Maestría en Ciencia de Datos

Visualización de Datos

Curso 2023

Obligatorio 3

11/08/2023

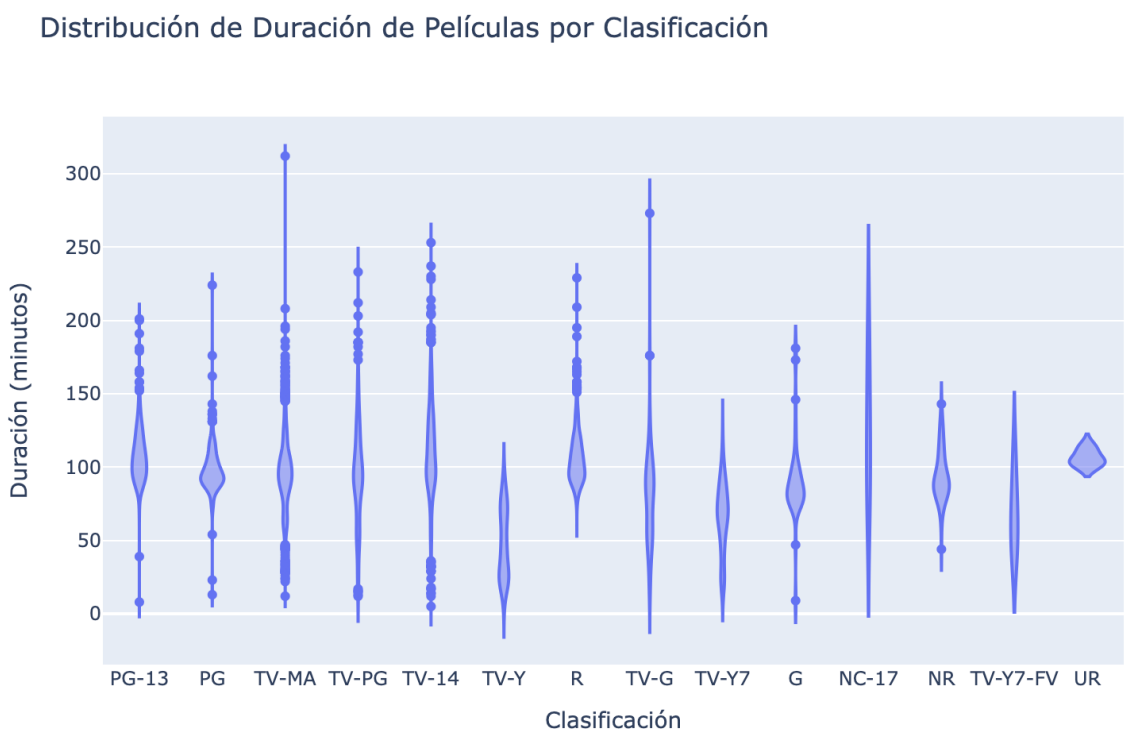
Camila Rojí

Diego Velasco

Desarrollo

Para el desarrollo de este laboratorio vamos a continuar trabajando con el dataset Netflix Movies and TV Shows utilizado en el laboratorio anterior.

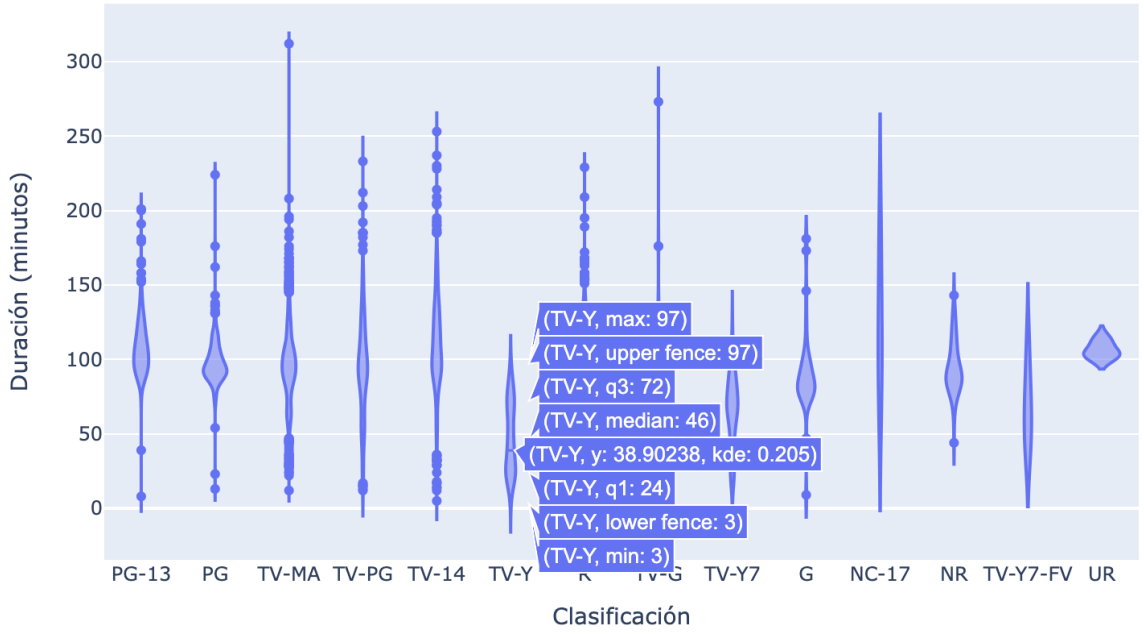
Partiendo del análisis realizado del trabajo anterior, ya estamos al tanto de que las películas con géneros infantiles tienen una duración menor que los dramas. En ese caso vamos a analizar cómo se distribuyen las duraciones de las películas con respecto a las distintas clasificaciones de las mismas.



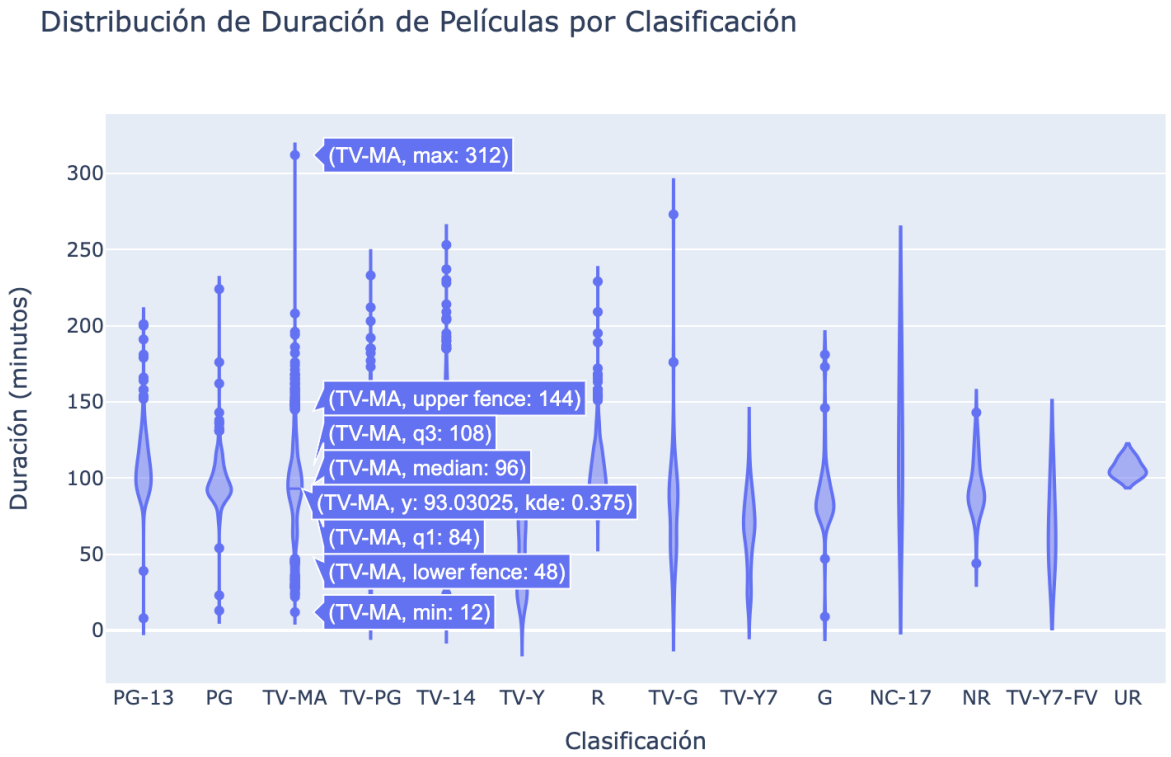
Observando la gráfica anterior parecería que la mayoría de las películas tienen una duración aproximada a los 100 minutos. Sin embargo las películas que tienen una menor duración promedio son las clasificadas con las las categorías TV-Y (Apta para niños de todas las edades), TV-Y7 (Recomendadas para niños mayores de 7 años) y TV-G (Adecuada para todo público en general).

Cómo la gráfica anterior es una gráfica interactiva, si posicionamos el mouse sobre las distintas categorías se puede apreciar el valor medio, máximo, mínimo, entre otros de la misma.

Por ejemplo, analizando particularmente la categoría TV-Y, se puede ver que el tiempo promedio es de 46 minutos y la película más larga dura 97 minutos.



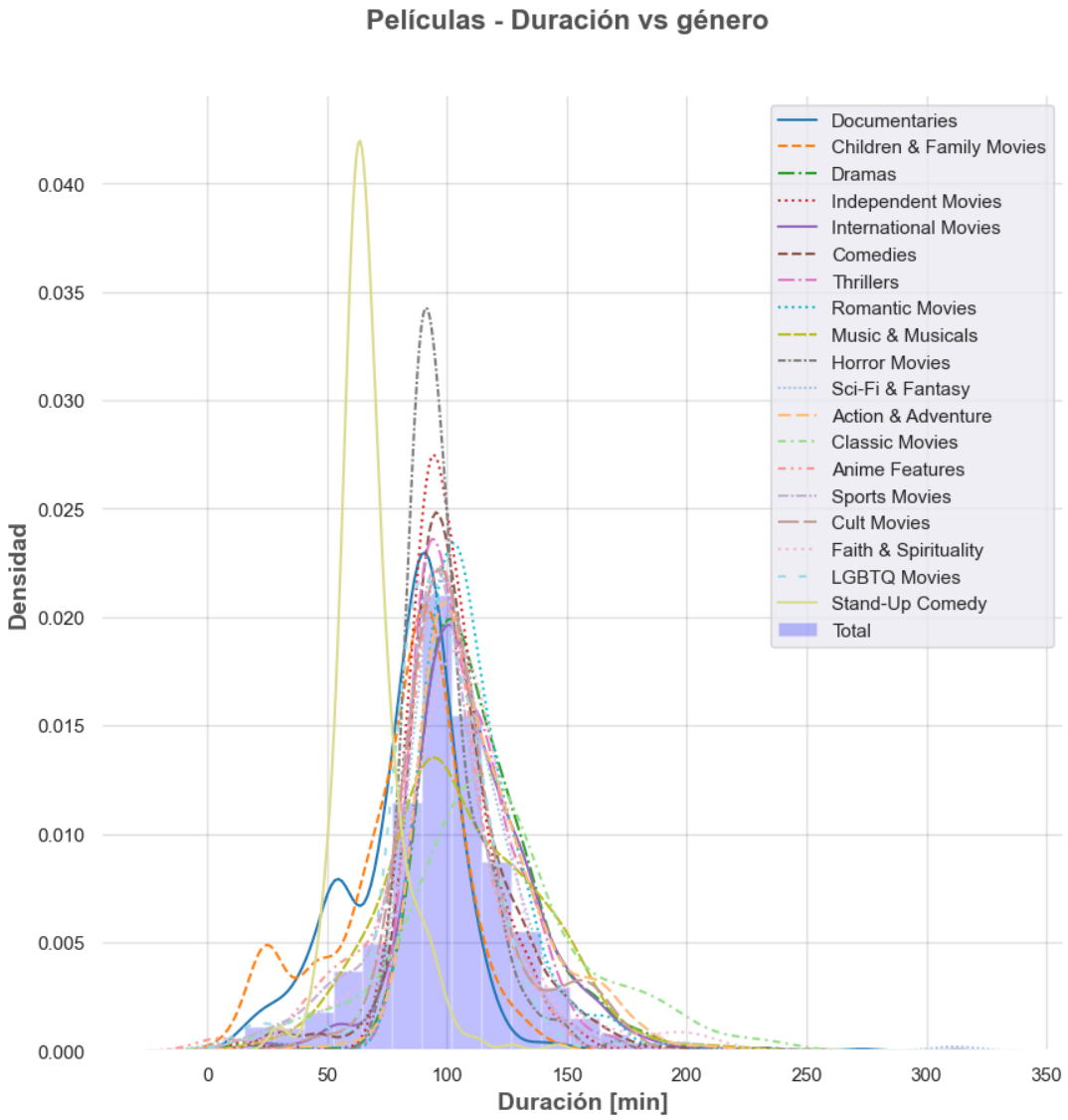
Ahora si analizamos películas con categorías TV-MA (Apta para solo adultos) vemos que la duración media es de 96 minutos existiendo películas de hasta 312 minutos.



Con esta gráfica se podría deducir que hay una relación entre la duración y la clasificación de la película ya que el tiempo promedio de una película solo para adultos coincide con el máximo de una película apta para niños de todas las edades.

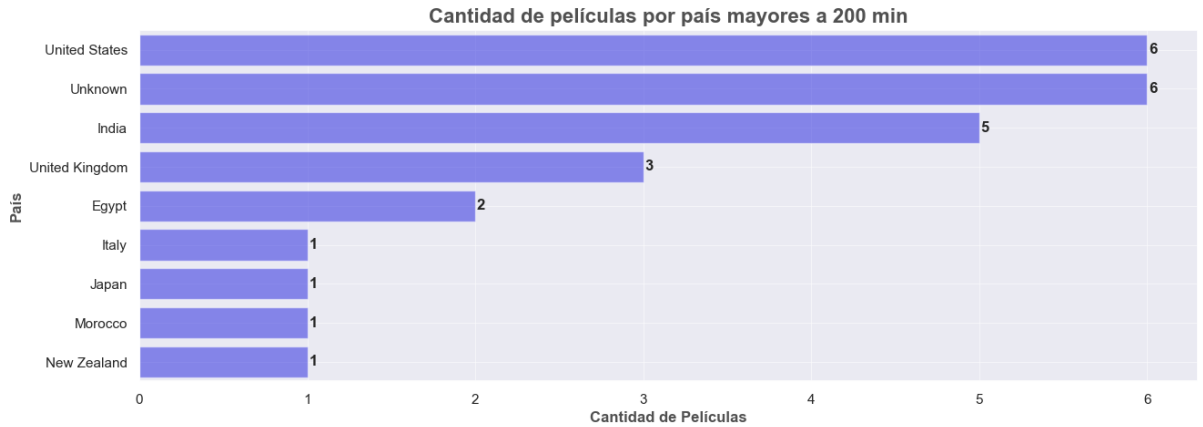
Para poder reproducir la visualización antes mencionada vamos a entregar un notebook en Python en el cual se estuvo trabajando para que se pueda realizar una mayor experimentación.

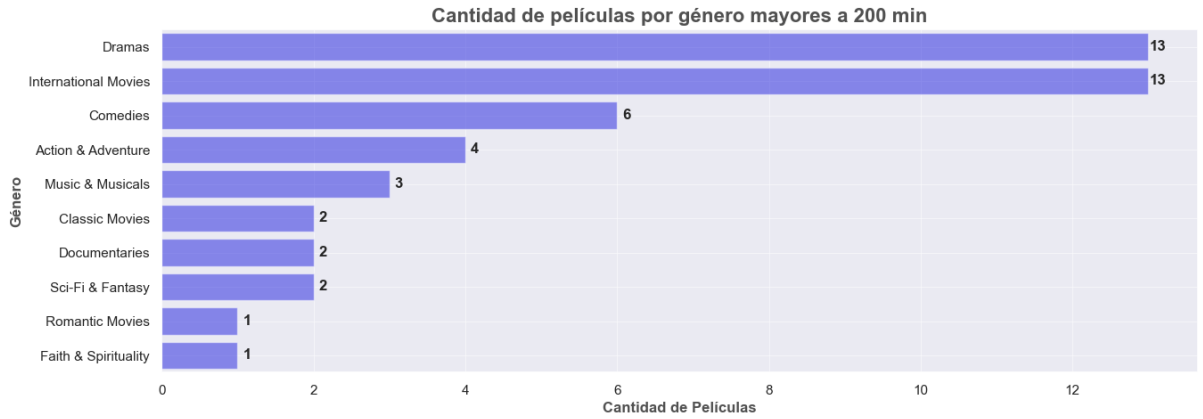
Asimismo, resulta interesante analizar la relación que tiene la duración de una película con el género de la misma. Para ello se realiza un histograma de la duración de todas las películas, y se lo superpone con gráficos de densidad para cada género. Esto es mostrado a continuación.



El histograma de duración para todas las películas es el indicado en barras llenas, mientras que las líneas de densidad corresponden a los distintos géneros. Se utilizan colores y tipos de líneas diferentes para facilitar la distinción entre varios géneros. Es posible ver que la duración media está en el entorno de 100 minutos, algo que parece ser coherente. Asimismo, los shows de stand up parecen tener las duraciones medias más bajas, ya que el pico se encuentra en el entorno de los 60 minutos, lo que tiene sentido ya que este tipo de shows tienden a durar una hora.
También es posible ver que géneros como acción y documentales tienen picos leves en el entorno de los 25 y 50 minutos respectivamente. Por el otro lado, las películas clásicas presentan varias películas en el entorno de los 175 minutos, teniendo al menos una de

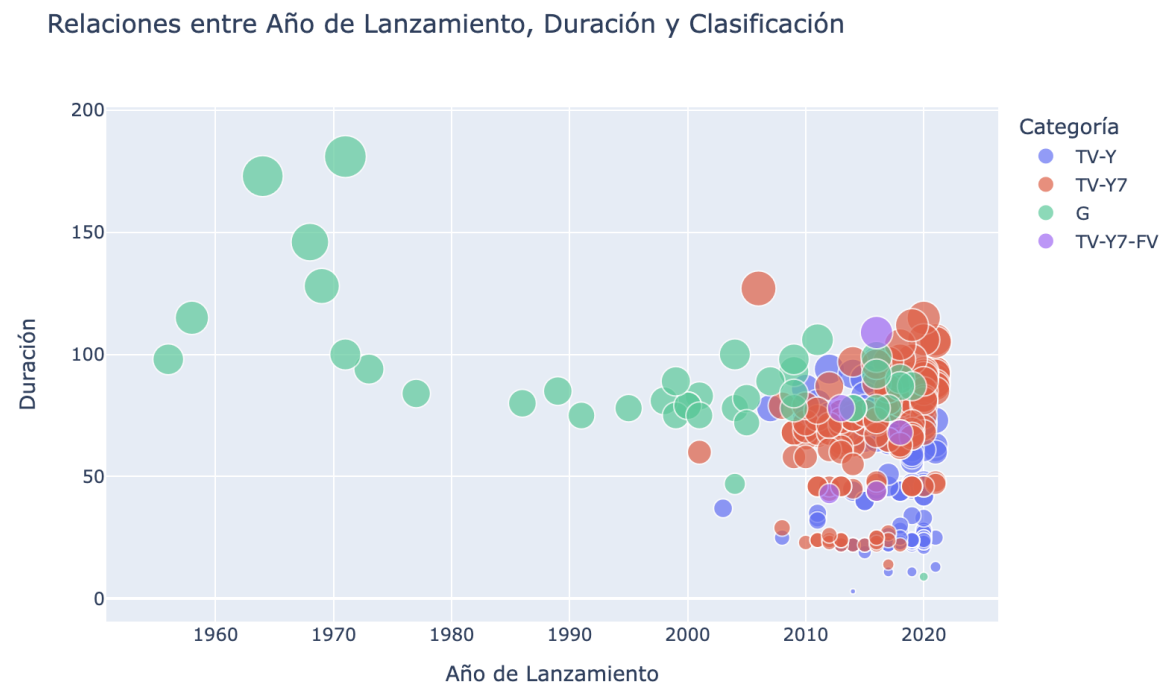
aproximadamente 300 minutos. Si bien las películas de menor duración corresponden a cortometrajes, resulta de interés evaluar el país y género de las de larga duración. Para ello, se toma un umbral de 200 minutos y se realizan las siguientes gráficas.





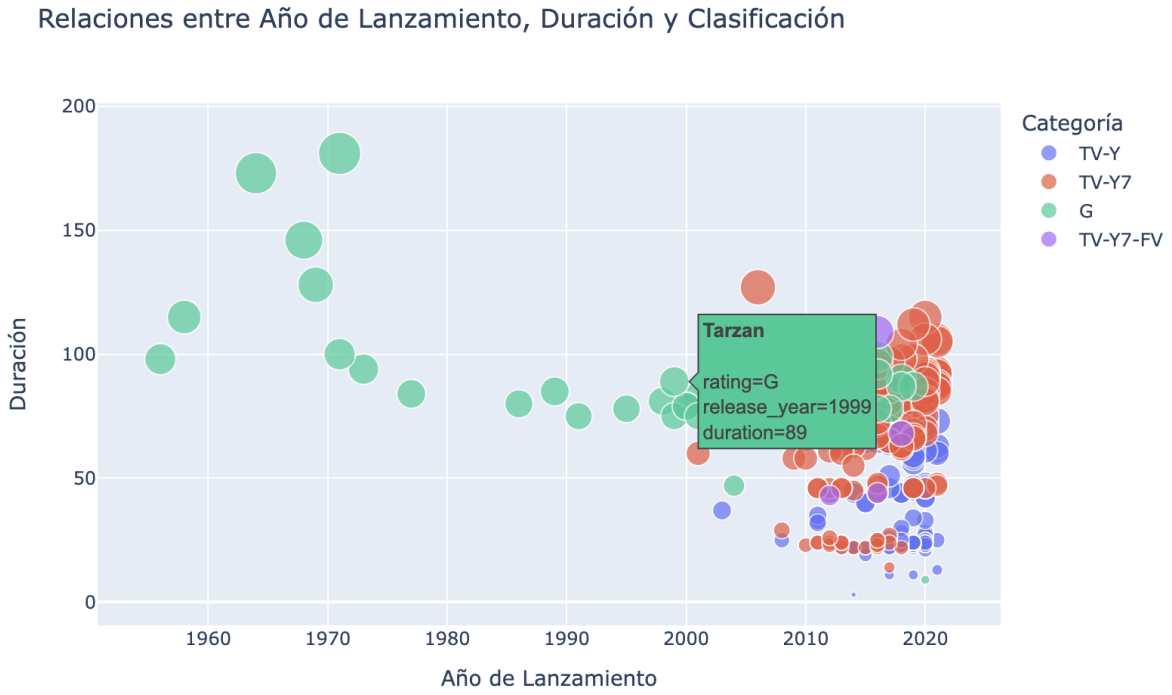
Se aprecia, entonces, que las películas más largas tienden a ser dramas e internacionales, y son producidas mayoritariamente por Estados Unidos, India y Reino Unido. Cabe destacar que estos países eran los mayores productores de películas de manera general, de acuerdo al gráfico presentado en la tarea anterior.

Continuando con el análisis de las películas de categorías infantiles, generamos un gráfico que muestra la relación entre el año de lanzamiento de las películas y las clasificaciones de las mismas. Cómo se puede ver realizamos un filtrado de las categorías que están dirigidas al público infantil.



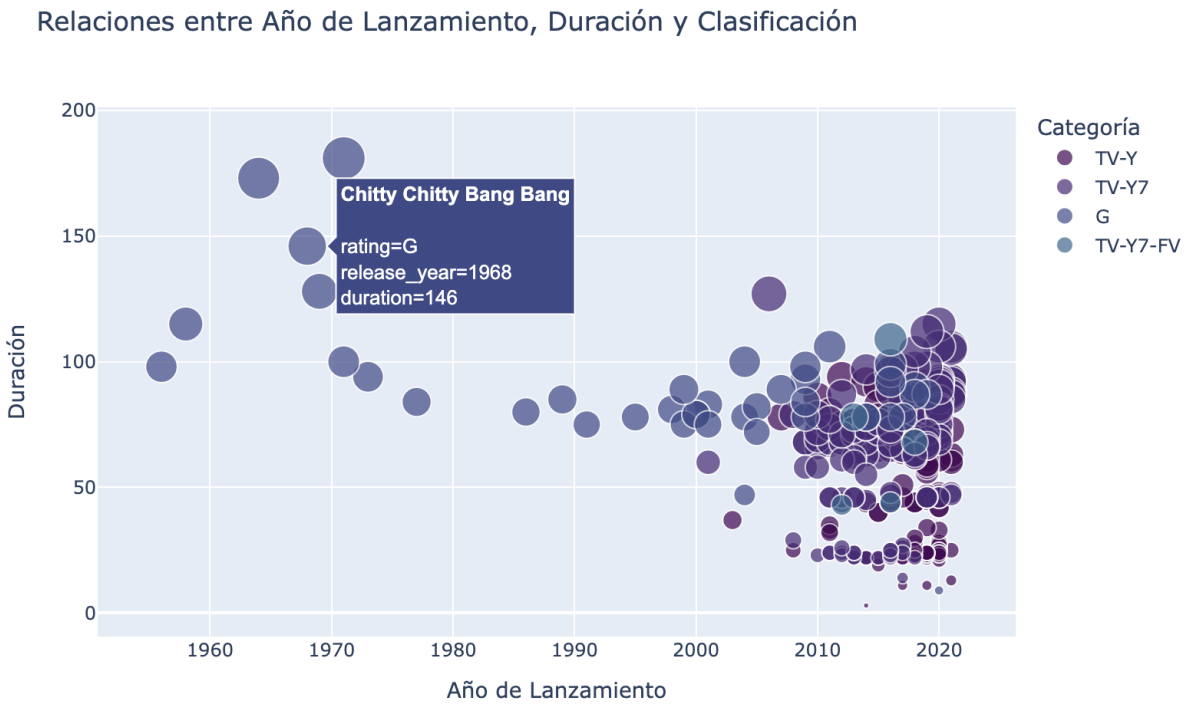
En este gráfico se puede deducir que la clasificación con categorías más específicas para el público infantil comenzó a partir del año 2000 ya que las anteriores están clasificadas con la categoría G (apta para todo público).

La gráfica realizada es también una gráfica interactiva en la que al posicionarse sobre un punto se puede ver el nombre de la película, el rating, el año de lanzamiento y duración.

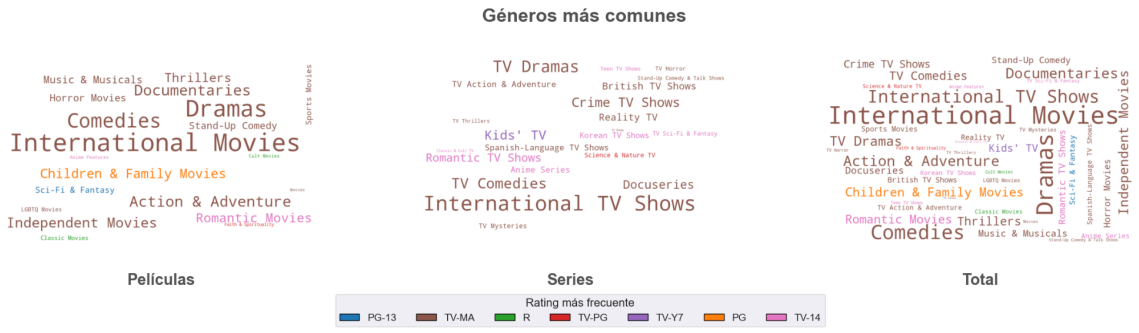


En este caso la diferencia de tamaños de los puntos está indicando la duración, es decir los puntos más pequeños tienen una duración más corta que aquellos puntos con un diámetro más grande.

Como notamos que los colores antes generados pueden ser difíciles de distinguir para personas daltónicas generamos el mimo gráfico pero con la paleta de color viridis para poder facilitarles la distinción de las categorías.

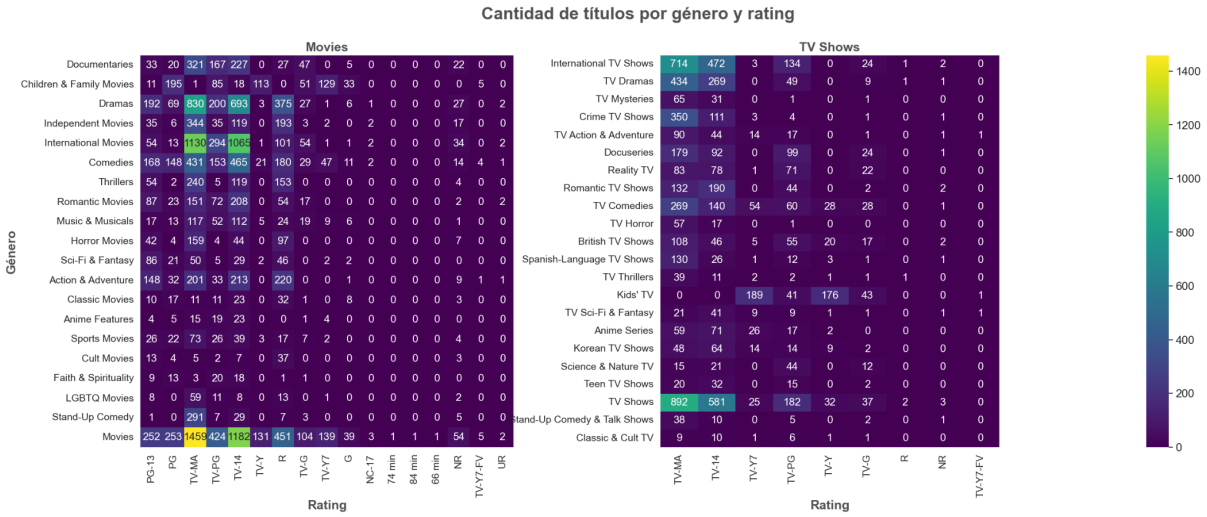


Asimismo, de manera de visualizar los géneros con mayor cantidad de películas, se realiza un gráfico tipo wordcloud, en el cual el tamaño de la palabra se corresponde con la frecuencia del género en el dataset original.



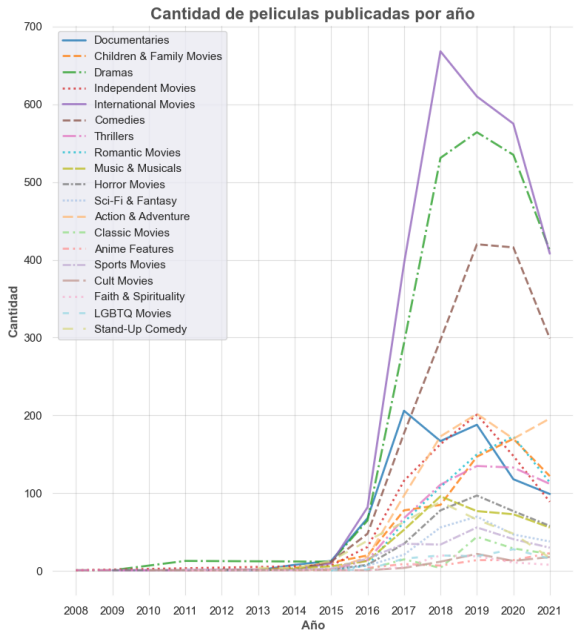
Es posible observar que los títulos internacionales son los más abundantes del catálogo, seguidos por dramas y comedias. Por otro lado, cada género es graficado con el color correspondiente al rating más frecuente para ese género. Es posible apreciar que el rating para adultos (TV-MA) es el más frecuente para una gran cantidad de géneros distintos, seguido por TV-14, el cuál se encuentra en comedias románticas mayoritariamente, y por PG, el cual se encuentra en películas infantiles.

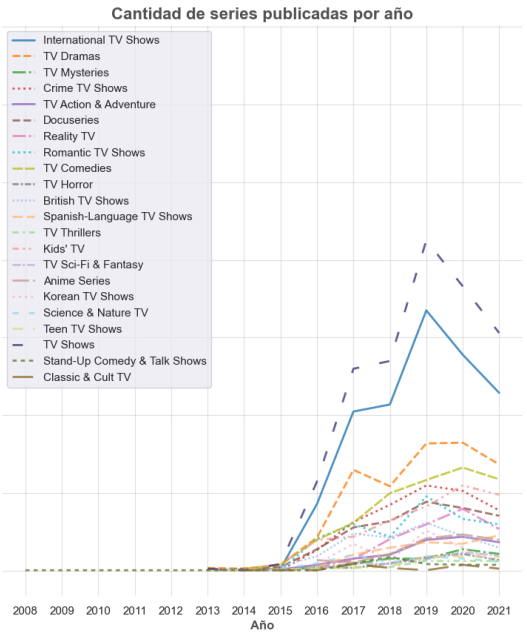
Con el objetivo de visualizar la relación entre género y rating de manera más precisa, se realiza un mapa de calor que indica la cantidad de películas y series para cada par género-rating. Se representa con una escala de color secuencial (viridis), la cual representa los valores más altos en amarillo y los más bajos en violeta.



Dejando de lado los géneros “Movies” y “TV Shows” debido a su generalidad, es posible ver como se confirma lo que sugería la imagen anterior: las películas y series dramáticas e independientes con rating PV-MA son las más frecuentes, seguidas por los mismos géneros pero con ratings PG-14.

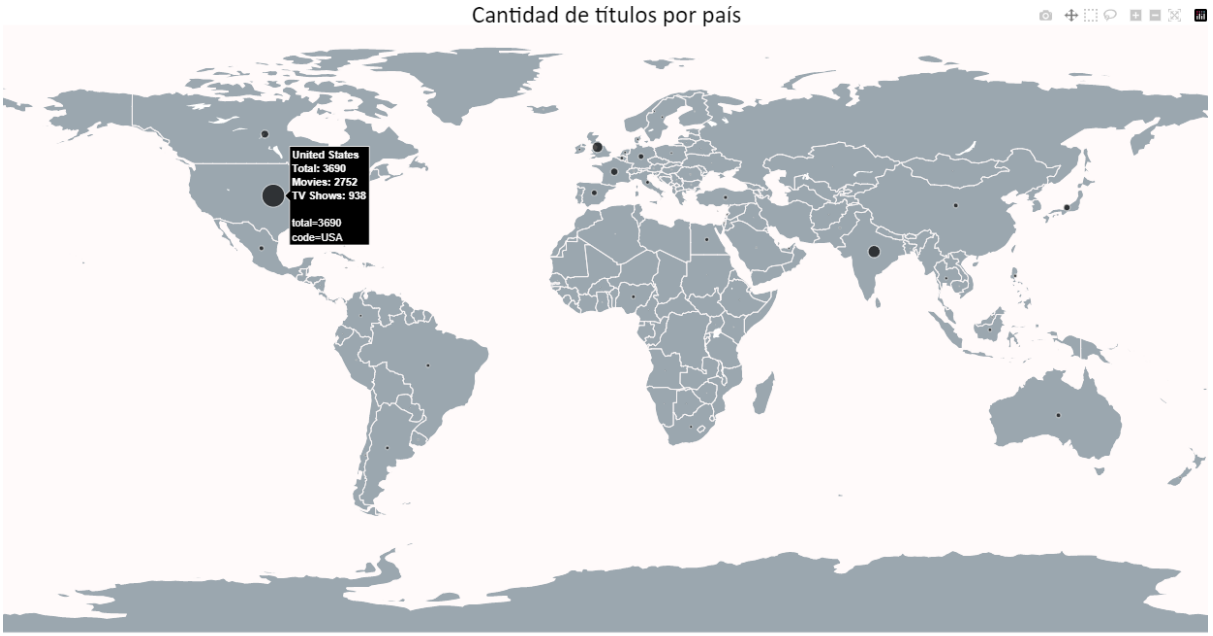
Siguiendo con el análisis de género, se busca visualizar la evolución de la cantidad de publicaciones por año discriminadas por género, tanto para series como para películas. Esto puede verse a continuación.





Recordando el gráfico presentado en la tarea anterior, es posible apreciar que se cumple la misma tendencia para la cantidad total que para cada género por separado, así como también se cumple para series y películas por separado. Asimismo, se confirma nuevamente que las películas y series internacionales y dramáticas son las que más abundan en el catálogo actual.

También resulta de interés visualizar, de manera interactiva, la cantidad de títulos publicados por cada país. Para ello se realiza un gráfico tipo mapa, en el cual se indica, para cada país incluído en el dataset, la cantidad de películas, shows, y total de títulos publicados. El marcador de cada país es proporcional a la cantidad total de títulos publicados por dicho país, razón por la cual Estados Unidos e India tienen los marcadores de mayor tamaño.



En este gráfico también se puede realizar zoom en el continente que se desee. Como se mencionó en la tarea anterior, Estados Unidos e India son actualmente los mayores contribuyentes de contenidos a la plataforma, seguidos por Reino Unido, lo que queda evidenciado por el tamaño de los marcadores en dichos países. A continuación se presenta un zoom de Europa y Sudamérica, a modo de ejemplo.


Romania
Total: 14
Movies: 14
TV Shows: 0
total=14
code=ROU


Uruguay
Total: 14
Movies: 13
TV Shows: 1
total=14
code=URY