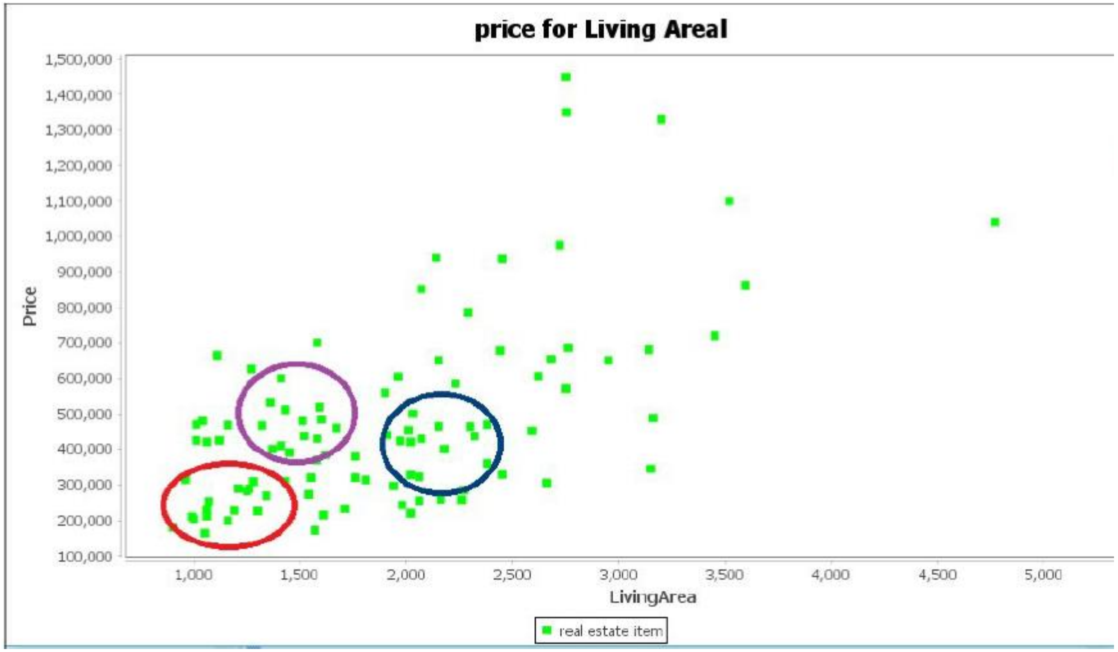
Clustering

Clustering is a popular form of an unsupervised learning algorithm and is used to discover and club similar entities (like movies, customers, products, likes, and so on) into groups or clusters

Within a group, the items are similar to each other based on certain attributes they possess

The similarity of the datasets is calculated based on the similarity finding mathematical formulas and approaches, like Euclidean distance, Pearson coefficient, Jaccard distance, cosine similarity, and so on.

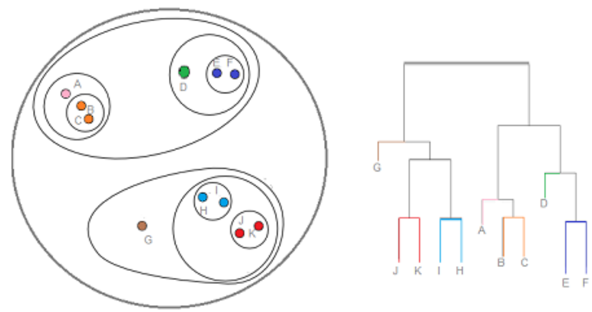


Clusters Example

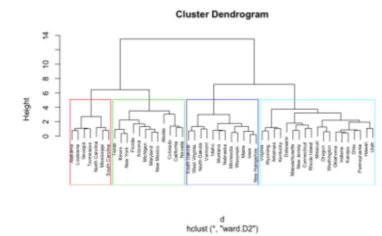
- Customer segmentation: Clustering can be used to segment customers, for example, of online e-commerce stores.
- Search engines: Clustering is used to form smaller groups of sub-search categories that can help users looking for specific items via a search engine.
- Data exploration: In the data exploration phase, clustering can be used.
- Finding epidemic breakout zones: There are studies where clustering has been applied to find zones of disease breakout areas based on the availability of statistical data.
- Biology: To find groups of genes that show similar behavior.
- News categorization: Google News does this to categorize news automatically into groups such as sports news, technology related news, and so on.
- Summarization of news: First, we can find custom groups in the news and then find the centroid and use it to summarize the news.

Types of clustering

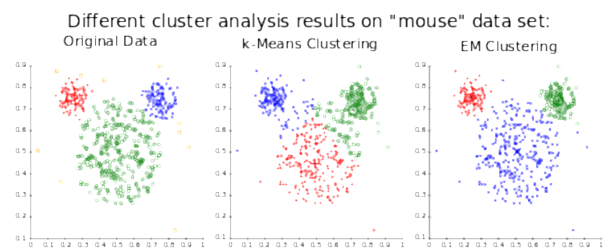
- Hierarchical
- *k*-means
- Bisecting *k*-means



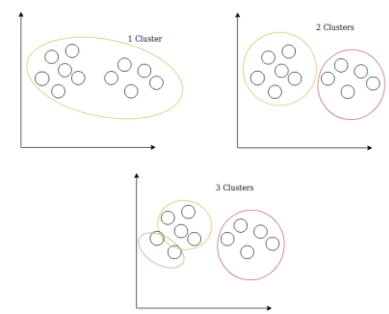
Hierarchical



Hierarchical (Dendogram)



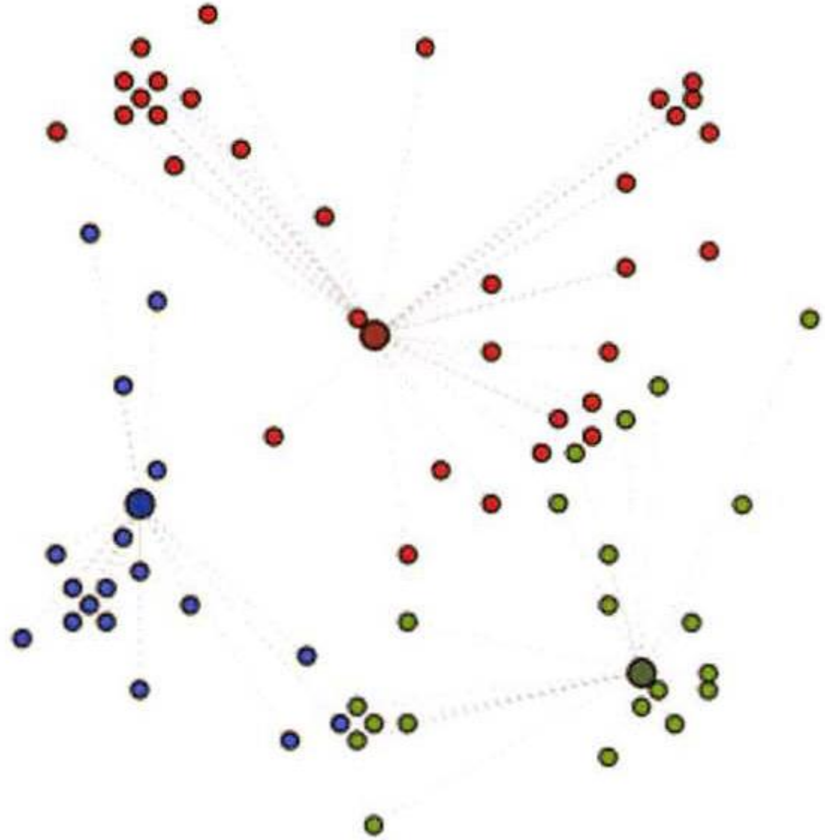
k-means



Bisecting *k*-means

k-means

- Analytical technique that, for a chosen value of *k*, identifies *k* clusters of objects based on the objects' proximity to the center of the *k* groups.
- The center is determined as the arithmetic average (mean) of each cluster's *n*-dimensional vector of attributes.



Algoritma *k*-means

1. Tentukan nilai *k*

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

2. Pilih *k* titik pusat (*centroid*) kluster
   - Misalnya:
     - Secara acak
     - Menghitung jarak instans terjauh

$$(x_c, y_c) = \left| \frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m} \right|$$

3. Isi setiap kluster dengan instans
   a. Didasari oleh <u>jarak</u> (mis. *Euclidean distance*) setiap instans dengan titik pusat dari kluster yang terbentuk
   - Hitung (<u>*update*</u>) nilai titik pusat kluster
   - *Update* nilai centroids (*means*) = rata-rata nilai atribut dalam kluster – *center of mass*

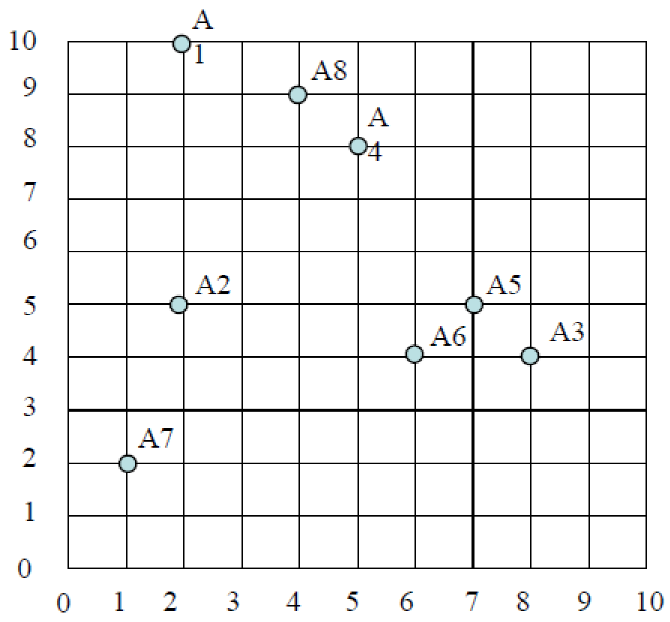
   b. Pada saat <u>iterasi</u>: Pindahkan instans ke kluster yang jarak titik pusatnya lebih dekat
4. <u>Iterasi:</u> Kembali ke langkah no. 3b
   - Sampai konvergen (tidak ada perubahan isi kluster, jarak telah optimal – perubahan statis)

Contoh *k*-means Clustering

- Gunakan metode *k*-means untuk mengelompokkan titik-titik berikut ini:

  A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

- Asumsikan *k*=3, dengan centroid awal adalah titik A1, A4 dan A7
- Seandainya *k*-means hanya dilakukan dalam satu iterasi saja.
  - Gunakan *euclidean distance*
  - Berikan *centroid* baru yang dihasilkan



Distance Matrix

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Iterasi 1

$$seed1=A1=(2,10),\ seed2=A4=(5,8),\ seed3=A7=(1,2)$$

A1:

d(A1, seed1)=0 as A1 is seed1

d(A1, seed2)= $\sqrt{13}$ >0

d(A1, seed3)= $\sqrt{65}$ >0

$\Rightarrow$ A1 $\in$ cluster1

A2:

d(A2,seed1)= $\sqrt{25}$ = 5

d(A2, seed2)= $\sqrt{18}$ = 4.24

d(A2, seed3)= $\sqrt{10}$ = 3.16 $\Leftarrow$ smaller

$\Rightarrow$ A2 $\in$ cluster3

A3:

d(A3, seed1)= $\sqrt{36}$ = 6

d(A3, seed2)= $\sqrt{25}$ = 5 $\Leftarrow$ smaller

d(A3, seed3)= $\sqrt{53}$ = 7.28

$\Rightarrow$ A3 $\in$ cluster2

A4:

d(A4, seed1)= $\sqrt{13}$

d(A4, seed2)=0 as A4 is seed2

d(A4, seed3)= $\sqrt{52}$ >0

$\Rightarrow$ A4 $\in$ cluster2

Iterasi 1 (cont'd)

A5:
d(A5, seed1)= $\sqrt{50}$ = 7.07
d(A5, seed2)= $\sqrt{13}$ = 3.60 ← smaller
d(A5, seed3)= $\sqrt{45}$ = 6.70
➔ A5 $\in$ cluster2

A6:
d(A6, seed1)= $\sqrt{52}$ = 7.21
d(A6, seed2)= $\sqrt{17}$ = 4.12 ← smaller
d(A6, seed3)= $\sqrt{29}$ = 5.38
➔ A6 $\in$ cluster2

A7:
d(A7, seed1)= $\sqrt{65}$ >0
d(A7, seed2)= $\sqrt{52}$ >0
d(A7, seed3)=0 as A7 is seed3
➔ A7 $\in$ cluster3

A8:
d(A8, seed1)= $\sqrt{5}$
d(A8, seed2)= $\sqrt{2}$ ← smaller
d(A8, seed3)= $\sqrt{58}$
➔ A8 $\in$ cluster2

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

C1= (2, 10)
C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)
C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

$$\left(x_c, y_c\right) = \left(\frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m} \right)$$

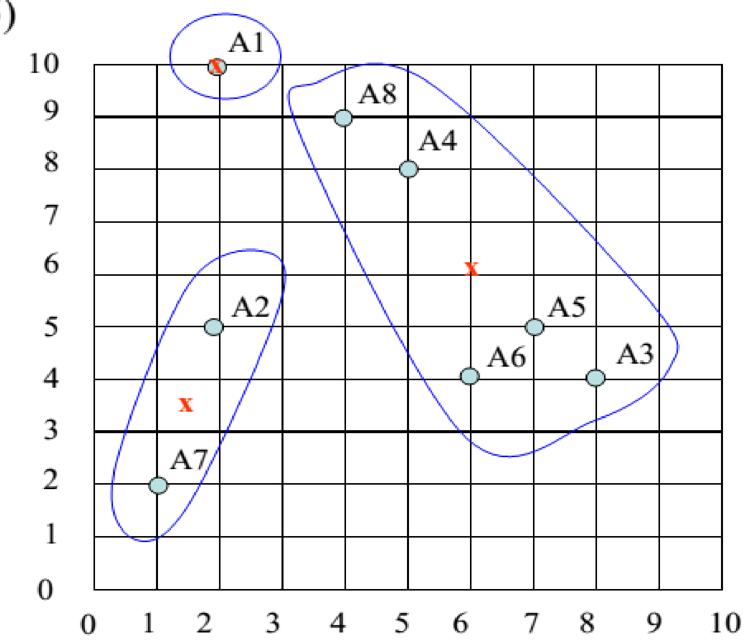
Akhir Iterasi 1

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

C1= (2, 10)

C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)

C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)



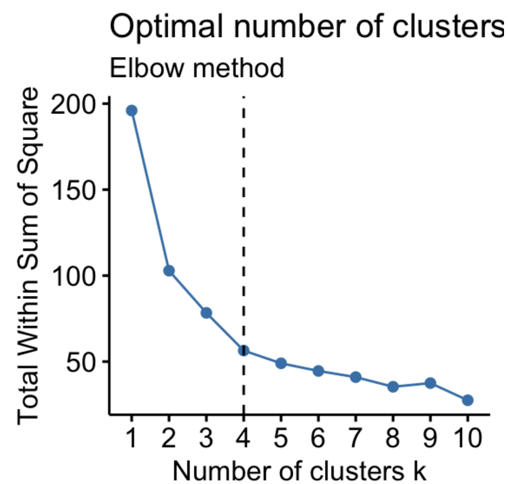
Determining the number of clusters

- With the preceding algorithm, *k* clusters can be identified in a given dataset, but what value of *k* should be selected?
- The value of *k* can be chosen based on a reasonable guess or some predefined requirement
- A <u>heuristic</u> using the **Within Sum of Squares** (WSS) metric is examined to determine a reasonably <u>optimal</u> value of k.

https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f

$$WSS = \sum_{i=1}^{M} d(p_i, q^{(i)})^2 = \sum_{i=1}^{M}\sum_{j=1}^{n}\left(p_{ij} - q_j^{(i)}\right)^2$$

The sum of the squares of the distances between each data point and the closest centroid.



Customer Segmentation

Attribute name	Description
Invoice number	Invoice number; a number uniquely assigned to each transaction
Stock code	Product (item) code; a 5-digit integral number uniquely assigned to each distinct product
Description	Product item name
Quantity	Quantity of items purchased in a single transaction
Invoice date	Date of the transaction
Unit price	Price of the item (in pounds)
Customer ID	Unique ID of the person making the transaction
Country	Country from where the transaction was made, such as United Kingdom or France

```
17/05/16 08:50:06 INFO CodeGenerator: Code generated in 32.169375 ms
+------+------+--------------------+---+---------------+----+-----+--------------+
|   _c0|   _c1|                 _c2|_c3|            _c4| _c5|  _c6|           _c7|
+------+------+--------------------+---+---------------+----+-----+--------------+
|536365|85123A|WHITE HANGING HEA...|  6|12/1/2010 8:26|2.55|17850|United Kingdom|
|536365| 71053| WHITE METAL LANTERN|  6|12/1/2010 8:26|3.39|17850|United Kingdom|
|536365|84406B|CREAM CUPID HEART...|  8|12/1/2010 8:26|2.75|17850|United Kingdom|
|536365|84029G|KNITTED UNION FLA...|  6|12/1/2010 8:26|3.39|17850|United Kingdom|
|536365|84029E|RED WOOLLY HOTTIE...|  6|12/1/2010 8:26|3.39|17850|United Kingdom|
|536365| 22752|SET 7 BABUSHKA NE...|  2|12/1/2010 8:26|7.65|17850|United Kingdom|
|536365| 21730|GLASS STAR FROSTE...|  6|12/1/2010 8:26|4.25|17850|United Kingdom|
|536366| 22633|HAND WARMER UNION...|  6|12/1/2010 8:28|1.85|17850|United Kingdom|
|536366| 22632|HAND WARMER RED P...|  6|12/1/2010 8:28|1.85|17850|United Kingdom|
|536367| 84879|ASSORTED COLOUR B...| 32|12/1/2010 8:34|1.69|13047|United Kingdom|
|536367| 22745|POPPY'S PLAYHOUSE...|  6|12/1/2010 8:34| 2.1|13047|United Kingdom|
|536367| 22748|POPPY'S PLAYHOUSE...|  6|12/1/2010 8:34| 2.1|13047|United Kingdom|
|536367| 22749|FELTCRAFT PRINCES...|  8|12/1/2010 8:34|3.75|13047|United Kingdom|
```

RFM and the terms are explained in the following table:

Recency	This value depicts how recently the customer has purchased. We will be using this as the most recent purchase date and calculating the difference between the current date and this date, so the net value for this will be an integer representing the number of days since last purchase.
Frequency	This value depicts how many times the customer has purchased. We will be using it as the count of the number of times the customer has purchased within a year.
Monetary	This value depicts the amount the customer has spent overall in the store. We will store this as the amount depicted by multiplying the cost of items and the quantity sold to the customer.