PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

Laboratorio 1
Inteligencia de Negocios [ICI6442]

Andrés Marcelo Vidal Soto
Valentina Paz San Martín Pacheco

Abril, 2024

Introducción y contexto

En la actualidad cada vez es más frecuente la suscripción a servicios de streaming de películas y series, donde el espectro de estos es cada vez más variado, aumentando la demanda de estos cada día más. El pionero de este tipo de servicios corresponde a Netflix, empresa que comenzó a transmitir contenido audiovisual desde el año 2007 teniendo tanto la ventaja de la innovación como la desventaja de entrar en el terreno desconocido de los servicios de streaming. Sin embargo, a pesar de la variedad de opiniones sobre Netflix, esta empresa sigue siendo hasta el día de hoy la plataforma de streaming más cotizada y es dado a esto, que se escogió analizar el contenido que esta empresa transmite a sus suscriptores.

Para el análisis del negocio de Netflix se utilizó un conjunto de datos de Títulos de Netflix, el cual corresponde a una recopilación completa de películas y programas de televisión disponibles en Netflix, que cubre varios aspectos como el tipo de contenido, el título, el director, el reparto, el país de producción, el año de estreno, la clasificación, la duración, los géneros y una breve descripción. Este conjunto de datos es fundamental para analizar las tendencias en el contenido de Netflix, comprender la popularidad de los géneros del contenido audiovisual y examinar la distribución del contenido en diferentes regiones y períodos de tiempo.

Atributos del Dataset

El dataset utilizado posee los siguientes atributos

1. Identificador “show_id”
2. Tipo de contenido“type”
3. Título “title”
4. Director “director”
5. Reparto “cast”
6. País de producción“country”
7. Fecha añadida a la plataforma“date_added”
8. Fecha de estreno “release_year”
9. Calificación de edad “rating”
10. Duración del contenido “duration”
11. Géneros “listed_in”
12. Descripción del contenido“description”

A continuación se explicarán más detalladamente los atributos en conjunto a los valores que estos pueden tener.

1. Identificador “show_id”: Corresponde simplemente al identificador del contenido audiovisual en la base de datos, en el dataset escogido, es decir, el original, este posee un formato alfanumérico donde se enumeran del 1 al 8809 con una S al comienzo.

2. Tipo de contenido“type”: Indica si el contenido audiovisual corresponde a una serie (TV Show) o a una película (Movie).

3. Título “title”: Corresponde al título del contenido audiovisual

4. Director “director”: Indica el nombre del o de los directores del contenido.

5. Reparto “cast”: Listado de actores y actrices principales que participan en el contenido.

6. País de producción“country”: País o países que produjeron el contenido.

7. Fecha añadida a la plataforma“date_added”: Fecha en la que el contenido fue añadido a Netflix.

8. Fecha de estreno “release_year”: Fecha en la que el contenido se estrenó a nivel mundial.

9. Clasificación de edad “rating”: Indica la clasificación de edad apropiada para la visualización del contenido, dado que esta varía dependiendo del país, el “rating” del dataset corresponde al de Estados Unidos cuya calificación será detallada a continuación:
    a. TV-G: “General Audiences”, contenido apto para todo público, es decir, no posee nada que ofenda a los padres si lo ven los niños.
    b. TV-Y: “Appropriate for all children” contenido diseñado específicamente para un público infantil, incluyendo niños de 6 y menores, en especial, menores de 6 años.
    c. “PG” & “TV-PG”: “Parental Guidance Suggested”, existe contenido que podría no ser adecuado para niños. Se recomienda a los padres dar "orientación parental", ya que puede contener material que a los padres quizás no les guste para sus hijos pequeños.
    d. TV-Y7: “For children age 7 and above” El contenido puede no ser considerado apropiado para algunos niños menores de 7 años.
    e. PG-13: “Parents Strongly Cautioned”, Parte del material puede ser inapropiado para niños menores de 13 años. Se insta a los padres a tener cuidado, dado que algunos contenidos pueden ser inapropiados para los preadolescentes.
    f. TV-14: “Unsuitable for children under 14 years of age” El contenido es solo para mayores de 14 años y los menores a esta edad deben estar acompañados por un padre o tutor adulto.
    g. R: “Restricted”, Los menores de 17 años deben estar acompañados por un padre o tutor adulto. Contiene material para adultos, por lo que se insta a los padres a aprender más sobre la película antes de llevarse a sus hijos con ellos.

h. “TV-Y17” & “TV-MA”: “Adults Only”/”Mature Audiences Only”, No se admiten menores de 17 años. Contenido sólo para adultos.
i. NR: “Not Rated” El contenido no posee una clasificación de edad definida o en caso de versiones sin cortes o sin censura se utiliza NR porque la clasificación difiere de la cinta estrenada previamente.

10. Duración del contenido “duration”: Duración en minutos del contenido.

11. Géneros “listed_in”: Género o géneros a los que se encuentra “enlistado” el contenido.

12. Descripción del contenido“description”: Descripción o sinopsis del contenido.

Si bien el dataset es bastante completo para el negocio de servicios de streaming, se encontraron en este los siguientes errores:

1. Indica tener 12 columnas (show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description), pero sus registros poseen más columnas sin valor o (NaN teniendo en cuenta que el lenguaje utilizado es python), dando como total 26 columnas donde 14 de estas tienen valor NaN.
2. Los registros tienen atributos sin valor (NaN), brindando poca información a la hora realizar reportes con este dataset.
3. Sus IDs poseen la letra S y posterior un número, dificultando su iteración con la letra. Los números que se almacenan como texto pueden producir resultados inesperados, como una fórmula no calculada que se muestra en lugar de un resultado.
4. Algunas filas del dataset contienen saltos de línea, por lo que se decidió eliminar las filas con estas características, ya que no se puede concluir que el contenido siguiente pertenece a la fila anterior o si se trata de un error más grande.

Procesamiento y transformación de datos

Dados los errores encontrados (descritos en la sección anterior), se consideró necesario una limpieza y normalización de datos para garantizar que los datos estén libres de valores faltantes, inconsistencias y que los datos estén en un formato consistente y estandarizado, para facilitar su análisis y procesamiento posterior. Las transformaciones realizadas serán detalladas a continuación:

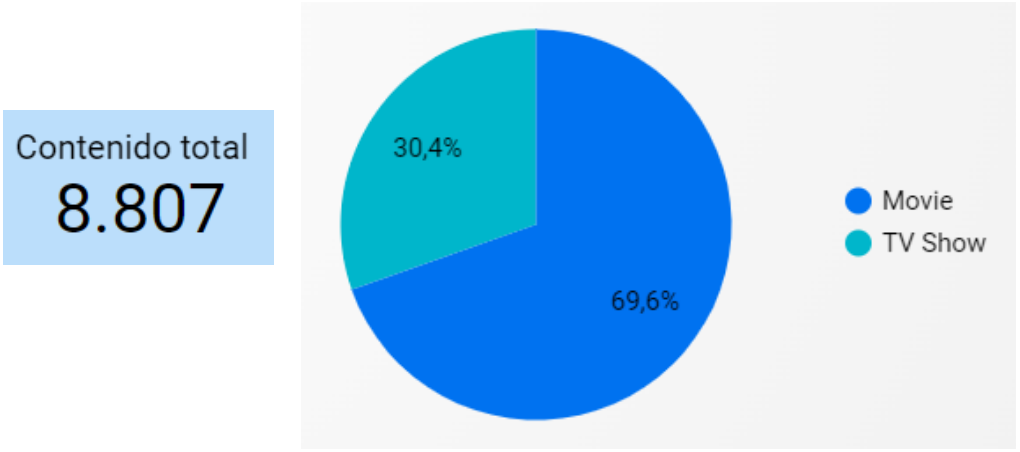
1. Limpieza de datos:
    a. Se eliminaron las columnas sin valor o NaN, ya que estas no aportan información valiosa para el análisis de datos mientras que aumentaría el tamaño de la base de datos sin proporcionar ningún beneficio y así optimizar el almacenamiento y mejorar el rendimiento en el procesamiento de datos.
    b. Se reemplazaron los atributos sin valor (NaN), o datos faltantes, por "no especificado" para denotar la ausencia de información, conservando así las filas o registros para futuros análisis, evitando errores o comportamientos inesperados en las herramientas de análisis de datos y además para facilitar la comprensión de los datos por parte de los usuarios.
    c. Se eliminaron filas que contenían saltos de línea. Los saltos de línea pueden causar problemas al procesar la información, por lo que se decidió eliminarlos, ya que no se puede asegurar a qué registro pertenece como para unirlo a este, preparando así los datos, garantizando su coherencia e integridad de los datos para su posterior procesamiento y análisis de una manera más fluida e inequívoca.
2. Normalización de datos:
    a. Se eliminó la S de los IDs, dejándolos 100% numéricos para facilitar su uso en operaciones, como ordenar, agrupar o realizar cálculos matemáticos basados en los IDs. Esto también busca mejorar la eficiencia en el almacenamiento y procesamiento de datos, ya que los valores numéricos suelen ocupar menos espacio y son por ende más rápidos de procesar, además de evitar posibles errores o inconsistencias al trabajar con los IDs, especialmente cuando se realizan operaciones que implican comparaciones o uniones de datos.

Lo anterior se realizó en un código de Python utilizando Visual Studio Code como herramienta de desarrollo, este código se encuentra en el archivo llamado ("Lab1_Vidal_SanMartín.py") en el zip enviado, una vez ejecutado este código debe generarse el archivo "netflix_titles_modified.csv" el cual fue ingresado en la plataforma Looker Studio para realizar el posterior analisis y visualizacion de los datos.

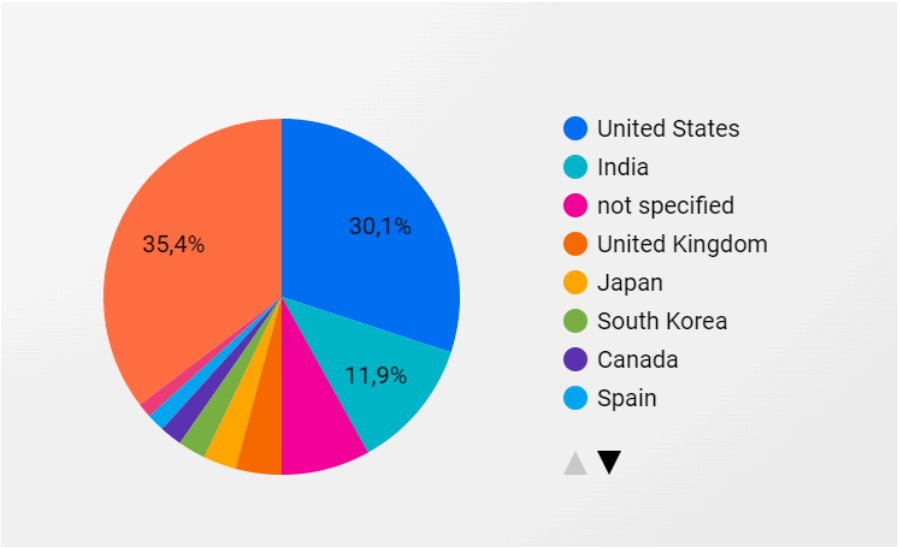
Análisis y visualización de datos

Dentro del análisis realizado se encontró que en total existen 6130 películas dentro de la plataforma, mientras que 2677 corresponden a series de televisión, además de esto se realizaron los siguientes insights durante el análisis de los datos:

- **Contenido Total y gráfico circular**: Este indicador junto a este gráfico muestran cuánto contenido se transmite en total y qué porcentaje se transmite de películas (Movie) y series (TV Show). Esto resulta útil si se sabe la demanda de cada tipo de contenido, ya que así se puede decidir sobre cuál tipo de contenido conviene más invertir.



- **Contenido total por país:** Casi un tercio del contenido se distribuye en contenido de Estados Unidos con un 30,1%, seguido por la india que contiene un 11,9%, esto puede deberse a que como netflix es una empresa estadounidense prioriza contenido proveniente de este país, además de que el dataset corresponde al contenido que se transmite en Estados Unidos, siendo que el contenido varía según el país de donde uno accede a Netflix, por lo que se podría inferir que además de la alta producción de contenido audiovisual estadounidense, este además es masificado por las empresas de este mismo país. Esto resulta importante de saber para el negocio debido a que, según los resultados en temas de ventas se puede evaluar qué contenido conviene más transmitir según el país donde fueron desarrollados.



- **Gráfico de barras horizontal con clasificación de edad:** Este gráfico indica el contenido que se transmite en cada país clasificado según la clasificación de edad. Analizando el gráfico se puede observar que la categoría que posee más contenido es TV-MA (contenido para mayores de 17) seguido por TV-14, si uno intersecciona estos dos, se podría inferir que el grupo etario que más contenido tiene disponible, y probablemente el que más utiliza Netflix, se encuentra entre los adolescentes y los jóvenes mayores de edad. La información sobre la clasificación de edad del contenido transmitido es útil para la empresa ya que si se sabe el grupo etario que más consume contenido en un país, se puede invertir en proveer más contenido de una clasificación de edad acorde al este grupo.

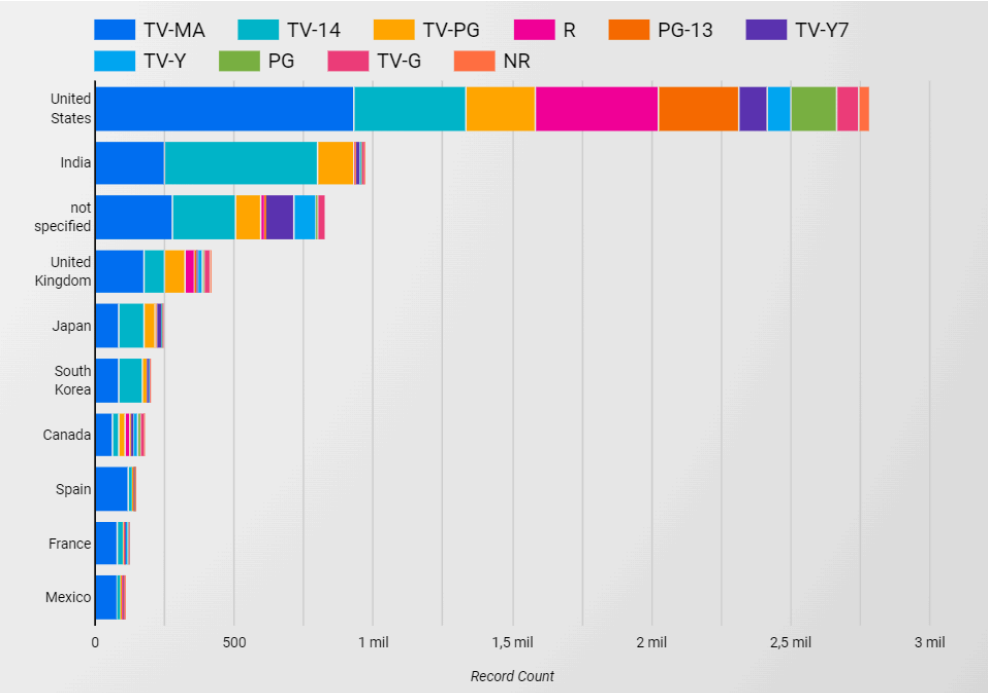


Grafico de calificación por edades de cada contenido

- **Gráfico de barras vertical de contenidos por país:** Dentro del análisis que se realizó se consideró evaluar el volumen de contenido que posee cada país entre series y películas, por lo que se llegó a la conclusión de que la mayoría de países tienen más películas que series, a excepción de japón y corea del sur. Esto resulta útil para el negocio si se sabe cuánto se consume de este contenido en cada país según su tipo (serie o película), ya que así se puede saber en cuál de los dos conviene invertir más por país.

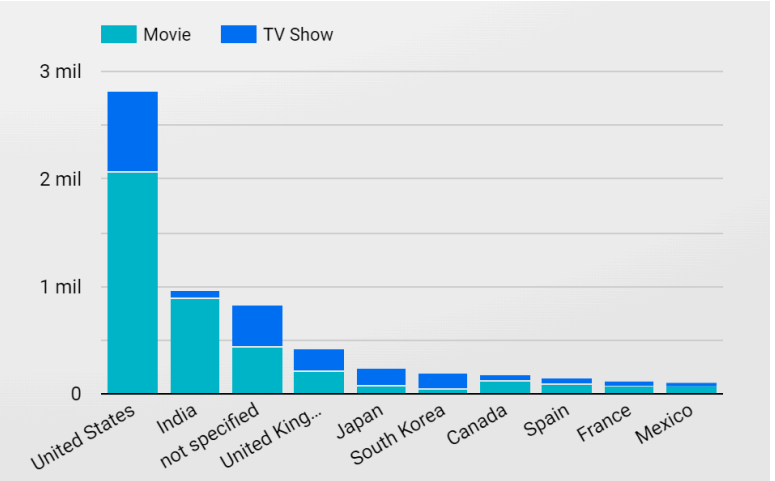
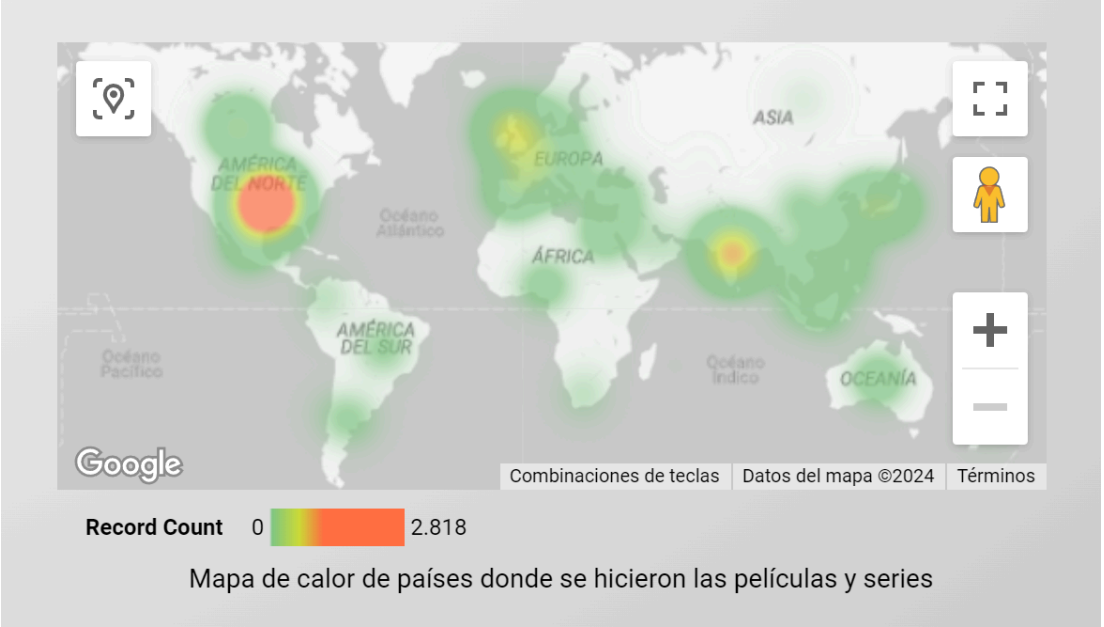


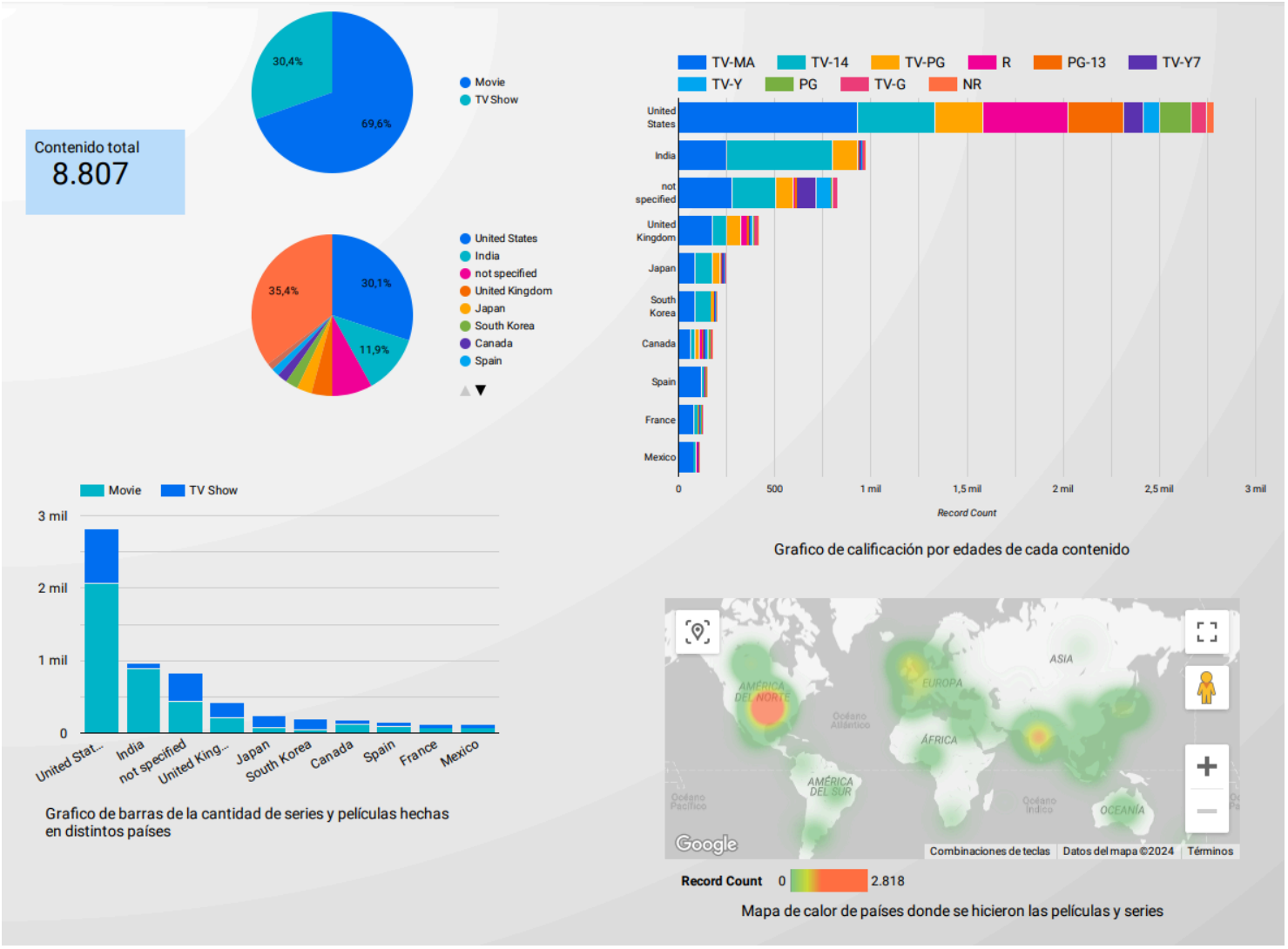
Grafico de barras de la cantidad de series y películas hechas en distintos países

- **Mapa de calor:** Finalmente se presenta un mapa de calor que facilita la visualización de la procedencia del contenido de una forma más general y fácil de ver. Esto si bien muestra la misma información que el gráfico circular de contenido total por país, resulta más fácil para ver más allá de solo los países, las zonas geográficas, como continentes o regiones de estos continentes, que poseen más contenido transmitido por Netflix. Esto es útil para saber en qué zonas podría invertirse más para obtener contenido si aumenta la demanda por contenido de alguna zona geográfica en particular, como por ejemplo en la actualidad que aumentó la demanda por contenido asiático, sobre todo surcoreano.



Mapa de calor de países donde se hicieron las películas y series

La vista general de los gráficos y el mapa en su conjunto, como tablero de control (dashboard) se ve de la siguiente manera:



Si se desea ver de forma detallada el análisis generado e interactuar con este, puede visualizarse en el siguiente enlace:
https://lookerstudio.google.com/reporting/52238424-1b2c-4c8d-93f7-9ac4362a2181

Conclusiones y hallazgos

Dentro de las conclusiones que se pudieron encontrar se tiene que el contenido de Netflix se basa principalmente en contenido de Estados Unidos, se cree que esto puede deberse a que el público objetivo de la empresa es principalmente gente que desea ver series o películas donde el idioma original sea inglés, por consecuente se llega a entender porque el contenido del Reino Unido tiene una importante presencia en los datos analizados. Además existe una tendencia entre los países, la cual consiste en tener más películas que series, pero existen ciertas excepciones que pueden deberse a el lugar geográfico ya que los dos países que evidencian esta característica son de la misma región (Japón y Corea del Sur).

También se encuentra una tendencia etaria de contenido para jóvenes mayores de 14 y de 17 dado a la cantidad de contenido disponible para su clasificación de edad.

Finalmente se puede evidenciar que existe un problema con los datos analizados y es que muchos datos no se encontraban ingresados, por lo que se tuvo que reemplazar por el dato “not specified” que dificulta generar una conclusión más precisa, ya que según el análisis existe un 8,1% de series y películas que no tienen un país de procedencia, que si bien con el resto de datos se podría evidenciar una tendencia tener esos datos faltantes dificulta tener un análisis sólido, por no mencionar del resto de columnas que presentaron el mismo problema.