Wrap-up on Clustering Methods

In this segment, an additional example of clustering will be provided using insurance data from the R package `CASdatasets`.

To run this example, you first need to install this package as follows:

```
install.packages("sp")
install.packages("xts")
install.packages("zoo")
install.packages("CASdatasets",
repos = "http://cas.uqam.ca/pub/", type="source")
library(CASdatasets)
```

In addition, the properties of the various linkage methods will be briefly compared.

PnC demand (1–9)

The "PnCdemand" data file contains indicators of the demand for property and liability insurance in terms of national economic and risk aversion characteristics.

```
data("PnCdemand")
PnC <- PnCdemand[PnCdemand[,3]==6,]
rownames(PnC) <- PnC[,1]
PnC <- PnC[,-c(1:11,14,19,20:22)]
PnC <- PnC[-which(is.na(PnC$Transport)),]
PnC[1:4,]
```

```
          Auto Transport  FireProp   PecLoss  GenLiab    AccSick
Australia 126.7004  38.20918  78.91767  8.133222 38.20920   7.328372
Austria   316.6039  48.97695 182.26715  5.844308 48.97695 219.426127
Belgium   285.4164  50.45312 157.03971 23.504401 50.45309 152.753762
Canada    191.0111  34.04791 121.92755  6.362195 34.04791 140.404066
```

PnC demand (2–9)

There are data for 22 variables and 22 countries from 1987 to 1993.

A complete list of variables can be found here:

`http://cas.uqam.ca/pub/web/CASdatasets-manual.pdf`

The following analysis uses only the premium densities (i.e., gross premiums per capita):

- ✓ automobile
- ✓ fire and other property damage
- ✓ pecuniary (credit loss, surety loss)
- ✓ transport (rail, air and ship)
- ✓ general liability
- ✓ accident and sickness.

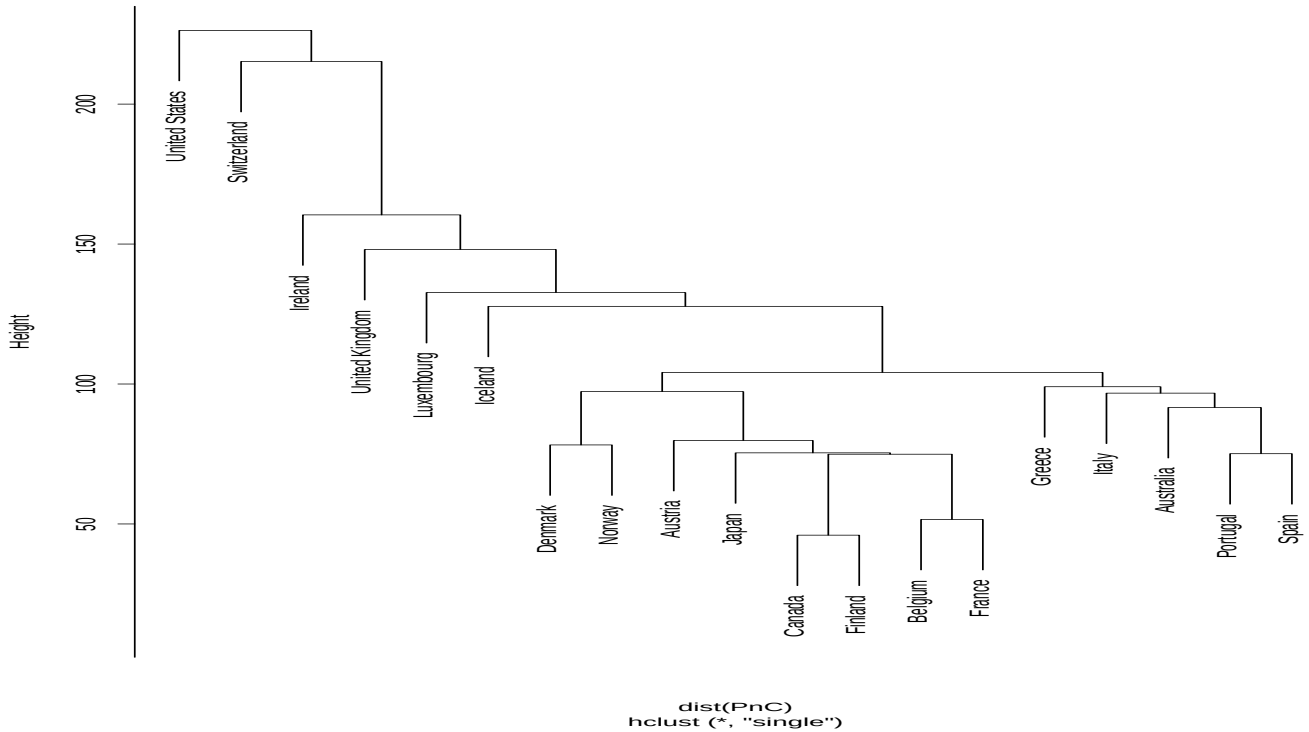
PnC demand (3–9)

```
dist(PnC[1:4,], method = "euclidean")
          Australia   Austria   Belgium
Austria  303.26072
Belgium  230.17052  79.81559
Canada   154.05400 161.57018 105.49807

dist(PnC.stand[1:4,], method = "euclidean")
          Australia  Austria  Belgium
Austria   2.697589
Belgium   2.207120 1.052350
Canada    1.299317 1.565903 1.374771
```

PnC demand (4–9)

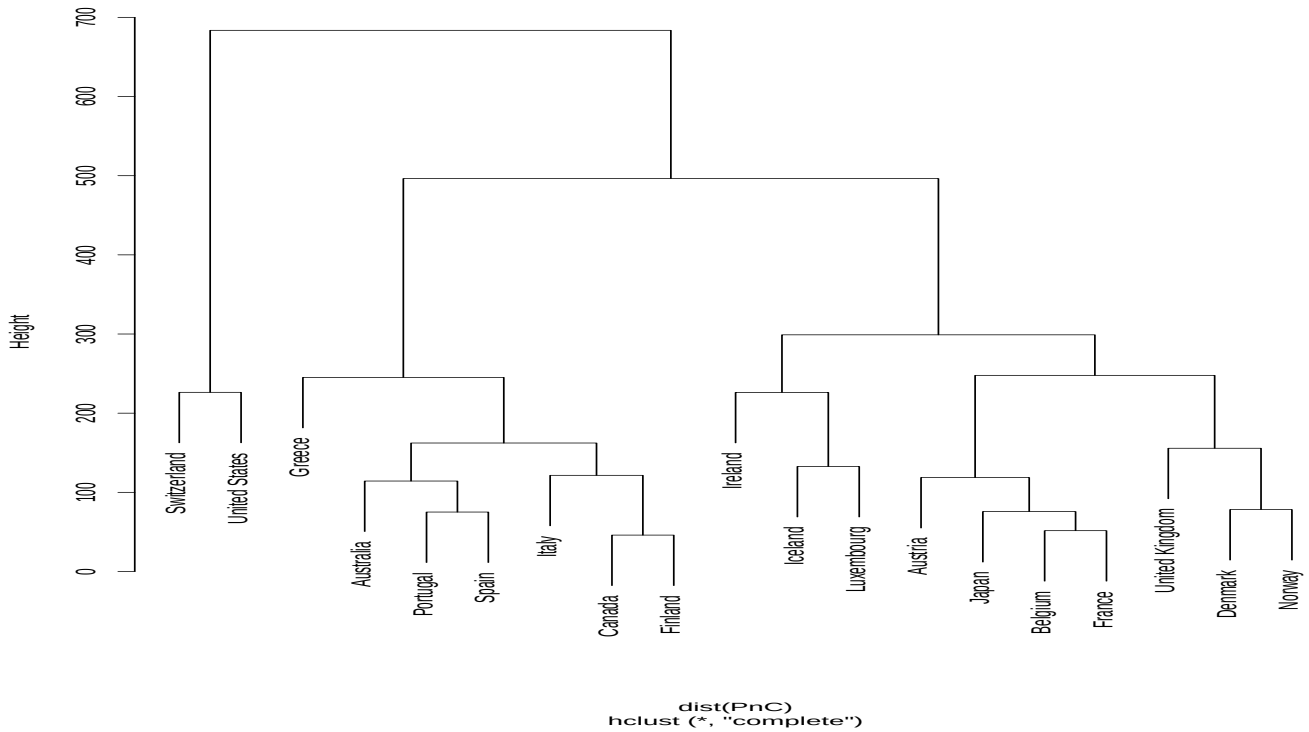
```
plot(hclust(dist(PnC), method = "single"), main = "")
```



This method tends to create large classes, adding variables one by one.

PnC demand (5–9)

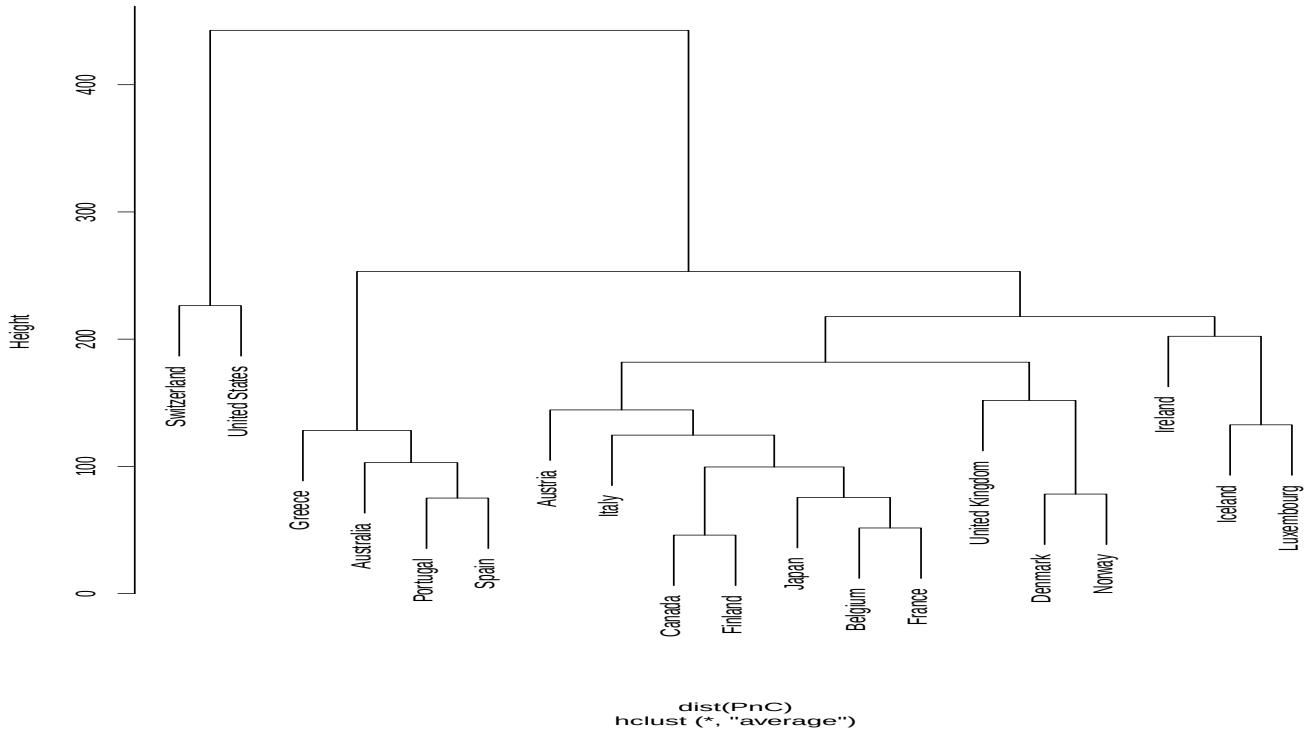
```
plot(hclust(dist(PnC),method="complete"),main="")
```



This method tends to create classes of homogeneous size.
It is sensitive to outliers.

PnC demand (6–9)

```
plot(hclust(dist(PnC),method="average"),main="")
```



This method tends to create classes of equal variance. The result is not invariant to monotone transformations of the distances.

PnC demand (7–9)

As seen before, one can use the command cutree to create classes, viz.

```
groups <- cutree(hc, 3)
names(which(groups==1))
```

```
[1] "Australia" "Canada"    "Finland"   "Greece"    "Italy"
[6] "Portugal"  "Spain"
```

```
names(which(groups==2))
```

```
 [1] "Austria"        "Belgium"        "Denmark"
 [4] "France"         "Iceland"        "Ireland"
 [7] "Japan"          "Luxembourg"     "Norway"
[10] "United Kingdom"
```

```
names(which(groups==3))
```

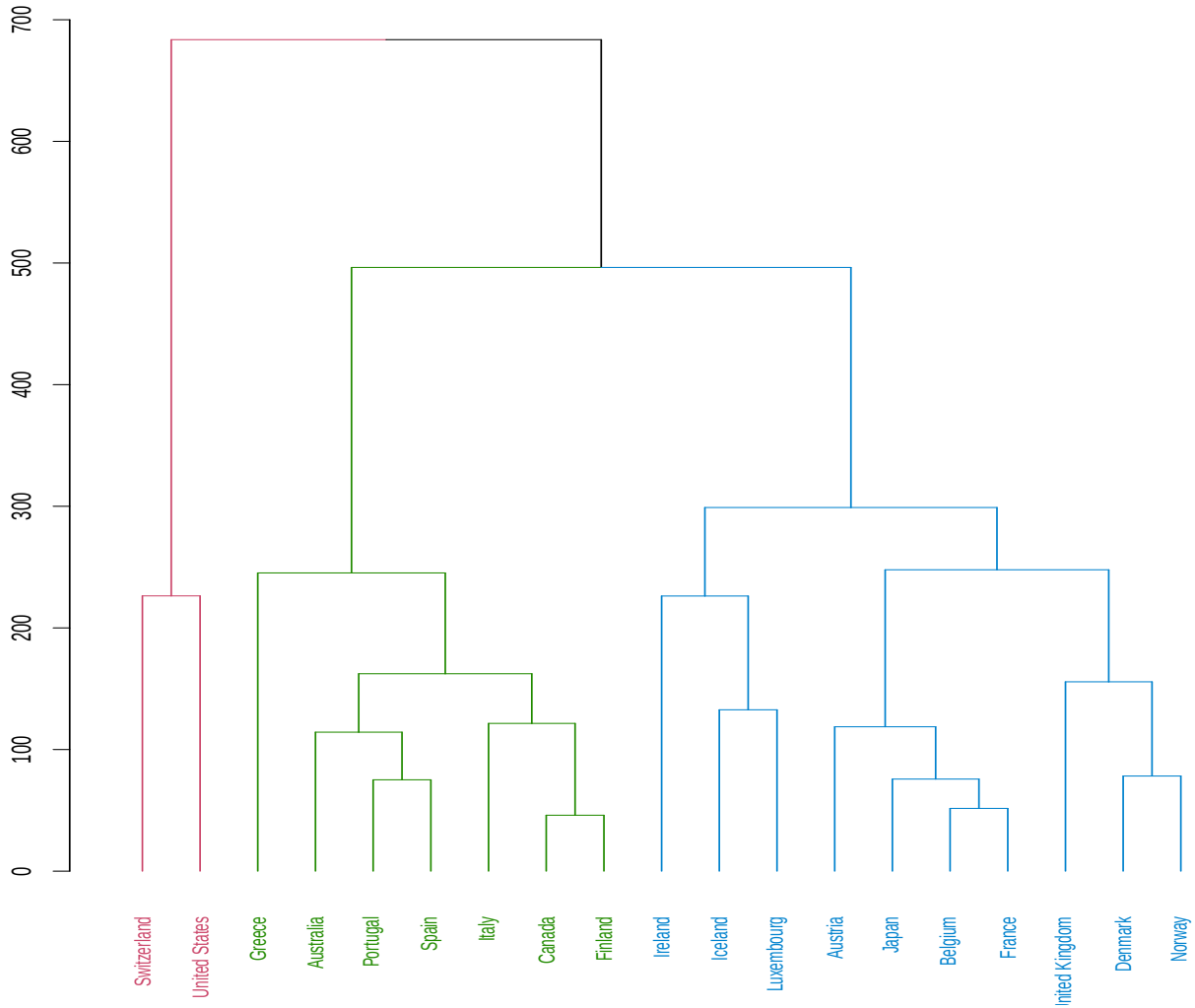
```
[1] "Switzerland"   "United States"
```

PnC demand (8–9)

Here is code that allows one to generate a nice tree:

```
library(dendextend)
dend <- as.dendrogram(hc)
dend <- set(dend,"branches_k_color", k = 3)
dend <- set(dend,"labels_color", k = 3)
dend <- set(dend,"labels_cex", c(.8))
plot(dend)
```

PnC demand (9–9)



Review of Properties: Single Linkage (1–4)

Pros:

- performs well when the variables are of different types;
- has good theoretical properties under certain conditions;
- makes it possible to create groups with irregular shapes;
- is robust to outliers.

Cons:

- tends to create a large group surrounded by small satellite groups;
- looses in efficiency if the underlying groups are regular in shape;
- is well behaved under conditions that are rarely met in practice.

Review of Properties: Complete Linkage (2–4)

Pros:

- performs well when the variables are of different types;
- tends to form groups of equal size.

Cons:

- tends to form groups of equal size;
- is very sensitive to outliers;
- is rarely used in practice.

Review of Properties: Other Methods (3–4)

Average linkage:

- **Pro:** tends to form groups with *small* variance;
- **Con:** tends to form groups with *equal* variance.

Centroid:

- **Pro:** is robust to outliers;
- **Con:** is not very efficient in the absence of outliers.

Median:

- **Pro:** is even more robust to outliers;
- **Con:** is very inefficient in the absence of outliers.

Review of Properties: Ward's Methods (4–4)

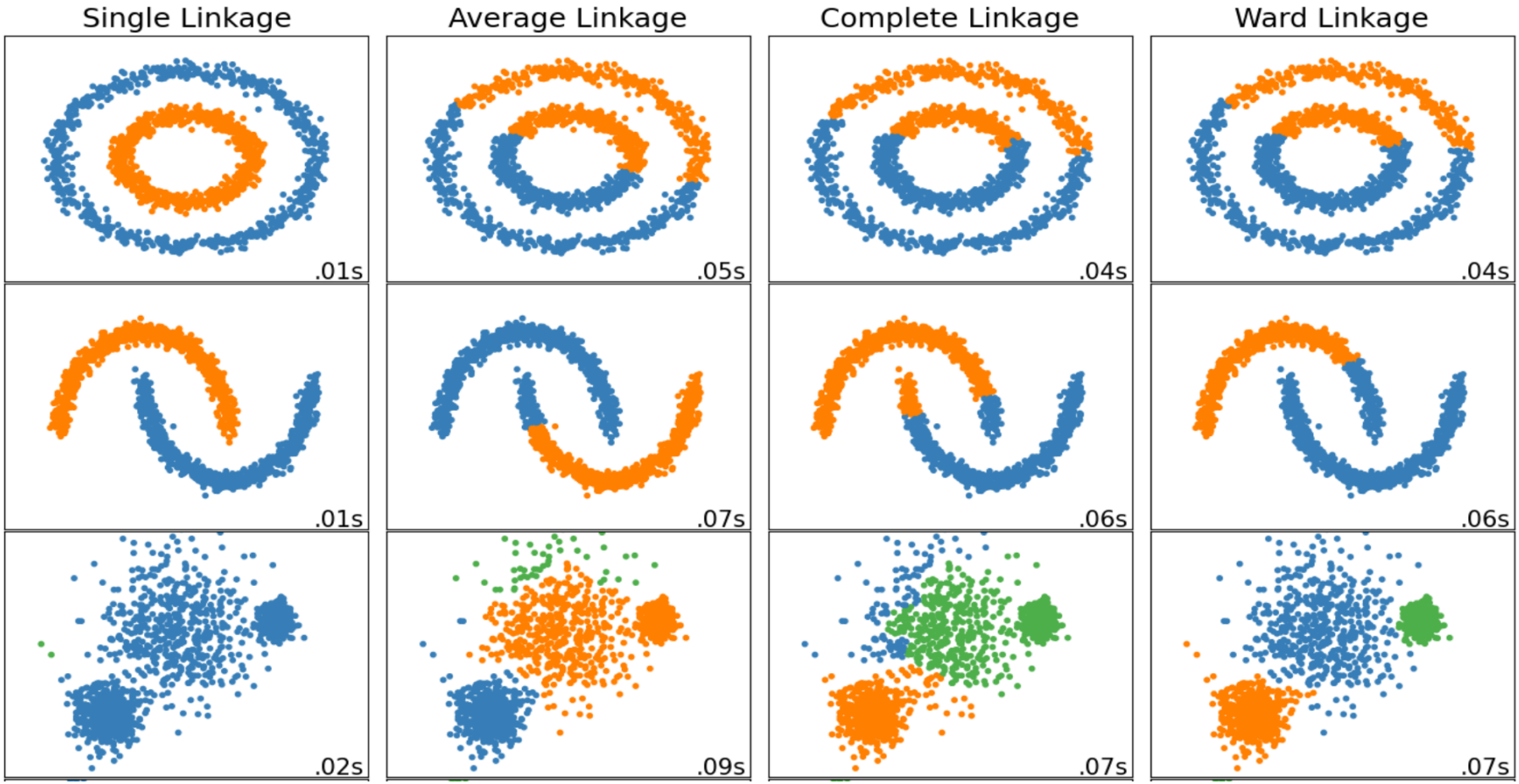
Pro:

- is optimal when the observations are multivariate normal with the same covariance matrix.

Cons:

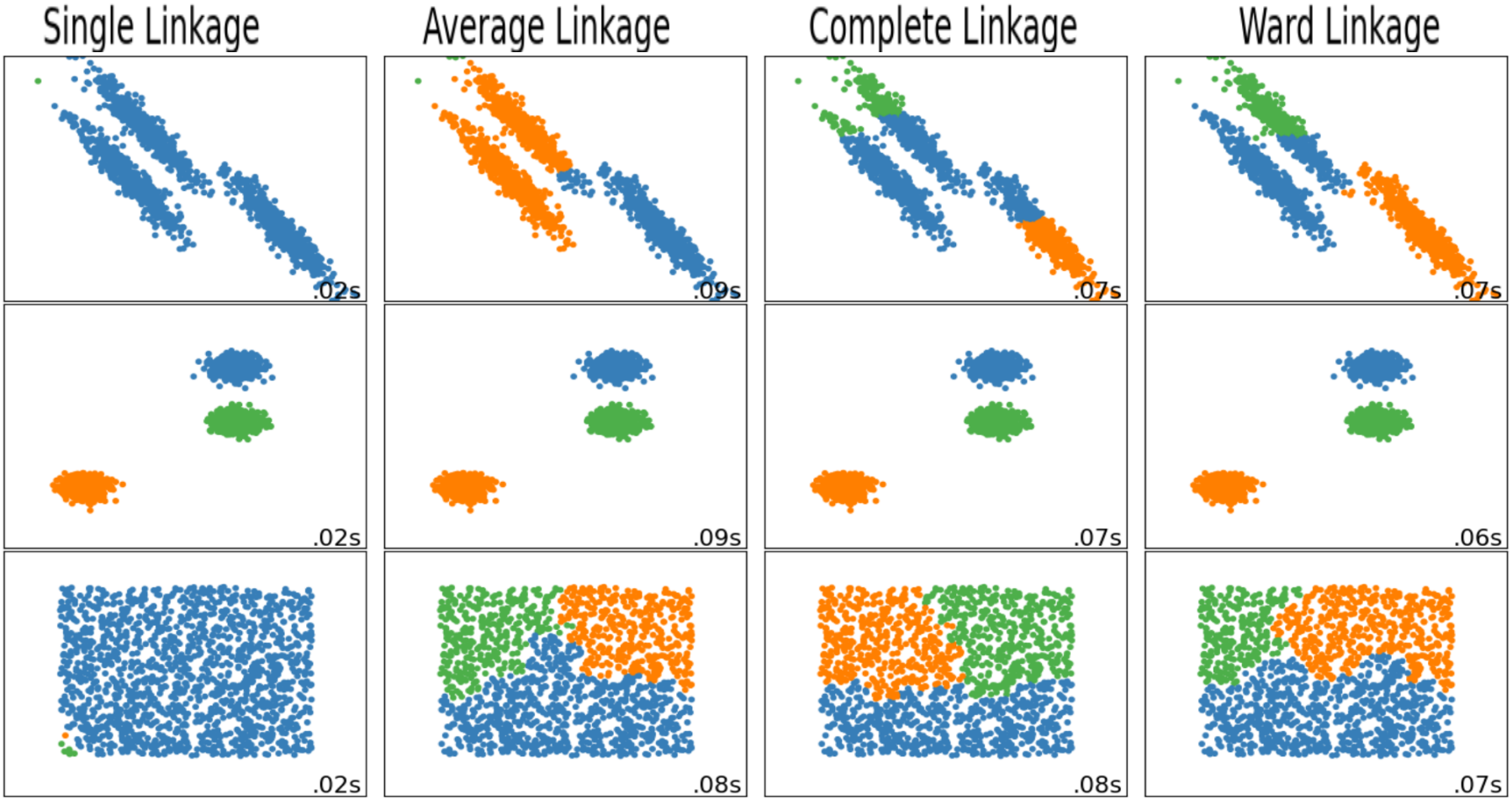
- tends to form small groups;
- tends to form groups of equal size;
- is sensitive to outliers.

Illustration (1–2)



Source: `https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html`

Illustration (2–2)



Total running time of the script: (0 minutes 3.219 seconds)

Source: `https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html`

References

G. James, D. Witten, T. Hastie, R. Tibshirani (2013).
An Introduction to Statistical Learning.
Springer, New York.

L. Kaufman, P.J. Rousseeuw (2005).
Finding Groups in Data: An Introduction to Cluster Analysis.
Wiley, New York.