



UNSUPERVISED LEARNING IN R

Welcome to the course!

Chapter 1 overview

- Unsupervised learning
- Three major types of machine learning
- Execute one type of unsupervised learning using R

Types of machine learning

- Unsupervised learning
 - Finding structure in unlabeled data
- Supervised learning
 - Making predictions based on labeled data
 - Predictions like regression or classification
- Reinforcement learning

Labeled vs. unlabeled data

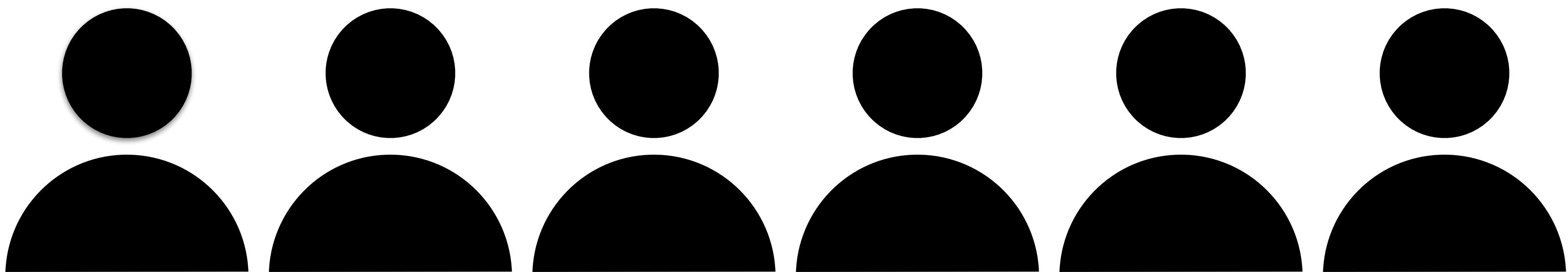
← Features →

Observations

Color	Shape	Size (cm)	Label
Blue	Square	10	Group
Red	Ellipse	2.4	1
Red	Ellipse	20.7	1
Unlabeled data			2

Unsupervised learning

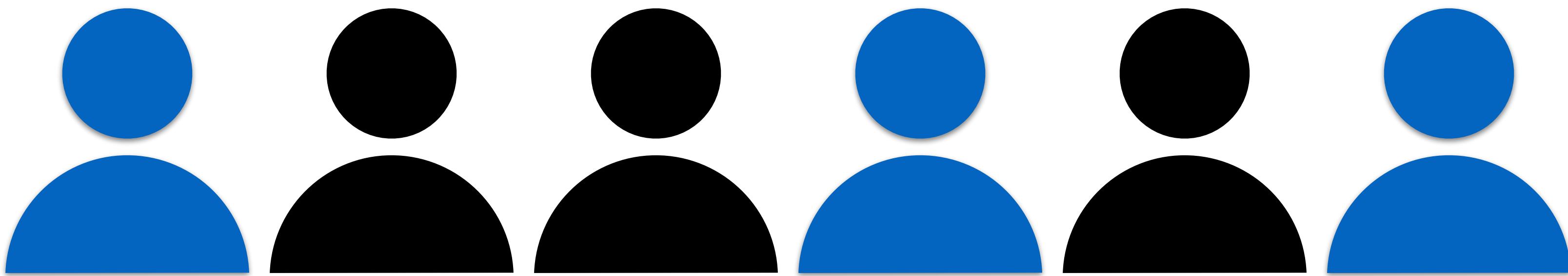
- Finding homogenous subgroups within larger group



People have features such as income, education attainment, and gender

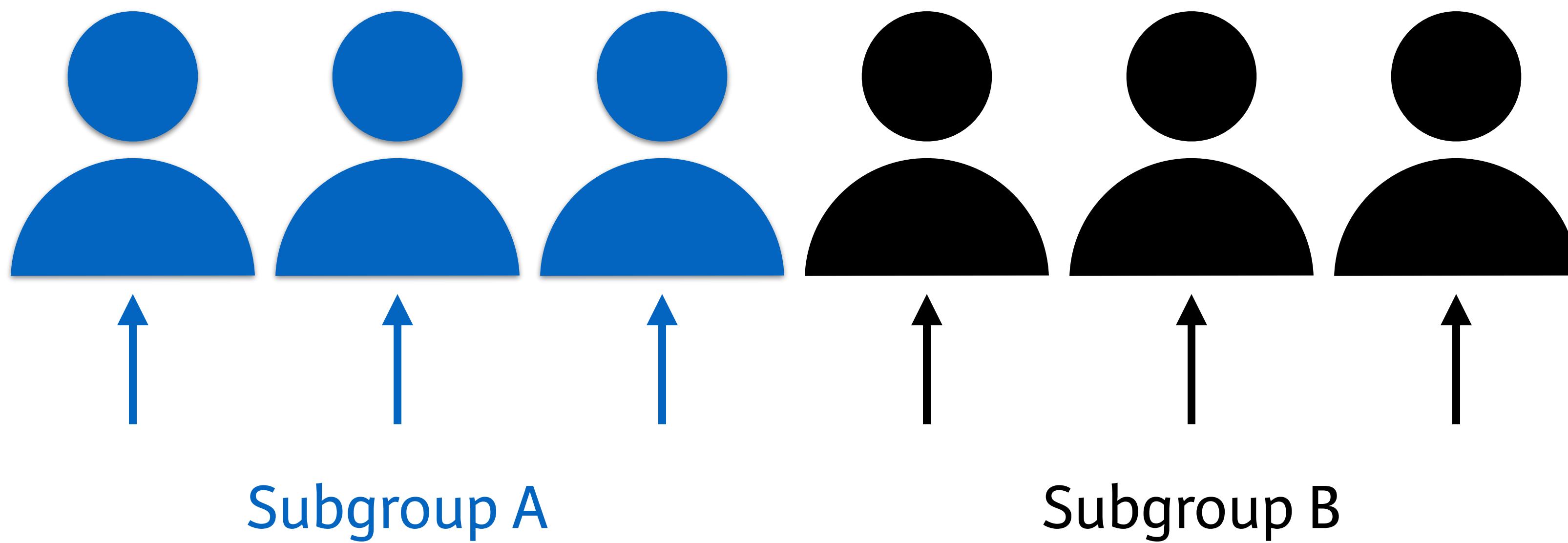
Unsupervised learning

- Finding homogenous subgroups within larger group



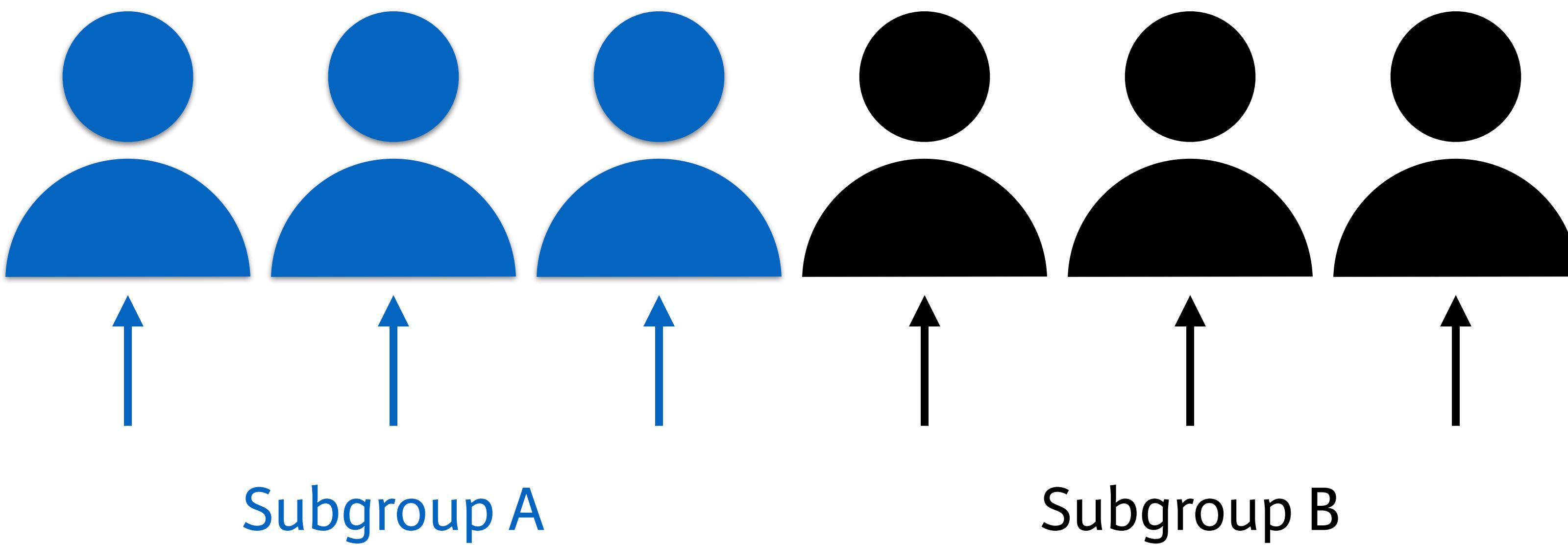
Unsupervised learning

- Finding homogenous subgroups within larger group



Unsupervised learning

- Finding homogenous subgroups within larger group
 - Clustering



Clustering examples



Unsupervised learning

- Finding homogenous subgroups within larger group
 - Clustering
- Finding patterns in the features of the data
 - Dimensionality reduction

Dimensionality reduction

- Find patterns in the features of the data
- Visualization of high dimensional data
- Pre-processing before supervised learning

Challenges and benefits

- No single goal of analysis
- Requires more creativity
- Much more unlabeled data available than cleanly labeled data



UNSUPERVISED LEARNING IN R

Let's practice!

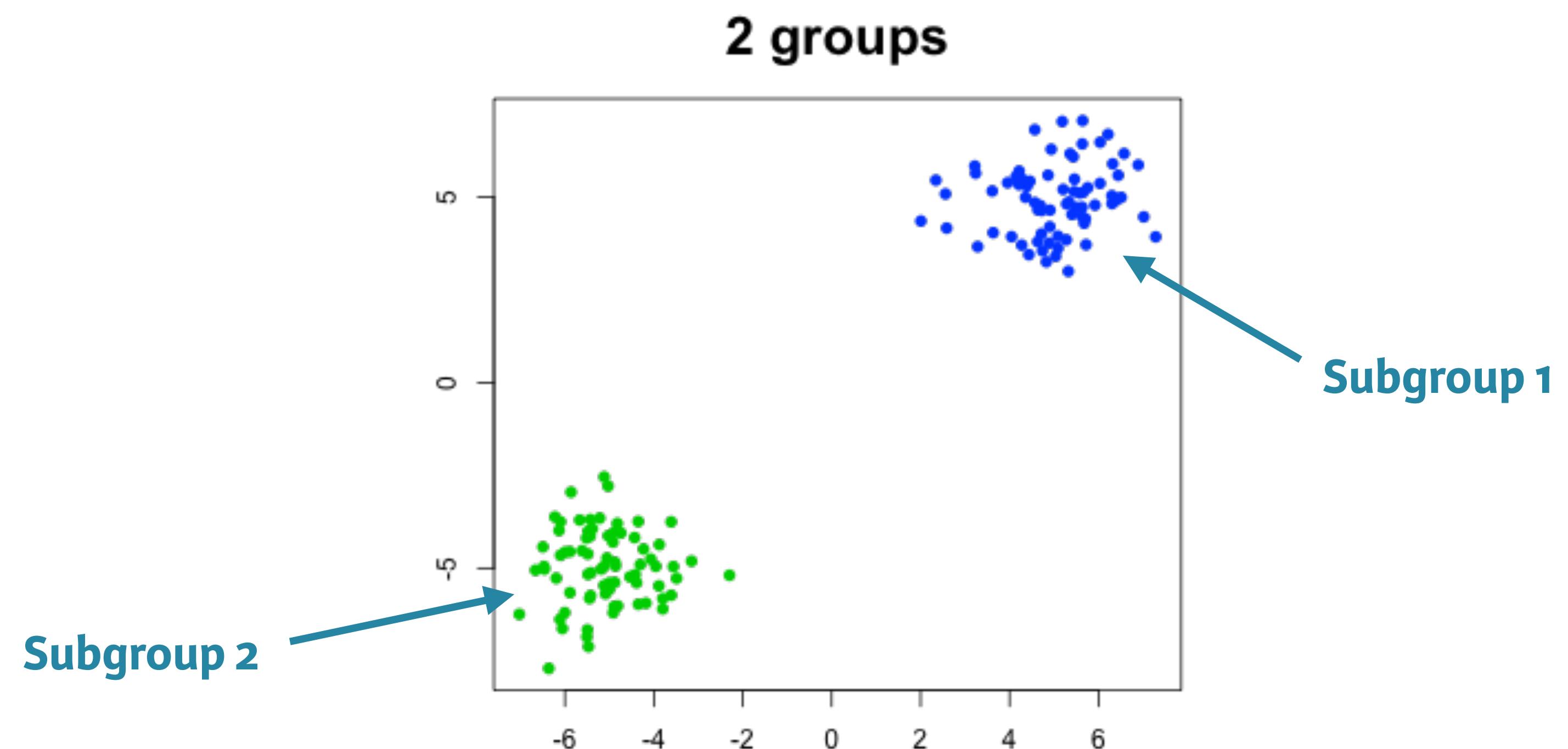


UNSUPERVISED LEARNING IN R

Introduction to k-means clustering

k-means clustering algorithm

- First of two clustering algorithms covered in this course
- Breaks observations into pre-defined number of clusters



k-means in R

```
> # k-means algorithm with 5 centers, run 20 times  
> kmeans(x, centers = 5, nstart = 20)
```

- One observation per row, one feature per column
- k-means has a random component
- Run algorithm multiple times to improve odds of the best model

First exercises

- First exercise uses synthetic data
- Synthetic data generated from 3 subgroups
- Selecting the best number of subgroups for k-means
- Example with more fun data later in the chapter



UNSUPERVISED LEARNING IN R

Let's practice!



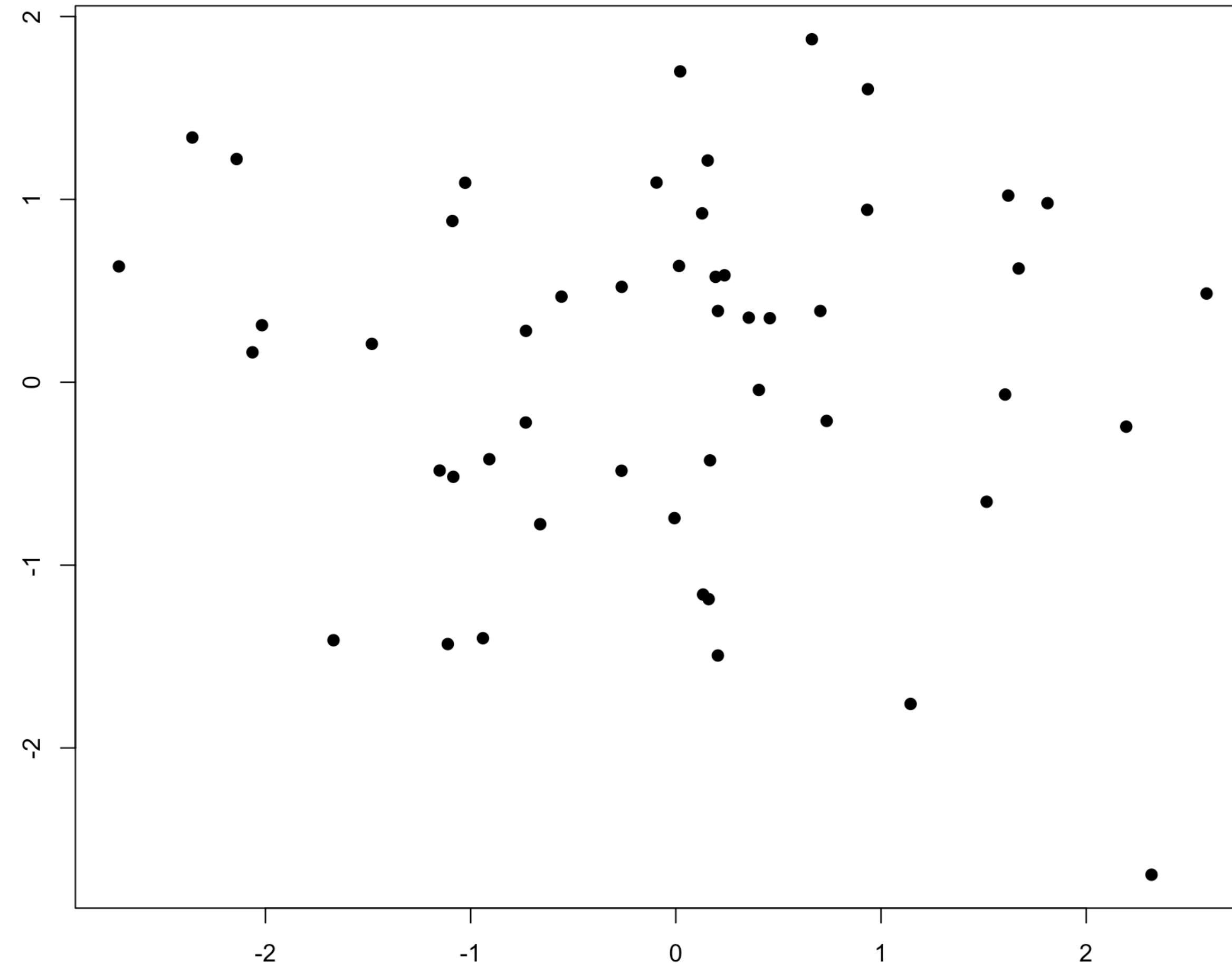
UNSUPERVISED LEARNING IN R

How kmeans() works and practical matters

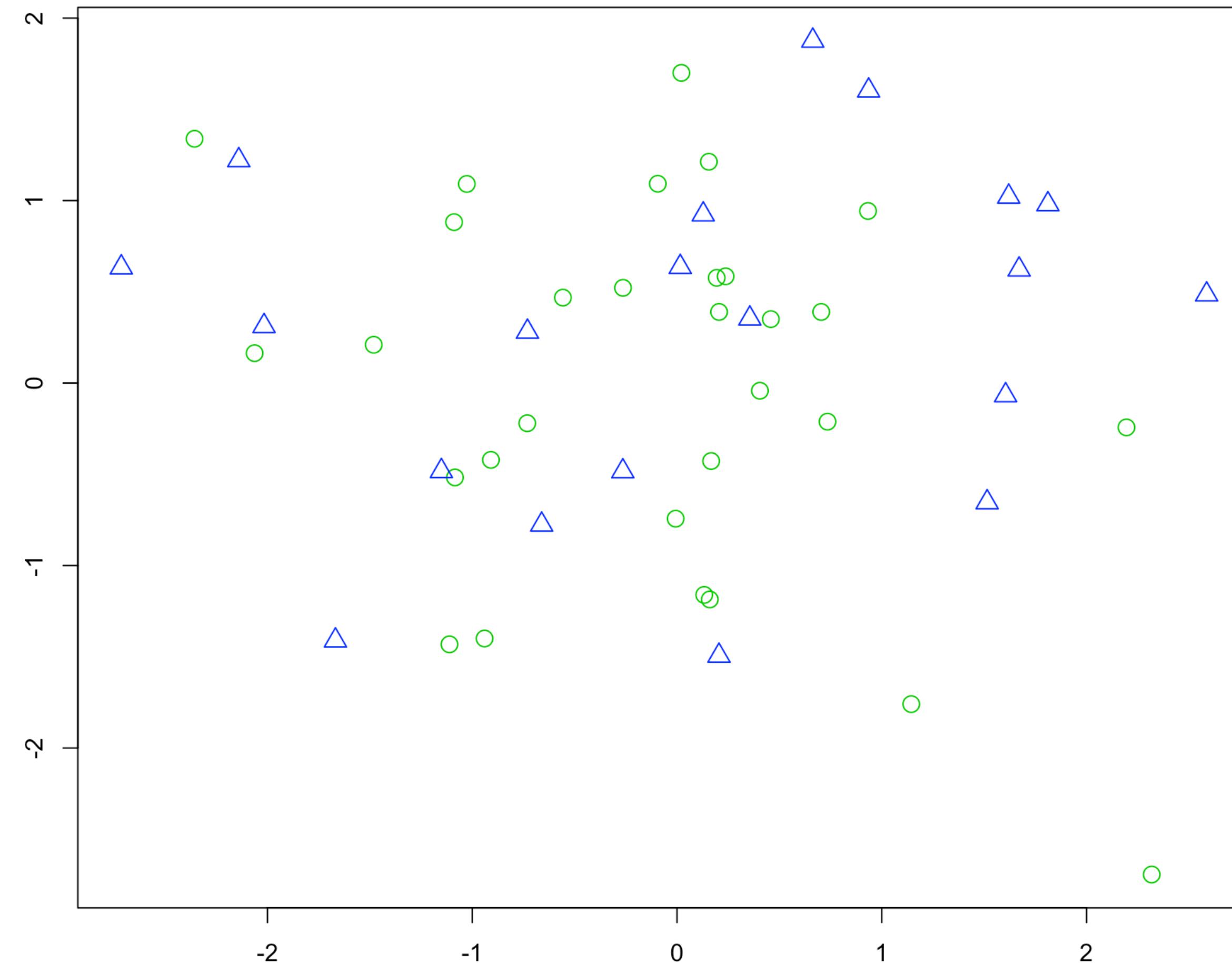
Objectives

- Explain how k-means algorithm is implemented visually
- Model selection: determining number of clusters

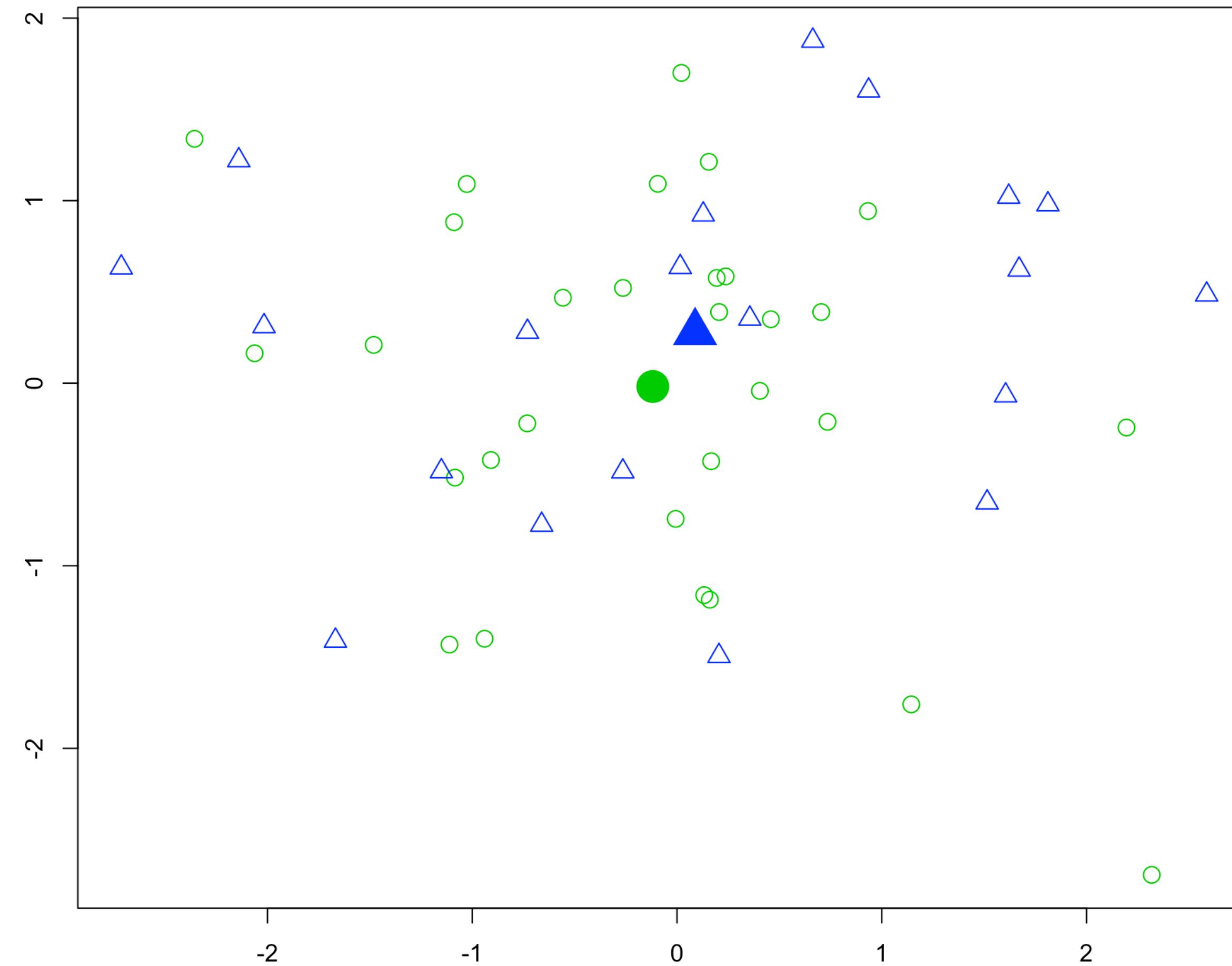
Observations



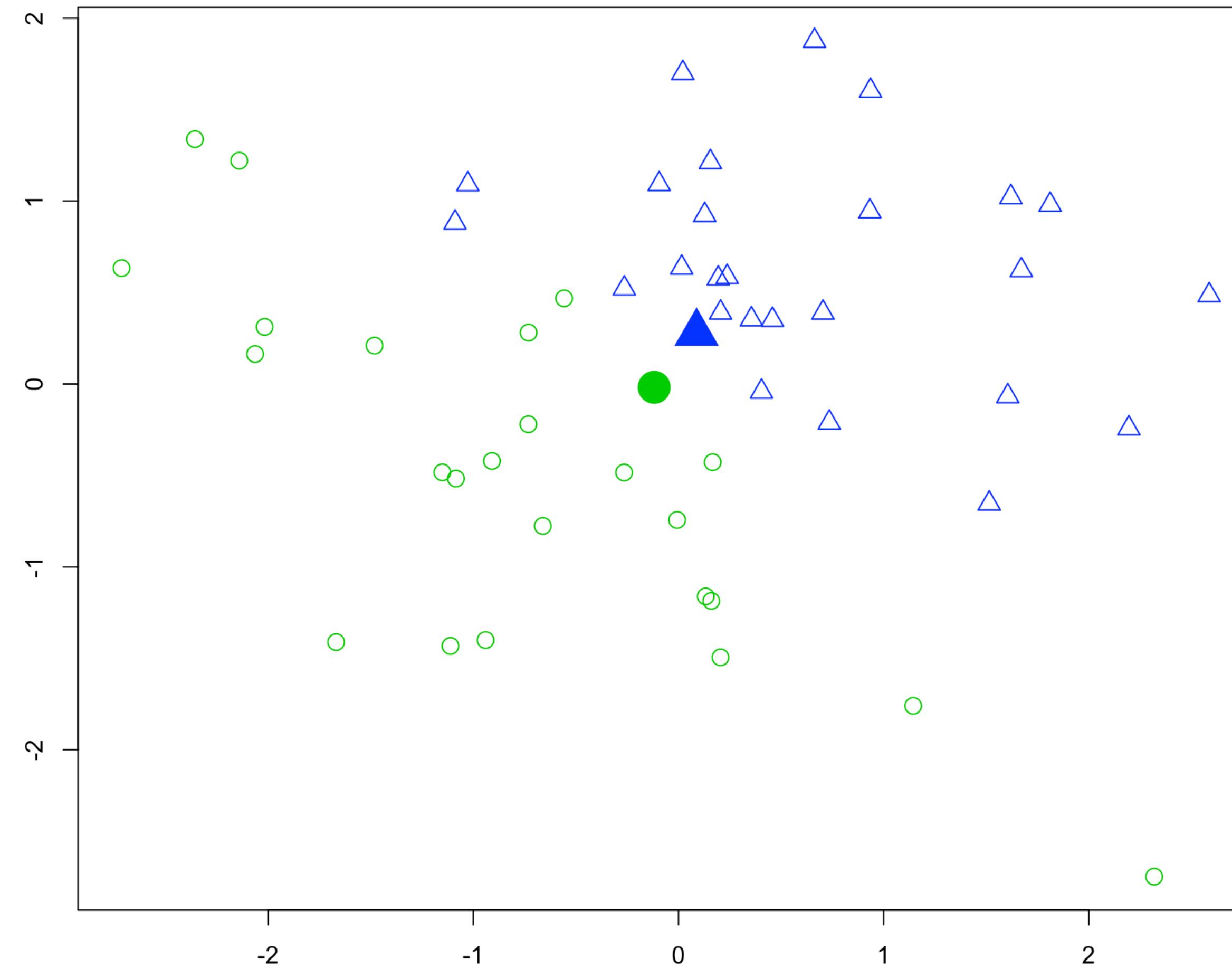
Random Cluster Assignment



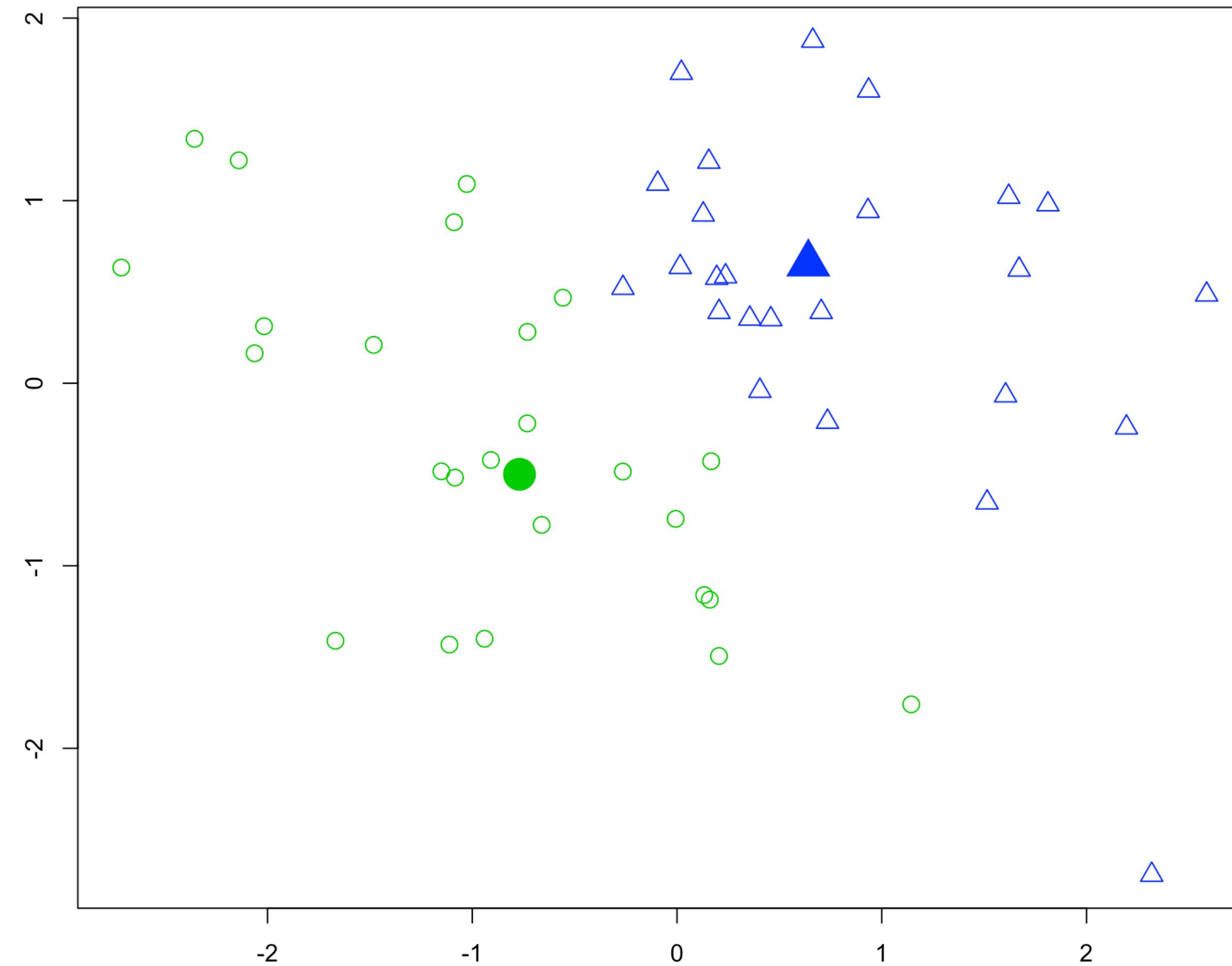
Cluster Centers Calculated



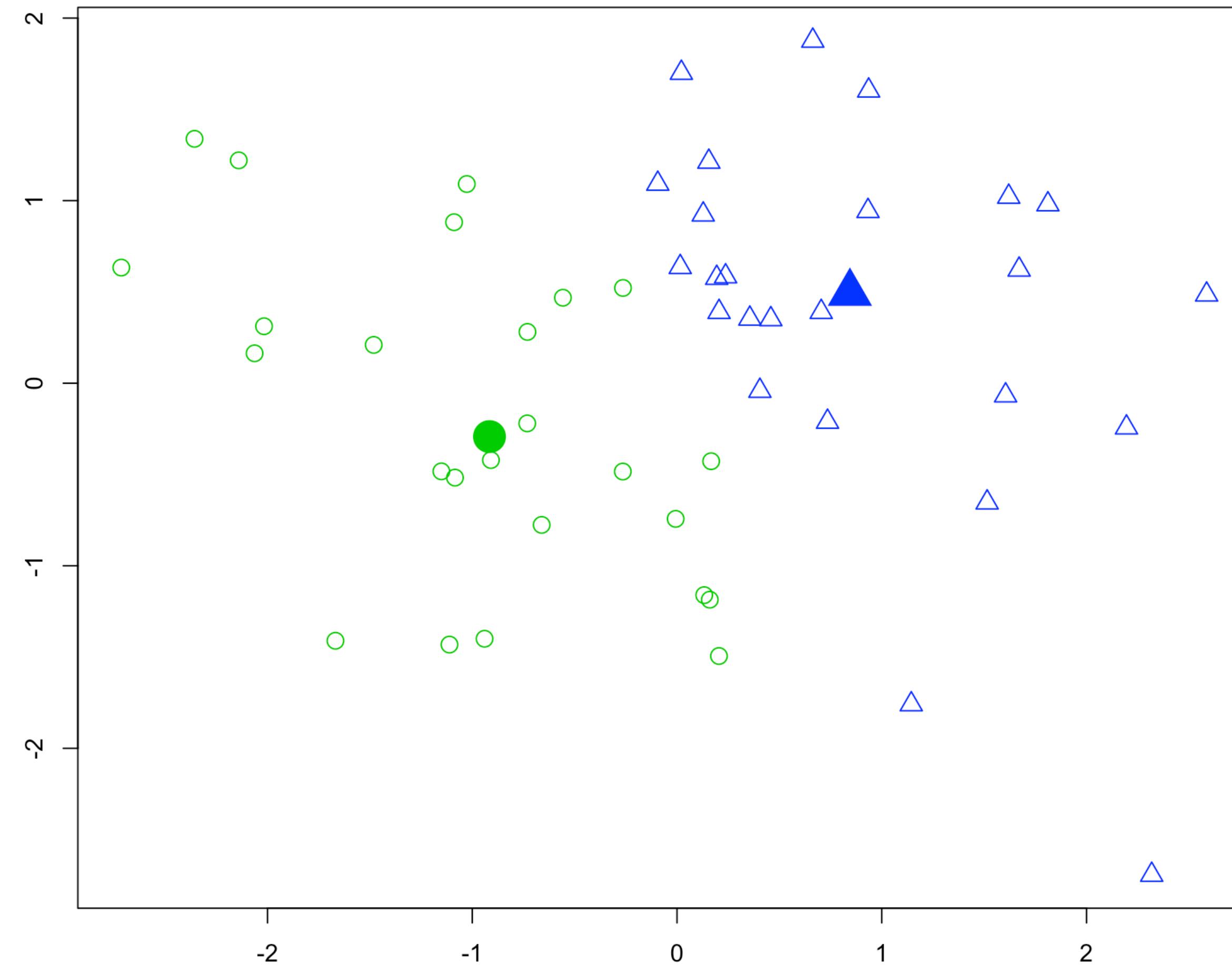
Iteration 1 - After Reassignment



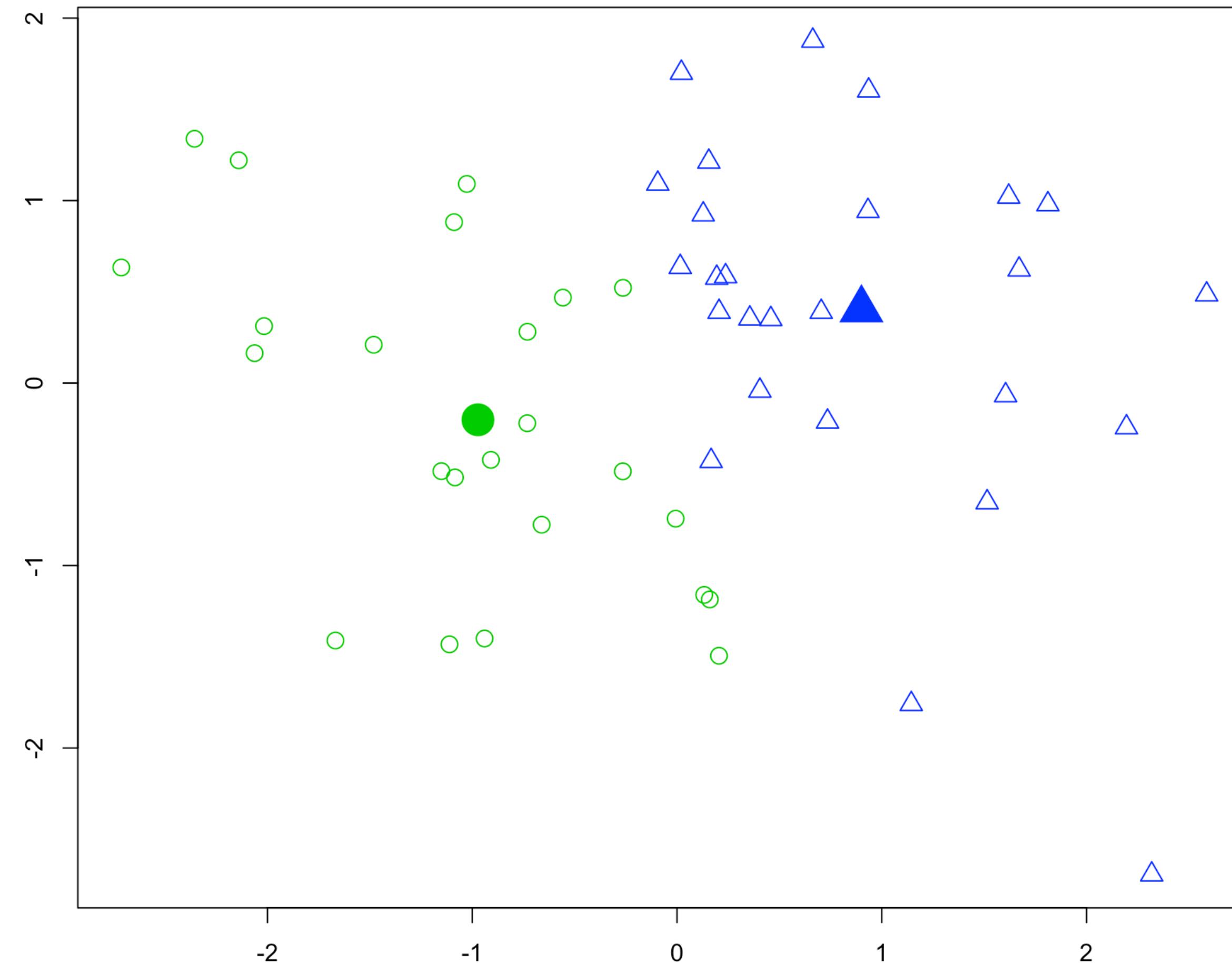
Iteration 2



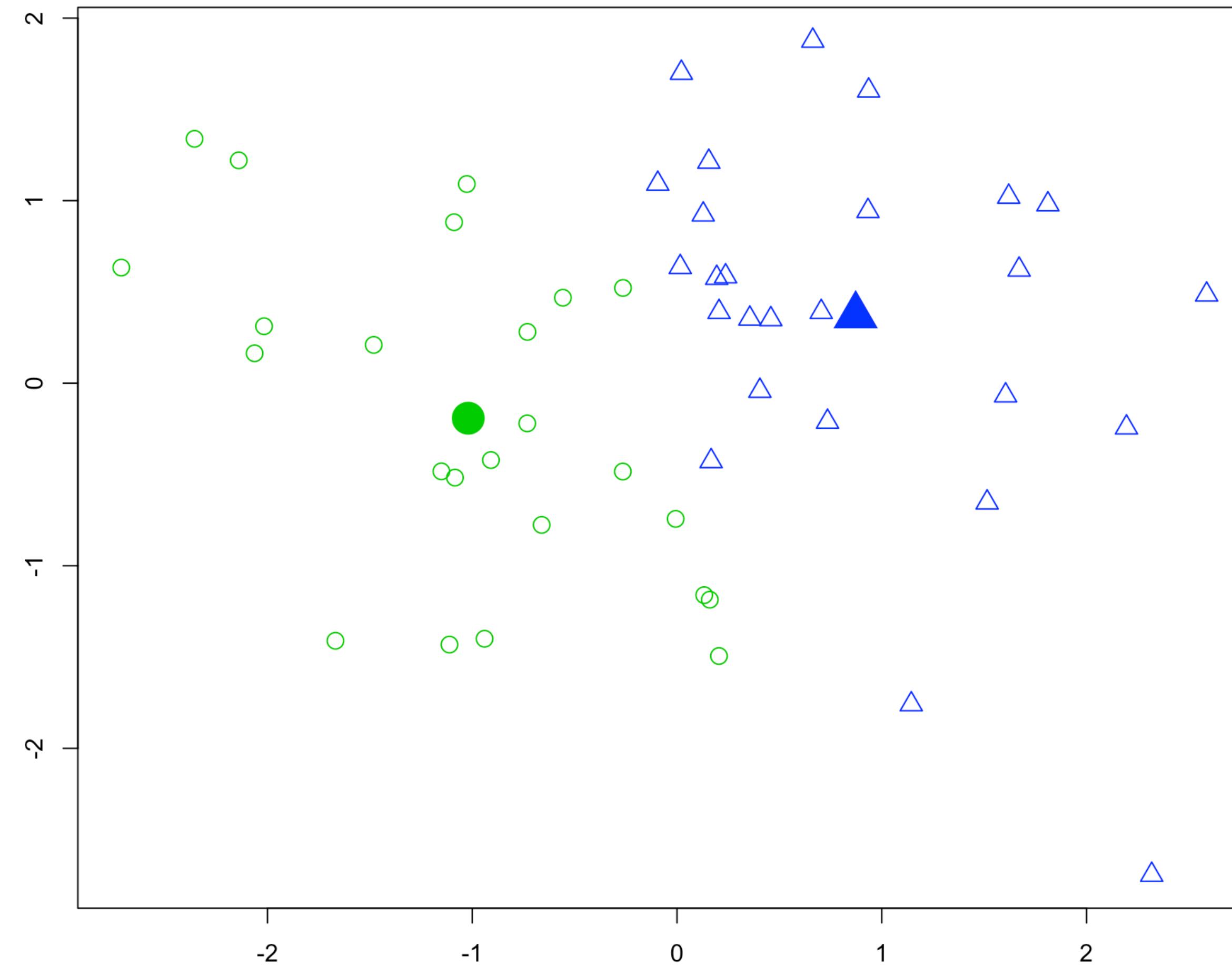
Iteration 3



Iteration 4



Iteration 5



Model selection

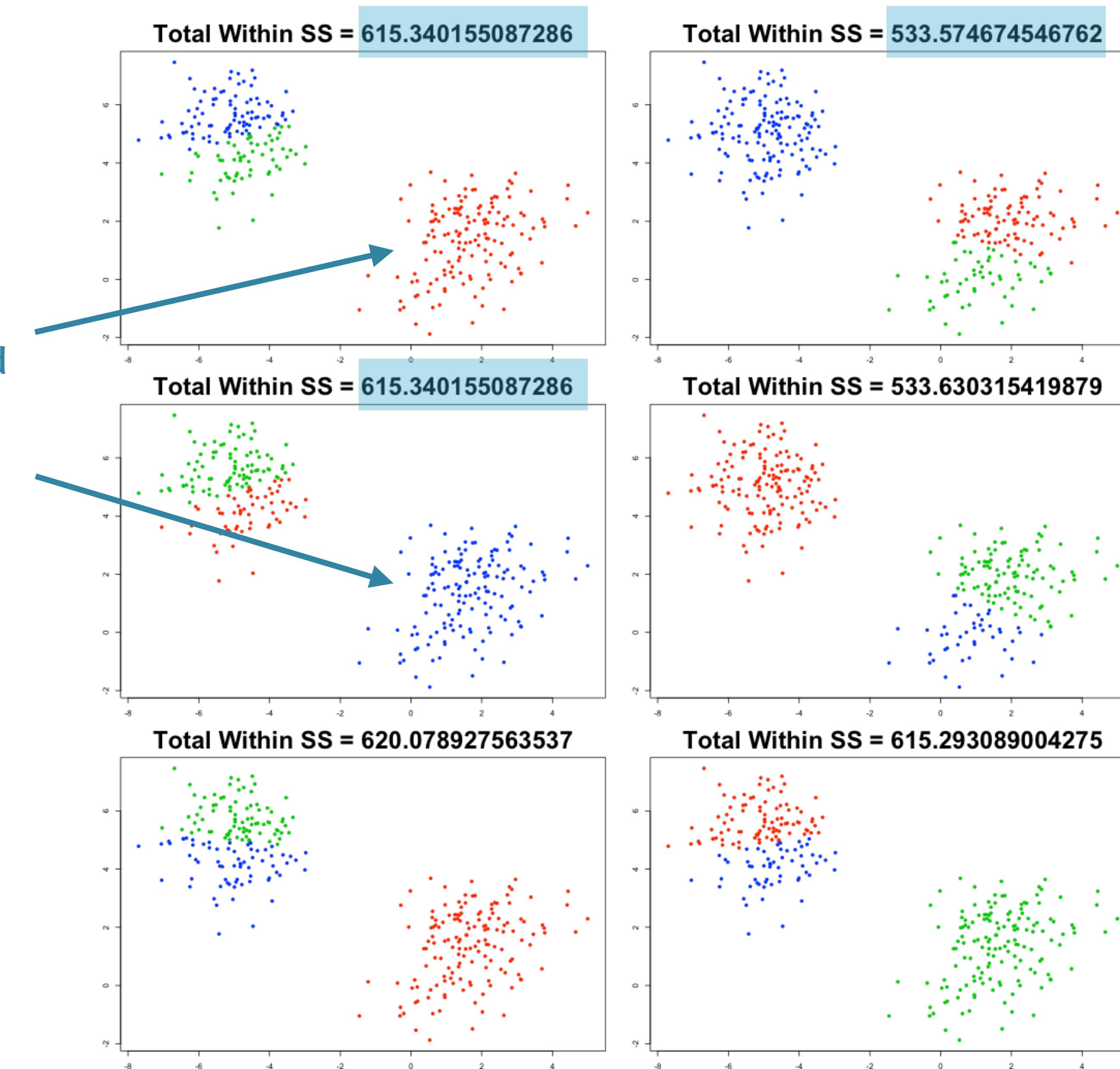
- Recall k-means has a random component
- Best outcome is based on *total within cluster sum of squares*:
 - For each cluster
 - For each observation in the cluster
 - Determine squared distance from observation to cluster center
 - Sum all of them together

Model selection

```
> # k-means algorithm with 5 centers, run 20 times  
> kmeans(x, centers = 5, nstart = 20)
```

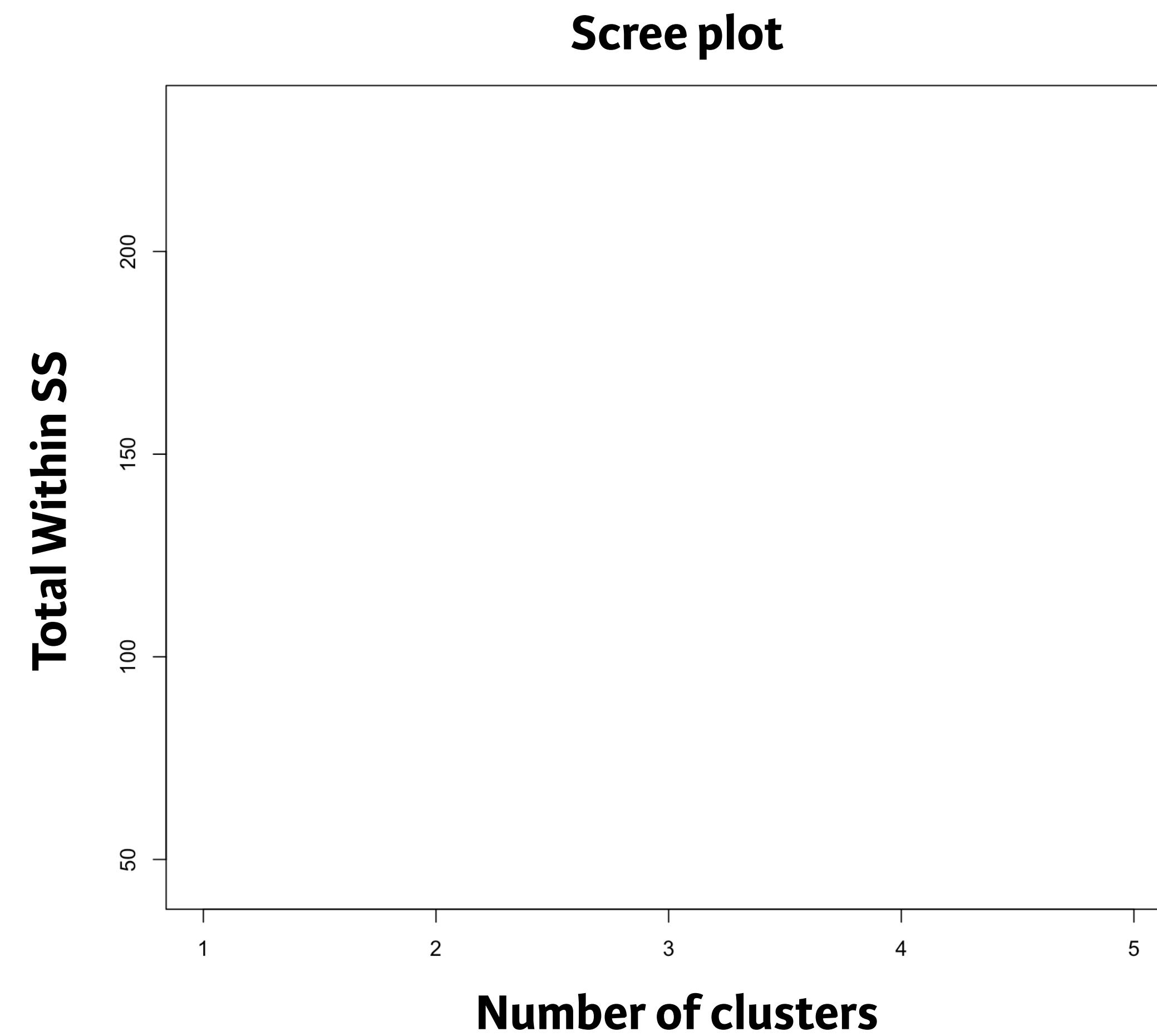
- Running algorithm multiple times helps find the global minimum total within cluster sum of squares
- You'll see an example in the exercises

Identical groupings and
Total Within SS, but
different cluster labels



Determining number of clusters

- Trial and error is not the best approach





UNSUPERVISED LEARNING IN R

Let's practice!



UNSUPERVISED LEARNING IN R

Introduction to Pokemon data

"Real" data exercise



The Pokemon dataset

```
> head(pokemon)
```

	HitPoints	Attack	Defense	SpecialAttack	SpecialDefense	Speed
[1,]	45	49	49	65	65	45
[2,]	60	62	63	80	80	60
[3,]	80	82	83	100	100	80
[4,]	80	100	123	122	120	80
[5,]	39	52	43	60	50	65
[6,]	58	64	58	80	65	80

- Hosted at <https://www.kaggle.com/abcsds/pokemon>
- More information on Pokemon and these features can be found at <http://pokemondb.net/pokedex>

Data challenges

- Selecting the variables to cluster upon
- Scaling the data (will handle in last chapter)
- Determining the number of clusters
 - Often no clean "elbow" in scree plot
 - This will be a core part of the exercises
- Visualize the results for interpretation



UNSUPERVISED LEARNING IN R

Let's practice!



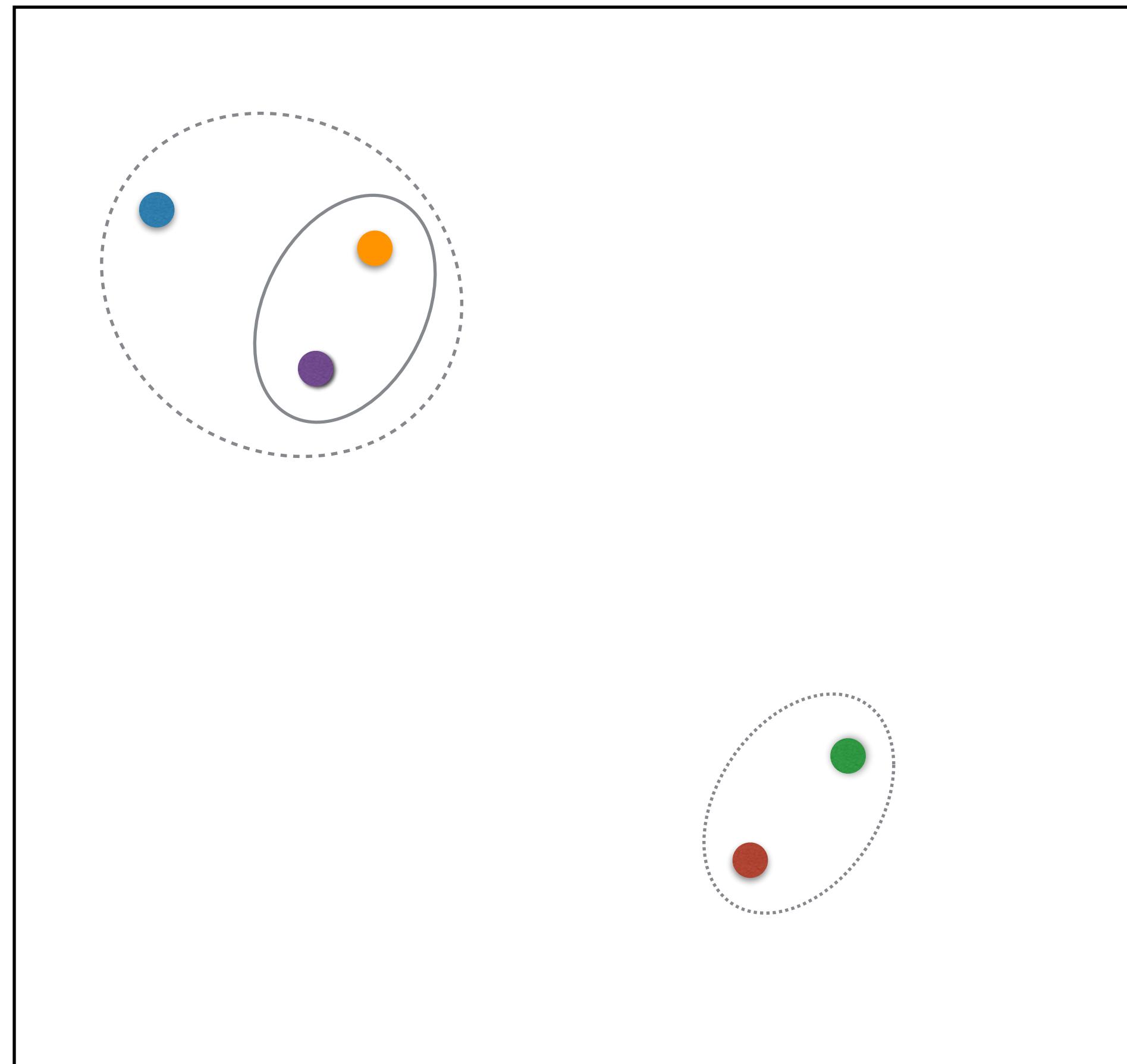
UNSUPERVISED LEARNING IN R

Review of k-means clustering

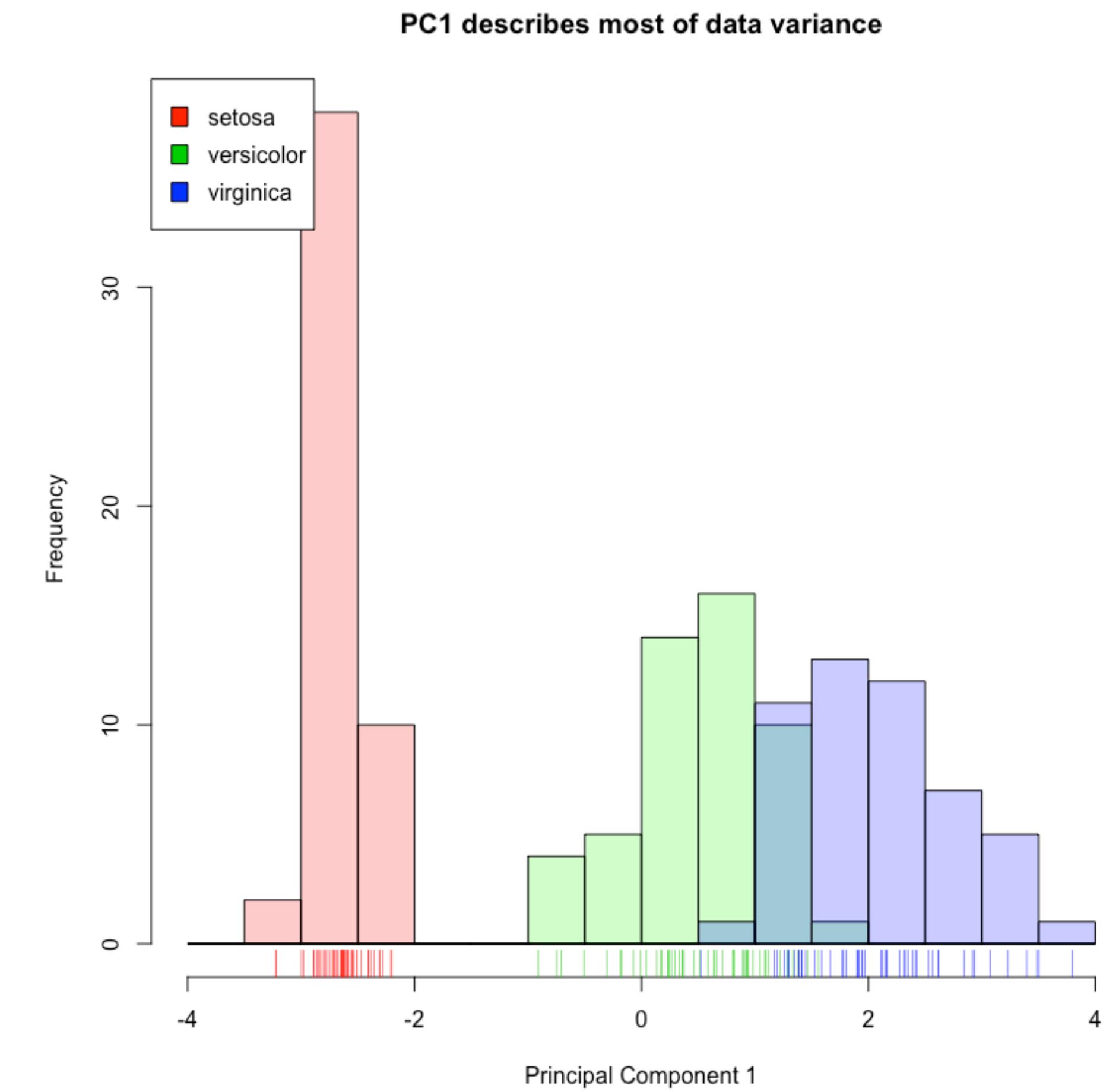
Chapter review

- Unsupervised vs. supervised learning
- How to create k-means cluster model in R
- How k-means algorithm works
- Model selection
- Application to "real" (and hopefully fun) dataset

Coming up: chapter 2

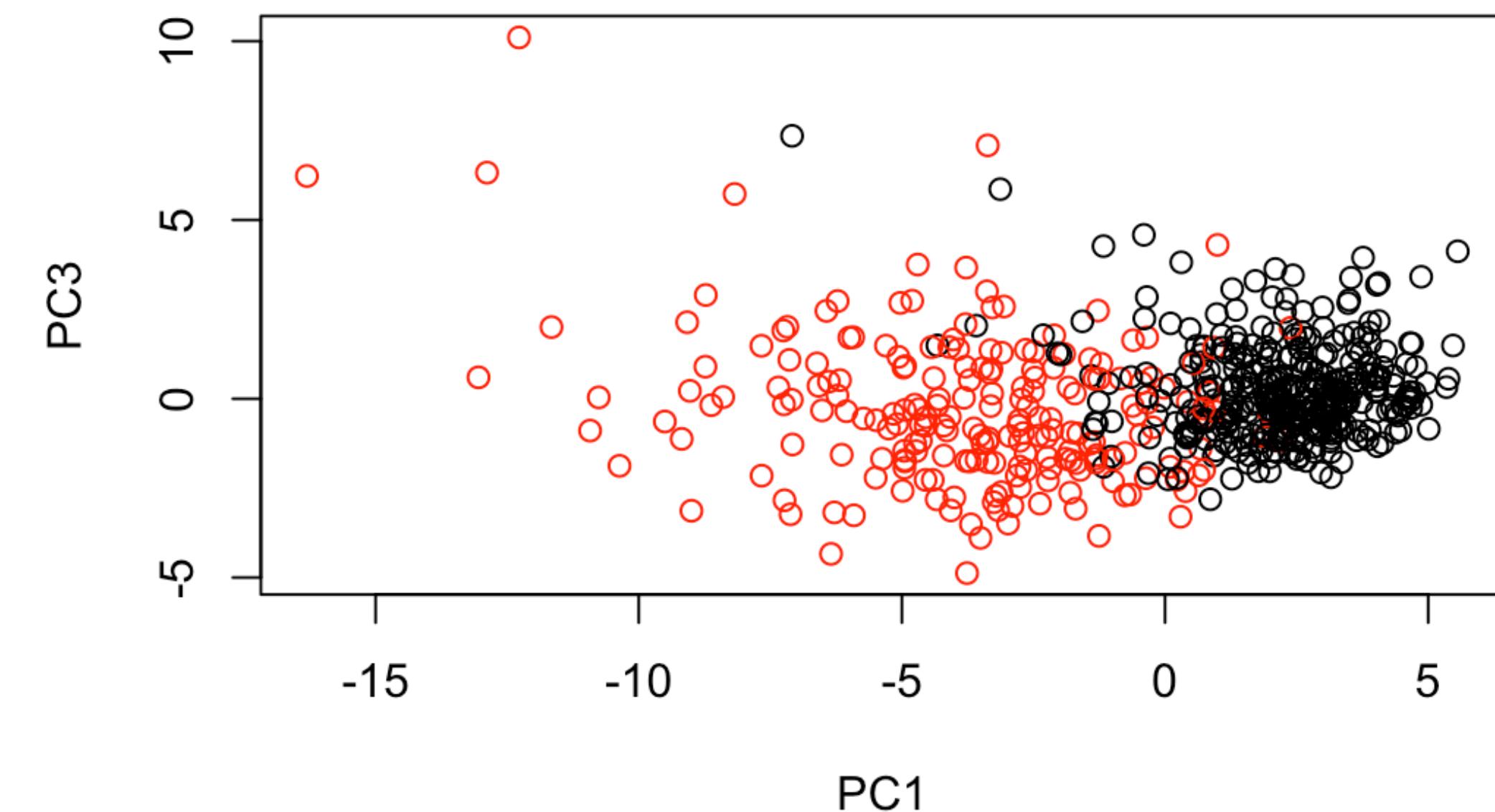


Coming up: chapter 3



Coming up: chapter 4

```
> # Repeat for components 1 and 3  
> plot(wisc.pr$x[, c(1, 3)], col = (diagnosis + 1),  
      xlab = "PC1", ylab = "PC3")
```





UNSUPERVISED LEARNING IN R

Let's practice!