

Proyecto Netflix Movies MADM

Laura Moreno, Josep Roman, Paul Ramírez

11/28/2020

Contenidos

1	Objetivo	2
2	Data wrangle	2
2.1	Importación de datos	2
2.2	Limpieza de los datos	3
3	Estadística descriptiva	4
3.1	Resumen	4
3.2	Tipo de variables	4
3.3	Estadísticos dataframe puntuaciones	5
3.4	Transformación variable fecha valoración	5
3.5	Análisis del número de valoraciones por mes y día de la semana	6
3.6	Distribución películas estrenadas por año	8
3.7	Análisis top 10 películas con más valoraciones por año de valoración	9
3.8	Evolución del score promedio de las 10 películas con más valoraciones	10
3.9	Comparación top 5 películas con más valoraciones	11
3.10	Estudios adicionales	12

1 Objetivo

2 Data wrangle

2.1 Importación de datos

2.1.1 Importación datos puntuaciones películas

Creamos una semilla específica para seleccionar aleatoriamente 250 películas con las que desarrollar el análisis exploratorio.

```
filas_ID_combined_all = read.csv(here("Data","filas_ID_combined_all.txt"))
set.seed(081034)
n_filas = nrow(filas_ID_combined_all)
muestra_grupo = sample(1:n_filas, 250, replace=F)
pelis <- filas_ID_combined_all[as.vector(muestra_grupo),]
```

Cargamos los 4 archivos originales con las puntuaciones, siguiendo el siguiente patrón:

```
data1 = read_tsv(here("Raw data","combined_data_1.txt"),col_names = FALSE)
```

Generamos un tibble vacío, y en función del archivo en el que se encuentre la película, vamos añadiendo en `scores` las filas correspondientes a nuestras películas:

```
scores = tibble()
for(i in 1:nrow(pelis)){
  if (data[i]==1){
    scores = rbind(scores,data1[filas[i]:filas_final[i],])
  }
  else if (data[i]==2){
    scores = rbind(scores,data2[filas[i]:filas_final[i],])
  }
  else if (data[i]==3){
    scores = rbind(scores,data3[filas[i]:filas_final[i],])
  }
  else {
    scores = rbind(scores,data4[filas[i]:filas_final[i],])
  }
}
```

Guardamos un csv con solo nuestras 250 películas en el formato original

```
write_csv(scores, here("Data", "nuestras_pelis_raw.csv"))
```

Carga archivo puntuaciones de nuestras 250 películas

Cargamos el csv generado en el paso anterior:

```
aux = read_csv(here("Data", "nuestras_pelis_raw.csv"), col_names = T)
```

2.1.2 Importación datos títulos películas

Carga archivo títulos películas

```
#algunas películas tienen una coma en su nombre, así que cargamos primero todo como una única columna, y  
titles = read_table(here("Data", 'movie_titles_raw.csv'), col_names=F) %>%  
  separate(col = 1, into = c("MovieID", "Release_Year", "Title"), sep = ",", extra = "merge")
```

2.2 Limpieza de los datos

2.2.1 Limpieza datos puntuaciones películas

Aplicamos el código de Ricardo para limpiar el dataframe `aux` y pasar al dataframe `scores` con una fila para cada valoración de usuario. A continuación, reorganizamos variables y asignamos tipos de variable:

```
scores %<>% relocate(MovieID, UserID, Date, Score)  
scores %<>% mutate(across(c(MovieID:UserID, Score), as.integer))  
scores %<>% mutate(Date = as.Date(Date))
```

2.2.2 Limpieza datos títulos películas

```
titles %<>% mutate(across(c(MovieID:Release_Year), as.integer))
```

Table 1: Daframe títulos películas

MovieID	Release_Year	Title
1	2003	Dinosaur Planet
2	2004	Isle of Man TT 2004 Review
3	1997	Character

2.2.3 Join de 'scores' con 'titles'

Hacemos un `left join` de `scores` con `titles` para añadir a la primera los títulos de cada película y el año en que se publicaron:

```
scores %<>% left_join(titles, by = 'MovieID')
```

2.2.4 Exportación datos limpios

Exportamos el archivo csv limpio para trabajar con el a partir de ahora:

```
write_csv(scores, here("Data", "nuestras_pelis.csv"))
```

2.2.5 Importación datos limpios para analizar en la sección *Estadística descriptiva*

```
scores = read_csv(here("Data","nuestras_pelis.csv"))
scores %<>% mutate(across(c(MovieID,UserID,Score,Release_Year), as.integer))
```

3 Estadística descriptiva

3.1 Resumen

- MovieIDs range from 1 to 17770 sequentially.
- UserIDs range from 1 to 2649429, with gaps. There are 480189 users.
- Ratings are on a five star (integral) scale from 1 to 5.
- Dates have the format YYYY-MM-DD.

Vemos que tenemos información de la películas 1 a la 15, y las puntuaciones se hicieron entre el 2000 y el 2005 (mayoritariamente en 2005). Distribución de los meses y días en que se puntuo es uniforme.

Veamos más información sobre los datos:

3.2 Tipo de variables

```
glimpse(scores)
```

```
## Rows: 1,508,892
## Columns: 6
## $ MovieID      <int> 515, 515, 515, 515, 515, 515, 515, 515, 515, 515, 515,...
## $ UserID       <int> 2295232, 1560318, 2550394, 1502043, 1507284, 771626, 1...
## $ Date         <date> 2005-08-16, 2005-10-04, 2005-11-01, 2005-08-15, 2005-...
## $ Score        <int> 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, ...
## $ Release_Year <int> 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, ...
## $ Title        <chr> "Avia Vampire Hunter", "Avia Vampire Hunter", "Avia Va...
```

Variables tipo *int*: MovieID, CustomerID, Score, Release_Year

- *UserID* & *MovieID* : Contiene un número entero, estos son objetos que contienen un único campo, un identificado ID para cada usuario (o película), no queremos duplicados. En cuanto al *MovieID*, será transformado en las gráficas a *chr* para visualizarlo mejor.
- *Release_Year*: No existen años con decimales, por lo tanto utilizar variables para datos enteros sería suficiente.
- *Score*: Las puntuaciones son números enteros del 1 - 5.

Variables tipo *date*: Date

- *Date* : esta variable incluye datos de tipo fecha (YY/MM/DD) por ello lo más adecuado es tratarlo como una variable de este tipo.

Variables tipo *chr*: Title

- *Title*: Utilizamos el tipo carácter porque nos interesan objetos que representan un conjunto de letras.

3.3 Estadísticos dataframe puntuaciones

```
movie_scores <- scores %>%
  group_by(MovieID) %>%
  summarise(Sum_Score = sum(Score), Mean_Score = mean(Score), SD_Score = sd(Score), Mode_Score = mlv(Score))
  left_join(titles, by = 'MovieID')

movie_scores_table <- movie_scores %>%
  ungroup() %>%
  select(-MovieID, -Release_Year) %>%
  relocate(Title, n, Sum_Score, Mean_Score, SD_Score, Mode_Score, Median_Score)

kable(head(movie_scores_table %>% arrange(desc(Mean_Score))), digits = 3, align = "c", caption = "Estadísticos puntuaciones")
kable_styling(latex_options = "hold_position", font_size = 8) %>%
  column_spec(1,width = "4cm")
```

Table 2: Estadísticos puntuaciones

Title	n	Sum_Score	Mean_Score	SD_Score	Mode_Score	Median_Score
Curb Your Enthusiasm: Season 3	12148	52674	4.336	1.000	5	5
Prime Suspect 3	2637	11222	4.256	0.899	5	4
Singin' in the Rain	29225	119852	4.101	0.947	5	4
VeggieTales: Dave and the Giant Pickle	2476	10102	4.080	1.125	5	4
The Thin Man	11095	44665	4.026	0.961	5	4
Felicity: Season 3	2820	11320	4.014	1.237	5	4

3.4 Transformación variable fecha valoración

Usamos la librería *lubridate* para generar variables separadas para año, número de mes, número de semana del año, número de día del mes, número de día de la semana, y una variable binaria que especifica si el día es fin de semana o entre semana.

```
scores_dates <- scores %>%
  mutate(
    Year = year(Date),
    n_month = month(Date),
    Week = week(Date),
    Day = day(Date),
    n_day_week = wday(Date,
      week_start = getOption("lubridate.week.start", 1)
    ),
    Is_Weekend = if_else(isWeekend(Date) == TRUE, "Weekend", "Weekday" )
  )
```

A partir de las variables de número de mes y número de día de la semana, creamos dos factores ordenados para el mes y el día de la semana.

```
scores_dates %<>% mutate(
  Month = ordered(n_month, levels = seq(1, 12, 1), labels = month.abb),
```

```

Day_Week = ordered(n_day_week, levels = seq(1, 7, 1), labels = day.abb)
)

scores_dates_table <- scores_dates %>%
  select(MovieID, UserID, Score, Date, Year, Month, Day, Day_Week, Is_Weekend)

kable(head(scores_dates_table, 3))

```

MovieID	UserID	Score	Date	Year	Month	Day	Day_Week	Is_Weekend
515	2295232	1	2005-08-16	2005	Aug	16	Tue	Weekday
515	1560318	3	2005-10-04	2005	Oct	4	Tue	Weekday
515	2550394	1	2005-11-01	2005	Nov	1	Tue	Weekday

3.5 Análisis del número de valoraciones por mes y día de la semana

```

coeff <- max(scores$Score) / max(month_scores$n)
n_max_limit <- max(month_scores$n) + 10^4
n_min_limit <- plyr::round_any((n_max_limit - min(month_scores$n))*2, 10^4)

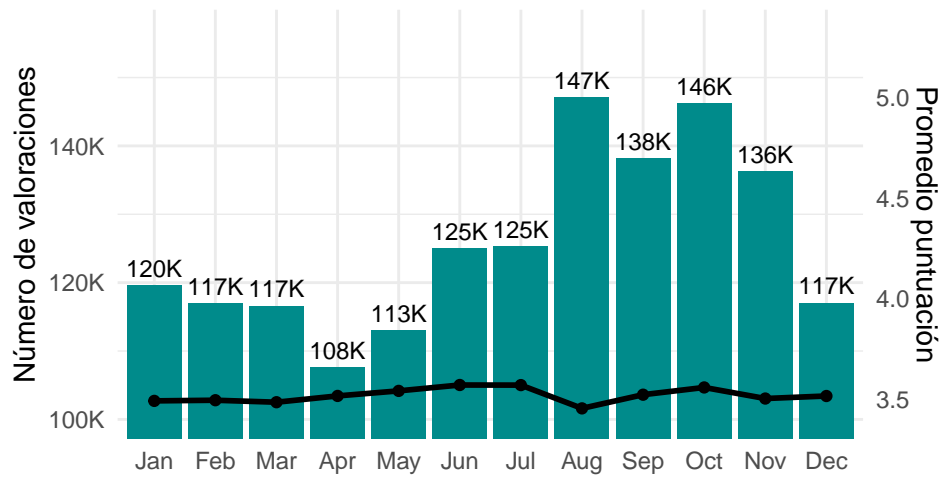
gg1 <- ggplot(data = month_scores, aes(x = Month))

gg2 <- gg1 +
  geom_bar(aes(y = n), fill = "darkcyan", stat = "identity") +
  coord_cartesian(ylim = c(n_min_limit, n_max_limit)) +
  geom_point(aes(y = Mean_Score/coeff)) +
  geom_line(aes(y = Mean_Score/coeff), size = 1, group = 1) +
  scale_y_continuous(
    name = "Número de valoraciones",
    labels = ks,
    sec.axis = sec_axis(~.*coeff, name = "Promedio puntuación")
  ) +
  labs(title = "Valoraciones por mes", x = "") +
  geom_text(aes(y = n, label = ks(n)), angle = 0, vjust = -0.5, size = 3) +
  theme_minimal()

gg2

```

Valoraciones por mes



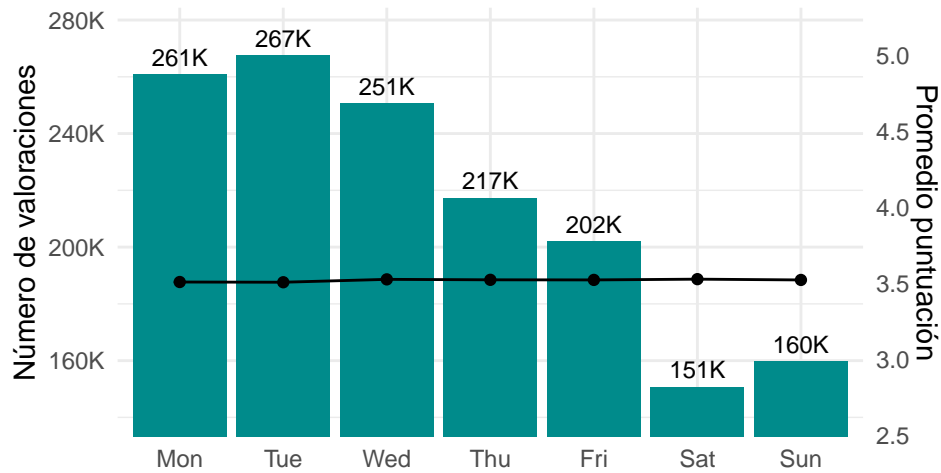
```
coeff <- max(scores$Score) / max(day_week_scores$n)
n_max_limit <- max(day_week_scores$n) + 10^4
n_min_limit <- plyr::round_any(min(day_week_scores$n) - 10^4, 10^4, f = floor)

gg1 <- ggplot(data = day_week_scores, aes(x = Day_Week))
```

```
gg2 <- gg1 +
  geom_bar(aes(y = n), fill = "darkcyan", stat = "identity") +
  coord_cartesian(ylim = c(n_min_limit, n_max_limit)) +
  geom_point(aes(y = Mean_Score/coeff)) +
  geom_line(aes(y = Mean_Score/coeff), group = 1) +
  scale_y_continuous(
    name = "Número de valoraciones",
    labels = ks,
    sec.axis = sec_axis(~.*coeff, name = "Promedio puntuación")
  ) +
  labs(title = "Valoraciones por día de la semana", x = "") +
  geom_text(aes(y = n, label = ks(n)), angle = 0, vjust = -0.5, size = 3) +
  theme_minimal()
```

```
gg2
```

Valoraciones por día de la semana



```
weekend_weekday_scores <- scores_dates %>%
  group_by(Is_Weekend) %>%
  summarise(Mean_Score = mean(Score), n = n())

kable(weekend_weekday_scores)
```

Is_Weekend	Mean_Score	n
Weekday	3.517167	1198343
Weekend	3.525031	310549

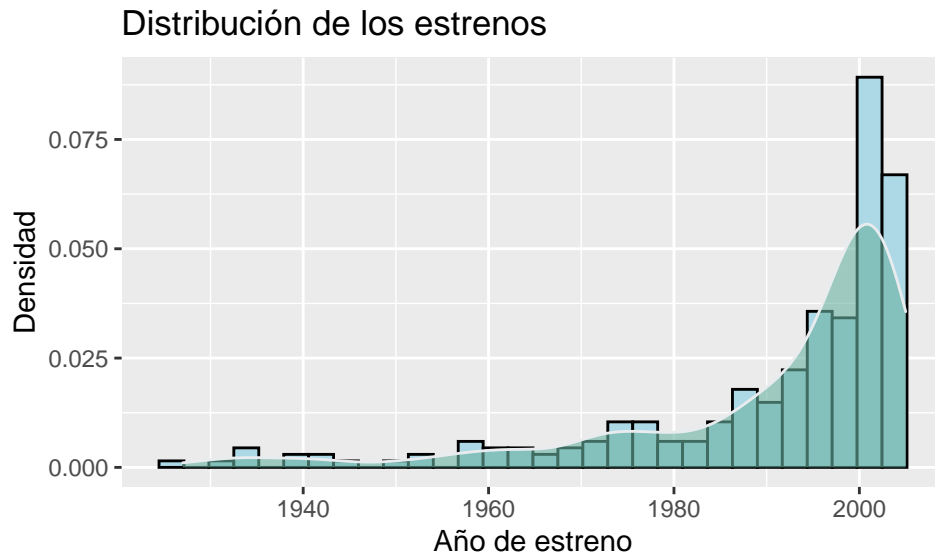
```
n_scores_weekend = weekend_weekday_scores %>% filter(Is_Weekend == TRUE) %>% select(n)
n_scores = sum(weekend_weekday_scores$n)
n_scores_weekend_weekday_ratio = n_scores_weekend / n_scores #el 18% de las valoraciones son en fin de
```

3.6 Distribución películas estrenadas por año

```
nuestros_movie_ids <- tibble(MovieID = unique(scores$MovieID))
nuestros_titles <- titles %>%
  right_join(nuestros_movie_ids, by = "MovieID")

movies_per_year <- nuestros_titles %>%
  group_by(Release_Year) %>%
  summarise(n = n())
```

```
ggplot(data = nuestros_titles, aes(x=Release_Year, y=..density..)) +
  geom_histogram(colour="black", fill="lightblue") +
  geom_density(fill="#69b3a2", color="#e9ecf", alpha=0.6) +
  labs(x='Año de estreno', y='Densidad', title='Distribución de los estrenos')
```

En nuestra muestra del 1.41% del total de las películas de netflix , el año con más estrenos en fue el 2000 con 25. Por otra parte en un 50% de los años se estrenaron como mucho 2 películas.

3.7 Análisis top 10 películas con más valoraciones por año de valoración

```
votaciones_ano <- group_by(scores, MovieID, Year=year(Date)) %>%
  summarise(Votos = n_distinct((UserID)), Mean = round(mean(Score),3)) #agrupamos por película y año e

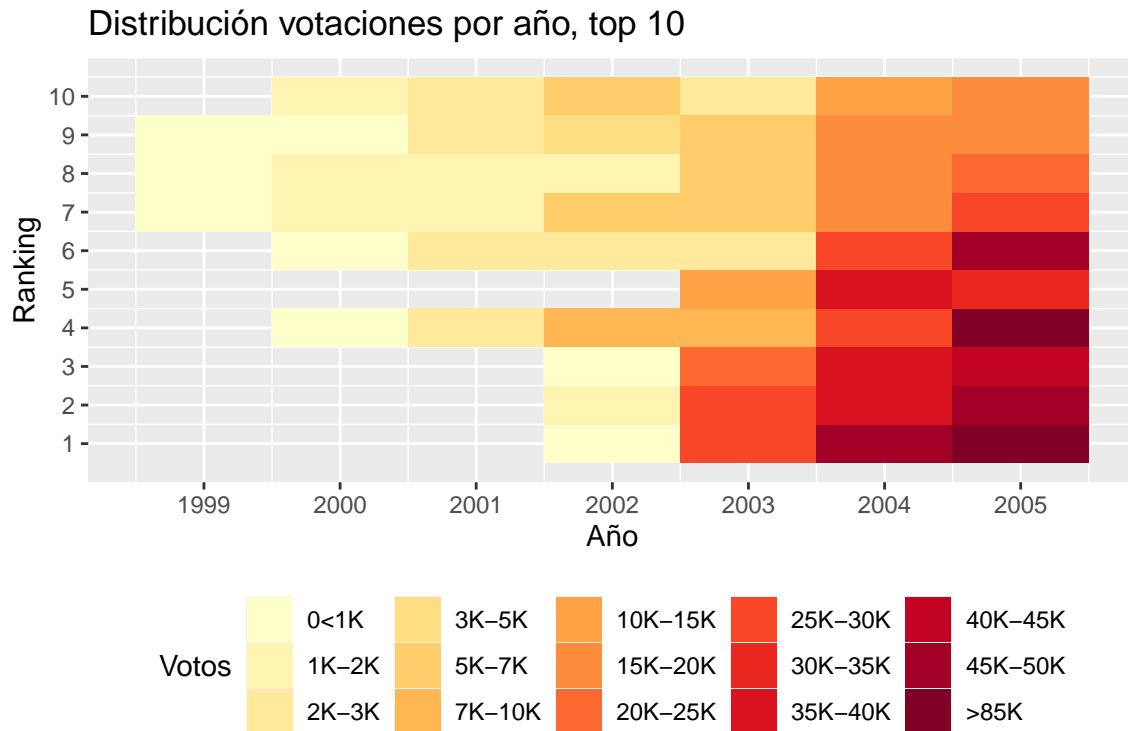
top10_votada <- head(arrange(movie_scores[,c('MovieID','n','Mean_Score','Title')], desc(n)), 10) #las
movies_onfire <- filter(votaciones_ano, MovieID %in% top10_votada$MovieID)

movies_onfire$Ranking <- movies_onfire$MovieID
for (i in 1:10) {
  movie <- top10_votada$MovieID[i]
  indexes <- which(movies_onfire$MovieID == movie)
  movies_onfire$Ranking <- replace(movies_onfire$Ranking, indexes, i)
  count=i
}
#Ordenamos las pelis según el top10
top10_votada$Ranking=1:10
```

Para visualizar la distribución de votaciones por año que obtubieron las 10 películas más votadas de *Netflix*, creamos un *Heatmap*

```
# 3r creamos el Heatmap
# movies_onfire <- arrange(movies_onfire, desc(Ranking))

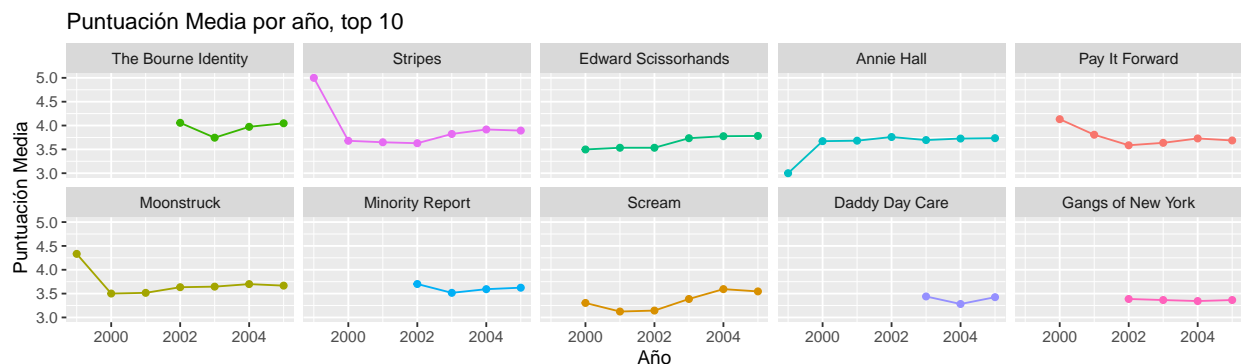
ggplot(movies_onfire, aes(text = paste('Votos:', Votos), ID = MovieID, y = Ranking, x = Year )) + #tex
  geom_tile(aes(fill = secuencia)) +
  scale_y_continuous(breaks=10:1) +
  scale_x_continuous(breaks=1999:2005) +
  scale_fill_manual(values = mycolors) + #secuencia de colores
  labs(fill = 'Votos', x='Año', title='Distribución votaciones por año, top 10') +
  theme(legend.position="bottom")
```



3.8 Evolución del score promedio de las 10 películas con más valoraciones

```
movies_onfire %<>% left_join(titles[, -2], by = 'MovieID')
orden_titulos <- arrange(top10_votada[, c('Title', 'Mean_Score')], desc(Mean_Score))
movies_onfire %<>% transform(Title=factor(Title, levels=as.vector(orden_titulos$Title)))

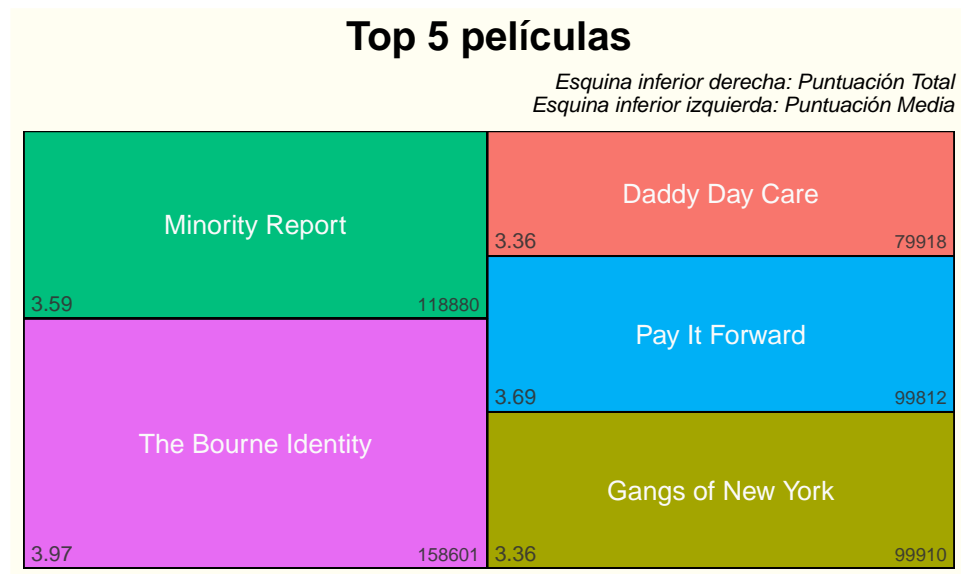
ggplot(movies_onfire, aes(Year, Mean, group=MovieID, colour=factor(MovieID))) +
  geom_point() +
  geom_line() +
  facet_wrap(~Title, nrow = 2, scale='fixed') +
  labs(y='Puntuación Media', x='Año', title='Puntuación Media por año, top 10') +
  theme(legend.position="none")
```



El 2000 parece que fue un punto de inflección, 3 películas se estrenaron ese año y las ya existentes sufrieron un cambio en su puntuación media ya sea para mal como fue el caso de *Moonstruck* y *Stripes*, o para bien como en el caso de Annie Hall.

3.9 Comparación top 5 películas con más valoraciones

A continuación, representamos las 5 películas que recogen más valoraciones en un **treemap**:



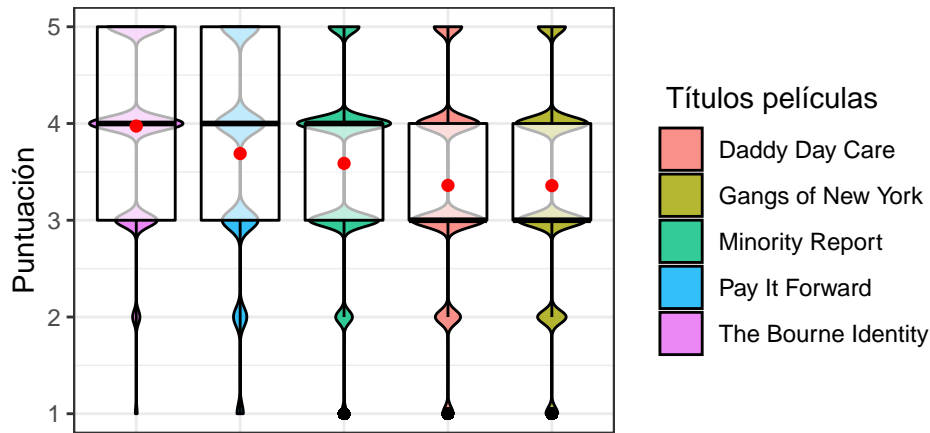
Utilizaremos el *diagrama de violin* combinado con un diagrama de cajas y bigotes para representar la distribución de la puntuación en las 5 películas seleccionadas. El punto rojo representa la media de cada película, mientras que la línea negra que atraviesa el diagrama de cajas es la moda.

```
## # A tibble: 5 x 9
## # Groups:   MovieID [5]
##   MovieID Sum_Score Mean_Score SD_Score Mode_Score Median_Score      n
##   <int>    <int>    <dbl>    <dbl>    <int>    <dbl>  <int>
## 1     6037    630194     3.97    0.845         4         4 158601
## 2     8387    426515     3.59    0.907         4         4 118880
## 3    10730    335467     3.36    1.05         4         3 99910
## 4      313    368241     3.69    1.09         4         4 99812
## 5     9645    268580     3.36    1.04         3         3 79918
## # ... with 2 more variables: Release_Year <int>, Title <chr>
```

```
plot2
```

Distribución de las puntuaciones en un diagrama de violín

El punto rojo sitúa la puntuación media



3.10 Estudios adicionales

Buscar los top 5 usuarios que más películas han puntuado. Luego, comparar con el top 1 usuario qué películas han dejado de evaluar el resto.

Primero vamos a buscar el número de total de películas que han sido evaluadas por usuario:

```
#número de veces que ha votado cada usuario
num_votos_por_usuario = aggregate(scores$UserID, by = list(Usuario=scores$UserID), length)
```

En segundo lugar, seleccionaremos el top 5 usuarios que más películas han puntuado,

```
df <- scores %>% group_by(UserID) %>% count()
df <- scores %>% group_by(UserID) %>% summarise(NN = n())
df <- scores %>% group_by(UserID) %>%
  summarise(NN = n(), percent = n()/nrow(.) ) #Añadir a la tabla el % que representa cada usuario en el
df <- scores %>% group_by(UserID) %>%
  summarise (NN = n()) %>%
  mutate(percent= NN / sum(NN))
top_5_users <- head(df[order(df$NN, decreasing = TRUE),],5)
knitr::kable(top_5_users, digits = 5, align = "c", caption = "Top 5 Usuarios") %>%
  kable_styling(latex_options = "hold_position", font_size = 8) %>%
  column_spec(1,width = "4cm")
```

Table 3: Top 5 Usuarios

UserID	NN	percent
305344	249	0.00017
387418	247	0.00016
2439493	232	0.00015
1664010	223	0.00015
2118461	208	0.00014

En tercer lugar, buscaremos qué películas han sido evaluadas por estos usuarios. Seguidamente, compararemos el total de películas evaluadas por el usuario top_1 con el resto:

El usuario que más películas ha puntuado es el 305344, entonces vamos a comparar las películas que este usuario con el resto de usuarios.

```
top_pelis_1$comp2 <- as.integer(top_pelis_1$MovieID %in% top_pelis_2$MovieID)
top_pelis_1$comp3 <- as.integer(top_pelis_1$MovieID %in% top_pelis_3$MovieID)
top_pelis_1$comp4 <- as.integer(top_pelis_1$MovieID %in% top_pelis_4$MovieID)
top_pelis_1$comp5 <- as.integer(top_pelis_1$MovieID %in% top_pelis_5$MovieID)
Dif_1_2= top_pelis_1 %>%
  filter(comp2 %in% c('0'))
Dif_1_3= top_pelis_1 %>%
  filter(comp3 %in% c('0')) # obtener las que no aparecen
Dif_1_4= top_pelis_1 %>%
  filter(comp4 %in% c('0')) # obtener las que no aparecen
Dif_1_5= top_pelis_1 %>%
  filter(comp5 %in% c('0')) # obtener las que no aparecen
```

Aquí tenemos las 18 películas el top_3 no ha puntuado pero el top_1 si:

```
films_3
```

```
## [1] "The Program"
## [2] "Jane Goodall's Wild Chimpanzees: IMAX"
## [3] "Federal Protection"
## [4] "Omagh"
## [5] "Britney and Kevin: Chaotic: The DVD"
## [6] "The Last Time I Committed Suicide"
## [7] "The Land Before Time XI: The Invasion of the Tinysauruses"
## [8] "Portland Street Blues"
## [9] "Playgirls and the Vampire"
## [10] "Lumiere and Company"
## [11] "The Adventures of Robin Hood: Bonus Material"
## [12] "A Crime of Passion"
## [13] "Hamlet"
## [14] "Hear No Evil"
## [15] "Halloween III: Season of the Witch"
## [16] "Carmen Electra's Fit to Strip"
## [17] "Jack O'Lantern"
## [18] "Female Yakuza Tale"
```

Este proceso debería realizarse con el resto de usuarios, sin embargo, por cuestión de espacio no vamos a ejecutar los códigos. Finalmente, obtendríamos que, la diferencia del top 2 con el top 1 son 3 películas, mientras que el top 4 con el top 1 se diferencia en 26 películas y, por último, la diferencia con el último usuario asciende a 41 películas.

Para terminar, hemos creado un correlograma de la tabla `scores_dates`, solamente con los valores numéricos. Luego, se ha incluido también un correlograma con los p-valores de las dimensiones anteriores, para saber si estas son o no son significantes. Las variables que presentan correlación son las sombreadas en color. Después, la tabla de al lado, señala aquellas variables que no son significantes en nuestro análisis.

