Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Análisis y creación de una tabla analítica de datos con información de películas de IMDB

Proyecto Final del Módulo 1 - Diplomado en Ciencia de Datos

Autor:
Juan Manuel Zapata López

Febrero 2021

Índice

1. Introducción

La historia del cine comenzó hace más de 200 años, cuando en 1895 se proyectó el primer filme en París. Desde entonces ha experimentado una serie de cambios en varios sentidos. Por un lado, la tecnología del cinematógrafo ha evolucionado mucho, desde sus inicios con el cine mudo de los hermanos Lumière hasta el cine digital del siglo XXI. Por otro lado, ha evolucionado el lenguaje cinematográfico, incluidas las convenciones del género, y han surgido así distintos géneros cinematográficos.

Estos dos hechos, han logrado que miles de personas alrededor del mundo se apasionen por este arte. Actualmente, según la base de datos IMBd, se han realizado más de 3 millones de títulos, de ellos 350,000 son películas y el resto son capítulos de series de televisión.

Ante esta inmensidad de contenido, las personas han encontrado una forma de otorgar una calificación a cada filme y con ello hacer qué más personas vean una película (si ésta tiene buenas notas y comentarios) o evitar que un filme tenga espectadores (cuando tiene malas notas).

Aunque existen diversas plataformas online para esta actividad, una de las más populares actualmente es IMDb, un sitio web en el que cualquier persona puede encontrar miles de reseñas de películas y calificación de las que ya la vieron.

Cada usuario puede otorgar a alguna película un voto de entre 1 y 10 puntos, con los votos de todos los usuarios se obtiene un promedio que se considera la nota. Películas con una nota alta generan que más personas quieran verla y cuando tienen menores notas, ocurre lo contrario.

si bien, es difícil cuantificar con una calificación lo buena o mala que puede ser una película, lo cierto es que miles de personas en cada rincón del mundo ocupan este sitio para decidir cuál sera la siguiente cinta con la que se entretendrán.

1.1. Descripción del dataset

En este contexto, surge la curiosidad de investigar que sucede actualmente en el sitio web. A través de la página https://www.kaggle.com/ se enontró una base de datos con características de cerca de 90 mil películas distintas producidas alrededor del mundo, con una pequeña reseña y la calificación promedio de los usuarios en iMDb.

Estos filmes han sido grabados en distintos años y por distintas compañías. Se pueden encontrar títulos en más de 5 idiomas.

1.2. Objetivo

El objetivo del presente texto, es aplicar las técnicas de limpieza de datos, hacer ingeniería de variables, detectar de outliers, imputar missings y aplicación algunas otrás técnicas para hacer un análisis exploratorio del dataset y obtener resultados interesantes.

También, se busca preparar una tabla analítica de datos con la que se pueda hacer inferencia sobre la calificación de una película basada en sus características.

2. Data Set

2.1. Descripción de la información

En el dataset que se seleccionó para analizar, encontramos inicialmente un total de 85,855 registros de películas con 22 variables que se describen a continuación.

Variable	Tipo	Descripción
imdb_title_id	Alfanúmerica	Id que identifica la pelicula en la plataforma IMDb
original_title	Texto	Título original de la película
year	Número entero	Año de estreno
date_published	Fecha	Fecha de estreno
genre	Texto	Género
duration	Número entero	Duración en minutos
country	Texto	País de origen
language	Texto	Idioma
director	Texto	Nombre del director o directores
writer	Texto	Nombre del escritor o escritores
production_company	Texto	Nombre de la compañía que produjo la película
actors	Texto	Diferentes actores que participaron, separados por coma
description	Texto	Descripción de la trama
avg_vote	Número flotante	Promedio de los votos recibidos
votes	Número entero	Número de votos recibidos
budget	Número flotante	Presupuesto (en distintas monedas)
usa_gross_income	Número flotante	Ingresos brutos generados en Estados Unidos
worlwide_gross_income	Número flotante	Ingresos brutos generados en todo el mundo
metascore	Número flotante	Rating de metascore
reviews_from_users	Número entero	Número de reseñas de usuarios
reviews_from_critics	Número entero	Número de reseñas de críticos

Cuadro 1: Descripción de variables

Para hacer más sencilla la identificación de cada tipo de variable, se ha añadido un prefijo al nombre de cada una de la siguiente forma:

Prefijo	Tipo	Variables
c_	Variable continua	year, duration, avg_vote, votes,budget, usa_gross_income, worlwide_gross_income, metascore,reviews_from_users, reviews_from_critics
v_	Variable nominal u ordinal	country, genre, language
d_	Variable de tipo fecha	date_published
t_	Variable de tipo texto	imdb_title_id, title, original_title, director, writer, production_company, actors, description

Cuadro 2: Clasificación de variables

3. Calidad de datos

3.1. Completitud

La completitud es la cantidad de registros que no tienen missings expresada en porcentaje. Para esta tabla, se calculó la completitud de cada variable y se obtuvo la siguiente gráfica.

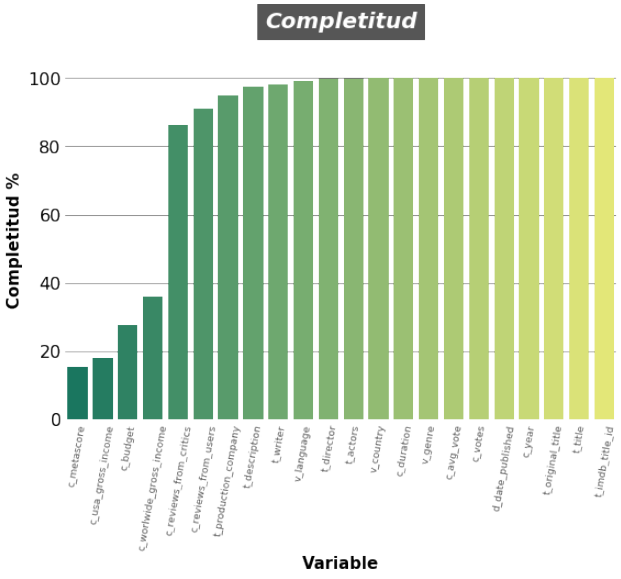

Figura 1: Completitud de las variables

Las variables con completitud menor a 80 % fueron eliminadas pues no son susceptibles a ser imputadas, ya que se podría modificar la distribución de las variables e inducir a resultados imprecisos en el modelo.

3.2. Duplicados

Se buscaron líneas duplicadas en la tabla y también registros duplicados por ID, pero no se encontraron.

3.3. Limpieza de variables

A continuación se describen las modificaciones realizadas para corregir el contenido de las variables y asegurar que se satisfagan las seis dimensiones de la calidad de datos: completitud, conformidad, consistencia, duplicación, integridad y precisión. Dos de estos puntos ya fueron tratados con anterioridad, nos enfocaremos en los cuatro restantes.

A continuación se presentan los tratamientos realizados para cada variable en la que se identificó algo erróneo en cuanto a la calidad.

- d_date_published: Se exploró con expresiones regulares que todo su contenido se acoplara a un formato de fecha, se encontraron 4,563 registros que sólo contenían el año, pero se decidió conservarlos (colocando 1 de enero). Adicionalmente se encontró un registro con formato de texto que fue eliminado:

t_imdb_title_id	t_title	t_original_title	c_year	d_date_published
tt8206668	Bad Education	Bad Education	TV Movie 2019	TV Movie 2019

Figura 2: registro no válido

- c_year: Esta variable se recalculó a partir del contenido de la variable d_date_published para asegurar su integridad.
- variables continuas: Se verificó que fueran de tipo int o float y se hizo un resumen, obteniendo lo siguiente:

	c_year	c_duration	c_avg_vote	c_votes	c_reviews_from_users	c_reviews_from_critics
count	85854.000000	85854.000000	85854.000000	8.585400e+04	78257.000000	74057.000000
mean	1993.897139	100.351329	5.898642	9.493321e+03	46.039690	27.479036
std	24.168842	22.553964	1.234988	5.357465e+04	178.512268	58.338977
min	1894.000000	41.000000	1.000000	9.900000e+01	1.000000	1.000000
25%	1979.000000	88.000000	5.200000	2.050000e+02	4.000000	3.000000
50%	2003.000000	96.000000	6.100000	4.840000e+02	9.000000	8.000000
75%	2013.000000	108.000000	6.800000	1.766000e+03	27.000000	23.000000
max	2021.000000	808.000000	9.900000	2.278845e+06	10472.000000	999.000000

Figura 3: resumen de variables continuas

- Variables de tipo texto: En este caso se limpió su contenido de tal forma que se eliminaron caracteres especiales, se dejó el texto en minúsculas, no se removieron números ni comas (puesto que algunas variables separan elementos con ellas y esto se usará para hacer ingenierías de variables posteriormente). Después, se removieron stopwords, es importante mencionar que dada la naturaleza del dataset, éstas variables se encuentran en varios idiomas por lo cual se decidió eliminar stopwords del inglés, español, alemán, italiano, francés y portugués. Como se observa en la figura 1, de la página 4, algunas de éstas variables contienen missings, se tuvo cuidado de no reescribirlos por error.

4. Análisis exploratorio

Con el objetivo de entender el contenido de la base, se analizó de forma gráfica cada una de las variables. Se inició haciendo gráficas de nubes de palabras para las variables más relevantes de tipo texto.

Para el título de las películas, observamos que la palabras más frecuentes (ya traducidas al español) son amor, chica, noche, muerte, vida, día, casa, etc. Con esto nos damos cuenta de que títulos son los más comerciales (pues muchas personas deciden ver películas si el título les parece atractivo).

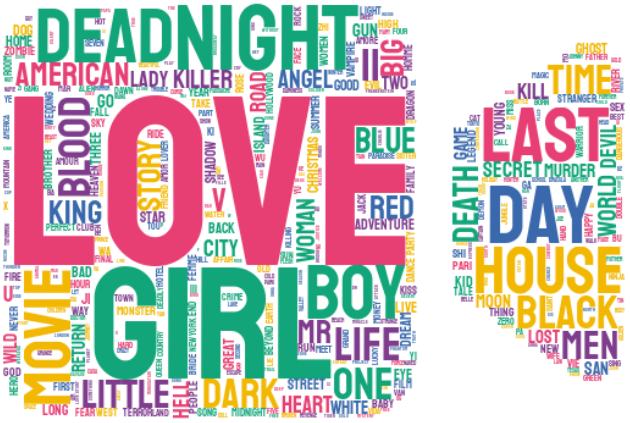

Figura 4: Nube de palabras para los títulos

También se analizó con la misma técnica a la variable que describe el contenido de los filmes, y se detectó que muchas películas hablan sobre la familia, encontrar el amor, el sentido de la vida, conocer amigos o a una pareja, sobre descubrir y encontrar cosas y sentimientos, etc. Es decir, muchas películas se apegan a los sentimientos de los espectadores.

Figura 5: Nube de palabras para las descripciones

Se encontró que las 10 compañías con más producciones realizadas son las siguientes, (sin embargo este top 10 representa sólo el 8,5 % de los registros, con lo cuál podemos deducir que existe una gran cantidad de compañías presentes en esta tabla).

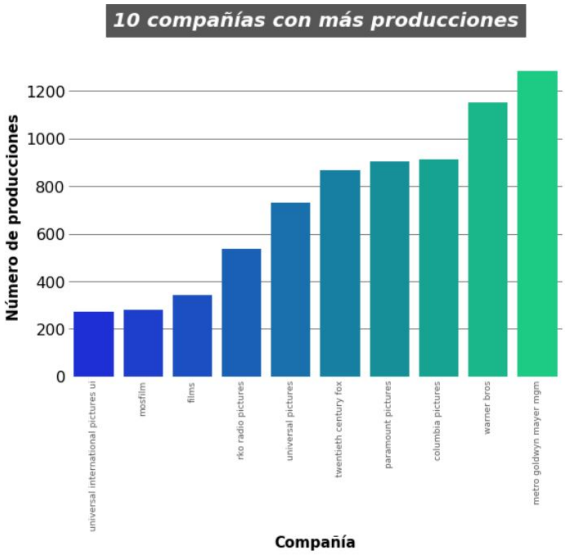

Figura 6: Compañías con más producciones

Si vemos la gráfica de la distribución del año de estreno, podemos notar que tenemos películas desde finales del siglo XIX, es decir, desde los comienzos de la industria cinematográfica. Sin embargo, la mayoría de registros son de los últimos 20 años.

Figura 7: Distribución del año de estreno

Aprovechando que se está realizando el análisis exploratorio de las variables, se crearon las siguientes variables a partir de la fecha de estreno: día de la semana (lunes - domingo), día del mes (1-31) y si es fin de semana (1 = Sí, 0 = No) y también vamos a explorarlas.

Como es de esperarse, la mayoría de películas se estrenaron en viernes, pues este día es en el que comúnmente, después de salir del trabajo o escuela, los grupos de personas se reúnen para realizar actividades juntos. De hecho, poco más de la mitad de los registros se estrenaron en jueves o viernes.

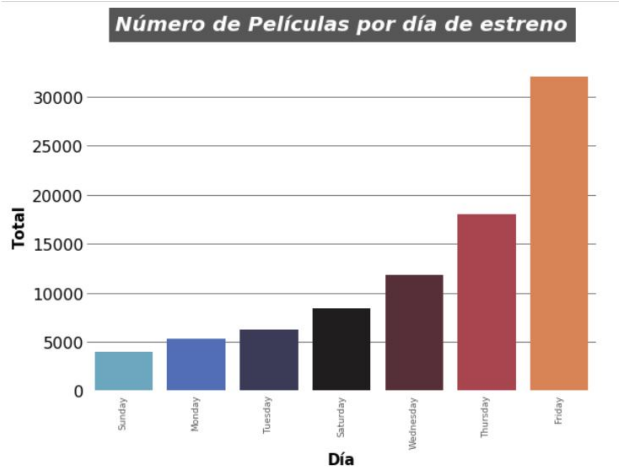

Figura 8: Número de estrenos por día

Del análisis anterior podemos deducir que no es muy popular estrenar una película en sábado o domingo, pero para ser más exactos se graficó esta situación y se obtuvo que sólo el 14.4 % de registros se estrenaron en estos dos días. Recordemos que 1 = Se estrenó en fin de semana y 0 en otro caso.

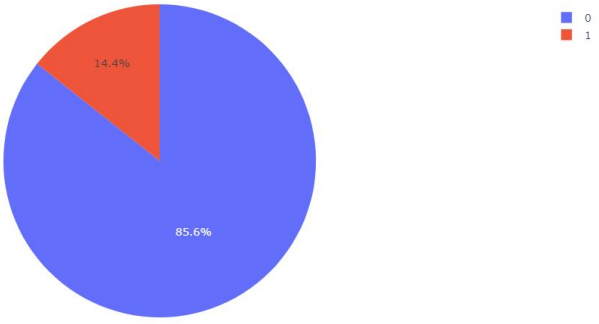

Figura 9: Estrenos en fin de semana

Si graficamos la distribución del día del estreno, percibimos que se comporta bastante uniforme durante los meses. Parece que los primeros días del mes tienen mayor distribución, pero esta situación se puede deber a que en la limpieza de la fecha había registros con solo el año y se decidió colocarles 1 de enero + año como su fecha de lanzamiento.

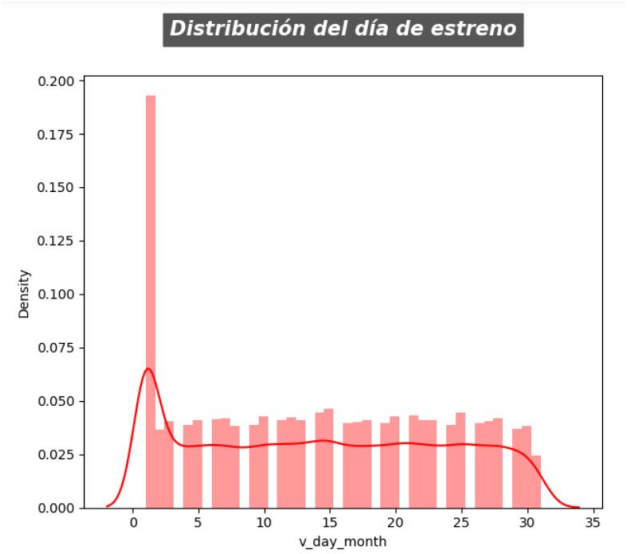

Figura 10: Distribución por día del mes de lanzamiento

Si pensamos en los géneros más frecuentes, al graficar percibimos que el drama, la comedia y la combinación de estos son los más frecuentes en nuestra tabla. Quizás este hecho se deba a que son los géneros que causan más ganancias, al ser los preferidos por el público alrededor de todo el mundo.

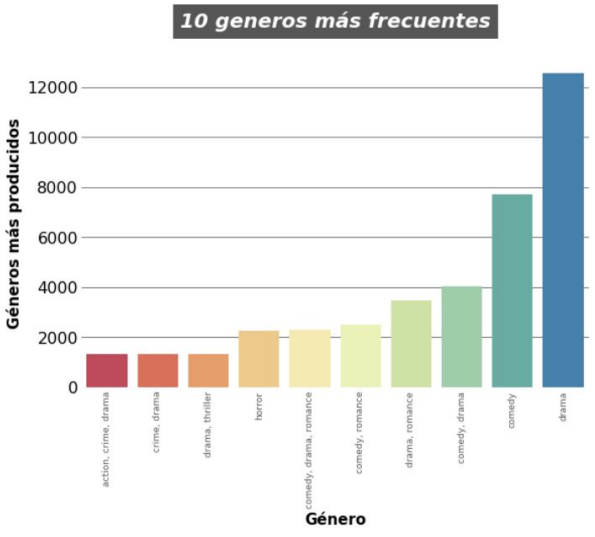

Figura 11: Géneros más populares

Popularmente se sabe que las películas duran al rededor de 120 minutos, sin embargo, existen casos donde la duración se prolonga mucho más allá de esto, se sabe que la película más larga del mundo tiene una duración de 5,220 minutos u 87 horas. Nuestra tabla también cuenta con esos registros extraños. A pesar de ello, notamos que la mayoría de películas dura menos de 200 minutos.

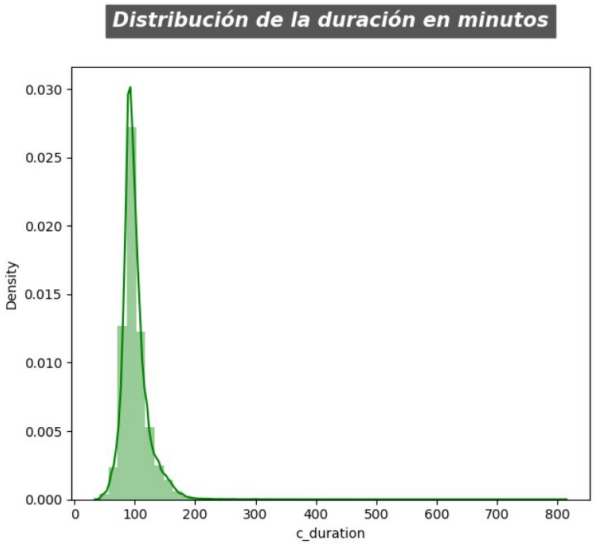

Figura 12: Duración en minutos

Sólo para ver mejor este hecho, se decidió hacer la gráfica en horas y sin considerar películas de duraciones extremas, se percibió que la mayoría de películas duran entre 1 y 2.5 horas.

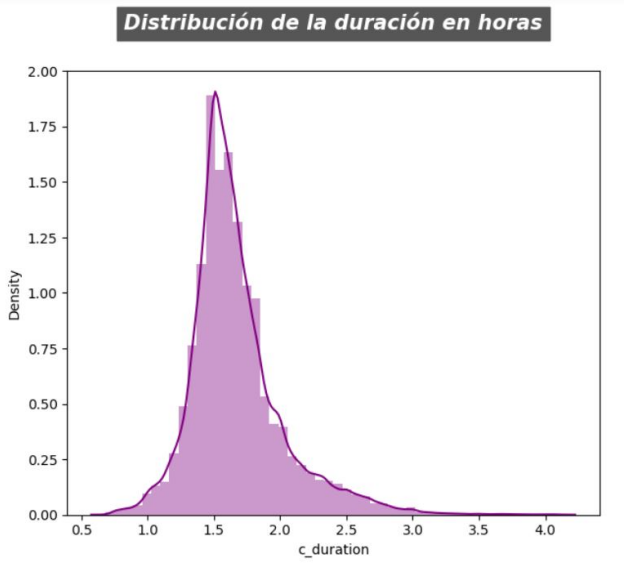

Figura 13: Duración en horas

Si ahora analizamos el país de origen de las películas, no es de sorprender que Estados Unidos sea el líder de la industria, pues ahí se encuentra el centro cinematográfico más famoso del mundo: Holliwood. Por el contrario, sí resulta sorprendente ver que la India está en segundo lugar y Reino Unido en tercero.

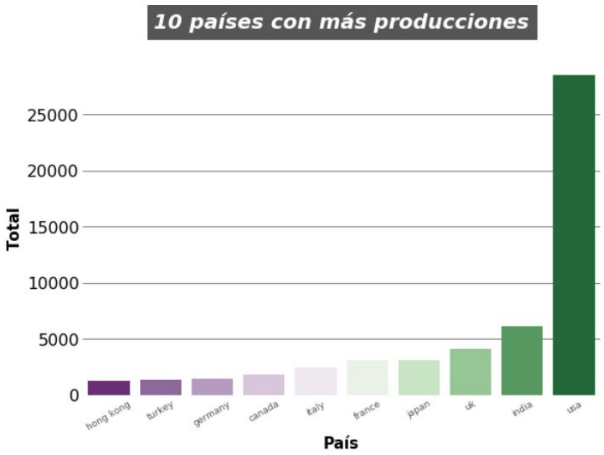

Figura 14: Países con más películas producidas

Relacionado a este último hecho, está que el inglés es el idioma original más frecuente, pues es el idioma universal por excelencia. Lo sigue el francés (conocido como el idioma del amor) y el español (que es el segundo idioma más hablado en el mundo).

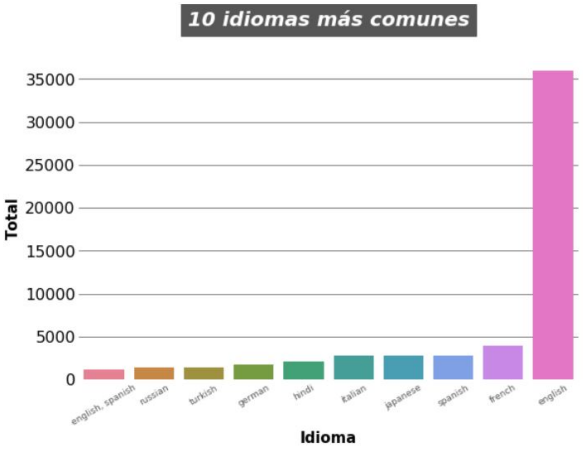

Figura 15: Idiomas más frecuentes

Finalmente, es interesante analizar el comportamiento de los votos de los usuarios de IMDb. Pesé a ser un sitio web muy popular, notamos que la mayoría de películas no recibieron más de mil votos.

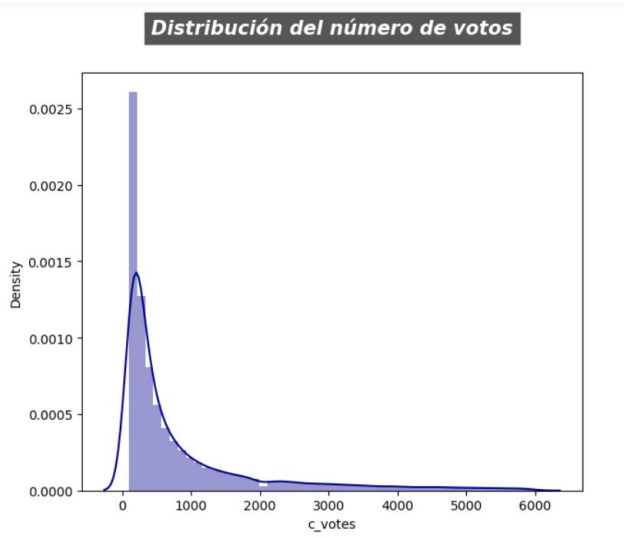

Figura 16: Distribución del número de votos

Y también se nota que las personas que votan, son duras con las calificaciones, pues el promedio se concentra al rededor de 6.

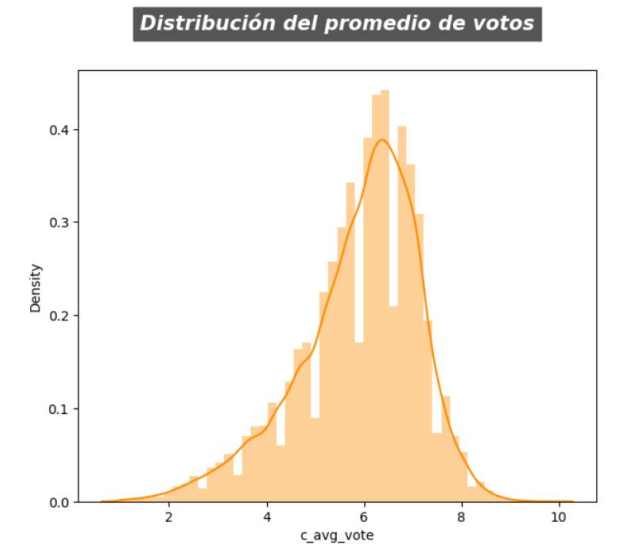

Figura 17: Distribución del promedio de voto

5. Ingeniería de variables

En esta sección se intentó sacar provecho del contenido de la tabla, obteniendo variables nuevas a partir de las existentes y con algunas técnicas de la ingeniería de variables.

La variable que contiene el título de la película (sin haber sufrido modificaciones por limpieza de texto), se dividió por palabras y con ello se obtuvieron las siguientes nuevas variables:

- c_num_words : Número de palabras presentes en el título
- c_num_letters : Número de letras presentes en el título
- c_long_avg_word : Promedio de letras por palabra en el título

En el caso de la variable género, recordemos que en la limpieza de texto no se eliminaron comas, esto es porque algunos registros tenían como valor más de un género (separados por comas), como se ilustra a continuación:

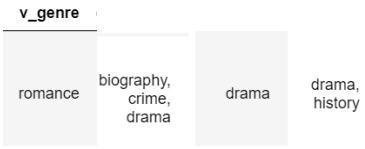

Figura 18: Ejemplo del contenido de la variable género

Se decidió hacer un corpus concatenando todo los valores y se obtuvo la frecuencia de las palabras, con ello notamos que en realidad todas las películas están clasificadas en 24 géneros y sus combinaciones. Con base en esta información, se determinó crear una variable dummy para cada género de la siguiente forma:

$$f_i(x) = \begin{cases} 1 & si\ el\ registro\ contiene\ el\ género\ i \\ 0 & en\ otro\ caso \end{cases} \qquad \forall x \in \{registros\}, i \in \{géneros\}$$

La forma de conseguir crear estas variables fue a través de sklearn.feature_extraction.text y la función CountVectorizer, que aunque no está diseñada para esto, funcionó muy bien.

Como la mayoría de películas provienen de USA, se creó la variable v_country_usa con los valores 1 = el país de origen es USA y 0 en otro caso. Similarmente, se creó la variable v _language_en con los valores 1 = la película está en inglés y 0 en otro caso.

Se creó la variable c_n_actors, contando la cantidad de actores presentes (y separados por coma) de la variable t_actors.

Dado que la variable v__week se encontraba en texto y en inglés, se transformó a tipo entero, donde mondey = 1, tuesday = 2, etc.

6. Identificación y remoción de outliers

En este caso, se creó una función que detecta outliers primero por rango intercuartil, después por percentiles .5 y .95 y finalmente por z-score. Para z-score, primero se aplicó una prueba de normalidad, si era rechazada se hacía una transformación de box-cox para intentar inducir la normalidad, si ésta era aceptada se hacía z-score, en otro caso, este método no se usaba.

Estos son los resultados:

Para la columna c_year se tiene: Número de outliers por IQR: 625. Número de outliers por percentiles .5 y .95: 5636. Los datos no distribuyen normal, se hizo una transformación de boxcox. Los datos transformados distribuyen normal. Número de outliers por z-score: 281. Hay 281 outliers comúnes = 0.003 %.

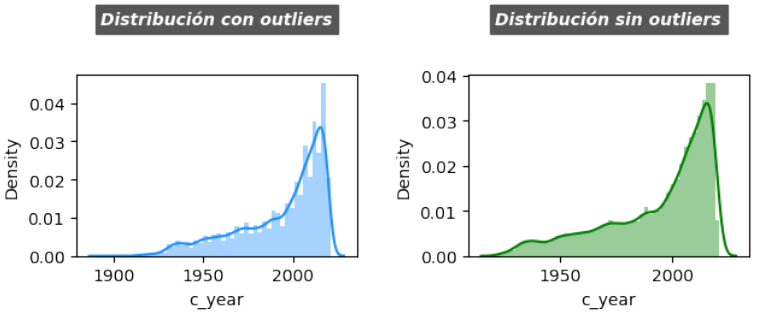

Figura 19: Variable c_year con y sin outliers

Para la columna c_duration se tiene: Número de outliers por IQR: 5752. Número de outliers por percentiles .5 y .95: 8190. Los datos no distribuyen normal, se hizo una transformación de boxcox. Los datos no se pudieron transformar a una normal, no se puede aplicar z-score. Hay 4807 outliers comúnes = 0.056 %

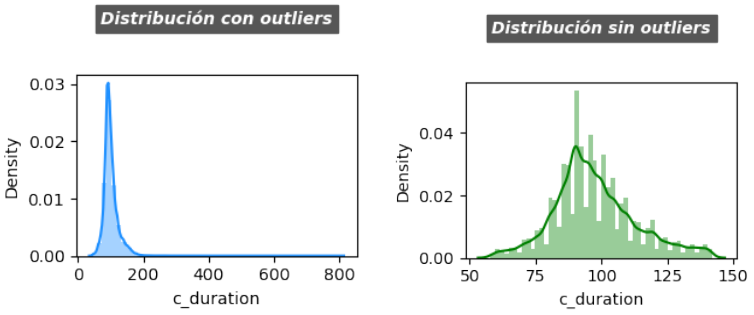

Figura 20: Variable c_duration con y sin outliers

Para la columna c_votes se tiene: Número de outliers por IQR: 12727. Número de outliers por percentiles .5 y .95: 7890. Los datos no distribuyen normal, se hizo una transformación de boxcox. Los datos no se pudieron transformar a una normal, no se puede aplicar z-score. Hay 4039 outliers comúnes = 0.05 %

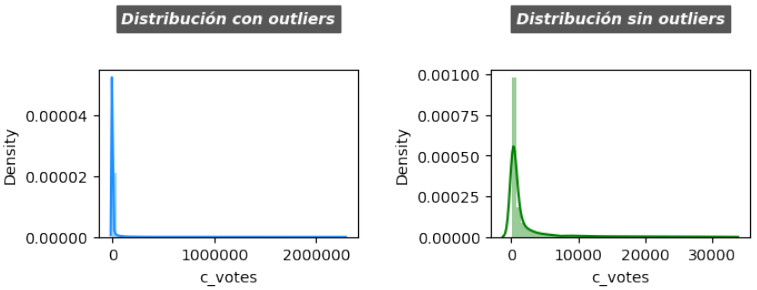

Figura 21: Variable c_votes con y sin outliers

Para la columna c_reviews_from_users se tiene: Número de outliers por IQR: 7170. Número de outliers por percentiles .5 y .95: 3482. Los datos distribuyen normal. Número de outliers por z-score: 882. Hay 882 outliers comúnes = 0.011 %

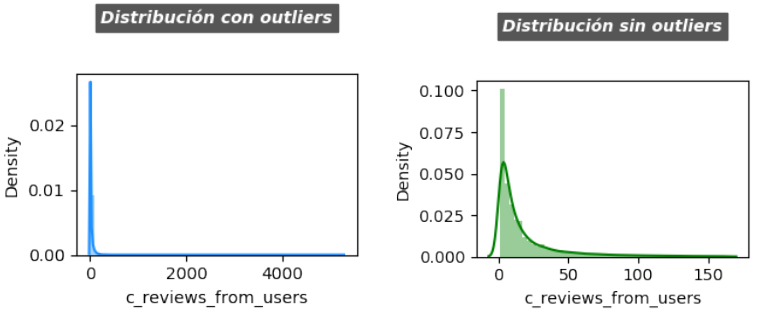

Figura 22: Variable c_reviews_from_users con y sin outliers

Para la columna c_reviews_from_critics se tiene: Número de outliers por IQR: 6036. Número de outliers por percentiles .5 y .95: 3253. Los datos distribuyen normal. Número de outliers por z-score: 1515. Hay 1515 outliers comúnes = 0.0192 %

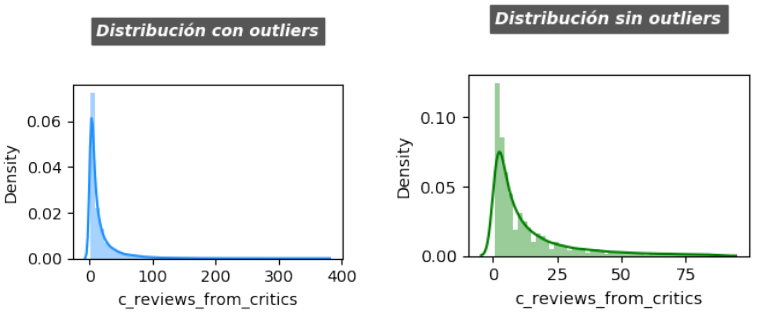

Figura 23: Variable c_reviews_from_critics con y sin outliers

Para la columna c_num_words se tiene: Número de outliers por IQR: 1035. Número de outliers por percentiles .5 y .95: 2067. Los datos no distribuyen normal, se hará una transformación de boxcox. los datos no se pudieron transformar a una normal, no se puede aplicar z-score. Hay 1035 outliers comúnes = 0.014 %

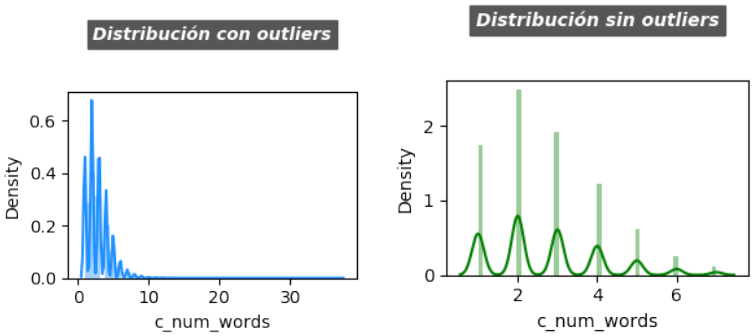

Figura 24: Variable c_num_words con y sin outliers

Para la columna c_num_letters se tiene: Número de outliers por IQR: 1651. Número de outliers por percentiles .5 y .95: 4775. Los datos no distribuyen normal, se hizo una transformación de boxcox. los datos no se pudieron transformar a una normal, no se puede aplicar z-score. Hay 1651 outliers comúnes = 0.023 %

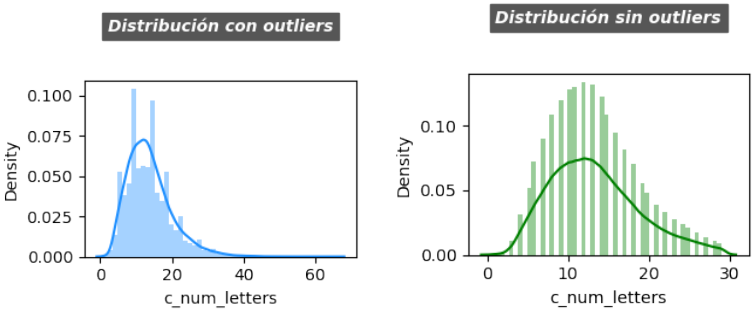

Figura 25: Variable c_num_letters con y sin outliers

Para la columna c_long_avg_word se tiene: Número de outliers por IQR: 2055. Número de outliers por percentiles .5 y .95: 5083. Los datos no distribuyen normal, se hará una transformación de boxcox. los datos no se pudieron transformar a una normal, no se puede aplicar z-score. Hay 2055 outliers comúnes = 0.029 %

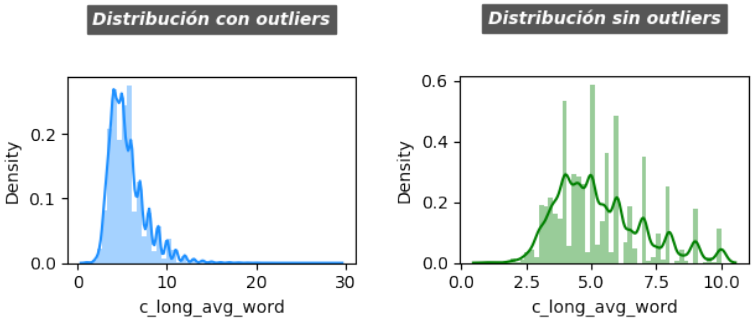

Figura 26: Variable c_long_avg_word con y sin outliers

Para la columna c_n_actors se tiene: Número de outliers por IQR: 8916. Número de outliers por percentiles .5 y .95: 3367. Los datos distribuyen normal. Número de outliers por z-score: 1311 Hay 1311 outliers comúnes = 0.019 %.

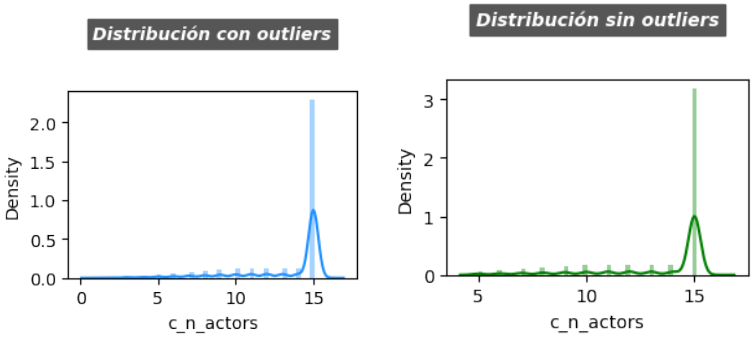

Figura 27: Variable c_n_actors con y sin outliers

Con estos cambios, la tabla redujo su tamaño al 80 %, lo cual es aceptable.

7. Tratamiento de missings

En este punto, tenemos las siguientes cinco variables con problemas de completitud:

	columna	total	completitud
0	c_reviews_from_critics	9337	86.325024
1	c_reviews_from_users	6173	90.959020
2	v_language_en	658	99.036293
3	v_country_usa	49	99.928235
4	c_n_actors	48	99.929699

Figura 28: Variables con missings

Al hacer un análisis exploratorio de las variables c_reviews_from_critics y c_reviews_from_users, notamos que se trata de valores que en realidad son cero, pues pese a recibir votos, no tuvieron reseñas, por esta razón se decidió imputarlos con 0.

Para las variables v_language_en y v_coutry_usa se tenía el problema de que provienen de ingenierías realizadas antes de limpiar missings. Para solucionar el problema, se eliminó su contenido, se recuperaron las variables originales (v_language, v_country) a través del id de los registros y se imputaron missings en ellas.

A partir de este punto, se dividió a la tabla en los conjuntos X_train y X_test para poder hacer las imputaciones.

- Para la variable v_language, se imputó con la moda (english), se realizó una prueba de chi cuadrado para comprobar que la distribución de las categorías se mantuviera y esta no fue rechazada.
- Para la variable v_country, se imputó con la moda (usa), se realizó una prueba de chi cuadrado para comprobar que la distribución de las categorías se mantuviera y esta no fue rechazada.
- Para la variable c_n_actors, al ser una variable continua, se intentó imputar con la media, moda y mediana y se aplicó una prueba de Kolmogorov Smirnof (que testea si la distribución se mantiene aún con la imputación). Se eligió a la moda = 15 como valor de imputación, pues con ella se tuvo un p-value mayor y un estadístico pequeño.

Las imputaciones se hicieron con información del X_train, pero también se aplicaron tanto al X_train como al X_test. Se verificó de nuevo que no hubiese variables con completitud menor a 100 y finalmente se volvieron a aplicar las ingenierías que generan las variables v_language_en y v_coutry_usa.

8. Reducción de dimensiones

En este punto, sólo contamos con variables de tipo numérico (aún siendo dummies). Para poder hacer la reducción, se creo una tabla auxiliar que junta la información del X_train y X_test. En esta tabla se hizó la reducción de dimensiones para finalmente replicar los resultados al X_train.

8.1. Filtro de baja varianza

Se obtuvo la varianza de cada variable, sin considerar la varianza de variables dummies, no se tenía alguna otra varianza baja, por lo cual ninguna variable fue eliminada en este filtro.

8.2. Filtro de alta correlación

Para iniciar, se graficó a la matriz de correlaciones de las variables explicativas, para tener una vista general de como se comportan las variables entre sí, ésta fue calculada con el método de pearson.

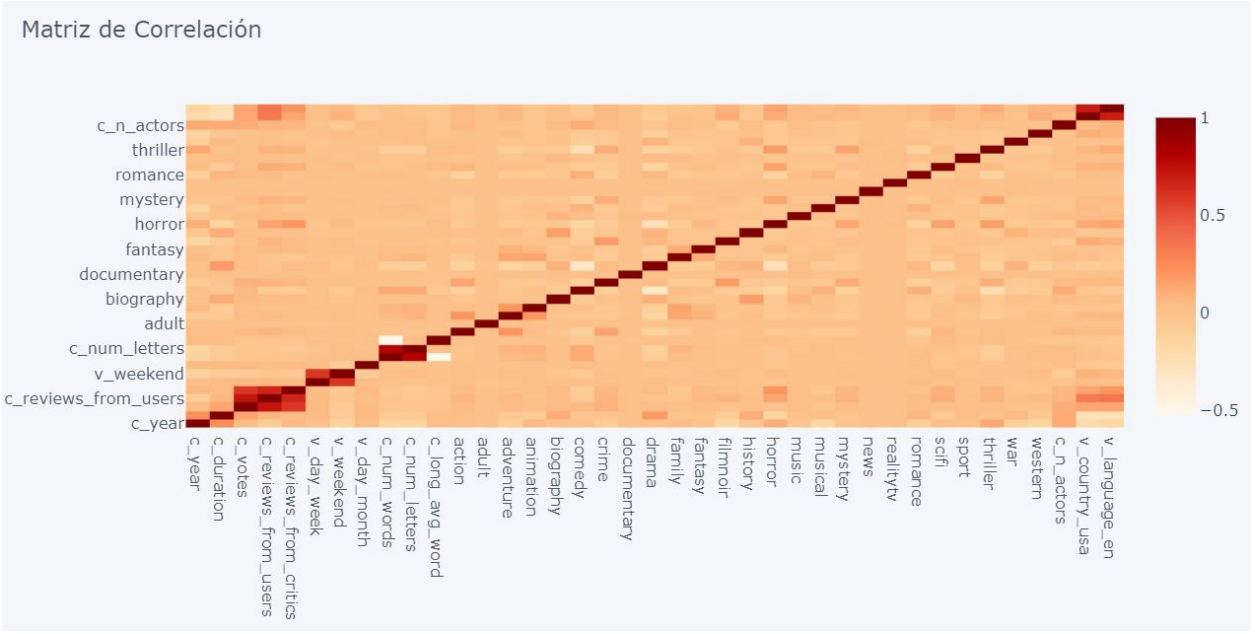

Figura 29: Matriz de correlaciones

Se aprecia que hay variables con alta correlación, por lo cual se procede a analizar cuales de ellas pueden ser eliminadas. Se obtienen las siguientes variables con una correlación en valor absoluto mayor a 0.5:

Variable 1	Variable 2	Correlación
c_num_letters	c_num_words	0.79748
c_votes	c_reviews_from_users	0.727953
v_country_usa	v_language_en	0.695458
c_reviews_from_critics	c_reviews_from_users	0.658746
v_day_week	v_weekend	0.598949
c_reviews_from_critics	c_votes	0.596924
c_long_avg_word	c_num_words	0.517699

Cuadro 3: Variables con alta correlación

Finalmente, se decide eliminar a las variables v_country_usa, c_num_letters,c_long_avg_word , v_weekend y c_votes. Esto se eligio teniendo en cuanta que las variables que no fueron eliminadas, pueden servir más al modelo para predecir. La causa de remover variables con este método, es que ay variables explicativas que al estar altamente correlaciones, en realidad pueden contener la misma información y tienen tendencias similares.

8.3. Filtro de baja correlación con la variable objetivo

Para las variables que aún tenemos en la tabla, se calculó su correlación de pearson con la variable objetivo, se tienen los siguiente resultados:

Variable	Corr con tgt	Variable	Corr con tgt	Variable	Corr con tgt
horror	0.3330	war	0.0733		
drama	0.3140	c_reviews_from_users	0.0652		
c_year	0.2591	filmnoir	0.0624	c_num_words	0.0255
v_language_en	0.2570	v_day_month	0.0513	mystery	0.0222
c_duration	0.2318	v_day_week	0.0494	western	0.0161
scifi	0.1724	c_n_actors	0.0462	sport	0.0111
thriller	0.1578	musical	0.0405	realitytv	0.0109
action	0.1286	music	0.0400	family	0.0076
c_reviews_from_critics	0.1090	fantasy	0.0362	documentary	0.0067
romance	0.1062	adventure	0.0353	adult	0.0055
biography	0.0840	animation	0.0348	news	0.0019
history	0.0800	crime	0.0335	comedy	0.0008

Figura 30: Correlaciones con la variable objetivo

Se observa que en realidad son muchas las variables que presentan baja correlación con la c_avg_vote, se decide eliminar aquellas cuya correlación sea menor a 0.03.

8.4. Multicolinealidad

Para ver si existen variables que puedan inducir a problemas de multicolinealidad, se calculó el factor de inflación de la varianza, a continuación se presentan las variables con $VIF > 5$:

	variables	VIF
0	c_year	85.713362
1	c_duration	51.604949
4	v_day_week	9.024287
22	c_n_actors	30.426688

Figura 31: Variables con $VIF > 5$

Se eliminó a las variables c_year y c_duration por se las dos con mayor VIF, sin embargo, sería buena idea no eliminarlas porque dada su naturaleza, pueden ser útiles en el modelo predictivo. Lo ideal sería correr un modelo con ellas y sin ellas para ver cuál resulta mejor.

8.5. Análisis de componentes principales

Antes de iniciar con esta sección y como las variables eliminadas anteriormente se hicieron en una tabla auxiliar, se hizo un listado con todas las columnas removidas y se eliminaron tanto de X_train como de X_test. También, se separó a la variable objetivo de cada uno de estos conjuntos para que al estandarizarlos ésta no se viera afectada (tampoco puede ser participe del PCA).

Así, se obtuvieron las 22 componentes principales, si graficamos la varianza explicada obtenemos lo siguiente:

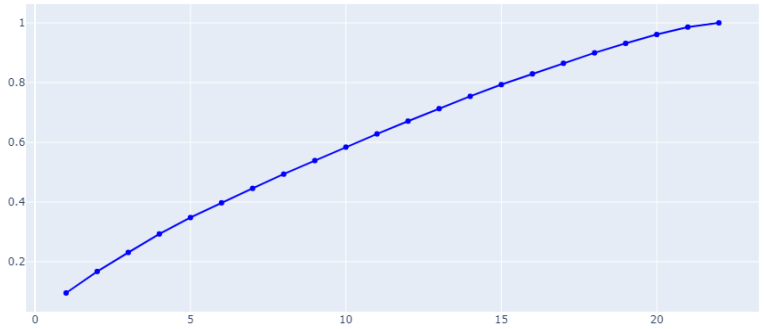

Figura 32: Varianza explicada por PCA

Se observa que con 15 componentes logramos explicar el 80 % de la varianza. Por lo tanto, la tabla con estas 15 componentes principales será nuestra tabla analítica de datos con la que iniciaremos (en otra etapa) a correr modelos de ciencias de datos.