

# Práctica Netflix

Daniel Ramos & Sergi Fornés

## Importación y limpieza de los datos

Vamos a realizar el análisis a partir de datos sobre valoraciones de películas de Netflix. Los datos que nos pueden resultar más interesantes son los valores de las valoraciones, la cantidad de valoraciones que tiene cada película, la fecha de las valoraciones 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 este 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 título.

## Obtención de las películas de interés

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

```
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 valorada.

```
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, teniendo 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 archivo .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, "./model.netflix/pelis.csv")
```

## Análisis exploratorio de los datos

Antes de empezar con el análisis, cargamos el data frame anteriormente creado y comprobamos que las variables están en el formato correcto.

```
data <- read_csv("./model.netflix/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 Alone" ...
## - 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**.

```

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") +
  scale_x_continuous(breaks = seq(1930, 2000, 10))

```



### Distribución de películas estrenadas por año

En el gráfico anterior vemos una tendencia creciente de las películas estrenadas por año, sobretodo desde mediados de los 80 hasta los 2000. Una posible causa es que se ha abaratado mucho el proceso de crear películas en las últimas décadas, por lo que en la actualidad hay muchas más películas.

**Creación de variables temporales** A continuación, creamos nuevas variables temporales a partir de la fecha de valoración `Date`. Para cada valoración obtenemos el año, el mes, el día, la semana y el día de la semana.

```

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",
                                    "Agosto", "Septiembre", "Octubre", "Noviembre", "Diciembre"))

```

```

    "Agosto", "Septiembre", "Octubre", "Noviembre", "Diciembre")),
Day_D = day(Date),
Week_D = week(Date),
Day_of_week_D = factor(wday(Date),
                        levels = c(1,2,3,4,5,6,7),
                        labels = c("Domingo", "Lunes", "Martes", "Miercoles", "Jueves",
                                  "Viernes", "Sabado")))

```

**Información estadística de las valoraciones por película** Creamos una tabla con la cantidad, la suma, la media, la mediana, la moda y la desviación típica de las valoraciones de cada película.

```

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))

```

Se ha usado el paquete `formattable` para personalizar la tabla:

- *Cantidad de Valoraciones*: Se ha colocado una barra para cada película, la cual aumenta de tamaño en función de la cantidad de valoraciones que tiene.
- *Media*: La media de la valoración de cada película cambia de color en función de su valor. El rango de colores está entre el rojo (para los valores más bajos) y el verde (para los valores más altos).
- *Desviación Típica*: El color de la desviación típica se verá con mayor intensidad cuando el valor sea alto, mientras que a penas será perceptible cuando el valor sea bajo.

```

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", "Medi

```

**Análisis de las 5 películas más valoradas** Se realizan una serie de gráficos con el fin de observar de la mejor manera posible la distribución de las puntuaciones de las 5 películas más valoradas. Se ha creado un gráfico de barras agrupadas, un diagrama de cajas y bigotes, y un mapa de calor.

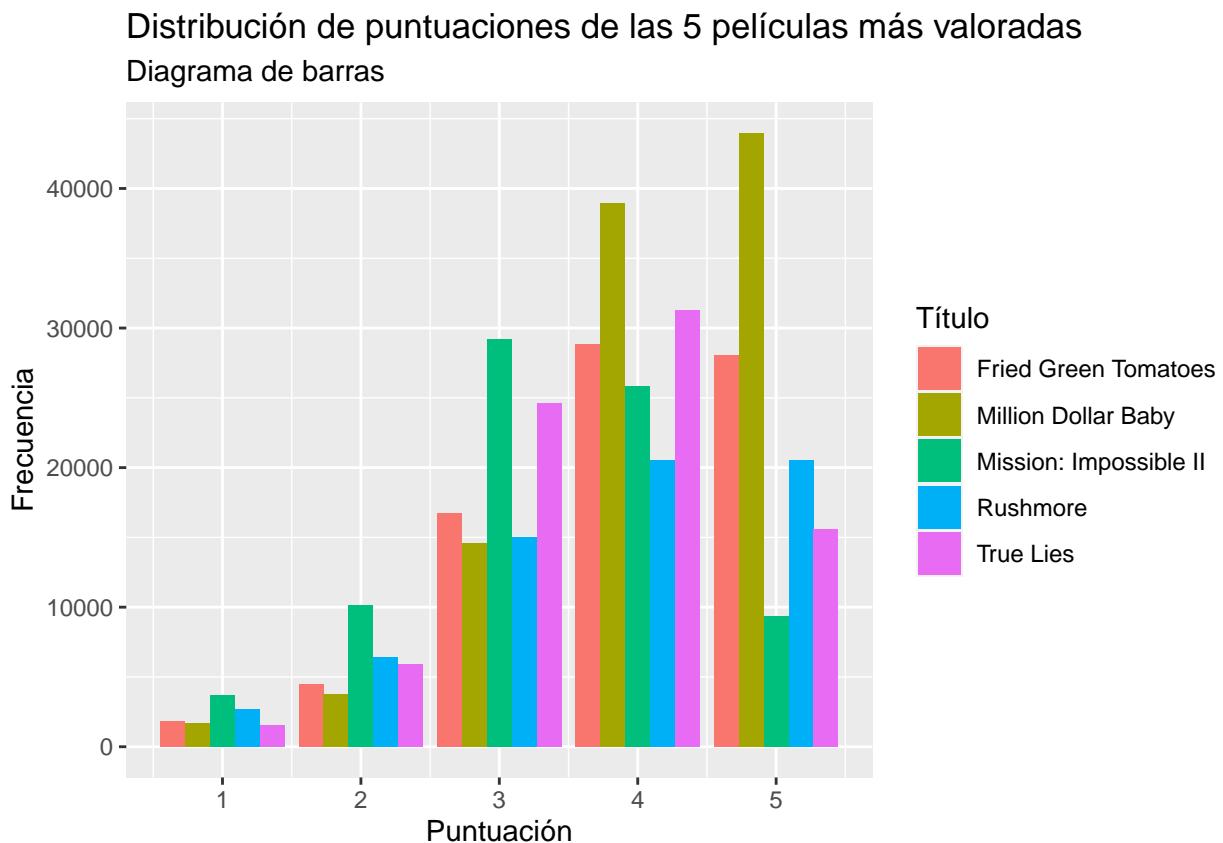
```

top5 <- films_table %>%
  top_n(5, count)

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"))
```

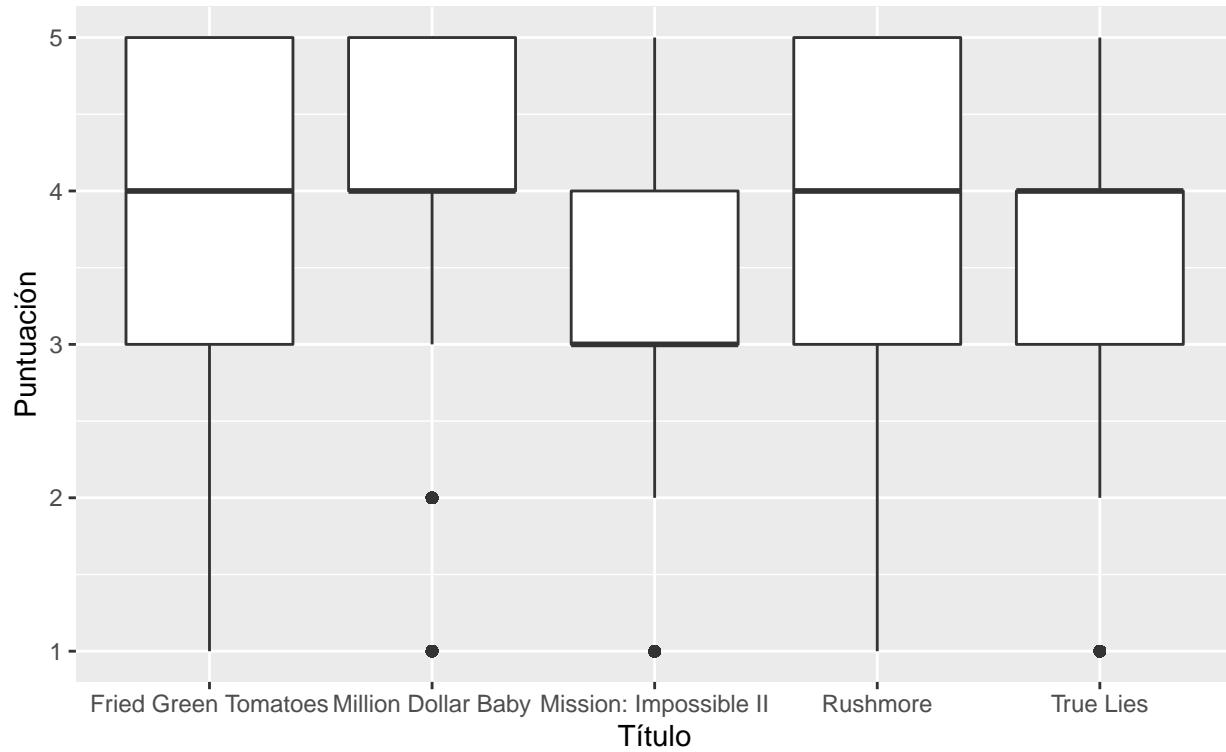


En el gráfico anterior se observa como las puntuaciones positivas predominan sobre las negativas. Sería más interesante realizar y comparar gráficos de densidades, pero la puntuación contiene valores discretos, por lo que no es posible realizarlos.

```
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")
```

## Distribución de puntuaciones de las 5 películas más valoradas

Diagrama de cajas y bigotes

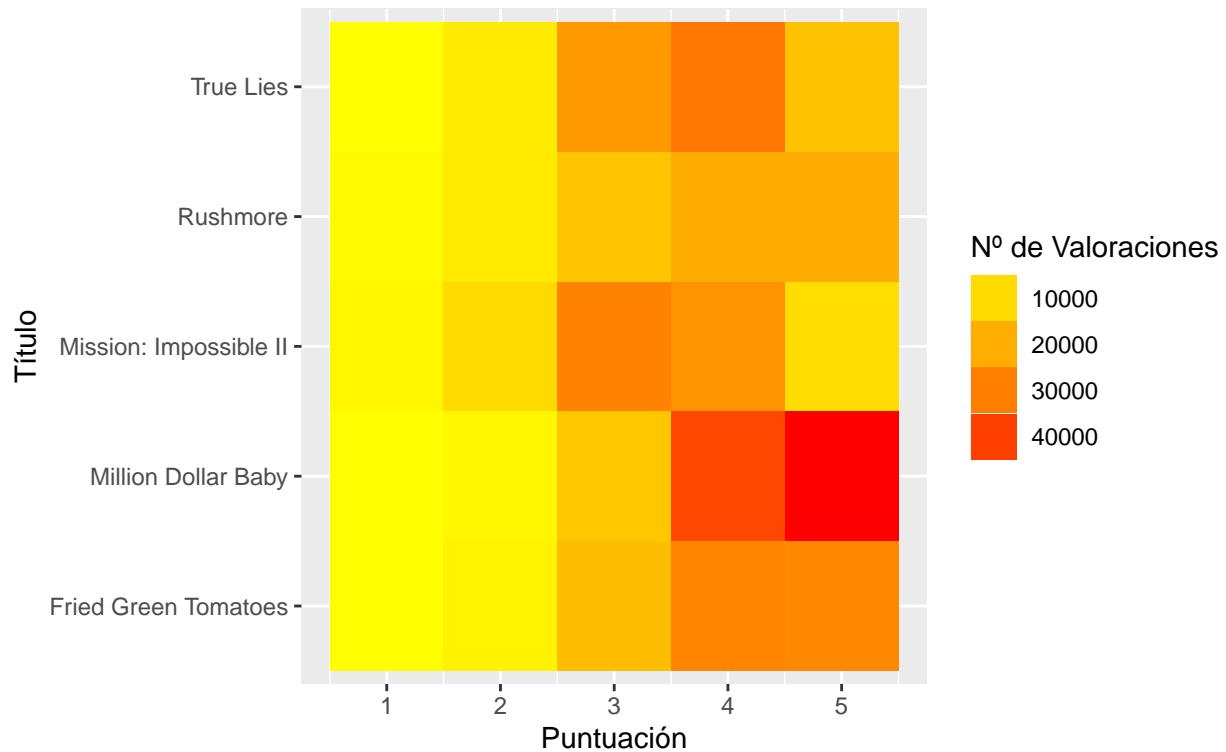


Con el diagrama de cajas y bigotes se pueden observar claramente las posiciones de los cuartiles, aunque la visualización es un poco extraña debido al problema anteriormente mencionado de valores discretos.

```
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"))
```

## Distribución de puntuaciones de las 5 películas más valoradas

### Mapa de calor

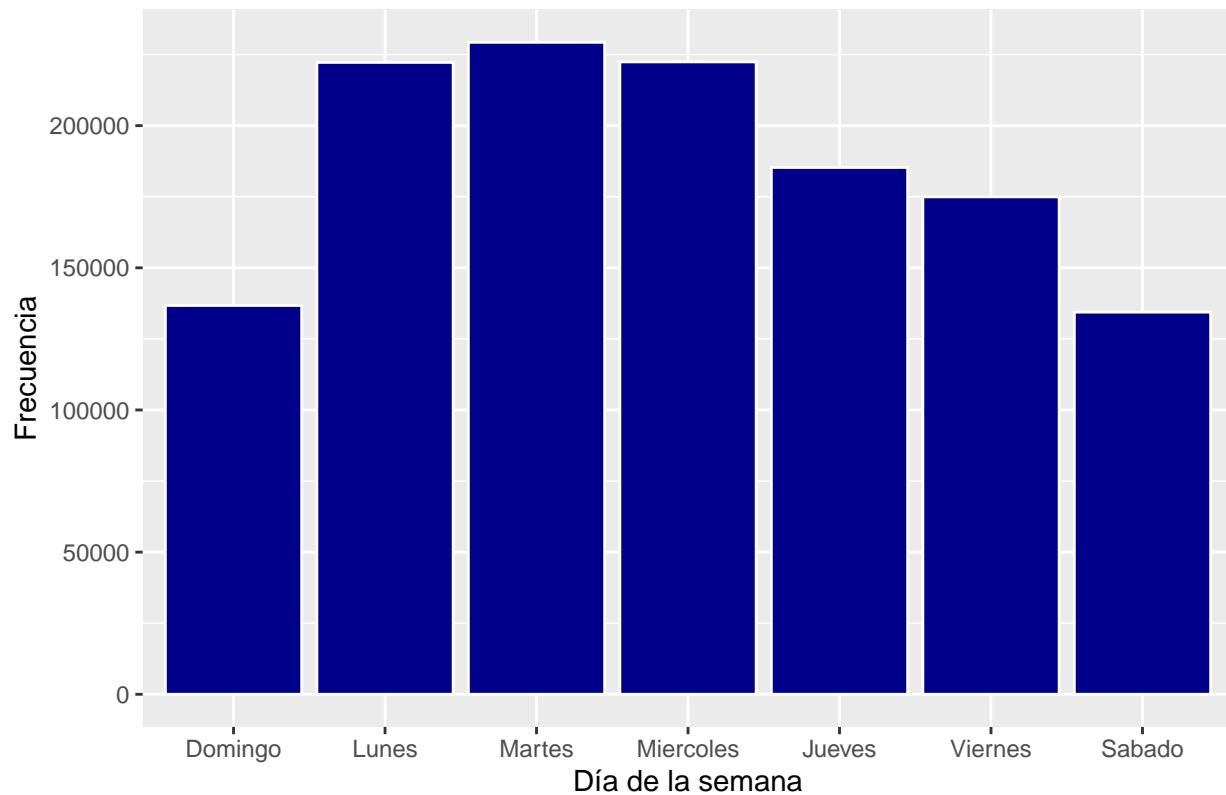


Usando un mapa de calor se visualiza la distribución de las puntuaciones de una manera muy sencilla y fácil de entender.

**Distribución de valoraciones por día de la semana y mes** En primer lugar se analiza la cantidad de valoraciones, y posteriormente si los valores de estas cambian según 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")
```

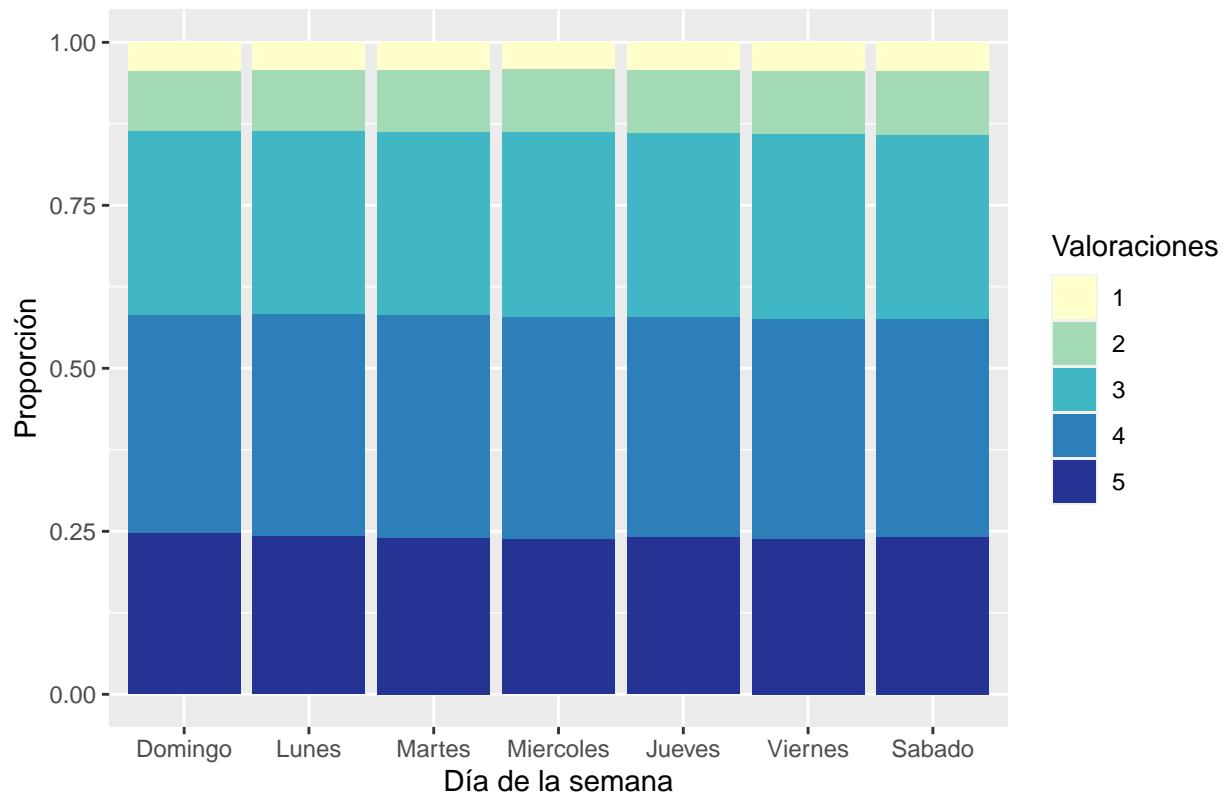
## Distribución de las puntuaciones por día de la semana



En el gráfico anterior se puede observar que de lunes a viernes se puntuhan más películas que durante el fin de semana. Una de las posibles causas es que durante el fin de semana los usuarios vean las películas acompañados por un grupo de amigos o tengan planes al acabar la película y por tanto no valoren la película al final de su visualización. Otra posible causa es que el uso de Netflix sea más frecuente entre semana, ya que la oferta de planes de ocio los fines de semana es mucho más amplia.

```
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")) +
  scale_fill_manual(values = brewer.pal(n = 5, name = "YlGnBu"))
```

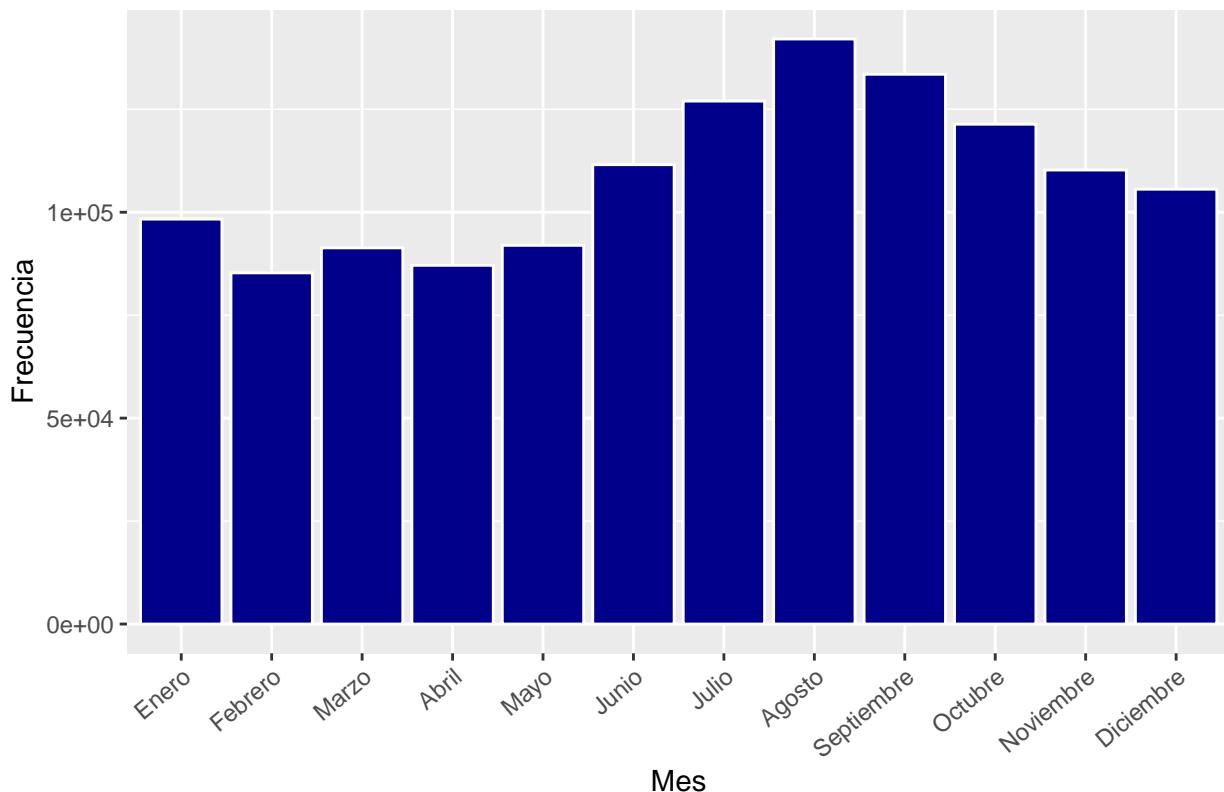
## Distribución de los valores de las puntuaciones por día de la semana



Como se puede ver, el dia de la semana en el cual se visiona la película no tiene efecto sobre el valor de la puntuación.

```
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))
```

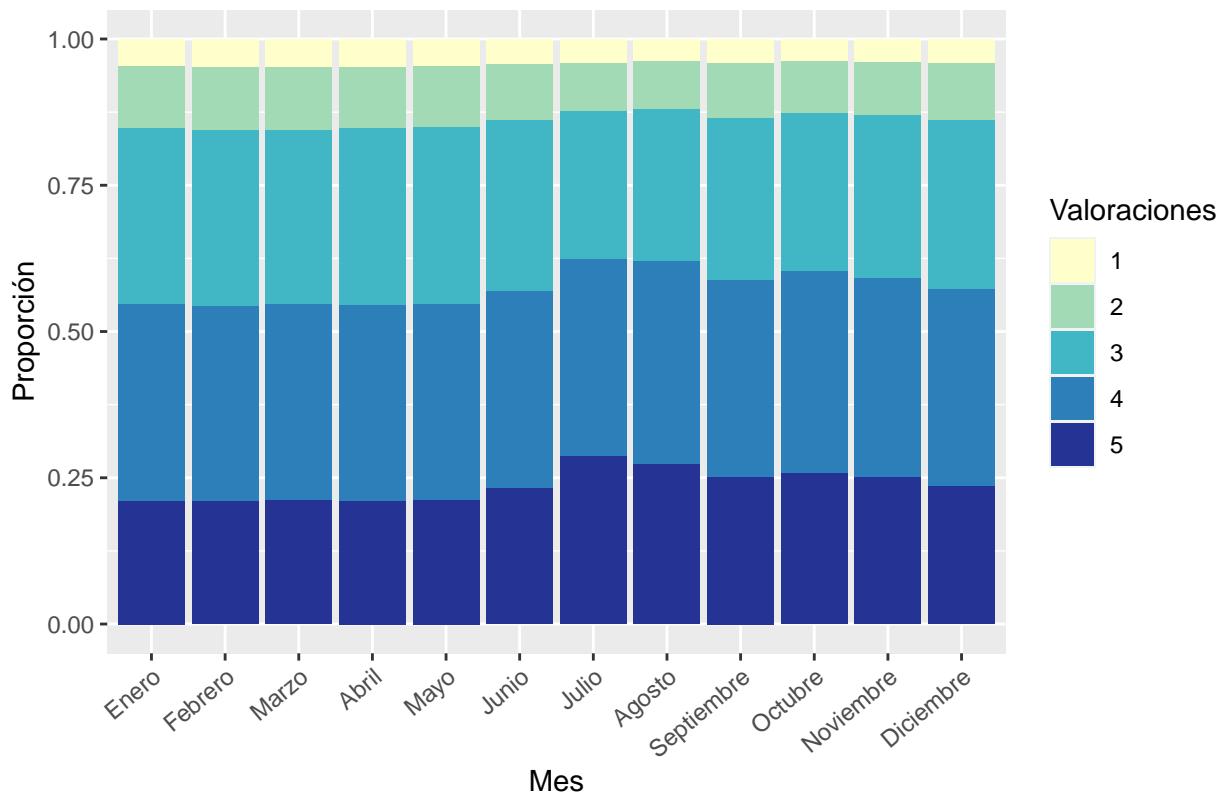
## Distribución de las puntuaciones por mes



En este gráfico se puede observar que el periodo en que más películas se ha puntuado es durante verano, este resultado es congruente ya que en verano normalmente la gente suele estar de vacaciones y tiene más tiempo para ver películas y puntuarlas.

```
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)) +
  scale_fill_manual(values = brewer.pal(n = 5, name = "YlGnBu"))
```

## Distribución de los valores de las puntuaciones por mes



Mientras que el día de la semana en que se puntuó la película no está correlacionado con la puntuación, aquí se observa que desde junio hasta final de año la puntuación de las películas es más generosa. Por tanto, el mes en el que se puntuó la película tiene relación con la puntuación.

**Tabla agrupada por película y año del número de valoraciones** Primero se genera una tabla con las 250 películas y sus puntuaciones distribuidas desde 1999 hasta 2005.

```
frec_table <- as.data.frame.matrix(table(data$title, data$Year_D)) %>%
  rownames_to_column("Título")

#frec_table %>%
#  formattable(align = c("l", "c", "c", "c", "c", "c", "c", "c"))
```

A continuación se escogen las diez películas con mayor número de valoraciones y se representa mediante un mapa de calor:

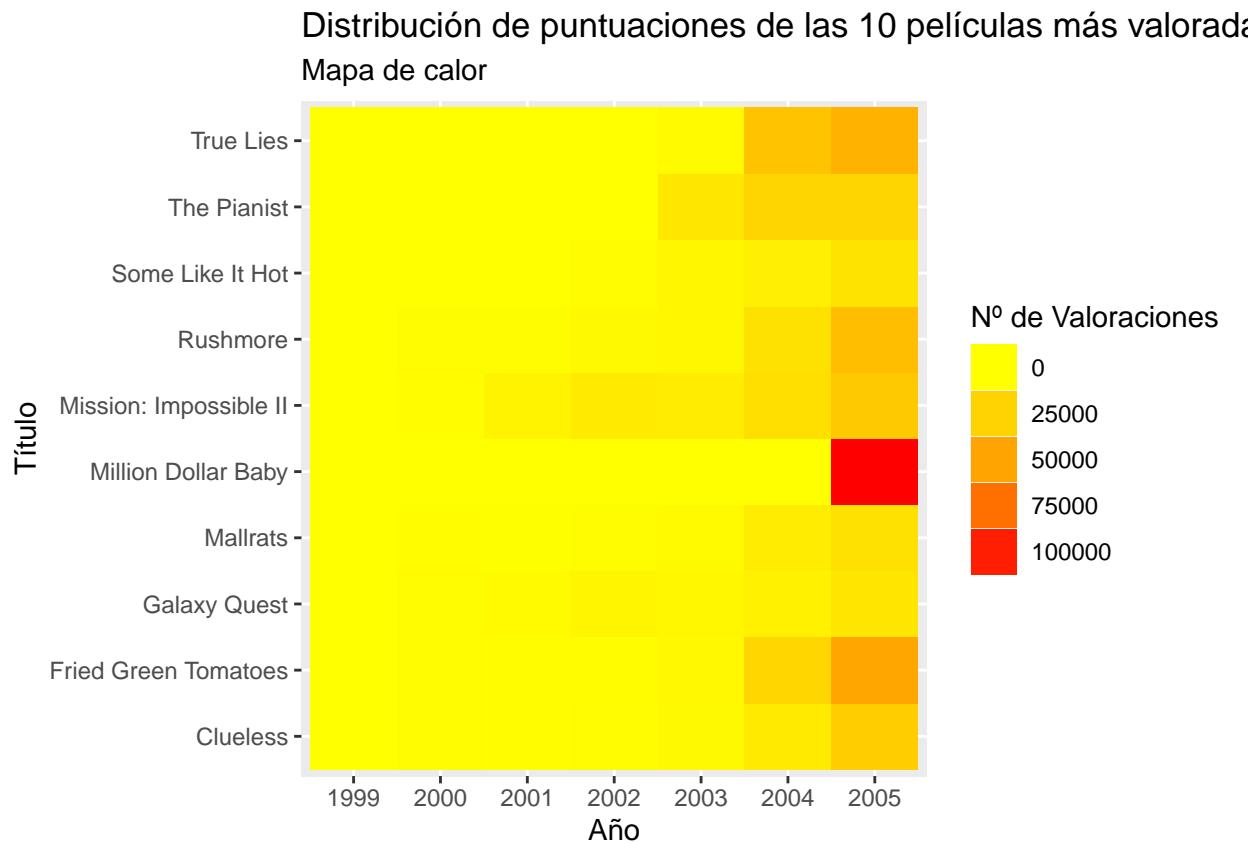
```
top10 <- frec_table %>%
  mutate(total = rowSums(.[2:8])) %>%
  top_n(10, total)

top10 %>%
  pivot_longer(c("1999", "2000", "2001", "2002", "2003", "2004", "2005"), names_to = "año", values_to =
```

```

xlab("Año") +
ylab("Título") +
ggtitle("Distribución de puntuaciones de las 10 películas más valoradas",
        "Mapa de calor") +
guides(fill = guide_legend(title = "Nº de Valoraciones"))

```



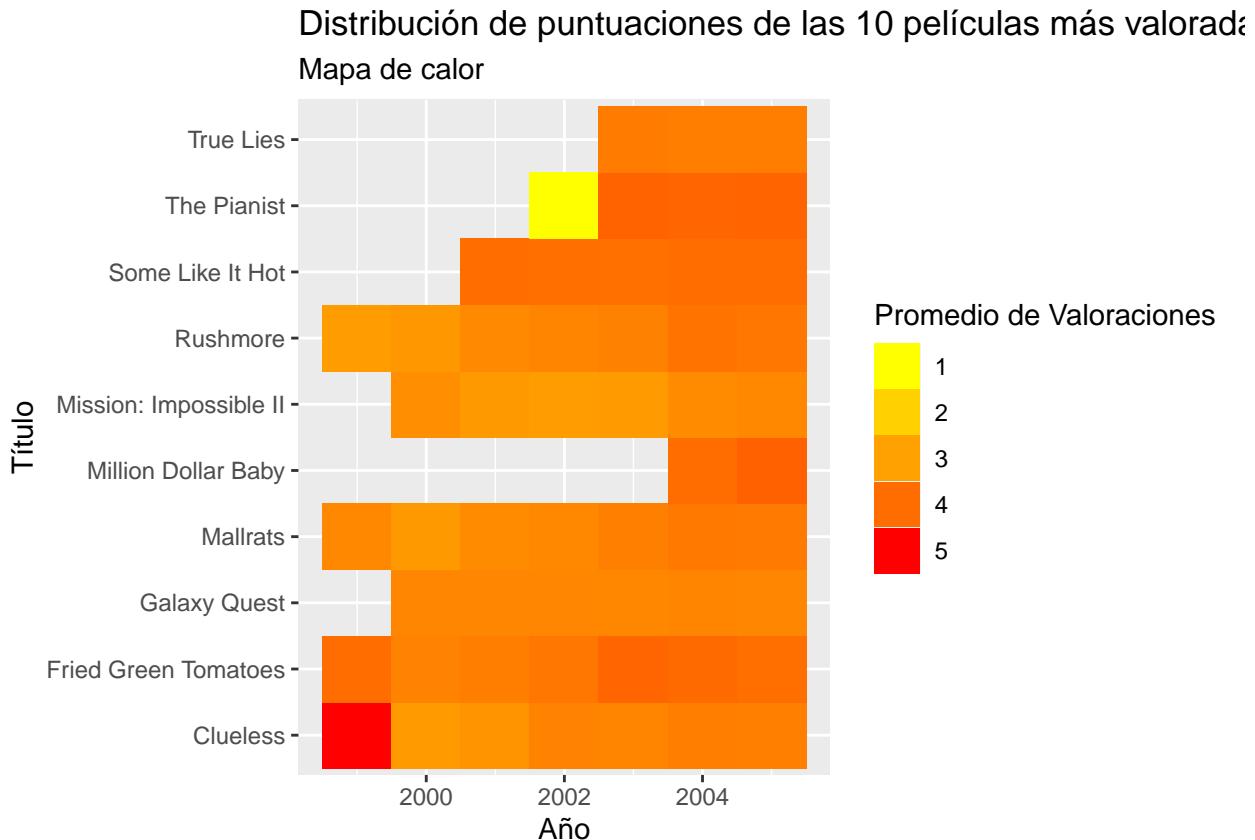
La cantidad de valoraciones de los primeros años es muy pequeña probablemente debido a dos factores: la película aun no estaba en Netflix y la cantidad de usuarios de la plataforma era mucho menor.

**Distribución del score promedio por año de las 10 películas más valoradas** Se crea un mapa de calor del promedio de las puntuaciones de las 10 películas anteriormente analizadas.

```

data %>%
  filter>Title %in% top10$Título) %>%
  group_by(Title, Year_D) %>%
  summarise(mean_score = mean(Score)) %>%
  ggplot() +
  geom_tile(aes(x = Year_D, y = Title, fill = mean_score)) +
  scale_fill_gradient(low="yellow", high="red") +
  xlab("Año") +
  ylab("Título") +
  ggtitle("Distribución de puntuaciones de las 10 películas más valoradas",
          "Mapa de calor") +
  guides(fill = guide_legend(title = "Promedio de Valoraciones"))

```



En el gráfico anterior se pueden observar 2 outliers, en concreto, en las películas “The Pianist” y “Clueless”. Nos fijamos en dichas observaciones:

```
data %>%
  filter((Title == "The Pianist" & Year_D == 2002) | (Title == "Clueless" & Year_D == 1999))

## # A tibble: 2 x 11
##   User Score Date      ID_film Release_Year Title Year_D Month_D Day_D Week_D
##   <dbl> <dbl> <date>    <dbl>       <dbl> <chr>  <dbl> <fct>   <int> <dbl>
## 1 423952     1 2002-12-08    2743       2002 The ~  2002 Diciem~     8    49
## 2 122223     5 1999-12-18    17482      1995 Clue~  1999 Diciem~    18    51
## # ... with 1 more variable: Day_of_week_D <fct>
```

Ambas son puntuaciones dadas por un único usuario y distorsionan la información del gráfico, por ello se procede a su eliminación.

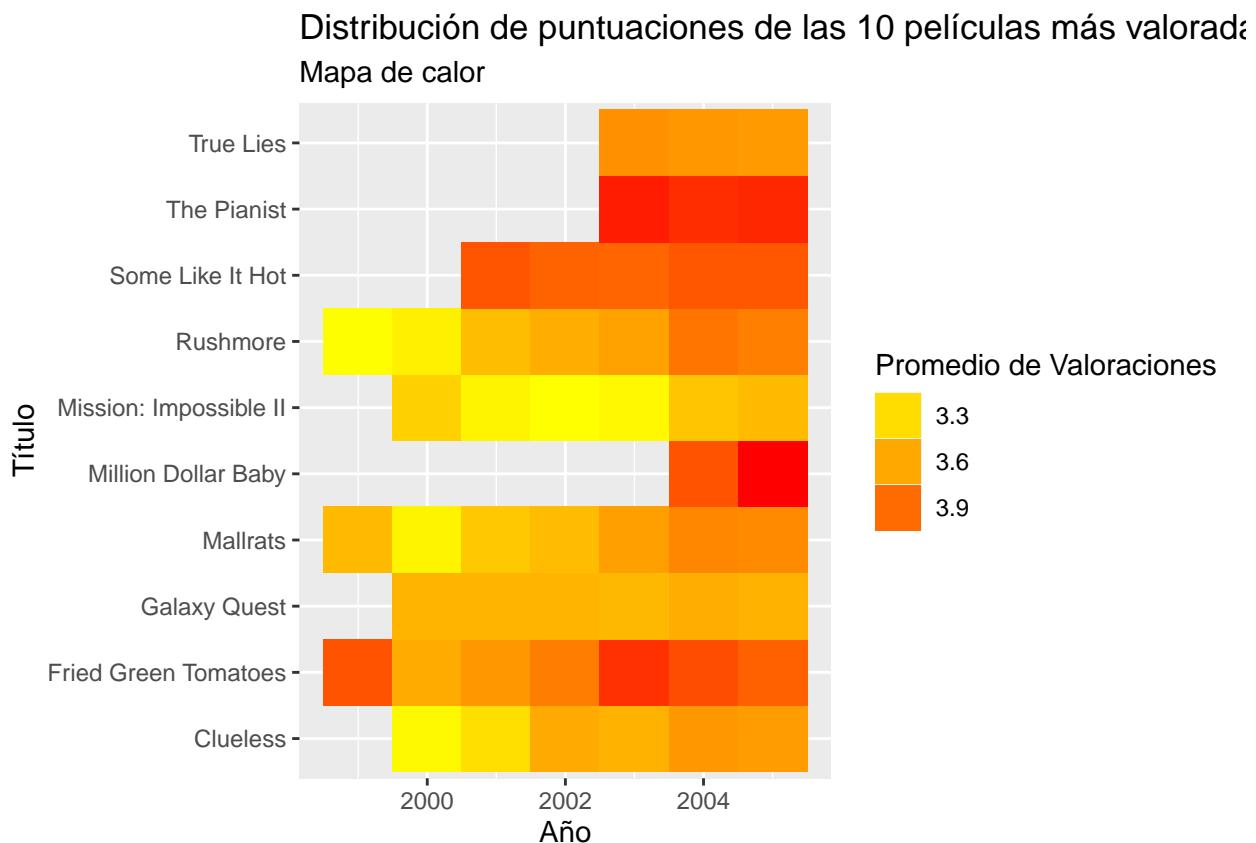
```
# Quitamos los outliers
data_2 <- data %>%
  filter((Title != "The Pianist" | Year_D != 2002), (Title != "Clueless" | Year_D != 1999))

# Rehacemos el gráfico
data_2 %>%
  filter(Title %in% top10$Título) %>%
  group_by(Title, Year_D) %>%
  summarise(mean_score = mean(Score)) %>%
```

```

ggplot() +
  geom_tile(aes(x = Year_D, y = Title, fill = mean_score)) +
  scale_fill_gradient(low="yellow", high="red") +
  xlab("Año") +
  ylab("Título") +
  ggtitle("Distribución de puntuaciones de las 10 películas más valoradas",
          "Mapa de calor") +
  guides(fill = guide_legend(title = "Promedio de Valoraciones"))

```



Ahora se puede observar con mayor precisión la distribución de las puntuaciones de las diez películas más valoradas.

**Análisis del usuario** A continuación, se analiza el comportamiento de los usuarios que han valorado al menos una película.

```

users <- data %>%
  group_by(User) %>%
  summarise(count = n(),
           mean_score = mean(Score))

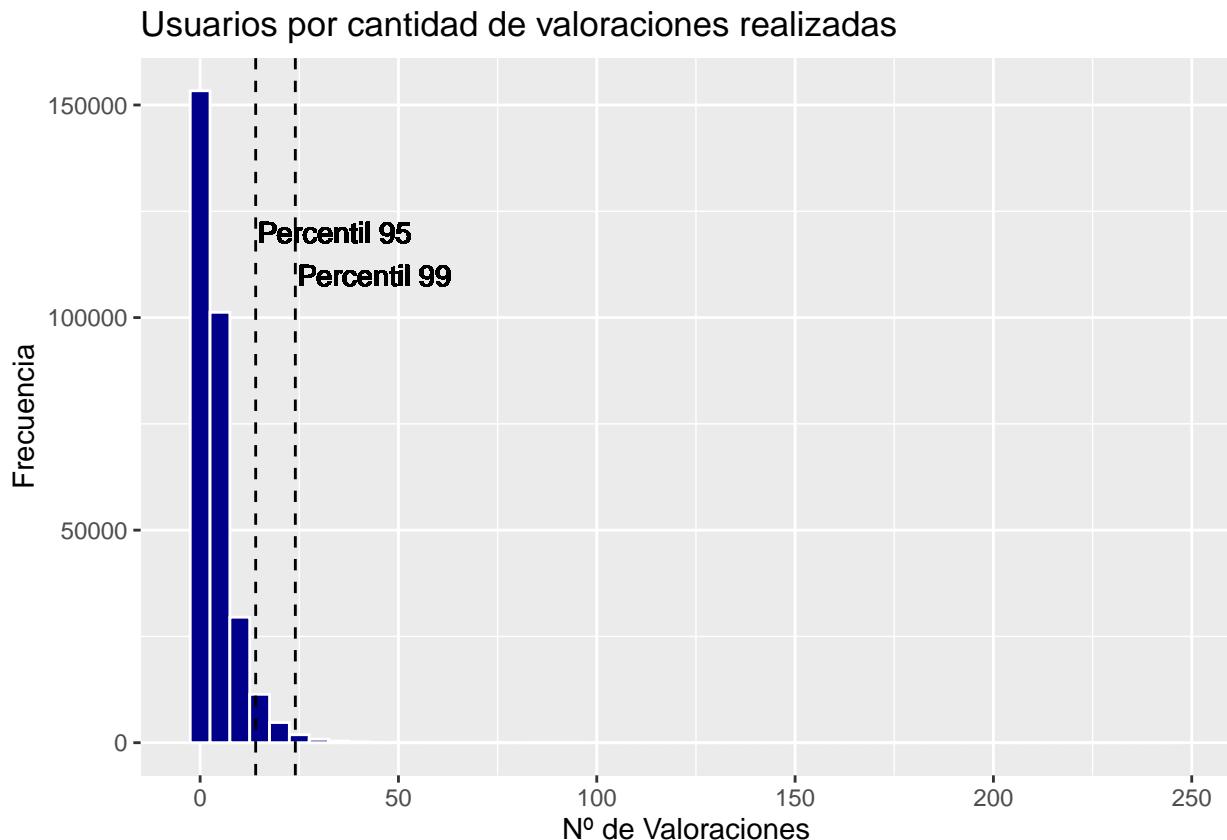
ggplot(users) +
  geom_histogram(mapping = aes(x = count), binwidth = 5, fill = "darkblue", col = "white") +
  geom_vline(xintercept = quantile(users$count, 0.95), linetype = "dashed") +
  geom_vline(xintercept = quantile(users$count, 0.99), linetype = "dashed") +
  geom_text(aes(x = quantile(count, 0.99) + 20, y = 110000, label = "Percentil 99"))

```

```

geom_text(aes(x = quantile(count, 0.95) + 20, y = 120000, label = "Percentil 95")) +
xlab("Nº de Valoraciones") +
ylab("Frecuencia") +
ggtitle("Usuarios por cantidad de valoraciones realizadas")

```



En el histograma se observa claramente que la gran mayoría de usuarios han valorado pocas películas, parece que hay outliers a partir de 50 películas valoradas. Las rectas verticales son los percentiles 95 y 99.

Vamos a ver como se distribuye la nota media de las valoraciones de los usuarios en función de la cantidad de películas valoradas.

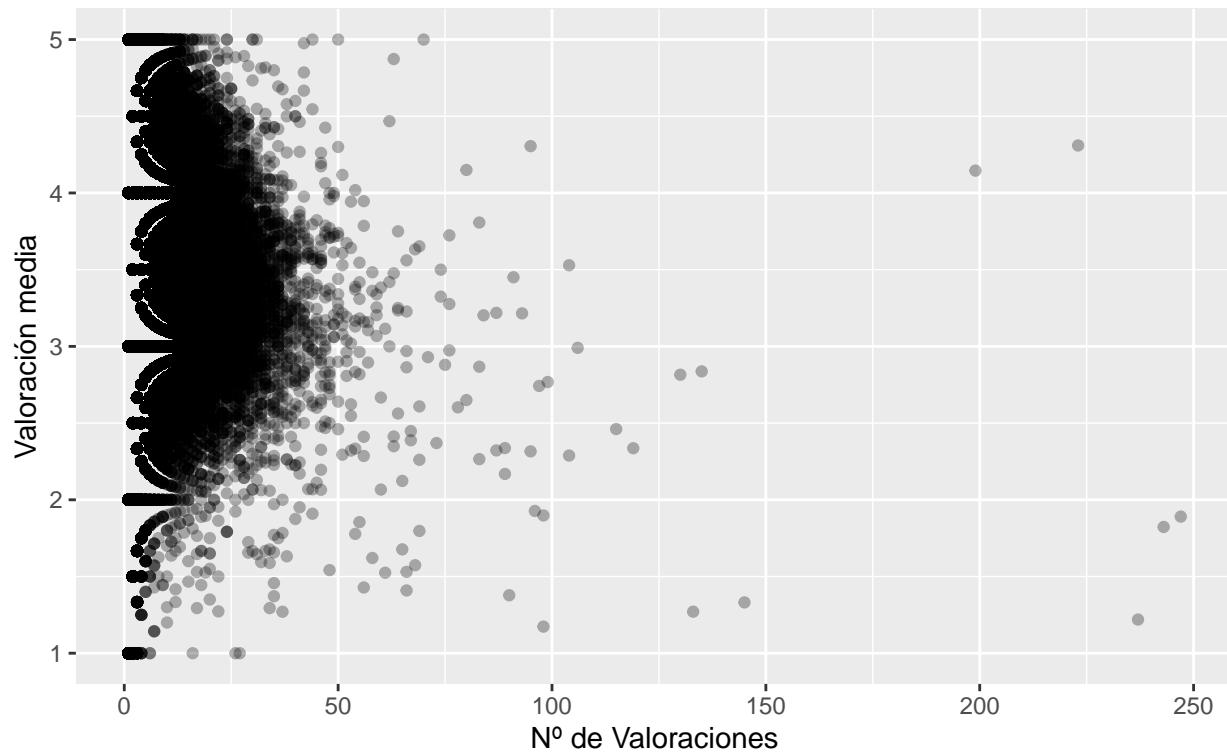
```

ggplot(users) +
  geom_point(mapping = aes(x = count, y = mean_score), alpha = 0.3) +
  xlab("Nº de Valoraciones") +
  ylab("Valoración media") +
  ggtitle("Valoración media de usuarios en función de su cantidad de valoraciones",
          "Diagrama de dispersión")

```

## Valoración media de usuarios en función de su cantidad de valoraciones

### Diagrama de dispersión

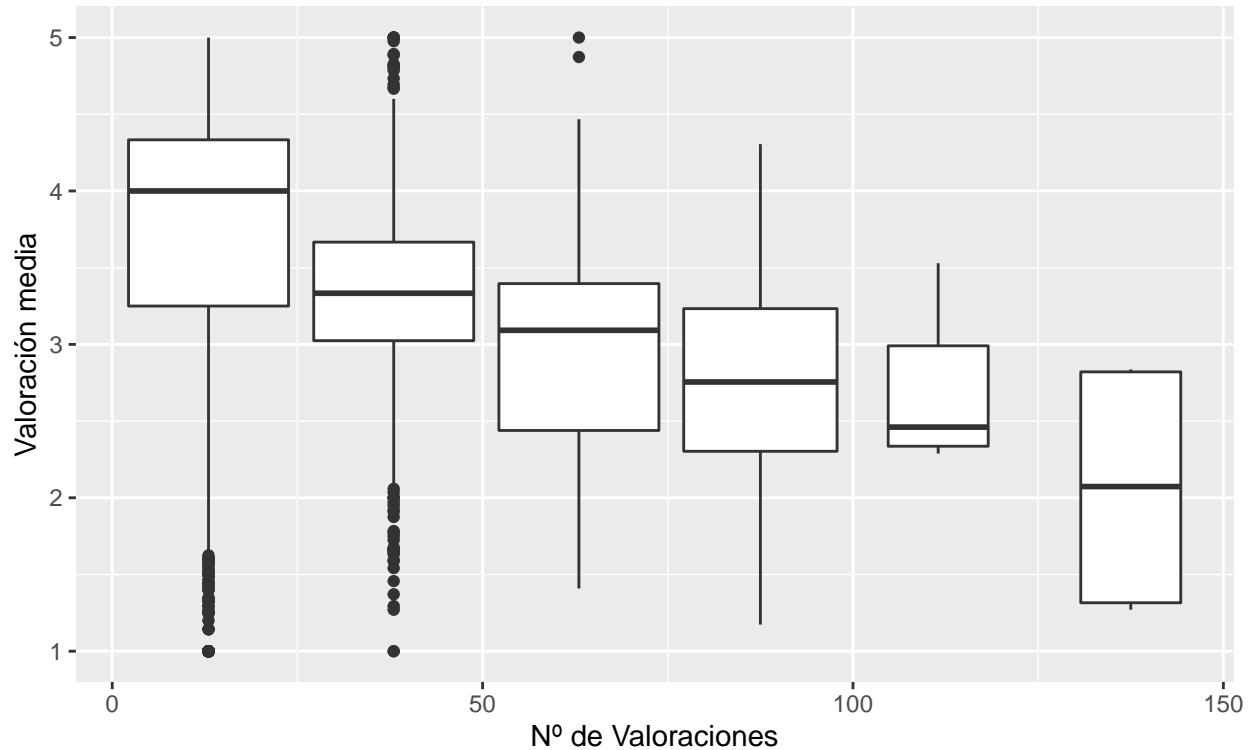


No se ve con claridad ninguna relación, aunque parece que la varianza disminuye con la cantidad de películas valoradas. Para buscar algún tipo de relación se realiza un diagrama de cajas y bigotes, hemos observado que únicamente tenemos 5 usuarios que hayan valorado más de 150 películas, por lo que no los incluiremos en el diagrama.

```
users %>%
  filter(count < 150) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = count, y = mean_score, group = cut(count, breaks = seq(0, 150, 25)))) +
  xlab("Nº de Valoraciones") +
  ylab("Valoración media") +
  ggtitle("Valoración media de usuarios en función de su cantidad de valoraciones",
          "Diagrama de cajas y bigotes")
```

## Valoración media de usuarios en función de su cantidad de valoraciones

### Diagrama de cajas y bigotes

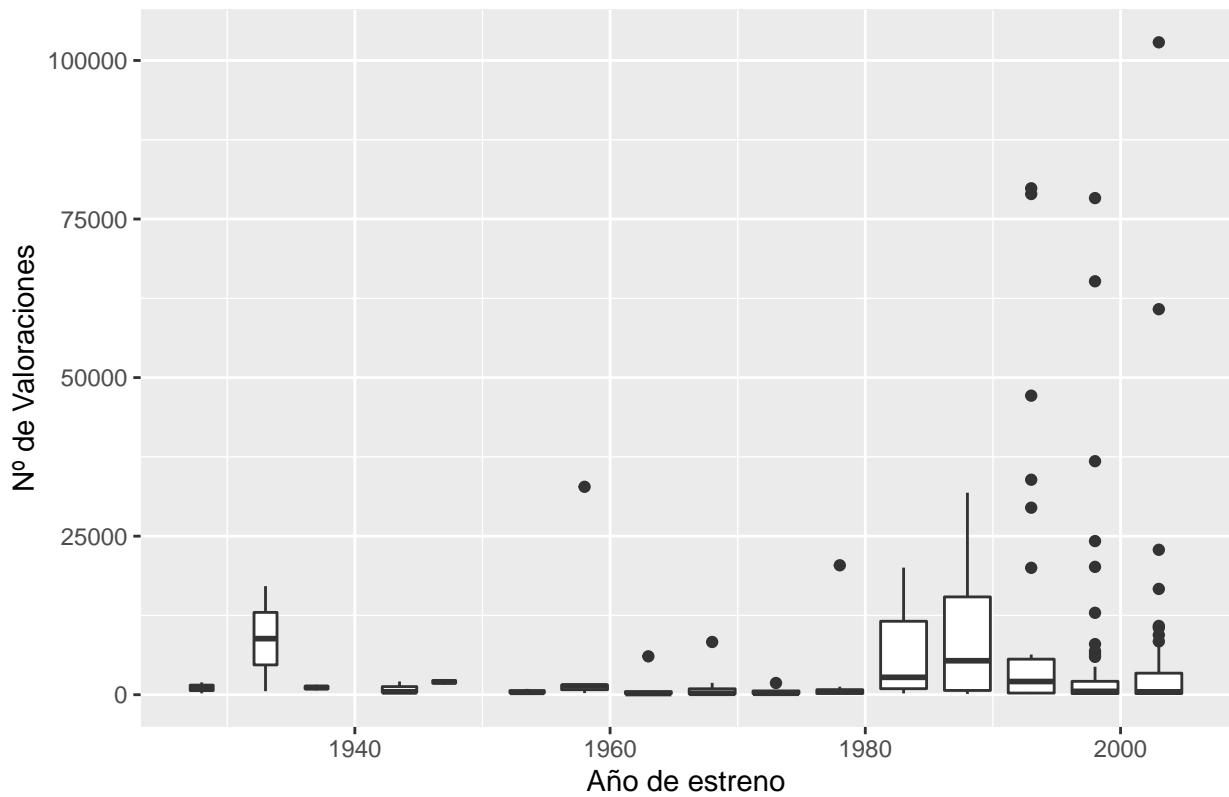


Con este gráfico si que vemos como a mayor número de valoraciones, más baja es la nota media. Esto puede ser debido a que los usuarios que más valoren sean más críticos, por lo que votarán más a la baja. También puede ser que los individuos que hayan valorado pocas películas hayan valorado las mejores, así que su valoración media es alta.

**Análisis de las películas por antigüedad** Vamos a ver si las películas antiguas se valoran más o menos que las nuevas. Evidentemente se valoran más las películas actuales debido a que hay más, por lo que nos fijaremos en la cantidad media de valoraciones que tienen las películas.

```
data %>%
  group_by(Title, Release_Year) %>%
  summarise(count = n(), mean_score = mean(Score)) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = Release_Year, y = count, group = cut(Release_Year, breaks = seq(1925, 2015, 10)))) +
  xlab("Año de estreno") +
  ylab("Nº de Valoraciones") +
  ggtitle("Cantidad de valoraciones de películas por año de estreno")
```

## Cantidad de valoraciones de películas por año de estreno

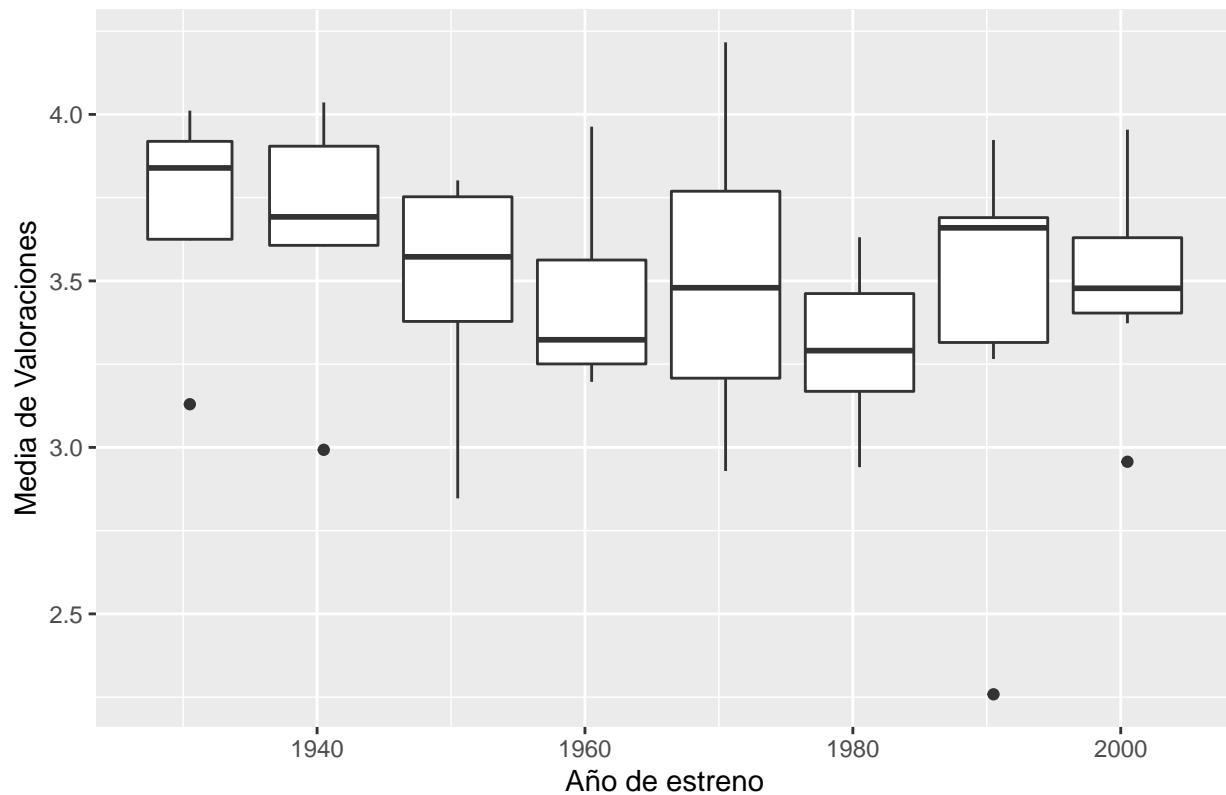


Se observa que la mayoría de las películas no están muy valoradas, aunque hay algunas modernas que han sido extremadamente valoradas, sobretodo películas de los años 90 a la actualidad. Es interesante que, en este intervalo de tiempo, la mediana es igualmente muy baja; seguramente sea debido a que las películas estrella son muy valoradas, pero también hay una gran cantidad de películas poco conocidas. Cabe señalar que entre el 1980 y 1995 la mediana de las valoraciones es relativamente alta, puede que sea por la muestra aleatoria de 250 películas o porque Netflix solo ha adquirido las más populares de la época.

Para terminar, buscamos alguna relación del valor de la valoración con el año de estreno de la película. La media se calcula a partir de todas las puntuaciones de las películas de ese año.

```
data %>%
  group_by(Release_Year) %>%
  summarise(mean_score = mean(Score)) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = Release_Year, y = mean_score, group = cut(Release_Year, breaks = seq(1935, 2005, 5)))) +
  xlab("Año de estreno") +
  ylab("Media de Valoraciones") +
  ggtitle("Valor medio de las valoraciones por año de estreno")
```

## Valor medio de las valoraciones por año de estreno



Se observa que las películas más antiguas están mejor valoradas que las más recientes, aunque esto puede ser debido a que son películas de calidad, ya que sino Netflix no las incluiría en su catálogo. Aun así, son pocas las películas antiguas y tienen pocas visualizaciones, por lo que la media de las puntuaciones podría estar sesgada. Anteriormente hemos visto que las películas entre 1980 y 1995 han sido muy valoradas, y aquí vemos una tendencia positiva en sus puntuaciones.

**Conclusiones** Se ha analizado desde diferentes perspectivas los datos de las valoraciones de Netflix. En un futuro, sería interesante incluir datos como el sexo, la edad, el país de los usuarios y otros datos socioeconómicos con el fin de poder diferenciar mejor sus preferencias. A nivel de las películas también estaría bien ver el número total de visualizaciones que no han sido puntuadas, la productora cinematográfica a la cual pertenecen, la duración de la película... con el fin de analizar con mayor precisión las valoraciones de las películas.

Para un futuro análisis, se podría aplicar algún modelo de agrupación o de predicción sobre los usuarios o las películas.