PROYECTO EDA – José Yaya

INTRODUCCION

El presente proyecto se centra en el análisis del cine de género drama y terror, con un enfoque particular en el cine de terror en lengua española. Surge de la inquietud por comprender las razones detrás de la escasa producción de películas de terror en este idioma. Esta situación podría estar influenciada por una serie de factores, como la historia de la producción cinematográfica, la demanda del mercado, la existencia de estereotipos, tabúes y estigmas culturales, así como la falta de promoción y el interés del público. Además, la influencia del idioma en la escritura de guiones podría desempeñar un papel crucial en la forma en que se desarrollan las historias y los personajes.

El lenguaje no solo afecta la construcción de diálogos, sino que también impacta en el uso de metáforas y en la creación de atmósferas que son esenciales para el género del terror. Por lo tanto, planteamos la hipótesis de que la lengua en la que se redacta el guion de una película podría influir significativamente en su clasificación dentro del género cinematográfico. ¿Hasta qué punto el idioma en el que se elabora el guion afecta la percepción y la recepción del género de terror en el cine hispanohablante? Esta pregunta nos guiará en el desarrollo de nuestro análisis.

Para ello, hemos utilizado una muestra de datos obtenida de Kaggle
https://www.kaggle.com/datasets/akashguna/netflix-prize-shows-information?resource=download

HIPÓTESIS

La hipótesis se puede resumir de la siguiente manera:

¿La lengua en la que se redacta el guion de una película influye significativamente en su clasificación dentro del género cinematográfico, afectando la percepción y la recepción del género de terror en el cine hispanohablante?

Esta hipótesis se basa en la idea de que factores como la historia de la producción cinematográfica, estereotipos culturales, tabúes, la promoción y el interés del público, así como el impacto del idioma en la narrativa, pueden contribuir a la escasa producción de películas de terror en lengua española.

1. IMPORTACION DE PAQUETES

En esta sección se importarán los paquetes esenciales para realizar el Análisis Exploratorio de Datos (EDA), incluyendo **matplotlib**, **numpy**, **pandas**, **os**, **seaborn**, **scipy. stats, chi2_contingency**. También se configurará para **ignorar advertencias**.

2. CARGA DE DATOS

En esta sección, cargaremos y examinaremos el conjunto de datos para comprender su estructura, características y calidad. Esta revisión inicial nos ayudará a identificar anomalías, valores faltantes y la distribución de los datos, lo que guiará los pasos siguientes en nuestro análisis. Se utiliza el comando **pd.read_csv** para cargar el archivo **imdb.csv**.

3. LIMPIEZA DE DATOS

La limpieza de datos es un proceso crucial para garantizar la calidad de la información en el análisis. En esta fase, se abordarán problemas como valores nulos y errores de formato. Se ha identificado que ciertas columnas (genre, country, language, cast, director, composer y writer) contienen datos con corchetes y comillas simples. Para solucionarlo, se creará una función que limpiará estas columnas y generará un nuevo DataFrame.

Las modificaciones incluirán:

- **Eliminación de caracteres no deseados**: Utilizando el método str.replace() para quitar comillas y corchetes, logrando un formato uniforme.
- **Eliminación de espacios en blanco**: Aplicando str.strip() para eliminar espacios al inicio o final de los valores.
- **Uso de expresiones regulares**: Implementando regex=True en las operaciones para identificar y eliminar caracteres no deseados de manera más eficiente.

Finalmente, se guardará el DataFrame limpio en un nuevo archivo CSV.

4. ANÁLISIS - TRATAMIENTO DE PELICULAS POR GÉNERO

4.1. Análisis de las películas del género horror

El proceso descrito consiste en cargar un conjunto de datos de películas desde un archivo CSV y filtrar la información relacionada con películas de horror. A continuación, se detallan los pasos:

4.1.1 **Carga del dataset**: Se carga el archivo CSV ubicado en "./data/df_imdb_cleaned.csv" en un DataFrame llamado df_imdb_2.
4.1.2 **Filtrado por tipo**: Se seleccionan las filas donde la columna 'kind' es 'movie' o 'tv movie', creando un nuevo DataFrame llamado movies.
4.1.3 **Filtrado por género**: Se filtran las filas del DataFrame movies para obtener solo aquellas donde el género es 'Horror' (ignorando mayúsculas), resultando en movie_horror.
4.1.4 **Selección de columnas**: Se seleccionan columnas específicas ('titleId', 'title', 'kind', 'genre', 'country', 'language') del DataFrame movie_horror, y se resetear el índice.
4.1.5 **Análisis de valores nulos**: Se evalúan los nulos en el DataFrame horror, mostrando tanto la cantidad como el porcentaje de valores nulos.
4.1.6 **Reemplazo de nulos**: Se reemplazan los valores nulos en la columna 'language' por el correspondiente valor de la columna 'country', y se verifica nuevamente la cantidad de nulos en el DataFrame.

4.2. Análisis de películas del género drama

Se utiliza el método '.loc()' para acceder al grupo de filas 'Drama' en la columna 'genre', creando un DataFrame llamado 'movie_drama'. Luego, se seleccionan las columnas relevantes y se crea un nuevo DataFrame 'drama'. Se verifica la cantidad y porcentaje de valores nulos en 'drama'. Finalmente, con el método '.fillna()' se reemplazan los valores nulos de la columna 'language' con los correspondientes de la columna 'country' en la misma fila, y se comprueba nuevamente la cantidad de valores nulos.

5. ANALISIS DESCRIPTIVO DE LAS PELÍCULAS DEL GENERO HORROR Y DRAMA

5.1. Análisis descriptivo de películas del género horror

El análisis descriptivo de películas del género horror incluye varios aspectos clave:

5.1.1 **Cantidad de películas por idioma**: Se cuenta el número de películas de horror en diferentes idiomas utilizando la función value_counts().
5.1.2 **Porcentaje de películas por idioma**: Se calcula la frecuencia relativa de cada idioma mediante value_counts(normalize=True) y se presenta en porcentaje, redondeado a dos decimales.
5.1.3 **Análisis por país**: Se genera un DataFrame que muestra la cantidad de películas de horror por país, contando las películas y renombrando las columnas para mayor claridad.

5.2. Análisis descriptivo de películas del género drama

Se realizó un análisis de las películas del género drama en relación a su idioma y país de origen:

5.2.1 Se contabilizó la cantidad de películas de drama por idioma utilizando el método value_counts().

5.2.2 Se visualizó la cantidad de películas por idioma en términos porcentuales, empleando value_counts(normalize=True) para obtener la frecuencia relativa multiplicada por 100 y redondeada a dos decimales.

5.2.3 Se creó un DataFrame que muestra el número de películas de drama clasificadas por país. Se utilizó value_counts() y reset_index() para transformar los datos en un formato más accesible, renombrando las columnas para mayor claridad.

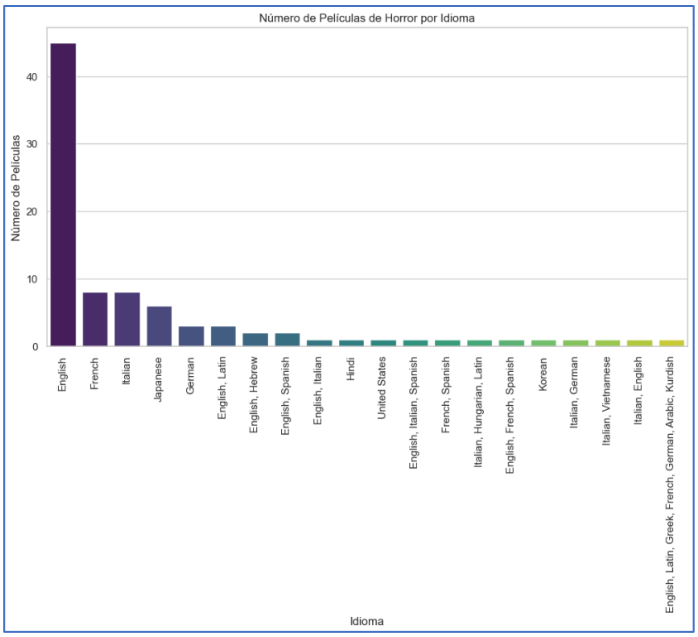
6. VISUALIZACION DE DATOS PARA PELÍCULAS DEL GÉNERO HORROR Y DRAMA

6.1. Visualización de datos para películas del género horror

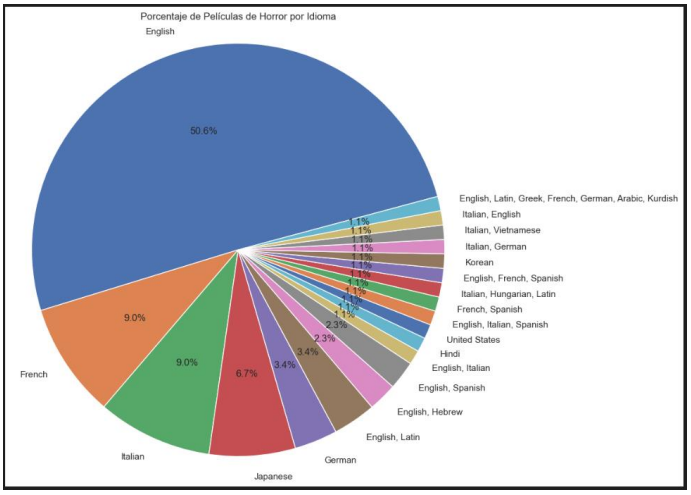
Se presentan dos secciones de visualización de datos sobre películas del género horror, clasificadas por idioma y país.

6.1.1 **Visualización por idioma:**

- **Gráfico de Barras:** Se utiliza un gráfico de barras para mostrar el número de películas de horror por idioma. Se configura el estilo con seaborn y se presenta con etiquetas adecuadas y rotación de las mismas para mejor visualización.

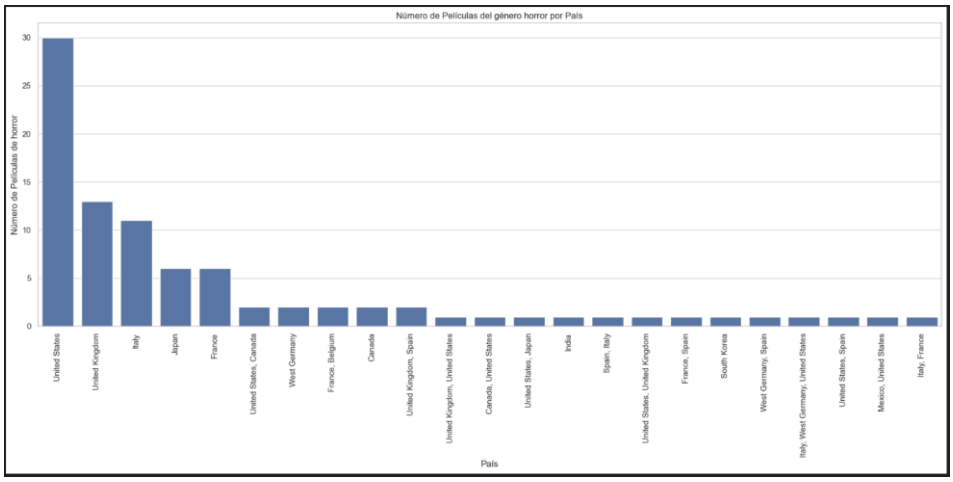


- **Gráfico de Pastel:** Se crea un gráfico de pastel para representar el porcentaje de películas de horror por idioma. Este gráfico también se configura con seaborn, mostrando los porcentajes en cada sección y manteniendo una forma circular.



6.1.2 **Visualización por país:**

- **Gráfico de Barras:** Se elabora un gráfico de barras para ilustrar el número de películas de horror por país. El gráfico se personaliza con títulos y etiquetas en los ejes, además de rotar los nombres de los países para facilitar su lectura. Se ajusta el layout para evitar superposiciones.



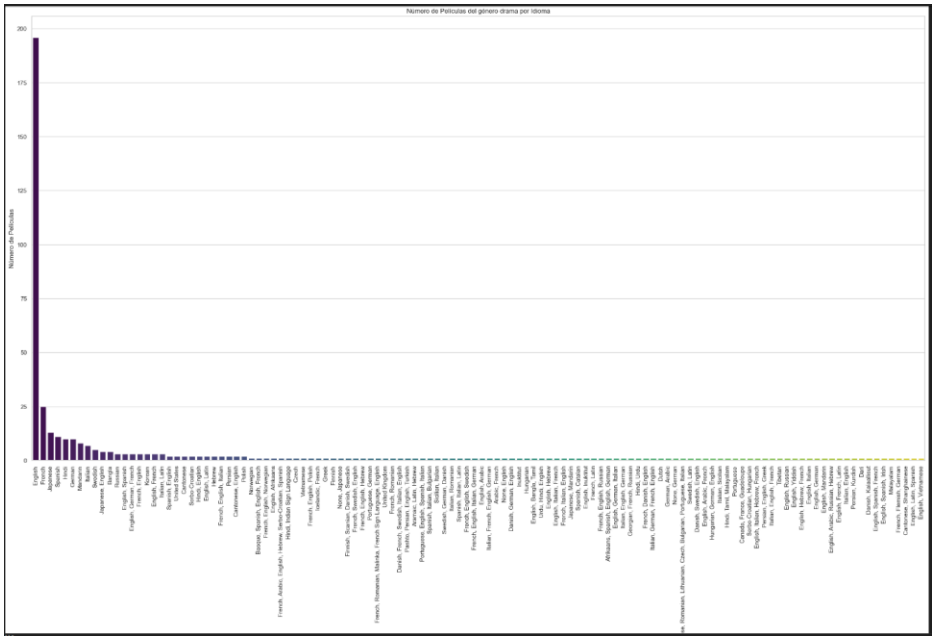
Ambas visualizaciones permiten analizar la distribución de películas de horror a través de diferentes idiomas y países.

6.2. Visualización de datos para películas del género drama

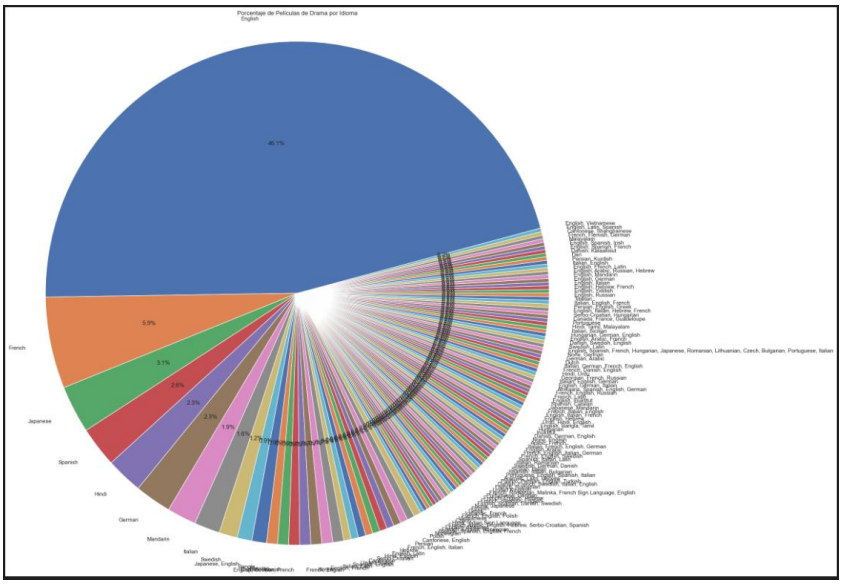
La sección 6.2 se centra en la visualización de películas del género drama, clasificadas por idioma y país, utilizando gráficos de barras y gráficos de pastel.

6.2.1 **Visualización por Idioma**

- **Gráfico de Barras**: Se utiliza Seaborn para crear un gráfico que muestra la cantidad de películas de drama por idioma. Se establecen etiquetas y se ajusta la rotación de los ejes para una mejor visualización.

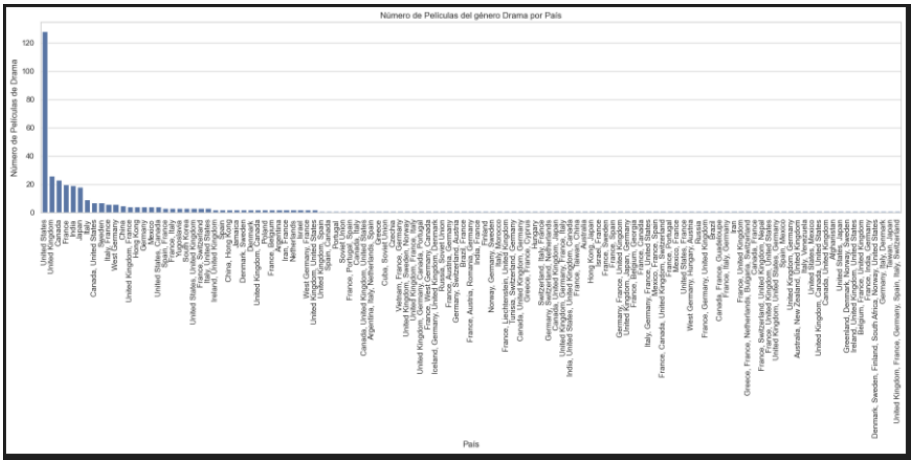


- **Gráfico de Pastel**: Se presenta un gráfico de pastel que ilustra el porcentaje de películas de drama por idioma, con porcentajes visibles y un ángulo de inicio ajustado para mejorar la visualización.

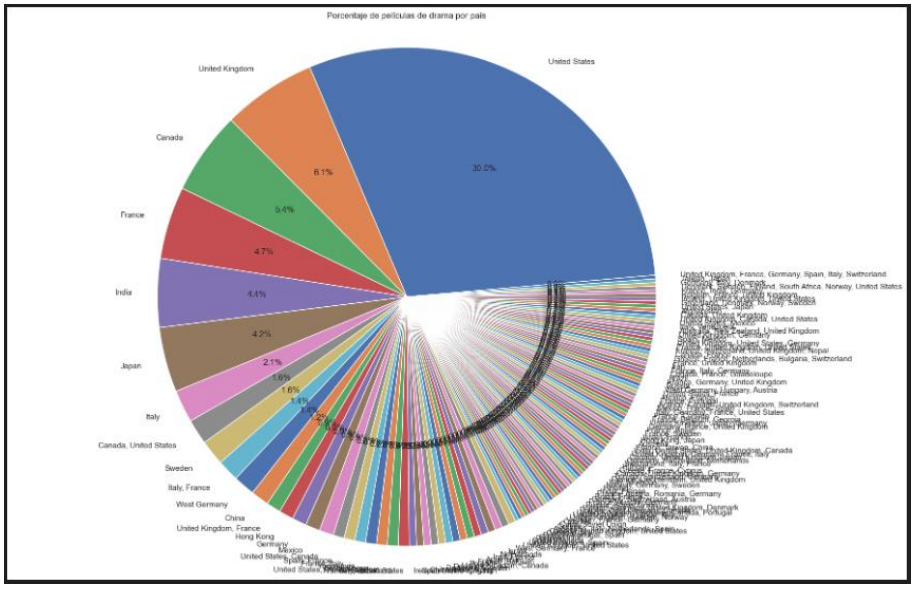


6.2.2 **Visualización por País**

- **Gráfico de Barras**: Se elabora un gráfico de barras que muestra el número de películas de drama por país, con personalización en las etiquetas y rotación de los nombres de los países para facilitar la lectura.



- **Gráfico de Pastel**: Se calcula el porcentaje de películas de drama por país y se representa en un gráfico de pastel, asegurando que el gráfico tenga forma circular y sea fácil de interpretar.



Ambas secciones utilizan visualizaciones efectivas para representar información relevante sobre las películas de drama, tanto por idioma como por país.

7. PRIMERAS CONCLUSIONES Y REFLEXIONES

Tras analizar los resultados de las visualizaciones, los análisis, y reflexiones sobre la hipótesis inicial. Resumimos lo siguiente en el análisis de idiomas en el género horror y drama:

1. **Tendencia del idioma en el género horror:**
    - El análisis revela que el inglés es el idioma más utilizado en películas de horror, con un 50.6%. Este idioma parece contener vocabulario que evoca miedo y tensión.
    - Otros idiomas prevalentes en este género son el francés y el italiano (9% cada uno), seguidos por el japonés (6.7%).
    - Pensamos que el estilo narrativo y las diferencias culturales, reflejadas en el idioma, pueden influir en cómo se percibe el miedo en las historias de horror.

- Se destaca una escasa producción de películas de horror en español.

2. **Idiomas en el género drama:**
    - En el análisis del género dramático, el inglés también es el idioma predominante, con un 46.1%. Le siguen el francés (5.9%), el japonés (3.1%) y el español (2.6%).
    - Se observa que los países de habla inglesa dominan la producción cinematográfica, con Estados Unidos liderando (30%), seguido de Inglaterra (6.1%) y Canadá (5.4%).
    - La producción de dramas en español es bastante limitada, con pocos países alcanzando cifras significativas.
3. **Comparativa entre géneros:**
    - Tanto en el género horror como en el drama, el inglés es el idioma dominante, aunque el horror tiene una mayor concentración de producciones en este idioma.
    - La producción de cine en otros idiomas, incluidos el español, es baja en ambos géneros, destacando una tendencia similar en la predominancia del inglés.

8. ANÁLISIS EXTRA: ANÁLISIS ESTADÍSTICOS MÁS PROFUNDOS

El análisis EXTRA se centra en realizar un estudio estadístico profundo sobre películas de los géneros horror y drama, con el fin de validar nuestra hipótesis. Se llevará a cabo un análisis descriptivo detallado y se aplicarán pruebas estadísticas, como la Chi-Cuadrado, para examinar relaciones y patrones en los datos. Este análisis incluirá variables adicionales como idioma, género, país, calificación y votos.

8.1. Limpieza de datos

En la fase de limpieza de datos, se llevará a cabo un proceso exhaustivo para asegurar la integridad y calidad de la información en las columnas country, year, rating y vote. Esto implica eliminar entradas erróneas, duplicados y valores atípicos para obtener un conjunto de datos preciso y confiable, esencial para el análisis posterior.

El proceso incluye cargar el dataset, filtrar las filas donde 'kind' es 'movie' y 'genre' es 'horror' o 'drama', y seleccionar las columnas necesarias, que incluyen country, year, rating, y vote para futuras operaciones.

8.2. Análisis descriptivo con las columnas rating y vote

Se llevará a cabo un análisis descriptivo de las columnas "rating" y "vote" para evaluar su influencia en la percepción de los géneros de las películas, según la hipótesis inicial. Utilizando el método describe(), se obtendrán estadísticas como la media, mediana, mínimo, máximo y cuartiles, que ayudarán a entender la distribución de estos factores. Se mostrará una descripción del "rating" y "vote" a través de un dataframe específico.

8.3. Comparaciones entre las cantidades de películas de Horror y Drama por idioma y país

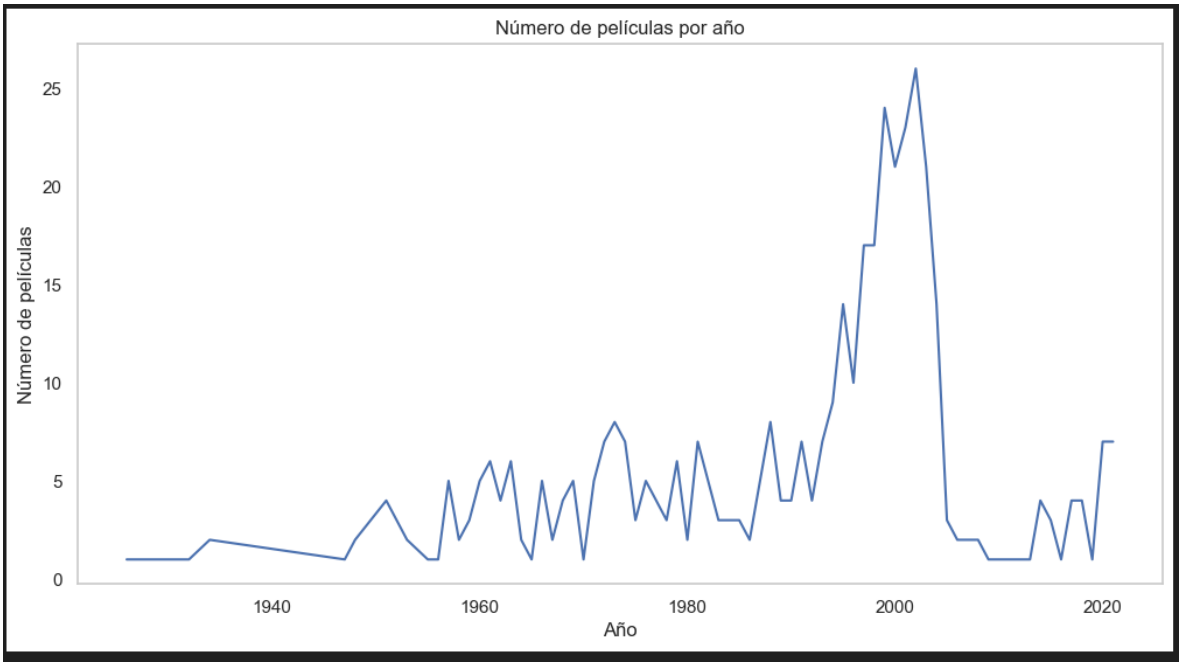
Se presenta un análisis comparativo de la cantidad de películas de los géneros Horror y Drama según idioma y país. Para ello, se cuentan las películas producidas anualmente y se grafican las tendencias a lo largo del tiempo.

Se utiliza la función groupby() para agrupar los datos por idioma y género, y al agrupar por ['country', 'language', 'genre'], se logra un conteo más detallado de las películas en cada combinación de país e idioma. La función unstack(fill_value=0) ayuda a rellenar las filas sin datos con ceros, facilitando así la visualización.

El resultado se presenta en una tabla que permite comparar las cantidades de películas en los dos géneros seleccionados.

8.4. Análisis de tendencias en el tiempo

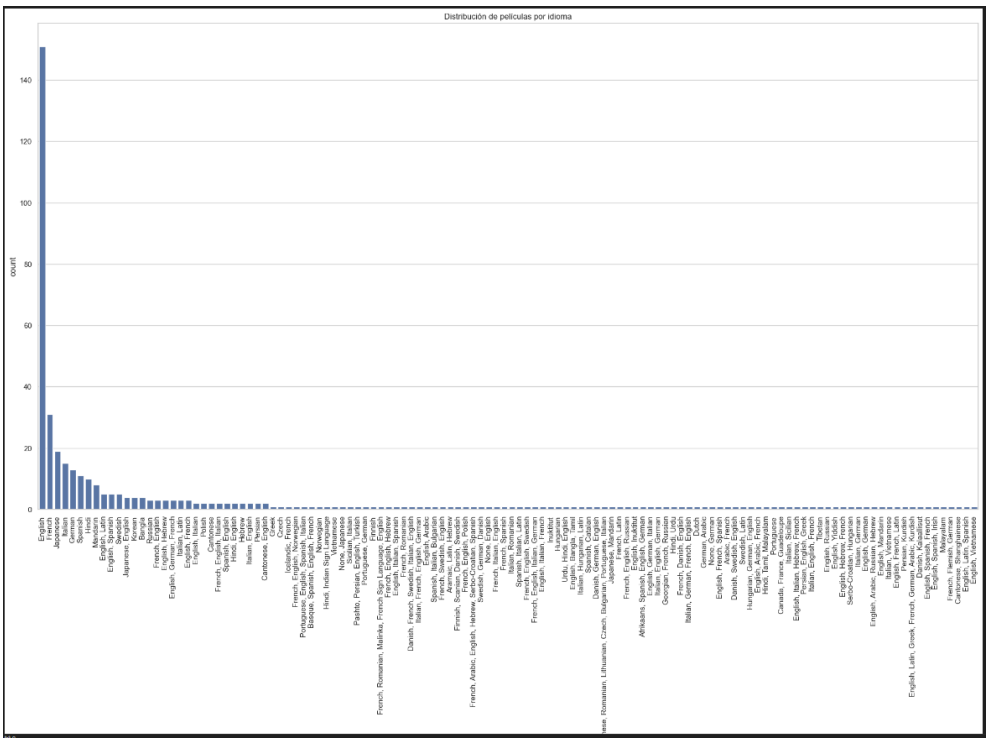
Se analizará la evolución de la producción de películas a lo largo de los años utilizando un conjunto de datos con información temporal. Se contará el número de películas producidas cada año y se graficarán estos datos para identificar tendencias. Se utilizará un gráfico de líneas para mostrar el número de películas por año, incluyendo títulos y etiquetas en los ejes.



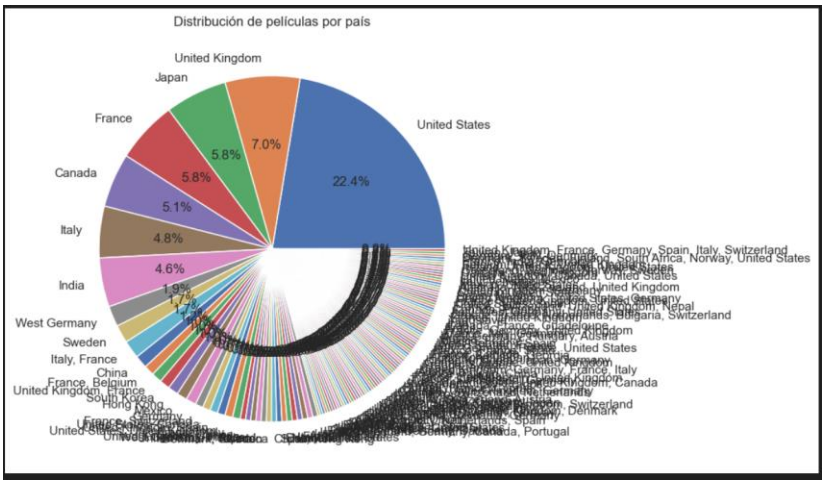
8.5. Visualización de la distribución de películas por idioma y país

En la sección 8.5 se aborda la visualización de la distribución de películas según idioma y país utilizando gráficos. Se emplea un gráfico de barras, generado con **countplot(),** para mostrar la cantidad de películas por idioma, y un gráfico de pastel para ilustrar la proporción de películas por país.

El gráfico de barras se configura con un tamaño de 25x13 y las etiquetas del eje x se rotan 90 grados para mayor claridad.



Por otro lado, el gráfico de pastel se presenta en un tamaño de 7x7, mostrando las proporciones de cada país con porcentajes.



8.6. Pruebas estadísticas para validar la hipótesis

8.6.1 Prueba Chi-Cuadrado

Se realizó un análisis estadístico utilizando la prueba Chi-Cuadrado para evaluar la influencia del idioma en la clasificación de películas de terror en el contexto hispanohablante, considerando variables como idioma, género, país, rating y votos.

1. **Filtrado de Datos**: Se filtraron las películas del género "Horror" en el DataFrame.
2. **Análisis Chi-Cuadrado**:
    - **Idioma y País**: Se construyó una tabla de contingencia y se calculó el Chi-Cuadrado, obteniendo un valor de **896.46** con un p-valor de **4.19e-44**, lo que indica una relación significativa entre el idioma y el país.
    - **Idioma y Rating**: Se categorizó el rating en grupos, resultando en un Chi-Cuadrado de **47.20** y un p-valor de **0.10**, sugiriendo que no hay una relación significativa.
    - **Idioma y Votos**: Se categorizó el número de votos, obteniendo un Chi-Cuadrado de **38.90** y un p-valor de **0.82**, lo que también indica una falta de relación significativa.

Resultados finales Chi-Cuadrado:

- **Idioma y País**: Relación significativa.
- **Idioma y Rating**: No significativo.
- **Idioma y Votos**: No significativo.

9. CONCLUSIONES GENERALES

Hemos compilado estos hallazgos en reportes que incluyen gráficos y análisis realizados, así como mis conclusiones sobre la influencia del idioma en el género cinematográfico de horror y drama.

Con los resultados, hemos realizado un análisis más profundo, como pruebas estadísticas para validar la hipótesis.

En las pruebas estadísticas se abordó la hipótesis planteada sobre la influencia del idioma en la clasificación dentro del género cinematográfico, especialmente en el cine de terror en el contexto hispanohablante, en este análisis estadístico se incluyeron las variables relevantes: idioma, género, país, rating y votos.

Chi-cuadrado:

- Idioma y País: Existe una fuerte asociación.

- Idioma y Rating: No hay suficiente evidencia para afirmar que existe una asociación significativa.

- Idioma y Votos: No hay suficiente evidencia para afirmar que existe una asociación significativa.

En conclusión, la hipótesis inicial se ve respaldada en cuanto a la influencia del idioma en la producción de películas de horror, especialmente en relación con el país de origen. Sin embargo, no se encontraron evidencias suficientes para afirmar que el idioma afecta significativamente la recepción en términos de ratings o votos en el género de terror en el cine hispanohablante. Esto sugiere que, aunque el idioma puede ser un factor determinante en la producción y la percepción cultural del horror, hay otros elementos que también juegan un papel importante en la recepción del género.