Complementos

Sesión 5

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Datos de anime

El archivo `anime.csv` contiene información sobre distintos anime:

- Código identificador del anime
- Nombre del anime
- Género del anime
- Tipo de anime (OVA, Película, TV, etcétera)
- Número de episodios
- Puntos de Rating que tuvo el anime
- Cantidad de seguidores del anime

Práctica 1: Análisis exploratorio

a) Utilizando la función `import()` del paquete *rio* cargue los datos en R. Analice la estructura de los datos. ¿Hay alguna variable que no tenga el formato correspondiente? ¿Cuál? ¿Por qué? Plantee una manera de solucionar este problema.

b) Verifique que ningún anime se repita. ¿Por qué es importante verificar que no hayan datos duplicados?

c) ¿Cuántos tipos de anime hay? ¿Cuántos anime hay en cada tipo/categoría? ¿Cuál es la mediana de la cantidad de episodios en los tipos de anime *Movie*, *OVA* y *TV*? ¿En qué categoría hay más animes con mayor cantidad de episodios?

d) ¿Cuál es el anime con mayor y cuál el con menor rating?

e) Entregue los 10 animes con mayor rating de la lista, ¿a qué género pertenecían estos anime? ¿Qué términos suelen repetirse en estos géneros de anime?

Práctica 1

Respuestas práctica 1

Respuesta a)

```
library(rio) #Cargue la librería rio (previamente debe ser instalada)

anime<-import(file.choose()) #Guardo los datos en un objeto llamado anime

View(anime) #Vista previa de los datos

#a) Analizar la estructura de los datos

nrow(anime) #Numero de filas
[1] 12294

ncol(anime) #Numero de columnas
[1] 7

names(anime) #Número de las columnas (variables generalmente)
[1] "anime_id" "name"     "genre"    "type"     "episodes" "rating"   "members"

str(anime) #Formato de las variables
'data.frame': 12294 obs. of  7 variables:
 $ anime_id: int  32281 5114 28977 9253 9969 32935 11061 820 15335 15417 ...
 $ name    : chr  "Kimi no Na wa." "Fullmetal Alchemist: Brotherhood" "GintamaÂ°" "Steins;Gate" ...
 $ genre   : chr  "Drama, Romance, School, Supernatural" "Action, Adventure, Drama, Fantasy, Magic, Military, Shounen" "Action, Comedy, Hi
 $ type    : chr  "Movie" "TV" "TV" "TV" ...
 $ episodes: num  1 64 51 24 51 10 148 110 1 13 ...
 $ rating  : num  9.37 9.26 9.25 9.17 9.16 9.15 9.13 9.11 9.1 9.11 ...
 $ members : int  200630 793665 114262 673572 151266 93351 425855 80679 72534 81109 ...
```

Respuesta a)

```
#La variable episodes corresponde a la cantidad de episodios pero su formato
#no es numérico:

class(anime$episodes) #Es de tipo character, ¿por qué?
[1] "character"

#Veamos qué valores toma esta variable:

unique(anime$episodes)

#Notar que un valor que toma esta variable es "Unknown" por lo tanto, por dicha
#razón, el vector de episodios es leído en formato caracter. ¿Qué hacer?

#Veamos cuántos registros toman este valor:

length(which(anime$episodes=="Unknown")) #Son 340 anime con este problema
[1] 340

#Una opción que suele utilizarse bastante es definir estos casos con valor numérico en particular
#por ejemplo, podría ser 0, donde 0 NO indicaría que el anime tuvo 0 episodios
#sino que, no fue especificado. Hay que tener cuidado de no utilizar estos ceros
# al calcular estadísticas

anime$episodes[which(anime$episodes=="Unknown")]<-0 #Se les asigna 0 a estos casos

anime$episodes<-as.numeric(anime$episodes) #Se le aplica el formato numerico

class(anime$episodes) #Ahora sí posee valor numerico! :)
[1] "numeric"
```

Respuesta b)

```
## b) Verificar que ningún anime se repita
# Notar que el nombre del anime o el anime_id funcionan como ID o identificador único
# por anime, hay que verificar que es unico cada ID, es decir, que solo se repite una vez

#FORMA 1:

nrow(anime) #Numero de filas, numero de registros
[1] 12294

length(unique(anime$anime_id))
[1] 12294

nrow(anime)==length(unique(anime$anime_id)) #Si son iguales ningún ID se repite
[1] TRUE

#Por lo tanto ningún anime está duplicado

#FORMA 2:

table(anime$anime_id) #Indica cuántas veces se repite cada ID

table(table(anime$anime_id)) #Indica un resumen de cuántas veces se repitió cada ID

    1
12294

#Si sólo aparece el número 1 entonces cada ID aparece sólo una vez

#Por lo tanto ningún anime está duplicado
```

¿Por qué es importante verificar que no hayan datos duplicados?

Porque esto sesgaría los resultados de las estadísticas obtenidas. Además no tiene sentido replicar cierto registro o información varias veces. Es súper importante realizar análisis de los ID o identificadores, si estos se repiten varias veces, podríamos estar en presencia de datos de seguimiento. Por ejemplo, seguimiento de pacientes o personas (se poseen distintos registros por persona).

Respuesta c)

```
#c) ¿Cuántos tipos de anime hay?

unique(anime$type) #Valores que toma la variable tipo de anime
 "Movie"   "TV"      "OVA"     "Special" "Music"   "ONA"     ""

#Notar que el último caso "" se denomina dato faltante! Sucede cuando no se especificó
#dicha información

#¿Cuántos anime hay en cada categoría:

table(anime$type)

          Movie   Music     ONA     OVA Special      TV
     25    2348     488     659    3311    1676    3787

#Hay 25 anime cuyo tipo es un dato faltante o no especificado.
# 3787 anime eran de tipo TV
# 3311 anime eran de tipo OVA
# 2348 anime de tipo pelicula, etcétera...

# Mediana de la cantidad de episodios en las categorías
# Movie, OVA y TV?

#Aplicamos filtros

median(anime$episodes[which(anime$type=="Movie")])
[1] 1

median(anime$episodes[which(anime$type=="OVA")])
[1] 2

median(anime$episodes[which(anime$type=="TV")])
[1] 24
```

Respuesta c)

```
#Si se quiere mostrar en una tabla:
data.frame(Tipo=c("Movie", "OVA", "TV"), Mediana=c(median(anime$episodes[which(anime$type=="Movie")]),
median(anime$episodes[which(anime$type=="OVA")]),
median(anime$episodes[which(anime$type=="TV")])))
   Tipo Mediana
1 Movie       1
2   OVA       2
3    TV      24

#En el tipo TV se esperarían más animes con más episodios
#Esto tiene sentido puesto que las categorías Movie y OVAS suelen ser como
#episodios "especiales" del anime, son episodios muy ocasionales
```

Respuesta d)

```
#d) ¿Cuál es el anime con mayor y cuál el con menor rating?

anime$name[which.max(anime$rating)]
[1] "Taka no Tsume 8: Yoshida-kun no X-Files"

anime$name[which.min(anime$rating)]
[1] "Platonic Chain: Ansatsu Jikkouchuu"
```

Respuesta e)

```
#e) Entregue los 10 animes con mayor rating de la lista

order(anime$rating, decreasing=TRUE) #Ordena de manera decreciente los anime por rating

order(anime$rating, decreasing=TRUE)[1:10] #Selecciona los 10 primeros
[1] 10465 10401  9596     1  9079     2     3 10787     4     5

anime[order(anime$rating, decreasing=TRUE)[1:10], c("name", "genre")]

#Términos que se repiten en estos géneros de anime:

#Action, Drama, Comedy, Historical, Shounen, Sci-Fi
```