Movies

2025-01-28

library(dplyr)

##   
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(readr)  
library(scales)

##   
## Adjuntando el paquete: 'scales'

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(stringr)

1. Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

movies <- read\_csv('./movies.csv')

## Rows: 10000 Columns: 27  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (14): genres, homePage, productionCompany, productionCompanyCountry, pr...  
## dbl (11): id, budget, revenue, runtime, popularity, voteAvg, voteCount, gen...  
## lgl (1): video  
## date (1): releaseDate  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(movies)

## # A tibble: 6 × 27  
## id budget genres homePage productionCompany productionCompanyCou…¹  
## <dbl> <dbl> <chr> <chr> <chr> <chr>   
## 1 5 4000000 Crime|Comedy https:/… Miramax|A Band A… US|US   
## 2 6 21000000 Action|Thril… <NA> Universal Pictur… US|US|JP   
## 3 11 11000000 Adventure|Ac… http://… Lucasfilm|20th C… US|US   
## 4 12 94000000 Animation|Fa… http://… Pixar US   
## 5 13 55000000 Comedy|Drama… <NA> Paramount|The St… US|   
## 6 14 15000000 Drama <NA> Jinks/Cohen Comp… |US   
## # ℹ abbreviated name: ¹​productionCompanyCountry  
## # ℹ 21 more variables: productionCountry <chr>, revenue <dbl>, runtime <dbl>,  
## # video <lgl>, director <chr>, actors <chr>, actorsPopularity <chr>,  
## # actorsCharacter <chr>, originalTitle <chr>, title <chr>,  
## # originalLanguage <chr>, popularity <dbl>, releaseDate <date>,  
## # voteAvg <dbl>, voteCount <dbl>, genresAmount <dbl>,  
## # productionCoAmount <dbl>, productionCountriesAmount <dbl>, …

summary(movies)

## id budget genres homePage   
## Min. : 5 Min. : 0 Length:10000 Length:10000   
## 1st Qu.: 12286 1st Qu.: 0 Class :character Class :character   
## Median :152558 Median : 500000 Mode :character Mode :character   
## Mean :249877 Mean : 18551632   
## 3rd Qu.:452022 3rd Qu.: 20000000   
## Max. :922260 Max. :380000000   
## productionCompany productionCompanyCountry productionCountry   
## Length:10000 Length:10000 Length:10000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## revenue runtime video director   
## Min. :0.000e+00 Min. : 0.0 Mode :logical Length:10000   
## 1st Qu.:0.000e+00 1st Qu.: 90.0 FALSE:9430 Class :character   
## Median :1.631e+05 Median :100.0 TRUE :84 Mode :character   
## Mean :5.674e+07 Mean :100.3 NA's :486   
## 3rd Qu.:4.480e+07 3rd Qu.:113.0   
## Max. :2.847e+09 Max. :750.0   
## actors actorsPopularity actorsCharacter originalTitle   
## Length:10000 Length:10000 Length:10000 Length:10000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## title originalLanguage popularity releaseDate   
## Length:10000 Length:10000 Min. : 4.258 Min. :1902-04-17   
## Class :character Class :character 1st Qu.: 14.578 1st Qu.:2004-11-10   
## Mode :character Mode :character Median : 21.906 Median :2013-03-08   
## Mean : 51.394 Mean :2009-01-09   
## 3rd Qu.: 40.654 3rd Qu.:2018-03-07   
## Max. :11474.647 Max. :2022-01-21   
## voteAvg voteCount genresAmount productionCoAmount  
## Min. : 1.300 Min. : 1 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 5.900 1st Qu.: 120 1st Qu.: 2.000 1st Qu.: 2.000   
## Median : 6.500 Median : 415 Median : 3.000 Median : 3.000   
## Mean : 6.483 Mean : 1342 Mean : 2.596 Mean : 3.171   
## 3rd Qu.: 7.200 3rd Qu.: 1316 3rd Qu.: 3.000 3rd Qu.: 4.000   
## Max. :10.000 Max. :30788 Max. :16.000 Max. :89.000   
## productionCountriesAmount actorsAmount castWomenAmount   
## Min. : 0.000 Min. : 0 Length:10000   
## 1st Qu.: 1.000 1st Qu.: 13 Class :character   
## Median : 1.000 Median : 21 Mode :character   
## Mean : 1.751 Mean : 2148   
## 3rd Qu.: 2.000 3rd Qu.: 36   
## Max. :155.000 Max. :919590   
## castMenAmount   
## Length:10000   
## Class :character   
## Mode :character   
##   
##   
##

1. Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)

str(movies)

## spc\_tbl\_ [10,000 × 27] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ id : num [1:10000] 5 6 11 12 13 14 15 16 18 19 ...  
## $ budget : num [1:10000] 4.0e+06 2.1e+07 1.1e+07 9.4e+07 5.5e+07 ...  
## $ genres : chr [1:10000] "Crime|Comedy" "Action|Thriller|Crime" "Adventure|Action|Science Fiction" "Animation|Family" ...  
## $ homePage : chr [1:10000] "https://www.miramax.com/movie/four-rooms/" NA "http://www.starwars.com/films/star-wars-episode-iv-a-new-hope" "http://movies.disney.com/finding-nemo" ...  
## $ productionCompany : chr [1:10000] "Miramax|A Band Apart" "Universal Pictures|Largo Entertainment|JVC" "Lucasfilm|20th Century Fox" "Pixar" ...  
## $ productionCompanyCountry : chr [1:10000] "US|US" "US|US|JP" "US|US" "US" ...  
## $ productionCountry : chr [1:10000] "United States of America" "Japan|United States of America" "United States of America" "United States of America" ...  
## $ revenue : num [1:10000] 4.26e+06 1.21e+07 7.75e+08 9.40e+08 6.77e+08 ...  
## $ runtime : num [1:10000] 98 110 121 100 142 122 119 141 126 149 ...  
## $ video : logi [1:10000] FALSE FALSE NA NA FALSE FALSE ...  
## $ director : chr [1:10000] "Allison Anders|Alexandre Rockwell|Robert Rodriguez|Quentin Tarantino" "Stephen Hopkins" "George Lucas" "Andrew Stanton" ...  
## $ actors : chr [1:10000] "Tim Roth|Jennifer Beals|Antonio Banderas|Valeria Golino|David Proval|Sammi Davis|Amanda de Cadenet|Madonna|Ione Skye|Lili Taylo "Emilio Estevez|Cuba Gooding Jr.|Denis Leary|Stephen Dorff|Jeremy Piven|Peter Greene|Michael DeLorenzo|Everlast|Michael Wiseman| "FALSE" "FALSE" ...  
## $ actorsPopularity : chr [1:10000] "22.225|23.519|17.816|19.893|9.027|7.147|7.769|6.476|7.906|12.838|17.988|20.027|4.47|0.6|1.286|23.95|22.082|3.44"| \_\_truncated\_\_ "9.008|6.383|10.757|18.295|11.772|14.777|9.669|0.98|4.369|0.6|1.38|2.835|3.807|1.4|1.111" "11.881|24.542|14.434|10.651|6.888|5.811|1.432|1.288|2.086|8.166|1.701|2.033|0.98|3.702|2.392|0.98|2.769|1.459|1"| \_\_truncated\_\_ "9.79|8.084|8.538|33.379|11.733|11.866|8.141|12.095|1.768|5.042|11.103|6.336|9.649|5.97|9.477|2.766|9.685|9.524|"| \_\_truncated\_\_ ...  
## $ actorsCharacter : chr [1:10000] "Ted the Bellhop|Angela|Man|Athena|Siegfried|Jezebel|Diana|Elspeth|Eva|Raven|Kiva|Wife|Sarah|Juancho|Corpse|TV Dancing Girl|Ches "Frank Wyatt|Mike Peterson|Fallon|John Wyatt|Ray Cochran|Sykes|Teddy, the Kid|Rhodes|Travis|Dre|Buck|Linda Wyatt|Clarissa|Rita|C "Luke Skywalker|Han Solo|Princess Leia Organa|Grand Moff Tarkin|Obi-Wan \"Ben\" Kenobi|C-3PO|R2-D2|Chewbacca|Darth Vader (perfor "Marlin (voice)|Dory (voice)|Nemo (voice)|Gill (voice)|Bloat (voice)|Peach (voice)|Gurgle (voice)|Bubbles (voice)|Deb / Flo (voi ...  
## $ originalTitle : chr [1:10000] "Four Rooms" "Judgment Night" "Star Wars" "Finding Nemo" ...  
## $ title : chr [1:10000] "Four Rooms" "Judgment Night" "Star Wars" "Finding Nemo" ...  
## $ originalLanguage : chr [1:10000] "en" "en" "en" "en" ...  
## $ popularity : num [1:10000] 20.9 9.6 100 134.4 58.8 ...  
## $ releaseDate : Date[1:10000], format: "1995-12-09" "1993-10-15" ...  
## $ voteAvg : num [1:10000] 5.7 6.5 8.2 7.8 8.5 8 8 7.9 7.5 8.2 ...  
## $ voteCount : num [1:10000] 2077 223 16598 15928 22045 ...  
## $ genresAmount : num [1:10000] 2 3 3 2 3 1 2 2 5 2 ...  
## $ productionCoAmount : num [1:10000] 2 3 2 1 2 2 2 26 2 1 ...  
## $ productionCountriesAmount: num [1:10000] 1 2 1 1 1 1 1 12 1 1 ...  
## $ actorsAmount : num [1:10000] 25 15 105 24 76 40 152 29 117 24 ...  
## $ castWomenAmount : chr [1:10000] "15" "3" "5" "5" ...  
## $ castMenAmount : chr [1:10000] "9" "9" "62" "18" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. id = col\_double(),  
## .. budget = col\_double(),  
## .. genres = col\_character(),  
## .. homePage = col\_character(),  
## .. productionCompany = col\_character(),  
## .. productionCompanyCountry = col\_character(),  
## .. productionCountry = col\_character(),  
## .. revenue = col\_double(),  
## .. runtime = col\_double(),  
## .. video = col\_logical(),  
## .. director = col\_character(),  
## .. actors = col\_character(),  
## .. actorsPopularity = col\_character(),  
## .. actorsCharacter = col\_character(),  
## .. originalTitle = col\_character(),  
## .. title = col\_character(),  
## .. originalLanguage = col\_character(),  
## .. popularity = col\_double(),  
## .. releaseDate = col\_date(format = ""),  
## .. voteAvg = col\_double(),  
## .. voteCount = col\_double(),  
## .. genresAmount = col\_double(),  
## .. productionCoAmount = col\_double(),  
## .. productionCountriesAmount = col\_double(),  
## .. actorsAmount = col\_double(),  
## .. castWomenAmount = col\_character(),  
## .. castMenAmount = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

1. Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados
2. Responda las siguientes preguntas

4.1 ¿Cuáles son las 10 películas que contaron con más presupuesto?

top\_budget\_movies <- movies %>%   
 arrange(desc(budget)) %>%   
 select(originalTitle, budget) %>%  
 head(10)  
  
print(top\_budget\_movies)

## # A tibble: 10 × 2  
## originalTitle budget  
## <chr> <dbl>  
## 1 Pirates of the Caribbean: On Stranger Tides 380000000  
## 2 Avengers: Age of Ultron 365000000  
## 3 Avengers: Endgame 356000000  
## 4 Pirates of the Caribbean: At World's End 300000000  
## 5 Justice League 300000000  
## 6 Avengers: Infinity War 300000000  
## 7 Superman Returns 270000000  
## 8 Tangled 260000000  
## 9 The Lion King 260000000  
## 10 Spider-Man 3 258000000

4.2 ¿Cuáles son las 10 películas que más ingresos tuvieron?

top\_revenue\_movies <- movies %>%  
 arrange(desc(revenue)) %>%  
 select(originalTitle, revenue) %>%  
 head(10)  
  
print(top\_revenue\_movies)

## # A tibble: 10 × 2  
## originalTitle revenue  
## <chr> <dbl>  
## 1 Avatar 2847246203  
## 2 Avengers: Endgame 2797800564  
## 3 Titanic 2187463944  
## 4 Star Wars: The Force Awakens 2068223624  
## 5 Avengers: Infinity War 2046239637  
## 6 Jurassic World 1671713208  
## 7 The Lion King 1667635327  
## 8 Spider-Man: No Way Home 1631853496  
## 9 The Avengers 1518815515  
## 10 Furious 7 1515047671

4.3 ¿Cuál es la película que más votos tuvo?

most\_voted\_movie <- movies %>%  
 arrange(desc(voteCount)) %>%  
 select(originalTitle, voteCount) %>%  
 head(1)  
  
print(most\_voted\_movie)

## # A tibble: 1 × 2  
## originalTitle voteCount  
## <chr> <dbl>  
## 1 Inception 30788

4.4 ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios?

worst\_movie <- movies %>%  
 arrange(voteAvg) %>%  
 select(originalTitle, voteAvg) %>%  
 head(1)  
  
print(worst\_movie)

## # A tibble: 1 × 2  
## originalTitle voteAvg  
## <chr> <dbl>  
## 1 <U+5287><U+5834><U+7248> <U+62B1><U+304B><U+308C><U+305F><U+3044><U+7… 1.3

4.5 ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras

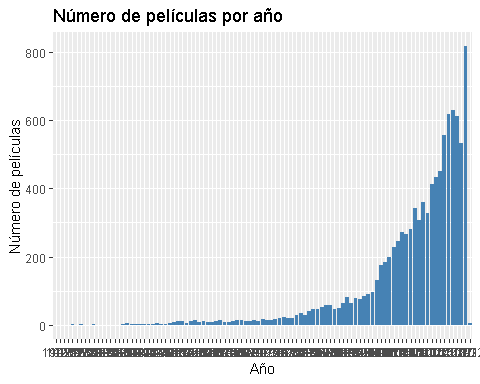
movies\_per\_year <- movies %>%  
 group\_by(year = format(as.Date(releaseDate), "%Y")) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count))  
  
movies\_per\_year\_with\_count <- movies %>%  
 group\_by(year = format(as.Date(releaseDate), "%Y")) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count)) %>%  
 select(year, count)  
 head(5)

## [1] 5

print(movies\_per\_year\_with\_count)

## # A tibble: 99 × 2  
## year count  
## <chr> <int>  
## 1 2021 816  
## 2 2018 629  
## 3 2017 618  
## 4 2019 612  
## 5 2016 557  
## 6 2020 533  
## 7 2015 450  
## 8 2014 432  
## 9 2013 412  
## 10 2011 361  
## # ℹ 89 more rows

ggplot(movies\_per\_year, aes(x=year, y=count)) +  
 geom\_bar(stat="identity", fill="steelblue") +  
 ggtitle("Número de películas por año") +  
 xlab("Año") +  
 ylab("Número de películas")



4.6 ¿Cuál es el género principal de las 20 películas más recientes? ¿Cuál es el género principal que predomina en el conjunto de datos? Represéntelo usando un gráfico. ¿A qué género principal pertenecen las películas más largas?

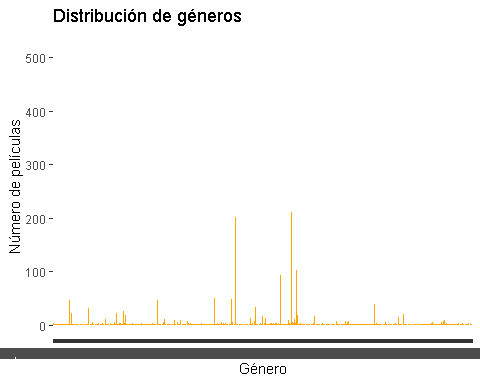
recent\_movies\_genres <- movies %>%  
 arrange(desc(releaseDate)) %>%  
 head(20) %>%  
 group\_by(genres) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count))  
  
print(recent\_movies\_genres)

## # A tibble: 17 × 2  
## genres count  
## <chr> <int>  
## 1 Comedy 2  
## 2 Drama 2  
## 3 <NA> 2  
## 4 Action|Adventure|Comedy 1  
## 5 Action|Drama 1  
## 6 Action|Thriller|Crime|Drama 1  
## 7 Animation|Adventure|Fantasy|Family 1  
## 8 Comedy|Drama 1  
## 9 Comedy|Horror 1  
## 10 Comedy|Romance 1  
## 11 Comedy|Romance|Fantasy 1  
## 12 Documentary 1  
## 13 Drama|Fantasy 1  
## 14 Drama|Mystery 1  
## 15 Horror|Thriller|Mystery 1  
## 16 Romance 1  
## 17 Thriller 1

predominant\_genre <- movies %>%  
 group\_by(genres) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count)) %>%  
 head(1)  
  
print(predominant\_genre)

## # A tibble: 1 × 2  
## genres count  
## <chr> <int>  
## 1 Drama 521

ggplot(movies, aes(x = genres)) +  
 geom\_bar(fill = "orange") +  
 ggtitle("Distribución de géneros") +  
 xlab("Género") +  
 ylab("Número de películas")



4.7 ¿Las películas de qué genero principal obtuvieron mayores ganancias?

genre\_revenue <- movies %>%  
 group\_by(genres) %>%  
 summarise(total\_revenue = sum(revenue, na.rm = TRUE)) %>%  
 arrange(desc(total\_revenue))  
  
print(genre\_revenue)

## # A tibble: 2,346 × 2  
## genres total\_revenue  
## <chr> <dbl>  
## 1 Action|Adventure|Science Fiction 19780390887  
## 2 Comedy 17074165367  
## 3 Comedy|Romance 13775242729  
## 4 Drama 12771394567  
## 5 Action|Adventure|Fantasy 11540955567  
## 6 Adventure|Action|Science Fiction 10014836400  
## 7 Drama|Romance 9703898926  
## 8 Adventure|Fantasy|Action 8191167898  
## 9 Science Fiction|Action|Adventure 8187440971  
## 10 Action|Adventure|Fantasy|Science Fiction 7931134107  
## # ℹ 2,336 more rows

4.8 ¿La cantidad de actores influye en los ingresos de las películas?¿se han hecho películas con más actores en los últimos años?

correlation <- cor(movies$actorsAmount, movies$revenue, use="complete.obs")  
print(correlation)

## [1] -0.01955488

recent\_movies\_actors <- movies %>%  
 arrange(desc(releaseDate)) %>%  
 head(100) %>%  
 summarise(mean\_actors=mean(actorsAmount, na.rm=TRUE))  
  
print(recent\_movies\_actors)

## # A tibble: 1 × 1  
## mean\_actors  
## <dbl>  
## 1 17.5

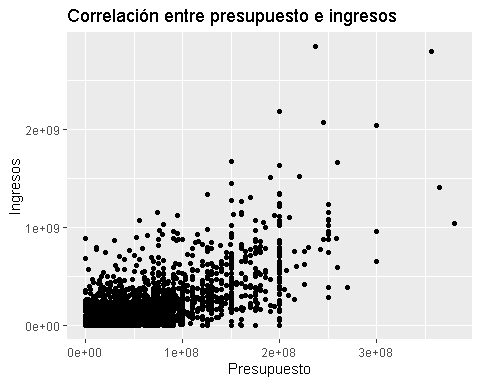
4.10 ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

top\_rated\_directors <- movies %>%  
 arrange(desc(voteAvg)) %>%  
 head(20) %>%  
 select(director, voteAvg)  
  
print(top\_rated\_directors)

## # A tibble: 20 × 2  
## director voteAvg  
## <chr> <dbl>  
## 1 "Thomas Coven" 10   
## 2 "V\xedctor Barba|Juan Olivares" 10   
## 3 "Rebecca Sugar" 10   
## 4 "Laurent Bouzereau" 10   
## 5 "Kaku Arakawa" 10   
## 6 "Christin Baker" 10   
## 7 <NA> 10   
## 8 "Miguel Angel Zavala" 10   
## 9 <NA> 9.8  
## 10 "Dave Bullock|Troy Adomitis|Victor Cook" 9.6  
## 11 "Samuel Leong" 9.5  
## 12 <NA> 9.5  
## 13 <NA> 9.5  
## 14 "Won Myeong-jun" 9.5  
## 15 "Selena Quintanilla" 9.4  
## 16 "Haruo Sotozaki" 9.3  
## 17 "Haruo Sotozaki" 9.3  
## 18 <NA> 9.2  
## 19 "Ulises Valencia" 9.2  
## 20 "Kim Nam-joon|Jeon Jung-kook|Kim Tae-hyung|Park Ji-min|Jung Ho-seok|… 9.2

4.11 ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión

ggplot(movies, aes(x = budget, y = revenue)) +  
 geom\_point() +  
 ggtitle("Correlación entre presupuesto e ingresos") +  
 xlab("Presupuesto") +  
 ylab("Ingresos")



correlation\_budget\_revenue <- cor(movies$budget, movies$revenue, use = "complete.obs")  
print(correlation\_budget\_revenue)

## [1] 0.757454

4.12 ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?

movies$releaseMonth <- format(as.Date(movies$releaseDate), "%m")  
  
monthly\_revenue <- movies %>%  
 group\_by(releaseMonth) %>%  
 summarise(mean\_revenue = mean(revenue, na.rm = TRUE)) %>%  
 arrange(desc(mean\_revenue))  
  
print(monthly\_revenue)

## # A tibble: 12 × 2  
## releaseMonth mean\_revenue  
## <chr> <dbl>  
## 1 06 94747108.  
## 2 05 87845442.  
## 3 07 76028696.  
## 4 12 74358880.  
## 5 11 71492112.  
## 6 04 52595654.  
## 7 03 51115942.  
## 8 02 42908353.  
## 9 10 38987332.  
## 10 08 35970079.  
## 11 01 33773691.  
## 12 09 31928917.

4.13 ¿En qué meses se han visto los lanzamientos con mejores ingresos? ¿cuantas películas, en promedio, se han lanzado por mes?

movies\_per\_month <- movies %>%  
 group\_by(releaseMonth) %>%  
 summarise(count = n()) %>%  
 summarise(mean\_movies = mean(count))  
  
print(movies\_per\_month)

## # A tibble: 1 × 1  
## mean\_movies  
## <dbl>  
## 1 833.

4.14 ¿Cómo se correlacionan las calificaciones con el éxito comercial?

correlation\_voteAvg\_revenue <- cor(movies$voteAvg, movies$revenue, use = "complete.obs")  
print(correlation\_voteAvg\_revenue)

## [1] 0.1412618

4.15 ¿Qué estrategias de marketing, como videos promocionales o páginas oficiales, generan mejores resultados?

movies %>%  
 group\_by(video) %>%  
 summarise(mean\_revenue = mean(revenue, na.rm = TRUE))

## # A tibble: 3 × 2  
## video mean\_revenue  
## <lgl> <dbl>  
## 1 FALSE 50343588.  
## 2 TRUE 506309.  
## 3 NA 190528171.

movies %>%  
 group\_by(homePage) %>%  
 summarise(mean\_revenue = mean(revenue, na.rm = TRUE))

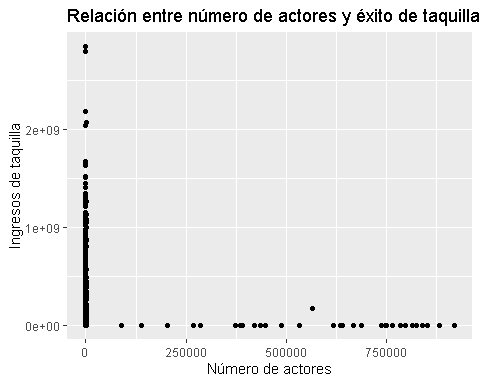
## # A tibble: 4,126 × 2  
## homePage mean\_revenue  
## <chr> <dbl>  
## 1 "Canal+ Espa\xf1a" 0  
## 2 "Ignition Creative" 0  
## 3 "Loose Screw Productions" 0  
## 4 "Purgatory Blues LLC" 5000  
## 5 "SET Film Workshop" 0  
## 6 "Six Ten Productions" 0  
## 7 "http://009.ph9.jp/" 0  
## 8 "http://1111themovie.com/" 0  
## 9 "http://12mightyorphans.com" 3615193  
## 10 "http://13films.net/projects/extortion/" 0  
## # ℹ 4,116 more rows

4.16 ¿La popularidad del elenco está directamente correlacionada con el éxito de taquilla?

# Calcular la correlación entre el número de actores y los ingresos  
correlation\_cast\_revenue <- cor(movies$actorsAmount, movies$revenue, use = "complete.obs")  
  
# Imprimir el resultado de la correlación  
print(correlation\_cast\_revenue)

## [1] -0.01955488

ggplot(movies, aes(x = actorsAmount, y = revenue)) +  
 geom\_point() +  
 ggtitle("Relación entre número de actores y éxito de taquilla") +  
 xlab("Número de actores") +  
 ylab("Ingresos de taquilla")

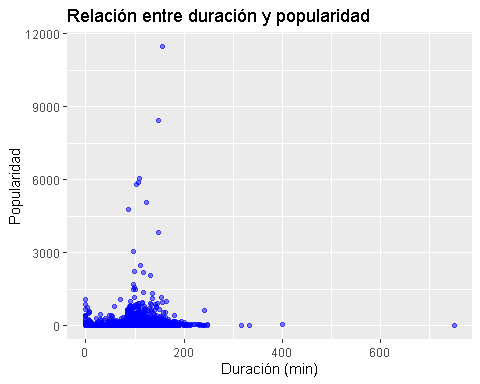


5.1 ¿Cuál es la relación entre la duración de las películas y su popularidad?

correlation\_runtime\_popularity <- cor(movies$runtime, movies$popularity, use = "complete.obs")  
print(correlation\_runtime\_popularity)

## [1] 0.03253848

ggplot(movies, aes(x = runtime, y = popularity)) +  
 geom\_point(alpha = 0.5, color = "blue") +  
 ggtitle("Relación entre duración y popularidad") +  
 xlab("Duración (min)") +  
 ylab("Popularidad")

 5.2 ¿Qué director ha producido más películas?

most\_frequent\_directors <- movies %>%  
 group\_by(director) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count)) %>%  
 head(10)  
  
print(most\_frequent\_directors)

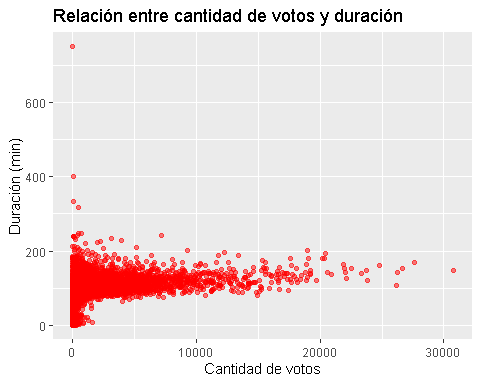
## # A tibble: 10 × 2  
## director count  
## <chr> <int>  
## 1 <NA> 74  
## 2 Steven Spielberg 29  
## 3 Clint Eastwood 28  
## 4 Ridley Scott 23  
## 5 Kunihiko Yuyama 19  
## 6 Martin Scorsese 19  
## 7 Steven Soderbergh 19  
## 8 Robert Zemeckis 18  
## 9 Ron Howard 18  
## 10 Woody Allen 17

5.3 ¿Las películas con más votos tienden a ser más largas?

correlation\_votes\_runtime <- cor(movies$voteCount, movies$runtime, use = "complete.obs")  
print(correlation\_votes\_runtime)

## [1] 0.282363

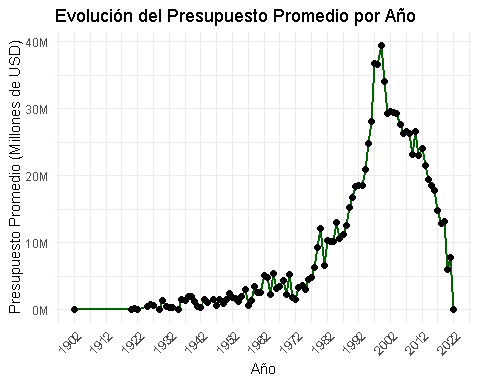
ggplot(movies, aes(x = voteCount, y = runtime)) +  
 geom\_point(alpha = 0.5, color = "red") +  
 ggtitle("Relación entre cantidad de votos y duración") +  
 xlab("Cantidad de votos") +  
 ylab("Duración (min)")



5.4 ¿Cuál ha sido la evolución del presupuesto promedio de las películas a lo largo de los años?

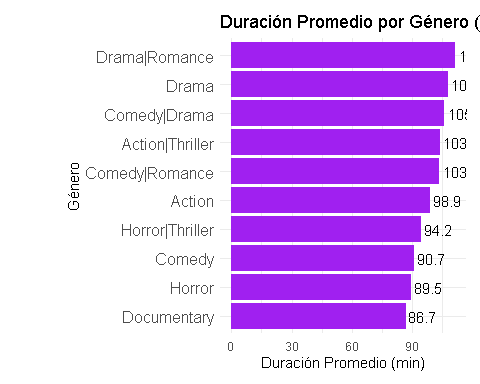
avg\_budget\_per\_year <- movies %>%  
 group\_by(year = as.numeric(format(as.Date(releaseDate), "%Y"))) %>%  
 summarise(avg\_budget = mean(budget, na.rm = TRUE))  
  
ggplot(avg\_budget\_per\_year, aes(x = year, y = avg\_budget)) +  
 geom\_line(color = "darkgreen", size = 1) +  
 geom\_point(color = "black", size = 2) +   
 scale\_x\_continuous(breaks = seq(min(avg\_budget\_per\_year$year), max(avg\_budget\_per\_year$year), by = 10)) + # Intervalos de 10 años  
 scale\_y\_continuous(labels = scales::label\_number(scale = 1e-6, suffix = "M")) + # Mostrar en millones de USD  
 ggtitle("Evolución del Presupuesto Promedio por Año") +  
 xlab("Año") +  
 ylab("Presupuesto Promedio (Millones de USD)") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotar etiquetas del eje X

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



5.5 ¿Existen diferencias en la duración de las películas según su género principal?

avg\_runtime\_by\_genre <- movies %>%  
 mutate(main\_genre = str\_extract(genres, "^[^,]+")) %>% # Extraer el primer género  
 group\_by(main\_genre) %>%  
 summarise(avg\_runtime = mean(runtime, na.rm = TRUE), count = n()) %>%  
 arrange(desc(count)) %>% # Ordenar por cantidad de películas  
 slice\_head(n = 10) %>% # Seleccionar los 10 géneros más populares  
 arrange(desc(avg\_runtime)) # Ordenar por duración promedio  
  
ggplot(avg\_runtime\_by\_genre, aes(x = reorder(main\_genre, avg\_runtime), y = avg\_runtime)) +  
 geom\_bar(stat = "identity", fill = "purple") +  
 geom\_text(aes(label = round(avg\_runtime, 1)), hjust = -0.1, size = 4) +  
 coord\_flip() +  
 ggtitle("Duración Promedio por Género (Top 10 más populares)") +  
 xlab("Género") +  
 ylab("Duración Promedio (min)") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(size = 12),  
 plot.margin = margin(10, 10, 10, 50))



5.6 ¿Cómo ha cambiado la popularidad de las películas en el tiempo?

popularity\_over\_years <- movies %>%  
 group\_by(year = as.numeric(format(as.Date(releaseDate), "%Y"))) %>%  
 summarise(avg\_popularity = mean(popularity, na.rm = TRUE))  
  
ggplot(popularity\_over\_years, aes(x = year, y = avg\_popularity)) +  
 geom\_line(color = "orange", size = 1) +  
 geom\_point(color = "black", size = 2) +  
 scale\_x\_continuous(  
 breaks = seq(min(popularity\_over\_years$year, na.rm = TRUE),   
 max(popularity\_over\_years$year, na.rm = TRUE),   
 by = 5) # Intervalos de 5 años  
 ) +  
 ggtitle("Evolución de la Popularidad de las Películas") +  
 xlab("Año") +  
 ylab("Popularidad Promedio") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotar etiquetas del eje X

