Visualización Avanzada IMDB

Javier Aos Aragonés

11/05/2022

Introducción

A lo largo de la historia del cine se han creado numerosas obras cinematográficas de diferentes características. Algunas de estas más largas, otras más cortas, pertenecientes a géneros distintos, con valoraciones y recaudaciones diferentes etc. Dentro de esta gran variedad el espectador debe elegir la obra de la que desea disfrutar, ya sea a la hora de ir al cine, de comprar la película físicamente o incluso de elegir entre las muchas opciones que las plataformas digitales ofrecen.

Todo esto requiere de un desembolso económico previo, lo que hace que la industria del cine se convierta en un negocio y por ello que dentro del propio negocio se pueda analizar qué tipo de películas funcionan mejor atendiendo a las características de cada una. En este trabajo se analizarán los datos de las 1000 películas mejor valoradas en IMDB (Internet Movie Data Base). A partir de los datos brindados por esta página, podremos tener un mejor entendimiento sobre "qué hace a una película una buena película".

Análisis exploratorio de los datos

Cargamos el dataset

```
df <- read.csv("https://raw.githubusercontent.com/Javieraos/data/main/imdb_top_1000.csv?raw=true")
```

Limpieza de datos

En primer lugar, vamos a realizar una breve limpieza y manipulación de los datos para posteriormente poder realizar las visualizaciónes de manera más óptima.

```
df$Poster_Link = NULL # Eliminamos las columnas innecesarias
df$Overview = NULL
df[df == ""] = NA # Sustituimos valores vacios por "NA"
df$Gross = gsub(",", "", df$Gross) # Transformamos las variables para poder trabajar con ellas
df$Gross = as.integer(df$Gross)
df$Runtime = gsub(" min$", "", df$Runtime)
df$Runtime = as.integer(df$Runtime)
df$Certificate = as.factor(df$Certificate)
df$Released_Year = as.integer(df$Released_Year)
df$Genre = gsub(" ", "", df$Genre)
str(df)
```

```
## 'data.frame':    1000 obs. of  14 variables:
```

```
##  $ Series_Title : chr  "The Shawshank Redemption" "The Godfather" "The Dark Knight" "The Godfather: I
##  $ Released_Year: int  1994 1972 2008 1974 1957 2003 1994 1993 2010 1999 ...
##  $ Certificate  : Factor w/ 16 levels "16","A","Approved",..: 2 2 15 2 13 13 2 2 15 2 ...
##  $ Runtime      : int  142 175 152 202 96 201 154 195 148 139 ...
##  $ Genre        : chr  "Drama" "Crime,Drama" "Action,Crime,Drama" "Crime,Drama" ...
##  $ IMDB_Rating  : num  9.3 9.2 9 9 9 8.9 8.9 8.9 8.8 8.8 ...
##  $ Meta_score   : int  80 100 84 90 96 94 94 94 74 66 ...
##  $ Director     : chr  "Frank Darabont" "Francis Ford Coppola" "Christopher Nolan" "Francis Ford Copp
##  $ Star1        : chr  "Tim Robbins" "Marlon Brando" "Christian Bale" "Al Pacino" ...
##  $ Star2        : chr  "Morgan Freeman" "Al Pacino" "Heath Ledger" "Robert De Niro" ...
##  $ Star3        : chr  "Bob Gunton" "James Caan" "Aaron Eckhart" "Robert Duvall" ...
##  $ Star4        : chr  "William Sadler" "Diane Keaton" "Michael Caine" "Diane Keaton" ...
##  $ No_of_Votes  : int  2343110 1620367 2303232 1129952 689845 1642758 1826188 1213505 2067042 185474
##  $ Gross        : int  28341469 134966411 534858444 57300000 4360000 377845905 107928762 96898818 29
```

Una vez realizada la limpieza de los datos podemos ver que el dataset consta de 1000 observaciones de las vuales tenemos 14 variables para cada una. Estas variables son:

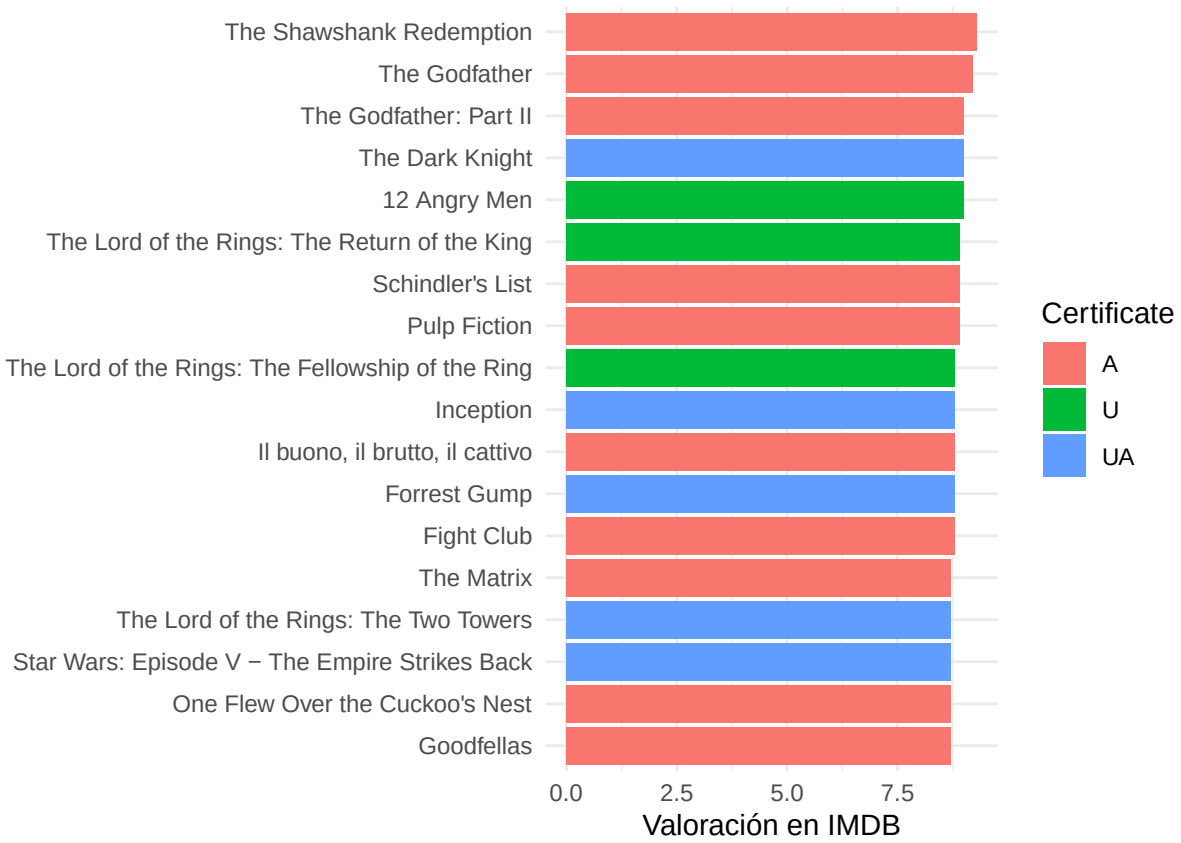
1. Series_Title: Nombre de la película
2. Released_Year: Año en el que se estrenó
3. Certificate: Certificado de la película
4. Runtime: Duración total
5. Genre: Géneros
6. IMDB_Rating: Valoración de la película en IMDB
7. Meta_score: Valoración de la película según críticos de cine
8. Director: Nombre del director
9. Star1, Star2, Star3, Star4: Nombre de las "estrellas"
10. No_of_Votes: Número total de votos
11. Gross: Recaudación total generada por la película

Como podemos ver, tenemos varios datos a tener en cuenta si queremos valorar a una película como mejor que otra. Dependiendo de los criterios que seleccionemos tendremos unas películas u otras como ganadoras.

Visualización de los datos

Veamos cuales son las mejores películas según su valoración en IMDB.

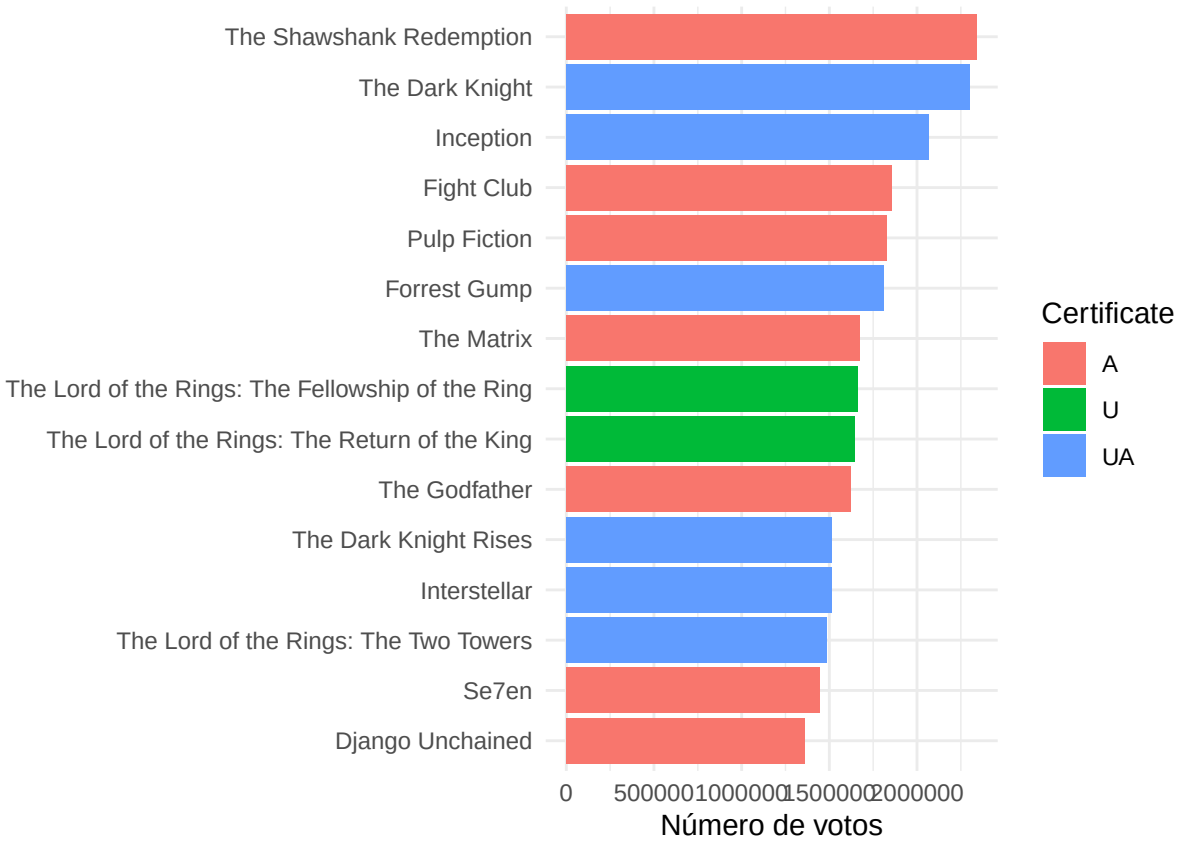
```
df %>% top_n(15, wt=IMDB_Rating) %>%
ggplot(aes(reorder(Series_Title, IMDB_Rating), IMDB_Rating, fill = Certificate)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x="", y="Valoración en IMDB") +
  theme_minimal()
```



Vemos que el gráfico de barras no muestra demasiada dispersión entre las 15 películas mejor valoradas, teniendo como ganadora a "The Shawshank Redemption" seguida muy de cerca por "The Godfather". También podemos observar que la categoría más común entre las películas mejor valoradas es la A, seguida de la UA y la U.

Veamos que ocurre si seleccionamos como criterio el número de votos.

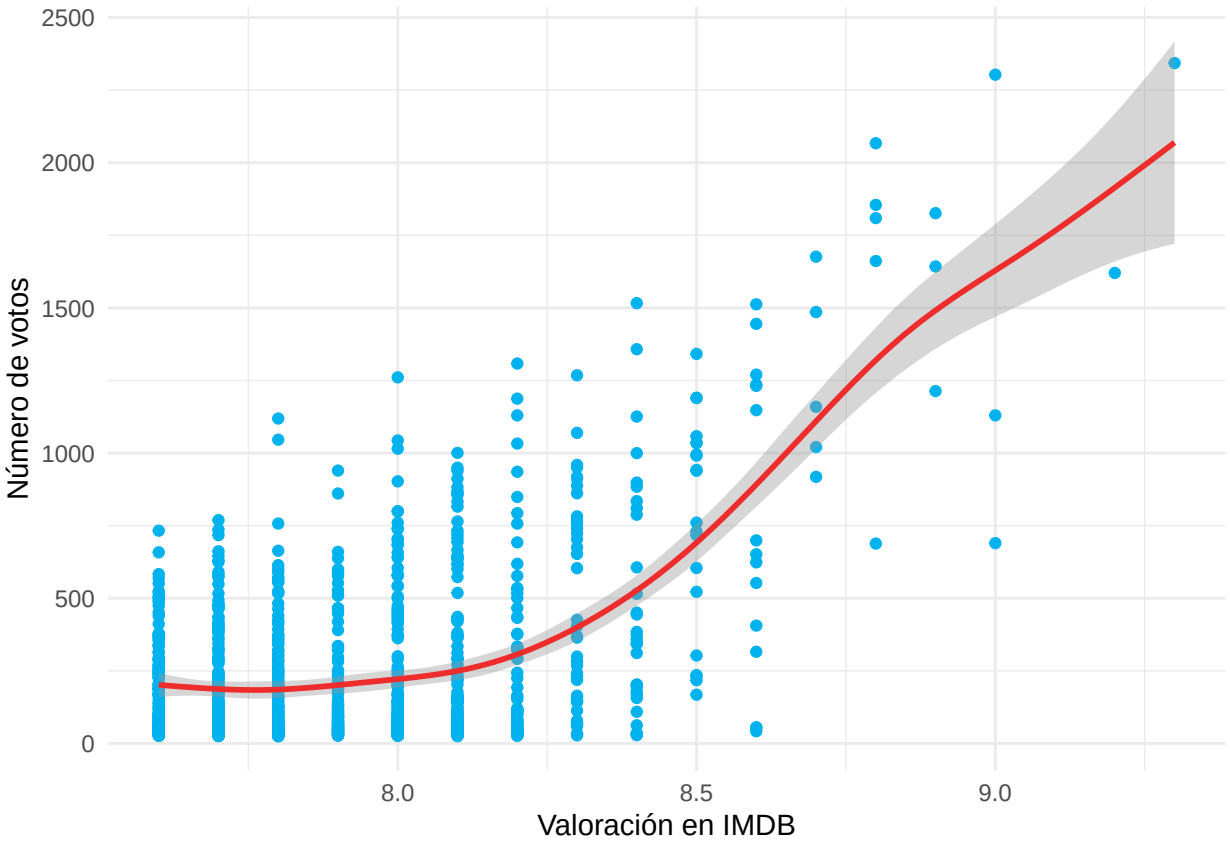
```
df %>% top_n(15, wt=No_of_Votes) %>%
ggplot(aes(reorder(Series_Title, No_of_Votes), No_of_Votes, fill = Certificate)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x="", y="Número de votos") +
  theme_minimal()
```



Como podemos ver en este gráfico de barras existe una dispersión algo mayor. Aunque tenemos como ganadora otra vez a "The Shawshank Redemption". Si nos fijamos, varias películas que aprecieron en el primer gráfico también aparecen aquí, como puede ser "The Dark Knight", "Pulp Fiction" o "Inception" entre otras. Esto nos indica que podría existir una relación positiva entre estas dos variables, en la que a mayor número de votos mayor valoración obtenrá la película y viceversa.

Vamos a intentar visualizar mejor esta relación con una gráfica de puntos.

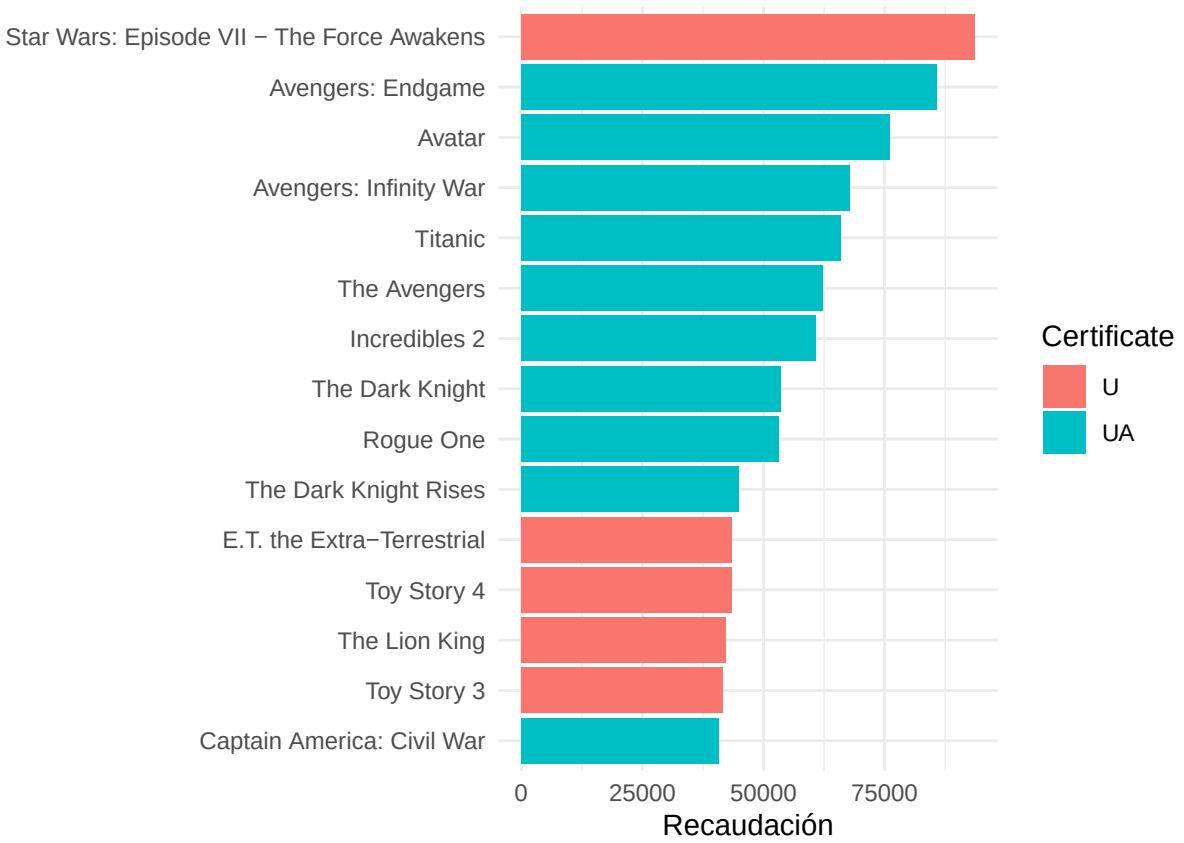
```
ggplot(df, aes(x=IMDB_Rating, y=No_of_Votes/1000)) +
  geom_point(color="deepskyblue2") +
  geom_smooth(color="firebrick2") +
  labs(
    x="Valoración en IMDB",
    y="Número de votos"
  ) +
  theme_minimal()
```



Ahora sí, podemos ver de manera más clara que existe una relación positiva entre la valoración en IMDB y el número de votos que ha recibido la película.

Pero, siendo honesto, para la industria cinematográfica, lo que realmente hace una película buena y por tanto, el criterio que se suele utilizar para medir el éxito o fracaso de los nuevos estrenos es la recaudación generada por la película. Si una película no "da dinero" desgraciadamente se considerará un fracaso. Por tanto, vamos a ver cuales son las películas con mayor recaudación.

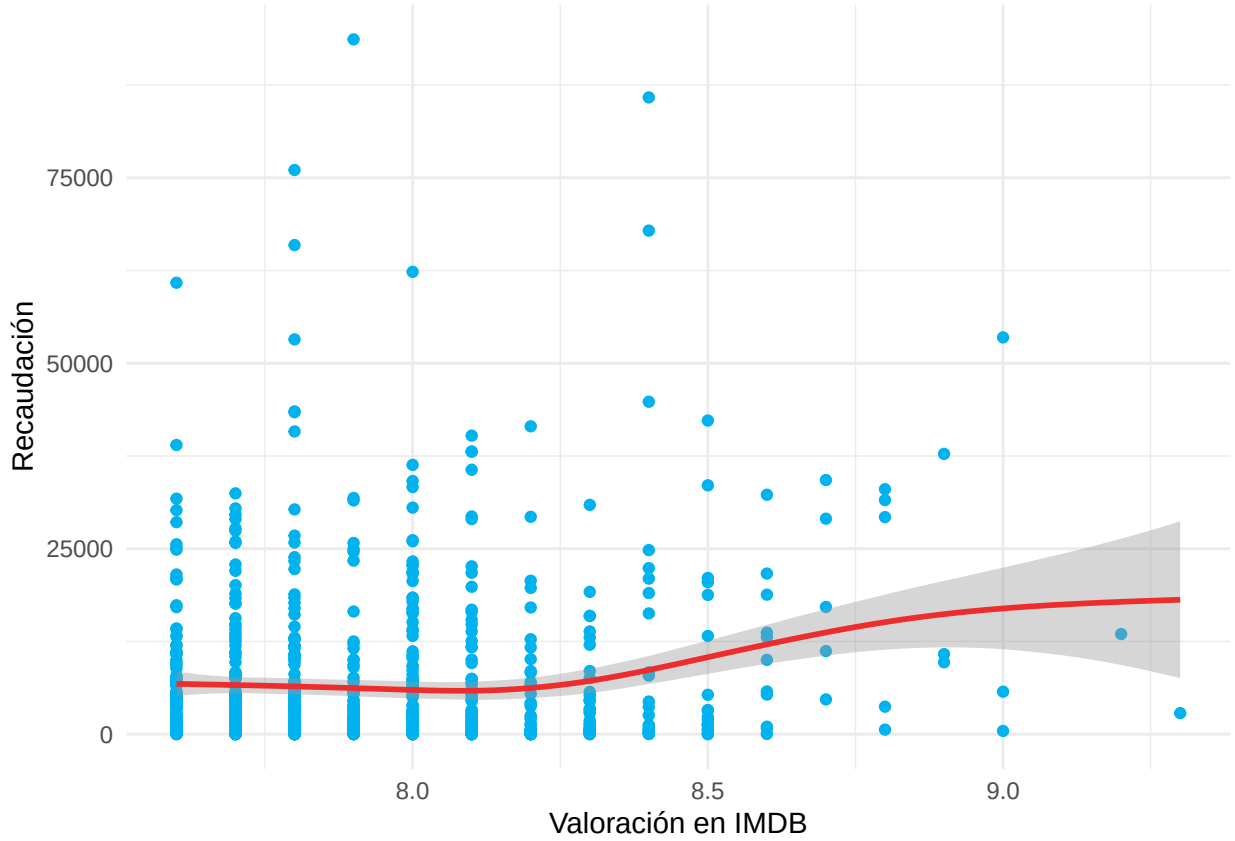
```
df %>% top_n(15, wt=Gross) %>%
ggplot(aes(reorder(Series_Title, Gross), Gross/10000, fill = Certificate)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x="", y="Recaudación") +
  theme_minimal()
```



En el gráfico podemos ver hay una dispersión incluso mayor que en los dos anteriores. Vemos que los certificados que más recaudan son el UA y el U. Además vemos que la mayoría de películas no coinciden con los otros gráficos. Es decir, una película puede estar bien valorada y además tener un buen número de votos pero eso no garantiza que consiga una buena recaudación, lo cual podría indicar que no existe correlación entre estas variables.

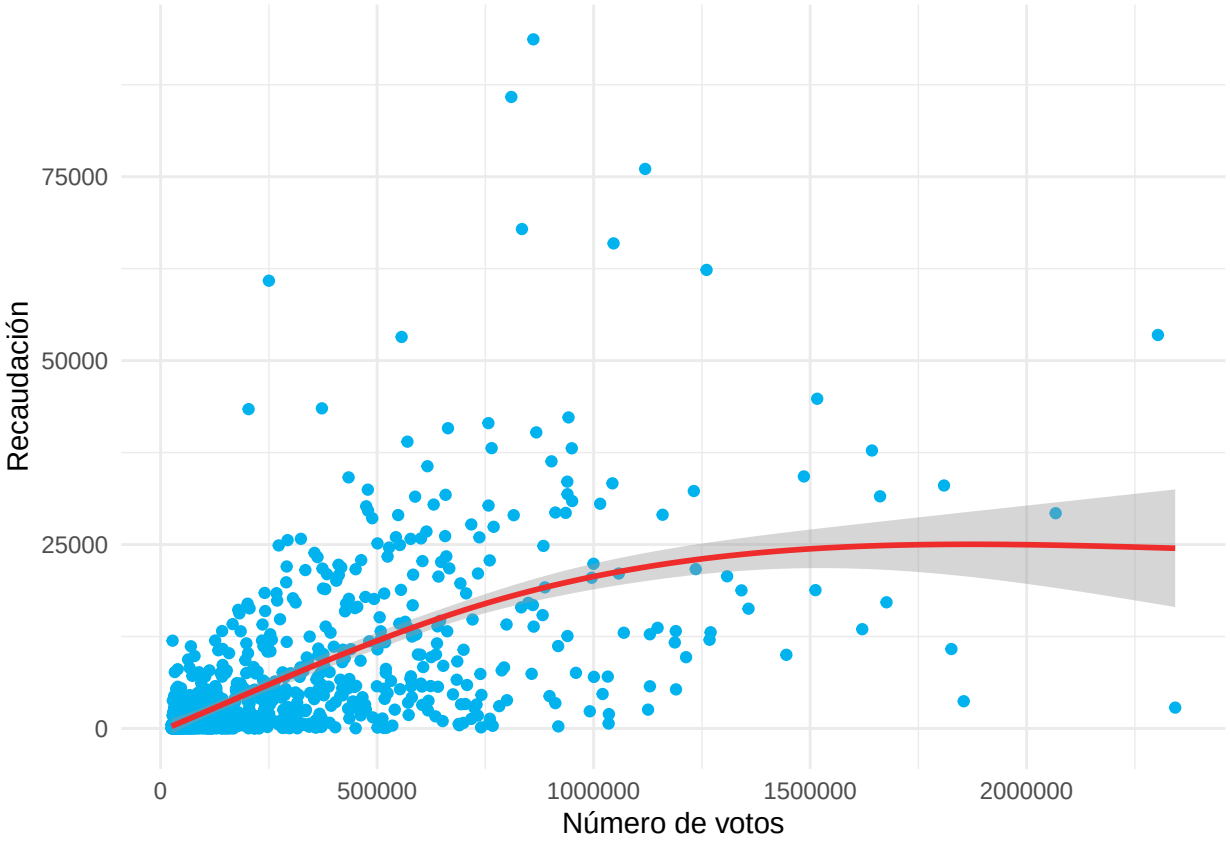
Vamos a verlo mejor gráficamente.

```
ggplot(df, aes(IMDB_Rating, Gross/10000)) +
  geom_point(color="deepskyblue2") +
  geom_smooth(color="firebrick2") +
  labs(
    x="Valoración en IMDB",
    y="Recaudación"
  ) +
  theme_minimal()
```



Vemos como en gráfico si que existe una pequeña diferencia a la alza entre la recaudación de las películas con una valoración mayor a 8.4 y las que tienen una menor valoración. Esta diferencia es mínima y en la gráfica no se puede apreciar una fuerte correlación entre ambas variables.

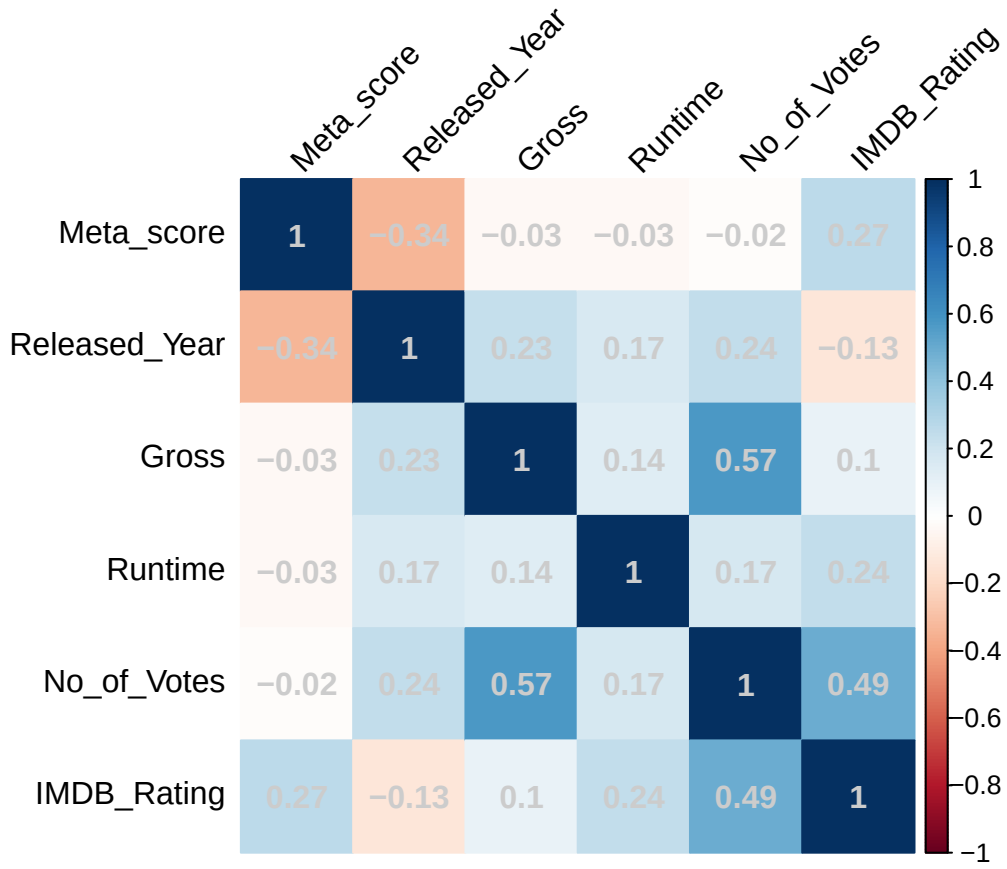
```
ggplot(df, aes(No_of_Votes, Gross/10000)) +
  geom_point(color="deepskyblue2") +
  geom_smooth(color="firebrick2") +
  labs(
    x="Número de votos",
    y="Recaudación"
  ) +
  theme_minimal()
```



En cambio, aquí sí que podemos apreciar una mayor correlación entre el número de votos y la recaudación de la película. Esta correlación se vuelve algo más debil a partir de más o menos los 1.25 millones de votos.

Para realmente ver la correlación de todas la variables a la vez y saber de manera exacta cual es su relación y si está es lo sufientemente significativa podemos hacer una matriz de correlación.

```
num_cols <- unlist(lapply(df, is.numeric))
df_num <- df[, num_cols]
m <- cor(df_num, use = "pairwise.complete.obs")
corrplot(m, order = "AOE", method = "color", addCoef.col = "gray80", tl.col="black", tl.srt=45)
```

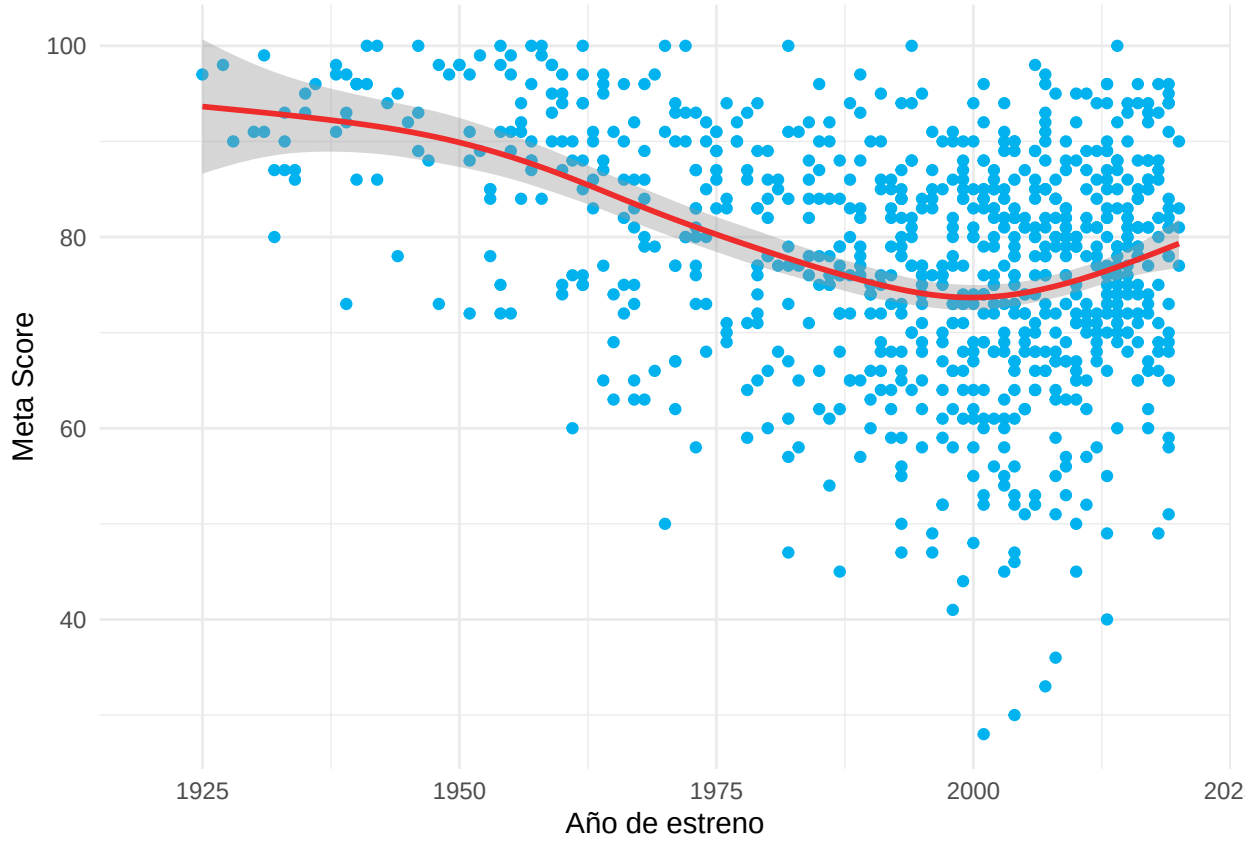


Como podemos ver existen dos correlaciones positivas que pueden llegar a ser significativas. La primera, entre el número de votos y la valoración en IMDB, con un 0.49, la cual ya habíamos visualizado con aterioridad. La segunda, entre el número de votos y la recaudación, con un 0.57, lo cual ya empieza a ser una correlación bastante fuerte.

En cuanto a las correlaciones negativas, solo cabe mencionar la existente entre el año de estreno y la valoración obtenida por los críticos de cine, con un -0.34, tenemos una correlación negativa moderada entre estas dos variables.

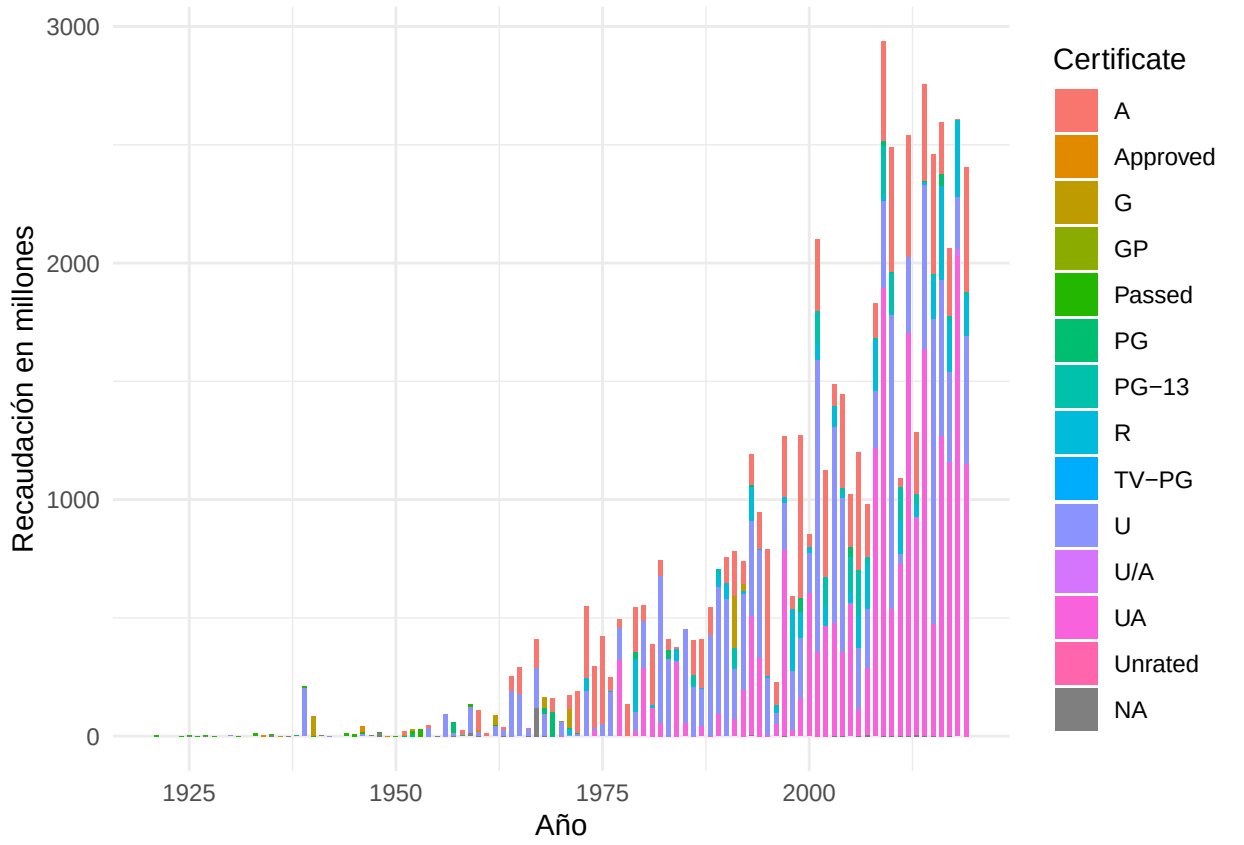
Vamos a representar la correlación negativa.

```
ggplot(df, aes(Released_Year, Meta_score)) +
  geom_point(color="deepskyblue2") +
  geom_smooth(color="firebrick2") +
  labs(
    x="Año de estreno",
    y="Meta Score"
  ) +
  theme_minimal()
```



Ahora sí, podemos ver visualmente como a medida que pasan los años, las películas van perdiendo calidad según los críticos de cine. Pero, recordemos que la industria cinematográfica tiene como criterio princiapal la recaudación. Por tanto, vamos a ver si este efecto negativo del paso del tiempo también afecta a la recaudación.

```
ggplot(df, aes(Released_Year, Gross/1000000, fill = Certificate)) +
  geom_bar(stat = "identity", width = 0.6) +
  labs(x="Año", y="Recaudación en millones") +
  theme_minimal()
```



En el gráfico vemos que no es así, de hecho, a medida que pasa el tiempo las películas y la indutria recauda más. Aunque como hemos visto antes la correlación entre estas dos variables es de 0.23, lo cual nos dice que su es muy leve y por tanto, no por que pase el tiempo se va a recaudar más.

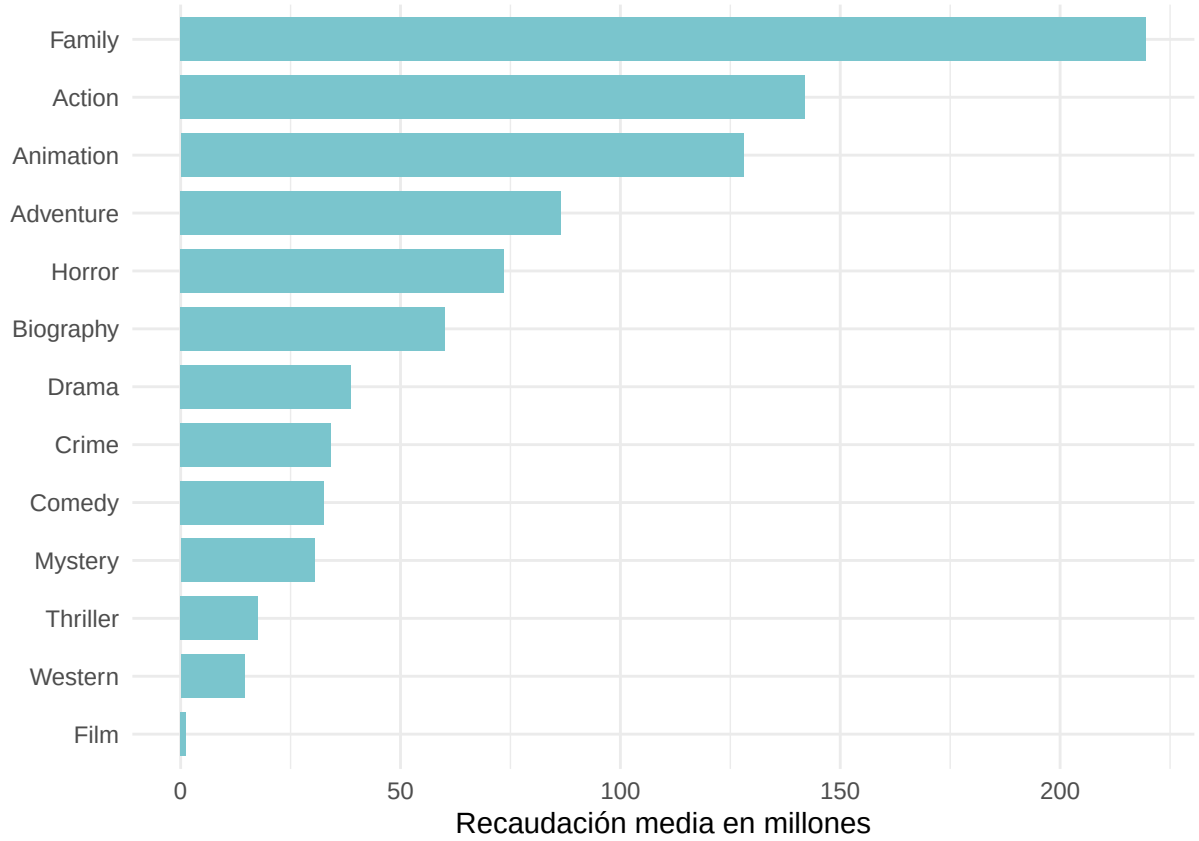
También hemos representado en el gráfico el certificado de cada película para poder apreciar su evolución, pero honestamente, al existir tantos tipos de certificados no se puede apreciar demasiado bien la evolución de estos en el gráfico, por tanto, no podemos extraer una conclusión o apreciación precisa.

Otra variable interesante para valorar si una película es buena o no, podría ser la recaduación media según el género de la película. Vamos a ver un gráfico sobre esto.

```
df$Main_Genre <- gsub("([A-Za-z]+).*", "\\1", df$Genre)

Main_Genre <- df %>%
  group_by(Main_Genre) %>%
  summarize(Average = mean(Gross/1000000, na.rm=TRUE))

ggplot(Main_Genre[-9,], aes(x=reorder(Main_Genre, Average), y=Average)) +
    geom_bar(stat='identity', fill = "#7AC5CD", width = 0.75) +
    coord_flip() +
    xlab('') +
    ylab('Recaudación media en millones') +
    theme_minimal()
```

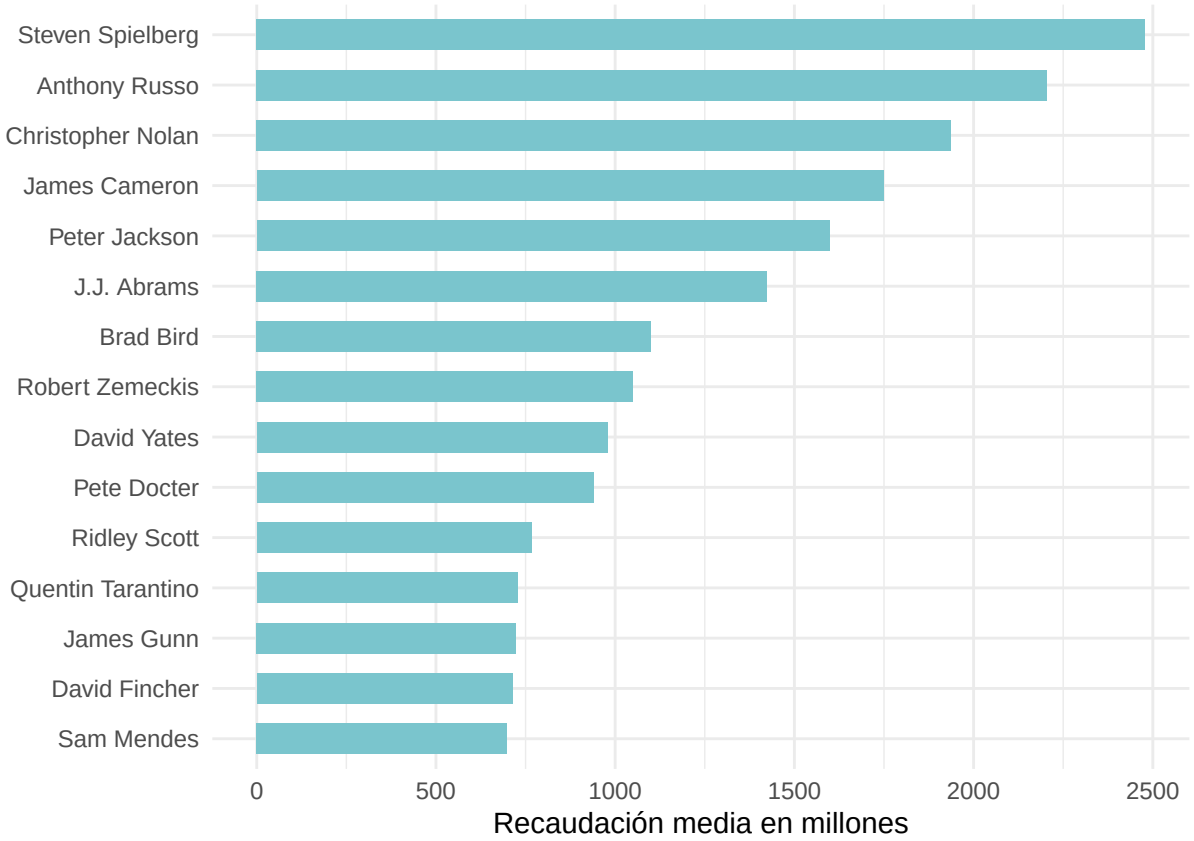


Como podemos ver la categoría que más recauda de media es "Family" con bastante diferencia a las demás. Por tanto, aunque el género no sea una variable continua y el programa no nos permita calcular un índice de correlación, vemos que existen algunos géneros que han recaudado más que otros a lo largo de la historia, con lo cual, se supone que al menos en un futuro a corto plazo esto seguirá siendo así.

Podemos hacer los mismo con los directores, pero esta vez sumando el total de la recaudación, ya que muchos directores tienes solo una película y quizás con tan solo una observación no tendría mucho sentido hacer la media.

```
Director_Avg = df %>% select(Director, Gross) %>% group_by(Director) %>% summarize(Sum = sum(Gross/1000
```

```
Director_Avg %>% top_n(15, wt=Sum) %>%
ggplot(aes(reorder(Director, Sum), Sum)) +
  geom_bar(stat = "identity", fill = "#7AC5CD", width = 0.6) +
  coord_flip() +
  labs(x="", y="Recaudación media en millones") +
  theme_minimal()
```

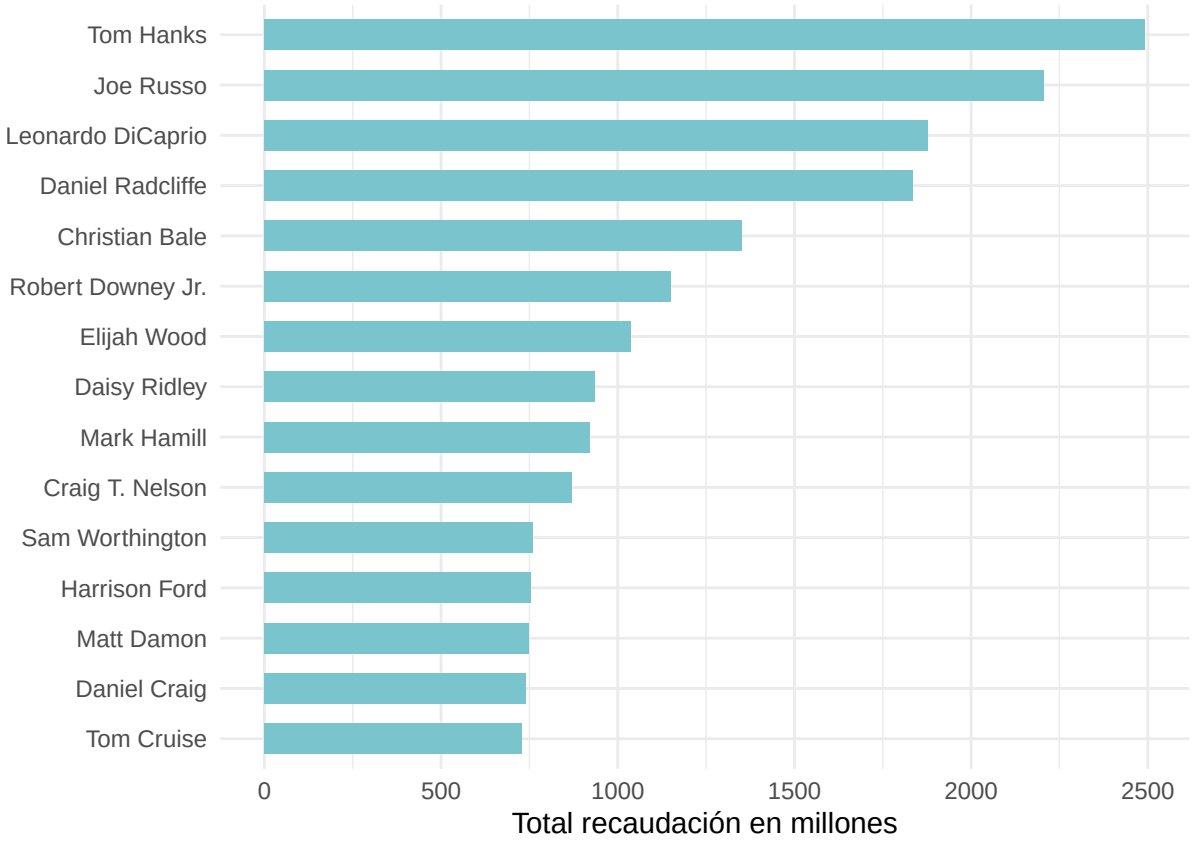


Vemos que de el director que más a recaudado a sido "Steven Spielberg".

Veamos los actores principales.

```
Star1_Avg = df %>% select(Star1, Gross) %>% group_by(Star1) %>% summarize(Sum = sum(Gross/1000000, na.r


Star1_Avg %>% top_n(15, wt=Sum) %>%
ggplot(aes(reorder(Star1, Sum), Sum)) +
  geom_bar(stat = "identity", fill = "#7AC5CD", width = 0.6) +
  coord_flip() +
  labs(x="", y="Total recaudación en millones") +
  theme_minimal()
```



Como vemos, el actor que más ha recaudado es "Tom Hanks".

Al igual que con los géneros, también vemos en los directores y los actores principales ciertas personas que hacen que una película recaude más que otra y por tanto, se conviertan en una "varibale" con una correlación positiva a la recaudación.

Conclusión

A lo largo del análisis hemos podido ver como dependiendo del criterio u objetivo que elijamos la concepción de "buena película" puede cambiar mucho. Hemos visto como una buena película según su recaudación es distinta a una buena película según los espectadores o incluso los críticos de cine. También hemos visto como ocurría lo mismo con el género de la película o los directores y actores principales. Aunque siendo realistas, y como hemos comentado a lo largo del análisis, el criterio más importante para la industrial del cine, por suerte o desgracia, es la recaudación. Atendiendo principiapalmente a este criterio, sabemos que existen ciertas variables correlacionadas (número de votos, géneros, directores y actores preferencia etc.) que hacen que la película tenga una mayor probabilidad de recaudar una buena cantidad en taquilla.

Por tanto, como conclusión, podemos decir que una película puede considerarse buena dependiendo de cuanto satisfaga los criterios "objetivo". En este caso para la recaudación hemos visto que existen algunos géneros preferencia, como el familiar o la acción, así como actores y directores, que consiguen una mayor recaudación que los demás.

```
tinytex::install_tinytex(force = TRUE)
```