# CSCI446/946 Big Data Analytics

## Week 4 Advanced Analytical Theory and Methods: Clustering

School of Computing and Information Technology
University of Wollongong Australia

# Advanced Analytical Theory and Methods: Clustering

- Overview of Clustering
- K-means clustering
  - Overview of the Method
  - Determining the Number of Clusters
  - Diagnostics
  - Reasons to Choose and Cautions
- Additional Algorithms

All the figures, tables and codes are from the book "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data" unless indicated otherwise.

# Overview of Clustering

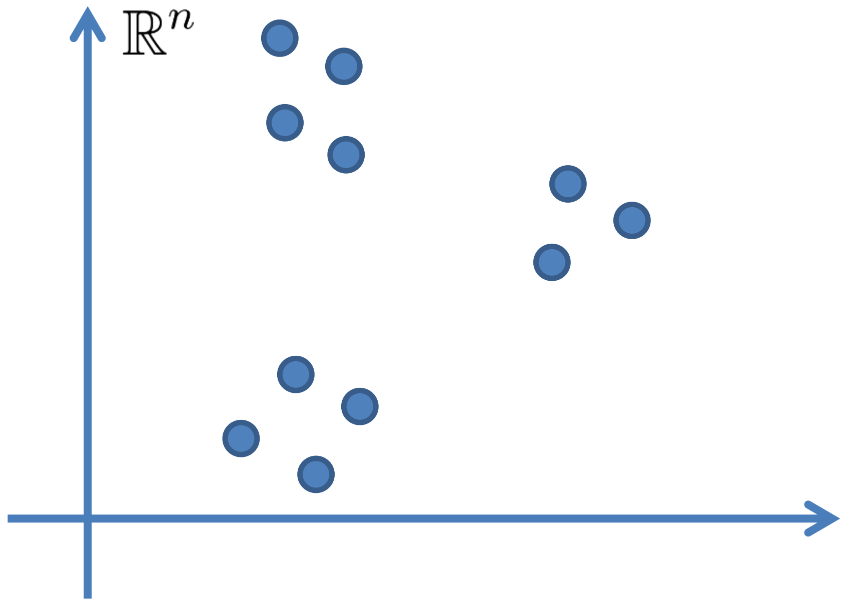
- Supervised vs. Unsupervised Techniques
  - Labelled data vs. Unlabelled data
- Unsupervised Techniques
  - Refers to the problem of finding hidden structure within unlabelled data
  - Clustering, density estimation, dimensionality reduction, etc.
- Clustering is an unsupervised technique

# Overview of Clustering


From http://mobblog.cs.ucl.ac.uk/category/paper/

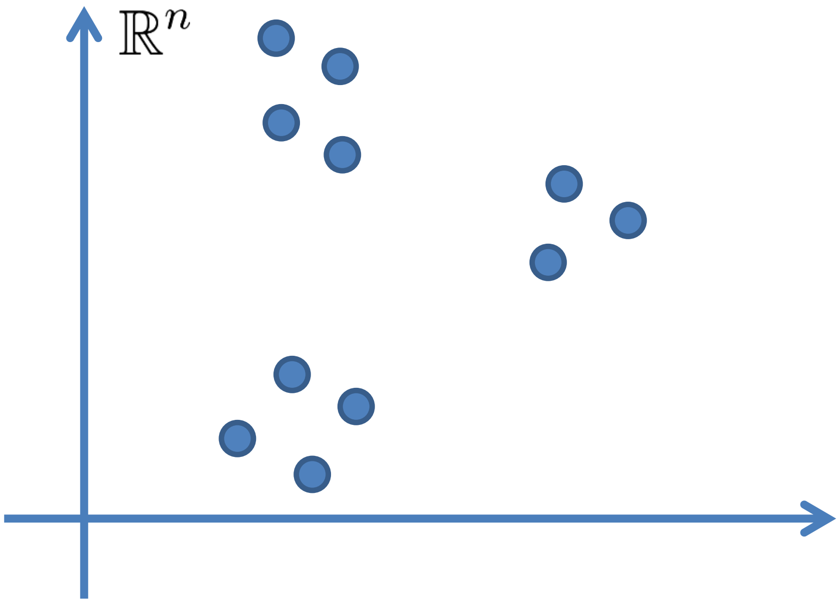
# K-means Clustering

- Given a collection of m objects each with n measurable attributes
  - Mathematically, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m \in \mathbb{R}^n$
  - Each object is a point in an n-dimensional space



# K-means Clustering

- For a chosen value of k, identify k clusters of objects based on the objects' proximity to the centre of the k groups



# K-means Clustering

- Use Cases
  - Often used as a lead-in to classification
  - Once clusters are identified, labels can be applied to each cluster to do classification
- Applications
  - Image Processing
  - Medical (Clustering patients)
  - Customer grouping (find similar customers)

# K-means Clustering

- Application to image processing

Original image

$K = 2$

$K = 3$

$K = 10$

# K-means Clustering

- Application to image processing

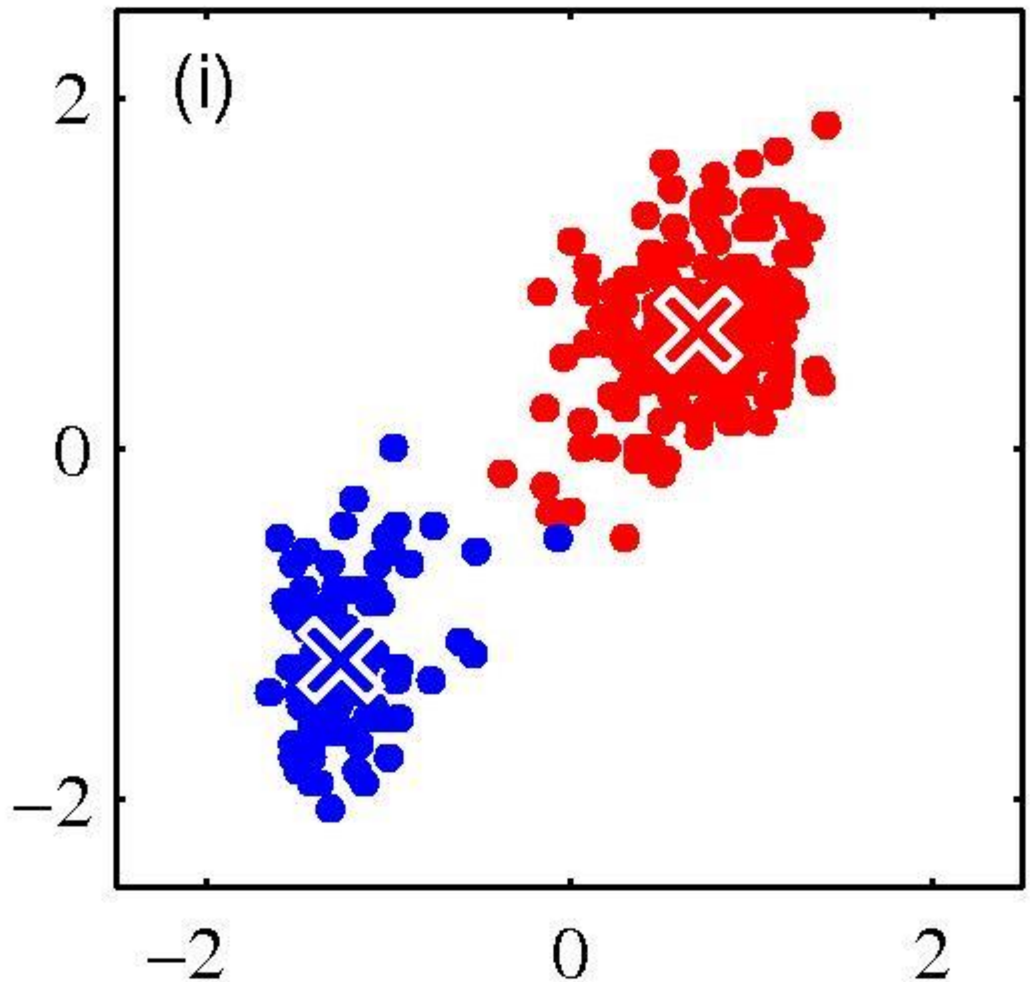
Original

K=2

K=3

K=10

# Overview of K-means Clustering



# Overview of K-means Clustering

- Four steps
  1. Choose the value of k and the k initial guess for the centriods
  2. Compute the distance from each data point to each centriod. Assign each point to the closest centriod.
  3. Update the centriod of each cluster
  4. Repeat Steps 2 and 3 until convergence

# Overview of K-means Clustering

- Compute the Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Compute the centriod for a cluster

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{m} \mathbf{x}_i}{m}$$

# Overview of K-means Clustering

- An optimization point of view
  - A combinatorial partition problem

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|_2^2; \quad r_{ij} \in \{0, 1\}$$

$$\{r_{ij}^*\} = \arg \min_{r_{ij} \in \{0,1\}} J$$

# Determine the Number of Clusters

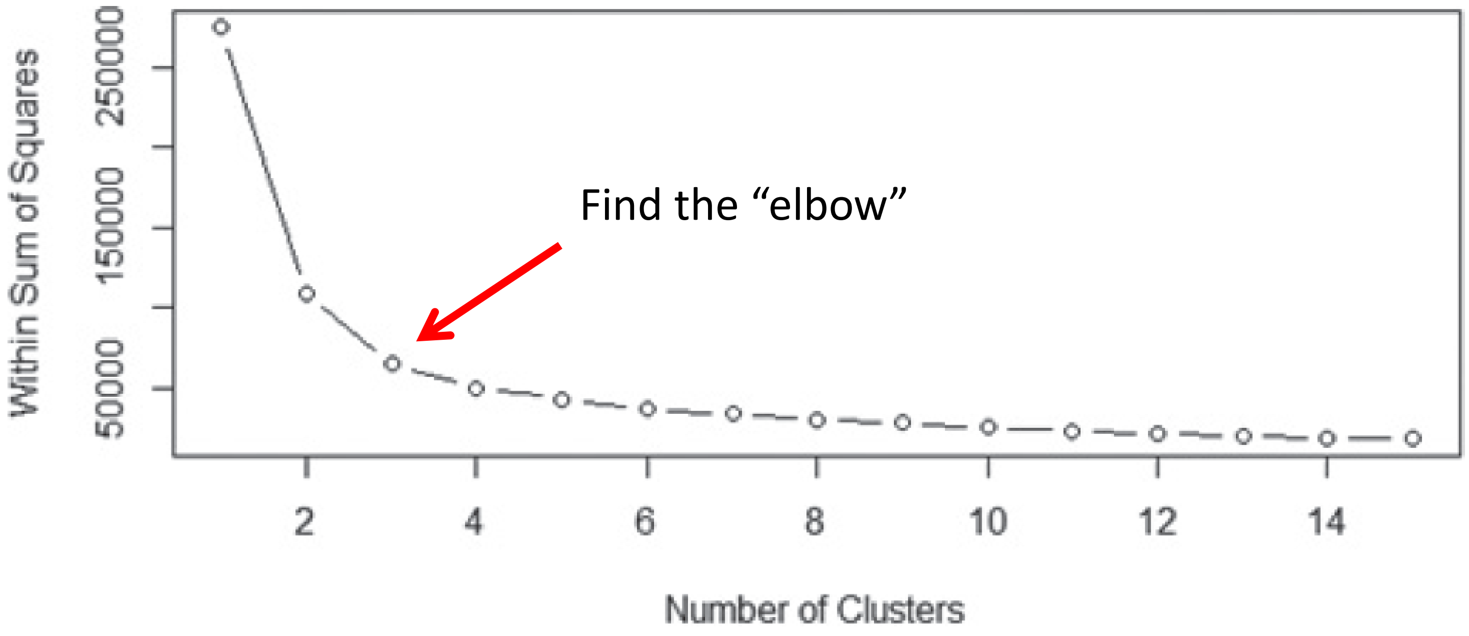
- What value of k shall be selected?
  - A reasonable guess, some predefined requirement
  - k-1, k, or k+1?
- Within Sum of Squares (WSS)
  - A heuristic
  - Sum of the squares of the distances between each data point and the closest centroid

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|_2^2; \quad r_{ij} \in \{0, 1\}$$

# Determine the Number of Clusters

- Within Sum of Squares (WSS)

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|_2^2; \quad r_{ij} \in \{0, 1\}$$



# Using R to Perform K-mean Clustering

- Task is to
  - Group 620 high school seniors based on their grades in “English”, “Math”, and “Science”

```
library(plyr)
library(ggplot2)
library(cluster)
library(lattice)
library(graphics)
library(grid)
library(gridExtra)

#import the student grades
grade_input = as.data.frame(read.csv("c:/data/grades_km_input.csv"))
```

# Using R to Perform K-mean Clustering

- Task is to
  - Group 620 high school seniors based on their grades in "English", "Math", and "Science"

```
kmdata_orig = as.matrix(grade_input[,c("Student","English", "Math","Science")])
kmdata <- kmdata_orig[,2:4]
```

```
kmdata[1:10,]
```

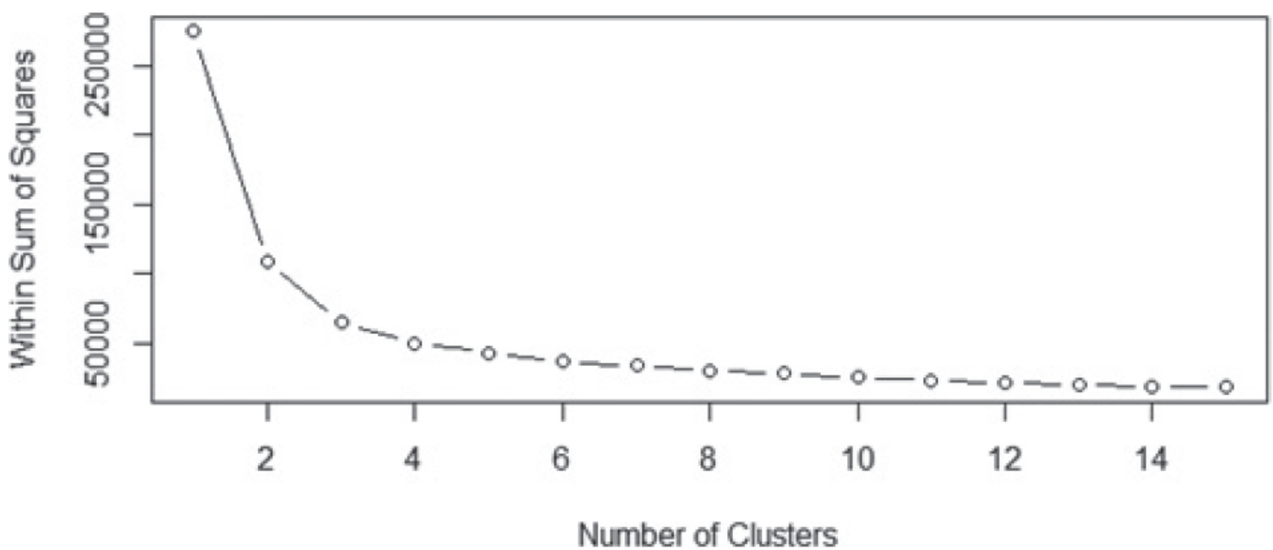
```
      English Math Science
 [1,]      99   96      97
 [2,]      99   96      97
 [3,]      98   97      97
 [4,]      95  100      95
 [5,]      95   96      96
 [6,]      96   97      96
 [7,]     100   96      97
 [8,]      95   98      98
 [9,]      98   96      96
[10,]      99   99      95
```

# Using R to Perform K-mean Clustering

- Compute and plot WSS to choose k value

```
wss <- numeric(15)
for (k in 1:15) wss[k] <- sum(kmeans(kmdata, centers=k, nstart=25)$withinss)
```

```
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within Sum of
Squares")
```



# Using R to Perform K-means Clustering

- Perform K-means Clustering

```
km = kmeans(kmdata,3, nstart=25)
km

K-means clustering with 3 clusters of sizes 158, 218, 244

Cluster means:
   English     Math  Science
1 97.21519 93.37342 94.86076
2 73.22018 64.62844 65.84862
3 85.84426 79.68033 81.50820

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 1 1
 [41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 1 1
 [81] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 1 1
[121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      3 3 3 3 3 3 3 3 3 3
```

# Using R to Perform K-means Clustering

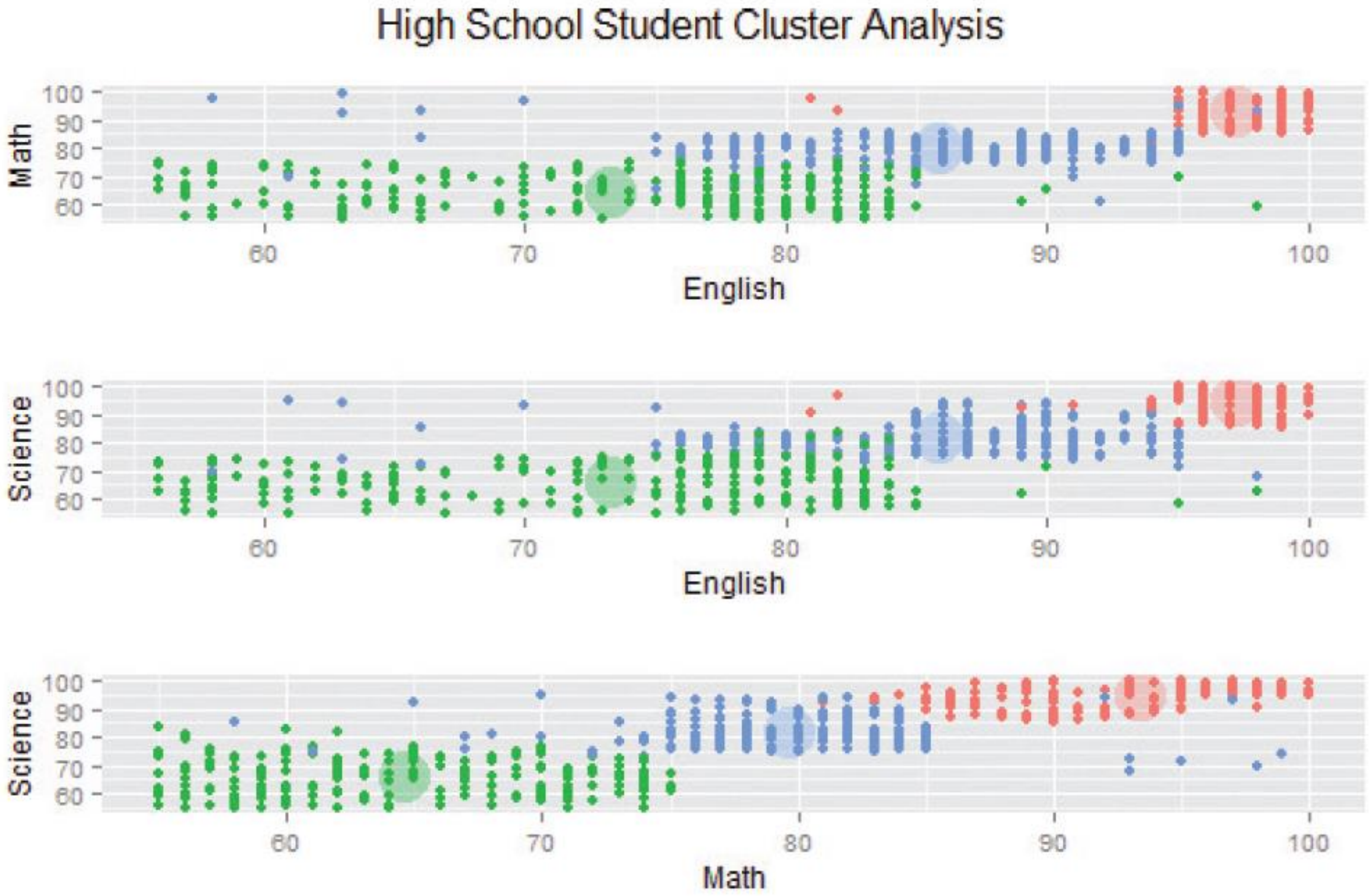
- Perform K-means Clustering

```
[521] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
      2 2 2 2 2 2 2 2 2 2
[561] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
      2 2 2 2 2 2 2 2 2 2
[601] 3 3 2 2 3 3 3 3 1 1 3 3 3 2 2 3 2 3 3 3

Within cluster sum of squares by cluster:
[1]  6692.589 34806.339 22984.131
 (between_SS / total_SS =  76.5 %)

Available components:

[1] "cluster"   "centers"  "totss"    "withinss"   "tot.withinss"
[6] "betweenss" "size"     "iter"     "ifault"
```

```
c( wss[3] , sum(km$withinss) )

[1] 64483.06 64483.06
```

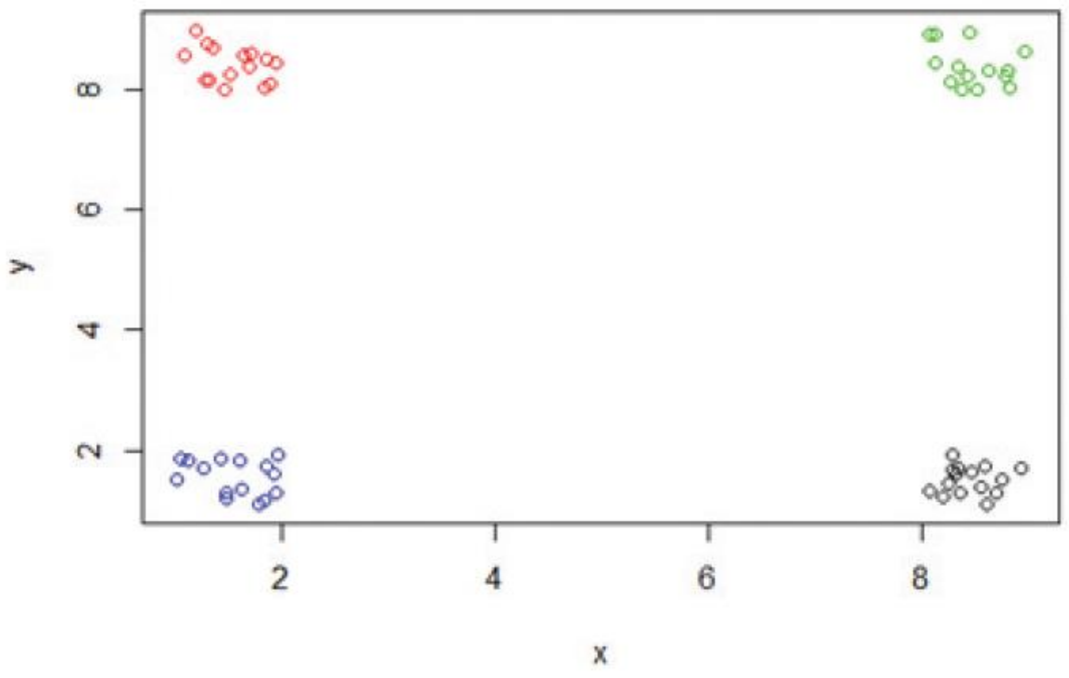
# Using R to Perform K-means Clustering

- Visualize the identified clusters and centroids



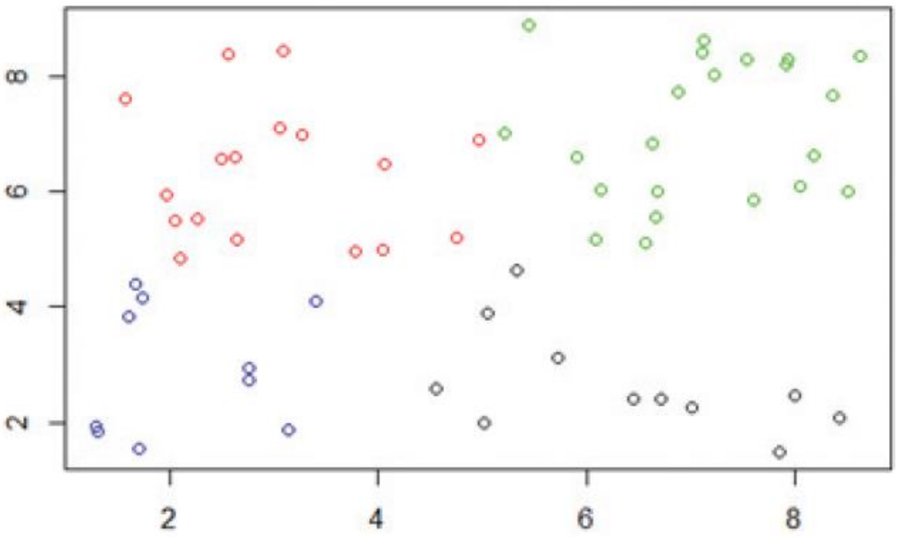
# Diagnostics

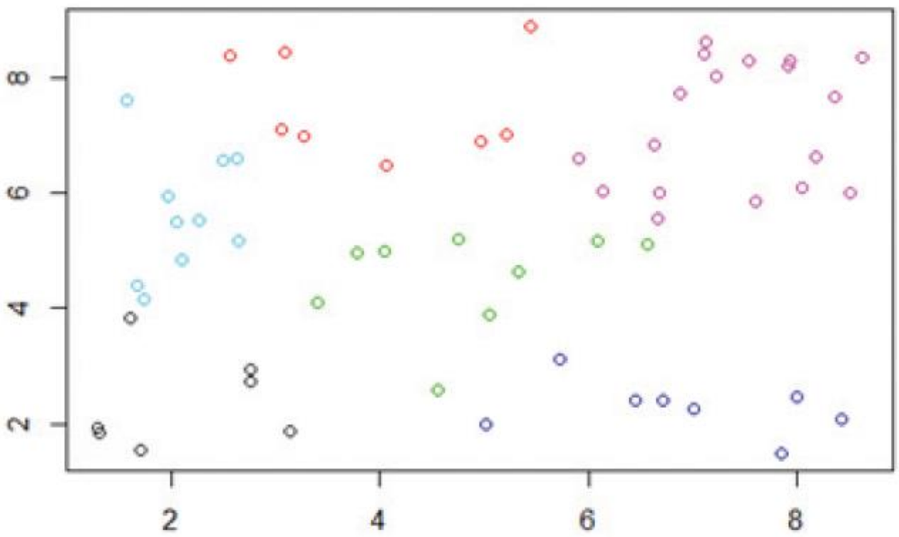
- The following questions shall be asked
  - Are the clusters well separated from each other?
  - Do any of the clusters have only a few points?
  - Do any of the centriods appear to be too close to each other?



# Diagnostics

- A principle
  - If using more clusters does not better distinguish the groups, it is almost certainly better to go with fewer clusters



# Reasons to Choose and Cautions

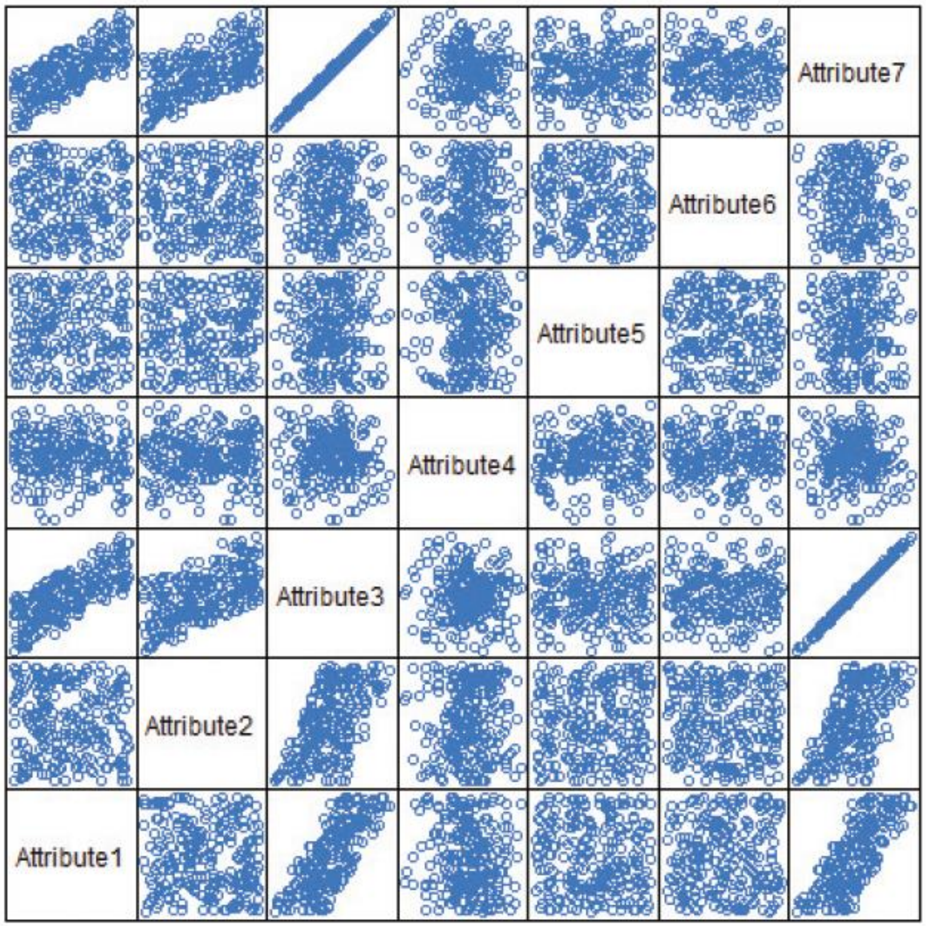
- Several decisions that must be made
  - What object attributions shall be included in clustering analysis?
  - What unit of measure shall be used for each attribute?
  - Do the attributes need to be rescaled?
    - One attribute could have a disproportionate effect

# Reasons to Choose and Cautions

- Object attributes
  - Whether it will be known for a new object?
  - Best to reduce the number of attributes to the extent of possible
    - Avoid using too many variables (Why?)
    - Avoid using several similar variables (Why?)
- Identify any highly correlated attributes
- Feature selection, PCA, etc.

# Reasons to Choose and Cautions

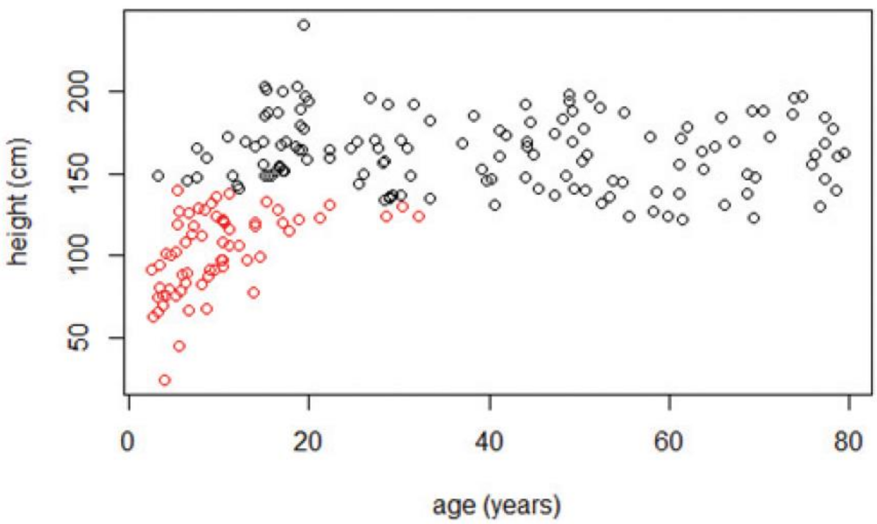
- Identify any highly correlated attributes

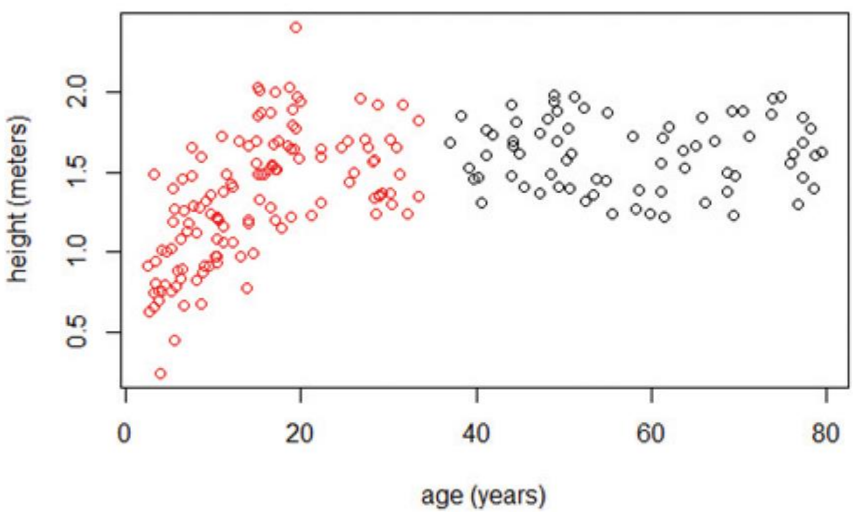



What is your observation?

# Reasons to Choose and Cautions

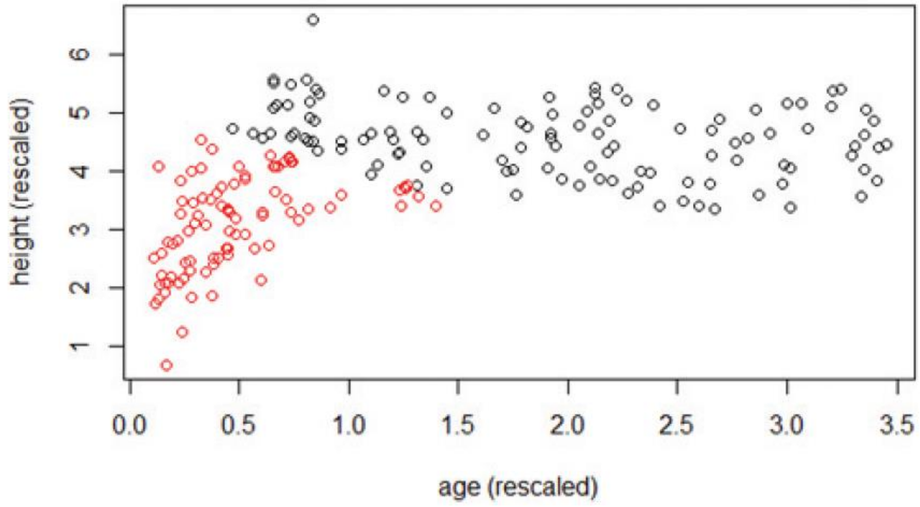
- Units of measure could affect clustering result

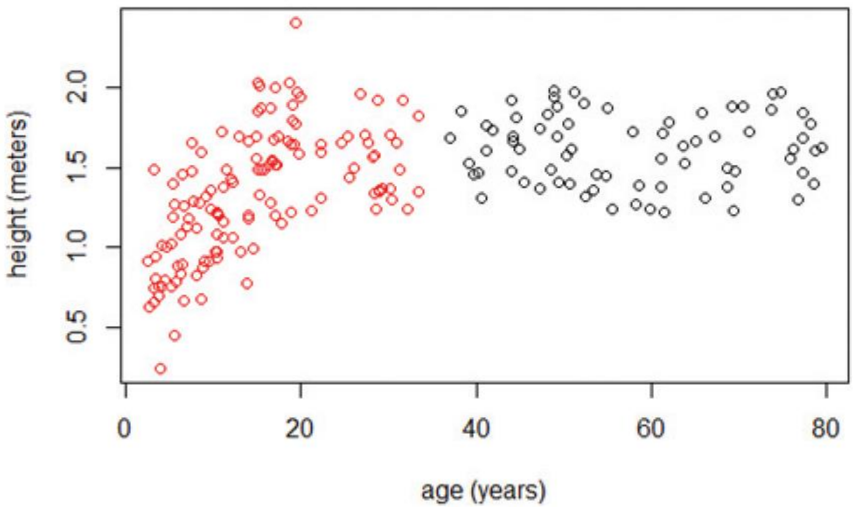


# Reasons to Choose and Cautions

- Rescaling attributes affect clustering result
  - Divide each attribute by its standard deviation



# Additional Considerations

- K-means clustering is sensitive to the starting positions of the initial centroids
  - Usually, we run the k-means clustering several times for a particular k value to choose the clustering result with the lowest WSS value
  - Implemented by the `nstart` option in `kmeans()`
- Other distances
  - Manhattan distance & the median of cluster

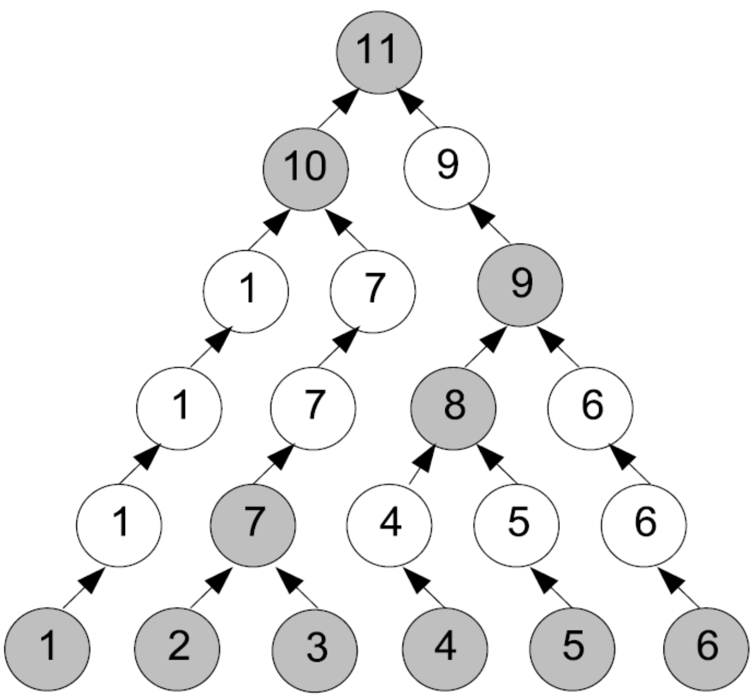
$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{n} |x_i - y_i|$$

# Additional Algorithms

- K-means clustering is easily applied to numeric data where the concept of distance can naturally be applied
- K-modes handles categorical data
  - Use the number of differences in the respective components of the attributes
    - What is the distance between (a,b,e,d) and (d,d,d,d)?
  - Implemented by the `kmode()` function

# Additional Algorithms

- Hierarchical Clustering (`hclust()`)
  - Hierarchical agglomerative clustering
  - Hierarchical divisive clustering




1. Each object is initially treated as a cluster
2. The clusters are then combined with the most similar cluster in each step
3. This process is repeated until one cluster (containing all objects) exists

# Recap: Advanced Analytical Theory and Methods: Clustering

- To use k-means properly
  - Properly select and scale the attribute values
  - Ensure that the distance between objects is meaningful
  - Choose the number of clusters, k
  - If k-means is not appropriate, consider others
  - Take advantage of visualization tools for diagnostics

# Q&A


**Images Courtesy of Google Image**