

Package ‘overviewR’

October 14, 2022

Type Package

Title Easily Extracting Information About Your Data

Version 0.0.11

Description Makes it easy to display descriptive information on a data set. Getting an easy overview of a data set by displaying and visualizing sample information in different tables (e.g., time and scope conditions). The package also provides publishable ‘LaTeX’ code to present the sample information.

License GPL-3

URL <https://github.com/cosimameyer/overviewR>

BugReports <https://github.com/cosimameyer/overviewR/issues>

Depends R (>= 3.5.0)

Imports data.table (>= 1.14.2), dplyr (>= 1.0.0), ggplot2 (>= 3.3.2), ggrepel (>= 0.8.2), ggvenn (>= 0.1.8), rlang, tibble (>= 3.0.1), tidyr

Suggests countrycode, covr, devtools, knitr, magrittr, pkgdown, rmarkdown, spelling, testthat, xtable

VignetteBuilder knitr, rmarkdown

Encoding UTF-8

Language en-US

LazyData true

RoxygenNote 7.1.2

NeedsCompilation no

Author Cosima Meyer [cre, aut],
Dennis Hammerschmidt [aut]

Maintainer Cosima Meyer <cosima.meyer@gmail.com>

Repository CRAN

Date/Publication 2022-08-06 08:50:02 UTC

R topics documented:

| | |
|------------------------------|----|
| .overview_heat | 2 |
| .overview_tab | 3 |
| calculate_share_non_row_wise | 4 |
| calculate_share_row_wise | 4 |
| find_int_runs | 5 |
| overview_add_na_output | 5 |
| overview_crossplot | 6 |
| overview_crosstab | 7 |
| overview_heat | 8 |
| overview_na | 9 |
| overview_overlap | 10 |
| overview_plot | 11 |
| overview_plot_absolute | 12 |
| overview_plot_percentage | 12 |
| overview_print | 13 |
| overview_tab | 15 |
| overview_tab_df | 16 |
| overview_tab_dt | 16 |
| theme_heat_plot | 17 |
| theme_na_plot | 17 |
| toydata | 18 |

Index 19

| | |
|-----------------------------|----------------------------|
| <code>.overview_heat</code> | <code>.overview_tab</code> |
|-----------------------------|----------------------------|

Description

Internal function that calculates the ‘overview_tab’ for data.table objects

Usage

```
.overview_heat(
  dat = NULL,
  id = NULL,
  time = NULL,
  label = FALSE,
  perc = FALSE,
  col_low = NULL,
  col_high = NULL,
  xaxis = NULL,
  yaxis = NULL,
  theme_plot = NULL,
  exp_total = NULL,
  col_names = NULL
)
```

Arguments

| | |
|-------------------------|--|
| <code>dat</code> | The data set |
| <code>id</code> | The scope (e.g., country codes or individual IDs). The axis is ordered in ascending order by default. |
| <code>time</code> | The time (e.g., time periods given by years, months, ...) |
| <code>label</code> | If TRUE (default), the total number of observations/percentages of observations are displayed. If FALSE, it returns no labels. |
| <code>perc</code> | If FALSE (default) plot returns the total number of observations per time-scope-unit. If TRUE, it returns the number of observations per time-scope-unit in percentage |
| <code>col_low</code> | Hex color code for the lowest value (default is "#dceaf2") |
| <code>col_high</code> | Hex color code for the highest value (default is "#2A5773") |
| <code>xaxis</code> | Label of your x axis ("Time frame" is default) |
| <code>yaxis</code> | Label of your y axis ("Sample" is default) |
| <code>theme_plot</code> | Previously generated theme |
| <code>exp_total</code> | Expected total number of observations (i.e. maximum) for time unit. |
| <code>col_names</code> | The column names (containing id and time) |

Value

A ggplot

`.overview_tab`
`.overview_tab`

Description

Internal function that calculates the 'overview_tab' for data.table objects

Usage

```
.overview_tab(dat = NULL, id = NULL, time = NULL, col_names = NULL)
```

Arguments

| | |
|------------------------|--|
| <code>dat</code> | Your data set |
| <code>id</code> | Scope (e.g., country codes or individual IDs) |
| <code>time</code> | Time (e.g., time periods given by years, months, ...). There are three options to add a date variable: 1) Time can be a character vector containing one time variable, 2) a time variable following the YYYY-MM-DD format, or 3) or a list containing multiple time variables ('time = list(year = NULL, month = NULL, day = NULL)'). |
| <code>col_names</code> | The column names (containing id and time) |

Value

A data.table

```
calculate_share_non_row_wise  
    calculate_share_non_row_wise
```

Description

Function used in 'overview_na' to calculate the column-wise share of NA

Usage

```
calculate_share_non_row_wise(dat = NULL)
```

Arguments

dat Data frame

Value

The function returns a data set that has the information on the column-wise NA share

```
calculate_share_row_wise  
    calculate_share_row_wise
```

Description

Function used in 'overview_na' to calculate the share of NA row-wise

Usage

```
calculate_share_row_wise(dat = NULL)
```

Arguments

dat Data frame

Value

The function returns a data set that has the information on the row-wise NA share

| | |
|---------------|----------------------|
| find_int_runs | <i>find_int_runs</i> |
|---------------|----------------------|

Description

Function used in 'overview_tab' to find running integers

Usage

```
find_int_runs(run = NULL)
```

Arguments

run Variable (integer) that should be checked for consecutive values

Value

The function returns a data set

| | |
|------------------------|-------------------------------|
| overview_add_na_output | <i>overview_add_na_output</i> |
|------------------------|-------------------------------|

Description

Function used in 'overview_na' to generate a new data frame with na_count and percentage share of NAs for each row

Usage

```
overview_add_na_output(dat_result = NULL, dat = NULL)
```

Arguments

dat_result Data.frame from 'overview_na'
dat Data frame

Value

The function returns a data set that has the information on the row-wise NA share

overview_crossplot *overview_crossplot*

Description

This function plots a ggplot to visualize a cross table plot.

Usage

```
overview_crossplot(  
  dat,  
  id,  
  time,  
  cond1,  
  cond2,  
  threshold1,  
  threshold2,  
  xaxis = "Condition 1",  
  yaxis = "Condition 2",  
  label = FALSE,  
  color = FALSE  
)
```

Arguments

| | |
|------------|---|
| dat | Your data set |
| id | Your scope (e.g., country codes or individual IDs). If the id variable contains NAs, they will not be included in the plot. |
| time | Your time (e.g., time periods given by years, months, ...) |
| cond1 | Variable that describes the first condition |
| cond2 | Variable that describes the second condition |
| threshold1 | A threshold for cond1 |
| threshold2 | A threshold for cond2 |
| xaxis | Label of the x axis ("Condition 1" is default) |
| yaxis | Label of the y axis ("Condition 2" is default) |
| label | Label of the observations. Overlapping labels are avoided by using 'ggrepel' |
| color | Color of the different observation groups |

Value

A ggplot figure that presents the sample information visually in a cross table

Examples

```
data(toydata)
overview_crossplot(
  dat = toydata,
  cond1 = gdp,
  cond2 = population,
  threshold1 = 25000,
  threshold2 = 27000,
  id = ccode,
  time = year
)
```

| | |
|-------------------|--------------------------|
| overview_crosstab | <i>overview_crosstab</i> |
|-------------------|--------------------------|

Description

Sorts a data set conditionally in a cross table. This can be helpful to get a sense of the time and scope conditions of a data set. Note, if used with a data set that has multiple observations on the id-time unit, the function automatically aggregates this information using the mean.

Usage

```
overview_crosstab(dat, cond1, cond2, threshold1, threshold2, id, time)
```

Arguments

| | |
|------------|---|
| dat | A data set object |
| cond1 | Variable that describes the first condition |
| cond2 | Variable that describes the second condition |
| threshold1 | A threshold for cond1 |
| threshold2 | A threshold for cond2 |
| id | Scope (e.g., country codes or individual IDs) |
| time | Time (e.g., time periods given by years, months, ...) |

Value

A data frame object that contains a summary of the data set that can later be converted to a 'LaTeX' output using `overview_print`

Examples

```

data(toydata)
overview_crosstab(
  dat = toydata,
  cond1 = gdp,
  cond2 = population,
  threshold1 = 25000,
  threshold2 = 27000,
  id = ccode,
  time = year
)

```

```
overview_heat
```

```
overview_heat
```

Description

This function plots a heat map to visualize the coverage of the time-scope-units of the data. Options include total number of cases per time-scope-unit or relative number in percentage.

Usage

```

overview_heat(
  dat,
  id,
  time,
  perc = FALSE,
  exp_total = NULL,
  xaxis = "Time frame",
  yaxis = "Sample",
  col_low = "#dceaf2",
  col_high = "#2A5773",
  label = TRUE
)

```

Arguments

| | |
|------------------------|--|
| <code>dat</code> | The data set |
| <code>id</code> | The scope (e.g., country codes or individual IDs). The axis is ordered in ascending order by default. |
| <code>time</code> | The time (e.g., time periods given by years, months, ...) |
| <code>perc</code> | If FALSE (default) plot returns the total number of observations per time-scope-unit. If TRUE, it returns the number of observations per time-scope-unit in percentage |
| <code>exp_total</code> | Expected total number of observations (i.e. maximum) for time unit. |
| <code>xaxis</code> | Label of your x axis ("Time frame" is default) |

| | |
|----------|--|
| yaxis | Label of your y axis ("Sample" is default) |
| col_low | Hex color code for the lowest value (default is "#dceaf2") |
| col_high | Hex color code for the lowest value (default is "#2A5773") |
| label | If TRUE (default), the total number of observations/percentages of observations are displayed. If FALSE, it returns no labels. |

Value

A ggplot figure that presents sample coverage visually

Examples

```
data(toydata)
overview_heat(toydata, ccode, year, perc = TRUE, exp_total = 12)
```

| | |
|-------------|--------------------|
| overview_na | <i>overview_na</i> |
|-------------|--------------------|

Description

This function plots a ggplot to visualize the distribution of NAs across all variables in the data set.

Usage

```
overview_na(
  dat,
  yaxis = "Variables",
  perc = TRUE,
  row_wise = FALSE,
  add = FALSE
)
```

Arguments

| | |
|----------|---|
| dat | Your data set |
| yaxis | Label of your y axis ("Variables" is default) |
| perc | If TRUE (default) plot returns the number of NAs in percentage |
| row_wise | If TRUE (FALSE is default) plot return the number of NAs per row |
| add | If TRUE (FALSE is default) it generates a new data frame with na_count and percentage share of NAs for each row |

Value

Depending on the selection, the function returns a ggplot figure that presents the distribution of NAs in the data set or adds the information on the row-wise NA share

Examples

```
data(toydata)
overview_na(toydata, perc = FALSE)
```

| | |
|------------------|-------------------------|
| overview_overlap | <i>overview_overlap</i> |
|------------------|-------------------------|

Description

Provides an overview of the overlap of two data sets. Cautionary note: This function is currently only preliminary workable and can only capture 2 data sets. We are working on an extension that allows to compare multiple data sets.

Usage

```
overview_overlap(  
  dat1,  
  dat2,  
  dat1_id,  
  dat2_id,  
  dat1_name = "Data set 1",  
  dat2_name = "Data set 2",  
  plot_type = "bar"  
)
```

Arguments

| | |
|------------------------|---|
| <code>dat1</code> | A first data set object |
| <code>dat2</code> | A second data set object |
| <code>dat1_id</code> | Scope (e.g., country codes or individual IDs) of <code>dat1</code> . It is important that both ID variables are exactly the same to generate the perfect match. |
| <code>dat2_id</code> | Scope (e.g., country codes or individual IDs) of <code>dat2</code> . It is important that both ID variables are exactly the same to generate the perfect match. |
| <code>dat1_name</code> | Name of <code>dat1</code> ("Data set 1" is the default) |
| <code>dat2_name</code> | Name of <code>dat2</code> ("Data set 2" is the default) |
| <code>plot_type</code> | Type of plot ("bar" and "venn" are the two options) "venn" relies on the <code>ggvenn</code> function |

Value

A `ggplot2` object (bar chart) that shows the overlap of two data sets.

Examples

```
## Not run:
data(toydata)
toydata2 <- toydata[which(toydata$year > 1992), ]
overview_overlap(
  dat1 = toydata, dat2 = toydata2, dat1_id = ccode,
  dat2_id = ccode
)

## End(Not run)
```

 overview_plot

 overview_plot

Description

This function plots a ggplot to visualize the distribution of scope objects across the time frame.

Usage

```
overview_plot(
  dat,
  id,
  time,
  xaxis = "Time frame",
  yaxis = "Sample",
  asc = TRUE,
  color,
  dot_size = 2
)
```

Arguments

| | |
|----------|---|
| dat | Your data set |
| id | Your scope (e.g., country codes or individual IDs). If the id variable contains NAs, they will not be included in the plot. |
| time | Your time (e.g., time periods given by years, months, ...) |
| xaxis | Label of the x axis ("Time frame" is default) |
| yaxis | Label of the y axis ("Sample" is default) |
| asc | Sorting the y axis in ascending order ("TRUE" is default) |
| color | Optional argument that defines the color |
| dot_size | Option argument that defines the dot size (default is 2) |

Value

A ggplot figure that presents the sample information visually

Examples

```
data(toydata)
overview_plot(dat = toydata, id = ccode, time = year)
```

```
overview_plot_absolute
      overview_plot_absolute
```

Description

Function used in ‘overview_na’ to plot the absolute share of NA values

Usage

```
overview_plot_absolute(
  dat_result = NULL,
  theme_plot = NULL,
  yaxis = NULL,
  xaxis = NULL
)
```

Arguments

| | |
|-------------------------|----------------------------------|
| <code>dat_result</code> | Data frame |
| <code>theme_plot</code> | Theme for the plot (pre-defined) |
| <code>yaxis</code> | Name for yaxis |
| <code>xaxis</code> | Name for xaxis |

Value

The function returns a ggplot

```
overview_plot_percentage
      overview_plot_percentage
```

Description

Function used in ‘overview_na’ to plot the percentage share of NA values

Usage

```
overview_plot_percentage(
  dat_result = NULL,
  theme_plot = NULL,
  yaxis = NULL,
  xaxis = NULL
)
```

Arguments

| | |
|------------|----------------------------------|
| dat_result | Data frame |
| theme_plot | Theme for the plot (pre-defined) |
| yaxis | Name for yaxis |
| xaxis | Name for xaxis |

Value

The function returns a ggplot

| | |
|----------------|-----------------------|
| overview_print | <i>overview_print</i> |
|----------------|-----------------------|

Description

Produces a 'LaTeX' output for output obtained via overview_tab and overview_crosstab

Usage

```
overview_print(
  obj,
  title = "Time and scope of the sample",
  id = "Sample",
  time = "Time frame",
  crosstab = FALSE,
  cond1 = "Condition 1",
  cond2 = "Condition 2",
  save_out = FALSE,
  path,
  file,
  label = "tab:tab1",
  fontsize
)
```

Arguments

| | |
|----------|--|
| obj | Overview object produced by overview_tab or overview_crosstab |
| title | Caption of the table (default is "Time and scope of the sample") |
| id | The name of the left column (default is "Sample"), will be ignored if crosstab is TRUE |
| time | The name of the right column (default is ("Time frame")), will be ignored if crosstab is TRUE |
| crosstab | Logical argument, if TRUE produces a crosstab output, default is FALSE |
| cond1 | Description for the first condition (character), will be ignored if crosstab is FALSE. This should correspond to the input for cond1 in overview_crosstab |
| cond2 | Description for the second condition (character), will be ignored if crosstab is FALSE. This should correspond to the input for cond2 in overview_crosstab |
| save_out | Optional argument, exports the output table as a .tex file, default is FALSE |
| path | Specifies the path where the output should be saved |
| file | Specifies name and file type (.tex) |
| label | Specifies the label (default is "tab:tab1") |
| fontsize | Specifies the font size (all 'LaTeX' font sizes such as "scriptsize" or "small" work) |

Value

A 'LaTeX' output that can either be copy-pasted in a text document or exported directed as a .tex file

Examples

```
data(toydata)

overview_object <- overview_tab(dat = toydata, id = ccode, time = year)
overview_print(
  obj = overview_object,
  title = "Some nice title",
  crosstab = FALSE
)

overview_ct_object <- overview_crosstab(
  dat = toydata,
  cond1 = gdp,
  cond2 = population,
  threshold1 = 25000,
  threshold2 = 27000,
  id = ccode,
  time = year
)
overview_print(
  obj = overview_ct_object,
  title = "Some nice title for a cross tab",
```

```

  crosstab = TRUE,
  cond1 = "Name of first condition",
  cond2 = "Name of second condition"
)

```

 overview_tab

 overview_tab

Description

Provides an overview table for the time and scope conditions of a data set. If a `data.table` object is provided, the function uses `data.table`'s syntax to perform the evaluation

Usage

```

overview_tab(
  dat,
  id,
  time = list(year = NULL, month = NULL, day = NULL),
  complex_date = FALSE
)

```

Arguments

| | |
|---------------------------|---|
| <code>dat</code> | A data frame or data table object |
| <code>id</code> | Scope (e.g., country codes or individual IDs) |
| <code>time</code> | Time (e.g., time periods given by years, months, ...). There are three options to add a date variable: 1) Time can be a character vector containing one time variable, 2) a time variable following the YYYY-MM-DD format, or 3) or a list containing multiple time variables (<code>'time = list(year = NULL, month = NULL, day = NULL)'</code>). |
| <code>complex_date</code> | Boolean argument identifying if there is a more complex (list-wise) <code>date_time</code> parameter (FALSE is the default) |

Value

A data frame object that contains a summary of a sample that can later be converted to a 'LaTeX' output using `overview_print`

Examples

```

# With version 1 (and also 2):

data(toydata)
output_table <- overview_tab(dat = toydata, id = ccode, time = year)

# With version 3:
overview_tab(dat = toydata, id = ccode, time = list(

```

```

year = toydata$year,
month = toydata$month, day = toydata$day
), complex_date = TRUE)

```

```

overview_tab_df      overview_tab_df

```

Description

Internal function that calculates the ‘overview_tab’ for data.frame objects

Usage

```
overview_tab_df(dat2 = NULL, dat = NULL, id = NULL, time = NULL)
```

Arguments

| | |
|------|--|
| dat2 | Your data set |
| dat | Your data set |
| id | Scope (e.g., country codes or individual IDs) |
| time | Time (e.g., time periods given by years, months, ...). There are three options to add a date variable: 1) Time can be a character vector containing one time variable, 2) a time variable following the YYYY-MM-DD format, or 3) or a list containing multiple time variables (‘time = list(year = NULL, month = NULL, day = NULL)’). |

Value

A data.frame

```

overview_tab_dt      overview_tab_dt

```

Description

Internal function that calculates the ‘overview_tab’ for data.table objects

Usage

```
overview_tab_dt(dat = NULL, id = NULL, time = NULL, col_names = NULL)
```


Arguments

| | |
|-----------|--|
| dat | Your data set |
| id | Scope (e.g., country codes or individual IDs) |
| time | Time (e.g., time periods given by years, months, ...). There are three options to add a date variable: 1) Time can be a character vector containing one time variable, 2) a time variable following the YYYY-MM-DD format, or 3) or a list containing multiple time variables ('time = list(year = NULL, month = NULL, day = NULL)'). |
| col_names | The column names (containing id and time) |

Value

A data.table

| | |
|-----------------|------------------------|
| theme_heat_plot | <i>theme_heat_plot</i> |
|-----------------|------------------------|

Description

Defines the theme for the 'overview_heat' plot function

Usage

```
theme_heat_plot()
```

Value

A theme for the 'overview_heat' plot

| | |
|---------------|----------------------|
| theme_na_plot | <i>theme_na_plot</i> |
|---------------|----------------------|

Description

Defines the theme for the 'overview_na' plot function

Usage

```
theme_na_plot()
```

Value

A theme for the 'overview_na' plot

`toydata`*Cross-sectional data for countries*

Description

Small, artificially generated toy data set that comes in a cross-sectional format where the unit of analysis is either country-year or country-year-month. It provides artificial information for five countries (Angola, Benin, France, Rwanda, and the UK) for a time span from 1990 to 1999 to illustrate the use of the package.

Usage

```
data(toydata)
```

Format

An object of class "data.frame"

ccode ISO3 country code (as character) for the countries in the sample (Angola, Benin, France, Rwanda, and UK)

year A value between 1990 and 1999

month An abbreviation (MMM) for month (character)

gpd A fake value for GDP (randomly generated)

population A fake value for population (randomly generated)

References

This data set was artificially created for the overviewR package.

Examples

```
data(toydata)
head(toydata)
```

Index

* datasets

- toydata, [18](#)
- .overview_heat, [2](#)
- .overview_tab, [3](#)

- calculate_share_non_row_wise, [4](#)
- calculate_share_row_wise, [4](#)

- find_int_runs, [5](#)

- overview_add_na_output, [5](#)
- overview_crossplot, [6](#)
- overview_crosstab, [7](#)
- overview_heat, [8](#)
- overview_na, [9](#)
- overview_overlap, [10](#)
- overview_plot, [11](#)
- overview_plot_absolute, [12](#)
- overview_plot_percentage, [12](#)
- overview_print, [13](#)
- overview_tab, [15](#)
- overview_tab_df, [16](#)
- overview_tab_dt, [16](#)

- theme_heat_plot, [17](#)
- theme_na_plot, [17](#)
- toydata, [18](#)