Package 'MSMU'

July 22, 2025

Title Descriptive Statistics Functions for Numeric Data

Version 0.1.2

Description Provides fundamental functions for descriptive statistics, including MODE(), estimate_mode(), center_stats(), position_stats(), pct(), spread_stats(), kurt(), skew(), and shape_stats(), which assist in summarizing the center, spread, and shape of numeric data. For more details, see McCurdy (2025), ``Introduction to Data Science with R'' <https://jonmccurdy.github.io/Introduction-to-Data-Science/>.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.2

Depends R (>= 3.5)

LazyData true

Suggests roxygen2

NeedsCompilation no

Author Luke Papayoanou [aut], Jon McCurdy [aut, cre]

Maintainer Jon McCurdy <j.r.mccurdy@msmary.edu>

Repository CRAN

Date/Publication 2025-07-22 11:01:57 UTC

Contents

`baseball_teams` *Professional baseball teams data*

Description

This dataset contains historical performance and statistics for professional baseball teams across multiple seasons from 2000-2020.

Usage

```
baseball_teams
```

Format

A data frame with 630 rows and 12 columns:

year Year (integer)

team_name Team (character)

games_played Number of games played (integer)

wins Number of wins (integer)

losses Number of losses (integer)

world_series World series winner that specific year (character)

runs_scored Number of total runs scored during season (integer)

hits Number of total hits during season (integer)

homeruns Number of total homeruns during season (integer)

earned_run_average Team earned run average per 9 innings (numeric)

fielding_percentage Team fielding percentage (numeric)

home_attendance Average home game attendance (integer)

Source

Data retrieved from Lahmans Baseball Database with alterations made for educational purposes

---

`basketball` *College basketball data*

Description

This dataset contains performance statistics for 363 men's college basketball teams from the 2022-23 season.

Usage

```
basketball
```

Format

A data frame with 363 rows and 18 columns:

School School (character)

State State (character)

W Wins (integer)

L Loss's (integer)

W.L. Win Loss percentage (numeric)

SRS Simple Rating System (numeric)

SOS Strength of Schedule (numeric)

Points.Scored Points scored (integer)

Points.Allowed Points allowed (integer)

FG. Team field goal percentage (numeric)

X3P. Three point percentage (numeric)

FT. Free throw percentage (numeric)

Rebounds Number of rebounds (integer)

AST Number of assists (integer)

STL Number of steals (integer)

Blocks Number of blocks (integer)

Turn.Overs Number of turn overs (integer)

Fouls Number of fouls (integer)

Source

Data retrieved from Sports Reference with alterations made for educational purposes.

`center_stats` *Summary of Central Tendency*

Description

Computes a variety of center statistics for a numeric vector, including: mean, median, trimmed means (10% and 25%), and estimated mode (via probability density function using `estimate_mode()`).

Usage

```
center_stats(x)
```

Arguments

`x`	A numeric vector.

Value

A named numeric vector with values for:

mean Arithmetic mean

median Median

trim25 25% trimmed mean

trim10 10% trimmed mean

est_mode Estimated mode from `estimate_mode()`

See Also

`estimate_mode`

Examples

```
# Center Stats of continuous random data
set.seed(123)
x <- rnorm(1000, mean=50, sd=10)
center_stats(x)

# Center Stats of Sepal Length in iris data set
data("iris")
center_stats(iris$Sepal.Length)
```

`christmas` *Christmas data*

Description

Santa's dataset, exploring if Santa gives children presents based a variety of variables!

Usage

```
christmas
```

Format

A data frame with 1000 rows and 15 columns:

Gender Gender (character)

Toy_Count Number of toys (integer)

Chores_Completed Number of Chores completed (numeric)

Favorite_Color Childs Favorite color (character)

Helping_Hand Childs helping hand number/score (integer)

Complaints_Received Number of complaints child says (numeric)

Tantrum_Count Number of Tantrums child has (integer)

Rule_Breaks Number of rule breaking child does (numeric)

Sharing_Behavior Childs willingness to share (numeric)

Hours_of_Sleep Childs average hours of sleep per night (numeric)

Screen_Time Childs average hours of screen time (numeric)

School_Grade Childs school grade (numeric)

Parent_Presence Childs parent presence (numeric)

Greed_Score Santas numeric system for labeling childrens greed (numeric)

Outcome Whether a child gets a present or coal (character)

Source

Santa

class_demographics *Class demographics*

Description

A sample dataset representing demographic and academic information for 50 college students.

Usage

```
class_demographics
```

Format

A data frame with 50 rows and 6 columns:

names Persons name (character)

ages Persons age (int)

state Persons state (character)

year Persons year in college (character)

majors Persons major (character)

sport Binary Sport, 1(yes) or 0(no) (integer)

Source

Synthetic Data

college_data *College data*

Description

This dataset provides detailed information on 777 U.S. colleges and universities from 1995, covering aspects of admissions, academics, finances, and student demographics.

Usage

```
college_data
```

Format

A data frame with 777 rows and 16 columns:

Name College name (character)

Region US region (character)

Accept Acceptance (integer)

Enroll Enrollment (integer)

Top10perc Percent of students that were top 10 in highschool class (integer)

Top25perc Percent of students that were top 25 in highschool class (integer)

F.Undergrad Full time undergrad (integer)

P.Undergrad Part time undergrad (integer)

Outstate Number of Out of state students (integer)

Room.Board Annual room and board price (integer)

PhD Percentage of Faculty with a PhD (integer)

Terminal Percentage of Faculty with a terminal degree (integer)

S.F.Ratio Student Faculty ratio (numeric)

perc.alumni Percent of alumni who donate to the college (integer)

Expend Instructional expenditure per student (integer)

Grad.Rate Graduation Rate (integer)

Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. Adapted from the College data set in the ISLR library with alterations made for educational purposes.

`county_data` *County data*

Description

Data for 3142 counties in the United States containing demographic, educational, economic, and technological statistics.

Usage

```
county_data
```

Format

A data frame with 3142 rows and 17 columns:

state State (character)

name County name (character)

fips County level FIPS code (integer)

pop County population (integer)

households Number of households (integer)

median_age Median age of people in county (numeric)

age_over_18 Percent age of people over 18 (numeric)

age_over_65 Percent age of people over 65 (numeric)

hs_grad Percent of highschool grads (numeric)

bachelors Percent of people with bachelors degrees (numeric)

white Percent of population that is white (numeric)

black Percent of population that is black (numeric)

hispanic Percent of population that is hispanic (numeric)

household_has_smartphone Percent of households who have a smartphone (numeric)

mean_household_income Average household income (integer)

median_household_income Median household income (integer)

unemployment_rate Unemployment rate (numeric)

Source

Adapted from the county_complete data set in the usdata library with alterations made for educational purposes.

`course_scores` *Course scores data*

Description

This dataset contains academic performance records for 200 students across four years of high school, with scores or letter grades in English and Math.

Usage

```
course_scores
```

Format

A data frame with 200 rows and 10 columns:

student Student ID (integer)

type Grade type (character)

Freshman_English Freshman English Score/letter grade (character)

Freshman_Math Freshman Math Score/letter grade (character)

Sophomore_English Sophomore English Score/letter grade (character)

Sophomore_Math Sophomore Math Score/letter grade (character)

Junior_English Junior English Score/letter grade (character)

Junior_Math Junior Math Score/letter grade (character)

Senior_English Senior English Score/letter grade (character)

Senior_Math Senior Math Score/letter grade (character)

Source

Synthetic Data

`data_210_census` *Synthetic Census dataset*

Description

A synthetic dataset containing demographic and socioeconomic information for 1,000 individuals.

Usage

```
data_210_census
```

Format

A data frame with 1000 rows and 5 columns:

age Persons Age (integer)

gender Persons Gender (character)

degree Persons level of education (character)

salary Persons Yearly Salary (integer)

height Persons Height in inches (integer)

Source

Synthetic Data

election_2020 *2020 election data*

Description

Dataset providing detailed results from the 2020 U.S. presidential election at the county level.

Usage

```
election_2020
```

Format

A data frame with 32177 rows and 7 columns:

state State (character)

state_ev State electoral votes (integer)

county County name (character)

candidate Candidate name (character)

party Candidate party (character)

total_votes Total number of votes (integer)

won True or false for the candidate to win the county (logical)

Source

Data retrieved from MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2020" with alterations made for educational purposes.

estimate_mode *Estimate Mode using Density function to find Mode of continuous data*

Description

Estimates the mode of a numeric vector by identifying the value corresponding to the peak of its estimated probability density function.

Usage

```
estimate_mode(x)
```

Arguments

x	A numeric vector. Missing values (NA) are removed.

Value

A single numeric value representing the estimated mode.

Examples

```
# Estimate the mode of continuous random data
set.seed(123)
x <- rnorm(1000, mean=5, sd=2)
estimate_mode(x)

# Estimate the mode of miles-per-gallon (mpg) in the mtcars dataset
data("mtcars")
estimate_mode(mtcars$mpg)
```

---

exam_data *Exam data*

Description

Synthetic dataset containing academic performance and background information for 1,000 students.

Usage

```
exam_data
```

Format

A data frame with 1000 rows and 8 columns:

gender Students gender (character)

race.ethnicity Students race/ethnicity (character)

parental.level.of.education Parents level of education (character)

lunch Students lunch plan (character)

test.preparation.course Student test prep level (character)

math.score Students math score (integer)

reading.score Students reading score (integer)

writing.score Students writing score (integer)

Source

Data retrieved from roycekimmons generated data

`football` *Football/Quarterback data*

Description

Dataset containing performance statistics for 106 football players who attempted a pass in the NFL for the 2022 season.

Usage

```
football
```

Format

A data frame with 106 rows and 17 columns:

Player Players name (character)

Tm Players team (character)

Age Players Age (integer)

Pos Players position (character)

G Number of games (integer)

GS Number of games starting (integer)

Wins Number of wins (integer)

Cmp Number of completions (integer)

Att Number of throwing attempts (integer)

Cmp. Completion percentage (numeric)

Yds Number of yards thrown (integer)

TD Number of touchdowns (integer)

Int Number of interceptions thrown (integer)

Y.A Yards per Attempt (numeric)

Y.G Yards per Game (numeric)

Rate Passer rating (numeric)

QBR Total Quarterback Rating (numeric)

Source

Data retrieved from Pro Football Reference with alterations made for educational purposes.

`heart` *Heart data*

Description

Dataset containing medical and diagnostic information for 303 patients, used to study the presence of Atherosclerotic Heart Disease (AHD).

Usage

```
heart
```

Format

A data frame with 303 rows and 14 columns:

Age Patients age (integer)

Sex Patients Sex (1 = Male, 0 = Female) (integer)

ChestPain Chest pain type (character)

RestBP Resting blood pressure (in mm Hg on admission to the hospital) (integer)

Chol Serum cholesterol in mg/dl (integer)

Fbs fasting blood sugar > 120 mg/dl (1 = true; 0 = false) (integer)

RestECG Resting electrocardiographic results (integer)

MaxHR Maximum heart rate achieved (integer)

ExAng Exercise induced angina (1 = yes; 0 = no) (integer)

Oldpeak ST depression induced by exercise relative to rest (numeric)

Slope The slope of the peak exercise ST segment (integer)

Ca Number of major vessels (0-3) colored by fluoroscopy (integer)

Thal Thal condition (character)

AHD Atherosclerosis Heart Disease condition (character)

Source

Data retrieved from UC Irvine Machine Learning Repository

housing_data *Housing data*

Description

Data on houses that were recently sold in the Duke Forest neighborhood of Durham, NC in November 2020.

Usage

```
housing_data
```

Format

A data frame with 98 rows and 6 columns:

price Home price (numeric)

bed Number of bedrooms (integer)

bath Number of bathrooms (numeric)

area Square footage (integer)

year_built Date house was built (integer)

lot lot size (numeric)

Source

Adapted from the duke_forest dataset in the openintro library with alterations made for educational purposes.

income_data *Income data*

Description

Dataset containing basic demographic and financial information for 20 individuals.

Usage

```
income_data
```

Format

A data frame with 20 rows and 5 columns:

ID ID (integer)

Ages age (integer)

Years_til_Retirement.65 Years until retirement at 65 (integer)

Salary Salary (integer)

Birth_weight Birth weight (integer)

Source

Synthetic Data

`kurt` *Compute Sample Kurtosis*

Description

Calculates the kurtosis of a numeric vector. A value near 0 suggests normal kurtosis (mesokurtic), positive values indicate heavier tails (leptokurtic), and negative values indicate lighter tails (platykurtic).

Usage

```
kurt(x)
```

Arguments

`x`	A numeric vector.

Details

The z-scores are computed as:

$$z_i = \frac{x_i - \bar{x}}{sd}$$

The kurtosis is then calculated as:

$$\text{Kurtosis} = \frac{1}{n} \sum_{i=1}^{n} z_i^4 - 3$$

Where:

- $\bar{x}$ is the mean of x,
- sd is the standard deviation of x,
- and n is the number of observations.

Value

A single numeric value representing the kurtosis

Examples

```
# Kurtosis of mpg in mtcars
data("mtcars")
kurt(mtcars$mpg)
```

`ledger_data` *Ledger data*

Description

Dataset mimicking a ledger showing the price an item was bought and sold for, the date it occurred, and the color of the product.

Usage

```
ledger_data
```

Format

A data frame with 4 rows and 104 columns:

color colors (character)

type age (integer)

Jan_08 Price on date (numeric)

Jan_15 Price on date (numeric)

Jan_16 Price on date (numeric)

Jan_31 Price on date (numeric)

Feb_02 Price on date (numeric)

Feb_03 Price on date (numeric)

Feb_04 Price on date (numeric)

Feb_14 Price on date (numeric)

Feb_20 Price on date (numeric)

Feb_22 Price on date (numeric)

Feb_25 Price on date (numeric)

Feb_27 Price on date (numeric)

Feb_28 Price on date (numeric)

Mar_01 Price on date (numeric)

Mar_05 Price on date (numeric)
Mar_09 Price on date (numeric)
Mar_12 Price on date (numeric)
Mar_16 Price on date (numeric)
Mar_20 Price on date (numeric)
Mar_21 Price on date (numeric)
Mar_22 Price on date (numeric)
Mar_24 Price on date (numeric)
Mar_27 Price on date (numeric)
Mar_28 Price on date (numeric)
Mar_31 Price on date (numeric)
Apr_06 Price on date (numeric)
Apr_08 Price on date (numeric)
Apr_10 Price on date (numeric)
Apr_18 Price on date (numeric)
Apr_19 Price on date (numeric)
Apr_24 Price on date (numeric)
Apr_26 Price on date (numeric)
Apr_29 Price on date (numeric)
May_01 Price on date (numeric)
May_04 Price on date (numeric)
May_12 Price on date (numeric)
May_17 Price on date (numeric)
May_24 Price on date (numeric)
May_25 Price on date (numeric)
May_28 Price on date (numeric)
Jun_01 Price on date (numeric)
Jun_04 Price on date (numeric)
Jun_11 Price on date (numeric)
Jun_16 Price on date (numeric)
Jun_25 Price on date (numeric)
Jun_28 Price on date (numeric)
Jul_03 Price on date (numeric)
Jul_04 Price on date (numeric)
Jul_08 Price on date (numeric)
Jul_10 Price on date (numeric)
Jul_11 Price on date (numeric)

Jul_13 Price on date (numeric)

Jul_18 Price on date (numeric)

Jul_23 Price on date (numeric)

Jul_25 Price on date (numeric)

Aug_05 Price on date (numeric)

Aug_12 Price on date (numeric)

Aug_13 Price on date (numeric)

Aug_24 Price on date (numeric)

Aug_26 Price on date (numeric)

Sep_02 Price on date (numeric)

Sep_06 Price on date (numeric)

Sep_07 Price on date (numeric)

Sep_08 Price on date (numeric)

Sep_16 Price on date (numeric)

Sep_21 Price on date (numeric)

Sep_22 Price on date (numeric)

Sep_23 Price on date (numeric)

Sep_27 Price on date (numeric)

Oct_07 Price on date (numeric)

Oct_09 Price on date (numeric)

Oct_10 Price on date (numeric)

Oct_15 Price on date (numeric)

Oct_16 Price on date (numeric)

Oct_17 Price on date (numeric)

Oct_19 Price on date (numeric)

Oct_20 Price on date (numeric)

Oct_21 Price on date (numeric)

Oct_22 Price on date (numeric)

Oct_29 Price on date (numeric)

Oct_30 Price on date (numeric)

Oct_31 Price on date (numeric)

Nov_03 Price on date (numeric)

Nov_04 Price on date (numeric)

Nov_12 Price on date (numeric)

Nov_13 Price on date (numeric)

Nov_14 Price on date (numeric)

Nov_16 Price on date (numeric)

Nov_18 Price on date (numeric)

Nov_23 Price on date (numeric)

Nov_24 Price on date (numeric)

Dec_02 Price on date (numeric)

Dec_03 Price on date (numeric)

Dec_06 Price on date (numeric)

Dec_11 Price on date (numeric)

Dec_12 Price on date (numeric)

Dec_13 Price on date (numeric)

Dec_16 Price on date (numeric)

Dec_17 Price on date (numeric)

Dec_18 Price on date (numeric)

Dec_19 Price on date (numeric)

Dec_26 Price on date (numeric)

Source

Synthetic Data

`mlb_eda` *MLB data*

Description

Batter statistics for 2018 Major League Baseball season

Usage

```
mlb_eda
```

Format

A data frame with 1270 rows and 13 columns:

name Players name (character)

team Players team (character)

position Players position (character)

games Number of games (integer)

AB Number of at bats (integer)

R Number of runs (integer)

H Number of hits (integer)

doubles Number of doubles (integer)

HR Number of Home runs (integer)

RBI Number of Runs Batted In (integer)

AVG Players batting average (numeric)

SLG Players Slugging percentage (numeric)

OPS Players On-base Plus Slugging (numeric)

Source

Data retrieved from MLB, with alterations made for educational purposes.

`MODE` *Find the Mode of a Numeric Vector*

Description

Calculates the mode (most frequent value) of a numeric vector. If there is a tie, returns all values that share the highest frequency.

Usage

```
MODE(x)
```

Arguments

`x`	A numeric vector.

Value

A numeric value (or vector) representing the mode(s) of `x`.

Examples

```
# Mode of a Numeric Vector
MODE(c(1,2,3,3,3,4,5,5,3,8))

# Mode of the number of cylinders in mtcars dataset
data("mtcars")
MODE(mtcars$cyl)
```

mount_dorms *Mount St.Mary's dorm data*

Description

Dataset summarizing the distribution of male and female students across various dormitories at Mount College, categorized by academic year.

Usage

```
mount_dorms
```

Format

A data frame with 4 rows and 11 columns:

year Students year (character)

m_Pangborn Males living in Pangborn (integer)

m_Sheridan Males living in Sheridan (integer)

m_Terrace Males living in Terrace (integer)

m_Powell Males living in Powell (integer)

m_Towers Males living in the Towers (integer)

f_Pangborn Females living in Pangborn (integer)

f_Sheridan Females living in Sheridan (integer)

f_Terrace Females living in Terrace (integer)

f_Powell Females living in Powell (integer)

f_Towers Females living in the Towers (integer)

Source

Synthetic Data

MSMU *MSMU: Fundamental Data Functions Package*

Description

The MSMU package provides core functions for descriptive statistics and exploratory data analysis. It includes functions for computing central tendency, spread, shape, and position statistics, along with utility functions for estimating modes and standardized ranges. The package contains

Functions

- MODE()
- estimate_mode()
- center_stats()
- position_stats()
- pct()
- spread_stats()
- skew()
- kurt()
- shape_stats()

Datasets

- data_210_census
- class_demographics
- mlb_eda
- housing_data
- football
- college_data
- basketball
- mount_dorms
- baseball_teams
- christmas
- heart
- county_data
- reaction_time
- election_2020
- soccer
- exam_data
- course_scores
- income_data
- ledger_data

Author(s)

Luke Papayoanou, Jon McCurdy

pct *Percent Within N Standard Deviations of the Mean*

Description

Calculates the percentage of values in a numeric vector that fall within n standard deviations of the mean.

Usage

```
pct(x, n)
```

Arguments

x	A numeric vector.
n	A positive numeric value indicating how many standard deviations from the mean to use as bounds.

Value

A single numeric value representing the percentage (0–100) of values within the specified range.

Examples

```
# Percentage of values that fall within 2 sds of the mean in random normal data
set.seed(123)
x <- rnorm(1000)
pct(x,2)

# Percentage of values that fall within 2 sds of the mean in iris Sepal Lengths
data("iris")
pct(iris$Sepal.Length, 2)
```

position_stats *Computes Position Statics, Quintiles and Quartiles*

Description

Calculates the quintiles, including quartiles(data is split in 4 equal parts) and quintiles(data is split in 5 equal parts) of a numeric vector using the 'quantile()' function. NA's are removed.

Usage

```
position_stats(x)
```

Arguments

x	A numeric vector.

Details

Percentiles are values that divide a dataset into 100 equal parts, each representing 1% of the distribution. For example, the 25th percentile is the value below which 25% of the data fall.

Quartiles are special percentiles that divide the data into four equal groups: Q1 (25th percentile), Q2 (50th percentile or median), Q3 (75th percentile).

Quintiles divide data into five equal groups, each representing 20% of the distribution: 20th percentile, 40th, 60th, 80th percentiles split the data into quintiles.

Value

A list with two elements:

quint Numeric vector of quintiles (0%, 20%, 40%, ..., 100%)

quart Numeric vector of quartiles (0%, 25%, 50%, 75%, 100%)

Examples

```
# Position stats of random data
set.seed(123)
x <- rnorm(1000)
position_stats(x)

# Position stats of MPG in mtcars data set
data("mtcars")
position_stats(mtcars$mpg)
```

---

reaction_time *Reaction Data*

Description

This dataset contains synthetic reaction time measurements for 100 individuals under different conditions.

Usage

```
reaction_time
```

Format

A data frame with 100 rows and 6 columns:

person Person id (integer)

color color (character)

left left (numeric)

right right (numeric)

age Person age (numeric)

gender Person gender (character)

Source

Synthetic Data

---

shape_stats *Computes Sample Skew and Kurtosis*

Description

Calculates the skewness of a numeric vector (via `skew()`). A positive value indicates right skew (long right tail), while a negative value indicates left skew (long left tail). A zero value represents symmetry. Calculates the kurtosis of a numeric vector (via `kurt()`). A value near 0 suggests normal kurtosis (mesokurtic), positive values indicate heavier tails (leptokurtic), and negative values indicate lighter tails (platykurtic).

Usage

```
shape_stats(x)
```

Arguments

`x`	A numeric vector.

Value

A list with two elements:

skew Skew of Data from `skew()`

kurt Kurtosis of Data from `kurt()`

Examples

```
# Shape stats of mpg in mtcars
data("mtcars")
shape_stats(mtcars$mpg)
```

skew *Compute Sample Skewness*

Description

Calculates the skewness of a numeric vector. A positive value indicates right skew (long right tail), while a negative value indicates left skew (long left tail). A zero value represents symmetry

Usage

```
skew(x)
```

Arguments

x	A numeric vector.

Value

A single numeric value representing the skewness of the distribution.

Examples

```
# Skew of Sepal Lengths in iris
data("iris")
skew(iris$Sepal.Length)
```

soccer *Historic soccer data*

Description

This dataset contains historical match results from various international soccer games between different countries for the years 1872-2024.

Usage

```
soccer
```

Format

A data frame with 13750 rows and 5 columns:

date Date of match (character)
home_team Home team name (character)
away_team Away team name (character)
home_score Home teams goal count (integer)
away_score Away teams goal count (integer)

Source

Data retrieved from Kaggle International football results dataset with alterations made for educational purposes.

`spread_stats` *Summary of Spread Statistics*

Description

Computes a variety of spread statistics for a numeric vector, including: standard deviation, iqr, the normalized minimum, maximum, and range as well as the percentage of data within 1, 2, and 3 standard deviations (via `pct()`)

Usage

```
spread_stats(x)
```

Arguments

`x`	A numeric vector

Value

sd Standard Deviation

iqr Inter Quartile Range

minz Normalized Minimum

maxz Normalized Maximum

diffz Normalized Range

pct1 Percent of data within 1 standard deviation from `pct()`

pct2 Percent of data within 2 standard deviation from `pct()`

pct3 Percent of data within 3 standard deviation from `pct()`

See Also

`pct`

Examples

```
# Spread stats of random normal data
set.seed(123)
x <- rnorm(1000)
spread_stats(x)

# Spread stats of mpg in mtcars
data("mtcars")
spread_stats(mtcars$mpg)
```

Index