

Proyecto DAT002 – Análisis de películas y sus regiones

February 27, 2019

ALFREDO ROJAS

Contenido

[Contenido 1](#_Toc3664492)

[Introducción 2](#_Toc3664493)

[Descripcion de las fuentes de datos 3](#_Toc3664494)

[Diccionario de datos 3](#_Toc3664495)

[Destino de datos 6](#_Toc3664496)

[Transformacion de los datos 7](#_Toc3664497)

[Herramienta de visualización – Power BI 9](#_Toc3664498)

[Nombre de la herramienta. 9](#_Toc3664499)

[Licenciamiento / costo. 9](#_Toc3664500)

[Versiones. 9](#_Toc3664501)

[Ventajas de la herramienta. 9](#_Toc3664502)

[Información de reportes 10](#_Toc3664503)

[Cantidad de películas por país 10](#_Toc3664504)

[Rating promedio por país 11](#_Toc3664505)

[Rating mínimo por país 12](#_Toc3664506)

[Mayor rating por país 13](#_Toc3664507)

[Origen de los datos 14](#_Toc3664508)

[Conclusiones 15](#_Toc3664509)

[Bibliografía 16](#_Toc3664510)

Introducción

Estados Unidos es el principal proveedor de películas a nivel mundial.

¿Qué países compiten contra Estados Unidos en cuanto a películas se refiere?

Tomaremos en cuenta los datos de IMDB para realizar esta comparativa.

Podemos tomar en cuenta el país de procedencia de las películas y su rating, basado en votos de usuarios, para comparar las películas de países que no sean Estados Unidos, contra las películas que si lo sean.

Finalmente podremos ver qué país tiene mayor posibilidad de competir contra Hollywood.

Descripcion de las fuentes de datos

Las fuentes de datos vendrán de IMDB y Kaggle.

Diccionario de datos

El diccionario de datos, también provisionado por IMDB, se distribuye de esta forma:

Datos de títulos de las películas:

* titleId: Un identificador único para cada película
* ordering: un numero para identificar filas para cada película. Se utiliza en caso de que una película tenga mas de una fila de información (Diferente región o idioma, etc)
* title: El título de la película, localizado dependiendo de la región
* región: Región para la versión de la película
* language: Idioma de la película
* types: Enumera el tipo de título, determina si es “alternativo”, “dvd”, entre otros.
* atributes: información adicional del título.
* isOriginalTitle: Nos dice si un titulo es original o no.

Datos de las películas:

* titleId: Un identificador único para cada película.
* titleType: Tipo de título (película, corto, serie de televisión, etc)
* primaryTitle: El título primario de la película utilizado en material promocional
* originalTitle: El título original de la película.
* isAdult: Nos dice si la película es clasificada para adultos o no.
* startYear: Año en que se estrenó una película.
* endYear: Año en el que finaliza una serie televisiva, no se utiliza para películas.
* runTimeMinutes: duración de la película, en minutos.
* genres: Generos de la película.

Datos de escritores y directores de la película:

* titleId: Un identificador único para cada película.
* Directors: directores de la película
* Writers: escritores de la película

Datos de personal principal de la película:

* titleId: Un identificador único para cada película.
* Ordering: un identificador para cada fila de información de cada película
* personId: un identificador único para cada persona
* category: la categoría del trabajo que desempeño la persona
* job: el trabajo específico que desempeño la persona
* characters: el nombre del personaje que desempeño la persona, si aplica.

Datos de ratings de las películas

* titleId: Un identificador único para cada película.
* averageRating: rating promedio de todos los ratings individuales
* numVotes: número de votos totales para cada película

Datos de personas

* personId: identificador único para cada persona
* primaryName: nombre por el que es conocida cada persona
* brithYear: año en el que nació la persona
* deathYear: año en el que murió la persona, si aplica
* primaryProfession: el top 3 de profesiones de la persona
* knownForTitles: películas por las que la persona es conocida

Además de estos datos, se utilizó un set de datos disponible en Kaggle, el cual contiene las siguientes columnas:

* color: indica si la película es a color o no
* num\_critic\_for\_reviews: número de expertos que criticaron la película
* duration: duración de la película en minutos
* director\_name: Nombre del director de la película
* director\_facebook\_likes: número de likes que tiene el director de la película en su página de Facebook
* actor\_3\_name: Segundo actor de reparto en la película
* actor\_3\_facebook\_likes: Cantidad de likes que tiene el segundo actor de reparto en su página de Facebook
* actor\_2\_name: Primer actor de reparto en la película
* actor\_2\_facebook\_likes: Cantidad de likes que tiene el primer actor de reparto en su página de Facebook
* actor\_1\_name: Actor principal en la película
* actor\_1\_facebook\_likes: Número de likes del actor principal en su pagina de Facebook
* gross: Cantidad de dinero recaudada
* genres: Géneros de la película, separados por pipes (“|”)
* movie\_title: Nombre de la película
* num\_voted\_users: Cantidad de usuarios que votaron por un rating para la película
* cast\_total\_facebook\_likes: Cantidad de likes del reparto de la película
* facenumber\_in\_poster: Número de caras de los actores que aparecen en el poster de la película
* plot\_keywords: Principales palabras utilizadas para identificar la película
* movie\_imdb\_link: link a la página de IMDB de la película
* num\_user\_for\_reviews: Cantidad de reviews de usuarios para la película
* language: Idioma de la película
* country: País de origen de la película
* content\_rating: Rating de edad para la película
* Budget: Presupuesto con el que se realizó la película
* title\_year: Año de estreno de la película
* imdb\_score: Rating final de la película en IMDB
* aspect\_ratio: Aspecto con el que se grabó la película
* movie\_facebook\_likes: Cantidad de likes que tiene la película en Facebook

De este set de datos de Kaggle solo nos interesan dos columnas: País de origen de la película y el link a la página de IMDB.

Destino de datos

Se creará una base de datos de SQL donde se podrán manipular los datos, en la misma se crearán views para tener un mejor manejo de estos.

De esta base de datos se podrán extraer los datos para ser manipulados en un programa de visualización como lo es Tableau o PowerBI.

Transformacion de los datos

Como una transformación general, se reemplazaron todos los valores nulos provenientes de la fuente de datos (“/N”) con “N/A”

Datos de Kaggle:

Para los datos de Kaggle, se realizó una simple transformación: Se tomó el link de la página de IMDB para cada película y se limpió el mismo, de forma que solamente nos queda una llave, la cual comienza con “tt” y termina con un número.

Esta llave es la misma que la columna “title\_id” que tenemos en el resto de los datos de IMDB, y servirá para conectar ambos sets de datos.

Datos de títulos:

Para los datos de las películas, se tuvieron que filtrar los títulos para tomar en cuenta solamente las películas.

Datos de géneros:

Para los datos de los géneros se hicieron las siguientes transformaciones:

* Se tomaron en cuenta solamente las películas, obtenidas en el paso anterior.
* Se separaron los géneros de cada película, puesto que estos vienen en una lista.
* Se excluyeron los siguientes géneros, puesto que eran los menos relevantes a nivel de ratings:
  + Adult
  + War
  + Music
  + Musical
  + Film-Noir
  + Family
* Se tomo en cuenta solamente el primer género para cada película, a tipo de “Género Principal” (Primary Genre), esto a fin de no tener películas duplicadas con más de un género.

Datos de ratings:

Para los datos de ratings, se tomaron en cuenta solamente aquellos ratings que pertenecieran a películas. Anteriormente, se tomaban en cuenta aquellas películas cuya cantidad de votos excediera los 3155 votos. Sin embargo, después de la inclusión del set de datos de Kaggle, la muestra ahora se conforma de alrededor de cinco mil películas, para las cuales se posee el rating.

Datos de directores y escritores:

Para estos datos, se realizó una transformación parecida a la antes mencionada para géneros. Se separaron los datos, puesto que vienen en listas, y se “pivotearon” los mismos, de forma que para cada película se tiene una o mas filas, dependiendo de que tantos directores o escritores tenga la misma.

Como nota adicional, todas estas transformaciones se realizaron utilizando un script de Python, con la librería Pandas.

Luego de esto, se procedió a la creación de la base de datos, y de las vistas de los datos, donde se tiene, por ejemplo, una vista que nos brinda los datos de una película, los ratings, y el género principal de la misma.

Herramienta de visualización – Power BI

Nombre de la herramienta.

La herramienta utilizada para este proyecto es Power BI, herramienta de visualización de Microsoft.

Licenciamiento / costo.

Power BI es una herramienta de uso gratuito, sin embargo, tiene una versión “Pro”, por la cual se deben pagar $9.99 por usuario, por mes.

Además de esto, se tiene una versión “Premium”, más orientada a empresas, y su precio varía según la cantidad de usuarios que la vayan a utilizar. Como ejemplo, para cinco mil usuarios, se deben cancelar cerca de $25,000 al mes

Versiones.

Power BI cuenta con una versión gratuita, una versión “Pro” y una versión “Premium”

Ventajas de la herramienta.

Power BI es una herramienta sumamente poderosa, en la cual se pueden realizar transformaciones de datos y visualizaciones en el mismo lugar.

Además de esto, da la capacidad de realizar más de un reporte por dashboard, y se pueden unir varias fuentes de datos.

Información de reportes

Cantidad de películas por país

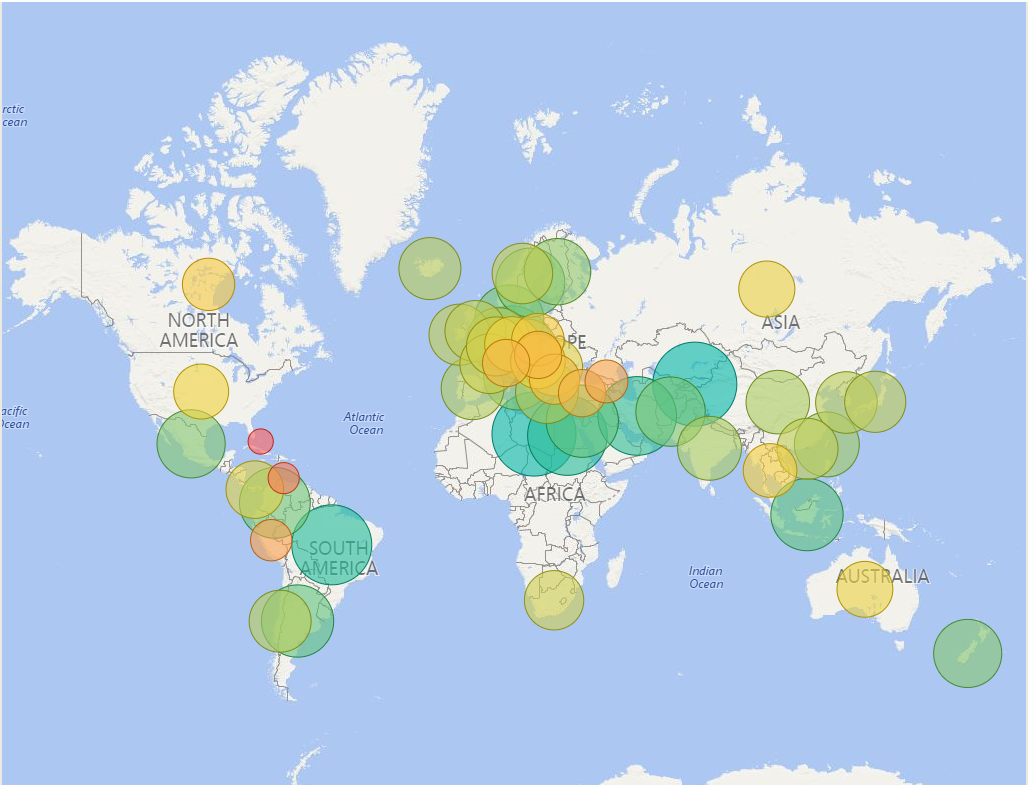


En este reporte se pueden ver la cantidad de películas por país.

Es notorio, y esperado, que Estados Unidos sea el país con la mayor cantidad de películas, con 3202 películas.

En esta muestra de datos, el segundo lugar se lo lleva el reino unido, con 373 películas.

Rating promedio por país

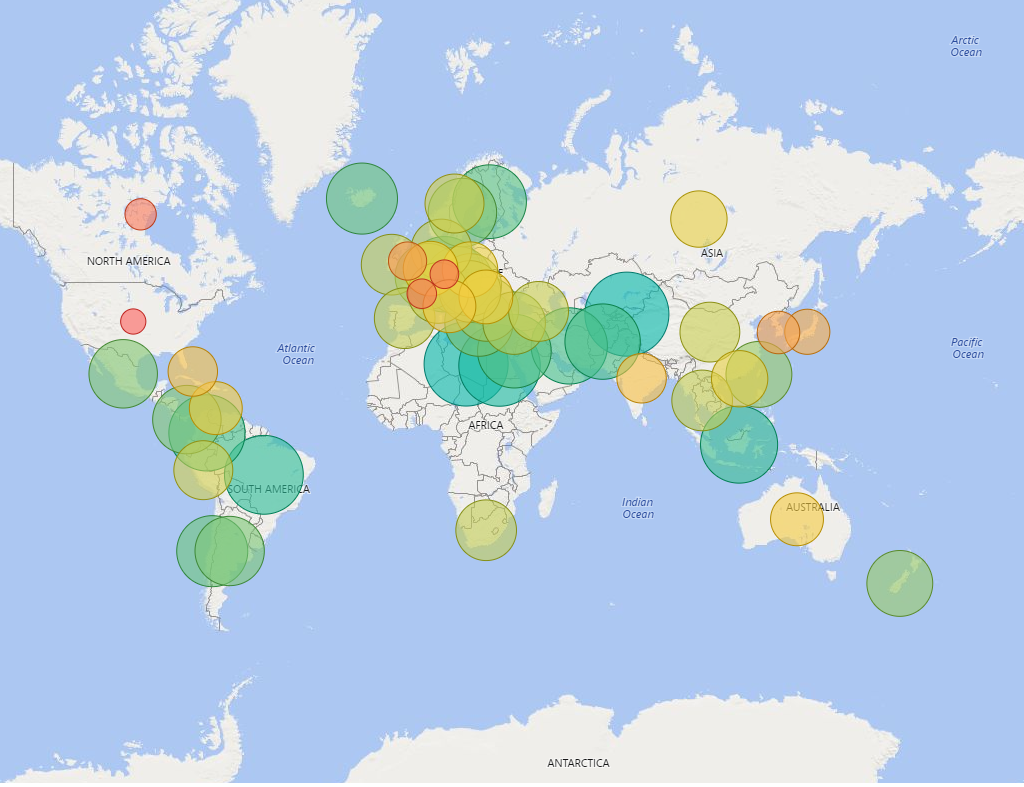


En este reporte podemos ver el rating promedio para todas las películas de cada país.

Podemos ver que Estados Unidos, aunque tiene una gran cantidad de películas, no tiene un rating promedio alto, con solo 6.4.

Sin embargo, esto es esperado por la gran cantidad de películas que posee, donde otros países, como Afganistán tienen un rating promedio de 7.4, pero con solamente una película.

Rating mínimo por país

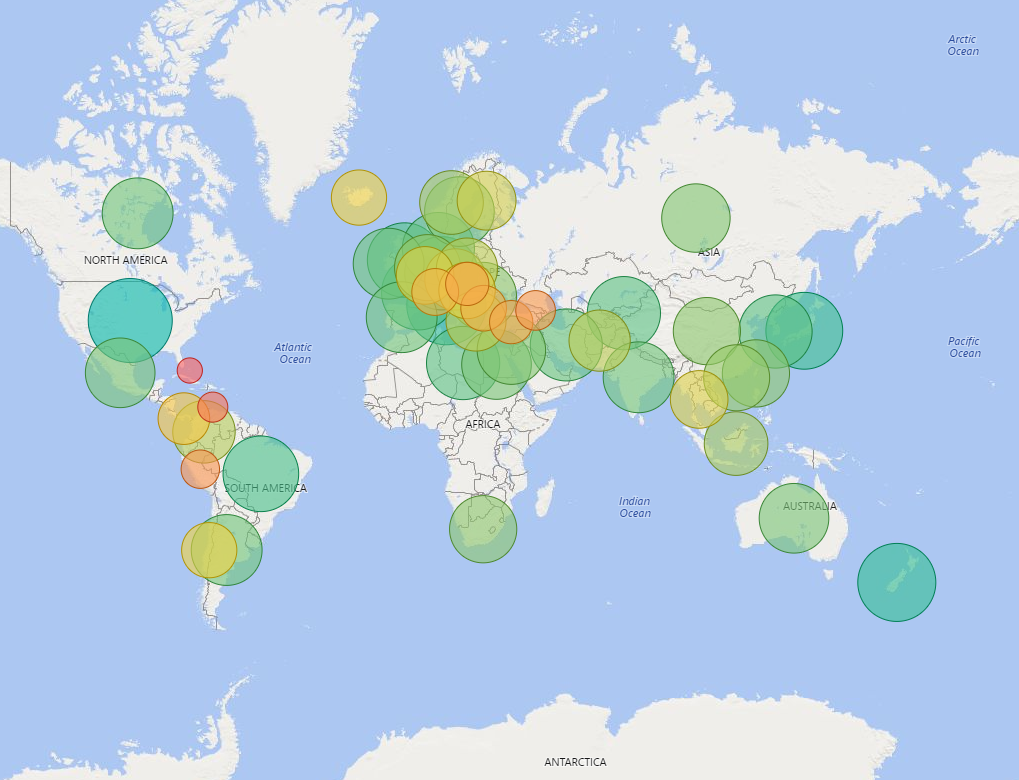


En este reporte podemos ver el menor rating por país.

En este podemos apreciar que Estados Unidos tiene la película peor calificada de la muestra, con una calificación de 1.6.

En este caso, el Reino Unido no esta lejos, puesto que su peor calificación es de 3.1

Mayor rating por país



Este reporte es la contraparte del anterior, donde podemos ver la puntuación mayor para cada país.

En este caso, y sin mucha sorpresa, Estados Unidos tiene la mayor puntuación con un 9.3

Sin embargo, otros países no están lejos, donde Reino Unido tiene un rating de 8.5, Japón uno de 8.7, y Nueva Zelanda uno de 8.8

Origen de los datos

Para todos los reportes antes mencionados se utilizó la misma fuente de datos: Una base de datos de SQL de la cual se extrajeron datos según se necesitaron.

Así pues, se hizo una extracción donde se tenía: El país, el conteo de películas, el promedio de calificaciones, el máximo de calificaciones, y el mínimo de calificaciones.

Conclusiones

La muestra de datos utilizada, desgraciadamente, limita muchísimo los resultados. Muchas de las películas en la muestra son de Estados Unidos, y esto causa que los resultados se vean muy decantados en su favor.

Sin embargo, los reportes nos pueden decir cosas provechosas. Es claro que ningún otro país se puede comparar en cantidad de películas con Estados Unidos.  
La única excepción a esta regla sea, quizás, la India, sin embargo, esto no se puede demostrar con la muestra de datos actual.

En cuanto a calidad, Estados Unidos esta cerca de la media, si no por debajo de la misma.  
Con un promedio de 6.4, hay países que lo superan como Japón, Nueva Zelanda o el Reino Unido.

Esto se demuestra aún mas con el reporte de la puntuación mínima, donde Estados Unidos tiene la película peor calificada de toda la muestra.

Caso contrario a este es el reporte de mayor calificación, donde Estados Unidos tiene la película con la mayor calificación de la muestra.  
Sin embargo, países como Japón, Nueva Zelanda, o Reino Unido, están muy cerca de Estados Unidos en cuanto a calificaciones.

Dicho todo esto, podemos decir que, aunque Estados Unidos tiene la mayor cantidad de películas de la muestra, no significa que Estados Unidos tenga la mejor calidad en cuanto a películas, con otros países compitiendo muy de cerca con ellos.

Así pues, Estados Unidos podrá hacer muchísimas películas, pero la calidad de las mismas es muy fluctuante, donde otros países tienen muchas menos películas, pero con mayor calidad.

Bibliografía

IMDB. (s.f.). *Alternative Interfaces - IMDb*. Recuperado el 27 de 2 de 2019, de http://imdb.com/interfaces

*Microsoft Power BI*. (s.f.). Obtenido de https://powerbi.microsoft.com/en-us/

Yueming. (2018). *IMDB 5000 Movie Dataset*. Obtenido de https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset