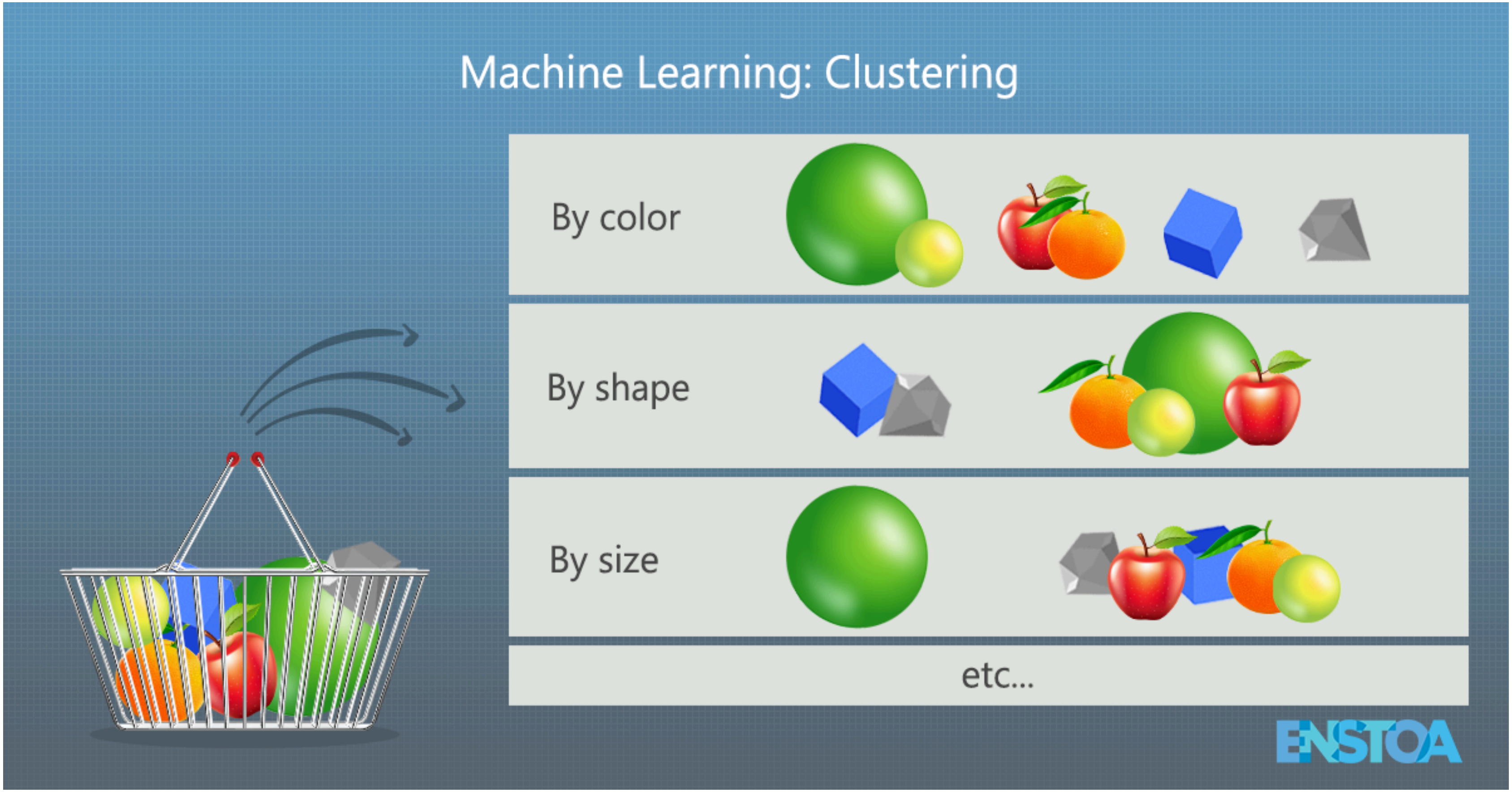
# K-means Clustering

Son Nguyen

# What is clustering?

Clustering is grouping data points into groups where data points in one group are `similar` to each other.

# What is clustering?



# Methods of Clustering

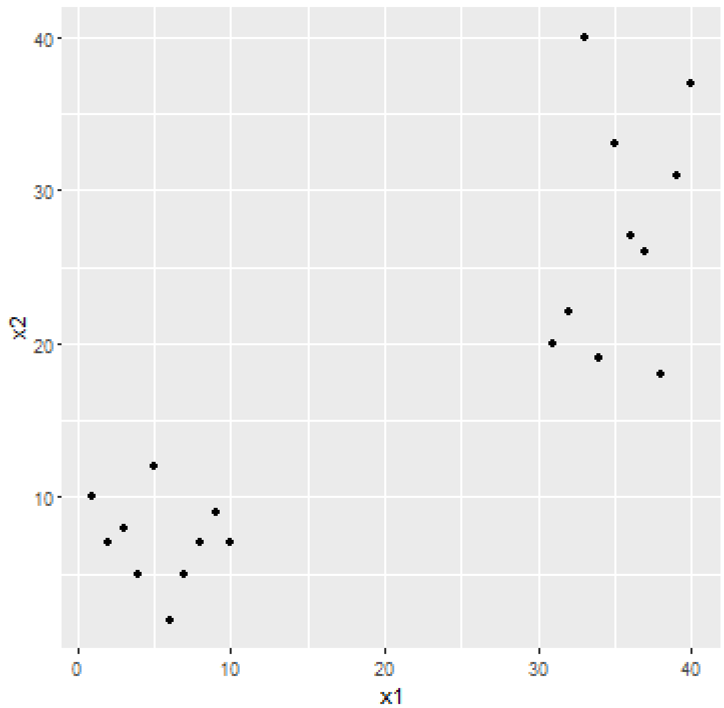
We will cover two clustering methods:

- K-means clustering and
- Hierarchical clustering

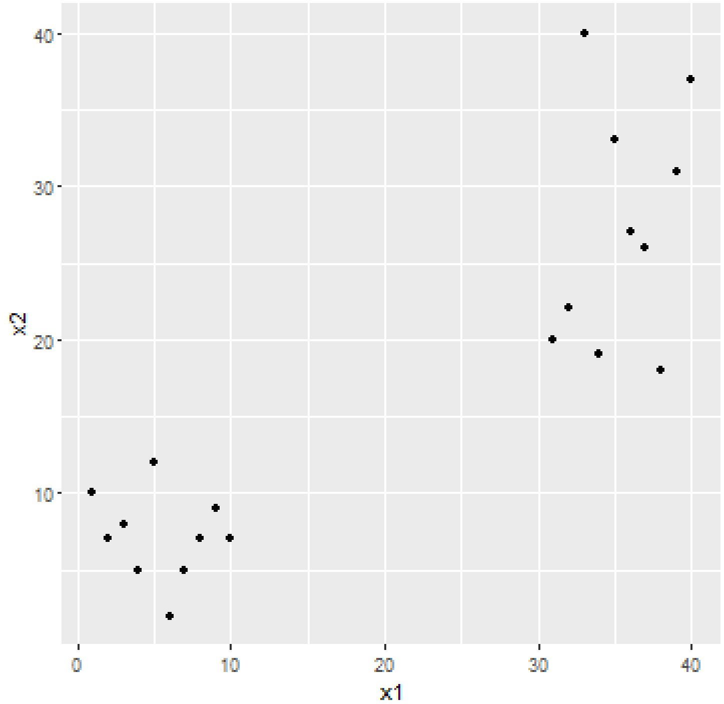
# K-means Clustering

- Data
- Visualize Data
- Result of K-means clustering

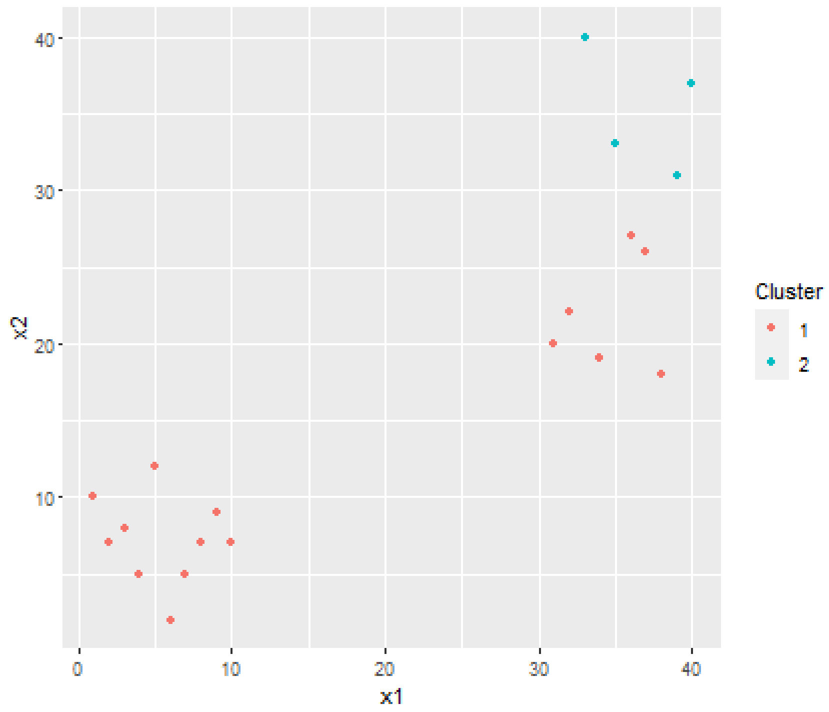
# Step 1



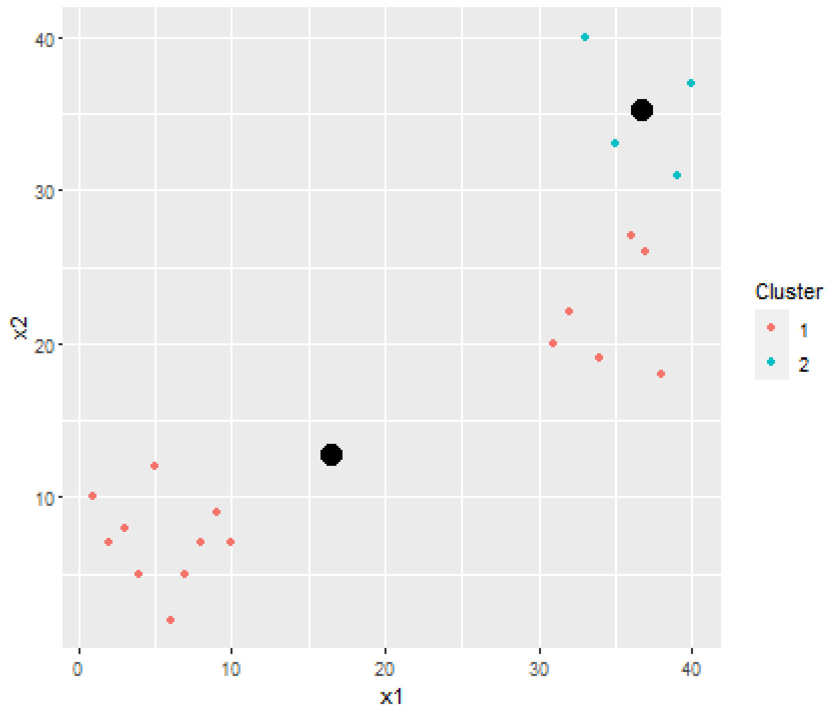
# Step 1: Randomly select centroids



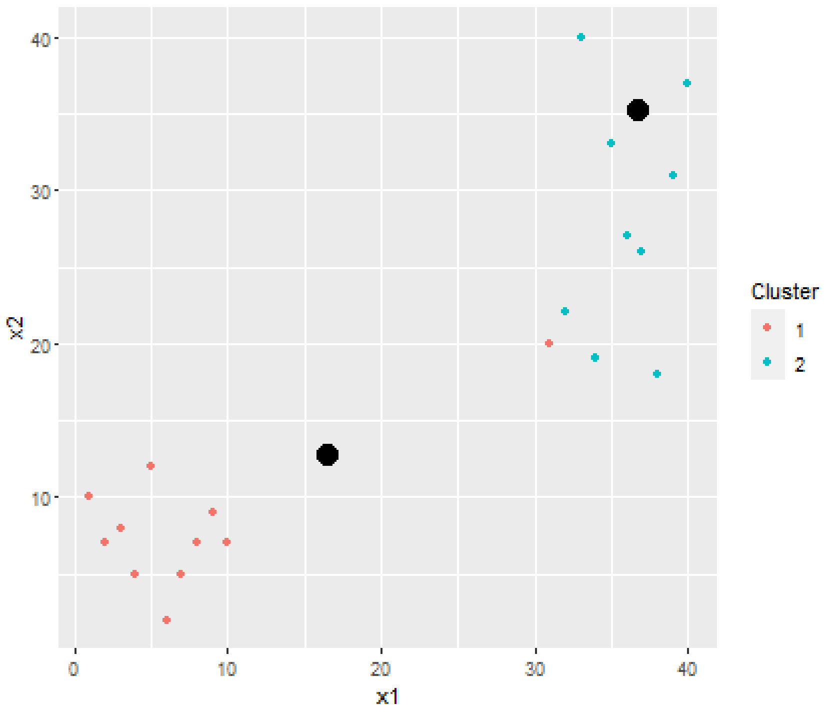
# Step 1: Collect points for each clusters



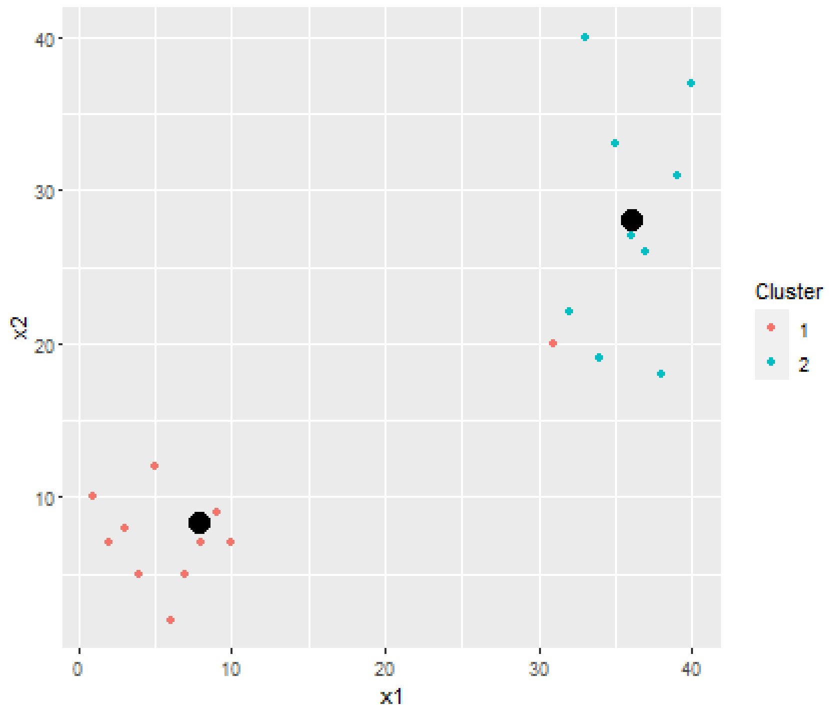
# Locate centroids



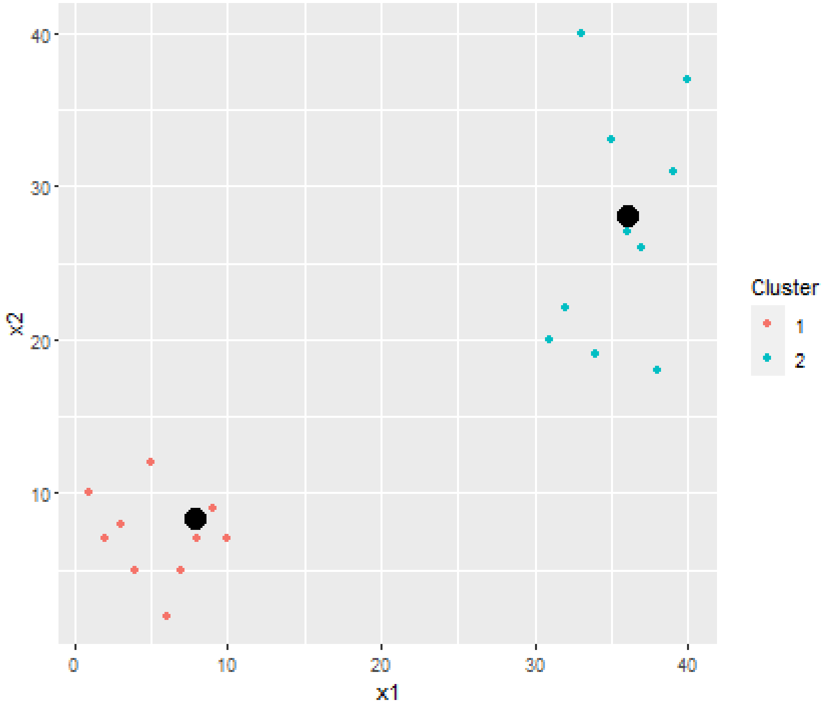
# Collect points for each clusters



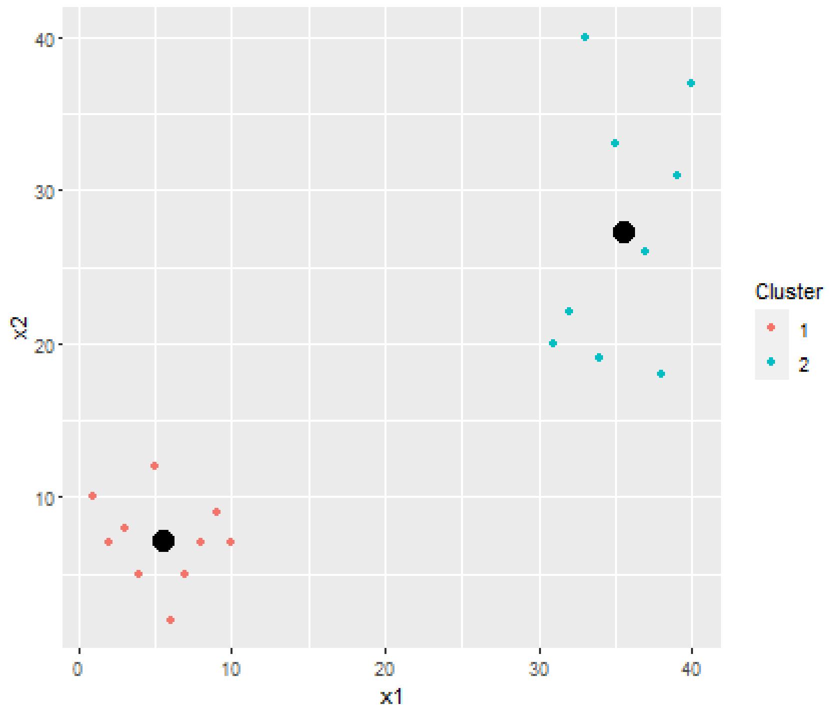
# Relocate centroids



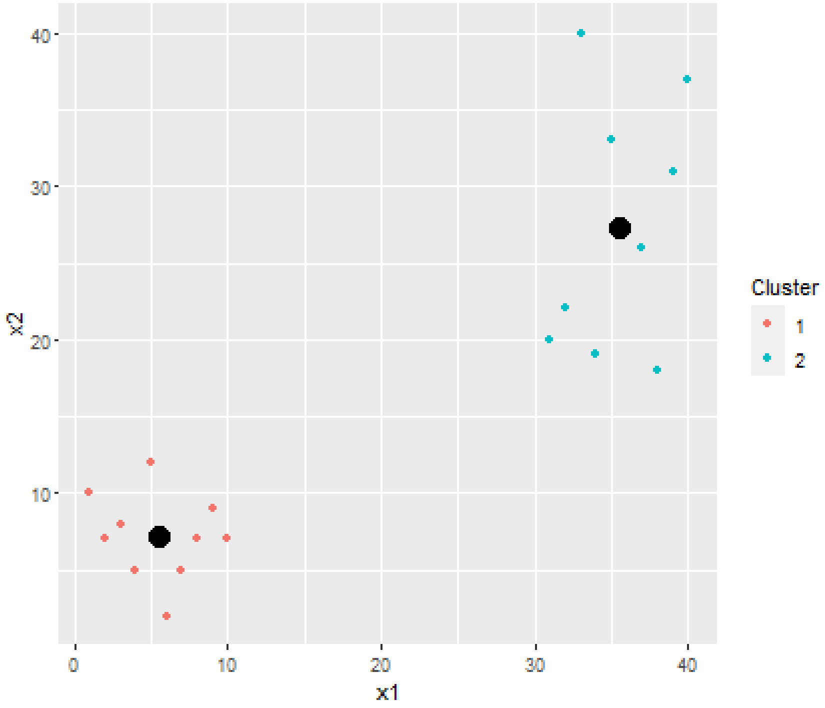
# Collect points for each clusters



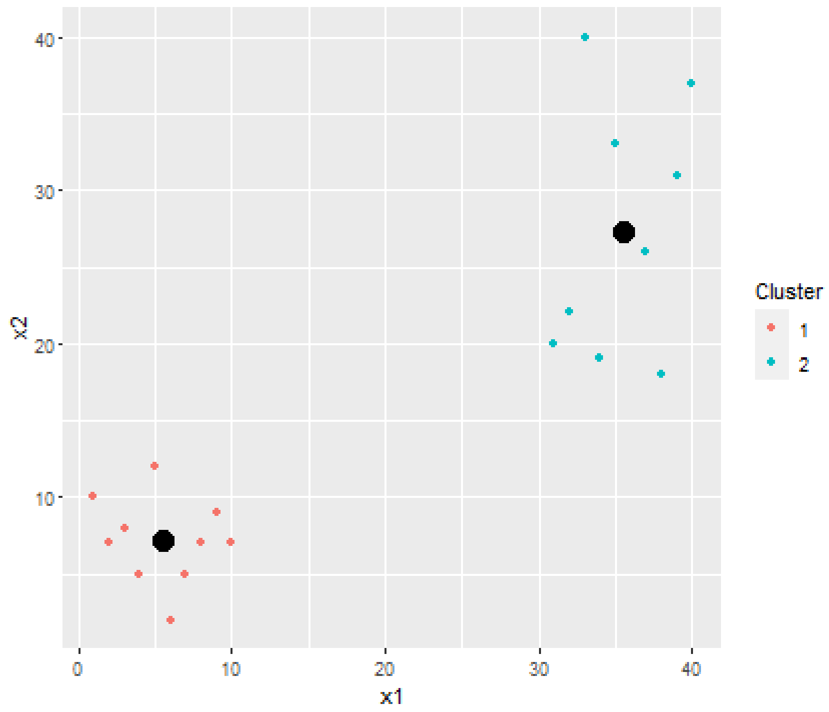
# Relocate centroids



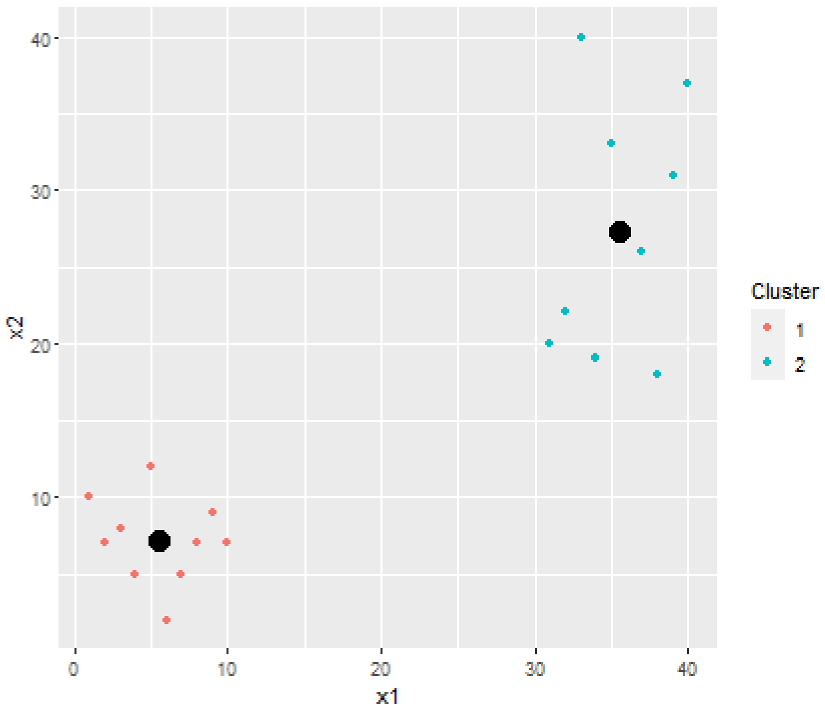
# Collect points for each clusters



# Relocate centroids

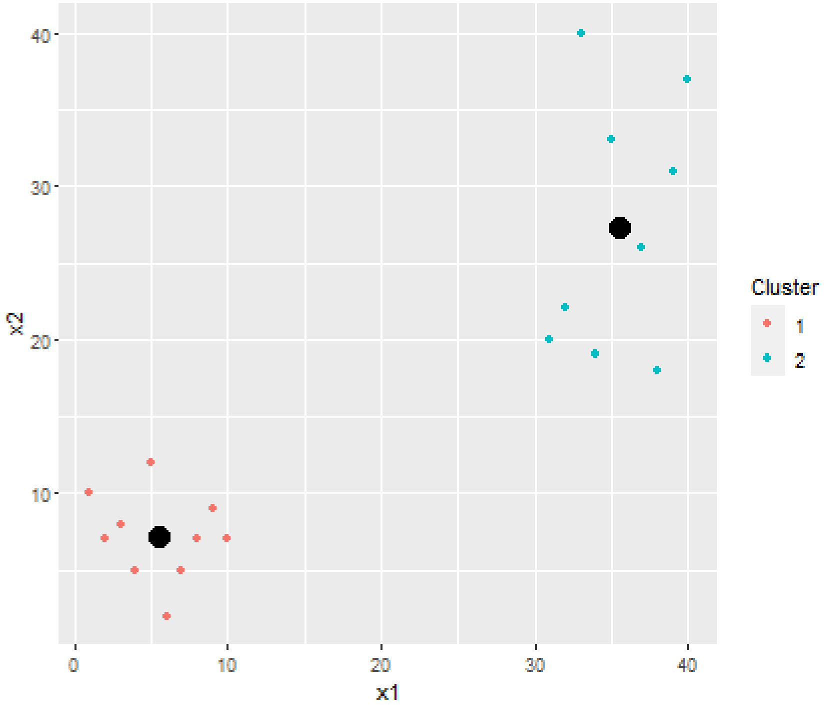


# Step 2: Collect points for each clusters

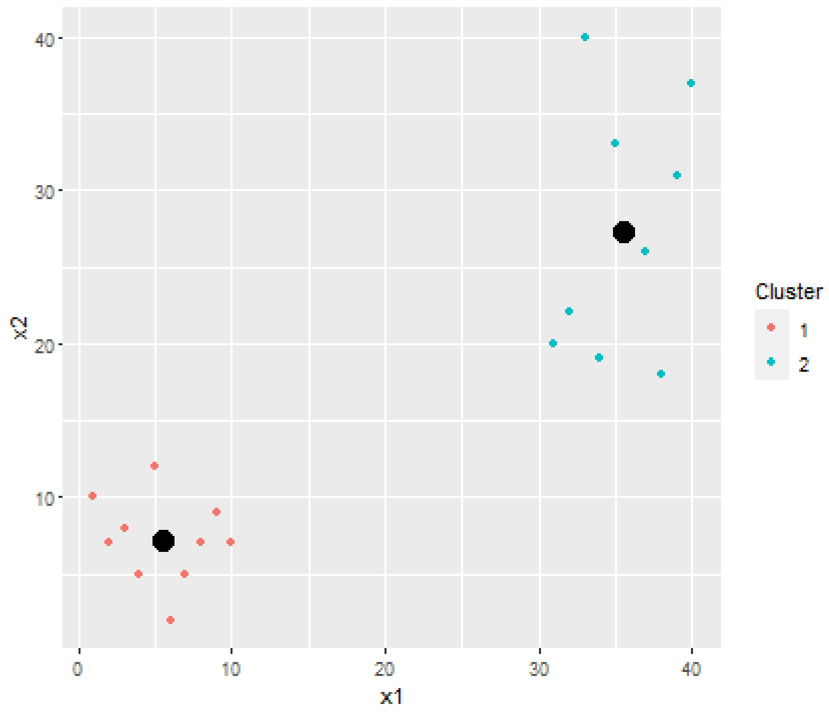


| GPA | Screen Time |
|---|---|
| 3.9 | 90 |
| 4.0 | 150 |
| 3.0 | 140 |
| 2.0 | 139 |

# Step 2: Relocate centroids



# Step 2: Collect points for each clusters



# Centroids

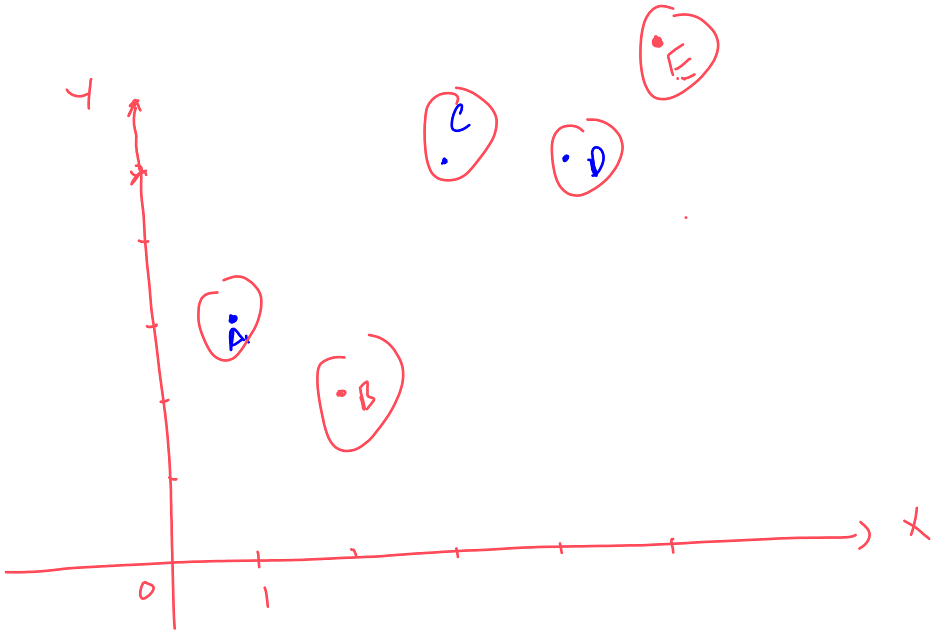
| Cluster | x1 | x2 |
|---|---|---|
| 1 | 5.5 | 7.2 |
| 2 | 35.5 | 27.3 |

# K-means Algorithm

- 1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
- 1. Iterate until the cluster assignments stop changing:
  - (a) For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
  - (b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

# Dataset

| Point | x | y |
|---|---|---|
| A | 1 | 3 |
| B | 2 | 2 |
| C | 3 | 5 |
| D | 4 | 5 |
| E | 5 | 6 |



# Randomly Assign Cluster to Points

| Cluster | Point | x | y |
|---|---|---|---|
| 1 | A | 1 | 3 |
| 2 | B | 2 | 2 |
| 1 | C | 3 | 5 |
| 1 | D | 4 | 5 |
| 2 | E | 5 | 6 |

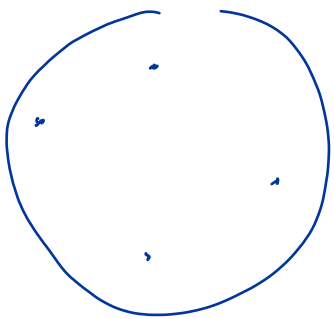
# Determine Centroids

| Cluster | Point | x | y | M_1x | M_1y | M_2x | M_2y |
|---|---|---|---|---|---|---|---|
| 1 | A | 1 | 3 | 2.67 | 4.33 | 3.5 | 4 |
| 2 | B | 2 | 2 | 2.67 | 4.33 | 3.5 | 4 |
| 1 | C | 3 | 5 | 2.67 | 4.33 | 3.5 | 4 |
| 1 | D | 4 | 5 | 2.67 | 4.33 | 3.5 | 4 |
| 2 | E | 5 | 6 | 2.67 | 4.33 | 3.5 | 4 |

- Centroid 1: $M_1 = \frac{A+C+D}{3} = \frac{(1,3)+(3,5)+(4,5)}{3} = \frac{(8,13)}{3} = (2.67, 4.33)$
- Centroid 2: $M_2 = \frac{B+E}{\underset{2}{\cancel{3}}} = \frac{(2,2)+(5,6)}{2} = \frac{(7,8)}{2} = (3.5, 4)$

# Distance to Centroids

| Cluster | Point | x | y | M_1x | M_1y | M_2x | M_2y | dc1 | dc2 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | A | 1 | 3 | 2.67 | 4.33 | 3.5 | 4 | 2.13 | 2.69 |
| 2 | B | 2 | 2 | 2.67 | 4.33 | 3.5 | 4 | 2.42 | 2.50 |
| 1 | C | 3 | 5 | 2.67 | 4.33 | 3.5 | 4 | 0.75 | 1.12 |
| 1 | D | 4 | 5 | 2.67 | 4.33 | 3.5 | 4 | 1.49 | 1.12 |
| 2 | E | 5 | 6 | 2.67 | 4.33 | 3.5 | 4 | 2.87 | 2.50 |



- $AM_1 = \sqrt{(1-2.67)^2 + (3-4.33)^2} = 2.13$ and so on
- Variance within Cluster 1 = $V_1 = AM_1^2 + CM_1^2 + DM_1^2 = 7.3195$
- Variance within Cluster 2 = $V_2 = BM_2^2 + DM_2^2 = 12.5$

- Total Variance = $V_1 + V_2 = 19.83$

# Compare Distances to Centroids

| Cluster | Point | x | y | dc1 | dc2 | min_distance |
|---|---|---|---|---|---|---|
| 1 | A | 1 | 3 | 2.13 | 2.69 | 2.13 |
| 2 | B | 2 | 2 | 2.42 | 2.50 | 2.42 |
| 1 | C | 3 | 5 | 0.75 | 1.12 | 0.75 |
| 1 | D | 4 | 5 | 1.49 | 1.12 | 1.12 |
| 2 | E | 5 | 6 | 2.87 | 2.50 | 2.50 |

# Reassign Clusters

| Cluster | Point | x | y | dc1 | dc2 | min_distance | New_Cluster |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | A | 1 | 3 | 2.13 | 2.69 | 2.13 | 1 |
| 2 | B | 2 | 2 | 2.42 | 2.50 | 2.42 | 1 |
| 1 | C | 3 | 5 | 0.75 | 1.12 | 0.75 | 1 |
| 1 | D | 4 | 5 | 1.49 | 1.12 | 1.12 | 2 |
| 2 | E | 5 | 6 | 2.87 | 2.50 | 2.50 | 2 |

- New cluster 1 = {A, B, C}
- New cluster 2 = {D, E}
- Total Variance = $AN_1^2 + BN_1^2 + CN_1^2 + DN_2^2 + EN_2^2 = 7.67$

# Reassign Clusters

| Cluster | Point | x | y | dc1 | dc2 | min_distance | New_Cluster |
|---|---|---|---|---|---|---|---|
| 1 | A | 1 | 3 | 2.13 | 2.69 | 2.13 | 1 |
| 2 | B | 2 | 2 | 2.42 | 2.50 | 2.42 | 1 |
| 1 | C | 3 | 5 | 0.75 | 1.12 | 0.75 | 1 |
| 1 | D | 4 | 5 | 1.49 | 1.12 | 1.12 | 2 |
| 2 | E | 5 | 6 | 2.87 | 2.50 | 2.50 | 2 |

- New cluster 1 = {A, B, C}
- New cluster 2 = {D, E}
- Total Variance = $AN_1^2 + BN_1^2 + CN_1^2 + DN_2^2 + EN_2^2 = 7.67$
- The process continues until there is no change in the total variance
- The total variance will be reduced to its minimum.

# Step 1: Total Variance within

| Cluster | Point | x | y |
|---|---|---|---|
| 1 | A | 1 | 3 |
| 2 | B | 2 | 2 |
| 1 | C | 3 | 5 |
| 1 | D | 4 | 5 |
| 2 | E | 5 | 6 |

- Total Variance within 19.83

# Step 2: Total Variance within

| New_Cluster | Point | x | y |
|---|---|---|---|
| 1 | A | 1 | 3 |
| 1 | B | 2 | 2 |
| 1 | C | 3 | 5 |
| 2 | D | 4 | 5 |
| 2 | E | 5 | 6 |

- Total Variance within 7.67

# Step 2: Total Variance within

| New_Cluster | Point | x | y |
|---|---|---|---|
| 1 | A | 1 | 3 |
| 1 | B | 2 | 2 |
| 1 | C | 3 | 5 |
| 2 | D | 4 | 5 |
| 2 | E | 5 | 6 |

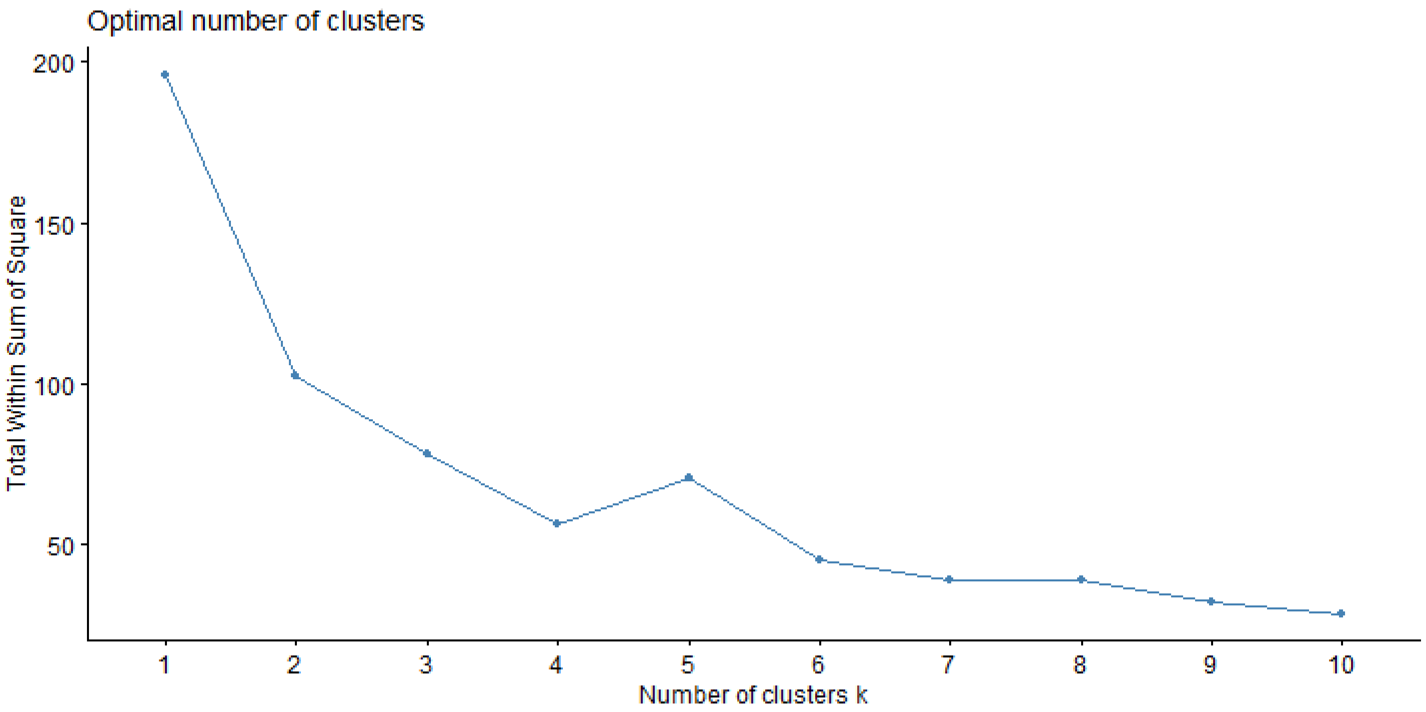
- Total Variance within 7.67
- The process continues until there is no change in the total variance
- The total variance will be reduced to its minimum.

# Terminology

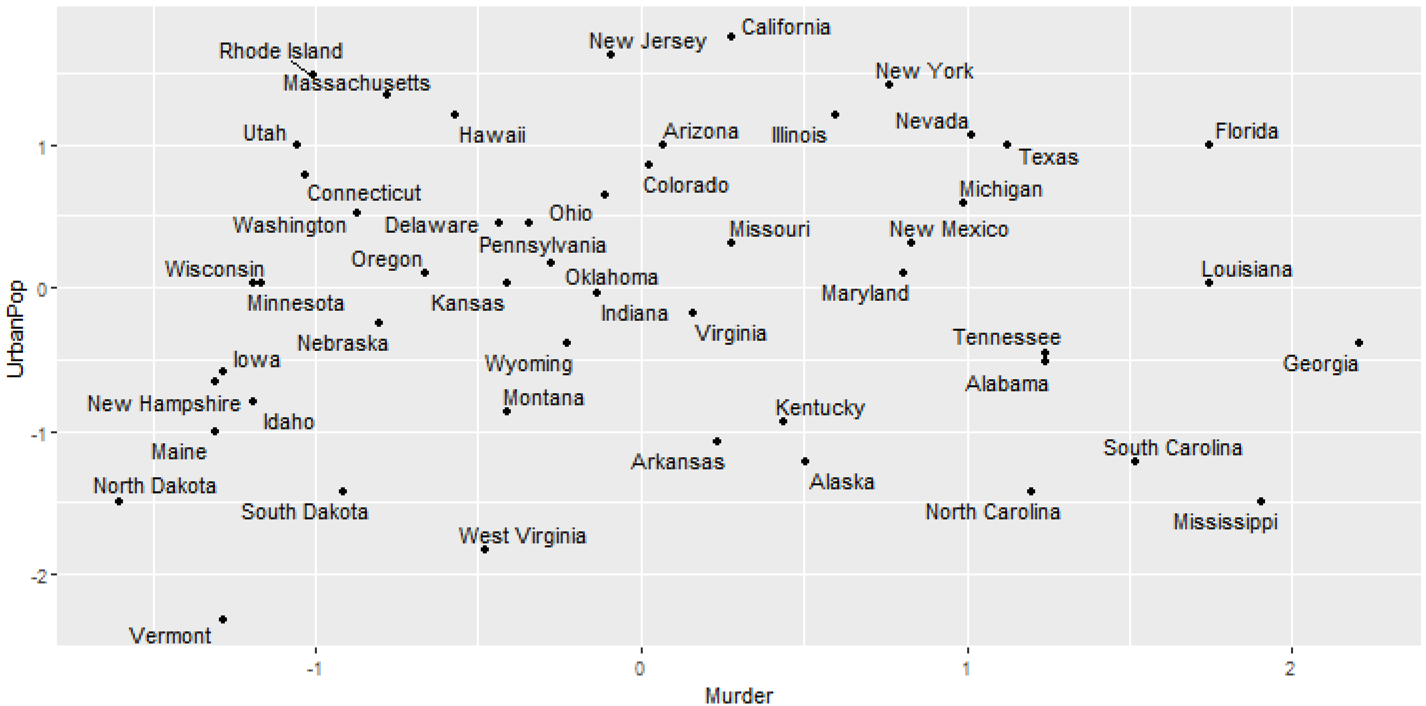
- Total Variance also called the total within sum square or the within-cluster sum of squares (WCSS) or WSS

# Mumber of clusters? Elbow method!

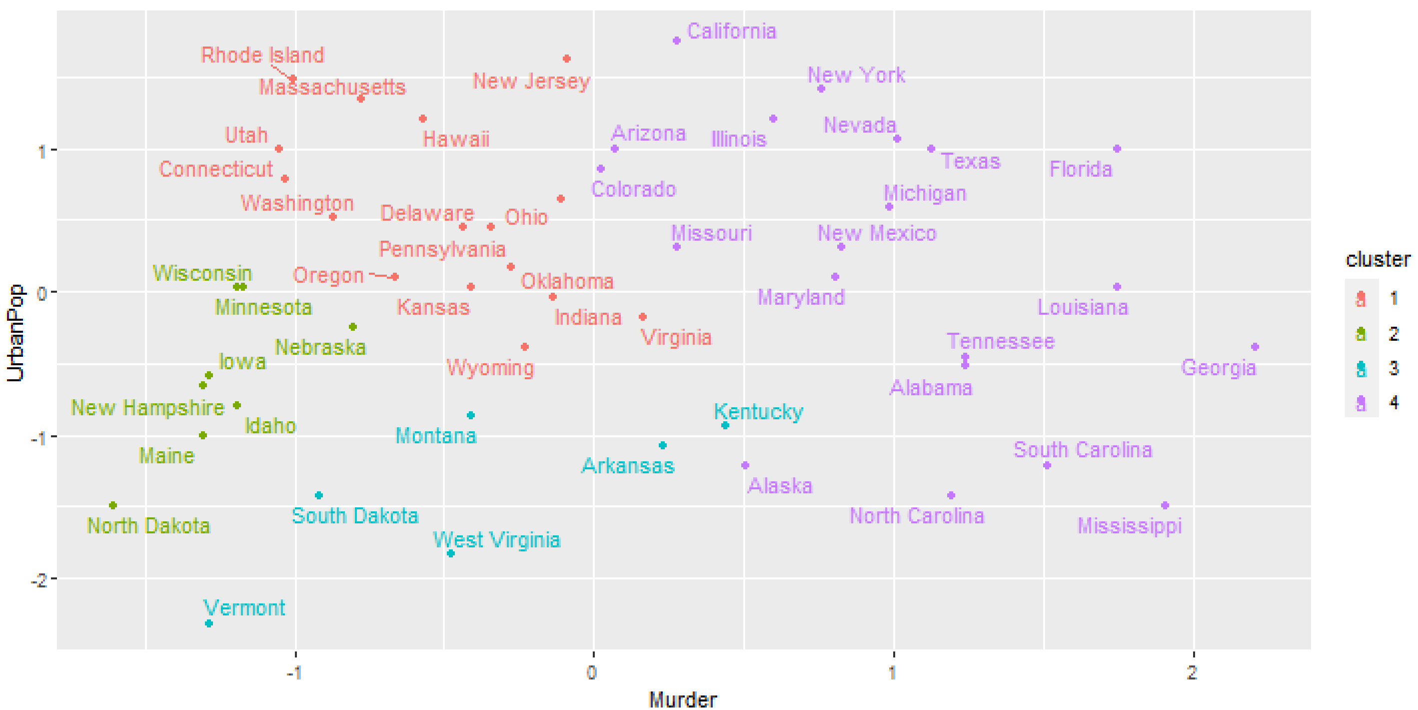
- Plot the WSS
- Decide the `elbow` of the graph



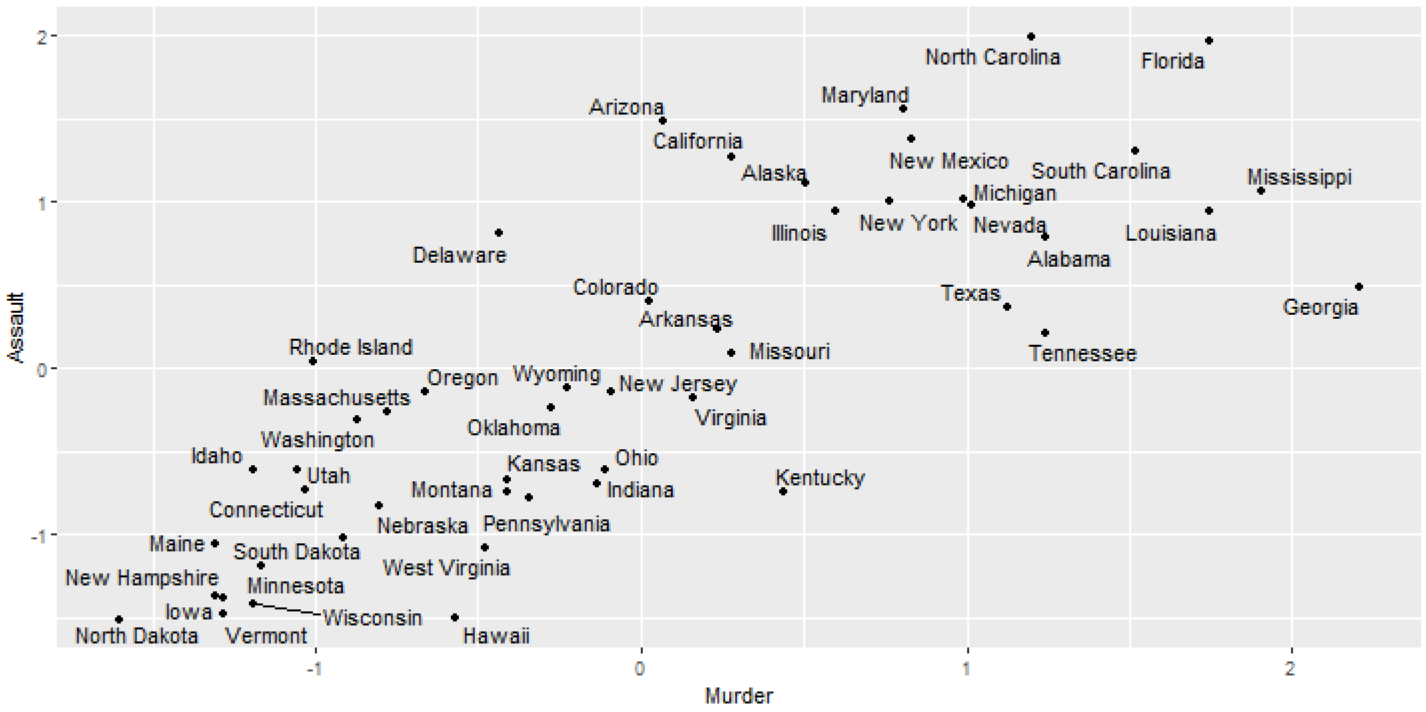
```
##                 Murder   Assault   UrbanPop         Rape cluster
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473       4
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941       4
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388       4
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602       3
## California 0.27826823 1.2628144  1.7589234  2.067820292       4
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207       4
```

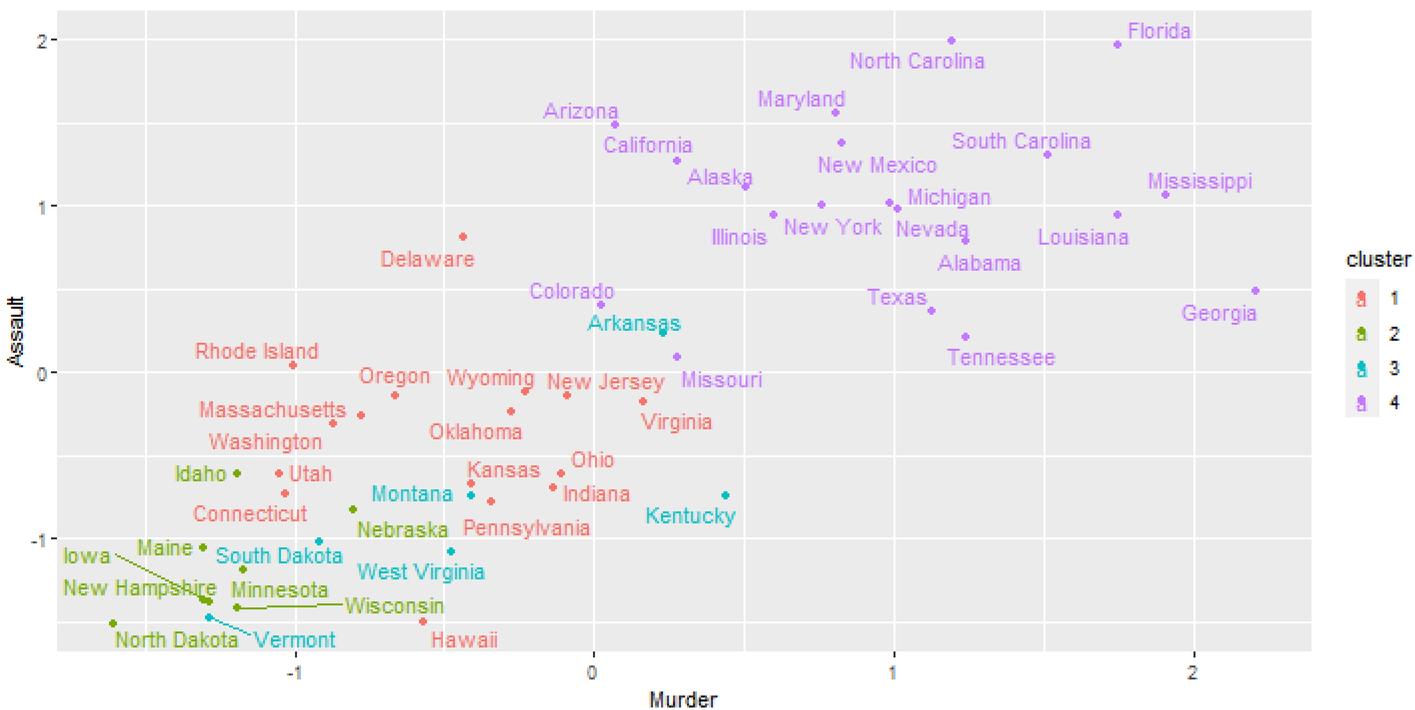


California
New Jersey
Rhode Island
Massachusetts
New York
Hawaii
Arizona
Illinois
Nevada
Utah
Florida
Texas
Colorado
Connecticut
Michigan
Ohio
Washington
Delaware
Missouri
New Mexico
Pennsylvania
Oregon
Wisconsin
Louisiana
Oklahoma
Maryland
Minnesota
Kansas
Indiana
Virginia
Nebraska
Tennessee
Iowa
Wyoming
Georgia
Alabama
Montana
New Hampshire
Kentucky
Idaho
Maine
South Carolina
Arkansas
Alaska
North Dakota
South Dakota
North Carolina
Mississippi
West Virginia
Vermont
UrbanPop
Murder
1
0
-1
-2
2

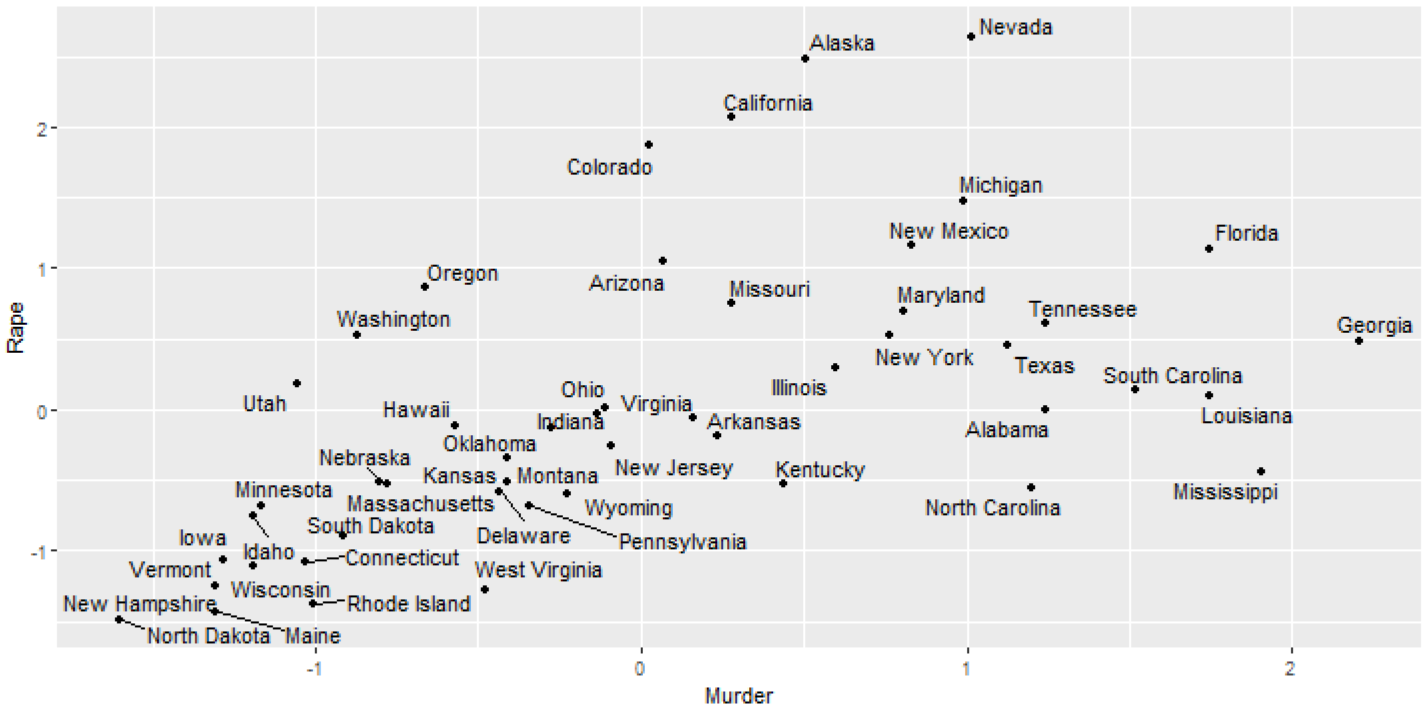


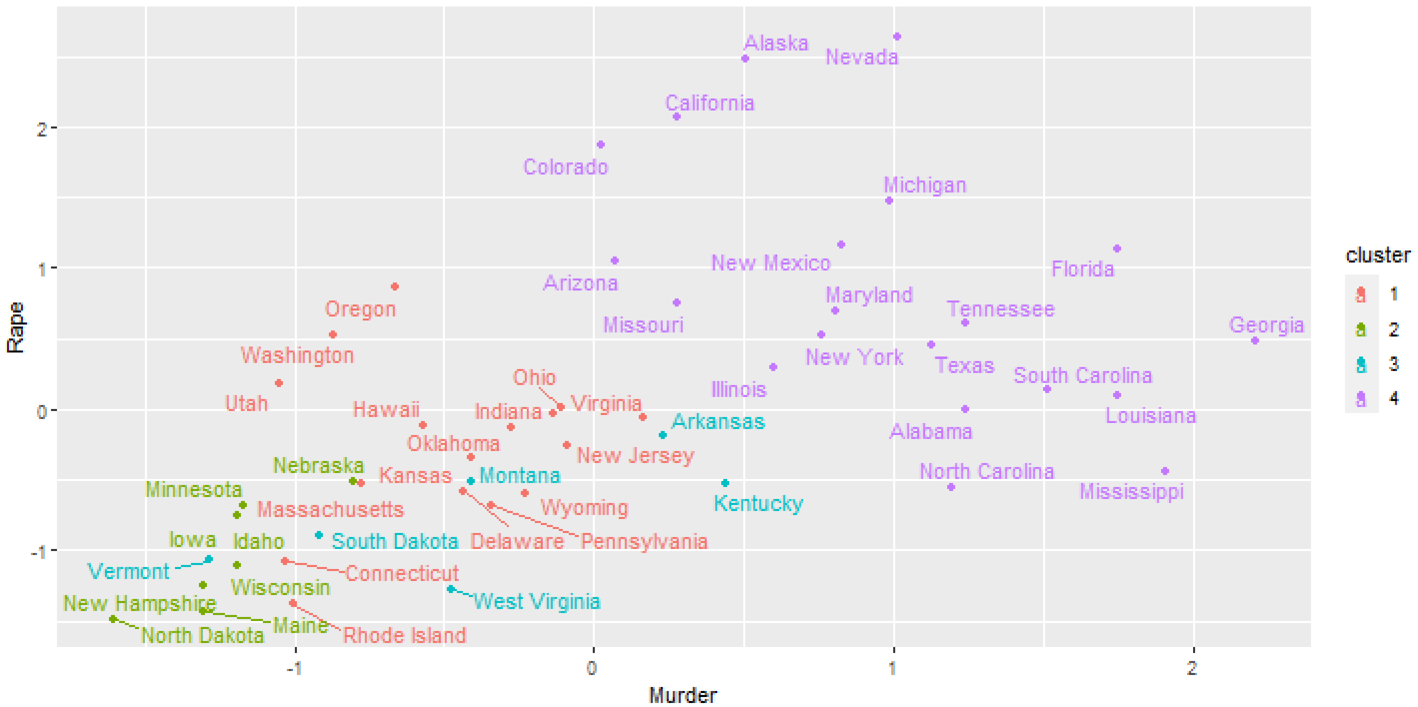
cluster
1
2
3
4
UrbanPop
Murder
California
Rhode Island
Massachusetts
New Jersey
New York
Utah
Hawaii
Arizona
Illinois
Nevada
Texas
Florida
Connecticut
Colorado
Michigan
Washington
Delaware
Ohio
Missouri
New Mexico
Pennsylvania
Wisconsin
Oregon
Oklahoma
Maryland
Minnesota
Kansas
Indiana
Louisiana
Nebraska
Virginia
Tennessee
Iowa
Wyoming
Georgia
Alabama
New Hampshire
Kentucky
Idaho
Montana
Maine
Arkansas
South Carolina
Alaska
North Carolina
North Dakota
South Dakota
Mississippi
West Virginia
Vermont

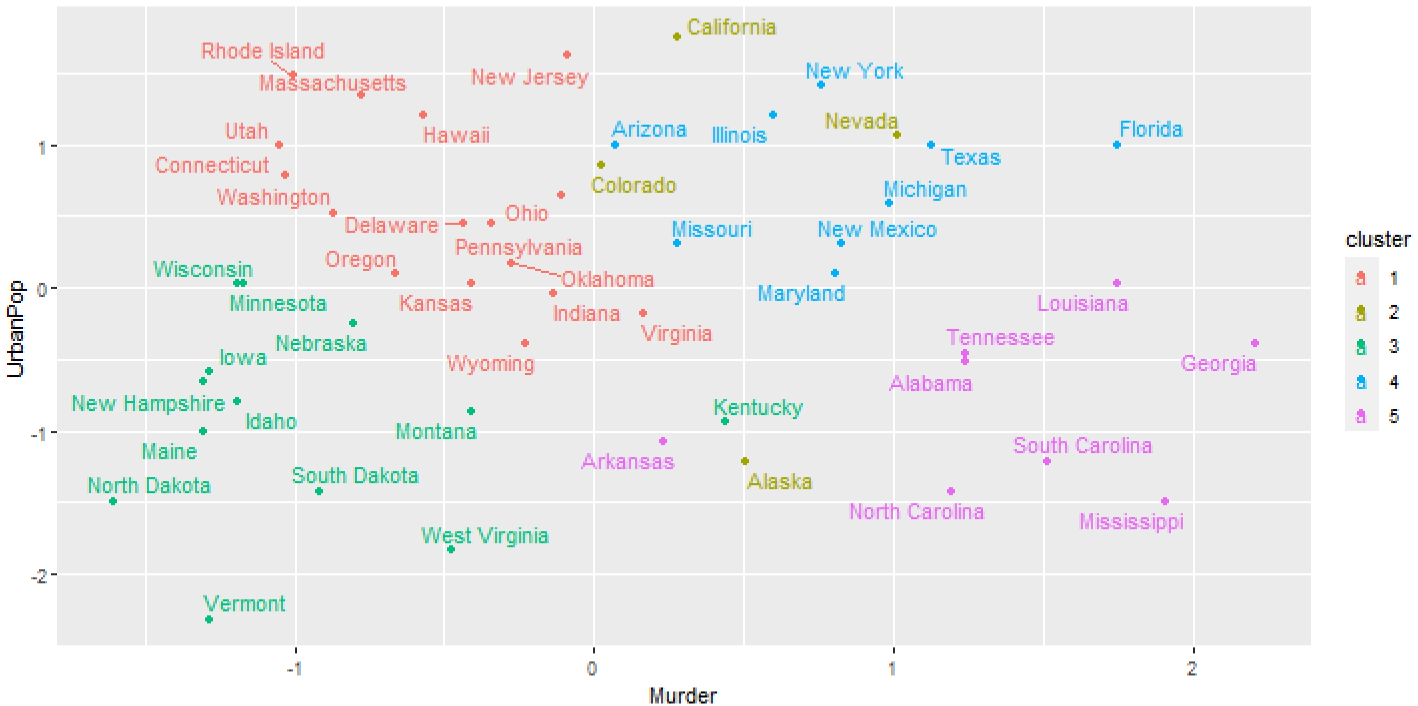

North Carolina
Florida
Maryland
Arizona
California
Alaska
New Mexico
South Carolina
Mississippi
Michigan
Illinois
New York
Nevada
Louisiana
Alabama
Delaware
Colorado
Texas
Georgia
Arkansas
Missouri
Tennessee
Rhode Island
Oregon
Wyoming
New Jersey
Massachusetts
Virginia
Oklahoma
Washington
Idaho
Kansas
Ohio
Utah
Kentucky
Montana
Indiana
Connecticut
Nebraska
Pennsylvania
Maine
South Dakota
West Virginia
New Hampshire
Minnesota
Iowa
Wisconsin
North Dakota
Vermont
Hawaii
Assault
Murder
2
1
0
-1

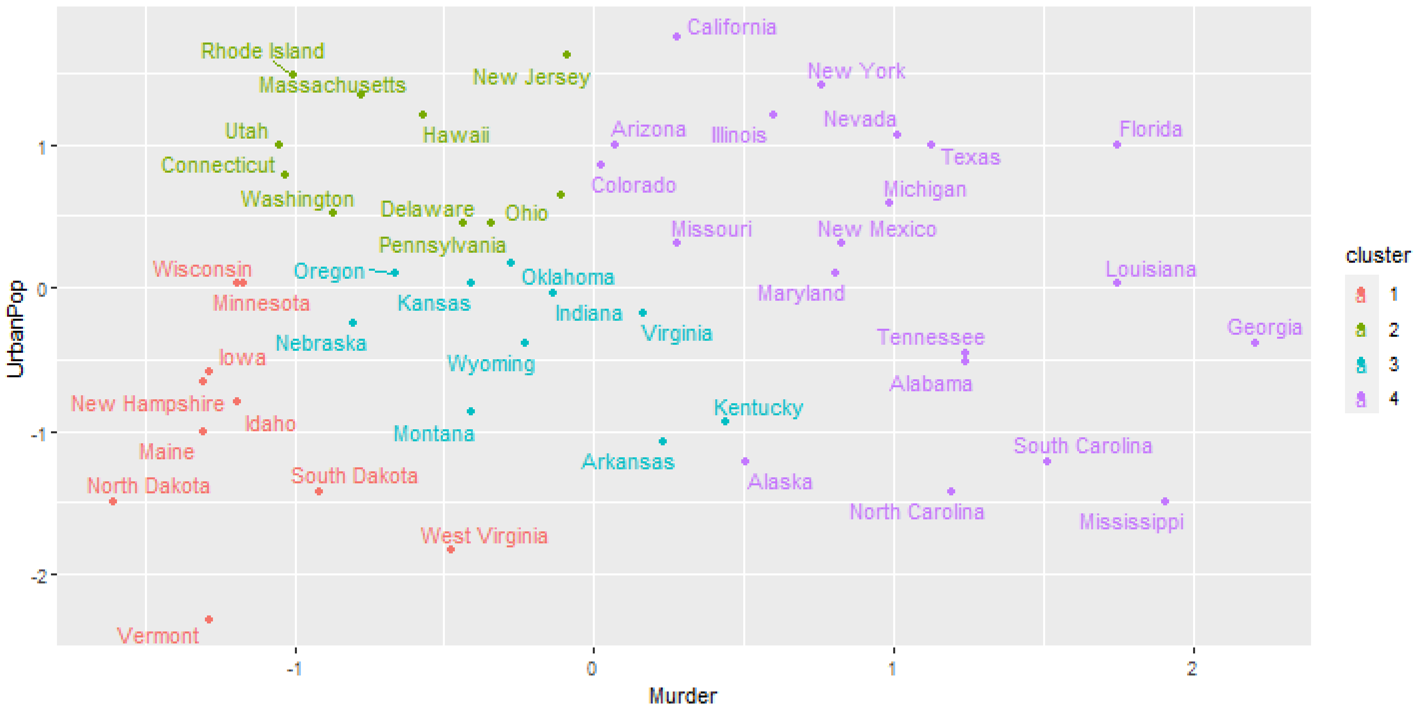


Florida
North Carolina
Maryland
Arizona
California
South Carolina
New Mexico
Alaska
Mississippi
Michigan
Illinois
New York
Nevada
Louisiana
Delaware
Alabama
Colorado
Texas
Georgia
Arkansas
Rhode Island
Tennessee
Oregon
Wyoming
New Jersey
Missouri
Massachusetts
Oklahoma
Virginia
Washington
Ohio
Kansas
Idaho
Utah
Montana
Indiana
Kentucky
Connecticut
Nebraska
Pennsylvania
Maine
Iowa
South Dakota
West Virginia
New Hampshire
Minnesota
Wisconsin
North Dakota
Vermont
Hawaii
Assault
Murder
cluster
1
2
3
4

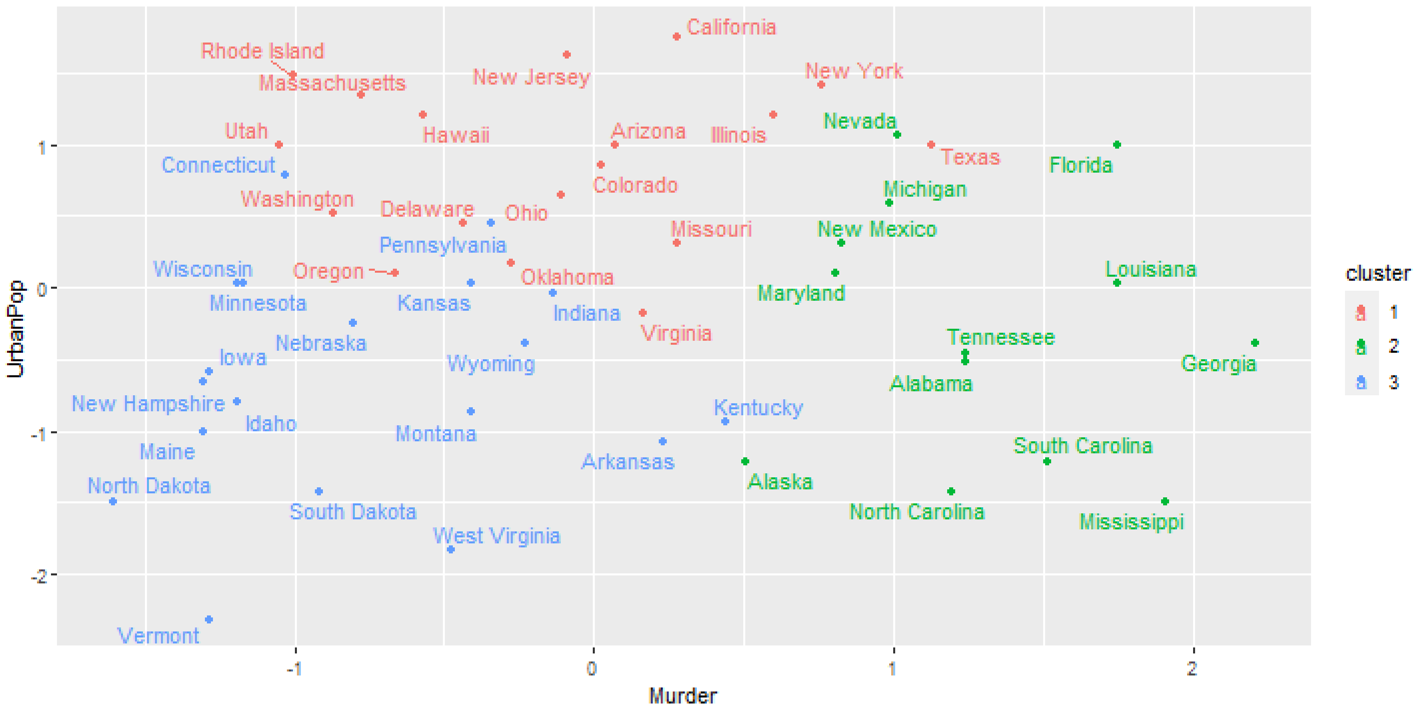

Nevada
Alaska
California
Colorado
Michigan
New Mexico
Florida
Oregon
Arizona
Missouri
Maryland
Tennessee
Georgia
Washington
New York
Texas
Illinois
South Carolina
Utah
Ohio
Virginia
Hawaii
Indiana
Arkansas
Louisiana
Alabama
Oklahoma
Nebraska
New Jersey
Kentucky
Kansas
Montana
Mississippi
Minnesota
Massachusetts
Wyoming
North Carolina
South Dakota
Delaware
Pennsylvania
Iowa
Idaho
Connecticut
Vermont
West Virginia
Wisconsin
New Hampshire
Rhode Island
North Dakota
Maine
Rape
Murder
2
1
0
-1
-1
0
1
2


Alaska
Nevada
California
Colorado
Michigan
New Mexico
Florida
Arizona
Maryland
Missouri
Tennessee
Georgia
Oregon
New York
Texas
Washington
South Carolina
Ohio
Illinois
Utah
Virginia
Hawaii
Indiana
Arkansas
Louisiana
Alabama
Oklahoma
New Jersey
Nebraska
Kansas
Montana
North Carolina
Minnesota
Mississippi
Kentucky
Massachusetts
Wyoming
Iowa
Idaho
South Dakota
Delaware
Pennsylvania
Vermont
Connecticut
Wisconsin
New Hampshire
West Virginia
Maine
North Dakota
Rhode Island
Rape
Murder
cluster
1
2
3
4
2
1
0
-1
-1
0
1
2


California
Rhode Island
Massachusetts
New Jersey
New York
Utah
Hawaii
Arizona
Illinois
Nevada
Florida
Texas
Connecticut
Colorado
Washington
Michigan
Delaware
Ohio
Missouri
New Mexico
Pennsylvania
Wisconsin
Oregon
Oklahoma
Maryland
Minnesota
Kansas
Indiana
Louisiana
Virginia
Nebraska
Tennessee
Iowa
Wyoming
Georgia
Alabama
New Hampshire
Kentucky
Idaho
Montana
Maine
South Carolina
Arkansas
Alaska
South Dakota
North Dakota
North Carolina
Mississippi
West Virginia
Vermont
UrbanPop
Murder
cluster
1
2
3
4
5



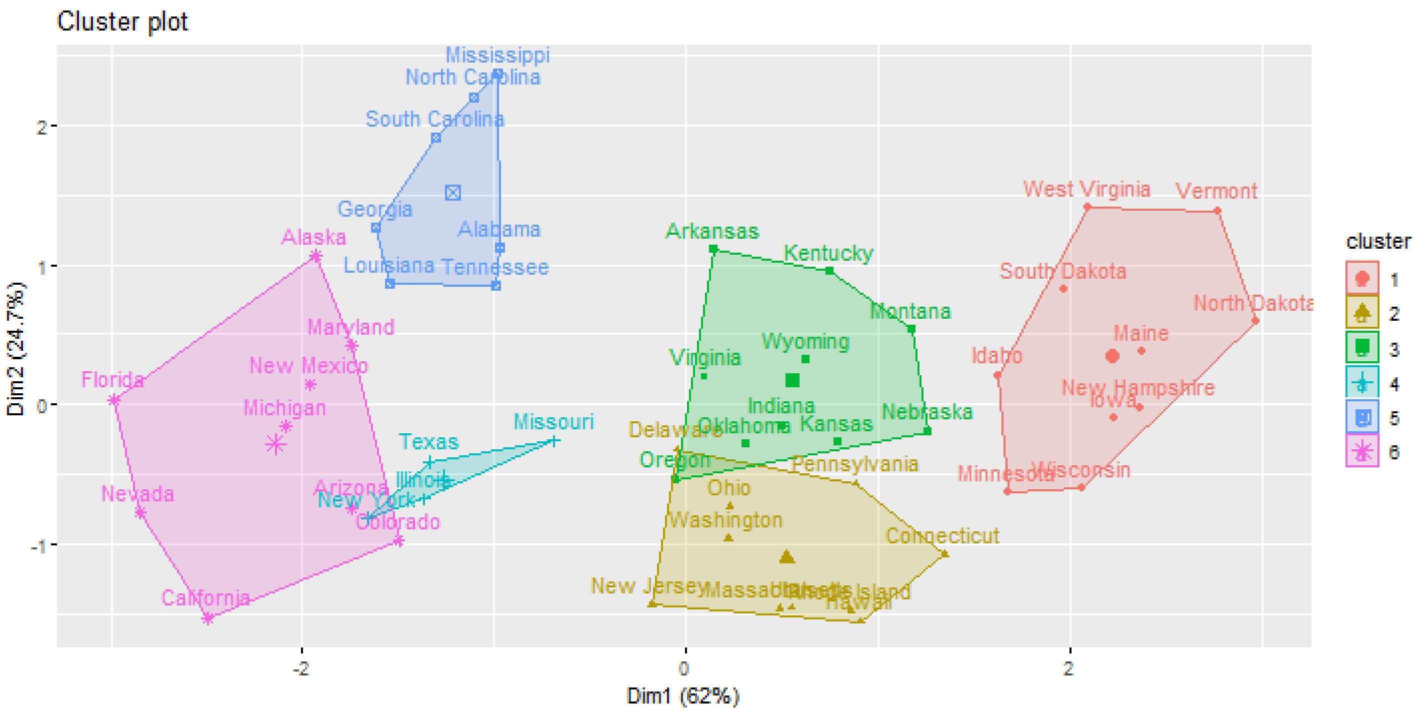
California
Rhode Island
Massachusetts
New Jersey
New York
Utah
Hawaii
Arizona
Illinois
Nevada
Florida
Texas
Connecticut
Colorado
Washington
Delaware
Ohio
Michigan
Missouri
New Mexico
Pennsylvania
Wisconsin
Oregon
Oklahoma
Louisiana
Maryland
Minnesota
Kansas
Indiana
Virginia
Georgia
Nebraska
Tennessee
Iowa
Wyoming
Alabama
New Hampshire
Kentucky
Idaho
Montana
South Carolina
Maine
Arkansas
Alaska
South Dakota
North Dakota
North Carolina
Mississippi
West Virginia
Vermont
UrbanPop
Murder
cluster
1
2
3
4



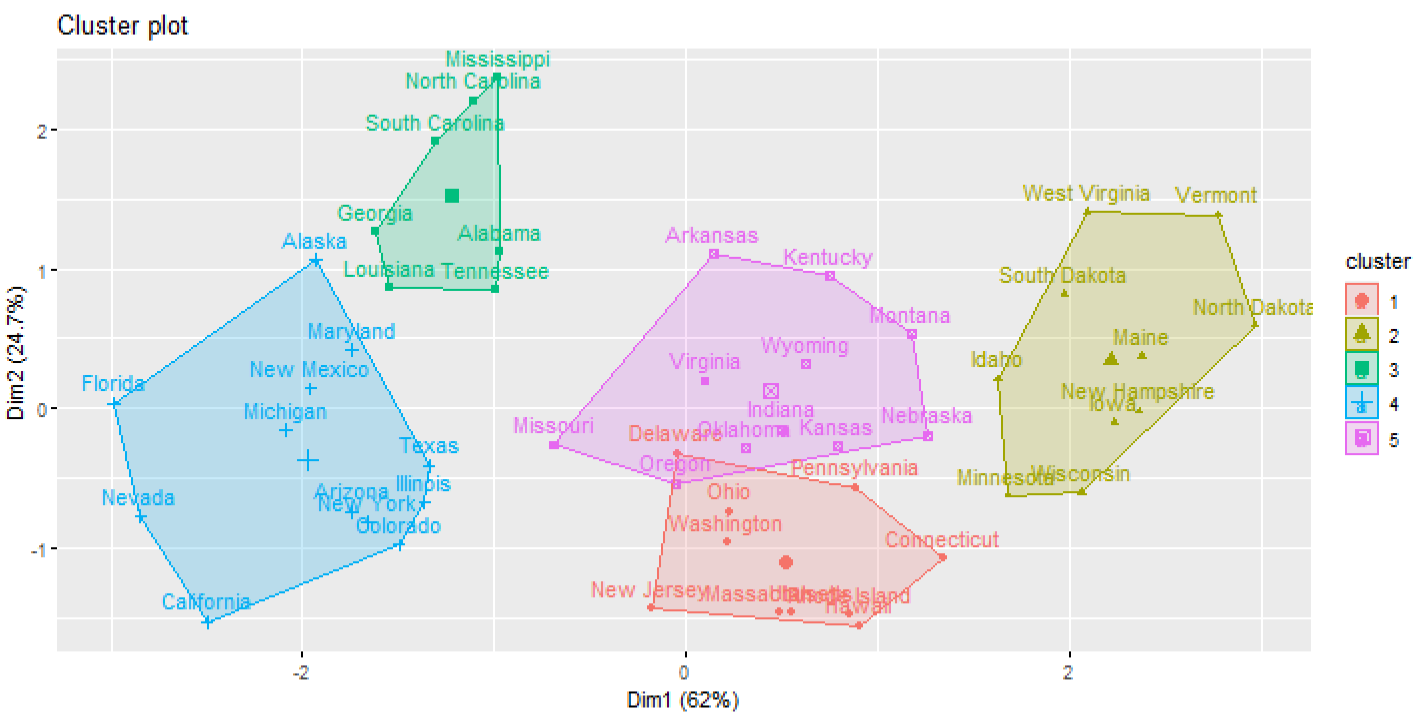
California
Rhode Island
Massachusetts
New Jersey
New York
Utah
Hawaii
Arizona
Illinois
Nevada
Texas
Florida
Connecticut
Colorado
Michigan
Washington
Delaware
Ohio
Missouri
New Mexico
Pennsylvania
Wisconsin
Oregon
Oklahoma
Louisiana
Maryland
Minnesota
Kansas
Indiana
Virginia
Nebraska
Tennessee
Iowa
Wyoming
Georgia
Alabama
New Hampshire
Kentucky
Idaho
Montana
Maine
South Carolina
Arkansas
Alaska
North Dakota
North Carolina
South Dakota
Mississippi
West Virginia
Vermont
UrbanPop
Murder
cluster
1
2
3

# Principal Components

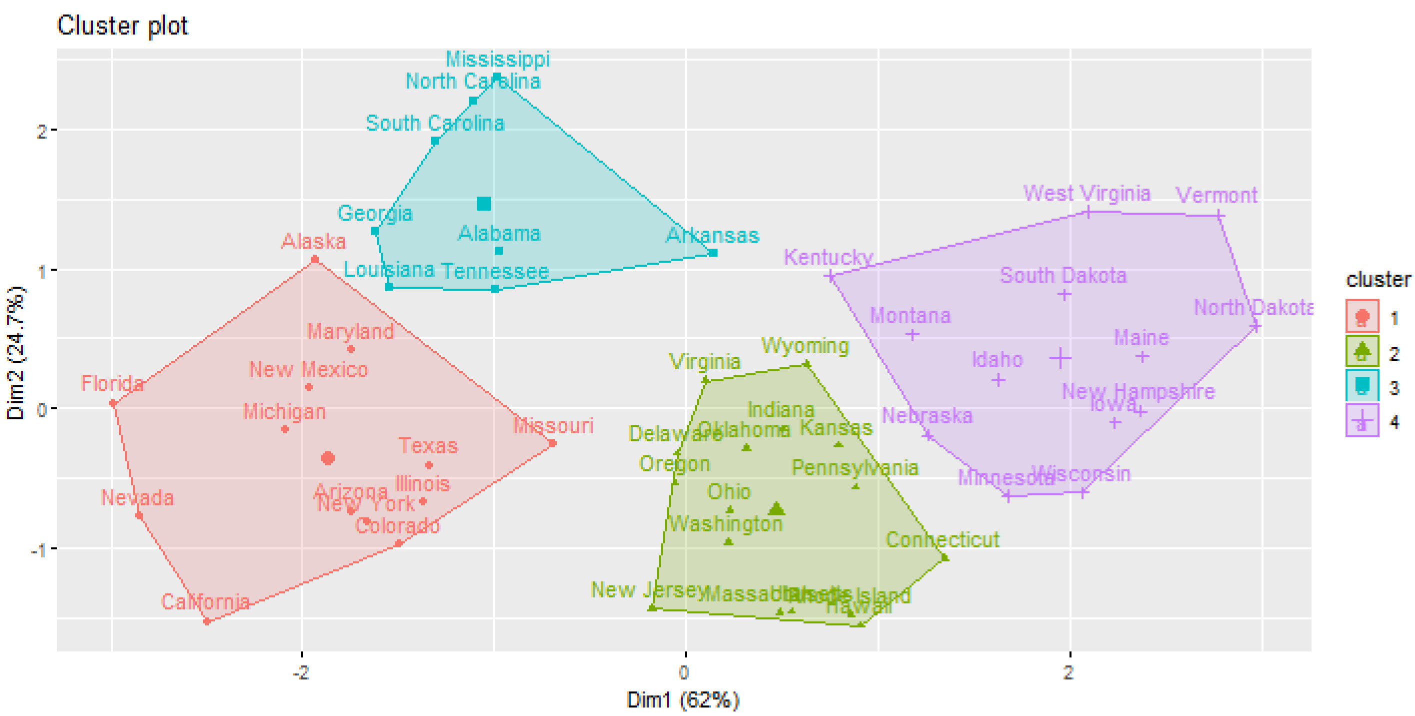
- Since we have four variables, we could have 6 different plots to visualize the clustring
- The more variables we have, the more plot we can have
- It is easier to contain all the variables in a few variables, then make plots.
- One way to do this is to use principal components analysis. The first two principals may contain most of the information of in the dataset.

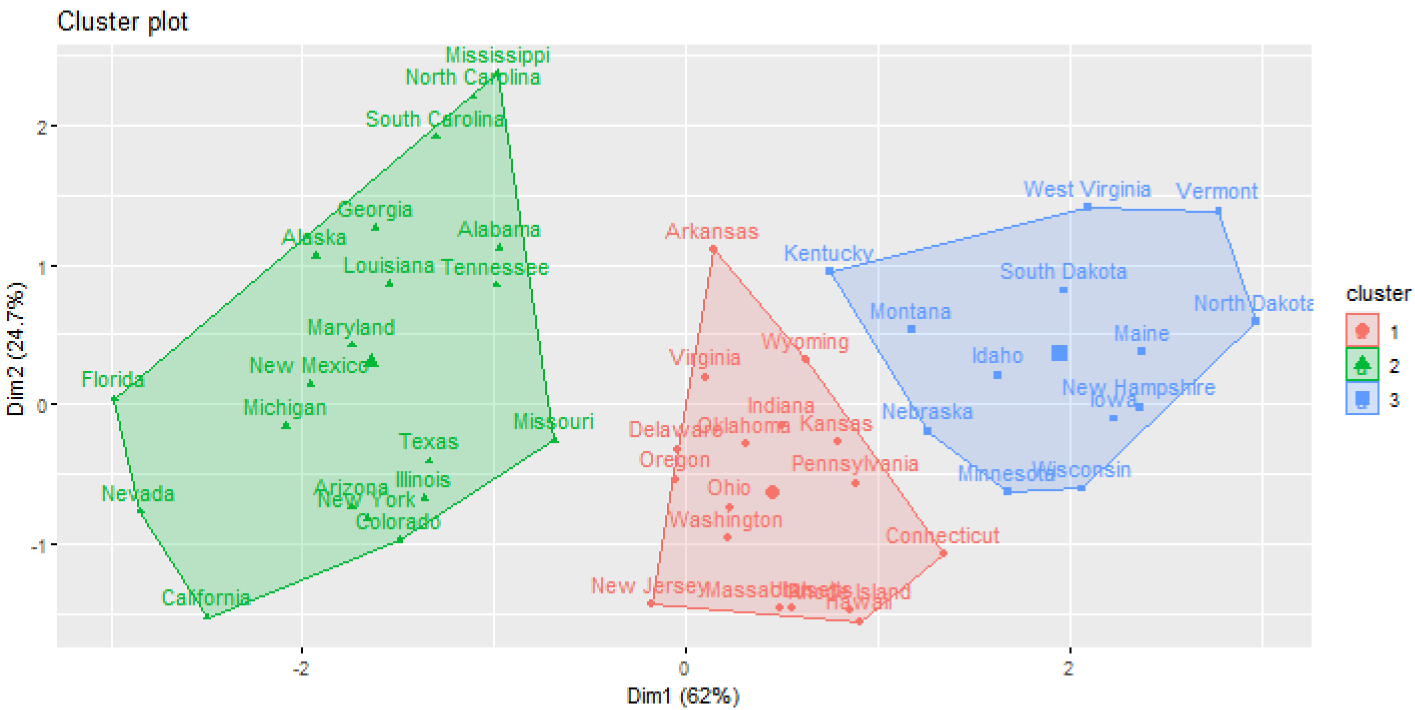


Cluster plot
Mississippi
North Carolina
South Carolina
Georgia
Alabama
Alaska
Louisiana
Tennessee
Maryland
New Mexico
Florida
Michigan
Texas
Missouri
Illinois
Arizona
New York
Nevada
Colorado
California
Arkansas
Kentucky
Montana
Wyoming
Virginia
Indiana
Oklahoma
Kansas
Nebraska
Delaware
Oregon
Pennsylvania
Ohio
Washington
Connecticut
New Jersey
Hawaii
West Virginia
Vermont
South Dakota
North Dakota
Maine
Idaho
New Hampshire
Iowa
Minnesota
Wisconsin
Dim1 (62%)
Dim2 (24.7%)
-2
0
2
-1
1
cluster
1
2
3
4
5
6

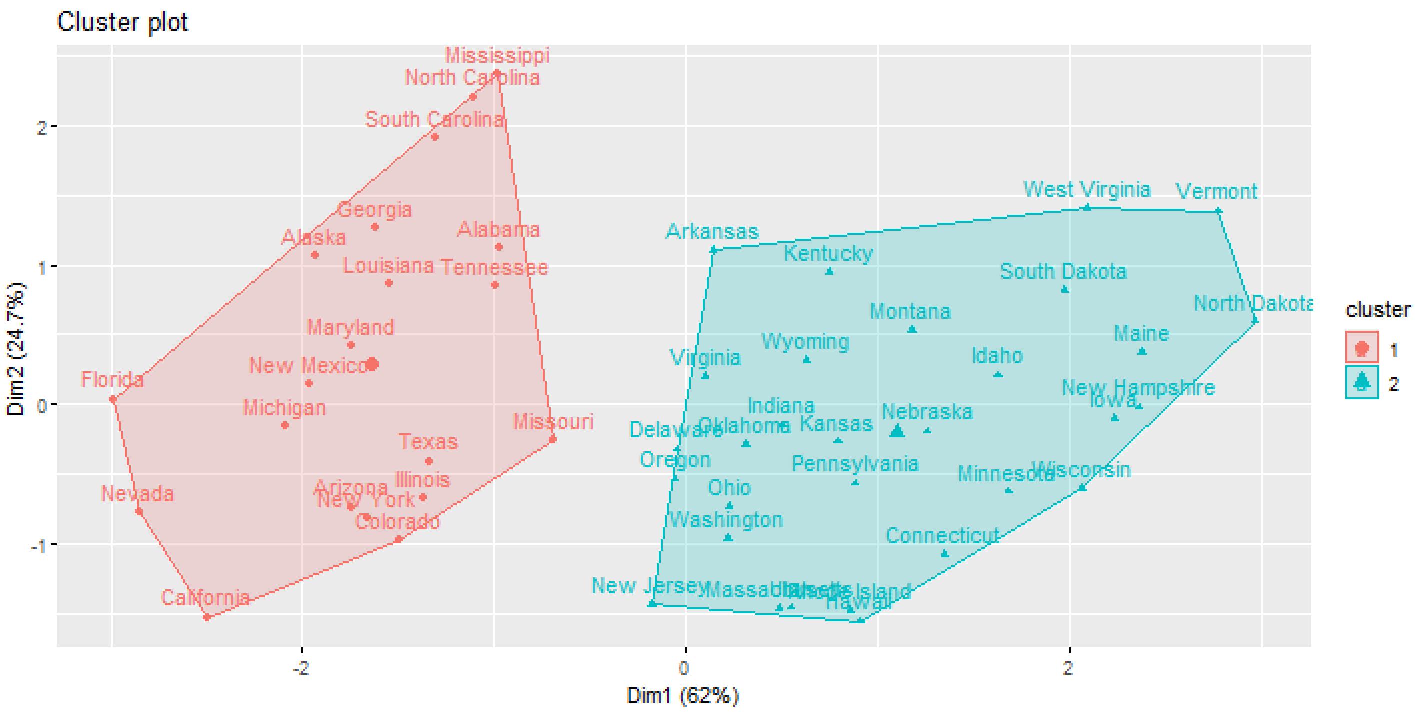


Cluster plot
Dim1 (62%)
Dim2 (24.7%)
cluster
1
2
3
4
5
Mississippi
North Carolina
South Carolina
Georgia
Alabama
Louisiana
Tennessee
Alaska
Maryland
New Mexico
Florida
Michigan
Texas
Illinois
Arizona
New York
Colorado
Nevada
California
Arkansas
Kentucky
Montana
Wyoming
Virginia
Indiana
Missouri
Delaware
Oklahoma
Kansas
Nebraska
Oregon
Pennsylvania
Ohio
Washington
Connecticut
New Jersey
Hawaii
West Virginia
Vermont
South Dakota
North Dakota
Maine
Idaho
New Hampshire
Iowa
Minnesota
Wisconsin


Cluster plot
Mississippi
North Carolina
South Carolina
Georgia
Alabama
Arkansas
Louisiana
Tennessee
Alaska
Maryland
New Mexico
Florida
Michigan
Missouri
Texas
Nevada
Arizona
Illinois
New York
Colorado
California
West Virginia
Vermont
Kentucky
South Dakota
North Dakota
Montana
Maine
Idaho
New Hampshire
Iowa
Nebraska
Minnesota
Wisconsin
Wyoming
Virginia
Indiana
Delaware
Oklahoma
Kansas
Oregon
Pennsylvania
Ohio
Washington
Connecticut
New Jersey
Hawaii
Dim1 (62%)
Dim2 (24.7%)
cluster
1
2
3
4
-2
0
2
-1
1



Cluster plot
Dim2 (24.7%)
Dim1 (62%)
2
1
0
-1
-2
0
2
Mississippi
North Carolina
South Carolina
Georgia
Alabama
Alaska
Louisiana
Tennessee
Maryland
New Mexico
Florida
Michigan
Texas
Missouri
Illinois
Arizona
New York
Nevada
Colorado
California
Arkansas
Kentucky
West Virginia
Vermont
South Dakota
Montana
North Dakota
Maine
Wyoming
Idaho
Virginia
New Hampshire
Iowa
Indiana
Nebraska
Delaware
Oklahoma
Kansas
Oregon
Pennsylvania
Minnesota
Wisconsin
Ohio
Washington
Connecticut
New Jersey
Hawaii
cluster
1
2
3



Cluster plot
Dim1 (62%)
Dim2 (24.7%)
cluster
1
2
-2
0
2
-1
1
Mississippi
North Carolina
South Carolina
Georgia
Alabama
Alaska
Louisiana
Tennessee
Maryland
New Mexico
Florida
Michigan
Missouri
Texas
Illinois
Arizona
New York
Nevada
Colorado
California
West Virginia
Vermont
Arkansas
Kentucky
South Dakota
North Dakota
Montana
Maine
Wyoming
Idaho
Virginia
New Hampshire
Iowa
Indiana
Nebraska
Delaware
Oklahoma
Kansas
Oregon
Pennsylvania
Minnesota
Wisconsin
Ohio
Washington
Connecticut
New Jersey
Hawaii