

Laboratorio 3

Francisco Acuña

2024-09-04

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

Carga de Datos

```
actors <- read.csv("actors.csv")  
directors_genres <- read.csv("directors_genres.csv")  
directors <- read.csv("directors.csv")  
movies_directors <- read.csv("movies_directors.csv")  
movies <- read.csv("movies.csv")  
roles <- read.csv("roles.csv")
```

1. Base de datos:

a. ¿Cuál es el número total de películas registradas?

```
total_peliculas <- nrow(movies)  
total_peliculas
```

```
## [1] 388269
```

b. ¿Cuántos directores únicos hay en la base de datos?

```
total_directores <- directors %>%
  distinct(id) %>%
  nrow()

total_directores
```

```
## [1] 86880
```

2. ¿Cuál es el promedio de géneros por cada director?

```
promedio_generos_director <- directors_genres %>%
  group_by(director_id) %>%
  summarise(total_generos = n_distinct(genre)) %>%
  summarise(promedio_generos = mean(total_generos))

promedio_generos_director
```

```
## # A tibble: 1 x 1
##   promedio_generos
##             <dbl>
## 1             2.41
```

3. Generar un informe basado en “Role” que incluya:

a. Número de películas asociadas a cada rol

```
peliculas_por_rol <- roles %>%
  group_by(role) %>%
  summarise(cantidad_peliculas = n_distinct(movie_id))

head(peliculas_por_rol,10)
```

```
## # A tibble: 10 x 2
##   role                                cantidad_peliculas
##   <chr>                                <int>
## 1 ""                                164782
## 2 " (1985)"                            1
## 3 " (1991 reissue only)"              1
## 4 " (episode \"Protest und Theori"    1
## 5 " (episode 4: The Criminal)"         1
## 6 " (episode Målbrott)"               1
## 7 " (episode one)"                    1
## 8 " (episode two)"                    1
## 9 " (segment \"A Boca\)"              1
## 10 " (segment \"A Suspeita\)"          1
```

b. Número de actores distintos por rol

```
actores_por_rol <- roles %>%  
  group_by(role) %>%  
  summarise(total_actores = n_distinct(actor_id))  
  
head(actores_por_rol,10)
```

```
## # A tibble: 10 x 2  
##   role                                total_actores  
##   <chr>                                <int>  
## 1 ""                                304819  
## 2 " (1985)"                            1  
## 3 " (1991 reissue only)"                1  
## 4 " (episode \"Protest und Theori"  
## 5 " (episode 4: The Criminal)"          1  
## 6 " (episode Målbrott)"                3  
## 7 " (episode one)"                     4  
## 8 " (episode two)"                     2  
## 9 " (segment \"A Boca\")"              2  
## 10 " (segment \"A Suspeita\")"          1
```

c. Número de actrices por rol

```
actrices_por_rol <- roles %>%  
  inner_join(actors %>% filter(gender == 'F'), by = c("actor_id" = "id")) %>%  
  group_by(role) %>%  
  summarise(total_actrices = n_distinct(actor_id))  
  
head(actrices_por_rol,10)
```

```
## # A tibble: 10 x 2  
##   role                                total_actrices  
##   <chr>                                <int>  
## 1 ""                                115354  
## 2 " (1991 reissue only)"                1  
## 3 " (episode 4: The Criminal)"          1  
## 4 " (segment \"La voce umana\")"          1  
## 5 " (segment Red Peppers) (segme"  
## 6 "\"Astoria\" Owner"                   1  
## 7 "\"Betsy Ross\""                     1  
## 8 "\"Frank\" Hickson"                   1  
## 9 "\"Fred\" Lincoln"                   1  
## 10 "\"Statue of Liberty\""              1
```

d. Cantidad de directores por rol

```
directores_por_rol <- movies_directors %>%
  inner_join(roles, by = "movie_id") %>%
  group_by(role) %>%
  summarise(total_directores = n_distinct(director_id))
```

```
## Warning in inner_join(., roles, by = "movie_id"): Detected an unexpected many-to-many relationship b
## i Row 2 of 'x' matches multiple rows in 'y'.
## i Row 119372 of 'y' matches multiple rows in 'x'.
## i If a many-to-many relationship is expected, set 'relationship =
## "many-to-many"' to silence this warning.
```

```
head(directores_por_rol,10)
```

```
## # A tibble: 10 x 2
##   role                                total_directores
##   <chr>                                <int>
## 1 ""                                42075
## 2 " (1985)"                          2
## 3 " (episode \"Protest und Theori"      1
## 4 " (episode 4: The Criminal)"        1
## 5 " (episode Målbrott)"              1
## 6 " (episode one)"                  1
## 7 " (episode two)"                  1
## 8 " (segment \"A Boca\")"              1
## 9 " (segment \"A Suspeita\")"          1
## 10 " (segment \"Head Like a Hole\")"    3
```

4. Crear un nuevo informe con la siguiente información:

a. Detalles del director (ID, nombre, apellido)

```
info_directores <- directors %>%
  select(director_id = id, nombre = first_name, apellido = last_name)
head(info_directores,10)
```

```
##   director_id      nombre  apellido
## 1           1        Todd         1
## 2           2        Les 12 Poissons
## 3           3      Lejaren  a'Hiller
## 4           4         Nian         A
## 5           5      Khairiya A-Mansour
## 6           6      Ricardo   A. Solla
## 7           8 Kodanda Rami Reddy     A.
## 8           9    Nageswara Rao     A.
## 9          10         Yuri        A.
## 10          11        Swamy    A.S.A.
```

b. Contar cuántas películas ha dirigido cada director

```
peliculas_dirigidas <- movies_directors %>%
  group_by(director_id) %>%
  summarise(total_peliculas = n_distinct(movie_id))

info_directores_completa <- info_directores %>%
  left_join(peliculas_dirigidas, by = "director_id")

head(info_directores_completa, 10)
```

| ## | director_id | nombre | apellido | total_peliculas |
|-------|-------------|--------------------|-----------|-----------------|
| ## 1 | 1 | Todd | 1 | 1 |
| ## 2 | 2 | Les 12 Poissons | | 1 |
| ## 3 | 3 | Lejaren | a'Hiller | 2 |
| ## 4 | 4 | Nian | A | 1 |
| ## 5 | 5 | Khairiya | A-Mansour | 1 |
| ## 6 | 6 | Ricardo | A. Solla | 1 |
| ## 7 | 8 | Kodanda Rami Reddy | A. | 35 |
| ## 8 | 9 | Nageswara Rao | A. | 1 |
| ## 9 | 10 | Yuri | A. | 1 |
| ## 10 | 11 | Swamy | A.S.A. | 1 |

c. ¿Cuántos actores han trabajado con cada director?

```
actores_por_director <- movies_directors %>%
  inner_join(roles, by = "movie_id") %>%
  group_by(director_id) %>%
  summarise(total_actores = n_distinct(actor_id))
```

```
## Warning in inner_join(., roles, by = "movie_id"): Detected an unexpected many-to-many relationship between
## i Row 2 of 'x' matches multiple rows in 'y'.
## i Row 119372 of 'y' matches multiple rows in 'x'.
## i If a many-to-many relationship is expected, set 'relationship =
## "many-to-many"' to silence this warning.
```

```
info_actores_por_director <- actores_por_director %>%
  inner_join(info_directores, by = "director_id")

head(info_actores_por_director, 10)
```

| ## | # A tibble: 10 x 4 | | | |
|------|----------------------------------|----------|--------------------|-------------|
| ## | director_id total_actores nombre | apellido | | |
| ## | <int> <int> <chr> | <chr> | | |
| ## 1 | 1 | 1 | Todd | 1 |
| ## 2 | 2 | 2 | Les | 12 Poissons |
| ## 3 | 3 | 15 | Lejaren | a'Hiller |
| ## 4 | 6 | 3 | Ricardo | A. Solla |
| ## 5 | 8 | 86 | Kodanda Rami Reddy | A. |

```
## 6          10          1 Yuri          A.
## 7          11          2 Swamy        A.S.A.
## 8          12         39 Per (I)       Aabel
## 9          13         23 Eivind       Aaeng
## 10         14          1 Mang         Aag
```

d. ¿Cuál es el género más común en las películas de cada director?

```
genero_mas_comun_director <- directors_genres %>%
  inner_join(movies_directors, by = "director_id") %>%
  group_by(director_id, genre) %>%
  summarise(apariciones = n()) %>%
  slice_max(apariciones, with_ties = FALSE) %>%
  ungroup()
```

```
## Warning in inner_join(., movies_directors, by = "director_id"): Detected an unexpected many-to-many :
## i Row 2 of 'x' matches multiple rows in 'y'.
## i Row 7 of 'y' matches multiple rows in 'x'.
## i If a many-to-many relationship is expected, set 'relationship =
## "many-to-many"' to silence this warning.
```

```
## 'summarise()' has grouped output by 'director_id'. You can override using the
## '.groups' argument.
```

```
info_genero_mas_comun <- genero_mas_comun_director %>%
  inner_join(info_directores, by = "director_id") %>%
  select(director_id, nombre, apellido, genre)

head(info_genero_mas_comun,10)
```

```
## # A tibble: 10 x 4
##   director_id nombre      apellido  genre
##   <int> <chr>          <chr>    <chr>
## 1         2 Les      12 Poissons Short
## 2         3 Lejaren   a'Hiller  Drama
## 3         5 Khairiya  A-Mansour Documentary
## 4         6 Ricardo   A. Solla  Drama
## 5         8 Kodanda Rami Reddy A.        Action
## 6        10 Yuri      A.        Comedy
## 7        11 Swamy     A.S.A.    Drama
## 8        12 Per (I)   Aabel     Comedy
## 9        16 Michael  Aaglund   Short
## 10       18 Astrid   Aakra     Animation
```

5. Analizar la distribución de “Roles” según las siguientes dimensiones:

a. Distribución por película

```
distribucion_rols_pelicula <- rols %>%
  group_by(movie_id) %>%
  summarise(cantidad_rols = n_distinct(role)) %>%
  group_by(cantidad_rols) %>%
  summarise(cantidad_peliculas = n()) %>%
  arrange(cantidad_rols)

head(distribucion_rols_pelicula,10)
```

```
## # A tibble: 10 x 2
##   cantidad_rols cantidad_peliculas
##         <int>         <int>
## 1             1          112552
## 2             2          26293
## 3             3          15283
## 4             4          11835
## 5             5          11508
## 6             6          10476
## 7             7          10043
## 8             8           9435
## 9             9           8723
## 10            10           8044
```

b. Distribución por director

```
distribucion_rols_director <- movies_directors %>%
  inner_join(rols, by = "movie_id") %>%
  group_by(director_id) %>%
  summarise(cantidad_rols = n_distinct(role)) %>%
  group_by(cantidad_rols) %>%
  summarise(cantidad_directores = n()) %>%
  arrange(cantidad_rols)
```

```
## Warning in inner_join(., rols, by = "movie_id"): Detected an unexpected many-to-many relationship between
## 'x' and 'y'.
## i Row 2 of 'x' matches multiple rows in 'y'.
## i Row 119372 of 'y' matches multiple rows in 'x'.
## i If a many-to-many relationship is expected, set 'relationship =
## "many-to-many"' to silence this warning.
```

```
head(distribucion_rols_director,10)
```

```
## # A tibble: 10 x 2
##   cantidad_rols cantidad_directores
```

| ## | <int> | <int> |
|-------|-------|-------|
| ## 1 | 1 | 16250 |
| ## 2 | 2 | 5858 |
| ## 3 | 3 | 4120 |
| ## 4 | 4 | 3080 |
| ## 5 | 5 | 2852 |
| ## 6 | 6 | 2380 |
| ## 7 | 7 | 1976 |
| ## 8 | 8 | 1724 |
| ## 9 | 9 | 1708 |
| ## 10 | 10 | 1668 |