Universidad del Valle de Guatemala
Minería de Datos
S20

Hoja de Trabajo #1

Análisis Exploratorio

Javier Chavez 21016
Andres Quezada 21085

(3 puntos) Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

Es un archivo de 10,000 datos sobre películas, y estadísticas/descripciones acerca de la película.

Las variables son:
- Id: Id de la película
- popularity: Índice de popularidad de la película calculado semanalmente
- budget: El presupuesto para la película.
- revenue: El ingreso de la película.
- original_title: El título original de la película, en su idioma original.
- originalLanguage: Idioma original en que se encuentra la película
- title: El título de la película traducido al inglés
- homePage: La página de inicio de la película
- video: Si tiene videos promocionales o no
- director: Director de la película
- runtime: La duración de la película.
- genres: El género de la película.
- genresAmount: Cantidad de géneros que representan la película
- productionCompany: Las compañías productoras de la película.
- productionCoAmount: Cantidad de compañías productoras que participaron en la película
- productionCompanyCountry: Países de las compañías productoras de la película
- productionCountry: Países en los que se llevó a cabo la producción de la película
- productionCountriesAmount: Cantidad de países en los que se rodó la película
- releaseDate: Fecha de lanzamiento de la película
- voteCount: El número de votos en la plataforma para la película.
- voteAvg: El promedio de los votos en la plataforma para la película
- actors: Actores que participan en la película (Elenco)
- actorsPopularity: Índice de popularidad del elenco de la película.
- actorsCharacter: Personaje que interpreta cada actor en la película
- actorsAmount: Cantidad de personas que actúan en la película
- castWomenAmount: Cantidad de actrices en el elenco de la película
- castMenAmount: Cantidad de actores en el elenco de la película.

Describe:

	id	budget	revenue	runtime	popularity	voteAvg	voteCount	genresAmount	productionCoAmount	productionCountriesAmount	actorsAmou
count	10000.000000	1.000000e+04	1.000000e+04	10000.000000	9999.000000	9999.000000	9999.000000	9999.000000	9999.000000	9999.000000	9999.0000(
mean	249876.829300	1.855163e+07	5.673793e+07	100.268100	51.397715	6.483468	1342.481748	2.596260	3.171617	1.751075	2147.8500{
std	257380.109004	3.662669e+07	1.495854e+08	27.777829	216.740056	0.984321	2564.305389	1.154373	2.539772	3.012235	37201.9316(
min	5.000000	0.000000e+00	0.000000e+00	0.000000	4.258000	1.300000	1.000000	0.000000	0.000000	0.000000	0.0000(
25%	12286.500000	0.000000e+00	0.000000e+00	90.000000	14.578500	5.900000	120.000000	2.000000	2.000000	1.000000	13.0000(
50%	152558.000000	5.000000e+05	1.631245e+05	100.000000	21.909000	6.500000	415.000000	3.000000	3.000000	1.000000	21.0000(
75%	452021.750000	2.000000e+07	4.479661e+07	113.000000	40.656000	7.200000	1316.000000	3.000000	4.000000	2.000000	36.0000(
max	922260.000000	3.800000e+08	2.847246e+09	750.000000	11474.647000	10.000000	30788.000000	16.000000	89.000000	155.000000	919590.0000(

Info:

```
---  ------                       --------------  -----
 0   id                           10000 non-null  int64
 1   budget                       10000 non-null  int64
 2   genres                       9947 non-null   object
 3   homePage                     4193 non-null   object
 4   productionCompany            9543 non-null   object
 5   productionCompanyCountry     8720 non-null   object
 6   productionCountry            9767 non-null   object
 7   revenue                      10000 non-null  int64
 8   runtime                      10000 non-null  int64
 9   video                        9514 non-null   object
 10  director                     9926 non-null   object
 11  actors                       9920 non-null   object
 12  actorsPopularity             9913 non-null   object
 13  actorsCharacter              9953 non-null   object
 14  originalTitle                9999 non-null   object
 15  title                        9999 non-null   object
 16  originalLanguage             9999 non-null   object
 17  popularity                   9999 non-null   float64
 18  releaseDate                  9999 non-null   object
 19  voteAvg                      9999 non-null   float64
 20  voteCount                    9999 non-null   float64
 21  genresAmount                 9999 non-null   float64
 22  productionCoAmount           9999 non-null   float64
 23  productionCountriesAmount    9999 non-null   float64
 24  actorsAmount                 9999 non-null   float64
 25  castWomenAmount              9999 non-null   object
 26  castMenAmount                9999 non-null   object
dtypes: float64(7), int64(4), object(16)
```

Shape:

10,000 Filas, 27 Columnas

```
(10000, 27)
```

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

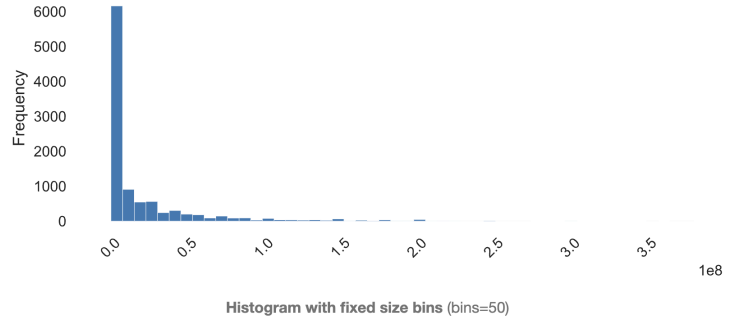
Variable	Tipo
Id	Cualitativa Ordinal
popularity	Cuantitativa Continua
budget	Cuantitativa Continua
revenue	Cuantitativa Continua
original_title	Cualitativa Nominal
originalLanguage	Cualitativa Nominal
title	Cualitativa Nominal
homePage	Cualitativa Nominal
video	Cualitativa Ordinal
director	Cualitativa Nominal
runtime	Cuantitativa Continua
genres	Cualitativa Nominal
genresAmount	Cuantitativa Discreta
productionCompany	Cualitativa Nominal
productionCoAmount	Cuantitativa Discreta
productionCompanyCountry	Cualitativa Nominal
productionCountry	Cualitativa Nominal

productionCountriesAmount	Cuantitativa Discreta
releaseDate	Cualitativa Nominal
voteCount	Cuantitativa Discreta
voteAvg	Cuantitativa Continua
actors	Cualitativa Nominal
actorsPopularity	Cuantitativa Continua
actorsCharacter	Cualitativa Nominal
actorsAmount	Cuantitativa Discreta
castWomenAmount	Cuantitativa Discreta
castMenAmount	Cuantitativa Discreta

(6 puntos) Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados.

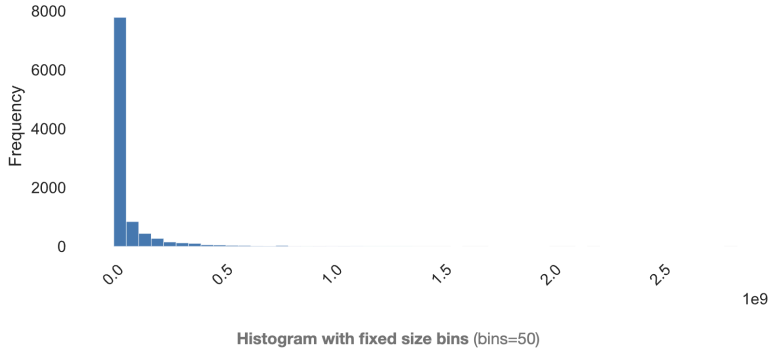
Variables cuantitativas:

budget



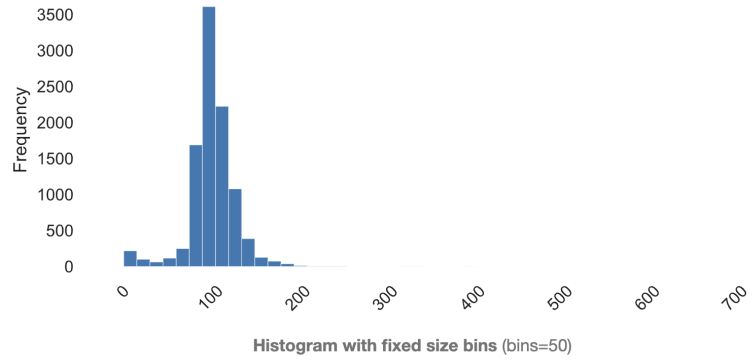
La distribución de budget está sesgada a la izquierda

revenue



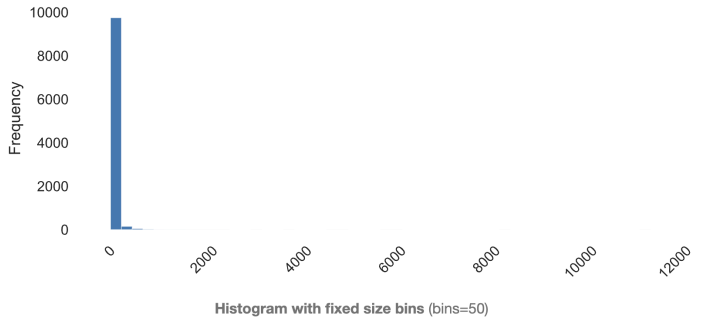
La distribución de revenue está sesgada a la izquierda

runtime



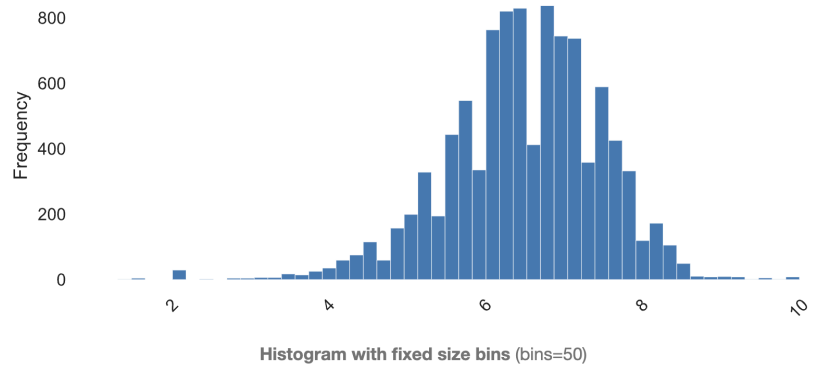
La distribución de runtime sigue un patrón normal

popularity



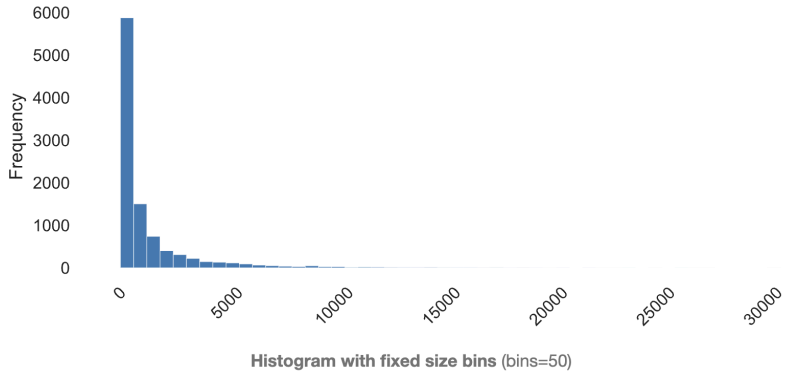
La distribución de popularity indica que la mayoría de datos tiene 0.

voteAvg



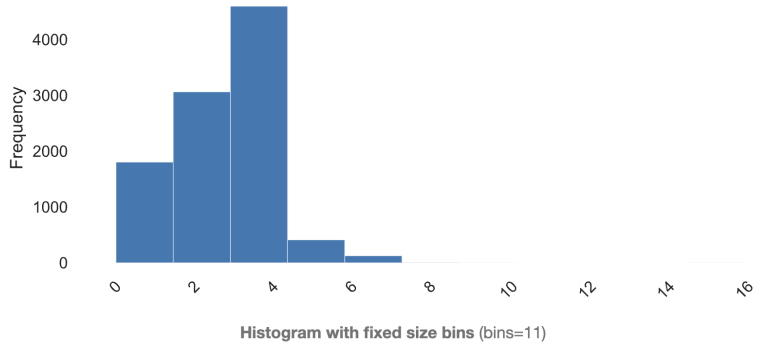
La distribución de voteAvg presenta la forma normal.

voteCount



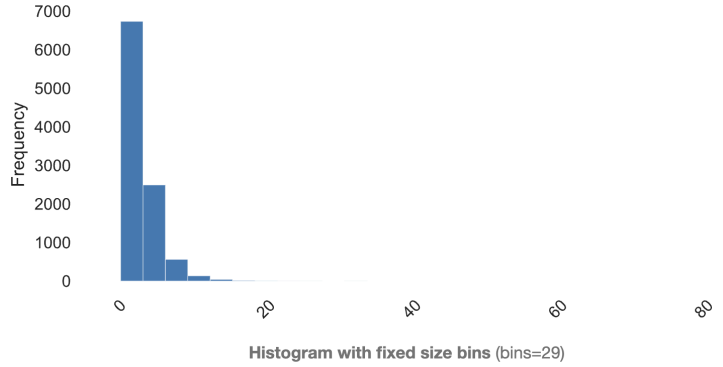
La distribución de voteCount presenta un sesgo hacia la izquierda.

genresAmount



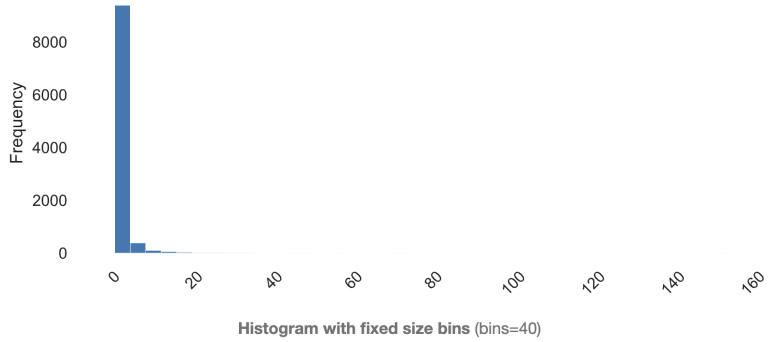
La distribución de genresAmount presenta un sesgo a la izquierda.

productionCoAmount



La distribución de productionCoAmount presenta un sesgo a la izquierda.

productionCountriesAmout



La distribución productionCountriesAmount tiene un sesgo a la izquierda.

Variables Cualitativas:

genres

Value	Count	Frequency (%)
fiction	592	5.2%
drama	521	4.5%
comedy	440	3.8%
science	305	2.7%
horror	230	2.0%
drama|romance	211	1.8%
horror|thriller	205	1.8%
comedy|roma…	201	1.7%
documentary	194	1.7%
movie	185	1.6%
Other values …	8406	73.2%

La tabla de frecuencias nos muestra variedad en los generos.

homepage
Todas las homepage son diferentes entonces no vale la pena hacer una tabla de frecuencias.

productionCompany:

Value	Count	Frequency (%)
pictures	1759	3.5%
films	1260	2.5%
entertainment	1231	2.5%
productions	1155	2.3%
film	919	1.8%
bros	670	1.3%
media	441	0.9%
disney	384	0.8%
century	371	0.7%
animation	361	0.7%
Other values …	41400	82.9%

La tabla de frecuencias nos muestra que la mayoria de compañias se agrupan en otros valores.

producionCompanyCountry:

Value	Count	Frequency (%)
us	1959	22.4%
us|us	1187	13.6%
us|us|us	579	6.6%
jp	253	2.9%
us|us|us|us	220	2.5%
	189	2.2%
us||us	184	2.1%
gb	125	1.4%
ca	94	1.1%
jp|jp	81	0.9%
Other values …	3858	44.2%

La tabla de frecuencias nos indica que la mayoría de películas fueron producidas en US.

productionCountry

Value	Count	Frequency (%)
of	6788	21.3%
states	6788	21.3%
america	6472	20.3%
united	5910	18.5%
japan	613	1.9%
kingdom|united	579	1.8%
kingdom	502	1.6%
canada|united	263	0.8%
south	195	0.6%
germany|united	164	0.5%
Other values …	3599	11.3%

La tabla de frecuencias indica que la producción se dio mas en United States of America

director

Value	Count	Frequency (%)
john	262	1.2%
david	249	1.1%
michael	202	0.9%
peter	131	0.6%
robert	130	0.6%
james	118	0.5%
paul	116	0.5%
scott	106	0.5%
lee	99	0.4%
steven	97	0.4%
Other values …	20805	93.2%

La tabla de frecuencias indica los nombres más repetidos en la columna de directores.

originalLanguage:

Value	Count	Frequency (%)
en	7771	77.7%
ja	644	6.4%
es	425	4.2%
fr	271	2.7%
ko	167	1.7%
zh	119	1.2%
it	100	1.0%
de	84	0.8%
cn	80	0.8%
ru	67	0.7%
Other values (30)	271	2.7%

La tabla de frecuencias indica que la mayoría de películas fueron grabadas originalmente en inglés.

Responda las siguientes preguntas:

(3 puntos) ¿Cuáles son las 10 películas que contaron con más presupuesto?

	title	budget
716	Pirates of the Caribbean: On Stranger Tides	380000000
4710	Avengers: Age of Ultron	365000000
5952	Avengers: Endgame	356000000
4953	Justice League	300000000
5953	Avengers: Infinity War	300000000
163	Pirates of the Caribbean: At World's End	300000000
607	Superman Returns	270000000
7134	The Lion King	260000000
3791	Tangled	260000000
280	Spider-Man 3	258000000

(3 puntos) ¿Cuáles son las 10 películas que más ingresos tuvieron?

	title	revenue
3210	Avatar	2847246203
5952	Avengers: Endgame	2797800564
307	Titanic	2187463944
4947	Star Wars: The Force Awakens	2068223624
5953	Avengers: Infinity War	2046239637
4914	Jurassic World	1671713208
7134	The Lion King	1667635327
9049	Spider-Man: No Way Home	1631853496
3397	The Avengers	1518815515
5087	Furious 7	1515047671

(3 puntos) ¿Cuál es la película que más votos tuvo?

	title	voteCount
3511	Inception	30788.0

(3 puntos) ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios?

	title	voteAvg
9786	DAKAICHI -I'm Being Harassed by the Sexiest Ma...	1.3

(8 puntos) ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras

La cantidad de películas hechas en cada año varía desde 1 hasta más de 800 películas:

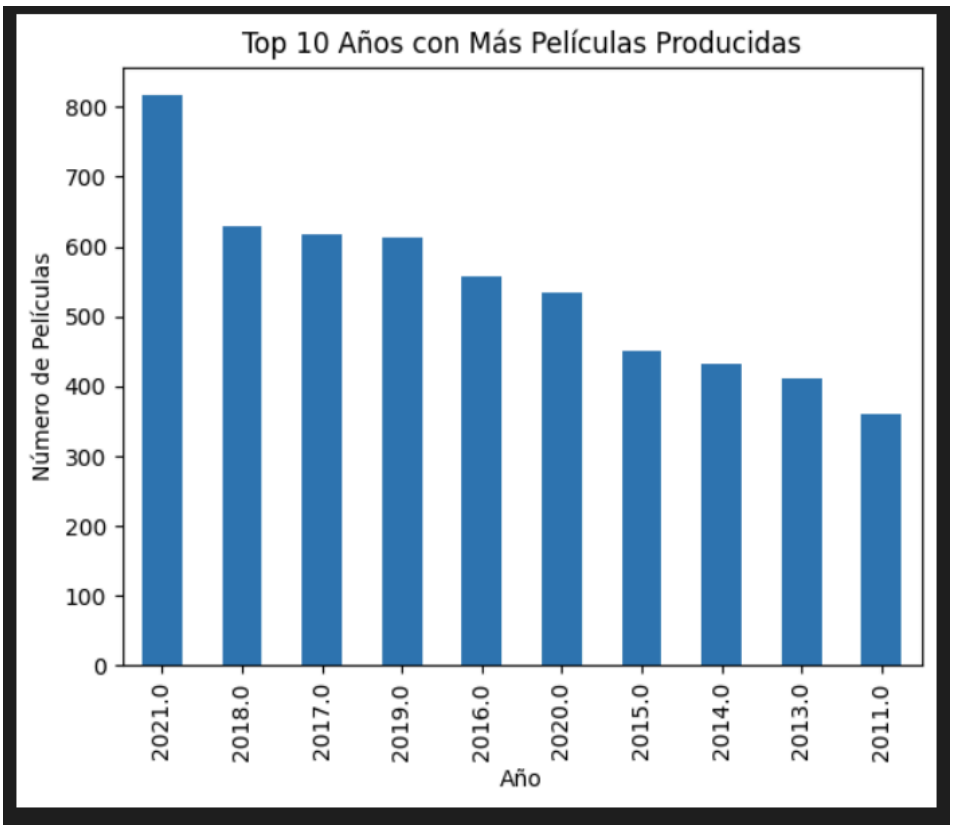
```
releaseDate
1902.0      1
1920.0      1
1921.0      1
1922.0      1
1925.0      2
           ...
2018.0    629
2019.0    612
2020.0    533
2021.0    816
2022.0      7
Length: 99, dtype: int64
```

El año con más películas fue el 2021, con 816 películas lanzadas.

```
releaseDate
2021.0    816
dtype: int64
```

Y el top 10 de años con más películas lanzadas son los siguientes:

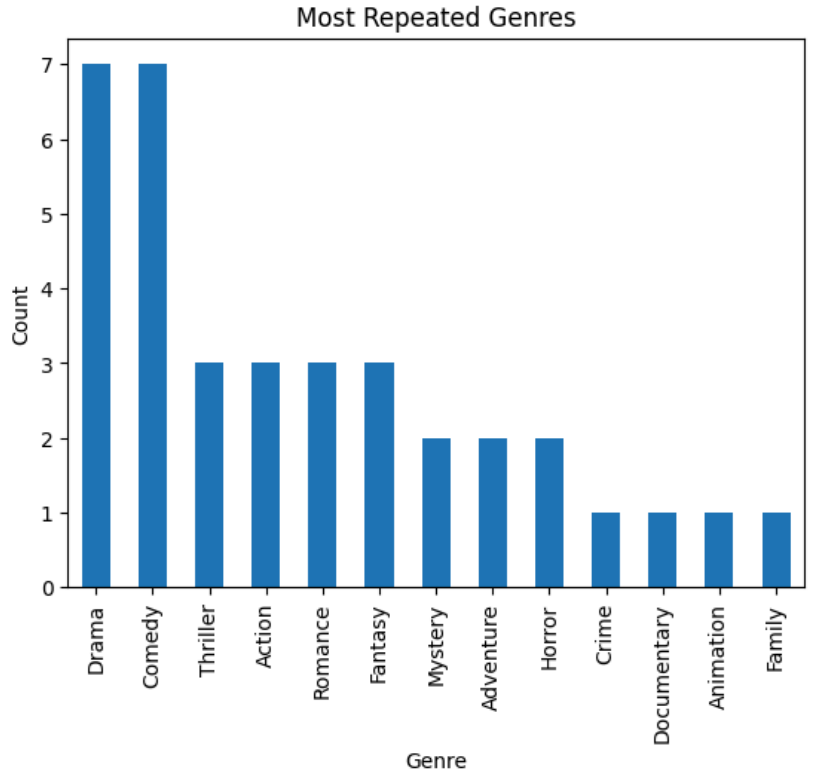
```
releaseDate
2021.0    816
2018.0    629
2017.0    618
2019.0    612
2016.0    557
2020.0    533
2015.0    450
2014.0    432
2013.0    412
2011.0    361
dtype: int64
```



Cómo se puede observar, todos son años después del 2010, lo que indica que en los últimos años la cantidad de películas producidas y lanzadas han incrementado y siguen esta tendencia alcista.

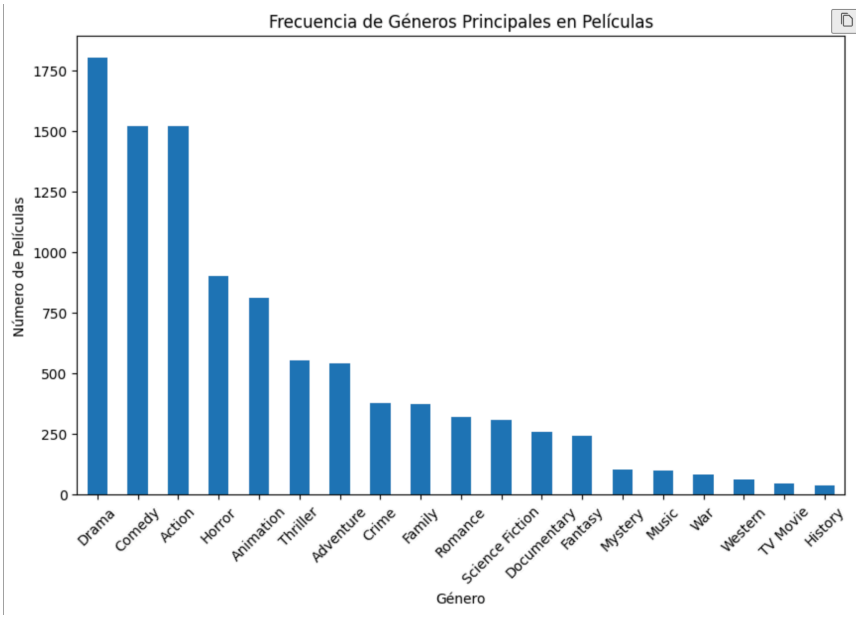
(9 puntos) ¿Cuál es el género principal de las 20 películas más recientes? ¿Cuál es el género principal que predomina en el conjunto de datos? Represéntelo usando un gráfico

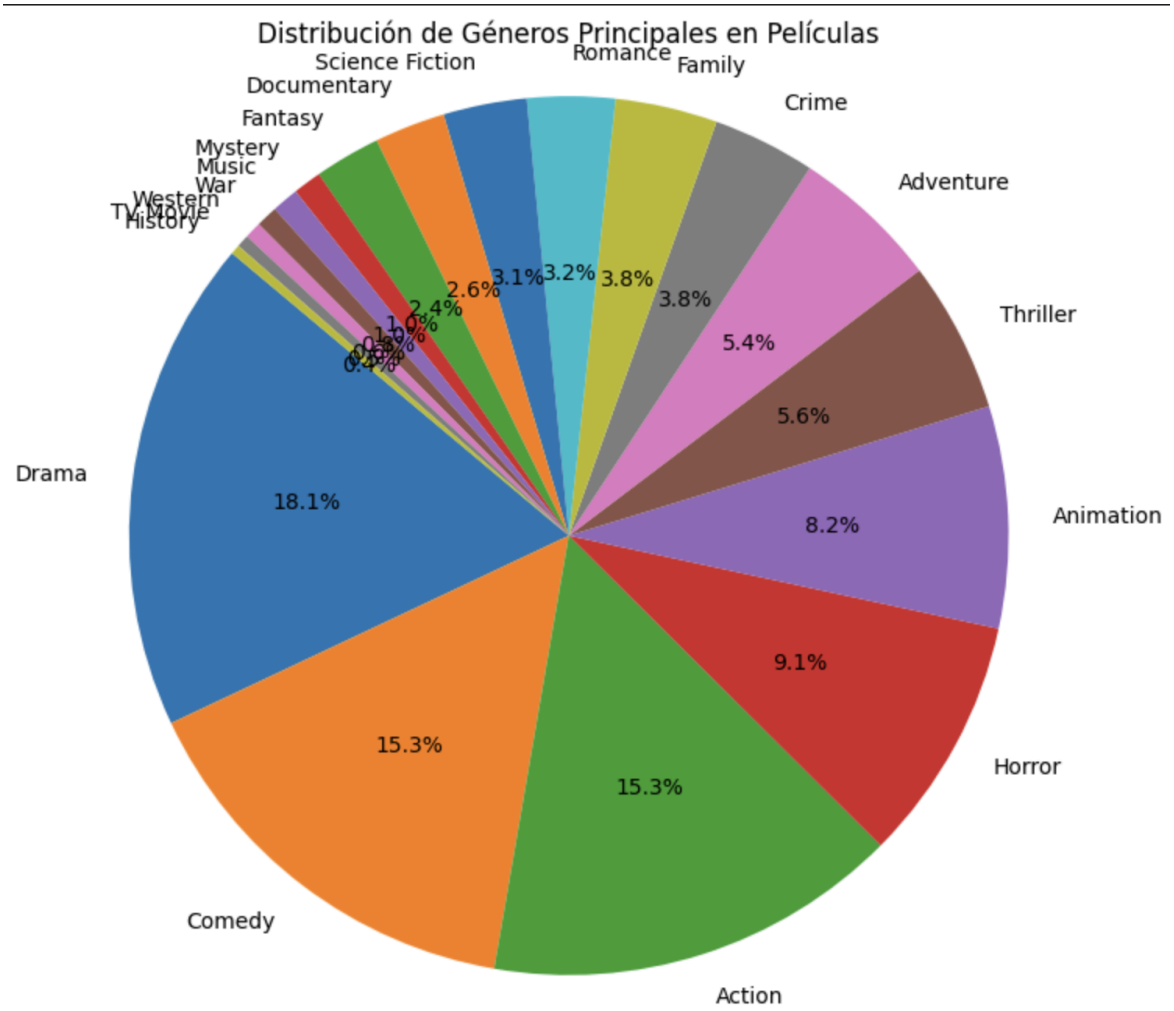
Entre las 20 películas más recientes, predomina el género de Drama y Comedia.



En cuánto al conjunto de datos general, el género dominante es Drama. Seguido por la comedia y la acción.

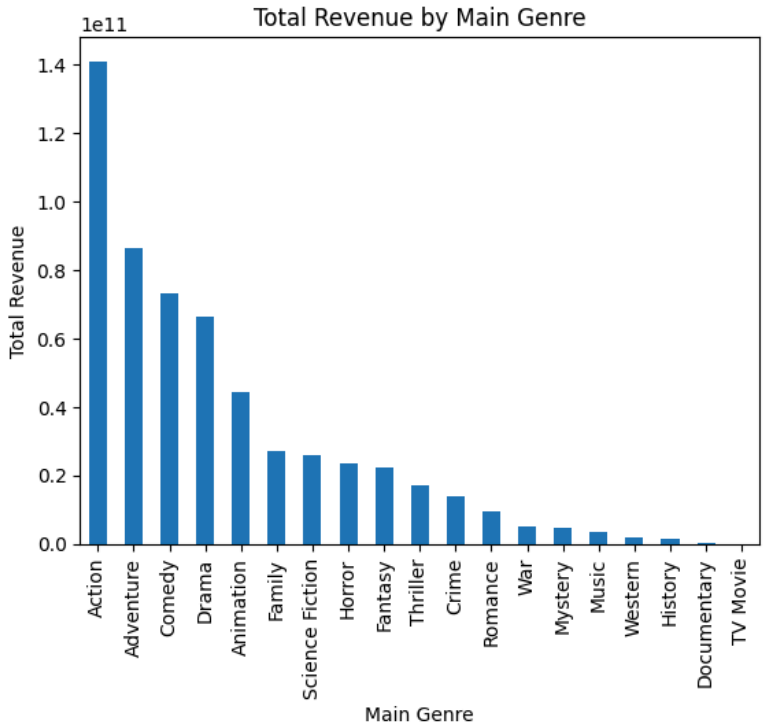
```
El género principal más predominante es: Drama
```




Distribución de Géneros Principales en Películas
Science Fiction
Romance
Family
Documentary
Crime
Fantasy
Adventure
Mystery
Music
War
Western
TV Movie
History
Thriller
Drama
Animation
Horror
Comedy
Action
18.1%
15.3%
15.3%
9.1%
8.2%
5.6%
5.4%
3.8%
3.8%
3.2%
3.1%
2.6%
2.4%

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

El género que mayor ganancia obtiene es la Acción, seguido por las películas de aventura y comedia. Los géneros que menos ganancias obtienen son las películas de televisión, los documentales y/o películas de historia.



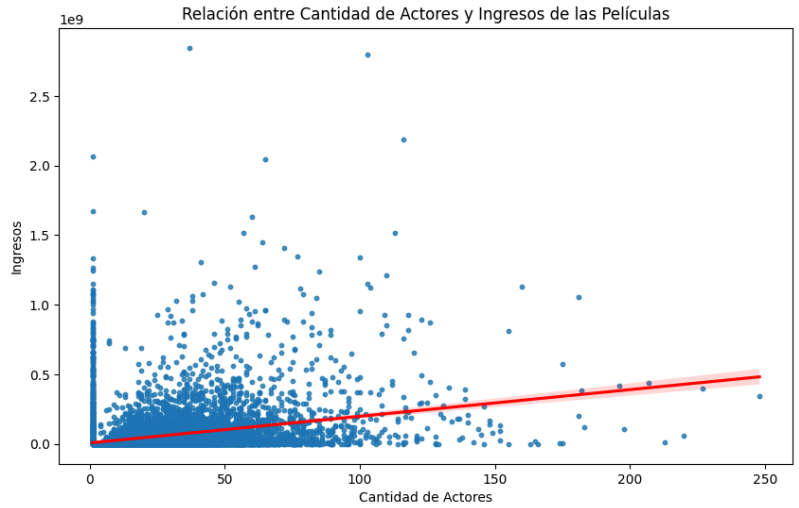

```
Ganancias totales por género principal:
main_genre
Action             140936671043
Adventure           86313291491
Comedy              72990070028
Drama               66415119599
Animation           44193668686
Family              27070466981
Science Fiction     25771021939
Horror              23483468549
Fantasy             22309474780
Thriller            17111426999
Crime               13764408000
Romance              9582546901
War                  5151630856
Mystery              4640756330
Music                3493810817
Western              1843584942
History              1481950337
Documentary           352161888
TV Movie                      0
Name: revenue, dtype: int64
```

(3 puntos) ¿La cantidad de actores influye en los ingresos de las películas?¿se han hecho películas con más actores en los últimos años?

Películas con más actores:

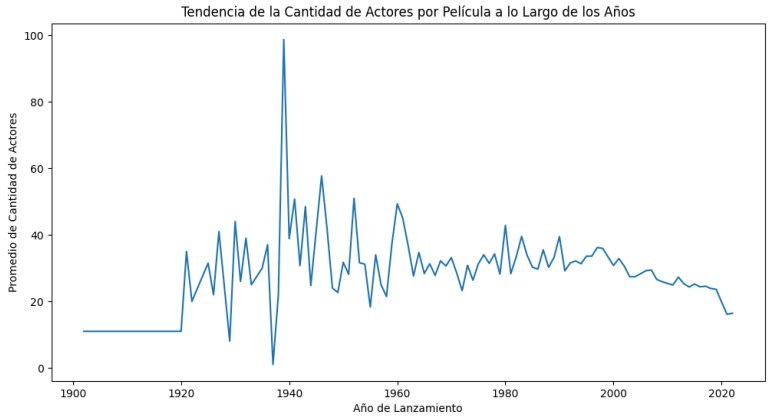
	title	actors_count
1015	Enchanted	248
7586	Mamma Mia! Here We Go Again	227
4543	Rock of Ages	220
919	Mr. Smith Goes to Washington	213
4578	Les Misérables	207

Los resultados indican una correlación positiva entre la cantidad de actores en una película y los ingresos que ésta genera. Sin embargo, el coeficiente de correlación es aproximadamente 0.28, lo que nos indica una influencia pobre. Aunque la relación es estadísticamente significativa, la fuerza de esta relación es débil. Esto significa que otros factores, además de la cantidad de actores, probablemente tengan un impacto más fuerte en los ingresos de estas películas.



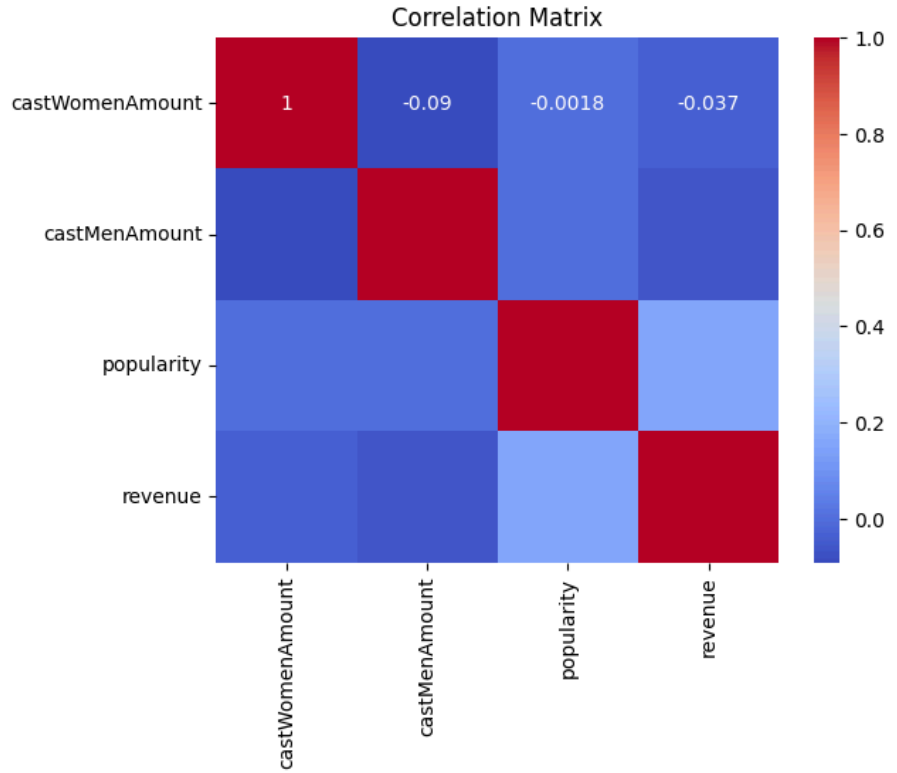

```
Correlación entre la cantidad de actores y los ingresos: 0.28389208800199184
```

Y no, en los últimos años la cantidad de actores por película ha disminuido, para tener actualmente un promedio alrededor de 15 actores por película. El pico histórico es de casi 100 actores y pertenece a la época de 1940.



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

Con base en la matriz de correlación presentada a continuación, se puede observar una correlación muy débil entre la cantidad de hombres o mujeres en el reparto y la popularidad o ingresos de las películas. La correlación entre la cantidad de mujeres en el reparto y la popularidad es casi insignificante (-0.001780), y la correlación entre la cantidad de hombres en el reparto y los ingresos es débil (-0.060145). Esto sugiere que, en general, la cantidad de hombres o mujeres en el reparto no tiene un impacto significativo en la popularidad o los ingresos de las películas según estos datos.



	castWomenAmount	castMenAmount	popularity	revenue
castWomenAmount	1.000000	-0.090478	-0.001780	-0.037398
castMenAmount	-0.090478	1.000000	0.002776	-0.060145
popularity	-0.001780	0.002776	1.000000	0.162545
revenue	-0.037398	-0.060145	0.162545	1.000000

(8 puntos) ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

Los directores responsables de las 20 películas mejor calificadas son:

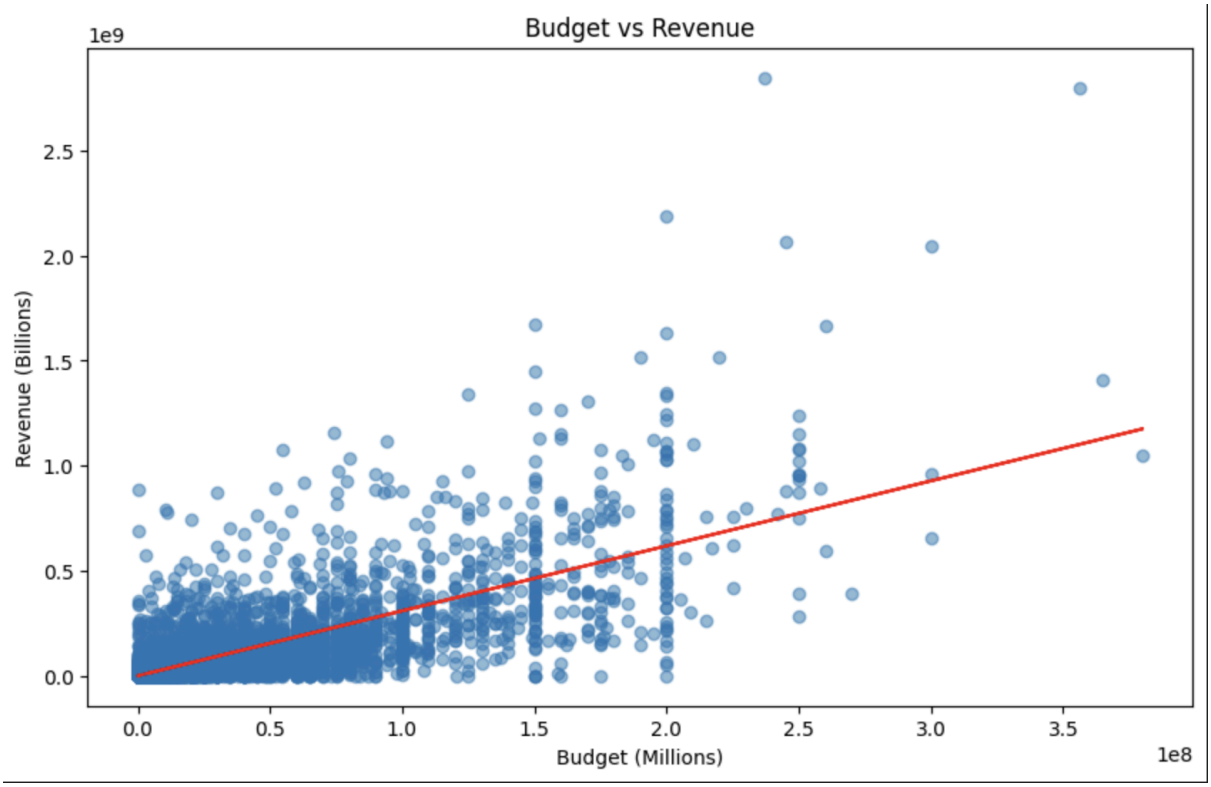
- Victor Barba y Juan Olivares
- Kaku Arakawa
- Laurent Bouzereau
- Miguel Angel Zavala
- Christin Baker
- Thomas Coven
- Rebecca Sugar
- Dave Bullock, Troy Adomitis y Victor Cook
- Won Myeong-jun
- …

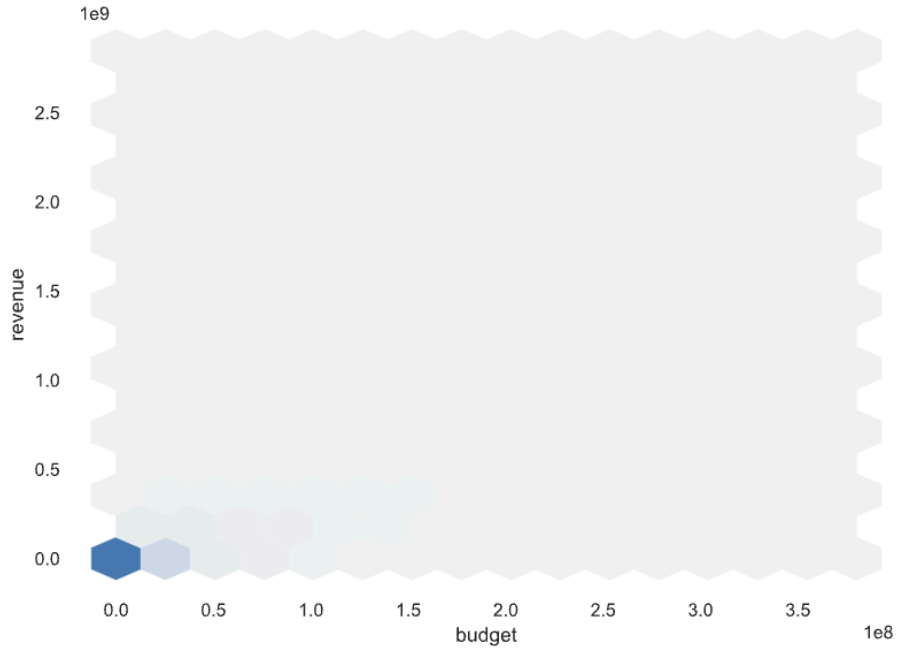
	director	voteAvg
16	Víctor Barba|Juan Olivares	10.0
5	Kaku Arakawa	10.0
7	Laurent Bouzereau	10.0
8	Miguel Angel Zavala	10.0
1	Christin Baker	10.0
14	Thomas Coven	10.0
11	Rebecca Sugar	10.0
2	Dave Bullock|Troy Adomitis|Victor Cook	9.6
17	Won Myeong-jun	9.5
12	Samuel Leong	9.5
13	Selena Quintanilla	9.4
3	Haruo Sotozaki	9.3
15	Ulises Valencia	9.2
9	Park Jun-soo	9.2
6	Kim Nam-joon|Jeon Jung-kook|Kim Tae-hyung|Park...	9.2
4	Igor Kopylov	9.2
10	Preston A. Whitmore II	9.0
0	Carlos Pérez Osorio	9.0

(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.

Se calculó una correlación positiva de 0.7575 entre los presupuestos y los ingresos sugiere una fuerte relación en la que, en general, los aumentos en los presupuestos tienden a asociarse con aumentos en los ingresos, y viceversa. Para visualizar esta relación, se puede utilizar un diagrama de dispersión para mostrar la tendencia de los datos y un histograma para examinar la distribución de los presupuestos y los ingresos por separado. El diagrama de dispersión demuestra la tendencia en esta relación, sin embargo es importante notar que existen valores distantes de la tendencia, que también difieren más conforme aumenta el presupuesto, ya que la cantidad de películas con presupuestos sumamente altos es baja. Pero en general se puede observar que es una correlación positiva y fuerte.

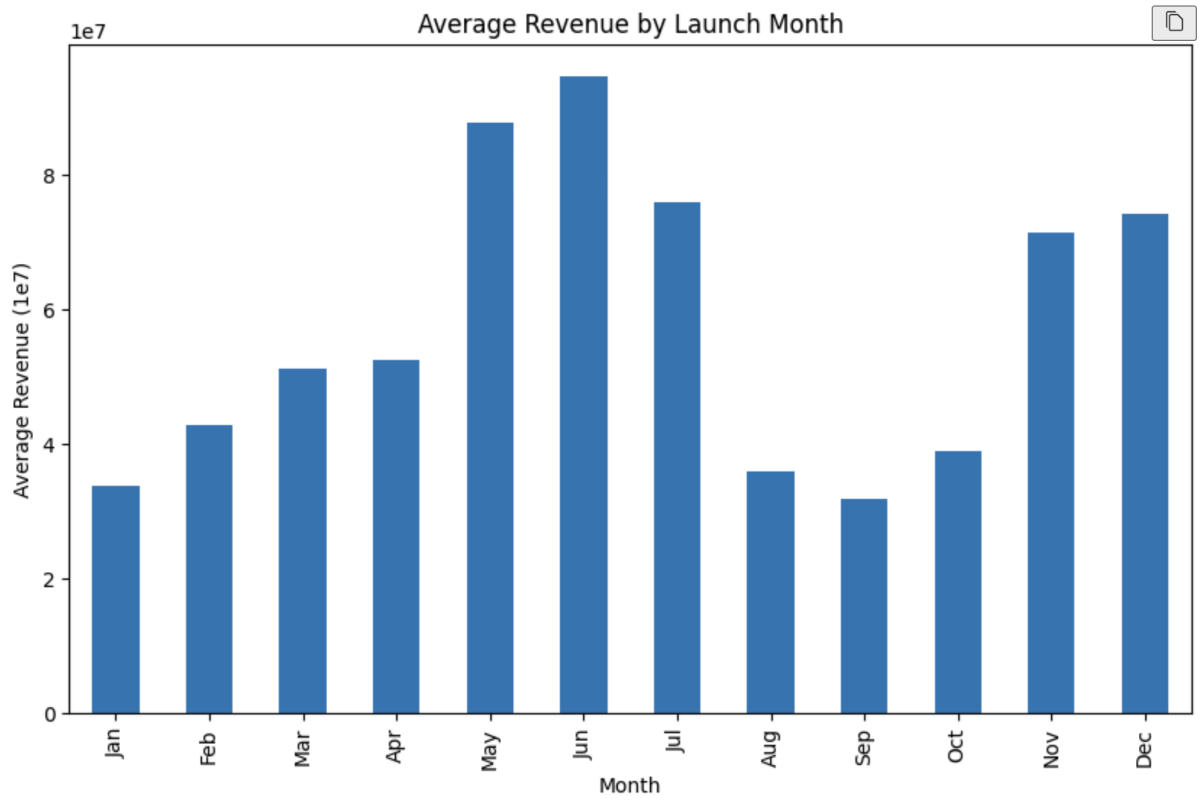
```
Correlation between budget and revenue: 0.757454042495599
```




1e9
2.5
2.0
1.5
1.0
0.5
0.0
revenue
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e8
budget

(7 puntos) ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?

Los ingresos parecen variar significativamente según el mes de lanzamiento, con mayo, junio y diciembre mostrando consistentemente los ingresos más altos, mientras que agosto, septiembre y enero tienden a tener ingresos más bajos en comparación. Por lo que si se tuviera que elegir un mes del año para lanzar una película y esperar los mejores ingresos, se recomendaría el período de tiempo entre Mayo y Junio.



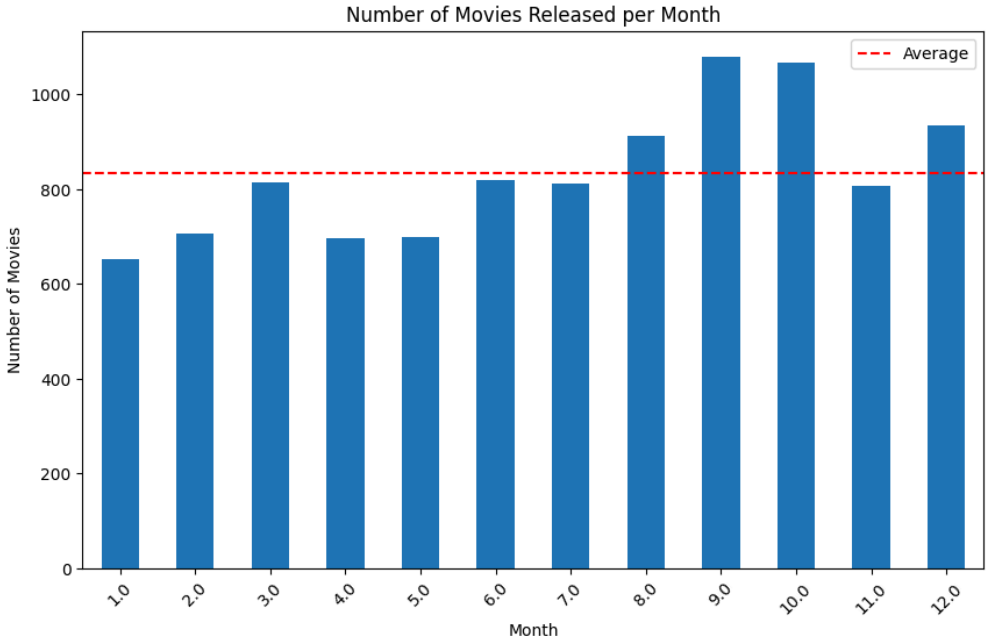
(8 puntos) ¿En qué meses se han visto los lanzamientos con mejores ingresos?¿cuantas películas, en promedio, se han lanzado por mes?

Las 10 películas con menores ingresos fueron lanzadas (en orden de menores ingresos a mayores) en los meses de:

- Junio
- Agosto
- Mayo
- Octubre
- Marzo
- Noviembre
- Diciembre
- Septiembre
- Julio

```
                            title      month  revenue
2805              Sex and Death 101       June        1
8325                    Brian Banks     August        4
825          Capturing the Friedmans        May        4
3666  Chestnut: Hero of Central Park    October       10
3010                    Hero Wanted      March       10
8447      Burn the Stage: The Movie   November       20
9228                The Last Avatar   December       27
2999    Rugrats in Paris: The Movie  September      103
2665                 Empire Records  September      303
7042                        Maligno       July      400
Average number of movies released per month: 833.25
```

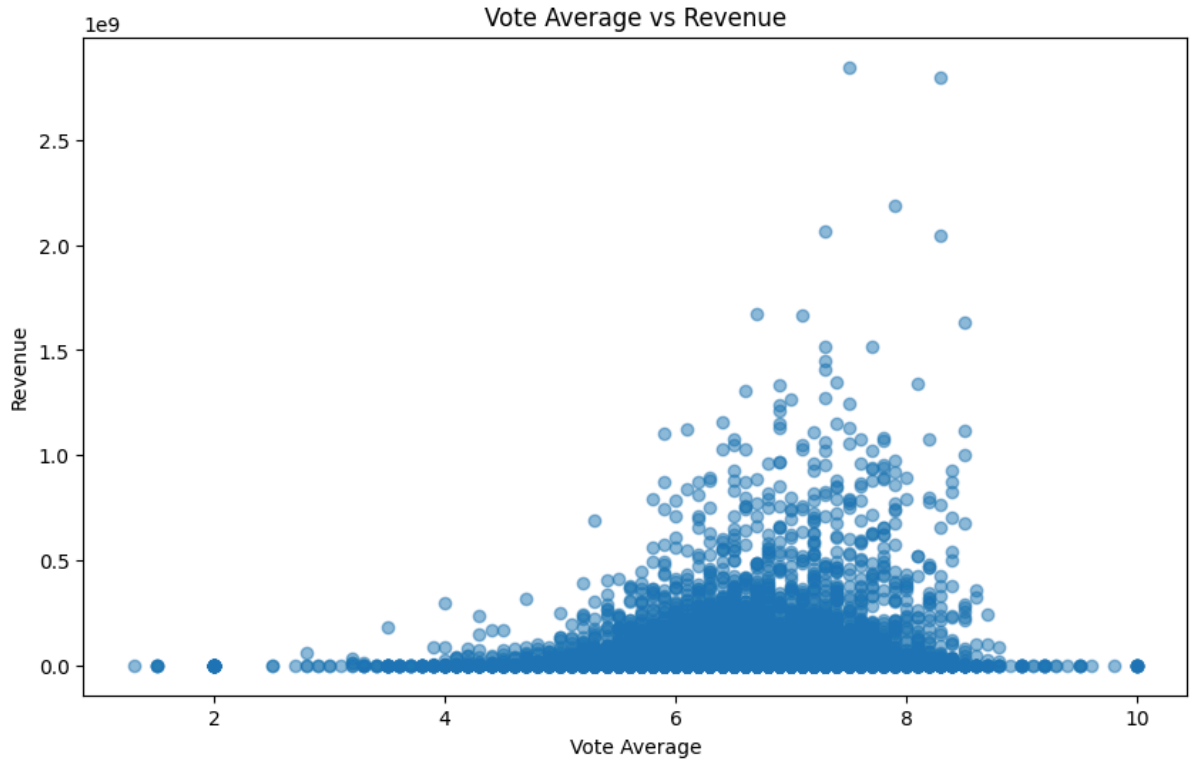
Y el promedio de películas lanzadas por mes es de 833 películas. Con los meses septiembre y octubre siendo los de mayor cantidad de lanzamientos.



(7 puntos) ¿Cómo se correlacionan las calificaciones con el éxito comercial?

La correlación entre las calificaciones y el éxito comercial, medida por los ingresos generados, es bastante baja, con un coeficiente de correlación de 0.141. Esto sugiere una relación débil entre estas dos variables. Mientras que una correlación positiva significativa podría indicar que las películas mejor valoradas tienden a generar más ingresos, esta correlación baja sugiere que la calidad percibida por los individuos calificadores, como se refleja en las calificaciones, no es un predictor fuerte del éxito comercial.

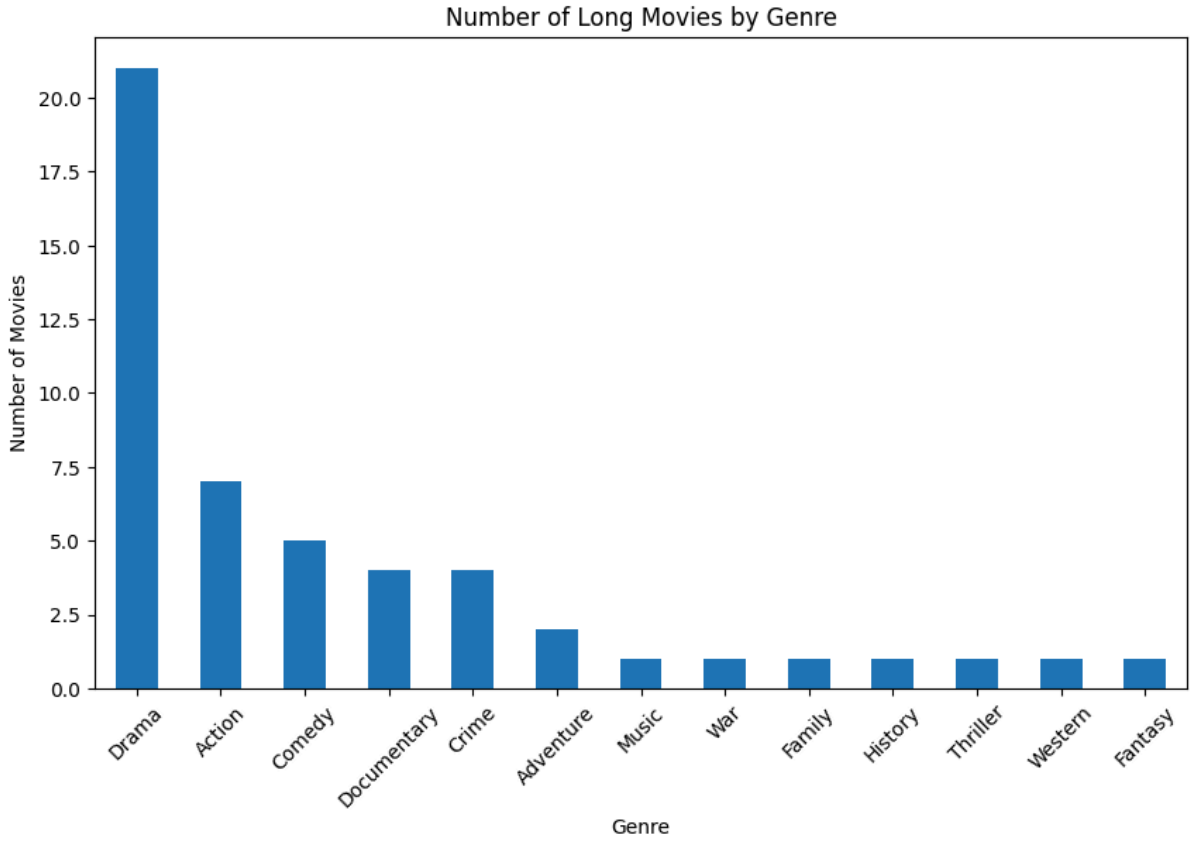
```
Correlation between voteAvg and revenue: 0.14126438971337793
```



(5 puntos) ¿A qué género principal pertenecen las películas más largas?

Tomando en cuenta las 50 películas con mayor duración, podemos evidenciar que el género dominante es el drama, con poco más del 40% de las películas más largas perteneciendo a este género, seguida por los géneros de acción y comedia.

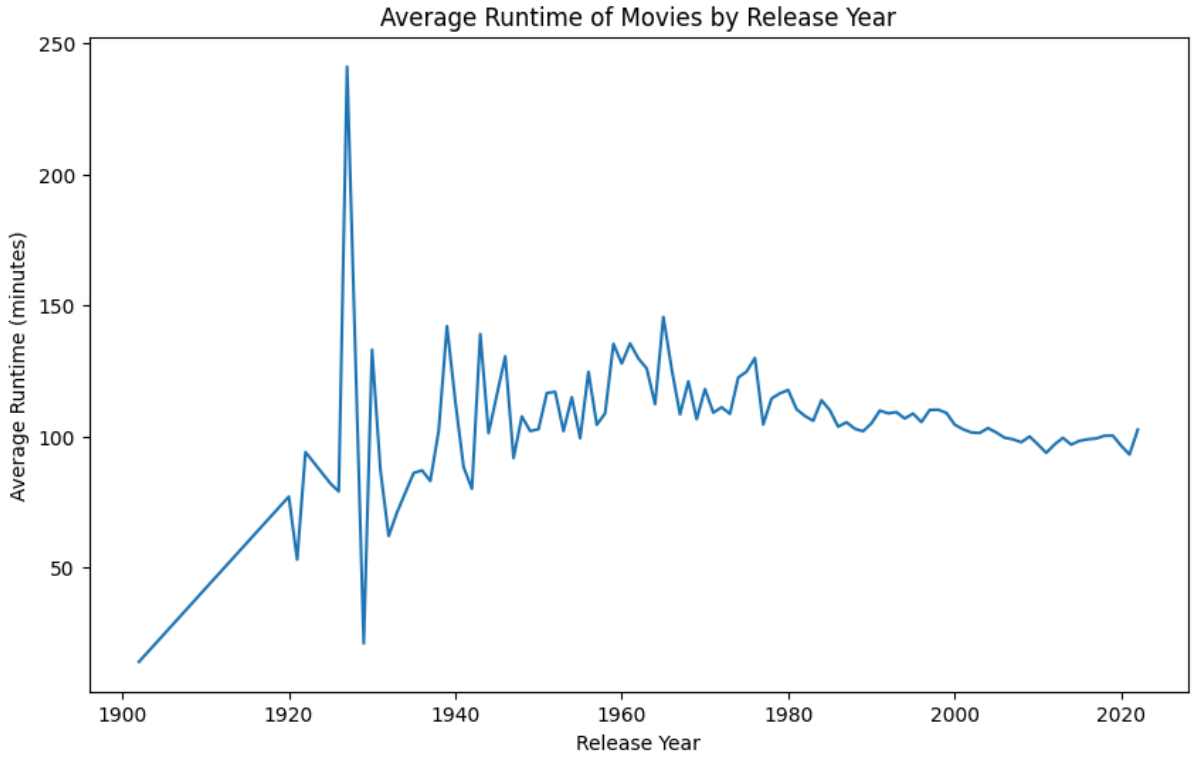
```
El género principal más común entre las películas más largas es: Drama
```



(¡10 puntos 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.

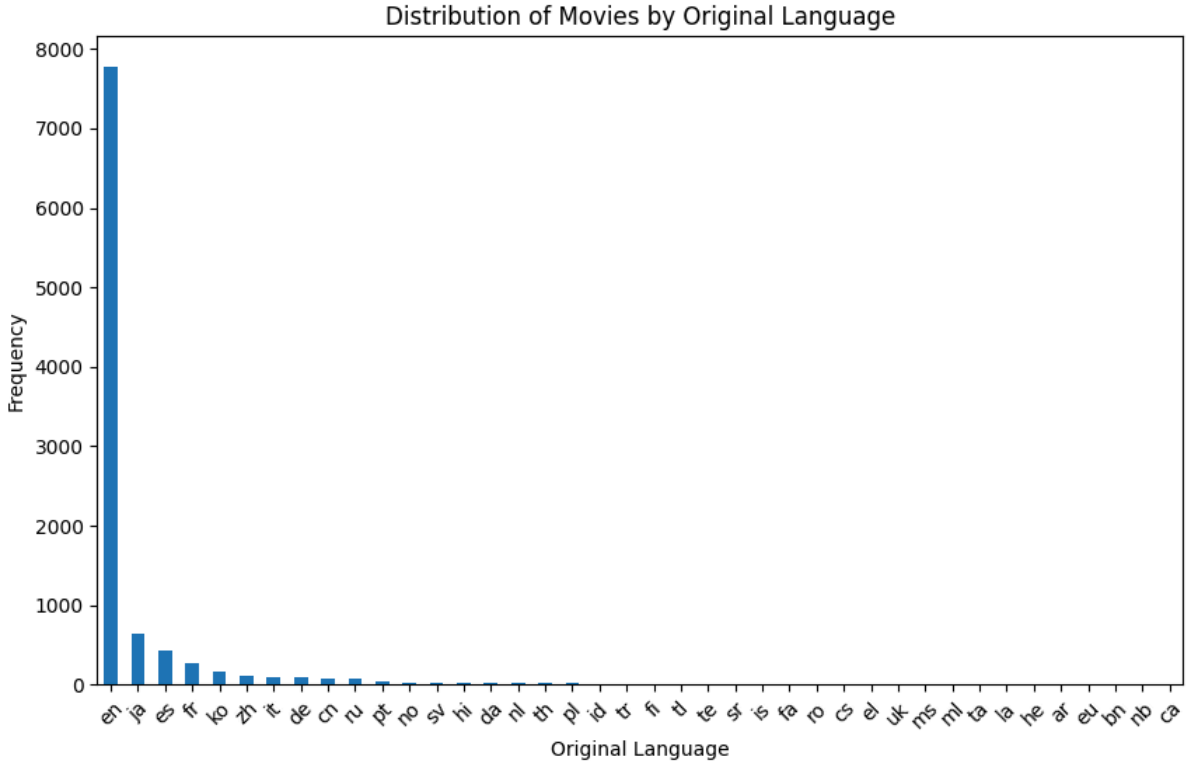
1) ¿Cómo ha variado la duración de las películas con el paso de los años?

Se puede observar y determinar que con el paso del tiempo, el promedio de duración de las películas se ha consolidado alrededor de los 120 minutos. Ya que lleva un rango en estos valores desde la década de los 80s. Es interesante ver que al principio del registro de los datos, las primeras películas de la historia duraban relativamente poco, y tuvieron una tendencia alcista que las llevó a tener un pico de casi 240 minutos de promedio de duración, que fue un all time high hasta reducirse nuevamente a lo que hoy en día es el promedio.



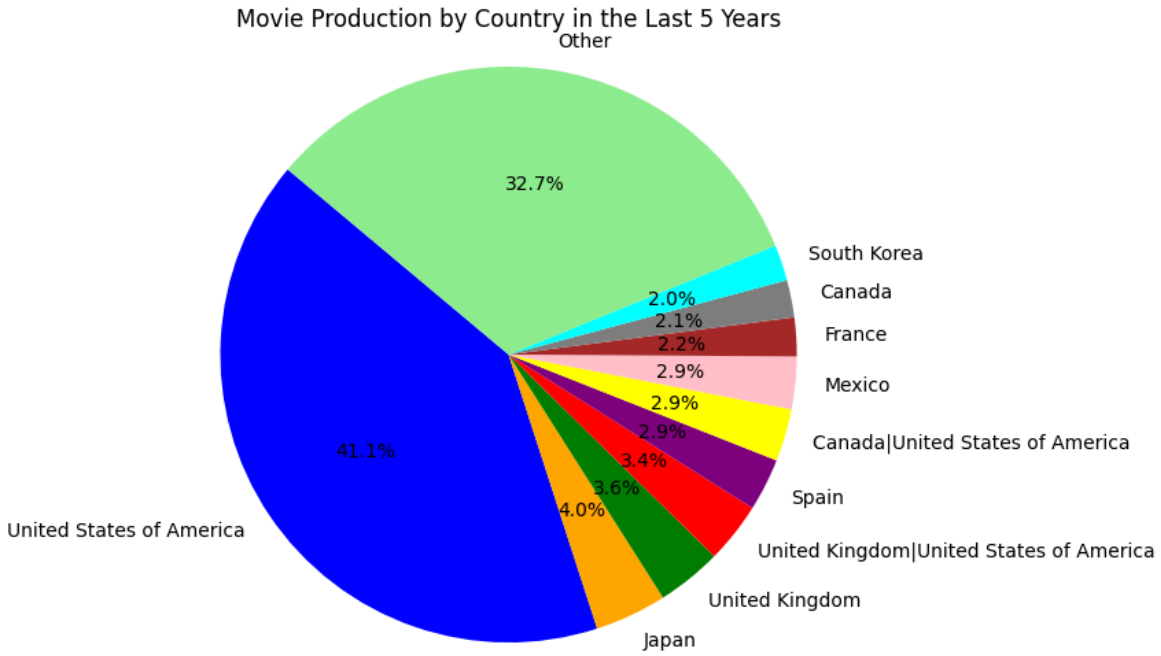
2) ¿Cómo se distribuyen las películas según el idioma original en el que fueron producidas?

Se puede observar que la total mayoría de las películas registradas fueron lanzadas originalmente en idioma inglés, seguidas del japonés que es interesante ya que este no es uno de los lenguajes más hablados del mundo, y en tercer lugar el español.



3) ¿Qué países han sido los principales productores de películas en los últimos 5 años?

Los principales países productores de películas han sido Estados Unidos, Japón, Inglaterra y España. Lo cuál demuestra que la respuesta anterior tiene validez, ya que los idiomas más utilizados son los correspondientes a los países más productores del mundo.



4) ¿Cuáles son los actores que más aparecen en las 50 películas con mayores ingresos?

   Los avengers 🦸

```
Top most repeated actors in the top 50 most grossing movies:
Stan Lee
Scarlett Johansson
Mark Ruffalo
Chris Evans
```

5) ¿Hay alguna correlación entre la cantidad de géneros de una película (genresAmount) y su éxito en términos de ingresos o votos?

No, la correlación entre estas dos variables es baja, con un valor de 0.13. Lo que indica que no por intentar cubrir más géneros con una misma película se conseguirán más ingresos. Por lo que se podría decir que es mejor centrarse en un género y tener un buen rendimiento en este.

```
Correlation between genresAmount and revenue: 0.13627831764306778
```

6) ¿Las compañías que más presupuesto invierten en sus películas son las mismas que más revenue generan?

Sí, como podemos observar al mostrar las 5 compañías que más invierten y las 5 compañías que más ingresos tienen, podemos observar que la mayoría entran en ambas categorías. Cómo Warner bros, Universal y Disney.

```
Top 5 Spending Production Companies:
productionCompany
Warner Bros. Pictures    25548927834
Universal Pictures       18182211864
Columbia Pictures        17011635000
Walt Disney Pictures     15955875223
Paramount                14740285844
Name: budget, dtype: int64
Total Budget of Top 5 Spending Production Companies: 91438935765

Top 5 Most Revenued Production Companies:
productionCompany
Warner Bros. Pictures    73640311482
Universal Pictures       60766812506
20th Century Fox         53680594765
Walt Disney Pictures     50618246120
Columbia Pictures        49716805742
Name: revenue, dtype: int64
Total Revenue of Top 5 Most Revenued Production Companies: 288422770615
```

Referencias

Equipo de desarrollo de Matplotlib. (2023). Matplotlib: A Comprehensive Library for Creating Static, Animated, and Interactive Visualizations in Python [Software]. Disponible en https://matplotlib.org

McKinney, W., & otros colaboradores de Pandas. (2023). pandas: Powerful data structures for data analysis, time series, and statistics [Software]. Disponible en https://pandas.pydata.org

Waskom, M. (2023). Seaborn: Statistical Data Visualization [Software]. Disponible en https://seaborn.pydata.org