Análisis de datos sobre películas, sus metadatos y premios

Javier Pozueco, Technical Analytics Director, Jellyfish,

1 de julio de 2017

Resumen

¿Es posible predecir si una película tendrá éxito, antes de que esta vea la luz? Puede ser que una película sea popular dependiendo únicamente de la calidad de la misma, pero puede también que determinados factores como su reparto, género, país o fecha de estreno, también influyan en el número de personas a las que esta les va a gustar. En este trabajo analizaremos si las caraterísticas de una película conocidas antes de que se realice, nos permiten saber si esta tendrá éxito entre los usuarios de IMDb.

Keywords: IMDb, Kaggle

1. Introducción y objeto del análisis

Para intentar predecir si una película tendrá éxito vamos a analizar las características de una película conocidas antes de que esta vea la luz y que están disponibles a través de la página web IMDb. Utilizaremos una base de datos ya existente e intentaremos utilizar estas características para saber si tienen alguna relación con la puntuación otorgada por los usuarios del portal web.

En primer lugar cargaremos la información de esta base de datos y comprobaremos cuál es su calidad. También analizaremos su distribución a lo largo del tiempo y cómo influye el país, el género, el director o el reparto en la variabilidad de las puntuaciones.

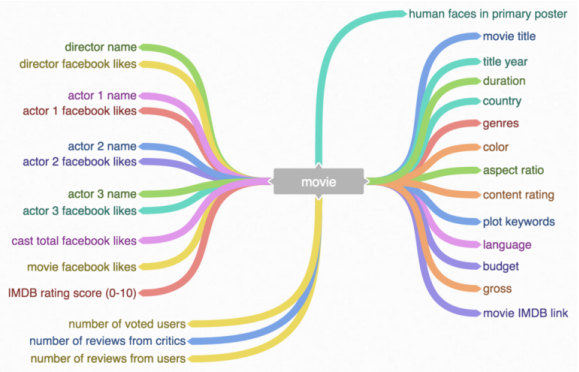
En base a esta información crearemos un modelo no supervisado que nos permita agrupar los datos de las películas si tener en cuenta su puntuación. A continuación crearemos varios modelos de regresión teniendo en cuenta su puntuación, para ver si somos capaces de predecir su éxito. Finalmente crearemos un sencillo recomendador, que dado el título de una película nos devuelva el título de películas similares que nos podrían llegar a gustar.

El análisis de los datos que van a ser utilizados se realizará utilizando Tableau. Para la limpieza y transformacíon de los datos se utilizará R, lenguaje también utilizado para crear los modelos no supervisados y supervisados. También se utilizará R para crear el sistema de recomendacíon.

2. Carga de los datos y análisis descriptivo

A pesar IMDb sólo presenta su información a través de su portal web, hay trabajos como [1] que ya han extraído los datos más relevantes utilizando, en este caso en particular, Python y su librería scrapy, siguiendo los siguientes pasos:

- Se ha utilizado scrapy para obtener un listado de 5043 películas desde el portal The Numbers.
- Se ha realizado una búsqueda en IMDb para obtener el enlace a cada una de las películas.
- Para cada uno de los enlaces, se ha descargado y parseado su contenido, con el fin de obtener los datos más relevantes que se pueden ver en la siguiente figura:



- Mediante el reconocimiento de imágenes, también se sabe para cada película el número de caras que aparecen en su póster.

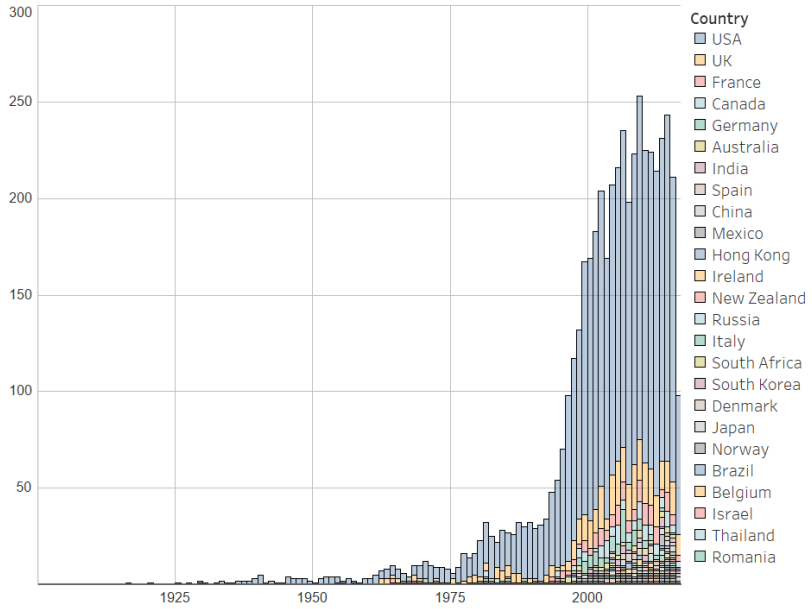
A continuación se incluyen notas importantes del autor, sobre cada uno de los atributos extraídos de las películas, que deberemos tener en cuenta en el análisis que se realizará más adelante:

- En aproximadamente 800 películas las ganacias son 0, debido a que esta información no está disponible o porque la herramienta utilizada para extraer el dato no devolvió ninguna respuesta en un tiempo razonable.
- En 908 directores de las películas descargadas, el número de likes en Facebook es 0 debido, como en el caso anterior, a que los valores aparecen en un marco que no se carga junto con el resto de la página.
- Hay películas para las que no se ha tenido en cuenta la moneda del país que las ha producido, y aunque en los datos se muestran dólares, en realidad es la moneda del país correspondiente.
- Para calcular los actores principales de cada película, se han tenido en cuenta todos los actores y actrices del reparto, y aquellos tres con mayor número likes en Facebook, son considerados los principales.

- Por último, cabe destacar también que el presupuesto y las ganancias de las películas no tienen en cuenta la inflación o el cambio de moneda que había en el año de su realización.

2.1. Datos generales sobre el dataset

Como primer paso vamos a analizar el dataset extraído de Kaggle, que será el utilizado en este trabajo. Para ello se muestra en la siguiente figura el número de películas por país en nuestro dataset cada año, como se puede ver en Tableau Public:



A continuación se muestra el número total de películas, incluyendo también series, según IMDb[2]. El número de películas descargado resulta ser inferior al 3 % del total, pero al ser un listado con los datos disponibles en The Numbers, podemos considerar que son las que tienen ganancias conocidas y mayores:

2.2. Rating otorgado por los usuarios

En el listado de películas considerado se incluye también el rating otorgado por los usuarios según IMDb. Este valor proviene de las puntuaciones que dan los usuarios con valores comprendidos entre 0 y 10, y que a través de diferentes métodos se pondera para evitar que un mismo usuario pueda votar varias veces.

Para que una película pueda estar dentro del top 250 de IMDb sólo se tienen en cuenta las votaciones de los usuarios más frecuentes, aunque para evitar cualquier tipo de engaño los requisitos necesarios para ser un usuario frecuente no se han publicado:

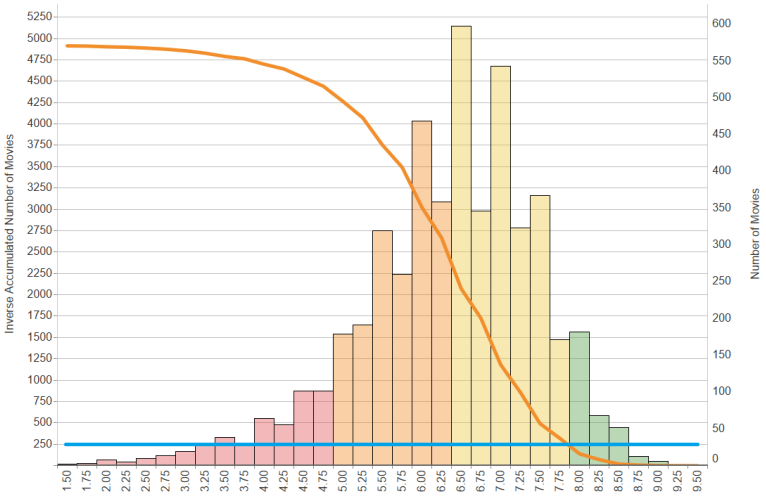
$$W = \frac{R_v + C_m}{v + m} \tag{1}$$

donde:

W = rating ponderado
R = media de las votaciones de 0 a 10
v = número de votos
m = mínimo número de votos para estar en el top 250
C = la media de todos los votos

En la figura que se muestra a continuación se puede observar la distribución del rating otorgado por los usuarios en el dataset considerado. En este gráfico se han destacado las películas según su calidad, en base a los siguientes criterios, como se puede ver en Tableau Public:

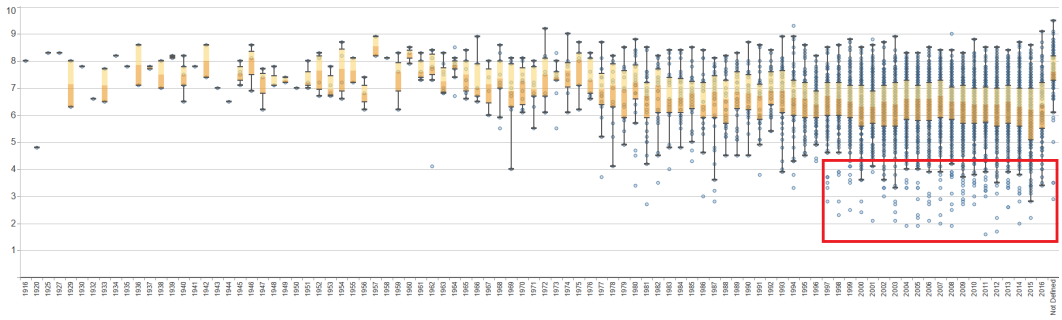
- Verde, si las películas son muy buenas y superan una puntuación de 8.
- Amarillo, si las películas son buenas y tienen una puntuación entre 6,5 y 8.
- Naranja, si las películas son regulares y tienen una puntuación entre 5 y 6,5.
- Rojo, si las películas son malas y bajan del 5.



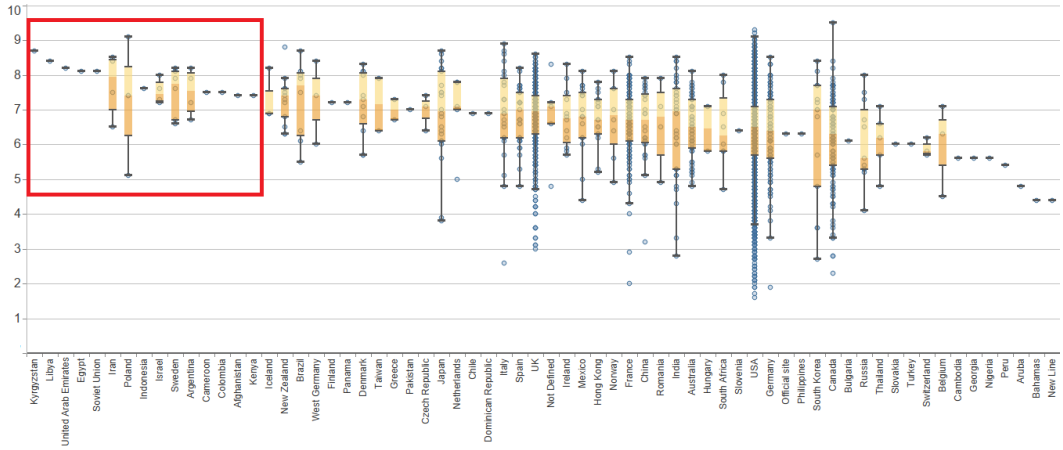
Cabe destacar también como el top 250 de IMDb se corresponde con una puntación ligeramente inferior a 8, según el gráfico anterior.

2.3. Distribucion de las puntuaciones

Desde finales de la segunda guerra mundial, se ha producido un incremento gradual de las películas producidas debido al avance de la tecnología para la realización de las mismas. Además, como se pudo ver anteriormente en el número de películas producidas cada año, desde finales del siglo pasado el avance de los medios digitales ha facilitado y abaratado su distribución para multitud de productoras independientes[3]. Por otro lado, en la siguiente figura también se puede observar que aunque el número de trabajos ha aumentado, también ha dismunido la calidad en algunos casos, como se puede ver en Tableau Public:

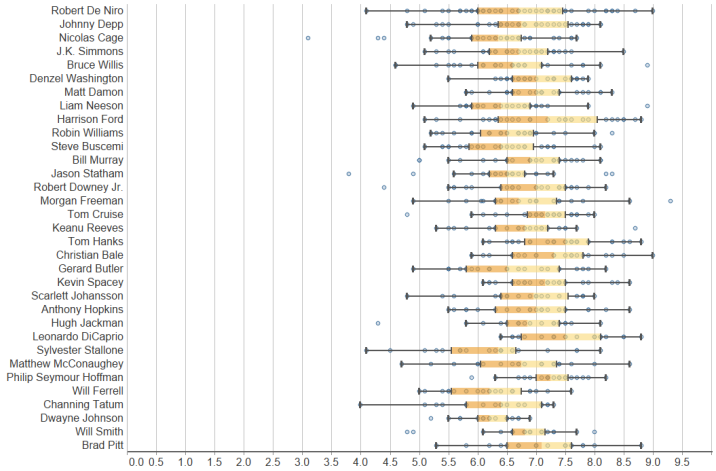


Hay países como Kirguistán, Libia, Egipto o Irán con un reducido número de películas pero cuya puntuación es superior al resto. Por otro lado Estados Unidos, Reino Unido y Francia son los países con más películas, pero cuya calidad no siempre es mejor que las demás. Cabe destacar que en el dataset que estamos estudiando aparecen películas de la India pero no en gran cantidad, siento este país uno de los productores actuales más grande, como se puede ver en Tableau Public:

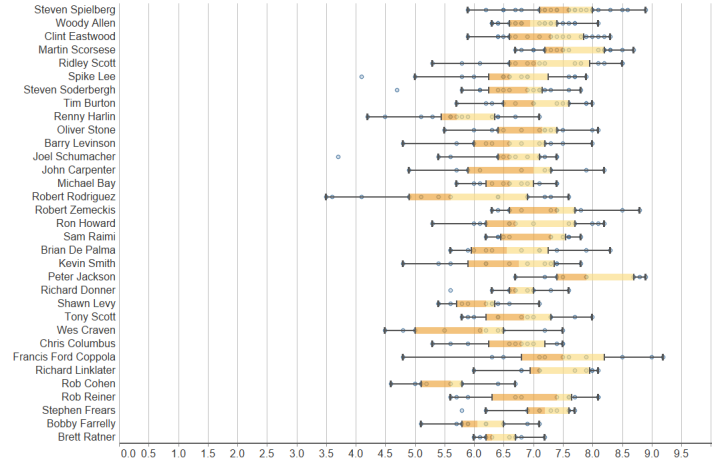


2.4. Análisis de actores y directores

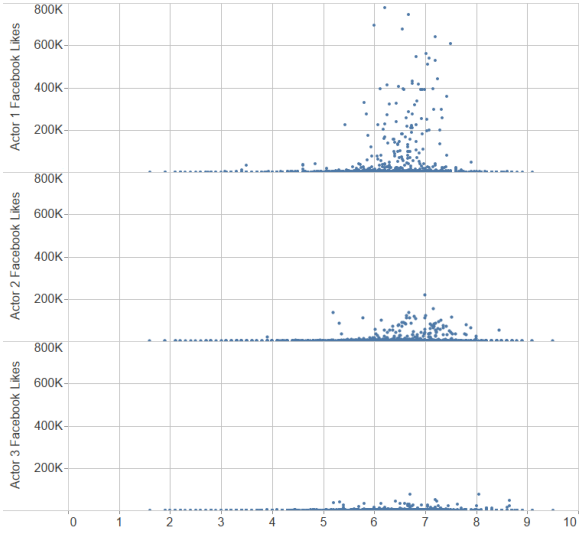
Resulta interesante también analizar las puntuaciones de los actores y directores con mayor número de películas. Se puede ver a continuación como la variación de las puntuaciones de los actores es mayor que la de los directores más populares. En el caso de los actores, algunos nombres como Robert De Niro, Sylvester Stallone o Channing Tatum tienen buenas películas, pero también películas de calidad mediocre. Otros actores como Matt Damon destacan por tener un gran número de películas de calidad notable y Nicolas Cage, por ejemplo, por haber realizado algunos papeles de mala calidad, como se puede ver en Tableau Public:



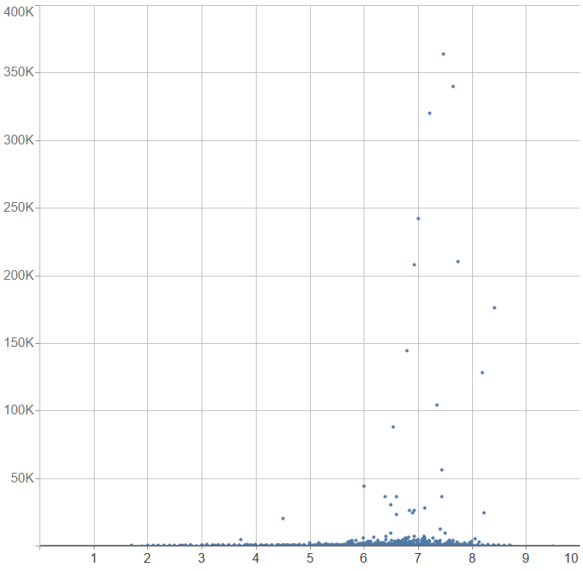
Con respecto a los directores destaca Steven Spielberg como el actor mas prolífico y siempre de gran calidad y Robert Rodríguez con diferentes producciones pésimas, como se puede ver en Tableau Public:



Para saber si influye el número de likes en la puntuacíon de las películas, podemos ver en los siguientes gráficos como es más importante en los actores tener más likes para obtener mejores resultados, y sobre todo el en actor principal de la película, como se puede ver en Tableau Public:

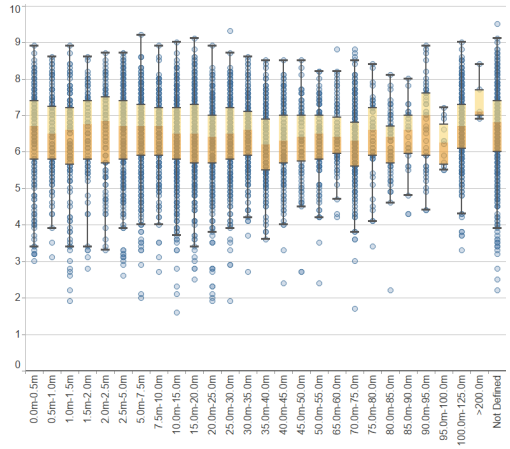


En los directores aunque con menos likes, las puntuaciones son superiores cuanto mayor es la gente a la que le gustan, como se puede ver en Tableau Public:

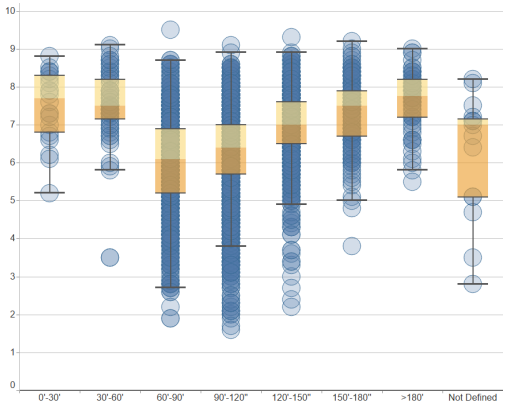


2.5. Análisis de presupuesto y duración

Para saber si influye el presupuesto y la duración de las películas en la puntuación obtenida en IMDb, se puede ver como a mayor dinero invertido menor es la dispersión en las puntuaciones, y aquellas películas que son mejores tienen una duración de menos de una hora o lo más larga posible, como se puede ver en Tableau Public 1 y Tableau Public 2:



(a) Rating por presupuesto

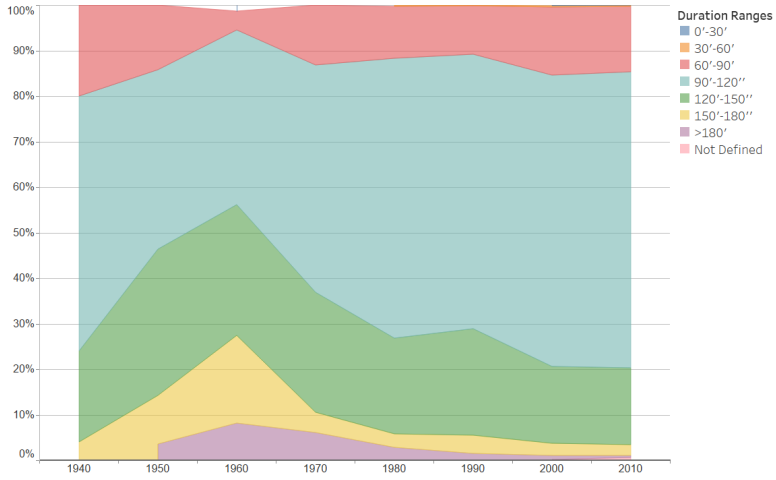


(b) Rating por duración

La combinación de estos dos factores nos muestra que una gran duración y presupuesto moderado resulta en mejores películas y como las películas de entre media hora y una hora también tienen gran aceptación entre el público, como se puede ver en Tableau Public:

Budget Ranges	0'-30'	30'-60'	60'-90'	90'-120"	120'-150"	150'-180"	>180'	Not Defined
0.0m-0.5m	6.80	7.50	6.49	6.45	5.97	8.30		5.10
0.5m-1.0m			6.56	6.37	7.50	8.30	8.10	
1.0m-1.5m		8.05	5.59	6.46	6.74	6.20		7.00
1.5m-2.0m	6.80		6.32	6.49	7.44	4.80		
2.0m-2.5m		8.20	6.28	6.43	7.25	7.83	8.70	
2.5m-5.0m	6.60	7.80	5.64	6.56	7.64	7.60	8.25	
5.0m-7.5m		6.47	5.62	6.50	7.08	7.20	7.90	6.70
7.5m-10.0m			6.33	6.32	7.13	7.44	7.80	
10.0m-15.0m			5.84	6.25	7.04	7.60	8.00	
15.0m-20.0m			5.46	6.29	7.28	7.56	8.08	
20.0m-25.0m			5.16	6.19	6.79	6.43	7.80	7.20
25.0m-30.0m			5.38	6.20	7.21	7.70	7.40	
30.0m-35.0m		8.10	6.18	6.26	7.07	7.00	7.90	
35.0m-40.0m			5.12	6.03	6.80	7.93	7.83	
40.0m-45.0m			5.99	5.92	7.02	7.20	7.57	
45.0m-50.0m		7.30	5.75	6.17	6.73	7.00	7.60	
50.0m-55.0m			5.50	6.15	6.61	7.90	7.43	
65.0m-60.0m			6.37	6.24	6.85	6.35	6.47	
70.0m-75.0m			5.55	5.97	6.71	7.43	7.55	5.10
75.0m-80.0m			6.00	6.15	6.79	7.20		
80.0m-85.0m			6.17	6.07	6.57	6.00		
85.0m-90.0m			5.53	6.18	7.01			
90.0m-95.0m			6.77	6.21	7.19	8.00	8.90	
95.0m-100.0m			6.73	6.13	5.95			
100.0m-125.0m			5.86	6.48	6.83	7.13	7.13	
>200.0m				7.20	7.75			
Not Defined	7.65	7.52	6.47	6.37	6.81	7.15	7.93	6.25

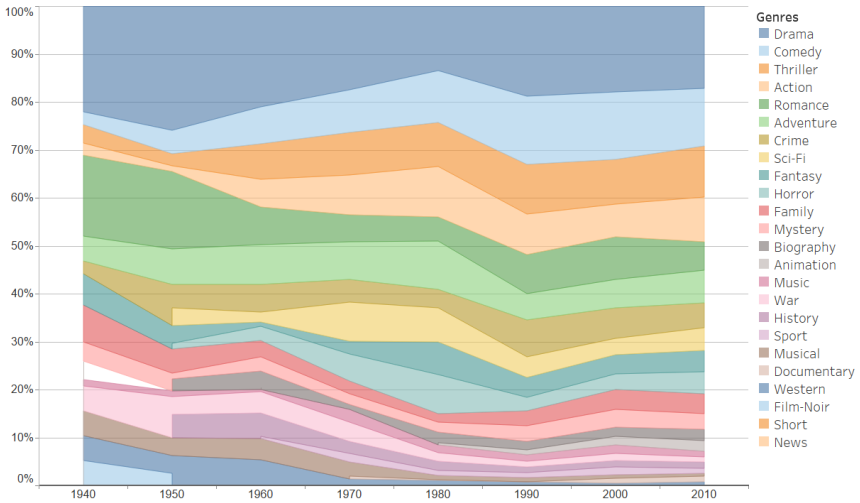
Como curiosidad se puede ver cómo ha variado la duración de las películas a lo largo del tiempo[4], podemos representar los ragos de duración anualmente y ver como son cada vez más populares las películas con una duración de entre 90 y 120 minutos, disponible en Tableau Public:



2.6. Análisis del género de las películas

Para saber si el género de las películas influye en el rating de las mismas se muestra a continuación cómo aquellos géneros menos populares son los que mayor puntuación reciben y el drama, que es el género más popular, se situa tras de estos, disponible en Tableau Public y el script utilizado para dividir el género en R Markdown:

Es interesante también ver como han evolucionado los géneros más populares a lo largo del tiempo, para ver como la acción, el thriller o la comedia son cada vez más populares y otros como el drama y el romance, ya no tanto, como se puede ver en Tableau Public:



3. Preparación y limpieza del dataset

En primer lugar se ha realizado un análisis de la calidad del dataset para saber la cantidad de datos no disponible para cada uno de los atributos. Como se puede ver en el anexo A, el presupuesto de las películas y las ganancias en mayor medida, arrojan una mayor falta de datos que el resto de características del dataset. Como las ganancias es un dato conocido únicamente despues de haberse realizado la película, no será tenido en cuenta en la creación de los modelos detallados en la siguientes secciones.

3.1. Modelo no supervisado

A continuación se seleccionan los datos para la construcción del modelo no supervisado solamente con las siguientes columnas y se eliminan las filas con datos no definidos con R Markdown. También se han eliminado todas las películas que no pertenecen a USA, ya que no siempre se realiza el cambio de moneda y el presupuesto no se puede utilizar para la construcción del modelo:

- Actor 1 Facebook Likes, Actor 2 Facebook Likes y Actor 3 Facebook Likes
- Director Facebook Likes
- Duration
- Budget

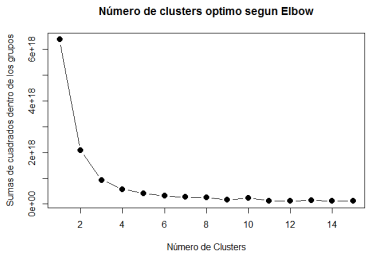
3.2. Modelo supervisado

Para la construción de los modelos de regresión se utilizarán las siguientes columnas y eliminándose como en el caso anterior las filas con datos no definidos y los países que no son USA, utilizando R Markdown. Cabe destacar, en base a [6], que se ha creado una columna por género de la película con el valor 0 o 1 si la película pertenece a ese género, una columna con el valor 0 o 1 si la película es el blanco y negro o en color, y una columna para el director, el actor 1, el actor 2 y el actor 3, con el número de películas en las que participan:

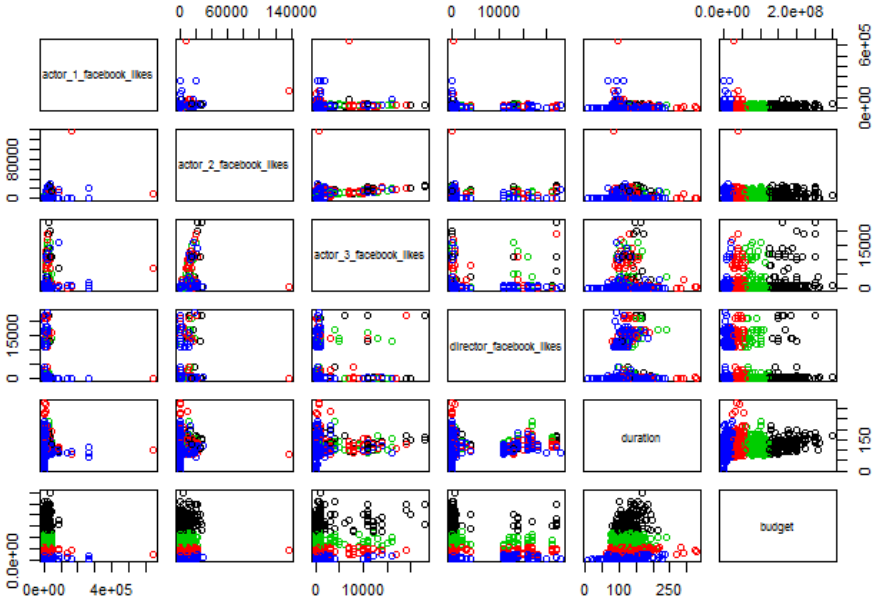
- Actor 1 Facebook Likes, Actor 2 Facebook Likes y Actor 3 Facebook Likes
- Director Facebook Likes
- Duration
- Budget
- Title Year
- Face Number in Poster
- Number Critic for Reviews
- Color
- Genre
- Director Movies
- Actor 1 Movies, Actor 2 Movies y Actor 3 Movies

4. Análisis exploratorio apoyado en algún método NO supervisado

Para la creación del modelo no supervisado se utilizará el conjunto de datos generado anteriormente y se calculará el número óptimo de clusters con R Markdown:



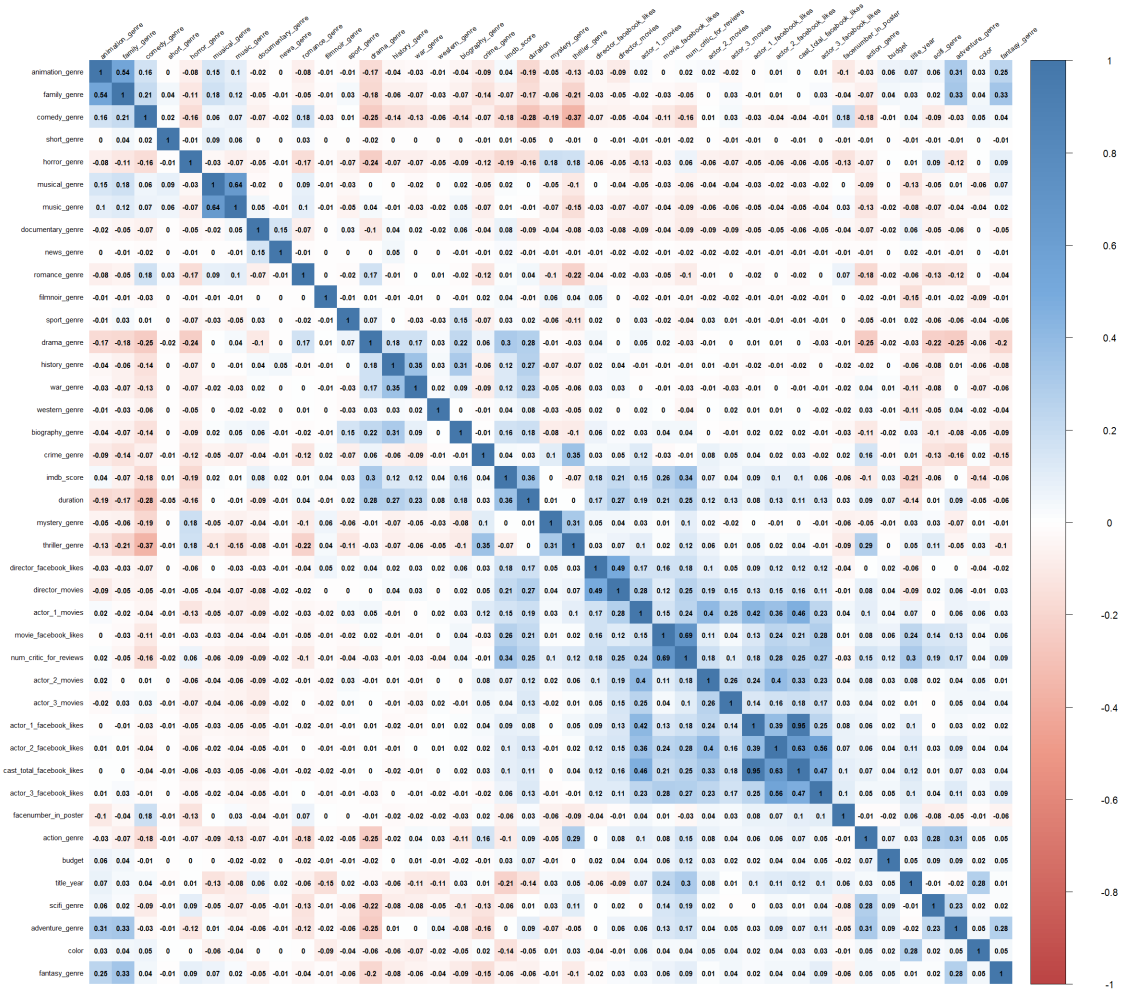
En la siguiente imagen se puede ver la relación entre los distintos clusters generados según el número óptimo calculado previamente, que en este caso es 4:



En base al análisis no supervisado, se han creados grupos de películas en base principalmente al presupuesto utilizado.

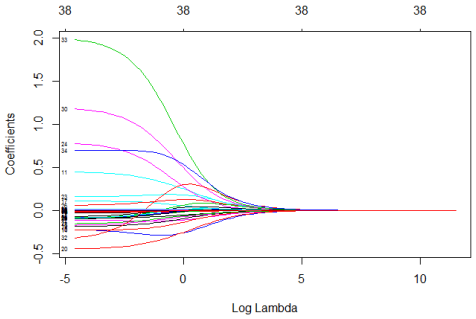
5. Modelos de Machine Learning supervisados

Para la creación del modelo no supervisado se utilizará el conjunto de datos generado anteriormente y se calculará el número óptimo de clusters según R Markdown:

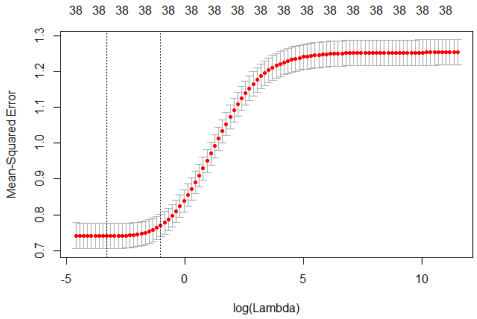


Según la matriz de correlación anterior, existe una gran influencia entre los atributos Cast Total Facebook Likes y Actor 1 Facebook Likes, por lo que se ha eliminado el primero para la construcción de los siguientes modelos de regresión.

El primero de ellos será un modelo de regresión Ridge[5] que será optimizado para obtener el valor del parámetro lambda que hará que se obtengan los mejores resultados en cuanto al valor del error MSE, utilizando R Markdown. En la primera imagen se pueden ver las diferentes curvas del modelo considerado con respecto a cada uno de los valores del parámetro a optimiar, y en la segunda imagen el error obtenido con cada valor del parámetro. El error del parámetro óptimo es 0,707641:

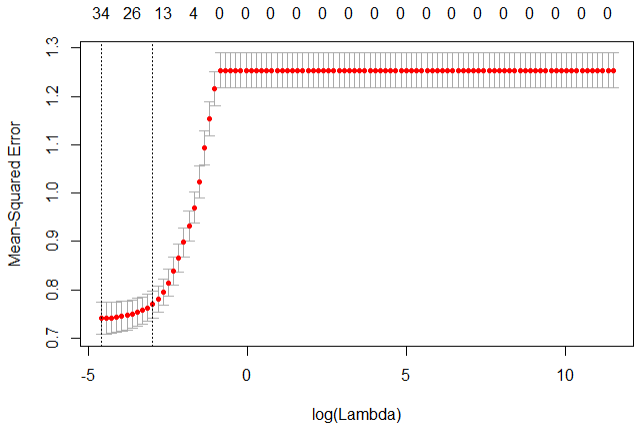


(c) Diferentes curvas del modelo



(d) Error de cada uno del parámetro lambda

El segundo que se va a construir será un modelo de regresión Lasso[5] que también será optimizado para obtener el mejor valor del parámetro lambda a utilizar, para minimizar el error MSE con R Markdown. Se puede observar en la figura mostrada a continuación que el mejor valor del parámetro lambda arroja un error MSE de 0,7100814, con lo cual se obtienen resultados ligeramente peores que con el anterior modelo:



En ambos modelos se puede observar como la popularidad del primer actor y del director, junto con su número de películas es importante a la hora de predecir los resultados. También como la duración de las películas influye en gran medida en su calidad y como ciertos géneros como el de guerra, drama, crimen, animación o noticias, pueden llegar a explicar los resultados.

6. Creación de un sistema de recomendación

Por último vamos a crear un sencillo recomendador en base al trabajo de [7]. Este recomendador se basa en el cálculo de la similitud de un conjunto de películas con una dada y teniendo en cuenta una serie de atributos:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\parallel A \parallel \parallel B \parallel} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{2}$$

En nuestro caso dado el título de una película, vamos a obtener aquellas películas que tengan en común algún actor o el director, y calcularemos su similitud teniendo en cuenta la puntuación en IMDb y su género, como se puede ver en R Markdown:

	imdb_score <dbl>	movie_title <fctr>	title_year <int>	sim <dbl>
126	7.2	The Matrix Reloaded	2003	0.9994619
124	6.7	The Matrix Revolutions	2003	0.9989014
62	5.4	Jupiter Ascending	2015	0.9798920
4177	7.1	My Own Private Idaho	1991	0.9773975
4822	8.1	Nothing But a Man	1964	0.9723357
2012	7.7	Dangerous Liaisons	1988	0.9708063
2771	7.7	Dangerous Liaisons	1988	0.9708063
4330	7.1	River's Edge	1986	0.9680281
876	5.6	Chain Reaction	1996	0.9676996
3452	7.0	The Neon Demon	2016	0.9674970

1-10 of 30 rows — Previous 1 2 3 Next

	imdb_score <dbl>	movie_title <fctr>	title_year <int>	sim <dbl>
232	6.1	Speed Racer	2008	0.9674015
1587	7.2	Speed	1994	0.9662175
1277	6.7	Sweet November	2001	0.9657648
975	6.5	The Replacements	2000	0.9644804
3100	6.9	Bill & Ted's Excellent Adventure	1989	0.9638154
2132	6.3	Hardball	2001	0.9630776
2221	6.8	Street Kings	2008	0.9629427
825	7.5	The Devil's Advocate	1997	0.9617314
1201	7.5	Bram Stoker's Dracula	1992	0.9617314
466	5.5	The Day the Earth Stood Still	2008	0.9611390

11-20 of 30 rows — Previous 1 2 3 Next

	imdb_score <dbl>	movie_title <fctr>	title_year <int>	sim <dbl>
4474	5.5	The Day the Earth Stood Still	2008	0.9611390
3367	7.4	Much Ado About Nothing	1993	0.9610696
3942	5.8	The Last Time I Committed Suicide	1997	0.9589497
85	6.3	47 Ronin	2013	0.9579403
2307	7.1	A Scanner Darkly	2006	0.9566994
1239	6.8	The Lake House	2006	0.9565035
450	6.7	Something's Gotta Give	2003	0.9556285
3838	7.4	Bound	1996	0.9528565
2194	6.2	Bill & Ted's Bogus Journey	1991	0.9457170
1529	5.3	The Watcher	2000	0.9234804

21-30 of 30 rows — Previous 1 2 3 Next

En el ejemplo anterior se pueden ver los resultados de nuestro recomendador para la película "The Matrix", en el que se muestra por order de similitud, las películas recomendadas con su título y año.

7. Conclusiones y próximos pasos

En este trabajo hemos hemos analízado una base de datos de películas en IMDb para la creación de un modelo no supervisado y dos modelos de regresión supervisados. También se ha creado un sencillo recomendador que en base al título de una película, calcula las películas semejantes con algún actor o el director en común, en base a su puntuación en IMDb y a su genéro.

Se ha conseguido predecir la puntuación de las películas con un margen de error relativamente pequeño, utilizando parte de los atributos proporcionados. Sería interesante en primer lugar completar aquella información en el conjunto de datos que no está disponible y en segundo lugar mejorar ciertos datos como el presupuesto y la recaudación, ya que la moneda utilizada no se ha unificado y tampoco se ha tenido en cuenta la inflación. Además, para películas antiguas y determinados países, parece que también faltan datos.

Con respecto al recomendador sólo se han tenido en cuenta películas con un reparto similar, pero también se podrían tener en cuenta el resto de atributos disponibles en el conjunto de datos.

Referencias

[1] C. Sun, *IMDB 5000 Movie Dataset*, 5000+ movie data scraped from IMDB website, Kaggle, 2016.

[2] Z. Brown, How many films are produced each year?, Quora, 2016.

[3] M. Lanzagorta, Horror Cinema By the Numbers, PopMatters, 2007.

[4] Z. Brown, Watching all the movies ever made, Justgeek, 2014.

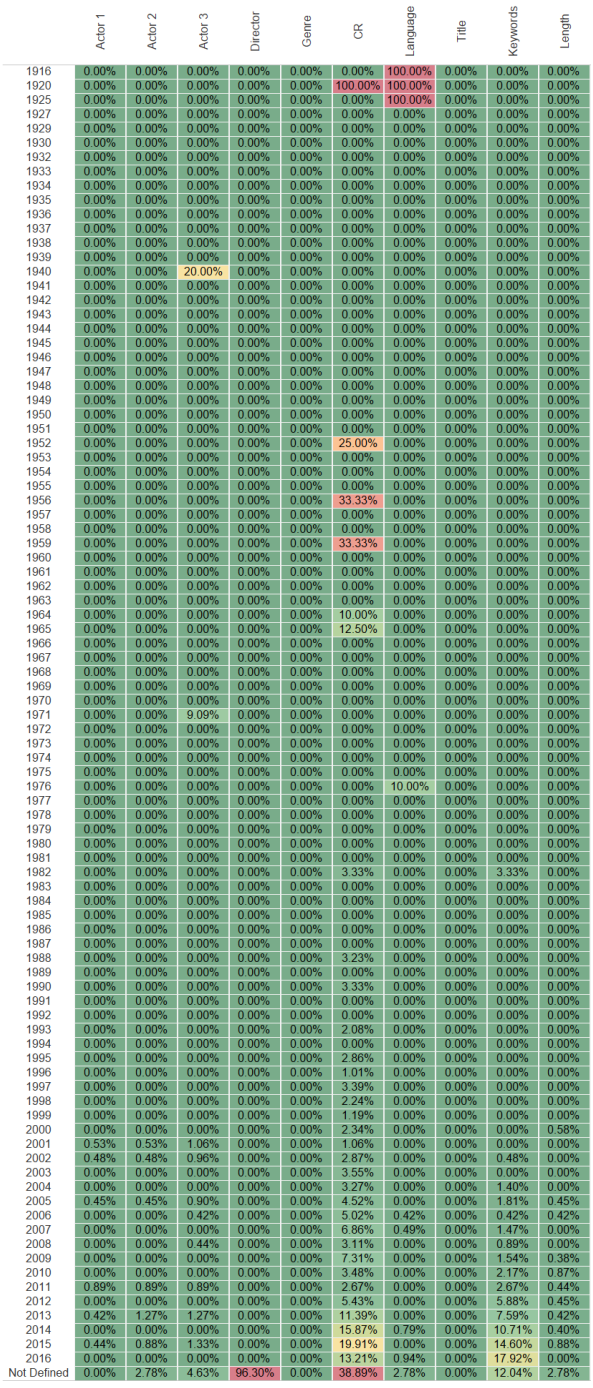
[5] T. Hastie and J. Qian, Glmnet Vignette, Stanford University, 2014.

[6] A. Sharma, Movie Recommendations, Kaggle, 2017.

[7] M. Chaware, Basic Recommender System using IMDb Data, Just Another Data Blog, 2015.

A. Calidad del dataset

Datos disponibles en Tableau Public:



	Actor 1	Actor 2	Actor 3	Director	Genre	CR	Language	Title	Keywords	Length
1916	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
1920	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%
1925	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
1927	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1929	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1930	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1932	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1933	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1934	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1935	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1936	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1937	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1938	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1939	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1940	0.00%	0.00%	20.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1941	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1942	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1943	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1944	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1945	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1946	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1947	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1948	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1949	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1950	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1951	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1952	0.00%	0.00%	0.00%	0.00%	0.00%	25.00%	0.00%	0.00%	0.00%	0.00%
1953	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1954	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1955	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1956	0.00%	0.00%	0.00%	0.00%	0.00%	33.33%	0.00%	0.00%	0.00%	0.00%
1957	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1958	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1959	0.00%	0.00%	0.00%	0.00%	0.00%	33.33%	0.00%	0.00%	0.00%	0.00%
1960	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1961	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1962	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1963	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1964	0.00%	0.00%	0.00%	0.00%	0.00%	10.00%	0.00%	0.00%	0.00%	0.00%
1965	0.00%	0.00%	0.00%	0.00%	0.00%	12.50%	0.00%	0.00%	0.00%	0.00%
1966	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1967	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1968	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1969	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1970	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1971	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1972	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1973	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1974	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1975	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1976	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.00%	0.00%	0.00%	0.00%
1977	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1978	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1979	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1980	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1981	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1982	0.00%	0.00%	0.00%	0.00%	0.00%	3.33%	0.00%	0.00%	3.33%	0.00%
1983	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1984	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1985	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1986	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1987	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1988	0.00%	0.00%	0.00%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%	0.00%
1989	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1990	0.00%	0.00%	0.00%	0.00%	0.00%	3.33%	0.00%	0.00%	0.00%	0.00%
1991	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1992	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1993	0.00%	0.00%	0.00%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%	0.00%
1994	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1995	0.00%	0.00%	0.00%	0.00%	0.00%	2.86%	0.00%	0.00%	0.00%	0.00%
1996	0.00%	0.00%	0.00%	0.00%	0.00%	1.01%	0.00%	0.00%	0.00%	0.00%
1997	0.00%	0.00%	0.00%	0.00%	0.00%	3.39%	0.00%	0.00%	0.00%	0.00%
1998	0.00%	0.00%	0.00%	0.00%	0.00%	2.24%	0.00%	0.00%	0.00%	0.00%
1999	0.00%	0.00%	0.00%	0.00%	0.00%	1.19%	0.00%	0.00%	0.00%	0.00%
2000	0.00%	0.00%	0.00%	0.00%	0.00%	2.34%	0.00%	0.00%	0.00%	0.58%
2001	0.53%	0.53%	1.06%	0.00%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%
2002	0.48%	0.48%	0.96%	0.00%	0.00%	2.87%	0.00%	0.00%	0.48%	0.00%
2003	0.00%	0.00%	0.00%	0.00%	0.00%	3.55%	0.00%	0.00%	0.00%	0.00%
2004	0.00%	0.00%	0.00%	0.00%	0.00%	3.27%	0.00%	0.00%	1.40%	0.00%
2005	0.45%	0.45%	0.90%	0.00%	0.00%	4.52%	0.00%	0.00%	1.81%	0.45%
2006	0.00%	0.00%	0.42%	0.00%	0.00%	5.02%	0.42%	0.00%	0.42%	0.42%
2007	0.00%	0.00%	0.00%	0.00%	0.00%	6.86%	0.49%	0.00%	1.47%	0.00%
2008	0.00%	0.00%	0.44%	0.00%	0.00%	3.11%	0.00%	0.00%	0.89%	0.00%
2009	0.00%	0.00%	0.00%	0.00%	0.00%	7.31%	0.00%	0.00%	1.54%	0.38%
2010	0.00%	0.00%	0.00%	0.00%	0.00%	3.48%	0.00%	0.00%	2.17%	0.87%
2011	0.89%	0.89%	0.89%	0.00%	0.00%	2.67%	0.00%	0.00%	2.67%	0.44%
2012	0.00%	0.00%	0.00%	0.00%	0.00%	5.43%	0.00%	0.00%	5.88%	0.45%
2013	0.42%	1.27%	1.27%	0.00%	0.00%	11.39%	0.00%	0.00%	7.59%	0.42%
2014	0.00%	0.00%	0.00%	0.00%	0.00%	15.87%	0.79%	0.00%	10.71%	0.40%
2015	0.44%	0.88%	1.33%	0.00%	0.00%	19.91%	0.00%	0.00%	14.60%	0.88%
2016	0.00%	0.00%	0.00%	0.00%	0.00%	13.21%	0.94%	0.00%	17.92%	0.00%
Not Defined	0.00%	2.78%	4.63%	96.30%	0.00%	38.89%	2.78%	0.00%	12.04%	2.78%

	Actor 1 Likes	Actor 2 Likes	Actor 3 Likes	Director Likes	Cast Likes	Critic Reviews	Movie Likes	User Reviews	Votes	Faces	Rating	Film Budget	Film Gross
1916	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1920	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1925	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1927	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1929	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%	0.00%
1930	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1932	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1933	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%
1934	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1935	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1936	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%
1937	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%	50.00%
1938	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1939	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	33.33%
1940	0.00%	0.00%	20.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	20.00%	60.00%
1941	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1942	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%	50.00%
1943	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1944	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1945	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	25.00%	100.00%
1946	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	33.33%
1947	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	66.67%
1948	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	33.33%	66.67%
1949	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1950	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1951	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1952	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	25.00%	75.00%
1953	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%
1954	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	60.00%
1955	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%	100.00%
1956	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1957	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%
1958	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
1959	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	66.67%
1960	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	66.67%
1961	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	80.00%
1962	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	75.00%
1963	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	62.50%
1964	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	40.00%
1965	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	12.50%	37.50%
1966	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	83.33%
1967	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	25.00%	75.00%
1968	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.09%	81.82%
1969	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.00%	70.00%
1970	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	66.67%
1971	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.09%	63.64%
1972	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	77.78%
1973	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	44.44%
1974	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	22.22%
1975	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%
1976	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.00%	80.00%
1977	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	56.25%
1978	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.25%	37.50%
1979	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	12.50%	56.25%
1980	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	8.33%	37.50%
1981	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.03%	48.48%
1982	0.00%	0.00%	0.00%	0.00%	0.00%	3.33%	0.00%	0.00%	0.00%	0.00%	0.00%	3.33%	46.67%
1983	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.09%	36.36%
1984	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	25.81%
1985	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	48.28%
1986	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.85%
1987	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.13%	3.13%
1988	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.23%
1989	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1990	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.33%	6.67%
1991	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.23%
1992	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.94%	0.00%
1993	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.08%	4.17%
1994	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.85%	1.85%
1995	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.29%	1.43%
1996	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.01%	4.04%
1997	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	8.47%	1.69%
1998	0.00%	0.00%	0.00%	0.00%	0.00%	1.49%	0.00%	0.00%	0.00%	0.00%	0.00%	8.21%	3.73%
1999	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.38%	1.19%
2000	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.51%	2.34%
2001	0.53%	0.53%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.19%	0.00%
2002	0.48%	0.48%	0.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.35%	4.31%
2003	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.69%	5.92%
2004	0.00%	0.00%	0.00%	0.00%	0.00%	0.93%	0.00%	0.00%	0.00%	0.00%	0.00%	7.94%	6.54%
2005	0.45%	0.45%	0.90%	0.00%	0.00%	0.90%	0.00%	0.00%	0.00%	0.00%	0.00%	8.14%	8.14%
2006	0.00%	0.00%	0.42%	0.00%	0.00%	0.42%	0.00%	0.00%	0.00%	0.00%	0.00%	10.46%	9.62%
2007	0.00%	0.00%	0.00%	0.00%	0.00%	1.47%	0.00%	0.98%	0.00%	0.00%	0.00%	9.80%	12.25%
2008	0.00%	0.00%	0.44%	0.00%	0.00%	0.44%	0.00%	0.44%	0.00%	0.00%	0.00%	10.67%	10.67%
2009	0.00%	0.00%	0.00%	0.00%	0.00%	1.54%	0.00%	0.38%	0.00%	0.00%	0.00%	16.92%	17.31%
2010	0.00%	0.00%	0.00%	0.00%	0.00%	0.87%	0.00%	0.43%	0.00%	0.43%	0.00%	14.35%	13.48%
2011	0.89%	0.89%	0.89%	0.00%	0.00%	0.89%	0.00%	0.00%	0.00%	0.00%	0.00%	8.89%	15.11%
2012	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.90%	0.00%	10.41%	18.10%
2013	0.42%	1.27%	1.27%	0.00%	0.00%	1.69%	0.00%	0.84%	0.00%	0.42%	0.00%	9.70%	21.10%
2014	0.00%	0.00%	0.00%	0.00%	0.00%	2.38%	0.00%	0.79%	0.00%	1.19%	0.00%	9.13%	33.73%
2015	0.44%	0.88%	1.33%	0.00%	0.00%	3.98%	0.00%	2.21%	0.00%	1.33%	0.00%	8.85%	36.73%
2016	0.00%	0.00%	0.00%	0.00%	0.00%	1.89%	0.00%	0.94%	0.00%	2.83%	0.00%	12.26%	30.19%
Not Defined	0.00%	2.78%	4.63%	96.30%	0.00%	8.33%	0.00%	5.56%	0.00%	0.00%	0.00%	92.59%	97.22%