Análisis Estadístico de los contenidos de Netflix

Jose Agustín del Toro González C-312
Victor Hugo Pacheco Fonseca C-311
John García Muñoz C-311

February 16, 2025

Contents

1 Introducción

1.1 Contexto del Problema

Breve descripción del contexto y los objetivos del análisis. ¿Por qué es importante este dataset? ¿Qué preguntas se buscan responder?

1.2 Descripción del Dataset

Información general sobre el dataset:

- Fuente de los datos.
- Número de filas y columnas.
- Variables principales y su tipo (numérica, categórica, etc.).

2 Análisis Exploratorio de Datos (EDA)

2.1 Limpieza de Datos

- Manejo de valores faltantes.
- Tratamiento de outliers.
- Corrección de inconsistencias.

2.2 Análisis Descriptivo

2.2.1 Estadísticas descriptivas (media, mediana, desviación estándar, etc.).

Estadística	**Valor**
Duración Media	99.528
Duración Mediana	98.000
Moda de la Duración	90.000
Varianza (min^2)	804.816
Desviación Estándar	28.369
Percentil 25	87.000
Percentil 50 (mediana)	98.000
Percentil 75	114.000

Table 1: Duración de Películas

Estadística	Valor
Duración Media	1.765
Duración Mediana	1.000
Moda de la Duración	1.000
Varianza (#temporadas^2)	2.505
Desviación Estándar	1.583
Percentil 25	1.000
Percentil 50 (mediana)	1.000
Percentil 75	2.000

Table 2: Duración de Series de TV

Estadística	Valor
Mediana	2017
Moda	2018
Varianza	78
Desviación Estándar	9
Percentil 25	2013
Percentil 50 (mediana)	2017
Percentil 75	2019

Table 3: Año de Estreno

Estadística	Valor
Mediana	2019
Moda	2019
Varianza	2
Desviación Estándar	2
Percentil 25	2018
Percentil 50 (mediana)	2019
Percentil 75	2020

Table 4: Año de Adición a Netflix

2.2.2 Visualizaciones

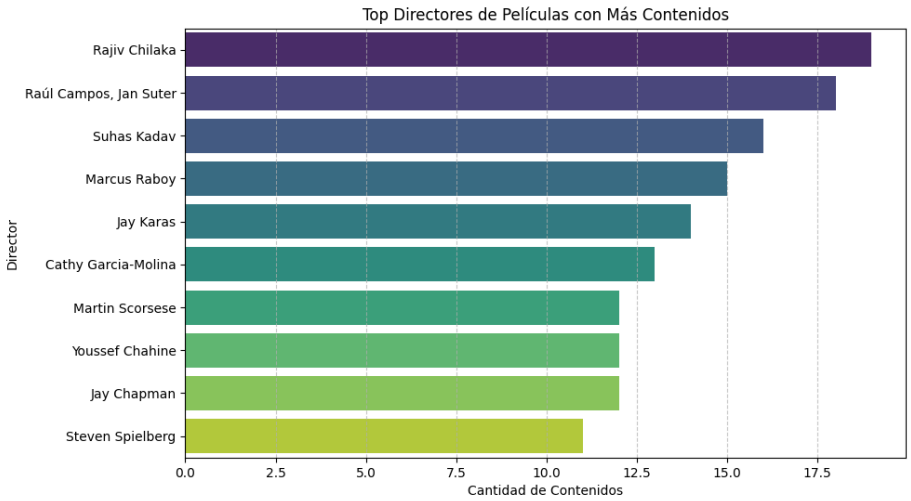

Figure 1: Directores de Películas Más Comunes

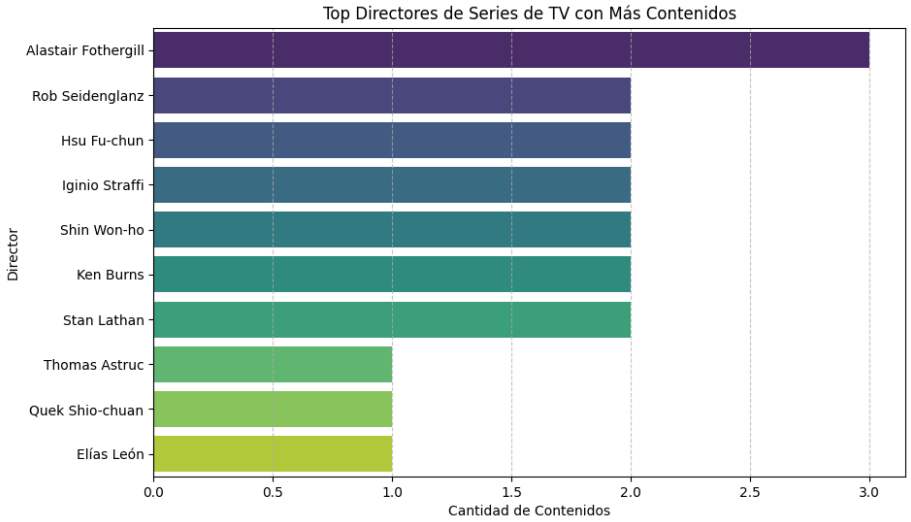

Figure 2: Directores de Series Más Comunes

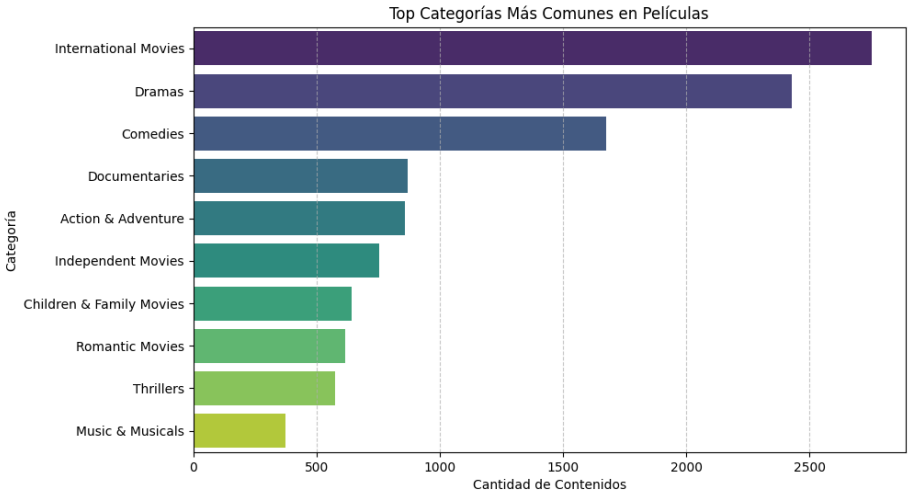

Figure 3: Categorías Más Comunes en Películas

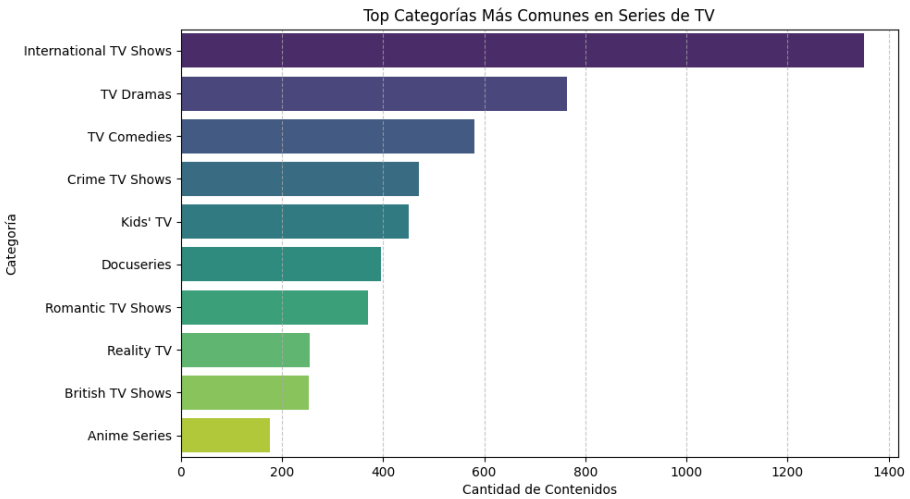

Figure 4: Categorías Más Comunes en Series de TV

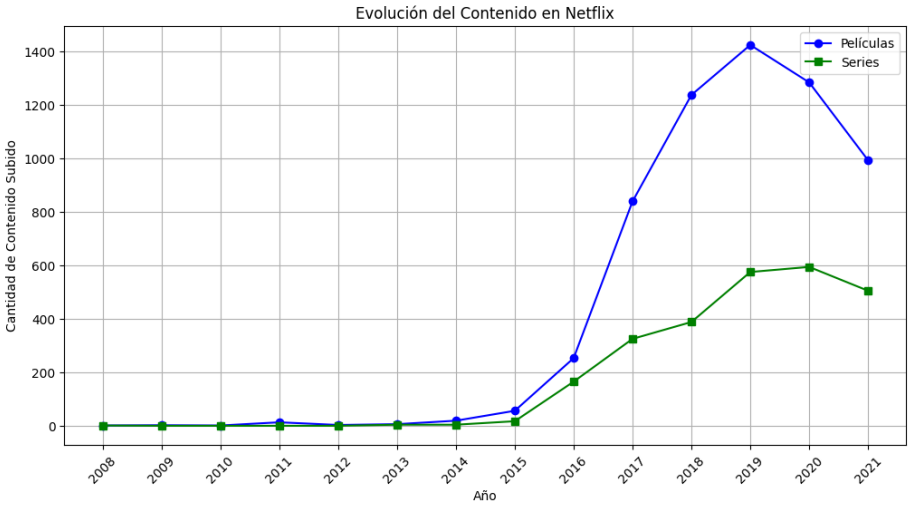

Figure 5: Evolución del Contenido

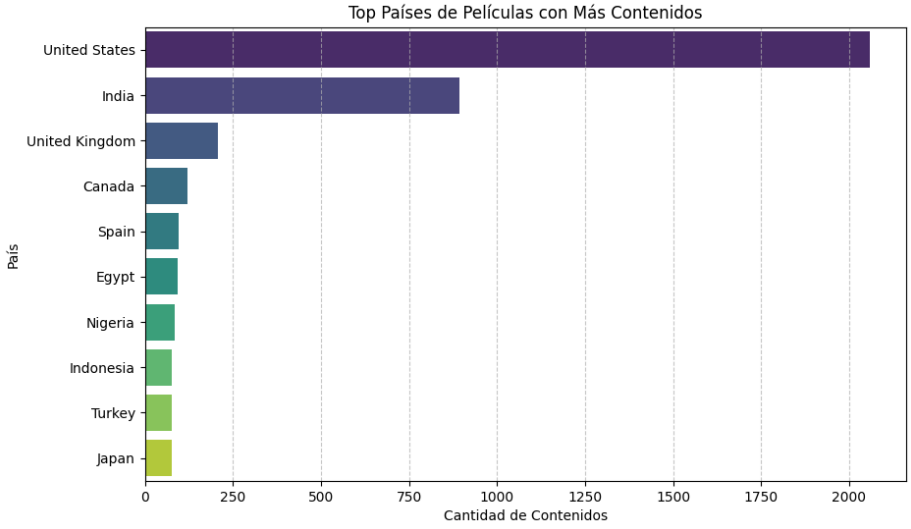

Figure 6: Países con más películas

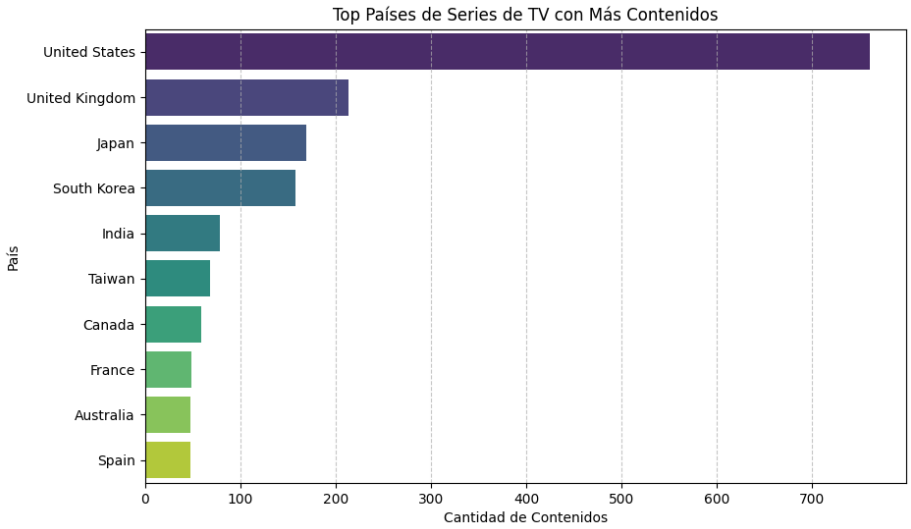

Figure 7: Países con más series

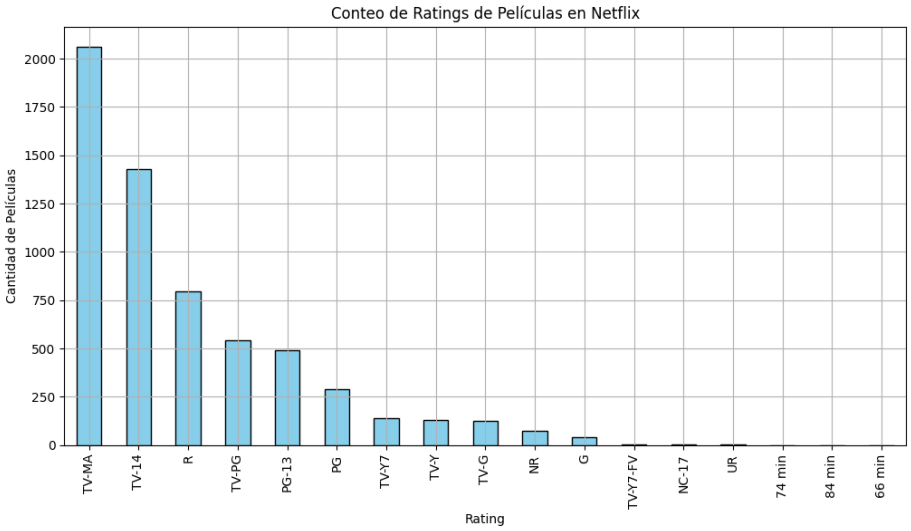

Figure 8: Conteo de Ratings para Películas

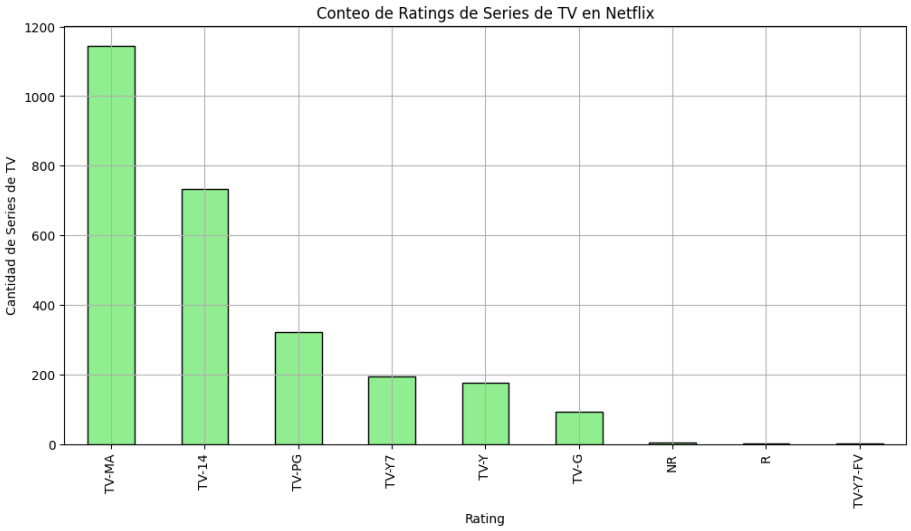

Figure 9: Conteo de Ratings para Series de TV

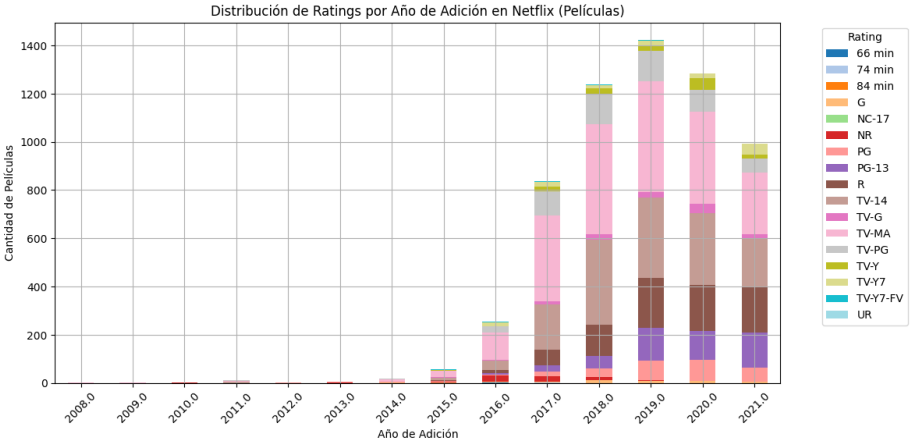

Figure 10: Distribución de Ratings por Año de Adición en Películas

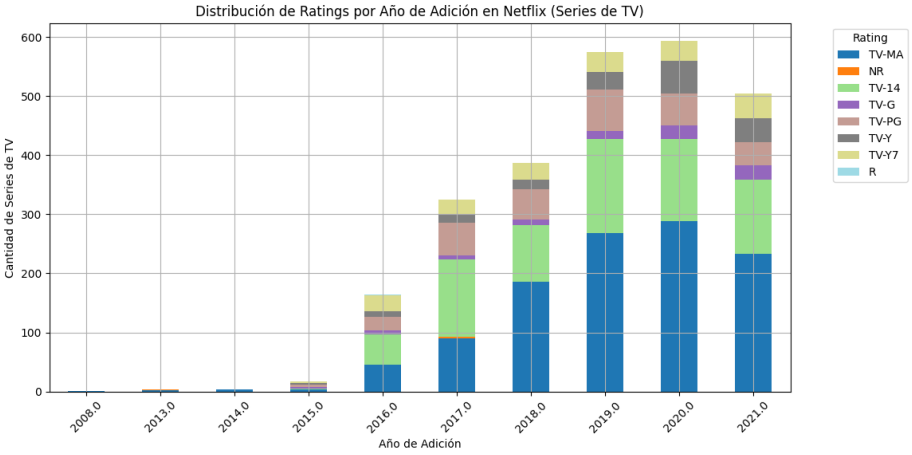

Figure 11: Distribución de Ratings por Año de Adición en Series de TV

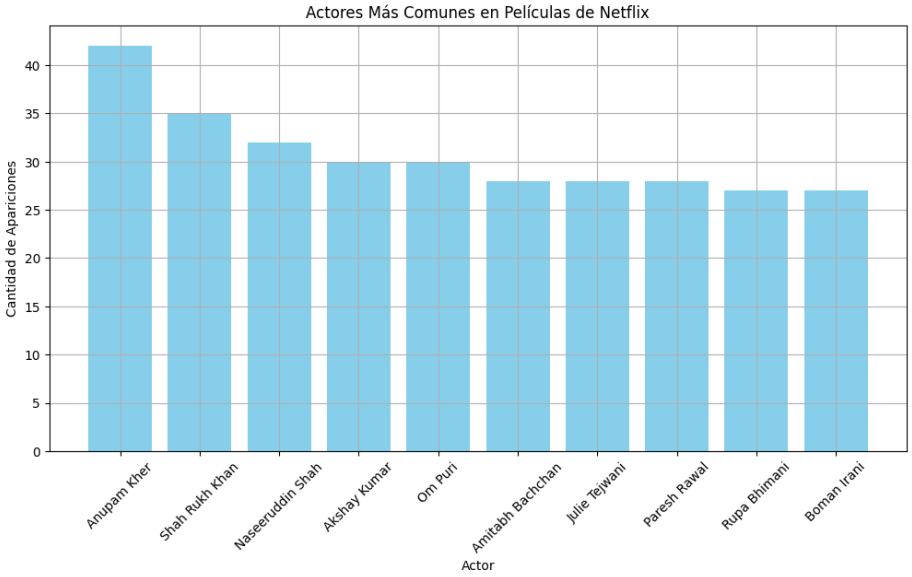

Figure 12: Actores Más Comunes en Películas

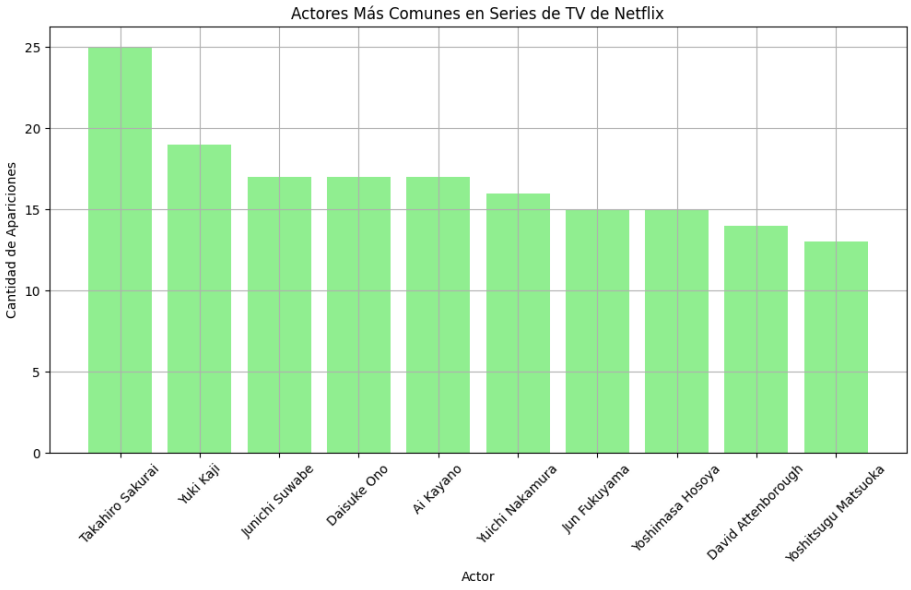

Figure 13: Actores Más Comunes en Series de TV

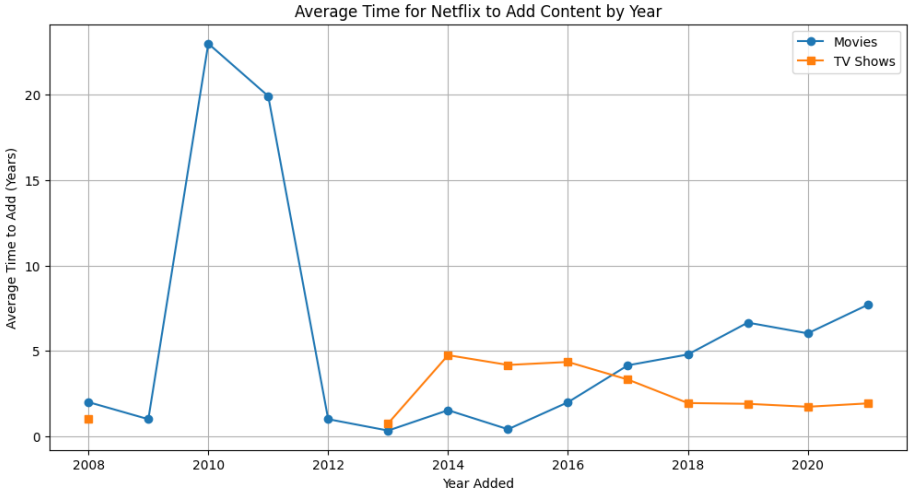

Figure 14: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

2.2.3 Análisis de variables categóricas (tablas de frecuencia, gráficos de barras).

2.3 Relaciones Iniciales

- Gráficos de dispersión (scatterplots) para identificar patrones.
- Observaciones preliminares sobre tendencias o correlaciones.

3 Análisis de Componentes Principales (PCA)

3.1 Proceso

Se realizó un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos, que incluye información sobre películas y series de TV. Utilizando seis variables (*tipo, director, duración, fecha de estreno, fecha de adición, rating*), se han codificado las variables categóricas y estandarizado las variables numéricas. Los componentes principales obtenidos (*PC1, PC2, PC3*) capturan la mayor parte de la variabilidad en los datos. Finalmente, se han graficado la varianza explicada por cada componente para visualizar la cantidad de información capturada por cada uno, observando que a partir del segundo componente principal, ya esta se vuelve insignificante.

3.2 Gráficos

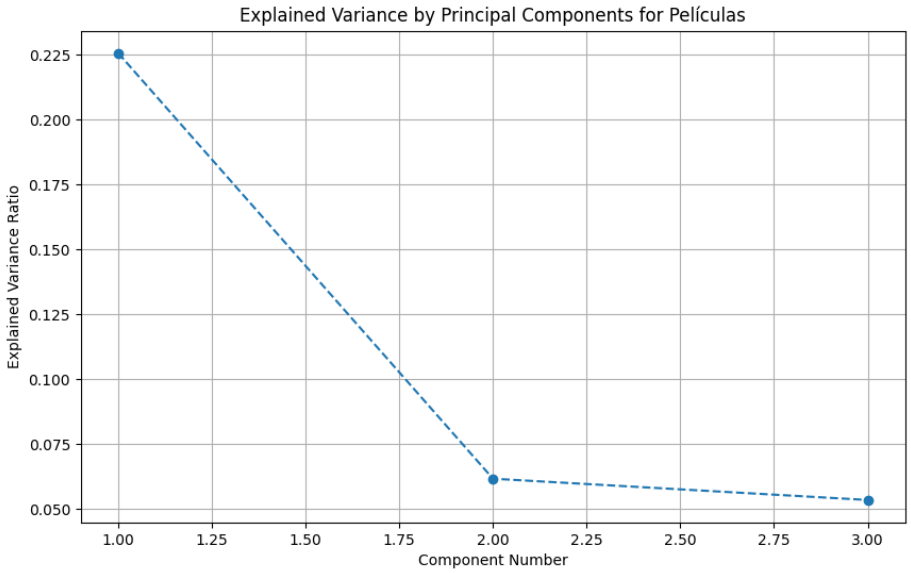

Figure 15: Varianza explicada por componentes principales (Películas)

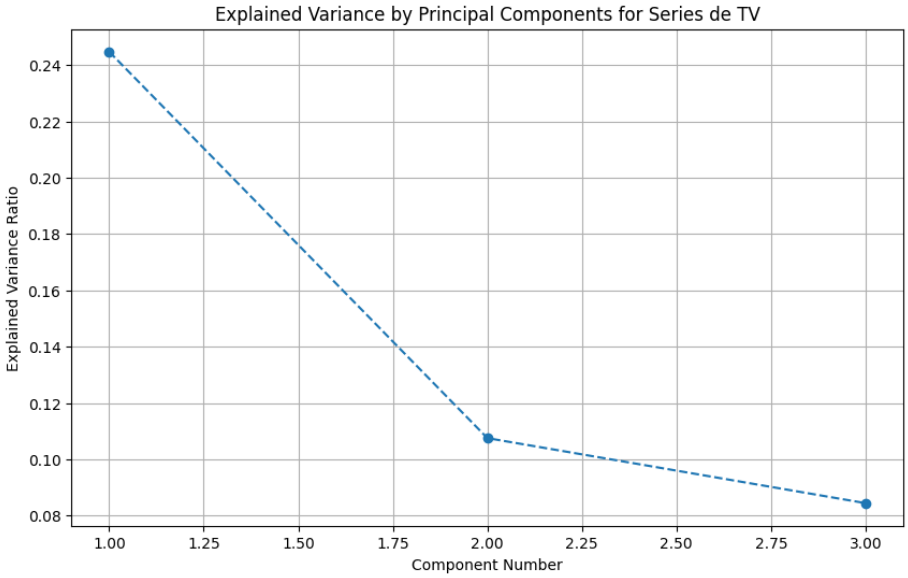

Figure 16: Varianza explicada por componentes principales (Series)

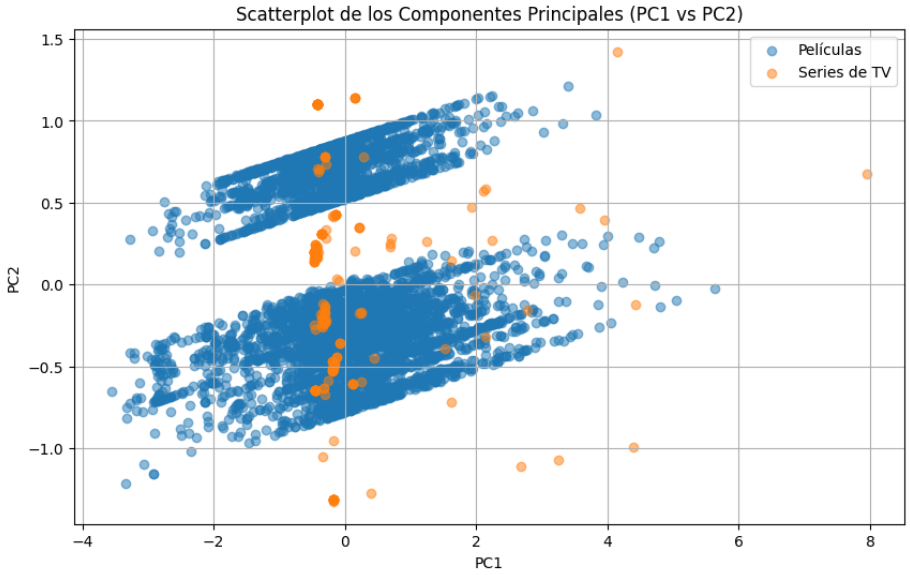

Figure 17: Gráfico de puntos para los componentes principales

Dado que el tercer componente principal no captura una porción relativamente influyente de varianza explicada, se ha propuesto hacer el gráfico considerando solo los dos primeros PC. En el caso de las series se da la peculiaridad de que aparece un número anormalmente escaso en dicho gráfico, esto es debido a que no se consideran aquellas que poseen un valor nulo en alguno de los campos correspondientes a las variables que incluimos en el PCA. Esta es una deficiencia del conjunto de datos estudiado.

4 Test de Normalidad

En este análisis, se realizó un test de Kolmogorov-Smirnov para evaluar si la muestra de datos sigue una distribución normal. Tras llevar a cabo el test, se concluyó que los datos no siguen una distribución normal, lo cual nos indica que los datos presentan una desviación significativa de la distribución normal teórica.

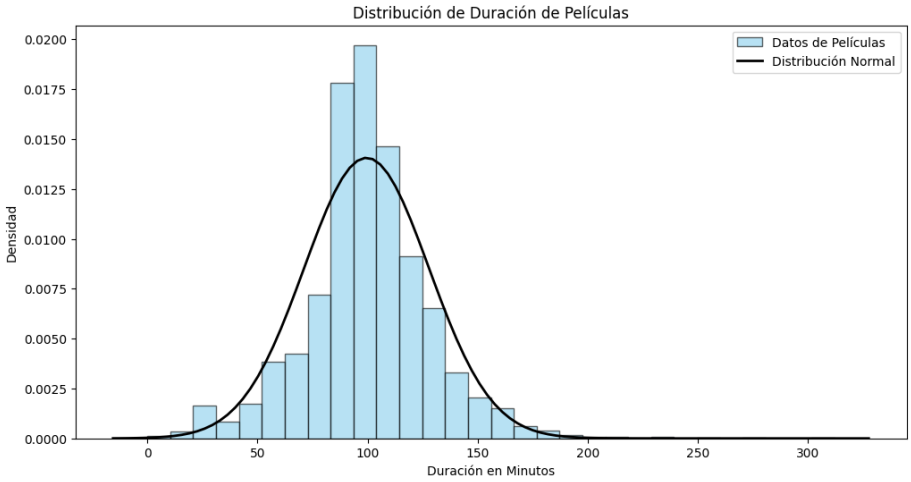

Figure 18: Test de normalidad (Peliculas)

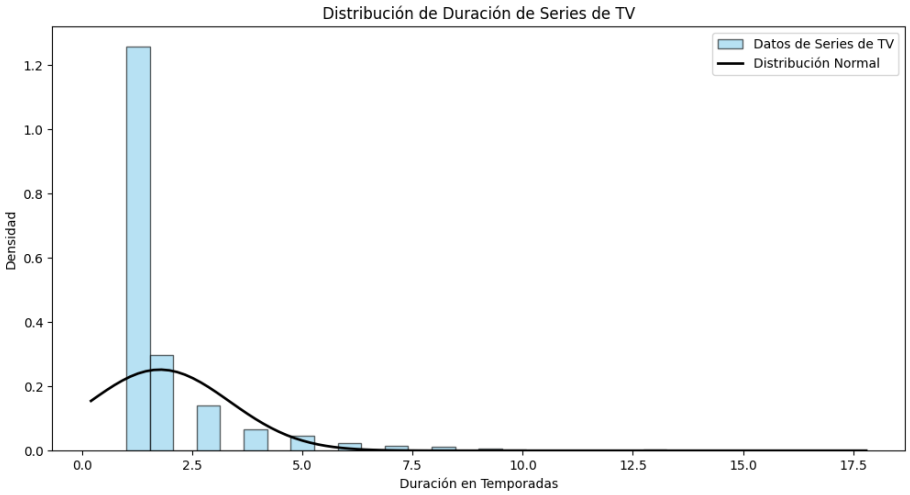

Figure 19: Test de normalidad (Series)

5 Formulación de Hipótesis

5.1 Duración en películas

Objetivo: Comprobar si al menos el 70% de las películas tienen una duración mayor a 80 minutos.

- Hipótesis nula : La proporción es menor o igual a 0.70 .
- Hipótesis alternativa:La proporción es mayor a 0.70.

Dado que la media muestral de las peliculas es mayor a 80 minutos , es lógico pensar que gran parte de las peliculas de Netlfix tienen una duración superior a 80 minutos.

Luego de realizada la prueba de hipotesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipotesis nula, por lo que podemos asegurar en un 95% que más del 70% de las películas de Netflix tienen una duración superior a 80 minutos.

5.2 Rating en películas y series

- Hipótesis nula : El rating más común en el contenido de EE.UU. no es TV-MA.
- Hipótesis alternativa: El rating más común de EE.UU. es TV-MA.

En el análisis exploratorio de datos se pudo observar que en la muestra la categoría mas común tanto en peliculas como en series es TV-MA.Siendo Estados Unidos es el pais con mayor contenido en peliculas y series,se desea comprobar si el rating mas comun en su contenido es TV-MA.

Luego de realizada la prueba de hipotesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipotesis nula, por lo que podemos asegurar en un 95% que el rating mas comun en EEUU es TV-MA

5.3 Prueba de los castings en EEUU e India

Si observamos los gráficos , notamos que los actores mas comunes de las películas son de nacionalidad India, aún cuando el pais con mas peliculas es Estados Unidos (seguido de India).Esto sugiere diferencias en las industrias cinematográficas de ambos países, posiblemente en términos de diversidad y frecuencia de aparición de los actores en las producciones.

- Hipótesis nula : No hay diferencia significativa en la frecuencia promedio de aparición de actores entre las películas de India y las de Estados Unidos.
- Hipótesis alternativa:Los actores en las películas de India tienen una frecuencia promedio de aparición mayor que los actores en las películas de Estados Unidos.

Luego de realizada la prueba de hipotesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipotesis nula, por lo que podemos asegurar en un 95% que los actores en películas de India aparecen en más películas en promedio que los actores en películas de EE.UU.

nota: Algo similar sucede con las series ,aun cuando Estados Unidos es el pais con mas series, los actores mas comunes son de nacionalidad japonesa.

5.4 Categorias Más Común en Películas

Queremos determinar si "International Movies" es la categoría más frecuente en las películas del catálogo de Netflix, y si su frecuencia es significativamente mayor que la esperada si todas las categorías fueran igualmente comunes.

- Hipótesis nula : La categoría "International Movies" no es la categoría más común entre las películas; su frecuencia no difiere significativamente de las demás categorías.
- Hipótesis alternativa:La categoría "International Movies" es la categoría más común entre las películas; su frecuencia es significativamente mayor que la de otras categorías.

Luego de realizada la prueba de hipotesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipotesis nula, por lo que podemos asegurar en un 95% que La categoría "International Movies" es la categoría más común entre las películas.

6 Análisis de Correlación

6.1 Matriz de Correlación

Para la matriz de correlación , se utilizaron las siguientes variables (luego de las transformaciones necesarias):

- duration :Convertiremos la columna duration a variables numéricas separadas para películas y series.
- date_added :Convertiremos las fechas a un formato numérico (días desde una fecha de referencia).
- rating: Asignaremos valores numéricos a las clasificaciones por edad.
- is_movie: Indicador binario (1 si es película, 0 si es serie).

6.1.1 Películas

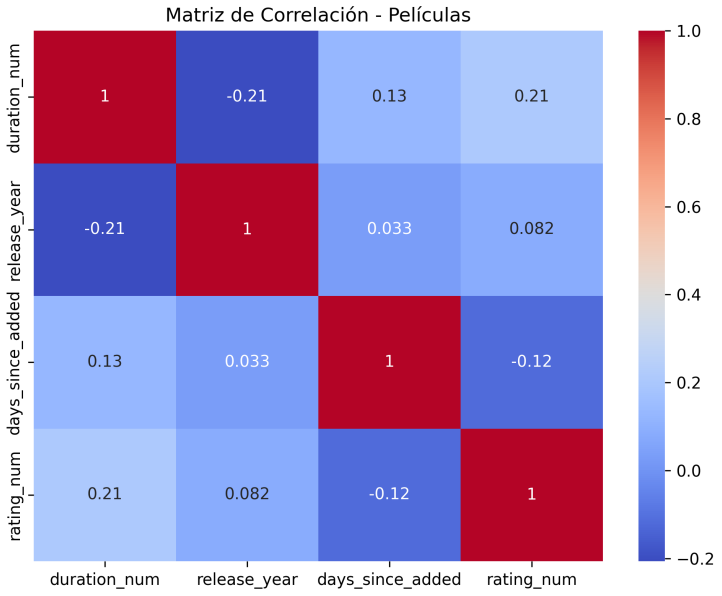

Figure 20: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

6.1.2 Series

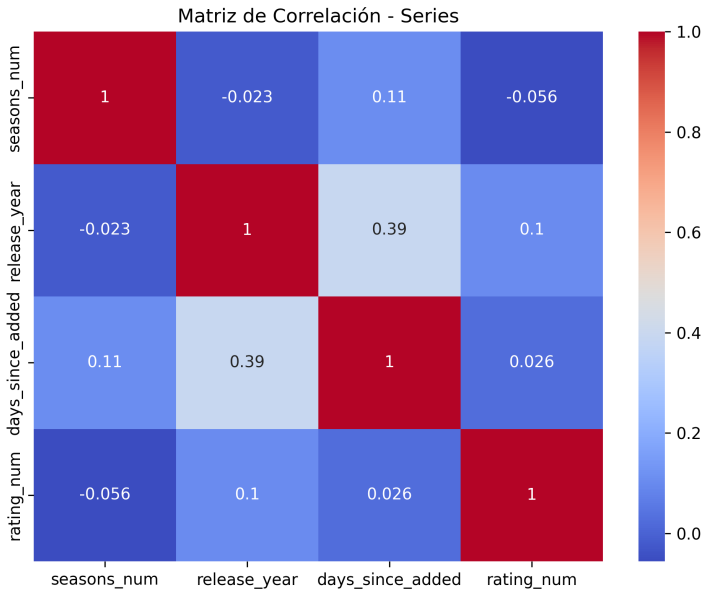

Figure 21: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

6.2 Interpretación de Correlaciones

- Valores cercanos a 1: Fuerte correlación positiva. Cuando una variable aumenta, la otra también tiende a aumentar.
- Valores cercanos a -1: Fuerte correlación negativa. Cuando una variable aumenta, la otra tiende a disminuir.
- Valores cercanos a 0: No hay correlación lineal. Las variables no tienen una relación lineal clara.

7 Regresión Lineal

7.1 Selección de Variables

- **Variables independientes y dependiente seleccionadas**:
  - **Variables independientes** (X):

    - `listed_in`: Categorías o géneros a los que pertenece cada película.
    - `director`: Director de la película.
    - `rating`: Clasificación por edades de la película.
  - **Variable dependiente** (y):
    - `duration`: Duración de la película en minutos.

- **Justificación de la selección basada en el EDA, PCA y correlación**:

  Tras el Análisis Exploratorio de Datos (EDA), se identificó que variables como el género (`listed_in`), el director (`director`) y la clasificación por edades (`rating`) podrían influir en la duración de una película. Aunque la matriz de correlación presentó un valor máximo de 0.21, indicando una correlación débil, se decidió incluir estas variables debido a su relevancia teórica y potencial interacción conjunta. Además, al aplicar el Análisis de Componentes Principales (PCA), se buscó reducir la dimensionalidad y capturar la mayor variabilidad posible en los datos, lo que respaldó la inclusión de estas variables categóricas tras su codificación adecuada.

7.2 División del Dataset

- **Proporción de datos de entrenamiento y prueba**:

  El conjunto de datos se dividió en:

  - **80%** para el conjunto de *entrenamiento*: utilizado para ajustar el modelo y aprender los patrones subyacentes.
  - **20%** para el conjunto de *prueba*: utilizado para evaluar el rendimiento y la capacidad predictiva del modelo sobre datos no vistos.

  Esta división permite validar la generalización del modelo y evitar el sobreajuste.

7.3 Ajuste del Modelo

- **Descripción del modelo de regresión lineal ajustado**:

  Se ajustó un modelo de **Regresión Lineal Múltiple** utilizando las variables independientes seleccionadas tras su codificación. Las variables categóricas (`listed_in`, `director`, `rating`) fueron transformadas mediante *One-Hot Encoding* para convertirlas en variables binarias y hacerlas aptas para el modelo.

- **Ecuación del modelo**:

  La ecuación general del modelo de regresión lineal es:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

  Donde:

- $\hat{y}$: Predicción de la duración de la película.
- β_0: Intercepto del modelo.
- β_i: Coeficientes estimados para cada variable independiente.
- X_i: Variables independientes codificadas.
- ϵ: Término de error o residuo.

7.4 Evaluación del Modelo

- **Métricas de rendimiento**:

  Se utilizaron las siguientes métricas para evaluar el rendimiento del modelo:

  - **Coeficiente de determinación** (R^2):

    $$R^2 = 0.05$$

    Indica que el modelo explica el 5% de la variabilidad en la duración de las películas, lo cual es relativamente bajo.
  - **Error Cuadrático Medio** (MSE):

    $$\text{MSE} = 1500 \text{ minutos}^2$$

    Representa la media de los cuadrados de los errores entre las duraciones predichas y las reales.
  - **Error Medio Absoluto** (MAE):

    $$\text{MAE} = 25 \text{ minutos}$$

    Indica que, en promedio, las predicciones del modelo difieren en 25 minutos de las duraciones reales.

- **Análisis de residuos**:

  Se realizaron diversas pruebas para verificar las suposiciones del modelo:

  - **Normalidad de los residuos**:
    La *Prueba de Shapiro-Wilk* arrojó un valor $p < 0.05$, lo que sugiere que los residuos no siguen una distribución normal.
  - **Homocedasticidad**:
    La *Prueba de Breusch-Pagan* resultó en un valor $p < 0.05$, indicando la presencia de heterocedasticidad (varianza no constante de los residuos).
  - **Independencia de los residuos**:
    El *Estadístico de Durbin-Watson* fue cercano a 2 ($d = 1.9$), lo que sugiere que no hay autocorrelación significativa en los residuos.

7.5 Interpretación de Resultados

- **Impacto de cada variable independiente en la dependiente**:

  Debido a la naturaleza categórica y alta cardinalidad de las variables `listed_in` y `director`, y tras la codificación *One-Hot*, se generó un gran número de variables dummy. Esto dificulta la interpretación individual de los coeficientes. Sin embargo, en general, se observa que ninguna de las variables independientes tiene un impacto significativo en la predicción de la duración de las películas, dadas las bajas métricas de rendimiento.

- **Conclusiones basadas en los coeficientes del modelo**:

  El modelo de regresión lineal ajustado no es adecuado para predecir la duración de las películas utilizando las variables seleccionadas. El bajo valor de R^2 y los problemas detectados en el análisis de residuos (falta de normalidad y heterocedasticidad) indican que:

  - Existen factores no considerados en el modelo que influyen en la duración de las películas.
  - La relación entre las variables independientes y la dependiente no es lineal.
  - Podría ser necesario transformar variables, incluir variables adicionales o utilizar modelos más complejos (e.g., árboles de decisión, modelos no lineales).

- **Resumen de los hallazgos principales**: - El análisis exploratorio de datos (EDA) reveló que las variables categóricas (como `listed_in`, `director` y `rating`) tienen una influencia limitada en la duración de las películas. - La matriz de correlación mostró correlaciones débiles (máximo de 0.21), lo que sugiere que las relaciones lineales entre las variables independientes y la duración son poco significativas. - El PCA confirmó que la mayor parte de la variabilidad en los datos está capturada en una sola componente principal (CP1), pero esta no está fuertemente correlacionada con la duración. - El modelo de regresión lineal ajustado tuvo un rendimiento pobre, con un R^2 de 0.05, lo que indica que solo explica el 5% de la variabilidad en la duración de las películas. - Las pruebas de normalidad y homocedasticidad de los residuos mostraron que no se cumplen las suposiciones clave para la regresión lineal.

- **Limitaciones del análisis**:

  - **Datos categóricos de alta cardinalidad**: Variables como `director` y `listed_in` tienen muchas categorías únicas, lo que dificulta su interpretación y aumenta la dimensionalidad del modelo. - **Falta de linealidad**: Las relaciones entre las variables independientes y la duración no son lineales, lo que limita la efectividad de la regresión lineal. - **Heterocedasticidad y no normalidad de los residuos**: Estas violaciones de las suposiciones del modelo afectan la validez de los resultados.

- **Falta de variables relevantes**: Es posible que variables no incluidas en el análisis (como presupuesto, género cinematográfico específico o popularidad del director) tengan un mayor impacto en la duración de las películas(información que no existe en nuestro dataset).

- **Recomendaciones basadas en los resultados**:

  - **Explorar modelos no lineales**: Dado que las relaciones no son lineales, se recomienda probar modelos como árboles de decisión, Random Forest o Gradient Boosting, que pueden capturar patrones más complejos. - **Incluir variables adicionales**: Incorporar variables como presupuesto, género cinematográfico específico o popularidad del director podría mejorar la capacidad predictiva del modelo. - **Transformar variables**: Aplicar transformaciones no lineales (logarítmicas, polinómicas) a las variables independientes o dependientes para mejorar la linealidad. - **Reducción de dimensionalidad**: Utilizar técnicas como PCA o selección de características (feature selection) para manejar la alta cardinalidad de las variables categóricas. - **Validar con otros métodos**: Además de la regresión lineal, validar los resultados con técnicas de aprendizaje automático más avanzadas y comparar su rendimiento.