Proyecto 1. Análisis Exploratorio

Andre Jo

2025-01-31

Leer CSV Movies y realizar importaciones de librerias externas

```
library(ggplot2)
datos <- read.csv("movies.csv")
```

4.7. (8 puntos) ¿Las películas de qué genero principal obtuvieron mayores ganancias?

```
datos_br <- datos[, c("genres", "budget", "revenue")]
profit <- cbind(datos_br, profit=datos$revenue-datos$budget)

#Quitar los ceros y generos en espacios blancos
profit <- profit[apply(profit!=0, 1, all), ]
profit <- profit[apply(profit!='', 1, all), ]
#Sacar el genero principal de la pelicula
genre_main = c()
for (genre in profit$genres) {
  main_genre <- strsplit(x=genre,split="|", fixed = TRUE)[[1]][1]
  genre_main <- append(genre_main, main_genre)
}
da <- cbind(genre_main, profit)
da <- aggregate(da$profit, list(da$genre_main), FUN=mean)
colnames(da) <- c("Generos", "Ganancia")
pregunta4.7 <- da[order(da$Ganancia, decreasing = TRUE), ]

head(pregunta4.7)
```

```
##            Generos  Ganancia
## 2        Adventure 173915372
## 3        Animation 173682706
## 8           Family 137238663
## 15 Science Fiction 120688751
## 1           Action 119572299
## 9          Fantasy 113629994
```

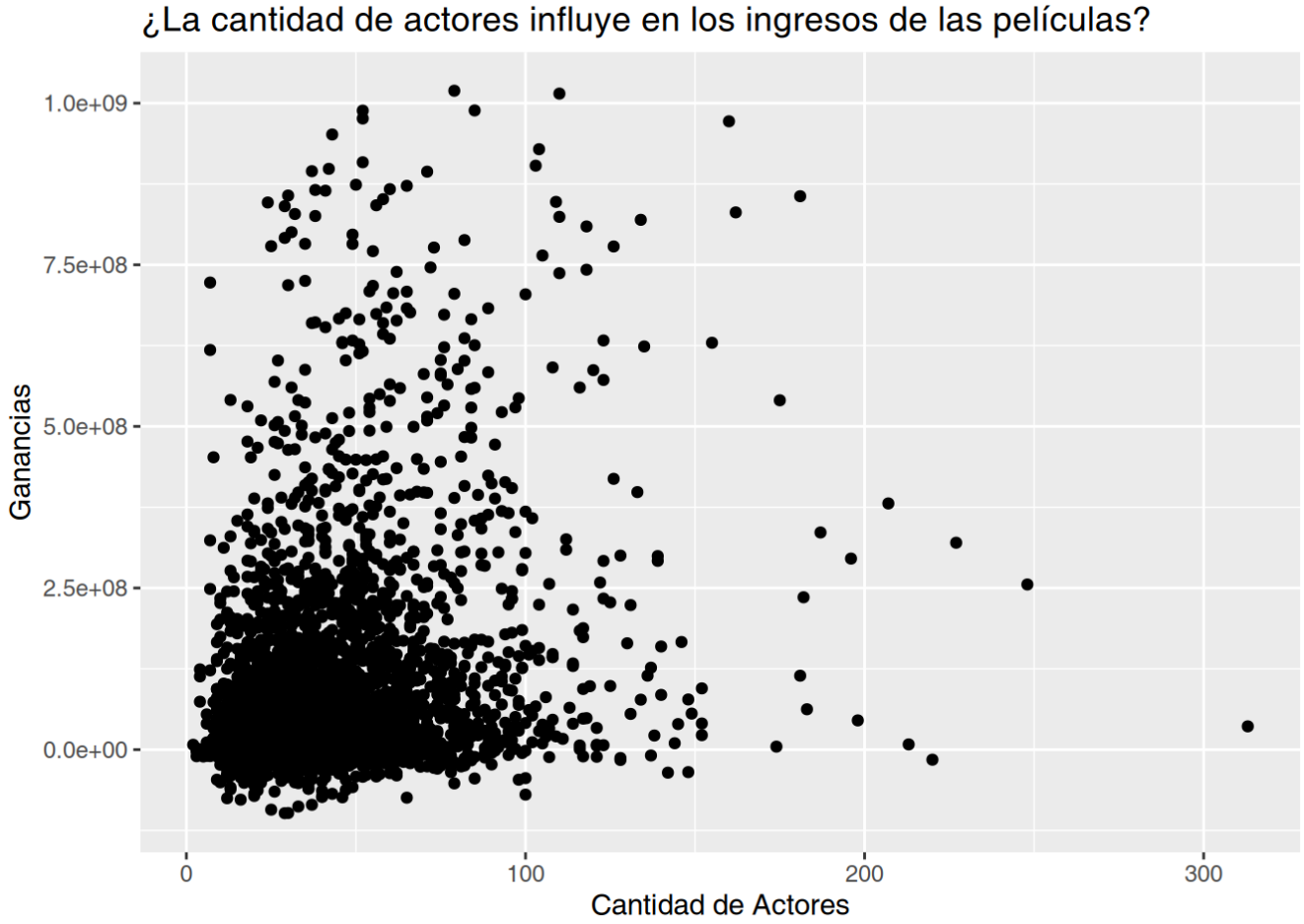
Se afirma que los generos que más ganancias obtuvieron fueron los siguientes: Adventure, Animation, Family, Science Fiction, Action y Fantasy con una ganancia mayor a 100,000,000

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

últimos años?

```
datos_br <- datos[, c( "title", "budget", "revenue", "actorsAmount")]
datos_br <- datos_br[apply(datos_br!=0, 1,all),]
datos_br <- datos_br[order(datos_br$actorsAmount, decreasing = TRUE), ]
profit <- cbind(datos_br, profit=datos_br$revenue-datos_br$budget)
a_graficar <- profit[ , c("profit", "actorsAmount") ]
a_graficar <- a_graficar[apply(a_graficar!=565916, 1,all),]
a_graficar <- a_graficar[apply(a_graficar!=0, 1,all),]
# Plot
ggplot(a_graficar, aes(x=actorsAmount, y=profit)) + geom_point() +
labs(title = "¿La cantidad de actores influye en los ingresos de las películas?",
x = "Cantidad de Actores", y = "Ganancias") +
ylim(-102566092, 1022566902)
```

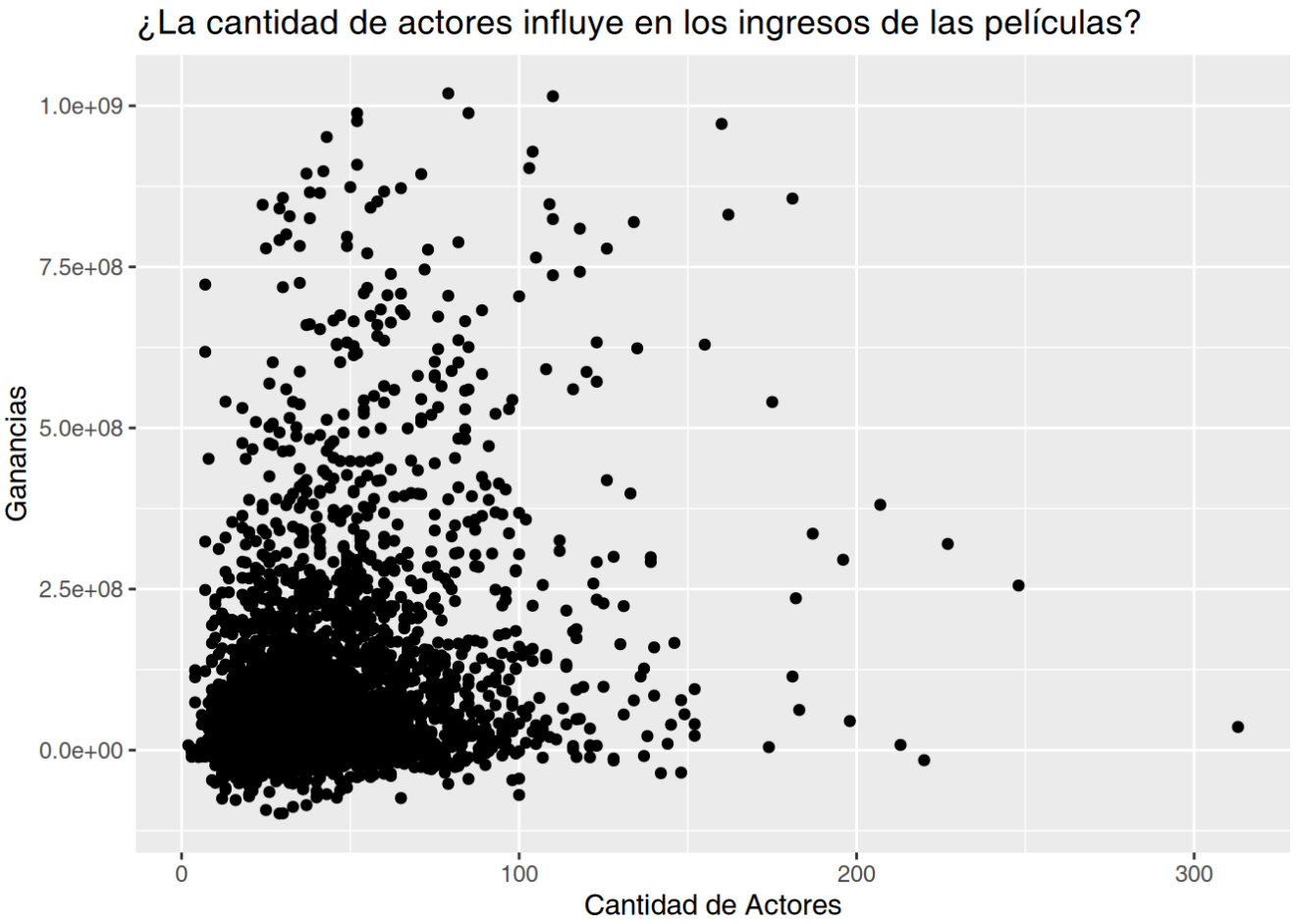
```
## Warning: Removed 26 rows containing missing values or values outside the scale
range
## (`geom_point()`).
```



No, como se observa en la gráfica hay solo hay una concentración en el rango de 100 actores donde se observa que las ganancias pueden alcanzar a 1022566902 mientras que cantidades en el rango de 200 y 300 no se esperan tantas ganancias.

```
datos_br <- datos[, c( "title", "budget", "revenue", "actorsAmount")]
datos_br <- datos_br[apply(datos_br!=0, 1,all),]
datos_br <- datos_br[order(datos_br$actorsAmount, decreasing = TRUE), ]
profit <- cbind(datos_br, profit=datos_br$revenue-datos_br$budget)
a_graficar <- profit[ , c("profit", "actorsAmount") ]
a_graficar <- a_graficar[apply(a_graficar!=565916, 1,all),]
a_graficar <- a_graficar[apply(a_graficar!=0, 1,all),]
# Plot
ggplot(a_graficar, aes(x=actorsAmount, y=profit)) + geom_point() +
labs(title = "¿La cantidad de actores influye en los ingresos de las películas?",
x = "Cantidad de Actores", y = "Ganancias") +
ylim(-102566092, 1022566902)
```

```
## Warning: Removed 26 rows containing missing values or values outside the scale
range
## (`geom_point()`).
```




```
datos_br <- datos[, c( "title", "releaseDate", "actorsAmount")]
datos_br <- datos_br[order(datos_br$releaseDate, decreasing = TRUE), ]

anios <- format(as.Date(datos_br$releaseDate), "%Y")

pregunta4.8 <- cbind( datos_br, anios)
pregunta4.8 <-aggregate(pregunta4.8$actorsAmount, by = list(Anios = pregunta4.8$an
ios), FUN=sum)
pregunta4.8 <- pregunta4.8[order(pregunta4.8$x, decreasing = TRUE), ]
```

Si, se observa que en los recientes años si participaron una mayor cantidad de actores a comparación con los años no recientes se observa menor cantidad de actores que han participado

##4.9. (3 puntos) ¿Es posible que la cantidad de hombres y mujeres en el reparto influya en la popularidad y los ingresos de las películas?

```
## 4.9. (3 puntos) ¿Es posible que la cantidad de hombres y mujeres en el reparto
influya en la popularidad y los ingresos de las películas?

pregunta4.9 <- datos[ , c( "castMenAmount", "castWomenAmount", "revenue", "popular
ity" ) ]

pregunta4.9 <- na.exclude(pregunta4.9)

#Por alguna razon alguien escribio strings en datos numericos esto solo los convie
rte las columnas a numericos
pregunta4.9$castMenAmount <- as.numeric(pregunta4.9$castMenAmount)
```

```
## Warning: NAs introduced by coercion
```

```
pregunta4.9$castWomenAmount <- as.numeric(pregunta4.9$castWomenAmount)
```

```
## Warning: NAs introduced by coercion
```

```
pregunta4.9 <- pregunta4.9[order(pregunta4.9$revenue, decreasing = TRUE), ]
pregunta4.9 <- pregunta4.9[apply(pregunta4.9[, c("castMenAmount", "castWomenAmoun
t", "revenue")]!=0, 1, all), ]

head(pregunta4.9)
```

```
##      castMenAmount castWomenAmount    revenue popularity
## 3211            21               9 2847246203    558.792
## 5953            62              28 2797800564    422.152
## 308             59              27 2187463944    186.300
## 4948            74              24 2068223624    105.189
## 5954            43              21 2046239637    648.960
## 4915            31              13 1671713208    152.090
```

No afecta la cantidad de hombres y mujeres, ya que se observa que no existe una relación significativa entre la popularidad y los ingresos, dado que no hay una diferencia notable.

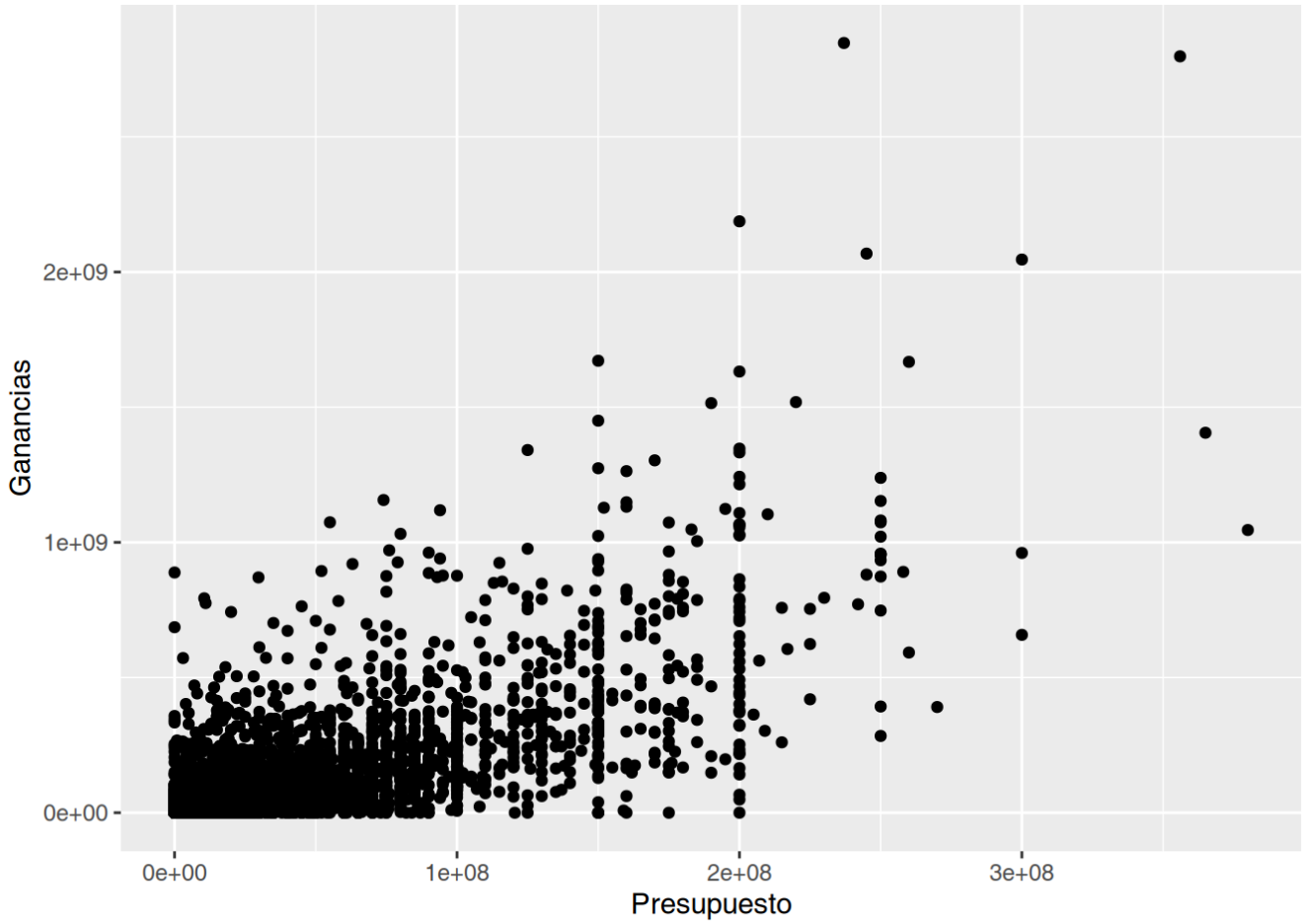
#4.10. (8 puntos) ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas? De acuerdo con la tabla propuesta se muestran los mejores directores de peliculas:

```
pregunta4.10 <- datos[ 1:20 , c( "director", "voteAvg") ]
pregunta4.10 <- pregunta4.10[order(pregunta4.10$voteAvg, decreasing = TRUE), ]
head(pregunta4.10 )
```

```
##                director voteAvg
## 5       Robert Zemeckis     8.5
## 15 Francis Ford Coppola     8.3
## 3          George Lucas     8.2
## 10           Fritz Lang     8.2
## 18        Michel Gondry     8.1
## 6            Sam Mendes     8.0
```

4.11. (8 puntos) ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión

```
pregunta4.11 <- datos[ , c( "revenue", "budget") ]
pregunta4.11 <- pregunta4.11[apply(pregunta4.10!=0, 1, all), ]
ggplot(pregunta4.11, aes(y=revenue, x=budget)) + geom_point() +
  labs(x = "Presupuesto", y = "Ganancias")
```



Se ve a inicios de la grafica hay una concentración de datos pero conforme se ve que el presupuesto aumenta se observa que las ganancias suben como se muestran en el rango de datos de 20000000

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

```
pregunta4.12 <- datos[ , c( "releaseDate", "revenue") ]
#Quitar los revenues que tienen 0
pregunta4.12 <- pregunta4.12[apply(pregunta4.12!=0, 1, all), ]
meses <- format(as.Date(pregunta4.12$releaseDate), "%m")

pregunta4.12 <- cbind( pregunta4.12, meses)

pregunta4.12 <- aggregate(pregunta4.12$revenue, list(meses), FUN=mean)
colnames(pregunta4.12) <- c("Fecha Lanzamiento", "Ingresos")
head(pregunta4.12)
```

```
##   Fecha Lanzamiento  Ingresos
## 1                01  69465131
## 2                02  82995335
## 3                03  97108375
## 4                04 101968175
## 5                05 165272557
## 6                06 165807439
```

Se observa el mes de abril,Mayo,Junio y Julio tienen mejores ingresos a comparación de otros meses.

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

```
pregunta4.13 <- pregunta4.12[order(pregunta4.12$Ingresos, decreasing = TRUE), ]
pregunta4.13[ 1:4 ,]
```

```
##    Fecha Lanzamiento  Ingresos
## 6                 06 165807439
## 5                 05 165272557
## 11                11 140375024
## 12                12 135527394
```

```
pregunta4.11 <- datos[ , c( "releaseDate", "revenue") ]
pregunta4.11 <- pregunta4.11[apply(pregunta4.11!=0, 1, all), ]
meses <- format(as.Date(pregunta4.11$releaseDate), "%m")
mese <- aggregate(pregunta4.11$revenue, list(meses), FUN=length)

mean(mese$x)
```

```
## [1] 435.25
```

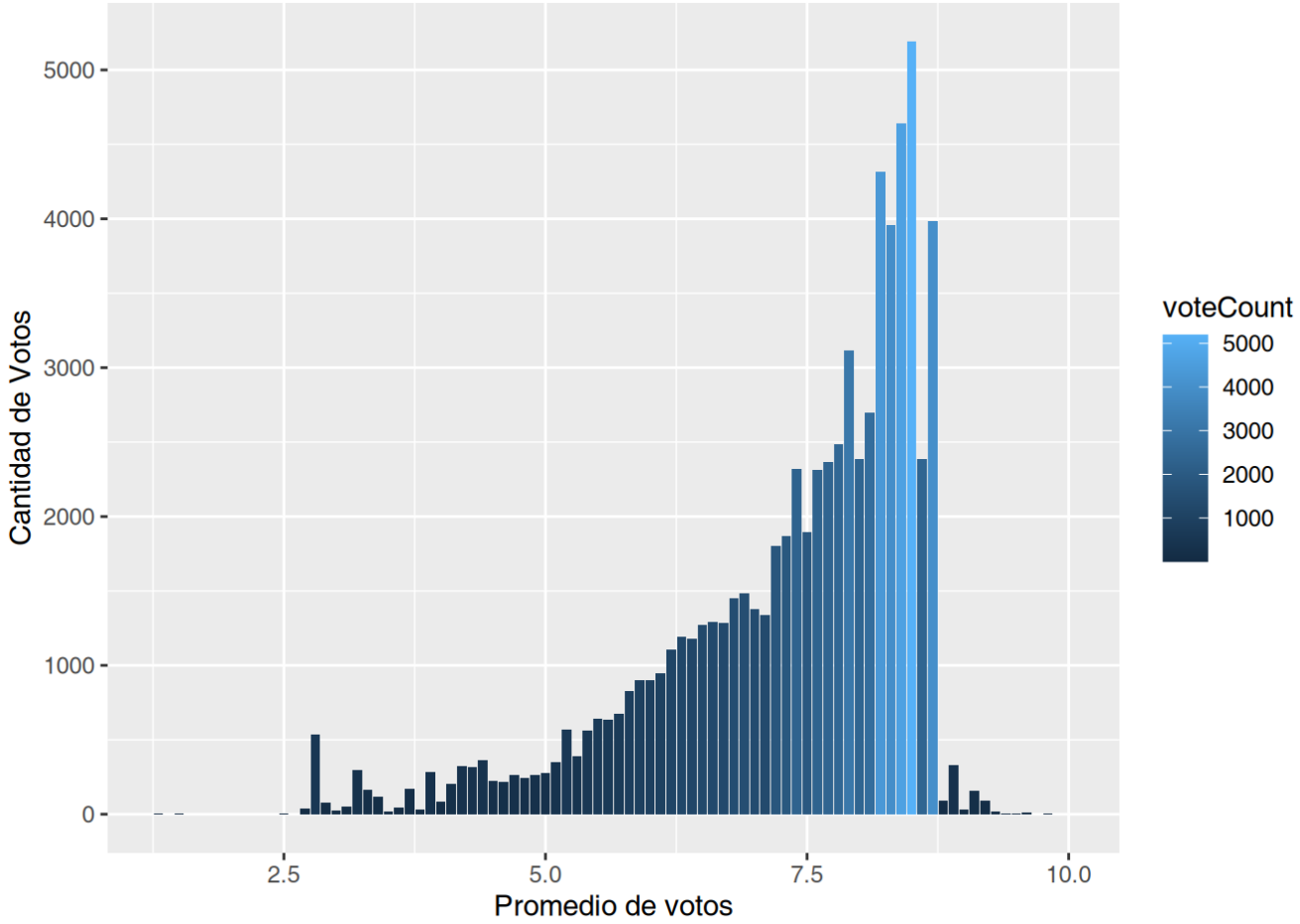
Se observa el mes de Abril, Mayo, Junio y Julio tienen mejores ingresos a comparación de otros meses. Además de eso su promedio por mes es de: 435.25.

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

```
pregunta4.14 <- datos[ , c( "voteAvg", "voteCount") ]

pregunta4.14 <- aggregate(pregunta4.14$voteCount, list(pregunta4.14$voteAvg), FUN=
mean)
colnames(pregunta4.14) <- c("voteAvg", "voteCount")

ggplot(pregunta4.14, aes(y=voteCount, x=voteAvg, fill=voteCount)) + geom_bar(stat
= "identity") +
  labs(y = "Cantidad de Votos", x = "Promedio de votos")
```



La relación entre las calificaciones y el éxito comercial en películas se infiere que calificaciones moderadamente altas (alrededor de 8) tienden a tener más éxito en términos de audiencia, con más de 5,000 personas otorgando esta puntuación. Esto sugiere que no siempre las películas mejor calificadas son las más exitosas comercialmente.

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

```
pregunta4.15 <- datos[ , c( "homePage", "video", "voteCount") ]
pregunta4.15$homePage <- !is.na(pregunta4.15$homePage) & pregunta4.15$homePage !=
""
pregunta4.15 <- pregunta4.15[order(-pregunta4.15$video), ]
pregunta4.15 <- aggregate(voteCount ~ video + homePage, data = pregunta4.15, FUN =
sum)
head(pregunta4.15)
```

```
##   video homePage voteCount
## 1 FALSE    FALSE   4488605
## 2  TRUE    FALSE      5259
## 3 FALSE     TRUE   7150421
## 4  TRUE     TRUE      1911
```

Como se puede observar con la siguiente tabla, se observa que al no tener un video o tener una pagina de promoción se esperan mejores resultados. Mientras que peliculas que si tienen páginas web y videos no generan mejores resultados

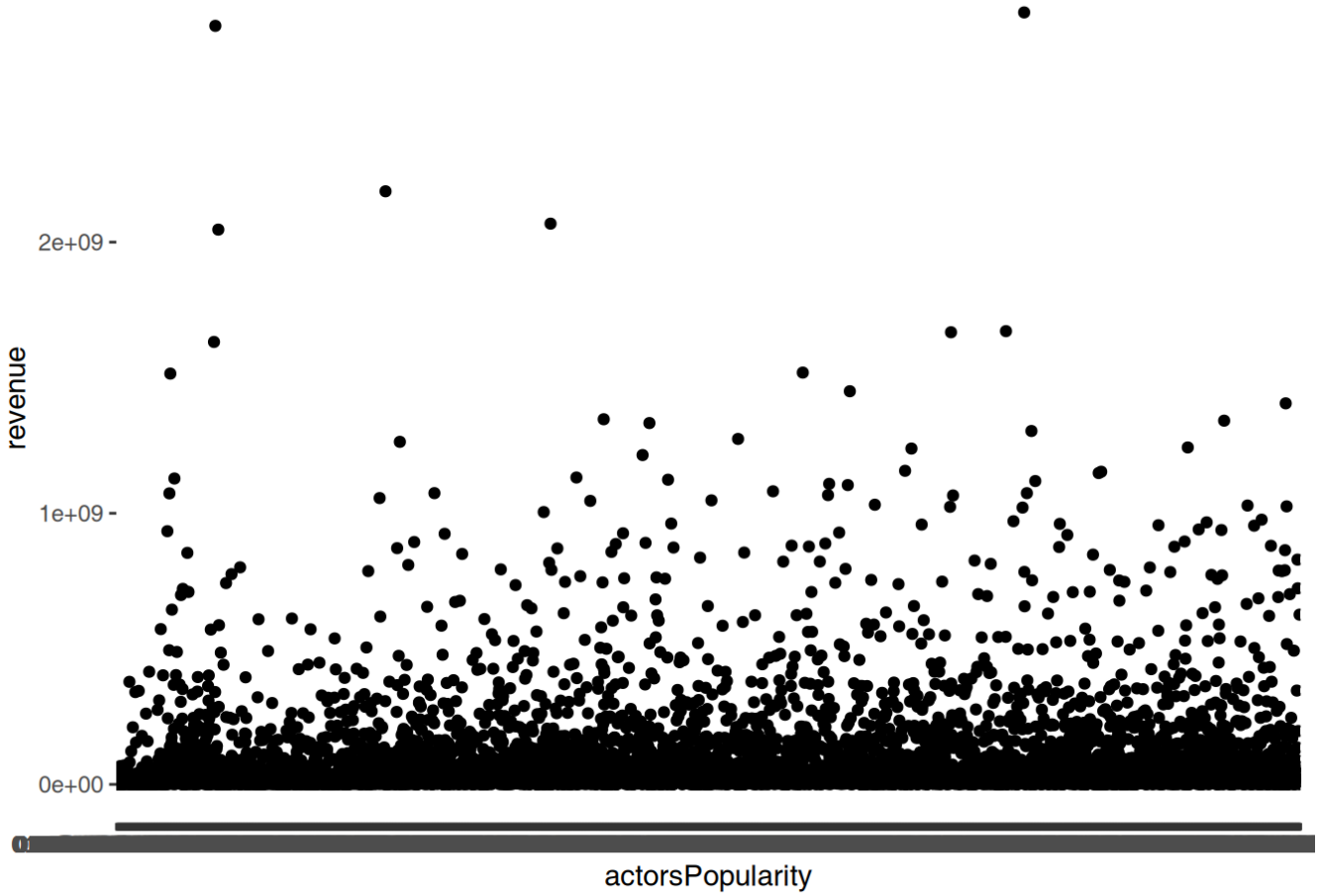
16: ¿La popularidad del elenco está directamente correlacionada con el éxito de taquilla?

```
pregunta4.16 <- datos[, c("actorsPopularity", "revenue")]
for (i in 1:nrow(pregunta4.16)) {
  for (j in 1:ncol(pregunta4.16)) {
    if (is.character(pregunta4.16[i, j])) {
      split_values <- strsplit(pregunta4.16[i, j], "\\|")[[1]]
      numeric_values <- as.numeric(split_values)
      mean_value <- mean(numeric_values, na.rm = TRUE)
      pregunta4.16[i, j] <- mean_value
    }
  }
}

pregunta4.16 <- pregunta4.16[order(pregunta4.16$revenue, decreasing = TRUE), ]
pregunta4.16 <- pregunta4.16[apply(pregunta4.16!="NaN", 1, all), ]
pregunta4.16 <- pregunta4.16[apply(pregunta4.16!=0, 1, all), ]
head(pregunta4.16)
```

```
##      actorsPopularity    revenue
## 3211  6.5532972972973 2847246203
## 5953 15.0504368932039 2797800564
## 308  3.27228448275862 2187463944
## 4948 4.10664361702128 2068223624
## 5954          17.2132 2046239637
## 4915        6.4104375 1671713208
```

```
ggplot(pregunta4.16, aes(x = actorsPopularity, y = revenue)) +
  geom_point() +
  labs( x = "actorsPopularity", y = "revenue")
```



Se observa en la gráfica que la popularidad del actor no afecta con el éxito comercial de la taquilla en la cual es generar mayor cantidad de ingresos.

##EXTRAS Genere usted otras seis preguntas que le parezcan interesantes porque le permitan realizar otras exploraciones y respóndalas. No puede repetir ninguna de las instrucciones anteriores.

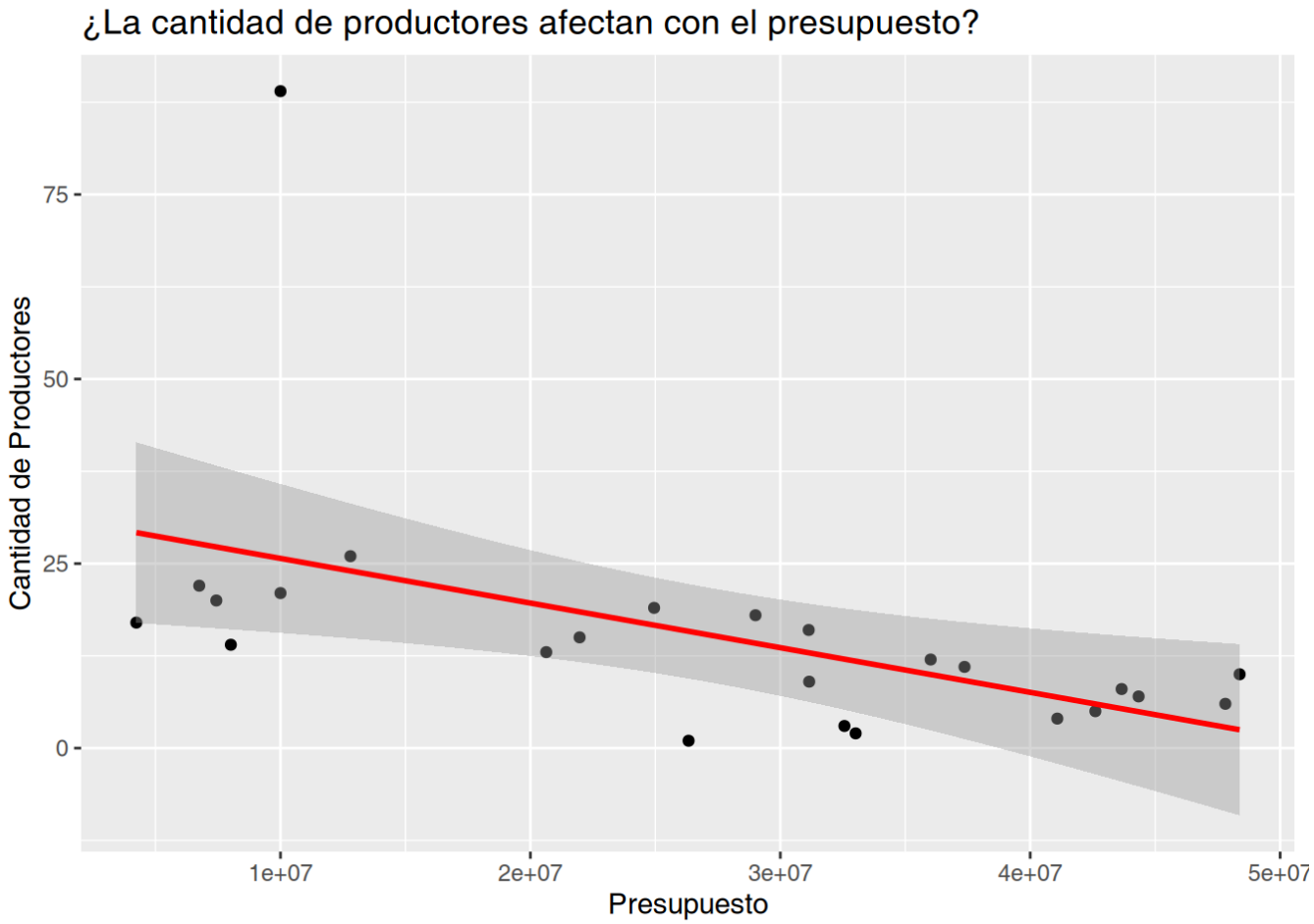
#4. ¿Existe una correlación entre el número de compañías productoras involucradas en una película y su presupuesto total?

```
company_x_budget  <- datos[, c("productionCoAmount" , "budget")]
company_x_budget <- company_x_budget[order(company_x_budget$productionCoAmount, de
creasing = TRUE), ]
company_x_budget <- company_x_budget[apply(company_x_budget!=0, 1, all), ]
company_x_budget <- aggregate(company_x_budget$budget, list(company_x_budget$produ
ctionCoAmount), FUN=mean)
colnames(company_x_budget) <- c("productionCoAmount", "budget")
company_x_budget
```

```
##    productionCoAmount   budget
## 1                   1 26329368
## 2                   2 33015173
## 3                   3 32565605
## 4                   4 41083853
## 5                   5 42605936
## 6                   6 47807805
## 7                   7 44336738
## 8                   8 43657015
## 9                   9 31155699
## 10                 10 48378894
## 11                 11 37371236
## 12                 12 36015753
## 13                 13 20635184
## 14                 14  8013982
## 15                 15 21969880
## 16                 16 31136576
## 17                 17  4233333
## 18                 18 29007434
## 19                 19 24950000
## 20                 20  7433333
## 21                 21 10000000
## 22                 22  6750000
## 23                 26 12800000
## 24                 89 10000000
```

```
ggplot(company_x_budget, aes(y=productionCoAmount, x=budget)) +   geom_point() +
  geom_smooth(method = "lm", col = "red") +
 labs(title = "¿La cantidad de productores afectan con el presupuesto?", x = "Pres
upuesto", y = "Cantidad de Productores")
```

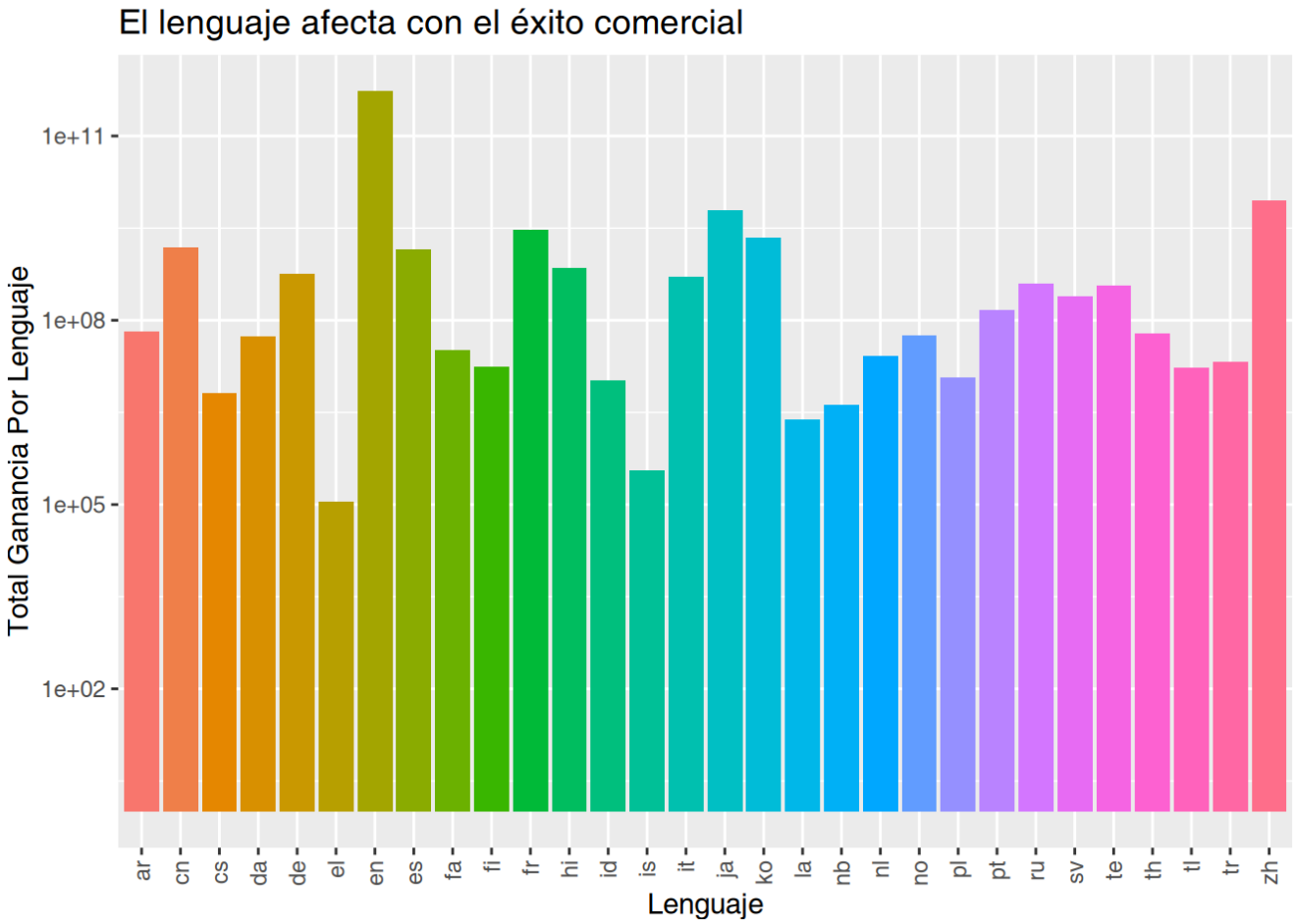
```
## `geom_smooth()` using formula = 'y ~ x'
```



Aunque se ve un punto atípico y con la cantidad de actores; se observa que entre mas compañias estan presentes, su presupuestos tiende a bajar como se muestra en esta linea de tendencia.

#5. El lenguaje afecta con el exito comercial ya que genera una mayor cantidad de ingresos?

```
lenguaje <- datos[, c("originalLanguage", "revenue")]
lenguaje <- aggregate(lenguaje$revenue, list(lenguaje$originalLanguage), FUN=sum)
colnames(lenguaje) <- c("Lenguaje", "Total_Ganancia_Por_Lenguaje")
lenguaje <- lenguaje[apply(lenguaje!=0, 1, all), ]
ggplot(lenguaje, aes(x=Lenguaje, y=Total_Ganancia_Por_Lenguaje, fill=Lenguaje)) +
  geom_bar(stat="identity") +
  labs(title = "El lenguaje afecta con el éxito comercial",
       x = "Lenguaje",
       y = "Total Ganancia Por Lenguaje") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  theme(legend.position = "none") +
  scale_y_log10()
```



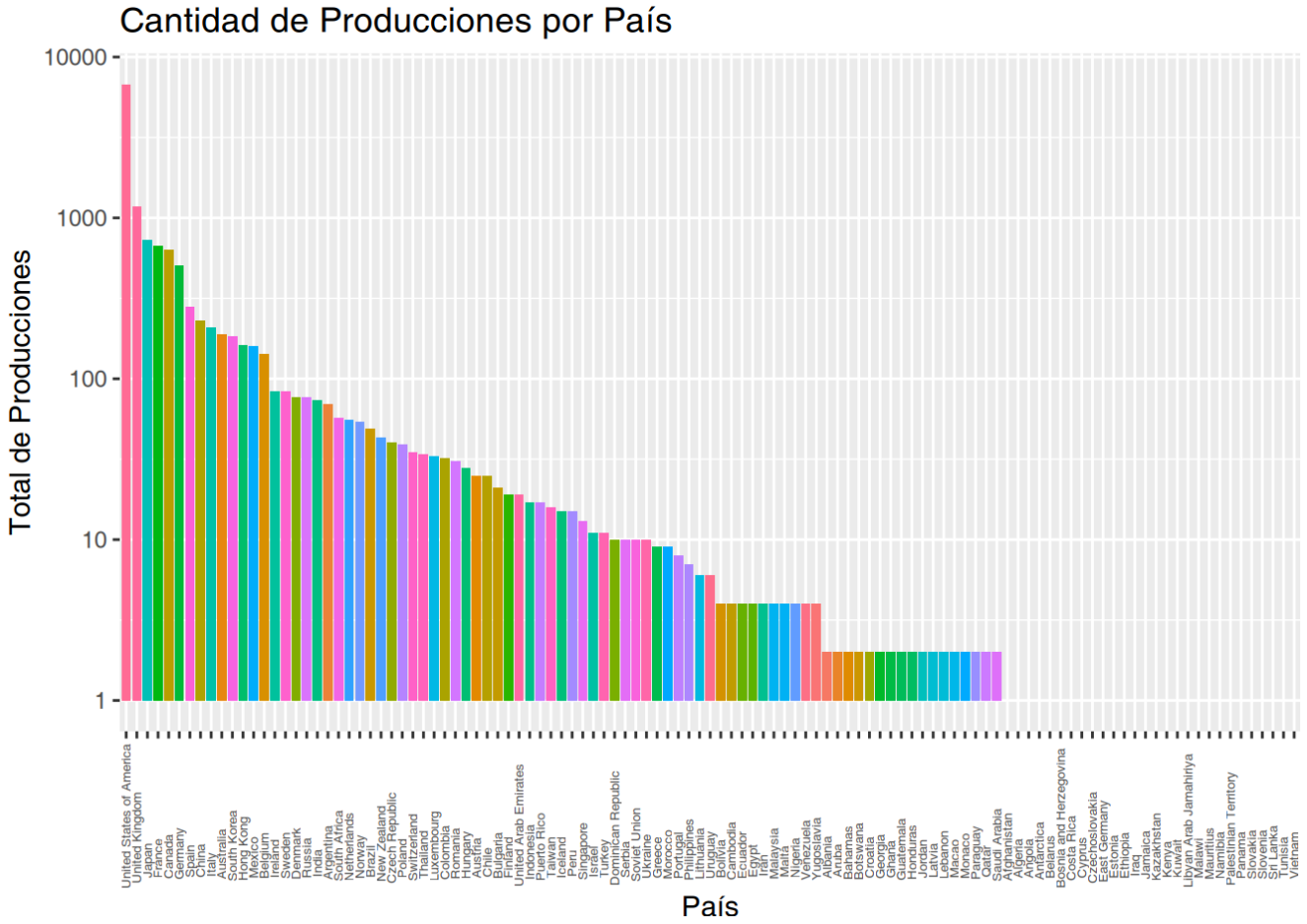
El idioma que ha generado mayores ingresos en las peliculas es el inglés, seguido por el chino en segundo lugar y el japonés en tercero. Esto se debe a que una gran parte de la población hablan el inglés o tiene una preferencia por ver y escuchar películas en este idioma, lo que contribuye a su éxito comercial.

##6.Que país es el más utilizado para producir una pelicula?

```
paises <- datos$productionCountry
paises_list <- strsplit(paises, split = "\\|", fixed = FALSE)
todos_paises <- unlist(paises_list)
todos_paises <- todos_paises[!is.na(todos_paises) & todos_paises != ""]
all_pais <- as.data.frame(table(todos_paises), stringsAsFactors = FALSE)
colnames(all_pais) <- c("Pais", "Total")
all_pais <- all_pais[apply(all_pais!=0, 1, all), ]
head(all_pais)
```

```
##          Pais Total
## 1 Afghanistan     1
## 2     Albania     2
## 3     Algeria     1
## 4      Angola     1
## 5  Antarctica     1
## 6   Argentina    70
```

```
ggplot(all_pais, aes(x = reorder(Pais, -Total), y = Total, fill = Pais)) +
  geom_bar(stat = "identity") +
  labs(title = "Cantidad de Producciones por País",
       x = "País",
       y = "Total de Producciones") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 0.1, size =
5)) +
  theme(legend.position = "none") +   scale_y_log10()
```



Se observa que los países con mayor producción cinematográfica son Estados Unidos, Reino Unido y Japón.