Universidad del Valle de Guatemala
Minería de Datos
Ing. Lynette García

Hoja de Trabajo 1.

Análisis Exploratorio

Estudiantes:
Oscar Juárez -17315
Rodrigo Samayoa - 17332
María Fernanda Estrada - 14198

Guatemala, 13 de febrero de 2020

1. **Resumen del conjunto de datos**

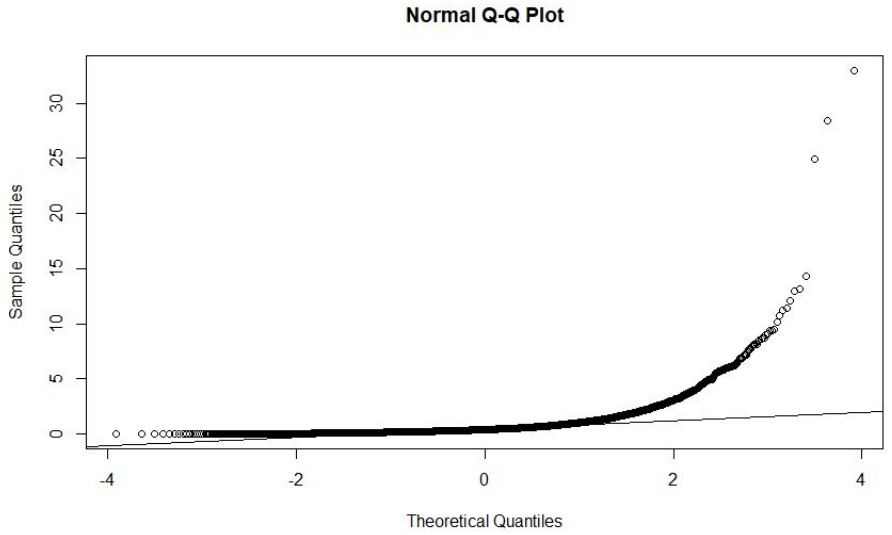
El conjunto de datos posee variables que detallan todas las características de las películas dentro del CSV. Con un total de 10,866 películas, el conjunto de datos ofrece información como la popularidad, el elenco de la película, el director, género, fecha de lanzamiento, etc.

2. **Tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)**

- *Id*: cualitativa nominal
- *imdb_id*: cualitativa nominal
- *popularity*: cuantitativa continua
- *budget*: cuantitativa discreta
- *revenue*: cuantitativa continua
- *original_title*: cualitativa nominal
- *cast*: cualitativa nominal
- *homepage*: cualitativa nominal
- *director*: cualitativa nominal
- *tagline*: cualitativa nominal
- *keywords*: cualitativa nominal
- *overview*: cualitativa nominal
- *runtime*: cuantitativa discreta
- *genres*: cualitativa nominal
- *production_companies*: cualitativa nominal
- *release_date*: cualitativa ordinal
- *vote_count*: cuantitativa discreta
- *vote_average*: cuantitativa continua
- *release_year*: cualitativa ordinal
- *Budget_adj:* Cuantitativa continua
- *Revenue_adj:* Cuantitativa continua

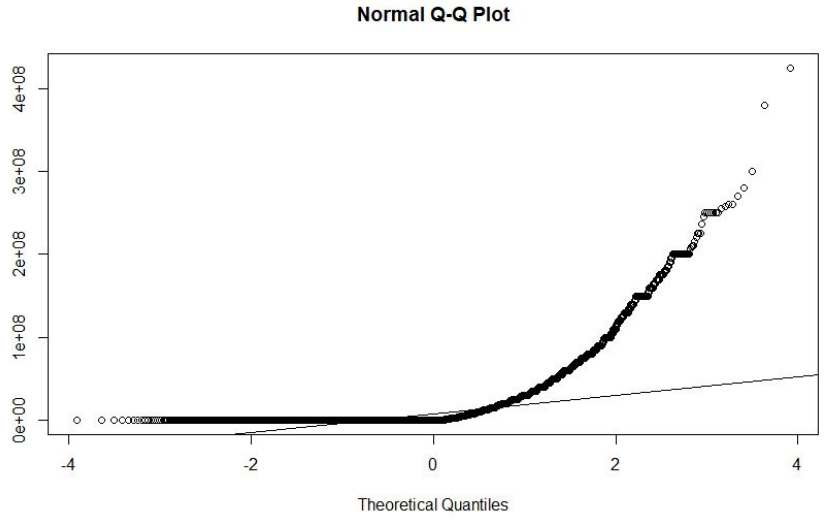
3. **¿Las variables cuantitativas siguen una distribución normal?**

3.1. *Popularidad:* Según el gráfico cuantil-cuantil, los datos no siguen una distribución estándar. Además, se acepta la hipótesis nula del test de Anderson-Darling, por lo que los datos no siguen una distribución estándar.



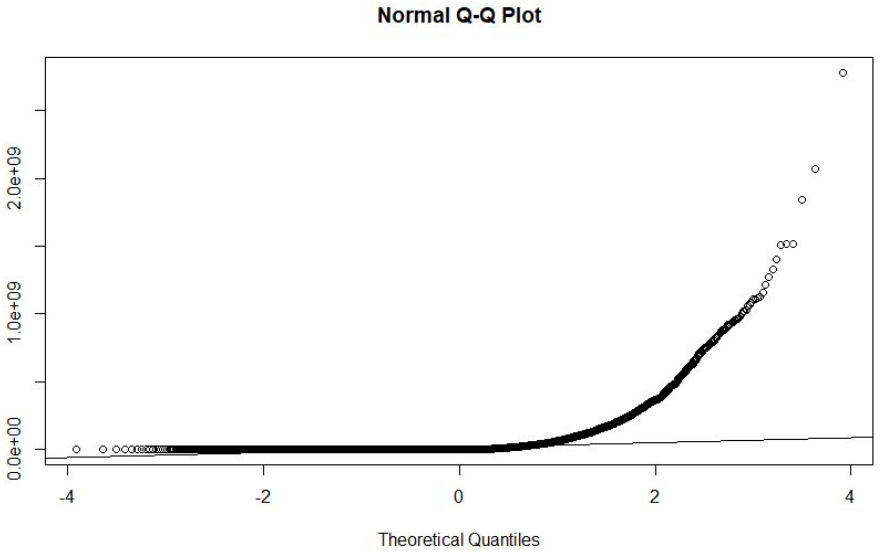
Resultado del test AD: A = 1338.9, p-value < 2.2e-16

3.2. *Presupuesto:* En base a la gráfica cuantil-cuantil y el hecho que se acepta la hipótesis nula del test de normalidad de Anderson-Darling, los datos no siguen una distribución estándar.



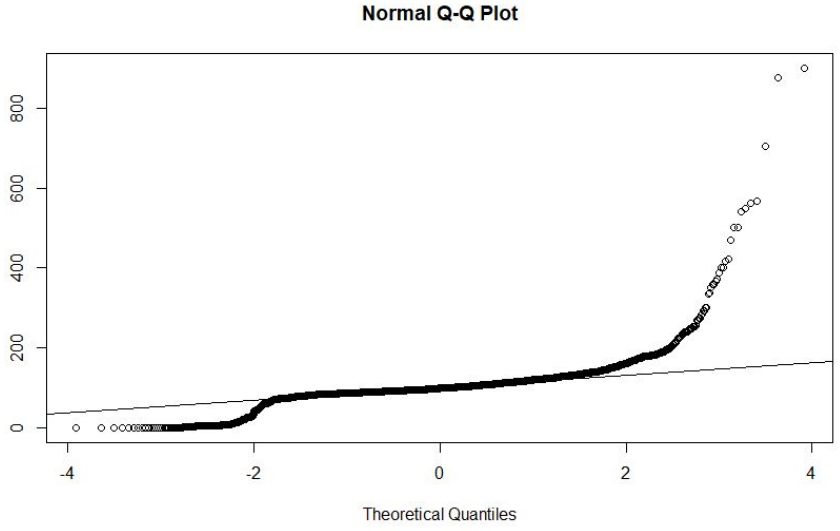
Resultado del test AD: A = 1722.6, p-value < 2.2e-16

3.3. *Ingreso:* En base a la gráfica cuantil-cuantil y el hecho que se acepta la hipótesis nula del test de normalidad de Anderson-Darling, los datos no siguen una distribución estándar.



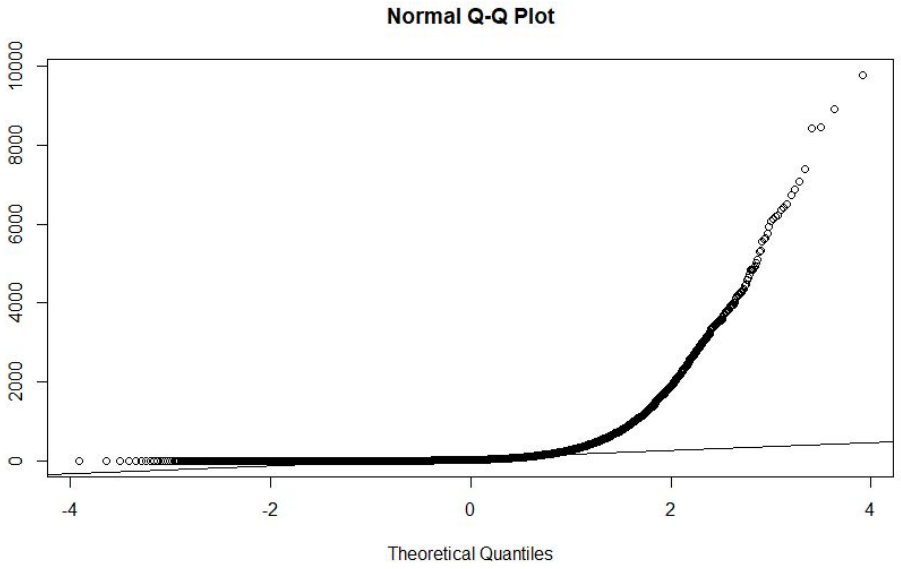
Resultado del test AD: A = 2299.7, p-value < 2.2e-16

3.4. *Duración:* En base a la gráfica cuantil-cuantil y el hecho que se acepta la hipótesis nula del test de normalidad de Anderson-Darling, los datos no siguen una distribución estándar.



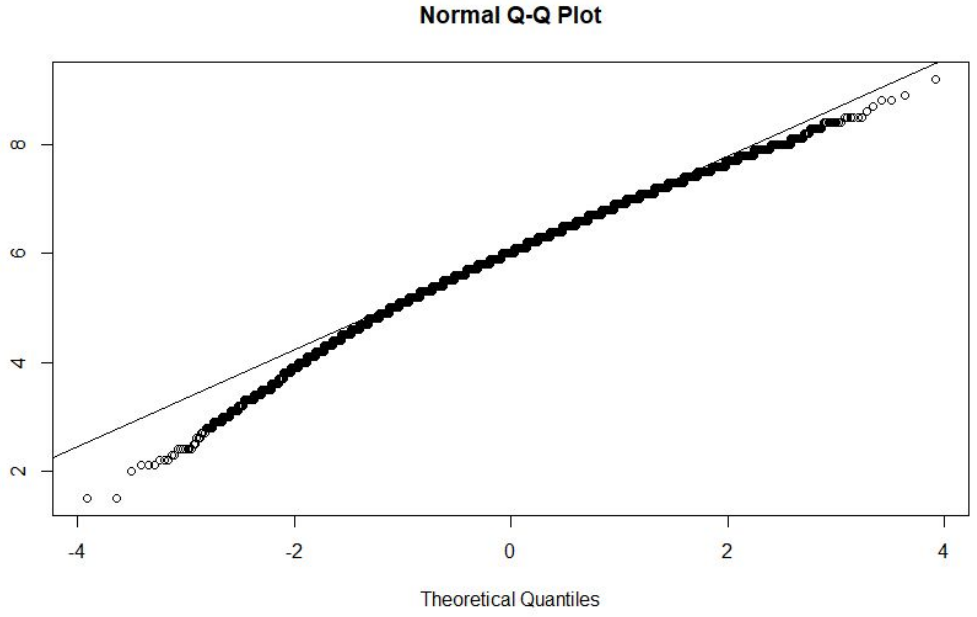
Resultado del test AD: A = 666.81, p-value < 2.2e-16

3.5. *Votos:* En base a la gráfica cuantil-cuantil y el hecho que se acepta la hipótesis nula del test de normalidad de Anderson-Darling, los datos no siguen una distribución estándar.



Resultado del test AD: A = 24.024, p-value < 2.2e-16

3.6. *Votos Promedio:* Dado el resultado de la gráfica cuantil-cuantil y el histograma del gráfico de votos promedio, se puede interpretar que los datos sí siguen una distribución normal.



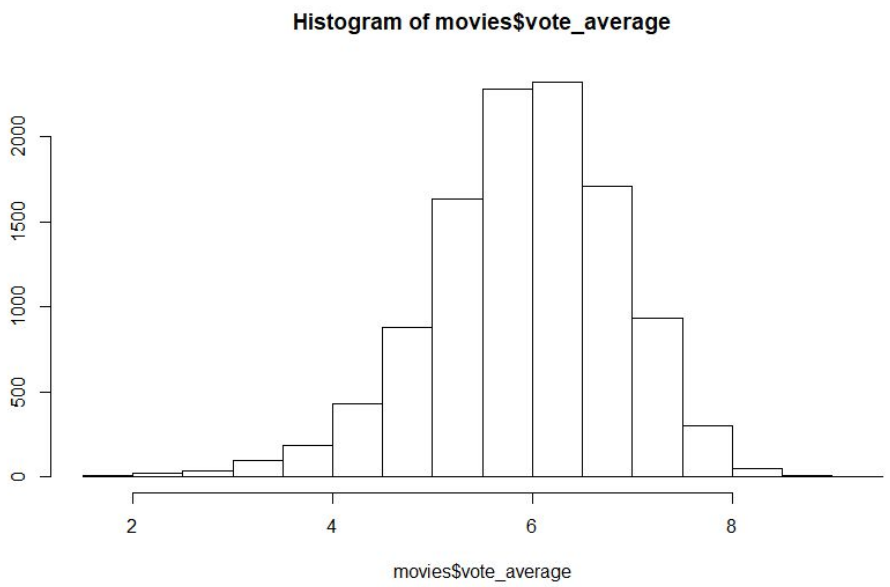
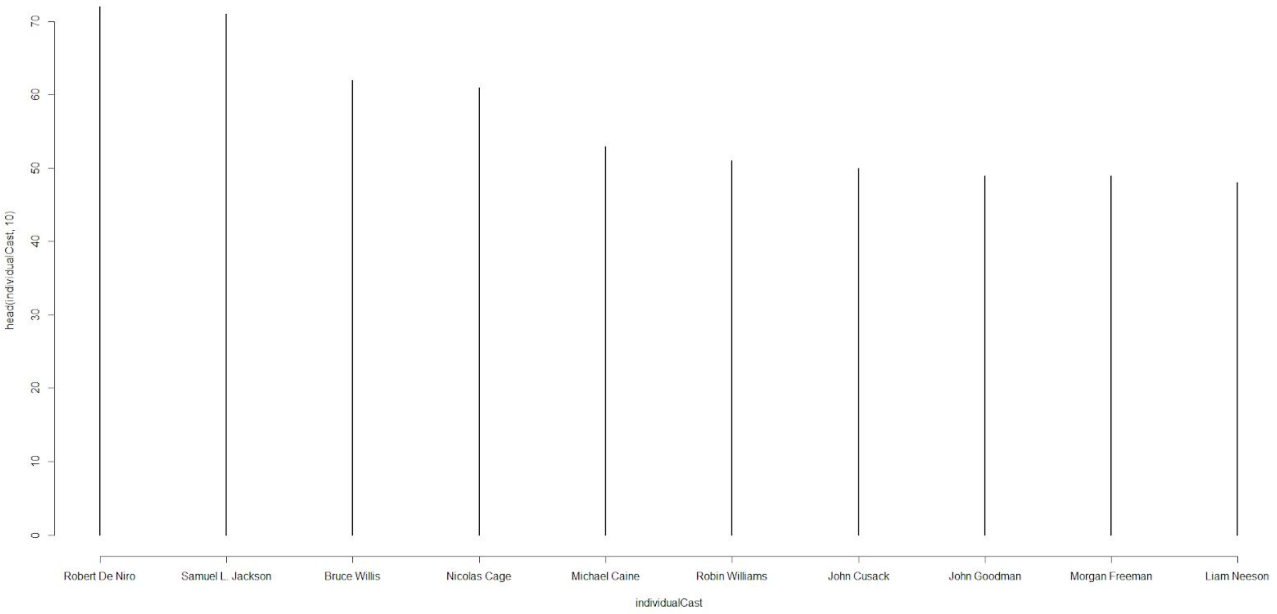
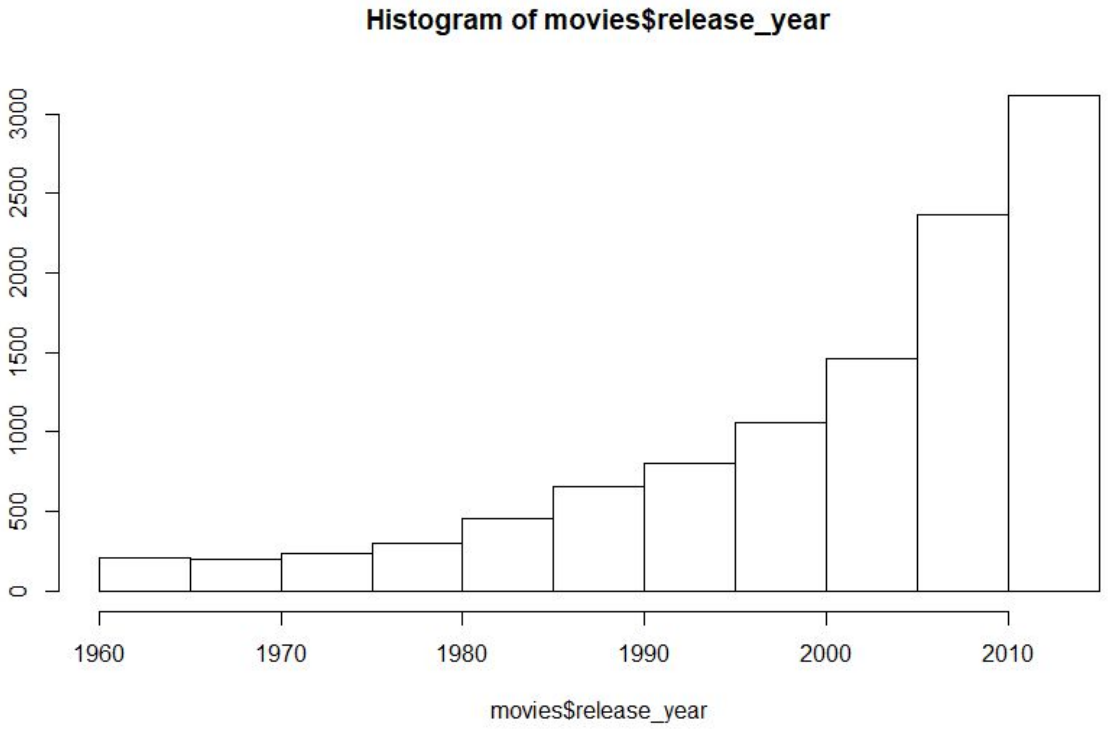

Histogram of movies$vote_average
0
500
1000
1500
2000
2
4
6
8
movies$vote_average

Tabla de frecuencias de las variables cualitativas

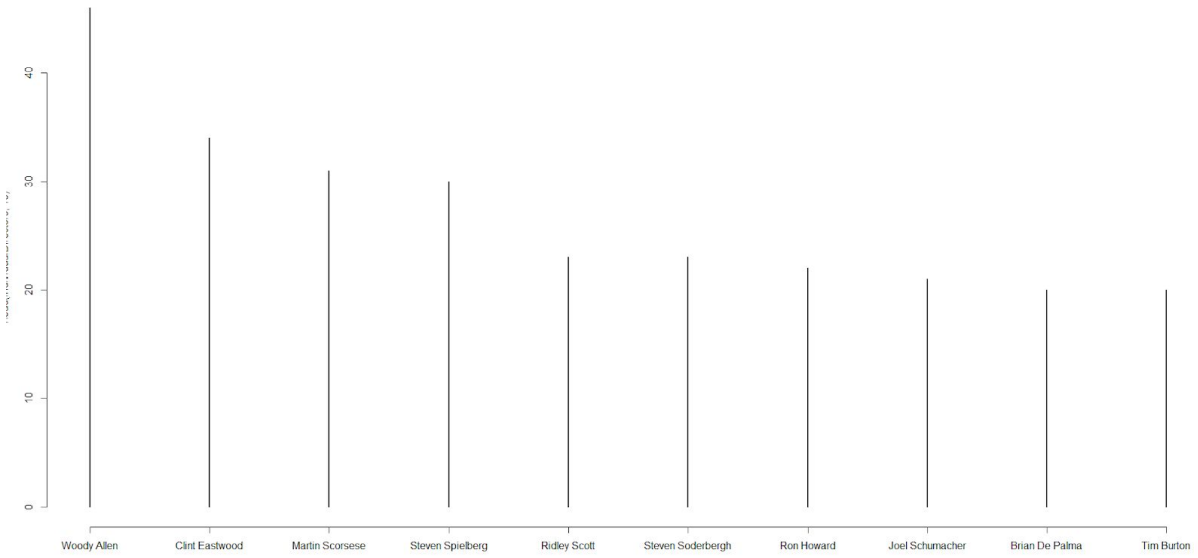
Actores:



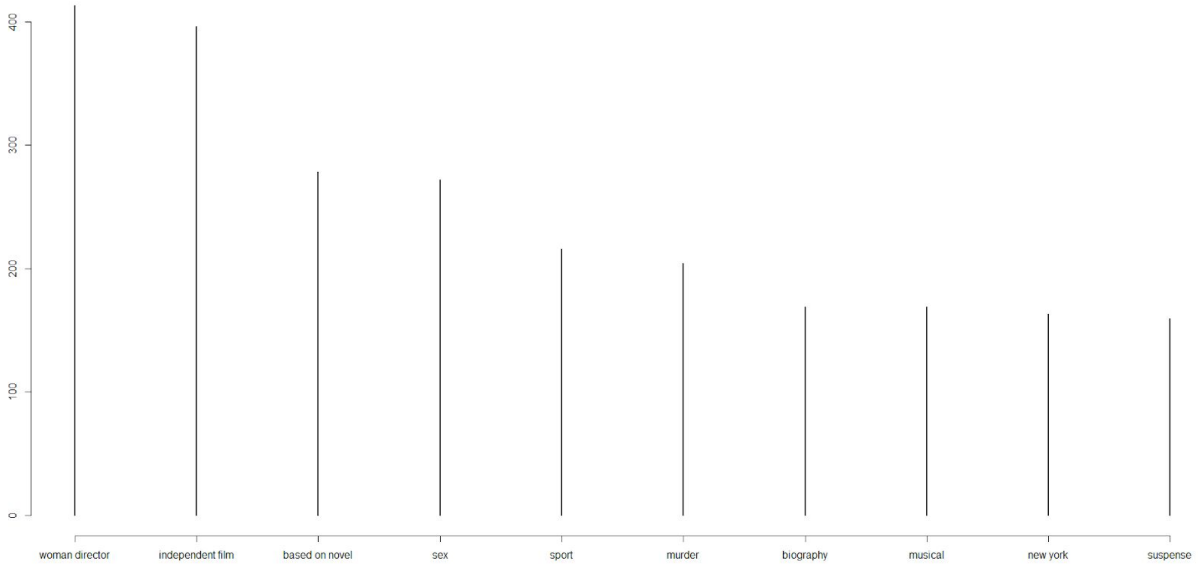
Año de estreno:



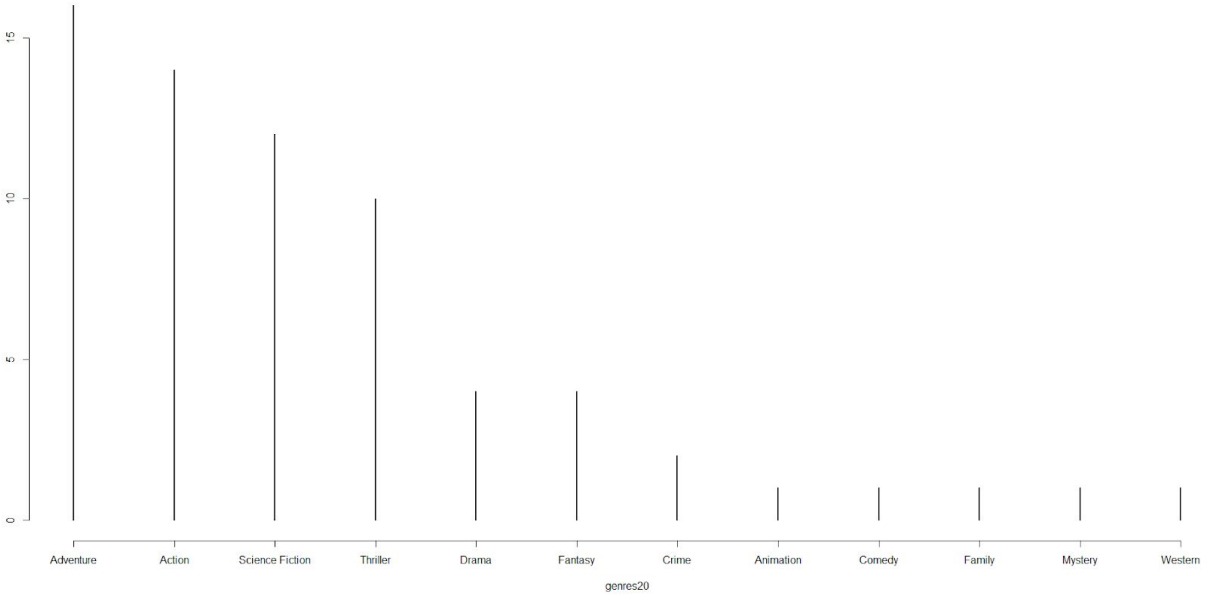
Director:



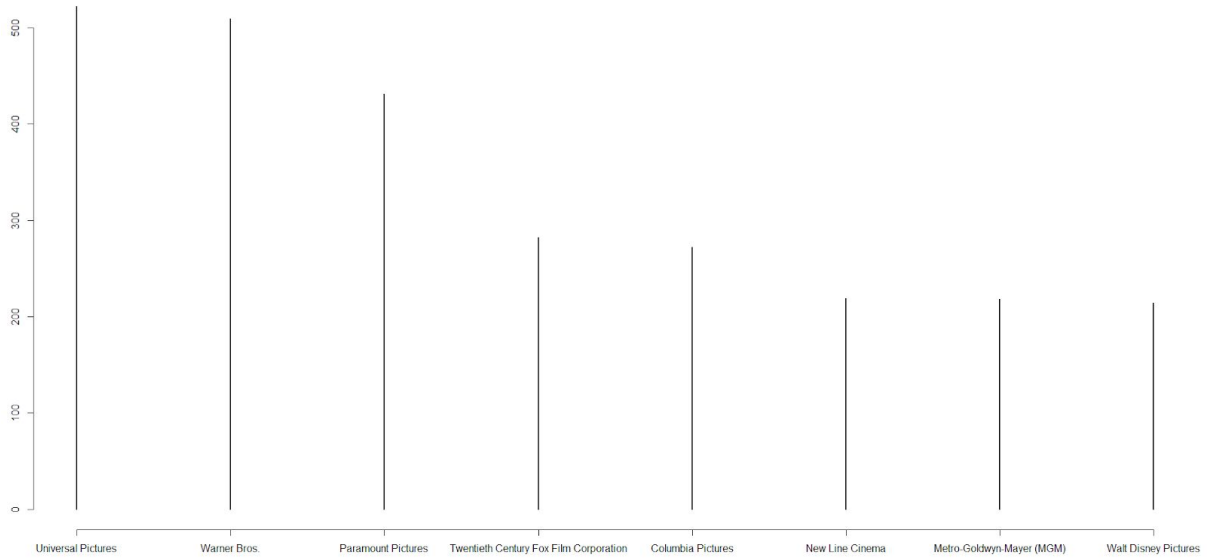
Palabras clave:



Géneros



Compañías de producción



4. Preguntas varias

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

Ordenadas de mayor a menor...

	id	imdb_id	popularity	budget	revenue	original_title	cast
2245	46528	tt1032751	0.250540	4.25e+08	11087569	The Warrior's Way	Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Danny Huston...
3376	1865	tt1298650	4.955130	3.80e+08	1021683000	Pirates of the Caribbean: On Stranger Tides	Johnny Depp|PenÃƒÂ©lope Cruz|Geoffrey Rush|Ian McShan...
7388	285	tt0449088	4.965391	3.00e+08	961000000	Pirates of the Caribbean: At World's End	Johnny Depp|Orlando Bloom|Keira Knightley|Geoffrey Rush|...
15	99861	tt2395427	5.944927	2.80e+08	1405035767	Avengers: Age of Ultron	Robert Downey Jr.|Chris Hemsworth|Mark Ruffalo|Chris Evan...
6571	1452	tt0348150	1.957331	2.70e+08	391081192	Superman Returns	Brandon Routh|Kevin Spacey|Kate Bosworth|James Marsden|...
1930	38757	tt0398286	2.865684	2.60e+08	591794936	Tangled	Zachary Levi|Mandy Moore|Donna Murphy|Ron Perlman|M....
4412	49529	tt0401729	1.588457	2.60e+08	284139100	John Carter	Taylor Kitsch|Lynn Collins|Mark Strong|Willem Dafoe|CiarÃƒ...
7395	559	tt0413300	2.520912	2.58e+08	890871626	Spider-Man 3	Tobey Maguire|Kirsten Dunst|James Franco|Thomas Haden ...
5509	57201	tt1210819	1.214510	2.55e+08	89289910	The Lone Ranger	Johnny Depp|Armie Hammer|William Fichtner|Helena Bonha...
635	122917	tt2310332	10.174599	2.50e+08	955119788	The Hobbit: The Battle of the Five Armies	Martin Freeman|Ian McKellen|Richard Armitage|Ken Stott|Gr...

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

Ordenadas de mayor a menor…

	id	imdb_id	popularity	budget	revenue	original_title
1387	19995	tt0499549	9.432768	2.37e+08	2781505847	Avatar
4	140607	tt2488496	11.173104	2.00e+08	2068178225	Star Wars: The Force Awakens
5232	597	tt0120338	4.355219	2.00e+08	1845034188	Titanic
4362	24428	tt0848228	7.637767	2.20e+08	1519557910	The Avengers
1	135397	tt0369610	32.985763	1.50e+08	1513528810	Jurassic World
5	168259	tt2820852	9.335014	1.90e+08	1506249360	Furious 7
15	99861	tt2395427	5.944927	2.80e+08	1405035767	Avengers: Age of Ultron
3375	12445	tt1201607	5.711315	1.25e+08	1327817822	Harry Potter and the Deathly Hallows: Part 2
5423	109445	tt2294629	6.112766	1.50e+08	1274219009	Frozen
5426	68721	tt1300854	4.946136	2.00e+08	1215439994	Iron Man 3

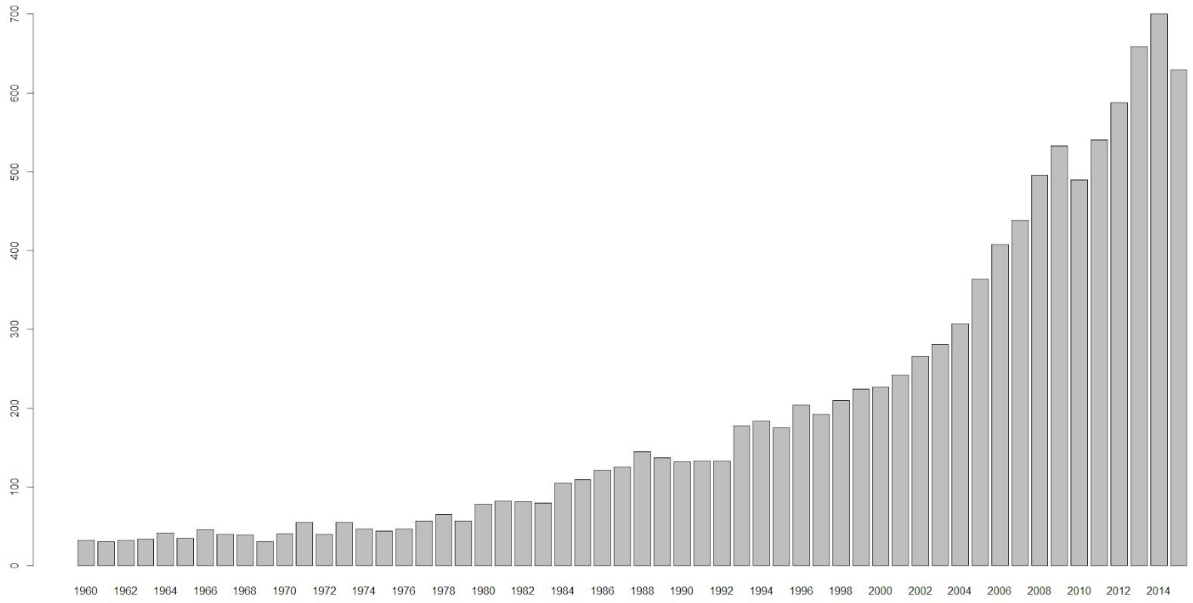
4.3. ¿Cuál es la película que más votos tuvo?

	id	imdb_id	popularity	budget	revenue	original_title
1920	27205	tt1375666	9.363643	1.6e+08	825500000	Inception

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

	id	imdb_id	popularity	budget	revenue	original_title
7773	25055	tt0960835	0.12112	0	0	Transmorphers

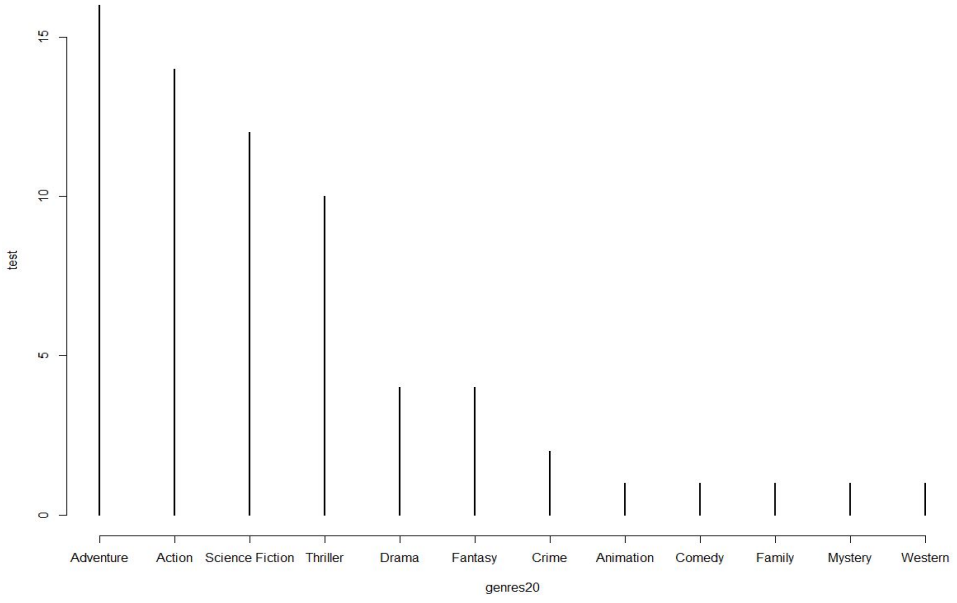
4.5. ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras



Se hicieron más películas en el año 2014.

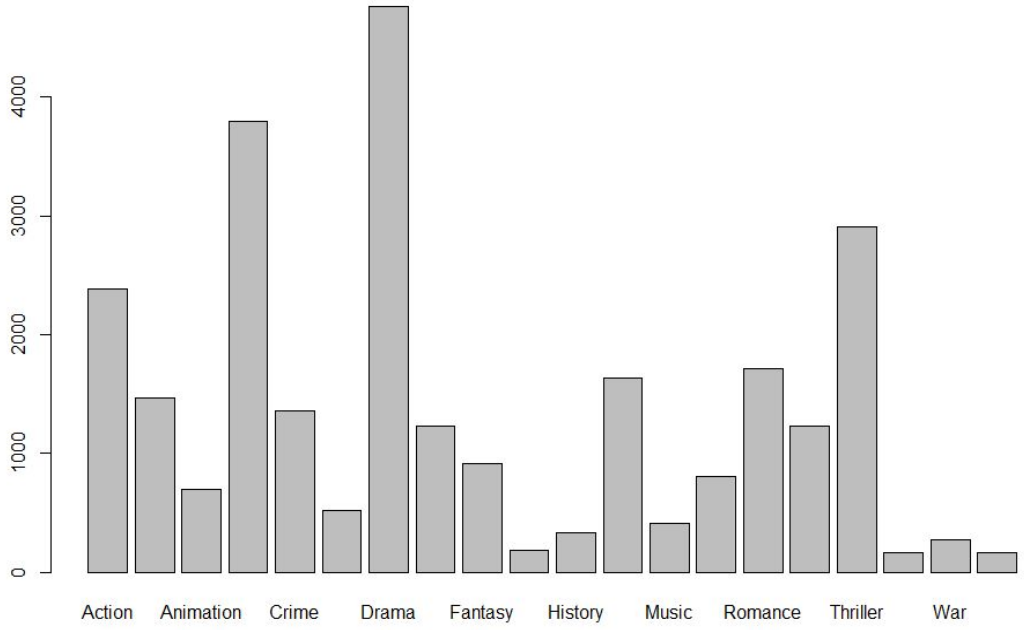
4.6. ¿Cuál es el género principal de las 20 películas más populares?

Para obtener el género principal, se utilizó una función para calcular la moda de los datos. Sin embargo, cada una de éstas 20 películas contaba con más de un género. Después de separar cada género en un ítem diferente y convertir la lista en un vector, se encontró que el género principal de las 20 películas más populares es "**Aventura**".



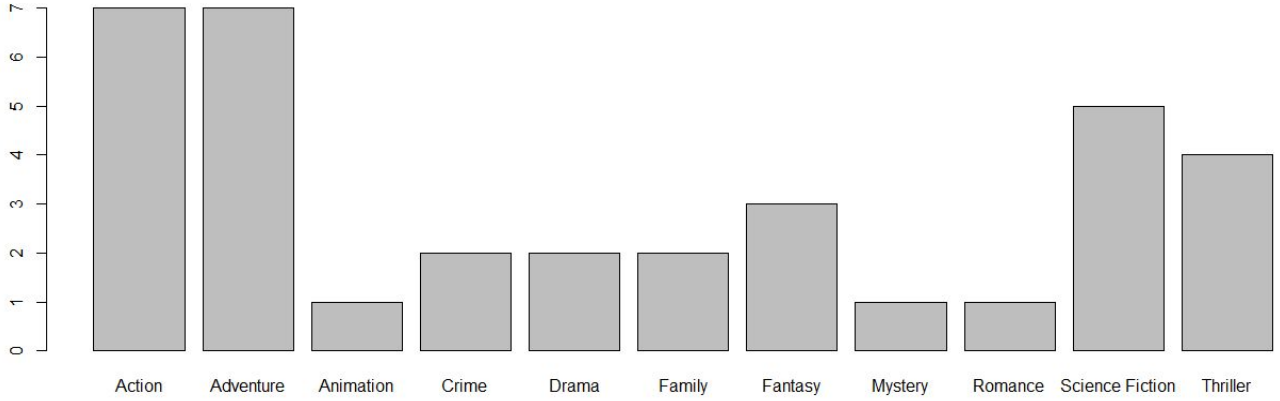
4.7. ¿Cuál es el género que predomina en el conjunto de datos? Represéntelo usando un gráfico

El género que predomina en el conjunto de datos es “**Drama**”.



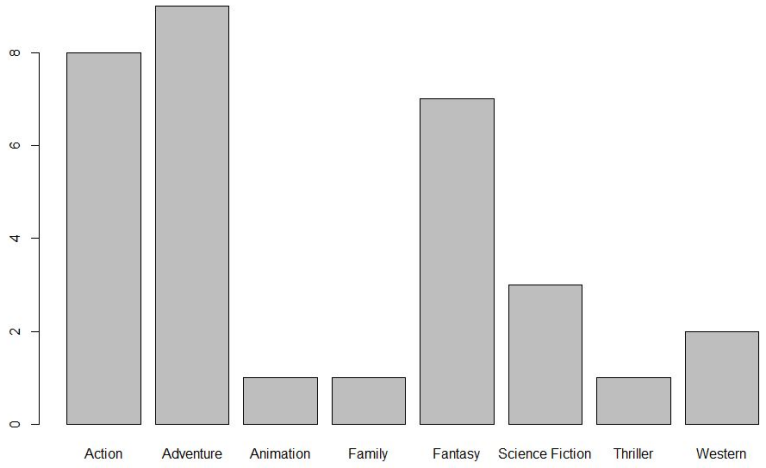
4.8. ¿Las películas de qué género principal obtuvieron mayores ganancias?

Para este inciso se toman en cuenta las ganancias (revenue - budget) y no el revenue como tal. Para ello, se restan ambas columnas y se crea una nueva con el resultado. Ahora con este nuevo subconjunto del dataset, se encontró que los géneros que tuvieron mayores ganancias fueron “**Acción**” y “**Aventura**”.



4.9. ¿Las películas de qué género principal necesitaron mayor presupuesto?

Similar al inciso 4.6, pero en lugar de usar todo el dataset, se utilizaron las 10 películas con más presupuesto (inciso 4.1). Las películas principalmente de "**Aventura**" necesitaron mayor presupuesto.



4.10. ¿Quiénes son los 20 mejores directores que hicieron películas altamente calificadas?

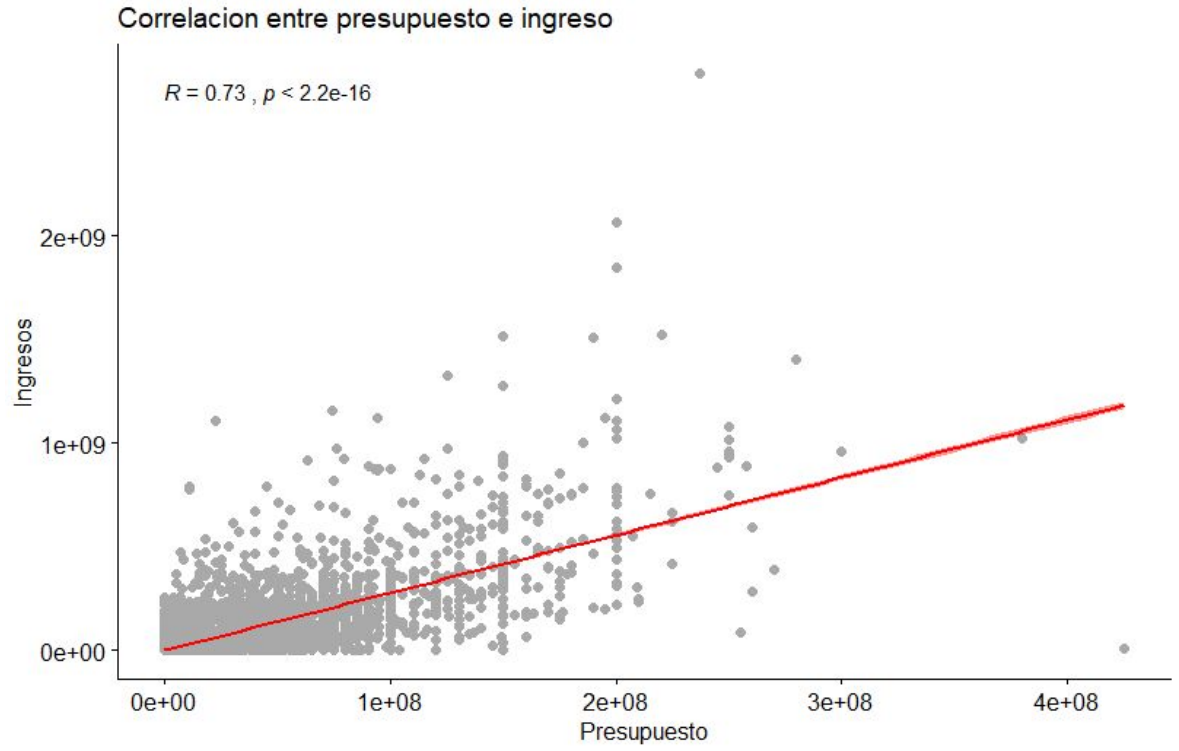
Ordenadas de mayor a menor…

	director	original_title	vote_average
3895	Mark Cousins	The Story of Film: An Odyssey	9.2
539	Jennifer Siebel Newsom	The Mask You Live In	8.9
1201	Carl Tibbetts	Black Mirror: White Christmas	8.8
2270	Derek Frankowski	Life Cycles	8.8
6912	David Mallet	Pink Floyd: Pulse	8.7
3691	Curt Morgan	The Art of Flight	8.5
5831	James Payne	Doctor Who: The Time of the Doctor	8.5
8222	Martin Scorsese|Michael Henry Wilson	A Personal Journey with Martin Scorsese Through American ...	8.5
8412	Saul Swimmer	Queen - Rock Montreal	8.5
8840	Stan Lathan	Dave Chappelle: Killin' Them Softly	8.5
610	Andrew Jarecki	The Jinx: The Life and Deaths of Robert Durst	8.4
2335	Sam Dunn|Scot McFadyen	Rush: Beyond the Lighted Stage	8.4
4179	Frank Darabont	The Shawshank Redemption	8.4
5924	Anthony Mandler	Tropico	8.4
5987	Jorge RamÃƒÂrez SuÃƒÂ¡rez	Guten Tag, RamÃƒÂ³n	8.4
7949	Jonathan Demme	Stop Making Sense	8.4
8371	Chris Bould	Bill Hicks: Relentless	8.4

9291	D.A. Pennebaker|David Dawkins|Chris Hegedus	Depeche Mode: 101	8.4
1323	Paul Dugdale	One Direction: Where We Are - The Concert	8.3
1865	Sam Dunn|Scot McFadyen	Iron Maiden: Flight 666	8.3

4.11. ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos?

Existe una fuerte relación entre el presupuesto y el ingreso de la película; mientras más presupuesto tiene, mayores ingresos puede obtener. No es una correlación perfecta, pero el coeficiente de 0.73 muestra una fuerte relación lineal ascendente. Los datos atípicos pueden afectar este resultado.



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

	Group.1	x
1	1	12968479332
2	2	19793785507
3	3	31395913991
4	4	26393260672
5	5	50454865815
6	6	61660585217
7	7	45382225481
8	8	25477766900
9	9	25731480293
10	10	29353709677
11	11	45896256786
12	12	58211863204

Los meses que más ingresos han tenido son junio, diciembre y mayo.

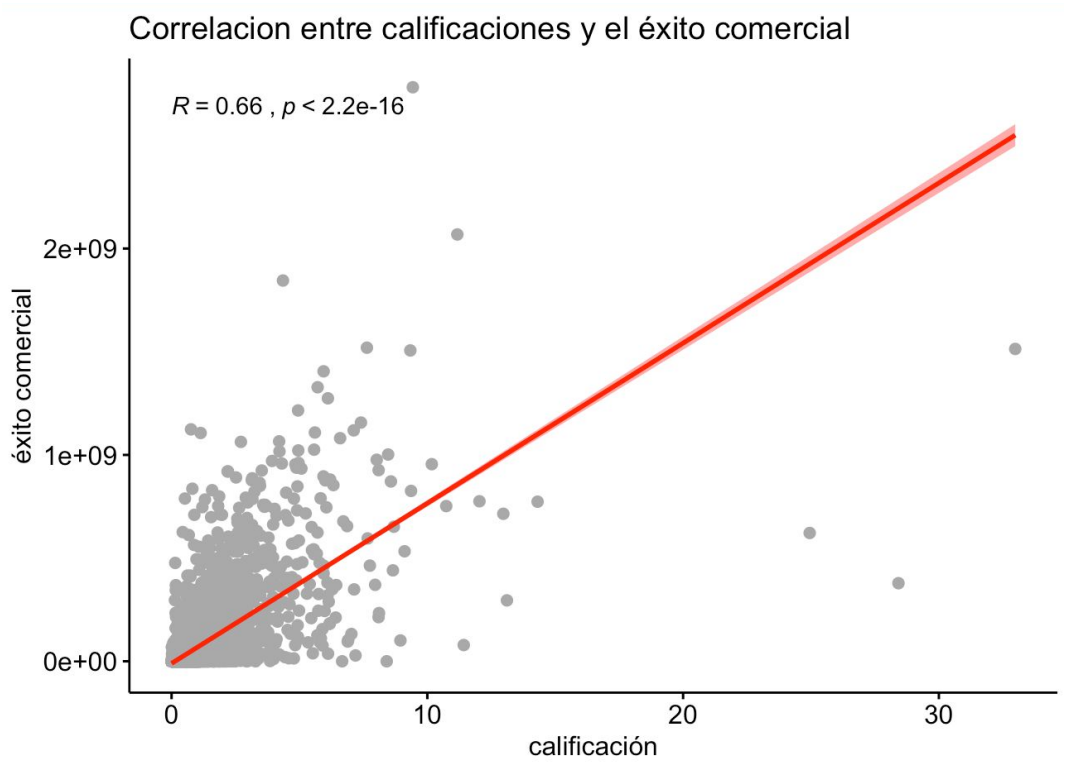
4.13. ¿En qué meses se han visto los lanzamientos máximos?

```
lazamientos_Maximos
  1   2   3   4   5   6   7   8    9   10  11  12
919 691 823 797 809 827 799 918 1331 1153 814 985
```

Los meses que más lanzamientos tienen son septiembre, luego octubre y luego diciembre.

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

Existe una relación entre el éxito comercial y la calificación de las películas, en la mayoría mientras más presupuesto tienen mejor calificación tienen. El coeficiente R de 0.66 muestra que existe una relación lineal ascendente.



4.15. ¿A qué género principal pertenecen las películas más largas?

	Var1	Freq
1	Drama	29
2	Documentary	13
3	History	12
4	Adventure	10
5	Fantasy	10
6	Science Fiction	8
7	Action	7
8	Horror	5
9	Romance	5
10	Thriller	5
11	War	4
12	Crime	3
13	Mystery	3
14	TV Movie	3
15	Comedy	2
16	Family	2
17	Western	2

El género con las películas más largas es el Drama, luego los documentales y luego la historia.

5. Preguntas extra

5.1. ¿La pelicula mas larga que tenga peor rating?

	id	imdb_id	popularity	budget	revenue	original_title	cast
6182	18729	tt0088583	0.000065	0	0	North and South, Book I	Patrick Swayze|Philip Casnoff|Kirstie Alley|Genie Fran...
7257	89049	tt0388644	0.001315	0	0	SoupÃ§ons	Michael Peterson
7268	203766	tt0403778	0.001531	0	0	Long Way Round	Ewan McGregor|Charley Boorman
4940	168219	tt2167393	0.003183	0	0	The Men Who Built America	
3895	125336	tt2044056	0.006925	0	0	The Story of Film: An Odyssey	Mark Cousins|Jean-Michel Frodon|Cari Beauchamp|A...

La pelicula mas larga que tiene peor rating es “North and South, Book 1”, con un 0% de rating.

5.2. ¿La película con más ganancia y con peor rating?

	id	imdb_id	popularity	budget	revenue	original_title
3523	38356	tt1399103	0.760503	1.95e+08	1123746996	Transformers: Dark of the Moon
8095	1642	tt0113957	1.136610	2.20e+07	1106279658	The Net
1931	10193	tt0435761	2.711136	2.00e+08	1063171911	Toy Story 3
5435	93456	tt1690953	3.928789	7.60e+07	970761885	Despicable Me 2
6556	58	tt0383574	4.205992	2.00e+08	1065659812	Pirates of the Caribbean: Dead Man's Chest
4368	49051	tt0903624	4.218933	2.50e+08	1017003568	The Hobbit: An Unexpected Journey
5232	597	tt0120338	4.355219	2.00e+08	1845034188	Titanic
5426	68721	tt1300854	4.946136	2.00e+08	1215439994	Iron Man 3
3376	1865	tt1298650	4.955130	3.80e+08	1021683000	Pirates of the Caribbean: On Stranger Tides
7388	285	tt0449088	4.965391	3.00e+08	961000000	Pirates of the Caribbean: At World's End
1922	12155	tt1014759	5.572950	2.00e+08	1025467110	Alice in Wonderland
4366	37724	tt1074638	5.603587	2.00e+08	1108561013	Skyfall

La película que más ha ganado que tiene el peor rating es “Transformers: Dark of the Moon”.

5.3. ¿La película con el peor rating del director que tiene más películas?

	original_title	popularity
10865	What's Up, Tiger Lily?	0.064317
7979	Broadway Danny Rose	0.133990
10805	Interiors	0.149259
7356	Stardust Memories	0.247924
9373	Shadows and Fog	0.267440
2552	Sweet and Lowdown	0.281948
9668	September	0.288278
8021	Zelig	0.388667
7279	Everything You Always Wanted to Know About Sex *B...	0.400301
9239	Crimes and Misdemeanors	0.415913

El director que tiene más películas hechas es “Woody Allen”, y su película que tiene peor rating es la de “What's Up, Tiger Lily?”