Package ‘overviewR’

February 15, 2023

Type Package

Title Easily Extracting Information About Your Data

Version 0.0.13

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 7.2.3

NeedsCompilation no

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

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

Repository CRAN

Date/Publication 2023-02-15 07:50:02 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_latex` ............................................. 9
- `overview_na` .................................................. 11
- `overview_overlap` ......................................... 12
- `overview_plot` ............................................. 13
- `overview_plot_absolute` ................................. 14
- `overview_plot_percentage` .......................... 14
- `overview_tab` ............................................... 15
- `overview_tab_df` ........................................ 16
- `overview_tab_dt` ........................................... 17
- `theme_heat_plot` ......................................... 17
- `theme_na_plot` ........................................... 18
- `toydata` ................................................... 18

Index 20

<table>
<thead>
<tr>
<th>.overview_heat</th>
<th>.overview_tab</th>
</tr>
</thead>
</table>

Description

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

Usage

```r
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

dat  The data set
id   The scope (e.g., country codes or individual IDs). The axis is ordered in ascending order by default.
time The time (e.g., time periods given by years, months, ...)
label If TRUE (default), the total number of observations/percentages of observations are displayed. If FALSE, it returns no labels.
perc 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

col_low Hex color code for the lowest value (default is "#dceaf2")
col_high Hex color code for the lowest value (default is "#2A5773")
xaxis Label of your x axis ("Time frame" is default)
yaxis Label of your y axis ("Sample" is default)
theme_plot Previously generated theme
exp_total Expected total number of observations (i.e. maximum) for time unit.
col_names The column names (containing id and time)

Value

A ggplot

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

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)
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

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

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

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
Description

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

Usage

```r
overview_crossplot(
  dat,
  id,
  time,
  cond1,
  cond2,
  threshold1,
  threshold2,
  xaxis = "Condition 1",
  yaxis = "Condition 2",
  label = FALSE,
  color = FALSE,
  dot_size = 2,
  fontsize = 2.5
)
```

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
dot_size: Option argument that defines the dot size (default is 2)
fontsize: If label is TRUE, the fontsize arguments allows to define the text of the labels (the default is 2.5)
Value

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

Examples

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

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

```r
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_latex`
Overview heat

Examples

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

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

```r
overview_heat(
  dat,
  id,
  time,
  perc = FALSE,
  exp_total = NULL,
  xaxis = "Time frame",
  yaxis = "Sample",
  col_low = "#dceaf2",
  col_high = "#2a5773",
  label = TRUE
)
```

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dat</td>
<td>The data set</td>
</tr>
<tr>
<td>id</td>
<td>The scope (e.g., country codes or individual IDs). The axis is ordered in ascending order by default.</td>
</tr>
<tr>
<td>time</td>
<td>The time (e.g., time periods given by years, months, ...)</td>
</tr>
<tr>
<td>perc</td>
<td>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</td>
</tr>
<tr>
<td>exp_total</td>
<td>Expected total number of observations (i.e. maximum) for time unit.</td>
</tr>
<tr>
<td>xaxis</td>
<td>Label of your x axis (&quot;Time frame&quot; is default)</td>
</tr>
</tbody>
</table>
Value

A ggplot figure that presents sample coverage visually

Examples

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

Description

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

Usage

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

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
The name of the right column (default is "Time frame"), will be ignored if `crosstab` is TRUE

Logical argument, if TRUE produces a crosstab output, default is FALSE

Description for the first condition (character), will be ignored if `crosstab` is FALSE. This should correspond to the input for `cond1` in `overview_crosstab`

Description for the second condition (character), will be ignored if `crosstab` is FALSE. This should correspond to the input for `cond2` in `overview_crosstab`

Optional argument, exports the output table as a .tex file, default is FALSE

Specifies the path and file name (.tex) where you store your output

Specifies the label (default is "tab:tab1")

Specifies the font size (all 'LaTeX' font sizes such as "scriptsize" or "small" work)

This argument is deprecated. Please use "file_path" instead and add the full path.

This argument is deprecated. Please use "file_path" instead and add the full path.

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

```r
data(toydata)

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

# overview_object <- overview_tab(dat = toydata, id = ccode, time = year)
overview_latex(
  obj = overview_object,
  title = "Some nice title",
  file_path = "some/path_to/your_output_file.tex"
)

overview_ct_object <- overview_crosstab(
  dat = toydata,
  cond1 = gdp,
  cond2 = population,
  threshold1 = 25000,
  threshold2 = 27000,
  id = ccode,
  time = year
)
overview_latex(
  obj = overview_object,
**overview_na**

```r
obj = overview_ct_object,
title = "Some nice title for a cross tab",
crosstab = TRUE,
cond1 = "Name of first condition",
cond2 = "Name of second condition"
```

Description

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

Usage

```r
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

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

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

**Arguments**

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

**Value**

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

**Examples**

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

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

### Usage

```r
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

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

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

dat   A data frame or data table object
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 YYYYY-MM-DD format, or 3) or a list containing multiple time variables (**time = list(year = NULL, month = NULL, day = NULL)**).
complex_date  Boolean argument identifying if there is a more complex (list-wise) date_time 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_latex
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)

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) a list containing multiple time variables (`time = list(year = NULL, month = NULL, day = NULL)`).

Value

A data.frame
**overview_tab_dt**

**Description**

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

**Usage**

```r
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) 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**

**Description**

Defines the theme for the ‘overview_heat’ plot function

**Usage**

```r
theme_heat_plot()
```

**Value**

A theme for the ‘overview_heat’ plot
theme_na_plot

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"

code 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)
gdp 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_latex, 9
  overview_na, 11
  overview_overlap, 12
  overview_plot, 13
  overview_plot_absolute, 14
  overview_plot_percentage, 14
  overview_tab, 15
  overview_tab_df, 16
  overview_tab_dt, 17

  theme_heat_plot, 17
  theme_na_plot, 18
  toydata, 18