

Package ‘xray’

December 8, 2017

Type Package

Title X Ray Vision on your Datasets

Version 0.2

URL <https://github.com/sicarul/xray/>

BugReports <https://github.com/sicarul/xray/issues>

Depends R (>= 3.4.0)

Imports dplyr (>= 0.7.0), scales, foreach, ggplot2, grid, lubridate

Description Tools to analyze datasets previous to any statistical modeling.

Has various functions designed to find inconsistencies and understanding the distribution of the data.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 6.0.1

NeedsCompilation no

Author Pablo Seibelt [aut, cre]

Maintainer Pablo Seibelt <pabloseibelt@sicarul.com>

Repository CRAN

Date/Publication 2017-12-08 05:15:59 UTC

R topics documented:

anomalies	2
distributions	3
timebased	3
xray	4

Index	5
--------------	----------

`anomalies`*Analyze a dataset and search for anomalies*

Description

If any anomalous columns are found, they are reported as a warning and returned in a `data.frame`. To interpret the output, we are getting these anomalies:

- NA values: NA
- 0 values: Zero
- Blank strings: Blank
- Infinite numbers: Inf

Usage

```
anomalies(data_analyze, anomaly_threshold = 0.8, distinct_threshold = 2)
```

Arguments

`data_analyze` a data frame or tibble to analyze

`anomaly_threshold`
the minimum percentage of anomalous rows for the column to be problematic

`distinct_threshold`
the minimum amount of distinct values the column has to have to not be problematic, usually you want to keep this at its default value.

Details

All of these value are reported in columns prefixed by q (quantity), indicating the rows with the anomaly, and p (percentage), indicating percent of total rows with the anomaly.

And, also any columns with only one distinct value, which means the column doesn't bring information to the table (If all rows are equal, why bother having that column?). We report the number of distinct values in `qDistinct`.

Examples

```
library(xray)
anomalies(mtcars, anomaly_threshold=0.5)
```

distributions	<i>Analyze each variable and generate a histogram describing it's distribution.</i>
---------------	---

Description

Also returns a table of all numeric variables describind it's percentiles 1, 10, 25, 50 (median), 75, 90 and 99.

Usage

```
distributions(data_analyze, outdir, charts = T)
```

Arguments

data_analyze	a data frame to analyze
outdir	an optional output directory to save the resulting plots as png images
charts	set this to false to avoid generating charts, useful for batch script usage

Examples

```
library(xray)
distributions(mtcars)
```

timebased	<i>Analyze each variable in respect to a time variable</i>
-----------	--

Description

Analyze each variable in respect to a time variable

Usage

```
timebased(data_analyze, date_variable, time_unit = "auto",
          nvals_num_to_cat = 2, outdir)
```

Arguments

data_analyze	a data frame to analyze
date_variable	the variable (length one character vector or bare expression) that will be used to pivot all other variables
time_unit	the time unit to use if not automatically
nvals_num_to_cat	numeric numeric values with this many or fewer distinct values will be treated as categorical
outdir	an optional output directory to save the resulting plots as png images

Examples

```
library(xray)
data(longley)
longley$Year=as.Date(paste0(longley$Year, '-01-01'))
timebased(longley, 'Year')
```

xray

xray *package*

Description

X-Ray - Dataset Analyzer

Index

anomalies, [2](#)

distributions, [3](#)

timebased, [3](#)

xray, [4](#)

xray-package (xray), [4](#)