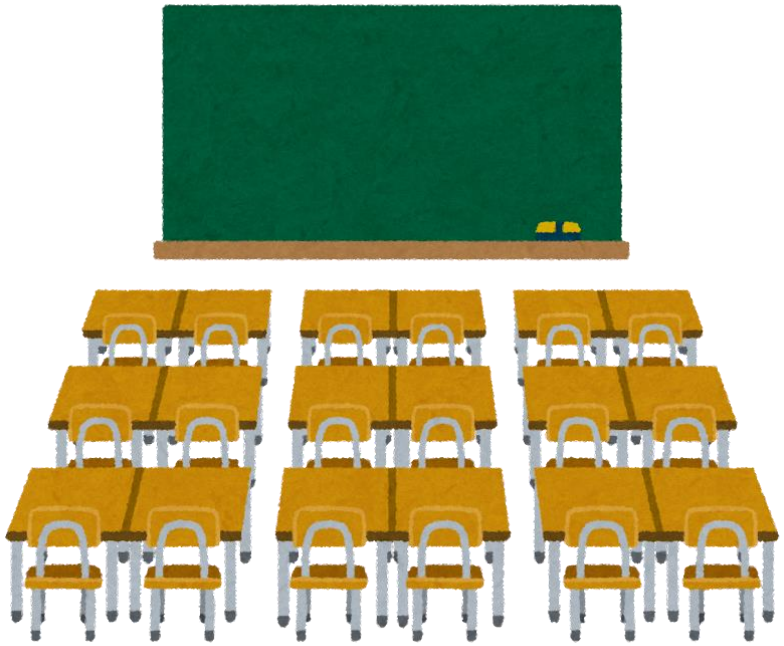
クラスタリング

自分と近い性格の人は誰？

クラスのみんな、気が合う人は誰だろう？



アンケートをやってみよう！

どれくらい好きか、選んで下さい。

5（とても好き） / 4(好き） / 3（普通）
2(あまり好きではない） / 1（好きではない）

Q1. プログラミング　5 4 3 2 1

Q2. 野球　5 4 3 2 1

Q3. 楽器演奏　5 4 3 2 1

・
・
・

回答結果

	洋楽	ドラマ	野球	JPOP	プログラミ	ライブ	バスケ	カラオケ	ペンギン	水泳	登山	舞台鑑賞	パン	ダンス	料理	楽器演奏	昼寝	動画鑑賞	犬	猫
生徒1	3	5	5	2	1	4	1	2	3	3	3	2	3	4	5	1	4	3	5	5
生徒2	3	3	3	2	3	1	5	1	2	5	3	5	5	4	2	1	2	2	3	5
生徒3	1	5	4	3	2	3	4	4	1	4	2	4	5	3	3	3	4	5	4	1
生徒4	4	3	5	2	1	3	5	4	1	2	2	2	4	5	3	2	4	4	1	2
生徒5	2	4	2	3	4	1	3	4	3	4	5	4	3	1	1	2	5	2	3	2
生徒6	3	5	1	2	5	2	5	5	2	1	3	5	2	2	2	5	3	3	4	4
生徒7	4	2	1	1	2	5	5	4	3	3	5	2	1	5	1	3	2	1	2	3
生徒8	2	3	3	3	1	3	1	3	4	3	4	5	5	5	4	1	4	1	4	3
生徒9	3	2	4	1	2	2	5	2	4	2	3	2	4	4	5	4	2	4	1	4
生徒10	4	5	4	2	1	2	4	3	3	4	4	4	3	5	5	4	3	4	5	1
生徒11	5	5	1	5	2	3	1	5	3	2	2	3	5	5	3	5	5	2	5	5
生徒12	5	1	1	5	5	4	1	5	4	2	1	4	3	5	1	5	4	4	2	1
生徒13	5	1	1	3	5	2	1	5	2	4	3	1	4	2	5	4	1	3	4	2
生徒14	4	2	4	1	3	1	2	3	4	2	3	5	3	3	5	3	2	1	1	1
生徒15	2	2	2	1	2	3	4	4	4	5	4	2	1	2	4	2	4	5	2	3
生徒16	5	1	4	4	2	3	4	4	3	4	3	1	4	1	4	3	3	2	5	2
生徒17	5	2	4	1	2	1	3	2	4	3	3	2	1	3	4	4	4	4	4	5
生徒18	2	5	2	3	1	3	5	5	1	1	3	3	2	5	3	4	2	3	4	2
生徒19	1	1	3	3	1	5	5	4	2	4	2	1	2	4	4	3	5	5	4	1
生徒20	1	2	2	1	5	2	1	4	2	3	2	1	5	5	5	3	1	3	2	3
生徒21	3	2	4	4	2	1	1	4	2	5	2	1	1	3	3	2	1	5	3	1
生徒22	2	3	5	4	2	2	2	2	2	5	4	5	2	2	4	1	2	4	4	4
生徒23	5	2	2	1	1	1	5	4	5	2	2	1	3	1	5	5	1	3	4	5
生徒24	4	5	2	5	1	4	3	5	4	3	2	3	1	5	5	5	2	5	4	3
生徒25	5	1	4	4	1	1	1	2	1	5	4	5	3	4	5	2	2	5	3	2
生徒26	5	5	2	3	1	2	5	5	1	3	4	3	2	3	2	2	5	2	4	2
生徒27	3	2	4	5	2	4	1	4	1	3	1	4	4	5	2	4	5	5	2	3
生徒28	5	3	2	5	4	2	3	3	2	1	3	5	1	2	5	4	3	5	3	4

どうやって似ている人を探す？

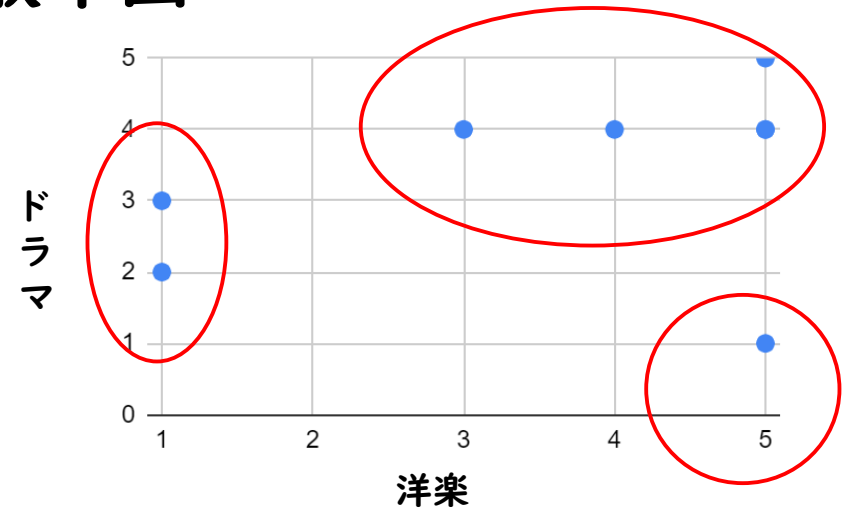
2変数だったら…（情報Ⅰでやったこと）

データ

	洋楽	ドラマ
生徒1	3	5
生徒2	3	3
生徒3	1	5
生徒4	4	3
生徒5	2	4
生徒6	3	5

⋮ ⋮ ⋮

散布図

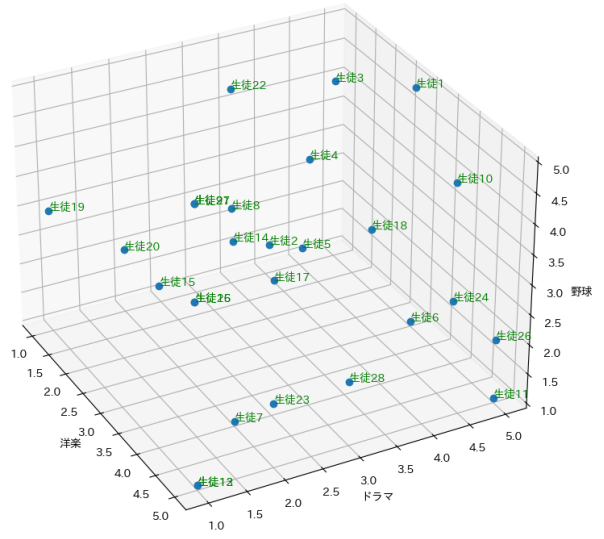


3変数だったらギリギリ…？

データ

	洋楽	ドラマ	野球
生徒1	3	5	5
生徒2	3	3	3
生徒3	1	5	4
生徒4	4	3	5
生徒5	2	4	2
生徒6	3	5	1
生徒7	4	2	1
生徒8	2	3	3
生徒9	3	2	4
生徒10	4	5	4
生徒11	5	5	1
生徒12	5	1	1
生徒13	5	1	1
生徒14	4	2	4

散布図



もっと変数があったらどうしよう

データ

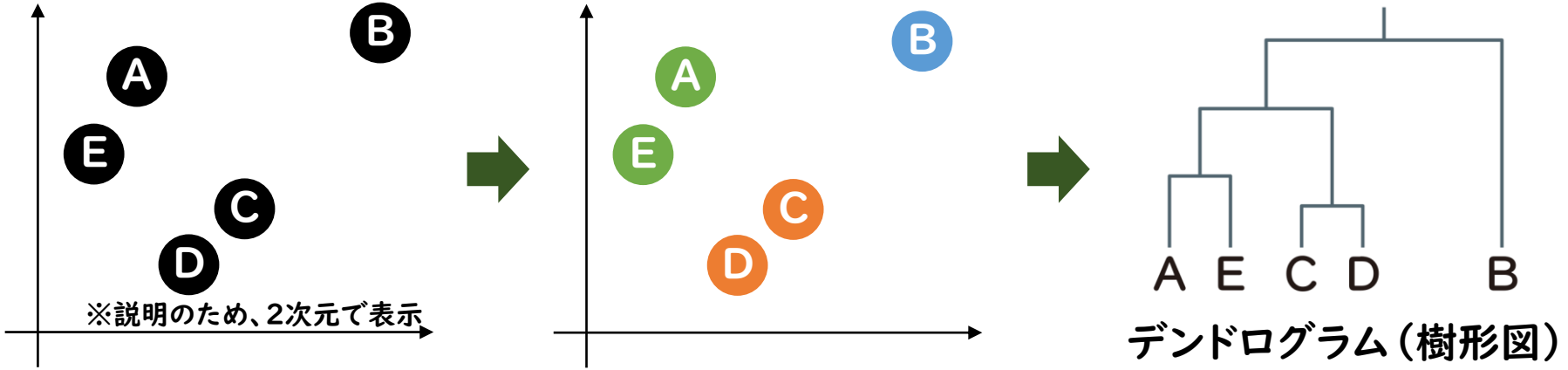
	洋楽	ドラマ	野球	JPOP	プログラミ	ライブ	バス
生徒1	3	5	5	2	1	4	
生徒2	3	3	3	2	3	1	
生徒3	1	5	4	3	2	3	
生徒4	4	3	5	2	1	3	
生徒5	2	4	2	3	4	1	
生徒6	3	5	1	2	5	2	
生徒7	4	2	1	1	2	5	
生徒8	2	3	3	3	1	3	
生徒9	3	2	4	1	2	2	
生徒10	4	5	4	2	1	2	
生徒11	5	5	1	5	2	3	
生徒12	5	1	1	5	5	4	
生徒13	5	1	1	3	5	2	
生徒14	4	2	4	1	3	1	
生徒15	2	2	2	1	2	3	

？

→こんな時に役立つ1つの方法が、クラスタリング

クラスタリング

データ間の距離を計算して、近いもの同士を結びつけて行く。



	洋楽	ドラマ	野球	JPOP	プログラミング	ライブ	バスケ	カラオケ	ペンギン	水泳	登山	舞台鑑賞	パン	ダンス
生徒1	3	5	5	2	1	4	1	2	3	3	3	2	3	
生徒2	3	3	3	2	3	1	5	1	2	5	3	5	5	
生徒3	1	5	4	3	2	3	4	4	1	4	2	4	5	
生徒4	4	3	5	2	1	3	5	4	1	2	2	2	4	
生徒5	2	4	2	3	4	1	3	4	3	4	5	4	3	
生徒6	3	5	1	2	5	2	5	5	2	1	3	5	2	
生徒7	4	2	1	1	2	5	5	4	3	3	5	2	1	
生徒8	2	3	3	3	1	3	1	3	4	3	4	5	5	
生徒9	3	2	4	1	2	2	5	2	4	2	3	2	4	
生徒10	4	5	4	2	1	2	4	3	3	4	4	4	3	
生徒11	5	5	1	5	2	3	1	5	3	2	2	3	5	
生徒12	5	1	1	5	5	4	1	5	4	2	1	4	3	
生徒13	5	1	1	3	5	2	1	5	2	4	3	1	4	
生徒14	4	2	4	1	3	1	2	3	4	2	3	5	3	
生徒15	2	2	2	1	2	3	4	4	4	5	4	2	1	

3次元以上のデータにも対応！

クラスタリングで似ている人を探してみよう！

	洋楽	ドラマ	野球	JPOP	プログラミ	ライブ	バスケ	カラオケ	ペンギン	水泳	登山	舞台鑑賞	パン	ダンス	料理	楽器演奏	昼寝	動画鑑賞	犬	猫
生徒1	3	5	5	2	1	4	1	2	3	3	3	2	3	4	5	1	4	3	5	5
生徒2	3	3	3	2	3	1	5	1	2	5	3	5	5	4	2	1	2	2	3	5
生徒3	1	5	4	3	2	3	4	4	1	4	2	4	5	3	3	3	4	5	4	1
生徒4	4	3	5	2	1	3	5	4	1	2	2	2	4	5	3	2	4	4	1	2
生徒5	2	4	2	3	4	1	3	4	3	4	5	4	3	1	1	2	5	2	3	2
生徒6	3	5	1	2	5	2	5	5	2	1	3	5	2	2	2	5	3	3	4	4
生徒7	4	2	1	1	2	5	5	4	3	3	5	2	1	5	1	3	2	1	2	3
生徒8	2	3	3	3	1	3	1	3	4	3	4	5	5	5	4	1	4	1	4	3
生徒9	3	2	4	1	2	2	5	2	4	2	3	2	4	4	5	4	2	4	1	4
生徒10	4	5	4	2	1	2	4	3	3	4	4	4	3	5	5	4	3	4	5	1
生徒11	5	5	1	5	2	3	1	5	3	2	2	3	5	5	3	5	5	2	5	5
生徒12	5	1	1	5	5	4	1	5	4	2	1	4	3	5	1	5	4	4	2	1
生徒13	5	1	1	3	5	2	1	5	2	4	3	1	4	2	5	4	1	3	4	2
生徒14	4	2	4	1	3	1	2	3	4	2	3	5	3	3	5	3	2	1	1	1
生徒15	2	2	2	1	2	3	4	4	4	5	4	2	1	2	4	2	4	5	2	3
生徒16	5	1	4	4	2	3	4	4	3	4	3	1	4	1	4	3	3	2	5	2
生徒17	5	2	4	1	2	1	3	2	4	3	3	2	1	3	4	4	4	4	4	5
生徒18	2	5	2	3	1	3	5	5	1	1	3	3	2	5	3	4	2	3	4	2
生徒19	1	1	3	3	1	5	5	4	2	4	2	1	2	4	4	3	5	5	4	1
生徒20	1	2	2	1	5	2	1	4	2	3	2	1	5	5	5	3	1	3	2	3
生徒21	3	2	4	4	2	1	1	4	2	5	2	1	1	3	3	2	1	5	3	1
生徒22	2	3	5	4	2	2	2	2	2	5	4	5	2	2	4	1	2	4	4	4
生徒23	5	2	2	1	1	1	5	4	5	2	2	1	3	1	5	5	1	3	4	5
生徒24	4	5	2	5	1	4	3	5	4	3	2	3	1	5	5	5	2	5	4	3
生徒25	5	1	4	4	1	1	1	2	1	5	4	5	3	4	5	2	2	5	3	2
生徒26	5	5	2	3	1	2	5	5	1	3	4	3	2	3	2	2	5	2	4	2
生徒27	3	2	4	5	2	4	1	4	1	3	1	4	4	5	2	4	5	5	2	3
生徒28	5	3	2	5	4	2	3	3	2	1	3	5	1	2	5	4	3	5	3	4

今回使用する環境

オープンソースの統計解析向けプログラミング言語

コマンドの解説(一時停止して見て下さい)

syumi <- read.csv(“syumi.csv",h=T,row.names=1)

変数「syumi」に「syumi.csv」ファイルを読み込む。
h=T ヘッダが存在する ／ row.names=1 1列目のデータを行の名前にする。

syumi.d <- dist(syumi)

dist関数を利用して、データ間の距離を計算する(distance=距離)。

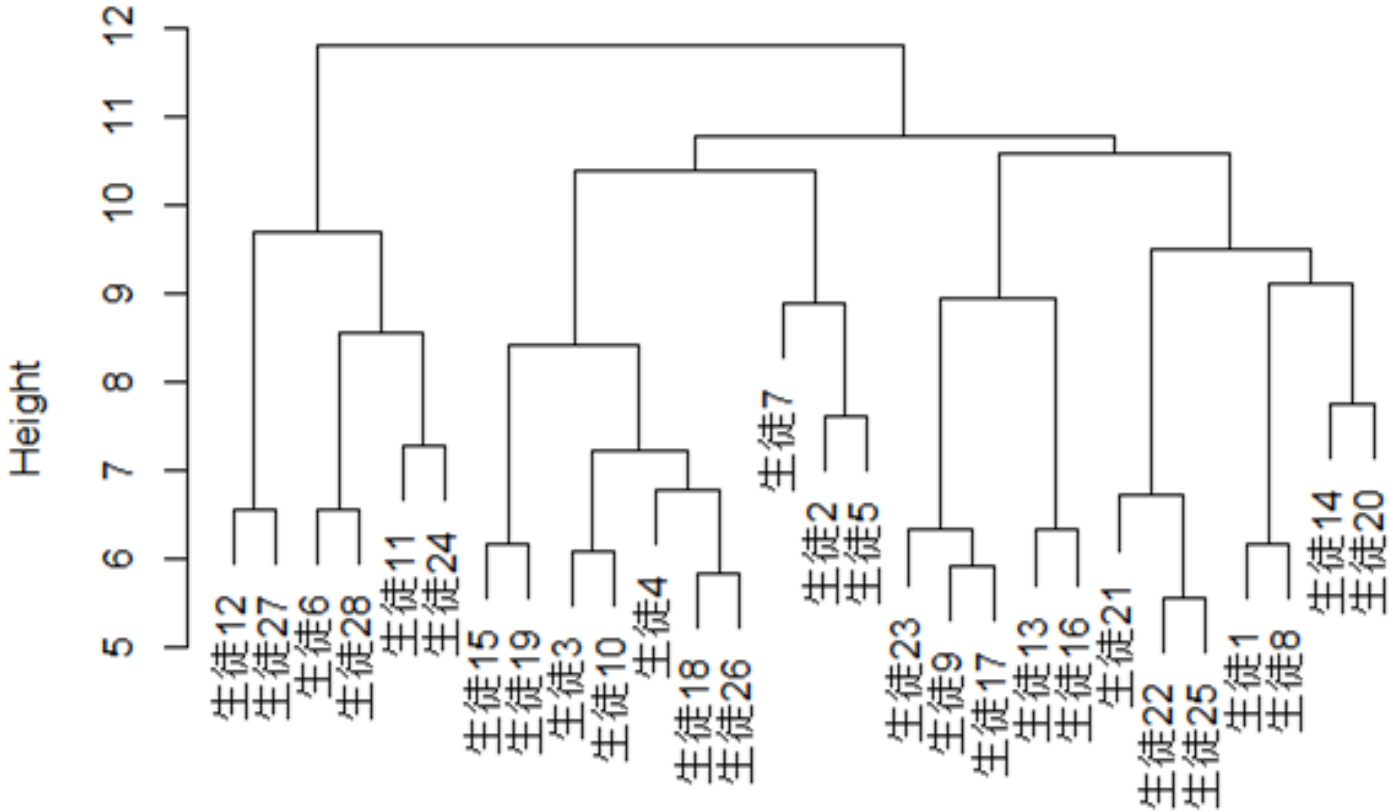
syumi.c <- hclust(syumi.d)

hclust関数を利用して、クラスタリングを行う。

ヘッダ

	洋楽	ドラマ	野球	JPOP	プログラミ	ライブ
生徒1	3	5	5	2	1	4
生徒2	3	3	3	2	3	1
生徒3	1	5	4	3	2	3
生徒4	4	3	5	2	1	3
生徒5	2	4	2	3	4	1
生徒6	3	5	1	2	5	2
生徒7	4	2	1	1	2	5
生徒8	2	3	3	3	1	3
生徒9	3	2	4	1	2	2

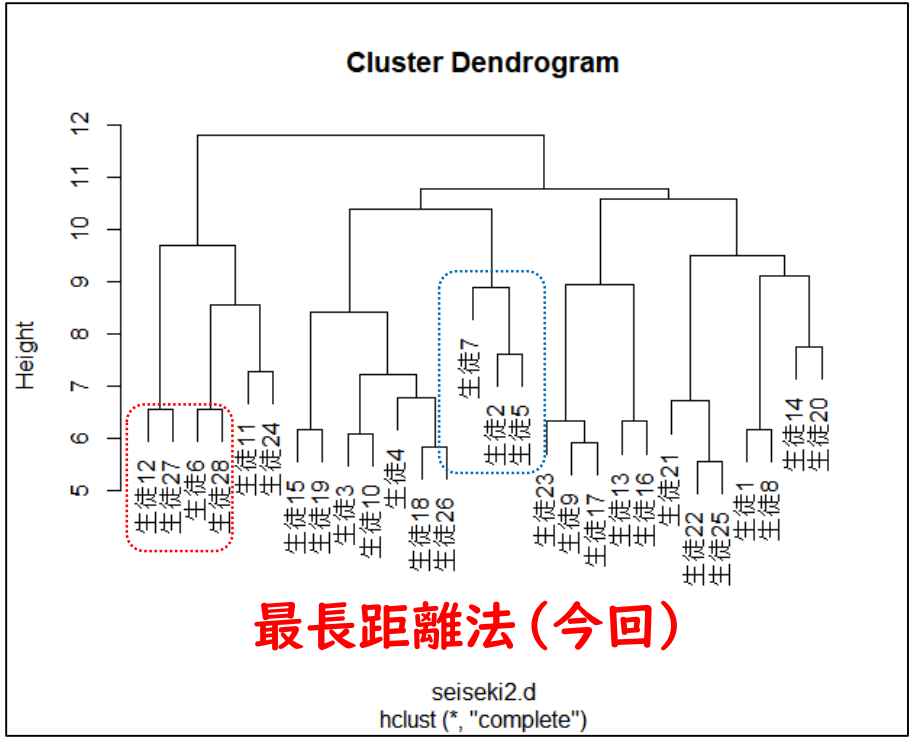
クラスタリングの結果


Height
5
6
7
8
9
10
11
12
生徒12
生徒27
生徒6
生徒28
生徒11
生徒24
生徒15
生徒19
生徒3
生徒10
生徒4
生徒18
生徒26
生徒7
生徒2
生徒5
生徒23
生徒9
生徒17
生徒13
生徒16
生徒21
生徒22
生徒25
生徒1
生徒8
生徒14
生徒20

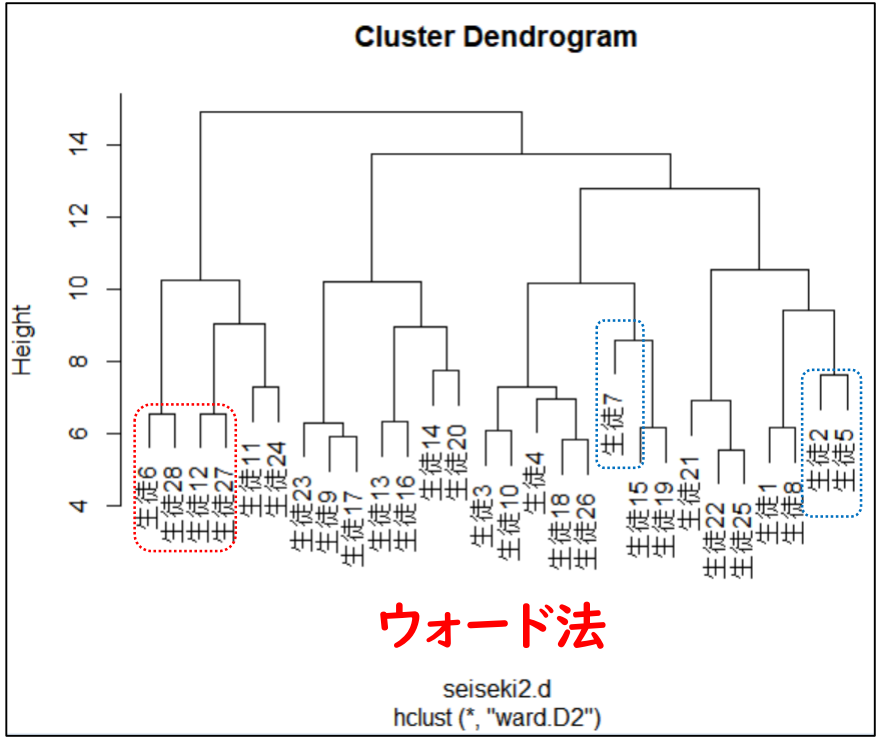
補足

クラスタリングには様々な手法がある。

hclust(〇〇〇〇,method=“ward.D2”)

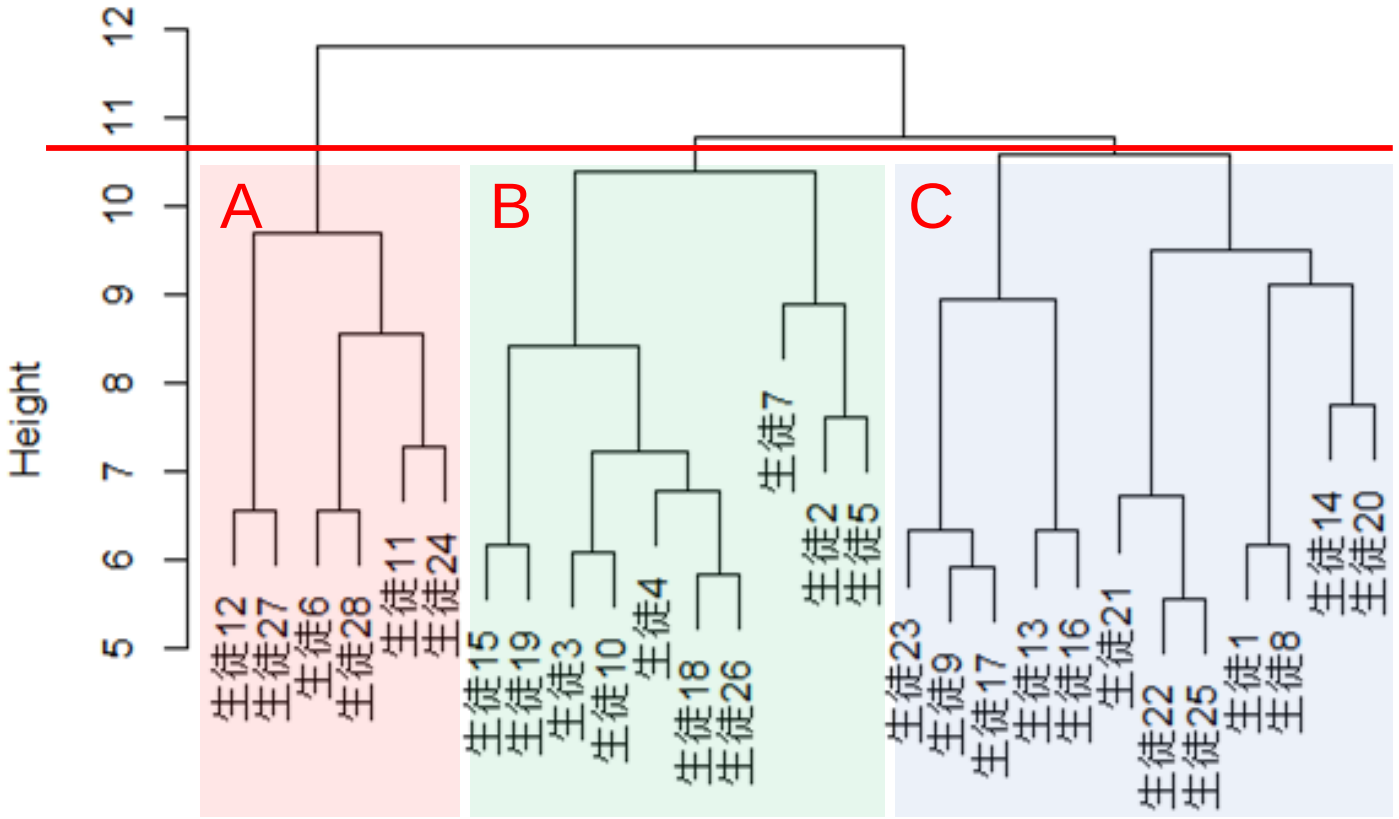


最長距離法（今回）



ウォード法

各グループの傾向を見てみよう



おわりに

他にも様々な例でやってみよう！

- ショッピングの傾向（好きなブランド）
- スポーツ選手の成績
- 音楽の好み（アーティスト・曲調）
- 品目別の売上データ
  -
  -
  -