ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Leticia Molina Mazarin

Madrid, 12 de mayo de 2022

RESUMEN

El objetivo del presente estudio es realizar un Análisis Exploratorio de Datos (EDA) para poner en práctica algunos de los conceptos aprendidos durante el bootcamp de Data Science en The Bridge.

Para ello, se ha elegido una base de datos de los títulos de Netflix a lo largo de los últimos años y se ha considerado un contexto inicial en el que productores de Netflix en España han encargado dicho estudio para comprobar algunas hipótesis de partida y ayudar en el próximo estreno de producción nacional: La Novia Gitana, basada en el primer libro de la trilogía de novelas negras de Carmen Mola.

El objetivo del equipo es conseguir con que el nuevo título entre en la lista de Top 10 semanales a nivel global por al menos tres semanas. Por lo tanto, quieren entender la oferta de títulos en Netflix a lo largo de los años para comprobar algunas de las siguientes hipótesis de partida:

- La oferta de películas es significativamente mayor que la oferta de series en la plataforma, pero las series vienen ganando importancia a lo largo de los años ya que logran mayor popularidad dentro de la plataforma.
- El mayor productor de contenidos en Netflix es EE. UU., primeramente, por su importante industria cinematográfica y segundo porque Netflix es una empresa norteamericana.
- El mejor momento para un estreno es el fin de semana.
- Los contenidos relacionados con el género Thriller y Horror tienen más estrenos en los meses de septiembre y octubre a causa del Halloween.

ÍNDICE DE CONTENIDOS

1 INTRODUCCIÓN

Netflix ha sido fundada en el año de 1997 en los Estados Unidos como un servicio de entrega de DVD a través de correos y actualmente es la mayor proveedora global de películas y series a través del sistema de *streaming*, siendo la precursora de dicha modalidad.

A partir de 2013, con la serie House of Cards, la compañía empezó a ofrecer contenidos de video producidos específicamente para su servicio de transmisión. Dicho contenido se convirtió en un foco importante de Netflix y, para finales de 2021, ya ofrece más de 2.400 títulos originales y producidos en diversos países.

En enero de 2016, Netflix anunciaba su expansión a 130 países. Los años de 2016 y 2017 marcan, por lo tanto, un antes y después para la empresa y para toda la industria cinematográfica a nivel mundial: la forma como las personas consumían entretenimiento empezaba a cambiar y no había marcha atrás. El *streaming* vino para quedarse y a días de hoy ya existen diversas otras plataformas como Amazon Prime Video, Disney+, Hulu, Apple, ofreciendo el mismo modelo.

Para la realización del presente trabajo de Análisis Exploratorio de Datos, se ha planteado un escenario en el que Netflix España ha comprado los derechos de reproducción de los libros de Carmen Mola y están en el momento de plantear el formato de dicha producción (serie o película) y el mejor momento para lanzarlo.

Para ello, se ha considerado una base de datos de los títulos de Netflix desde 2008 hasta septiembre de 2021 y se procederá a su análisis para sacar conclusiones y comprobar algunas hipótesis de partida.

La primera parte del presente estudio se destinará a presentar los detalles de la base de datos, para que se pueda entender los tipos de datos que están disponibles, como fueron tratados y que el tipo de análisis que se pudo realizar a partir de ellos.

Después se presentará de manera rápida la metodología utilizada para la realización del análisis de los datos y, finalmente, se procederá a la presentación de las conclusiones sacadas a partir del análisis.

2 DATOS DE REFERENCIA

Para la realización del presente EDA, se ha considerado una base de datos de Netflix disponible en Kaggle con los títulos añadidos a la plataforma desde el año de 2008 hasta septiembre de 2021.

La base cruda cuenta con 8.806 filas de títulos de Netflix y 12 columnas con diferentes informaciones. Toda la base de datos está en inglés y se ha optado por mantenerla en el idioma original. Las columnas disponibles en la base de datos son las siguientes:

- La columna 'show_id' no nos aporta información adicional, se optará por eliminarla.
- La columna 'type' enseña si determinado título es una película o una serie.
- La columna 'title' enseña el nombre de la serie o película en inglés.
- También es posible ver la información sobre el director o directora y el cast en las columnas ‘directo’ y ‘cast’ respectivamente.
- La columna 'production_country' enseña el país de producción de dicho título.
- En la columna 'date_added' es posible ver la fecha en la que determinado título fue añadido a la plataforma de Netflix, mientras que la columna 'release_year' enseña el año en el que ese título tuvo su lanzamiento dentro o fuera de la plataforma. Sabemos que hay muchos títulos que ya son originales de Netflix y se lanzan directamente en la plataforma, pero hay algunas películas o series que no son originales de Netflix y tuvieron un lanzamiento antes de entrar en la plataforma.
- En la columna 'rating' se puede ver la clasificación etaria de los títulos y tener una idea del público a que se destina determinada serie o película.
- La columna 'duration' enseña cuánto tiempo dura una película en minutos y una serie en cantidad de temporadas.
- La columna 'listed_in' enseña las categorías de cada título y 'description' nos trae una descripción de la película o serie.

Con esta base de datos será posible realizar una serie de comprobaciones como las que se presentarán más adelante.

3 METODOLOGÍA

El lenguaje utilizado para el tratamiento y análisis de los datos ha sido Python y todo desarrollo del código se hizo a través del editor de código Visual Code Studio. Se utilizaron algunas librerías de Python como Pandas, Numpy, Matplotlib, entre otras que vienen especificadas en el fichero de librerías.

Antes de empezar con el análisis y visualización de los datos, un primer paso imprescindible es el entendimiento de la base de datos y la limpieza y tratamiento de los datos. Algunos de los pasos de la limpieza han sido:

- Mirar el tamaño de la base de datos original.
- Entender que columnas son las más importantes para el análisis.
- Mirar si existen valores nulos y eliminarlos o tratarlos.
- Tratar la columna de fecha para que tenga el formato adecuado, separando también en distintas columnas los meses, años y días de la semana de los estrenos.
- Tratar la columna de país de producción porque algunos títulos cuentan con más de un país de producción.
- Tratar la columna de categorías ya que los títulos se encajan en más de una categoría.

Después de la limpieza y tratamiento de los datos, se procede al análisis de los mismo para sacar conclusiones acerca de la hipótesis de partida comentadas anteriormente.

4 ANÁLISIS DE LOS DATOS

Este capítulo se dedica a la presentación de las conclusiones sacadas del análisis y está dividido en los siguientes apartados: Netflix a lo largo de los años, estacionalidad de la oferta de contenidos, principales países productores de contenido, análisis del formato de los contenidos películas vs. series, público-objetivo, las principales categorías de las películas y series y, finalmente, un análisis más al detalle de las producciones españolas en Netflix.

4.1. Netflix a lo largo de los años

Netflix se hizo global en el año de 2016 con su expansión a más de 130 países. Por lo tanto, es esperado que la oferta de contenidos haya crecido a lo largo de los años. De hecho, en el gráfico 1 es posible ver un importante salto en la oferta de contenidos entre los años de 2016 y 2017.

Sin embargo, a causa de la pandemia del coronavirus en el 2020, muchas producciones tuvieron que detenerse y muchos lanzamientos se pospusieron. Podemos también comprobar la caída en la oferta de contenidos en dicho año.

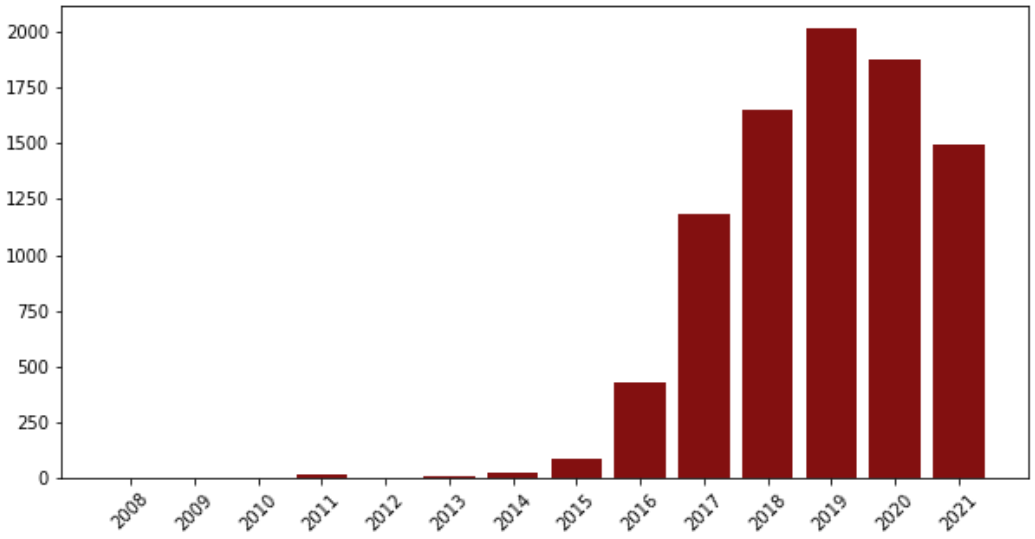

Gráfico 1: evolución de los contenidos en Netflix a lo largo de los años (contando con 2021)

El gráfico 1 muestra la evolución en la oferta de contenidos a lo largo de los años contando también con los datos del año de 2021. Dicha oferta venía creciendo y su tendencia al alza efectivamente cambió en el año 2020 muy probablemente a causa de la pandemia.

Sin embargo, es importante mencionar que la base de datos utilizada solo cuenta con títulos subidos hasta septiembre del 2021, por lo que faltan 3 meses de información y no es posible concluir si la oferta de contenidos ha sido mayor o menor en el 2021 que en el 2020.

Para ello, se crea un nuevo gráfico considerando solamente la oferta de contenidos hasta septiembre de cada año:

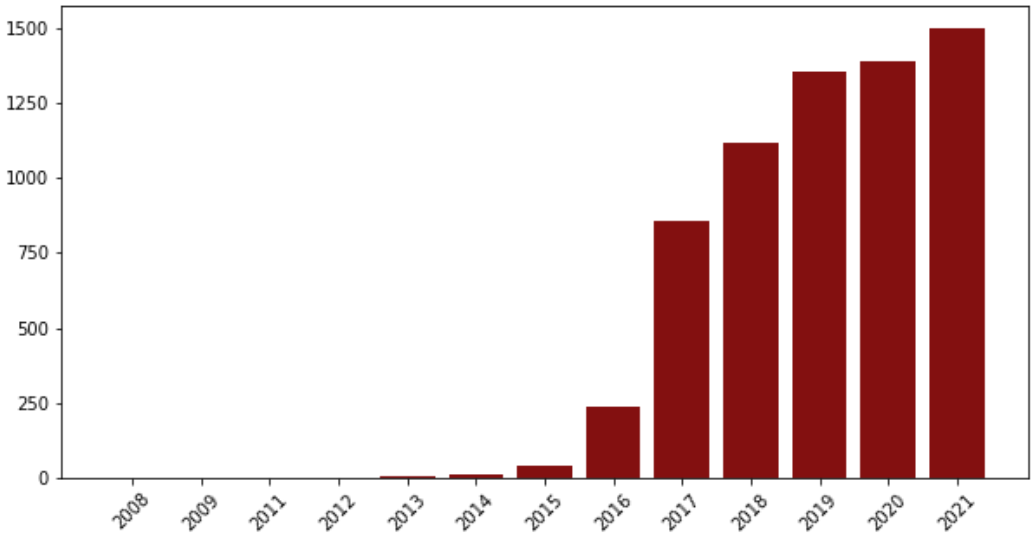

Gráfico 2: evolución de los contenidos en Netflix a lo largo de los años (contando solo con datos hasta septiembre de cada año)

En el gráfico 2 se ve un aumento en la oferta de contenidos dentro de la plataforma en el 2021 (por lo menos hasta septiembre). Mientras que en el año 2020 es posible ver claramente que la tendencia de aumento en la oferta de contenidos ya empieza a caer y se podría concluir que el último trimestre del año ha sido el más afectado.

De hecho, al hacer un tercer gráfico solo con los datos del último trimestre de cada año, queda bastante claro que el cuarto trimestre del 2020 fue el más afectado por la pandemia:

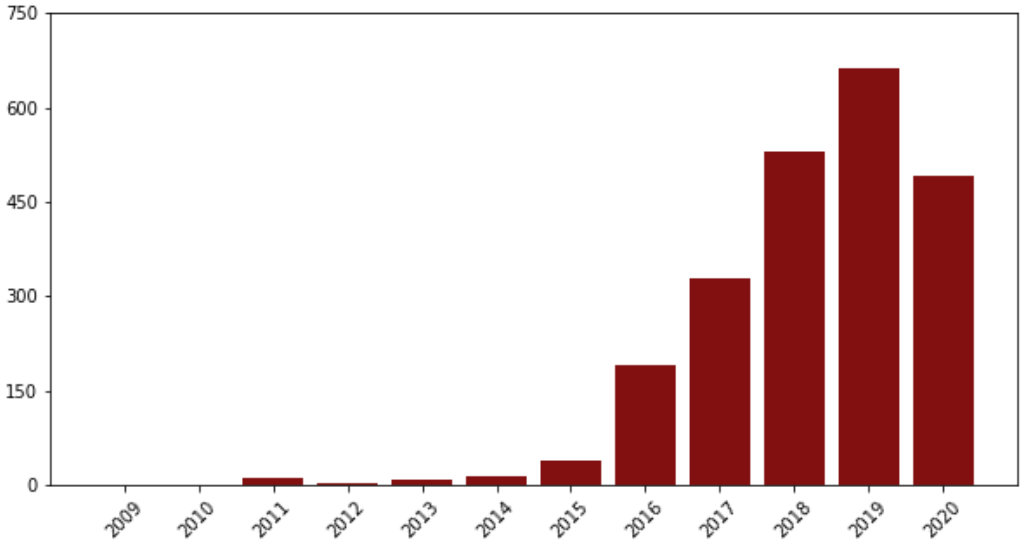

Gráfico 3: evolución de los contenidos en Netflix a lo largo de los años (contando solo con datos del último trimestre de cada)

4.2. Estacionalidad en la oferta de contenidos en Netflix

Después de ver la evolución en la oferta de contenidos a lo largo de los años, se verifica la estacionalidad de la oferta de contenidos en Netflix: ¿hay meses en los que la oferta de contenidos aumenta o disminuye?

El gráfico 4 muestra la oferta por mes de toda la base de datos. Sin embargo, como mencionado anteriormente, el hecho de no contar con los datos del último trimestre del 2021 podría estar afectando la información, por lo que se decide también realizar un segundo análisis descartando el año de 2021.

El gráfico 5 muestra la oferta de contenidos por mes descartando los datos de 2021 y entonces es posible concluir que efectivamente el último trimestre de los años suele ser el más importante para la inclusión de nuevos contenidos. De hecho, para poder concluir esto con seguridad, se realiza un test de hipótesis con chi cuadrado en Python y se concluye que es posible rechazar la hipótesis nula y afirmar que existe una estacionalidad en la oferta de nuevos contenidos en Netflix.

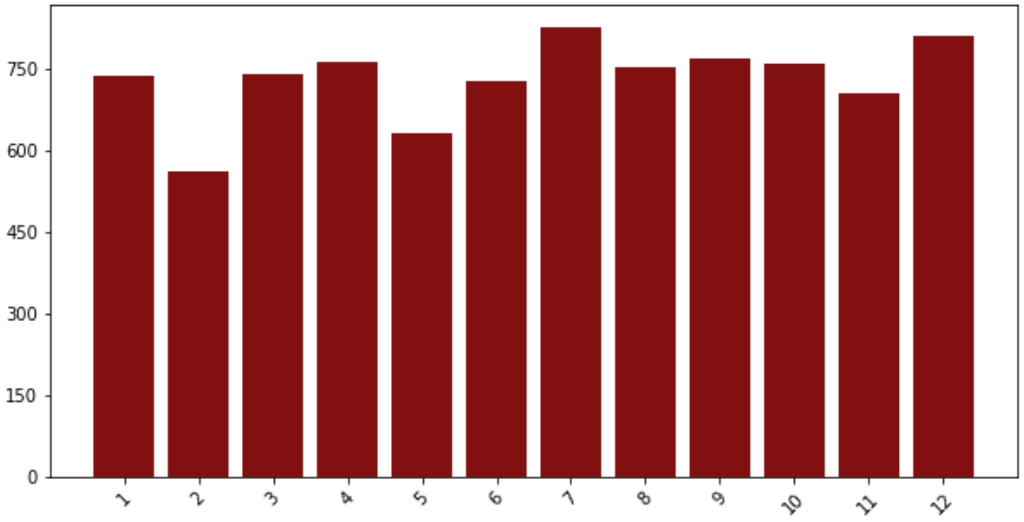

Gráfico 4: oferta de contenidos en Netflix a lo largo de los meses (contando el año de 2021)

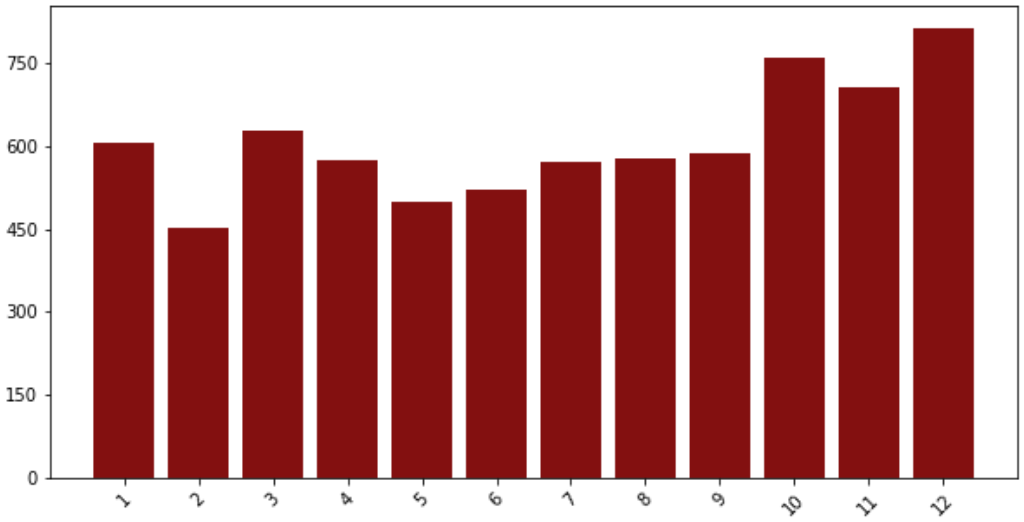

Gráfico 5: oferta de contenidos en Netflix a lo largo de los meses (descartando el año de 2021)

Además de verificar la estacionalidad por meses, se verifica también los días de la semana en los que se suele realizar los estrenos. La hipótesis de partida de los productores era de que los estrenos se hacían los fines de semana. Se concluye que los viernes son el día de la semana

preferido para los estrenos seguido del jueves, mientras que el día con menos estrenos es el domingo.

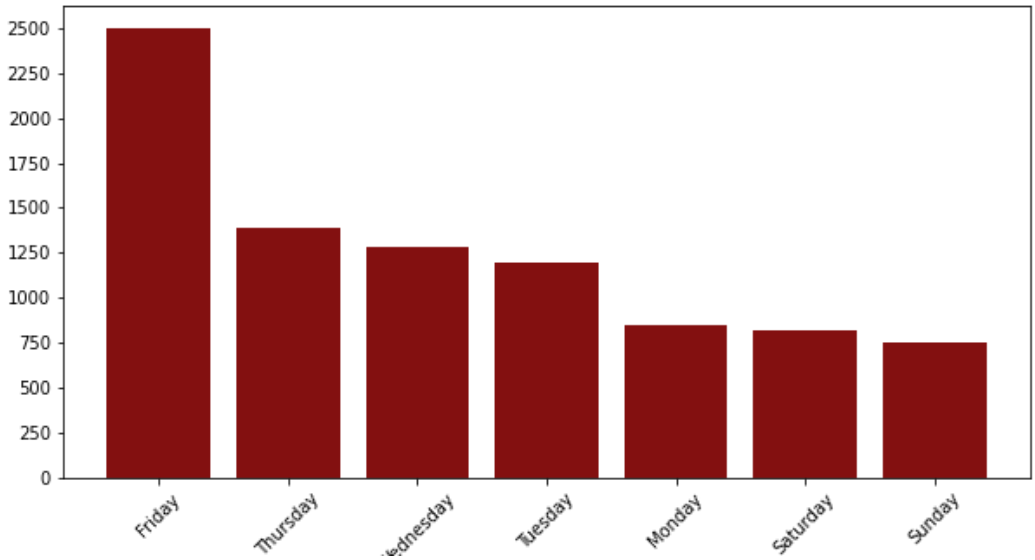

Gráfico 5: oferta de contenidos en Netflix por días de la semana

4.3. Principales países productores de contenido para Netflix

Este apartado está dedicado al entendimiento de los principales países productores de contenido para Netflix. El modelo de *streaming* ha democratizado mucho la oferta y el consumo de contenidos de distintos países, con Netflix y otras plataformas de *streaming* muchos países están teniendo la oportunidad de romper fronteras y llevar sus producciones a otras zonas geográficas.

Sin embargo, como es esperado, el principal país productor de contenido en Netflix sigue siendo EE. UU.. Primeramente, porque Netflix es una empresa norteamericana y después porque la industria cinematográfica del país es la más fuerte del mundo. También es muy conocida la industria cinematográfica de India (Bollywood) y aquí se refleja esto, ya que este es el segundo país con más producciones en Netflix, seguido de UK, Canadá y Japón. España aparece en octavo lugar, siendo que la base de datos analizada cuenta con 84 valores únicos para la columna de países.

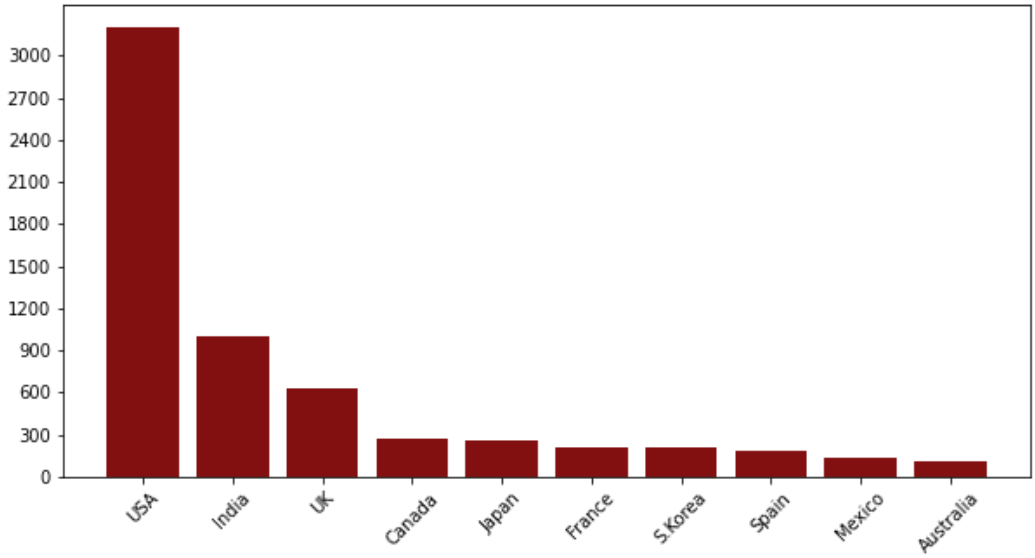

Gráfico 6: oferta de contenidos en Netflix por país de producción

4.4. Películas vs. series

La oferta de películas sigue siendo mucho mayor que la oferta de series dentro de Netflix. Esto se da porque las películas existen desde hace más tiempo, seguramente cuestan menos tiempo y menos dinero para producirse y son más rápidas de ver, por lo que la demanda de películas probablemente es más dinámica.

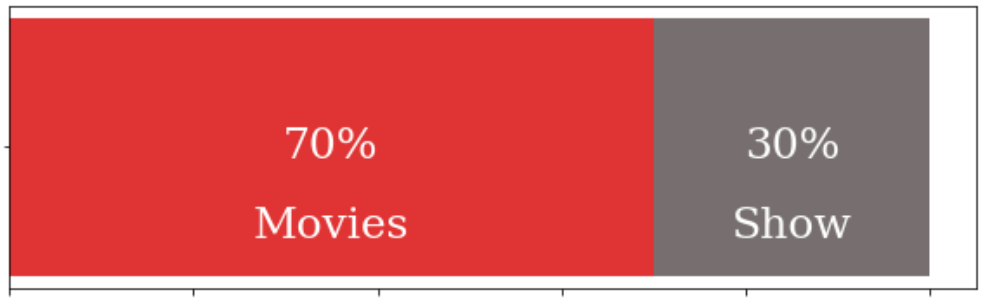

Gráfico 7: distribución de películas y series dentro de Netflix

Otra información interesante que se puede sacar es la distribución de tipo de contenido por países. En el siguiente gráfico es posible ver que India enfoca más del 90% de su producción a las películas y EE. UU. más de un 73%. España también cuenta con más películas que series en Netflix, con una distribución en proporción bastante similar a la de EE. UU.. Por otro lado, Japón y Corea del Sur invierten esa tendencia y cuentan con 67% y 78% respectivamente de sus contenidos en formato de series.

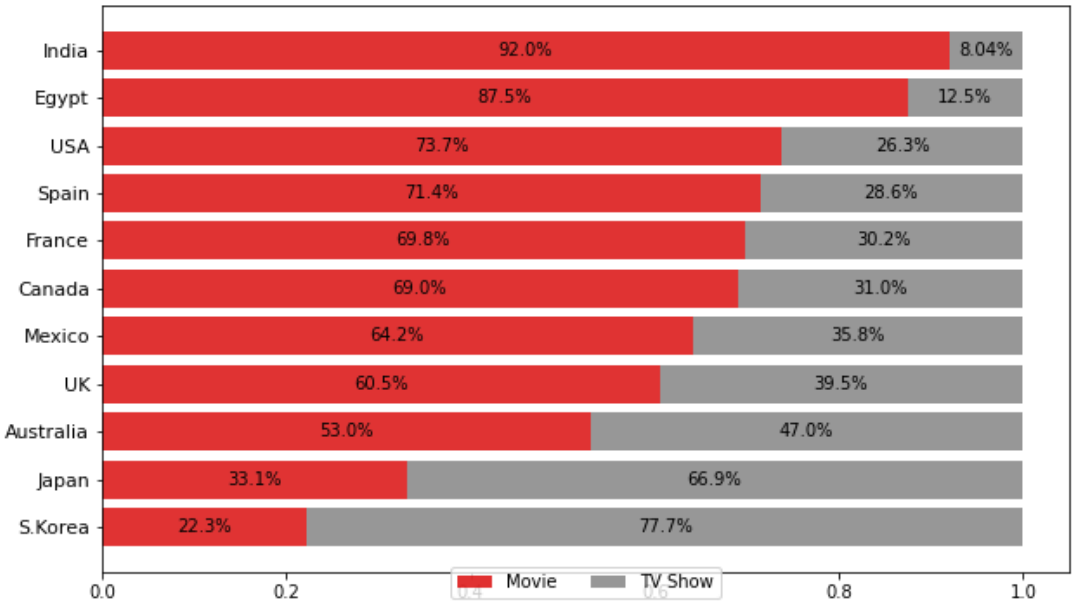

Gráfico 7: distribución de películas y series dentro de Netflix por país

4.5. Público-Objetivo

La clasificación por edades también es otro aspecto interesante que analizar. Aunque Netflix ha lanzado una parte de la plataforma exclusiva para el público infantil, el público adulto y adolescente sigue siendo el principal público-objetivo de la plataforma.

A continuación, es posible ver la distribución de público-objetivo por país productor, teniendo en cuenta los 10 principales países productores de la plataforma. La gran mayoría de las

producciones españolas está enfocada al público adulto mientras que India se dedica a atender al público adolescente. El país que cuenta con más producciones para el público infantil (kids y toddlers) es Australia seguido de Canadá.

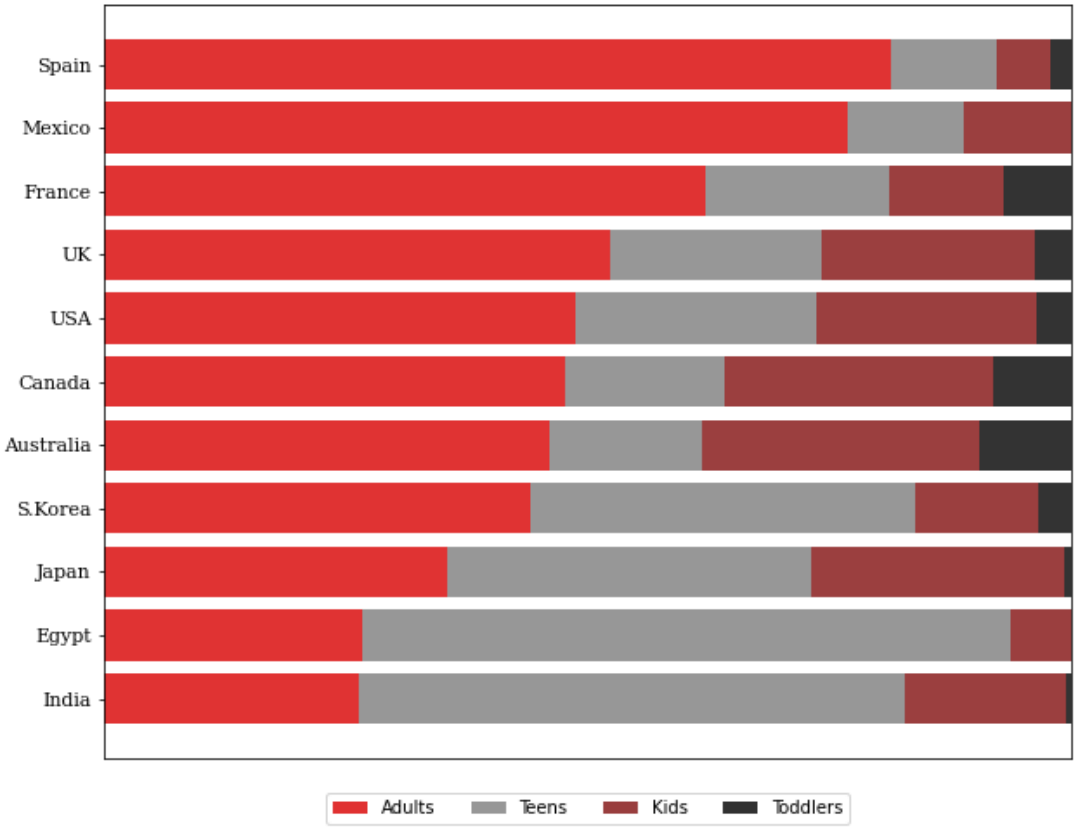

Gráfico 7: distribución de películas y series dentro de Netflix por país

4.6. Categorías

A partir de este punto se empieza a afilar un poco más sobre las categorías de series y películas para finalmente mirar más al detalle información sobre el género Thriller y Horror en el mundo y en España.

El gráfico 7 enseña las diez principales categorías por película y se concluye que los dos principales géneros son Drama y Comedia. Los géneros de Horror y Thriller se quedan en sexto y octavo lugares respectivamente.

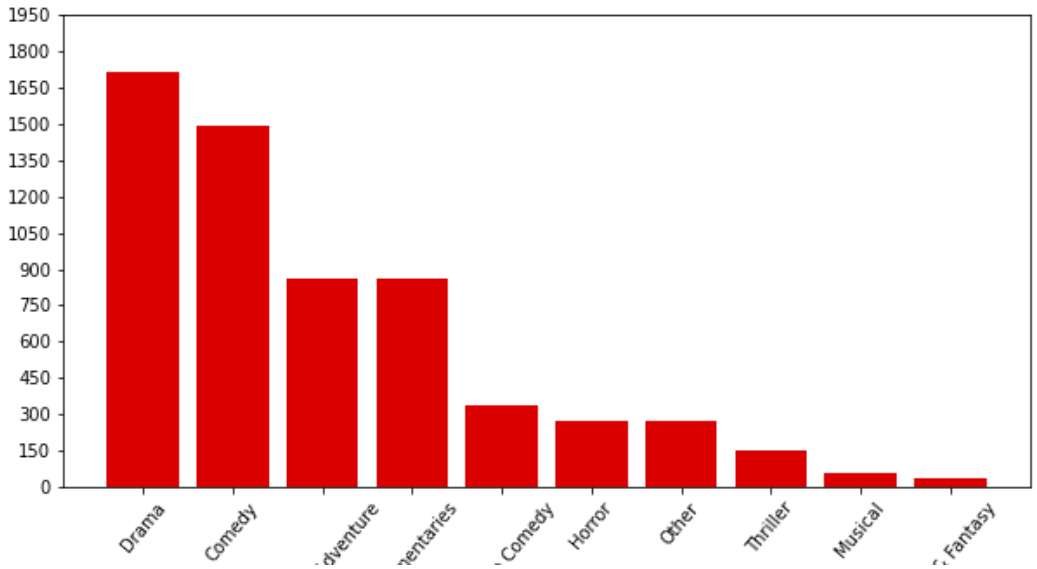

Gráfico 7: principales categorías por película

Por otro lado, cuando se miran las categorías por series, en el gráfico 8, el género Thriller es el que más éxito tiene. Una explicación puede estar en el concepto de *cliffhanger* (final en suspenso, en traducción libre), que es un recurso narrativo que consiste en colocar una situación extrema al final de un capítulo, generando una tensión en el espectador que aumenta su deseo de avanzar en la historia. Dicho recurso puede ser muy bien explorado en los géneros Thriller y debe de ser muy importante en las series porque hay más posibilidades de que el espectador abandone una serie que una película, ya que esta última demanda menos inversión de tiempo por parte del espectador, y el *cliffhanger* es una manera muy eficiente de evitar esto.

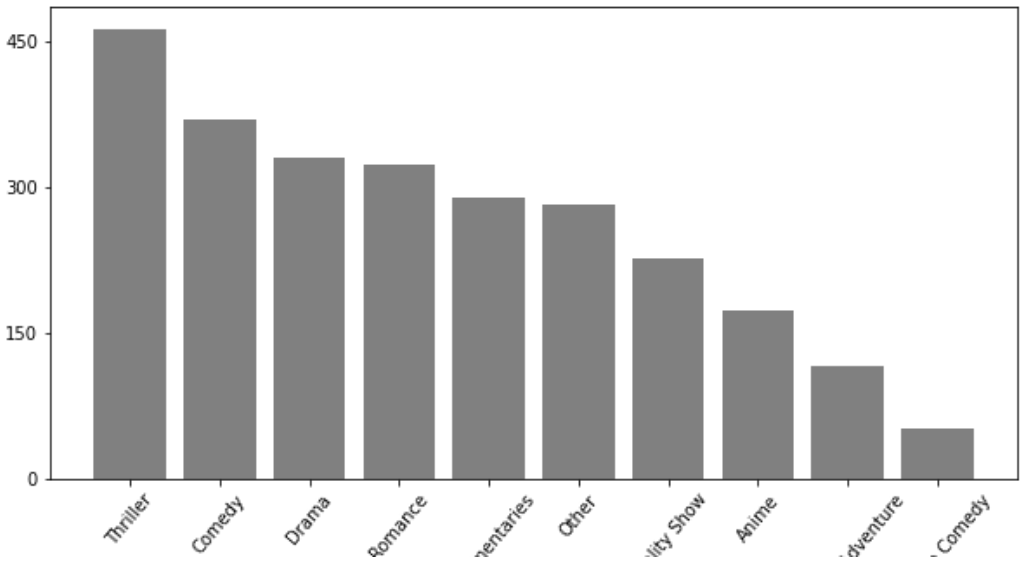

Gráfico 8: principales categorías por series

Entrando más en detalle acerca de la producción de series y películas de los géneros Thriller y Horror, los productores de Netflix que detienen el derecho de reproducción de los libros de Carmen Mola creen que los estrenos de estos géneros suelen hacerse en los meses de septiembre y octubre a causa de Halloween.

Sin embargo, después de algunos análisis de los datos, es posible concluir que no hay una correlación entre los meses de los años y el lanzamiento de contenidos de esas dos categorías. Para llegar a esa conclusión, primeramente, se hizo un gráfico (gráfico 9) de los estrenos de las dos categorías mencionadas a lo largo de los meses, eliminando los datos de 2021.

A simple vista parece que sí hay un aumento en la oferta de thrillers a partir de agosto, pero como los números no son muy significativos, es necesario aplicar un test de hipótesis para comprobarlo y, al hacerlo, se halla un p valor de 0.25 y, por lo tanto, se acepta la hipótesis nula de que no hay una relación entre los meses y los estrenos de Thriller y Horror.

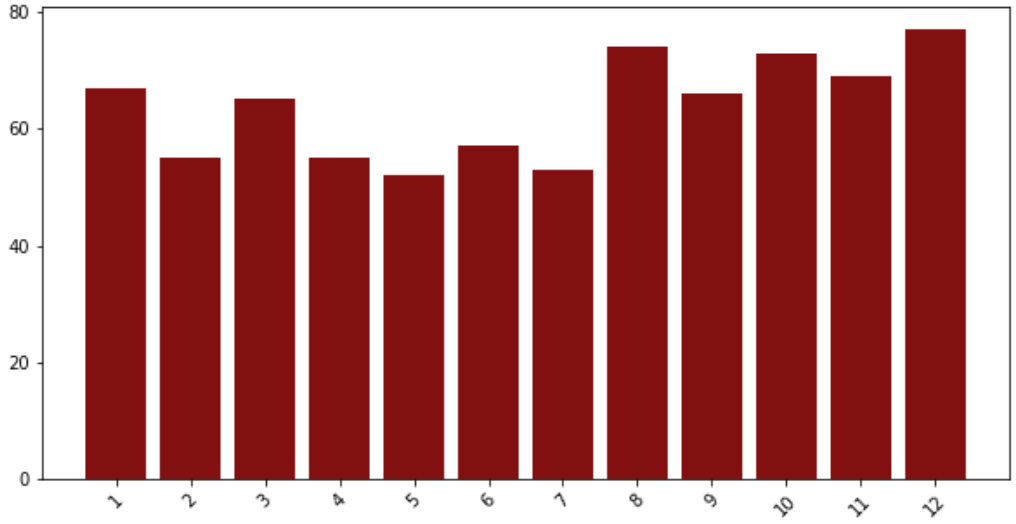

Gráfico 9: estrenos de Thriller y Horror por mes

4.7. Producciones Españolas

Por fin, se analiza la oferta de contenidos producidos en España y se concluye que la mayoría de las películas españolas son de Drama y Comedia en ese orden, véase el gráfico 10. Por otro lado, la gran mayoría de las series son de Thriller, siguiendo la tendencia general ya vista anteriormente (gráfico 11).

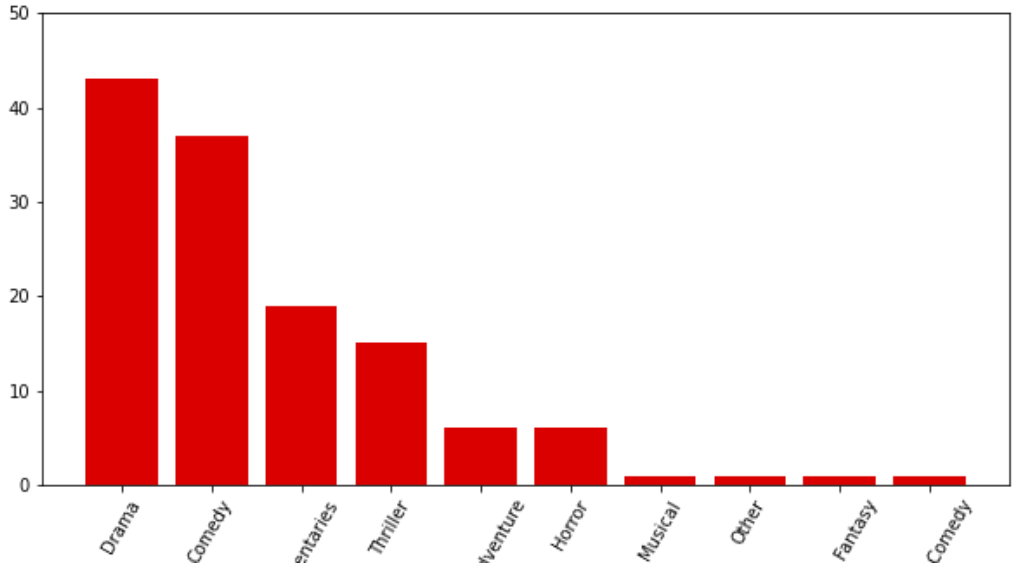

Gráfico 10: género de películas españolas en Netflix

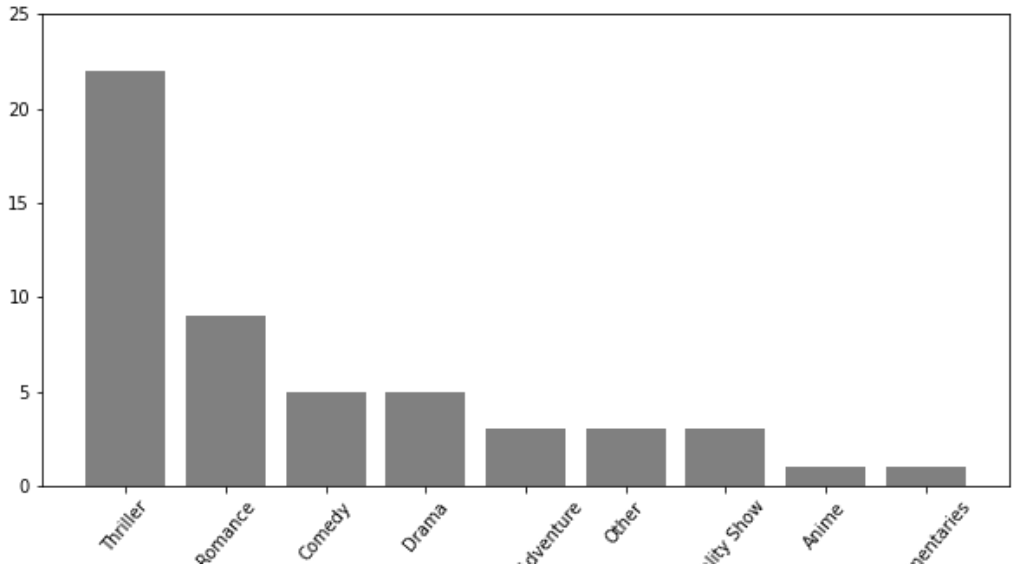

Gráfico 11: género de series españolas en Netflix

Para entrar en más detalles acerca de la popularidad de los títulos españoles en Netflix, se utiliza una segunda base de datos, disponible en la página de Netflix, que consiste en una relación de los títulos que estuvieron en el listado de Top 10 semanales globales desde junio de 2021 hasta abril de 2022.

Después de realizar el tratamiento de los datos y proceder con su análisis, se concluye que casi un 70% de las veces en las que un contenido de producción nacional estuvo en el listado de top 10 semanales a nivel global, ese contenido era en formato de serie. Eso puede darse muy probablemente al hecho de que las series demandan más tiempo para verse y por eso logran quedarse más semanas en los top 10.

Gráfico 12: distribución entre películas y series españolas en la lista de top 10 semanales a nivel mundial

También es posible concluir que el género de producción española que más logra estar en las listas de top 10 globales es el género Thriller, que cuenta con series y películas con temas relacionados a crímenes y misterios como Elite y La Casa de Papel.

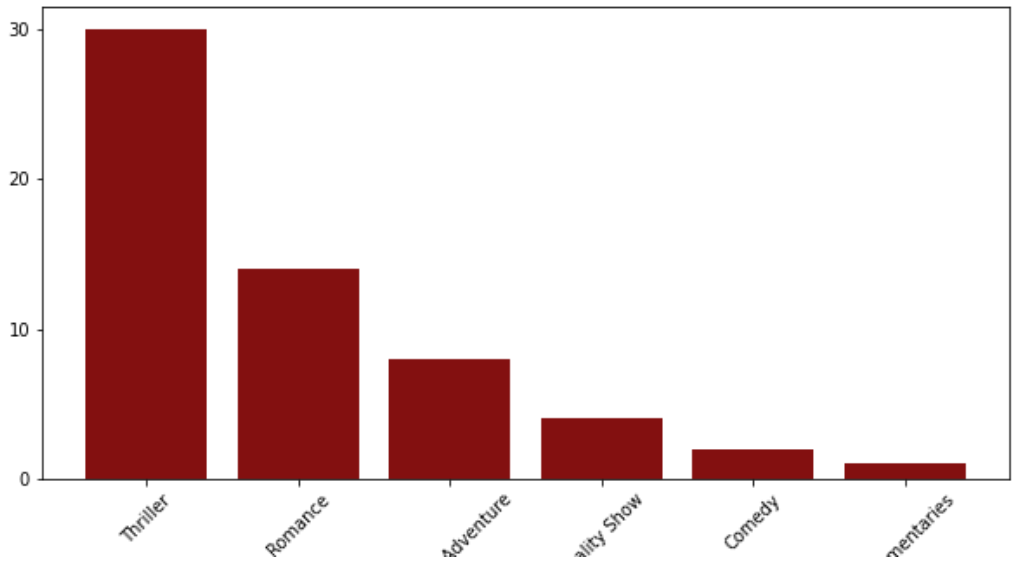

Gráfico 13: top 10 producciones españolas por categoría

Finalmente, en el gráfico 14 se presentan los 10 títulos españoles más populares en Netflix desde junio de 2021 hasta abril de 2022. La Casa de Papel sigue siendo la producción española más popular en la plataforma, llevando 14 semanas en la lista de top 10 a nivel mundial, seguida de Élite, que lleva 8 semanas.

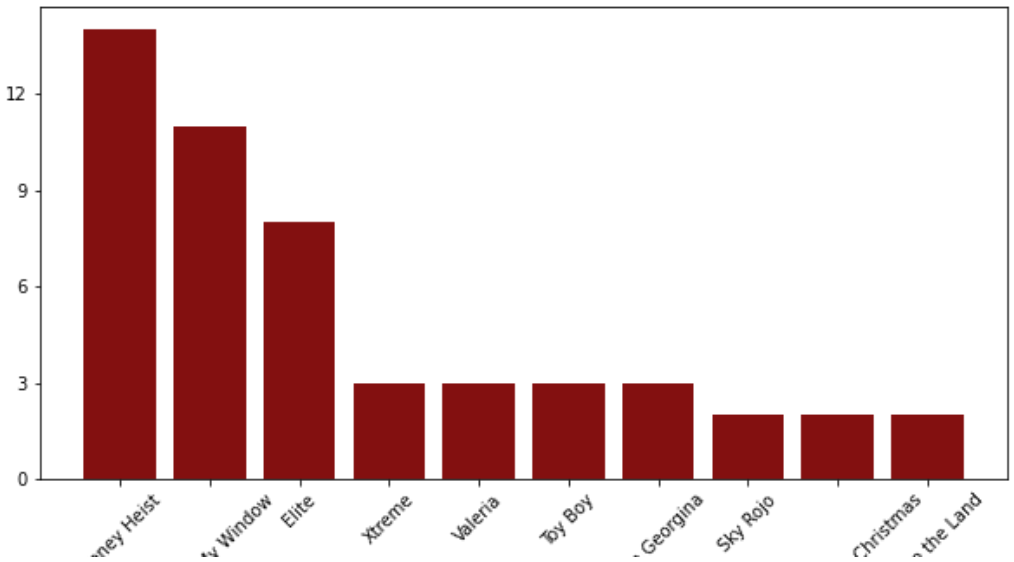

Gráfico 13: top 10 producciones españolas

CONCLUSIÓN

Después de la realización del EDA acerca de los contenidos disponibles en Netflix, se puede concluir que la gran mayoría de los contenidos sigue siendo películas, pero que países asiáticos como Japón y Corea del Sur se dedican a producir más series.

El principal público-objetivo es el público adulto seguido por el público adolescente. Las producciones españolas son en su gran mayoría, más del 80%, dedicadas al público adulto mientras que las producciones asiáticas como India, Japón y Corea del Sur tienen a los adolescentes como sus principales públicos-objetivo.

Existe una estacionalidad en la oferta de nuevos contenidos dentro de la plataforma de Netflix siendo el último trimestre del año el que más cuenta con estrenos. Además, la gran mayoría de los estrenos ocurren los viernes y el domingo es el día de la semana que cuenta con menos estrenos.

Aunque haya una estacionalidad en la oferta de contenidos dentro de Netflix, cuando se analiza solo la oferta de series y películas de los géneros Thriller y Horror no es posible afirmar que haya una correlación con los meses del año, por lo que no es posible afirmar que en los meses de septiembre y octubre se estrenen más contenidos de esos dos géneros a causa del Halloween.

El género Thriller es la principal categoría de las series disponibles en Netflix a nivel mundial y la tendencia sigue la misma para las series producidas en España. Esto se debe muy probablemente al hecho de que dicho género cuenta mucho con el recurso conocido como *cliffhanger* que hace con que el espectador quiera seguir viendo el próximo episodio y que funciona muy bien con el concepto de series.

Después de este análisis es posible concluir que el mejor formato para lanzar ‘La novia gitana’ es en formato serie y que su estreno sea un viernes. No es posible recomendar con seguridad el mejor mes, aunque por estrategia (para tener menos competencia), se recomendaría que el lanzamiento fuese entre julio y septiembre, durante las vacaciones de verano, por ejemplo.