**Version:** 1.0

**Date:** 2016-12-20

**Title:** clean()

**Author:**  Jeremy Guinta

**Depends:** gdata

January 19, 2017

**Description:** Performs standard cleaning steps on one or more data frames. If more than one data frame is provided, then the function will append all provided data frames into a single object. All objects must be of the same structure.

R topics documented:

Contents

[Description 2](#_Toc470186427)

[Usage 2](#_Toc470186428)

[Arguments 2](#_Toc470186429)

[Example 3](#_Toc470186430)

clean() Standard data frame cleaning

# Description

This process standardizes the data frame cleaning process by standardizing field names, field contents, converting dates and date times and replacing specified values with NAs. The following cleaning steps are performed:

1. trim, lowercase, and remove all punctuation from field names
2. convert all factor variables into character
3. trim and lowercase all character variables
4. replaces all blank values in every field with NA

If specified

1. convert dates into R formatted dates (using as.Date)
2. convert date times into R formatted date times (using as.POSIXct)

# Usage

clean(dta=c(), datefmt=(), datecol=c(), datetimefmt=c(), datetimecol=c() , na.strings=c(“”) )

# Arguments

dta a character vector of the data frame names in memory. If more than one is specified, than the data frames will be combined.

datefmt a character vector of the format of each date field (e.g. “%m/%d/%Y”)

datecol a character vector of the variable names from dta that are date fields. The order of these should be the same order as date formats in datefmt.

datetimefmt a character vector of the format of each datetime field (e.g. “%m/%d/%Y %H:%M:%S”)

datetimecol a character vector of the variable names from dta that are date time fields. The order of these should be the same order as date formats in datetimefmt.

na.stings a character vector of values to replace with NA values in the dataframe. Defaults to “”.

# Example

#Create Data

cll1<-data.frame(

DATE=c("5/11/2010","5/11/2010","5/11/2010","5/11/2010","5/11/2010","5/11/2010","5/11/2010","5/11/2010","5/11/2010","5/11/2010")

,TIME=c("10:58:38","10:58:38","10:58:38","10:59:05","10:59:05","10:59:05","11:00:08","11:00:08","11:00:35", "")

,datetime=c("5/11/2010 10:58:38","5/11/2010 10:58:38","5/11/2010 10:58:38","5/11/2010 10:59:05","5/11/2010 10:59:05","5/11/2010 10:59:05","5/11/2010 11:00:08","5/11/2010 11:00:08","5/11/2010 11:00:08","5/11/2010 11:00:35")

,ANI=c("8003758211","8003758211","8003758211","8003758211","8003758211","8003758211","8003758211","8003758211","8003758211","8003758211")

,DISPOSITION\_CODE=c("No Answer","","Operator Intercept","Left Message","Hangup","No Answer","No Answer","No Answer","No Answer","Dial Error")

,CALL.TYPE=c("Outbound","Outbound","Outbound","Outbound","Outbound"," ","Outbound","Outbound","Outbound","Outbound")

)

cll2<-data.frame(

DATE=c("6/1/2010","6/1/2010","6/1/2010","6/1/2010","6/1/2010","6/1/2010","6/1/2010","6/1/2010","6/1/2010","6/1/2010")

,TIME=c("9:02:09","9:05:18","9:06:58","9:13:19","9:13:19","9:13:20","9:13:20","9:13:20","9:13:20","9:13:42")

,datetime=c("6/1/2010 9:02:09","6/1/2010 9:05:18","6/1/2010 9:06:58","6/1/2010 9:13:19","6/1/2010 9:13:19","6/1/2010 9:13:20","6/1/2010 9:13:20","6/1/2010 9:13:20","6/1/2010 9:13:20","6/1/2010 9:13:42")

,ANI=c("9492418026","9492418026","9492418026","9492418026","9492418026","9492418026","9492418026","9492418026","9492418026","9492418026")

,DISPOSITION\_CODE=c("Call Back","Left Message","Not Interested","No Answer","No Answer","N/A","No Answer","Left Message","No Answer","No Answer")

,CALL.TYPE=c("Agent","Agent","Agent","Outbound","Outbound","Outbound","Outbound","Outbound","Outbound",".")

)

head(cll1)

# DATE TIME datetime ANI DISPOSITION\_CODE CALL.TYPE

#1 5/11/2010 10:58:38 5/11/2010 10:58:38 8003758211 No Answer Outbound

#2 5/11/2010 10:58:38 5/11/2010 10:58:38 8003758211 Outbound

#3 5/11/2010 10:58:38 5/11/2010 10:58:38 8003758211 Operator Intercept Outbound

#4 5/11/2010 10:59:05 5/11/2010 10:59:05 8003758211 Left Message Outbound

#5 5/11/2010 10:59:05 5/11/2010 10:59:05 8003758211 Hangup Outbound

#6 5/11/2010 10:59:05 5/11/2010 10:59:05 8003758211 No Answer

head(cll2)

# DATE TIME datetime ANI DISPOSITION\_CODE CALL.TYPE

#1 6/1/2010 9:02:09 6/1/2010 9:02:09 9492418026 Call Back Agent

#2 6/1/2010 9:05:18 6/1/2010 9:05:18 9492418026 Left Message Agent

#3 6/1/2010 9:06:58 6/1/2010 9:06:58 9492418026 Not Interested Agent

#4 6/1/2010 9:13:19 6/1/2010 9:13:19 9492418026 No Answer Outbound

#5 6/1/2010 9:13:19 6/1/2010 9:13:19 9492418026 No Answer Outbound

#6 6/1/2010 9:13:20 6/1/2010 9:13:20 9492418026 N/A Outbound

#Clean

cll\_cln<-clean(dta=c("cll1", "cll2"), datefmt=c("%m/%d/%Y"),

datecol=c("DATE"), datetimefmt=c("%m/%d/%Y %H:%M:%S"), datetimecol=c("datetime"), na.strings=c(".",""," ","N/A","n/a"))

# [1] "1. Processing cll1..."

# [1] " --Standardizing fields..."

# [1] " --Standardizing field names..."

# [1] " --Removing blanks and replacing with NA..."

# [1] " --Converting date values..."

# [1] " --Converting datetime values"

# [1] "2. Processing cll2..."

# [1] " --Standardizing fields..."

# [1] " --Standardizing field names..."

# [1] " --Removing blanks and replacing with NA..."

# [1] " --Converting date values..."

# [1] " --Converting datetime values"

# [1] " --Combining data..."

# [1] "3. Summary of Output..."

# File Observation Count

# 1 cll1 10

# 2 cll2 10

# 3 Total 20

# 'data.frame': 20 obs. of 7 variables:

# $ date : Date, format: "2010-05-11" "2010-05-11" "2010-05-11" ...

# $ time : chr "10:58:38" "10:58:38" "10:58:38" "10:59:05" ...

# $ datetime : POSIXct, format: "2010-05-11 10:58:38" "2010-05-11 10:58:38" "2010-05-11 10:58:38" ...

# $ ani : chr "8003758211" "8003758211" "8003758211" "8003758211" ...

# $ dispositioncode: chr "no answer" NA "operator intercept" "left message" ...

# $ calltype : chr "outbound" "outbound" "outbound" "outbound" ...

# $ src : chr "cll1" "cll1" "cll1" "cll1" ...

# NULL

head(cll\_cln)

# date time datetime ani dispositioncode calltype src

#1 2010-05-11 10:58:38 2010-05-11 10:58:38 8003758211 no answer outbound cll1

#2 2010-05-11 10:58:38 2010-05-11 10:58:38 8003758211 <NA> outbound cll1

#3 2010-05-11 10:58:38 2010-05-11 10:58:38 8003758211 operator intercept outbound cll1

#4 2010-05-11 10:59:05 2010-05-11 10:59:05 8003758211 left message outbound cll1

#5 2010-05-11 10:59:05 2010-05-11 10:59:05 8003758211 hangup outbound cll1

#6 2010-05-11 10:59:05 2010-05-11 10:59:05 8003758211 no answer <NA> cll1