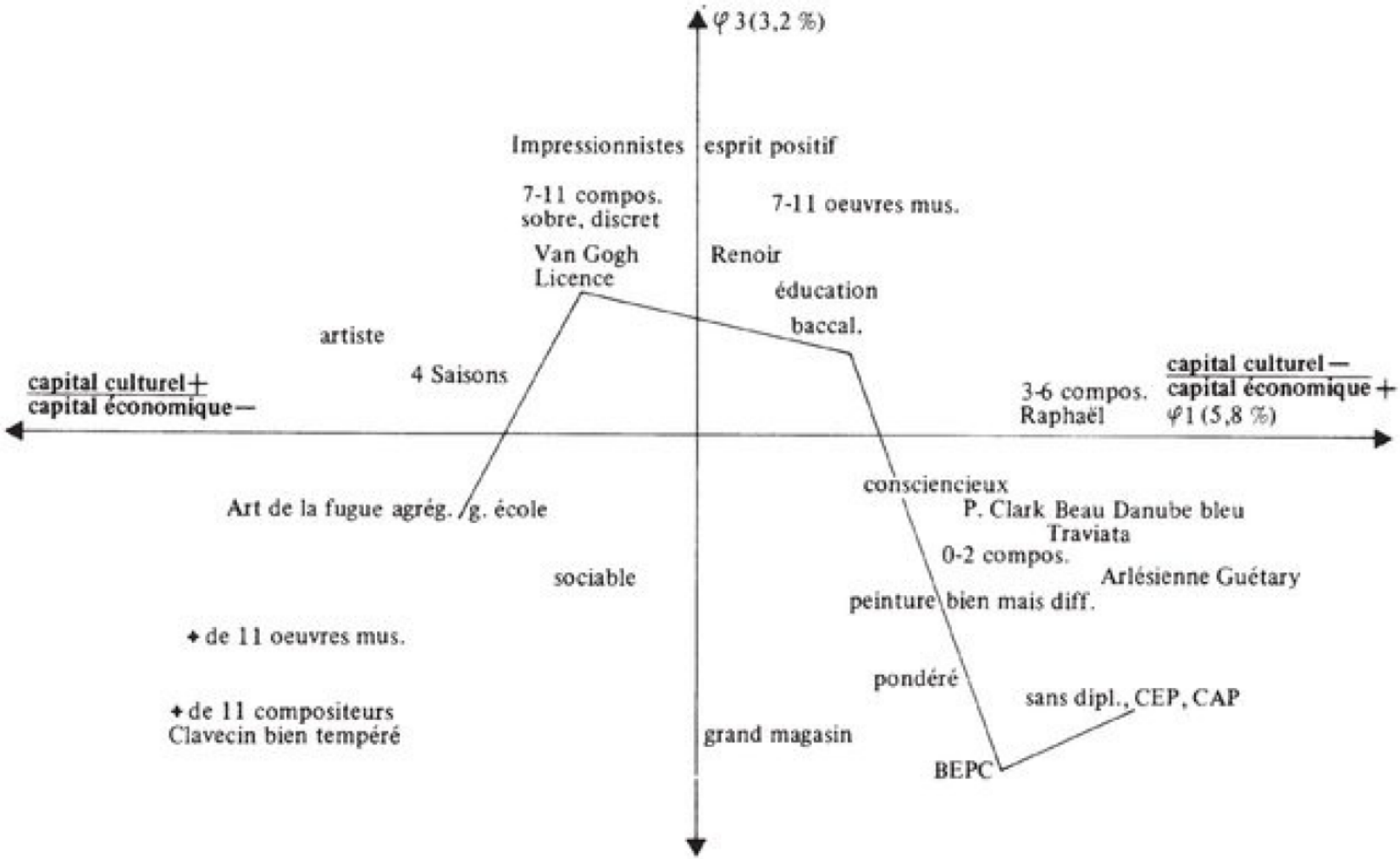
Introduction to data science & artificial intelligence (IF7100)

Arthur Charpentier

#272 Multivariate Analysis: Clusters

été 2020

Clusters


φ 3(3,2 %)
Impressionnistes
esprit positif
7-11 compos.
sobre, discret
7-11 oeuvres mus.
Van Gogh
Licence
Renoir
éducation
baccal.
artiste
4 Saisons
capital culturel+
capital économique−
capital culturel−
capital économique+
3-6 compos.
Raphaël
φ1 (5,8 %)
conscienciеux
Art de la fugue agrég. g. école
P. Clark Beau Danube bleu
Traviata
0-2 compos.
sociable
Arlésienne Guétary
peinture bien mais diff.
• de 11 oeuvres mus.
pondéré
sans dipl., CEP, CAP
• de 11 compositeurs
Clavecin bien tempéré
grand magasin
BEPC

X

k-Means

Consider n observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ in $\mathbb{R}^d$
Given k points $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k$ in $\mathbb{R}^d$ (center of clusters), consider the associated Voronoi diagram.

$$C(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k) = \sum_{i=1}^{n} \left(\min_{j=1,\cdots,k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\| \right)^2$$

But find $\min\{C(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k)\}$ this is a (very) difficult problem

k-Means

xxx

Algorithm 1: k-Means (Lloyd's algorithm)

initialization : draw k centers $\boldsymbol{\mu}_1, \cdot, \boldsymbol{\mu}_k$;

for $b = 1, 2, ..., T$ **do**

  assign: **for** $j = 1, 2, ..., k$ **do**

    $C_j \leftarrow \{i : \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\| \leq \|\boldsymbol{x}_i - \boldsymbol{\mu}_{j'}\|, \ \forall j'\}$;

  update: **for** $j = 1, 2, ..., k$ **do**

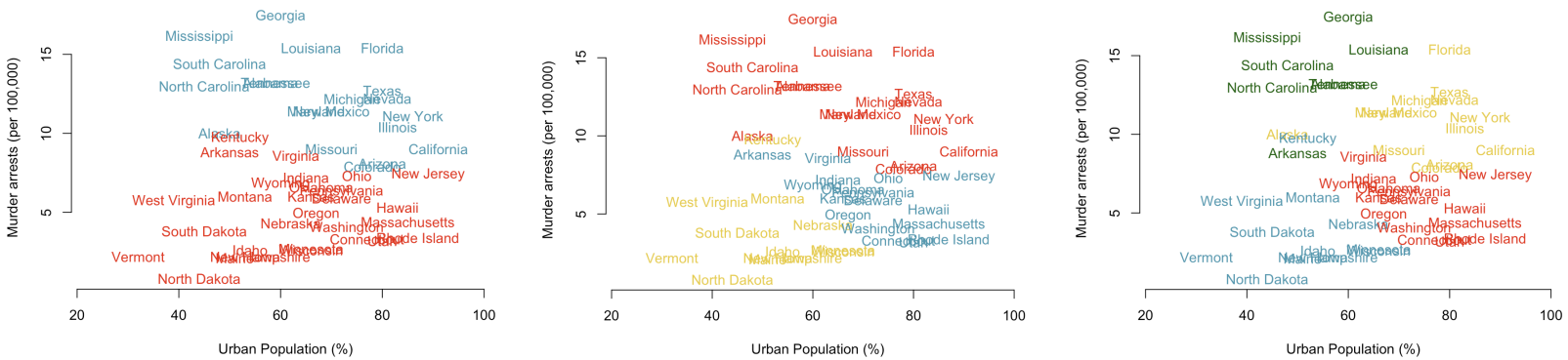
    $$\boldsymbol{\mu}_j \leftarrow \frac{1}{\#C_j} \sum_{i:\boldsymbol{x}_i \in C_j} \boldsymbol{x}_i;$$

Comparing US States

Use normalized variables (see Mahalanobis distance, #421)

```
1                Murder  Assault  UrbanPop        Rape
2  Alabama     1.242564  0.78283  -0.52090  -0.0034164
3  Alaska      0.507862  1.10682  -1.21176   2.4842029
4  Arizona     0.071633  1.47880   0.99898   1.0428783
5  Arkansas    0.232349  0.23086  -1.07359  -0.1849166
6  California  0.278268  1.26281   1.75892   2.0678202
7  Colorado    0.025714  0.39885   0.86080   1.8649672
```

$k = 2$, 3 and 4 (see https://uc-r.github.io

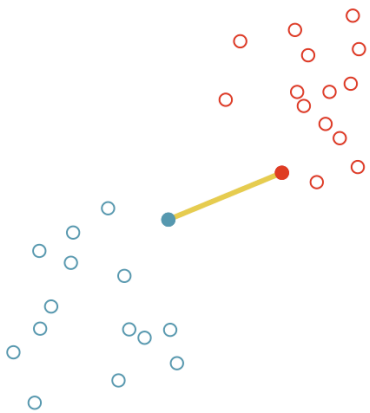


Linkage Methods

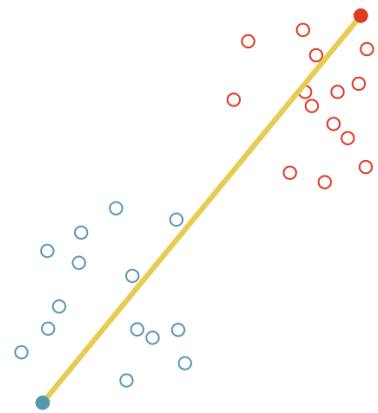
Algorithm 2: (Hierarchical) Linkage Algorithm

initialization : $C_1 = \{x_1\}, \cdots, C_n = \{x_n\}$;
for $i = 1, 2, ..., n-1$ **do**
  $(j^*, k^*) \leftarrow \text{argmin}\{d(C_j, C_k)\}$;
  $C_{j^*} \leftarrow C_{j^*} \cup C_{k^*}$ (and $C_{k^*} \leftarrow \emptyset$);

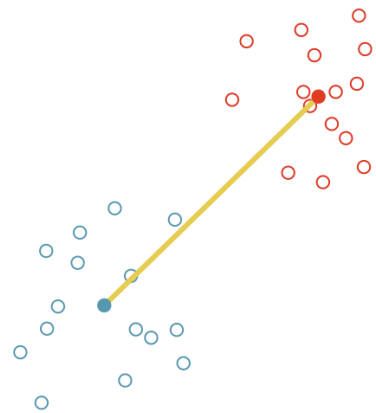
Problem here: what is the distance between two clusters ?



$$\min_{x \in C_1, y \in C_2} \{\|x - y\|\}$$
(single)



$$\max_{x \in C_1, y \in C_2} \{\|x - y\|\}$$
(complete)



$$\|\overline{x} - \overline{y}\|$$

Linkage Methods

Average pairwise distance

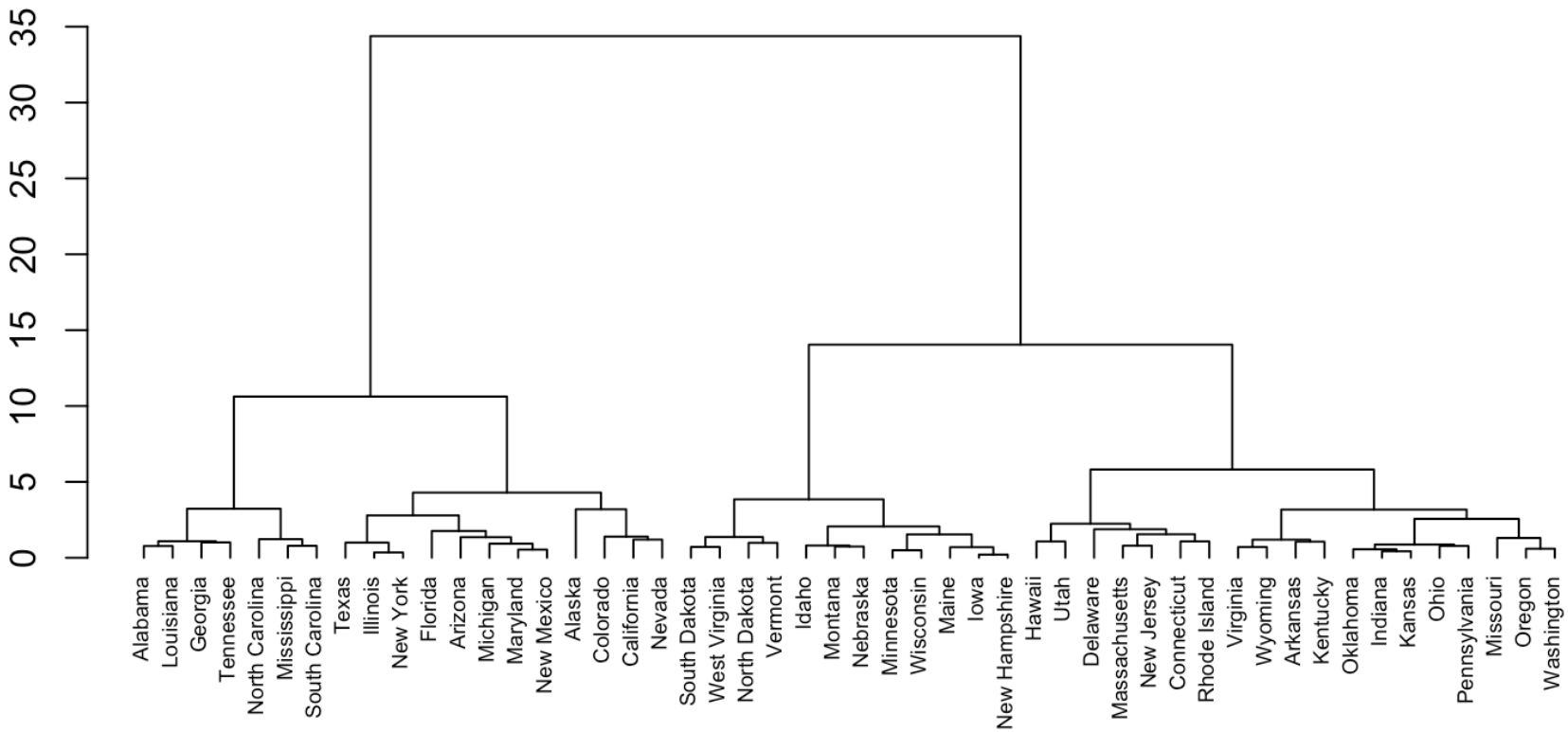
$$d(C_1, C_2)^2 = \frac{1}{n_1 n_2} \sum_{i, \boldsymbol{x}_i \in C_1} \sum_{j, \boldsymbol{y}_j \in C_2} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|$$

Ward's method: take into account the size of merged groups,

$$d(C_1, C_2)^2 = \frac{n_1 n_2}{n_1 + n_2} \cdot \|\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}}\|^2, \ \overline{\boldsymbol{x}} = \frac{1}{n_1} \sum_{i, \boldsymbol{x}_i \in C_1} \boldsymbol{x}_i, \ \overline{\boldsymbol{y}} = \frac{1}{n_2} \sum_{j, \boldsymbol{y}_j \in C_2} \boldsymbol{y}_j.$$

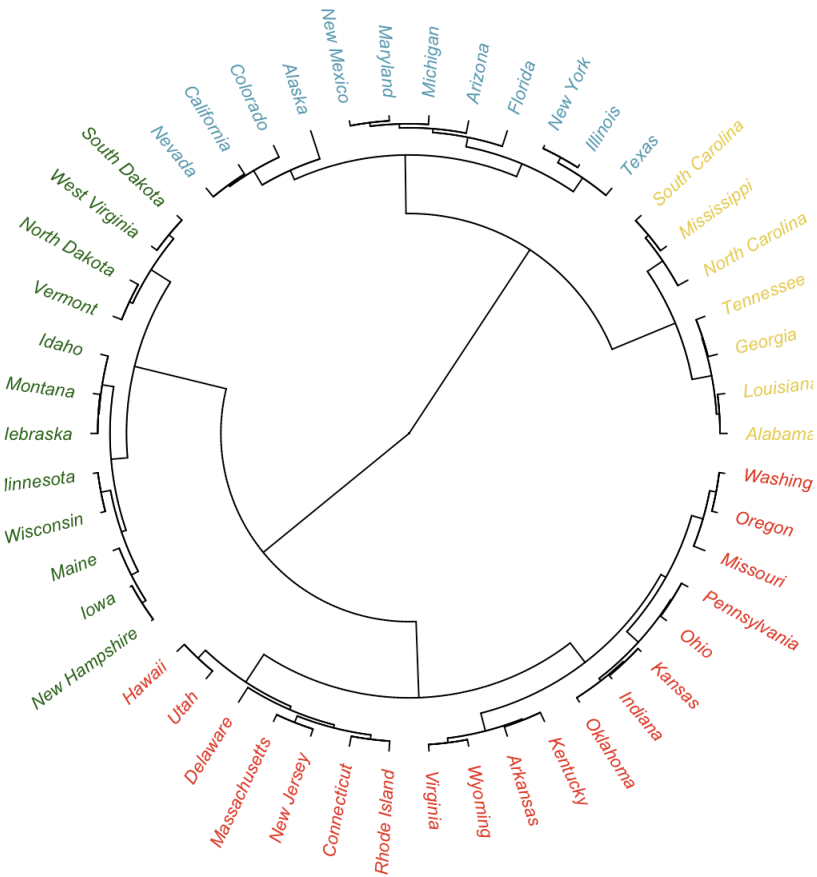
Comparing US States

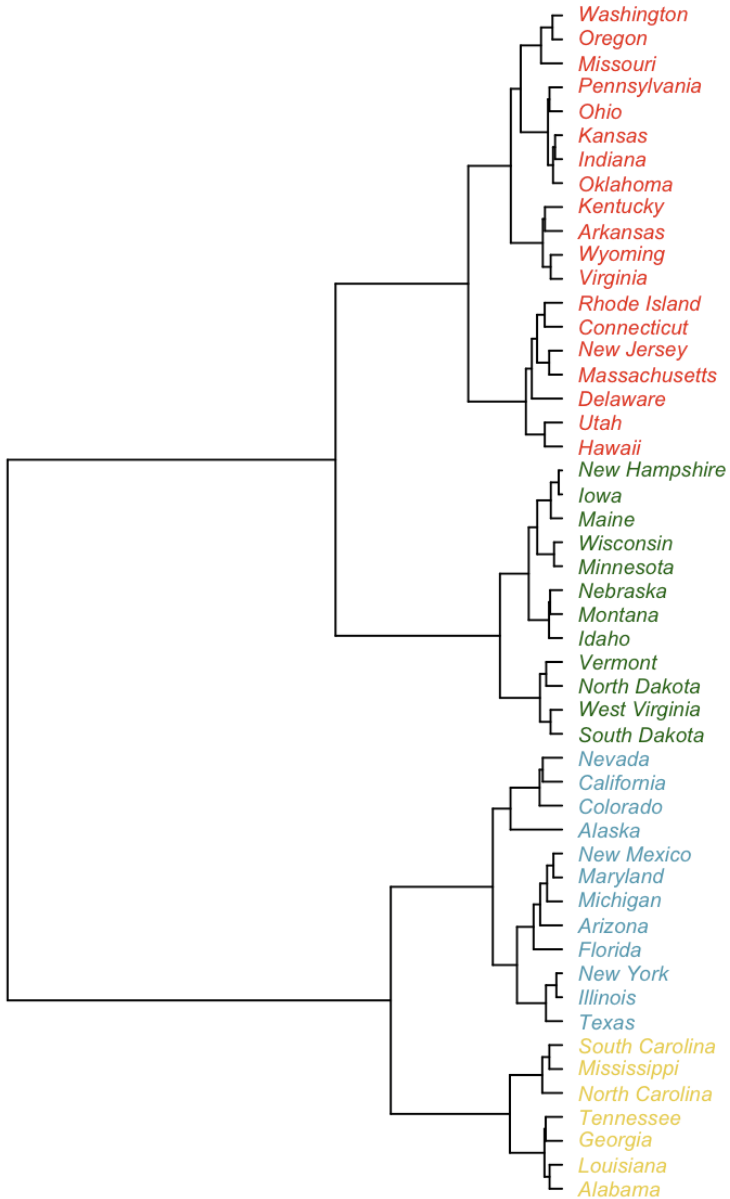
Using Ward's method,



Comparing US States

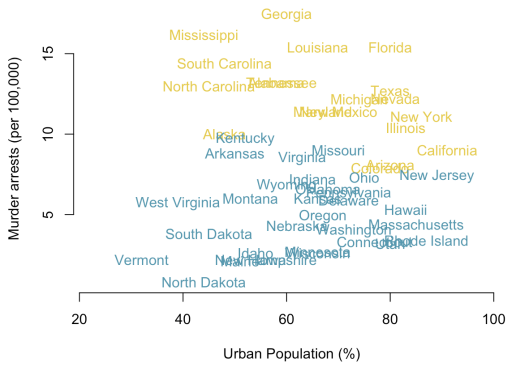
One can use a phylogenetic tree

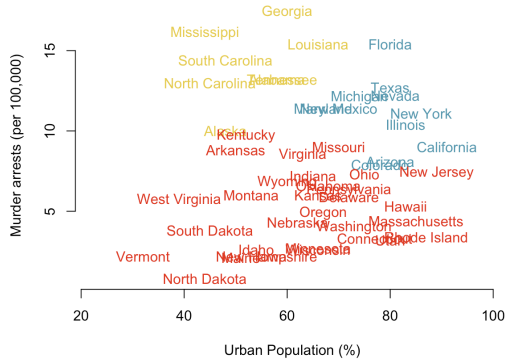


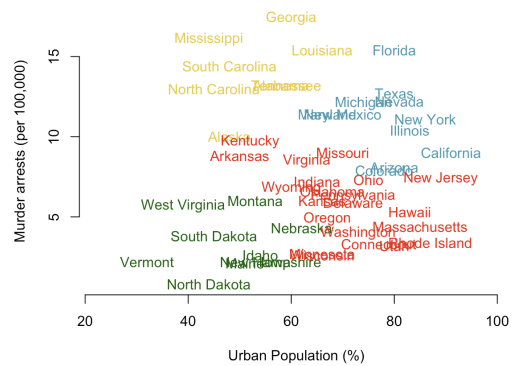


Comparing US States

On the two dimensional representation (urban population, murder rate), when obtain, as *optimal* 2, 3 or 4 classes

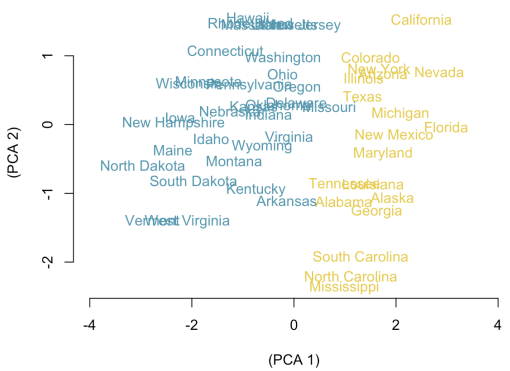


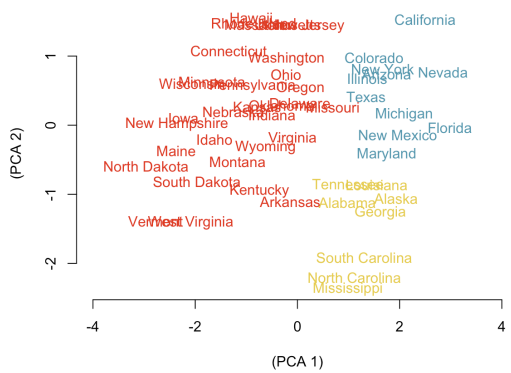


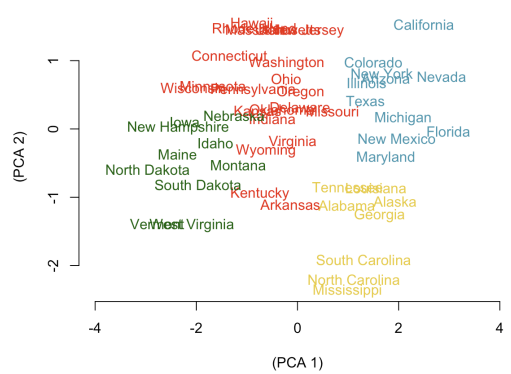


.2cm

or on the first 2 principal components (PCA projection)

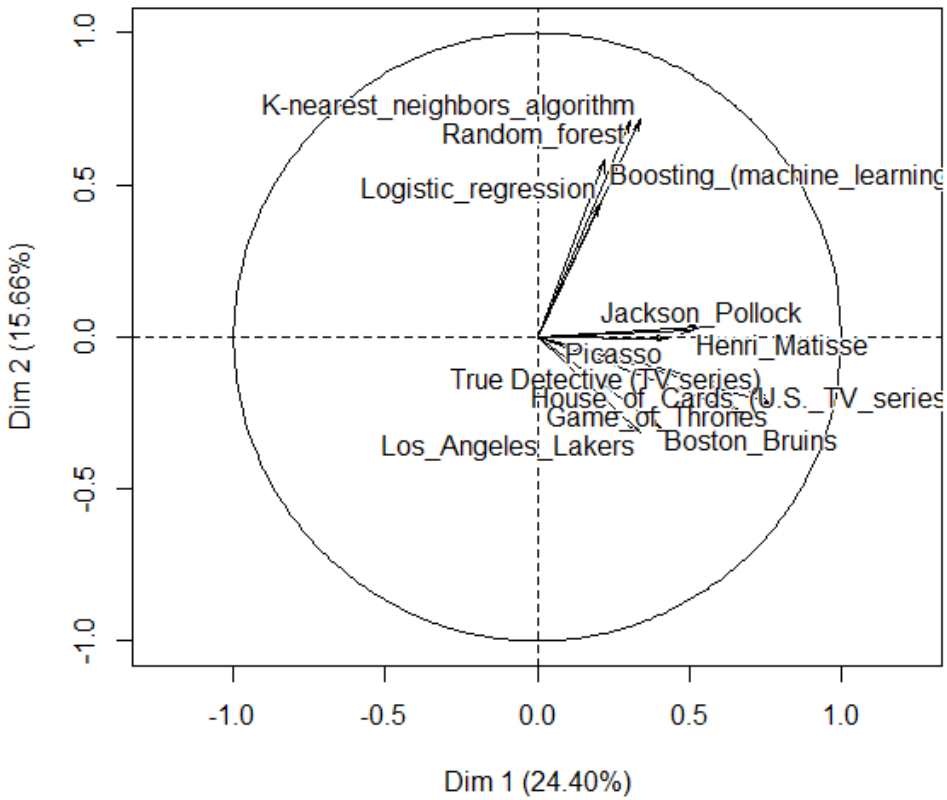


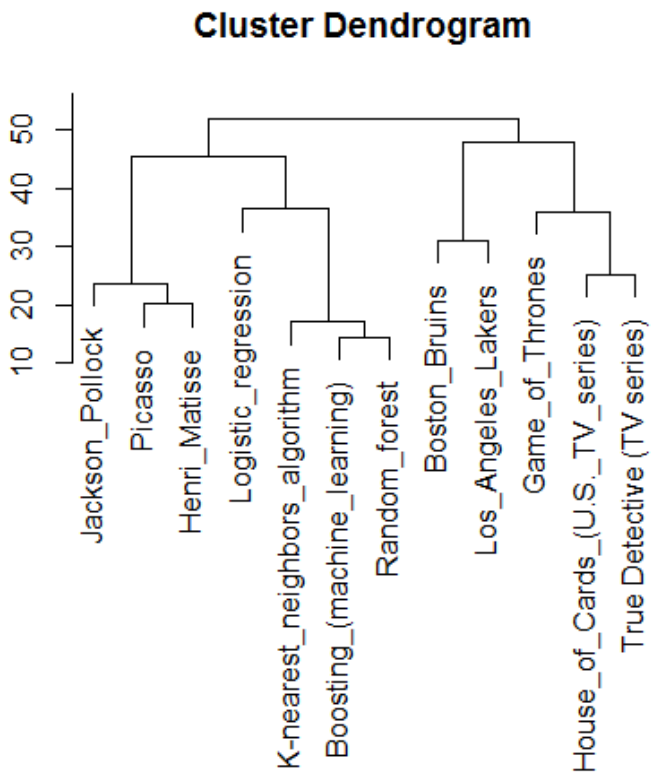




Comparing Wikipedia Pages

Bag of works from some pages Boosting_(machine_learning), Random_forest), K-nearest_neighbors), Logistic_regression), Boston_Bruins), Los_Angeles_Lakers), Game_of_Thrones), House_of_Cards), True_Detective), Picasso), Henri_Matisse), Jackson_Pollock).





Comparing 15 Cities Temperatures (in France)