Package 'datasetsICR'

October 13, 2022

Type Package

Title Datasets from the Book ``An Introduction to Clustering with R''

Version 1.0

Date 2020-05-31

Author Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Maintainer Paolo Giordani <paolo.giordani@uniroma1.it>

Description Companion to the book ``An Introduction to Clustering with R'' by P. Giordani, M.B. Ferraro and F. Martella (Springer, Singapore, 2020). The datasets are used in some case studies throughout the text.

Depends R (>= 3.5.0)

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2020-06-04 11:40:10 UTC

R topics documented:

`datasetsICR-package` *The R Package datasetsICR*

Description

This package contains some datasets analyzed in the book "An Introduction To Clustering With R".

Details

For a complete list, use `library(help = "datasetsICR")`.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

`butterfly` *butterfly dataset*

Description

Synthetic dataset.

Usage

```
data(butterfly)
```

Format

A matrix with 15 rows and 2 columns.

Details

The butterfly dataset motivates the need for the fuzzy approach to clustering. Two clusters are clearly visible, but unit n.8 is between the two clusters, and hence its assignment is complex for standard clustering methods.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)
Ruspini, E.H.: Numerical methods for fuzzy clustering. Inf. Sci. 2, 319-350 (1970)

Examples

```
data(butterfly)
```

`customers` *customers dataset*

Description

Annual spending of a sample of consumers.

Usage

```
data(customers)
```

Format

A data.frame with 440 rows on 8 variables.

Details

The dataset is a sample of 440 customers characterized by 6 continuous variables, giving the annual spending related to different types of goods. The variables are `Fresh`, `Milk`, `Grocery`, `Frozen`, `Detergents_Paper` and `Delicassen`. Two more variables are categorical and provide information on the customer channel (`Channel` with 2 levels: Horeca, i.e., Hotel/Restaurant/Cafe, Retail) and the region (`Region` with 3 levels: Lisbon, Oporto, Other). The categorical variables should not play an active role in the clustering process, but they can be used ex-post to aid cluster interpretation.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

http://archive.ics.uci.edu/ml

References

Abreu, N.: Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon (2011)
Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(customers)
X <- customers[,3:8]
```

`Economics` *Economics dataset*

Description

Performance indicators of Italian Economics faculties.

Usage

```
data(Economics)
```

Format

A data.frame with 55 rows on 13 variables.

Details

55 Italian Economics faculties in the academic year 2009/2010 evaluated by 12 indicators (6 productivity indicators, `P1`, `P2`, `P3A`, `P3B`, `P4A` and `P4B`; 6 teaching indicators, `D1`, `D2`, `D3`, `D4`, `D5` and `D6`). The dataset contains an additional variable, `University_Type`, distinguishing the faculties in `Private` and `Public` type. In the following, the variable description.
`P1`: Rate of persistence between the first and the second academic year.
`P2`: Achieved credits.
`P3A`: Rate of regular students enrolled in the three-year bachelor-level programmes.
`P3B`: Rate of regular students enrolled in the two-year master-level programmes.
`P4A`: Rate of regular graduated students in the three-year bachelor-level programmes.
`P4B`: Rate of regular graduate-students in the two-year master-level programmes.
`D1`: Permanent professors per credits.
`D2`: Permanent professors per enrolled student.
`D3`: Seats per enrolled student in the academic year 2009/2010.
`D4`: Seats per student enrolled in the academic year 2008/2009.
`D5`: Researchers to professors ratio.
`D6`: Monitored teaching activities.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)
Raponi, V., Martella, F., Maruotti, A.: A biclustering approach to university performances: an Italian case study. J. Appl. Stat. 43(1), (2015)

Examples

```
data(Economics)
X <- Economics[,1:12]
class <- Economics[,13]
```

Eurostat *Eurostat dataset*

Description

Economic indicators observed on some European countries in 2018.

Usage

```
data(Eurostat)
```

Format

A data.frame with 29 rows on 4 variables.

Details

The dataset refers to 4 economic indicators for 29 European countries in 2018. The indicators are:
`Inflation`: Annual HICP inflation rate (in percentage);
`Poverty`: People at risk of poverty or social exclusion;
`Unemployment`: Total unemployment rate;
`GDP`: GDP per capita in PPS.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://ec.europa.eu/eurostat/data/database

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(Eurostat)
```

FIFA *FIFA dataset*

Description

Football analytics from the FIFA 19 database.

Usage

```
data(FIFA)
```

Format

A data.frame with 18207 rows on 80 variables.

Details

The dataset contains the detailed attributes for every player registered in the latest edition of FIFA 19 database. Note that some player names display incorrectly because non-ASCII characters have been removed.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://www.kaggle.com/karangadiya/fifa19

References

Gadiya, K.: FIFA 19 complete player dataset (2018)
Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(FIFA)
```

flags *flags dataset*

Description

Survey on university faculty perceptions and practices of using Wikipedia as a teaching resource.

Usage

```
data(flags)
```

Format

A data.frame with 194 rows on 29 variables.

Details

The dataset contains details on flags in terms of quantitative and categorical variables, `bars`, `stripes`, `colours`, `red`, `green`, `blue`, `gold`, `white`, `black`, `orange`, `mainhue`, `circles`, `crosses`, `saltires`, `quarters`, `sunstars`, `crescent`, `triangle`, `icon`, `animate`, `text`, `topleft`, `botright`. The dataset also contains additional variables, `landmass`, `zone`, `area`, `population`, `language` and `religion`, that can be used for interpreting the clusters once they are found.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

http://archive.ics.uci.edu/ml

References

Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019)
Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(flags)
X <- flags[,7:29]
```

german *german dataset*

Description

Bank customers described by a set of attributes.

Usage

```
data(german)
```

Format

A data.frame with 1000 rows on 9 variables.

Details

The dataset contains 1000 bank consumers with 9 mixed measurements. Each row represents a person who takes a bank credit. Each person is either classified as good or bad customer according to her/his failure to repay. This information is described by the variable `Class Risk`. The variables are described below.

`Age`: Age (in years).
`Gender`: `female`, `male`.
`Housing`: `free`, `own`, `rent`.
`Saving accounts`: `little` (< 100 Deutsch Mark), `moderate` (100 <= ... < 500 Deutsch Mark), `quite rich` (500 <= ... < 1000 Deutsch Mark) `rich` (>= 1000 Deutsch Mark).
`Checking account`: `little` (< 0 Deutsch Mark), `moderate` (0 <= ... < 200 Deutsch Mark), `rich`(>= 200 Deutsch Mark). It represents the status of the existing checking account.
`Credit amount`: Credit amount (in Deutsch Mark).
`Duration`: Credit duration (in month).
`Purpose`: `car`, `furniture/equipment`, `radio/TV`, `domestic appliances`, `repairs`, `education`, `business`, `vacation/others`.
`Class Risk`: 1 (Good), 2 (Bad).

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

http://archive.ics.uci.edu/ml

References

Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019)
Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(german)
X <- german[,1:8]
class <- german[,9]
```

`lasvegas.trip` *lasvegas.trip dataset*

Description

Hotels in Las Vegas

Usage

```
data(lasvegas.trip)
```

Format

A data.frame with 21 rows on 9 variables.

Details

The dataset refers to 21 hotels in Las Vegas characterized by 9 variables. 3 of them are quantitative: `Score` (average TripAdvisor scores), `Hotel.stars` and `Nr..rooms`. The remaining 6 are categorical, in particular, binary and concern the presence of a give service: `Pool`, `Gym`, `Tennis.court`, `Spa`, `Casino`, `Free.internet`.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)
Moro, S., Rita, P., Coelho, J.: Stripping customers' feedback on hotels through data mining: the case of Las Vegas strip. Tourism Manage. Persp. 23, 41-52 (2017)

Examples

```
data(lasvegas.trip)
```

NBA.48 *NBA.48 dataset*

Description

Basketball analytics (NBA regular season 2018-19)

Usage

```
data(NBA.48)
```

Format

A data.frame with 530 rows on 29 variables.

Details

The dataset refers to the statistics of 530 players registered in the NBA regular season 2018-19. Note that statistics are normalized per 48 minutes.
The variables are: PLAYER, TEAM, AGE, GP (Games Played), W (Wins), L (Losses), MIN (Minutes Played), PTS (Points), FGM (Field Goals Made), FGA (Field Goals Attempted), FG. (Field Goal Percentage), X3PM (3 Point Field Goals Made), X3PA (3 Point Field Goals Attempted), X3P. (3 Point Field Goals Percentage), FTM (Free Throws Made), FTA (Free Throws Attempted), FT. (Free Throw Percentage), OREB (Offensive Rebounds), DREB (Defensive Rebounds), REB (Rebounds), AST (Assists), TOV (Turnovers), STL (Steals), BLK (Blocks), PF (Personal Fouls), FP (Fantasy Points), DD2 (Double doubles), TD3 (Triple doubles), X... (Plus Minus).

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://stats.nba.com/

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

See Also

NBA.game, NBA.external, NBA.efficiency

Examples

```
data(NBA.48)
```

NBA.efficiency *NBA.efficiency dataset*

Description

Basketball analytics (NBA regular season 2018-19)

Usage

```
data(NBA.efficiency)
```

Format

A data.frame with 258 rows on 2 variables.

Details

The dataset refers to the efficiency values of 258 players for the NBA regular season 2018-19. They can be used ex-post to aid interpretation of clusters obtained by using the statistics in NBA.48. Note that efficiency is observed for a subset of players of NBA.48.
The variables are: Player, EFF.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://stats.nba.com/

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

See Also

NBA.48, NBA.game, NBA.external

Examples

```
data(NBA.efficiency)
```

NBA.external *NBA.external dataset*

Description

Basketball analytics (NBA regular season 2018-19)

Usage

```
data(NBA.external)
```

Format

A data.frame with 530 rows on 10 variables.

Details

The dataset refers to the characteristics of 530 players for the NBA regular season 2018-19. They can be used ex-post to aid interpretation of clusters obtained by using the statistics in NBA.48.
The variables are: PLAYER, FORWARD, CENTER, GUARD, ROOKIE, SOPHOMORE, VETERAN, 1ST ROUND, 2ND ROUND, UNDRAFTED.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://stats.nba.com/

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

See Also

NBA.48, NBA.game, NBA.efficiency

Examples

```
data(NBA.external)
```

NBA.game *NBA.game dataset*

Description

Basketball analytics (NBA regular season 2018-19)

Usage

```
data(NBA.game)
```

Format

A data.frame with 530 rows on 29 variables.

Details

The dataset refers to the statistics of 530 players registered in the NBA regular season 2018-19.
The variables are: PLAYER, TEAM, AGE, GP (Games Played), W (Wins), L (Losses), MIN (Minutes Played), PTS (Points), FGM (Field Goals Made), FGA (Field Goals Attempted), FG. (Field Goal Percentage), X3PM (3 Point Field Goals Made), X3PA (3 Point Field Goals Attempted), X3P. (3 Point Field Goals Percentage), FTM (Free Throws Made), FTA (Free Throws Attempted), FT. (Free Throw Percentage), OREB (Offensive Rebounds), DREB (Defensive Rebounds), REB (Rebounds), AST (Assists), TOV (Turnovers), STL (Steals), BLK (Blocks), PF (Personal Fouls), FP (Fantasy Points), DD2 (Double doubles), TD3 (Triple doubles), X... (Plus Minus).

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://stats.nba.com/

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

See Also

NBA.48,NBA.external, NBA.efficiency

Examples

```
data(NBA.game)
```

seeds *seeds dataset*

Description

Measurements of geometrical properties of kernels belonging to three different varieties of wheat.

Usage

```
data(seeds)
```

Format

A data.frame with 210 rows on 8 variables (including 1 classification variable).

Details

The dataset is about 210 wheat grains belonging to three different varieties on which 7 quantitative variables related to the internal kernel structure detected by using a soft X-ray technique are observed. The information on the varietes is given by `variety` and the remaining quantitative variables are `area`, `perimeter`, `compactness`, `length of kernel`, `width of kernel`, `asymmetry coefficient`, `length of kernel groove`, `variety`.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

http://archive.ics.uci.edu/ml

References

Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019)
Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(seeds)
X <- seeds[, 1:7]
class <- seeds[,8]
```

USstate *US state dataset*

Description

FIPS codes for the US states

Usage

```
data(USstate)
```

Format

A data.frame with 50 rows on 3 variables.

Details

The dataset refers to the FIPS codes for the US states. The variables are `fips` (FIPS State Numeric Code), `usps` (Official USPS Code) and `name` (Name).

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

https://www.census.gov/

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(USstate)
```

wiki4HE *wiki4HE dataset*

Description

Survey on university faculty perceptions and practices of using Wikipedia as a teaching resource.

Usage

```
data(wiki4HE)
```

Format

A data.frame with 913 rows on 53 variables.

Details

The dataset contains socio-demographic characteristics and the answers of 913 university faculty members to questions on the use of Wikipedia as a teaching resource (5-level Likert scale). The variables referring to the socio-demographic characteristics are AGE, GENDER, DOMAIN, PhD, YEARSEXP, UNIVERSITY, UOC_POSITION, OTHER_POSITION OTHERSTATUS and USERWIKI. The variables referring to the survey are PU1, PU2, PU3, PEU1, PEU2, PEU3, ENJ1, ENJ2, Qu1, Qu2, Qu3, Qu4, Qu5, Vis1, Vis2, Vis3, Im1, Im2, Im3, SA1, SA2, SA3, Use1, Use2, Use3, Use4, Use5, Pf1, Pf2, Pf3, JR1, JR2, BI1, BI2, Inc1, Inc2, Inc3, Inc4, Exp1, Exp2, Exp3, Exp4, Exp5.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

http://archive.ics.uci.edu/ml

References

Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Meseguer, A., Aibar, E., Llados, J., Minguillon, J., Lerga, M.: Factors that influence the teaching use of Wikipedia in higher education. J. Assoc. Inf. Sci. Tech. 67, 1224-1232 (2015)

Examples

```
data(wiki4HE)
```

wine *wine dataset*

Description

Chemical analysis of wines grown in the same region in Italy but derived from 3 different cultivars.

Usage

```
data(wine)
```

Format

A data.frame with 178 rows on 14 variables (including 1 classification variable).

Details

The dataset includes 178 Italian wines characterized by 13 constituents (quantitative variables). The dataset contains an additional variable, Class, distinguishing the wines in 3 groups according to the cultivar. The quantitative variables are Class, Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline.

Author(s)

Paolo Giordani, Maria Brigida Ferraro, Francesca Martella

Source

http://archive.ics.uci.edu/ml

References

Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019)
Giordani, P., Ferraro, M.B., Martella, F.: An Introduction to Clustering with R. Springer, Singapore (2020)

Examples

```
data(wine)
X <- wine[,-1]
class <- wine[,1]
```

Index