Metriplica

RaÃºl

5 de noviembre de 2019

```
chooseCRANmirror(graphics=FALSE, ind=1)
knitr::opts_chunk$set(echo = TRUE)
```

Proyecto : Conversion Rate

Data Wrangling

Instalamos paquetes y leemos el Dataset

```
install.packages("pacman")
```

```
## package 'pacman' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raul9\AppData\Local\Temp\RtmpstFRQR\downloaded_packages
```

```
pacman::p_load(readr, lubridate, dplyr, ggplot2, plotly)

df <- read_csv("~/Raúl Vázquez/Personal/Test Metriplica/results_df.csv")
```

Comprobamos si existen NA's

```
sapply(df, function(x) sum(is.na(x)))
```

```
##            date channelGrouping  userAgeBracket        userType
##               0               0               0               0
##        sessions    transactions
##               0               0
```

Entendemos los datos

```
str(df)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 48852 obs. of  6 variables:
##  $ date           : num  20190101 20190101 20190101 20190101 20190101 ...
##  $ channelGrouping: chr  "(Other)" "(Other)" "(Other)" "(Other)" ...
##  $ userAgeBracket : chr  "25-34" "25-34" "35-44" "35-44" ...
##  $ userType       : chr  "New Visitor" "Returning Visitor" "New Visitor" "Returning Visitor" ...
##  $ sessions       : num  17 30 11 71 11 10 48 90 288 651 ...
##  $ transactions   : num  0 2 0 2 0 0 0 2 3 14 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_double(),
##   ..   channelGrouping = col_character(),
##   ..   userAgeBracket = col_character(),
##   ..   userType = col_character(),
##   ..   sessions = col_double(),
##   ..   transactions = col_double()
##   .. )
```

```r
summary(df)
```

```
##       date          channelGrouping    userAgeBracket
##  Min.   :20190101   Length:48852       Length:48852
##  1st Qu.:20190315   Class :character   Class :character
##  Median :20190531   Mode  :character   Mode  :character
##  Mean   :20190560
##  3rd Qu.:20190812
##  Max.   :20191027
##    userType            sessions       transactions
##  Length:48852       Min.   :   6.0   Min.   :  0.000
##  Class :character   1st Qu.:  44.0   1st Qu.:  0.000
##  Mode  :character   Median : 147.0   Median :  1.000
##                     Mean   : 334.6   Mean   :  3.676
##                     3rd Qu.: 395.0   3rd Qu.:  3.000
##                     Max.   :6578.0   Max.   :150.000
```

Limpieza de datos general

```r
df$date <- lubridate:: ymd(df$date) #Fecha

df$channelGrouping <- sub("(^[^-]+)-.*", "\\1", df$channelGrouping) # Nos quedamos con el primer grupo

col_names <- c("channelGrouping","userAgeBracket","userType")
df[,col_names] <- lapply(df[,col_names] , factor) # Convertimos varias columnas a factor
```