Package 'dlookr'

February 10, 2019

Type Package

Title Tools for Data Diagnosis, Exploration, Transformation

Version 0.3.8

Description A collection of tools that support data diagnosis, exploration, and transformation. Data diagnostics provides information and visualization of missing values and outliers and unique and negative values to help you understand the distribution and quality of your data. Data exploration provides information and visualization of the descriptive statistics of univariate variables, normality tests and outliers, correlation of two variables, and relationship between target variable and predictor. Data transformation supports binning for categorizing continuous variables, imputates missing values and outliers, resolving skewness. And it creates automated reports that support these three tasks.

License GPL-2 | file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 3.2.0)

Imports dplyr, magrittr, tidyr, ggplot2, RcmdrMisc, corrplot, rlang, purrr, tibble, tidyselect, classInt, moments, kableExtra, prettydoc, smbinning, xtable, knitr, rmarkdown, RColorBrewer, gridExtra, tinytex, methods, DMwR, mice, rpart

Suggests ISLR, nycflights13, randomForest, dbplyr, DBI, RSQLite, testthat

Author Choonghyun Ryu [aut, cre]

Maintainer Choonghyun Ryu <choonghyun.ryu@gmail.com>

BugReports https://github.com/choonghyunryu/dlookr/issues

VignetteBuilder knitr

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-02-10 05:23:20 UTC

R topics documented:

`dlookr-package` *dlookr: Tools for Data Diagnosis, Exploration, Transformation*

Description

dlookr provides data diagnosis, data exploration and transformation of variables during data analysis.

Details

It has three main goals:

- When data is acquired, it is possible to judge whether data is erroneous or to select a variable to be corrected or removed through data diagnosis.
- Understand the distribution of data in the EDA process. We can also understand the relationship between target variables and predictor variables for the prediction model.
- Imputates including missing value and outlier, standardization and resolving skewness, and binning of continuous variables.

To learn more about dlookr, start with the vignettes: ‘browseVignettes(package = "dlookr")‘

Author(s)

Maintainer: Choonghyun Ryu <choonghyun.ryu@gmail.com>

See Also

Useful links:

- Report bugs at https://github.com/choonghyunryu/dlookr/issues

binning *Binning the Numeric Data*

Description

The binning() converts a numeric variable to a categorization variable.

Usage

```
binning(x, nbins, type = c("quantile", "equal", "pretty", "kmeans",
  "bclust"), ordered = TRUE, labels = NULL)
```

Arguments

x	numeric vector for binning.
nbins	number of classes. required. if missing, nclass.Sturges is used.
type	binning method. one of "quantile", "equal", "equal", "pretty", "kmeans", "bclust" The "quantile" style provides quantile breaks. The "equal" style divides the range of the variable into nbins parts. The "pretty" style chooses a number of breaks not necessarily equal to nbins using base::pretty function. The "kmeans" style uses stats::kmeans function to generate the breaks. The "bclust" style uses e1071::bclust function to generate the breaks using bagged clustering. "kmeans" and "bclust" type logic was implemented by classInt::classIntervals function.
ordered	whether to build an ordered factor or not.
labels	the label names to use for each of the bins.

Details

This function is useful when used with the mutate/transmute function of the dplyr package.

Value

An object of bins class. Attributes of bins class is as follows.

- type : binning type, "quantile", "equal", "pretty", "kmeans", "bclust".
- breaks : the number of intervals into which x is to be cut.
- levels : levels of binned value.
- raw : raw data, x argument value.

"bins" class attributes information

Attributes of the "bins" classs that is as follows.

- class : "bins".
- levels : factor or ordered factor levels

- type : binning method
- breaks : breaks for binning
- raw : before the binned the raw data

See vignette("transformation") for an introduction to these concepts.

See Also

binning_by, print.bins, summary.bins.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA
# Binning the carat variable. default type argument is "quantile"
bin <- binning(carseats$Income)
# Print bins class object
bin
# Summarise bins class object
summary(bin)
# Plot bins class object
plot(bin)
# Using labels argument
bin <- binning(carseats$Income, nbins = 4,
               labels = c("LQ1", "UQ1", "LQ3", "UQ3"))
bin
# Using another type argument
bin <- binning(carseats$Income, nbins = 5, type = "equal")
bin
bin <- binning(carseats$Income, nbins = 5, type = "pretty")
bin
bin <- binning(carseats$Income, nbins = 5, type = "kmeans")
bin
bin <- binning(carseats$Income, nbins = 5, type = "bclust")
bin

# -------------------------
# Using pipes & dplyr
# -------------------------
library(dplyr)

carseats %>%
 mutate(Income_bin = binning(carseats$Income)) %>%
 group_by(ShelveLoc, Income_bin) %>%
 summarise(freq = n()) %>%
 arrange(desc(freq)) %>%
 head(10)
```

binning_by *Optimal Binning for Scoring Modeling*

Description

The binning_by() finding class intervals for numerical variable using optical binning. Optimal binning categorizes a numeric characteristic into bins for ulterior usage in scoring modeling.

Usage

```
binning_by(df, y, x, p = 0.05, ordered = TRUE, labels = NULL)
```

Arguments

df	a data frame.
y	binary response variable (0,1). Integer(int) is required. Name of y must not have a dot. Name "default" is not allowed.
x	continuous characteristic. At least 5 different values. Value Inf is not allowed. Name of x must not have a dot.
p	percentage of records per bin. Default 5% (0.05). This parameter only accepts values greater that 0.00 (0%) and lower than 0.50 (50%).
ordered	whether to build an ordered factor or not.
labels	the label names to use for each of the bins.

Details

This function is useful when used with the mutate/transmute function of the dplyr package. And this function is implemented using smbinning() function of smbinning package.

Value

an object of optimal_bins class. Attributes of optimal_bins class is as follows.

- type : binning type, "optimal".
- breaks : the number of intervals into which x is to be cut.
- levels : levels of binned value.
- raw : raw data, x argument value.
- ivtable : information value table
- iv : information value
- flag : information value

"optimal_bins" class attributes information

Attributes of the "optimal_bins" classs that is as follows.

- class : "optimal_bins".
- levels : factor or ordered factor levels
- type : binning method
- breaks : breaks for binning
- raw : before the binned the raw data
- ivtable : information value table
- iv : information value
- target : binary response variable

See vignette("transformation") for an introduction to these concepts.

See Also

binning, smbinning.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# optimal binning
bin <- binning_by(carseats, "US", "Advertising")
bin
# summary optimal_bins class
summary(bin)
# visualize optimal_bins class
plot(bin, sub = "bins of Advertising variable")
```

---

correlate *Compute the correlation coefficient between two numerical data*

Description

The correlate() compute pearson's the correlation coefficient of the numerical data.

Usage

```
correlate(.data, ...)

## S3 method for class 'data.frame'
correlate(.data, ...)
```

Arguments

.data	a data.frame or a tbl_df.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, correlate() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing. See vignette("EDA") for an introduction to these concepts.

Details

This function is useful when used with the group_by() function of the dplyr package. If you want to compute by level of the categorical data you are interested in, rather than the whole observation, you can use grouped_df as the group_by() function. This function is computed stats::cor() function by use = "pairwise.complete.obs" option.

Correlation coefficient information

The information derived from the numerical data compute is as follows.

- var1 : names of numerical variable
- var2 : name of the corresponding numeric variable
- coef_corr : pearson's correlation coefficient

See Also

cor, correlate.tbl_dbi.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Correlation coefficients of all numerical variables
correlate(carseats)

# Select the variable to compute
correlate(carseats, Sales, Price)
correlate(carseats, -Sales, -Price)
correlate(carseats, "Sales", "Price")
correlate(carseats, 1)

# Using dplyr::grouped_dt
library(dplyr)

gdata <- group_by(carseats, ShelveLoc, US)
correlate(gdata, "Sales")
```

```
correlate(gdata)

# Using pipes ---------------------------------
# Correlation coefficients of all numerical variables
carseats %>%
 correlate()
# Positive values select variables
carseats %>%
 correlate(Sales, Price)
# Negative values to drop variables
carseats %>%
 correlate(-Sales, -Price)
# Positions values select variables
carseats %>%
 correlate(1)
# Positions values select variables
carseats %>%
 correlate(-1, -2, -3, -5, -6)
# ---------------------------------------------
# Correlation coefficient
# that eliminates redundant combination of variables
carseats %>%
 correlate() %>%
 filter(as.integer(var1) > as.integer(var2))

carseats %>%
 correlate(Sales, Price) %>%
 filter(as.integer(var1) > as.integer(var2))

# Using pipes & dplyr -------------------------
# Compute the correlation coefficient of Sales variable by 'ShelveLoc'
# and 'US' variables. And extract only those with absolute
# value of correlation coefficient is greater than 0.5
carseats %>%
 group_by(ShelveLoc, US) %>%
 correlate(Sales) %>%
 filter(abs(coef_corr) >= 0.5)

# extract only those with 'ShelveLoc' variable level is "Good",
# and compute the correlation coefficient of 'Sales' variable
# by 'Urban' and 'US' variables.
# And the correlation coefficient is negative and smaller than 0.5
carseats %>%
 filter(ShelveLoc == "Good") %>%
 group_by(Urban, US) %>%
 correlate(Sales) %>%
 filter(coef_corr < 0) %>%
 filter(abs(coef_corr) > 0.5)
```

`correlate.tbl_dbi` *Compute the correlation coefficient between two numerical data*

Description

The correlate() compute pearson's the correlation coefficient of the numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
correlate(.data, ..., in_database = FALSE,
  collect_size = Inf)
```

Arguments

.data	a tbl_dbi.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, correlate() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE. See vignette("EDA") for an introduction to these concepts.

Details

This function is useful when used with the group_by() function of the dplyr package. If you want to compute by level of the categorical data you are interested in, rather than the whole observation, you can use grouped_df as the group_by() function. This function is computed stats::cor() function by use = "pairwise.complete.obs" option.

Correlation coefficient information

The information derived from the numerical data compute is as follows.

- var1 : names of numerical variable
- var2 : name of the corresponding numeric variable
- coef_corr : pearson's correlation coefficient

See Also

correlate.data.frame, cor.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Correlation coefficients of all numerical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate()

# Positive values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate(Sales, Price)

# Negative values to drop variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate(-Sales, -Price, collect_size = 200)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate(1)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate(-1, -2, -3, -5, -6)

# ---------------------------------------------
# Correlation coefficient
# that eliminates redundant combination of variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate() %>%
  filter(as.integer(var1) > as.integer(var2))

con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  correlate(Sales, Price) %>%
  filter(as.integer(var1) > as.integer(var2))
```

```
# Using pipes & dplyr -------------------------
# Compute the correlation coefficient of Sales variable by 'ShelveLoc'
# and 'US' variables. And extract only those with absolute
# value of correlation coefficient is greater than 0.5
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  group_by(ShelveLoc, US) %>%
  correlate(Sales) %>%
  filter(abs(coef_corr) >= 0.5)

# extract only those with 'ShelveLoc' variable level is "Good",
# and compute the correlation coefficient of 'Sales' variable
# by 'Urban' and 'US' variables.
# And the correlation coefficient is negative and smaller than 0.5
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  filter(ShelveLoc == "Good") %>%
  group_by(Urban, US) %>%
  correlate(Sales) %>%
  filter(coef_corr < 0) %>%
  filter(abs(coef_corr) > 0.5)
```

`describe` *Compute descriptive statistic*

Description

The describe() compute descriptive statistic of numeric variable for exploratory data analysis.

Usage

```
describe(.data, ...)

## S3 method for class 'data.frame'
describe(.data, ...)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, describe() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing. See vignette("EDA") for an introduction to these concepts.

Details

This function is useful when used with the group_by function of the dplyr package. If you want to calculate the statistic by level of the categorical data you are interested in, rather than the whole statistic, you can use grouped_df as the group_by() function.

Value

An object of the same class as .data.

Descriptive statistic information

The information derived from the numerical data describe is as follows.

- n : number of observations excluding missing values
- na : number of missing values
- mean : arithmetic average
- sd : standard devation
- se_mean : standrd error mean. sd/sqrt(n)
- IQR : interquartile range (Q3-Q1)
- skewness : skewness
- kurtosis : kurtosis
- p25 : Q1. 25% percentile
- p50 : median. 50% percentile
- p75 : Q3. 75% percentile
- p01, p05, p10, p20, p30 : 1%, 5%, 20%, 30% percentiles
- p40, p60, p70, p80 : 40%, 60%, 70%, 80% percentiles
- p90, p95, p99, p100 : 90%, 95%, 99%, 100% percentiles

See Also

describe.tbl_dbi, diagnose_numeric.data.frame.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Describe descriptive statistics of numerical variables
describe(carseats)

# Select the variable to describe
describe(carseats, Sales, Price)
describe(carseats, -Sales, -Price)
describe(carseats, 5)
```

```
# Using dplyr::grouped_dt
library(dplyr)

gdata <- group_by(carseats, ShelveLoc, US)
describe(gdata, "Income")

# Using pipes ---------------------------------
# Positive values select variables
carseats %>%
 describe(Sales, CompPrice, Income)

# Negative values to drop variables
carseats %>%
 describe(-Sales, -CompPrice, -Income)

# Using pipes & dplyr -------------------------
# Find the statistic of all numerical variables by 'ShelveLoc' and 'US',
# and extract only those with 'ShelveLoc' variable level is "Good".
carseats %>%
 group_by(ShelveLoc, US) %>%
   describe() %>%
   filter(ShelveLoc == "Good")

# extract only those with 'Urban' variable level is "Yes",
# and find 'Sales' statistics by 'ShelveLoc' and 'US'
carseats %>%
 filter(Urban == "Yes") %>%
 group_by(ShelveLoc, US) %>%
 describe(Sales)
```

`describe.tbl_dbi` *Compute descriptive statistic*

Description

The describe() compute descriptive statistic of numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi for exploratory data analysis.

Usage

```
## S3 method for class 'tbl_dbi'
describe(.data, ..., in_database = FALSE,
  collect_size = Inf)
```

Arguments

`.data`	a tbl_dbi.

...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, describe() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE. See vignette("EDA") for an introduction to these concepts.

Details

This function is useful when used with the group_by function of the dplyr package. If you want to calculate the statistic by level of the categorical data you are interested in, rather than the whole statistic, you can use grouped_df as the group_by() function.

Value

An object of the same class as .data.

Descriptive statistic information

The information derived from the numerical data describe is as follows.

- n : number of observations excluding missing values
- na : number of missing values
- mean : arithmetic average
- sd : standard devation
- se_mean : standrd error mean. sd/sqrt(n)
- IQR : interquartile range (Q3-Q1)
- skewness : skewness
- kurtosis : kurtosis
- p25 : Q1. 25% percentile
- p50 : median. 50% percentile
- p75 : Q3. 75% percentile
- p01, p05, p10, p20, p30 : 1%, 5%, 20%, 30% percentiles
- p40, p60, p70, p80 : 40%, 60%, 70%, 80% percentiles
- p90, p95, p99, p100 : 90%, 95%, 99%, 100% percentiles

See Also

describe.data.frame, diagnose_numeric.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Positive values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  describe(Sales, CompPrice, Income)

# Negative values to drop variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  describe(-Sales, -CompPrice, -Income, collect_size = 200)

# Using pipes & dplyr -------------------------
# Find the statistic of all numerical variables by 'ShelveLoc' and 'US',
# and extract only those with 'ShelveLoc' variable level is "Good".
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  group_by(ShelveLoc, US) %>%
  describe() %>%
  filter(ShelveLoc == "Good")

# extract only those with 'Urban' variable level is "Yes",
# and find 'Sales' statistics by 'ShelveLoc' and 'US'
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  filter(Urban == "Yes") %>%
  group_by(ShelveLoc, US) %>%
  describe(Sales)
```

---

diagnose *Diagnose data quality of variables*

Description

The diagnose() produces information for diagnosing the quality of the variables of data.frame or tbl_df.

Usage

```
diagnose(.data, ...)

## S3 method for class 'data.frame'
diagnose(.data, ...)
```

Arguments

.data	a data.frame or a tbl_df.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.

Details

The scope of data quality diagnosis is information on missing values and unique value information. Data quality diagnosis can determine variables that require missing value processing. Also, the unique value information can determine the variable to be removed from the data analysis.

Value

An object of tbl_df.

Diagnostic information

The information derived from the data diagnosis is as follows.:

- variables : variable names
- types : data type of the variable or to select a variable to be corrected or removed through data diagnosis.
  - integer, numeric, factor, ordered, character, etc.
- missing_count : number of missing values
- missing_percent : percentage of missing values
- unique_count : number of unique values
- unique_rate : ratio of unique values. unique_count / number of observation

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose.tbl_dbi, diagnose_category.data.frame, diagnose_numeric.data.frame.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Diagnosis of all variables
diagnose(carseats)

# Select the variable to diagnose
diagnose(carseats, Sales, Income, Age)
diagnose(carseats, -Sales, -Income, -Age)
diagnose(carseats, "Sales", "Income", "Age")
diagnose(carseats, 1, 3, 8)

# Using pipes ---------------------------------
library(dplyr)

# Diagnosis of all variables
carseats %>%
  diagnose()
# Positive values select variables
carseats %>%
  diagnose(Sales, Income, Age)
# Negative values to drop variables
carseats %>%
  diagnose(-Sales, -Income, -Age)
# Positions values select variables
carseats %>%
  diagnose(1, 3, 8)
# Positions values select variables
carseats %>%
  diagnose(-8, -9, -10)

# Using pipes & dplyr -------------------------
# Diagnosis of missing variables
carseats %>%
  diagnose() %>%
  filter(missing_count > 0)
```

`diagnose.tbl_dbi` *Diagnose data quality of variables in the DBMS*

Description

The diagnose() produces information for diagnosing the quality of the column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
diagnose(.data, ..., in_database = TRUE,
  collect_size = Inf)
```

Arguments

.data	a tbl_dbi.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
in_database	a logical. Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.

Details

The scope of data quality diagnosis is information on missing values and unique value information. Data quality diagnosis can determine variables that require missing value processing. Also, the unique value information can determine the variable to be removed from the data analysis.

Value

An object of tbl_df.

Diagnostic information

The information derived from the data diagnosis is as follows.:

- variables : column names
- types : data type of the variable or to select a variable to be corrected or removed through data diagnosis.
  - integer, numeric, factor, ordered, character, etc.
- missing_count : number of missing values
- missing_percent : percentage of missing values
- unique_count : number of unique values
- unique_rate : ratio of unique values. unique_count / number of observation

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose.data.frame, diagnose_category.tbl_dbi, diagnose_numeric.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Diagnosis of all columns
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose()

# Positive values select columns
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose(Sales, Income, Age)

# Negative values to drop columns
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose(-Sales, -Income, -Age)

# Positions values select columns, and In-memory mode
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose(1, 3, 8, in_database = FALSE)

# Positions values select columns, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose(-8, -9, -10, in_database = FALSE, collect_size = 200)

# Using pipes & dplyr -------------------------
# Diagnosis of missing variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose() %>%
  filter(missing_count > 0)
```

`diagnose_category` *Diagnose data quality of categorical variables*

Description

The diagnose_category() produces information for diagnosing the quality of the variables of data.frame or tbl_df.

Usage

```
diagnose_category(.data, ...)

## S3 method for class 'data.frame'
diagnose_category(.data, ..., top = 10,
  add_character = TRUE)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose_category() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
`top`	an integer. Specifies the upper top rank to extract. Default is 10.
`add_character`	logical. Decide whether to include text variables in the diagnosis of categorical data. The default value is TRUE, which also includes character variables.

Details

The scope of the diagnosis is the occupancy status of the levels in categorical data. If a certain level of occupancy is close to 100 then the removal of this variable in the forecast model will have to be considered. Also, if the occupancy of all levels is close to 0 variable is likely to be an identifier.

Value

an object of tbl_df.

Categorical diagnostic information

The information derived from the categorical data diagnosis is as follows.

- variables : variable names
- levels: level names
- N : number of observation

- freq : number of observation at the levles
- ratio : percentage of observation at the levles
- rank : rank of occupancy ratio of levels

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_category.tbl_dbi, diagnose.data.frame, diagnose_numeric.data.frame, diagnose_outlier.data.fra

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Diagnosis of categorical variables
diagnose_category(carseats)

# Select the variable to diagnose
diagnose_category(carseats, ShelveLoc, Urban)
diagnose_category(carseats, -ShelveLoc, -Urban)
diagnose_category(carseats, "ShelveLoc", "Urban")
diagnose_category(carseats, 7)

# Using pipes ---------------------------------
library(dplyr)

# Diagnosis of all categorical variables
carseats %>%
  diagnose_category()
# Positive values select variables
carseats %>%
  diagnose_category(Urban, US)
# Negative values to drop variables
carseats %>%
  diagnose_category(-Urban, -US)
# Positions values select variables
carseats %>%
  diagnose_category(7)
# Positions values select variables
carseats %>%
  diagnose_category(-7)
# Top rank levels with top argument
carseats %>%
  diagnose_category(top = 2)

# Using pipes & dplyr -------------------------
# Extraction of level that is more than 60% of categorical data
carseats %>%
  diagnose_category()  %>%
```

```
  filter(ratio >= 60)
```

`diagnose_category.tbl_dbi` *Diagnose data quality of categorical variables in the DBMS*

Description

The diagnose_category() produces information for diagnosing the quality of the character(CHAR, VARCHAR, VARCHAR2, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
diagnose_category(.data, ..., top = 10,
  in_database = TRUE, collect_size = Inf)
```

Arguments

`.data`	a tbl_dbi.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose_category() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
`top`	an integer. Specifies the upper top rank to extract. Default is 10.
`in_database`	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory.
`collect_size`	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.

Details

The scope of the diagnosis is the occupancy status of the levels in categorical data. If a certain level of occupancy is close to 100 then the removal of this variable in the forecast model will have to be considered. Also, if the occupancy of all levels is close to 0 variable is likely to be an identifier.

Value

an object of tbl_df.

Categorical diagnostic information

The information derived from the categorical data diagnosis is as follows.

- variables : variable names
- levels: level names
- N : number of observation
- freq : number of observation at the levles
- ratio : percentage of observation at the levles
- rank : rank of occupancy ratio of levels

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_category.data.frame, diagnose.tbl_dbi, diagnose_category.tbl_dbi, diagnose_numeric.tbl_dbi, diagnose_outlier.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Diagnosis of all categorical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_category()

# Positive values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_category(Urban, US)

# Negative values to drop variables, and In-memory mode
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_category(-Urban, -US, in_database = FALSE)

# Positions values select variables, and In-memory mode and collect size is 200
con_sqlite %>%
```

```
  tbl("TB_CARSEATS") %>%
  diagnose_category(7, in_database = FALSE, collect_size = 200)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_category(-7)

# Top rank levels with top argument
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_category(top = 2)

# Using pipes & dplyr -------------------------
# Extraction of level that is more than 60% of categorical data
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_category()  %>%
  filter(ratio >= 60)
```

`diagnose_numeric` *Diagnose data quality of numerical variables*

Description

The diagnose_numeric() produces information for diagnosing the quality of the numerical data.

Usage

```
diagnose_numeric(.data, ...)

## S3 method for class 'data.frame'
diagnose_numeric(.data, ...)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose_numeric() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.

Details

The scope of the diagnosis is the calculate a statistic that can be used to understand the distribution of numerical data. min, Q1, mean, median, Q3, max can be used to estimate the distribution of data. If the number of zero or minus is large, it is necessary to suspect the error of the data. If the number of outliers is large, a strategy of eliminating or replacing outliers is needed.

Value

an object of tbl_df.

Numerical diagnostic information

The information derived from the numerical data diagnosis is as follows.

- variables : variable names
- min : minimum
- Q1 : 25 percentile
- mean : arithmetic average
- median : median. 50 percentile
- Q3 : 75 percentile
- max : maximum
- zero : count of zero values
- minus : count of minus values
- outlier : count of outliers

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_numeric.tbl_dbi, diagnose.data.frame, diagnose_category.data.frame, diagnose_outlier.data.fra

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Diagnosis of numerical variables
diagnose_numeric(carseats)

# Select the variable to diagnose
diagnose_numeric(carseats, Sales, Income)
diagnose_numeric(carseats, -Sales, -Income)
diagnose_numeric(carseats, "Sales", "Income")
diagnose_numeric(carseats, 5)

# Using pipes ---------------------------------
```

```
library(dplyr)

# Diagnosis of all numerical variables
carseats %>%
  diagnose_numeric()
# Positive values select variables
carseats %>%
  diagnose_numeric(Sales, Income)
# Negative values to drop variables
carseats %>%
  diagnose_numeric(-Sales, -Income)
# Positions values select variables
carseats %>%
  diagnose_numeric(5)
# Positions values select variables
carseats %>%
  diagnose_numeric(-1, -5)

# Using pipes & dplyr -------------------------
# Information records of zero variable more than 0
carseats %>%
  diagnose_numeric()  %>%
  filter(zero > 0)
```

`diagnose_numeric.tbl_dbi` *Diagnose data quality of numerical variables in the DBMS*

Description

The diagnose_numeric() produces information for diagnosing the quality of the numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
diagnose_numeric(.data, ..., in_database = FALSE,
  collect_size = Inf)
```

Arguments

`.data`	a tbl_dbi.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose_numeric() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.

in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.

Details

The scope of the diagnosis is the calculate a statistic that can be used to understand the distribution of numerical data. min, Q1, mean, median, Q3, max can be used to estimate the distribution of data. If the number of zero or minus is large, it is necessary to suspect the error of the data. If the number of outliers is large, a strategy of eliminating or replacing outliers is needed.

Value

an object of tbl_df.

Numerical diagnostic information

The information derived from the numerical data diagnosis is as follows.

- variables : variable names
- min : minimum
- Q1 : 25 percentile
- mean : arithmetic average
- median : median. 50 percentile
- Q3 : 75 percentile
- max : maximum
- zero : count of zero values
- minus : count of minus values
- outlier : count of outliers

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_numeric.data.frame, diagnose.tbl_dbi, diagnose_category.tbl_dbi, diagnose_outlier.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
```

```
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Diagnosis of all numerical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_numeric()

# Positive values select variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_numeric(Sales, Income, collect_size = 200)

# Negative values to drop variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_numeric(-Sales, -Income)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_numeric(5)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_numeric(-1, -5)

# Using pipes & dplyr -------------------------
# Information records of zero variable more than 0
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_numeric()  %>%
  filter(zero > 0)
```

`diagnose_outlier` *Diagnose outlier of numerical variables*

Description

The diagnose_outlier() produces outlier information for diagnosing the quality of the numerical data.

Usage

```
diagnose_outlier(.data, ...)
```

```
## S3 method for class 'data.frame'
diagnose_outlier(.data, ...)
```

Arguments

.data	a data.frame or a tbl_df.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose_outlier() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.

Details

The scope of the diagnosis is the provide a outlier information. If the number of outliers is small and the difference between the averages including outliers and the averages not including them is large, it is necessary to eliminate or replace the outliers.

Value

an object of tbl_df.

Outlier Diagnostic information

The information derived from the numerical data diagnosis is as follows.

- variables : variable names
- outliers_cnt : count of outliers
- outliers_ratio : percent of outliers
- outliers_mean : arithmetic average of outliers
- with_mean : arithmetic average of with outliers
- without_mean : arithmetic average of without outliers

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_outlier.tbl_dbi, diagnose.data.frame, diagnose_category.data.frame, diagnose_numeric.data.fra

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Diagnosis of numerical variables
```

```
diagnose_outlier(carseats)

# Select the variable to diagnose
diagnose_outlier(carseats, Sales, Income)
diagnose_outlier(carseats, -Sales, -Income)
diagnose_outlier(carseats, "Sales", "Income")
diagnose_outlier(carseats, 5)

# Using pipes ---------------------------------
library(dplyr)

# Diagnosis of all numerical variables
carseats %>%
  diagnose_outlier()
# Positive values select variables
carseats %>%
  diagnose_outlier(Sales, Income)
# Negative values to drop variables
carseats %>%
  diagnose_outlier(-Sales, -Income)
# Positions values select variables
carseats %>%
  diagnose_outlier(5)
# Positions values select variables
carseats %>%
  diagnose_outlier(-1, -5)

# Using pipes & dplyr -------------------------
# outlier_ratio is more than 1%
carseats %>%
  diagnose_outlier()  %>%
  filter(outliers_ratio > 1)
```

`diagnose_outlier.tbl_dbi` *Diagnose outlier of numerical variables in the DBMS*

Description

The diagnose_outlier() produces outlier information for diagnosing the quality of the numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
diagnose_outlier(.data, ..., in_database = FALSE,
  collect_size = Inf)
```

Arguments

.data	a tbl_dbi.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, diagnose_outlier() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.

Details

The scope of the diagnosis is the provide a outlier information. If the number of outliers is small and the difference between the averages including outliers and the averages not including them is large, it is necessary to eliminate or replace the outliers.

Value

an object of tbl_df.

Outlier Diagnostic information

The information derived from the numerical data diagnosis is as follows.

- variables : variable names
- outliers_cnt : count of outliers
- outliers_ratio : percent of outliers
- outliers_mean : arithmetic average of outliers
- with_mean : arithmetic average of with outliers
- without_mean : arithmetic average of without outliers

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_outlier.data.frame, diagnose.tbl_dbi, diagnose_category.tbl_dbi, diagnose_numeric.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Diagnosis of all numerical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_outlier()

# Positive values select variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_outlier(Sales, Income, collect_size = 200)

# Negative values to drop variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_outlier(-Sales, -Income)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_outlier(5)
# Positions values select variables

con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_outlier(-1, -5)

# Using pipes & dplyr -------------------------
# outlier_ratio is more than 1%
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_outlier()  %>%
  filter(outliers_ratio > 1)
```

---

diagnose_report *Reporting the information of data diagnosis*

Description

The diagnose_report() report the information for diagnosing the quality of the data.

Usage

```
diagnose_report(.data, output_format, output_file, output_dir, ...)

## S3 method for class 'data.frame'
diagnose_report(.data, output_format = c("pdf",
  "html"), output_file = NULL, output_dir = tempdir(),
  font_family = NULL, browse = TRUE, ...)
```

Arguments

.data	a data.frame or a tbl_df.
output_format	report output type. Choose either "pdf" and "html". "pdf" create pdf file by knitr::knit(). "html" create html file by rmarkdown::render().
output_file	name of generated file. default is NULL.
output_dir	name of directory to generate report file. default is tempdir().
...	arguments to be passed to methods.
font_family	charcter. font family name for figure in pdf.
browse	logical. choose whether to output the report results to the browser.

Details

Generate generalized data diagnostic reports automatically. You can choose to output to pdf and html files. This is useful for diagnosing a data frame with a large number of variables than data with a small number of variables. For pdf output, Korean Gothic font must be installed in Korean operating system.

Reported information

Reported from the data diagnosis is as follows.

- Diagnose Data
  - Overview of Diagnosis
    - List of all variables quality
    - Diagnosis of missing data
    - Diagnosis of unique data(Text and Category)
    - Diagnosis of unique data(Numerical)
  - Detailed data diagnosis
    - Diagnosis of categorical variables
    - Diagnosis of numerical variables
    - List of numerical diagnosis (zero)
    - List of numerical diagnosis (minus)

- Diagnose Outliers
  - Overview of Diagnosis
    - Diagnosis of numerical variable outliers
    - Detailed outliers diagnosis

See vignette("diagonosis") for an introduction to these concepts.

Examples

```
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# reporting the diagnosis information -------------------------
# create pdf file. file name is DataDiagnosis_Report.pdf
diagnose_report(carseats)
# create pdf file. file name is Diagn.pdf
diagnose_report(carseats, output_file = "Diagn.pdf")
# create pdf file. file name is ./Diagn.pdf and not browse
diagnose_report(carseats, output_dir = ".", output_file = "Diagn.pdf",
  browse = FALSE)
# create html file. file name is Diagnosis_Report.html
diagnose_report(carseats, output_format = "html")
# create html file. file name is Diagn.html
diagnose_report(carseats, output_format = "html", output_file = "Diagn.html")
```

---

diagnose_report.tbl_dbi

Reporting the information of data diagnosis for table of the DBMS

Description

The diagnose_report() report the information for diagnosing the quality of the DBMS table through tbl_dbi

Usage

```
## S3 method for class 'tbl_dbi'
diagnose_report(.data, output_format = c("pdf",
  "html"), output_file = NULL, output_dir = tempdir(),
  font_family = NULL, in_database = FALSE, collect_size = Inf, ...)
```

Arguments

.data	a tbl_dbi.
output_format	report output type. Choose either "pdf" and "html". "pdf" create pdf file by knitr::knit(). "html" create html file by rmarkdown::render().
output_file	name of generated file. default is NULL.
output_dir	name of directory to generate report file. default is tempdir().
font_family	charcter. font family name for figure in pdf.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.
...	arguments to be passed to methods.

Details

Generate generalized data diagnostic reports automatically. You can choose to output to pdf and html files. This is useful for diagnosing a data frame with a large number of variables than data with a small number of variables. For pdf output, Korean Gothic font must be installed in Korean operating system.

Reported information

Reported from the data diagnosis is as follows.

- Diagnose Data
  - Overview of Diagnosis
    - List of all variables quality
    - Diagnosis of missing data
    - Diagnosis of unique data(Text and Category)
    - Diagnosis of unique data(Numerical)
  - Detailed data diagnosis
    - Diagnosis of categorical variables
    - Diagnosis of numerical variables
    - List of numerical diagnosis (zero)
    - List of numerical diagnosis (minus)
- Diagnose Outliers
  - Overview of Diagnosis
    - Diagnosis of numerical variable outliers
    - Detailed outliers diagnosis

See vignette("diagonosis") for an introduction to these concepts.

See Also

diagnose_report.data.frame.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# reporting the diagnosis information -------------------------
# create pdf file. file name is DataDiagnosis_Report.pdf
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_report()

# create pdf file. file name is Diagn.pdf, and collect size is 350
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_report(collect_size = 350, output_file = "Diagn.pdf")

# create html file. file name is Diagnosis_Report.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_report(output_format = "html")

# create html file. file name is Diagn.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  diagnose_report(output_format = "html", output_file = "Diagn.html")
```

---

eda_report *Reporting the information of EDA*

Description

The eda_report() report the information of Exploratory data analysis for object inheriting from data.frame.

Usage

```
eda_report(.data, ...)

## S3 method for class 'data.frame'
eda_report(.data, target = NULL,
  output_format = c("pdf", "html"), output_file = NULL,
  output_dir = tempdir(), font_family = NULL, browse = TRUE, ...)
```

Arguments

.data	a data.frame or a tbl_df.
...	arguments to be passed to methods.
target	target variable.
output_format	report output type. Choose either "pdf" and "html". "pdf" create pdf file by knitr::knit(). "html" create html file by rmarkdown::render().
output_file	name of generated file. default is NULL.
output_dir	name of directory to generate report file. default is tempdir().
font_family	charcter. font family name for figure in pdf.
browse	logical. choose whether to output the report results to the browser.

Details

Generate generalized data EDA reports automatically. You can choose to output to pdf and html files. This is useful for EDA a data frame with a large number of variables than data with a small number of variables. For pdf output, Korean Gothic font must be installed in Korean operating system.

Reported information

The EDA process will report the following information:

- Introduction
  - Information of Dataset
  - Information of Variables
  - About EDA Report
- Univariate Analysis
  - Descriptive Statistics
  - Normality Test of Numerical Variables
    - Statistics and Visualization of (Sample) Data
- Relationship Between Variables
  - Correlation Coefficient
    - Correlation Coefficient by Variable Combination
    - Correlation Plot of Numerical Variables
- Target based Analysis

- Gruoped Descriptive Statistics
  - Gruoped Numerical Variables
  - Gruoped Categorical Variables
- Gruoped Relationship Between Variables
  - Grouped Correlation Coefficient
  - Grouped Correlation Plot of Numerical Variables

See vignette("EDA") for an introduction to these concepts.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

## target variable is categorical variable
# reporting the EDA information
# create pdf file. file name is EDA_Report.pdf
eda_report(carseats, US)

# create pdf file. file name is EDA.pdf
eda_report(carseats, "US", output_file = "EDA.pdf")

# create pdf file. file name is EDA.pdf and not browse
eda_report(carseats, "US", output_dir = ".", output_file = "EDA.pdf", browse = FALSE)

# create html file. file name is EDA_Report.html
eda_report(carseats, "US", output_format = "html")

# create html file. file name is EDA.html
eda_report(carseats, US, output_format = "html", output_file = "EDA.html")


## target variable is numerical variable
# reporting the EDA information
eda_report(carseats, Sales)

# create pdf file. file name is EDA2.pdf
eda_report(carseats, "Sales", output_file = "EDA2.pdf")

# create html file. file name is EDA_Report.html
eda_report(carseats, "Sales", output_format = "html")

# create html file. file name is EDA2.html
eda_report(carseats, Sales, output_format = "html", output_file = "EDA2.html")

## target variable is null
# reporting the EDA information
```

```
eda_report(carseats)


# create pdf file. file name is EDA2.pdf
eda_report(carseats, output_file = "EDA2.pdf")

# create html file. file name is EDA_Report.html
eda_report(carseats, output_format = "html")

# create html file. file name is EDA2.html
eda_report(carseats, output_format = "html", output_file = "EDA2.html")
```

`eda_report.tbl_dbi` *Reporting the information of EDA for table of the DBMS*

Description

The eda_report() report the information of Exploratory data analysis for object inheriting from the DBMS table through tbl_dbi

Usage

```
## S3 method for class 'tbl_dbi'
eda_report(.data, target = NULL,
  output_format = c("pdf", "html"), output_file = NULL,
  font_family = NULL, output_dir = tempdir(), in_database = FALSE,
  collect_size = Inf, ...)
```

Arguments

`.data`	a tbl_dbi.
`target`	target variable.
`output_format`	report output type. Choose either "pdf" and "html". "pdf" create pdf file by knitr::knit(). "html" create html file by rmarkdown::render().
`output_file`	name of generated file. default is NULL.
`font_family`	charcter. font family name for figure in pdf.
`output_dir`	name of directory to generate report file. default is tempdir().
`in_database`	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
`collect_size`	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.
`...`	arguments to be passed to methods.

Details

Generate generalized data EDA reports automatically. You can choose to output to pdf and html files. This is useful for EDA a data frame with a large number of variables than data with a small number of variables. For pdf output, Korean Gothic font must be installed in Korean operating system.

Reported information

The EDA process will report the following information:

- Introduction
  - Information of Dataset
  - Information of Variables
  - About EDA Report
- Univariate Analysis
  - Descriptive Statistics
  - Normality Test of Numerical Variables
    - Statistics and Visualization of (Sample) Data
- Relationship Between Variables
  - Correlation Coefficient
    - Correlation Coefficient by Variable Combination
    - Correlation Plot of Numerical Variables
- Target based Analysis
  - Gruoped Descriptive Statistics
    - Gruoped Numerical Variables
    - Gruoped Categorical Variables
  - Gruoped Relationship Between Variables
    - Grouped Correlation Coefficient
    - Grouped Correlation Plot of Numerical Variables

See vignette("EDA") for an introduction to these concepts.

See Also

eda_report.data.frame.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA
```

```
# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

## target variable is categorical variable
# reporting the EDA information
# create pdf file. file name is EDA_Report.pdf
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(US)

# create pdf file. file name is EDA.pdf
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report("US", output_file = "EDA.pdf")

# create html file. file name is EDA_Report.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report("US", output_format = "html")

# create html file. file name is EDA.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(US, output_format = "html", output_file = "EDA.html")

## target variable is numerical variable
# reporting the EDA information, and collect size is 350
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(Sales, collect_size = 350)

# create pdf file. file name is EDA2.pdf
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report("Sales", output_file = "EDA2.pdf")

# create html file. file name is EDA_Report.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report("Sales", output_format = "html")

# create html file. file name is EDA2.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(Sales, output_format = "html", output_file = "EDA2.html")

## target variable is null
# reporting the EDA information
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
```

```
  eda_report()

# create pdf file. file name is EDA2.pdf
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(output_file = "EDA2.pdf")

# create html file. file name is EDA_Report.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(output_format = "html")

# create html file. file name is EDA2.html
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  eda_report(output_format = "html", output_file = "EDA2.html")
```

`find_class` *Extract variable names or indices of a specific class*

Description

The find_class() extracts variable information having a certain class from an object inheriting data.frame.

Usage

```
find_class(df, type = c("numerical", "categorical", "categorical2"),
  index = TRUE)
```

Arguments

`df`	a data.frame or objects inheriting from data.frame
`type`	character. Defines a group of classes to be searched. "numerical" searches for "numeric" and "integer" classes, "categorical" searches for "factor" and "ordered" classes. "categorical2" adds "character" class to "categorical".
`index`	logical. If TRUE is return numeric vector that is variables index. and if FALSE is return character vector that is variables name. defalut is TRUE.

Value

character vector or numeric vector. The meaning of vector according to data type is as follows.

- character vector : variables name
- numeric vector : variables index

See Also

get_class.

Examples

```
## Not run:
# data.frame
find_class(iris, "numerical")
find_class(iris, "numerical", index = FALSE)
find_class(iris, "categorical")
find_class(iris, "categorical", index = FALSE)

# tbl_df
find_class(ISLR::Carseats, "numerical")
find_class(ISLR::Carseats, "numerical", index = FALSE)
find_class(ISLR::Carseats, "categorical")
find_class(ISLR::Carseats, "categorical", index = FALSE)

# type is "categorical2"
iris2 <- data.frame(iris, char = "chars",
                    stringsAsFactors = FALSE)
find_class(iris2, "categorical", index = FALSE)
find_class(iris2, "categorical2", index = FALSE)

## End(Not run)
```

---

find_na *Finding variables including missing values*

Description

Find the variable that contains the missing value in the object that inherits the data.frame or data.frame.

Usage

```
find_na(.data, index = TRUE, rate = FALSE)
```

Arguments

.data	a data.frame or a tbl_df.
index	logical. When representing the information of a variable including missing values, specify whether or not the variable is represented by an index. Returns an index if TRUE or a variable names if FALSE.
rate	logical. If TRUE, returns the percentage of missing values in the individual variable.

Value

Information on variables including missing values.

See Also

imputate_na, find_na.

Examples

```
## Not run:
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

find_na(carseats)

find_na(carseats, index = FALSE)

find_na(carseats, rate = TRUE)

## using dplyr ----------------------------------
library(dplyr)

# Perform simple data quality diagnosis of variables with missing values.
carseats %>%
  select(find_na(.)) %>%
  diagnose()

## End(Not run)
```

---

find_outliers *Finding variables including outliers*

Description

Find the numerical variable that contains outliers in the object that inherits the data.frame or data.frame.

Usage

```
find_outliers(.data, index = TRUE, rate = FALSE)
```

Arguments

.data	a data.frame or a tbl_df.
index	logical. When representing the information of a variable including outliers, specify whether or not the variable is represented by an index. Returns an index if TRUE or a variable names if FALSE.
rate	logical. If TRUE, returns the percentage of outliers in the individual variable.

Value

Information on variables including outliers.

See Also

find_na, imputate_outlier.

Examples

```
## Not run:
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

find_outliers(carseats)

find_outliers(carseats, index = FALSE)

find_outliers(carseats, rate = TRUE)

## using dplyr ----------------------------------
library(dplyr)

# Perform simple data quality diagnosis of variables with outliers.
carseats %>%
  select(find_outliers(.)) %>%
  diagnose()

## End(Not run)
```

---

find_skewness *Finding skewed variables*

Description

Find the numerical variable that skewed variable that inherits the data.frame or data.frame.

Usage

```
find_skewness(.data, index = TRUE, value = FALSE, thres = NULL)
```

Arguments

.data	a data.frame or a tbl_df.
index	logical. When representing the information of a skewed variable, specify whether or not the variable is represented by an index. Returns an index if TRUE or a variable names if FALSE.
value	logical. If TRUE, returns the skewness value in the individual variable.
thres	Returns a skewness threshold value that has an absolute skewness greater than thres. The default is NULL to ignore the threshold. but, If value = TRUE, default to 0.5.

Value

Information on variables including skewness.

See Also

find_na, find_outliers.

Examples

```
## Not run:
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

find_skewness(carseats)

find_skewness(carseats, index = FALSE)

find_skewness(carseats, thres = 0.1)

find_skewness(carseats, value = TRUE)

find_skewness(carseats, value = TRUE, thres = 0.1)

## using dplyr ----------------------------------
library(dplyr)

# Perform simple data quality diagnosis of variables with outliers.
carseats %>%
  select(find_skewness(.)) %>%
  diagnose()

## End(Not run)
```

---

get_class *Extracting a class of variables*

Description

The get_class() gets class of variables in data.frame or tbl_df.

Usage

```
get_class(df)
```

Arguments

df	a data.frame or objects inheriting from data.frame

Value

a data.frame Variables of data.frame is as follows.

- variable : variables name
- class : class of variables

See Also

find_class.

Examples

```
## Not run:
# data.frame
get_class(iris)

# tbl_df
get_class(ggplot2::diamonds)

library(dplyr)
get_class(ggplot2::diamonds) %>%
  filter(class %in% c("integer", "numeric"))

## End(Not run)
```

---

get_column_info *Describe column of table in the DBMS*

Description

The get_column_info() retrieves the column information of the DBMS table through the tbl_bdi object of dplyr.

Usage

```
get_column_info(df)
```

Arguments

df	a tbl_dbi.

Value

An object of data.frame.

Column information of the DBMS table

- SQLite DBMS connected RSQLite::SQLite():
  - name: column name
  - type: data type in R
- MySQL/MariaDB DBMS connected RMySQL::MySQL():
  - name: column name
  - Sclass: data type in R
  - type: data type of column in the DBMS
  - length: data length in the DBMS
- Oracle DBMS connected ROracle::dbConnect():
  - name: column name
  - Sclass: column type in R
  - type: data type of column in the DBMS
  - len: length of column(CHAR/VARCHAR/VARCHAR2 data type) in the DBMS
  - precision: precision of column(NUMBER data type) in the DBMS
  - scale: decimal places of column(NUMBER data type) in the DBMS
  - nullOK: nullability

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  get_column_info
```

---

get_os *Finding Users Machine's OS*

Description

Get the operating system that users machines.

Usage

```
get_os()
```

Value

OS names. "windows" or "osx" or "linux"

Examples

```
get_os()
```

`imputate_na` *Imputate Missing values*

Description

Missing values are imputated with some representative values and statistical methods.

Usage

```
imputate_na(.data, xvar, yvar, method, seed, print_flag)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`xvar`	variable name to replace missing value.
`yvar`	target variable.
`method`	method of missing values imputation.
`seed`	integer. the random seed used in mice. only used "mice" method.
`print_flag`	logical. If TRUE, mice will print history on console. Use print_flag=FALSE for silent computation. Used only when method is "mice".

Details

imputate_na () creates an imputation class. The ‘imputation‘ class includes missing value position, imputated value, and method of missing value imputation, etc. The ‘imputation‘ class compares the imputated value with the original value to help determine whether the imputated value is used in the analysis.

See vignette("transformation") for an introduction to these concepts.

Value

An object of imputation class. Attributes of imputation class is as follows.

- var_type : the data type of predictor to replace missing value.
- method : method of missing value imputation.
  - predictor is numerical variable
    - "mean" : arithmetic mean
    - "median" : median
    - "mode" : mode
    - "knn" : K-nearest neighbors
    - "rpart" : Recursive Partitioning and Regression Trees
    - "mice" : Multivariate Imputation by Chained Equations
  - predictor is categorical variable
    - "mode" : mode
    - "rpart" : Recursive Partitioning and Regression Trees
    - "mice" : Multivariate Imputation by Chained Equations
- na_pos : position of missing value in predictor.
- seed : the random seed used in mice. only used "mice" method.
- type : "missing values". type of imputation.

See Also

imputate_outlier.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Replace the missing value of the Income variable with median
imputate_na(carseats, Income, method = "median")

# Replace the missing value of the Income variable with rpart
# The target variable is US.
imputate_na(carseats, Income, US, method = "rpart")

# Replace the missing value of the Urban variable with median
imputate_na(carseats, Urban, method = "mode")

# Replace the missing value of the Urban variable with mice
# The target variable is US.
imputate_na(carseats, Urban, US, method = "mice")

## using dplyr ----------------------------------
```

```
library(dplyr)

# The mean before and after the imputation of the Income variable
carseats %>%
  mutate(Income_imp = imputate_na(carseats, Income, US, method = "knn")) %>%
  group_by(US) %>%
  summarise(orig = mean(Income, na.rm = TRUE),
    imputation = mean(Income_imp))

# If the variable of interest is a numarical variable
income <- imputate_na(carseats, Income, US, method = "rpart")
income
summary(income)
plot(income)

# If the variable of interest is a categorical variable
urban <- imputate_na(carseats, Urban, US, method = "mice")
urban
summary(urban)
plot(urban)
```

`imputate_outlier` *Imputate Outliers*

Description

Outliers are imputated with some representative values and statistical methods.

Usage

```
imputate_outlier(.data, xvar, method)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`xvar`	variable name to replace missing value.
`method`	method of missing values imputation.

Details

imputate_outlier() creates an imputation class. The 'imputation' class includes missing value position, imputated value, and method of missing value imputation, etc. The 'imputation' class compares the imputated value with the original value to help determine whether the imputated value is used in the analysis.

See vignette("transformation") for an introduction to these concepts.

Value

An object of imputation class. Attributes of imputation class is as follows.

- method : method of missing value imputation.
  - predictor is numerical variable
    - "mean" : arithmetic mean
    - "median" : median
    - "mode" : mode
    - "capping" : Imputate the upper outliers with 95 percentile, and Imputate the bottom outliers with 5 percentile.
- outlier_pos : position of outliers in predictor.
- outliers : outliers. outliers corresponding to outlier_pos.
- type : "outliers". type of imputation.

See Also

imputate_na.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Replace the missing value of the Price variable with median
imputate_outlier(carseats, Price, method = "median")

# Replace the missing value of the Price variable with rpart
# The target variable is US.
imputate_outlier(carseats, Price, method = "capping")

## using dplyr -----------------------------------
library(dplyr)

# The mean before and after the imputation of the Price variable
carseats %>%
  mutate(Price_imp = imputate_outlier(carseats, Price, method = "capping")) %>%
  group_by(US) %>%
  summarise(orig = mean(Price, na.rm = TRUE),
    imputation = mean(Price_imp, na.rm = TRUE))

# If the variable of interest is a numarical variable
price <- imputate_outlier(carseats, Price)
price
summary(price)
plot(price)
```

`normality` *Performs the Shapiro-Wilk test of normality*

Description

The normality() performs Shapiro-Wilk test of normality of numeric values.

Usage

```
normality(.data, ...)

## S3 method for class 'data.frame'
normality(.data, ..., sample = 5000)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, normality() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
`sample`	the numer of samples to perform the test.
	See vignette("EDA") for an introduction to these concepts.

Details

This function is useful when used with the `group_by` function of the dplyr package. If you want to test by level of the categorical data you are interested in, rather than the whole observation, you can use group_tf as the group_by function. This function is computed `shapiro.test` function.

Value

An object of the same class as .data.

Normality test information

The information derived from the numerical data test is as follows.

- statistic : the value of the Shapiro-Wilk statistic.
- p_value : an approximate p-value for the test. This is said in Roystion(1995) to be adequate for p_value < 0.1.
- sample : the numer of samples to perform the test. The number of observations supported by the stats::shapiro.test function is 3 to 5000.

See Also

normality.tbl_dbi, diagnose_numeric.data.frame, describe.data.frame, plot_normality.data.frame.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Normality test of numerical variables
normality(carseats)

# Select the variable to describe
normality(carseats, Sales, Price)
normality(carseats, -Sales, -Price)
normality(carseats, 1)
normality(carseats, Sales, Price, sample = 300)

# Using dplyr::grouped_dt
library(dplyr)

gdata <- group_by(carseats, ShelveLoc, US)
normality(gdata, "Sales")
normality(gdata, sample = 250)

# Using pipes ---------------------------------
# Normality test of all numerical variables
carseats %>%
 normality()

# Positive values select variables
carseats %>%
 normality(Sales, Price)

# Positions values select variables
carseats %>%
 normality(1)

# Using pipes & dplyr -------------------------
# Test all numerical variables by 'ShelveLoc' and 'US',
# and extract only those with 'ShelveLoc' variable level is "Good".
carseats %>%
 group_by(ShelveLoc, US) %>%
 normality() %>%
 filter(ShelveLoc == "Good")

# extract only those with 'Urban' variable level is "Yes",
# and test 'Sales' by 'ShelveLoc' and 'US'
carseats %>%
 filter(Urban == "Yes") %>%
```

```
 group_by(ShelveLoc, US) %>%
 normality(Sales)

# Test log(Income) variables by 'ShelveLoc' and 'US',
# and extract only p.value greater than 0.01.
carseats %>%
 mutate(log_income = log(Income)) %>%
 group_by(ShelveLoc, US) %>%
 normality(log_income) %>%
 filter(p_value > 0.01)
```

`normality.tbl_dbi` *Performs the Shapiro-Wilk test of normality*

Description

The normality() performs Shapiro-Wilk test of normality of numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
normality(.data, ..., sample = 5000,
  in_database = FALSE, collect_size = Inf)
```

Arguments

`.data`	a tbl_dbi.
`...`	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, normality() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
`sample`	the numer of samples to perform the test.
`in_database`	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
`collect_size`	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE. See vignette("EDA") for an introduction to these concepts.

Details

This function is useful when used with the group_by function of the dplyr package. If you want to test by level of the categorical data you are interested in, rather than the whole observation, you can use group_tf as the group_by function. This function is computed shapiro.test function.

Value

An object of the same class as .data.

Normality test information

The information derived from the numerical data test is as follows.

- statistic : the value of the Shapiro-Wilk statistic.
- p_value : an approximate p-value for the test. This is said in Roystion(1995) to be adequate for p_value < 0.1.
- sample : the numer of samples to perform the test. The number of observations supported by the stats::shapiro.test function is 3 to 5000.

See Also

normality.data.frame, diagnose_numeric.tbl_dbi, describe.tbl_dbi, plot_normality.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Normality test of all numerical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  normality()

# Positive values select variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  normality(Sales, Price, collect_size  = 200)

# Positions values select variables
```

```
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  normality(1)

# Using pipes & dplyr -------------------------
# Test all numerical variables by 'ShelveLoc' and 'US',
# and extract only those with 'ShelveLoc' variable level is "Good".
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  group_by(ShelveLoc, US) %>%
  normality() %>%
  filter(ShelveLoc == "Good")

# extract only those with 'Urban' variable level is "Yes",
# and test 'Sales' by 'ShelveLoc' and 'US'
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  filter(Urban == "Yes") %>%
  group_by(ShelveLoc, US) %>%
  normality(Sales)

# Test log(Income) variables by 'ShelveLoc' and 'US',
# and extract only p.value greater than 0.01.

# SQLite extension functions for log
RSQLite::initExtension(con_sqlite)

con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  mutate(log_income = log(Income)) %>%
  group_by(ShelveLoc, US) %>%
  normality(log_income) %>%
  filter(p_value > 0.01)
```

`plot.bins` *Visualize Distribution for an "bins" object*

Description

Visualize both plots on a single screen. The plot at the top is a histogram representing the frequency of the level. The plot at the bottom is a bar chart representing the frequency of the level.

Usage

```
## S3 method for class 'bins'
plot(x, ...)
```

Arguments

x	an object of class "bins", usually, a result of a call to binning().
...	arguments to be passed to methods, such as graphical parameters (see par).

See Also

binning, print.bins, summary.bins.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Binning the carat variable. default type argument is "quantile"
bin <- binning(carseats$Income, nbins = 5)
plot(bin)
# Using another type argument
bin <- binning(carseats$Income, nbins = 5, type = "equal")
plot(bin)
bin <- binning(carseats$Income, nbins = 5, type = "pretty")
plot(bin)
bin <- binning(carseats$Income, nbins = 5, type = "kmeans")
plot(bin)
bin <- binning(carseats$Income, nbins = 5, type = "bclust")
plot(bin)
```

---

plot.imputation *Visualize Information for an "imputation" Object*

Description

Visualize two kinds of plot by attribute of ‘imputation‘ class. The imputation of a numerical variable is a density plot, and the imputation of a categorical variable is a bar plot.

Usage

```
## S3 method for class 'imputation'
plot(x, ...)
```

Arguments

x	an object of class "imputation", usually, a result of a call to imputate_na() or imputate_outlier().
...	arguments to be passed to methods, such as graphical parameters (see par). only applies when the model argument is TRUE, and is used for ... of the plot.lm () function.

See Also

imputate_na, imputate_outlier, summary.imputation.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Imputate missing values ----------------------------
# If the variable of interest is a numarical variable
income <- imputate_na(carseats, Income, US, method = "rpart")
income
summary(income)
plot(income)

# If the variable of interest is a categorical variable
urban <- imputate_na(carseats, Urban, US, method = "mice")
urban
summary(urban)
plot(urban)

# Imputate outliers ----------------------------------
# If the variable of interest is a numarical variable
price <- imputate_outlier(carseats, Price, method = "capping")
price
summary(price)
plot(price)
```

---

plot.optimal_bins *Visualize Distribution for an "optimal_bins" Object*

Description

It generates plots for distribution, bad rate, and weight of evidence after running smbinning and saving its output.

See vignette("transformation") for an introduction to these concepts.

Usage

```
## S3 method for class 'optimal_bins'
plot(x, type = c("dist", "goodrate", "badrate",
  "WoE"), sub = "", ...)
```

Arguments

x	an object of class "optimal_bins", usually, a result of a call to binning_by().
type	options for visualization. Distribution ("dist"), Good Rate ("goodrate"), Bad Rate ("badrate"), and Weight of Evidence ("WoE").
sub	subtitle for the chart (optional).
...	arguments to be passed to methods, such as graphical parameters (see par). only applies to the first graph that is implemented with the boxplot() function.

See Also

binning_by, plot.bins, smbinning.plot.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# optimal binning
bin <- binning_by(carseats, "US", "Advertising")
bin

# summary optimal_bins class
summary(bin)

# information value
attr(bin, "iv")

# information value table
attr(bin, "ivtable")

# visualize optimal_bins class
plot(bin, sub = "bins of Advertising variable")
```

---

plot.relate *Visualize Information for an "relate" Object*

Description

Visualize four kinds of plot by attribute of relate class.

Usage

```
## S3 method for class 'relate'
plot(x, model = FALSE, hex_thres = 1000,
  pal = RColorBrewer::brewer.pal(7, "YlOrRd"), ...)
```

Arguments

x	an object of class "relate", usually, a result of a call to relate().
model	logical. This argument selects whether to output the visualization result to the visualization of the object of the lm model to grasp the relationship between the numerical variables.
hex_thres	an integer. Use only when the target and predictor are numeric variables. Used when the number of observations is large. Specify the threshold of the observations to draw hexabin plots that are not scatterplots. The default value is 1000.
pal	Color palette to paint hexabin. Use only when the target and predictor are numeric variables. Applied only when the number of observations is greater than hex_thres.
...	arguments to be passed to methods, such as graphical parameters (see par). only applies when the model argument is TRUE, and is used for ... of the plot.lm () function.

See Also

relate, print.relate.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# If the target variable is a categorical variable
categ <- target_by(carseats, US)

# If the variable of interest is a numarical variable
cat_num <- relate(categ, Sales)
cat_num
summary(cat_num)
plot(cat_num)

# If the variable of interest is a categorical variable
cat_cat <- relate(categ, ShelveLoc)
cat_cat
summary(cat_cat)
plot(cat_cat)

##----------------------------------------------------
# If the target variable is a categorical variable
num <- target_by(carseats, Sales)

# If the variable of interest is a numarical variable
num_num <- relate(num, Price)
num_num
summary(num_num)
plot(num_num)
```

```
plot(num_num, hex_thres = 400)

# If the variable of interest is a categorical variable
num_cat <- relate(num, ShelveLoc)
num_cat
summary(num_cat)
plot(num_cat)
```

`plot.transform` *Visualize Information for an "transform" Object*

Description

Visualize two kinds of plot by attribute of ‘transform‘ class. The Transformation of a numerical variable is a density plot.

Usage

```
## S3 method for class 'transform'
plot(x, ...)
```

Arguments

x	an object of class "transform", usually, a result of a call to transform().
...	arguments to be passed to methods, such as graphical parameters (see par).

See Also

`transform`, `summary.transform`.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Standardization ------------------------------
advertising_minmax <- transform(carseats$Advertising, method = "minmax")
advertising_minmax
summary(advertising_minmax)
plot(advertising_minmax)

# Resolving Skewness  --------------------------
advertising_log <- transform(carseats$Advertising, method = "log")
advertising_log
summary(advertising_log)
plot(advertising_log)
```

plot_correlate *Visualize correlation plot of numerical data*

Description

The plot_correlate() visualize correlation plot for find relationship between two numerical variables.

Usage

```
plot_correlate(.data, ...)

## S3 method for class 'data.frame'
plot_correlate(.data, ...)
```

Arguments

.data	a data.frame or a tbl_df.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, plot_correlate() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing. See vignette("EDA") for an introduction to these concepts.

Details

The scope of the visualization is the provide a correlation information. Since the plot is drawn for each variable, if you specify more than one variable in the ... argument, the specified number of plots are drawn.

See Also

plot_correlate.tbl_dbi, plot_outlier.data.frame.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Visualize correlation plot of all numerical variables
plot_correlate(carseats)

# Select the variable to compute
plot_correlate(carseats, Sales, Price)
plot_correlate(carseats, -Sales, -Price)
```

```
plot_correlate(carseats, "Sales", "Price")
plot_correlate(carseats, 1)

# Using dplyr::grouped_dt
library(dplyr)

gdata <- group_by(carseats, ShelveLoc, US)
plot_correlate(gdata, "Sales")
plot_correlate(gdata)

# Using pipes ---------------------------------
# Visualize correlation plot of all numerical variables
carseats %>%
  plot_correlate()
# Positive values select variables
carseats %>%
  plot_correlate(Sales, Price)
# Negative values to drop variables
carseats %>%
  plot_correlate(-Sales, -Price)
# Positions values select variables
carseats %>%
  plot_correlate(1)
# Positions values select variables
carseats %>%
  plot_correlate(-1, -2, -3, -5, -6)

# Using pipes & dplyr -------------------------
# Visualize correlation plot of 'Sales' variable by 'ShelveLoc'
# and 'US' variables.
carseats %>%
group_by(ShelveLoc, US) %>%
plot_correlate(Sales)

# Extract only those with 'ShelveLoc' variable level is "Good",
# and visualize correlation plot of 'Sales' variable by 'Urban'
# and 'US' variables.
carseats %>%
 filter(ShelveLoc == "Good") %>%
 group_by(Urban, US) %>%
 plot_correlate(Sales)
```

`plot_correlate.tbl_dbi` *Visualize correlation plot of numerical data*

Description

The plot_correlate() visualize correlation plot for find relationship between two numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
plot_correlate(.data, ..., in_database = FALSE,
  collect_size = Inf)
```

Arguments

.data	a tbl_dbi.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, plot_correlate() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE. See vignette("EDA") for an introduction to these concepts.

Details

The scope of the visualization is the provide a correlation information. Since the plot is drawn for each variable, if you specify more than one variable in the ... argument, the specified number of plots are drawn.

See Also

plot_correlate.data.frame, plot_outlier.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Visualize correlation plot of all numerical variables
con_sqlite %>%
```

```
  tbl("TB_CARSEATS") %>%
  plot_correlate()

# Positive values select variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_correlate(Sales, Price, collect_size = 200)

# Negative values to drop variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_correlate(-Sales, -Price)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_correlate(1)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_correlate(-1, -2, -3, -5, -6)

# Using pipes & dplyr -------------------------
# Visualize correlation plot of 'Sales' variable by 'ShelveLoc'
# and 'US' variables.
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  group_by(ShelveLoc, US) %>%
  plot_correlate(Sales)

# Extract only those with 'ShelveLoc' variable level is "Good",
# and visualize correlation plot of 'Sales' variable by 'Urban'
# and 'US' variables.
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  filter(ShelveLoc == "Good") %>%
  group_by(Urban, US) %>%
  plot_correlate(Sales)
```

`plot_normality` *Plot distribution information of numerical data*

Description

The plot_normality() visualize distribution information for normality test of the numerical data.

Usage

```
plot_normality(.data, ...)

## S3 method for class 'data.frame'
plot_normality(.data, ...)
```

Arguments

.data	a data.frame or a tbl_df.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, plot_normality() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing. See vignette("EDA") for an introduction to these concepts.

Details

The scope of the visualization is the provide a distribution information. Since the plot is drawn for each variable, if you specify more than one variable in the ... argument, the specified number of plots are drawn.

Distribution information

The plot derived from the numerical data vizualization is as follows.

- histogram by original data
- q-q plot by original data
- histogram by log transfer data
- histogram by square root transfer data

See Also

plot_normality.tbl_dbi, plot_outlier.data.frame.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Visualization of all numerical variables
plot_normality(carseats)

# Select the variable to plot
plot_normality(carseats, Income, Price)
```

```
plot_normality(carseats, -Income, -Price)
plot_normality(carseats, 1)

# Using dtplyr::grouped_dt
library(dplyr)

gdata <- group_by(carseats, ShelveLoc, US)
plot_normality(carseats)
plot_normality(carseats, "Sales")

# Using pipes ---------------------------------
# Visualization of all numerical variables
carseats %>%
 plot_normality()

# Positive values select variables
carseats %>%
plot_normality(Income, Price)

# Positions values select variables
carseats %>%
 plot_normality(1)

# Using pipes & dplyr -------------------------
# Plot 'Sales' variable by 'ShelveLoc' and 'US'
carseats %>%
 group_by(ShelveLoc, US) %>%
 plot_normality(Sales)

# extract only those with 'ShelveLoc' variable level is "Good",
# and plot 'Income' by 'US'
carseats %>%
 filter(ShelveLoc == "Good") %>%
 group_by(US) %>%
 plot_normality(Income)
```

`plot_normality.tbl_dbi` *Plot distribution information of numerical data*

Description

The plot_normality() visualize distribution information for normality test of the numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
plot_normality(.data, ..., in_database = FALSE,
  collect_size = Inf)
```

Arguments

.data	a tbl_dbi.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, plot_normality() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE. See vignette("EDA") for an introduction to these concepts.

Details

The scope of the visualization is the provide a distribution information. Since the plot is drawn for each variable, if you specify more than one variable in the ... argument, the specified number of plots are drawn.

Distribution information

The plot derived from the numerical data vizualization is as follows.

- histogram by original data
- q-q plot by original data
- histogram by log transfer data
- histogram by square root transfer data

See Also

plot_normality.data.frame, plot_outlier.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
```

```
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# Using pipes ---------------------------------
# Visualization of all numerical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_normality()

# Positive values select variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_normality(Income, Price, collect_size = 200)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_normality(1)

# Using pipes & dplyr -------------------------
# Plot 'Sales' variable by 'ShelveLoc' and 'US'
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  group_by(ShelveLoc, US) %>%
  plot_normality(Sales)

# extract only those with 'ShelveLoc' variable level is "Good",
# and plot 'Income' by 'US'
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  filter(ShelveLoc == "Good") %>%
  group_by(US) %>%
  plot_normality(Income)
```

---

plot_outlier *Plot outlier information of numerical data diagnosis*

Description

The plot_outlier() visualize outlier information for diagnosing the quality of the numerical data.

Usage

```
plot_outlier(.data, ...)

## S3 method for class 'data.frame'
plot_outlier(.data, ..., col = "lightblue")
```

Arguments

.data	a data.frame or a tbl_df.
...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, plot_outlier() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
col	a color to be used to fill the bars. The default is "lightblue".

Details

The scope of the diagnosis is the provide a outlier information. Since the plot is drawn for each variable, if you specify more than one variable in the ... argument, the specified number of plots are drawn.

Outlier diagnostic information

The plot derived from the numerical data diagnosis is as follows.

- With outliers box plot
- Without outliers box plot
- With outliers histogram
- Without outliers histogram

See vignette("diagonosis") for an introduction to these concepts.

See Also

plot_outlier.tbl_dbi, diagnose_outlier.data.frame.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Visualization of all numerical variables
plot_outlier(carseats)

# Select the variable to diagnose
plot_outlier(carseats, Sales, Price)
plot_outlier(carseats, -Sales, -Price)
plot_outlier(carseats, "Sales", "Price")
plot_outlier(carseats, 6)

# Using the col argument
plot_outlier(carseats, Sales, col = "gray")
```

```
# Using pipes ---------------------------------
library(dplyr)

# Visualization of all numerical variables
carseats %>%
  plot_outlier()
# Positive values select variables
carseats %>%
  plot_outlier(Sales, Price)
# Negative values to drop variables
carseats %>%
  plot_outlier(-Sales, -Price)
# Positions values select variables
carseats %>%
  plot_outlier(6)
# Positions values select variables
carseats %>%
  plot_outlier(-1, -5)

# Using pipes & dplyr -------------------------
# Visualization of numerical variables with a ratio of
# outliers greater than 1%
carseats %>%
  plot_outlier(carseats %>%
      diagnose_outlier() %>%
      filter(outliers_ratio > 1) %>%
      select(variables) %>%
      pull())
```

`plot_outlier.tbl_dbi` *Plot outlier information of numerical data diagnosis in the DBMS*

Description

The plot_outlier() visualize outlier information for diagnosing the quality of the numerical(INTEGER, NUMBER, etc.) column of the DBMS table through tbl_dbi.

Usage

```
## S3 method for class 'tbl_dbi'
plot_outlier(.data, ..., col = "lightblue",
  in_database = FALSE, collect_size = Inf)
```

Arguments

`.data`	a tbl_dbi.

...	one or more unquoted expressions separated by commas. You can treat variable names like they are positions. Positive values select variables; negative values to drop variables. If the first expression is negative, plot_outlier() will automatically start with all variables. These arguments are automatically quoted and evaluated in a context where column names represent column positions. They support unquoting and splicing.
col	a color to be used to fill the bars. The default is "lightblue".
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.

Details

The scope of the diagnosis is the provide a outlier information. Since the plot is drawn for each variable, if you specify more than one variable in the ... argument, the specified number of plots are drawn.

Outlier diagnostic information

The plot derived from the numerical data diagnosis is as follows.

- With outliers box plot
- Without outliers box plot
- With outliers histogram
- Without outliers histogram

See vignette("diagonosis") for an introduction to these concepts.

See Also

plot_outlier.data.frame, diagnose_outlier.tbl_dbi.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)
```

```
# Using pipes --------------------------------
# Visualization of all numerical variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_outlier()

# Positive values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_outlier(Sales, Price)

# Negative values to drop variables, and In-memory mode and collect size is 200
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_outlier(-Sales, -Price, collect_size = 200)

# Positions values select variables
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_outlier(6)

# Positions values select variables
carseats %>%
  plot_outlier(-1, -5)

# Using pipes & dplyr ------------------------
# Visualization of numerical variables with a ratio of
# outliers greater than 1%
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  plot_outlier(con_sqlite %>%
                 tbl("TB_CARSEATS") %>%
                 diagnose_outlier() %>%
                 filter(outliers_ratio > 1) %>%
                 select(variables) %>%
                 pull())
```

`print.relate` *Summarizing relate information*

Description

print and summary method for "relate" class.

Usage

```
## S3 method for class 'relate'
print(x, ...)
```

Arguments

x	an object of class "relate", usually, a result of a call to relate().
...	further arguments passed to or from other methods.

Details

print.relate tries to be smart about formatting four kinds of relate. summary.relate tries to be smart about formatting four kinds of relate.

See Also

plot.relate.

Examples

```
## Not run:
# Generate data for the example
diamonds2 <- diamonds
diamonds2[sample(seq(NROW(diamonds2)), 250), "price"] <- NA
diamonds2[sample(seq(NROW(diamonds2)), 20), "clarity"] <- NA

# Binning the carat variable. default type argument is "quantile"
bin <- binning(diamonds2$carat)

# Print bins class object
bin

# Summarise bins class object
summary(bin)

## End(Not run)


# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# If the target variable is a categorical variable
categ <- target_by(carseats, US)

# If the variable of interest is a numarical variable
cat_num <- relate(categ, Sales)
cat_num
summary(cat_num)
plot(cat_num)

# If the variable of interest is a categorical variable
cat_cat <- relate(categ, ShelveLoc)
cat_cat
summary(cat_cat)
```

```
plot(cat_cat)

##------------------------------------------------
# If the target variable is a categorical variable
num <- target_by(carseats, Sales)

# If the variable of interest is a numarical variable
num_num <- relate(num, Price)
num_num
summary(num_num)
plot(num_num)

# If the variable of interest is a categorical variable
num_cat <- relate(num, ShelveLoc)
num_cat
summary(num_cat)
plot(num_cat)
```

`relate` *Relationship between target variable and variable of interest*

Description

The relationship between the target variable and the variable of interest (predictor) is briefly analyzed.

Usage

```
relate(.data, predictor)
```

Arguments

`.data`	A target_df.
`predictor`	variable of interest. predictor. See vignette("relate") for an introduction to these concepts.

Details

Returns the four types of results that correspond to the combination of the target variable and the data type of the variable of interest.

- target variable: categorical variable
  - predictor: categorical variable
    - contegency table
    - c("xtabs", "table") class
  - predictor: numerical variable
    - descriptive statistic for each levles and total observation.

- target variable: numerical variable
  - predictor: categorical variable
    - ANOVA test. "lm" class.
  - predictor: numerical variable
    - simple linear model. "lm" class.

Value

An object of the class as relate. Attributes of relate class is as follows.

- target : name of target variable
- predictor : name of predictor
- model : levels of binned value.
- raw : table_df with two variables target and predictor.

Descriptive statistic information

The information derived from the numerical data describe is as follows.

- mean : arithmetic average
- sd : standard devation
- se_mean : standrd error mean. sd/sqrt(n)
- IQR : interqurtle range (Q3-Q1)
- skewness : skewness
- kurtosis : kurtosis
- p25 : Q1. 25% percentile
- p50 : median. 50% percentile
- p75 : Q3. 75% percentile
- p01, p05, p10, p20, p30 : 1%, 5%, 20%, 30% percentiles
- p40, p60, p70, p80 : 40%, 60%, 70%, 80% percentiles
- p90, p95, p99, p100 : 90%, 95%, 99%, 100% percentiles

See Also

print.relate, plot.relate.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# If the target variable is a categorical variable
categ <- target_by(carseats, US)
```

```
# If the variable of interest is a numarical variable
cat_num <- relate(categ, Sales)
cat_num
summary(cat_num)
plot(cat_num)

# If the variable of interest is a categorical variable
cat_cat <- relate(categ, ShelveLoc)
cat_cat
summary(cat_cat)
plot(cat_cat)

##------------------------------------------------
# If the target variable is a categorical variable
num <- target_by(carseats, Sales)

# If the variable of interest is a numarical variable
num_num <- relate(num, Price)
num_num
summary(num_num)
plot(num_num)

# If the variable of interest is a categorical variable
num_cat <- relate(num, ShelveLoc)
num_cat
summary(num_cat)
plot(num_cat)
```

`summary.bins` *Summarizing Binned Variable*

Description

summary method for class "bins" and "optimal_bins".

Usage

```
## S3 method for class 'bins'
summary(object, ...)

## S3 method for class 'bins'
print(x, ...)
```

Arguments

`object`	an object of class "bins" and "optimal_bins", usually, a result of a call to binning().
`...`	further arguments passed to or from other methods.

x — an object of class "bins" and "optimal_bins", usually, a result of a call to binning().

Details

print.bins() tries to be smart about formatting the frequency of bins, binned type, number of bins. summary.bins tries to be smart about formatting the levles, frequency of levels(bins), the ratio of levels in total observations. And this information is data.frame object.

See vignette("transformation") for an introduction to these concepts.

Value

The function summary.bins() computes and returns a data.frame of summary statistics of the binned given in object. Variables of data frame is as follows.

- levels : levles of factor.
- freq : frequency of levels.
- rate : ratio of levels in total observations. it is not percentage.

See Also

binning

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Binning the carat variable. default type argument is "quantile"
bin <- binning(carseats$Income)

# Print bins class object
bin

# Summarise bins class object
summary(bin)
```

---

summary.imputation *Summarizing imputation information*

Description

print and summary method for "imputation" class.

Usage

```
## S3 method for class 'imputation'
summary(object, ...)
```

Arguments

object	an object of class "imputation", usually, a result of a call to imputate_na() or imputate_outlier().
...	further arguments passed to or from other methods.

Details

summary.imputation tries to be smart about formatting two kinds of imputation.

See Also

imputate_na, imputate_outlier, summary.imputation.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Imputate missing values ----------------------------
# If the variable of interest is a numarical variable
income <- imputate_na(carseats, Income, US, method = "rpart")
income
summary(income)
plot(income)

# If the variable of interest is a categorical variable
urban <- imputate_na(carseats, Urban, US, method = "mice")
urban
summary(urban)
plot(urban)

# Imputate outliers ----------------------------------
# If the variable of interest is a numarical variable
price <- imputate_outlier(carseats, Price, method = "capping")
price
summary(price)
plot(price)
```

summary.transform *Summarizing transformation information*

Description

print and summary method for "transform" class.

Usage

```
## S3 method for class 'transform'
summary(object, ...)
```

Arguments

object	an object of class "transform", usually, a result of a call to transform().
...	further arguments passed to or from other methods.

Details

summary.transform compares the distribution of data before and after data conversion.

See Also

transform, summary.transform.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Standardization ------------------------------
advertising_minmax <- transform(carseats$Advertising, method = "minmax")
advertising_minmax
summary(advertising_minmax)
plot(advertising_minmax)

# Resolving Skewness  --------------------------
advertising_log <- transform(carseats$Advertising, method = "log")
advertising_log
summary(advertising_log)
plot(advertising_log)
```

`target_by` *Target by one variables*

Description

In the data analysis, a target_df class is created to identify the relationship between the target variable and the other variable.

Usage

```
target_by(.data, target, ...)

## S3 method for class 'data.frame'
target_by(.data, target, ...)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`target`	target variable.
`...`	arguments to be passed to methods.

Details

Data analysis proceeds with the purpose of predicting target variables that correspond to the facts of interest, or examining associations and relationships with other variables of interest. Therefore, it is a major challenge for EDA to examine the relationship between the target variable and its corresponding variable. Based on the derived relationships, analysts create scenarios for data analysis.

target_by() inherits the `grouped_df` class and returns a target_df class containing information about the target variable and the variable.

See vignette("EDA") for an introduction to these concepts.

Value

an object of target_df class. Attributes of target_df class is as follows.

- type_y : the data type of target variable.

See Also

`relate`.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# If the target variable is a categorical variable
categ <- target_by(carseats, US)

# If the variable of interest is a numarical variable
cat_num <- relate(categ, Sales)
cat_num
summary(cat_num)
plot(cat_num)

# If the variable of interest is a categorical variable
cat_cat <- relate(categ, ShelveLoc)
cat_cat
summary(cat_cat)
plot(cat_cat)

##-----------------------------------------------
# If the target variable is a categorical variable
num <- target_by(carseats, Sales)

# If the variable of interest is a numarical variable
num_num <- relate(num, Price)
num_num
summary(num_num)
plot(num_num)

# If the variable of interest is a categorical variable
num_cat <- relate(num, ShelveLoc)
num_cat
summary(num_cat)
plot(num_cat)
```

---

target_by.tbl_dbi *Target by one column in the DBMS*

Description

In the data analysis, a target_df class is created to identify the relationship between the target column and the other column of the DBMS table through tbl_dbi

Usage

```
## S3 method for class 'tbl_dbi'
```

```
target_by(.data, target, in_database = FALSE,
  collect_size = Inf, ...)
```

Arguments

.data	a tbl_dbi.
target	target variable.
in_database	Specifies whether to perform in-database operations. If TRUE, most operations are performed in the DBMS. if FALSE, table data is taken in R and operated in-memory. Not yet supported in_database = TRUE.
collect_size	a integer. The number of data samples from the DBMS to R. Applies only if in_database = FALSE.
...	arguments to be passed to methods.

Details

Data analysis proceeds with the purpose of predicting target variables that correspond to the facts of interest, or examining associations and relationships with other variables of interest. Therefore, it is a major challenge for EDA to examine the relationship between the target variable and its corresponding variable. Based on the derived relationships, analysts create scenarios for data analysis.

target_by() inherits the grouped_df class and returns a target_df class containing information about the target variable and the variable.

See vignette("EDA") for an introduction to these concepts.

Value

an object of target_df class. Attributes of target_df class is as follows.

- type_y : the data type of target variable.

See Also

target_by.data.frame, relate.

Examples

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)
```

```
# If the target variable is a categorical variable
categ <- target_by(con_sqlite %>% tbl("TB_CARSEATS") , US)

# If the variable of interest is a numarical variable
cat_num <- relate(categ, Sales)
cat_num
summary(cat_num)
plot(cat_num)

# If the variable of interest is a categorical column
cat_cat <- relate(categ, ShelveLoc)
cat_cat
summary(cat_cat)
plot(cat_cat)

##-------------------------------------------------
# If the target variable is a categorical column,
# and In-memory mode and collect size is 350
num <- target_by(con_sqlite %>% tbl("TB_CARSEATS"), Sales, collect_size = 350)

# If the variable of interest is a numarical column
num_num <- relate(num, Price)
num_num
summary(num_num)
plot(num_num)
plot(num_num, hex_thres = 400)

# If the variable of interest is a categorical column
num_cat <- relate(num, ShelveLoc)
num_cat
summary(num_cat)
plot(num_cat)
```

`transform` *Data Transformations*

Description

Performs variable transformation for standardization and resolving skewness of numerical variables.

Usage

```
transform(x, method = c("zscore", "minmax", "log", "log+1", "sqrt",
  "1/x", "x^2", "x^3"))
```

Arguments

`x`	numeric vector for transformation.
`method`	method of transformations.

Details

transform() creates an transform class. The 'transform' class includes original data, transformed data, and method of transformation.

See vignette("transformation") for an introduction to these concepts.

Value

An object of transform class. Attributes of transform class is as follows.

- method : method of transformation data.
  - Standardization
    - "zscore" : z-score transformation. (x - mu) / sigma
    - "minmax" : minmax transformation. (x - min) / (max - min)
  - Resolving Skewness
    - "log" : log transformation. log(x)
    - "log+1" : log transformation. log(x + 1). Used for values that contain 0.
    - "sqrt" : square root transformation.
    - "1/x" : 1 / x transformation
    - "x^2" : x square transformation
    - "x^3" : x^3 square transformation

See Also

summary.transform, plot.transform.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# Standardization ------------------------------
advertising_minmax <- transform(carseats$Advertising, method = "minmax")
advertising_minmax
summary(advertising_minmax)
plot(advertising_minmax)

# Resolving Skewness  --------------------------
advertising_log <- transform(carseats$Advertising, method = "log")
advertising_log
summary(advertising_log)
plot(advertising_log)

# Using dplyr ----------------------------------
library(dplyr)

carseats %>%
  mutate(Advertising_log = transform(Advertising, method = "log+1")) %>%
  lm(Sales ~ Advertising_log, data = .)
```

`transformation_report` *Reporting the information of transformation*

Description

The transformation_report() report the information of transformate numerical variables for object inheriting from data.frame.

Usage

```
transformation_report(.data, target = NULL, output_format = c("pdf",
  "html"), output_file = NULL, output_dir = tempdir(),
  font_family = NULL, browse = TRUE)
```

Arguments

`.data`	a data.frame or a `tbl_df`.
`target`	target variable. If the target variable is not specified, the method of using the target variable information is not performed when the missing value is imputed. and Optimal binning is not performed if the target variable is not a binary class.
`output_format`	report output type. Choose either "pdf" and "html". "pdf" create pdf file by knitr::knit(). "html" create html file by rmarkdown::render().
`output_file`	name of generated file. default is NULL.
`output_dir`	name of directory to generate report file. default is tempdir().
`font_family`	charcter. font family name for figure in pdf.
`browse`	logical. choose whether to output the report results to the browser.

Details

Generate transformation reports automatically. You can choose to output to pdf and html files. This is useful for Binning a data frame with a large number of variables than data with a small number of variables. For pdf output, Korean Gothic font must be installed in Korean operating system.

Reported information

The transformation process will report the following information:

- Imputation
  - Missing Values
    - * Variable names including missing value
  - Outliers
    - * Variable names including outliers
- Resolving Skewness
  - Skewed variables information

        * * Variable names with an absolute value of skewness greater than or equal to 0.5
- Binning
  - Numerical Variables for Binning
  - Binning
    * Numeric variable names
  - Optimal Binning
    * Numeric variable names

See vignette("transformation") for an introduction to these concepts.

Examples

```
# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# reporting the Binning information -------------------------
# create pdf file. file name is Transformation_Report.pdf & No target variable
transformation_report(carseats)
# create pdf file. file name is Transformation_Report.pdf
transformation_report(carseats, US)
# create pdf file. file name is Transformation.pdf
transformation_report(carseats, "US", output_file = "Transformation.pdf")
# create html file. file name is Transformation_Report.html
transformation_report(carseats, "US", output_format = "html")
# create html file. file name is Transformation.html
transformation_report(carseats, US, output_format = "html", output_file = "Transformation.html")
```

Index