Netflix enunciado práctica MD 20_21. GIN2

Grupo y nombre de cada usuario

22 diciembre, 2020

Contenidos

1 Taller evaluable en grupos datos netflix

Enlace a estos datos de Netflix Generad un proyecto nuevo. Bajad lo datos de netflix a un carpeta/directorio que se llame `netflix` y dentro de `netflix` crear una carpeta/directorio que se llame `model_netflix`.

Podéis (tenéis) que utilizar las ayudas del taller de estos datos.

1.1 Instrucciones

- Entregad en grupos de 4 ó 5 estudiantes.
- Se puede hacer con R o python.
- Hay que entregar el Rmd/notebook junto con su salida en html/pdf
- Máxima longitud: 10 páginas en pdf para cada apartado.
- Hay que cuidar la presentación, ortografía y redacción.
- Fecha entrega 20 de enero a las 23:55 2021.

1.2 Cuestión 1: Contexto del problema y modelo de datos (50%)

Como el problema es de datos masivos vamos cada grupo hará un muestreo de los 4 ficheros. Para facilitar la labor os proporcionamos un fichero en el que de cada película

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
filas_ID_combined_all=read_csv("data/filas_ID_combined_all.txt")
```

```
##
## -- Column specification --------------------------------------------------------
```

```
## cols(
##   X1 = col_character(),
##   fila = col_double(),
##   ID = col_double(),
##   fila_final = col_double(),
##   data = col_double()
## )
```

```
glimpse(filas_ID_combined_all)
```

```
## Rows: 17,770
## Columns: 5
## $ X1         <chr> "1:", "2:", "3:", "4:", "5:", "6:", "7:", "8:", "9:", "1...
## $ fila       <dbl> 1, 549, 695, 2708, 2851, 3992, 5012, 5106, 20017, 20113,...
## $ ID         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ fila_final <dbl> 548, 694, 2707, 2850, 3991, 5011, 5105, 20016, 20112, 20...
## $ data       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
```

En total hay 17750 películas con ID entero de 1 a 17750

- La columna `X1` es de tipo character contiene el identificador original en el fichero `1:` un entero seguido de ":"
- La columna `fila` es de tipo integer contiene el número de fila que contiene el identificador de la película en el fichero `combinen_data_x.txt` el valor de x es viene determinado por la columna `file_num`.
- La columna `ID` es de tipo integer el identificado de la película sin `:`
- La columna `fila_final` es de tipo integer contiene el número de fila que contiene ella última entra de la película `ID`
- La columna `file_num` es de tipo integer contiene un entero de 1 a 4 que indica si los datos de esa película están en el fichero `combinen_data_1.txt`, `combinen_data_2.txt`, `combinen_data_3.txt` o `combinen_data_4.txt`

Cada fichero contiene una cierta cantidad de películas

```
table(filas_ID_combined_all$file_num)
```

```
## Warning: Unknown or uninitialised column: `file_num`.
```

```
## < table of extent 0 >
```

ATENCIÓN Selecciona de las 1 a 17750 250 películas Semilla de grupo concatenar los dos últimos dígitos numéricos de vuestro DNI o tarjeta de residente

```
# dos últimos dígitos 51 52 53 de cada miembro ordenados de menor a mayor 515253
#  y si hay ceros  segid este ejemplo
#  si las terminación del dni son 01 02 03 ordenadas de menor a mayor

set.seed(01003)
runif(4)
```

```
## [1] 0.4720480 0.9390508 0.1033403 0.8906890
```

```
muestra_grupo=sample(1:12000,250,replace = FALSE)
head(muestra_grupo)
```

```
## [1] 7897 5036 3874 2263 4340 9851
```

Tenéis que localizar en el fichero `filas_ID_combined_all`que películas son en que fichero de `combined_data_?.txt` están y las lineas que tenéis que leer.

1. Contextualiza a partir de la información de Kaggle los datos de que disponemos. Qué datos contiene cada uno de los ficheros y para que´nos pueden resultar importantes para Netflix.

2. Leer cada película del fichero correspondiente y guardarlas, adecuadamente, en un mismo fichero para futuro tratamiento.
3. Construir el modelo de datos siguiendo las indicaciones de la taller ejemplo de netflix y generar la tibble netflix.
4. Leer el fichero de nombres y año y film que es `movie_titles.csv` y hacer un `inner_join` para disponer del título y año de estreno de cada película.
5. Guardar los datos procesado en un fichero csv, con el formato adecuado para utilizarlo en el siguiente apartado.

1.3 Cuestión 2: Análisis exploratorio (EDA). (50%)

En las siguientes preguntas aplica todo lo que hemos visto acerca de la documentación en el EDA: Título de gráficos, etiquetas de los ejes, coloreado con información, leyendas, tablas bien presentadas (knitr)...

1. Justifica para cada una de las variables de la tabla anterior el tipo de dato que mejor se ajusta a cada una de ellas: numérico, ordinal, categórico....
2. Estudia la distribución del numero de películas estrenadas por año. Realiza un gráfico de muestre esta distribución haciendo los ajustes necesarios (agrupaciones, cambios de escala, transformaciones...)
3. Investiga la librería `lubridate` (o la que consideréis para manipulación de datos) y utilízala para transformar la columna de la fecha de la valoración en varias columnas por ejemplo `year`,`month`, `week`, `day_of_week`.
4. Genera un tabla que para cada película nos dé el número total de valoraciones, la suma de las valoraciones, la media las valoraciones, y otras estadísticos de interés (desviación típica, moda , mediana).
5. De las cinco películas con más número total de valoraciones, compara sus estadísticos y distribuciones (histogramas, boxplot, violin plot,...)
6. Investiga la distribución de valoraciones por día de la semana y por mes.¿Qué meses y días de la semana se valoran más películas en netflix?
7. Genera una tabla agrupada por película y año del número de valoraciones. Representa la tabla gráficamente para de las 10 películas con mayor número de valoraciones .
8. Distribución del `score` promedio por año de las 10 películas con mayor número de valoraciones.
9. Realiza algún gráfico o estudió de estadísticos adicional que consideres informativo en base al análisis exploratorio anterior.