Análisis exploratorio de las mejores 1000 películas según IMBD

Indice de contenidos

Introducción

En tiempos de pandemia, debido a que todo el mundo se encontraba bajo confinamiento para prevenir contagios, muchas empresas sufrieron pérdidas masivas debido a la imposibilidad de trabajar con normalidad. Sin embargo existieron rubros que lograron generar ganancias récord, sobre todo aquellas que se convirtieron en las principales fuentes de entretenimiento durante el covid-19 como lo fueron los servicios de streaming. Según datos de Motion Picture Association en el año 2020 lograron superar un billón de suscripciones llegando a alcanzar $68.8 billones de dólares en ganancias globales, un aumento del 23% con respecto al año 2019. [Ver artículo]

A raíz de lo anterior, fueron que páginas como Internet Movie Database (IMBD) servían como principal fuente de búsqueda de películas o series para pasar nuestro aburrimiento encerrados en la casa, pero primero que todo ¿Qué es IMBD?

En este proyecto se analizarán las mejores mil películas según IMBD, tomando en cuenta algunas de las variables de interés disponibles en el sitio como género, ganancias, año de lanzamiento, número de votos, calificación, entre otros.

El análisis tiene como enfoque resolver principalmente las siguientes preguntas:

- ¿Cómo han ido variando las ganancias a través de las décadas?
- ¿Cómo a ido variando la duración a través de las décadas?

- ¿Son las películas más antiguas las que tienen mejor calificación?
- ¿Qué géneros son los más populares a través de las décadas?

Las preguntas anteriores tienen el objetivo principal de darnos a entender de mejor forma cómo han ido variando ciertos elementos de las películas a través de las décadas, ver si existe algún patrón que se mantiene durante el tiempo y obtener interesantes conclusiones a través de las décadas.

Se comenzará con una descripción de las variables a utilizar en el proyecto para ya luego ir resolviendo las preguntas establecidas a través del análisis de gráficos y tablas, y así presentar los resultados obtenidos. Se finalizará con una conclusión destacando los principales patrones encontrados.

La base de datos de IMBD

IMBD corresponde a una página web principalmente de películas y series, donde la principal característica es poder ver la calificación de estas, donde tanto los críticos profesionales como cualquier persona le pueden dar una nota del 1 al 10 para luego obtener el promedio de las calificaciones. También existen otros apartados como poder ver la trama, el elenco, reseñas de los usuarios y críticos, el equipo de producción, entre otros.

La base de datos a utilizar corresponde a las mejores mil películas de IMBD, la cual fue obtenida del sitio web Kaggle [Base de datos] e incluye datos hasta el 2020. Antes de comenzar con el análisis exploratorio es importante hacer un preprocesamiento a la base de datos, para trabajar con las variables de nuestro interés y ver si existen datos extraños, NAs, hacer cambios a la base para nuestra comodidad, etc.

Preprocesamiento

Lo primero que vamos a hacer es eliminar todas las variables que no son de nuestro interés, en nuestro caso serían: `Poster_Link`, `Certificate`, `Overview` y `Metascore` . Luego la variable `Runtime` es de tipo caracter ya que en la base de datos viene de la forma “89 min”, “100 min”,“120 min”, etc, por lo que eliminamos la palabra “min” para ya luego transformar la variable a numérica. Lo siguiente es cambiar el nombre de las variables del inglés al español para facilitar la comprensión de las variables.

Existe también una película que esta etiquetada como “PG” en su año de lanzamiento, por lo que buscamos el año de lanzamiento en Google de la película y reemplazamos “PG” por 1995. Por último para la variable `Genero` tenemos películas que se agrupan de la forma “Action, Drama, Crime” o “Drama, Action”. Para estos casos, donde existen mas de dos géneros que identifican a una película, tomamos el primer género del lado izquierdo es decir en el caso de “Action, Drama, Crime” tomamos solo “Action” y en el caso de “Drama, Action” tomamos solo “Drama”.

Las variables a utilizar en el reporte son las siguientes:

Variable	Tipo	Descripción
`Nombre_pelicula`	caracter	Nombre de la película
`Año_lanzamiento`	numérica	Año de lanzamiento de la película
`Duracion`	numérica	Duración en minutos de la película
`Genero`	caracter	Género de la película
`IMBD_Rating`	numérica	Calificación o "rating" de la película recibida en el sitio web IMBD (nota del 1 al 10)
`Numero_de_votos`	numérica	Número de personas que calificaron la película
`Ganancias`	numérica	Dinero recaudado por la película en dólares sin ajustar por inflación

Análisis Exploratorio

Ya que contamos con los años de lanzamiento de las películas, un buen punto de partida para este análisis exploratorio es ver cómo han variado las ganancias, la duración y el rating promedio a través del tiempo agrupando los años en décadas. Empezaremos con las ganancias, que se muestra en la Figura 1. Hay que tener en consideración que para este gráfico en específico se toma en cuenta solo hasta la década del 2010 (es decir de 2010 a 2019) ya que para las películas del año 2020 no se registraron sus ganancias.

Tal y como vemos en la Figura 1, notamos que en general a medida que han pasado las décadas las ganancias han aumentado considerablemente, especialmente si notamos que de la década del 1920 al 2010 las ganancias han aumentado casi 100 veces mas. También podemos ver que de 1920 a 1930 existe un gran salto en las ganancias, para luego en 1940 decaer nuevamente. Cabe mencionar que las ganancias de estos años pueden estar mal representadas ya que estamos hablando de películas de hace bastante tiempo por lo cual no se tienen sus ganancias exactas o simplemente no se registraron en la base de datos. Otra posibilidad es que durante estas décadas ocurrieron dos guerras mundiales lo cual pudo haber afectado a la producción de películas y las ganancias obtenidas.

Veamos ahora como han ido variando la duración de las películas a lo largo de los años.

De la Figura 2 lo más relevante es que en la década de 1920 se encuentran las películas de menor duración y en los siguientes años comenzaron a aumentar hasta llegar a la década de los 60. Desde ese entonces en adelante la tendencia es bastante uniforme, donde la duración varía entre 121 y 128 minutos, es decir, durante los últimos 60 años no han ocurrido cambios notorios en cuanto a la extensión de las películas se refiere.

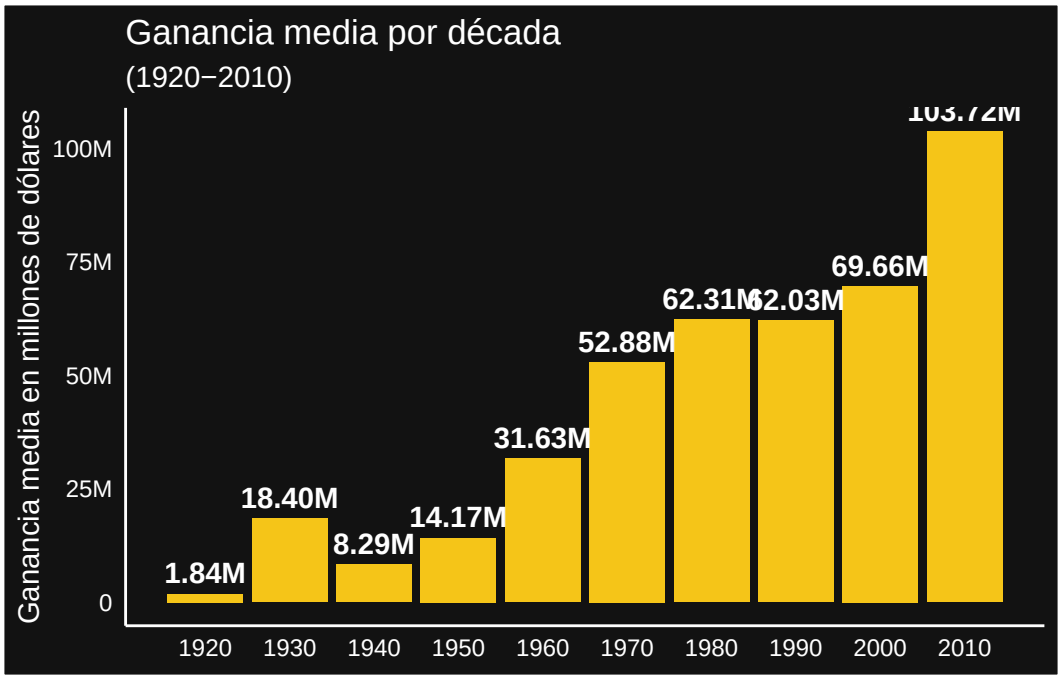

Figura 1: Ganancia media por década (1920-2010)

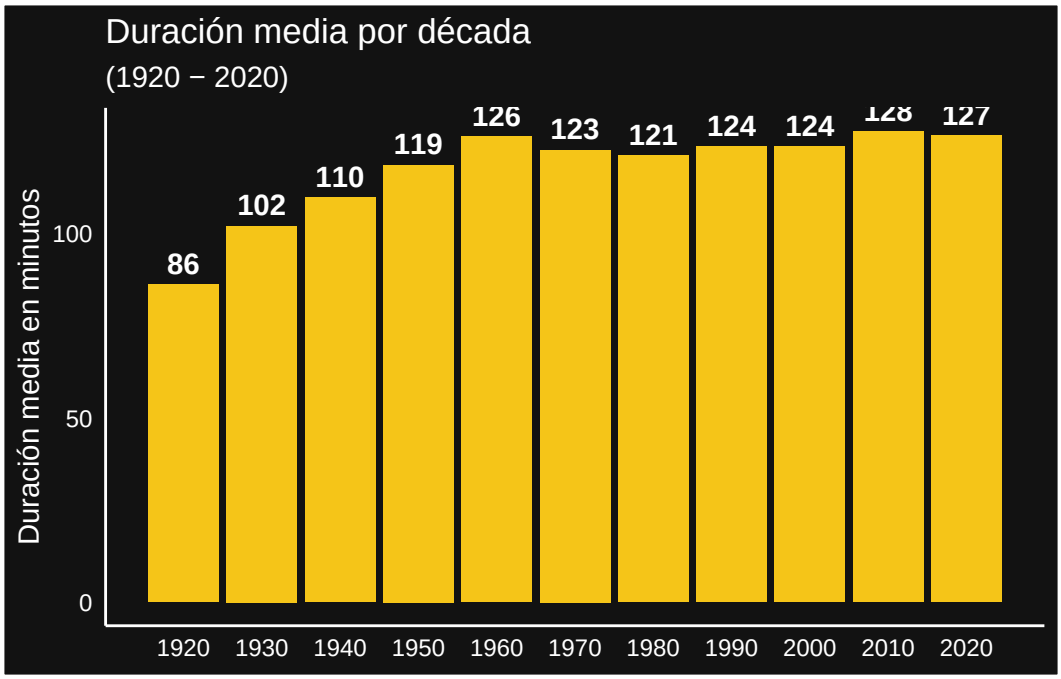

Figura 2: Duración media por década (1920-2020)

Ya vimos la variación de las ganancias (Figura 1) y de la duración (Figura 2) media a través del tiempo, por lo que nos queda el rating que se presenta en la siguiente figura:

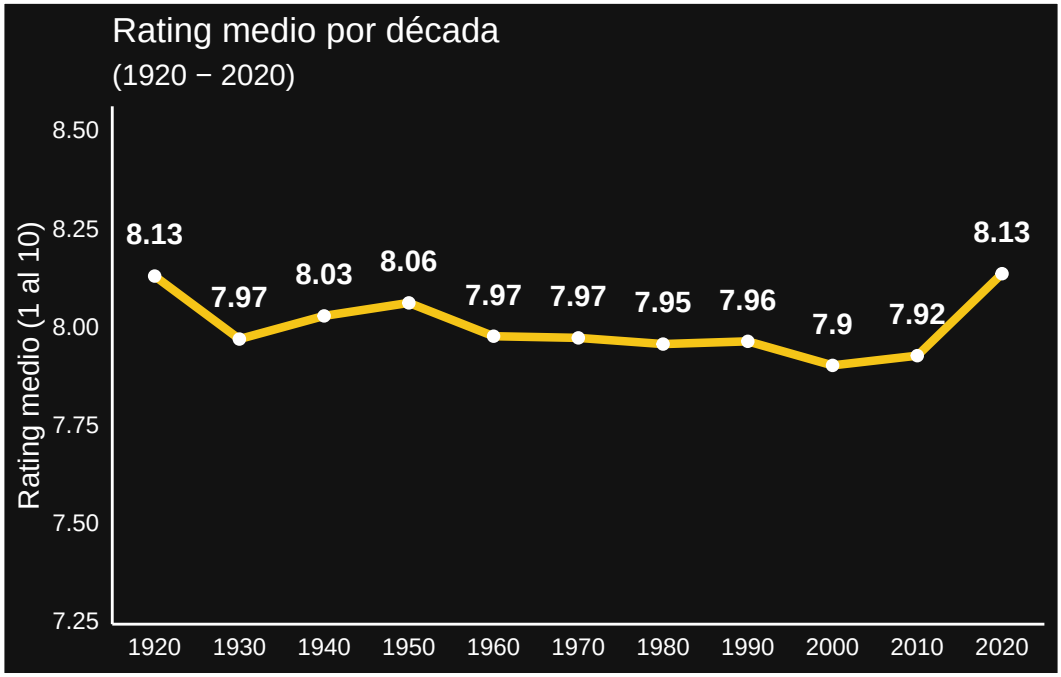

Figura 3: Rating medio por década (1920-2020)

Notamos de la Figura 3 que si bien los ratings son bastante parecidos entre sí, las notas mas altas se alcanzan en la década de 1920 y 2020. Esto es interesante ya que si contrastamos esta información con la Figura 1 y Figura 2 vemos que a pesar de que la década de 1920 tiene la menor duración y las ganancias más bajas aun así tiene el rating más alto junto a 2020, en cambio la década de 2010, el cual tiene las ganancias más altas y mayor duración, tiene uno de los ratings más bajos en comparación a las otras décadas. Vemos también que de la década del 60 al 90 se mantuvo el rating medio para luego decaer en el 2000, siendo esta la peor década en cuanto a rating para ya luego aumentar en las siguientes dos décadas.

A simple vista pareciera ser que las ganancias y la duración de las películas no tienen ninguna influencia en el rating según el análisis hecho anteriormente, pero surge otra posibilidad, quizás el número de votos puede estar afectando al rating por lo que es interesante ver la cantidad de votos por década.

Vemos de la Figura 4 que la popularidad de las películas han ido en aumento desde 1920 hasta 1990, donde en la década del 2000 y 2010 disminuye levemente respecto a 1990 para luego caer de forma abrupta en el año 2020. Esto puede estar sucediendo ya que comenzó la pandemia por lo que tanto la producción de películas como el marketing que se le realiza a las películas previo a su estreno se vieron afectadas. A raíz de lo anterior se realizaron muy pocas películas en 2020 y al no haber un marketing en condiciones normales afectó al número de votos ya

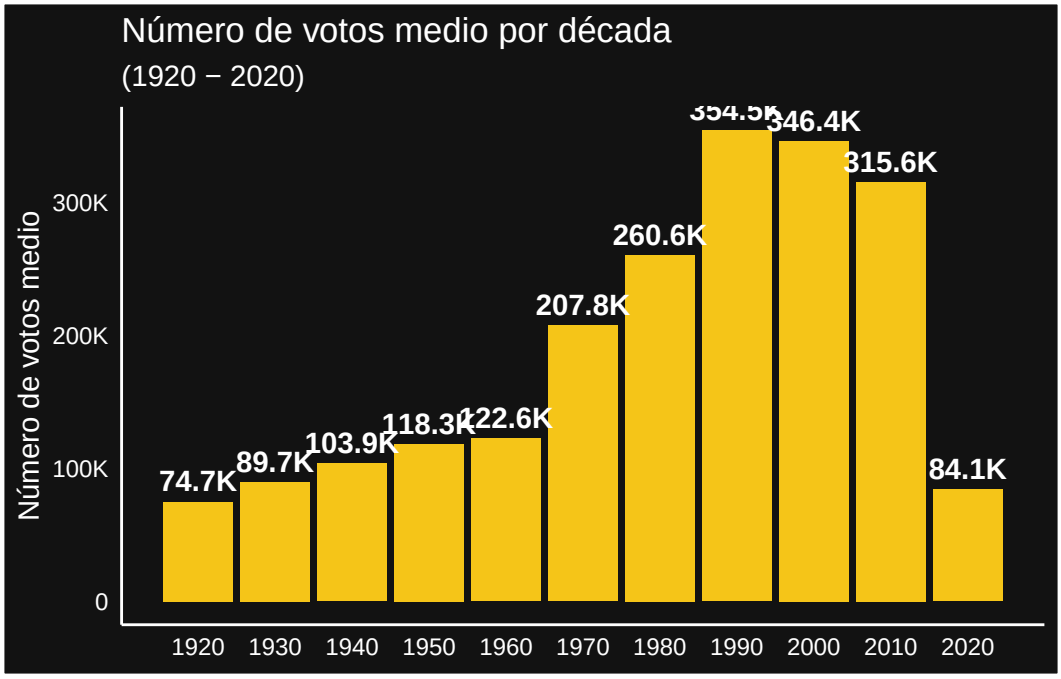

Figura 4: Número de votos medio por década (1920 - 2020)

que, las personas no sabían que cierta película había sido estrenada o las películas fueron estrenadas en cines antes de comenzar la pandemia y no la publicaron luego en plataformas de streaming.

Aún así como vimos en la Figura 3 las películas realizadas en este año obtuvieron los mejores ratings junto a 1920 que también tiene un cantidad de votos similar al 2020. Esto nos habla sobre la población nicho que existe para cierto tipo de películas, donde a pesar que en la década de 1920 tenga la popularidad más baja, tiene el rating más alto. Lo que es interesante es que la popularidad parece coincidir con las ganancias vistas en la Figura 1 donde las décadas más populares (1990, 2000 y 2010) tienen las ganancias más altas. Más adelante hablaremos sobre si existe una correlación entre estas dos variables, número de votos y ganancias.

Una variable que hemos pasado por alto es el género de la película y con esta variable pueden surgir algunas preguntas tales como ¿Cuál es el género más repetido por década? ¿Cuánto es la ganancia y rating medio de las películas según su género? ¿Qué tan populares son los diversos géneros?

Veamos primero qué género es el que más se repite por década.

Géneros más repetidos por década
(1920 - 2020)

Década	Género	Conteo
1920	Drama	3
1930	Comedy	9
1940	Drama	14
1950	Drama	23
1960	Drama	27
1970	Drama	22
1980	Action	19
1980	Comedy	19
1990	Drama	43

Pareciera entonces que las películas que mejor les va en cuanto a rating son drama pero ¿Será también que el género de drama es el que más ganancias obtiene? Antes de contestar esta pregunta veamos primero el top 10 de películas según su ganancia.

Top 10 películas según sus ganancias

Película	Año lanzamiento	Género	Rating	Número de votos	G
Star Wars: Episode VII - The Force Awakens	2015	Action	7.9	860823	9
Avengers: Endgame	2019	Action	8.4	809955	8
Avatar	2009	Action	7.8	1118998	7
Avengers: Infinity War	2018	Action	8.4	834477	6
Titanic	1997	Drama	7.8	1046089	6
The Avengers	2012	Action	8.0	1260806	6
Incredibles 2	2018	Animation	7.6	250057	6
The Dark Knight	2008	Action	9.0	2303232	5
Rogue One	2016	Action	7.8	556608	5
The Dark Knight Rises	2012	Action	8.4	1516346	4

Figura 6: Top 10 películas según su ganancia

Vemos de la Figura 6 que solo existe una película de drama que está en el top 10, donde 8 de las 10 películas son de acción. Nótese que las películas con mejores ganancias pertenecen a sagas muy conocidas como lo son “Star Wars” y las famosas películas de Marvel “Avengers”, las cuales al ser sagas que llevan mucho tiempo en el mundo del entretenimiento tienen una gran cantidad de fanáticos que se han ido acumulando a lo largo del tiempo lo que lleva a ganancias altísimas.

Una vez dicho esto, veamos entonces las ganancias y el rating medio según el género.

Si nos fijamos en la Figura 7 vemos que el género de acción tiene una ventaja absoulta si lo comparamos con el género de drama en cuanto a ganancias se refiere, sin embargo vemos que, a pesar de tener menos ganancias drama, aun así tiene prácticamente el mismo rating que las películas de acción por lo que una cosa es clara, que tan bien le vaya a una película en cuanto a sus ganancias no tiene nada que ver con que tan bien su rating final es y esto queda aún mas claro si nos fijamos en el género “western” que tiene un rating de 8.4 aproximadamente y tiene casi 10 veces menos ganancias con respecto al género de acción que tiene 8 de rating aproximadamente.

Aquí el género “family” es el que tiene las mayores ganancias ya que corresponde a dos películas que son “ET” y “Charlie en la fábrica de chocolate”. A pesar de las ganancias gigantescas que tienen estas dos películas, tienen el peor rating de todos los géneros lo que deja en evidencia la poca influencia que tiene las ganancias en el rating.

Ganancias y rating medio según su género

Género	Ganancias	Rating
Action	141963092	7.95
Adventure	86454990	7.94
Animation	127967528	7.93
Biography	60128732	7.94
Comedy	32537590	7.90
Crime	34191231	8.02
Drama	38677281	7.96
Family	219555277	7.80
Fantasy	[1] NaN	8.00
Film-Noir	1278626	7.97
Horror	73585773	7.91
Mystery	30439534	7.97
Thriller	17550741	7.80
Western	14555377	8.35

[1]Los datos faltantes corresponden a dos películas de los años 1920 y 1922

Figura 7: Tabla de ganancias y rating medio por género

Dijimos con anterioridad que íbamos a ver si existe alguna correlación entre el número de votos y las ganancias, por lo que realizamos el gráfico de correlación que se ve en la Figura 8

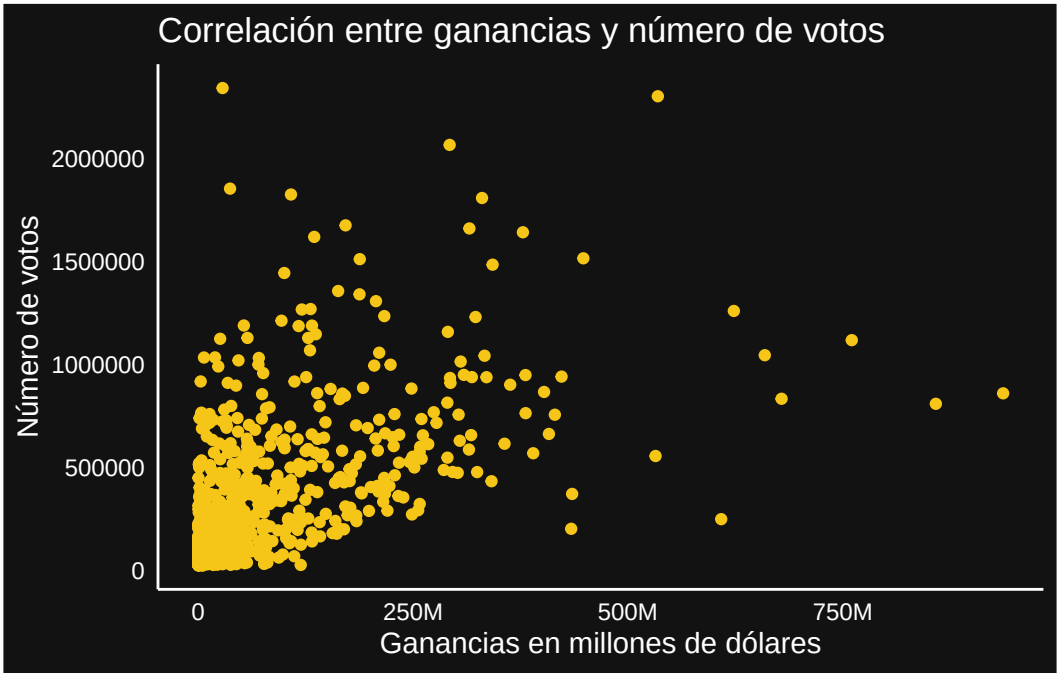

Figura 8: Gráfico de correlación entre ganancias y número de votos

Como muestra la Figura 8 pareciera existir una leve correlación positiva entre estas dos variables, sin embargo, existen películas que obtienen muchas ganancias pero menos de 500 mil votos y películas que tienen más de 1 millón de votos pero que obtienen pocas ganancias. Existe una relación entre el número de votos y las ganancias, pensando en que si una película tiene muchas ganancias claramente tendrá más votos pero lo contrario puede no suceder ya que si una película tiene muchos votos no significa que tendrá muchas ganancias ya que, ahora mas que nunca con los servicios de streaming, existe la posibilidad de ver películas de años anteriores y al ser justamente un servicio de streaming el dinero recaudado va para la empresa que ofrece el servicio, o al menos un gran porcentaje, y no va directamente a la productora de la película.

Por último, veamos la tabla de la Figura 7 y agreguemos la columna de número de votos.

De la Figura 9 vemos que el género de acción es el que mayor número de votos tiene y esto no es sorprendente, pues justamente hablabamos de que sagas como “Star Wars” y “Avengers” tienen una gran cantidad de fanáticos lo que se ve claramente reflejado en la tabla. Lo que si es curioso de esta tabla es que el género de animación tiene unas ganancias similares a las de acción pero tienen menos votos comparada con este, por lo que puede estar ocurriendo lo que mencionábamos antes de las poblaciones nicho, donde hay películas que tienen buena valoración pero poco número de votos y ganancias, y películas con ganancias altísimas y en

Ganancias,rating y número de votos medio según su género

Género	Ganancias	Rating	Número de votos
Action	141963092	7.95	420247
Adventure	86454990	7.94	313558
Animation	127967528	7.93	268032
Biography	60128732	7.94	272805
Comedy	32537590	7.90	178196
Crime	34191231	8.02	313398
Drama	38677281	7.96	212344
Family	219555277	7.80	275610
Fantasy	[1] NaN	8.00	73111
Film-Noir	1278626	7.97	122405
Horror	73585773	7.91	340232
Mystery	30439534	7.97	350250
Thriller	17550741	7.80	27733
Western	14555377	8.35	322416

[1]Los datos faltantes corresponden a dos películas de los años 1920 y 1922

Figura 9: Tabla de ganancias, rating y número de votos según su género

consecuencia un gran número de votos donde al tomar el promedio nos puede generar que el género de animación tenga ganancias muy altas pero poco número de votos en comparación a las películas de acción.

Conclusión