Análisis de datos: Netfliz Prize Program

Laura Basalo Tur, Camila Pérez

20/11/2020

Contents

Introducción

TODO: Comentar…

Carga de datos

TODO: Comentar…

```
#Se importan 'n' observaciones del conjunto de datos 'Netflix Prize data'
n = 10000
dataframe = read_tsv("data/combined_data_1.txt", col_names=FALSE, n_max = n)

#Se importa el dataframe de los datos de las películas
n_mov = 10
df_movies = read_csv("data/movie_titles.csv", col_names = FALSE, n_max = n_mov)
```

Limpieza de datos

TODO: Comentar…

```
#Se asigna una posición a cada observación para posteriormente indicar el id de película de cada una de
dataframe = dataframe %>%
            mutate(row=row_number())
rows = grep(":", dataframe$X1)
rows_ID = dataframe %>%
         filter( row %in% rows )
IDs = unique(rows_ID$X1)
reps = diff(c(rows_ID$row,max(dataframe$row)+1))

df = dataframe %>%
     mutate(ID1 = rep(rows_ID$X1,times= reps)) %>%
     filter(!(row %in% rows_ID))

#Se definen las columnas del dataframe
df = df %>%
     separate(X1,into = c("ID_film","Score","Data"), sep = ",") %>%
     separate(Data,into = c("Year","Month","Day"), sep = "-") %>%
     na.omit(df) %>%
     mutate(row=row_number())
```

```
## Warning: Expected 3 pieces. Missing pieces filled with ‘NA‘ in 8 rows [1, 549,
## 695, 2708, 2851, 3992, 5012, 5106].
```

```
#Se eliminan las variables auxiliares
rm(dataframe,rows,rows_ID,IDs,reps)

#Se ordenan las posiciones de las columnas y se indican su nuevo nombre
df = df[, c(6, 7, 1, 2, 3,4,5)]
df = df %>%
     rename(
      MovieID = ID1,
      CustomerID = ID_film,
      Rating = Score,
      Idx = row
     )

#Se elimina el caracter ':' de la columna del MovieID
df$MovieID = unlist(strsplit(df$MovieID , split = ':', fixed=FALSE))
```

Conociendo el dataframe y sus variables

TODO: Comentar…

First 10 rows

```r
# First 10 rows
knitr::kable(head(df))
```

Idx	MovieID	CustomerID	Rating	Year	Month	Day
1	1	1488844	3	2005	09	06
2	1	822109	5	2005	05	13
3	1	885013	4	2005	10	19
4	1	30878	4	2005	12	26
5	1	823519	3	2004	05	03
6	1	893988	3	2005	11	17

Last 10 rows

```r
# Last 10 rows
knitr::kable(tail(df))
```

Idx	MovieID	CustomerID	Rating	Year	Month	Day
9987	8	809074	4	2005	05	09
9988	8	2142408	1	2005	05	10
9989	8	2231367	3	2005	05	10
9990	8	1304395	4	2005	05	10
9991	8	1468830	3	2005	05	13
9992	8	1369078	1	2005	05	15

Summary

```r
# Summary
knitr::kable(summary(df))
```

	Idx	MovieID	CustomerID	Rating	Year	Month	Day
	Min. : 1	Length:9992	Length:9992	Length:9992	Length:9992	Length:9992	Length:9992
	1st Qu.:2499	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character
	Median :4996	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character
	Mean :4996	NA	NA	NA	NA	NA	NA
	3rd Qu.:7494	NA	NA	NA	NA	NA	NA
	Max. :9992	NA	NA	NA	NA	NA	NA

Structure

```r
# Structure
str(df)
```

```
## tibble [9,992 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Idx       : int [1:9992] 1 2 3 4 5 6 7 8 9 10 ...
##  $ MovieID   : chr [1:9992] "1" "1" "1" "1" ...
##  $ CustomerID: chr [1:9992] "1488844" "822109" "885013" "30878" ...
##  $ Rating    : chr [1:9992] "3" "5" "4" "4" ...
##  $ Year      : chr [1:9992] "2005" "2005" "2005" "2005" ...
##  $ Month     : chr [1:9992] "09" "05" "10" "12" ...
##  $ Day       : chr [1:9992] "06" "13" "19" "26" ...
##  - attr(*, "na.action")= ’omit’ Named int [1:8] 1 549 695 2708 2851 3992 5012 5106
##   ..- attr(*, "names")= chr [1:8] "1" "549" "695" "2708" ...
```

```r
str(df_movies)
```

```
## tibble [10 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ X1: num [1:10] 1 2 3 4 5 6 7 8 9 10
##  $ X2: num [1:10] 2003 2004 1997 1994 2004 ...
##  $ X3: chr [1:10] "Dinosaur Planet" "Isle of Man TT 2004 Review" "Character" "Paula Abdul’s Get Up &
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   X2 = col_double(),
##   ..   X3 = col_character()
##   .. )
```

```r
#Se transforman la variable 'Rating' a númerica
df$Rating = as.numeric(df$Rating)
```