Package 'dataQualityR'

February 19, 2015

Type Package

Title Performs variable level data quality checks and generates summary statistics

Version 1.0

Date 2013-09-15

Author Madhav Kumar <madhavkumar2005@gmail.com> and Shreyes Upadhyay

<shreyes.upadhyay@gmail.com>

Maintainer Madhav Kumar <madhavkumar2005@gmail.com>

Description The package performs variable level data quality checks including missing values, unique values, frequency tables, and generates summary statistics

License MIT | file LICENSE

Collate 'checkDataQuality.R'

NeedsCompilation no

Repository CRAN

Date/Publication 2013-09-21 23:03:18

R topics documented:

dataQualityR-package *Performs variable level data quality checks and generates summary statistics*

Description

The package performs variable level data quality checks including missing values, unique values, frequency tables, and generates sumamry statistics

Details

Package:	dataQualityR
Type:	Package
Version:	1.0
Date:	2013-09-15
License:	MIT License

Author(s)

Madhav Kumar and Shreyes Upadhyay

Maintainer: Madhav Kumar <madhavkumar2005@gmail.com>

Examples

```
data(crx)
num.file <- paste(tempdir(), "/dq_num.csv", sep= "")
cat.file <- paste(tempdir(), "/dq_cat.csv", sep= "")
checkDataQuality(data= crx, out.file.num= num.file, out.file.cat= cat.file)
```

checkDataQuality *checkDataQuality*

Description

The function takes in a data frame object, runs data quality checks on each variable, generates summary statistics, and outputs two csv files containing the data quality report – one for numeric variables and the other for categorical variables

Usage

```
checkDataQuality(data,
 out.file.num,
 out.file.cat,
 numeric.cutoff = -1)
```

Arguments

data	An object of class data.frame
out.file.num	Filename for saving data quality report of numeric variables
out.file.cat	Filename for saving data quality report of categoric variables
numeric.cutoff	The minimum number of unique values needed for a numeric variable to be treated as continous. This feature is included to account for binary or multi-category variables, with small number of unique values, which are stored as numeric. Default is -1 which does not place any cut-off and all numeric variables are treated as continuous

Value

Returns csv files stored directly on disk

Author(s)

Madhav Kumar and Shreyes Upadhyay

Examples

```
data(crx)
num.file <- paste(tempdir(), "/dq_num.csv", sep= "")
cat.file <- paste(tempdir(), "/dq_cat.csv", sep= "")
checkDataQuality(data= crx, out.file.num= num.file, out.file.cat= cat.file)
```

---

crx *Sample data.frame object*

Description

Multi-variate data set with information on credit card approvals. The data set contains numeric and categorical variables with some missing values The variable names and values have been changed to meaningless symbols to protect confidentiality of the data.

Usage

```
data(crx)
```

Format

A data frame with 690 observations with the following variables and their types.

`V1` b, a

`V2` continuous

`V3` continuous

`V4` u, y, l, t

`V5` g, p, gg

`V6` c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff

`V7` v, h, bb, j, n, z, dd, ff, o

`V8` continuous

`V9` t, f

`V10` t, f

`V11` continuous

`V12` t, f

`V13` g, p, s

`V14` continuous

`V15` continuous

`V16` 0, 1

Details

There are no more details required

Source

http://archive.ics.uci.edu/ml/datasets/Credit+Approval

References

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Examples

```
data(crx)
```

Index