Pràctica 2 - Data Cleaning

Autor: Rem Blanch Torras

Maig 2020

Detalls de l'activitat

Descripció

En aquesta activitat s'elabora un cas pràctic, consistent en el tractament d'un conjunt de dades (en anglès, dataset), orientat a aprendre a identi fi car les dades rellevants per a un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes

Objectiu

Els objectius que es persegueixen mitjançant el desenvolupament d'aquesta activitat pràctica són els següents:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contextos més amplis o multidisciplinaris.
- Saber identi fi car les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que permeti continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de recerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Competències

Així, les competències de l'Màster en Data Science que es desenvolupen són:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

Objectiu del estudi

Actualment, és normal que les empreses que decideixen produir pel·lícules s'interessin per saber com les han de generar per obtenir els millors resultats en termes d'ingressos, popularitat i aclamacions. Per això, solen realitzar estudis basats en les dades obtingudes de produccions anteriors. En el següent estudi, s'analitza la informació recollida en IMDB de 1000 pel·lícules durant l'època 2006 - 2016. La tasca és presentar les característiques de les pel·lícules que guanyen:

- Els ingressos més alts.
- Popularitat (valoració IMDB).
- Aclaració crítica (valoració metacrítica).
- Estratègia màrketing.
- Época més òptima per presentar el títol.

Descripció del dataset

El conjunt de dades està pres de la base de dades IMDB i emmagatzemat a Kaggle. Conté dades de 1000 pel·lícules més populars (segons valoració IMDB) del període 2006-2016. IMDB (Internet Movie Database) és una base de dades en línia d'informació relacionada amb pel·lícules, programes de televisió, vídeos domèstics i videojocs i fluxos d'internet, incloent biografies de repartiment, personal de producció i personal, resums de parcel·les, trivies i ressenyes i valoracions de fan.

Els usuaris registrats en aquest lloc són convidats a puntuar qualsevol pel·lícula en una escala d'1 a 10 i els totals es converteixen en una classificació mitjana ponderada que es mostra al costat de cada títol.

També mostra el Metascore de cada títol. Metascore és la qualificació donada per una altra companyia de qualificació de pel·lícules anomenada Metacritic. Tanmateix, a diferència d'IMDB, obtenen qualificacions d'agències de qualificació conegudes i calculen una mitjana ponderada d'aquestes valoracions.

Està constituït per 12 característiques (columnes) que presenten 1000 pel·lícules (files o registres). Entre els camps d'aquest conjunt de dades, trobem els següents:

- **Rank**: Identificador de la pel·lícula o série al dataset.
- **Title**: Títol de la pel·lícula.
- **Genre**: Llista de gèneres separats per comes utilitzats per classificar la pel·lícula.
- **description**: Breu descripció del títol.
- **Director**: Nom del director.
- **Actors**: Una llista separada per comes de les principals estrelles de la pel·lícula
- **Year**: Any en què la pel·lícula es va estrenar com a sencer.
- **Runtime (Minutes)**: La duració de la pel·lícula en minuts.
- **Rating**: Valoració dels usuaris de la pel·lícula 0-10
- **Votes**: Nombre de vots
- **Revenue (milions)**: Ingressos de pel·lícules en milions
- **Metascore**: una mitjana agregada de puntuacions de crítica. Els valors són entre 0 i 100. Les puntuacions més altes representen comentaris positius.

Integració i selecció de les dades d'interès a analitzar.

Començem per realitzar una vista rápida a les dades disponibles:

Data summary

Name	movies
Number of rows	1000
Number of columns	12

_____

Column type frequency:

character	5
numeric	7

_____

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Title	0	1	2	61	0	999	0
Genre	0	1	5	26	0	207	0
Description	0	1	42	421	0	1000	0
Director	0	1	3	32	0	644	0

Actors	0	1	43	77	0	996	0		

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Rank	0	1.00	500.50	288.82	1.0	250.75	500.50	750.25	1000.00	▇▇▇▇▇
Year	0	1.00	2012.78	3.21	2006.0	2010.00	2014.00	2016.00	2016.00	▂▂▂▇▅
Runtime (Minutes)	0	1.00	113.17	18.81	66.0	100.00	111.00	123.00	191.00	▁▆▇▂▁
Rating	0	1.00	6.72	0.95	1.9	6.20	6.80	7.40	9.00	▁▁▃▇▆
Votes	0	1.00	169808.26	188762.65	61.0	36309.00	110799.00	239909.75	1791916.00	▇▁▁▁▁
Revenue (Millions)	128	0.87	82.96	103.25	0.0	13.27	47.98	113.72	936.63	▇▁▁▁▁
Metascore	64	0.94	58.99	17.19	11.0	47.00	59.50	72.00	100.00	▁▅▇▆▅

Disposem de 5 variables categòriques i 7 numériques.

Curiosament, dels 1000 títols disponibles, tenim 999 valors únics. Revisem quin és el repetit

```
# Es busquen duplicats.
movies[duplicated(movies$Title),]
```

```
## # A tibble: 1 x 12
##    Rank Title Genre Description Director Actors  Year `Runtime (Minut~
Rating
##   <dbl> <chr> <chr> <chr>       <chr>    <chr>  <dbl>            <dbl>
<dbl>
## 1   633 The ~ Come~ A monster ~ Bong Jo~ Kang-~  2006              120
7
## # ... with 3 more variables: Votes <dbl>, `Revenue (Millions)` <dbl>,
## #   Metascore <dbl>
```

```
# Filtrat per el títol duplicat.
filter(movies, Title == 'The Host')
```

```
## # A tibble: 2 x 12
##    Rank Title Genre Description Director Actors  Year `Runtime (Minut~
Rating
##   <dbl> <chr> <chr> <chr>       <chr>    <chr>  <dbl>            <dbl>
<dbl>
## 1   240 The ~ Acti~ When an un~ Andrew ~ Saoir~  2013              125
5.9
## 2   633 The ~ Come~ A monster ~ Bong Jo~ Kang-~  2006              120
7
## # ... with 3 more variables: Votes <dbl>, `Revenue (Millions)` <dbl>,
## #   Metascore <dbl>
```

No hi ha pel·lícules duplicades. La pel·lícula amb el títol "The Host" es va estrenar el 2006 i el 2013. Per tant, són dues pel·lícules diferents.

Neteja de les dades.

Els valors que falten poden ser problemes per als propers passos. Per tant, hem de revisar-les abans de l'anàlisi i després podrem omplir els valors que falten d'algunes variables si és necessari.

```
# Dades amb valors nuls.
skim(movies) %>%
  dplyr::select(skim_variable, n_missing)
```

```
## # A tibble: 12 x 2
##    skim_variable      n_missing
##  * <chr>                  <int>
##  1 Title                      0
##  2 Genre                      0
##  3 Description                0
##  4 Director                   0
##  5 Actors                     0
##  6 Rank                       0
##  7 Year                       0
##  8 Runtime (Minutes)          0
##  9 Rating                     0
## 10 Votes                      0
## 11 Revenue (Millions)       128
## 12 Metascore                 64
```

Observem valors buits en:

- Revenue (Millions) = 128 (~ 12.8%)

- Metascore = 64 (~ 6.4%)

"Revenue (Millions)" i "Metascore" seran importants per a l'EDA. Per tant, aquestes columnes no es poden eliminar. El percentatge de files que falta valors per a "Ingressos (milions) és al voltant del 13%. Això és elevat, per tant, omplirem els valors que falten. El percentatge de files que tenen valors per a"Metascore" és al voltant del 6%. Això no és tan alt, de manera que anem a eliminar les files que falten valors per a aquestes columnes

```
# Omplir els valors que falten amb el valor mitjà.
movies$`Revenue (Millions)`[is.na(movies$`Revenue (Millions)`)] <-
median(movies$`Revenue (Millions)`, na.rm=TRUE)
# Fer drop de les files que falten valors per a "Metascore"
movies_clean <- na.omit(movies)
```

Per acabar, actualitzem els noms de Revenue i Runtime per facilitar operar amb ambdós atributs.

```
# Realització del canvi de nom als atributs Revenue i Runtime.
movies_clean <- movies_clean %>%
  dplyr::rename(
    Revenue = `Revenue (Millions)`,
    Runtime = `Runtime (Minutes)`
    )
```

Anàlisi de les dades.

En aquesta secció, anirem a explorar el conjunt de dades. Utilitzarem estadístiques descriptives i també visualització de dades per ajudar-nos a explorar el conjunt de dades.

Impacte dels directors sobre una pel·lícula

Les pel·lícules d'un director en concret reben ingressos més alts?

```
# Agrupem les ocurrencies de l'atribut Director y ordenem per aquells que
apareixen més vegades.
movies_clean %>%
  group_by(Director) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  top_n(n=10)
```

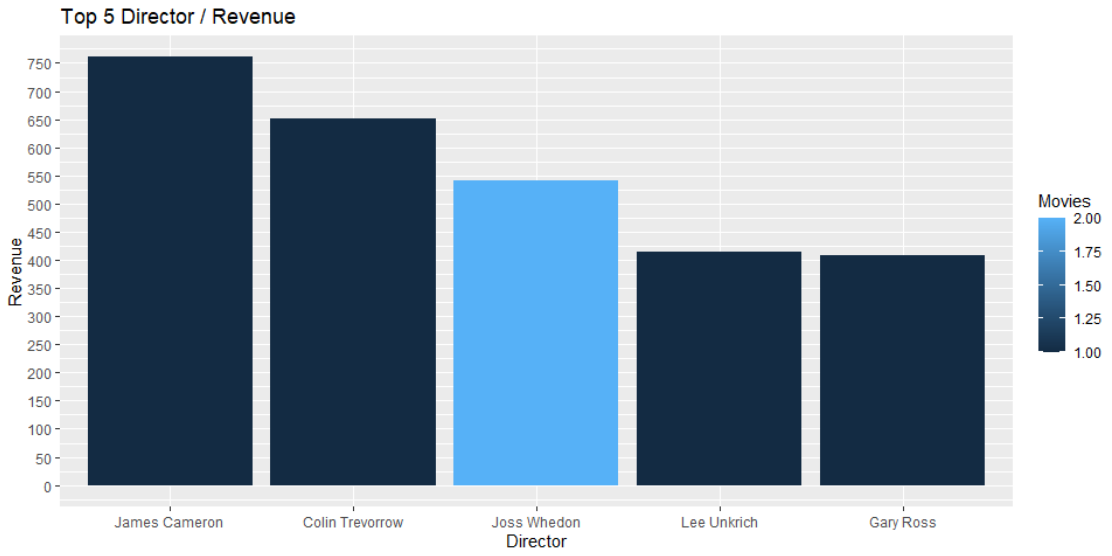
```
## # A tibble: 16 x 2
##    Director              n
##    <chr>             <int>
##  1 Ridley Scott          8
```

```
##  2 David Yates            6
##  3 M. Night Shyamalan     6
##  4 Michael Bay            6
##  5 Paul W.S. Anderson     6
##  6 Antoine Fuqua          5
##  7 Christopher Nolan      5
##  8 Danny Boyle            5
##  9 David Fincher          5
## 10 Denis Villeneuve       5
## 11 J.J. Abrams            5
## 12 Justin Lin             5
## 13 Martin Scorsese        5
## 14 Peter Berg             5
## 15 Woody Allen            5
## 16 Zack Snyder            5
```

```
# Descobrir el top 5 directors en funció de la mitjana de la recaudació.
top5Director <- movies_clean %>%
  group_by(Director) %>%
  summarise(Movies = n(), Revenue = mean(Revenue)) %>%
  arrange(desc(Revenue)) %>%
  top_n(n=5)

top5Director
```

```
## # A tibble: 5 x 3
##   Director        Movies Revenue
##   <chr>            <int>   <dbl>
## 1 James Cameron        1    761.
## 2 Colin Trevorrow      1    652.
## 3 Joss Whedon          2    541.
## 4 Lee Unkrich          1    415.
## 5 Gary Ross            1    408
```

```
# Plot del resultat anterior
ggplot(top5Director, aes(x=reorder(Director,desc(Revenue)), y=Revenue,
fill=Movies)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0,1000,50)) +
  ggtitle('Top 5 Director / Revenue') +
  labs(x = 'Director')
```

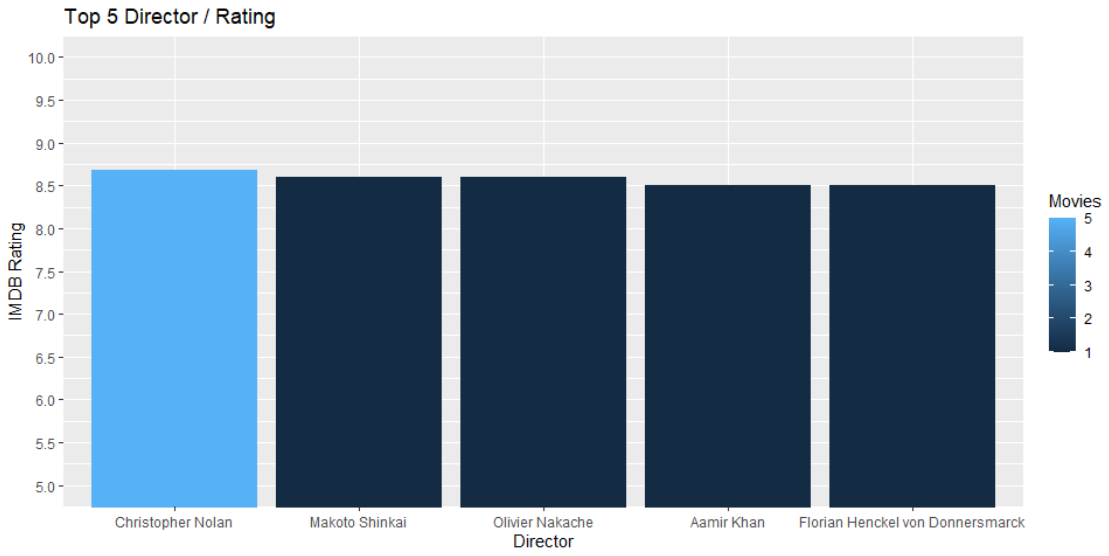


Les pel·lícules d'un director en particular reben una qualificació IMDB més alta?

```
# Descobrir el top 5 directors en funció de la mitjana de la qualificació a IMDB.
top5DirectorRating <- movies_clean %>%
  group_by(Director) %>%
  summarise(Movies = n(), Rating = mean(Rating)) %>%
  arrange(desc(Rating)) %>%
  top_n(n=5)

top5DirectorRating
```

```
## # A tibble: 5 x 3
##   Director                          Movies Rating
##   <chr>                              <int>  <dbl>
## 1 Christopher Nolan                      5   8.68
## 2 Makoto Shinkai                         1   8.6
## 3 Olivier Nakache                        1   8.6
## 4 Aamir Khan                             1   8.5
## 5 Florian Henckel von Donnersmarck       1   8.5
```

```
# Plot del resultat anterior
ggplot(top5DirectorRating, aes(x=reorder(Director,desc(Rating)),
y=Rating, fill=Movies)) +
  geom_bar(stat = "identity") +
  coord_cartesian(ylim = c(5, 10)) +
  scale_y_continuous(breaks = seq(5, 10, .5)) +
  ggtitle('Top 5 Director / Rating') +
  labs(x = 'Director', y = 'IMDB Rating')
```

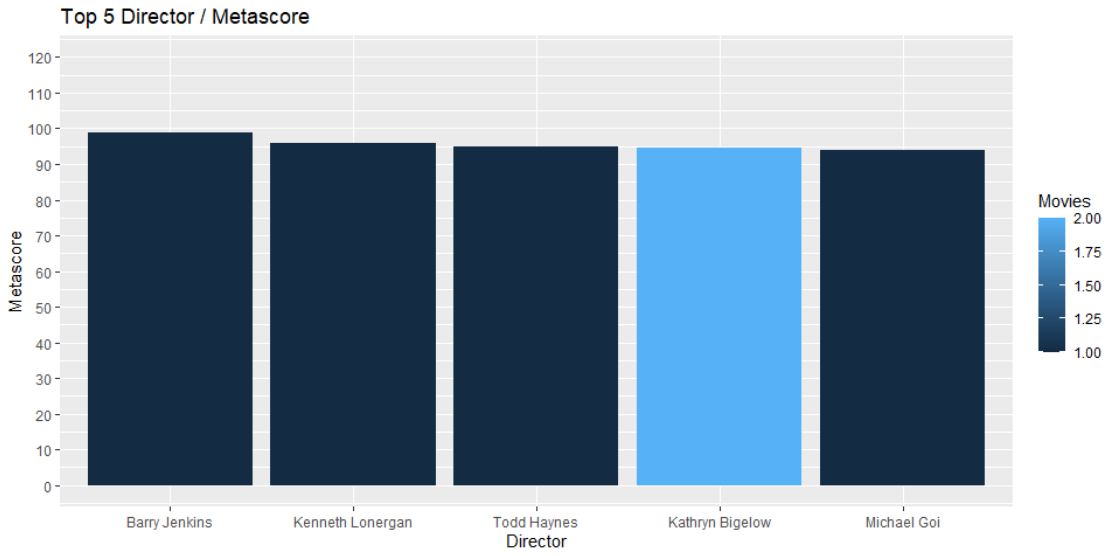


Les pel·lícules d'un director en particular reben una puntuació més alta de Metacritic?

```
# Descobrir el top 5 directors en funció de la mitjana de la puntuació a Metacrític.
top5DirectorMetascore <- movies_clean %>%
  group_by(Director) %>%
  summarise(Movies = n(), Metascore = mean(Metascore)) %>%
  arrange(desc(Metascore)) %>%
  top_n(n=5)

top5DirectorMetascore
```

```
## # A tibble: 5 x 3
##   Director         Movies Metascore
##   <chr>             <int>     <dbl>
## 1 Barry Jenkins         1      99
## 2 Kenneth Lonergan      1      96
## 3 Todd Haynes           1      95
## 4 Kathryn Bigelow       2      94.5
## 5 Michael Goi           1      94
```

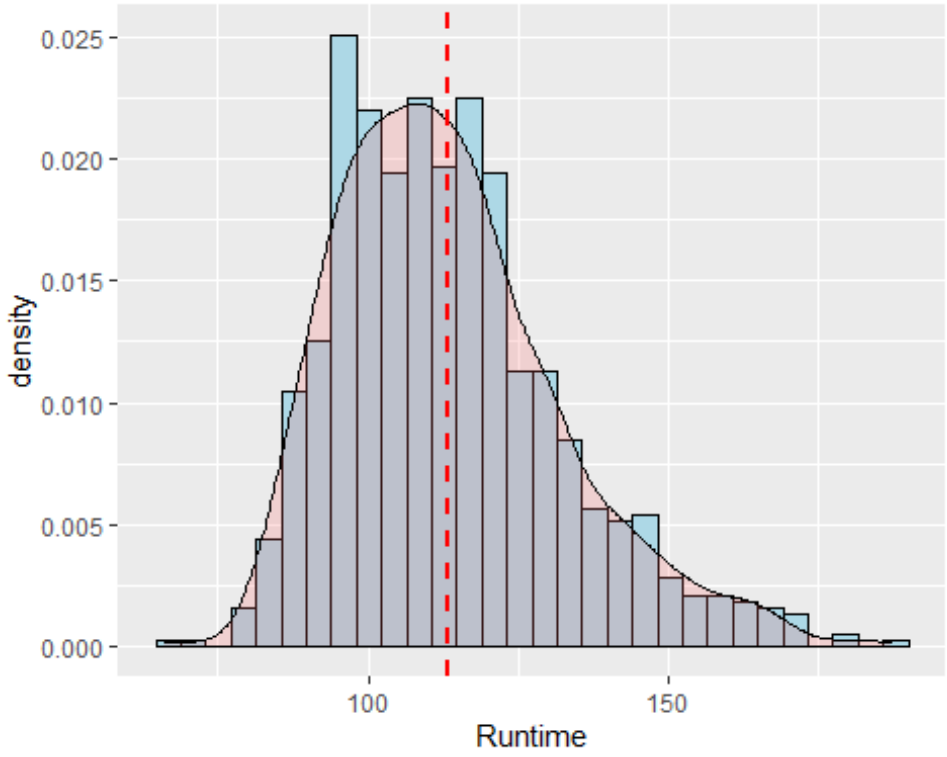
```
# Plot del resultat anterior
ggplot(top5DirectorMetascore, aes(x=reorder(Director,desc(Metascore)),
y=Metascore, fill=Movies)) +
  geom_bar(stat = "identity") +
  coord_cartesian(ylim = c(0, 120)) +
  scale_y_continuous(breaks = seq(0, 120, 10)) +
  ggtitle('Top 5 Director / Metascore') +
  labs(x = 'Director', y = 'Metascore')
```



Acabem de trobar els màxims directors en termes d'ingressos, classificació i metascore de pel·lícules.

Com afecta el temps d'execució de les pel·lícules?

```
# Esbrinem els percentils de la columna Runtime.
summary(movies_clean$Runtime)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    66.0   100.0   111.0   113.3   123.0   187.0

# Histograma de la distribució de les pel·lícules en funció de la seva duració.
ggplot(movies_clean, aes(x=Runtime)) +
 geom_histogram(aes(y=..density..), colour="black", fill="lightblue")+
 geom_density(alpha=.2, fill="#FF6666") +
 geom_vline(aes(xintercept=mean(Runtime)),
            color="red", linetype="dashed", size=1)
```



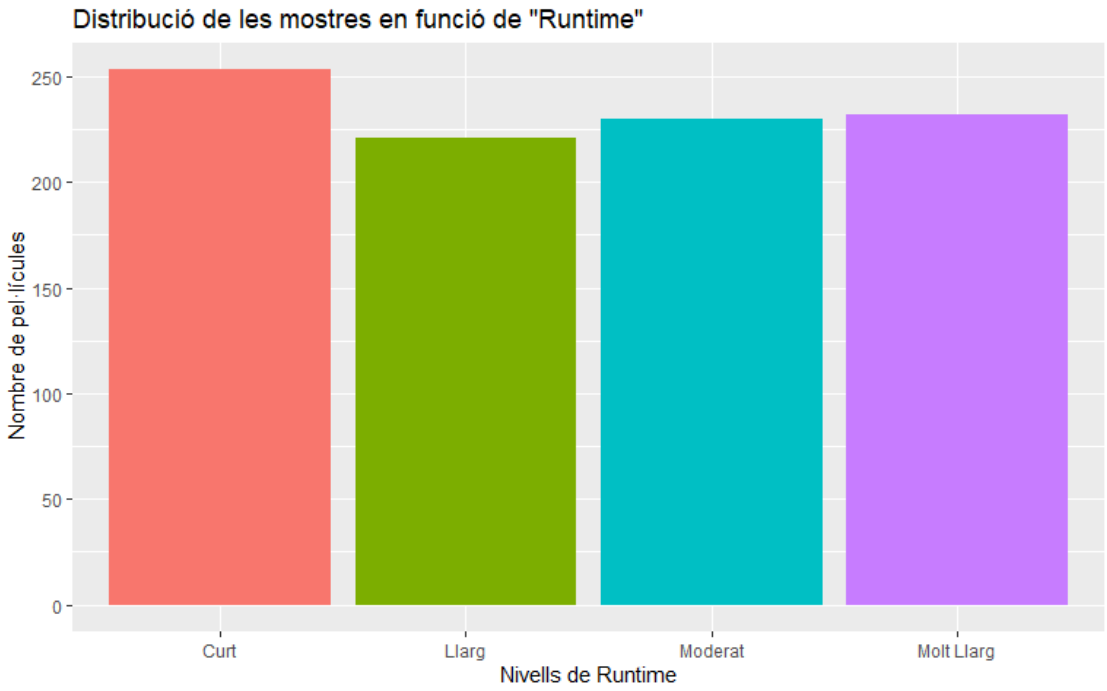
Classificarem el temps d'execució en diferents nivells en funció de la sortida del mètode de descripció:

Runtime	Nivell
0 - 100	Curt
101 - 111	Moderat
112 - 123	Llarg
124 - 187	Molt Llarg

```
# Apliquem els canvis anteriorment mencionats.
movies_clean <- mutate(movies_clean,
       Runtime_levels = ifelse(Runtime <= 100, "Curt",
                        ifelse(Runtime > 100 & Runtime <= 111, "Moderat",
                               ifelse(Runtime > 111 & Runtime <= 123,
"Llarg",
                                      ifelse(Runtime > 123 & Runtime <=
187, "Molt Llarg", 'Error')
                                      )
                               )
                        )
)

# La variable ha de ser categòrica.
movies_clean$Runtime_levels <- as.factor(movies_clean$Runtime_levels)
# Visualització del resultat.
head(movies_clean)
```

```
## # A tibble: 6 x 13
##    Rank Title Genre Description Director Actors  Year Runtime Rating
Votes
##   <dbl> <chr> <chr> <chr>       <chr>    <chr>  <dbl>   <dbl>  <dbl>
<dbl>
## 1     1 Guar~ Acti~ A group of~ James G~ Chris~  2014     121    8.1
757074
## 2     2 Prom~ Adve~ Following ~ Ridley ~ Noomi~  2012     124    7
485820
## 3     3 Split Horr~ Three girl~ M. Nigh~ James~  2016     117    7.3
157606
## 4     4 Sing  Anim~ In a city ~ Christo~ Matth~  2016     108    7.2
60545
## 5     5 Suic~ Acti~ A secret g~ David A~ Will ~  2016     123    6.2
393727
## 6     6 The ~ Acti~ European m~ Yimou Z~ Matt ~  2016     103    6.1
56036
## # ... with 3 more variables: Revenue <dbl>, Metascore <dbl>,
## #   Runtime_levels <fct>
```

```r
# Plot del resultat anterior
ggplot(movies_clean, aes(x=Runtime_levels, fill= Runtime_levels)) +
  geom_bar(show.legend = FALSE) +
  labs(x = 'Nivells de Runtime', y = ' Nombre de pel·lícules') +
  ggtitle('Distribució de les mostres en funció de "Runtime"')
```



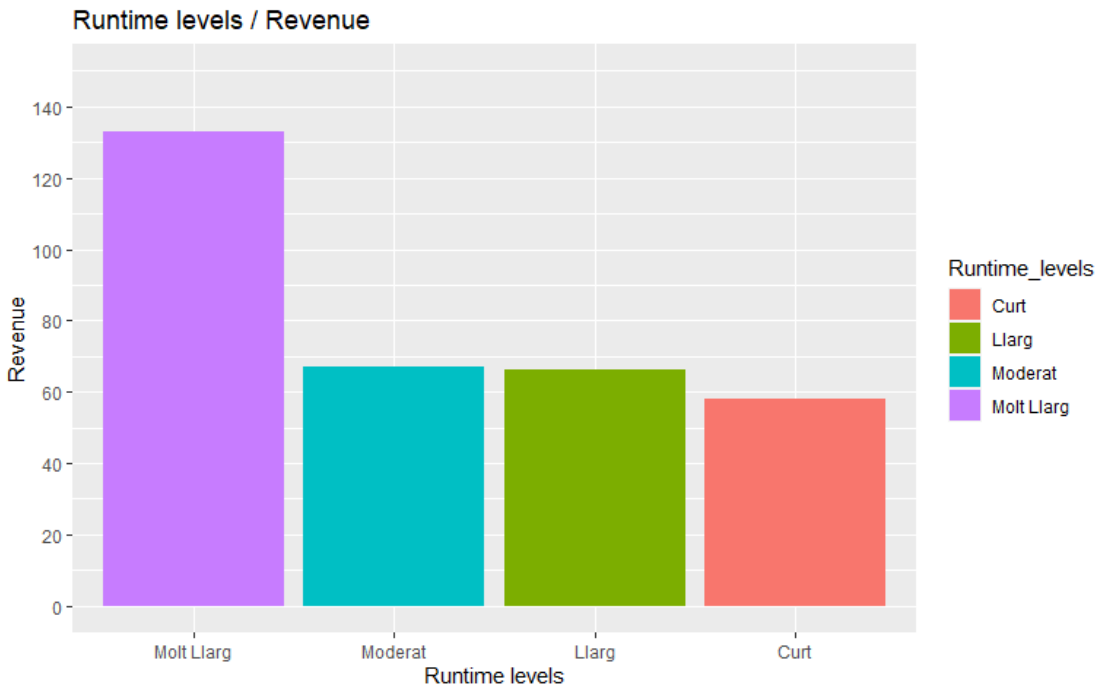
Quin interval de Runtime obté ingressos més alts?

```r
# Extreiem estadístiques de Runtime en un nou data frame.
runtime_stats <- movies_clean %>%
```

```
  group_by(Runtime_levels) %>%
  summarise(Revenue = mean(Revenue),
            Rating = mean(Rating),
            Metascore = mean(Metascore))

runtime_stats
```

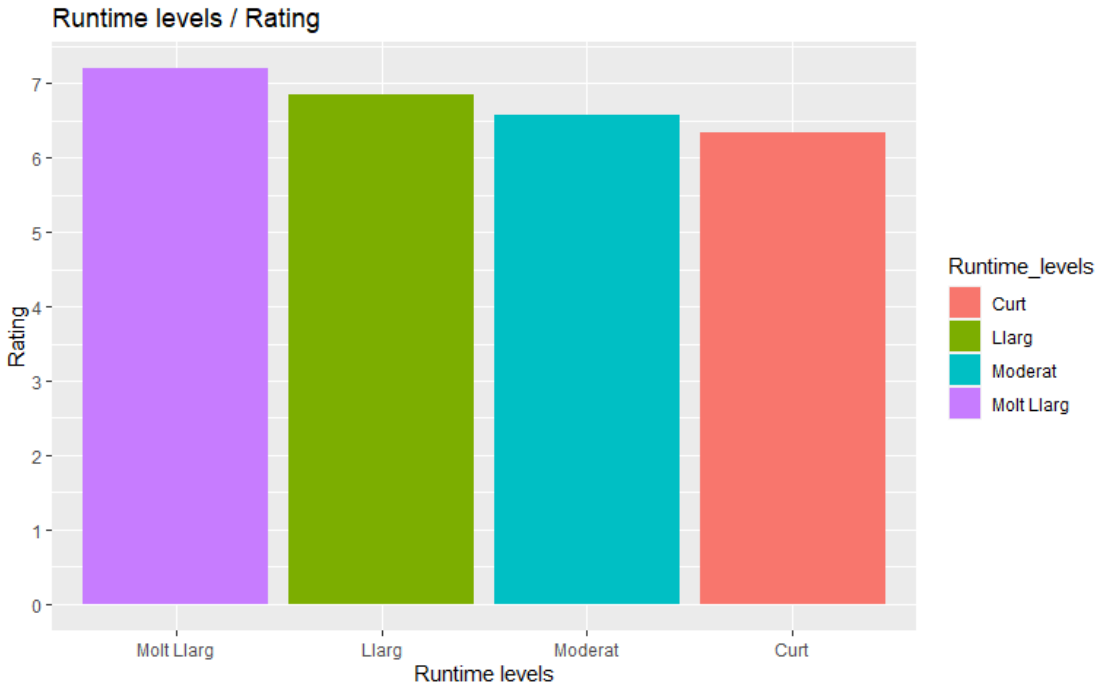
```
## # A tibble: 4 x 4
##   Runtime_levels Revenue Rating Metascore
##   <fct>            <dbl>  <dbl>     <dbl>
## 1 Curt              57.9   6.33      56.6
## 2 Llarg             66.4   6.85      59.3
## 3 Moderat           67.0   6.58      55.3
## 4 Molt Llarg       133.    7.20      65.0
```

```
# Plot de les seccions de Runtime amb els ingressos obtinguts
ggplot(runtime_stats, aes(x=reorder(Runtime_levels, desc(Revenue)),
y=Revenue, fill=Runtime_levels)) +
  geom_bar(stat = "identity") +
  coord_cartesian(ylim = c(0, 150)) +
  scale_y_continuous(breaks = seq(0, 150, 20)) +
  ggtitle('Runtime levels / Revenue') +
  labs(x = 'Runtime levels', y = 'Revenue')
```

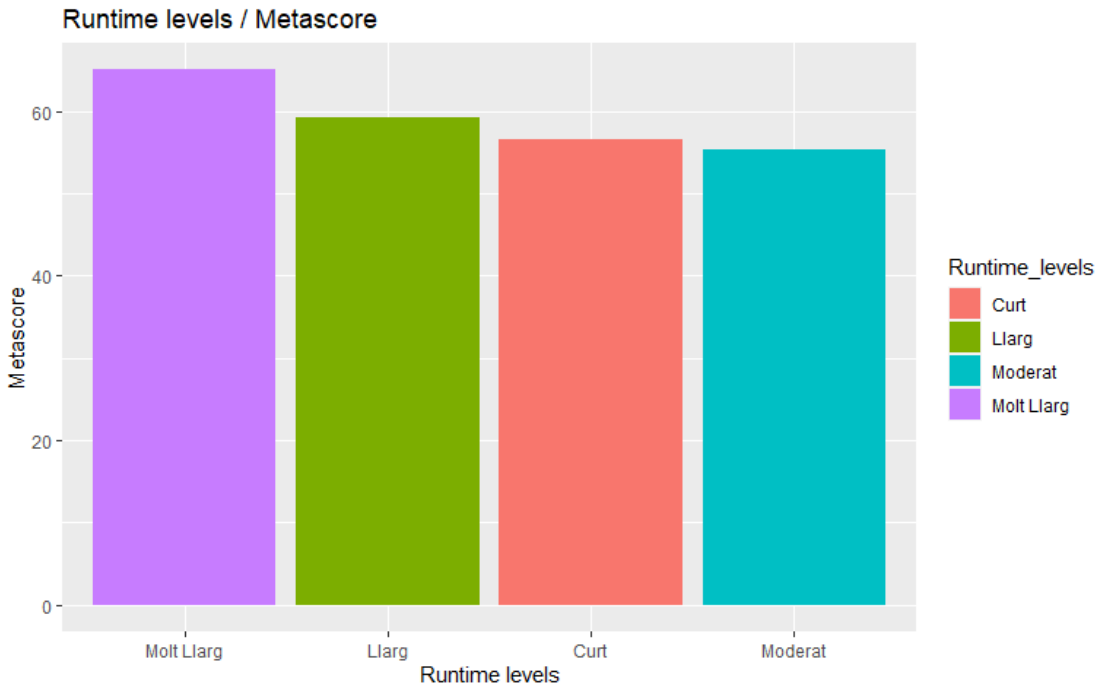



```
# Plot de les seccions de Runtime amb les qualificacions obtingudes.
ggplot(runtime_stats, aes(x=reorder(Runtime_levels, desc(Rating)),
y=Rating, fill=Runtime_levels)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 10, 1)) +
```

```
  ggtitle('Runtime levels / Rating') +
  labs(x = 'Runtime levels', y = 'Rating')
```




```
# Plot de les seccions de Runtime amb les qualificacions obtingudes a
Metacrític.
ggplot(runtime_stats, aes(x=reorder(Runtime_levels, desc(Metascore)),
y=Metascore, fill=Runtime_levels)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 100, 20)) +
  ggtitle('Runtime levels / Metascore') +
  labs(x = 'Runtime levels', y = 'Metascore')
```



Conclusions:

1. En general, les pel·lícules que tinguin una llarga durada (superior a 123 minuts) obtenen ingressos més alts, són més populars i aclamats per la crítica.
2. A mesura que augmenta el temps d'execució, les pel·lícules tendeixen a obtenir ingressos, popularitat i aclamacions més importants.
3. Cal tenir en compte que, les pel·lícules amb Runtime (superior a 123 minuts) superen altres pel·lícules en termes d'ingressos per un marge significatiu.

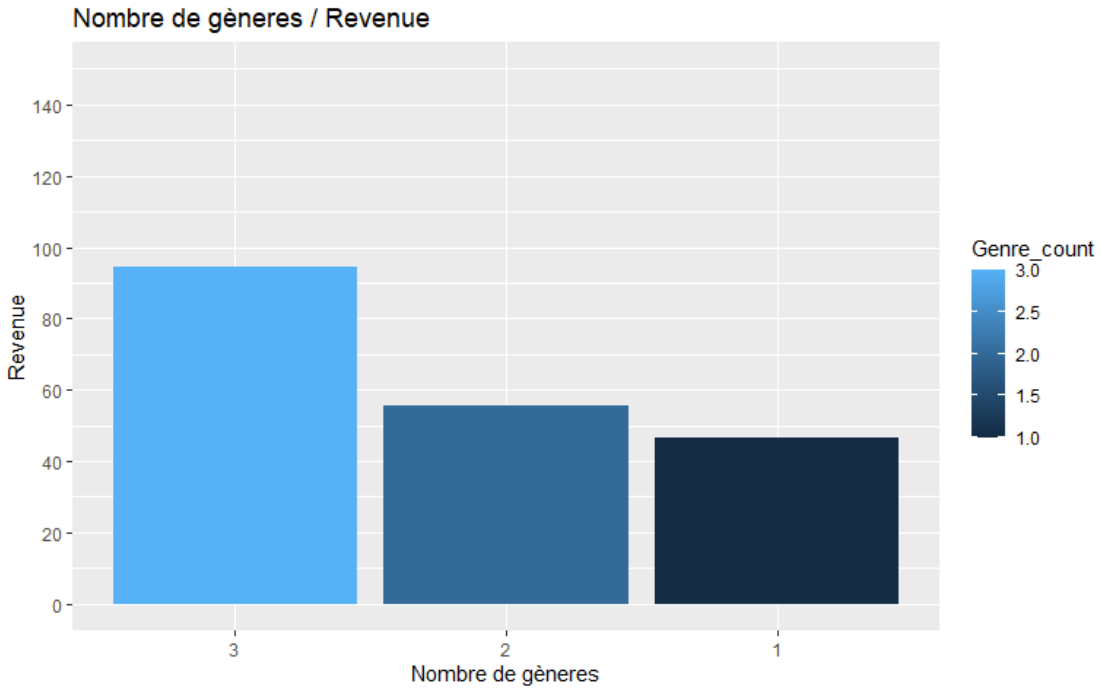
Com afecta el gènere d'una pel·lícula al resultat de la pel·lícula?

El nombre total de gèneres de la pel·lícula afecta els ingressos, la popularitat i l'aclamació crítica de la pel·lícula?

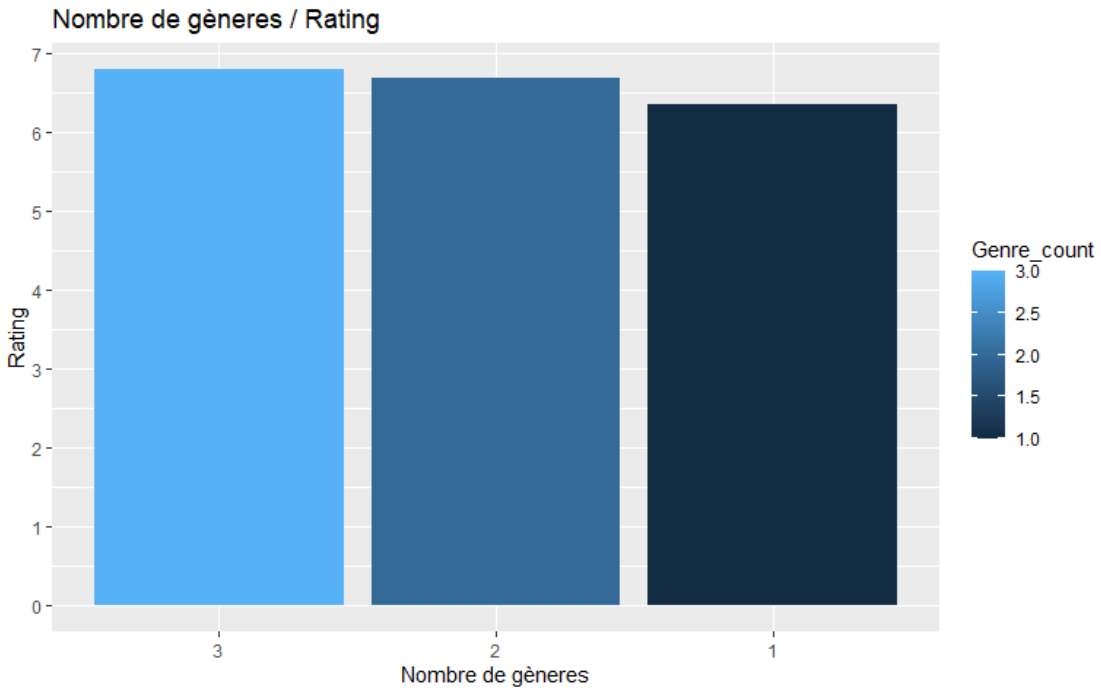
```
# La columna género conté la combinació de gèneres en format string
separat per comes.
movies_clean$Genre_count <-
sapply(strsplit(as.character(movies_clean$Genre), ","), length)
# Es crea un nou data set amb les estadístiques en funció del gènere de
la pel·lícula.
genre_count_stats <- movies_clean %>%
  group_by(Genre_count) %>%
  summarise(Revenue = mean(Revenue),
            Rating = mean(Rating),
            Metascore = mean(Metascore))

# Visualització del resultat.
genre_count_stats
```

```
## # A tibble: 3 x 4
##   Genre_count Revenue Rating Metascore
##         <int>   <dbl>  <dbl>     <dbl>
## 1           1    46.6   6.37      58.8
## 2           2    55.7   6.70      58.7
## 3           3    94.6   6.79      59.1
```

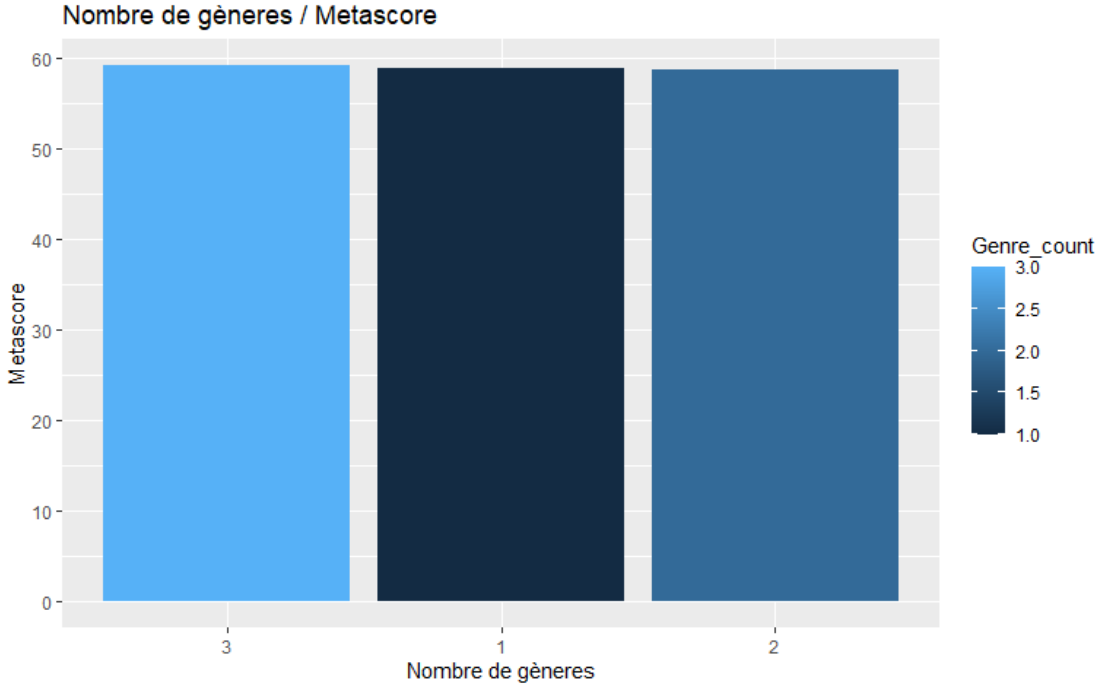
```r
# Plot de la relació del nombre de gèneres d'un títol i els seus ingressos.
ggplot(genre_count_stats, aes(x=reorder(Genre_count, desc(Revenue)), y=Revenue, fill=Genre_count)) +
  geom_bar(stat = "identity") +
  coord_cartesian(ylim = c(0, 150)) +
  scale_y_continuous(breaks = seq(0, 150, 20)) +
  ggtitle('Nombre de gèneres / Revenue') +
  labs(x = 'Nombre de gèneres', y = 'Revenue')
```




```r
# Plot de la relació del nombre de gèneres d'un títol i la seva qualificació.
ggplot(genre_count_stats, aes(x=reorder(Genre_count, desc(Rating)), y=Rating, fill=Genre_count)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 10, 1)) +
  ggtitle('Nombre de gèneres / Rating') +
  labs(x = 'Nombre de gèneres', y = 'Rating')
```




```r
# Plot de la relació del nombre de gèneres d'un títol i la seva qualificació a Metacrític.
ggplot(genre_count_stats, aes(x=reorder(Genre_count, desc(Metascore)), y=Metascore, fill=Genre_count)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  ggtitle('Nombre de gèneres / Metascore') +
  labs(x = 'Nombre de gèneres', y = 'Metascore')
```



Conclusions:

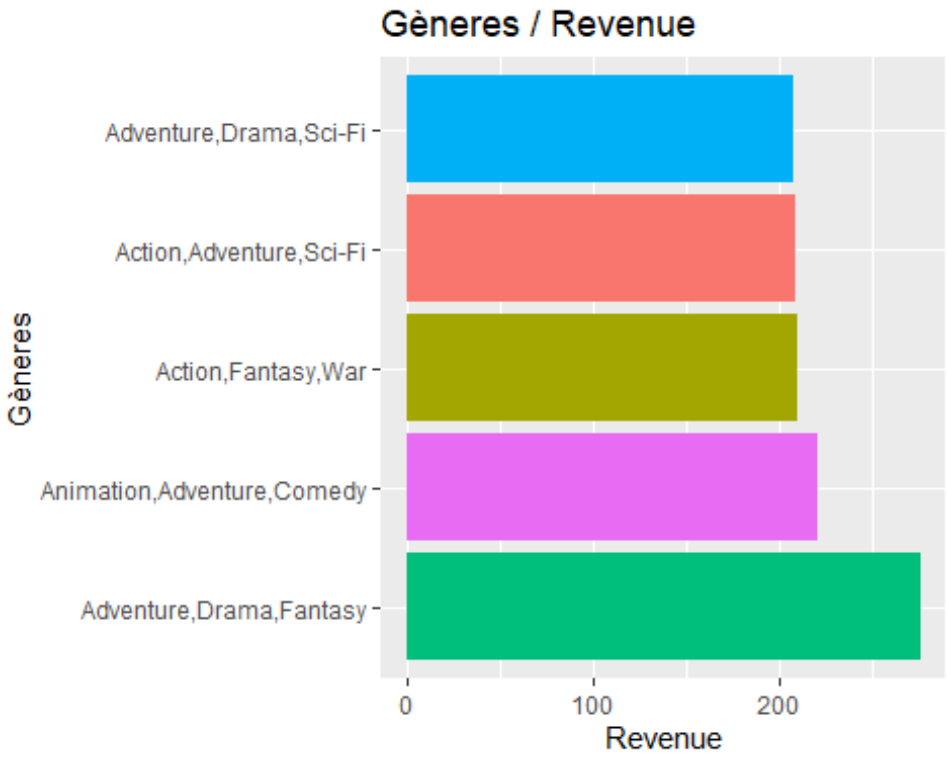
1. El nombre de gèneres de la pel·lícula augmenten significativament els ingressos de la pel·lícula. De mitjana, una pel·lícula amb 3 gèneres tendeix a guanyar el doble dels ingressos que una pel·lícula amb només 1 gènere.

2. Curiosament, el nombre de gèneres d'una pel·lícula no afecta significativament la qualificació IMDB ni el Metascore. Tot i això, tant la qualificació IMDB com Metascore van augmentant quan el nombre de gèneres augmenta.

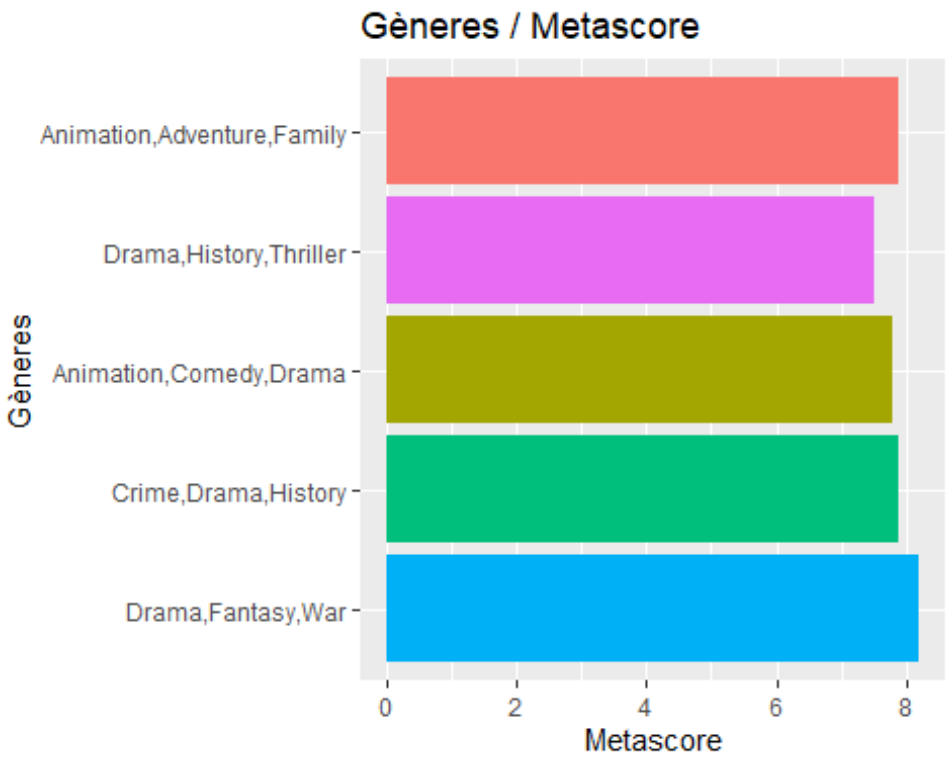
3. Si pensem en la troballa una mica detinguda, els gèneres no apareixen explícitament en una pel·lícula. Els espectadors no ho saben mentre veuen la pel·lícula. Però el nombre de gèneres té un impacte en general per a l'espectador. Tanmateix, quan un espectador s'asseu per valorar la pel·lícula, no valora segons el nombre de gèneres de la pel·lícula. La valora segons la seva semblança amb la pel·lícula en general.

Com que sabem que un nombre de 3 gèneres ofereix la millor pel·lícula, descobrim quina combinació de 3 gèneres obté més ingressos, qualificació i aclamacions crítiques?

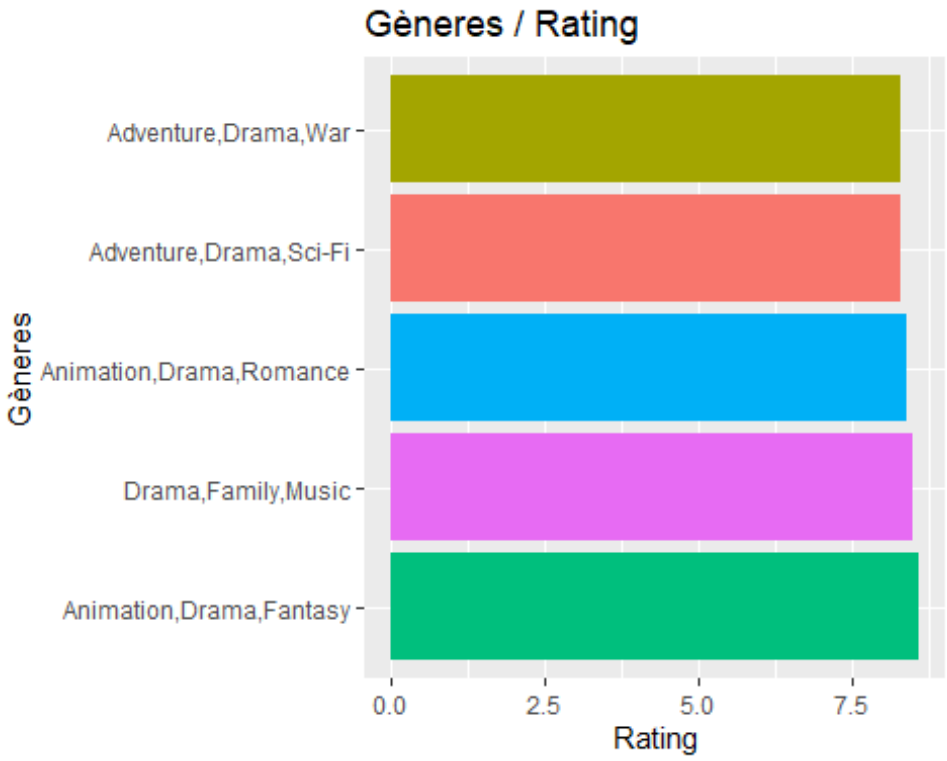
```
# Extreiem les pel·lícules amb 3 gèneres.
movies_3_genres <- subset(movies_clean, subset = Genre_count == 3)

# Generem estadístiques del data frame anterior.
genre3_count_stats <- movies_3_genres %>%
  group_by(Genre) %>%
  summarise(Revenue = mean(Revenue),
            Rating = mean(Rating),
            Metascore = mean(Metascore))

genre3_top_revenue <- genre3_count_stats %>%
  arrange(desc(Revenue)) %>%
  top_n(n=5,wt=Revenue)

# Top 5 combinació de gèneres amb millors Ingressos.
ggplot(genre3_top_revenue, aes(x=reorder(Genre, desc(Revenue)),
y=Revenue, fill=Genre)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +
  ggtitle('Gèneres / Revenue') +
  labs(x = 'Gèneres', y = 'Revenue')
```




```r
genre3_top_metascore <- genre3_count_stats %>%
  arrange(desc(Metascore)) %>%
  top_n(n=5,wt=Metascore)

# Top 5 combinació de gèneres amb millor qualificació.
ggplot(genre3_top_metascore, aes(x=reorder(Genre, desc(Metascore)),
y=Rating, fill=Genre)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +
  ggtitle('Gèneres / Metascore') +
  labs(x = 'Gèneres', y = 'Metascore')
```




```r
genre3_top_rating <- genre3_count_stats %>%
  arrange(desc(Rating)) %>%
  top_n(n=5,wt=Rating)

# Top 5 combinació de gèneres amb millor qualificació a Metacrític.
ggplot(genre3_top_rating, aes(x=reorder(Genre, desc(Rating)), y=Rating,
fill=Genre)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +
  ggtitle('Gèneres / Rating') +
  labs(x = 'Gèneres', y = 'Rating')
```

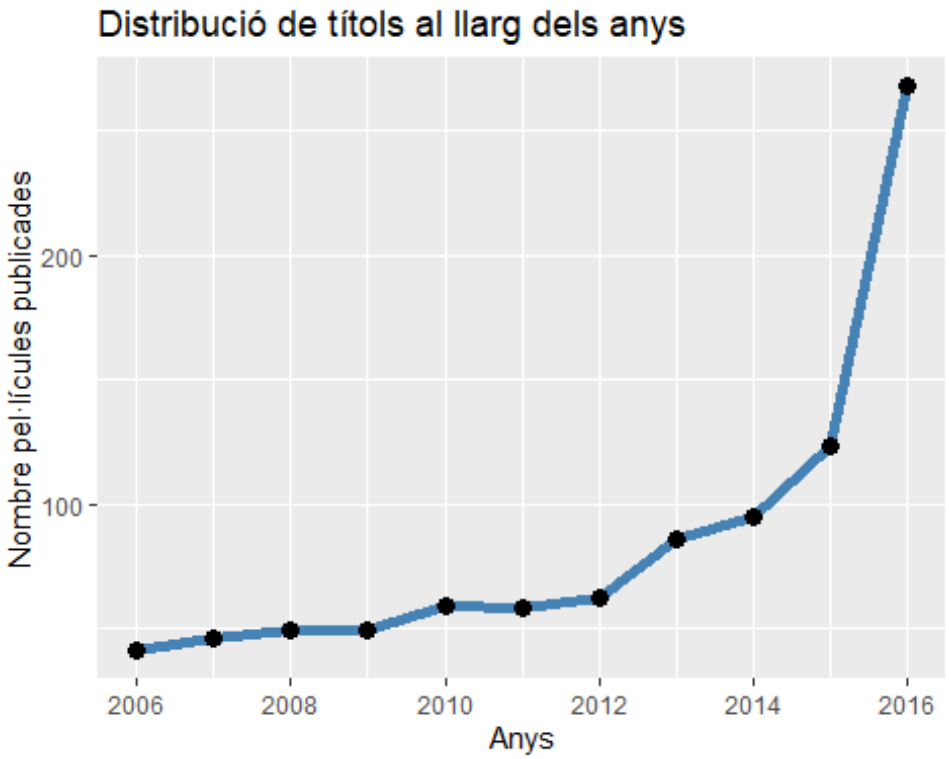


1. La combinació de gèneres: animació, drama i fantasia produeix els ingressos més alts.
2. La combinació de gèneres: animació, drama i fantasia produeix la qualificació més alta.
3. Tanmateix, als crítics els agrada principalment pel·lícules que continguin la combinació de Drama, Fantasia i Guerra

El creixement de la indústria cinematogràfica està en augment?

El nombre de pel·lícules augmenta amb els anys?

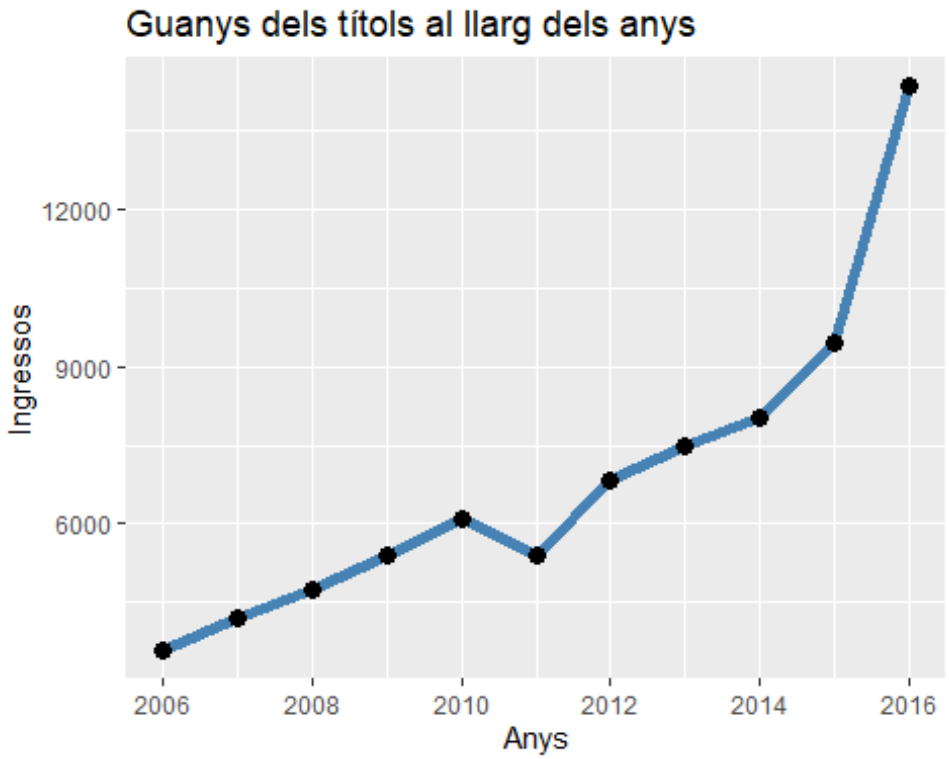
```
years_movie <- movies_clean %>%
  group_by(Year) %>%
  summarise(n = n())

# Plot del  recompte de pel·lícules per any.
ggplot(data = years_movie, aes(x=Year, y=n)) +
  geom_line(color='steelblue', size=2) +
  geom_point(size=3) +
  labs(x='Anys', y='Nombre pel·lícules publicades') +
  ggtitle('Distribució de títols al llarg dels anys')
```



1. El nombre de pel·lícules estrenades durant els darrers deu anys mostra una tendència a l'alça.
2. Hi ha un augment dramàtic (més del doble) del nombre de pel·lícules estrenades el 2016, en comparació amb el nombre de pel·lícules estrenades el 2015.

Els ingressos de pel·lícules augmenten amb els anys?

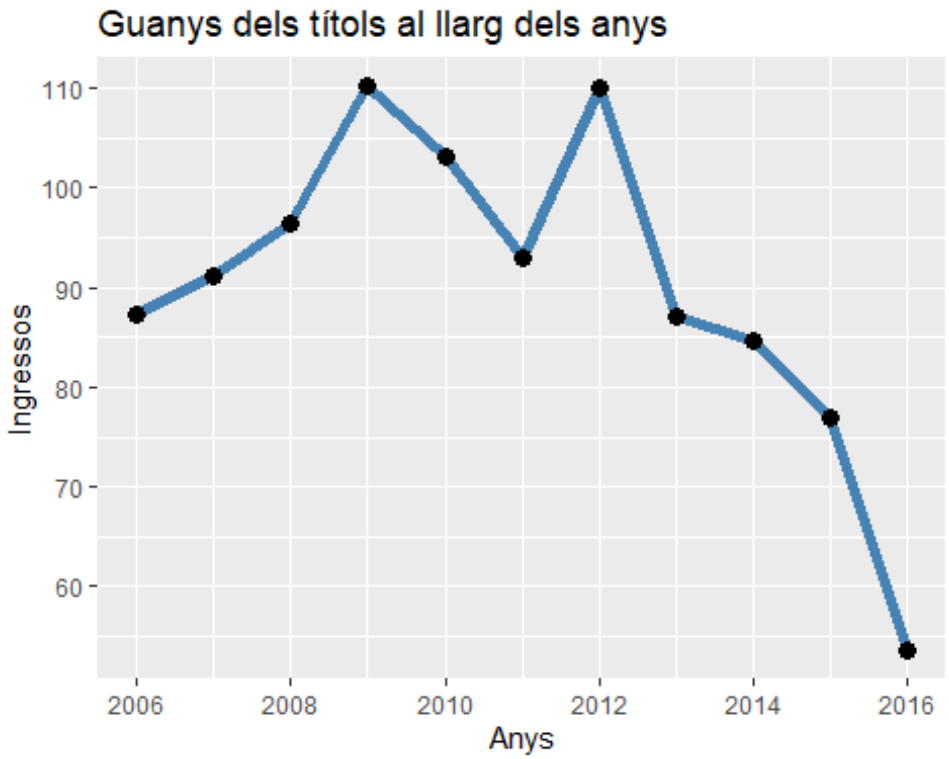
```
revenue_movie <- movies_clean %>%
  group_by(Year) %>%
  summarise(Revenue = sum(Revenue))

# Plot d'ingressos generats per pel·lícules en cada any.
ggplot(data = revenue_movie, aes(x=Year, y=Revenue)) +
  geom_line(color='steelblue', size=2) +
  geom_point(size=3) +
  labs(x='Anys', y='Ingressos') +
  ggtitle('Guanys dels títols al llarg dels anys')
```



1. Els ingressos de pel·lícules dels darrers deu anys mostren una tendència a l'alça.
2. Hi ha un augment espectacular dels ingressos de pel·lícules estrenades el 2016, en comparació amb els ingressos de pel·lícules estrenades el 2015. Això correlaciona amb l'augment dramàtic del nombre de pel·lícules estrenades el 2016 en comparació amb el 2015, tal com es mostra a la apartat anterior.

Els ingressos mitjans de les pel·lícules augmenten amb els anys?

```
revenue_movie <- movies_clean %>%
  group_by(Year) %>%
  summarise(Revenue = mean(Revenue))


ggplot(data = revenue_movie, aes(x=Year, y=Revenue)) +
  geom_line(color='steelblue', size=2) +
  geom_point(size=3) +
  labs(x='Anys', y='Ingressos') +
  ggtitle('Guanys dels títols al llarg dels anys')
```



1. Per què disminueix la recaptació mitjana de pel·lícules per any, mentre que el recaptat total de pel·lícules augmenta amb els anys?
2. Probablement és degut al fet que el nombre total de pel·lícules llançades a l'any augmenta. Aquesta tendència indica, cada cop hi ha més productors que entren a la indústria i hi ha una gran competència en el sector. Com a resultat, en un any, diverses pel·lícules es publiquen al mateix temps i es reparteixen els ingressos entre elles. Així, els ingressos mitjans de les pel·lícules en un any també disminueixen.
3. Per tant, probablement el fet de trobar el director només per ingressos mitjans, valoració mitjana i mitjana de Metascore pot no ser una bona idea. Per exemple, James Cameron només ha dirigit una sola pel·lícula (Avatar el 2009) en els darrers deu anys. En aquell any (2009), només es van llançar 50 pel·lícules, segons el gràfic de la línia de "Distribució de títols al llarg dels anys" (anterior). Però el 2016, la tendència ha canviat: el nombre de pel·lícules estrenades el 2016 és al voltant de 5 vegades la de les pel·lícules el 2009.
4. Cal esbrinar els directors que són més actius pel que fa a la direcció de més pel·lícules. Entre ells, hem d'esbrinar els directors que obtinguin més ingressos, qualificació i metascore. Així que anem a fer això a continuació

```
# Prenem la llista de tots els directors que han dirigit un nombre total
de pel·lícules de 5 o més.
director_stats <- movies_clean %>%
  group_by(Director) %>%
  summarise(Movies = n(), Rating = mean(Rating), Metascore =
mean(Metascore), Revenue = mean(Revenue))
```

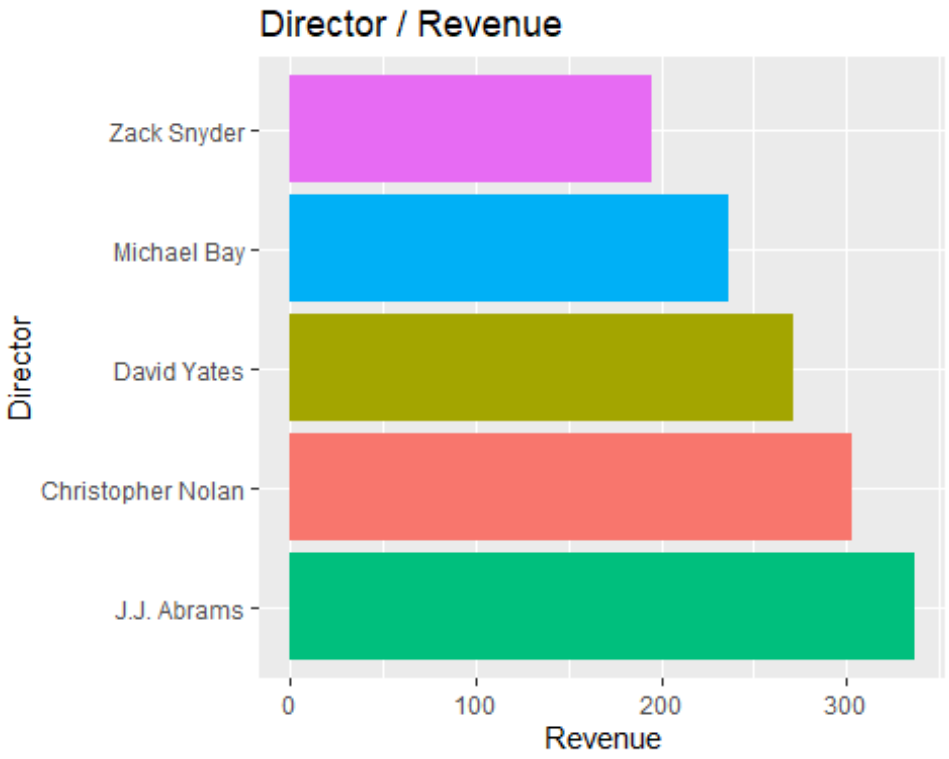
```r
director_5_stats <- director_stats[director_stats$Movies >= 5,]

head(director_5_stats)
```

```
## # A tibble: 6 x 5
##   Director          Movies Rating Metascore Revenue
##   <chr>              <int>  <dbl>     <dbl>   <dbl>
## 1 Antoine Fuqua          5   7.04      52.4    78.6
## 2 Christopher Nolan      5   8.68      74.8   303.
## 3 Danny Boyle            5   7.42      75      36.7
## 4 David Fincher          5   7.82      78.6   106.
## 5 David Yates            6   7.43      68.5   272.
## 6 Denis Villeneuve       5   7.76      75.6    43.2
```

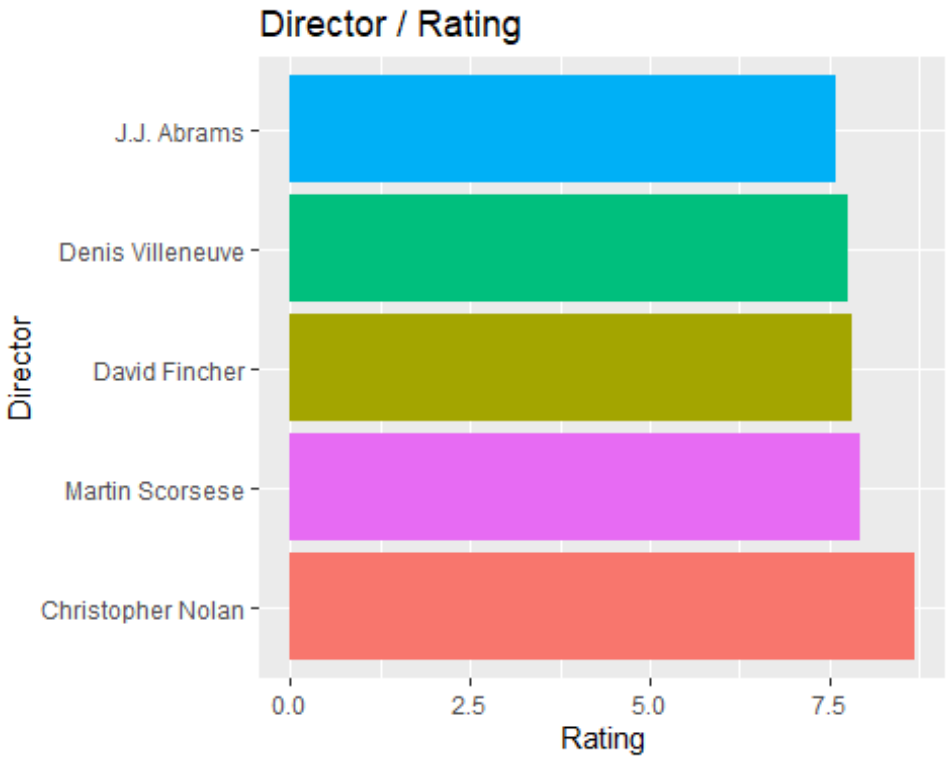
```r
# Prenem la llista de tots els directors que han dirigit un nombre total de pel·lícules de 5 o més.
director_5_stats <- director_5_stats %>%
  group_by(Director) %>%
  summarise(Revenue = mean(Revenue),
            Rating = mean(Rating),
            Metascore = mean(Metascore))
head(director_5_stats)
```

```
## # A tibble: 6 x 4
##   Director          Revenue Rating Metascore
##   <chr>               <dbl>  <dbl>     <dbl>
## 1 Antoine Fuqua        78.6   7.04      52.4
## 2 Christopher Nolan   303.    8.68      74.8
## 3 Danny Boyle          36.7   7.42      75
## 4 David Fincher       106.    7.82      78.6
## 5 David Yates         272.    7.43      68.5
## 6 Denis Villeneuve     43.2   7.76      75.6
```

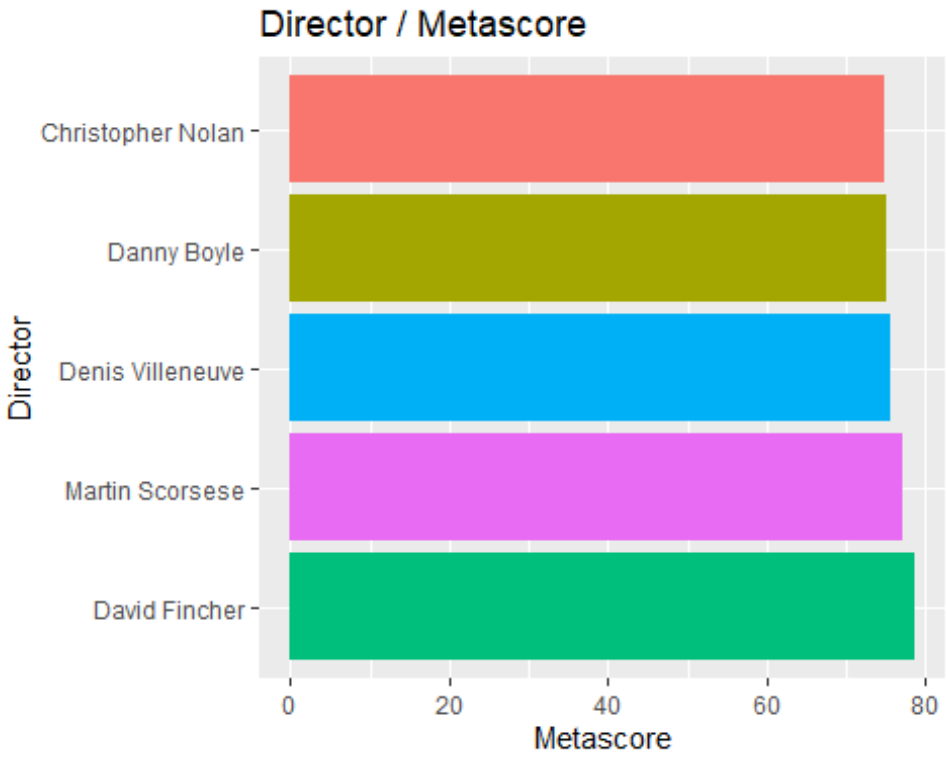
```r
# Plot dels 5 directos amb les pel·lícules més "taquilleres".
director_5_stats %>%
arrange(desc(Revenue)) %>%
  top_n(n=5,wt=Revenue) %>%
  ggplot(., aes(x=reorder(Director, desc(Revenue)), y=Revenue,
fill=Director)) +
    geom_bar(stat = "identity", show.legend = FALSE) +
    coord_flip() +
    ggtitle('Director / Revenue') +
    labs(x = 'Director', y = 'Revenue')
```




```r
# Plot dels 5 directos amb les pel·lícules amb millor qualificació.
director_5_stats %>%
arrange(desc(Rating)) %>%
  top_n(n=5,wt=Rating) %>%
  ggplot(., aes(x=reorder(Director, desc(Rating)), y=Rating,
fill=Director)) +
    geom_bar(stat = "identity", show.legend = FALSE) +
    coord_flip() +
    ggtitle('Director / Rating') +
    labs(x = 'Director', y = 'Rating')
```




```
# Plot dels 5 directos amb les pel·lícules amb millor qualificació a
Metacrític.
director_5_stats %>%
arrange(desc(Metascore)) %>%
  top_n(n=5,wt=Metascore) %>%
  ggplot(., aes(x=reorder(Director, desc(Metascore)), y=Metascore,
fill=Director)) +
    geom_bar(stat = "identity", show.legend = FALSE) +
    coord_flip() +
    ggtitle('Director / Metascore') +
    labs(x = 'Director', y = 'Metascore')
```

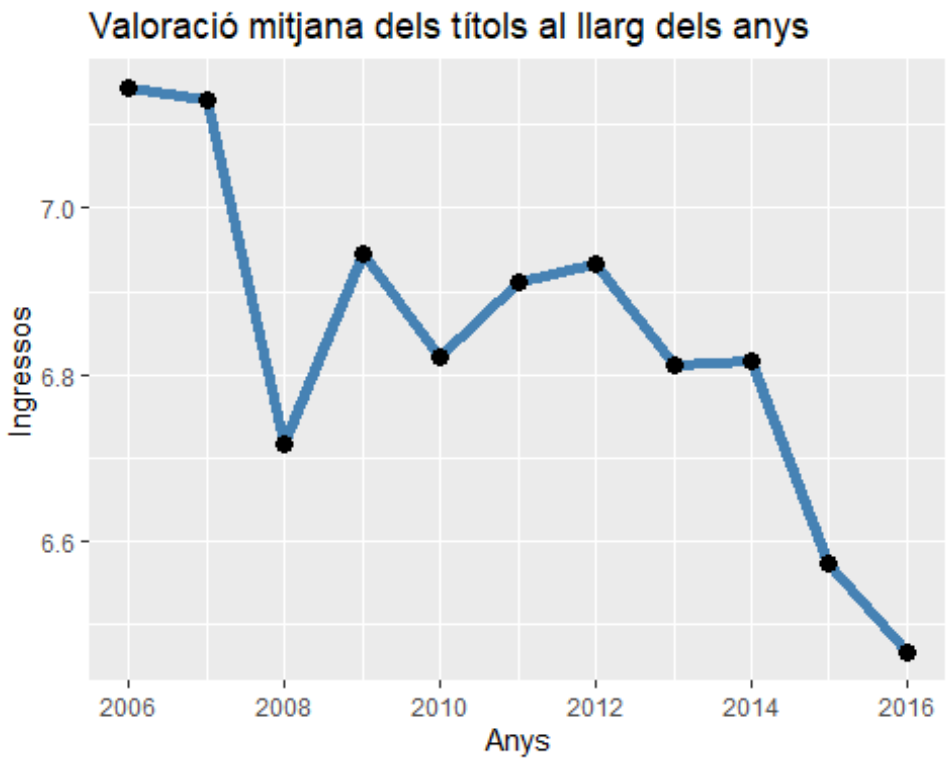


La popularitat de les pel·lícules augmenta amb els anys?

```
movies_rating <- movies_clean %>%
  group_by(Year) %>%
  summarise(Rating = mean(Rating))

movies_rating
```

```
## # A tibble: 11 x 2
##     Year Rating
##    <dbl>  <dbl>
##  1  2006   7.14
##  2  2007   7.13
##  3  2008   6.72
##  4  2009   6.94
##  5  2010   6.82
##  6  2011   6.91
##  7  2012   6.93
##  8  2013   6.81
##  9  2014   6.82
## 10  2015   6.57
## 11  2016   6.47
```

```
ggplot(movies_rating, aes(x=Year, y=Rating)) +
    geom_line(color='steelblue', size=2) +
    geom_point(size=3) +
    labs(x='Anys', y='Ingressos') +
    ggtitle('Valoració mitjana dels títols al llarg dels anys')
```



1. Per què la popularitat de les pel·lícules en termes de classificació IMDB disminueix amb els anys, tot i que els ingressos per a les pel·lícules augmenten amb els anys?
2. Analitzem les característiques de les pel·lícules el 2016.

Analitzem els gèneres, els nivells d'execució del cinema el 2016.

```
movies_clean %>%
  filter(Year == 2016) %>%
  count(Genre_count)
```

```
## # A tibble: 3 x 2
##   Genre_count     n
##         <int> <int>
## 1           1    47
## 2           2    64
## 3           3   157
```

```
movies_clean %>%
  filter(Year == 2016) %>%
  count(Genre) %>%
  arrange(desc(n)) %>%
  top_n(n=5,wt=n)
```

```
## # A tibble: 5 x 2
##   Genre                       n
##   <chr>                   <int>
```

```
## 1 Drama                          23
## 2 Comedy                         13
## 3 Comedy,Drama                   10
## 4 Horror,Thriller                10
## 5 Animation,Adventure,Comedy      9
```

Les 3 combinacions de gènere més populars són (tal com es troba a la secció 4.3.2):

'Animació, Drama, Fantasia' 'Drama, família, música' 'Animació, comèdia, drama'
Analitzem el nombre de pel·lícules d'aquestes combinacions de gènere el 2016

```
movies_clean %>%
  filter(Genre == 'Animation,Drama,Fantasy') %>%
  filter(Year == 2016)
```

```
## # A tibble: 1 x 14
##    Rank Title Genre Description Director Actors  Year Runtime Rating
Votes
##   <dbl> <chr> <chr> <chr>       <chr>    <chr>  <dbl>   <dbl>  <dbl>
<dbl>
## 1    97 Kimi~ Anim~ Two strang~ Makoto ~ Ryûno~  2016     106    8.6
34110
## # ... with 4 more variables: Revenue <dbl>, Metascore <dbl>,
## #   Runtime_levels <fct>, Genre_count <int>
```

```
movies_clean %>%
  filter(Genre == 'Drama,Family,Music') %>%
  filter(Year == 2016)
```

```
## # A tibble: 0 x 14
## # ... with 14 variables: Rank <dbl>, Title <chr>, Genre <chr>,
## #   Description <chr>, Director <chr>, Actors <chr>, Year <dbl>,
Runtime <dbl>,
## #   Rating <dbl>, Votes <dbl>, Revenue <dbl>, Metascore <dbl>,
## #   Runtime_levels <fct>, Genre_count <int>
```

```
movies_clean %>%
  filter(Genre == 'Animation,Comedy,Drama') %>%
  filter(Year == 2016)
```

```
## # A tibble: 1 x 14
##    Rank Title Genre Description Director Actors  Year Runtime Rating
Votes
##   <dbl> <chr> <chr> <chr>       <chr>    <chr>  <dbl>   <dbl>  <dbl>
<dbl>
## 1   794 Ma v~ Anim~ After losi~ Claude ~ Gaspa~  2016      66    7.8
4370
## # ... with 4 more variables: Revenue <dbl>, Metascore <dbl>,
## #   Runtime_levels <fct>, Genre_count <int>
```

A la secció 6.2, vam veure que les pel·lícules amb temps de durada llargs solen obtenir una qualificació de IMDB més. Analitzem el nivell d'execució de les pel·lícules el 2016

```
movies_clean %>%
  filter(Year == 2016) %>%
  count(Runtime_levels)
```

```
## # A tibble: 4 x 2
##   Runtime_levels     n
##   <fct>          <int>
## 1 Curt             101
## 2 Llarg             53
## 3 Moderat           69
## 4 Molt Llarg        45
```

Comparem els nivells d'execució de la pel·lícula del 2006

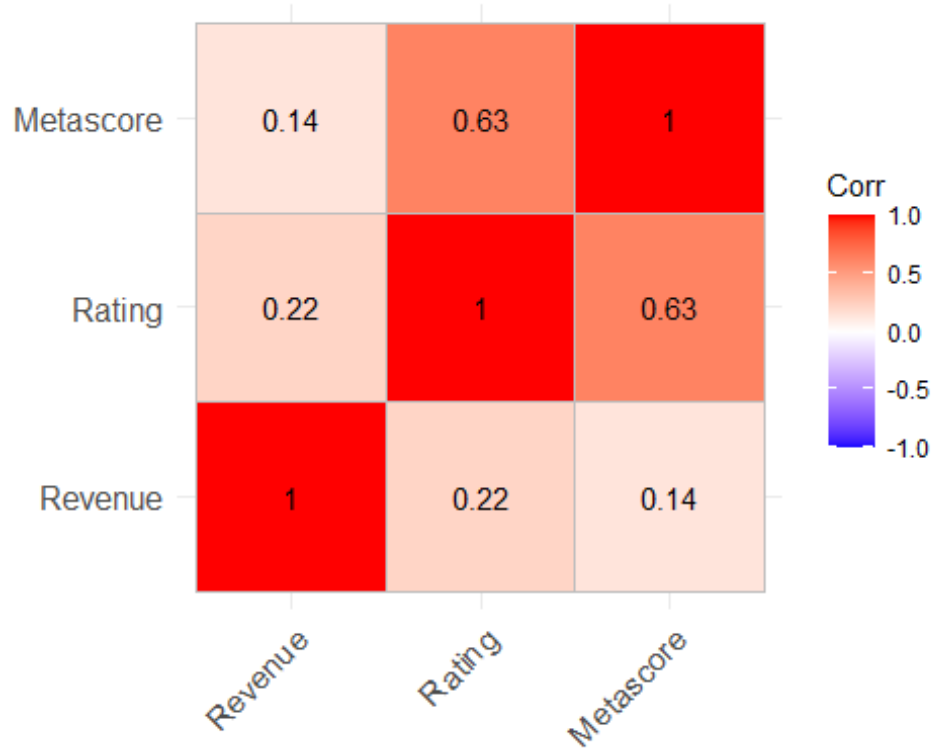
```
movies_clean %>%
  filter(Year == 2006) %>%
  count(Runtime_levels)
```

```
## # A tibble: 4 x 2
##   Runtime_levels     n
##   <fct>          <int>
## 1 Curt               3
## 2 Llarg              9
## 3 Moderat           16
## 4 Molt Llarg        13
```

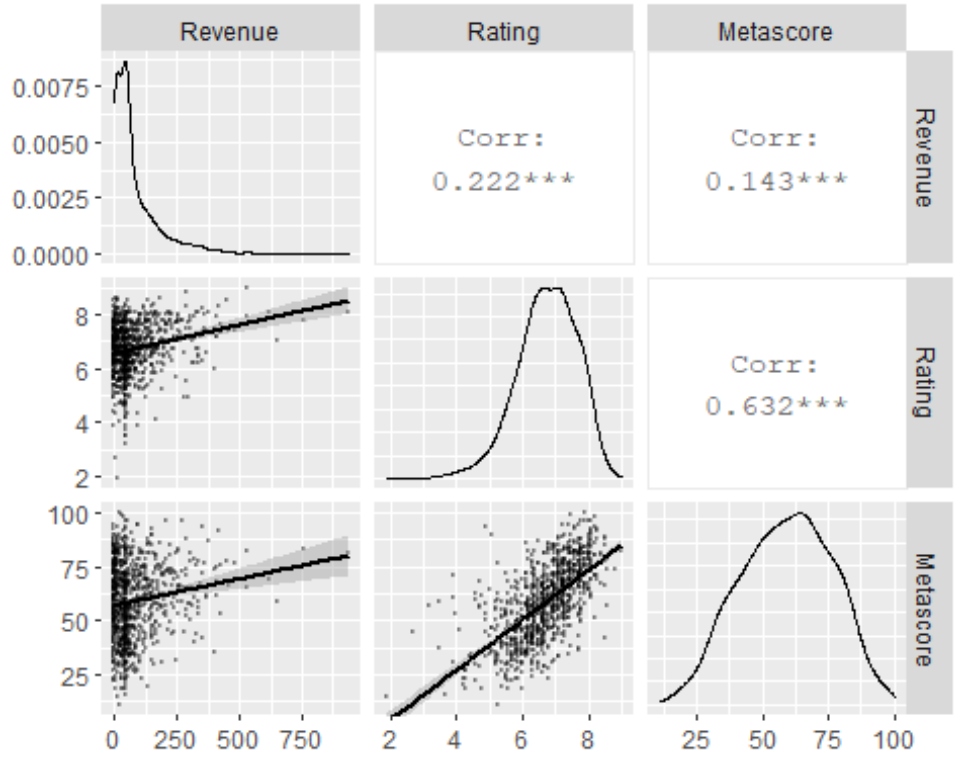
1. Sembla ser que la popularitat de les pel·lícules en termes de qualificació IMDB està disminuint a causa del nombre menor de pel·lícules amb la combinació de gènere que podria obtenir més valoració.
2. A més, els nivells de Runtime de pel·lícules del 2016 són més baixos. Hem vist a la secció 6.2 que les pel·lícules amb nivell d'execució "llarga" aporten més qualificació.

Quina relació hi ha entre Ingressos, Classificació i Metascore de les pel·lícules?

```
library(ggcorrplot)
library(GGally)
movies_clean %>%
  select(Revenue, Rating, Metascore) %>%
  cor %>%
  ggcorrplot(.,hc.order = TRUE, lab = TRUE)
```




```
movies_clean %>%
  select(Revenue, Rating, Metascore) %>%
  ggpairs(.,
          lower = list(continuous = wrap("smooth", alpha = 0.3, size=0.1)
          ))
```



Classificació i Metascore tenen una forta correlació. Per tant, significa que els usuaris registrats a IMDB i els crítics de Metacritic tendeixen a estar d'acord entre ells per a la majoria de pel·lícules.

Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Aquesta secció treu la conclusió de l'exploració feta al conjunt de dades de la secció 4.

Director

1. El director que ha obtingut els ingressos mitjans més alts és James Cameron. Tot i això, només ha dirigit 1 pel·lícula durant els 10 anys.
2. Les pel·lícules de Christopher Nolan són les més populars entre els espectadors, ja que la nota mitjana de les seves pel·lícules és la més alta dels 10 anys.
3. Els crítics aprecien molt les pel·lícules de Barry Jenkins. El segueixen de prop Kenneth Lonergan i Todd Haynes.
4. Pel que fa a la tendència de la indústria cinematogràfica, es mostren més afavorits els directors més actius.

5. Entre els directors més actius, les pel·lícules de J.J Abrams guanyen més quant a ingressos mitjans.
6. Christopher Nolan és el director actiu més popular pel que fa a la valoració mitjana entre el públic.
7. Pel que fa als directors que són més actius, els crítics afavoreixen David Fincher en termes de Metascore mitjà.

Temps d'execució

1. Les pel·lícules amb llarg temps (> 123 minuts) guanyen més quant a ingressos, qualificació i metascore.
2. Els ingressos són dramàticament alts per a les pel·lícules amb llargs temps d'execució.

Gènere

1. A mesura que el nombre de gèneres augmenta en una pel·lícula, els seus ingressos, qualificació i Metascore van augmentant.
2. Tot i això, els ingressos són significativament alts per a les pel·lícules amb un nombre de 3 gèneres.
3. L'aventura com a gènere és un factor habitual per a les pel·lícules que aporten més ingressos, amb la combinació de gèneres de "Aventura, Drama, Fantasia" que obté el màxim ingressos
4. El gènere més popular és Drama i la combinació de gèneres "Animació, Drama, Fantasia" obté la qualificació més alta.
5. De nou, el drama també és popular entre els crítics, amb la combinació de gènere de "Drama, Fantasia, Guerra" que obté el Metascore més alt.

Creixement de la indústria

1. La indústria creix respecte al nombre de pel·lícules estrenades i els ingressos totals guanyats any rere any.
2. Tot i això, els ingressos mitjans de les pel·lícules al llarg dels any mostra una tendència negativa. Probablement a causa de la competència creixent i de més pel·lícules llançades a la indústria.
3. La popularitat de les pel·lícules també mostra una tendència negativa. Probablement a causa de més pel·lícules llançades amb combinacions de gènere que no són populars entre els espectadors.

Classificació per a pel·lícules en general

1. Des del gràfic de correlació, queda clar que el públic i els crítics valoren les pel·lícules d'una manera similar.
2. En general, les pel·lícules de major qualificació i metascore solen obtenir més ingressos.

Perspectives d'acció

En aquesta secció es descriuen les accions que l'empresa ABC de producció pot dur a terme per assolir el seu objectiu. Això es basa en les conclusions extretes de l'EDA.

1. **Produir múltiples pel·lícules amb millors característiques**: La indústria cinematogràfica està creixent molt ràpidament. Més pel·lícules s'estrenen any rere any i la competència és molt elevada provocant la repartició d'ingressos entre moltes pel·lícules. No seria una bona idea esperar una 'ONE BIG MOVIE' com Avatar (de James Cameron) que aporti els ingressos, la qualificació i el metascore més alts. Per tant, cal produir més pel·lícules utilitzant les millors funcions (explicades a continuació) per obtenir els màxims beneficis.

2. **Produir pel·lícules amb realitzadors actius**: Aquelles persones que dirigeixen pel·lícules de pressupost molt moderat. Per exemple, les pel·lícules de Christopher Nolan aporten més ingressos, qualificació i metascore.

3. **Produïu pel·lícules que tinguin una durada d'execució llarga** - Runtime superior a 2 hores.

4. **Crea pel·lícules amb 3 combinacions de gènere**: Inclou una combinació de drama, animació, aventura, ciència-ficció, fantasia. Les pel·lícules haurien de relacionar-se amb el públic amb incidents de la vida real, haurien de fer-los 'sentir' l'emoció plasmada en la cinta i haurien de potenciar la seva imaginació.

5. **Produïu pel·lícules que busquin satisfer tant persones com crítics**: aquestes pel·lícules obtindran ingressos més sovint.