

Análisis Estadístico de un Dataset

Tu Nombre

February 15, 2025

Contents

| | | |
|----------|---|-----------|
| 1 | Introducción | 3 |
| 1.1 | Contexto del Problema | 3 |
| 1.2 | Descripción del Dataset | 3 |
| 2 | Análisis Exploratorio de Datos (EDA) | 3 |
| 2.1 | Limpieza de Datos | 3 |
| 2.2 | Análisis Descriptivo | 3 |
| 2.2.1 | Estadísticas descriptivas (media, mediana, desviación estándar, etc.) | 3 |
| 2.2.2 | Visualizaciones | 5 |
| 2.2.3 | Análisis de variables categóricas (tablas de frecuencia, gráficos de barras). | 12 |
| 2.3 | Relaciones Iniciales | 12 |
| 3 | Análisis de Componentes Principales (PCA) | 12 |
| 3.1 | Proceso | 12 |
| 3.2 | Gráficos | 12 |
| 4 | Test de Normalidad | 14 |
| 5 | Formulación de Hipótesis | 15 |
| 5.1 | Duración en películas | 15 |
| 5.2 | Categorías Más Común en Películas | 15 |
| 5.3 | Rating en películas y series | 16 |
| 5.4 | Prueba de los castings en EEUU e India | 16 |
| 6 | Análisis de Correlación | 17 |
| 6.1 | Matriz de Correlación | 17 |
| 6.1.1 | Películas | 17 |
| 6.1.2 | Series | 18 |
| 6.2 | Interpretación de Correlaciones | 18 |

| | | |
|----------|--|-----------|
| 7 | Regresión Lineal | 18 |
| 7.1 | Selección de Variables | 18 |
| 7.2 | División del Dataset | 19 |
| 7.3 | Ajuste del Modelo | 19 |
| 7.4 | Evaluación del Modelo | 19 |
| 7.5 | Interpretación de Resultados | 19 |
| 8 | Validación y Conclusiones | 19 |
| 8.1 | Validación Cruzada | 19 |
| 8.2 | Conclusiones | 19 |
| 9 | Apéndices | 19 |
| 9.1 | Código Utilizado | 19 |
| 9.2 | Tablas y Figuras Adicionales | 19 |

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

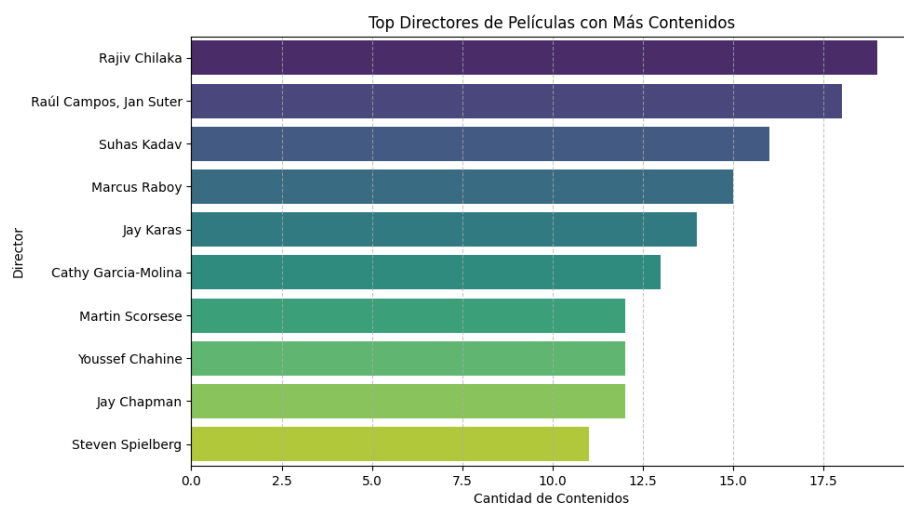


Figure 1: Directores de Películas Más Comunes

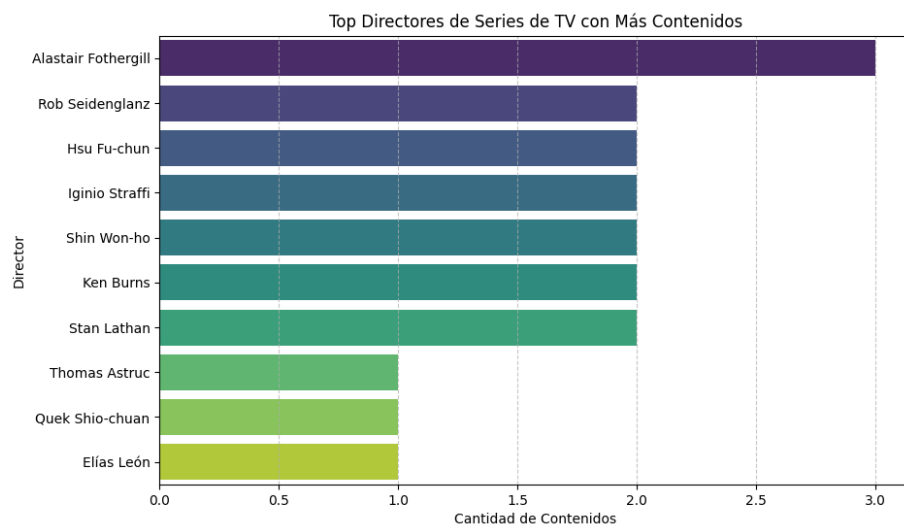


Figure 2: Directores de Series Más Comunes

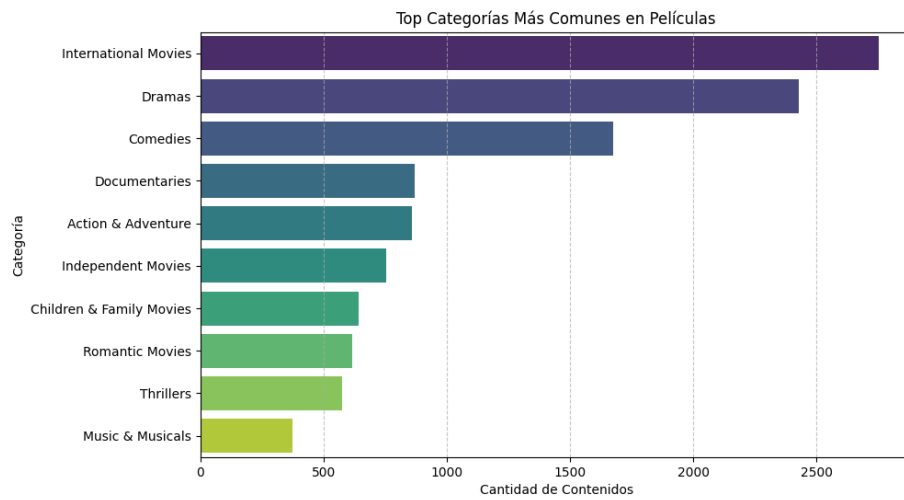


Figure 3: Categorías Más Comunes en Películas

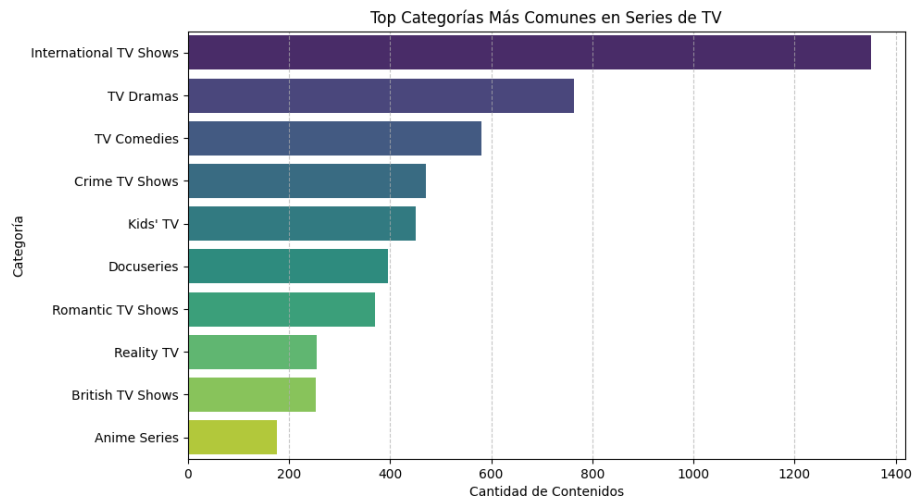


Figure 4: Categorías Más Comunes en Series de TV

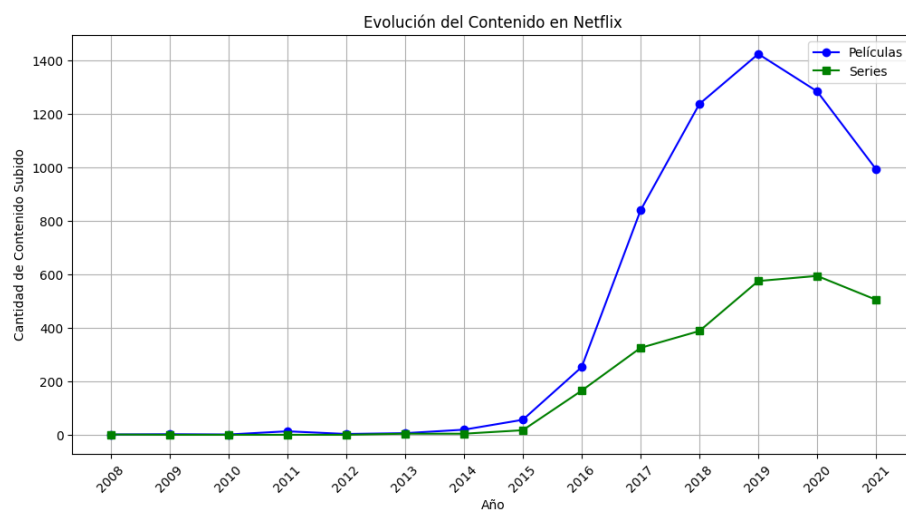


Figure 5: Evolución del Contenido

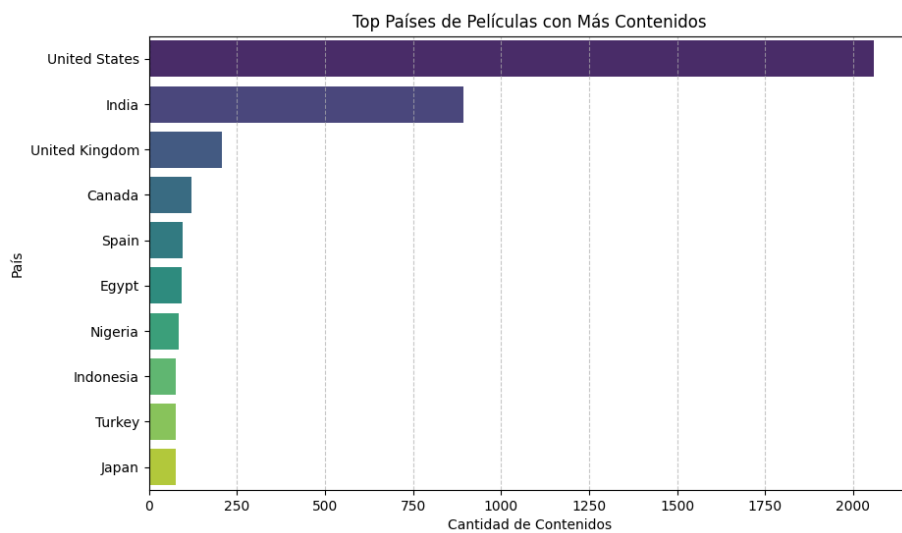


Figure 6: Países con más películas

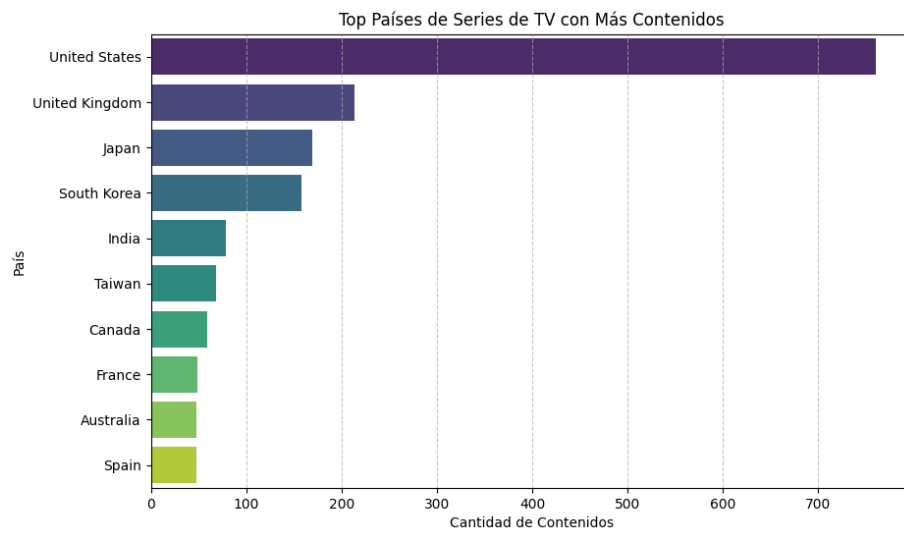


Figure 7: Países con más series

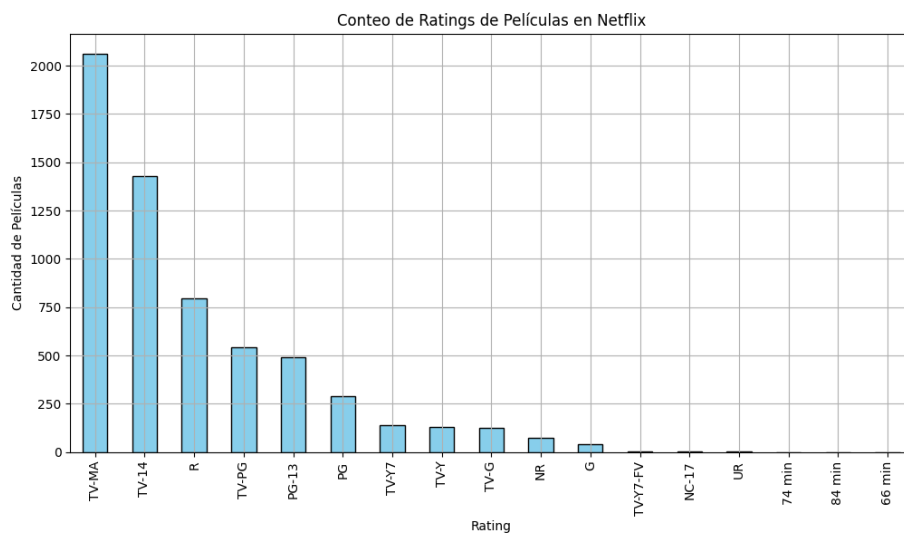


Figure 8: Conteo de Ratings para Películas

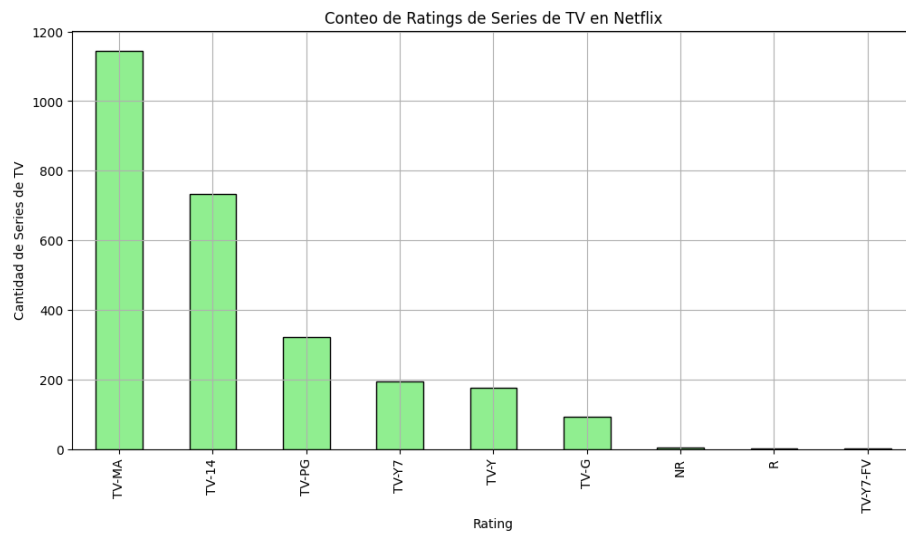


Figure 9: Conteo de Ratings para Series de TV

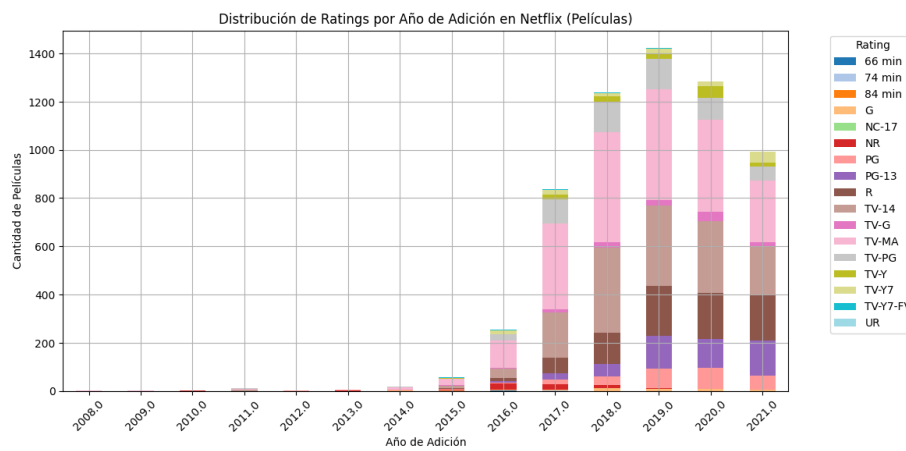


Figure 10: Distribución de Ratings por Año de Adición en Películas

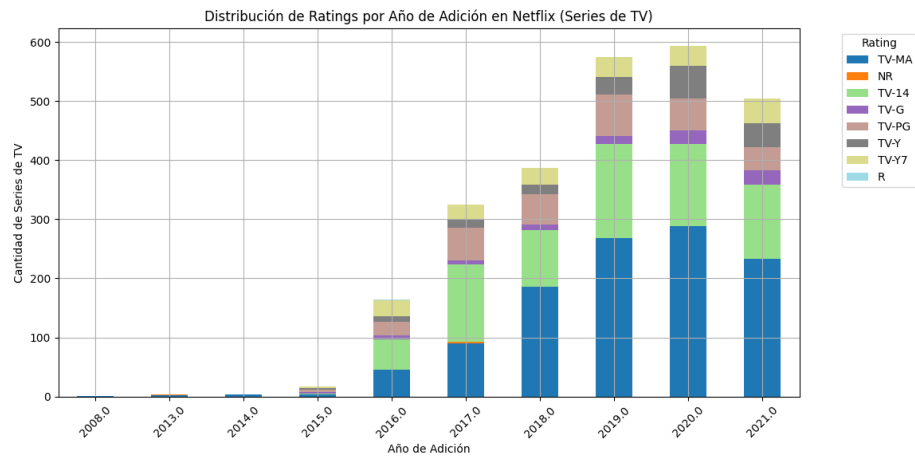


Figure 11: Distribución de Ratings por Año de Adición en Series de TV

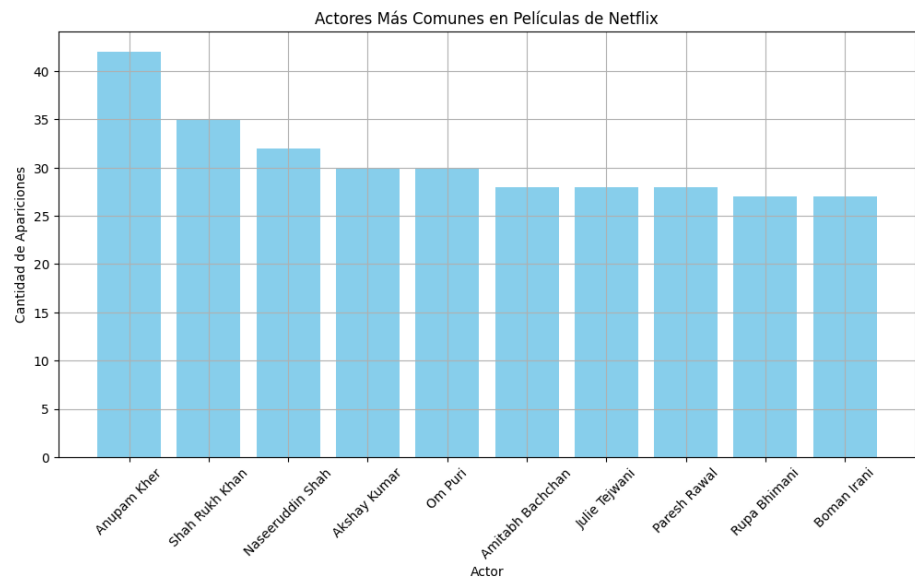


Figure 12: Actores Más Comunes en Películas

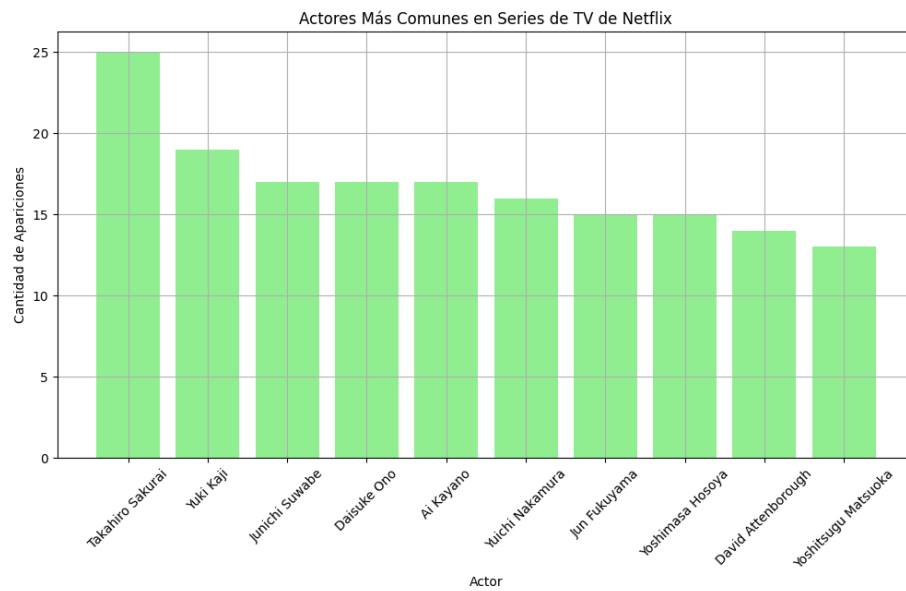


Figure 13: Actores Más Comunes en Series de TV

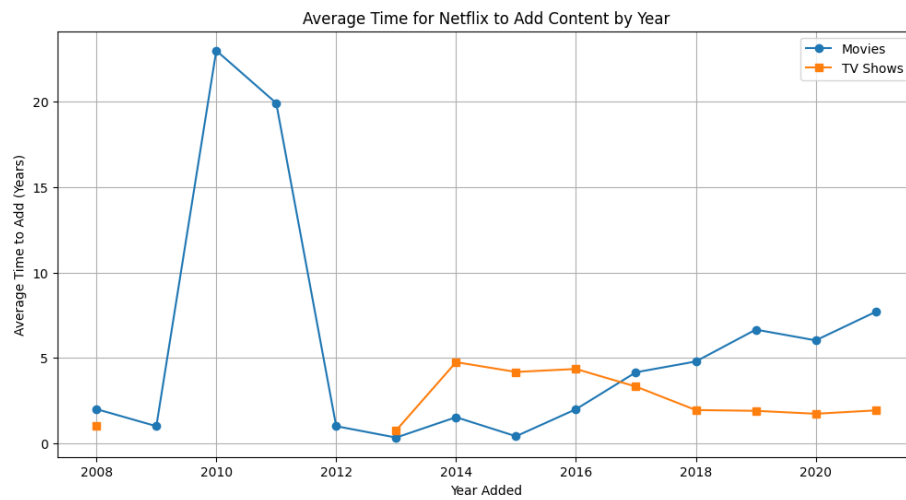


Figure 14: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

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

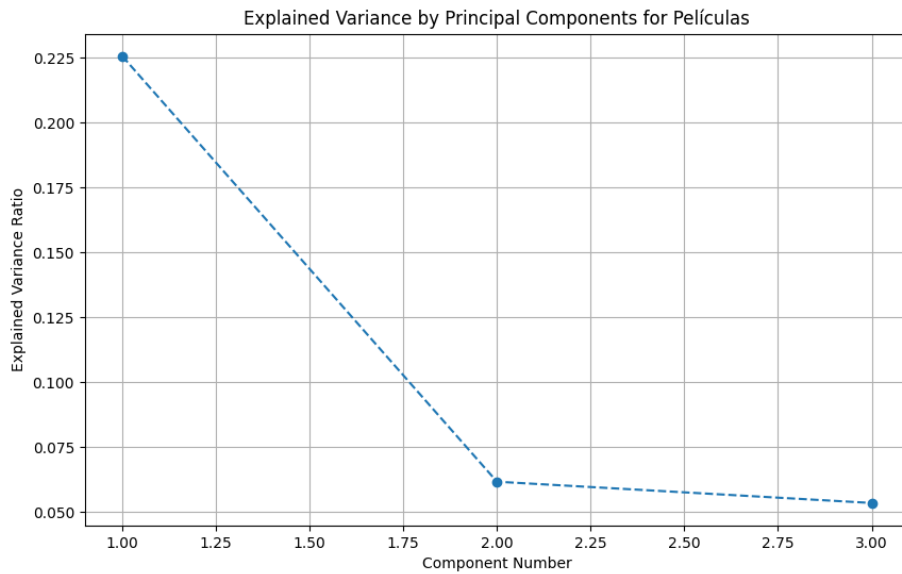


Figure 15: Varianza explicada por componentes principales (Películas)

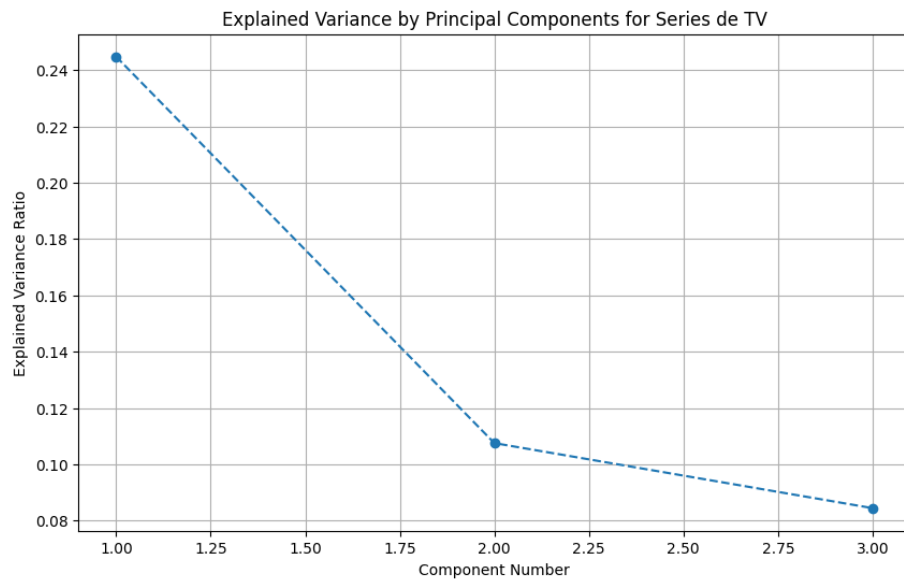


Figure 16: Varianza explicada por componentes principales (Series)

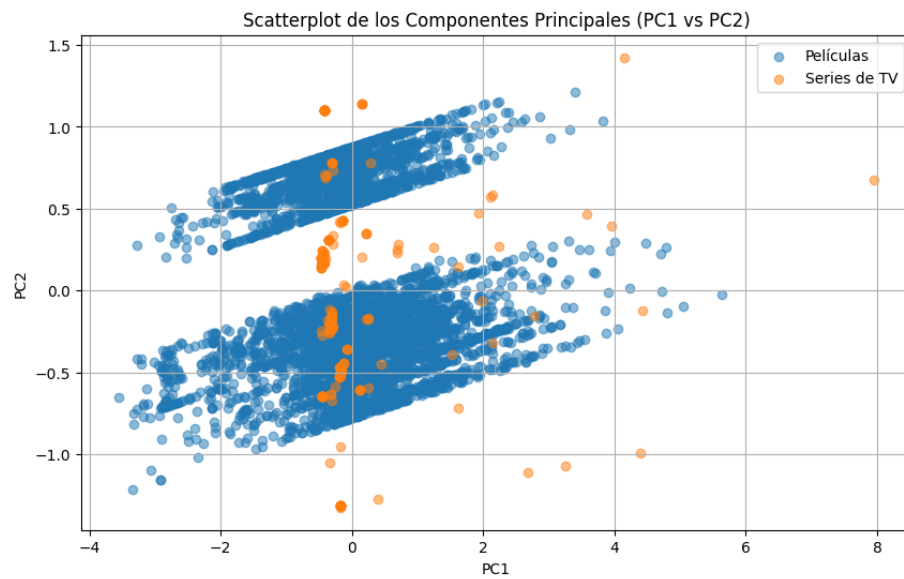


Figure 17: Gráfico de puntos para los componentes principales

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.

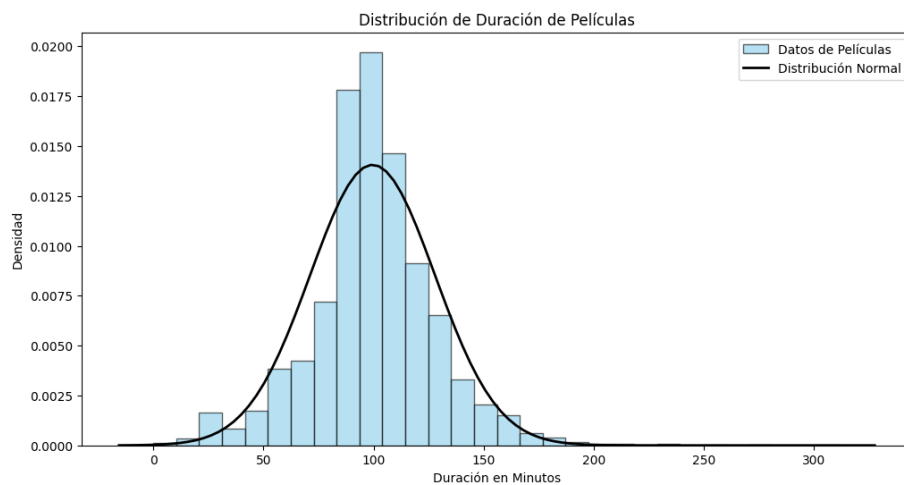


Figure 18: Test de normalidad (Películas)

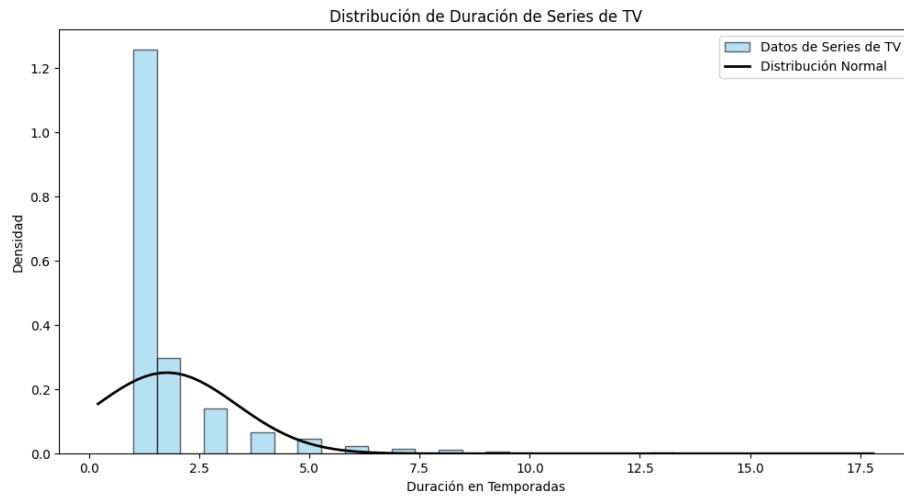


Figure 19: Test de normalidad (Series)

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 películas es mayor a 80 minutos , es lógico pensar que gran parte de las películas de Netflix tienen una duración superior a 80 minutos.

Luego de realizada la prueba de hipótesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipótesis 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 películas como en series es TV-MA. Siendo Estados Unidos es el país con mayor contenido en películas y series, se desea comprobar si el rating mas comun en su contenido es TV-MA.

Luego de realizada la prueba de hipótesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipótesis 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 país con mas películas 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 hipótesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipótesis 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 país con mas series, los actores mas comunes son de nacionalidad japonesa.

5.4 Categorías 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 hipótesis(ver en el archivo Hipotesis.ipynb) se rechaza la hipótesis 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

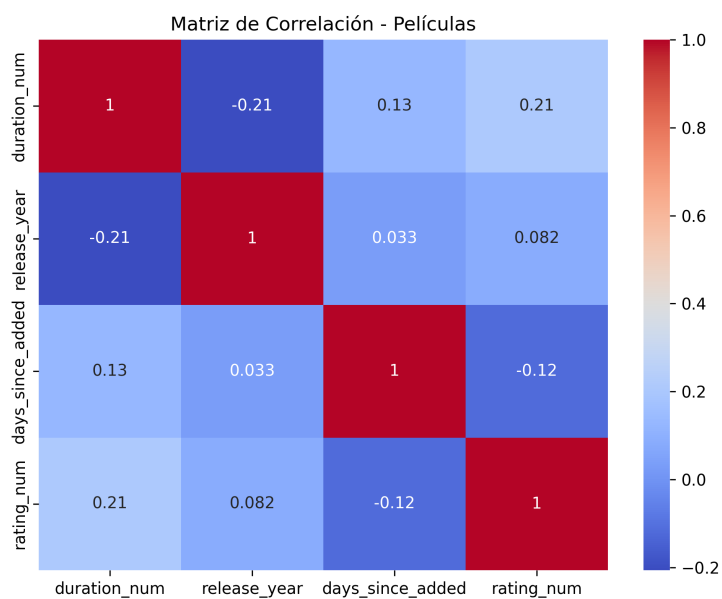


Figure 20: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

6.1.2 Series

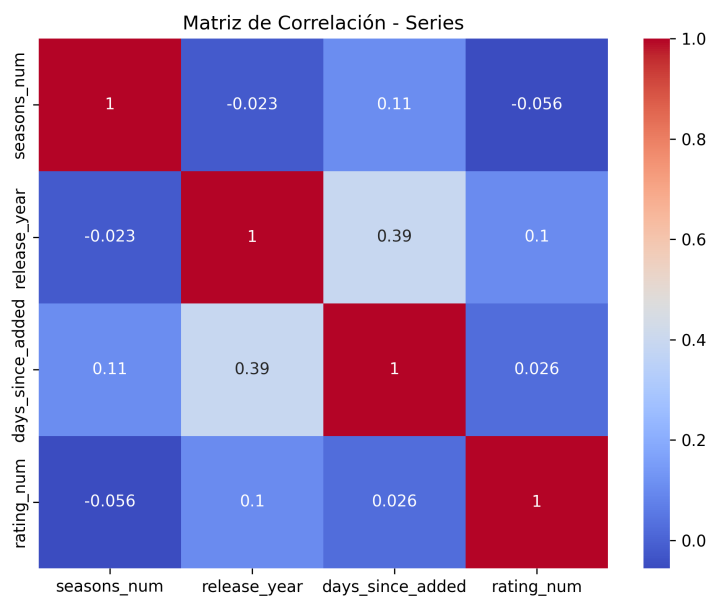


Figure 21: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

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 dependientes seleccionadas.
- Justificación de la selección basada en el EDA, PCA y correlación.

7.2 División del Dataset

- Proporción de datos de entrenamiento y prueba (ej. 80%-20%).

7.3 Ajuste del Modelo

- Descripción del modelo de regresión lineal ajustado.
- Ecuación del modelo.

7.4 Evaluación del Modelo

- Métricas de rendimiento (R^2 , MSE, MAE).
- Análisis de residuos (normalidad, homocedasticidad, independencia).

7.5 Interpretación de Resultados

- Impacto de cada variable independiente en la dependiente.
- Conclusiones basadas en los coeficientes del modelo.

8 Validación y Conclusiones

8.1 Validación Cruzada

- Descripción del proceso de validación cruzada.
- Resultados de la validación.

8.2 Conclusiones

- Resumen de los hallazgos principales.
- Limitaciones del análisis.
- Recomendaciones basadas en los resultados.

9 Apéndices

9.1 Código Utilizado

Incluye el código utilizado para el análisis (si es relevante).

9.2 Tablas y Figuras Adicionales

- Tablas y gráficos que no se incluyeron en el cuerpo principal pero que son relevantes.

Referencias

- Libros, artículos o recursos utilizados para el análisis.