

Package ‘mangoTraining’

October 13, 2022

Title Mango Solutions Training Datasets

Version 1.1.1

Contact

Description Datasets to be used primarily in conjunction with Mango Solutions training materials but also for the book 'SAMS Teach Yourself R in 24 Hours' (ISBN: 978-0-672-33848-9).

Version 1.0-

7 is largely for use with the book; however, version 1.1 has a much greater focus on use with training materials, whilst retaining compatibility with the book.

URL <https://www.mango-solutions.com/>

Depends R (>= 3.5.0)

Imports tibble

Suggests testthat

License GPL-2

LazyLoad yes

LazyData yes

RoxygenNote 7.1.1

BugReports <https://github.com/MangoTheCat/mangoTraining/issues>

NeedsCompilation no

Author Mango Solutions [aut],
Andrew Little [aut, cre, ctb, dtc, rev],
Karina Marks [aut, ctb],
Aimee Gott [aut],
Owen Jones [ctb]

Maintainer Andrew Little <alittle@mango-solutions.com>

Repository CRAN

Date/Publication 2021-04-28 08:00:02 UTC

R topics documented:

mangoTraining-package	2
auto_mpg	3
bbc_articles	4
bbc_articles_full	4
bbc_business_123	5
bbc_politics_123	5
body_image	6
book_sections	7
boston	7
breast_cancer	8
breast_cancer_clean_features	9
breast_cancer_clean_target	11
carriers	12
commute	12
demo_data	13
dow_jones_data	14
drugs	15
emax_data	16
emax_fun	16
logistic_fun	17
messy_data	18
missing_pk	19
pk_data	19
policy_data	20
qtpk2	21
run_data	22
students	23
tube_data	24
xp_data	25
x_iris	26
y_iris	27
Index	28

mangoTraining-package *Mango Solutions Training Datasets*

Description

Datasets designed to be used in conjunction with Mango Solutions training materials.

Details

Datasets designed to be used in conjunction with Mango Solutions' training materials and book, SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9). The data covers a range of applications and has been collected together from a number of sources. The airquality dataset, from the Core R datasets package is also provided in xlsx format in the extdata directory of this package.

Author(s)

Mango Solutions

Contact: Mango Solutions <rin24hours@mango-solutions.com>

auto_mpg

Auto MPG Data Set

Description

Data concerns city-cycle fuel consumption - revised from CMU StatLib library.

Usage

auto_mpg

Format

A matrix containing 398 observations and 10 attributes.

mpg Miles per gallon of the engine. Predictor attribute

cylinders Number of cylinders in the engine

displacement Engine displacement

horsepower Horsepower of the car

weight Weight of the car (lbs)

acceleration Acceleration of the car (seconds taken for 0-60mph)

model_year Model year of the car in the 1900s

origin Car origin

make Car manufacturer

car_name Name of the car

Source

<http://archive.ics.uci.edu/ml/datasets/Auto+MPG>

References

Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

bbc_articles	<i>BBC articles data</i>
--------------	--------------------------

Description

A collection of BBC news articles from the business or politics sections. There are a total of 927 articles used.

Usage

bbc_articles

Format

A tibble with 201,571 observations, each a word on a document.

word A word in an article

document The document/article ID where the word was taken from

Source

- <https://www.bbc.co.uk/news>

bbc_articles_full	<i>Full BBC Articles data</i>
-------------------	-------------------------------

Description

Full BBC Articles data

Usage

bbc_articles_full

Format

A tibble, with 927 observations of separate documents and their contents. This results in two columns.

words The words from a given article

document The 'document' (article) ID

Details

A collection of business and politics BBC news articles. Each row represents each article (document), with a document ID and a string of the text content with stop words removed. This is a 'dirty' version of the `bbc_articles` dataset, where we now have a string of text for each observation, as opposed to a single word.

Source

- <https://www.bbc.co.uk/news>

bbc_business_123	<i>BBC Business article data</i>
------------------	----------------------------------

Description

A single BBC Business article (not included in the full BBC articles dataset), given in tidy, one word per row format.

Usage

bbc_business_123

Format

A tibble with 107 observations, each a word on a document.

word A word in an article

document The document/article ID where the word was taken from. Note: this only has one unique value, however the column is kept for comparison with other BBC datasets.

Source

- <https://www.bbc.co.uk/news>

bbc_politics_123	<i>BBC Politics article data</i>
------------------	----------------------------------

Description

A single BBC Politics article (not included in the full BBC articles dataset), given in tidy, one word per row format.

Usage

bbc_politics_123

Format

A tibble with 86 observations, each a word on a document.

word A word in an article

document The document/article ID where the word was taken from. Note: this only has one unique value, however the column is kept for comparison with other BBC datasets.

Source

- <https://www.bbc.co.uk/news>

body_image

Body image dataset

Description

Body image dataset

Usage

body_image

Format

A tibble of 246 observations on 8 attributes.

ethnicity Subject's ethnicity (Asian, Europn, Maori, Pacific)

married How many times have they been married?

bodyim Subject's rating of themselves (slight.uw, right, slight.ow, mod.ow, very.ow)

sm.ever Have they ever smoked?

weight Weight in kilograms

height Height in centimetres

age Age in years

stressgp What stress group are they in?

Details

A simulated dataset containing data on the self-image of subjects with differing body aesthetics

Source

Simulated data

book_sections	<i>Gutenberg Project books dataset</i>
---------------	--

Description

A mixed up collection of words from different book sections of two books.

Usage

book_sections

Format

A tibble with 108,657 observations, each a word on a document. This data set is designed to show how LDA can be used to separate a set of mixed documents into two distinct "topics" (or books).

word Words from a given section within a book.

document The book section ID that the word came from.

Source

Data taken from two books of the Gutenberg Project

- <https://www.gutenberg.org/>

boston	<i>Boston housing dataset</i>
--------	-------------------------------

Description

Dataset containing housing values in the suburbs of Boston.

Usage

boston

Format

This data frame contains the following columns:

tract Census tract

medv Median value of owner-occupied homes in \$1,000s.

crim Per capita crime rate by town.

zn Proportion of residential land zoned for lots over 25,000 sq.ft.

indus Proportion of non-retail business acres per town.

chas Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
 nox Nitrogen oxides concentration (parts per 10 million).
 rm Average number of rooms per dwelling.
 age Proportion of owner-occupied units built prior to 1940.
 dis Weighted mean of distances to five Boston employment centres.
 rad Index of accessibility to radial highways.
 tax Full-value property-tax rate per \$10,000.
 ptratio Pupil-teacher ratio by town.
 b $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.
 lstat Lower status of the population (percent).

Details

The boston data frame has 506 rows and 15 columns.

Source

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics and Management* **5**, 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

breast_cancer

Wisconsin Diagnostic Breast Cancer (WDBC)

Description

The data contain measurements on cells in suspicious lumps in a women's breast. Features are computed from a digitised image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. All samples are classified as either *benign* or *malignant*.

Usage

breast_cancer

Format

breast_cancer is a tibble with 22 columns. The first column is an ID column. The second indicates whether the sample is classified as benign or malignant. The remaining columns contain measurements for 20 features. Ten real-valued features are computed for each cell nucleus. The references listed below contain detailed descriptions of how these features are computed. The mean, and "worst" (or largest - mean of the three largest values) of these features were computed for each image, resulting in 20 features. Below are descriptions of these features where * should be replaced by either mean or worst.

- *_radius mean of distances from center to points on the perimeter
- *_texture standard deviation of gray-scale values
- *_perimeter perimeter value
- *_area area value
- *_smoothness local variation in radius lengths
- *_compactness $\text{perimeter}^2 / \text{area} - 1.0$
- *_concavity severity of concave portions of the contour
- *_concave_points number of concave portions of the contour
- *_symmetry symmetry value
- *_fractal_dimension "coastline approximation" - 1

Note

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

Source

[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

References

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

breast_cancer_clean_features

Wisconsin Breast Cancer Database

Description

Wisconsin Breast Cancer Database

Usage

breast_cancer_clean_features

Format

A list containing a training and test dataset. These come from a data frame with 699 observations on 11 variables, however the ID and class columns have been removed. There is a train to test ratio of 0.8.

Cl.thickness Clump Thickness
 Cell.size Uniformity of Cell Size
 Cell.shape Uniformity of Cell Shape
 Marg.adhesion Marginal Adhesion
 Epith.c.size Single Epithelial Cell Size
 Bare.nuclei Bare Nuclei
 Bl.cromatin Bland Chromatin
 Normal.nucleoli Normal Nucleoli
 Mitoses Mitoses

Source

- Creator: Dr. William H. Wolberg (physician); University of Wisconsin Hospital ;Madison; Wisconsin; USA
- Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu)
- Received: David W. Aha (aha@cs.jhu.edu)

These data have been taken from the UCI Repository Of Machine Learning Databases at

- <http://archive.ics.uci.edu/ml/datasets.php>
- <http://www.ics.uci.edu/~mlearn/MLRepository.html>

and were converted to R format by Evgenia Dimitriadou.

References

1. Wolberg,W.H., \& Mangasarian,O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193-9196.
 - Size of data set: only 369 instances (at that point in time)
 - Collected classification results: 1 trial only
 - Two pairs of parallel hyperplanes were found to be consistent with 50% of the data
 - Accuracy on remaining 50% of dataset: 93.5%
 - Three pairs of parallel hyperplanes were found to be consistent with 67% of data
 - Accuracy on remaining 33% of dataset: 95.9%
2. Zhang,J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470-479). Aberdeen, Scotland: Morgan Kaufmann.
 - Size of data set: only 369 instances (at that point in time)
 - Applied 4 instance-based learning algorithms
 - Collected classification results averaged over 10 trials

- Best accuracy result:
- 1-nearest neighbor: 93.7%
- trained on 200 instances, tested on the other 169
- Also of interest:
- Using only typical instances: 92.2% (storing only 23.1 instances)
- trained on 200 instances, tested on the other 169

Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.

breast_cancer_clean_target

Wisconsin Breast Cancer Database

Description

Wisconsin Breast Cancer Database

Usage

breast_cancer_clean_target

Format

A list containing a training and test dataset. These come from a data frame with 699 observations on 11 variables, however only the target classes have been kept. There is a train to test ratio of 0.8.

Class.Benign Whether the sample was classified as benign

Class.malignant Whether the sample was classified as malignant

2. Zhang,J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470-479). Aberdeen, Scotland: Morgan Kaufmann.

- Size of data set: only 369 instances (at that point in time)
- Applied 4 instance-based learning algorithms
- Collected classification results averaged over 10 trials
- Best accuracy result:
- 1-nearest neighbor: 93.7%
- trained on 200 instances, tested on the other 169
- Also of interest:
- Using only typical instances: 92.2% (storing only 23.1 instances)
- trained on 200 instances, tested on the other 169

Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.

Source

- Creator: Dr. William H. Wolberg (physician); University of Wisconsin Hospital ;Madison; Wisconsin; USA
- Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu)
- Received: David W. Aha (aha@cs.jhu.edu)

These data have been taken from the UCI Repository Of Machine Learning Databases at

- <http://archive.ics.uci.edu/ml/datasets.php>
- <http://www.ics.uci.edu/~mllearn/MLRepository.html>

and were converted to R format by Evgenia Dimitriadou.

carriers

Carrier data

Description

This data comes from the RITA/Transtats database

Usage

carriers

Format

A dataframe with 1492 observations and 2 variables

Code A character string giving the IATA code for the carrier

Description Carrier name/description

commute

R For Data Science tidyuesday commute dataset

Description

Data from the ACS Survey detailing the use of different transport modes

Usage

commute

Format

A tibble containing 3,496 observations of 9 variables

city City

state State

city_size City Size -

- Small = 20K to 99,999
- Medium = 100K to 199,999
- Large = \geq 200K

mode Mode of transport, either walk or bike

n Number of individuals

percent Percent of total individuals

moe Margin of Error (percent)

state_abb Abbreviated state name

state_region ACS State region

Source

American Community Survey, United States Census Bureau

- R For Data Science repository: <https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-11-05>
- Article and underlying data can be found at: <https://www.census.gov/library/publications/2014/acs/acs-25.html?#>

demo_data

Demographics data

Description

A simulated dataset containing demographic data about a number of subjects.

Usage

demo_data

demoData

Format

A data frame with 33 observations on the following 7 demographic variables. This data is designed so that it can be merged with the dataset `pk_data`.

`Subject` A numeric vector giving the subject identifier

`Sex` A factor with levels F M

`Age` A numeric vector giving the age of the subject

`Weight` A numeric vector giving weight in kg

`Height` A numeric vector giving height in cm

`BMI` A numeric vector giving the subject body mass index

`Smokes` A factor with levels No Yes

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated data

dow_jones_data	<i>Dow Jones Index Data</i>
----------------	-----------------------------

Description

Dataset containing the Dow Jones Index between 2014-01-01 and 2015-01-01, which is a stock market index that measures the stock performance of 30 large companies listed on stock exchanges in the United States.

Usage

`dow_jones_data`

`dowJonesData`

Format

A data frame with 252 observations on the following 7 variables containing data from 2014-01-01 to 2015-01-01.

`Date` Date of observation in character string format "%m/%d/%Y"

`DJI.Open` Opening value of DJI on the specified date

`DJI.High` High value of the DJI on the specified date

DJI.Low Low value of the DJI on the specified date
DJI.Close Closing value of the DJI on the specified date
DJI.Volume the number of shares or contracts traded
DJI.Adj.Close Close price adjusted for dividends and splits

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Data obtained using yahooSeries from the fImport package.

drugs	<i>Repeated Measures Drug data</i>
-------	------------------------------------

Description

Repeated Measures Drug data

Usage

drugs

Format

A data frame with 20 observations on the following 3 variables.

Subj A numeric vector, giving the subject ID

Drug A numeric vector giving the drug ID, numbered 1 to 4

Value A numeric vector, giving the observation value

Source

Generated from example data used in <https://www.stat.tutoria.ls.com/SAS/TUTORIAL-PROC-GLM-REPEAT.htm>

emax_data	<i>Data that can be used to fit or plot Emax models</i>
-----------	---

Description

Data that can be used to fit or plot Emax models

Usage

emax_data

emaxData

Format

A data frame with 64 observations on the following 6 variables.

Subject a numeric vector giving the unique subject ID

Dose a numeric vector giving the dose group

E a numeric vector giving the Emax

Gender a numeric vector giving the gender

Age a numeric vector giving the age of the subject

Weight a numeric vector giving the weight

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated data

emax_fun	<i>Function to calculate Emax</i>
----------	-----------------------------------

Description

Calculation used for Emax in Mango Training materials. Note: This function has been renamed using tidyverse-style snake_case naming conventions. However the original name of the function has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Usage

```
emax_fun(Dose, E0 = 0, ED50 = 50, Emax = 100)
```

Arguments

Dose	User provided dose values
E0	Effect at time 0
ED50	50% of maximum effect
Emax	Maximum effect

Value

Numeric value/vector representing the response value.

Examples

```
emax_fun(Dose = 100)
```

logistic_fun	<i>Function to fit logistic model</i>
--------------	---------------------------------------

Description

Simple logistic function as used in Mango training materials. Note: This function has been renamed using tidyverse-style snake_case naming conventions. However the original name of the function has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Usage

```
logistic_fun(Dose, E0 = 0, EC50 = 50, Emax = 1, rc = 5)
```

Arguments

Dose	The dose value to calculate at
E0	Effect at time 0
EC50	50% of maximum effect
Emax	Maximum effect
rc	rate constant

Value

Numeric value/vector representing the response value.

Examples

```
logistic_fun(Dose = 50)
```

messy_data

Messy clinical trial data

Description

Simulated dataset for examples of reshaping data

Usage

```
messy_data
```

```
messyData
```

Format

A data frame with 33 observations on the following 7 variables. This data has been designed to show reshaping/tidying of data.

Subject A numeric vector giving the subject ID

Placebo.1 A numeric vector giving the subjects observed value on treatment Placebo at time 1

Placebo.2 A numeric vector giving the subjects observed value on treatment Placebo at time 2

Drug1.1 A numeric vector giving the subjects observed value on treatment Drug1 at time 1

Drug1.2 A numeric vector giving the subjects observed value on treatment Drug1 at time 2

Drug2.1 A numeric vector giving the subjects observed value on treatment Drug2 at time 1

Drug2.2 A numeric vector giving the subjects observed value on treatment Drug2 at time 2

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated data

missing_pk	<i>Clinical trial data</i>
------------	----------------------------

Description

Clinical trial data

Usage

missing_pk

missingPk

Format

A data frame with 165 observations on the following 4 variables.

Subject a numeric vector giving the subject identifier

Dose a numeric vector giving the dose group

Time a numeric vector giving the observation times

Conc a numeric vector giving the observed concentration

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated from 'pk_data'

pk_data	<i>Typical PK data</i>
---------	------------------------

Description

Typical PK data

Usage

pk_data

pkData

Format

A data frame with 165 observations on the following 4 variables.

Subject a numeric vector giving the subject identifier

Dose a numeric vector giving the dose group

Time a numeric vector giving the observation times

Conc a numeric vector giving the observed concentration

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated data

policy_data

Insurance Policy Data

Description

Insurance Policy Data

Usage

policy_data

policyData

Format

A data frame with 926 observations on the following 13 variables.

Year The four digit year of the policy

PolicyNo The policy number

TotalPremium The total insurance premium

BonusMalus Discount level

WeightClass The weight class of the car

Region Region of the car owner

Age Age of the main driver

Mileage Estimated annual mileage

Usage Car usage

PremiumClass Class of the car
 NoClaims Number of previous claims
 GrossIncurred Claim cost
 Exposure How long they have been driving

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated based on details of how to simulate car insurance data in Modern Actuarial Risk Theory Using R 2nd Edition (Rob Kaas, Marc Goovaerts, Jan Dhaene, Michel Denuit)

qtpk2	<i>Typical PK data</i>
-------	------------------------

Description

Typical PK data

Usage

qtpk2

Format

A data frame with 2061 observations on the following 8 variables.

subjid A numeric vector giving the subject ID
 treat A factor giving the treatment
 time A numeric vector giving the observation times
 qt A numeric vector giving the QT interval value
 qtcB A numeric vector giving corrected QT interval
 hr A numeric vector giving the heart rate
 rr A numeric vector giving the R-R interval
 sex A factor giving the subject sex

Source

A subset of the data qtpk originally provided in the QT package

`run_data`*An example of NONMEM run data*

Description

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Usage`run_data``runData`**Format**

A data frame with 73 observations on the following 10 variables.

ID a numeric vector giving the subject ID

DAY a numeric vector giving the day of the observation

CL a numeric vector giving the clearance value

V a numeric vector giving the volume of distribution

WT a numeric vector giving the weight

DV a numeric vector giving the dependent variable

IPRE a numeric vector giving the individual prediction

PRED a numeric vector giving the population prediction

RES a numeric vector giving the residual

WRES a numeric vector giving the weighted residual

Source

Simulated Data

students

Students simulated data

Description

Students simulated data

Usage

students

Format

A tibble with 146 observations of 15 variables.

Grade Final grade (A, B, C, D)

Pass Did they pass the course? (No, Yes)

Exam Mark in final exam (out of 100)

Degree The degree type undertaken by the student

Gender Gender of the student

Attend Did they regularly attend class? (No, Yes)

Assign Score obtained in mid-term assignment (out of 20)

Test Score obtained in previous term test (out of 20)

B Mark for short answer section (out of 20)

C Mark for long answer section (out of 20)

MC Mark for multiple choice sectionC (out of 30)

Colour Colour of exam booklet (Blue, Green, Pink, Yellow)

Stage1 Stage one grade (A, B, C)

Years.Since Number of years since doing Stage 1

Repeat Where they repeating the paper? (No, Yes)

Source

Simulated data

tube_data	<i>London Tube Performance data</i>
-----------	-------------------------------------

Description

London Tube Performance data

Usage

tube_data

tubeData

Format

A data frame with 1050 observations on the following 9 variables.

Line A factor with 10 levels, one for each London tube line

Month A numeric vector indicating the month of the observation

Scheduled A numeric vector giving the scheduled running time

Excess A numeric vector giving the excess running time

TOTAL A numeric vector giving the total running time

Opened A numeric vector giving the year the line opened

Length A numeric vector giving the line length

Type A factor indicating the type of tube line

Stations A numeric vector giving the number of stations on the line

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

This data was taken from "<https://data.london.gov.uk/dataset/tube-network-performance-data-transport-committee-report>"

xp_data

Typical NONMEM data

Description

Typical NONMEM data

Usage

xp_data

xpData

Format

A data frame with 1061 observations on the following 23 variables.

ID a numeric vector giving the subject ID

SEX a numeric vector giving the subject sex

RACE a numeric vector giving the subject race

SMOK a numeric vector giving the subject smoking status

HCTZ a numeric vector giving the treatment status

PROP a numeric vector giving the treatment status

CON a numeric vector giving the treatment status

DV a numeric vector giving the dependent variable

PRED a numeric vector giving population prediction

RES a numeric vector giving the residual

WRES a numeric vector giving the weighted residual

AGE a numeric vector giving the subject age

HT a numeric vector giving the subject height

WT a numeric vector giving the subject weight

SECR a numeric vector giving the serum creatinine value

OCC a numeric vector giving the occasion

TIME a numeric vector giving the time of the observation time

IPRE a numeric vector giving individual prediction

IWRE a numeric vector giving the individual weighted residual

SID a numeric vector giving the site ID

CL a numeric vector giving the clearance

V a numeric vector giving the volume of distribution

KA a numeric vector giving the absorption rate constant

Details

This dataset has been renamed using tidyverse-style snake_case naming conventions. However the original name of the dataset has been kept to ensure backwards compatibility with the book SAMS Teach Yourself R in 24 Hours (ISBN: 978-0-672-33848-9).

Source

Simulated Data

x_iris	<i>Iris predictors data for Species classification</i>
--------	--

Description

This data was taken from Edgar Anderson's famous iris data set. This gives the measurements (in centimeters) of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*. However, the species is seen as the target variable, and as such has been removed from this dataset, whilst being added to the counterpart `y_iris` dataset. Furthermore, the 4 remaining 'predictor' variables have been separated into a training and test set with a ratio of 4:1, followed by centering and scaling.

Usage

x_iris

Format

A list of two named matrices, 'train' and 'test', representing the training and test sets for the predictors. These have 4 columns each, with 120 and 30 rows respectively.

Sepal.Length Sepal length

Sepal.Width Sepal width

Petal.Length Petal length

Petal.Width Petal width

Source

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179-188. The data were collected by Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, 59, 2-5

- <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html>

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

y_iris

Iris class data for Species classification

Description

This data was taken from Edgar Anderson's famous iris data set. This gives the measurements (in centimeters) of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*. This is the target dataset (as a counterpart to the `x_iris` dataset) and thus only retains the Species information. As with the `x_iris` dataset, the data has been split into a training and test set with a ratio of 4:1. Following this the species class has been one-hot encoded to give three columns, one for each species level.

Usage

```
y_iris
```

Format

A list of two named matrices, 'train' and 'test', representing the training and test sets for the predictors. These have 3 indicator columns each, with 120 and 30 rows respectively.

`Species.setosa` Indicator column for the species class *setosa*

`Species.versicolor` Indicator column for the species class *versicolor*

`Species.virginica` Indicator column for the species class *virginica*

Source

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179-188. The data were collected by Anderson, Edgar (1935). The irises of the Gaspé Peninsula, *Bulletin of the American Iris Society*, 59, 2-5

- <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html>

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

Index

* datasets

- auto_mpg, 3
- bbc_articles, 4
- bbc_articles_full, 4
- bbc_business_123, 5
- bbc_politics_123, 5
- body_image, 6
- book_sections, 7
- boston, 7
- breast_cancer, 8
- breast_cancer_clean_features, 9
- breast_cancer_clean_target, 11
- carriers, 12
- commute, 12
- demo_data, 13
- dow_jones_data, 14
- drugs, 15
- emax_data, 16
- messy_data, 18
- missing_pk, 19
- pk_data, 19
- policy_data, 20
- qtpk2, 21
- run_data, 22
- students, 23
- tube_data, 24
- x_iris, 26
- xp_data, 25
- y_iris, 27

auto_mpg, 3

bbc_articles, 4

bbc_articles_full, 4

bbc_business_123, 5

bbc_politics_123, 5

body_image, 6

book_sections, 7

boston, 7

breast_cancer, 8

breast_cancer_clean_features, 9

breast_cancer_clean_target, 11

carriers, 12

commute, 12

demo_data, 13

dow_jones_data, 14

drugs, 15

emax_data, 16

emax_fun, 16

emaxData (emax_data), 16

logistic_fun, 17

mangoTraining (mangoTraining-package), 2

mangoTraining-package, 2

messy_data, 18

messyData (messy_data), 18

missing_pk, 19

missingPk (missing_pk), 19

pk_data, 19

pkData (pk_data), 19

policy_data, 20

policyData (policy_data), 20

qtpk2, 21

run_data, 22

runData (run_data), 22

students, 23

tube_data, 24

tubeData (tube_data), 24

x_iris, 26

xp_data, 25

`xpData(xp_data)`, [25](#)

`y_iris`, [27](#)