Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
Minería de Datos

Hoja de Trabajo 1

Gabriel Quiroz 19255
Jose Pablo Ponce 19092
Alejandra Gudiel 19232

1. (3 puntos) Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

	id	budget	...	castWomenAmount	castMenAmount
0	5	4000000	...	15	9
1	6	21000000	...	3	9
2	11	11000000	...	5	62
3	12	94000000	...	5	18
4	13	55000000	...	18	48
...	...	...	...	...	...
9995	920081	0	...	2	4
9996	920143	0	...	1	1
9997	922017	0	...	0	922017
9998	922162	0	...	922162	The Witcher: Fireplace
9999	922260	254056	...	4	3

[10000 rows x 27 columns]

El total de filas es 10000 y el total de columnas es 27

	id	budget	...	productionCountriesAmount	actorsAmount
count	10000.000000	1.000000e+04	...	10000.000000	10000.000000
mean	249876.829300	1.855163e+07	...	1.751000	2147.666600
std	257380.109004	3.662669e+07	...	3.012093	37200.075802
min	5.000000	0.000000e+00	...	0.000000	0.000000
25%	12286.500000	0.000000e+00	...	1.000000	13.000000
50%	152558.000000	5.000000e+05	...	1.000000	21.000000
75%	452021.750000	2.000000e+07	...	2.000000	36.000000
max	922260.000000	3.800000e+08	...	155.000000	919590.000000

[8 rows x 11 columns]

2. (5 puntos) Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)

- **Id:** Cualitativa ordinal
- **popularity:** Cuantitativa continua
- **budget:** Cuantitativa discreta
- **revenue:** Cuantitativa discreta
- **original_title:** Cualitativa nominal
- **originalLanguage**: Cualitativa nominal
- **title:** Cualitativa nominal
- **homePage:** Cualitativa nominal
- **video:** Cualitativa nominal
- **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**: cuantitativa discreta
- **voteCount:** cuantitativa discreta
- **voteAvg:** cuantitativa continua
- **actors:** cualitativa nominal
- **actorsPopularity:** cualitativa nominal
- **actorsCharacter:** cualitativa ordinal
- **actorsAmount:** cuantitativa discreta
- **castWomenAmount:** cuantitativa discreta
- **castMenAmount:** cuantitativa discreta

3. (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.

```
popularity
Estadisticos=26112.258, p=0.000
La variable popularity no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
budget
Estadisticos=6883.659, p=0.000
La variable budget no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
revenue
Estadisticos=11425.057, p=0.000
La variable revenue no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
genresAmount
Estadisticos=1150.580, p=0.000
La variable genresAmount no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
productionCoAmount
Estadisticos=12245.725, p=0.000
La variable productionCoAmount no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
productionCountriesAmount
Estadisticos=21809.083, p=0.000
La variable productionCountriesAmount no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
voteCount
Estadisticos=8610.621, p=0.000
La variable voteCount no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
voteAvg
Estadisticos=810.999, p=0.000
La variable voteAvg no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

```
actorsAmount
Estadisticos=21197.011, p=0.000
La variable actorsAmount no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)
```

4. Responda las siguientes preguntas:

4.1. (3 puntos) ¿Cuáles son las 10 películas que contaron con más presupuesto?

```
7    a = movies.nlargest(10, 'budget')['title']
8    print(a)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

alegudiel@Ales-MacBook-Pro analisis-exploratorio-1 % /L
rk/Versions/3.9/bin/python3 /Users/alegudiel/Documents/
movies.py
716     Pirates of the Caribbean: On Stranger Tides
4710                        Avengers: Age of Ultron
5952                              Avengers: Endgame
163         Pirates of the Caribbean: At World's End
4953                                    Justice League
5953                         Avengers: Infinity War
607                                Superman Returns
3791                                        Tangled
7134                                 The Lion King
280                                   Spider-Man 3
```

4.2. (3 puntos) ¿Cuáles son las 10 películas que más ingresos tuvieron?

```
10   b = movies.nlargest(10, 'revenue')['title']
11   print(b)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

alegudiel@Ales-MacBook-Pro analisis-exploratorio-1 % /Li
rk/Versions/3.9/bin/python3 /Users/alegudiel/Documents/M
movies.py
3210                          Avatar
5952                Avengers: Endgame
307                           Titanic
4947    Star Wars: The Force Awakens
5953          Avengers: Infinity War
4914                  Jurassic World
7134                   The Lion King
9049         Spider-Man: No Way Home
3397                    The Avengers
5087                        Furious 7
```

4.3. (3 puntos) ¿Cuál es la película que más votos tuvo?

```
13    c = movies.nlargest(10, 'voteCount')['title'
14    print(c)

PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL

alegudiel@Ales-MacBook-Pro analisis-exploratorio-1 % /Li
rk/Versions/3.9/bin/python3 /Users/alegudiel/Documents/M
movies.py
3511                  Inception
5027               Interstellar
87              The Dark Knight
3397               The Avengers
5889                   Deadpool
3210                     Avatar
4825    Guardians of the Galaxy
5953     Avengers: Infinity War
275                  Fight Club
374                Pulp Fiction
```

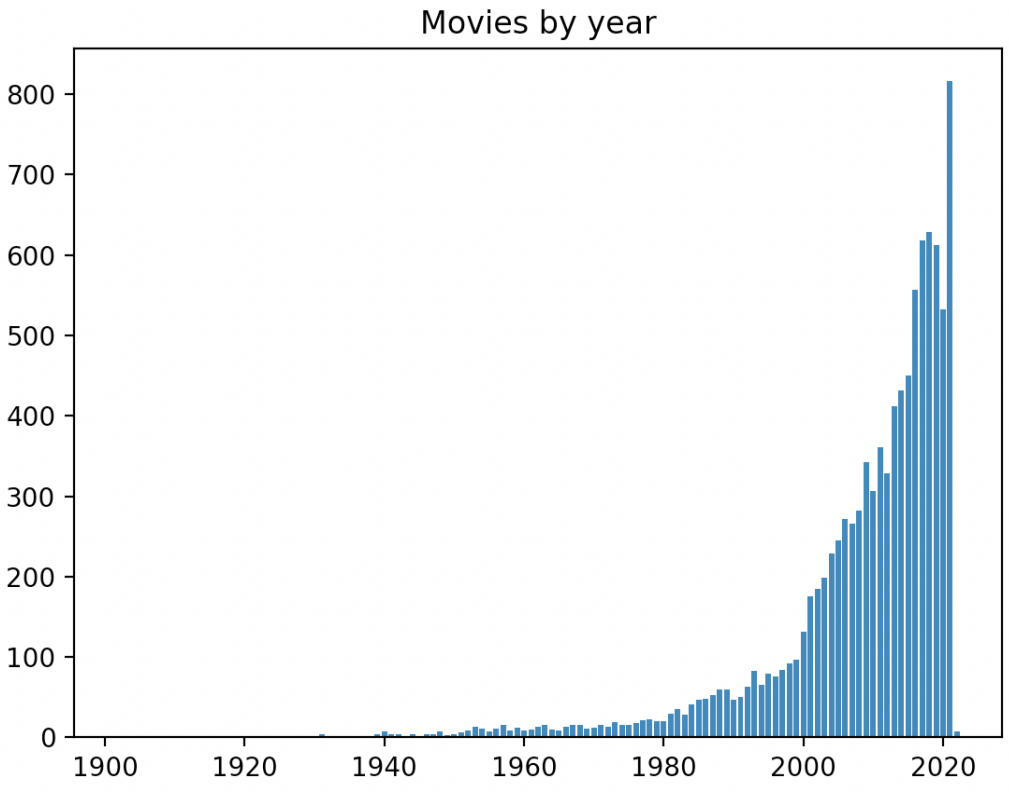
4.4. (3 puntos) ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios?

```
17    d = movies.nlargest(10, 'voteAvg')['title']
18    print(d)
19

PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL

alegudiel@Ales-MacBook-Pro analisis-exploratorio-1 % /Librar
k/Versions/3.9/bin/python3 /Users/alegudiel/Documents/Mineri
vies.py
8632                              Hot Naked Sex & the City
9084                                              Holidays
9246       Steven Universe: The Movie: Behind the Curtain
9298     Spirit of Vengeance: The Making of 'Ghost Rider'
9347    How Ponyo was Born ~Hayao Miyazaki's Thought P...
9733                                 Christmas at the Ranch
9875            El Chavo Del Ocho: Lo Mejor De Don Ramon
9990                              Los Vengadores Chiflados
6750                  Dragon Ball Kami BEST [Bonus DVD]
6885      The Spectacular Spider-Man Attack of the Lizard
```

4.5. (8 puntos) ¿Cuántas películas se hicieron en cada año?
¿En qué año se hicieron más películas? En 2021.

```
#4.5 ¿Cuántas películas se hicieron en cada año?
# ¿En qué año se hicieron más películas?
# Haga un gráfico de barras
movies['releaseDate'] = pd.to_datetime(movies['releaseDate'])
releaseDt = movies['releaseDate'].dt.year.value_counts()
#graph
ejeX = np.array(pd.value_counts(pd.to_datetime(movies.releaseDate).dt.year).keys())
ejeY = pd.value_counts(pd.to_datetime(movies.releaseDate).dt.year)
plt.bar(ejeX, ejeY)
plt.title("Movies by year")
plt.rcParams['figure.figsize'] = (10, 10)
print(releaseDt)
plt.show()
```

```
alegudiel@Ales-MacBook-Pro analisis-exploratori
thon.framework/Versions/3.9/bin/python3 /Users/
analisis-exploratorio-1/movies.py
2021    816
2018    629
2017    618
2019    612
2016    557
       ...
1922      1
1937      1
1929      1
1921      1
1932      1
Name: releaseDate, Length: 99, dtype: int64
```



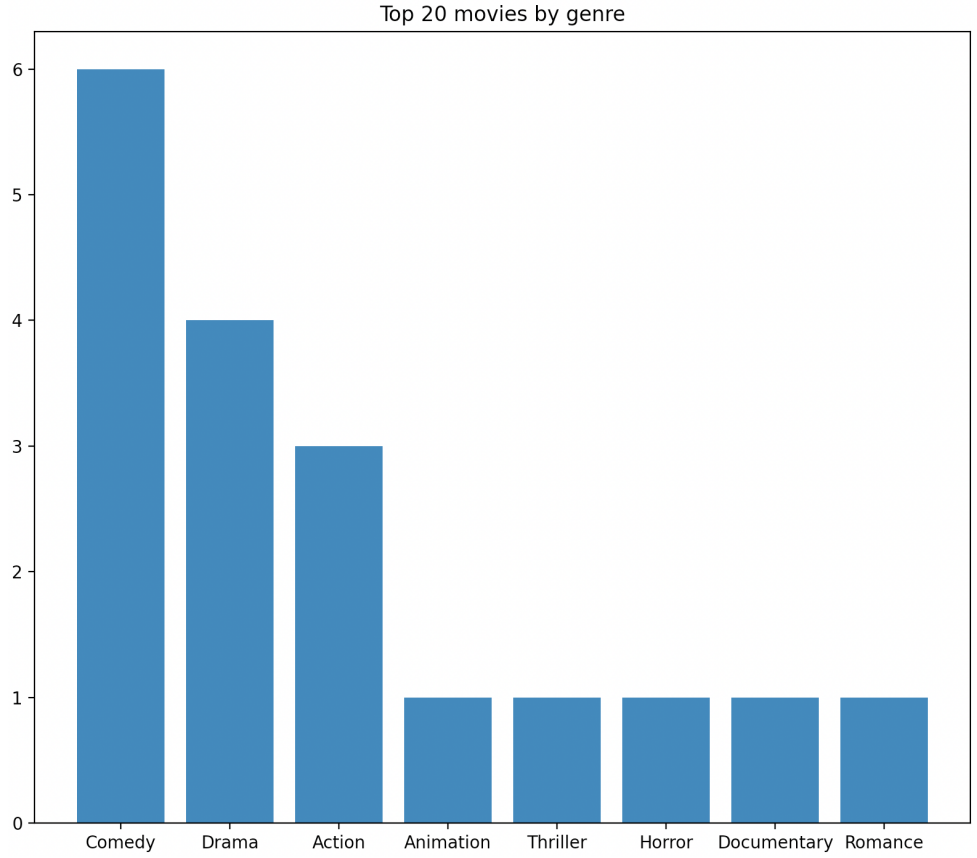
4.6. (9 puntos) ¿Cuál es el género principal de las 20 películas más recientes? ¿Cuál es el género principal que predomina en el conjunto de datos? Represéntelo usando un gráfico

```
37  #4.6 ¿Cuál es el género principal de las 20 películas más recientes?
38  # ¿Cuál es el género principal que predomina en el conjunto de datos?
39  # Represéntelo usando un gráfico
40  movies['genre1'] = movies['genres'].str.split('|', n=-1).str[0]
41  top_movies = movies.sort_values('releaseDate', ascending=False)[['title', 'genre1', 'releaseDate']].head(20)
42  print(top_movies)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

Python
Python

```
alegudiel@Ales-MacBook-Pro analisis-exploratorio-1 % /Library/Frameworks/Python.framework/Versions/3.9/bin/python
3 /Users/alegudiel/Documents/Mineria/analisis-exploratorio-1/movies.py
                                          title       genre1 releaseDate
9516                          A Shot Through the Wall       Drama  2022-01-21
9545                                       Presque      Comedy  2022-01-19
9197                                Italian Studies       Drama  2022-01-14
9808                                     See for Me    Thriller  2022-01-07
9586                                American Siege      Action  2022-01-07
9982  Harry Potter 20th Anniversary: Return to Hogwarts  Documentary  2022-01-01
9951                                    WWE Day 1 2022      Action  2022-01-01
9257                                Return of Chucky         NaN  2021-12-31
9241                               Karem, La Posesi⊡n         NaN  2021-12-31
9866                     Gabriel's Rapture: Part One     Romance  2021-12-31
9683                     Hilda and the Mountain King   Animation  2021-12-30
9877                                   The Kindred      Horror  2021-12-28
9988                                Death to 2021      Comedy  2021-12-27
9346                                        Lulli      Comedy  2021-12-26
9253                               American Underdog       Drama  2021-12-25
9996                                      El Paseo 6      Comedy  2021-12-25
9479                                    The ExorSIS      Comedy  2021-12-25
9560                                The Lost Girls       Drama  2021-12-25
9170                                  Minnal Murali      Action  2021-12-24
9872                  1000 Miles From Christmas      Comedy  2021-12-24
```



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

```
54  # 4.7 ¿Las películas de qué genero principal obtuvieron mayores ganancias?
55  topRevenue = movies.sort_values('revenue', ascending=False)[['title', 'genre1', 'revenue']].head(30)
56  print(topRevenue)
```

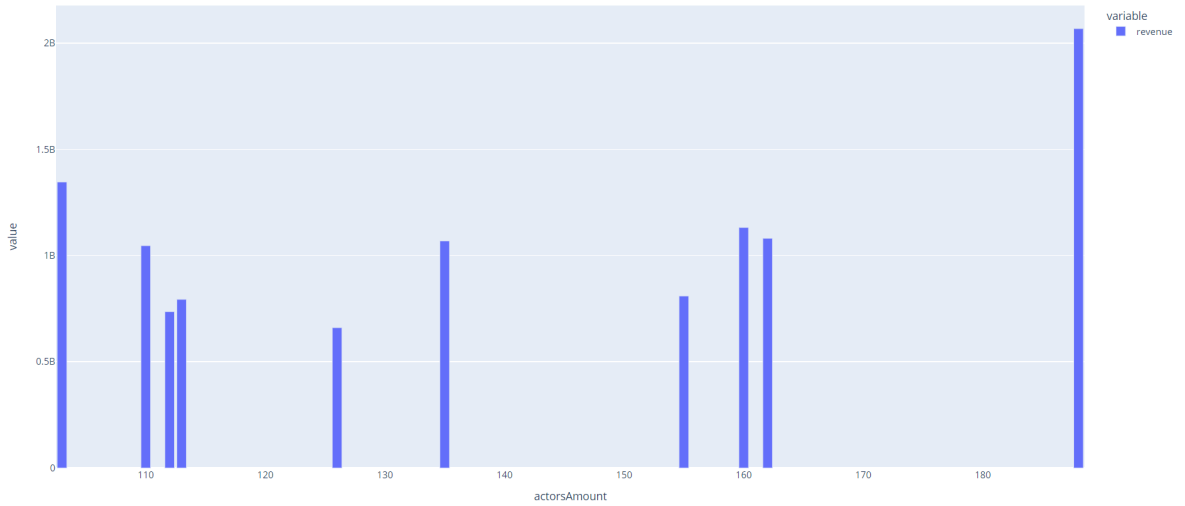
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

```
alegudiel@Ales-MacBook-Pro analisis-exploratorio-1 % /Library/Frameworks/Python.framework/Versions/3.9/bin/python
3 /Users/alegudiel/Documents/Mineria/analisis-exploratorio-1/movies.py
                                        title          genre1      revenue
3210                                   Avatar          Action  2.847246e+09
5952                         Avengers: Endgame       Adventure  2.797801e+09
307                                   Titanic           Drama  2.187464e+09
4947                 Star Wars: The Force Awakens          Action  2.068224e+09
5953                       Avengers: Infinity War       Adventure  2.046240e+09
4914                               Jurassic World          Action  1.671713e+09
7134                                The Lion King       Adventure  1.667635e+09
9049                      Spider-Man: No Way Home          Action  1.631853e+09
3397                                The Avengers  Science Fiction  1.518816e+09
5087                                   Furious 7          Action  1.515048e+09
6180                                   Frozen II          Family  1.450027e+09
4710                      Avengers: Age of Ultron          Action  1.405404e+09
5798                                Black Panther          Action  1.346739e+09
2509   Harry Potter and the Deathly Hallows: Part 2         Fantasy  1.341511e+09
5148                      Star Wars: The Last Jedi  Science Fiction  1.332699e+09
6428                 Jurassic World: Fallen Kingdom          Action  1.303460e+09
4765                                      Frozen       Animation  1.274219e+09
6108                         Beauty and the Beast          Family  1.263521e+09
5625                                Incredibles 2          Action  1.242805e+09
6271                         The Fate of the Furious          Action  1.238765e+09
4338                                    Iron Man 3          Action  1.214811e+09
5301                                      Minions          Family  1.156731e+09
5701                    Captain America: Civil War       Adventure  1.153296e+09
5939                                      Aquaman          Action  1.148462e+09
7239                     Spider-Man: Far From Home          Action  1.131928e+09
5954                                Captain Marvel          Action  1.128276e+09
3771                  Transformers: Dark of the Moon          Action  1.123794e+09
68    The Lord of the Rings: The Return of the King       Adventure  1.118889e+09
3735                                      Skyfall          Action  1.108561e+09
4665                 Transformers: Age of Extinction  Science Fiction  1.104000e+09
```

4.8. (3 puntos) ¿La cantidad de actores influye en los ingresos de las películas?¿se han hecho películas con más actores en los últimos años?

```
count     10000.000000
mean       2147.666600
std       37200.075802
min           0.000000
25%          13.000000
50%          21.000000
75%          36.000000
max      919590.000000
Name: actorsAmount, dtype: float64
```

Sacando la descripción de los datos de la cantidad de actores podemos ver que la media es de 2147.66

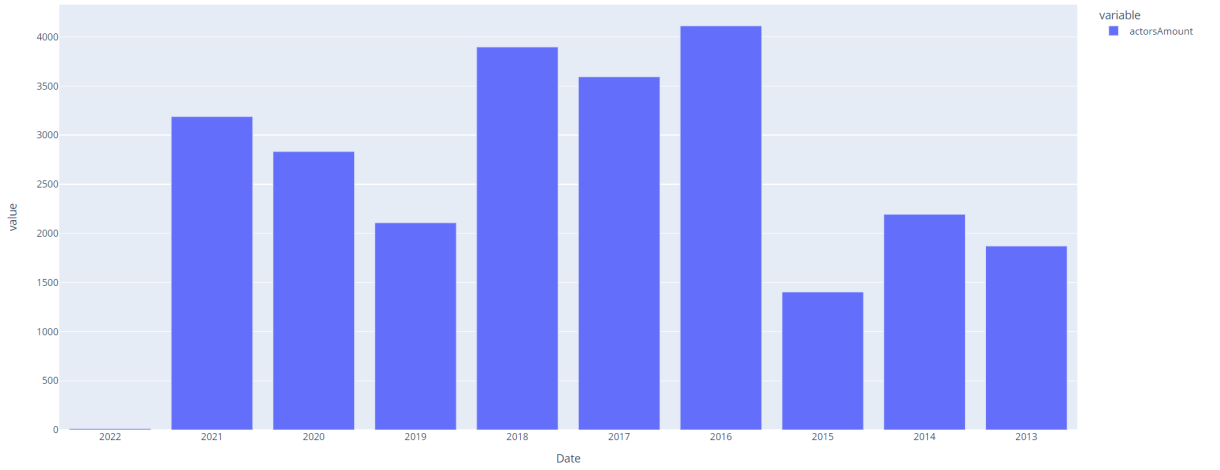


Obteniendo las 10 películas que mas ingresos tienen vemos que están en un rango de 100 a 188 actores

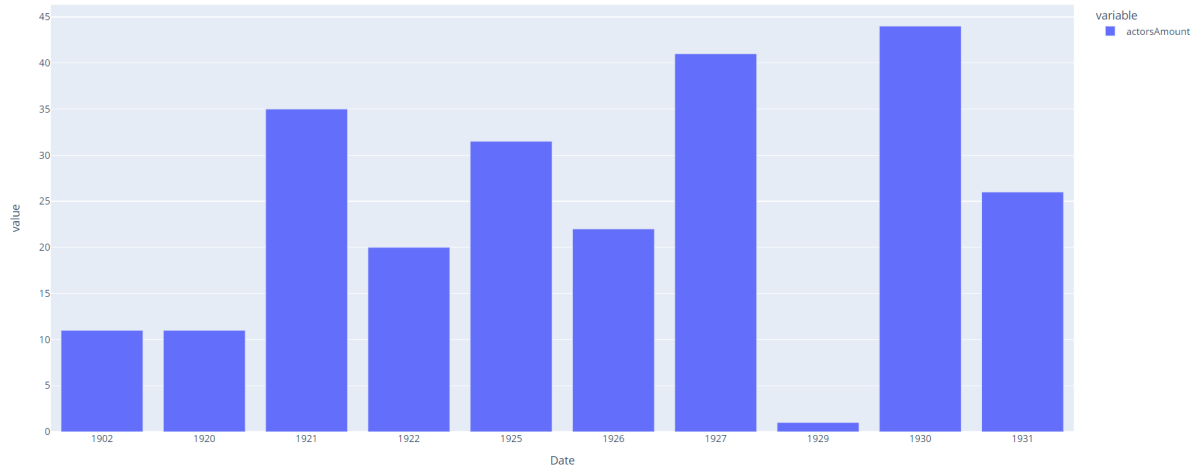
actorsAmount	
668297	2.700000e+01
2	1.305971e+05
1	1.412042e+05
3	1.589831e+05
5	7.755319e+05
0	2.734985e+06
4	2.872344e+06
6	2.963727e+06
174	7.200000e+06
8	7.669301e+06

Obteniendo las películas que menos ingresos tienen (diferentes de 0) podemos observar que un 80% están en un rango de 0-10 actores.

En esta gráfica se observa la cantidad de actores en los años más recientes (2013-2022)



En esta gráfica se observa la cantidad de actores en los años más antiguos (1902-1931)



Hay una clara diferencia entre el primer y el segundo rango, vemos que la cantidad de actores ha aumentado muchísimo. Sin embargo en estas graficas se toma el promedio de actores por ano por lo que es importante mencionar que que hayan mas actores en a;os recientes es en parte porque ahora se hacen muchas mas películas que antes.

title	Date	actorsAmount
Phantastische Harry Potter Momente - Das gro\ufffde...	2021	919590
Ben 10 Alien X-tinction	2021	882878
El Chavo Del Ocho: Lo Mejor De Don Ramon	2006	853253
Yumbina: La droga del sexo	2006	840964
Sex Education Films	2017	825491
El sexo me da risa 8	2018	815762
DRagON BALL P2 2wice dropda bbeet	2008	796822
Live Spectacle NARUTO ~Song of the Akatsuki~	2021	784594
The Roots of Wolverine: A Conversation with X-...	2009	765207
Dangerous Lesson	2020	750209
Hot Hair Salon	2020	748501
Game of Thrones: The Story So Far	2017	738562
Sexo e a Metr\ufffdpole	2004	687920
The Last Avatar	2014	668297
Holidays	2016	641026
Miraculous: Tales of Ladybug and Cat Noir - Gi...	2019	640103
Miraculous: Tales of Ladybug and Cat Noir - Pr...	2018	640100
Miraculous: Tales of Ladybug and Cat Noir: Lad...	2017	640097
The Scream	2019	634646
John Wick: Assassin's Code (Extra)	2015	619022

En esta tabla se muestran las 20 películas que más actores tienen, y observamos que de esas 20, 15 pertenecen al rango de 2013-2022 por lo que sí se puede afirmar que la cantidad de actores ha aumentado en los últimos años con respecto a muchos años atrás.

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?

Tabla que muestra las 10 películas con mejores ingresos y su cantidad de actores hombres y mujeres

revenue	castWomenAmount	castMenAmount
2.847246e+09	9	21
2.797801e+09	28	62
2.187464e+09	27	59
2.068224e+09	24	74
2.046240e+09	21	43
1.671713e+09	13	31
1.667635e+09	6	13
1.631853e+09	13	33
1.518816e+09	25	74
1.515048e+09	15	28

Tabla que muestra las 10 películas con menores ingresos y su cantidad de actores hombres y mujeres

revenue	castWomenAmount	castMenAmount
1.0	15	9
4.0	4	7
4.0	0	0
10.0	10	11
10.0	4	11
20.0	0	7
27.0	The Last Avatar	The Last Avatar
103.0	11	11
303.0	4	8
400.0	1	0
486.0	3	7

Tabla que muestra las 11 películas con menor popularidad y su cantidad de actores hombres y mujeres

popularity	castWomenAmount	castMenAmount
5.165	3	18
5.936	5	5
6.643	3	24618
6.781	4	3
6.839	3	9
7.074	5	7
7.164	6	13
7.242	3	6
7.269	6	25
7.526	8	26
7.545	16	36

Tabla que muestra las 11 películas con más popularidad y su cantidad de actores hombres y mujeres.

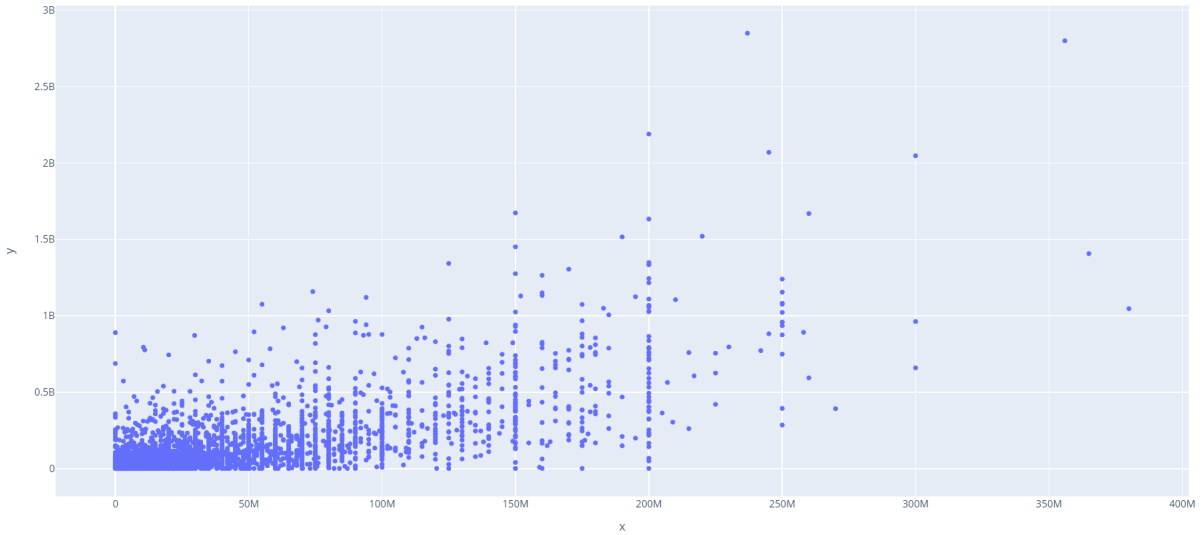
popularity	castWomenAmount	castMenAmount
11474.647	11	25
8443.740	13	33
6055.643	8	13
5887.379	9	10
5804.441	9	9
5051.222	12	18
3828.374	15	34
3062.764	11	26
2179.912	2	19
2066.867	13	22
1553.397	5	9

Observando los datos vemos que hay un aspecto general que tienen las 4 tablas y es que de 42 películas en 36 hay mayor cantidad de actores masculinos y solo en 1 hay mayor cantidad femenina.

4.10. (8 puntos) ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

voteAvg	director
10.0	Thomas Coven
9.0	Preston A. Whitmore II
8.9	Park Jun-soo
8.8	Taichi Ishidate
8.8	Amp Wong
8.7	Frank Darabont
8.7	Francis Ford Coppola
8.7	Aditya Chopra
8.6	Makoto Shinkai
8.6	MTJJ
8.6	Francis Ford Coppola
8.6	Steven Spielberg
8.5	Quentin Tarantino
8.5	Hideaki Anno
8.5	Christopher Nolan
8.5	Sergio Leone
8.5	Peter Jackson
8.5	Jon Watts
8.5	Morgan Spurlock
8.5	Robert Zemeckis

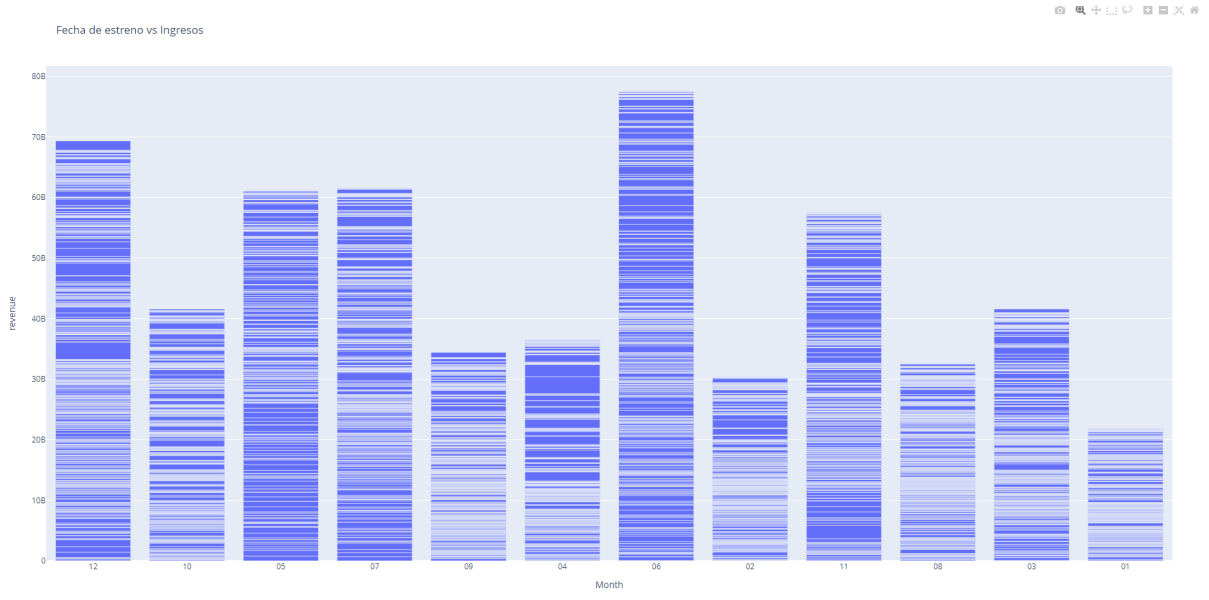
4.11. (8 puntos) ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión.

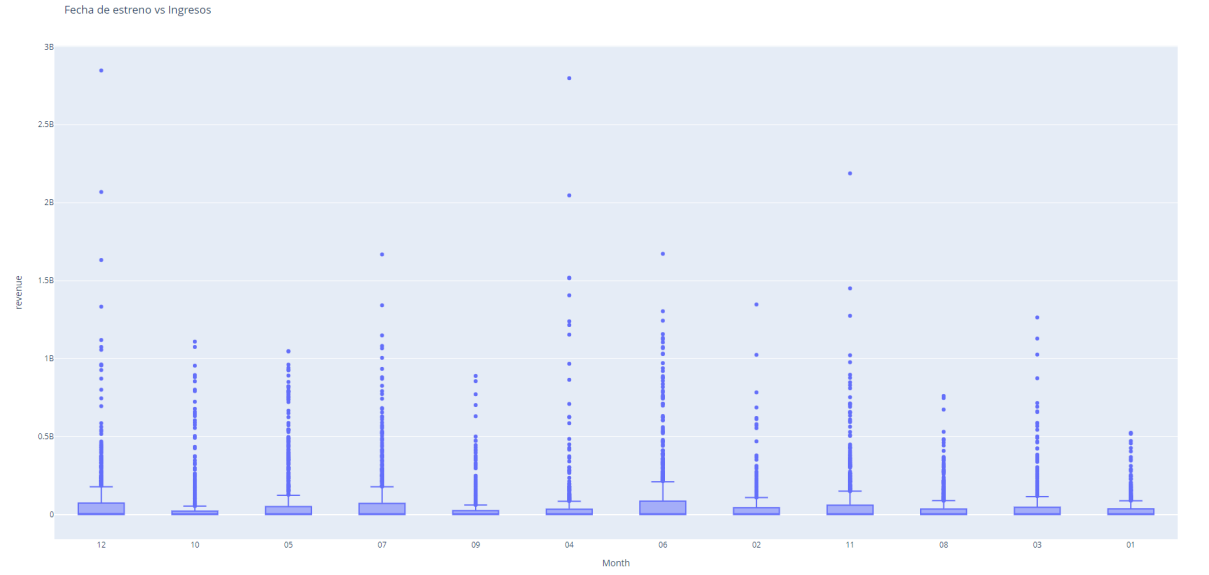


title	budget	revenue
Pirates of the Caribbean: On Stranger Tides	380000000	1.045714e+09
Avengers: Age of Ultron	365000000	1.405404e+09
Avengers: Endgame	356000000	2.797801e+09
Justice League	300000000	6.579243e+08
Avengers: Infinity War	300000000	2.046240e+09
Pirates of the Caribbean: At World's End	300000000	9.610000e+08
Superman Returns	270000000	3.910812e+08
The Lion King	260000000	1.667635e+09
Tangled	260000000	5.924617e+08
Spider-Man 3	258000000	8.908716e+08
The Dark Knight Rises	250000000	1.081041e+09
The Hobbit: The Desolation of Smaug	250000000	9.584000e+08
John Carter	250000000	2.841391e+08
Batman v Superman: Dawn of Justice	250000000	8.736349e+08
Captain America: Civil War	250000000	1.153296e+09
Star Wars: The Rise of Skywalker	250000000	1.074144e+09
The Fate of the Furious	250000000	1.238765e+09
Harry Potter and the Half-Blood Prince	250000000	9.339592e+08
Solo: A Star Wars Story	250000000	3.929524e+08
Harry Potter and the Deathly Hallows: Part 1	250000000	9.543059e+08

title	budget	revenue
Avatar	237000000	2.847246e+09
Avengers: Endgame	356000000	2.797801e+09
Titanic	200000000	2.187464e+09
Star Wars: The Force Awakens	245000000	2.068224e+09
Avengers: Infinity War	300000000	2.046240e+09
Jurassic World	150000000	1.671713e+09
The Lion King	260000000	1.667635e+09
Spider-Man: No Way Home	200000000	1.631853e+09
The Avengers	220000000	1.518816e+09
Furious 7	190000000	1.515048e+09
Frozen II	150000000	1.450027e+09
Avengers: Age of Ultron	365000000	1.405404e+09
Black Panther	200000000	1.346739e+09
Harry Potter and the Deathly Hallows: Part 2	125000000	1.341511e+09
Star Wars: The Last Jedi	200000000	1.332699e+09
Jurassic World: Fallen Kingdom	170000000	1.303460e+09
Frozen	150000000	1.274219e+09
Beauty and the Beast	160000000	1.263521e+09
Incredibles 2	200000000	1.242805e+09
The Fate of the Furious	250000000	1.238765e+09

4.12. (7 puntos) ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?

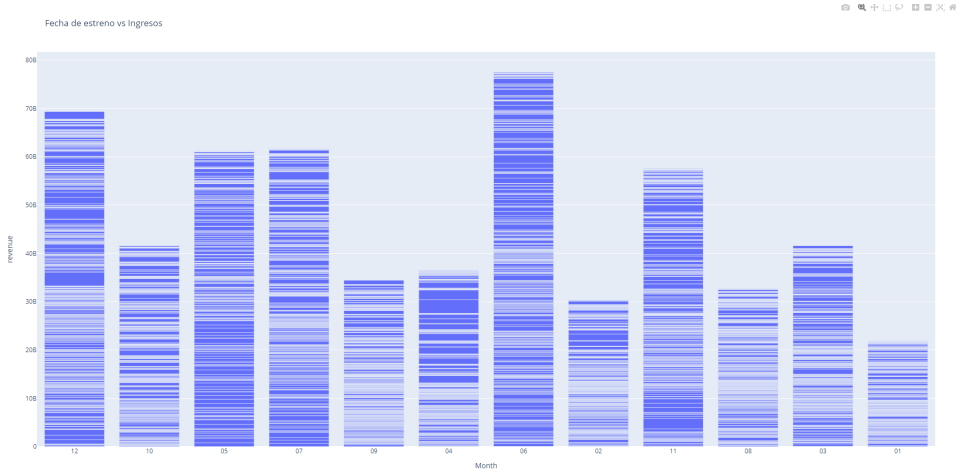




4.12 En la primera gráfica de barras se puede observar que los meses que predominaban con mayores ingresos eran diciembre, junio, mayo y julio. Luego en la gráfica de cajas y bigotes se puede observar si hay algún sesgo en los ingresos que tuvieron las películas y se puede observar que de igual manera estos meses siguen teniendo los mejores ingresos.

4.13. (8 puntos) ¿En qué meses se han visto los lanzamientos con mejores ingresos?¿Cuántas películas, en promedio, se han lanzado por mes?

En la siguiente gráfica de barras se puede notar que los meses con mayores ingresos suelen ser diciembre, junio y los que les siguen son mayo y julio.

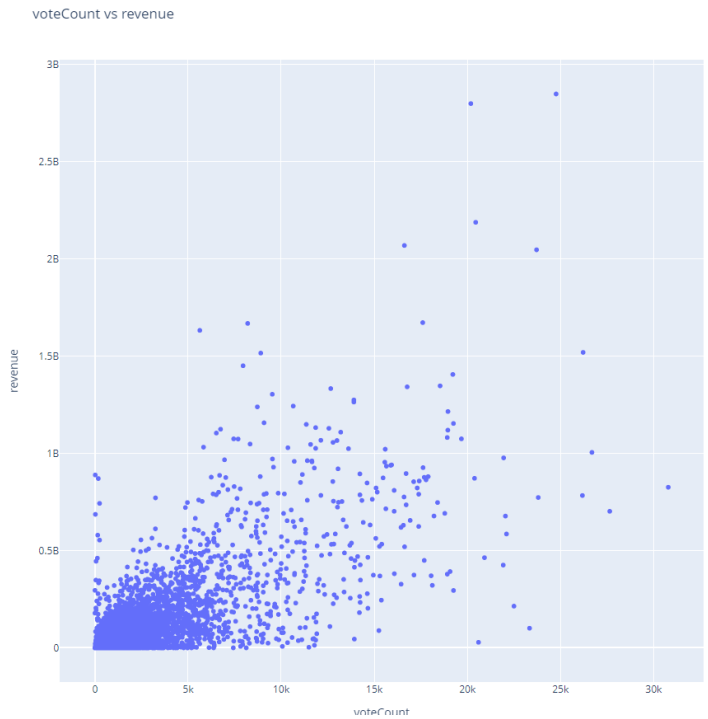


Aquí se muestra la media, el minimo y maximo de la cantidad de estrenos que se dan por mes, siendo septiembre el mes con más estrenos en promedio y abril el que tiene en promedio menos estrenos.

```
             mean min max
month
1       11.642857   1  48
2       11.031250   1  64
3       14.051724   1  66
4       10.545455   1  63
5       11.633333   1  67
6       11.219178   1  81
7       12.492308   1  82
8       14.265625   1  74
9       18.288136   1  80
10      15.257143   1  91
11      12.809524   1  68
12      11.987179   1  74
PS C:\Users\joopa\Desktop\hdt1>
```

4.14. (7 puntos) ¿Cómo se correlacionan las calificaciones con el éxito comercial?

Al analizar el diagrama de dispersión se nota una correlación positiva débil entre la cantidad de ingreso que representa si tuvo éxito y la cantidad de votos que obtuvo, de esta manera se puede confirmar que a mayor cantidad de votos se observa una tendencia de crecer en la cantidad de ingresos que recibió.



4.15. (5 puntos) ¿A qué género principal pertenecen las películas más largas?

Al obtener las 100 películas más largas, se puede notar que el género más dominante fue el de drama con 36 películas y aventura y acción le siguen con 12 películas, aunque representan muy poco a comparación de la cantidad de películas que tuvo el drama.