Práctica Netflix

Tecnologías para el Análisis de Datos Masivos

Daniel Ramos & Sergi Fornés

Importación y limpieza de los datos

Información sobre el Raw data

Vamos a realizar el análisis a partir de datos sobre valoraciones de películas de Netflix. Los datos que nos pueden resultar más interesantes son las notas de las valoraciones, la cantidad de valoraciones que tiene cada película y la identificación de quien hace estas valoraciones.

Obtenemos los datos a partir de 5 ficheros `.txt` y un `.csv`.

`combined_data_x.txt`

Tenemos a nuestra disposición 4 ficheros de este tipo: `combined_data_1.txt`, `combined_data_2.txt`, `combined_data_3.txt` y `combined_data_4.txt`. En estos ficheros tenemos información sobre valoraciones numéricas puestas por usuarios a películas de Netflix. Cada bloque de valoraciones está precedido por un número que indica la película (ID de la película). Para cada película existe un conjunto de valoraciones, de las cuales tenemos información sobre la fecha de la valoración, la nota (del 1 al 5) y un identificador del usuario.

`filas_ID_combined_all.txt`

En el fichero tenemos la localización de los bloques de valoraciones por películas en los ficheros `combined_data_x.txt`.

Columnas del fichero:

- `X1`: Identificación del bloque. Es el mismo nombre que aparece en los ficheros `combined_data_x.txt`.
- `fila`: Número de fila en el que se encuentra la identificación del bloque.
- `ID`: ID de la película asociada al bloque.
- `fila_final`: Última fila que contiene una valoración de la película `ID`.
- `data`: Número de fichero `combined_data_x.txt` al que pertenece la película `ID`. Tiene 4 valores posibles.

`movie_titles.csv`

En este archivo podemos encontrar información sobre las películas. Tenemos el ID de la película, su año de estreno y su nombre en inglés.

Obtención de las peliculas de interés

Decidimos las películas que vamos a analizar usando como semilla aleatoria nuestras dos últimas cifras del DNI.

```
set.seed(3793)
rand_pelis <- sort(sample(1:17770, 250, replace = FALSE))
```

Obtenemos del fichero `filas_ID_combined_all.txt` únicamente la información de las películas que nos ha tocado analizar.

```
loc_pelis <- read_csv("../data/filas_ID_combined_all.txt", col_types = "ciiii") %>%
  filter(ID %in% rand_pelis)
```

Creación del data frame

De cada uno de los archivos `combined_data_x.txt`, buscamos los datos de las películas que analizaremos y los metemos en un data frame, especificando en cada observación cual es el ID de la película.

```
i <- 1
df <- tibble()
for(comb_data in 1:4){
  file <- paste("../data/combined_data_", comb_data, ".txt", sep = "")
  while(loc_pelis[i,"data"] == comb_data & i <= 250){
    df_aux <- read_csv(file,
                       col_names = c("User","Score","Date"),
                       col_types = "iiD",
                       skip = loc_pelis[[i,"fila"]],
                       n_max = loc_pelis[[i,"fila_final"]] - loc_pelis[[i,"fila"]]) %>%
      mutate(ID_film = rand_pelis[i])
    df <- rbind(df, df_aux)
    i <- i + 1
  }
}
```

Cargamos el fichero con los nombres de las películas, tenemos en cuenta que el nombre de la película puede llevar , y arreglamos los valores `NA`. Este nuevo data frame cuenta con el ID de cada película, por lo que podemos unirlo con el data frame de las valoraciones de los usuarios para ampliarlo con información sobre las películas. Después lo guardamos en un fichero `.csv` para poder cargarlo y analizarlo en la siguiente sección.

```
names_pelis <- read_tsv("../data/movie_titles.csv",
                        locale = readr::locale(encoding = "ISO-8859-1"),
                        col_names = FALSE) %>%
  separate(col = X1,
           sep = ",",
           into = c("ID_film","Release_Year","Title"),
           extra = "merge",
           convert = TRUE) %>%
  mutate(Release_Year = ifelse(Release_Year == "NULL", NA, as.integer(Release_Year)))

df <- inner_join(df, names_pelis)

write_csv(df, "../data/pelis.csv")
```

Análisis exploratorio de los datos

```
data <- read_csv("../data/pelis.csv")

str(data)
```

```
## tibble [1,305,391 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ User        : num [1:1305391] 712664 2336678 2422606 1241149 672200 ...
##  $ Score       : num [1:1305391] 3 3 3 3 2 3 2 1 3 2 ...
##  $ Date        : Date[1:1305391], format: "2004-09-02" "2004-08-23" ...
##  $ ID_film     : num [1:1305391] 26 26 26 26 26 26 26 26 26 26 ...
##  $ Release_Year: num [1:1305391] 2004 2004 2004 2004 2004 ...
##  $ Title       : chr [1:1305391] "Never Die Alone" "Never Die Alone" "Never Die Alone" "Never Die Al
##  - attr(*, "spec")=
##   .. cols(
##   ..   User = col_double(),
##   ..   Score = col_double(),
##   ..   Date = col_date(format = ""),
##   ..   ID_film = col_double(),
##   ..   Release_Year = col_double(),
##   ..   Title = col_character()
##   .. )
```

Tipología de las variables:

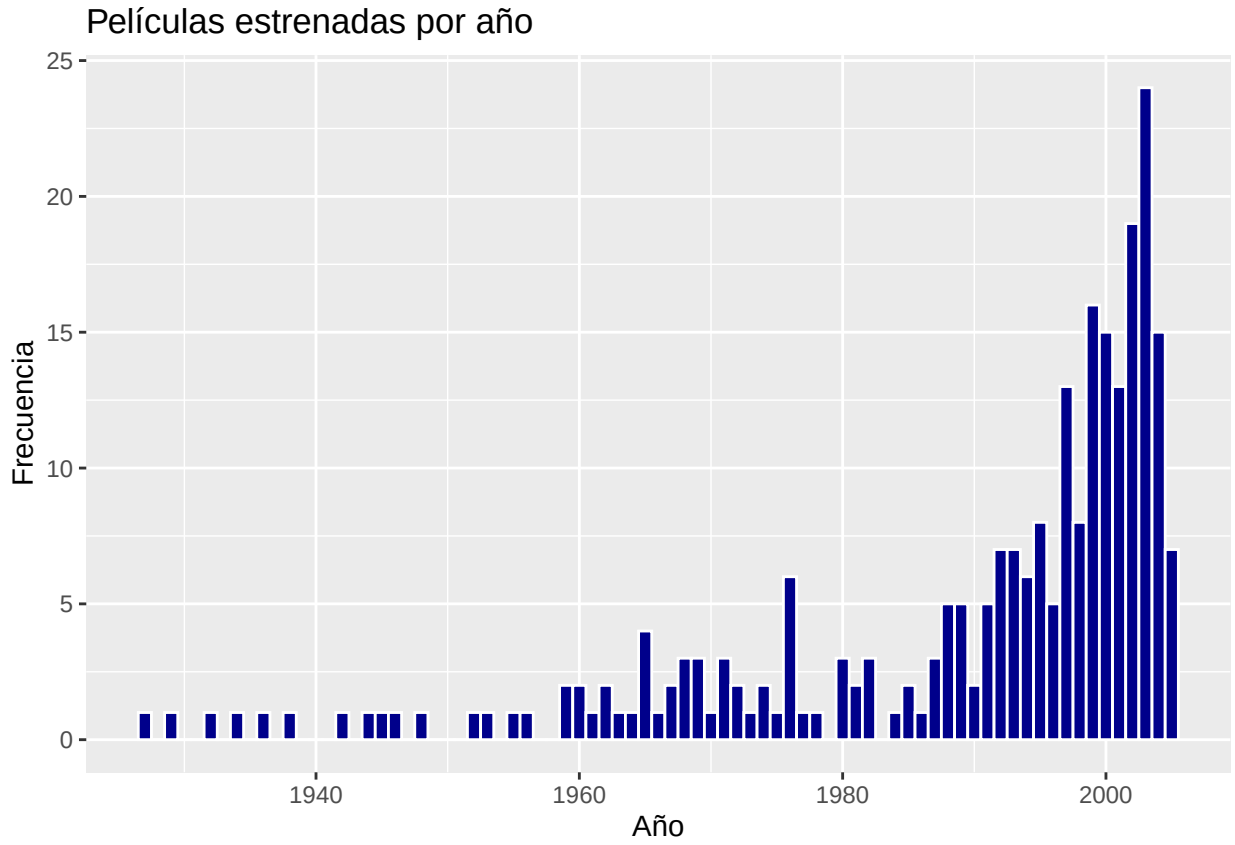
- `User`: Se puede considerar esta variable como categórica, ya que es un identificador del usuario que ha realizado la valoración. Pero al haber tantos usuarios diferentes, la dejaremos como variable tipo `num`.
- `Score`: Es una variable numérica que refleja la puntuación de la película de peor a mejor con valores enteros del 1 al 5. También podría ser considerada como una variable ordinal ya que sus valores son discretos, aunque para trabajar con ella es mejor dejarla tipo `num`.
- `Date`: Es una variable tipo fecha, representa el día que se valoró la película.
- `ID_film`: Esta variable tiene la misma forma que `User`, es decir, es una variable categórica que refleja la película valorada, pero la dejamos en formato `num`. Ya tenemos la variable `Title` para identificar las películas, por lo que realmente podríamos desechar esta variable. Pero es una manera más sencilla para acceder a las películas.
- `Release_Year`: Podríamos tenerla en formato `Date`, pero al ser únicamente el año, es más cómodo usarla como una variable tipo `num`.
- `Title`: Es claramente una variable categórica que representa el título de la película en formato `character`.

PREGUNTA2

```
data %>%
  group_by(Release_Year, Title) %>%
  summarise() %>%
  ggplot() +
```

```
  geom_histogram(aes(x = Release_Year), binwidth = 1, fill = "darkblue", col = "white") +
  xlab("Año") +
  ylab("Frecuencia") +
  ggtitle("Películas estrenadas por año")
```

```
## `summarise()` regrouping output by 'Release_Year' (override with `.groups` argument)
```




```
# FALTA METER MÁS AÑOS EN EL EJE X (CADA 10 AÑOS)
```

PREGUNTA3

```
data <- data %>%
  mutate(Year_D = year(Date),
         Month_D = factor(month(Date),
                          levels = c(1,2,3,4,5,6,7,8,9,10,11,12),
                          labels = c("Enero", "Febrero", "Marzo", "Abril", "Mayo", "Junio", "Julio", "Ag
         Day_D = day(Date),
         Week_D = week(Date),
         # tenemos en cuenta que el primer dia de la semana es Domingo en vez de Lunes
         Day_of_week_D = factor(wday(Date),
                                levels = c(1,2,3,4,5,6,7),
                                labels = c("Domingo", "Lunes", "Martes", "Miercoles", "Jueves", "Viernes
```

PREGUNTA4

```
films_table <- data %>%
  group_by(Title) %>%
  summarise(count = n(),
            sum_scores = sum(Score),
            mean_scores = round(mean(Score), 2),
            median_scores = median(Score),
            mode_scores = unique(Score)[which.max(tabulate(match(Score, unique(Score))))],
            sd_scores = round(sd(Score), 2))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
  # es necesario instalar el paquete "formattable"
films_table %>%
  formattable(align = c("l","c","c","c","c", "c", "c"),
              list(mean_scores = color_tile("#FF7F7F", "#71CA97"),
                   count = color_bar("lightgrey"),
                   sd_scores = color_tile("white", "lightblue")),
              col.names = c("Título","Cantidad de Valoraciones", "Suma de Valoraciones", "Media", "Media
```

Título

Cantidad de Valoraciones

Suma de Valoraciones

Media

Mediana

Moda

Desviación Típica

.Com for Murder

153

355

2.32

2

2

1.09

32 Short Films About Glenn Gould

2083

7396

3.55

4

4

1.14

99 Women (Unrated Director's Cut)

240

442

1.84

2

1

1.02

A Decade Under the Influence

3405

11588

3.40

3

4

1.02

A Murder of Crows

6407

23581

3.68

4

4

0.99

A Nightmare on Elm Street 4: The Dream Master

10515

33563

3.19

3

3

1.18

A Walk in the Sun

274

820

2.99

3

3

1.15

After Life

593

2193

3.70

4

4

1.12

Alias: Season 1

16683

72478

4.34

5

5

1.01

Alice's Adventures in Wonderland

127

377

2.97

3

3

1.15

Alien Nation

5146

18118

3.52

4

4

0.86

Amelie: Bonus Material

568

2150

3.79

4

5

1.18

An American Werewolf in London

19599

70477

3.60

4

4

0.95

Angels in the Endzone

704

2432

3.45

3

3

1.15

Anne of Green Gables: The Sequel

701

3128

4.46

5

5

0.84

Another Man's Poison

378

1277

3.38

3

4

1.11

Aqua Teen Hunger Force: Vol. 1

6890

28722

4.17

5

5

1.15

Baby the Rain Must Fall

171

509

2.98

3

3

0.88

Barefoot in the Park

8318

29987

3.61

4

4

0.93

Barney: Best Manners

275

856

3.11

3

3

1.24

Batman: Mask of the Phantasm

2823

10637

3.77

4

4

1.00

Battle Athletes Victory: Vol. 5: No Looking Back

389

1197

3.08

3

1

1.54

Bauhaus: Gotham

173

597

3.45

4

4

1.30

Bear in the Big Blue House: A Bear for All Seasons

163

519

3.18

3

3

1.32

Ben & Arthur

655

997

1.52

1

1

0.92

Big Girls Don’t Cry

491

1548

3.15

3

3

1.08

Black Sunday

1380

4145

3.00

3

3

1.09

Blackwoods

250

611

2.44

2

3

1.09

Body of Evidence

2317

6298

2.72

3

3

1.15

Bread and Roses

2113

6963

3.30

3

3

1.00

Brief Encounter

1605

6102

3.80

4

4

1.01

Bug

257

768

2.99

3

3

1.20

Caillou: Caillou's Treasure Hunt and Other Adventures

209

687

3.29

4

4

1.30

Callas Forever

774

2607

3.37

3

4

1.06

Candide

192

692

3.60

4

4

1.23

Carnal Crimes

153

306

2.00

2

1

1.06

Cedric the Entertainer: Starting Lineup 2

767

2497

3.26

3

3

1.18

Charles Mingus: Triumph of the Underdog

251

945

3.76

4

4

1.12

Chuck & Buck

3404

10275

3.02

3

3

1.15

Cindy Crawford: Shape Your Body Workout

104

307

2.95

3

3

1.28

Circuitry Man / Circuitry Man 2: Plughead Rewired

184

435

2.36

2

2

1.11

Circus World

102

328

3.22

3

3

1.12

City by the Sea

22851

69255

3.03

3

3

0.88

CKY: Infiltrate, Destroy, Rebuild: The Video Album

271

1012

3.73

4

5

1.40

Clash of the Titans

20036

73440

3.67

4

4

0.96

Clueless

47152

171161

3.63

4

4

0.98

Colosseum: A Gladiator's Story

148

522

3.53

4

4

1.08

Dance with the Devil

522

1284

2.46

2

3

1.16

Dangerous Evidence: The Lori Jackson Story

195

635

3.26

3

3

1.13

Dead Presidents

6328

21265

3.36

3

3

1.01

Discovering Alaska

90

241

2.68

3

3

1.00

Disney Princess Sing-Along Songs: Once Upon a Dream

499

1706

3.42

4

4

1.18

DJ Shadow: In Tune and On Time

243

887

3.65

4

4

1.25

Dora the Explorer: Dora's Halloween

1488

5753

3.87

4

4

1.05

Double Whammy

1081

2975

2.75

3

3

1.00

Drugstore Cowboy

21063

75936

3.61

4

4

0.96

Ellis Island

172

532

3.09

3

3

1.25

Faster Than Sound: Nova

175

575

3.29

3

3

0.99

Final Stab

96

205

2.14

2

2

1.04

Firefox

8906

29105

3.27

3

3

0.91

Footballers Wives: Season 1

334

1169

3.50

4

4

1.35

Fourplay

99

249

2.52

3

3

0.87

Frasier: Season 2

3311

12877

3.89

4

4

1.09

Fried Green Tomatoes

79845

316386

3.96

4

4

0.99

Frosty's Winter Wonderland / 'Twas the Night Before Christmas

601

2390

3.98

4

4

0.98

Galaxy Quest

36824

130378

3.54

4

4

0.99

Gentlemen of Fortune

164

645

3.93

4

5

1.26

Gerry

3861

8496

2.20

2

1

1.21

Gorillaz: Phase One Celebrity Take Down

97

350

3.61

4

3

1.12

Gothic Industrial Madness

111

292

2.63

3

3

1.24

Guide

184

656

3.57

4

4

1.22

H.R. Pufnstuf

466

1467

3.15

3

3

1.31

Harold and the Purple Crayon: The Complete Series

122

430

3.52

4

5

1.31

Head in the Clouds

6434

20162

3.13

3

3

0.96

Hiroshima Mon Amour

1537

5762

3.75

4

4

1.09

Hocus Pocus

20010

66933

3.34

3

3

1.08

Homicide: The Movie

1325

4531

3.42

3

3

1.12

Hum Dil De Chuke Sanam

949

3407

3.59

4

4

1.16

I Capture the Castle

8089

26904

3.33

3

3

0.95

Inspirations

388

1191

3.07

3

3

1.05

Introducing Dorothy Dandridge

5973

21235

3.56

4

3

0.98

Iron Maiden: The Early Days

163

607

3.72

4

5

1.28

It Happened One Night

17123

68685

4.01

4

4

0.91

Ivan the Terrible

455

1680

3.69

4

4

1.10

Jackson Pollock: Love and Death on Long Island

444

1475

3.32

3

3

1.04

Jaws 2

20407

60010

2.94

3

3

1.00

Jimi Hendrix: Rainbow Bridge

214

593

2.77

3

3

1.29

Jimmy Neutron: Boy Genius

10846

37923

3.50

4

4

1.00

Jin-Roh: The Wolf Brigade

1696

5965

3.52

4

4

1.11

Jonah: A VeggieTales Movie

7775

27968

3.60

4

4

1.25

Josh Groban: Live at the Greek

378

1478

3.91

4

4

1.13

Khakee

515

1771

3.44

4

4

1.03

Kirby: A Dark & Stormy Knight

138

425

3.08

3

1

1.49

Kiss: Unauthorized Kiss

93

218

2.34

2

1

1.24

Kull the Conqueror

2615

7256

2.77

3

3

1.10

L'Enfer

443

1262

2.85

3

3

1.02

Labyrinth

31853

122731

3.85

4

4

1.03

Laserhawk

69

158

2.29

2

2

1.02

Laurence Olivier's Hamlet

2450

9194

3.75

4

4

1.03

Lazytown: New Superhero

202

662

3.28

3

4

1.24

Lilo and Stitch 2

5730

22132

3.86

4

4

0.98

Little House on the Prairie: The Pilot

1431

5967

4.17

4

5

1.04

Lone Wolf and Cub: Sword of Vengeance

631

2520

3.99

4

4

1.00

Lost in Space: Season 3: Vol. 2

230

818

3.56

4

4

1.26

Love and a .45

1295

3949

3.05

3

3

1.13

Love Don't Cost a Thing

9413

30118

3.20

3

3

1.07

Love Stories

146

488

3.34

3

4

1.17

Mahalia Jackson: The Power and the Glory

117

401

3.43

4

4

1.32

Maid to Order

5569

16607

2.98

3

3

1.09

Mallrats

33900

125865

3.71

4

4

1.07

Married to It

206

576

2.80

3

3

1.00

Martian Successor Nadesico

920

3519

3.83

4

4

1.18

McKenzie Break

127

372

2.93

3

3

0.96

Mesmer

265

639

2.41

2

2

1.01

Million Dollar Baby

102861

428401

4.16

4

5

0.91

Mission: Impossible II

78304

261963

3.35

3

3

1.00

Mississippi Burning

25292

99502

3.93

4

4

0.80

Monkeybone

4648

11427

2.46

2

2

1.11

Monsieur Ibrahim

7542

29158

3.87

4

4

0.88

Morlang

117

356

3.04

3

3

1.09

Mr. Baseball

2457

8123

3.31

3

3

0.89

Mulan

1505

5194

3.45

4

4

1.23

Mutant X: Season 2

334

1173

3.51

4

5

1.30

My Life So Far

3008

9593

3.19

3

3

0.98

My Name Is Nobody

288

1051

3.65

4

4

1.05

Najica Blitz Tactics

412

1405

3.41

4

4

1.20

National Geographic: Vietnam's Unseen War: Pictures from the Other Side

278

889

3.20

3

3

1.05

National Lampoon's Christmas Vacation: Special Edition

15002

54656

3.64

4

4

1.13

Neil Gaiman's A Short Film About John Bolton

141

418

2.96

3

3

1.22

Never a Dull Moment

120

343

2.86

3

3

1.15

Never Die Alone

5861

16374

2.79

3

3

1.09

Nine Queens

7978

29461

3.69

4

4

0.98

No Deposit No Return

723

2586

3.58

4

4

1.05

Obsession

742

2189

2.95

3

3

1.02

One Hundred and One Nights

116

306

2.64

3

3

1.11

Original Gangstas

413

1134

2.75

3

3

1.17

Oscar

2684

8070

3.01

3

3

1.25

Pan Tadeusz

122

396

3.25

3

3

1.36

Piglet's Big Movie

6888

24425

3.55

4

4

1.02

Pippi's Adventures on the South Seas

205

590

2.88

3

3

1.13

Place Vendome

795

2305

2.90

3

3

0.97

Poetic Justice

5606

18189

3.24

3

3

1.13

Poirot: Murder in Mesopotamia

1495

6040

4.04

4

4

0.91

Pokemon 4Ever

3138

9318

2.97

3

3

1.38

Pornografia

71

167

2.35

2

1

1.29

Rabid Grannies

128

305

2.38

2

2

1.16

Recipe for Disaster

134

383

2.86

3

3

1.06

Red Green's: We Can't Help It – We're Men

112

373

3.33

4

4

1.46

Renegade: Season 1

132

400

3.03

3

4

1.42

Reservoir Dogs: Bonus Material

765

3030

3.96

4

5

1.09

Riding the Bullet

1507

3895

2.58

3

3

0.98

Road House

16685

54925

3.29

3

3

1.08

Robot Monster

222

632

2.85

3

3

1.29

Rocky & Bullwinkle & Friends: Season 2

794

3049

3.84

4

5

1.13

Rocky & Bullwinkle: The Best of Fractured Fairy Tales: Vol. 1

86

287

3.34

3

3

1.20

Rolie Polie Olie: The Baby 'Bot Chase

381

1287

3.38

3

4

1.14

Roughing It

417

1382

3.31

3

4

1.16

Roujin Z

371

1138

3.07

3

3

1.22

Run the Wild Fields

438

1459

3.33

3

3

0.97

Rushmore

65185

245349

3.76

4

5

1.12

Saaya

102

295

2.89

3

3

1.22

Saboteur

2093

7549

3.61

4

4

0.93

Saikano

183

679

3.71

4

4

1.16

Sasquatch Hunters

97

219

2.26

2

2

1.23

Satin Rouge

588

1785

3.04

3

3

1.05

Seinfeld: Seasons 1 & 2

11184

49050

4.39

5

5

0.93

Sherlock Holmes: The Sign of Four

1216

4711

3.87

4

4

1.02

Simon & Garfunkel: Old Friends Live on Stage

248

969

3.91

4

4

1.05

Simpatico

1231

2995

2.43

2

2

0.94

Sleepover

10545

35472

3.36

3

3

1.04

Some Like It Hot

32781

130248

3.97

4

4

0.92

Spawn 2

2037

6696

3.29

3

3

1.28

Spider-Man: The ’67 Classic Collection

1093

4003

3.66

4

4

1.15

Star Trek: Voyager: Season 1

6007

23681

3.94

4

5

1.18

Storefront Hitchcock

178

567

3.19

3

4

1.42

Storytelling

8408

25396

3.02

3

3

1.12

Stranger than Paradise

4331

15395

3.55

4

4

1.15

Surfin' Shorts

84

263

3.13

3

3

1.21

Swami

105

319

3.04

3

4

1.29

Sweeney Todd: The Demon Barber of Fleet Street

1158

4527

3.91

4

5

1.09

Sweepers

225

571

2.54

2

2

1.07

Tess

833

2850

3.42

3

3

1.02

The Avengers

4412

10301

2.33

2

2

1.07

The Best of Riverdance

67

232

3.46

3

3

1.26

The Best of Thunderbirds

189

568

3.01

3

3

1.29

The Big Green

1784

6356

3.56

4

3

1.02

The Bridge at Remagen

1872

6770

3.62

4

4

0.91

The Cardinal

310

991

3.20

3

3

1.05

The Clash: Westway to the World

934

3631

3.89

4

4

1.01

The Cuckoo

3378

12694

3.76

4

4

0.96

The Dead Zone

4854

18570

3.83

4

4

0.96

The Desperate Trail

148

448

3.03

3

3

0.98

The Fabulous Story of the Cuban Cigar

122

365

2.99

3

3

1.12

The Fire That Burns

322

911

2.83

3

3

1.10

The Hallelujah Trail

497

1663

3.35

3

3

1.06

The Haunting of Morella

168

360

2.14

2

2

1.01

The Heroic Trio

855

2610

3.05

3

3

1.10

The Killing of a Chinese Bookie

459

1503

3.27

3

3

1.18

The Lathe of Heaven

1264

4045

3.20

3

3

1.09

The Lion King II: Simba's Pride

12914

48645

3.77

4

4

1.00

The Lon Chaney Collection: The Unknown

224

849

3.79

4

4

1.01

The Lucy Show

597

2049

3.43

3

5

1.34

The Man Who Never Was

256

906

3.54

4

4

1.00

The Man with the Golden Arm: 50th Anniversary Special Edition

919

3283

3.57

4

4

0.99

The Newcomers

102

291

2.85

3

3

1.15

The Newsroom: The Complete Series

111

341

3.07

3

4

1.48

The Odyssey of Life: The Photographer's Secrets: Nova

168

521

3.10

3

3

1.14

The Pact of Silence

417

1211

2.90

3

3

1.05

The Pianist

60787

250533

4.12

4

4

0.90

The Prophecy: Uprising

1866

5334

2.86

3

3

1.03

The Sailor Who Fell from Grace with the Sea

211

590

2.80

3

3

1.08

The Scarlet Pimpernel

906

3197

3.53

4

4

1.08

The Seven Percent Solution

497

1537

3.09

3

3

0.95

The Sex Pistols: The Great Rock ‘n’ Roll Swindle

133

389

2.92

3

3

1.22

The Siege

20160

70649

3.50

4

4

0.91

The Singing Detective

3193

6915

2.17

2

1

1.08

The Singing Forest

243

514

2.12

2

1

1.17

The Snapper

3868

12712

3.29

3

3

1.00

The Thin Blue Line

1350

5266

3.90

4

4

1.08

The Three Stooges: Cops and Robbers

1571

6134

3.90

4

5

1.20

The Three Stooges: Stooged and Confoosed

697

2813

4.04

4

5

1.14

The Trojan Women

119

346

2.91

3

3

1.27

The Twilight Zone: Vol. 22

6053

22114

3.65

4

4

1.15

The Volcano Disaster

186

459

2.47

2

2

1.10

The Winter Guest

194

567

2.92

3

3

1.05

The World of Narue

206

754

3.66

4

4

1.17

There's Something About McConkey

126

408

3.24

3

3

1.38

Tin Cup

24225

82289

3.40

3

3

0.96

Tom and Jerry: The Movie

1698

5354

3.15

3

3

1.22

Trapped in Paradise

2479

8331

3.36

3

3

0.99

True Lies

78944

290337

3.68

4

4

0.94

Turk 182!

1094

3819

3.49

4

3

0.93

UFC Hits: Ultimate Fighting Championship

1541

5416

3.51

4

4

1.16

Un Chien Andalou

1944

7559

3.89

4

4

1.08

Uncle Saddam

531

1558

2.93

3

3

1.04

Uncovered: The Whole Truth About the Iraq War

3526

13346

3.79

4

4

1.14

Under Siege

29488

97741

3.31

3

3

1.02

Upstairs, Downstairs: Season 5

1848

7792

4.22

5

5

1.08

Vagabond

525

1736

3.31

3

3

1.06

Vampyr

563

1762

3.13

3

3

1.22

Vietnam: We Were Heroes 1st Cavalry Division Airmobile

106

259

2.44

2

2

1.18

Violence in a Women's Prison

213

435

2.04

2

1

1.13

Warren Miller's: Storm

608

2227

3.66

4

4

1.08

Wild Things: Diamonds in the Rough

1774

4971

2.80

3

3

1.10

Wilder Days

79

220

2.78

3

3

1.05

Wiseguy: Season 1: Part 2

590

2165

3.67

4

4

1.23

Woyzeck

183

636

3.48

4

4

1.06

WWE: The Monday Night War

266

1133

4.26

5

5

1.01

Young Torless

109

340

3.12

3

3

1.06

Zanjeer

233

837

3.59

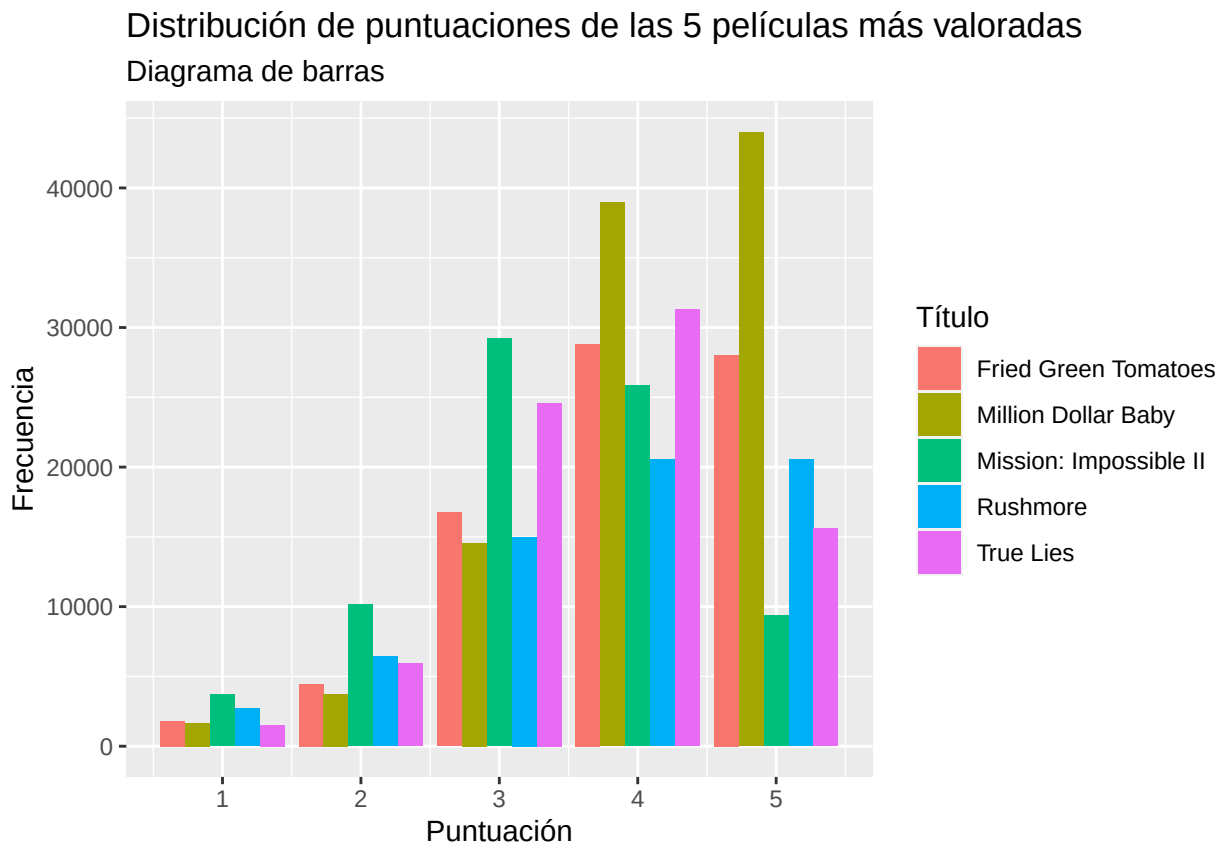
4

4

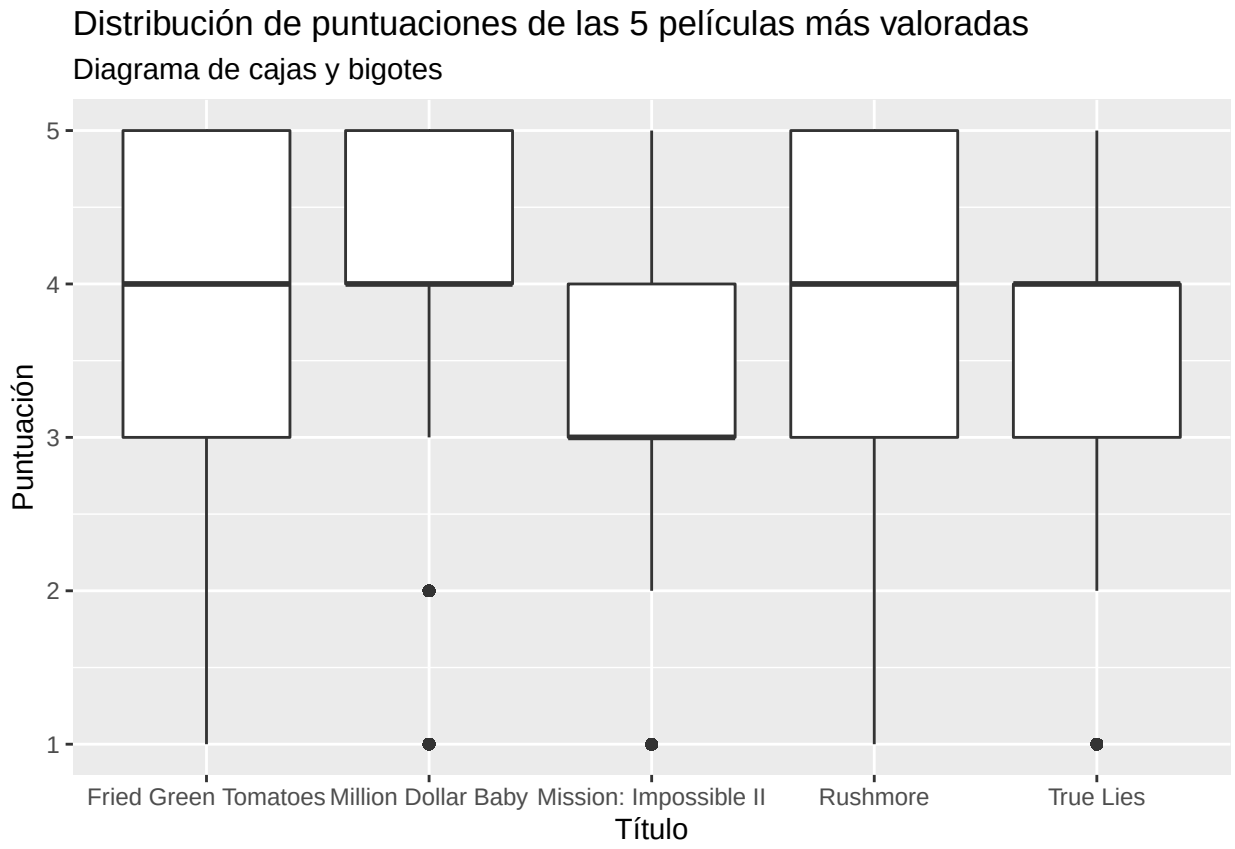
1.17

PREGUNTA5

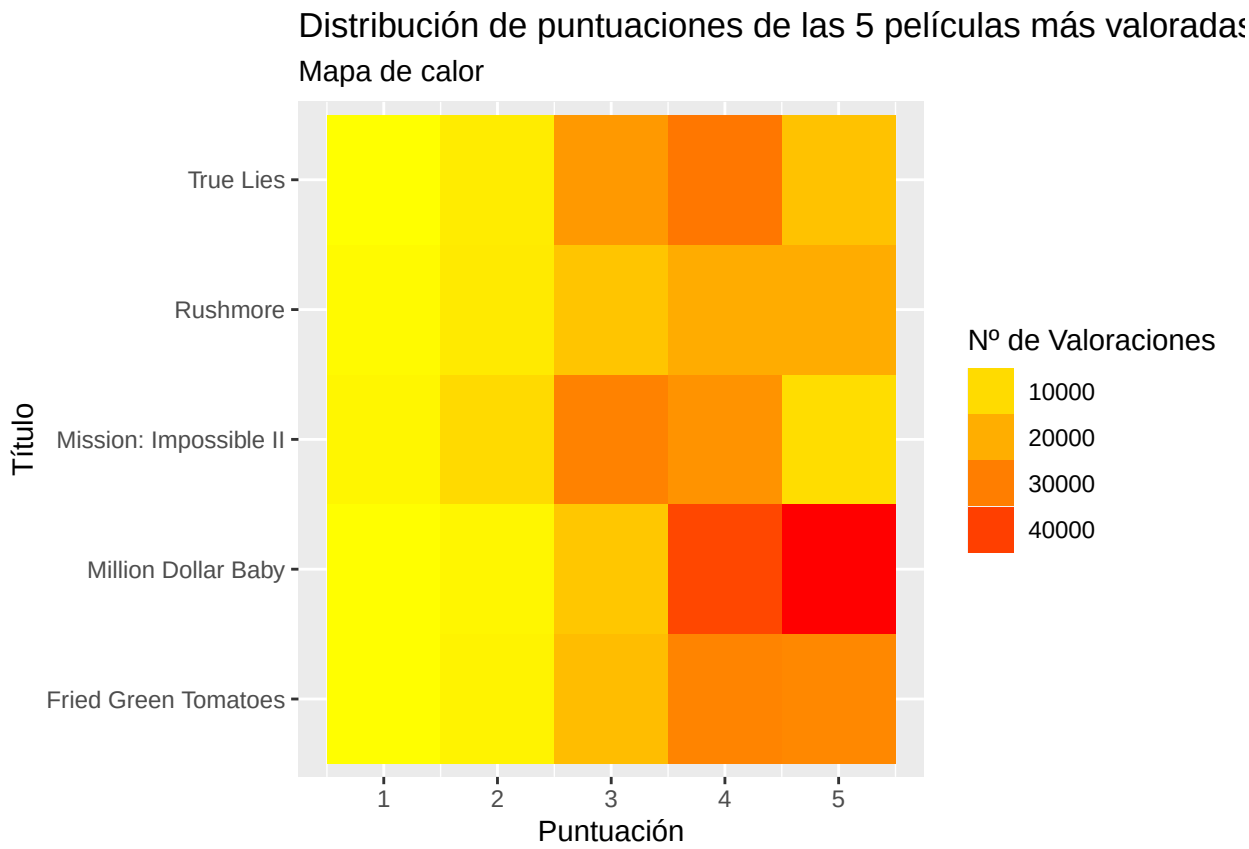
```
top5 <- films_table %>%
  top_n(5, count) %>%
  select("Title")

# barras
data %>%
  filter(Title %in% top5$Title) %>%
  ggplot() +
  geom_bar(aes(x = Score, fill = Title), position = "dodge") +
  xlab("Puntuación") +
  ylab("Frecuencia") +
  ggtitle("Distribución de puntuaciones de las 5 películas más valoradas",
          "Diagrama de barras") +
  guides(fill = guide_legend(title = "Título"))
```




```
# cajas y bigotes
data %>%
  filter(Title %in% top5$Title) %>%
  ggplot() +
  geom_boxplot(aes(x = Title, y = Score)) +
  xlab("Título") +
  ylab("Puntuación") +
  ggtitle("Distribución de puntuaciones de las 5 películas más valoradas",
          "Diagrama de cajas y bigotes")
```




```
# mapa de calor
data %>%
  filter(Title %in% top5$Title) %>%
  count(Title, Score) %>%
  ggplot() +
  geom_tile(aes(x = Score, y = Title, fill = n)) +
  scale_fill_gradient(low="yellow", high="red") +
  xlab("Puntuación") +
  ylab("Título") +
  ggtitle("Distribución de puntuaciones de las 5 películas más valoradas",
          "Mapa de calor") +
  guides(fill = guide_legend(title = "Nº de Valoraciones"))
```

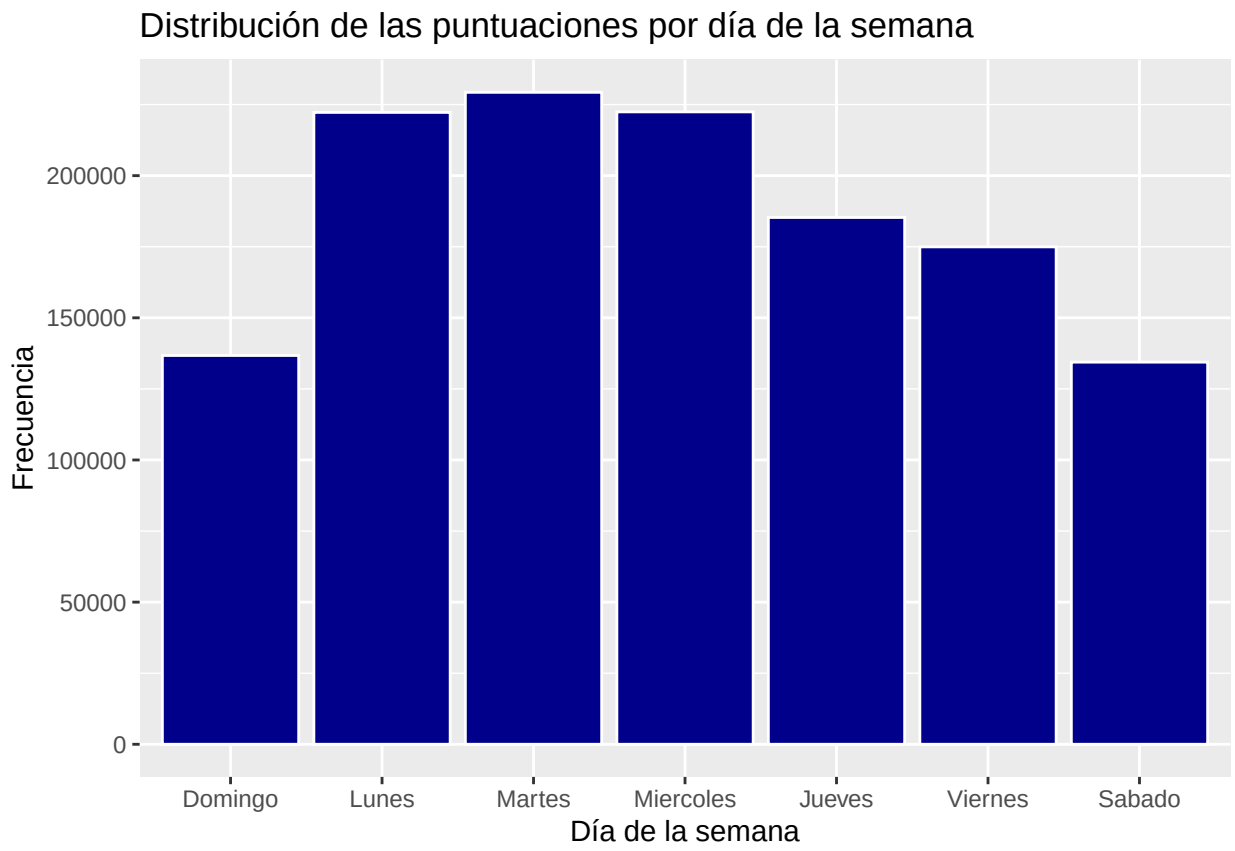


COMENTARIO: Los 2 primeros gráficos se ven mal debido a que la variable puntuación es discreta. Con el mapa de calor se visualiza mejor.

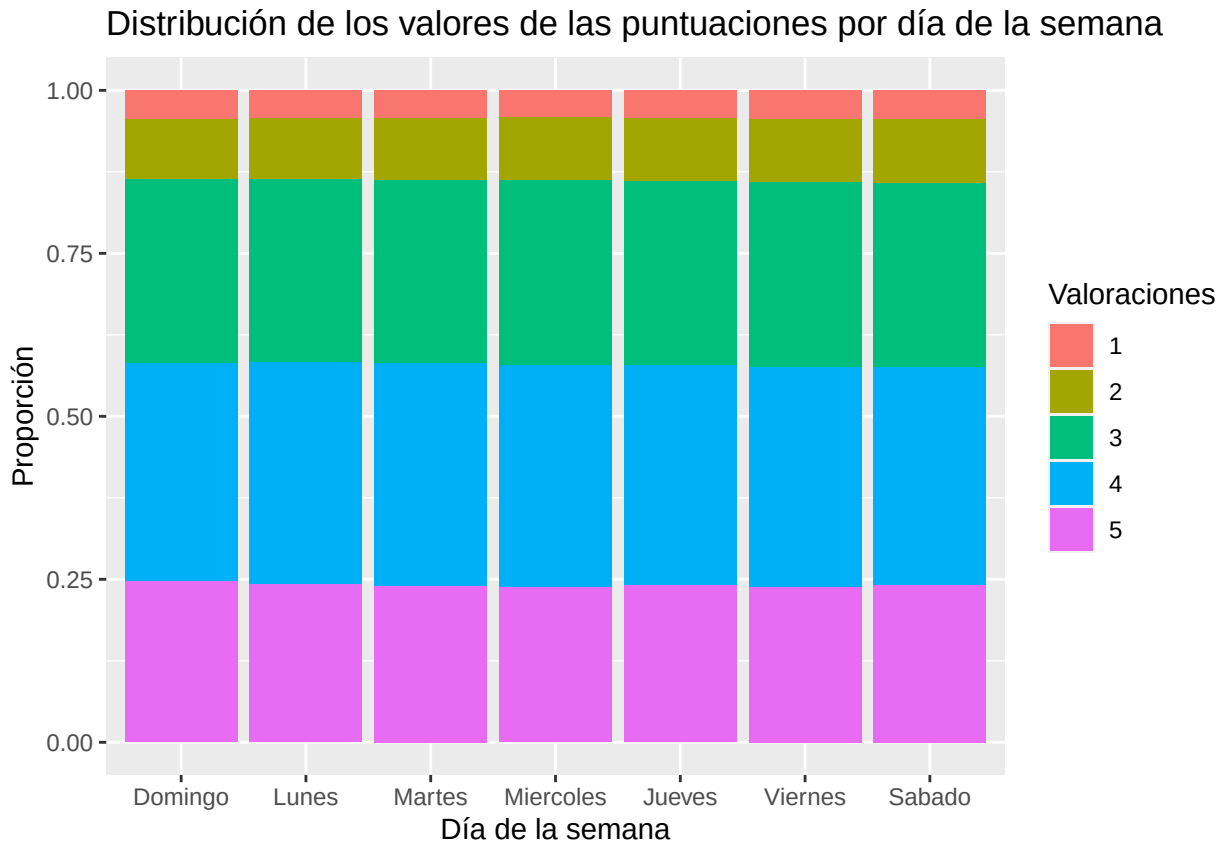
PREGUNTA6

COMENTARIO: Primero analizaremos la cantidad de valoraciones, y luego si los valores de estas cambian segun la fecha.

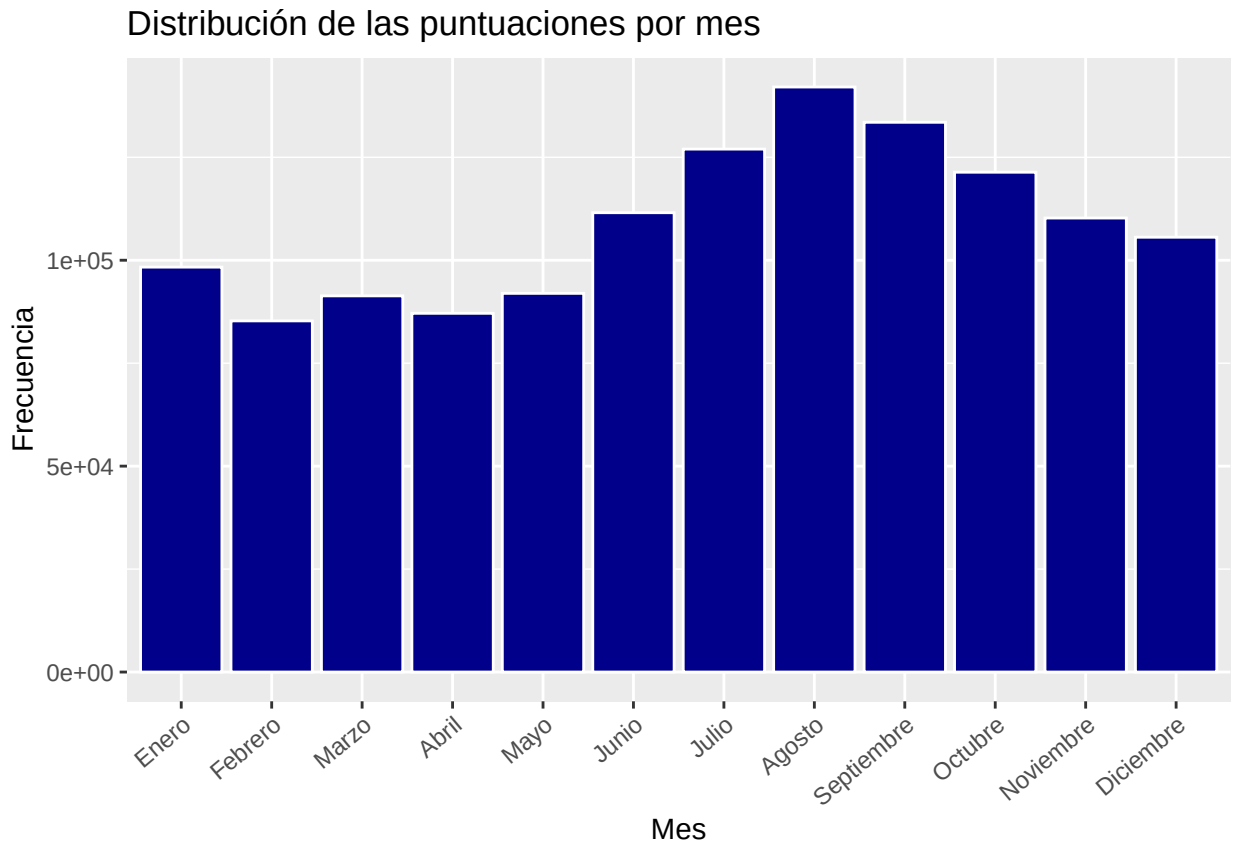
```
ggplot(data) +
  geom_bar(aes(x = Day_of_week_D), fill = "darkblue", col = "white") +
  xlab("Día de la semana") +
  ylab("Frecuencia") +
  ggtitle("Distribución de las puntuaciones por día de la semana")
```



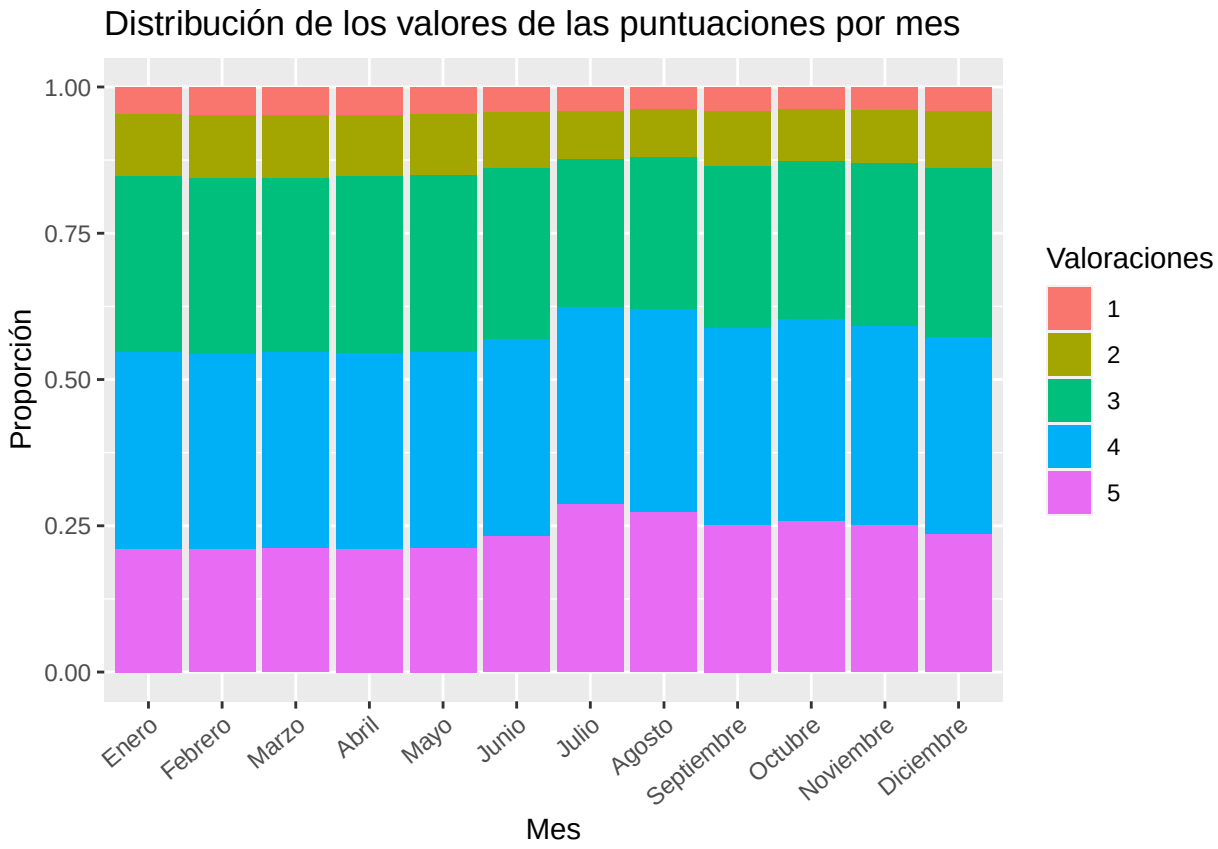

```
ggplot(data) +
  geom_bar(aes(x = Day_of_week_D, fill = as.factor(Score)), position = "fill") +
  xlab("Día de la semana") +
  ylab("Proporción") +
  ggtitle("Distribución de los valores de las puntuaciones por día de la semana") +
  guides(fill = guide_legend(title = "Valoraciones"))
```




```
ggplot(data) +
  geom_bar(aes(x = Month_D), fill = "darkblue", col = "white") +
  xlab("Mes") +
  ylab("Frecuencia") +
  ggtitle("Distribución de las puntuaciones por mes") +
  theme(axis.text.x = element_text(angle = 40, hjust = 1))
```




```
ggplot(data) +
  geom_bar(aes(x = Month_D, fill = as.factor(Score)), position = "fill") +
  xlab("Mes") +
  ylab("Proporción") +
  ggtitle("Distribución de los valores de las puntuaciones por mes") +
  guides(fill = guide_legend(title = "Valoraciones")) +
  theme(axis.text.x = element_text(angle = 40, hjust = 1))
```


Distribución de los valores de las puntuaciones por mes
Proporción
1.00
0.75
0.50
0.25
0.00
Enero
Febrero
Marzo
Abril
Mayo
Junio
Julio
Agosto
Septiembre
Octubre
Noviembre
Diciembre
Mes
Valoraciones
1
2
3
4
5