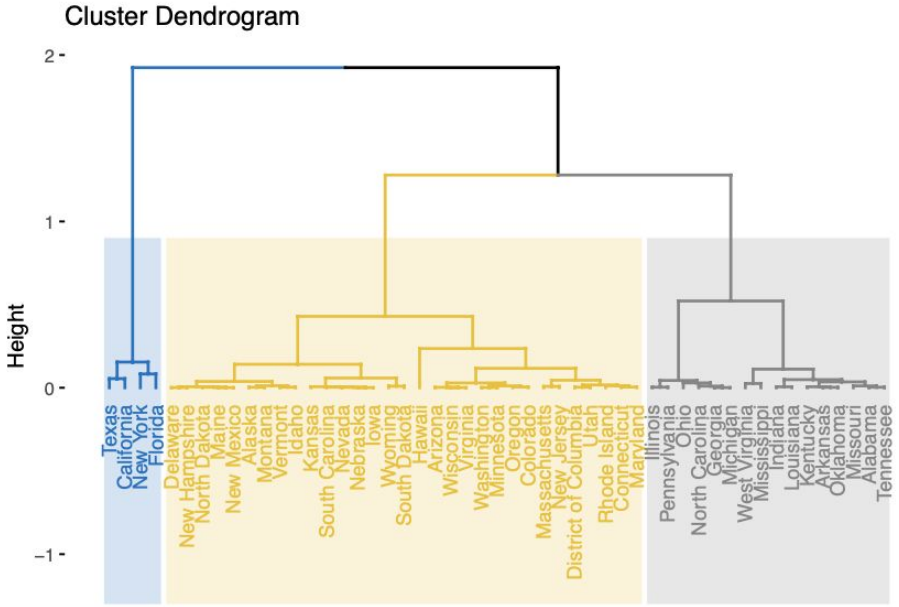


- New York and Florida also have relatively better health infrastructure than other states.
- Hawaii also forms a different cluster as its data points were much different than the average for **all** the variables.
- Texas, California and Florida had much lower spreads and fatality rate → importance of healthcare infrastructure.

Clustering Validation using HCPC analysis

- Clustering methods employed are now validated again using HCPC analysis. → combines PCA, Partitioning and hierarchical methods.
- 3 clusters are the optimal number of clusters obtained here. Same is performed in k-means clustering.



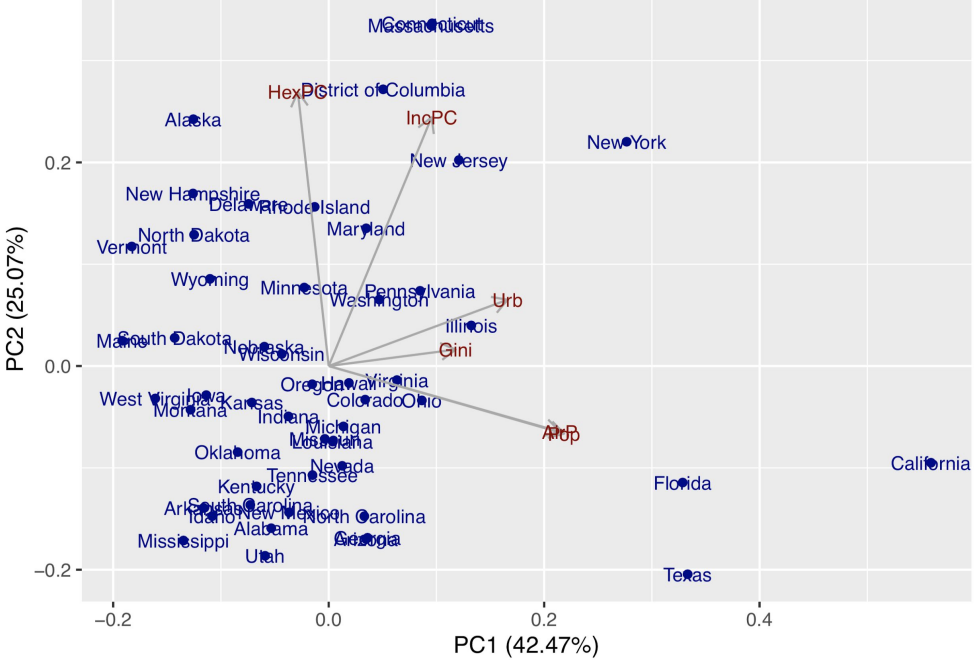


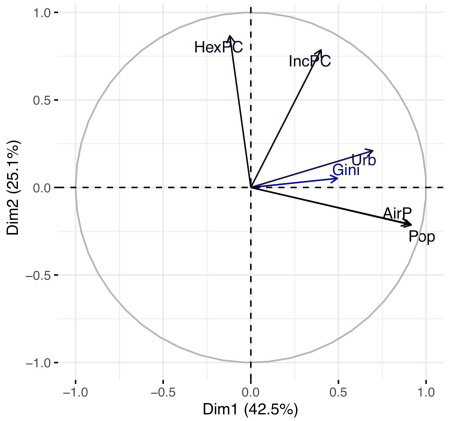
- In both the cases, we observe that Texas, New York, California and Florida form a separate cluster.

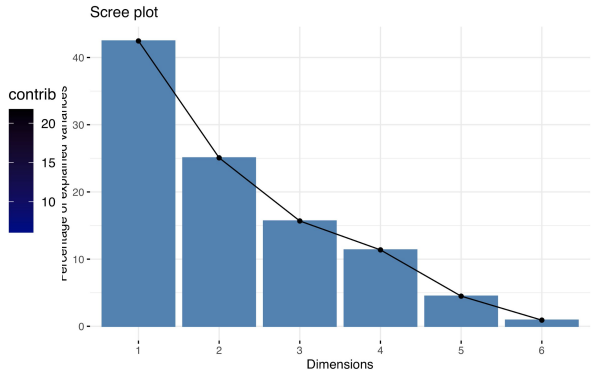
Analysis of Economic structure in the data

- The first two principal components explain about 68% of the variation in data.
- The states that are close to each other in the PC Plane have similar values in terms of the economic parameters.
- Eg. Missouri and Louisiana. In the PC plane, we see that these two states are close to each other.
- CA, TX, FL, are seen to retain their significance in terms of high population and number of airports.
- Variables that are closer to the origin like Gini are poorly represented in the factor map by the PCA.
- DC by proximity has the highest health expenditure and Utah the least.

State	Pop	Gini	IncPC	HexPC	AirP	Urb
Louisiana	4645184	0.499	45542	7815	1	0.732
Missouri	6169270	0.4646	46635	8107	2	0.704

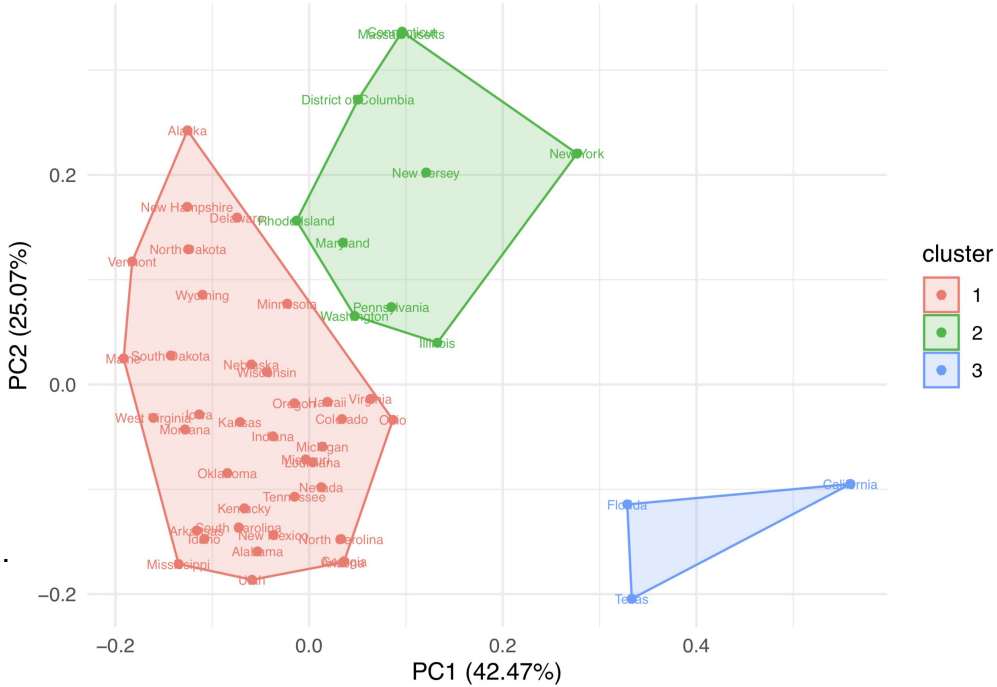


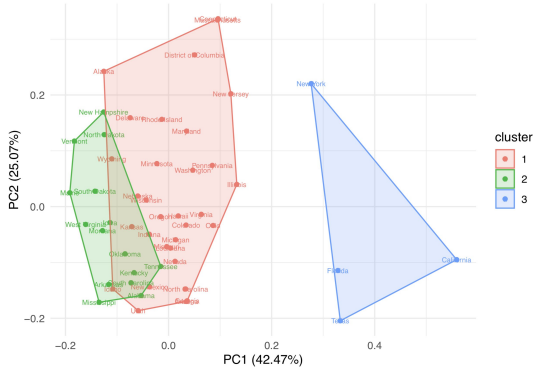


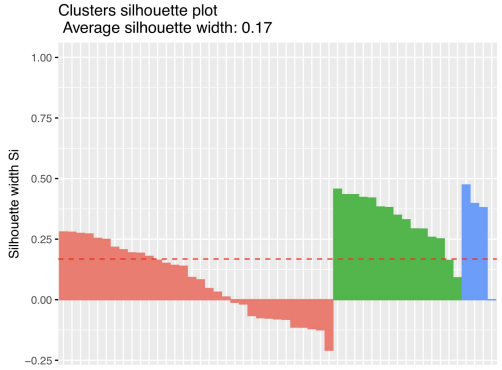


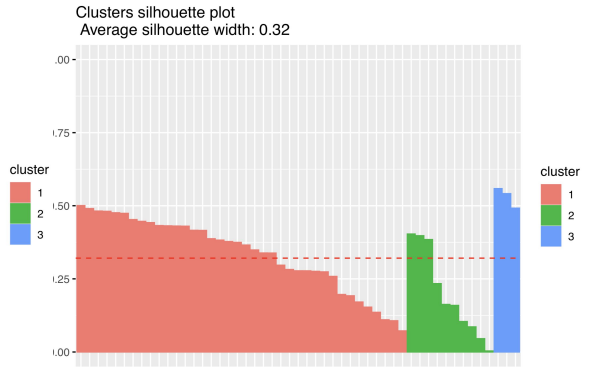
Partitioned Clustering Economic variables.

- The optimal value of k was found to be 3. This gave a Silhouette coefficient of 0.32.
- SC = 0.32 indicates the presence of a weak structure.
- However, on trying alternate algorithms such as pam algorithm, fanny, etc, we observed that the silhouette value decreased further.
- Further PAM showed cluster overlap.
- Hence, we prefer K-means clustering for the given case. The low silhouette value, is rather high for real world data.
- CA, TX, FL economic bloc revealed.
- NY, DC, NJ, and other upper East Coast bloc next in line.



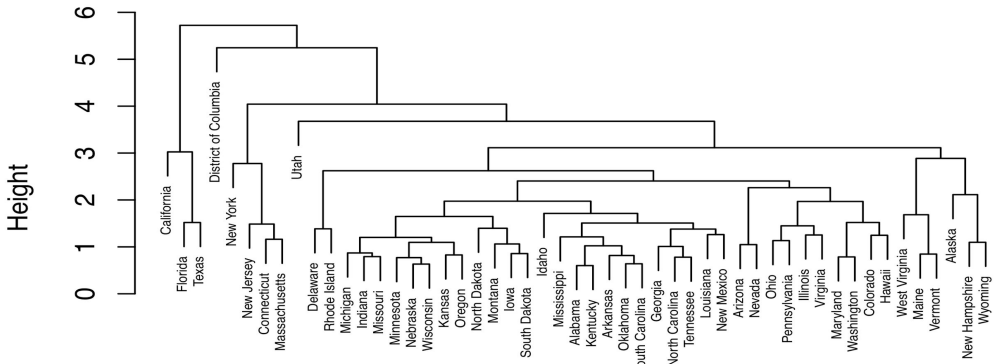


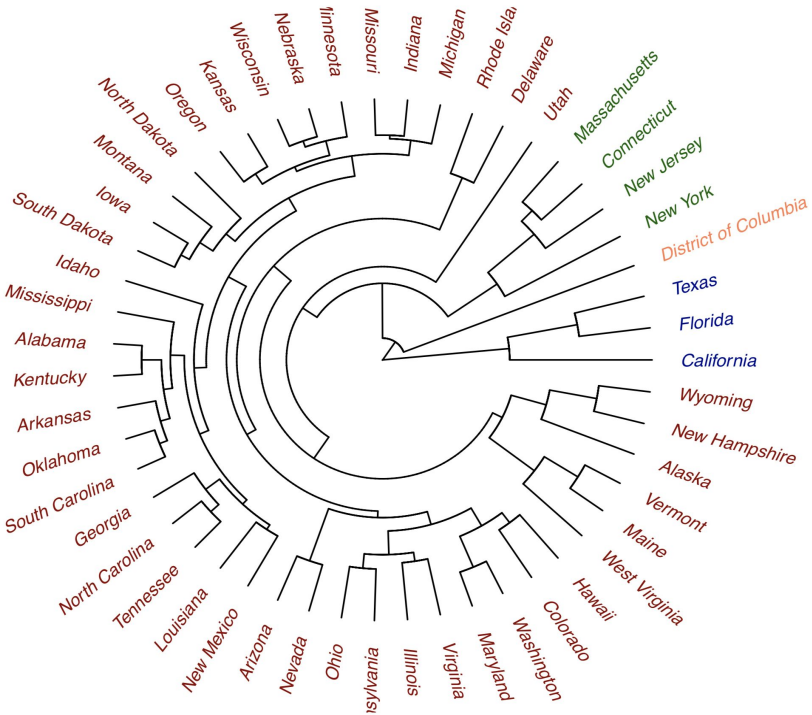




Hierarchical Clustering

- We use average linkage algorithm for hierarchical clustering. Problem of crowding and chaining resolved.
- We get four major clusters with PCA plot outliers forming separate clusters in Hierarchical Clustering.
- CA, TX, FL bloc seen here too.
- DC is a separate cluster owing to its small population and high IncPC. It has no large airports.
- Special status of Capitol Hill in the US economy.



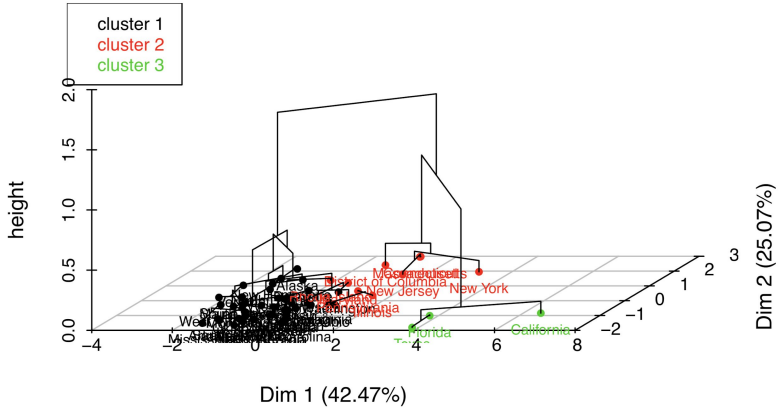


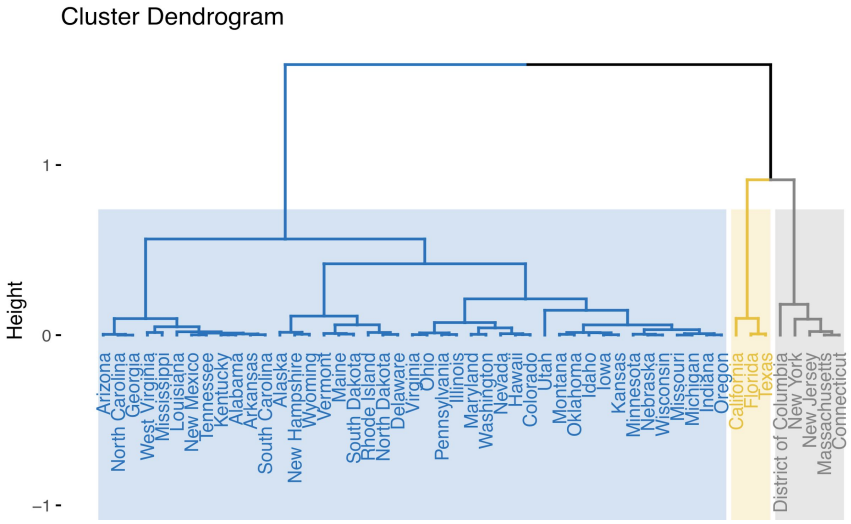
- NY, NJ → large pandemic spread
  Also, Connecticut & Massachusetts → significant spread.
- Economic centres - a zone for spread of the pandemic, relative to others.

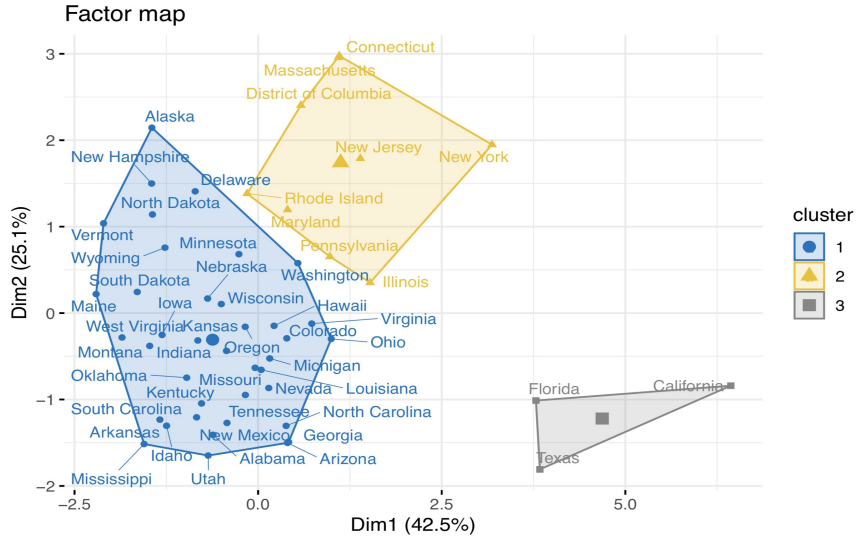
Clustering Validation using HCPC analysis

- Clustering methods employed are now validated again using HCPC analysis. → combines PCA, Partitioning and hierarchical methods.
- 3 clusters are the optimal number of clusters obtained here. Same is performed in k-means clustering.
- In hierarchical too 3 is the optimal cuttree level. However the selection of 4, highlights the special status of DC.

Hierarchical clustering on the factor map

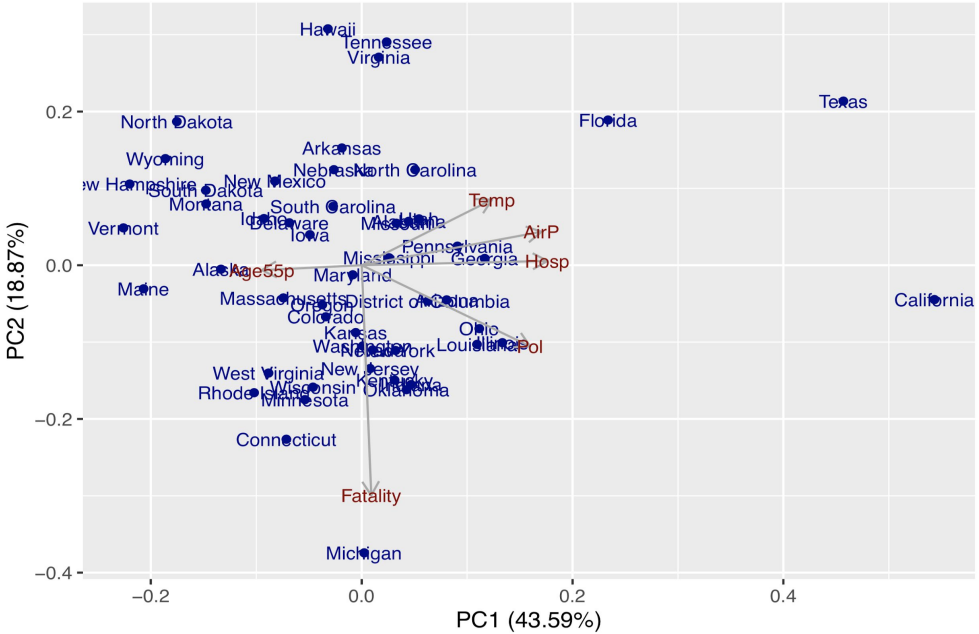






What determines fatality?

- Perform variable selection on the regression model with Fatality as the response variable. Results in: AirP, Pol, Temp, Age55p and Hosp as key determinants of Fatality;
- Perform PCA on the selected model with a smaller feature space. PC1 and PC2 explain 62% variance in data.
- CA, TX, FL are still outliers. NY is not.



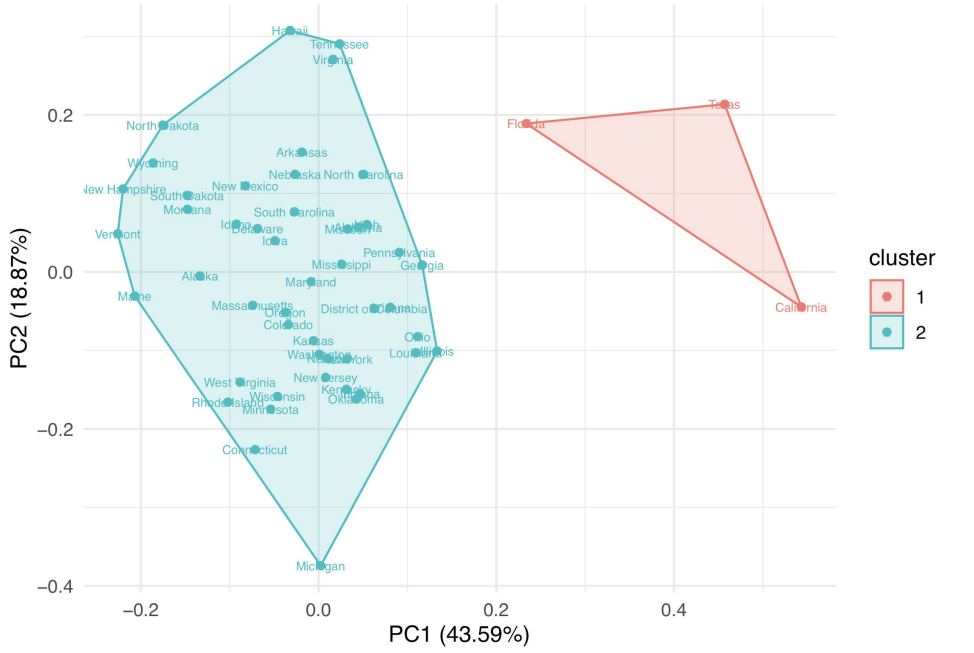

```
lmf01<-lm(Fatality~., data=dff)
lmf02 <- stepAIC(lmf01,direction="both", trace=FALSE)
summary(lmf02)
```

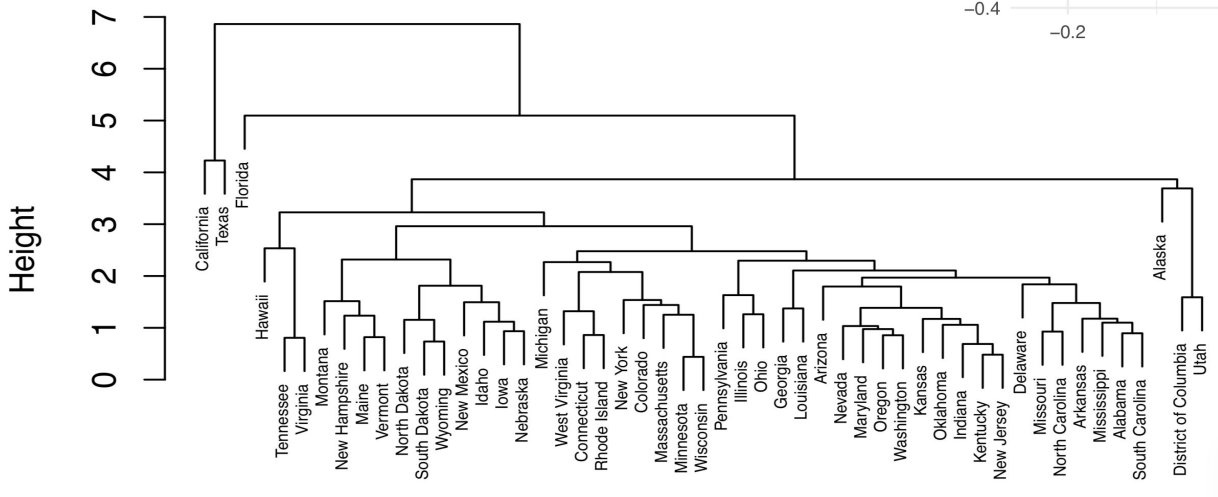
```
##
## Call:
## lm(formula = Fatality ~ Gini + AirP + Urb + Pol + Temp + Age55p +
##     Hosp, data = dff)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0280486 -0.0073689  0.0008818  0.0064282  0.0248928
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.259e-01  4.336e-02  -2.905  0.00579 **
## Gini         1.440e-01  8.807e-02   1.635  0.10943
## AirP        -4.871e-03  1.846e-03  -2.638  0.01155 *
## Urb          3.216e-02  1.606e-02   2.003  0.05151 .
## Pol          3.580e-03  1.630e-03   2.196  0.03354 *
## Temp        -1.118e-03  4.333e-04  -2.579  0.01341 *
## Age55p       1.862e-01  6.944e-02   2.682  0.01034 *
## Hosp         7.374e-05  3.579e-05   2.060  0.04545 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01217 on 43 degrees of freedom
## Multiple R-squared:  0.3554, Adjusted R-squared:  0.2505
## F-statistic: 3.388 on 7 and 43 DF,  p-value: 0.005733
```

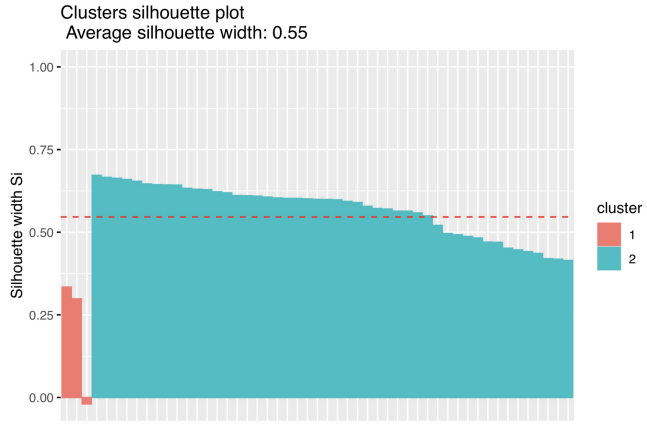
```
dffnew <- dselect(dff,c(Fatality,AirP,Pol,Temp,Age55p,Hosp))
```

Clustering

- The optimal value of k was found to be 2.
- The silhouette coefficient turned out to be a whopping 0.55, thereby indicating a reasonable structure in the clustering model.
- Hierarchical clustering too results in the same.
- CA, TX, FL turn out to be a separate group and our previous reasoning holds good.

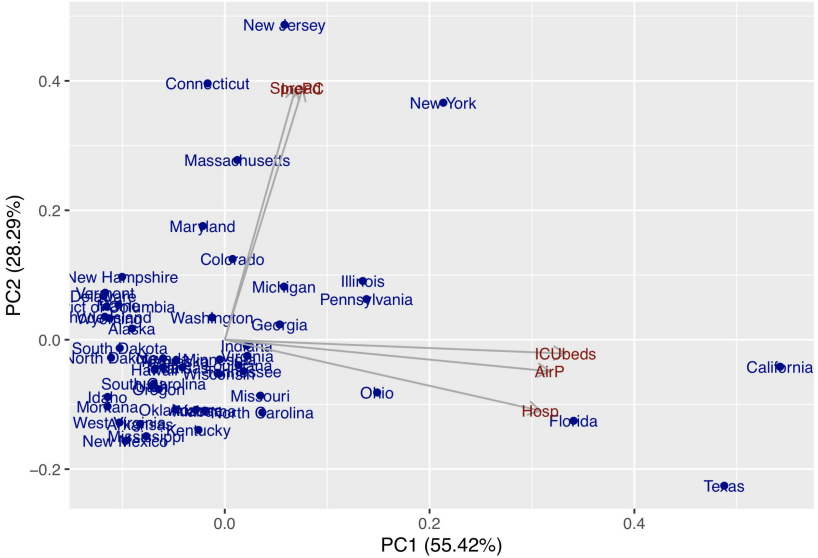






What determines spread?

- Perform variable selection on the regression model with Spread as the response variable.
- Results in: AirP, IncPC, ICUbeds and Hosp as key determinants of Spread;
- Perform PCA on the selected model with a smaller feature space. PC1 and PC2 explain 84% variance in data.
- CA, TX, FL are still outliers. NY is not.



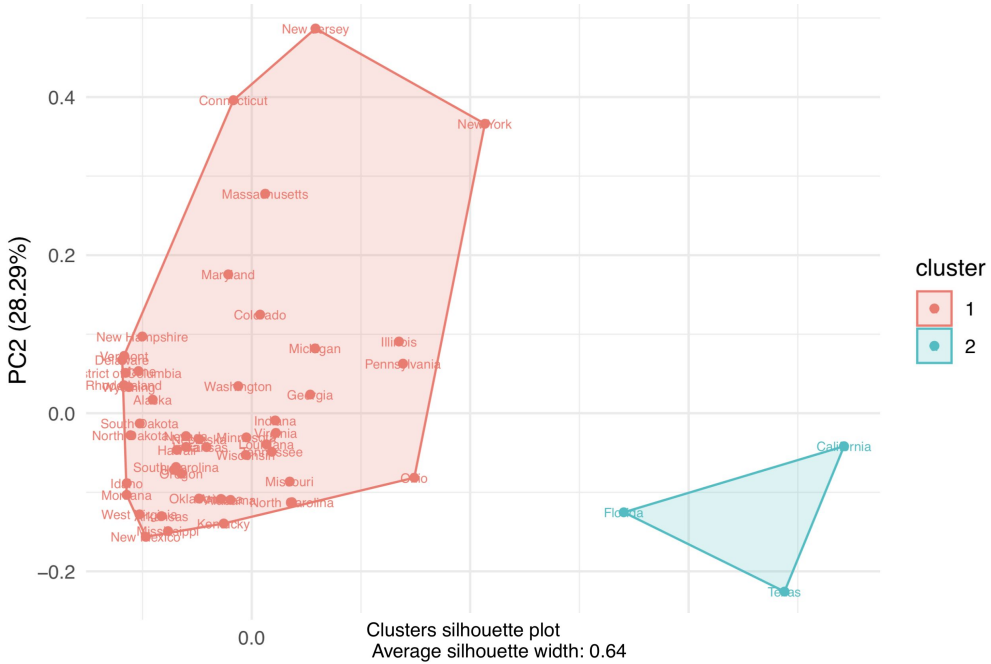

```
lms01<-lm(Spread~., data=dfs)
lms02 <- stepAIC(lms01,direction="both", trace=FALSE)
summary(lms02)
```

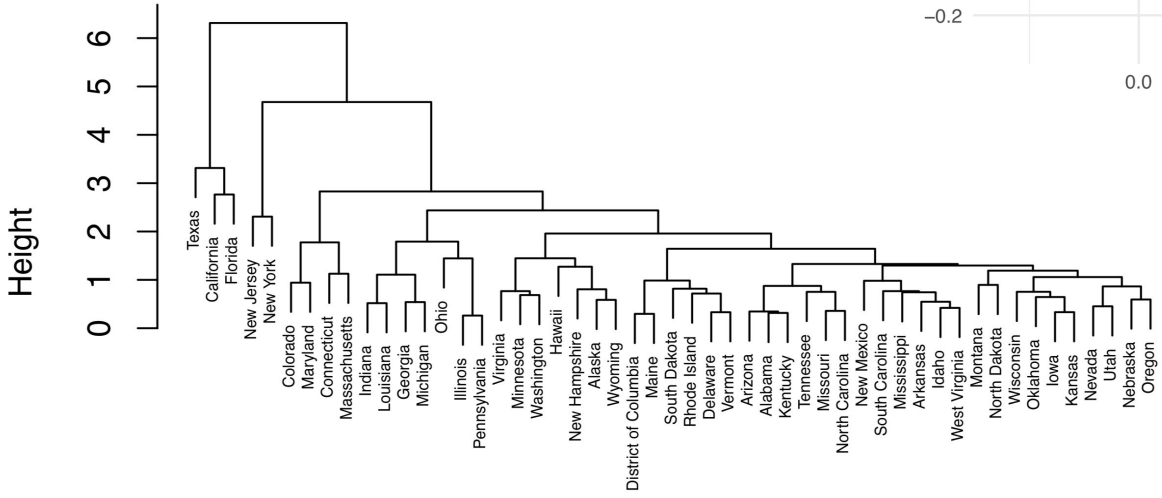
```
##
## Call:
## lm(formula = Spread ~ Gini + IncPC + AirP + ICUbeds + RespD +
##     Hosp, data = dfs)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.113536 -0.041535  0.002819  0.028312  0.202597
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.889e-01  2.498e-01  -1.557  0.12668
## Gini         7.838e-01  4.442e-01   1.765  0.08455 .
## IncPC        4.111e-06  1.355e-06   3.035  0.00403 **
## AirP        -7.494e-02  1.430e-02  -5.241 4.34e-06 ***
## ICUbeds      1.067e-04  2.222e-05   4.800 1.86e-05 ***
## RespD       -1.611e-03  1.094e-03  -1.472  0.14806
## Hosp        -5.150e-04  2.277e-04  -2.262  0.02870 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05978 on 44 degrees of freedom
## Multiple R-squared:  0.6331, Adjusted R-squared:  0.583
## F-statistic: 12.65 on 6 and 44 DF,  p-value: 3.067e-08
```

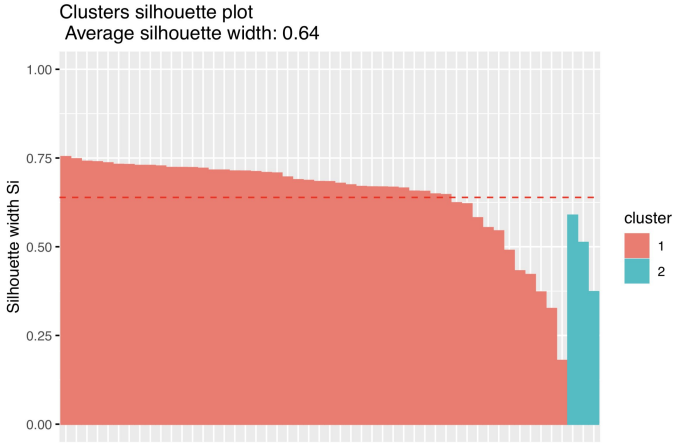
```
dfsnew <- dselect(dfs,c(Spread,IncPC,AirP,ICUbeds,Hosp))
```

Clustering

- The optimal value of k was found to be 2.
- The silhouette coefficient turned out to be a whopping 0.64, thereby indicating a reasonable structure in the clustering model.
- CA, TX, FL turn out to be a separate group and our previous reasoning holds good.







Conclusions

- Thus we have observed several patterns in the United States COVID-19 data using methods of PCA and clustering.
- Many of our analysis plots identify New York State to be an outlier amongst the rest of the states. This is majorly due to the state's distinction as ground zero for the COVID-19 spread in the US.
- All our analyses cluster California, Texas and Florida together - health and economic parameters particularly.
- States like Michigan, NJ and Connecticut also had a high spread and fatality of the pandemic.
- It is also to be noted that the states - California, Texas have a higher number of physicians, hospitals as well as ICU beds.
- We speculate that the reasons for the low spread and fatality rate in the states of CA, TX and FL are because of their above-average health infrastructure.