Bootcamp Data Science

GABRIELA CONTE MARCOS

ANÁLISIS EXPLORATORIO DE DATOS TENDENCIAS DE NETFLIX

INDICE

ANÁLISIS DEL CATÁLOGO DE NETFLIX: EXPLORACIÓN Y CONCLUSIONES

1. Introducción

Los datos analizados en este trabajo provienen de un conjunto descargado de Kaggle, que incluye información sobre más de 8.000 títulos disponibles en Netflix. Este dataset contiene múltiples columnas categóricas y algunas numéricas, proporcionando información sobre títulos, años de lanzamiento, géneros, duración, países de producción, entre otros.

En un análisis preliminar, observamos que Estados Unidos domina ampliamente en términos de cantidad de títulos, seguido por otros países como India, Reino Unido, Canadá y Francia. Sin embargo, el datset presentaba varios desafíos:

- Altos valores nulos en variables como director y cast, que se solucionaron de manera estratégica.
- Problemas de cardinalidad y redundancia en columnas como listed_in, lo que requería ajustes para extraer información relevante.
- La variable duration mezclaba unidades de tiempo (minutos para películas y temporadas para series), dificultando el análisis directo.

Dado este contexto, nuestro análisis busca responder a la hipótesis de si Netflix adapta su catálogo según factores socioculturales de cada país o si, por el contrario, mantiene una estrategia global uniforme. Para ello, tratamos los datos y desarrollamos un análisis enfocado en patrones de consumo y producción a nivel regional y global.

2. Hipótesis y Enfoque del Análisis

La hipótesis principal que guía este trabajo es que Netflix ajusta su contenido según factores demográficos y socioculturales de cada país. Este enfoque se basó en identificar patrones específicos de consumo a través de variables clave como país "country", género, tipo y duración. Además, evaluamos la evolución temporal del catálogo para entender su estrategia de expansión.

Sub-hallazgos esperados:

- Los géneros más populares varían según la región.
- La duración promedio de películas y series se mantiene estable, independientemente del género o país.
- La oferta inicial de Netflix estuvo dominada por películas, con un cambio hacia series tras 2015.

3. Tratamiento y Limpieza de Datos

El tratamiento y limpieza de datos es una etapa crucial en cualquier análisis, ya que garantiza la calidad y consistencia de los datos antes de realizar cualquier exploración o inferencia. A continuación, se detalla cómo abordamos esta etapa en el conjunto de datos de Netflix, enfrentándonos a problemas comunes como valores nulos, formatos inconsistentes y redundancias en las columnas.

3.1. Manejo de valores nulos

El dataset presentaba valores nulos en varias columnas importantes. En particular:

Columna director: Esta columna contenía numerosos valores nulos, probablemente porque muchos títulos no especifican un director. Decidimos sustituir estos valores por el término "Desconocido". Aunque esto limita el análisis basado en directores, permitía mantener la integridad del conjunto de datos sin eliminar registros completos.

Columna cast: Similar a la columna anterior, tenía una gran proporción de valores faltantes. También optamos por reemplazar los valores nulos con "Desconocido", ya que esta información no era crítica para la hipótesis principal.

Columna rating: Aquí, encontramos valores nulos en ciertas filas. En este caso, realizamos una imputación manual basada en títulos similares o contexto, para clasificar todos los títulos en una de las cinco categorías que definimos posteriormente (por ejemplo, "Todos los públicos", "Mayores de 18", etc. Esto ayudó a simplificar el análisis.

Columna country: (10% nulos, 5.84% cardinalidad): Imputamos condicionalmente por Listed_In para preservar tendencias culturales y evitar el sesgo que genera imputar con la moda general.

Columna date_added: (0.1% nulos): hemos generado una columna nueva para saber aquellos que tenían nulos y les hemos imputado la moda.

3.2. Transformación de columnas para análisis

Algunas columnas requerían transformaciones adicionales para hacerlas más útiles y manejables:

Columna listed_in (géneros):

Esta columna originalmente contenía múltiples géneros separados por comas, lo que dificultaba un análisis claro. Implementamos una función personalizada que asignaba a cada título un único género dominante. Esto se logró mediante un diccionario predefinido, en el cual agrupamos los géneros principales. Por ejemplo, títulos marcados como "Comedia, Drama" fueron clasificados como "Drama" si este género predominaba en la clasificación inicial. Así, creamos una nueva columna, "genre", que contenía un único género por título, simplificando su análisis.

Columna duration (duración):

La variable duration incluía datos mezclados: minutos para películas y temporadas para series. Para resolver esto, la dividimos en dos nuevas columnas:

- duration_minutes: Solo para películas, indicando la duración en minutos.
- duration_seasons: Solo para series, indicando el número de temporadas.

Esta transformación permitió realizar análisis separados entre películas y series, identificando patrones específicos en cada tipo de contenido y sobre todo conseguir una variable numérica ya que no teníamos muchas para el análisis.

3.3. Manejo de cardinalidad y redundancia

Algunas columnas presentaban problemas de cardinalidad (número excesivo de categorías) o redundancia y baja cardinalidad respecto a la información que podían proporcionar:

Columna country:

Los títulos podían estar asociados a múltiples países, lo que dificultaba evaluar la representación por país. Para resolverlo, seleccionamos únicamente el primer país listado, asumiendo que era el principal responsable del título. Esto permitió identificar claramente los países con mayor número de títulos y analizar patrones regionales.

Columna release_year (año de lanzamiento):

Aseguramos la uniformidad del formato en esta columna y la preparamos para análisis temporales. Posteriormente, dividimos esta variable en componentes de año y mes, lo que permitió crear mapas de calor para observar patrones estacionales en las producciones de Netflix.

3.4. Estrategias para garantizar datos consistentes

Además de los pasos anteriores, tomamos medidas adicionales para asegurar que los datos estuvieran listos para el análisis:

- Eliminamos duplicados para evitar sesgos en los resultados.
- Ordenamos los datos por año de lanzamiento para facilitar su análisis cronológico.
- Estandarizamos los nombres de las categorías para evitar inconsistencias en su interpretación.

Gracias a este tratamiento exhaustivo, obtuvimos un conjunto de datos limpio y estructurado, listo para responder a las preguntas planteadas en nuestra hipótesis y llevar a cabo un análisis global y específico de las tendencias del catálogo de Netflix.

4. Análisis Global

Análisis Univariante

El análisis univariante se centró en examinar las variables del dataset de manera aislada, lo que nos permitió obtener una visión preliminar y determinar patrones clave dentro de las distintas categorías. Al tratar con un conjunto de datos donde predominan las variables categóricas, este análisis se enfocó principalmente en calcular y visualizar las frecuencias absolutas para identificar el camino a seguir por el análisis.

En la variable country, se confirmó que Estados Unidos lidera ampliamente en términos de títulos disponibles, seguido por países como India, Reino Unido y Canadá. Esto estableció una base sólida para explorar cómo la oferta de contenido puede variar según la región, destacando posibles factores socioculturales.

Por otro lado, la variable type mostró que las películas son la principal oferta de Netflix, superando en número a las series. Este hallazgo resultó interesante, especialmente al considerar cómo el crecimiento de las series ha sido más reciente y estratégico en la expansión de Netflix.

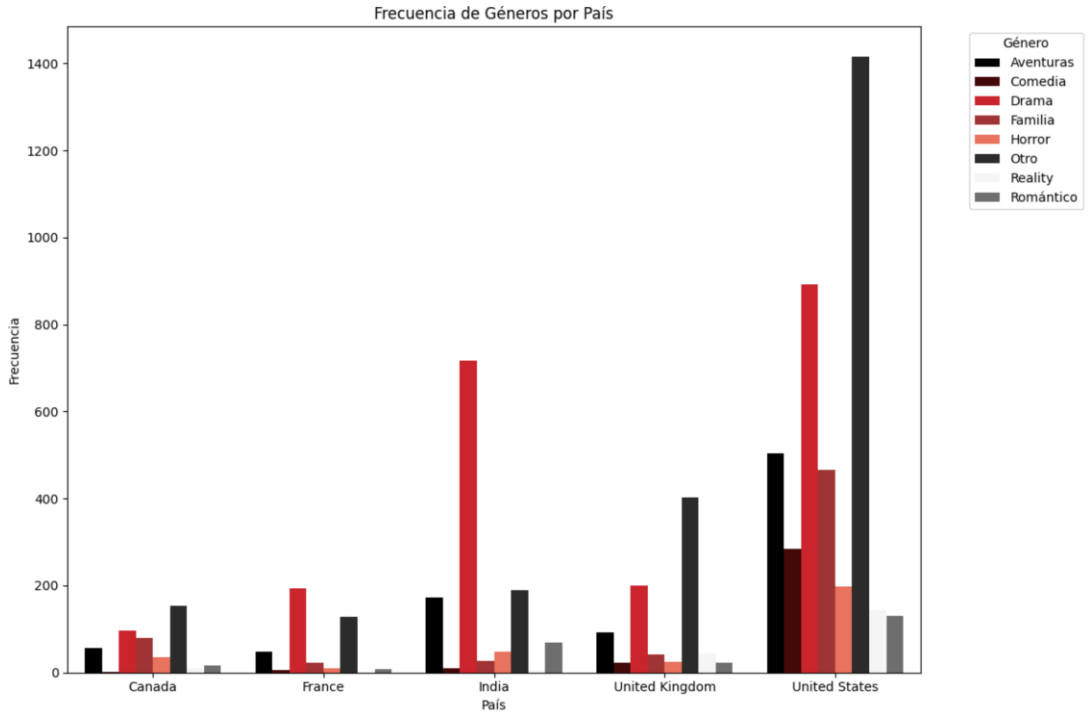
La variable rating, una vez reclasificada en cinco categorías principales, reveló una concentración significativa de títulos clasificados como contenido para audiencias adultas, especialmente en mercados clave como Estados Unidos. Este dato subrayó una estrategia orientada hacia consumidores adultos, en contraste con otros mercados más homogéneos en términos de contenido para toda la familia.

Finalmente, en las variables numéricas, como la duración, observamos valores promedio consistentes: las películas duran aproximadamente 100 minutos, mientras que las series tienen una media de dos temporadas. Estas observaciones reflejan una estrategia que responde a las expectativas generales de consumo de contenido.

Análisis Bivariante

El análisis bivariante buscó entender cómo se relacionan las variables categóricas entre sí y cómo estas relaciones pueden apoyar nuestra hipótesis inicial. Se exploraron múltiples combinaciones de variables, destacando patrones significativos y útiles para profundizar en la dinámica del catálogo de Netflix.

La relación entre país y género fue una de las más reveladoras. Países como Estados Unidos y Reino Unido compartieron un patrón de consumo centrado en el drama y la comedia, mientras que India mostró una preferencia marcada por el drama, alineándose con los factores culturales del país. Este hallazgo resaltó que Netflix parece adaptar su catálogo en función de los gustos locales.



En el cruce entre tipo y clasificación por edad, se observó que las películas abarcan una mayor variedad de clasificaciones por edades, mientras que las series tienden a concentrarse en clasificaciones maduras como “TV-MA”. Este patrón puede sugerir que las series son diseñadas para audiencias específicas y menos diversificadas.

La relación entre género y tipo mostró cómo ciertos géneros, como la comedia, son más comunes en películas, mientras que otros, como el drama, tienen una distribución más equilibrada entre películas y series. Esto reflejó una estrategia de contenido diversificada, adaptada a diferentes formatos.

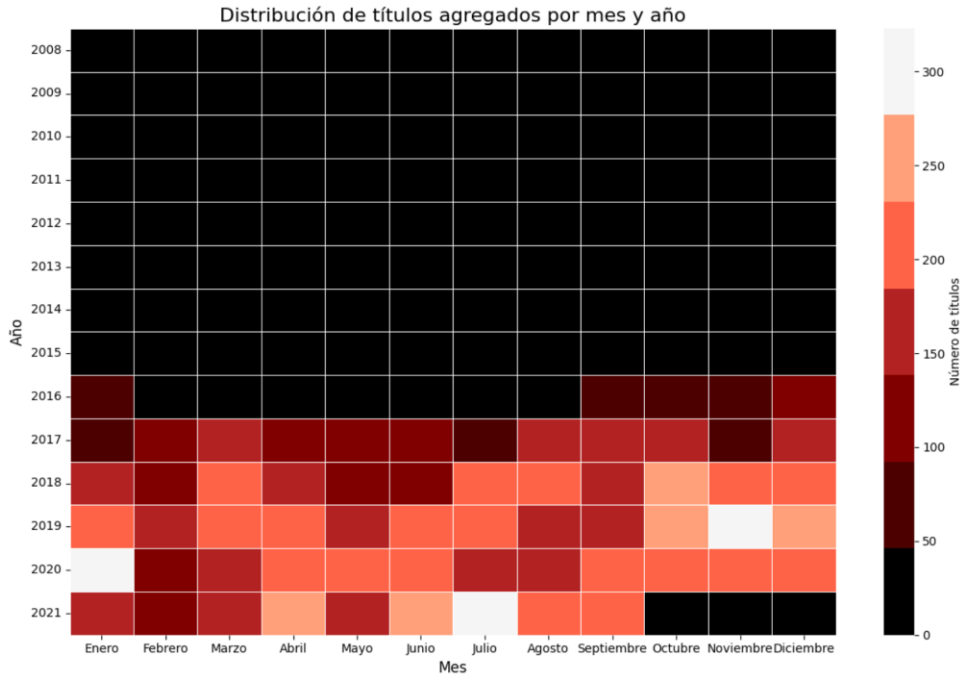
Finalmente, el análisis bivariante nos permitió identificar qué combinaciones de variables ofrecían los insights más significativos, priorizando aquellas como país y género, donde las diferencias culturales y estratégicas resultaron más evidentes.

Análisis Multivariante

El análisis multivariante permitió profundizar en la interacción de múltiples variables simultáneamente. Se utilizaron mapas de calor como herramienta principal, analizando combinaciones por país y género y la distribución temporal del contenido en función de las fechas de lanzamiento.

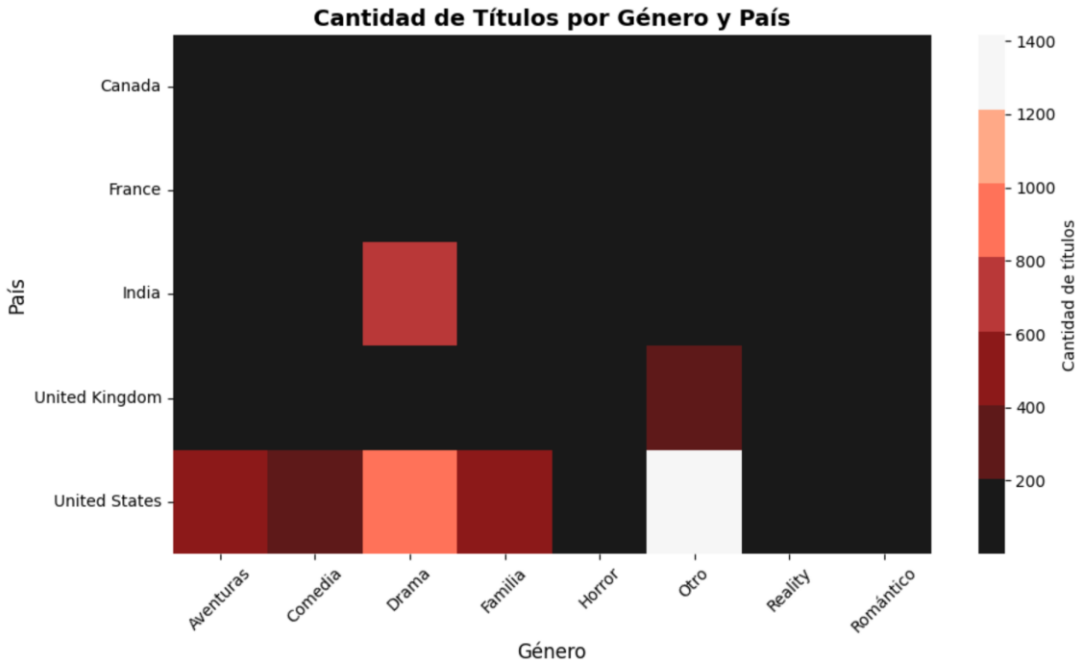
En el mapa de calor por país y género, se destacó cómo Estados Unidos lidera en diversidad y volumen de títulos, con categorías como “Drama” y “Otro” dominando.

En contraste, India mostró una inclinación exclusiva hacia el drama, mientras que Canadá presentó una distribución más uniforme entre géneros. Estas observaciones respaldaron la idea de que Netflix adapta su catálogo a las preferencias locales, aunque sigue manteniendo una oferta globalizada en ciertos mercados principales.



El mapa de calor por evolución temporal reveló que la estrategia de Netflix ha cambiado drásticamente a lo largo de los años. Antes de 2015, el contenido era escaso, especialmente en series. Sin embargo, a partir de 2017, se evidenció un crecimiento exponencial, alcanzando picos en 2020 y 2021. Este análisis reflejó cómo Netflix ha

evolucionado de una plataforma de nicho a un gigante global, adaptando su catálogo no solo a las necesidades locales, sino también a las tendencias estacionales y estratégicas.



Relevancia Global del Análisis

Este análisis global confirmó la importancia de considerar tanto las variables individuales como las interacciones entre ellas para comprender cómo Netflix adapta y gestiona su contenido. La relación entre país y género fue particularmente significativa, ofreciendo insights clave sobre cómo factores socioculturales influyen en la estrategia de contenido de la plataforma.

Conclusiones de los Análisis

Estos análisis específicos permitieron validar nuestra hipótesis inicial. Netflix no mantiene una estrategia uniforme, sino que adapta su contenido a las preferencias culturales y temporales de sus mercados principales. La interacción entre género y país, así como la evolución temporal del catálogo, revelaron patrones claros que destacan la capacidad de Netflix para equilibrar una oferta global con adaptaciones locales estratégicas.

5. Conclusiones

A través de un análisis exhaustivo de las variables del dataset, hemos obtenido resultados claros que nos permiten responder a esta hipótesis de manera fundamentada.

La relación entre el catálogo de Netflix y los factores demográficos

Nuestro análisis demuestra que, aunque Netflix posee un catálogo amplio y global, los patrones de consumo y la oferta de contenido están fuertemente influenciados por factores específicos de cada región. Esto quedó evidenciado especialmente en el análisi de la variable país que subraya, que, aunque Netflix tiene un catálogo globalizado, existen adaptaciones que responden a las preferencias locales de ciertos mercados.

Resultados del análisis de género y tipo de contenido

Uno de los hallazgos más relevantes fue el predominio absoluto del género drama en casi todos los países analizados. Sin embargo, en países como Canadá, la diversidad en la distribución de géneros resalta una excepción interesante: aquí no solo el drama es importante, sino que otros géneros tienen una representación significativa.

Además, al analizar la variable tipo (películas vs. series), descubrimos que Netflix comenzó ofreciendo principalmente películas y no fue hasta 2015 cuando las series comenzaron a tener un papel predominante. Este cambio coincide con el auge de producciones originales de Netflix, especialmente en Estados Unidos, marcando un punto de inflexión en su estrategia de contenido.

Temporalidad en la producción y oferta de contenido

El análisis de las fechas de lanzamiento y producción, complementado con el mapa de calor temporal, proporcionó información valiosa sobre la evolución de la estrategia de Netflix. Observamos un crecimiento lento en la oferta de títulos antes de 2015, seguido de una expansión acelerada entre 2017 y 2019, cuando Netflix consolidó su presencia global. Los picos observados en 2020 y 2021 coinciden con los períodos de mayor inversión en

contenido, aunque en los últimos meses de 2021 se percibe una posible desaceleración estratégica.

Asimismo, la estacionalidad también juega un papel relevante: enero y julio destacan como los meses con mayor volumen de estrenos, lo que podría estar alineado con períodos de mayor demanda por parte de los usuarios.

Implicaciones para la hipótesis

En relación con nuestra hipótesis principal, podemos concluir que Netflix no mantiene una estrategia de contenido uniforme. En su lugar, adapta su oferta a las características de cada región, aunque ciertos patrones globales, como el predominio del drama, siguen siendo evidentes. Este balance entre globalización y localización permite a Netflix satisfacer tanto las expectativas de una audiencia global como las de mercados específicos.

Reflexión final

En general, este análisis confirma que Netflix ha logrado un equilibrio entre la creación de un catálogo global atractivo y la adaptación a los mercados locales. Este enfoque dual es fundamental para su éxito continuo como plataforma de streaming líder en un mundo con preferencias tan diversas. Los resultados obtenidos sientan las bases para análisis más profundos, como explorar las razones detrás de las excepciones observadas en países como Canadá, o evaluar cómo se relacionan estos hallazgos con factores demográficos como la población o la economía.

En resumen, las conclusiones obtenidas no solo responden a nuestra hipótesis, sino que también abren nuevas preguntas para explorar en futuros análisis.

Fuente: *https://www.kaggle.com/*