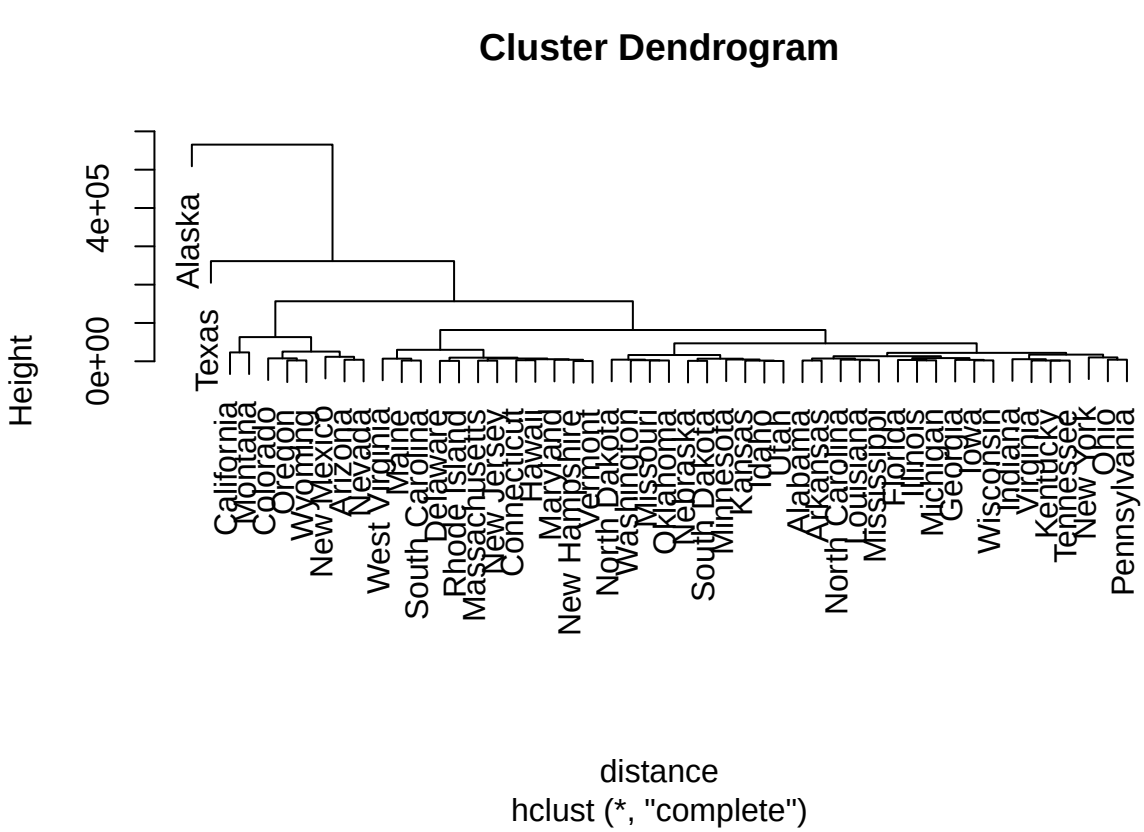
Clustering.R

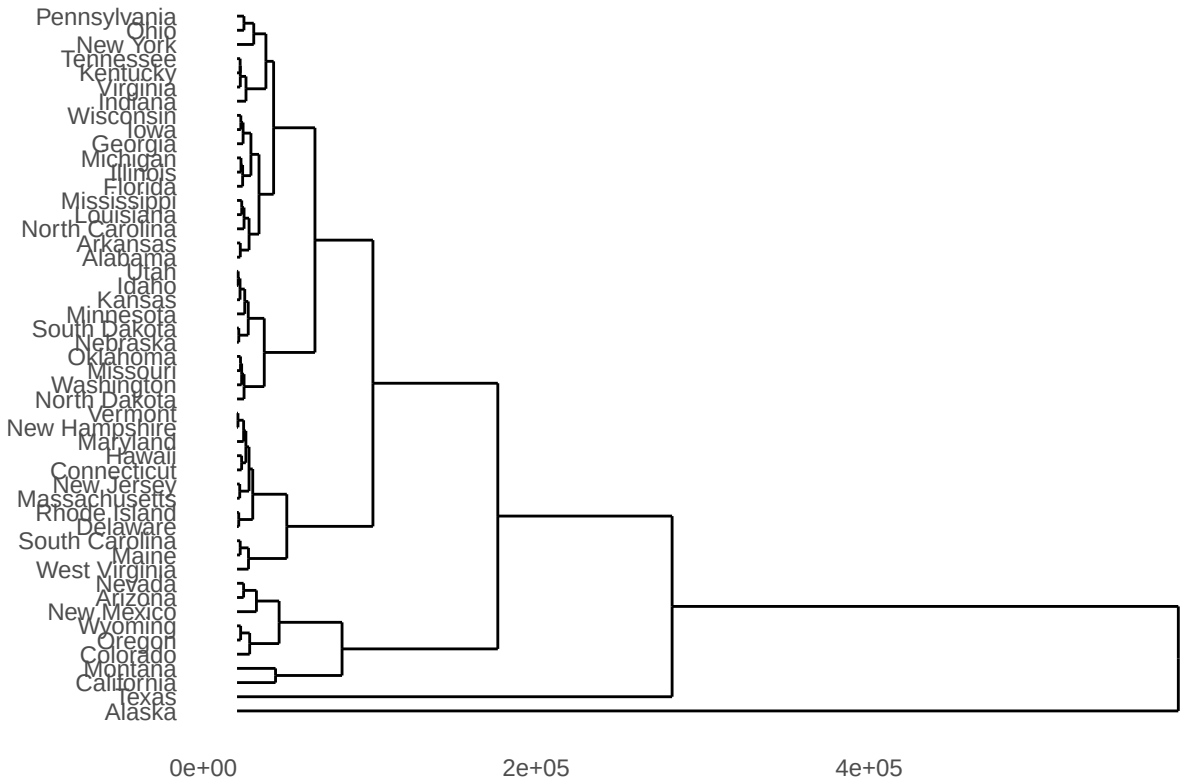
jasho

2020-03-14

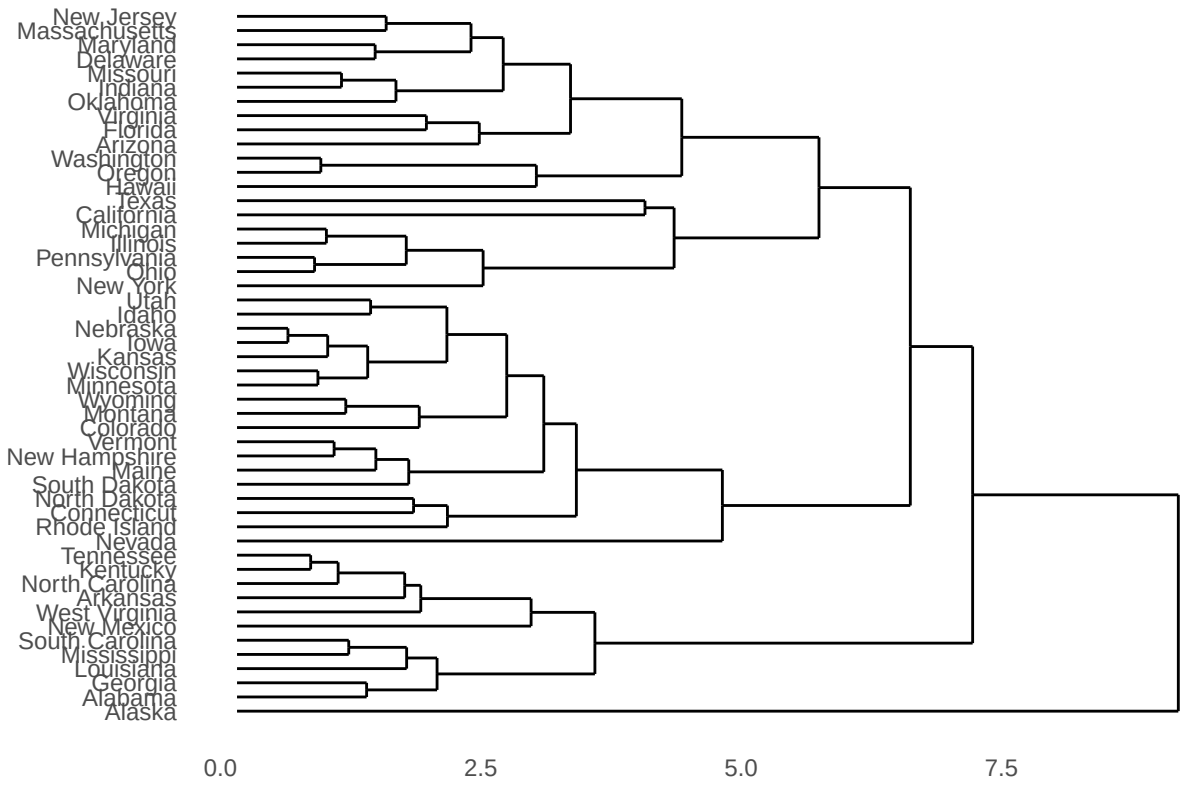
```
pacman::p_load(datasets,tidyverse,ggdendro, pander, cluster)
dat = state.x77

distance <- dist(as.matrix(dat))
hc <- hclust(distance)
plot(hc)
```



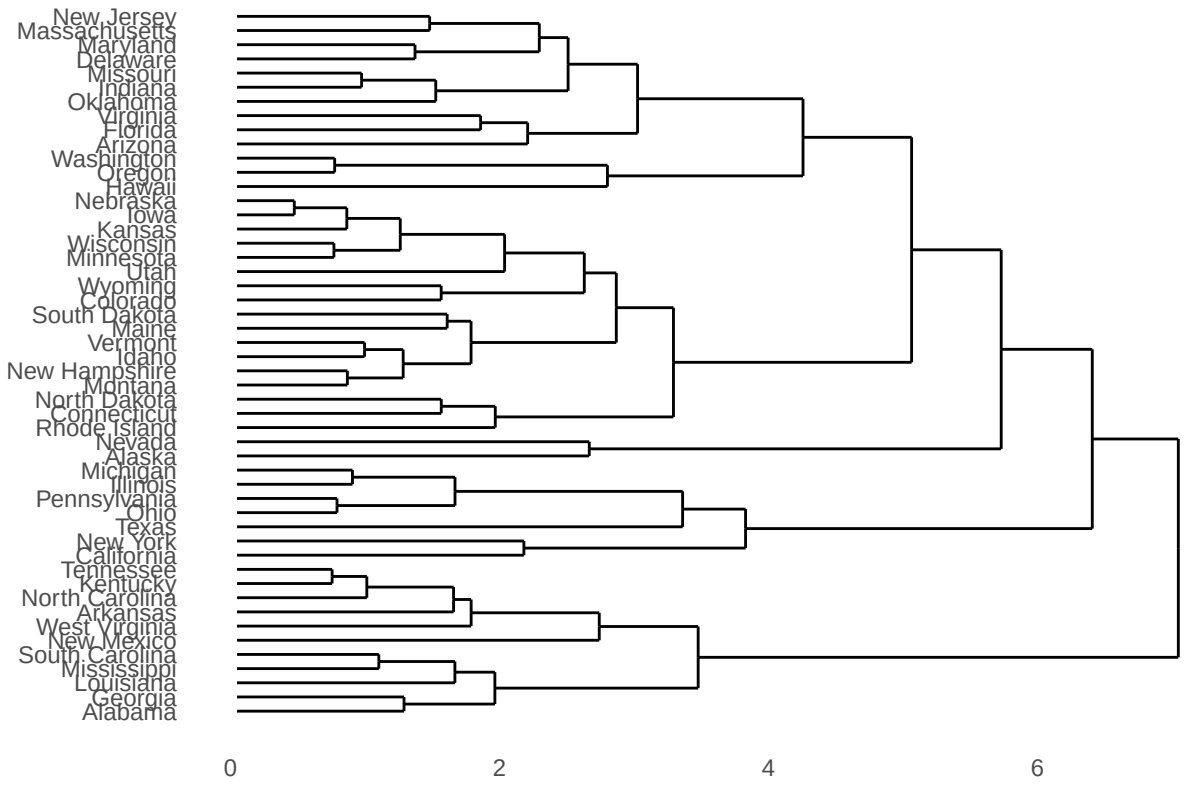
```
ggdendro::ggdendrogram(hc, rotate = 1)
```




```
#Scaling eliminated the effects or population
#and area that overpowered the other variables.
#this allowed the population to have asimilar
#effect to area. i.e cali and texas being the
#closest to eachother.

datsc <- as.data.frame(scale(dat))
scdist <- dist(as.matrix(datsc))
hcsc <- hclust(scdist)
ggdendro::ggdendrogram(hcsc, rotate = 1)
```

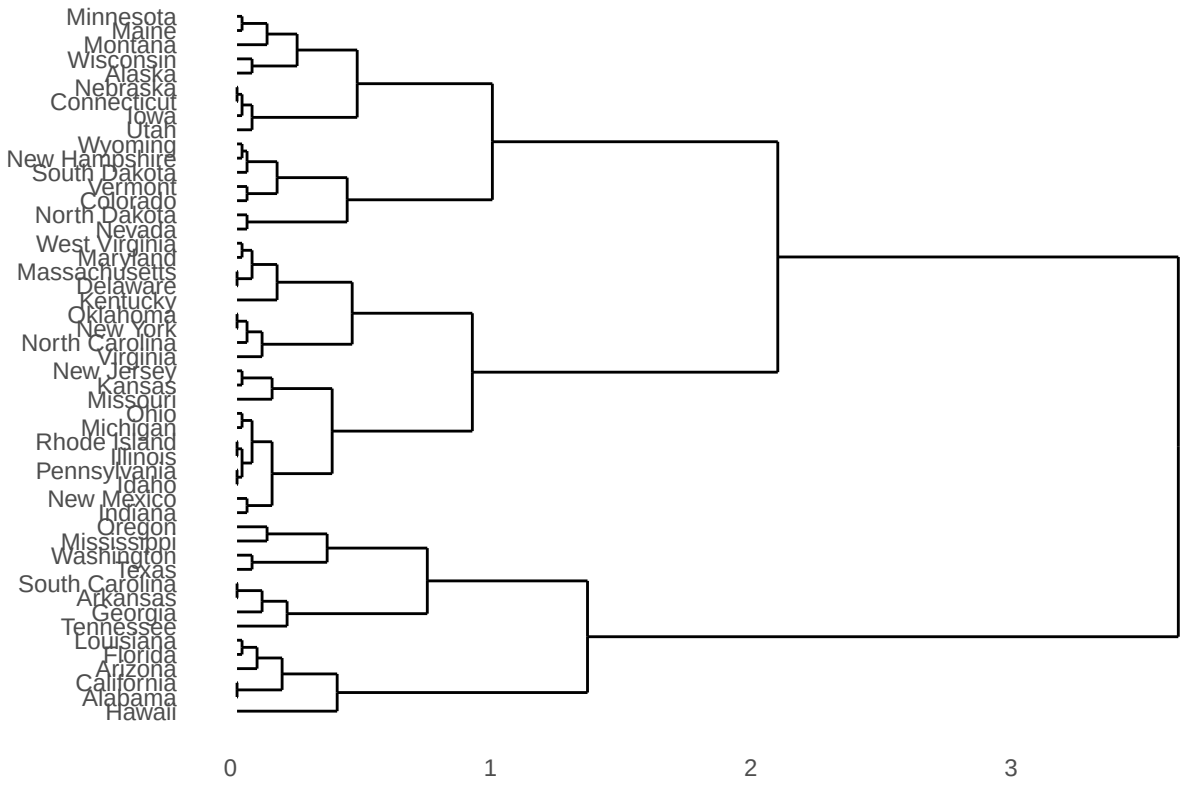



```
#Taking out area really starts to show the
#important stuff like HS grad and literacy,
#these
#things were not nearly as important when area
#was involved

namos <- dat[,0]
datAsc <- datsc %>%
  select(-Area)
Ascdist <- dist(as.matrix(datAsc))
Ahc <- hclust(Ascdist)
ggdendro::ggdendrogram(Ahc, rotate = 1)
```

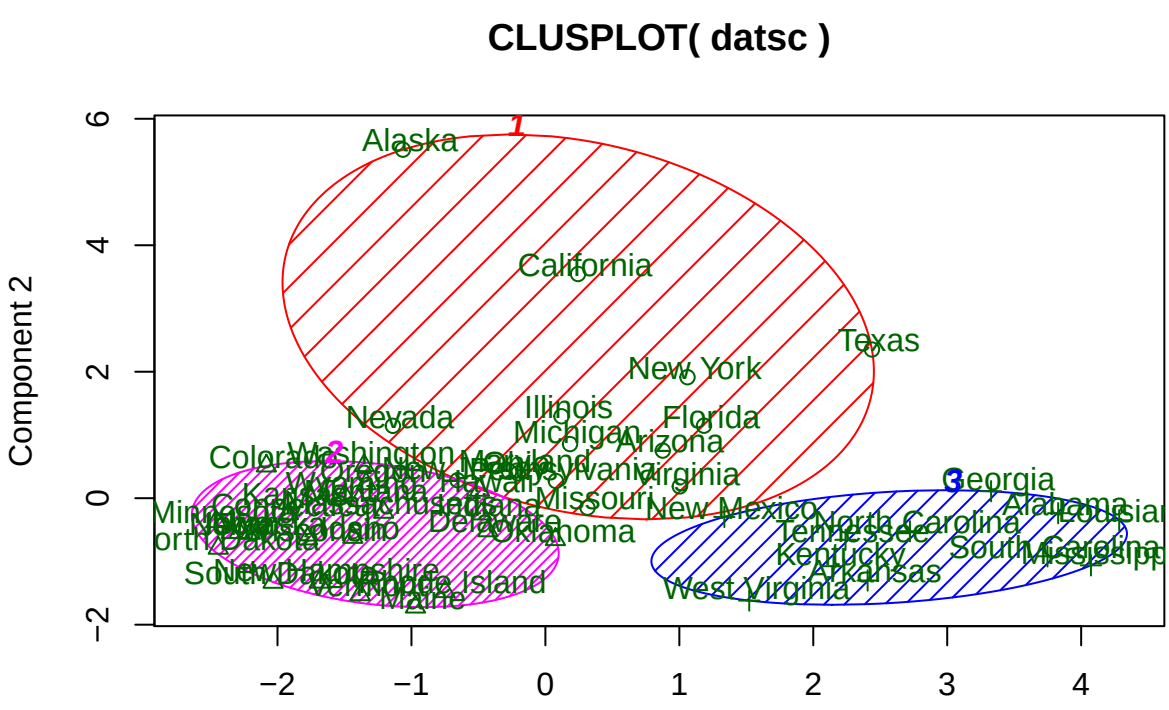



```
#Frost only is super cool, and since I lived in
#Texas and Washington can confirm that Frost
```

```
Fdat <- datsc %>%
  select(Frost)
Fdist <- dist(as.matrix(Fdat))
Fhc <- hclust(Fdist)
ggdendrogram(Fhc, rotate = 1)
```

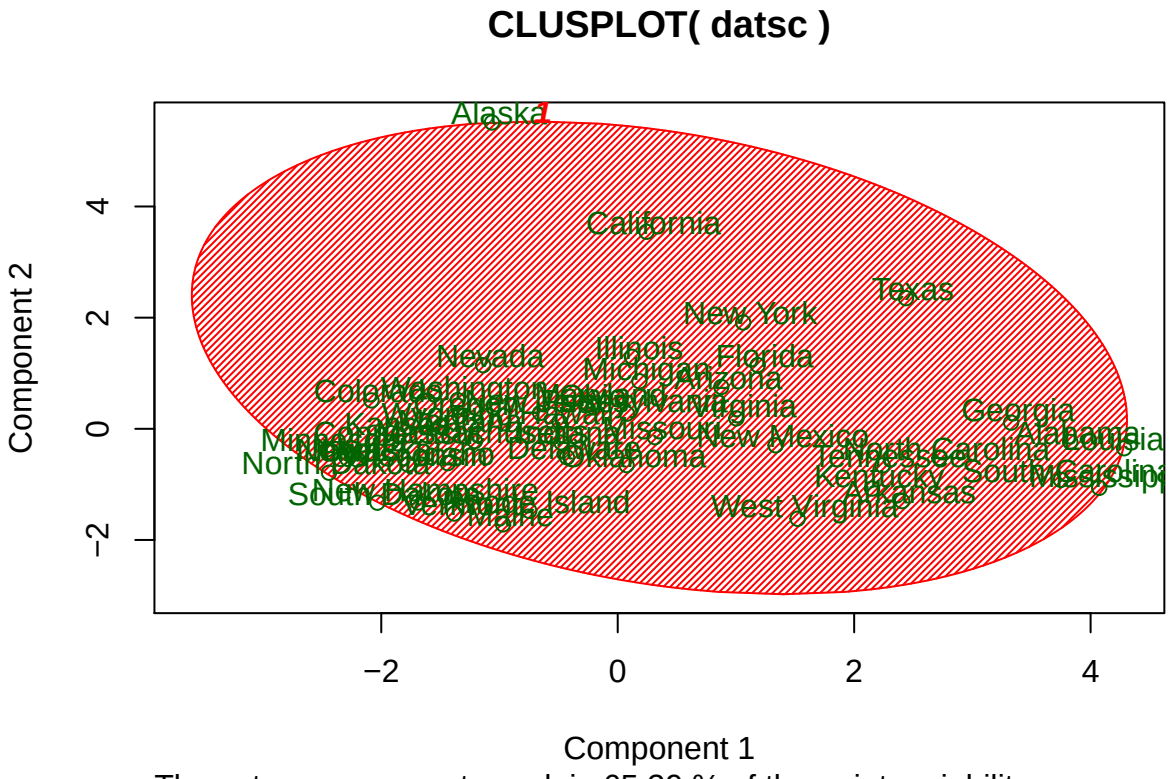


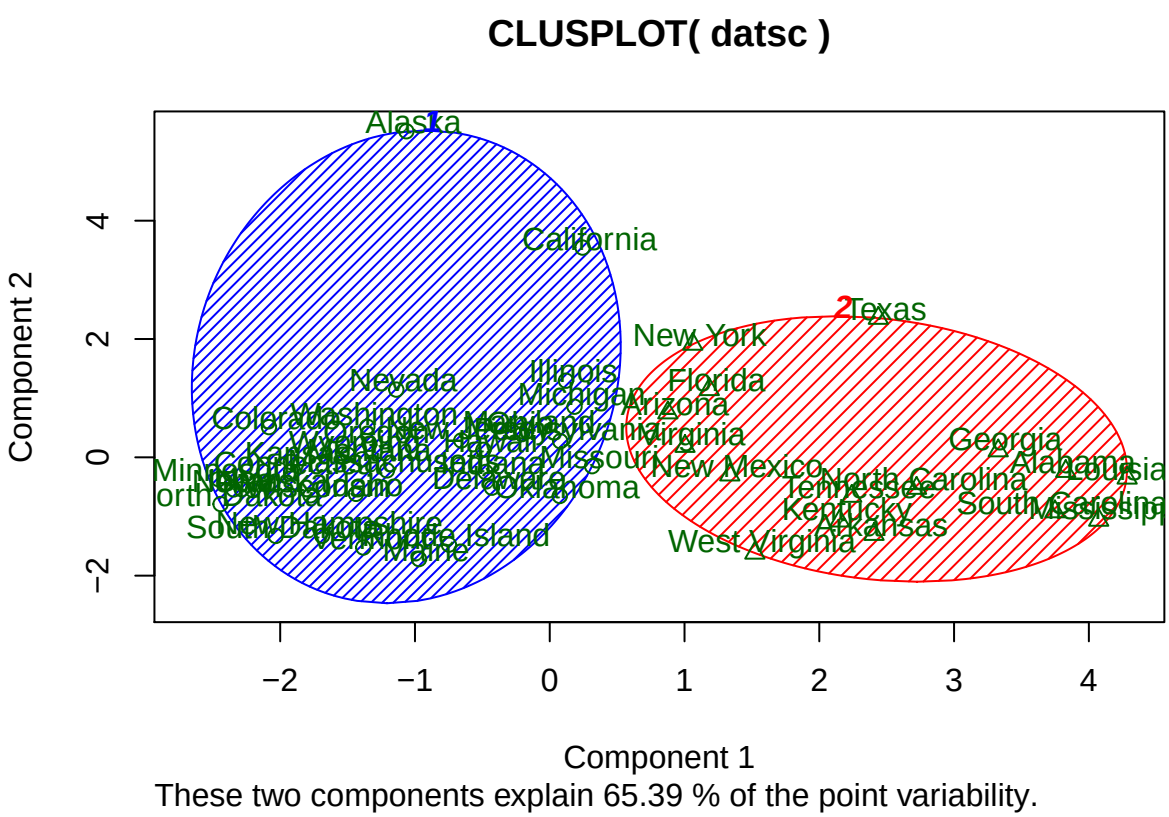

```
#K-means

clus1 <- kmeans(datsc, 3)
clusplot(datsc, clus1$cluster, color = TRUE, shade = TRUE, labels = 2, lines =0)
```



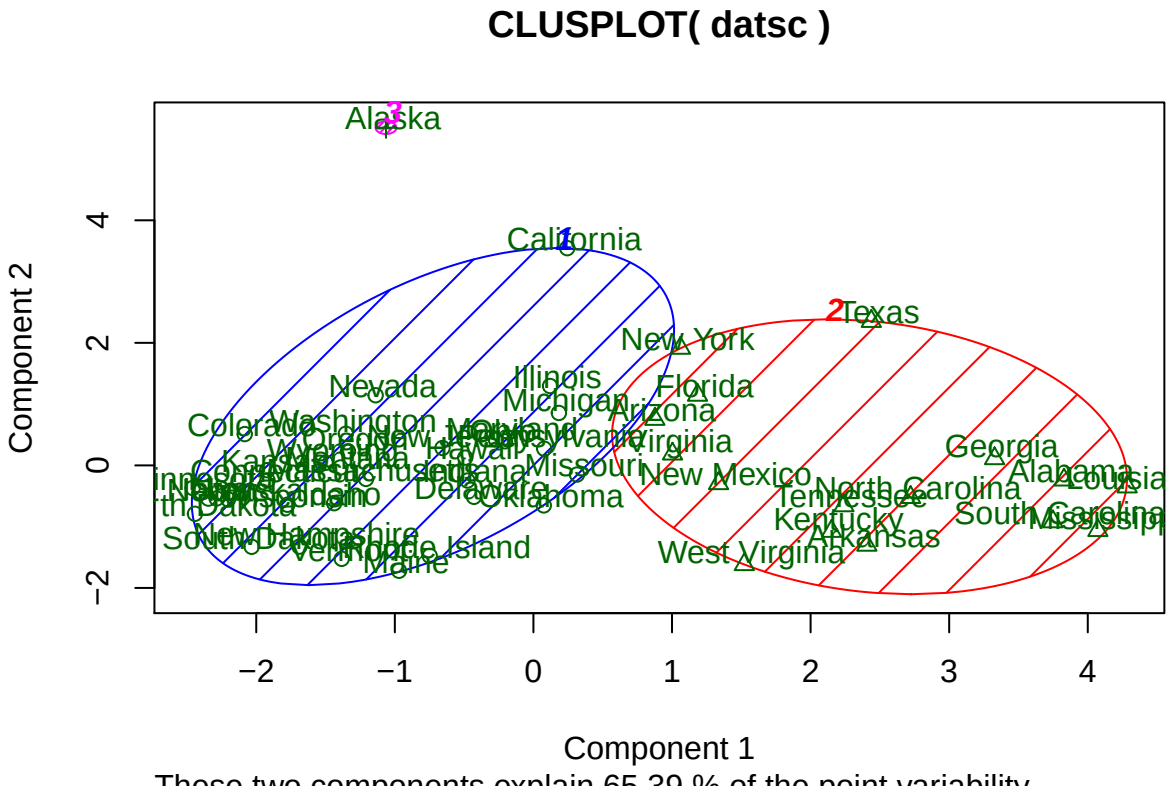
These two components explain 65.39 % of the point variability.

```
#large population/large area is the dominating
#factor for the clusters

for (i in 1:25) {
thisclus <- kmeans(datsc, i)
clusplot(datsc, thisclus$cluster, color = TRUE, shade = TRUE, labels = 2, lines =0)
}
```

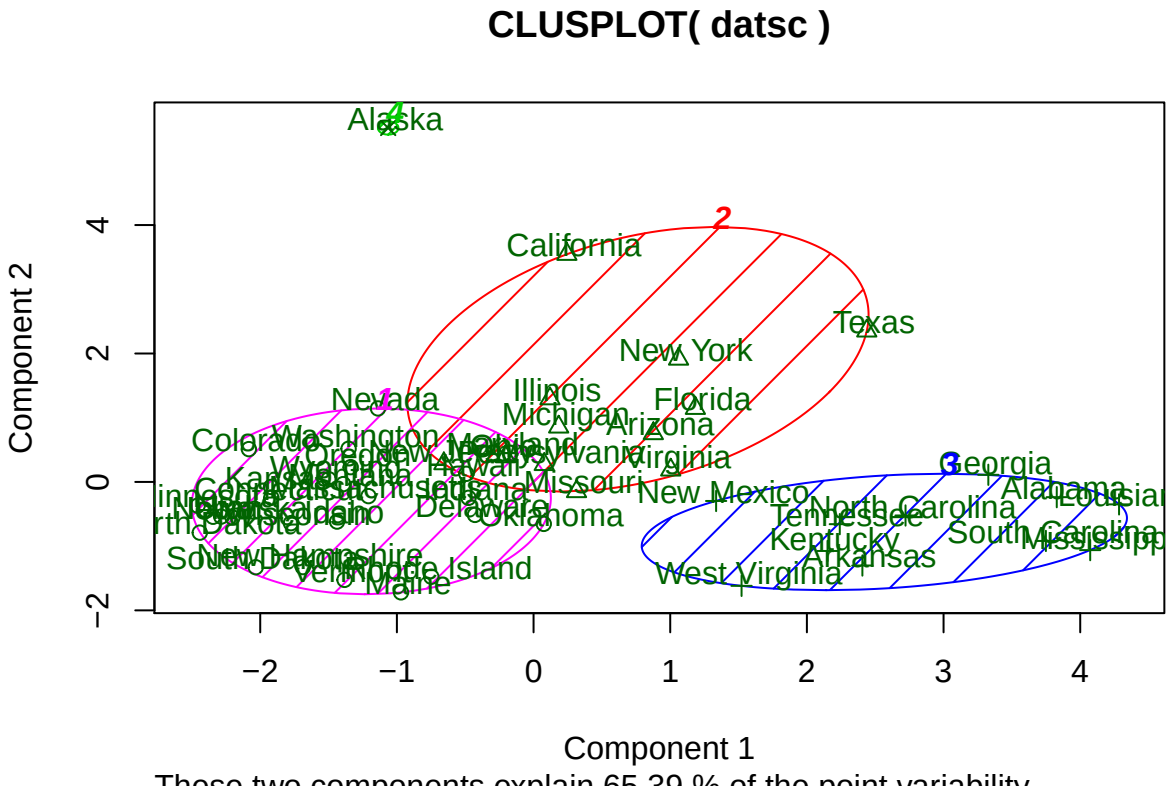

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Nevada
Illinois
Florida
Michigan
Georgia
New Mexico
Oklahoma
West Virginia
Arkansas
Kentucky
−2
0
2
4



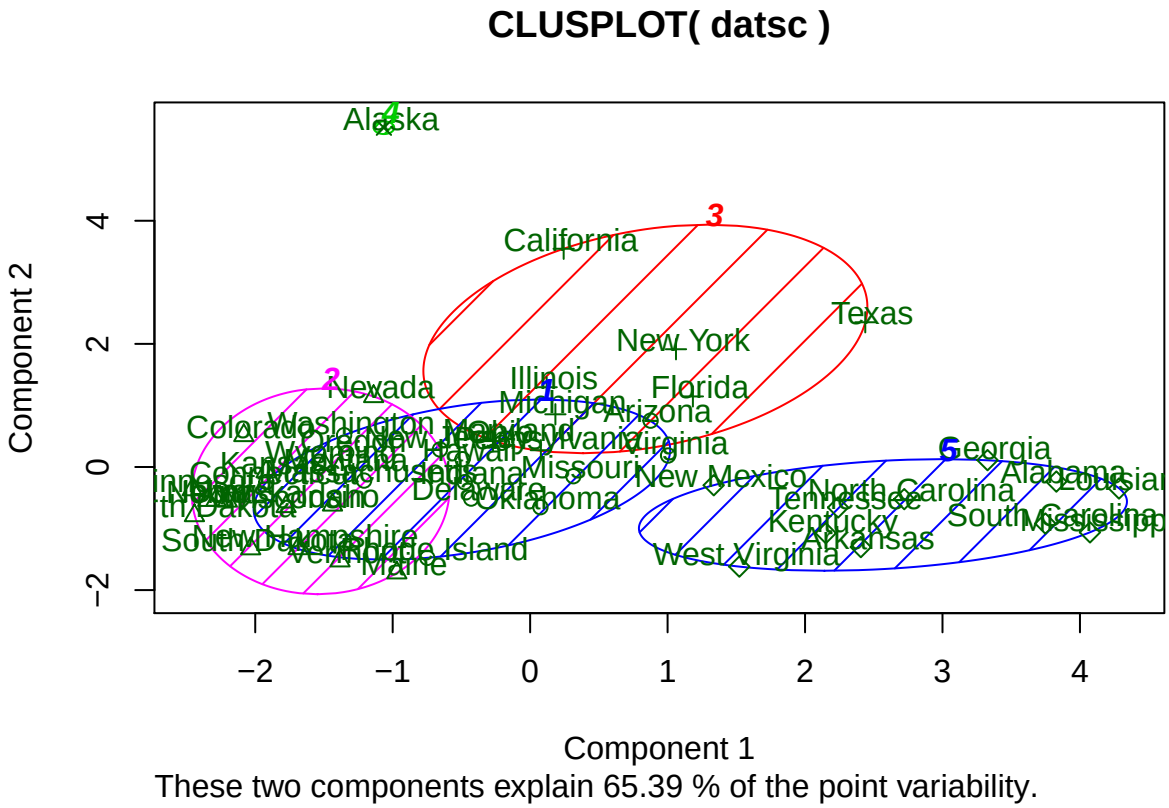
These two components explain 65.39 % of the point variability.



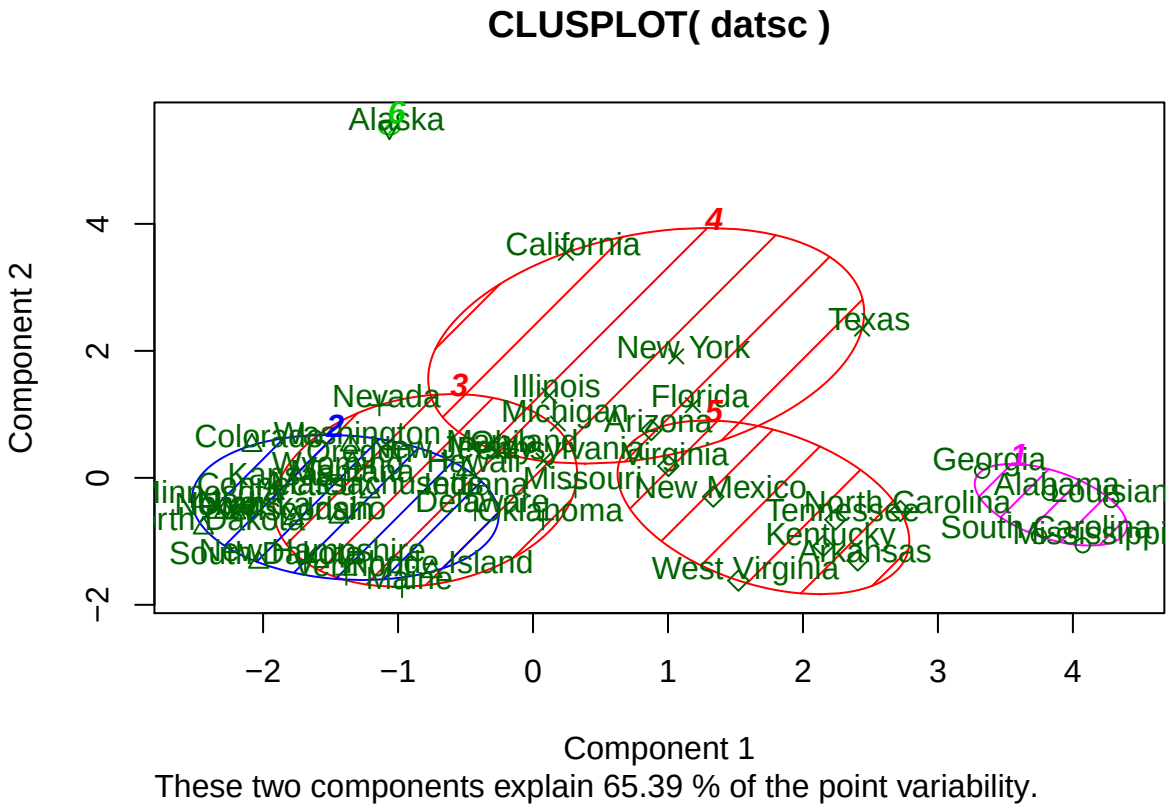
These two components explain 65.39 % of the point variability.

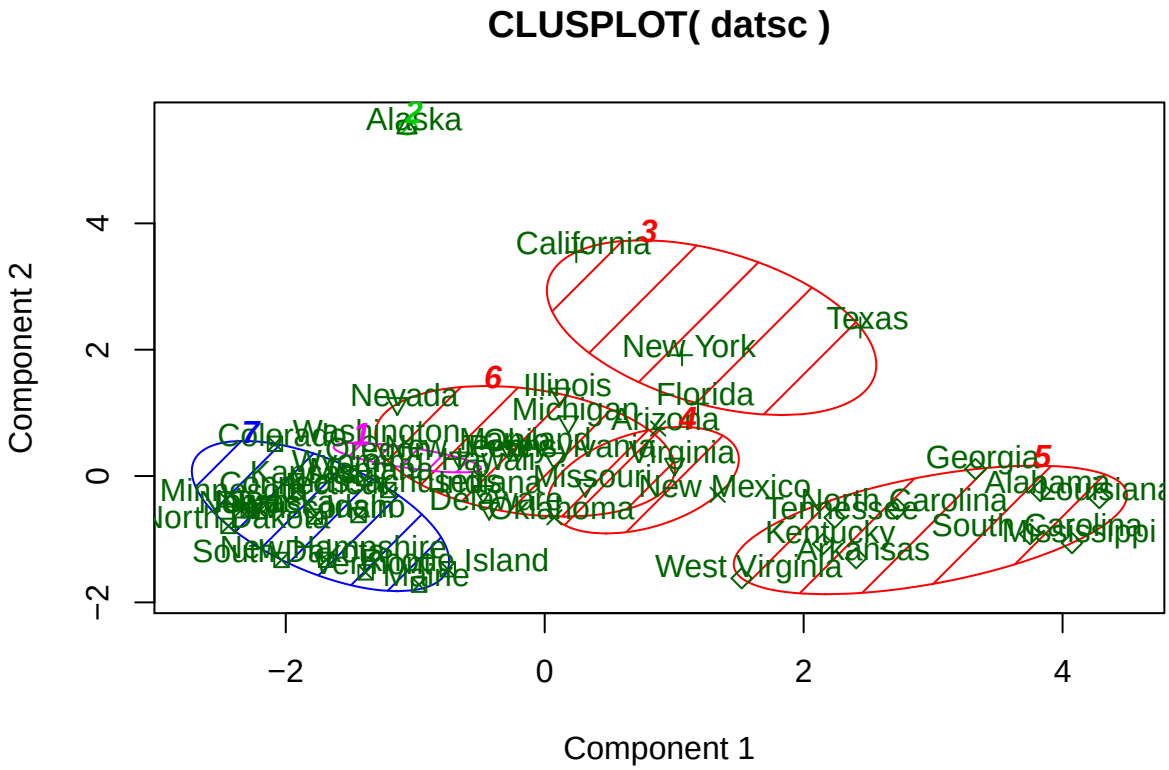


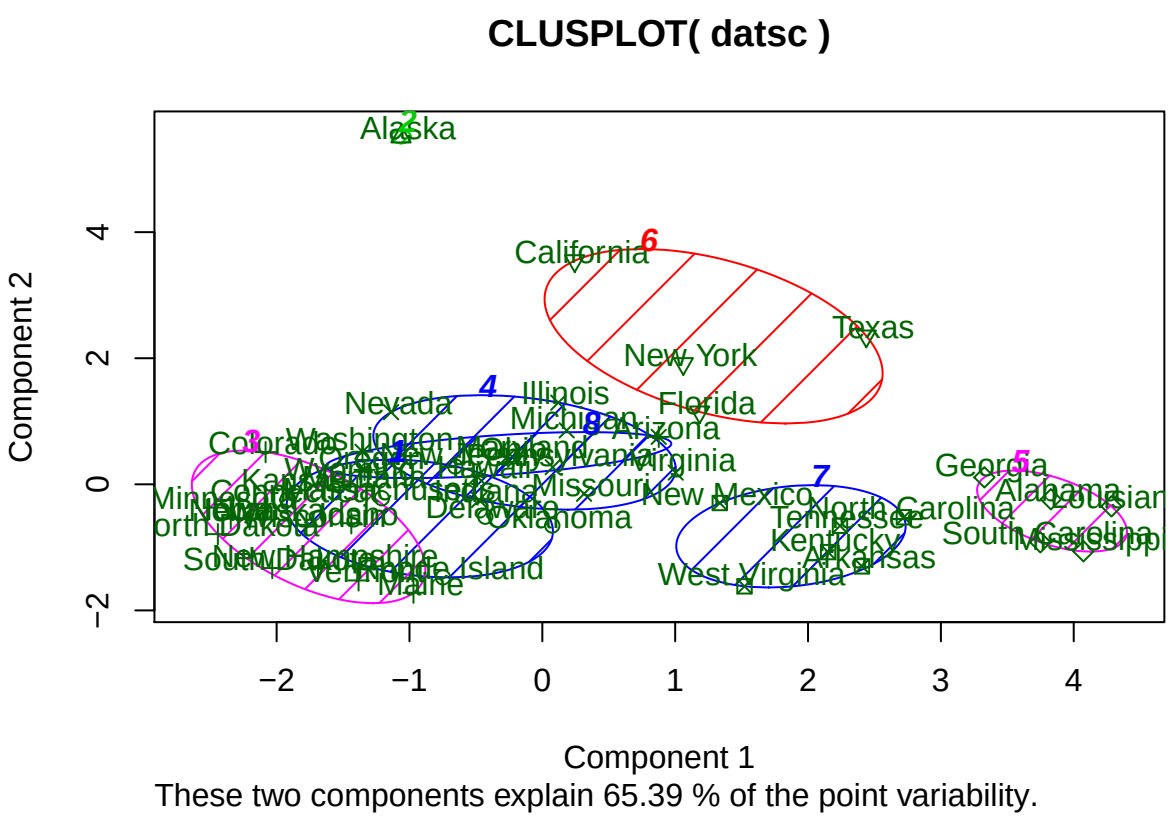
These two components explain 65.39 % of the point variability.

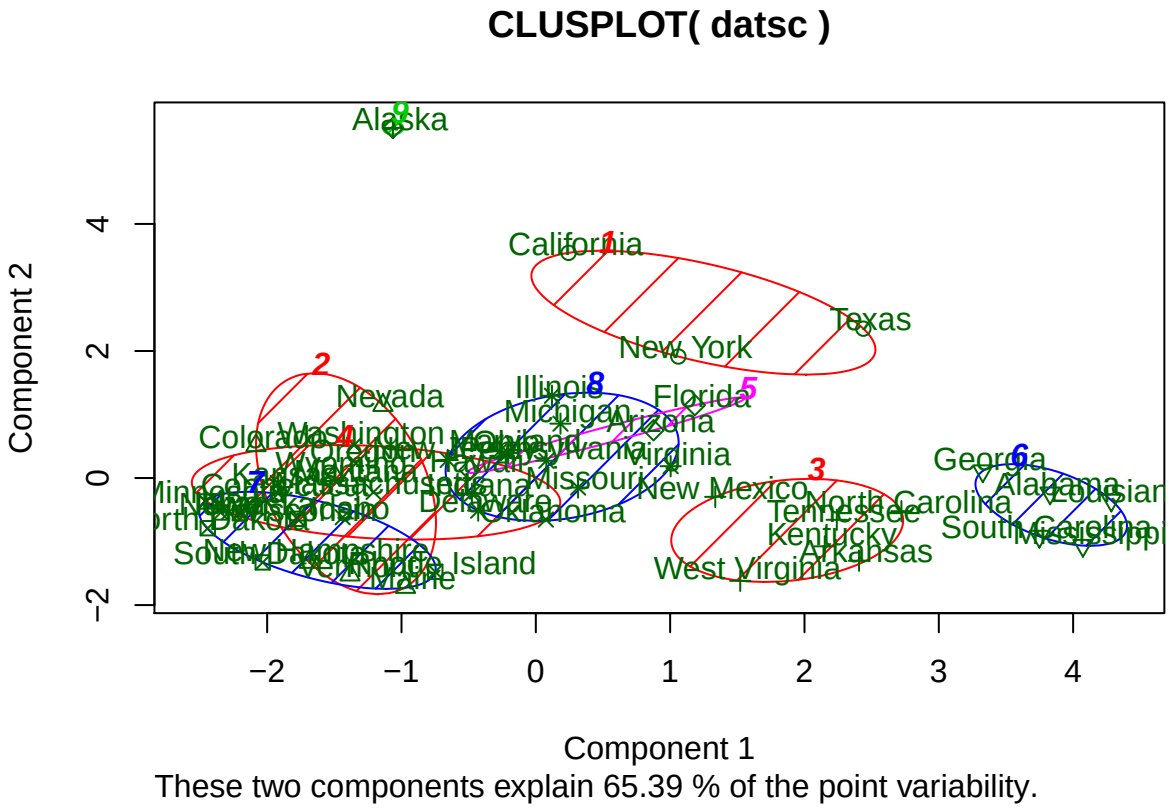


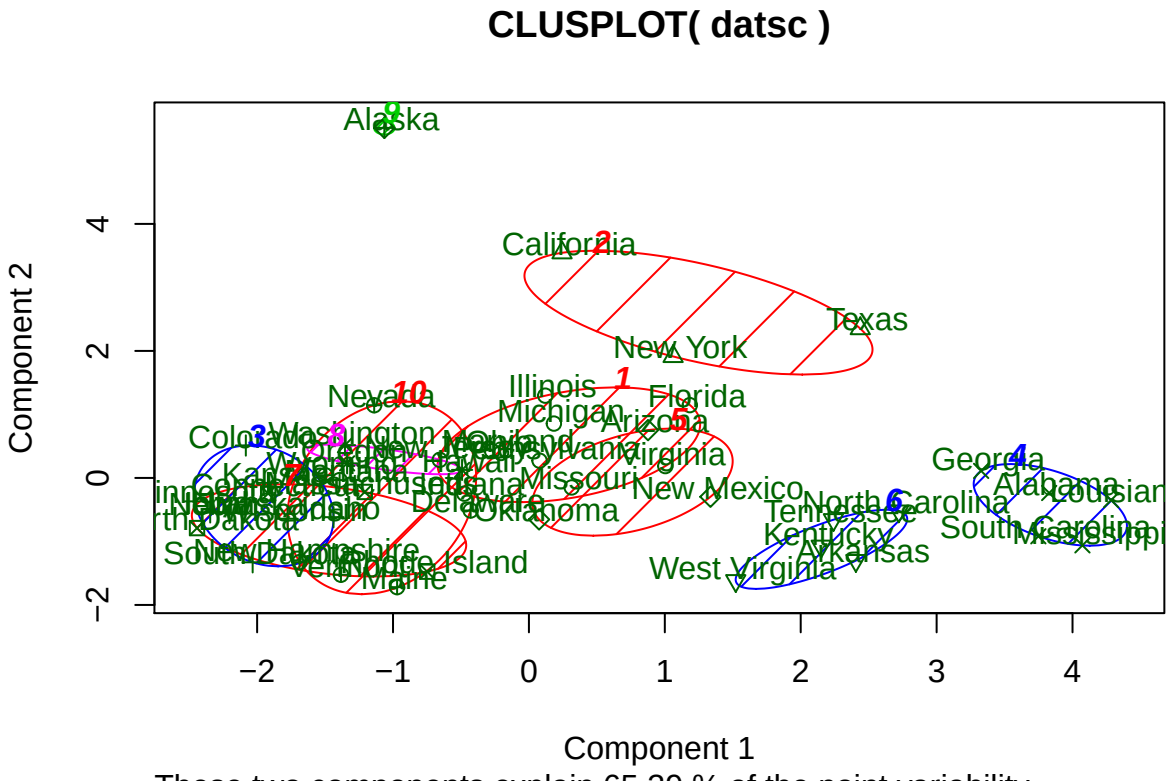
These two components explain 65.39 % of the point variability.

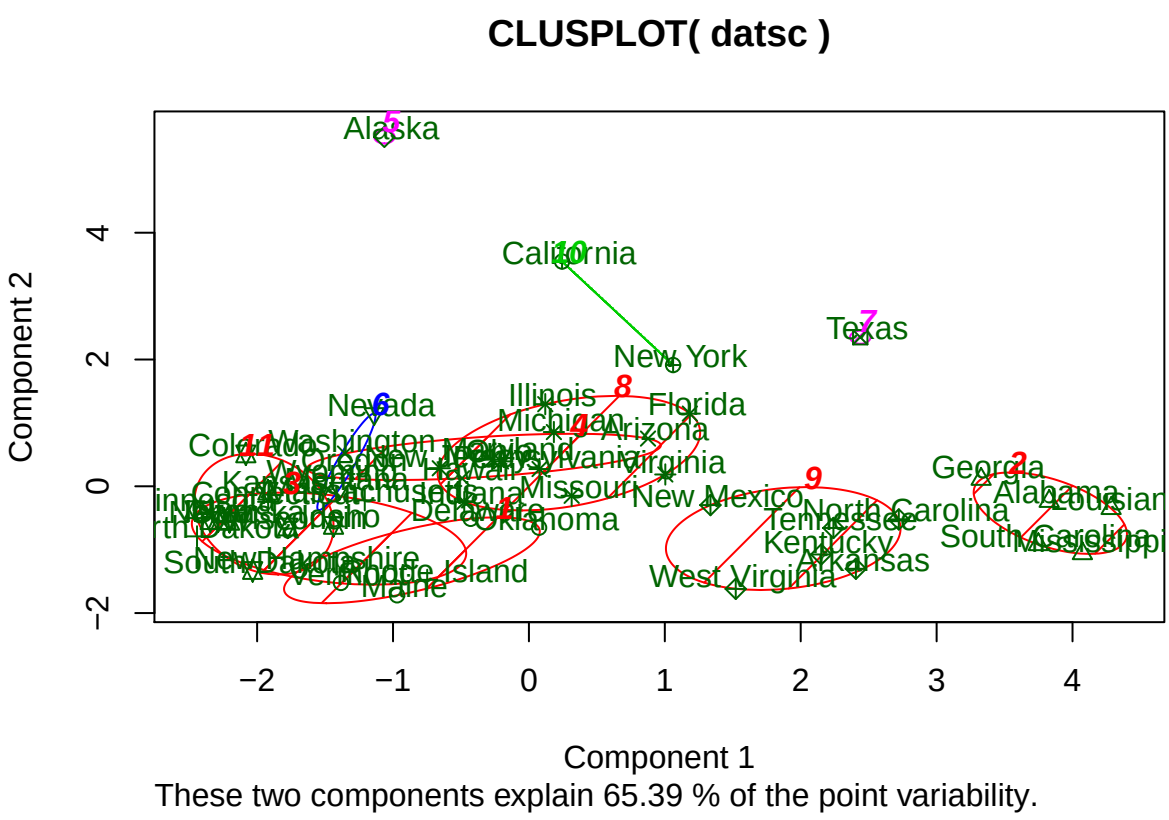

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Illinois
Florida
Nevada
Michigan
Arizona
Virginia
Georgia
Alabama
New Mexico
Tennessee
North Carolina
Kentucky
Arkansas
West Virginia
Oklahoma
Delaware
Missouri
Maine
1
2
3
4
5
6


CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Florida
Nevada
Illinois
Michigan
Arizona
Virginia
Georgia
Alabama
New Mexico
Missouri
Oklahoma
North Carolina
Tennessee
Kentucky
South Carolina
Arkansas
West Virginia
Rhode Island
Maine
1
2
3
4
5
6
7

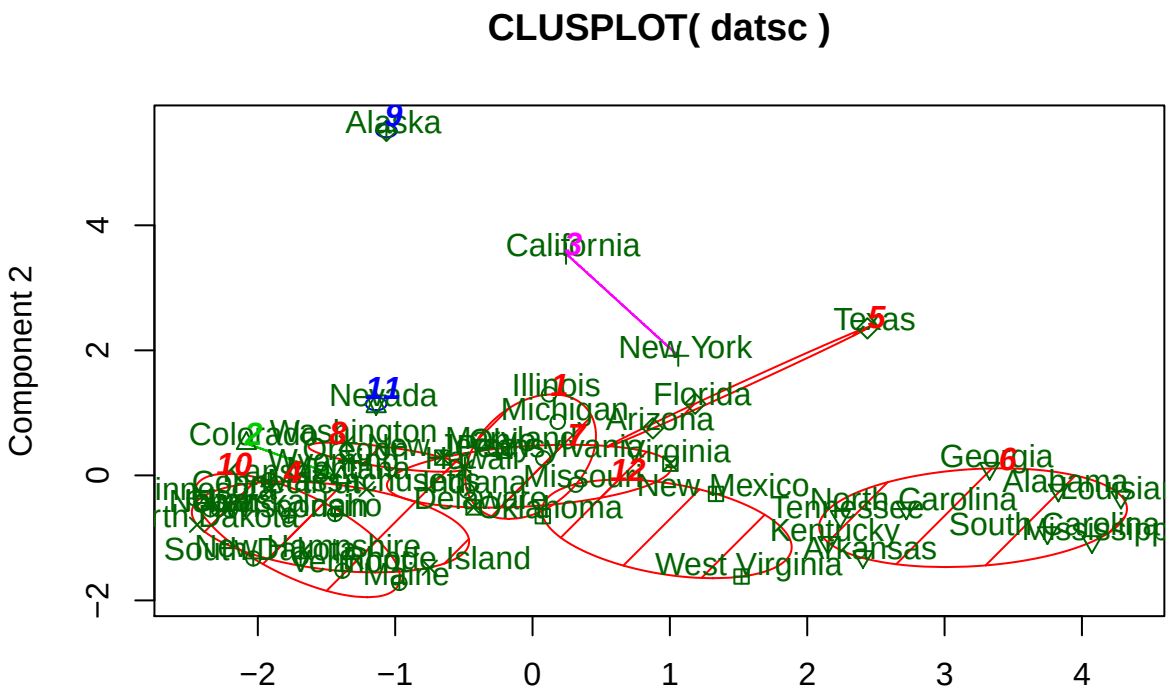

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Florida
Nevada
Illinois
Michigan
Arizona
Virginia
Missouri
New Mexico
Oklahoma
Georgia
Alabama
North Carolina
Tennessee
Kentucky
Arkansas
West Virginia
Maine
−2
−1
0
1
2
3
4

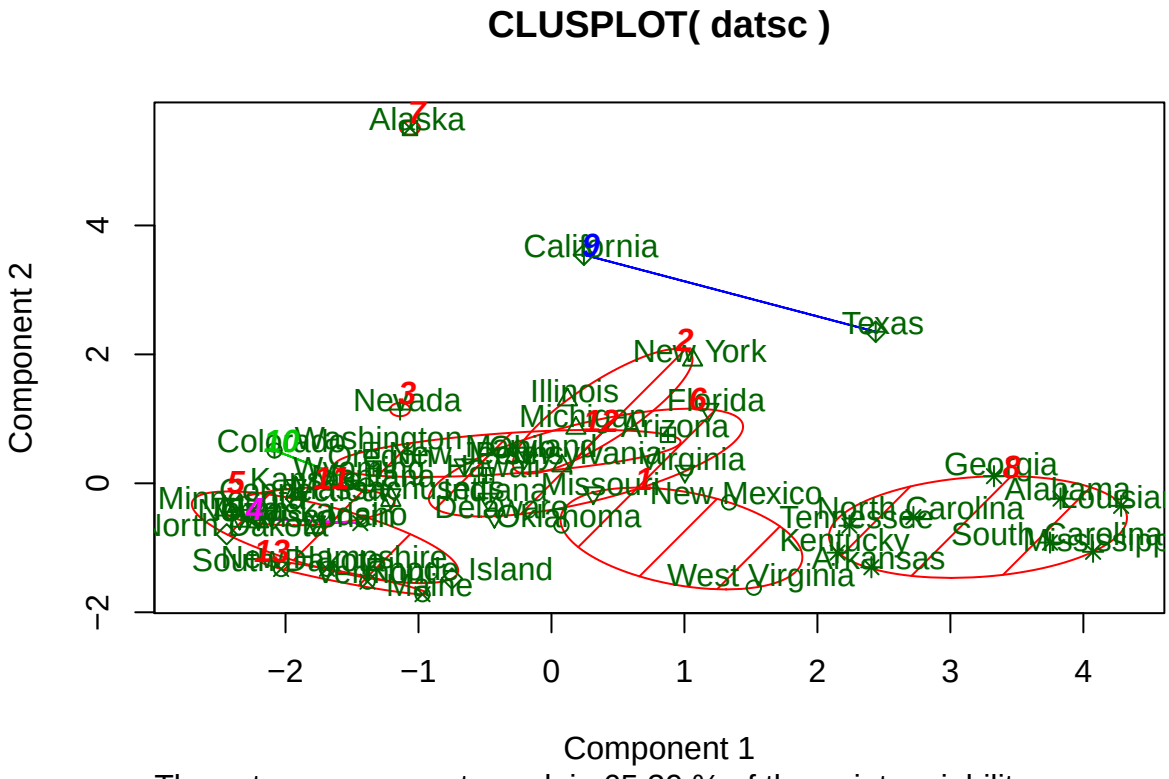

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Nevada
Florida
Arizona
Virginia
New Mexico
Georgia
Alabama
Tennessee
Kentucky
Arkansas
West Virginia
Missouri
Oklahoma
Maine
1
2
3
4
5
6
7
8


CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Illinois
Florida
Nevada
Michigan
Arizona
Virginia
Missouri
New Mexico
Georgia
Alabama
Oklahoma
Delaware
Kentucky
Arkansas
West Virginia
Maine
−2
−1
0
1
2
3
4

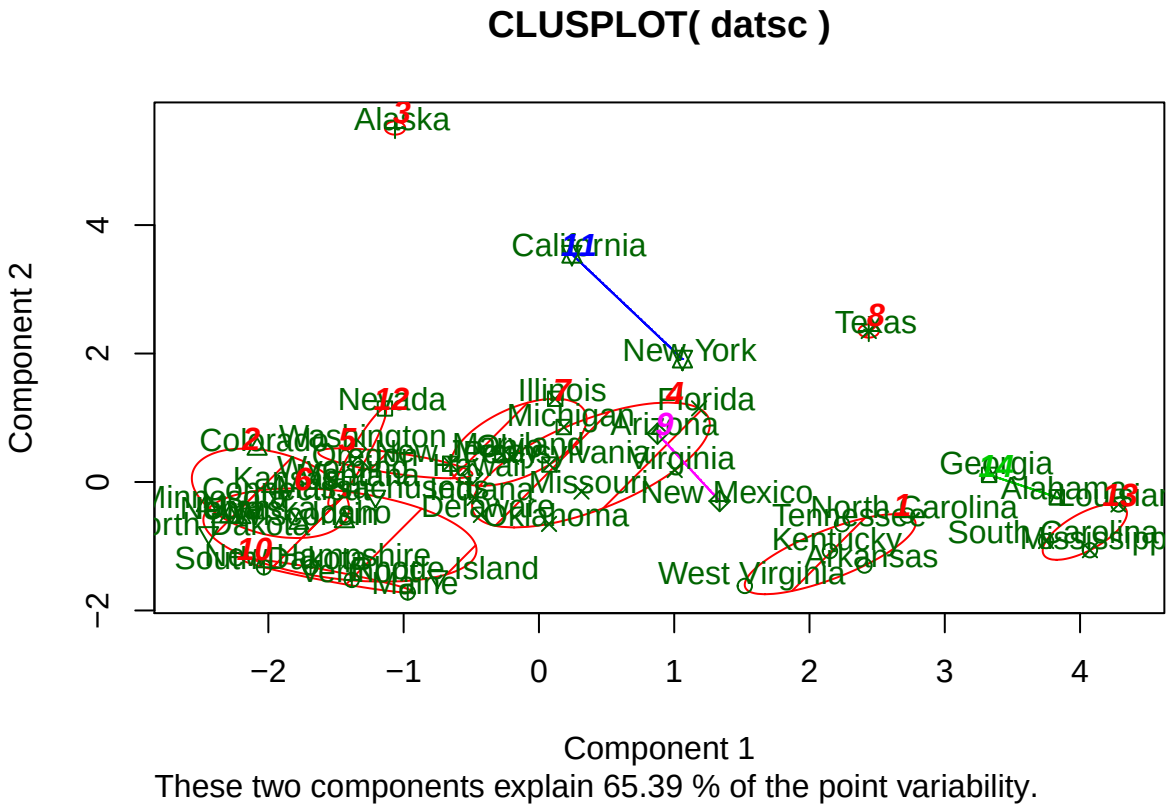


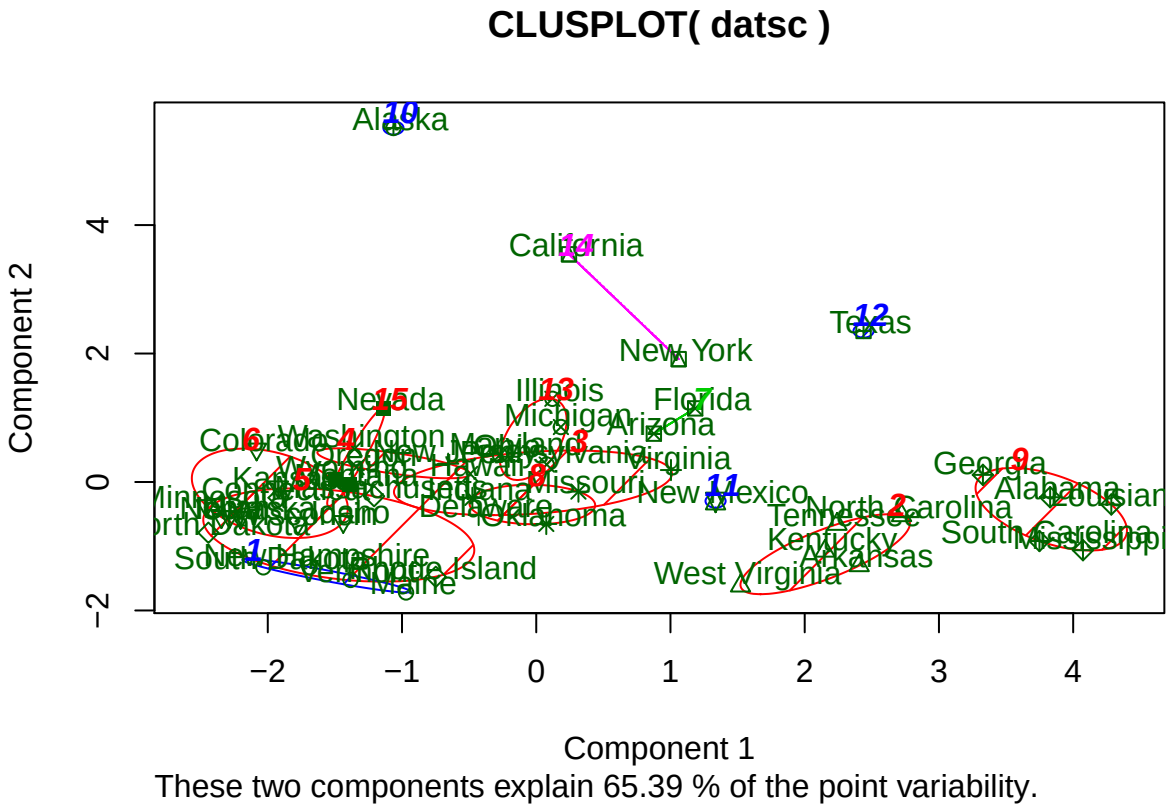
These two components explain 65.39 % of the point variability.


CLUSPLOT(datsc)
Alaska
California
New York
Texas
Nevada
Illinois
Michigan
Florida
Arizona
Georgia
Alabama
New Mexico
North Carolina
Tennessee
Kentucky
South Carolina
Arkansas
West Virginia
Maine
Component 2
Component 1
These two components explain 65.39 % of the point variability.

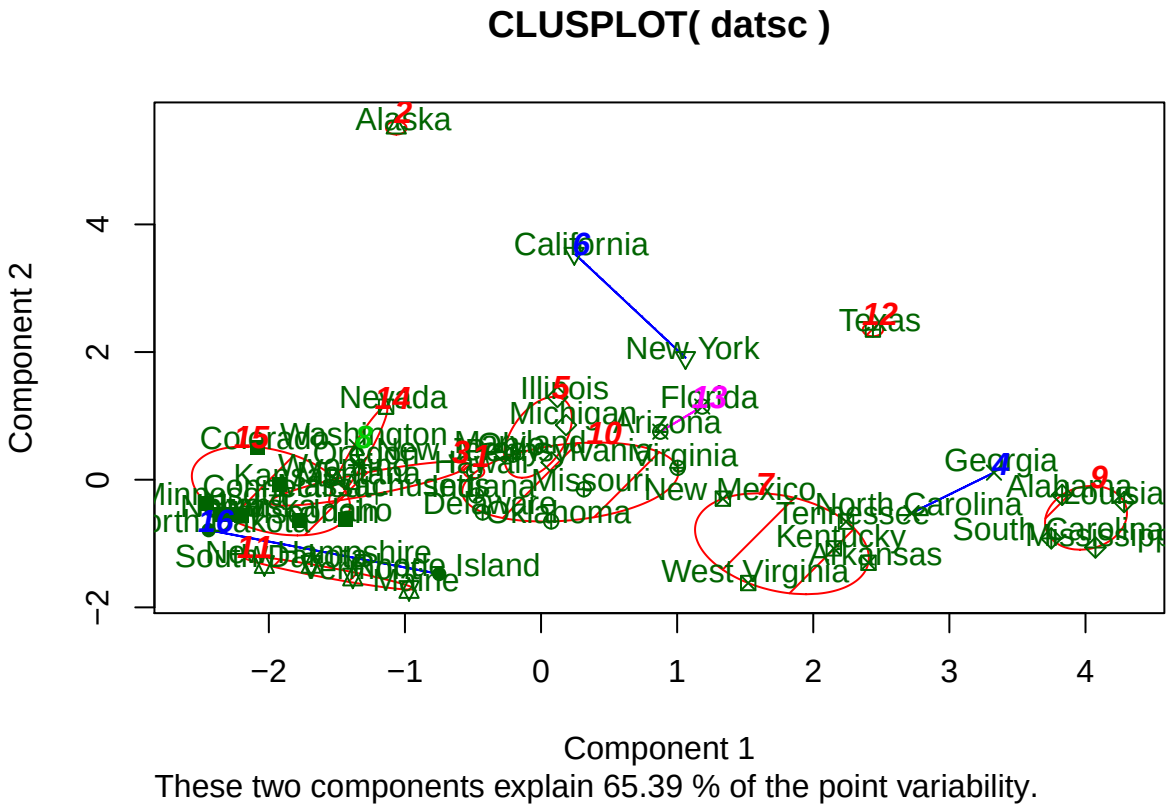


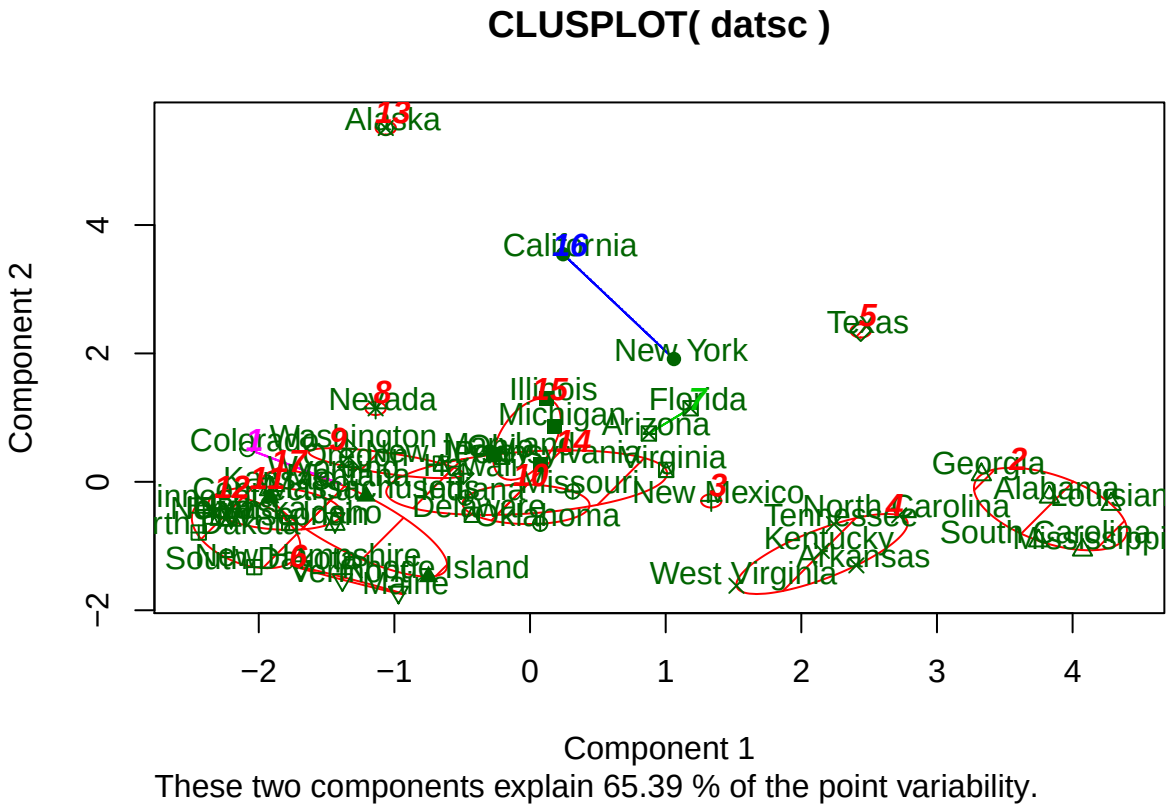
These two components explain 65.39 % of the point variability.


CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Nevada
Illinois
Florida
Michigan
Arizona
Virginia
Missouri
New Mexico
North Carolina
Tennessee
Kentucky
Arkansas
West Virginia
South Carolina
Georgia
Alabama
Louisiana
Mississippi
Colorado
Washington
Oklahoma
Delaware
Rhode Island
Maine
1
2
3
4
5
6
7
8
9
10
11
12
13
14

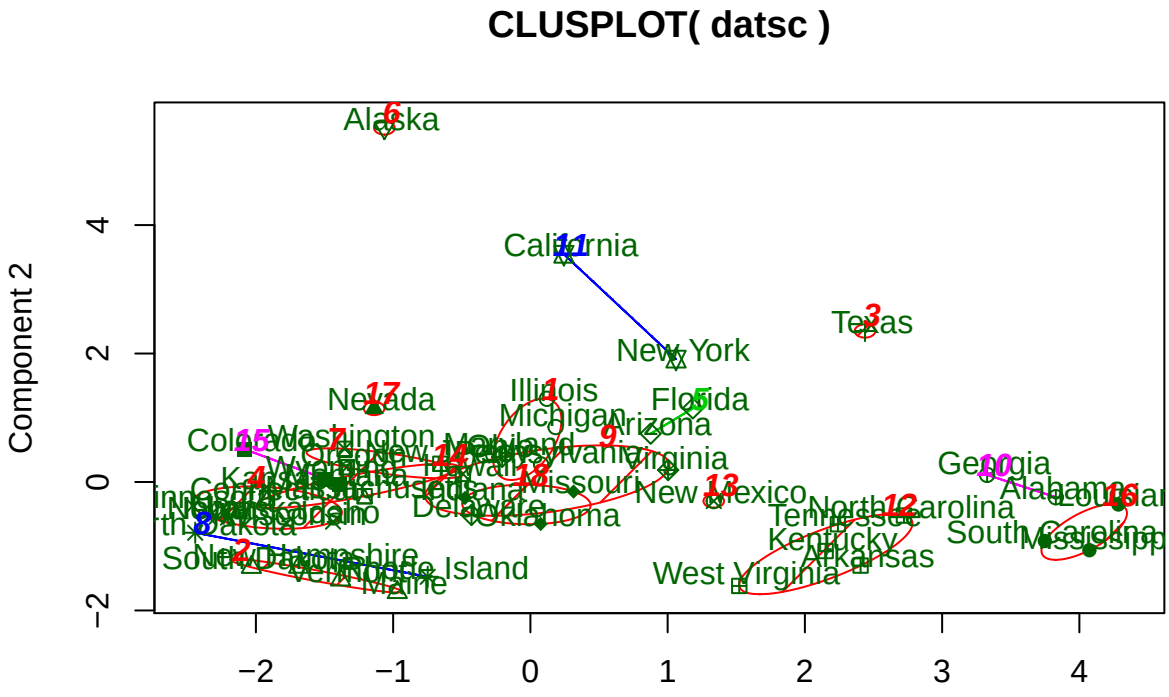


These two components explain 65.39 % of the point variability.

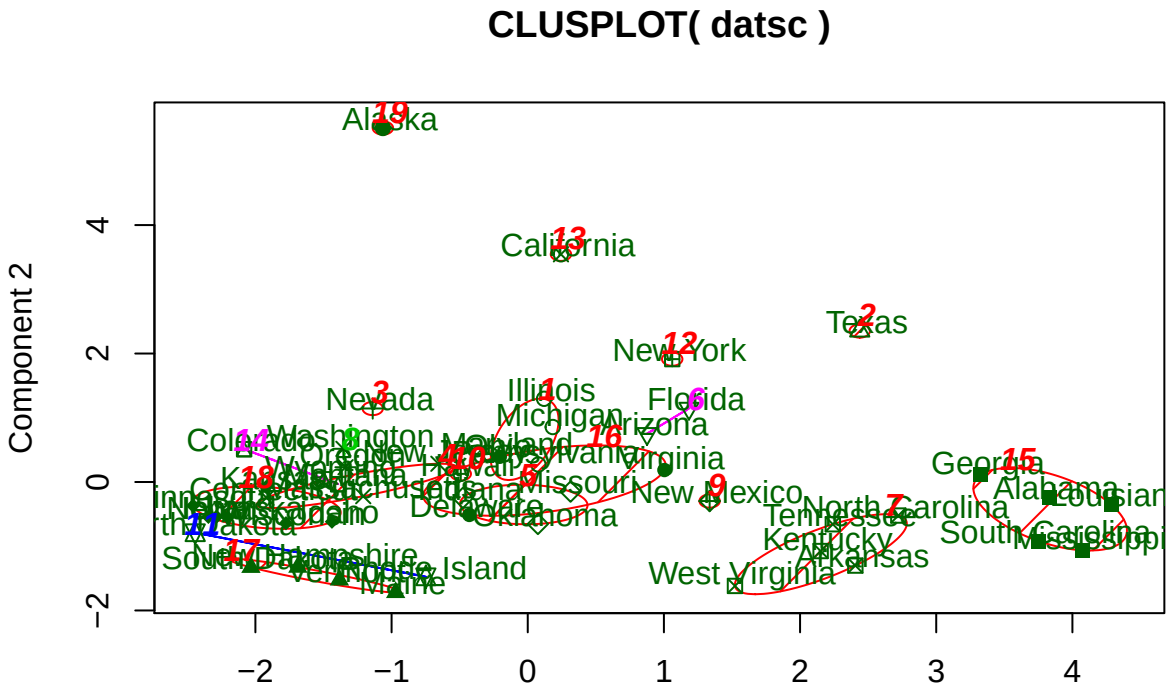

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
New York
Texas
Nevada
Illinois
Michigan
Florida
Arizona
Georgia
Alabama
Missouri
New Mexico
Tennessee
Kentucky
Arkansas
West Virginia
Delaware
Oklahoma
Rhode Island
Maine

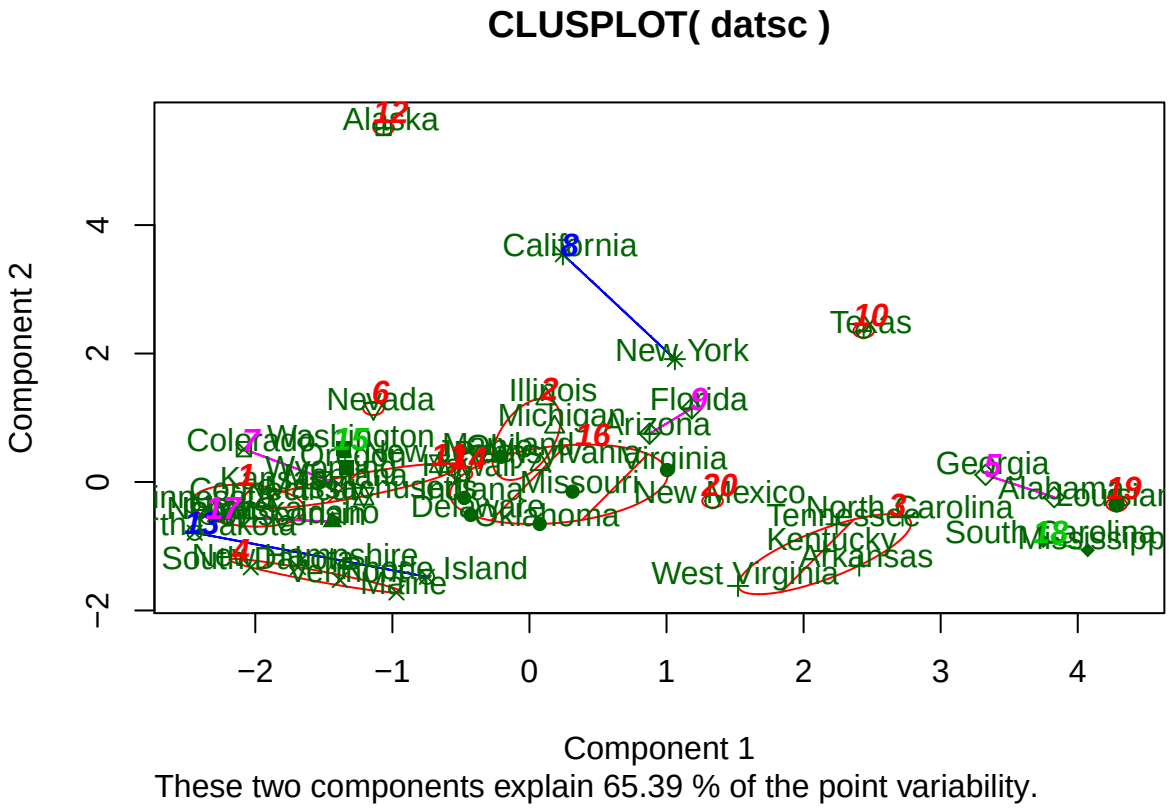


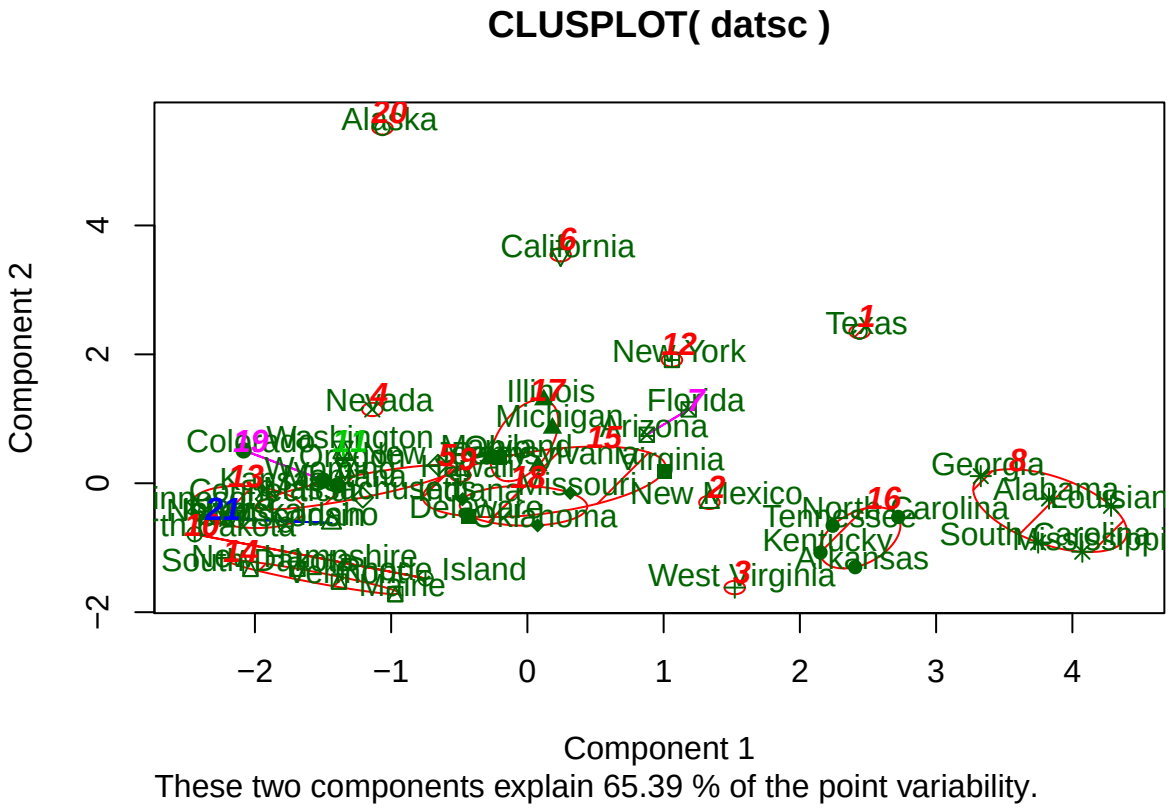
These two components explain 65.39 % of the point variability.



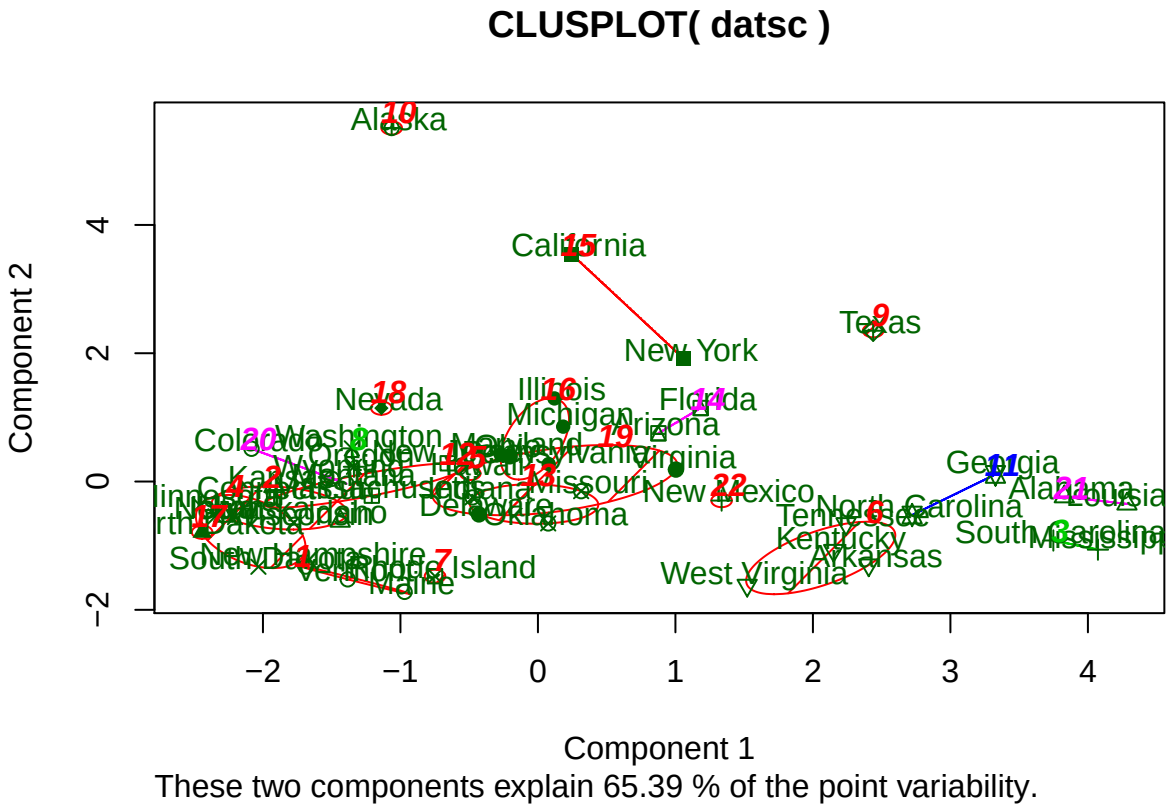
These two components explain 65.39 % of the point variability.

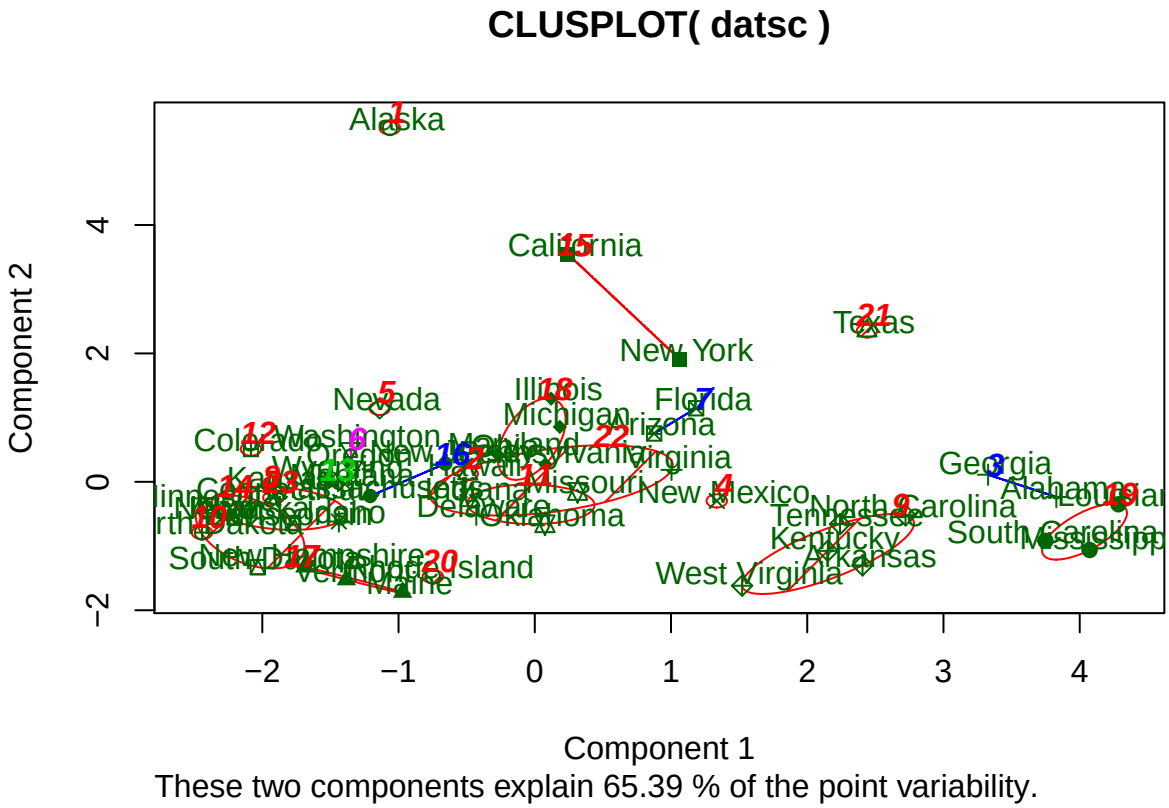

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
Texas
New York
Nevada
Illinois
Michigan
Florida
Arizona
Virginia
New Mexico
Georgia
Alabama
North Carolina
Tennessee
Kentucky
Arkansas
West Virginia
Oklahoma
Delaware
Missouri
Maine
Rhode Island

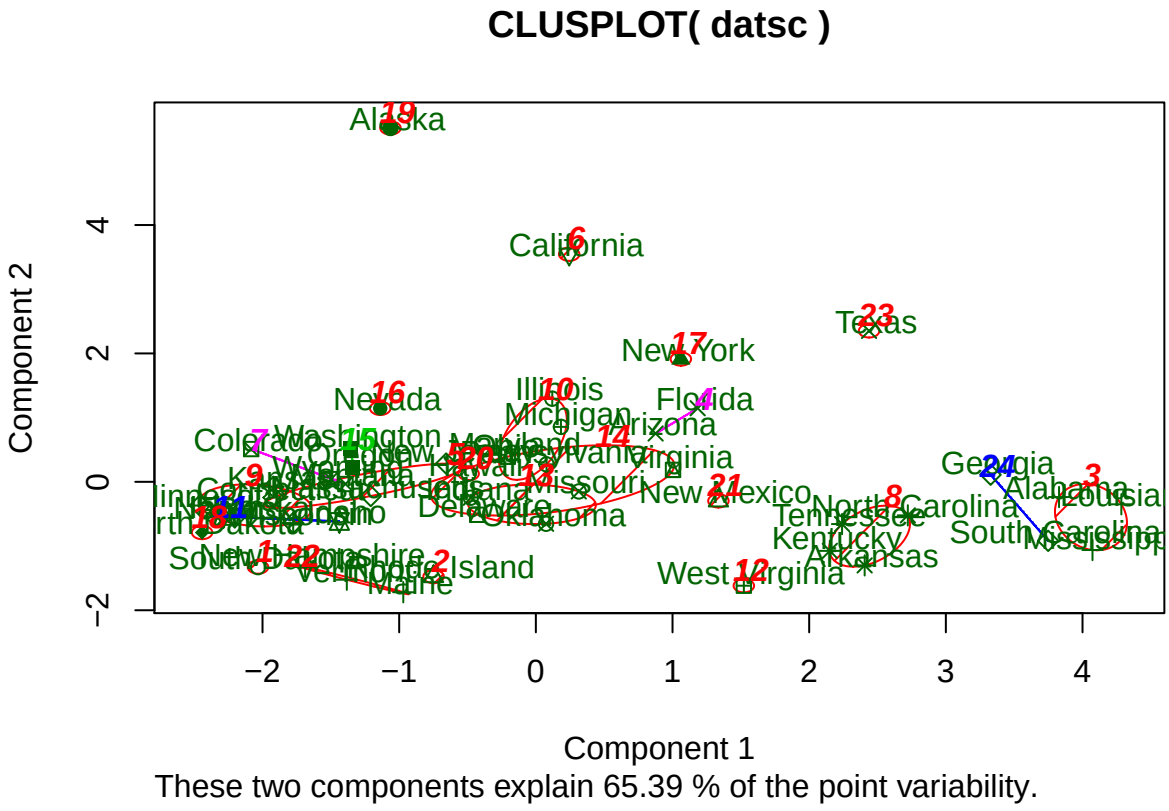

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
New York
Texas
Nevada
Florida
Illinois
Michigan
Arizona
Virginia
Missouri
New Mexico
Delaware
Oklahoma
Georgia
Alabama
North Carolina
Tennessee
Kentucky
Arkansas
South Carolina
Mississippi
West Virginia
Maine
Rhode Island
Colorado
Washington
−2
−1
0
1
2
3
4

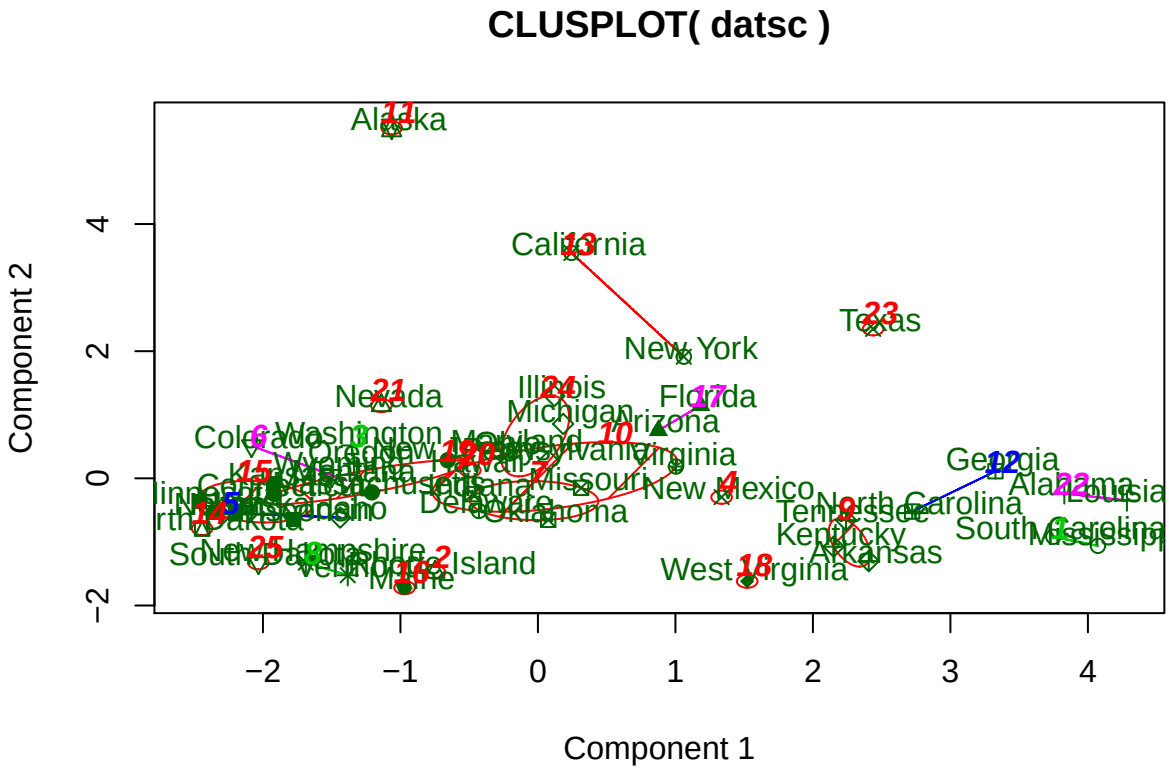


These two components explain 65.39 % of the point variability.

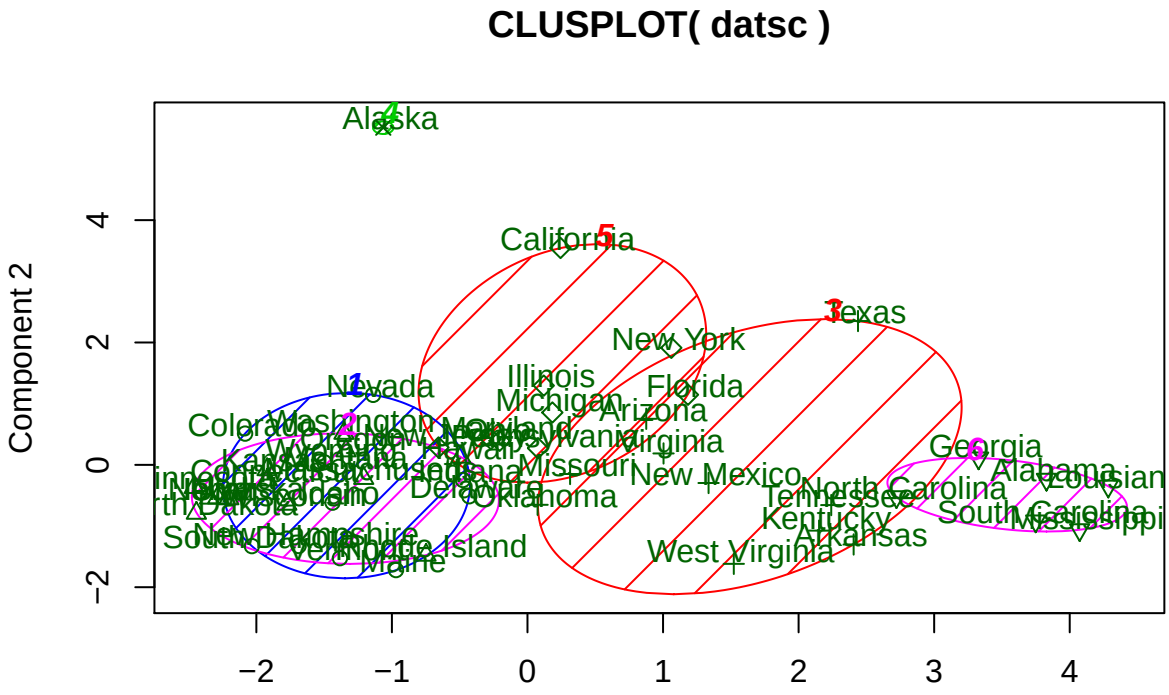

CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
New York
Texas
Nevada
Illinois
Michigan
Florida
Arizona
Virginia
Georgia
Alabama
New Mexico
North Carolina
South Carolina
Tennessee
Kentucky
Arkansas
West Virginia
Oklahoma
Delaware
Missouri
Washington
Colorado
New Hampshire
Maine
Rhode Island
South Dakota
−2
−1
0
1
2
3
4


CLUSPLOT(datsc)
Component 2
Component 1
These two components explain 65.39 % of the point variability.
Alaska
California
New York
Texas
Nevada
Illinois
Florida
Michigan
Arizona
Virginia
Georgia
Alabama
Louisiana
New Mexico
North Carolina
South Carolina
Tennessee
Kentucky
Arkansas
West Virginia
Missouri
Oklahoma
Delaware
Maine
Rhode Island
Colorado
Washington
−2
−1
0
1
2
3
4


CLUSPLOT(datsc)
Component 2
Component 1
−2
−1
0
1
2
3
4
Alaska
California
Texas
New York
Nevada
Illinois
Michigan
Florida
Arizona
Colorado
Washington
Georgia
Alabama
Missouri
New Mexico
North Carolina
Tennessee
Kentucky
Arkansas
South Carolina
Oklahoma
Delaware
West Virginia
Rhode Island
Maine
These two components explain 65.39 % of the point variability.




```
#I like 6

clus6 <- kmeans(datsc, 6)
clusplot(datsc, clus6$cluster, color = TRUE, shade = TRUE, labels = 2, lines =0)
```




```
pander(clus6$cluster)
```

Table 1: Table continues below

Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut
6	4	3	3	5	1	2

Table 2: Table continues below

Delaware	Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa
1	5	6	2	1	5	1	2

Table 3: Table continues below

Kansas	Kentucky	Louisiana	Maine	Maryland	Massachusetts	Michigan
2	3	6	1	5	2	5

Table 4: Table continues below

Minnesota	Mississippi	Missouri	Montana	Nebraska	Nevada
2	6	3	1	2	1

Table 5: Table continues below

New Hampshire	New Jersey	New Mexico	New York	North Carolina
1	5	3	5	6

Table 6: Table continues below

North Dakota	Ohio	Oklahoma	Oregon	Pennsylvania	Rhode Island
2	5	3	2	5	2

Table 7: Table continues below

South Carolina	South Dakota	Tennessee	Texas	Utah	Vermont	Virginia
6	1	3	3	2	1	3

Washington	West Virginia	Wisconsin	Wyoming
2	3	2	1

```
#New york texas and cali are clearly grouped
#due to poulation. Alaska is alone due to area,
#A lot of the middle sized states are grouoed
#together in the bottom right, and these ultr
#small states are all in the bottom left,
```