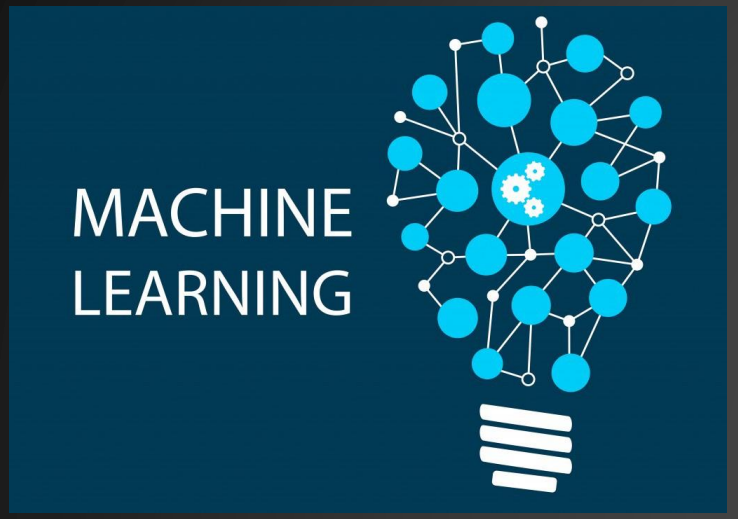
INTRODUCTION TO



FOR

USING

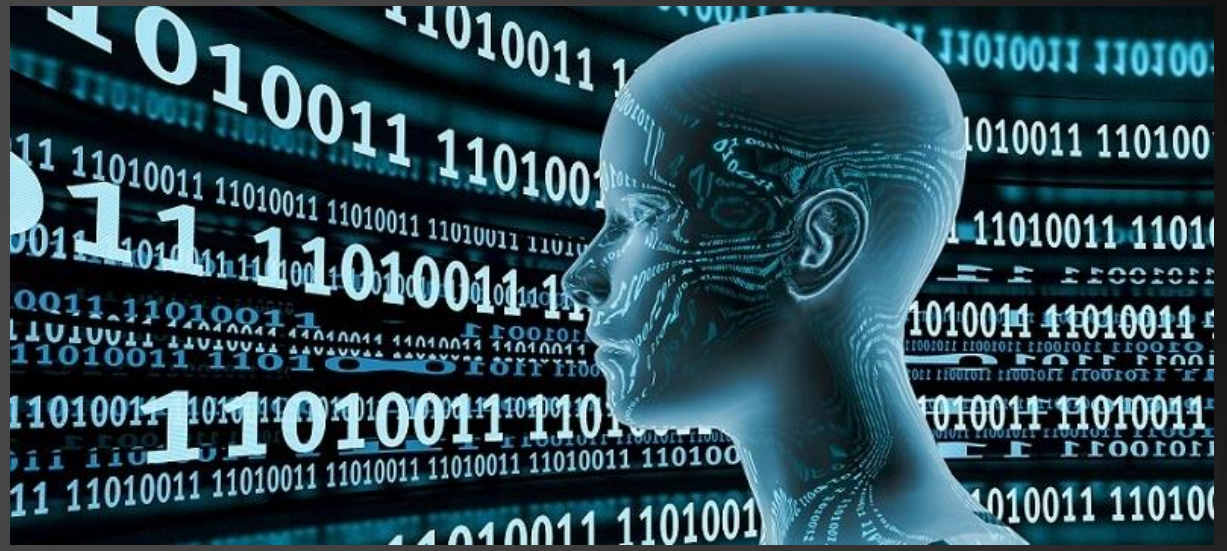
TUE VU

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CCIT\CITI

OUTLINES

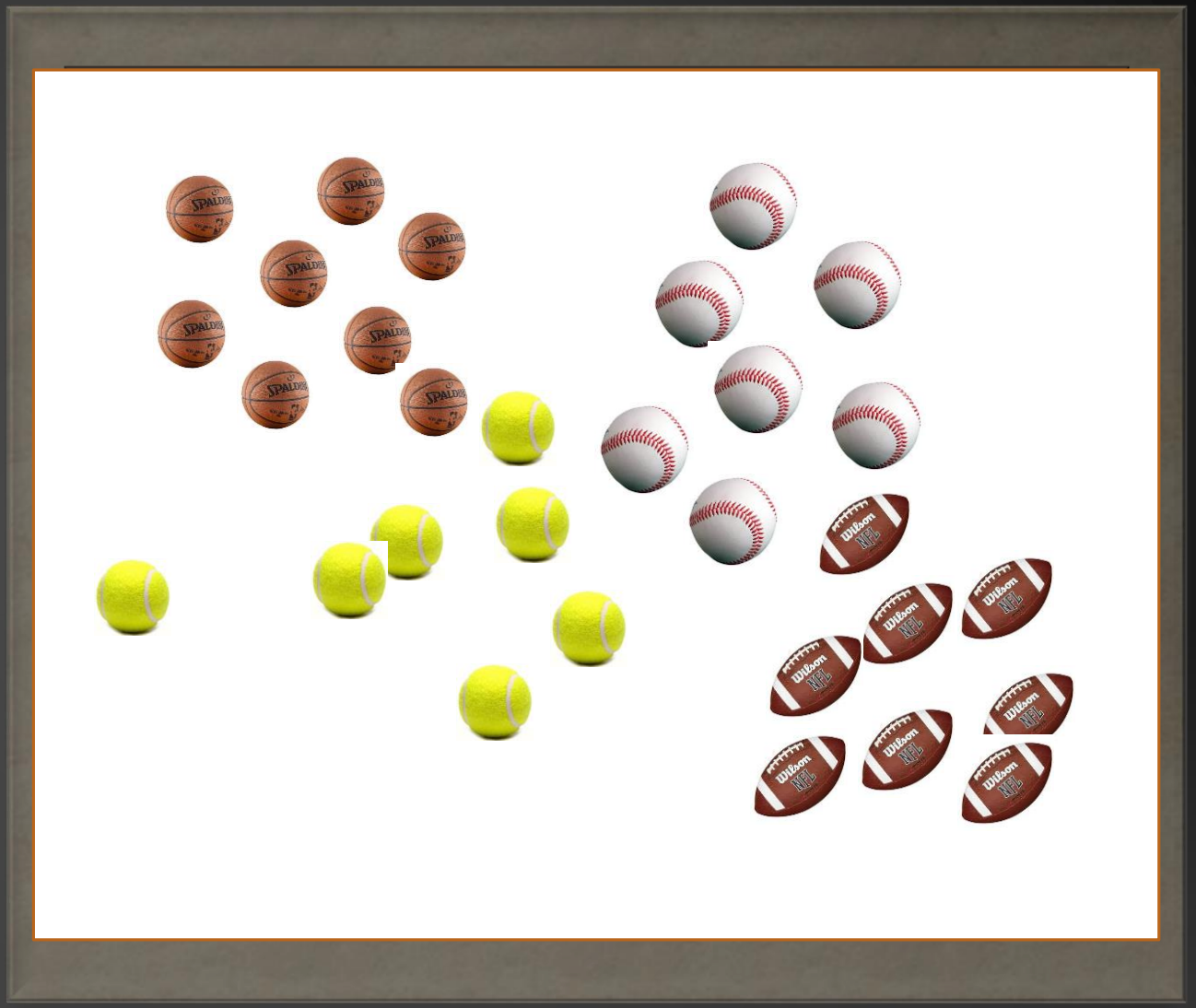
1. Introduction to Machine Learning
2. Why R
3. Types of Machine Learning
4. Caret package
5. Supervised Learning
6. Unsupervised Learning



6.1. Introduction

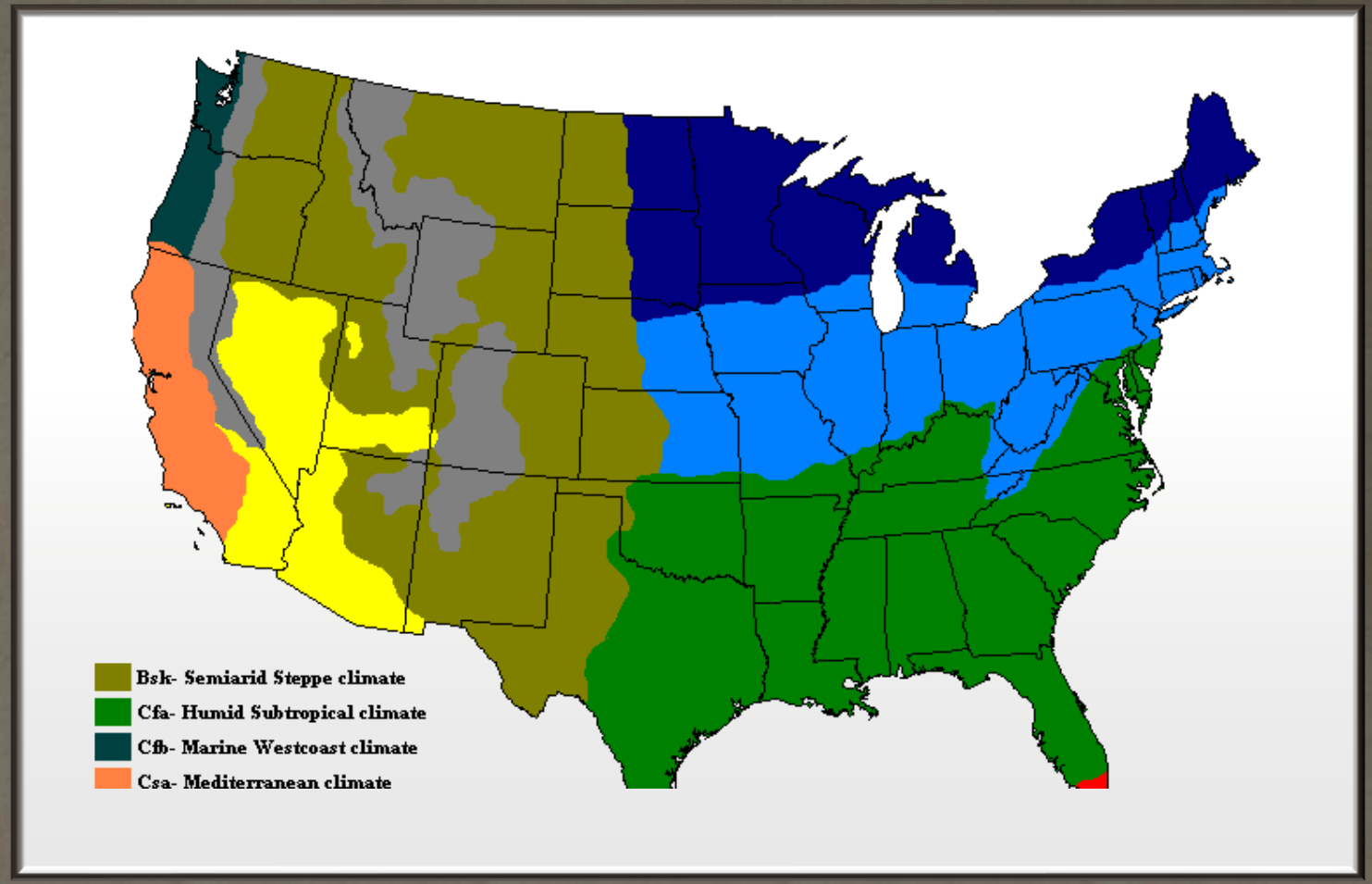
- Used when no feature output data
- Often used for clustering data

- K-means clustering
- Hierarchical clustering
- Ward clustering
- Partition Around Median (PAM)



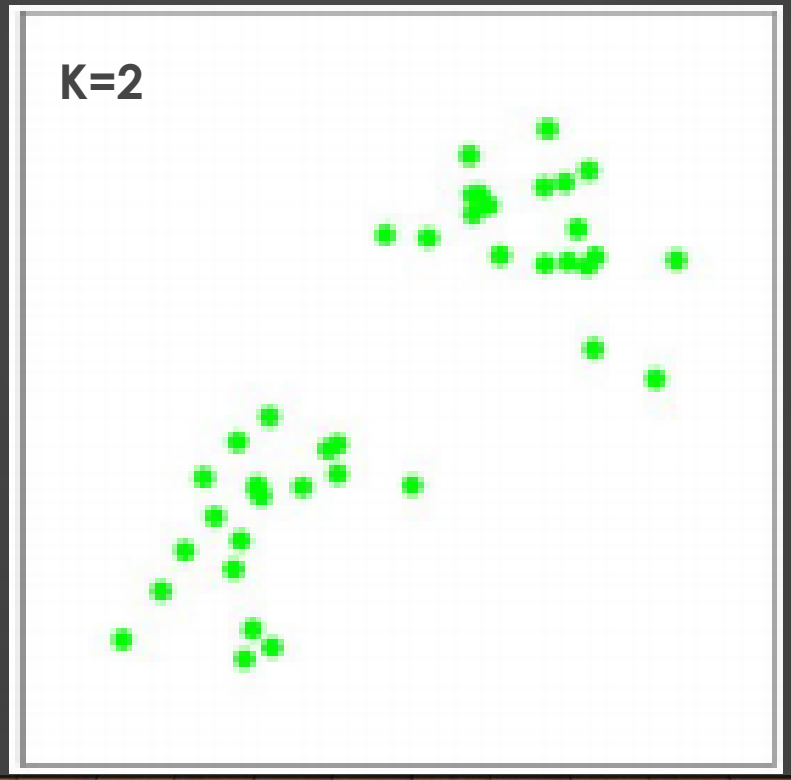
6. Unsupervised Learning

6.1. Introduction



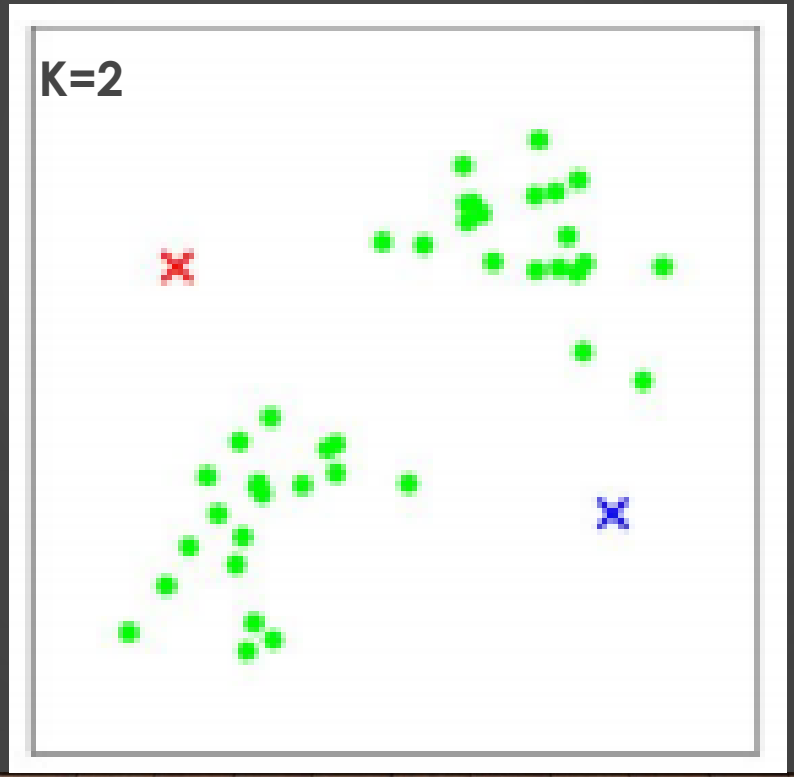
6. Unsupervised Learning

6.2. K-means clustering



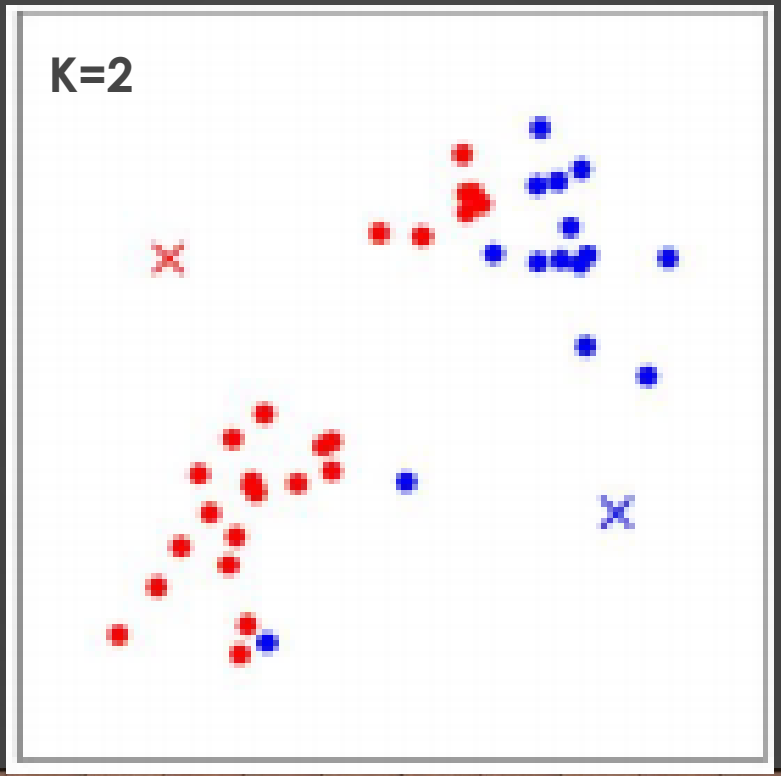
6. Unsupervised Learning

6.2. K-means clustering



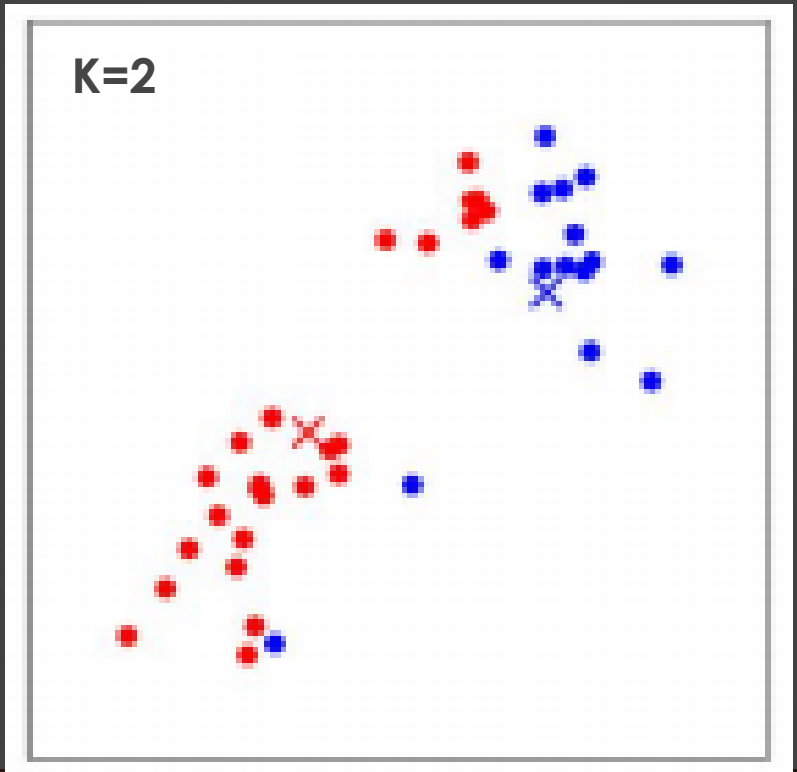
6. Unsupervised Learning

6.2. K-means clustering



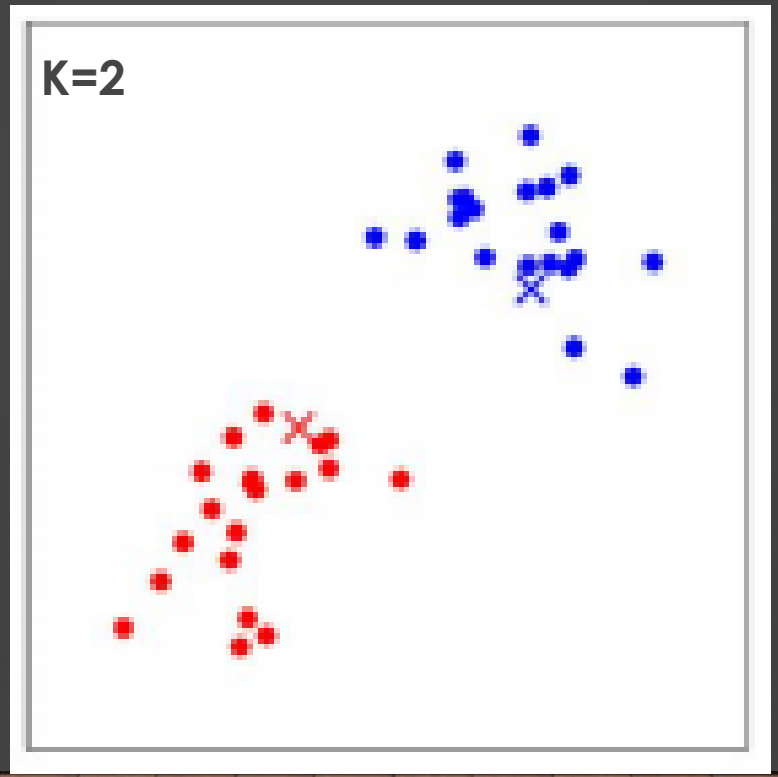
6. Unsupervised Learning

6.2. K-means clustering



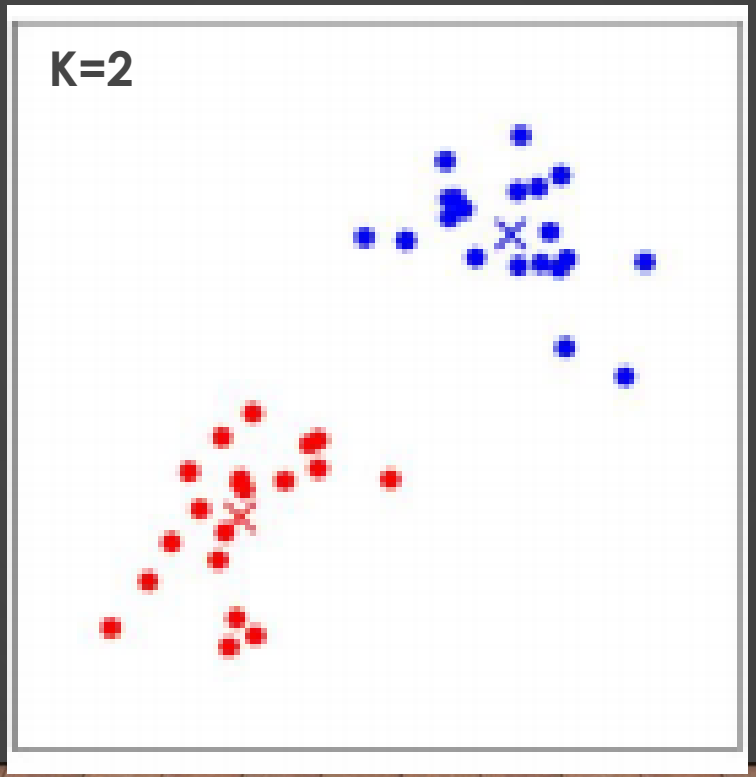
6. Unsupervised Learning

6.2. *K-means clustering*



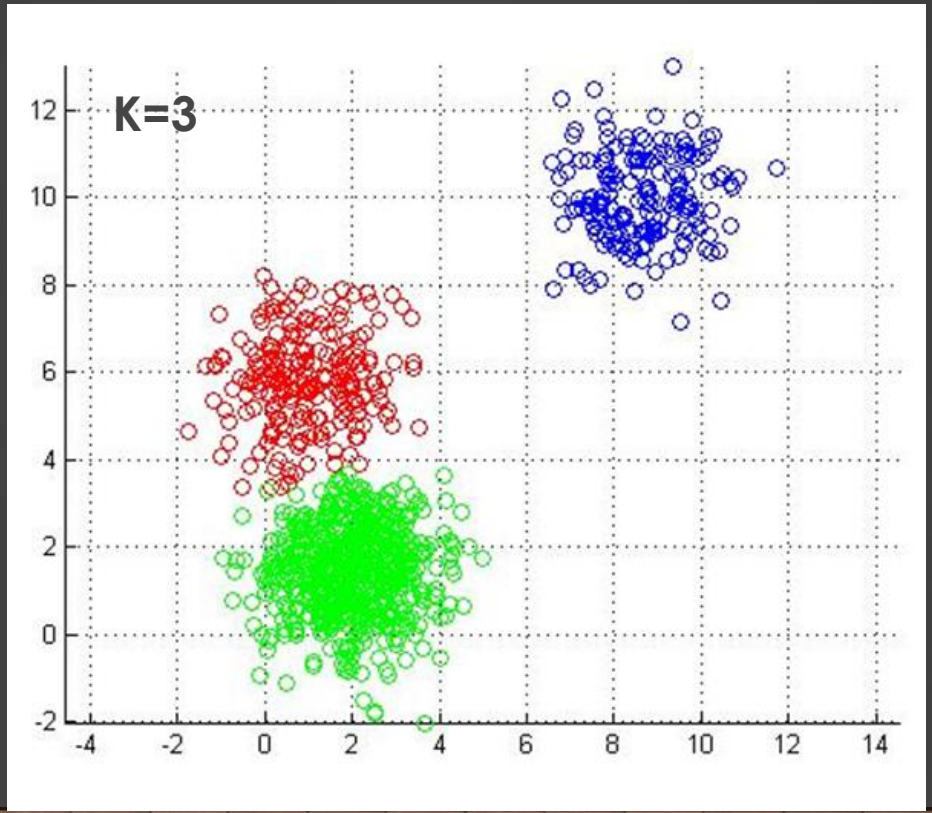
6. Unsupervised Learning

6.2. K-means clustering



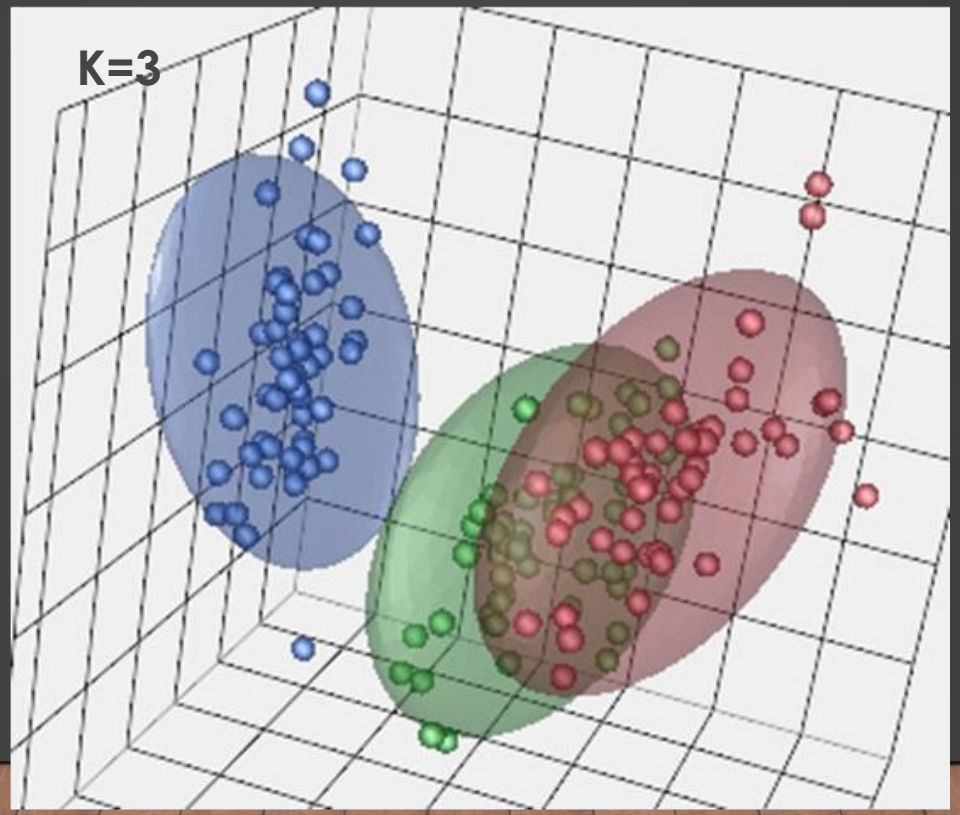
6. Unsupervised Learning

6.2. K-means clustering



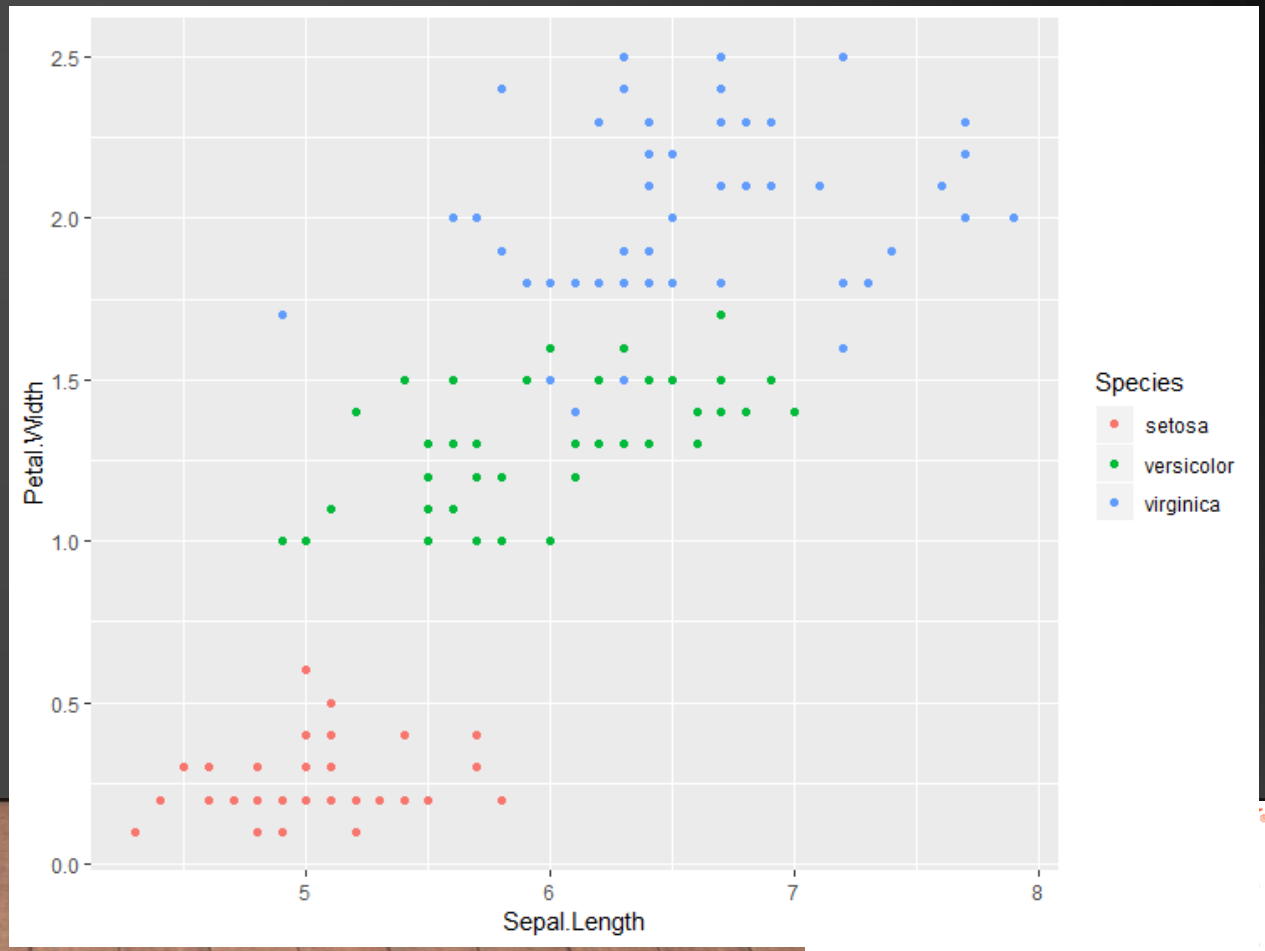
6. Unsupervised Learning

6.2. K-means clustering



6.2. *K-means clustering*

```
library(ggplot2)
library(factoextra)
data(iris)
ggplot(iris,aes(x=Sepal.Length,y=Petal.Width))+
      geom_point(aes(color=Species))
```



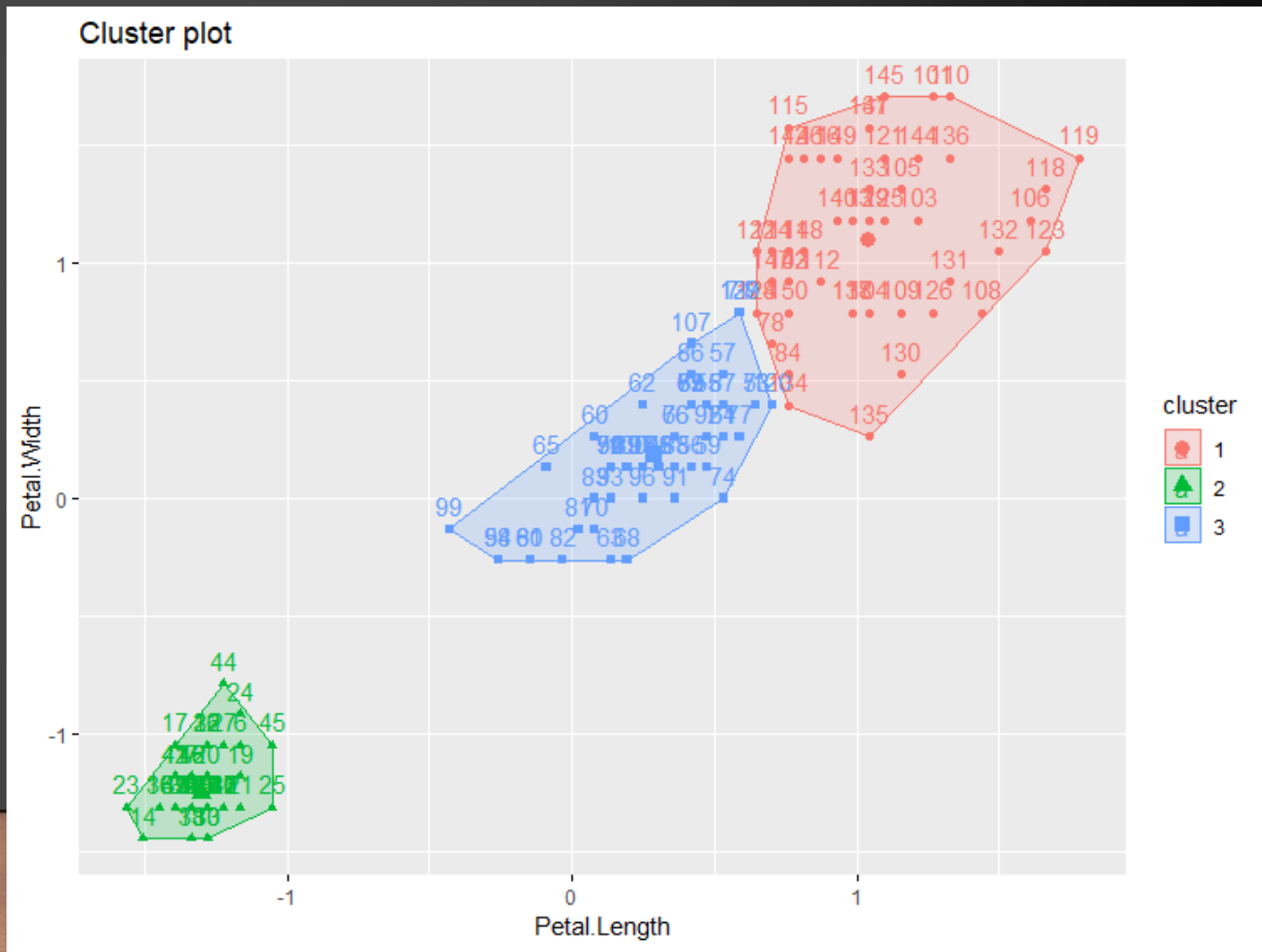
6.2. K-means clustering

```
set.seed(123)
km <- kmeans(iris[,3:4],3,nstart=20)

table(km$cluster,iris$Species)
fviz_cluster(km,data=iris[,3:4])
```

6.2. K-means clustering

```
library(factoextra)

table(km$cluster,iris$Species)
fviz_cluster(km,data=iris[,3:4])
```

```
  setosa versicolor virginica
1      0          2        46
2     50          0         0
3      0         48         4
```



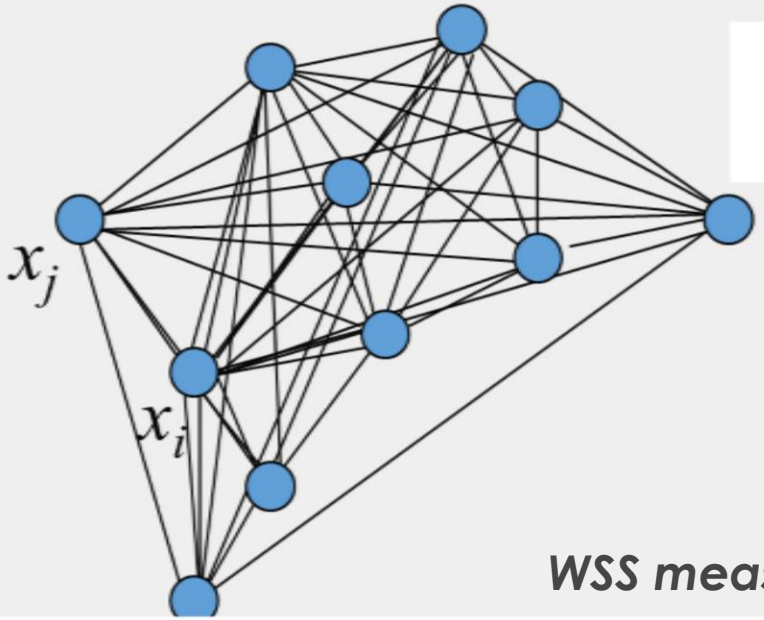
6. Unsupervised Learning

6.2. K-means clustering

6.3. Find Optimal K

6.3.1. Elbow Approach

Within-Cluster Sum of Squares



$$D = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| x_i - x_j \right\|^2$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

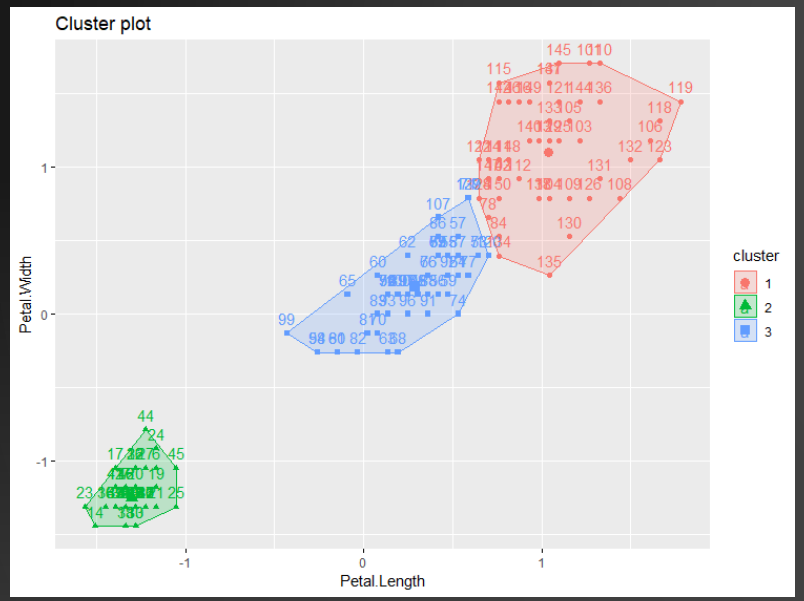
WSS measures the compactness of clusters

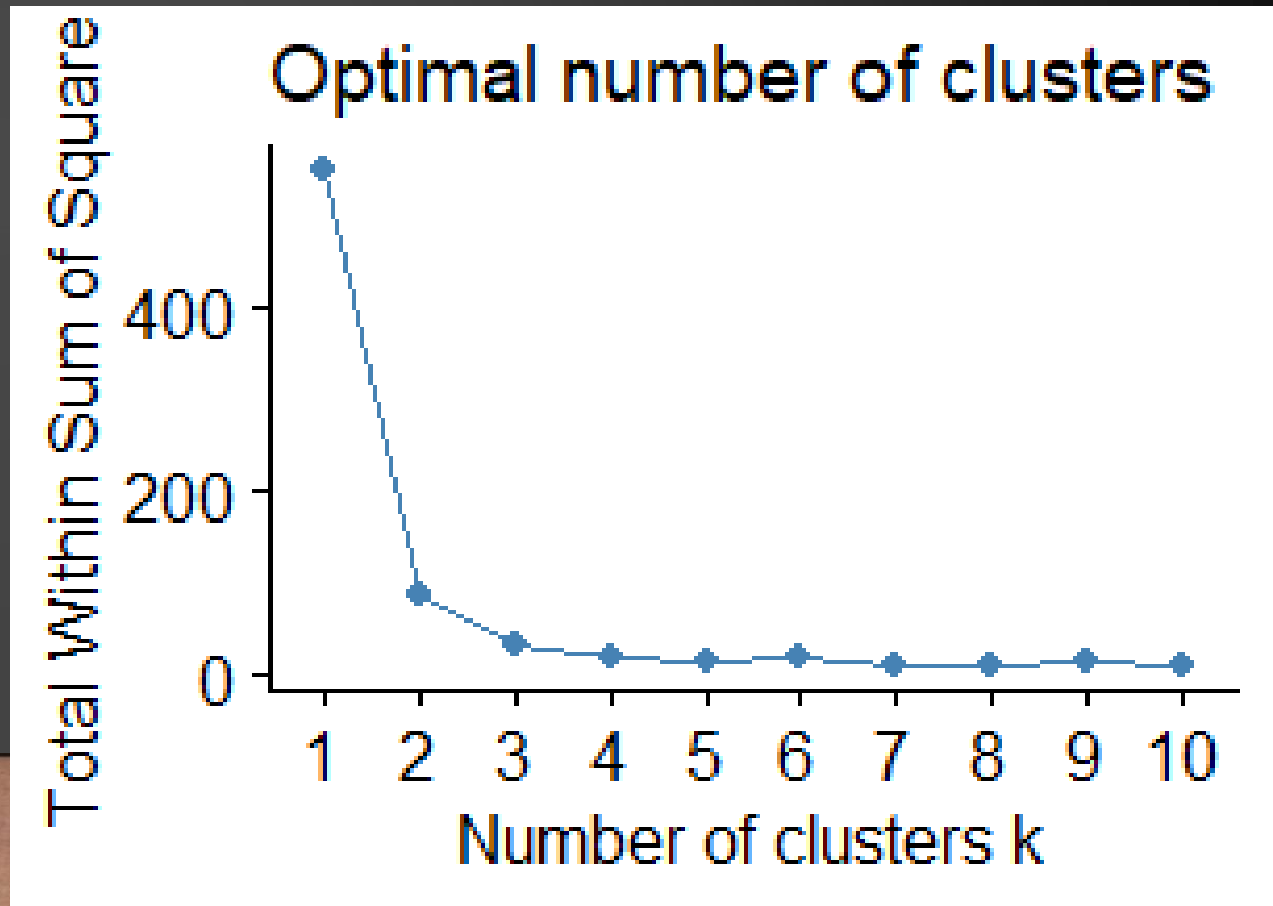
6. Unsupervised Learning

6.3. Find Optimal K

6.3.1. Elbow Approach

```
fviz_nbclust(iris[,3:4], kmeans, method = "wss")
```





6.3. Find Optimal K

6.3.2. Gap-Statistics

- Developed by Prof. Tibshirani et al in Stanford
- Applied to any clustering method (K-means, Hierarchical)
- Maximize the Gap function:

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

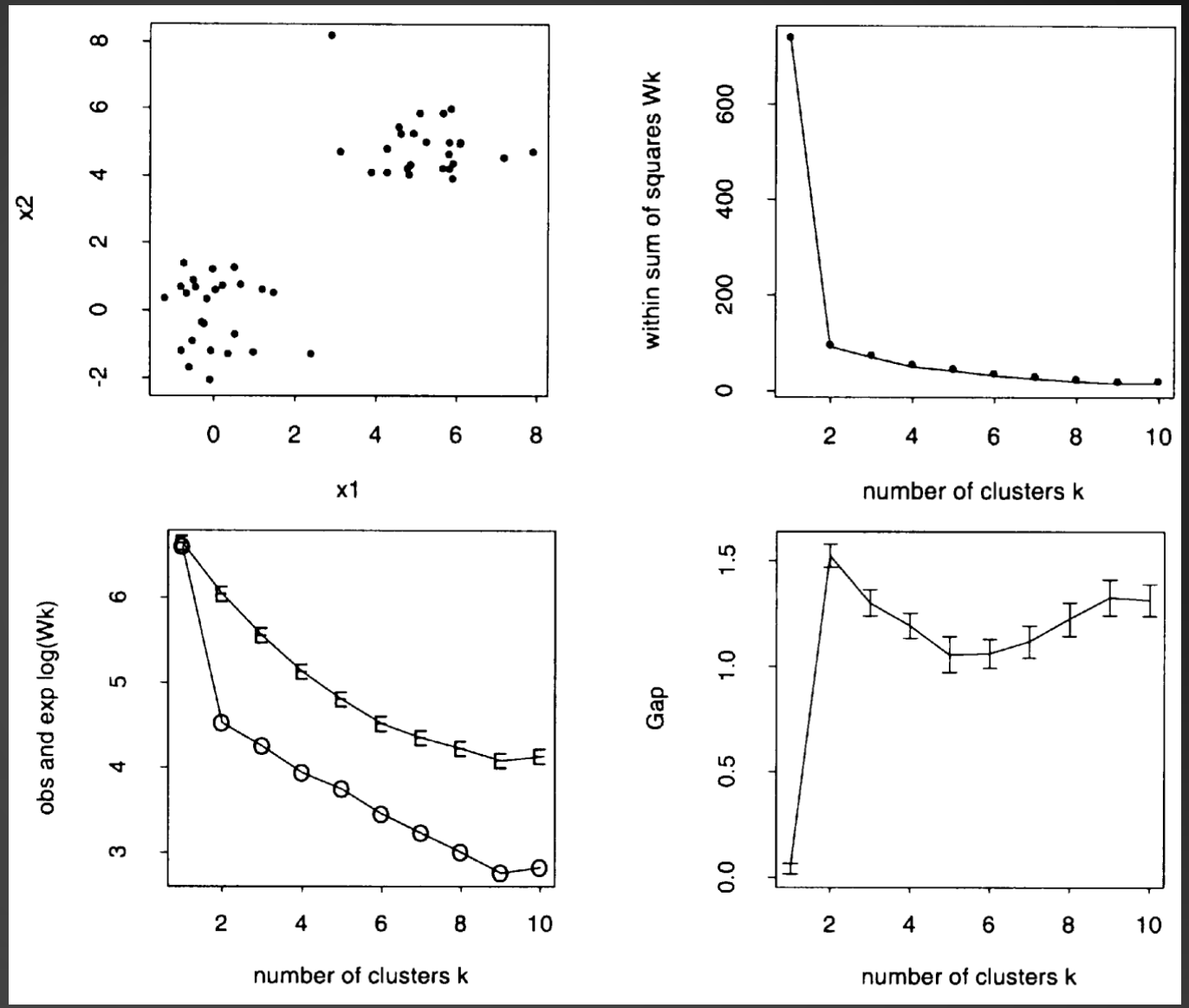
E*n: expectation under a sample size of n from the reference distribution

$$E_n^*\{\log(W_k)\} \approx \log(p\,n/12) - (2/p)\log(k) + const$$

6. Unsupervised Learning

6.3. Find Optimal K

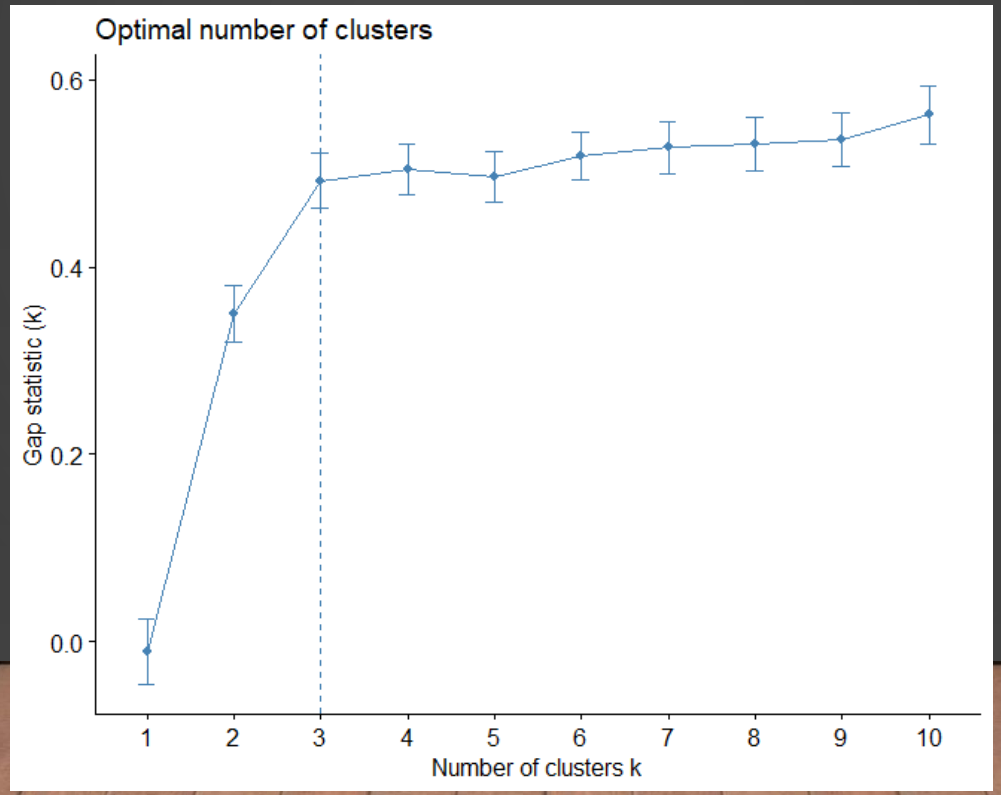
6.3.2. Gap-Statistics



6.3. Find Optimal K

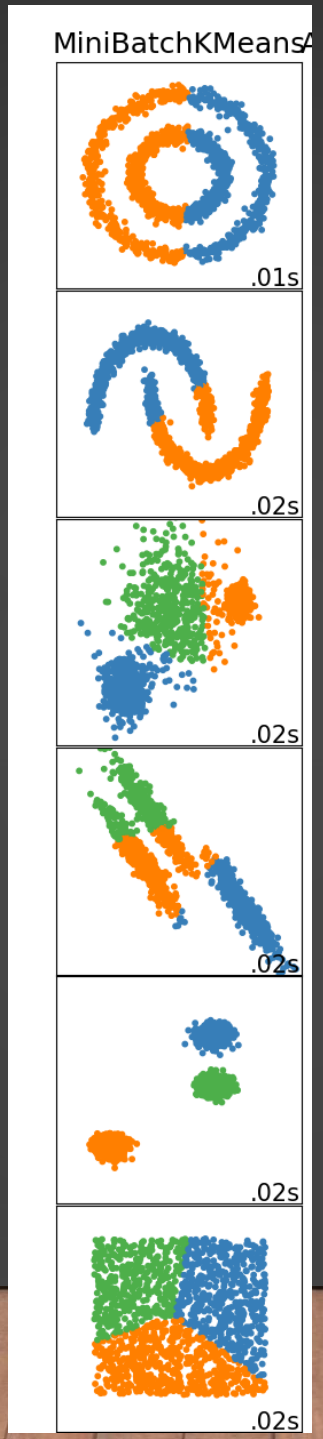
6.3.2. Gap-Statistics

```
library(cluster)

gap_stat <- clusGap(iris[,3:4], FUN = kmeans, nstart=20, K.max = 10, B = 50)
fviz_gap_stat(gap_stat)
```



6. Unsupervised Learning

6.4. Other clustering methods



6. Unsupervised Learning

6.4. Other clustering methods