Laboratorio#3

Isaac Cyrman

2024-09-03

```
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
```

```
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. Información general sobre la base de datos:

a. ¿Cuántas películas existen en la base de datos?

```
num_peliculas <- movies %>%
  summarise(total_peliculas = n())

num_peliculas
```

```
##   total_peliculas
## 1          388269
```

b. ¿Cuántos directores?

```
num_directores <- directors %>%
  summarise(total_directores = n())

num_directores
```

```
##   total_directores
## 1            86880
```

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

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

promedio_generos_por_director
```

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

3. Genere un nuevo reporte por "Role" con la siguiente información:

a. Número de películas

```
num_peliculas_por_rol <- roles %>%
  group_by(role) %>%
  summarise(num_peliculas = n_distinct(movie_id))

head(num_peliculas_por_rol,10)
```

```
## # A tibble: 10 x 2
##    role                              num_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

```
num_actores <- roles %>%
  group_by(role) %>%
  summarise(total_actores = n_distinct(actor_id))

head(num_actores,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"              3
```

```
##  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

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

head(num_actrices,10)
```

```
## # A tibble: 10 x 2
##    role                              num_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"              1
##  6 "\"Astoria\" Owner"                          1
##  7 "\"Betsy Ross\""                             1
##  8 "\"Frank\" Hickson"                          1
##  9 "\"Fred\" Lincoln"                           1
## 10 "\"Statue of Liberty\""                      1
```

d. Número de directores

```
num_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 be
## 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(num_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. Genere un nuevo reporte con la siguiente información:

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

```
informacion_directores <- directors %>%
  select(id, first_name, last_name) %>%
  rename(director_id = id, nombre = first_name, apellido = last_name)

head(informacion_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. Número de películas que ha dirigido

```
num_peliculas_dirigidas <- movies_directors %>%
  group_by(director_id) %>%
  summarise(num_peliculas = n_distinct(movie_id))

informacion_directores_completa <- informacion_directores %>%
  left_join(num_peliculas_dirigidas, by = "director_id")

head(informacion_directores_completa,10)
```

```
##    director_id             nombre    apellido num_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. Número de actores con los que ha trabajado

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

```
## Warning in inner_join(., roles, by = "movie_id"): Detected an unexpected many-to-many relationship be
## 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.
```

```
actores_por_director_con_nombre <- actores_por_director %>%
  inner_join(directors, by = c("director_id" = "id")) %>%
  select(director_id, first_name, last_name, num_actores)

head(actores_por_director_con_nombre,10)
```

```
## # A tibble: 10 x 4
##    director_id first_name         last_name   num_actores
##          <int> <chr>              <chr>             <int>
##  1           1 Todd               1                     1
##  2           2 Les                12 Poissons           2
##  3           3 Lejaren            a'Hiller             15
##  4           6 Ricardo            A. Solla              3
##  5           8 Kodanda Rami Reddy A.                   86
##  6          10 Yuri               A.                    1
##  7          11 Swamy              A.S.A.                2
##  8          12 Per (I)            Aabel                39
##  9          13 Eivind             Aaeng                23
## 10          14 Mang               Aag                   1
```

d. Género más común de sus películas

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

```
## Warning in inner_join(., directors_genres, by = "director_id"): Detected an unexpected many-to-many r
## i Row 7 of `x` matches multiple rows in `y`.
## i Row 2 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.
```

```
genero_mas_comun_con_nombre <- genero_mas_comun_por_director %>%
  inner_join(directors, by = c("director_id" = "id")) %>%
  select(director_id, first_name, last_name, genre)

head(genero_mas_comun_con_nombre,10)
```

```
## # A tibble: 10 x 4
##    director_id first_name         last_name   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. Encuentre la distribución de "Roles" por las siguientes dimensiones:

a. Película

```
hist_role_movie <- movies %>%
  left_join(roles, by = c("id" = "movie_id")) %>%
  group_by(id) %>%
  summarise(n_roles = n_distinct(role)) %>%
  ungroup() %>%
  group_by(n_roles) %>%
  summarise(n_movies = n()) %>%
  arrange(n_roles)

head(hist_role_movie,10)
```

```
## # A tibble: 10 x 2
##    n_roles n_movies
##      <int>    <int>
##  1       1   200569
##  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. Director

```
hist_role_director <- movies_directors %>%
  left_join(roles, by = "movie_id") %>%
  group_by(director_id) %>%
  summarise(n_roles = n_distinct(role)) %>%
  ungroup() %>%
  group_by(n_roles) %>%
```

```
  summarise(n_directores = n()) %>%
  arrange(n_roles)
```

```
## Warning in left_join(., roles, by = "movie_id"): Detected an unexpected many-to-many relationship be
## 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(hist_role_director,10)
```

```
## # A tibble: 10 x 2
##    n_roles n_directores
##      <int>        <int>
##  1       1        27248
##  2       2         7687
##  3       3         4499
##  4       4         3316
##  5       5         2892
##  6       6         2420
##  7       7         2006
##  8       8         1789
##  9       9         1722
## 10      10         1676
```