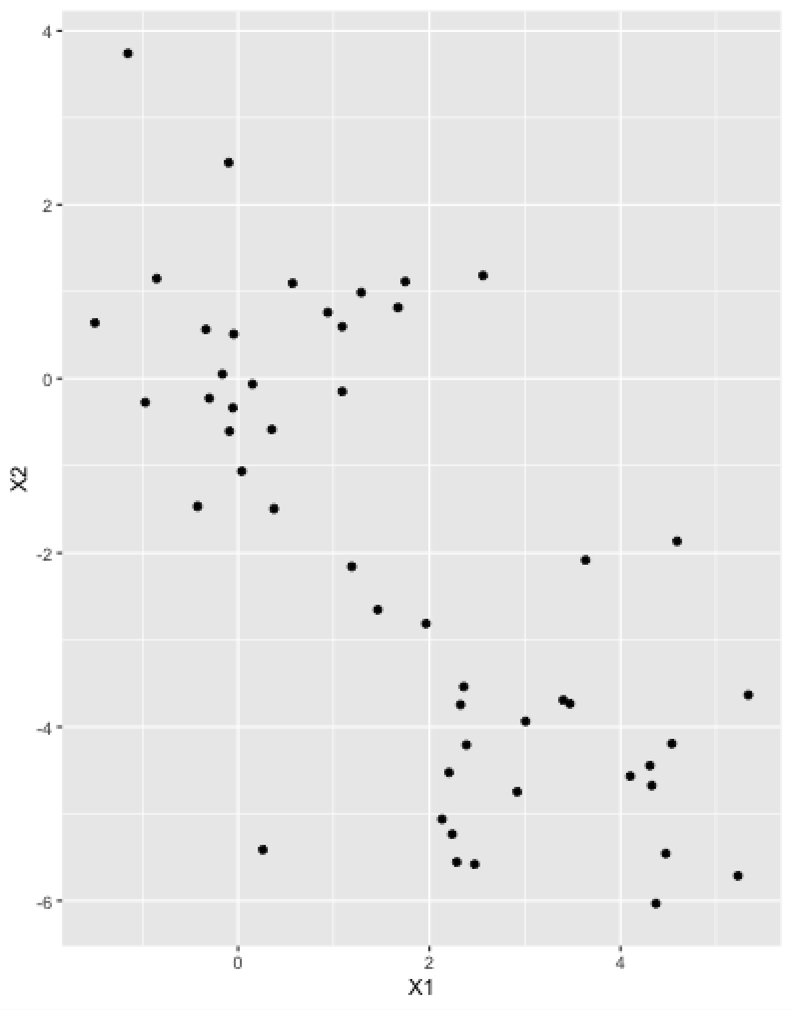
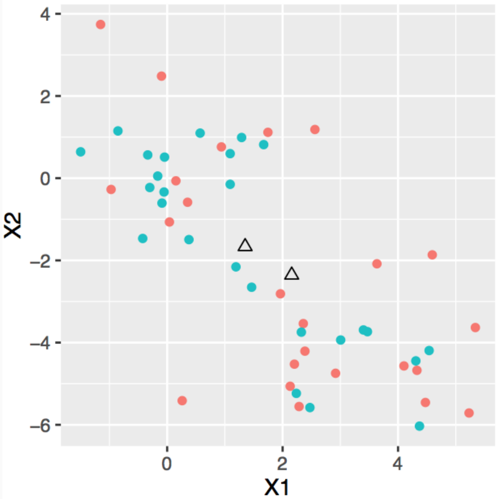
Clustering: k-means



4
2
0
-2
-4
-6
X2
0
2
4
X1

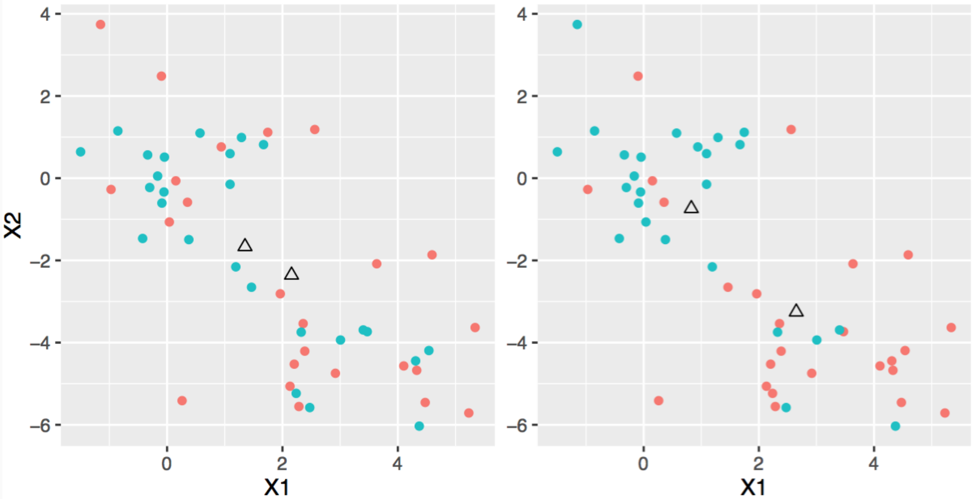
Three initial partitions



Partition 1

sum(W(C_k)) = 487.02

Three initial partitions



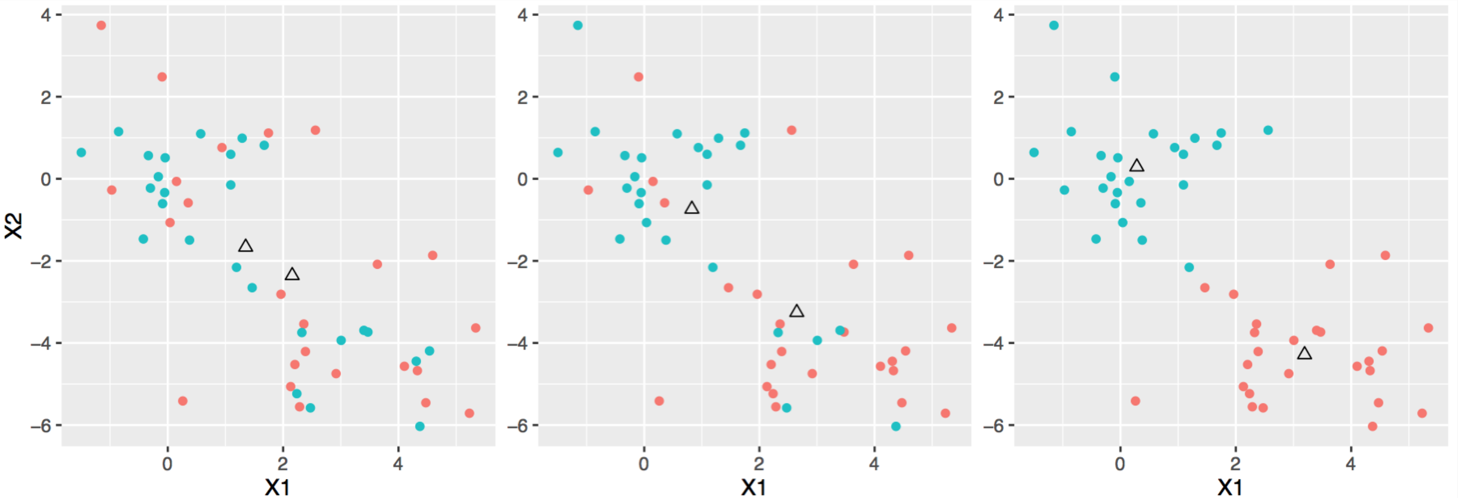
Partition 1

sum(W(C_k)) = 487.02

Partition 2

sum(W(C_k)) = 377.27

Three initial partitions



Partition 1

sum(W(C_k)) = 487.02

Partition 2

sum(W(C_k)) = 377.27

Partition 3

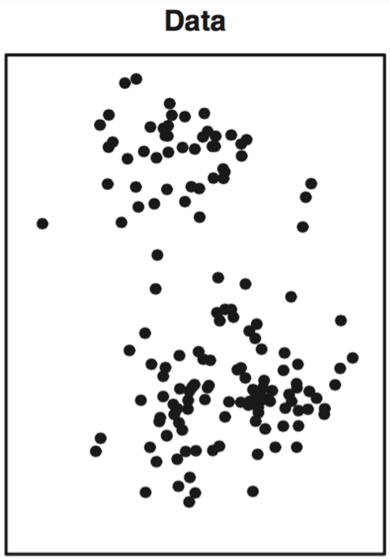
sum(W(C_k)) = 130.53

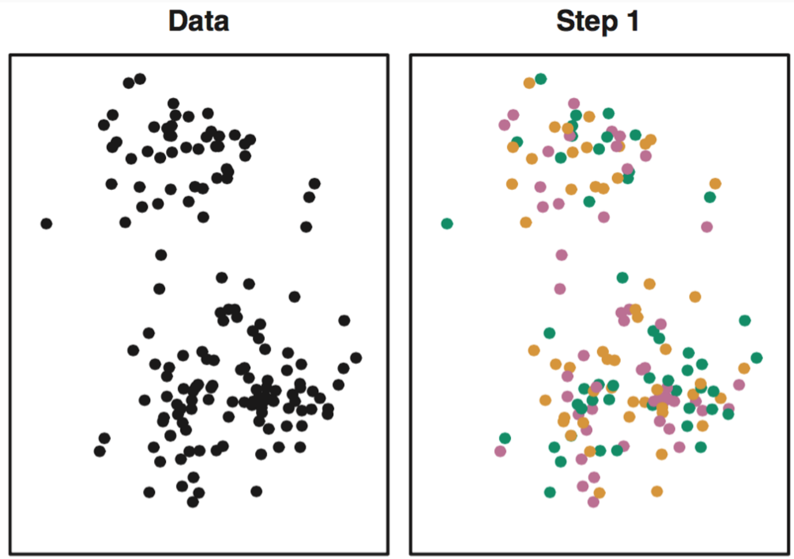
Algorithm 10.1

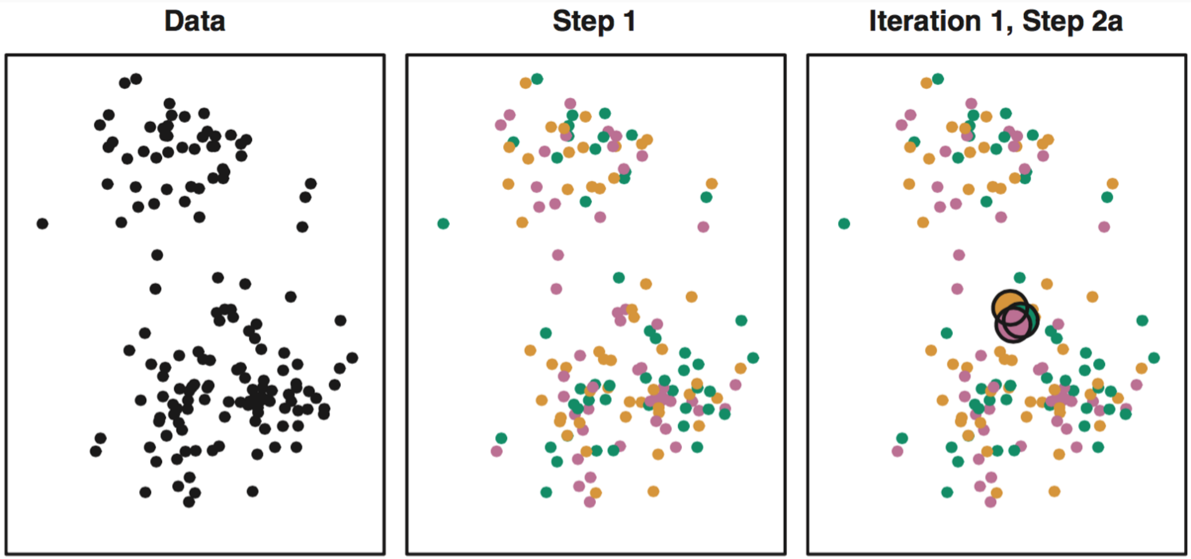
1. Randomly assign each obs. to 1 of K clusters.

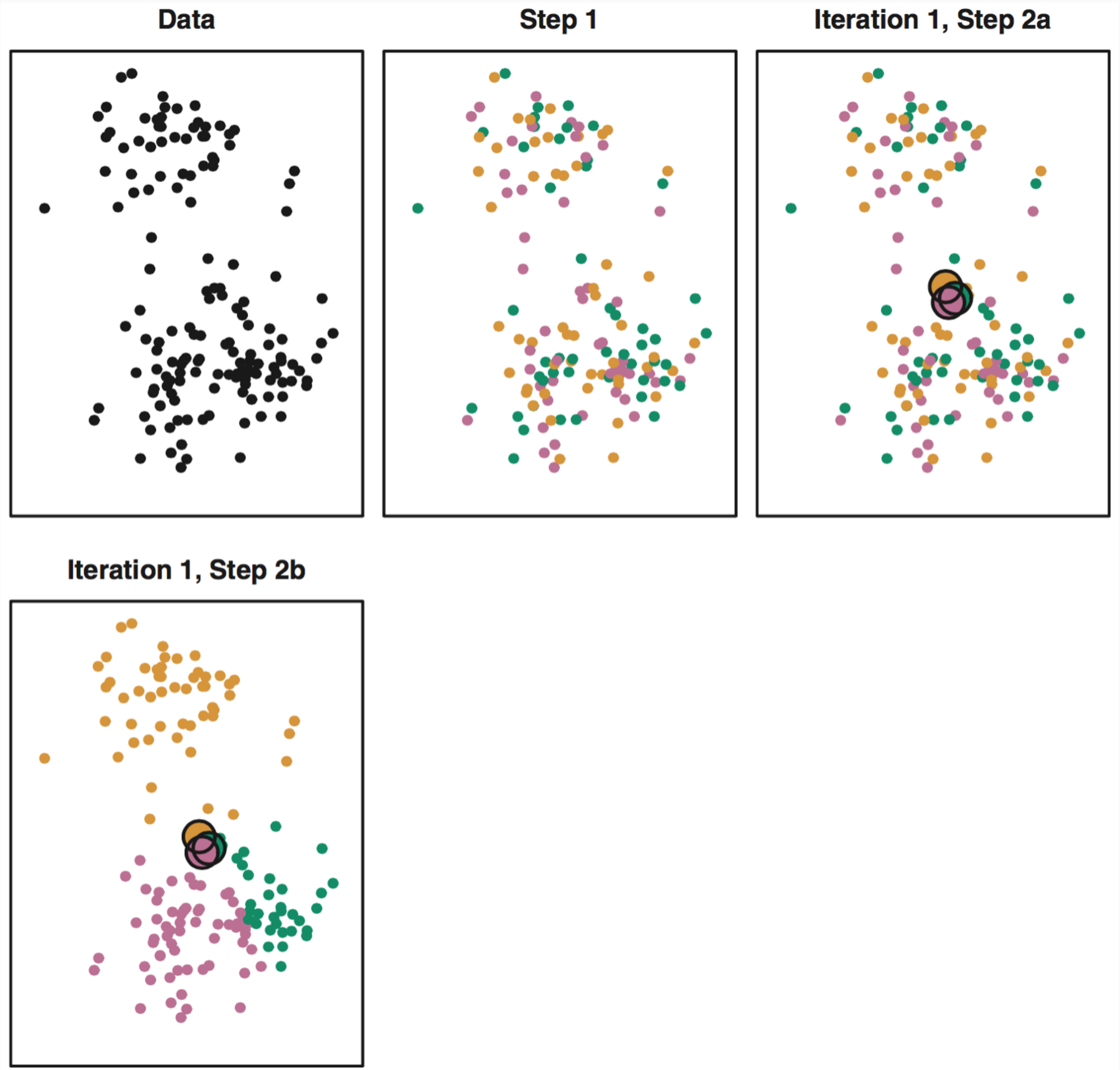
2. Iterate until the clusters stop changing:

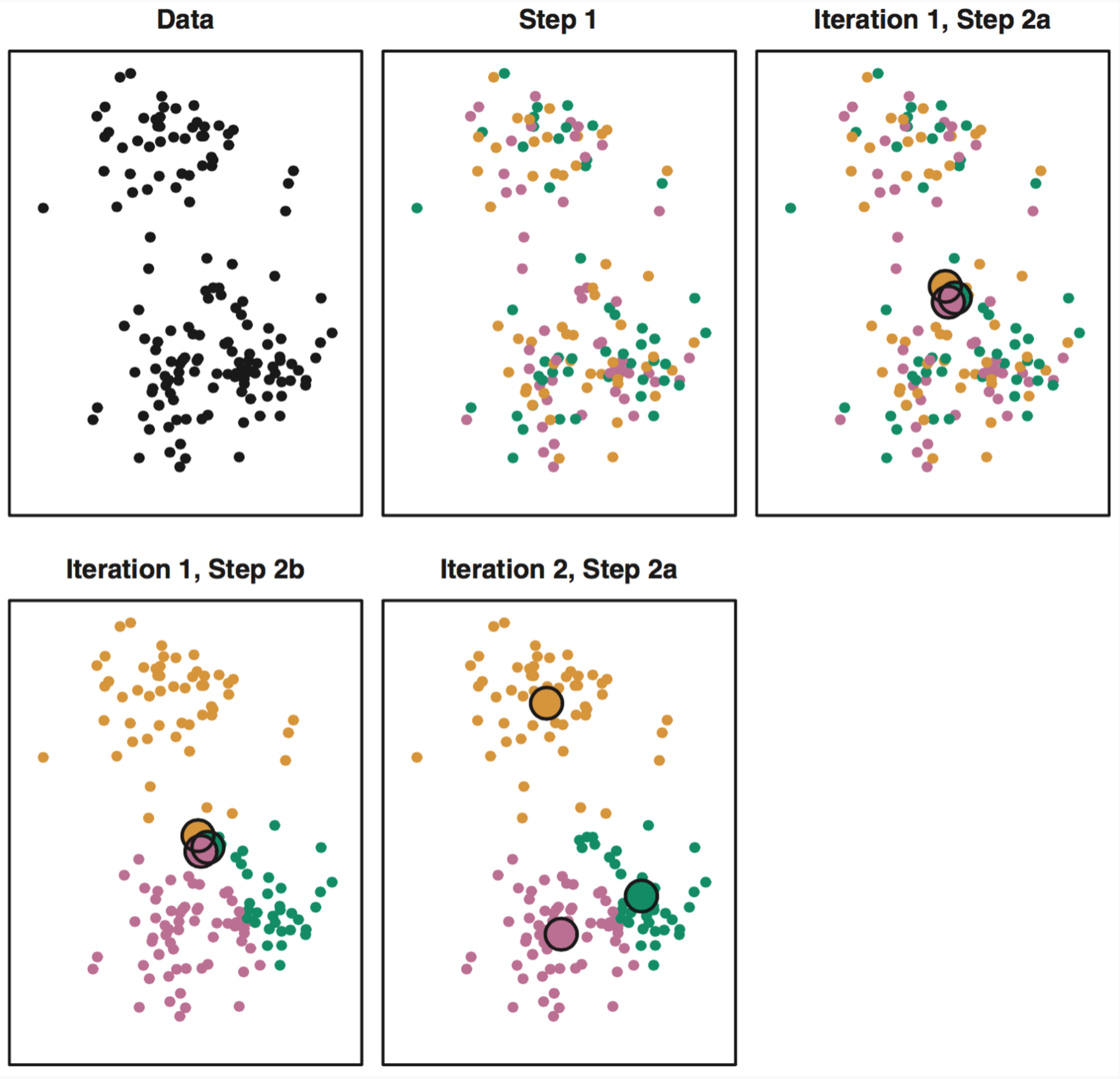
    - For each of the K clusters, compute the centroid (i.e. mean vector).
    - Assign each observation to the cluster whose centroid is closest (by Euclidean distance).

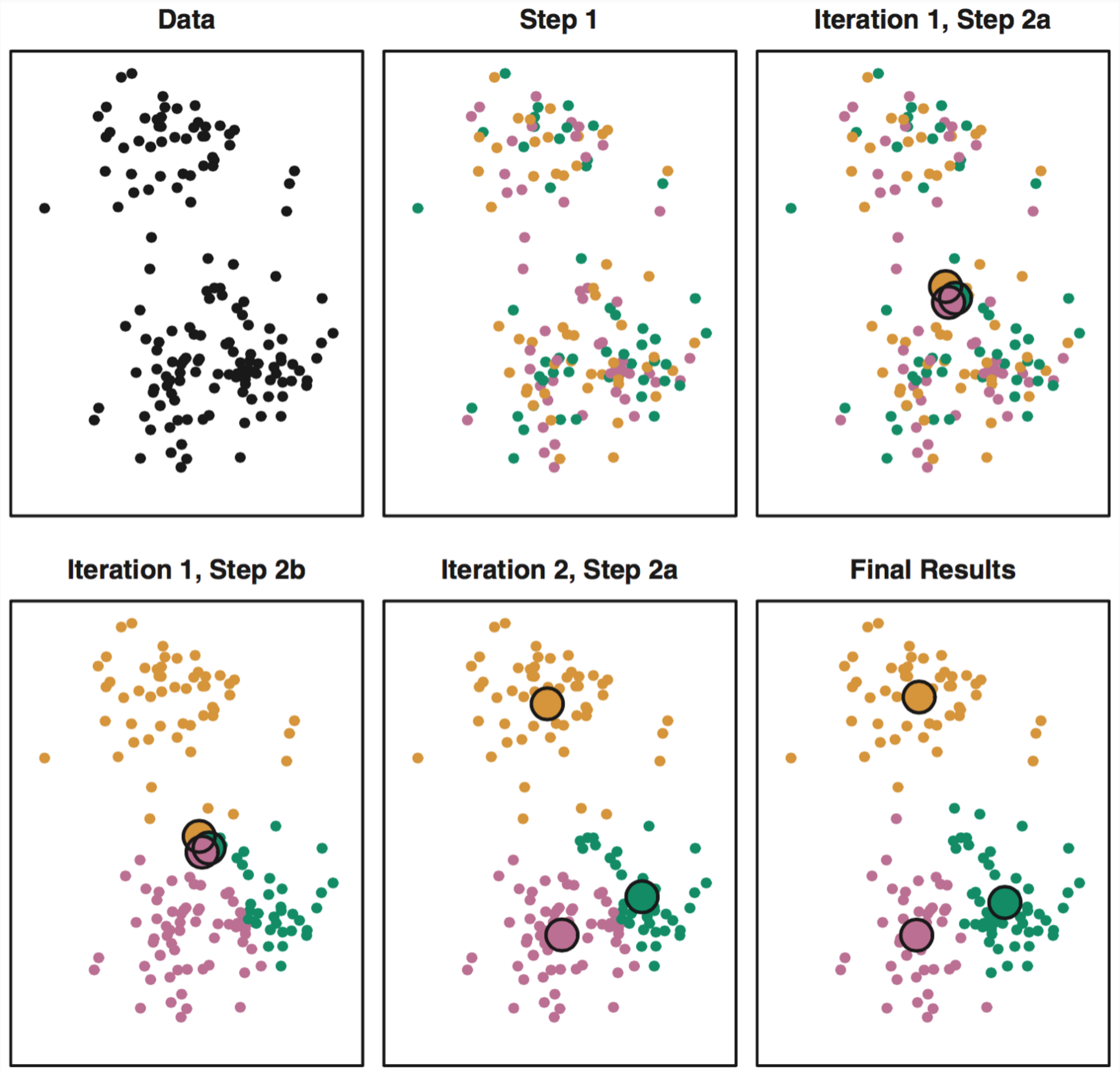

Data


Data
Step 1


Data
Step 1
Iteration 1, Step 2a


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Important considerations

1. The final partition is dependent on initial assignments.
    - *Solution*: run the algorithm several times with different starting conditions and select best.
2. Consider scaling the variables
    - Scale if you want "similar" to mean close w.r.t. all variables.

Activity 5

Use K-means clustering to identify the best 2, 3, and 4 clusterings of US states based on the data in the `poverty`. Use Euclidean distance for your similarity measure.

```
poverty <- read.delim("https://bit.ly/381pd5e")
```

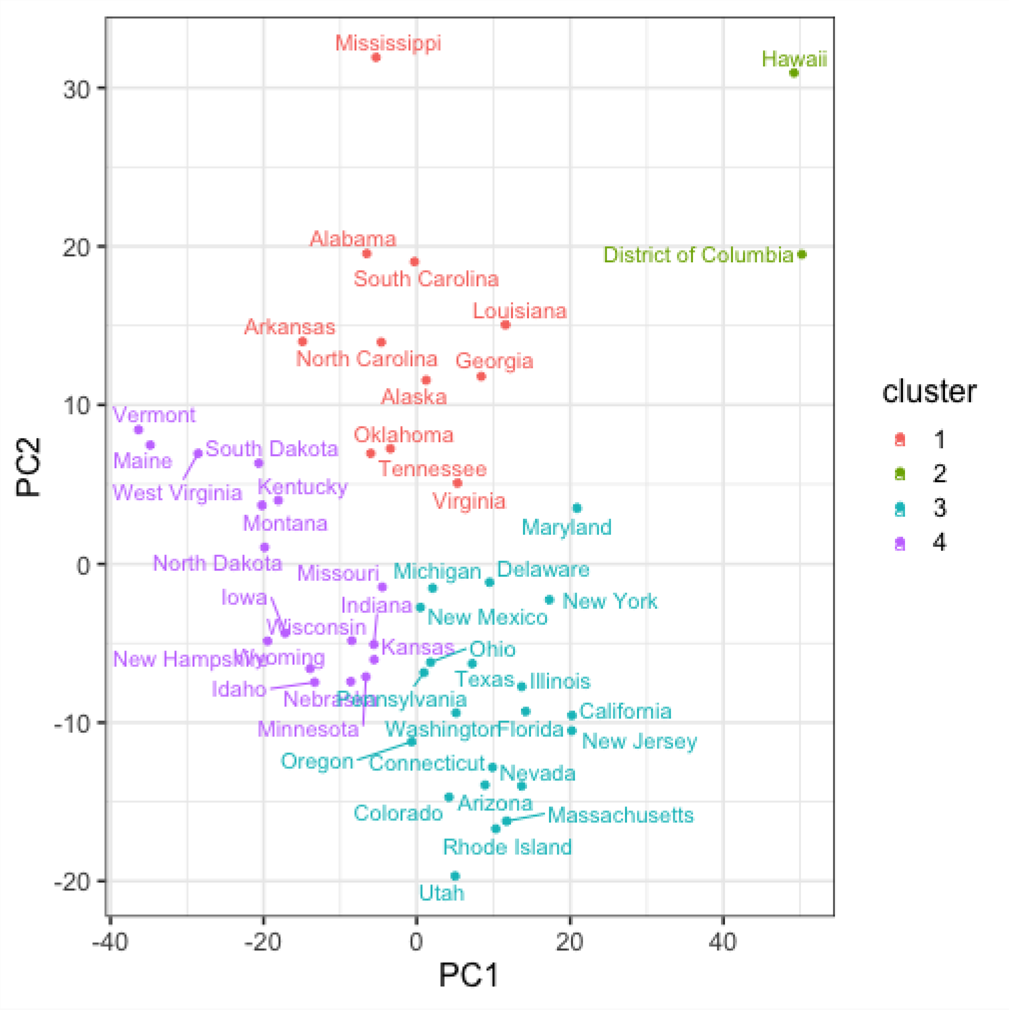
Useful functions:

- `kmeans()`
- `set.seed()`
- `geom_text()` or `ggrepel::geom_text_repel()`

1. What do the variables seem to mean?
2. Find best cluster assignments of size K.
3. Generate a scatterplot of the 51 obs and their first two PCs.
4. Color code each with their cluster assignment.

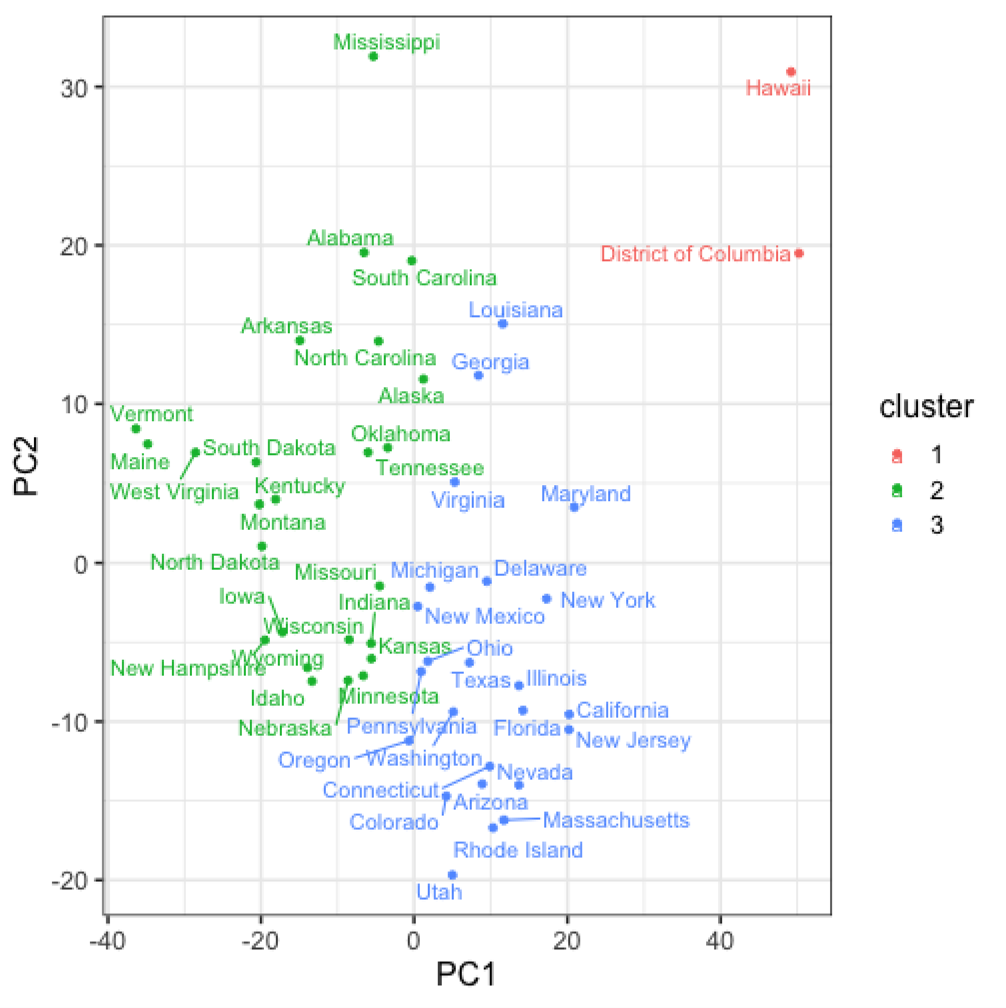
Choosing K

K = 4



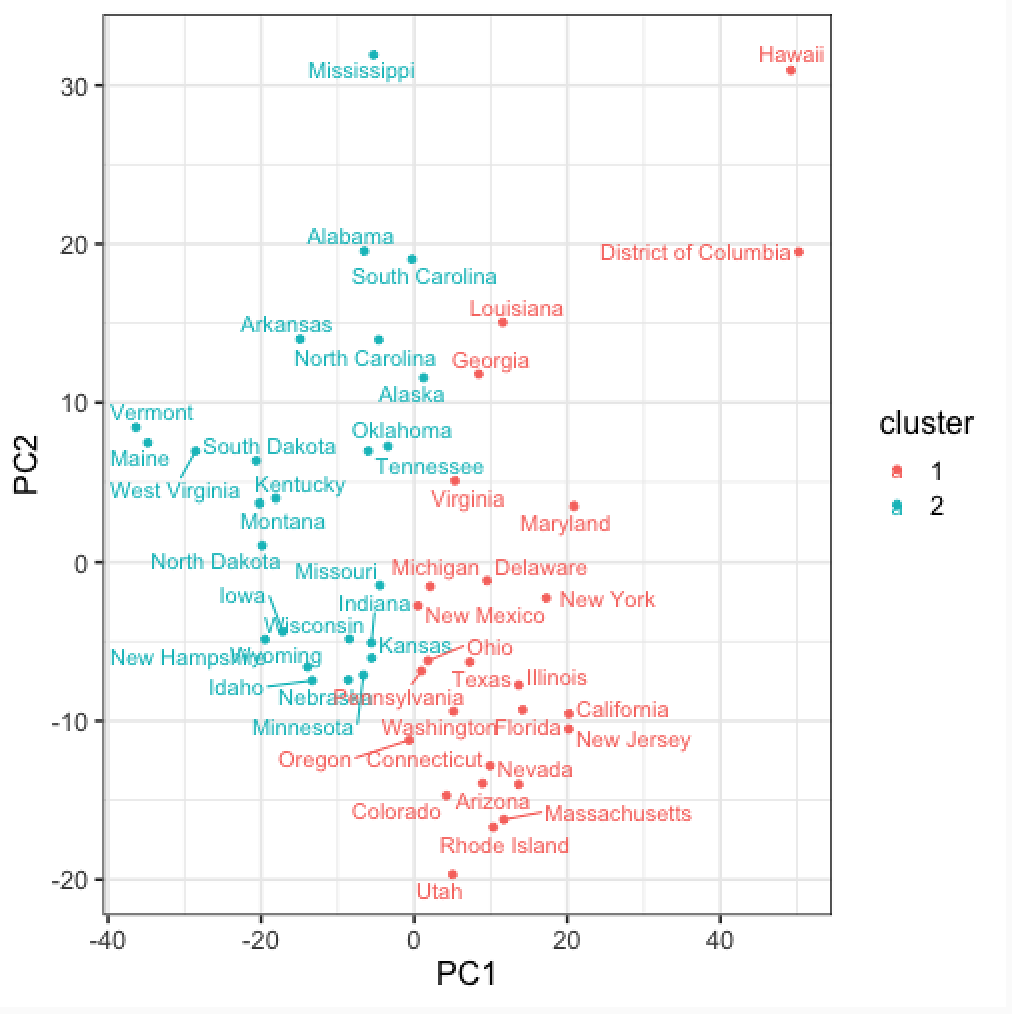
Mississippi
Hawaii
Alabama
South Carolina
District of Columbia
Louisiana
Arkansas
North Carolina
Georgia
Alaska
Vermont
Oklahoma
South Dakota
Maine
Tennessee
West Virginia
Kentucky
Virginia
Maryland
Montana
North Dakota
Missouri
Michigan
Delaware
Iowa
Indiana
New York
New Mexico
Wisconsin
Kansas
Ohio
New Hampshire
Wyoming
Texas
Illinois
Idaho
Nebraska
Pennsylvania
California
Minnesota
Washington
Florida
New Jersey
Oregon
Connecticut
Nevada
Arizona
Colorado
Massachusetts
Rhode Island
Utah
PC2
PC1
30
20
10
0
-10
-20
-40
-20
0
20
40
cluster
1
2
3
4

K = 3



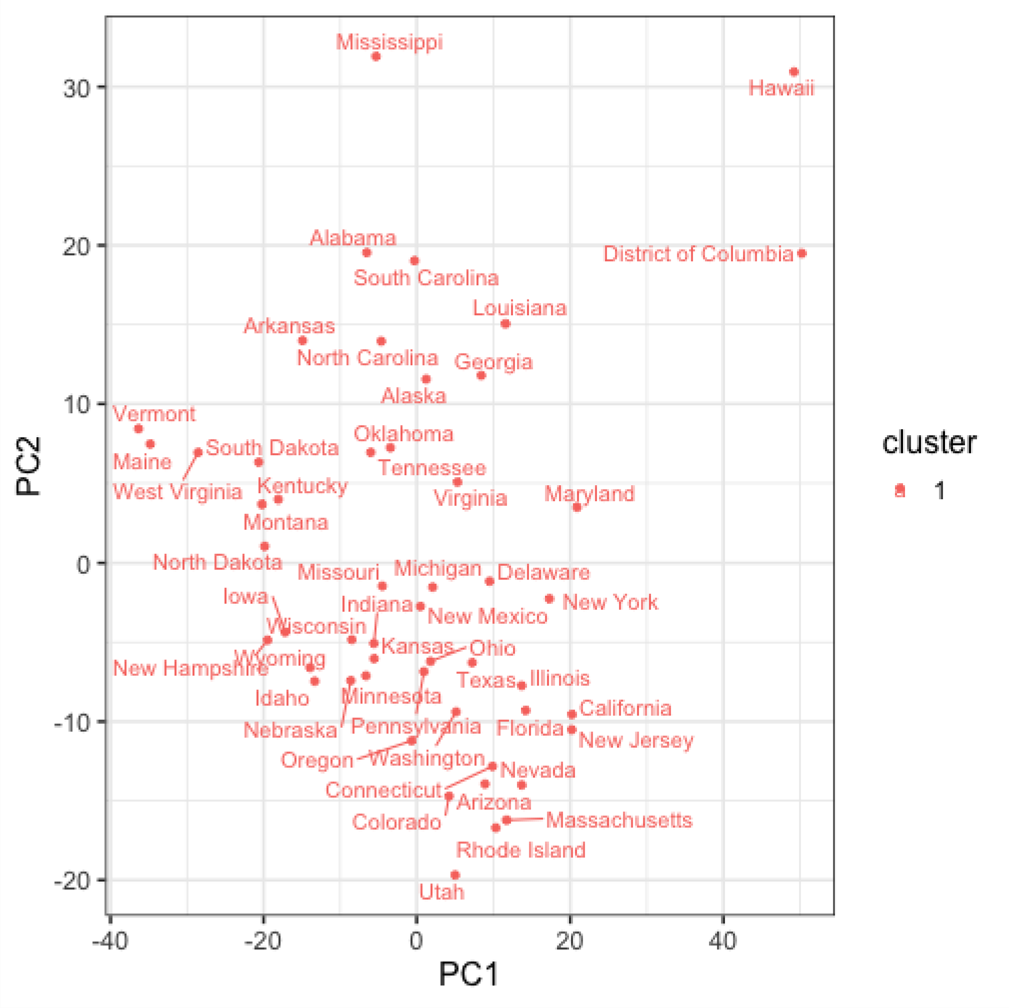
cluster
1
2
3
PC1
PC2
Mississippi
Hawaii
Alabama
District of Columbia
South Carolina
Louisiana
Arkansas
North Carolina
Georgia
Alaska
Vermont
Oklahoma
South Dakota
Maine
Tennessee
West Virginia
Kentucky
Virginia
Maryland
Montana
North Dakota
Missouri
Michigan
Delaware
Iowa
Indiana
New York
New Mexico
Wisconsin
Kansas
Ohio
New Hampshire
Wyoming
Texas
Illinois
Idaho
Minnesota
California
Nebraska
Pennsylvania
Florida
New Jersey
Oregon
Washington
Nevada
Connecticut
Arizona
Colorado
Massachusetts
Rhode Island
Utah

K = 2



PC2
PC1
cluster
1
2
30
20
10
0
-10
-20
-40
-20
0
20
40
Mississippi
Hawaii
Alabama
South Carolina
District of Columbia
Louisiana
Arkansas
North Carolina
Georgia
Alaska
Vermont
Oklahoma
South Dakota
Maine
Tennessee
West Virginia
Kentucky
Virginia
Maryland
Montana
North Dakota
Missouri
Michigan
Delaware
Iowa
Indiana
New York
New Mexico
Wisconsin
Kansas
Ohio
New Hampshire
Wyoming
Texas
Illinois
Idaho
Nebraska
Pennsylvania
California
Minnesota
Washington
Florida
New Jersey
Oregon
Connecticut
Nevada
Colorado
Arizona
Massachusetts
Rhode Island
Utah

K = 1



Mississippi
Hawaii
Alabama
South Carolina
District of Columbia
Louisiana
Arkansas
North Carolina
Georgia
Alaska
Vermont
Maine
South Dakota
Oklahoma
Tennessee
West Virginia
Kentucky
Virginia
Maryland
Montana
North Dakota
Missouri
Michigan
Delaware
Iowa
Indiana
New York
New Mexico
Wisconsin
Kansas
Ohio
New Hampshire
Wyoming
Texas
Illinois
Idaho
Minnesota
California
Nebraska
Pennsylvania
Florida
New Jersey
Oregon
Washington
Nevada
Connecticut
Arizona
Colorado
Massachusetts
Rhode Island
Utah
PC1
PC2
cluster
1

Variation with K = 1

```
names(km1)
```

```
## [1] "cluster"      "centers"      "totss"        "with
## [5] "tot.withinss" "betweenss"    "size"         "iter
## [9] "ifault"
```

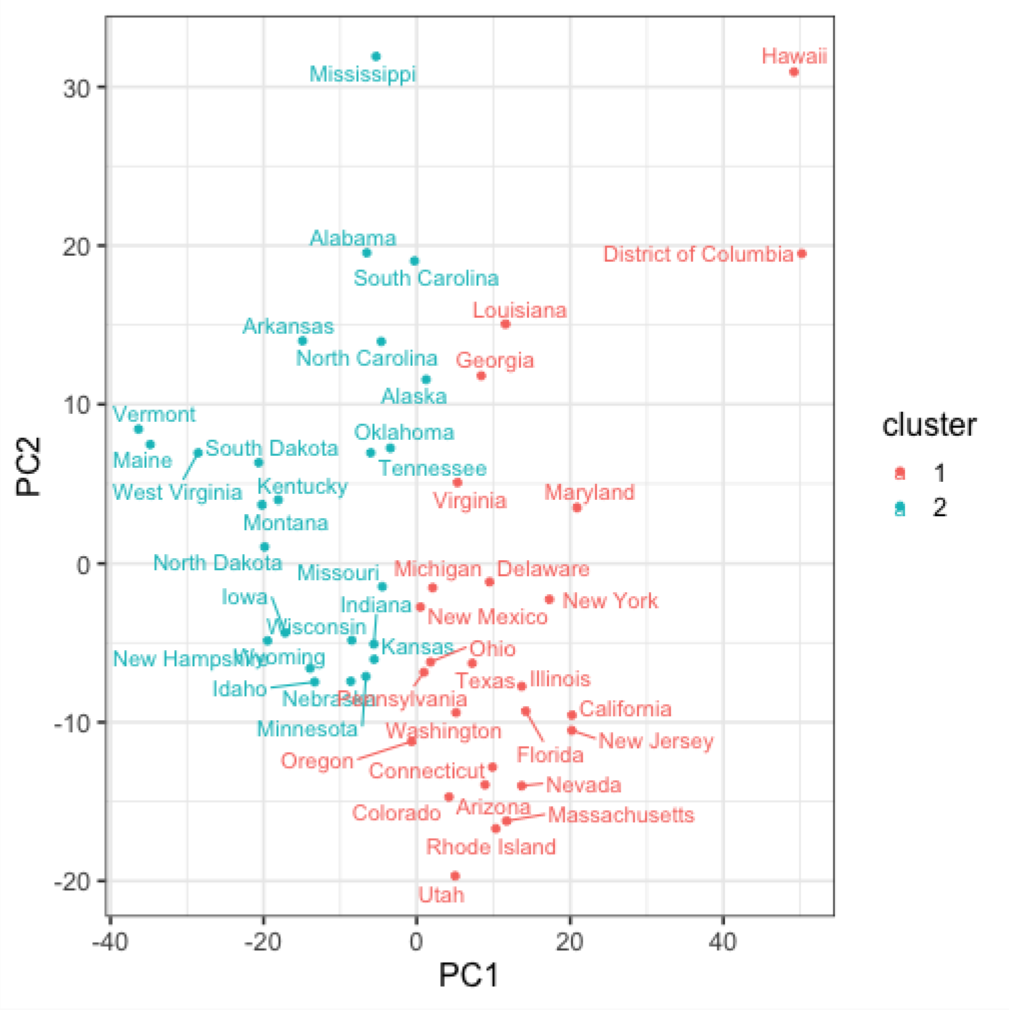
```
km1$withinss
```

```
## [1] 22776.26
```

```
km1$tot.withinss
```

```
## [1] 22776.26
```

K = 2



Mississippi
Hawaii
Alabama
South Carolina
District of Columbia
Louisiana
Arkansas
North Carolina
Georgia
Alaska
Vermont
Oklahoma
Maine
South Dakota
Tennessee
West Virginia
Kentucky
Virginia
Maryland
Montana
North Dakota
Missouri
Michigan
Delaware
Iowa
Indiana
New York
New Mexico
Wisconsin
Kansas
Ohio
New Hampshire
Wyoming
Texas
Illinois
Idaho
Nebraska
Pennsylvania
California
Minnesota
Washington
New Jersey
Florida
Oregon
Connecticut
Nevada
Colorado
Arizona
Massachusetts
Rhode Island
Utah
30
20
10
0
-10
-20
-40
-20
0
20
40
PC2
PC1
cluster
1
2

Variation with K = 2

```
km2$withinss
```

```
## [1] 8257.379 5480.290
```

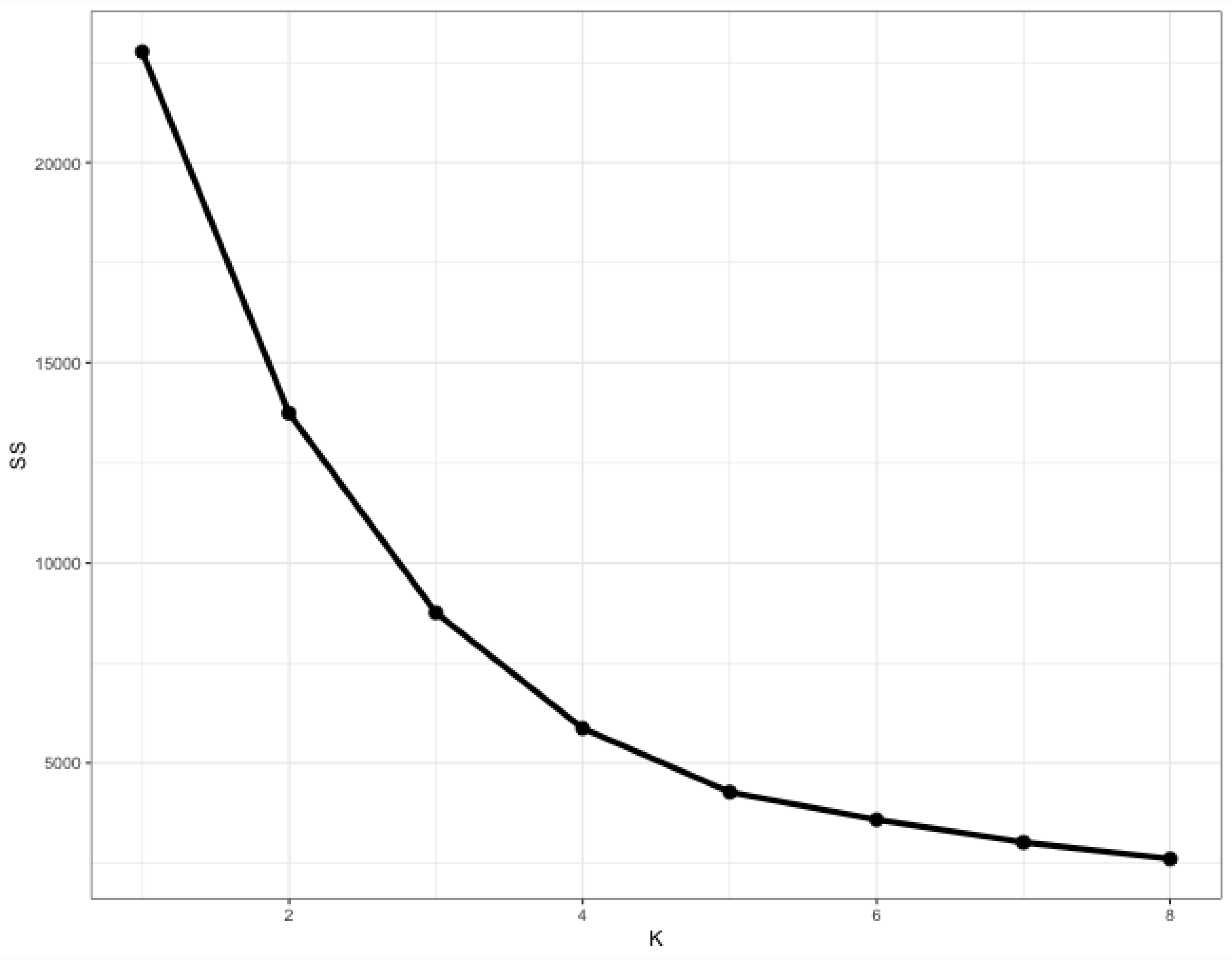
```
km2$tot.withinss
```

```
## [1] 13737.67
```

```
km2$totss
```

```
## [1] 22776.26
```

TWSS and K



Selecting K

- Use domain area knowledge.
- Look for "elbow" in a scree plot.
- Formalize "elbow" with Gap statistic (Tibshirani, 2001).

The number of clusters is often ambiguous, which shouldn't be surprising in an unsupervised setting.

Choice of K is choosing where on the spectrum between complete aggregation (K = 1) and no aggregation (K = n).