Package 'dataSDA'

June 19, 2023

Type Package

Title Data Sets for Symbolic Data Analysis

Version 0.1.0

Date 2023-06-18

Author Po-Wei Chen [aut],
Han-Ming Wu [cre]

Maintainer Han-Ming Wu <wuhm@g.nccu.edu.tw>

Description Collects a diverse range of symbolic data and offers a comprehensive set of functions that facilitate the conversion of traditional data into the symbolic data format.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Depends R (>= 4.0.0)

Suggests testthat (>= 2.1.0), knitr, rmarkdown, RSDA

VignetteBuilder knitr

Imports magrittr, tidyr

NeedsCompilation no

Repository CRAN

Date/Publication 2023-06-19 15:30:14 UTC

R topics documented:

`Abalone.iGAP` *Abalone iGAP format Dataset*

Description

A interval-valued data set containing 24 units, created from from the Abalone dataset (UCI Machine Learning Repository), after aggregating by sex and age.

Usage

```
data(Abalone.iGAP)
```

Format

An object of class `data.frame` with 24 rows and 7 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(Abalone.iGAP)
```

`age_cholesterol_weight.int` *Age-cholesterol-weight Interval-Valued Dataset*

Description

Age-cholesterol-weight Interval-Valued Dataset.

Usage

```
data(age_cholesterol_weight.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 7 rows and 4 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(age_cholesterol_weight.int)
```

`airline_flights` *Airline Flights Dataset*

Description

Airline Flights Dataset.

Usage

```
data(airline_flights)
```

Format

An object of class `data.frame` with 16 rows and 17 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(airline_flights)
```

---

airline_flights2 *Airline Flights Modal-Valued Dataset*

Description

Airline Flights Modal-Valued Dataset.

Usage

```
data(airline_flights2)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 16 rows and 6 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(airline_flights2)
```

baseball.int *Baseball Interval-Valued Dataset*

Description

Baseball Interval-Valued Dataset.

Usage

```
data(baseball.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 19 rows and 3 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(baseball.int)
```

bird.int *Bird Interval-Valued Dataset*

Description

Bird Interval-Valued Dataset.

Usage

```
data(bird.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 20 rows and 2 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(bird.int)
```

blood_pressure.int *Blood Pressure Interval-Valued Dataset*

Description

blood pressure Interval-Valued Dataset.

Usage

```
data(blood_pressure.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 15 rows and 3 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(blood_pressure.int)
```

car.int *Car Interval-Valued Dataset*

Description

Car Interval-Valued Dataset.

Usage

```
data(car.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 8 rows and 5 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(car.int)
```

Cars.int *Cars Interval Dataset*

Description

Cars Interval Dataset generated from Cars dataset. This data set consist of the intervals for four characteristics (Price, EngineCapacity, TopSpeed and Acceleration) of 27 cars models partitioned into four different classes (Utilitarian, Berlina, Sportive and Luxury).

Usage

```
data(Cars.int)
```

Format

A data frame containing 27 observations on 5 variables, the first five with the interval characteristics for 27 car models, the last one a factor indicating the model class.

Source

https://CRAN.R-project.org/package=MAINT.Data

Examples

```
data(Cars.int)
```

ChinaTemp.int *China Temperatures Interval Dataset*

Description

China Temperatures Interval Dataset generated from ChinaTemp dataset. This data set consist of the intervals of observed temperatures (Celsius scale) in each of the four quarters, Q_1 to Q_4, of the years 1974 to 1988 in 60 chinese meteorologic stations; one outlier observation (YinChuan_1982) has been discarded. The 60 stations belong to different regions in China, which therefore define a partition of the 899 stations-year combinations.

Usage

```
data(ChinaTemp.int)
```

Format

A data frame containing 899 observations on 5 variables, the first four with the temperatures by quarter in the 899 stations-year combinations, the last one a factor indicating the geographic region of each station.

Source

https://CRAN.R-project.org/package=MAINT.Data

Examples

```
data(ChinaTemp.int)
```

`clean_colnames` *clean_colnames*

Description

This function is used to clean up variable names to conform to the RSDA format.

Usage

```
clean_colnames(data)
```

Arguments

`data`	The conventional data.

Value

Data after cleaning variable names.

Examples

```
data(mushroom)
mushroom.clean <- clean_colnames(data = mushroom)
```

`crime` *Crime demographics Dataset*

Description

Crime demographics Dataset.

Usage

```
data(crime)
```

Format

An object of class `data.frame` with 15 rows and 7 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(crime)
```

`crime2` *Crime demographics Modal-Valued Dataset*

Description

Crime demographics Modal-Valued Dataset.

Usage

```
data(crime2)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 15 rows and 3 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(crime2)
```

`finance.int` *Finance Interval-Valued Dataset*

Description

Finance Interval-Valued Dataset.

Usage

```
data(finance.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 14 rows and 7 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(finance.int)
```

---

fuel_consumption *Fuel Consumption Dataset*

Description

Fuel Consumption Dataset.

Usage

```
data(fuel_consumption)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 10 rows and 3 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(fuel_consumption)
```

health_insurance *Health Insurance Dataset*

Description

Health Insurance Dataset.

Usage

```
data(health_insurance)
```

Format

An object of class `data.frame` with 51 rows and 30 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(health_insurance)
```

health_insurance2 *Health Insurance Modal-Valued Dataset*

Description

Health Insurance Modal-Valued Dataset.

Usage

```
data(health_insurance2)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 6 rows and 6 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(health_insurance2)
```

hierarchy *Hierarchy Dataset*

Description

Hierarchy Dataset.

Usage

```
data(hierarchy)
```

Format

An object of class `data.frame` with 20 rows and 6 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(hierarchy)
```

hierarchy.int *Hierarchy Interval-Valued Dataset*

Description

Hierarchy Interval-Valued Dataset.

Usage

```
data(hierarchy.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 20 rows and 6 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(hierarchy.int)
```

horses.int *Horses Interval-Valued Dataset*

Description

Horses Interval-Valued Dataset.

Usage

```
data(horses.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 8 rows and 7 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(horses.int)
```

iGAP_to_MM *iGAP to MM*

Description

To convert iGAP files to CSV files.

Usage

```
iGAP_to_MM(data, location)
```

Arguments

`data`	The iGAP file.
`location`	The location of the symbolic variable in the data.

Value

A CSV data file.

Examples

```
data(Abalone.iGAP)
Abalone <- iGAP_to_MM(Abalone.iGAP, c(1, 2, 3, 4, 5, 6, 7))
```

`lackinfo.int` *Lack of information questionnaire interval dataset.*

Description

Lack of information questionnaire interval dataset generated from lackinfo dataset. A dataset containing some biographical data and the responses to 5 items measuring the perception of lack of information in a questionnaire.

Usage

```
data(lackinfo.int)
```

Format

A data frame with 50 observations of the following 8 variables:

- `id`: identification number.
- `sex`: sex of the respondent (`male` or `female`).
- `age`: respondent's age (in years).
- `item1`: respondent's interval-valued answer to item 1.
- `item2`: respondent's interval-valued answer to item 2.
- `item3`: respondent's interval-valued answer to item 3.
- `item4`: respondent's interval-valued answer to item 4.
- `item5`: respondent's interval-valued answer to item 5.

Details

An educational innovation project was carried out for improving teaching-learning processes at the University of Oviedo (Spain) for the 2020/2021 academic year. A total of 50 students have been requested to answer an online questionnaire about some biographical data (sex and age) and their perception of lack of information by selecting the interval that best represents their level of agreement to the statements proposed in a interval-valued scale bounded between 1 and 7, where 1 represents the option 'strongly disagree' and 7 represents the option 'strongly agree'.

These are the 5 items used to measure the perception of lack of information:

- I1: I receive too little information from my classmates.
- I2: It is difficult to receive relevant information from my classmates.
- I3: It is difficult to receive relevant information from the teacher.
- I4: The amount of information I receive from my classmates is very low.
- I5: The amount of information I receive from the teacher is very low.

Source

https://CRAN.R-project.org/package=IntervalQuestionStat

Examples

```
data(lackinfo.int)
```

`LoansbyPurpose.int` *Loans by purpose: Interval Dataset*

Description

Loans by purpose interval dataset generated from LoansbyPurpose dataset. This data set consist of the lower and upper bounds of the intervals for four interval characteristics of the loans aggregated by their purpose. The original microdata is available at the Kaggle Data Science platform and consists of 887 383 loan records characterized by 75 descriptors. Among the large set of variables available, we focus on borrowers' income and account and loan information aggregated by the 14 loan purposes, wich are considered as the units of interest.

Usage

```
data(LoansbyPurpose.int)
```

Format

A data frame containing 14 observations on the following 4 variables:

- `ln-inc`: The current loan purpose of natural logarithm of the self-reported annual income provided by the borrower during registration
- `ln-revolbal`: The current loan purpose of natural logarithm of the total credit revolving balance
- `open-acc`: The current loan purpose of the number of open credit lines in the borrower's credit file
- `total-acc`: The current loan purpose, of the total number of credit lines currently in the borrower's credit file

Source

https://CRAN.R-project.org/package=MAINT.Data

Examples

```
data(LoansbyPurpose.int)
```

mushroom *Mushroom Data Set*

Description

The mushroom data set consists of a set of 23 species described by 3 interval variables. These mushroom species are members of the genus Agaricies. The specific variables and their values are extracted from the Fungi of California Species.

Usage

```
data(mushroom)
```

Format

A data frame with 23 observations and 5 variables named Species, Pileus Cap Width, Stipe Length, Stipe Thickness, and Edibility.

- `Species`: The class of mushroom.
- `Pileus Cap Width`: The pileus cap width of the mushroom.
- `Stipe Length`: The stipe length of the mushroom.
- `Stipe Thickness`: The stipe thickness of the mushroom.
- `Edibility`: The edibility of mushroom (U: unknown, Y: Yes, N: No, T: Toxic).

Source

Billard, L. and Diday, E. (2006) Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley & Sons, Ltd.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(mushroom)
```

mushroom.int *Mushroom Interval Dataset*

Description

Mushroom interval dataset generated from mushroom dataset. The mushroom data set consists of a set of 23 species described by 3 interval variables. These mushroom species are members of the genus Agaricies. The specific variables and their values are extracted from the Fungi of California Species.

Usage

```
data(mushroom.int)
```

Format

A data frame with 23 observations and 5 variables named Species, Pileus Cap Width, Stipe Length, Stipe Thickness, and Edibility.

- `Species`: The class of mushroom.
- `Pileus Cap Width`: The pileus cap width of the mushroom.
- `Stipe Length`: The stipe length of the mushroom.
- `Stipe Thickness`: The stipe thickness of the mushroom.
- `Edibility`: The edibility of mushroom (U: unknown, Y: Yes, N: No, T: Toxic).

Source

Billard, L. and Diday, E. (2006) Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley & Sons, Ltd.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(mushroom.int)
```

nycflights.int *New York City flights Interval Dataset*

Description

New York City flights interval dataset generated from nycflights dataset. A interval-valued data set containing 142 units and four interval-valued variables (dep_delay, arr_delay, air_time and distance), created from from the flights data set in the R package nycflights13 (on-time data for all flights that departed the JFK, LGA or EWR airports in 2013), after removing all rows with missing observations, and aggregating by month and carrier.

Usage

```
data(nycflights.int)
```

Format

FlightsDF A data frame containing the original 327346 valid (i.e. with non missing values) flights from the nycflights13 package, described by the 4 variables: dep_delay, arr_delay, air_time and distance.

FlightsUnits A factor with 327346 observations and 142 levels, indicating the month by carrier combination to which each orginal flight belongs to.

FlightsIdt An IData object with 142 observations and 4 interval-valued variables, describing the intervals formed by agregating the FlightsDF microdata by the 0.05 and 0.95 quantiles of the subsamples formed by FlightsUnits factor.

Source

https://CRAN.R-project.org/package=MAINT.Data

References

Duarte Silva, A. P., Brito, P., Filzmoser, P., & Dias, J. G. (2021). MAINT. Data: Modelling and Analysing Interval Data in R. R Journal, 13(2).

Examples

```
data(nycflights.int)
```

occupations *Occupation Salaries Dataset*

Description

Occupation Salaries Dataset.

Usage

```
data(occupations)
```

Format

An object of class `data.frame` with 9 rows and 11 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(occupations)
```

occupations2 *Occupation Salaries Modal-Valued Dataset*

Description

Occupation Salaries Modal-Valued Dataset.

Usage

```
data(occupations)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 9 rows and 4 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(occupations2)
```

`ohtemp.int` *30 year trimmed mean daily temperatures interval dataset for the Ohio river basin.*

Description

30 year trimmed mean daily temperatures interval dataset for the Ohio river basin generated from ohtemp dataset. Intervals are defined by the mean daily maximum and minimum temperatures for the Ohio river basin from January 1, 1988 - December 31, 2018. The 116 observations in this dataset all had at least 300 daily observations of temperature in at least 30 of the 31 considered years. The mean was calculated after trimming 10 influence of potential outliers.

Usage

```
data(ohtemp.int)
```

Format

A data frame with 161 rows and 7 variables:

- `ID`: The global historical climatological network (GHCN) station identifier
- `NAME`: The GHCN station name
- `STATE`: The two-digit designation for the state in which each station resides
- `LATITUDE`: Latitude coordinate position
- `LONGITUDE`: Longitude coordinate position
- `ELEVATION`: Elevation of the measurement location (meters)
- `TEMPERATURE`: The 30 year mean daily temperature (tenths of degrees Celsius)

Source

https://CRAN.R-project.org/package=intkrige

Examples

```
data(ohtemp.int)
```

profession.int *Profession Work Salary Time Interval-Valued Dataset*

Description

Profession Work Salary Time Interval-Valued Dataset.

Usage

```
data(profession.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 15 rows and 4 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(profession.int)
```

RSDA_format *RSDA Format*

Description

This function changes the format of the data to conform to RSDA format.

Usage

```
RSDA_format(data, sym_type1 = NULL, location = NULL, sym_type2 = NULL, var = NULL)
```

Arguments

`data`	A conventional data.
`sym_type1`	The labels I means an interval variable and $S means set variable.
`location`	The location of the sym_type in the data.
`sym_type2`	The labels I means an interval variable and $S means set variable.
`var`	The name of the symbolic variable in the data.

Value

Return a dataframe with a label added to the previous column of symbolic variable.

Examples

```
data("mushroom")
mushroom.set <- set_variable_format(data = mushroom, location = 8, var = "Species")
mushroom.tmp <- RSDA_format(data = mushroom.set, sym_type1 = c("I", "S"),
                            location = c(25, 31), sym_type2 = c("S", "I", "I"),
                           var = c("Species", "Stipe.Length_min", "Stipe.Thickness_min"))
```

`set_variable_format` *Set Variable Format*

Description

This function changes the format of the set variables in the data to conform to the RSDA format.

Usage

```
set_variable_format(data, location, var)
```

Arguments

`data`	A conventional data.
`location`	The location of the set variable in the data.
`var`	The name of the set variable in the data.

Value

Return a dataframe in which a set variable is converted to one-hot encoding.

Examples

```
data("mushroom")
mushroom.set <- set_variable_format(data = mushroom, location = 8, var = "Species")
```

soccer.bivar.int *Soccer bivar Interval Data Set*

Description

Soccer bivar interval dataset generated from soccer.bivar dataset. A real interval-valued data set.

Usage

```
soccer.bivar.int
```

Format

A data frame with 20 rows and 3 variables:

- y: The response variable Y (weight)
- t1: The explanatory variable T1 (height)
- t2: The explanatory variable T2 (age)

Details

This data set concerns the record of the Weight (Y), Height (T1) and Age (T2) from 20 soccer teams of the premiere French championship.

Source

https://CRAN.R-project.org/package=iRegression

References

Lima Neto, E. A., Cordeiro, G. and De Carvalho, F.A.T. (2011). Bivariate symbolic regression models for interval-valued variables. Journal of Statistical Computation and Simulation (Print), 81, 1727–1744.

Examples

```
data(soccer.bivar.int)
```

veterinary.int *Veterinary Interval-Valued Dataset*

Description

Veterinary Interval-Valued Dataset.

Usage

```
data(veterinary.int)
```

Format

An object of class `symbolic_tbl` (inherits from `tbl_df`, `tbl`, `data.frame`) with 10 rows and 3 columns.

References

Billard L. and Diday E. (2006).Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Examples

```
data(veterinary.int)
```

write_csv_table *Write Symbolic Data Table*

Description

This function write (save) a symbolic data table from a CSV data file.

Usage

```
write_csv_table(data, file, output)
```

Arguments

`data`	The conventional data.
`file`	The name of the CSV file.
`output`	This is an experimental argument, with default TRUE, and can be ignored by most users.

Value

Write in CSV file the symbolic data table.

Examples

```
data(mushroom)
mushroom.set <- set_variable_format(data = mushroom, location = 8, var = "Species")
mushroom.tmp <- RSDA_format(data = mushroom.set, sym_type1 = c("I", "S"),
                            location = c(25, 31), sym_type2 = c("S", "I", "I"),
                           var = c("Species", "Stipe.Length_min", "Stipe.Thickness_min"))
mushroom.clean <- clean_colnames(data = mushroom.tmp)
# We can save the file in CSV to RSDA format as follows:
write_csv_table(data = mushroom.clean, file = "mushroom_interval.csv", output = FALSE)
```

Index