Análisis Estadístico de un Dataset

Tu Nombre

February 14, 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

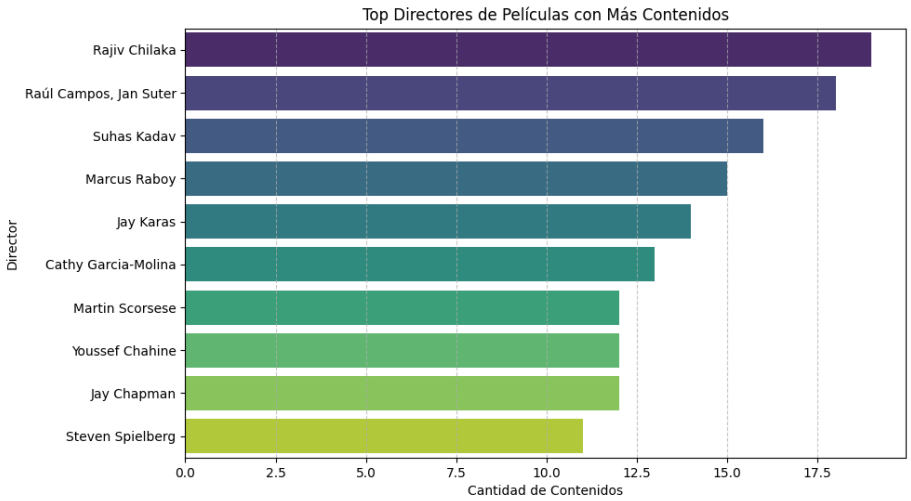

Figure 1: Directores de Películas Más Comunes

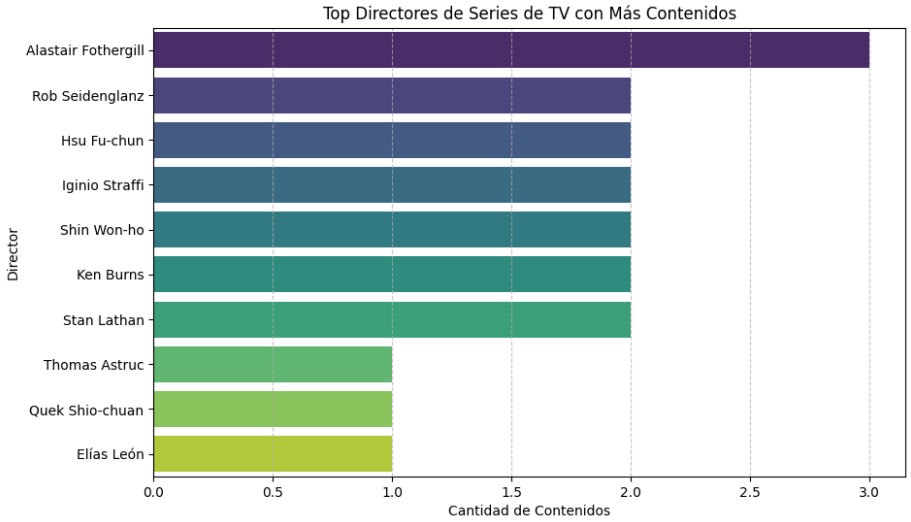

Figure 2: Directores de Series Más Comunes

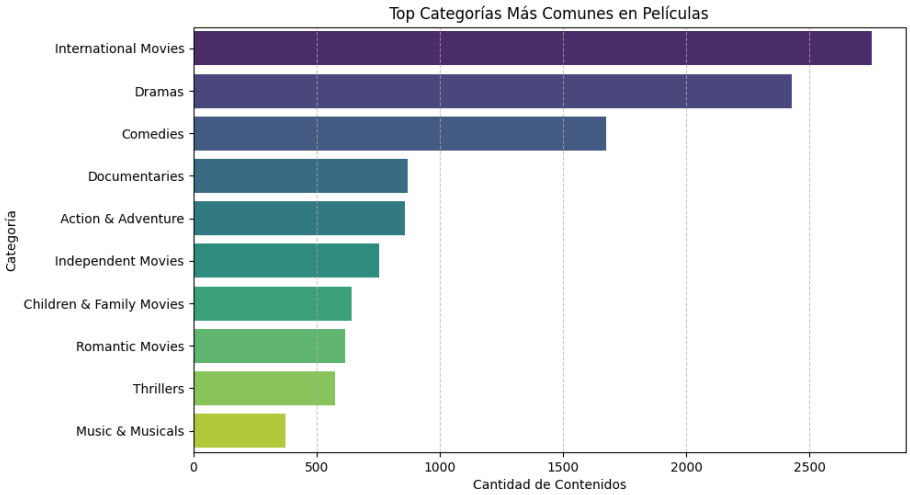

Figure 3: Categorías Más Comunes en Películas

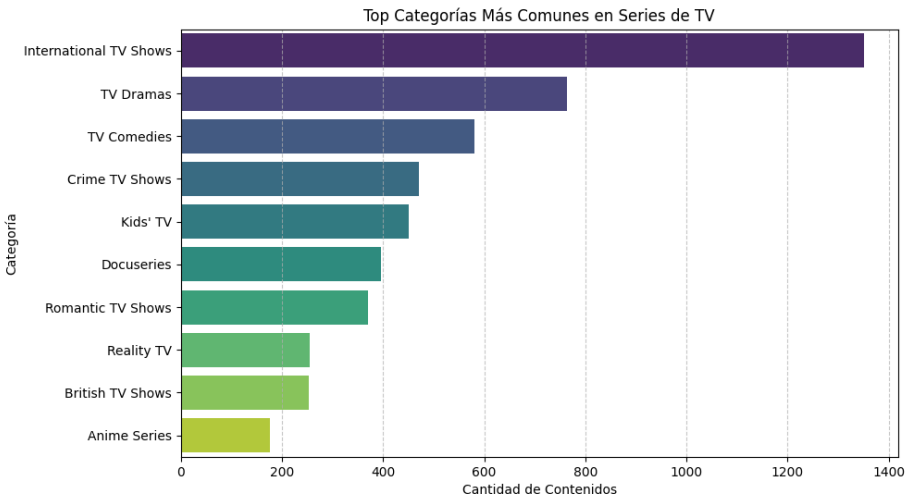

Figure 4: Categorías Más Comunes en Series de TV

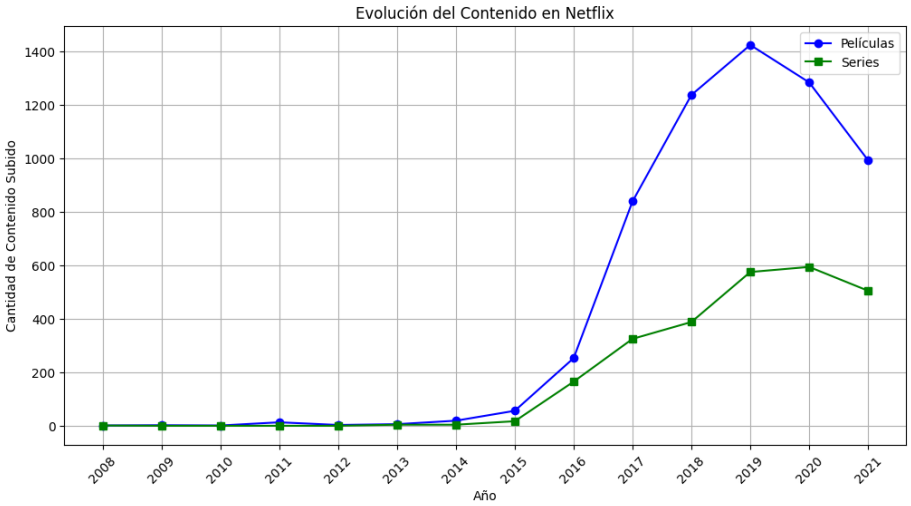

Figure 5: Evolución del Contenido

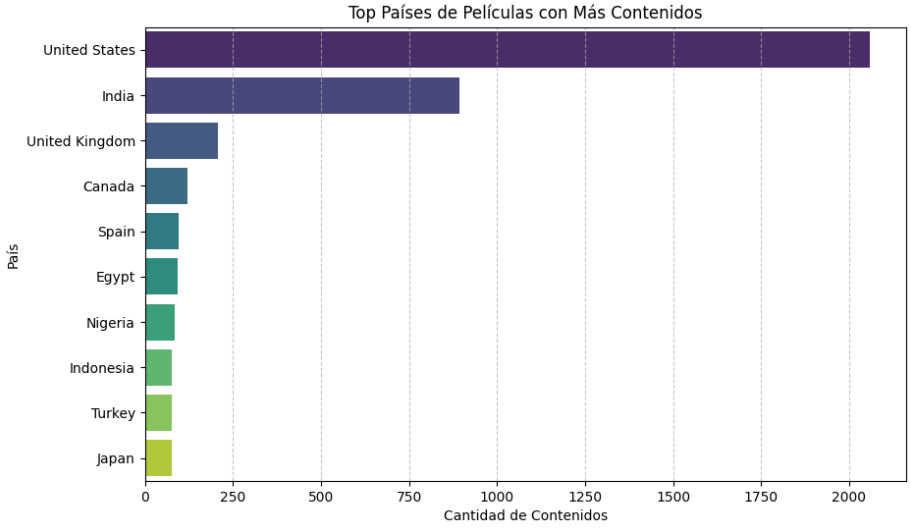

Figure 6: Países con más películas

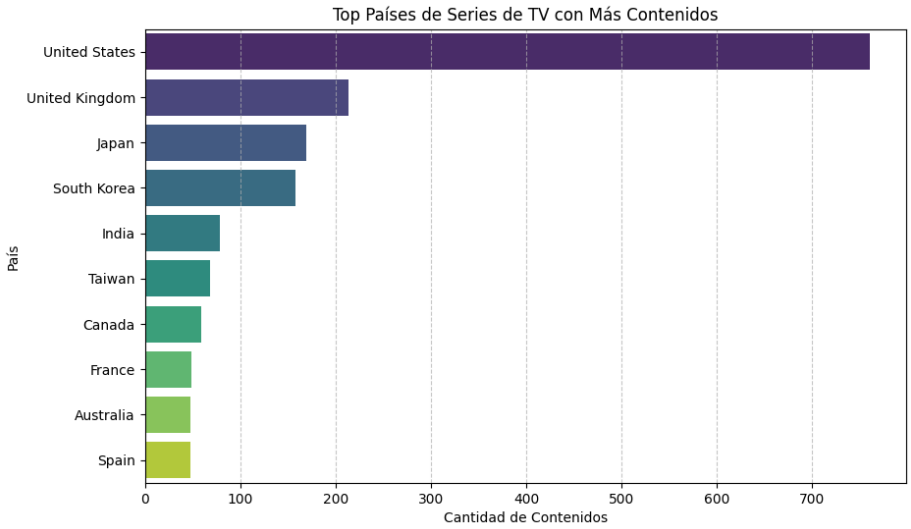

Figure 7: Países con más series

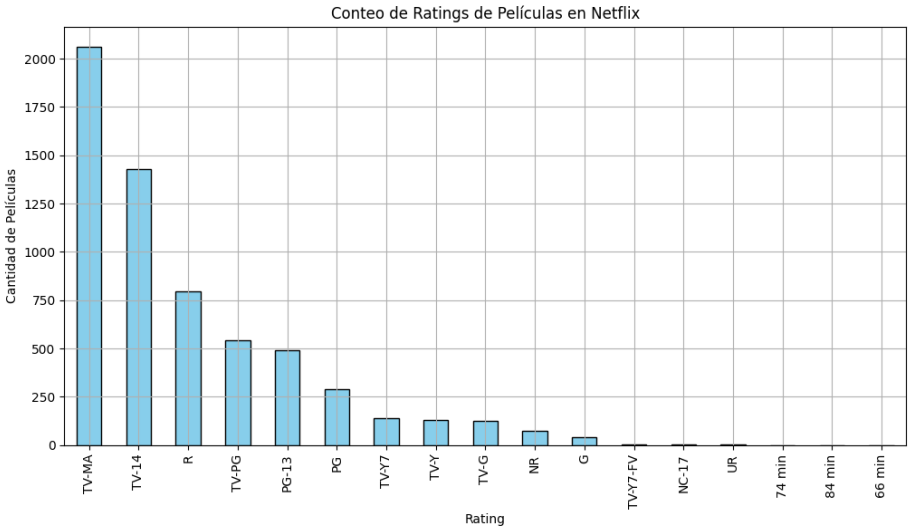

Figure 8: Conteo de Ratings para Películas

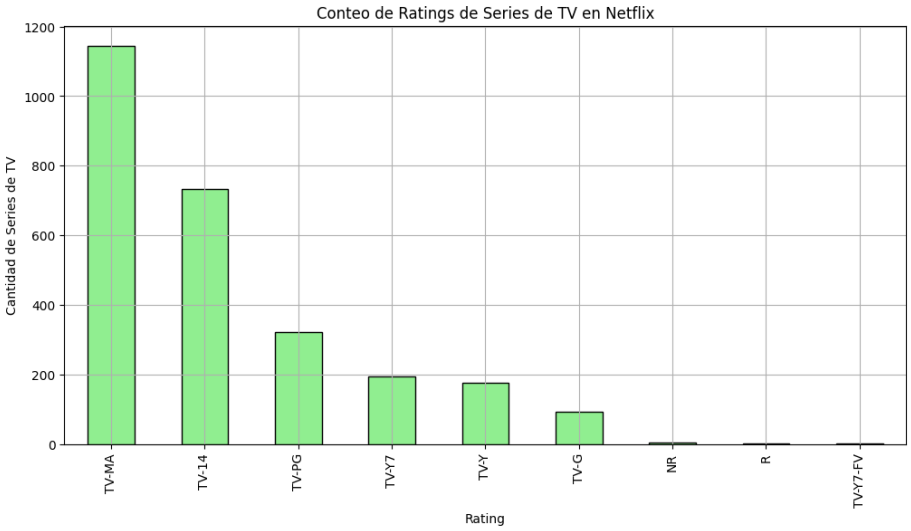

Figure 9: Conteo de Ratings para Series de TV

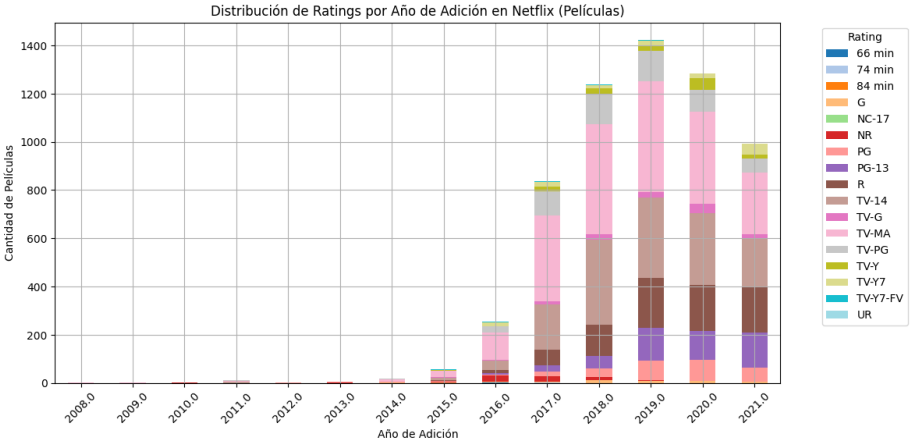

Figure 10: Distribución de Ratings por Año de Adición en Películas

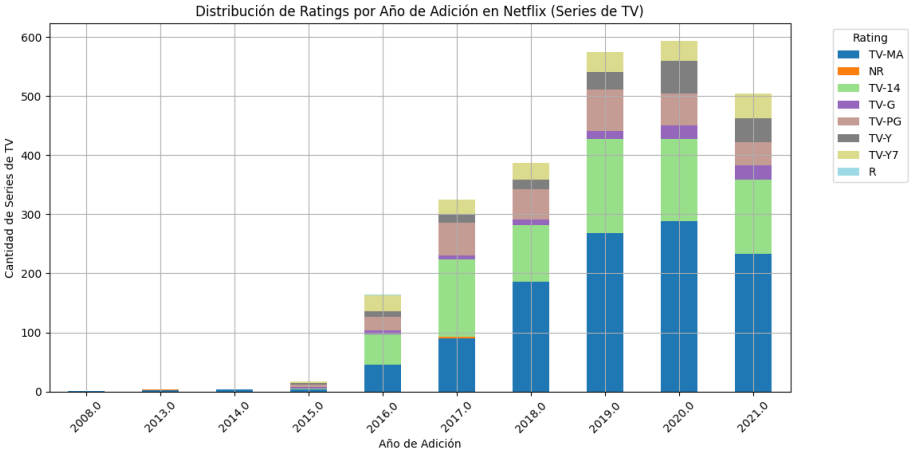

Figure 11: Distribución de Ratings por Año de Adición en Series de TV

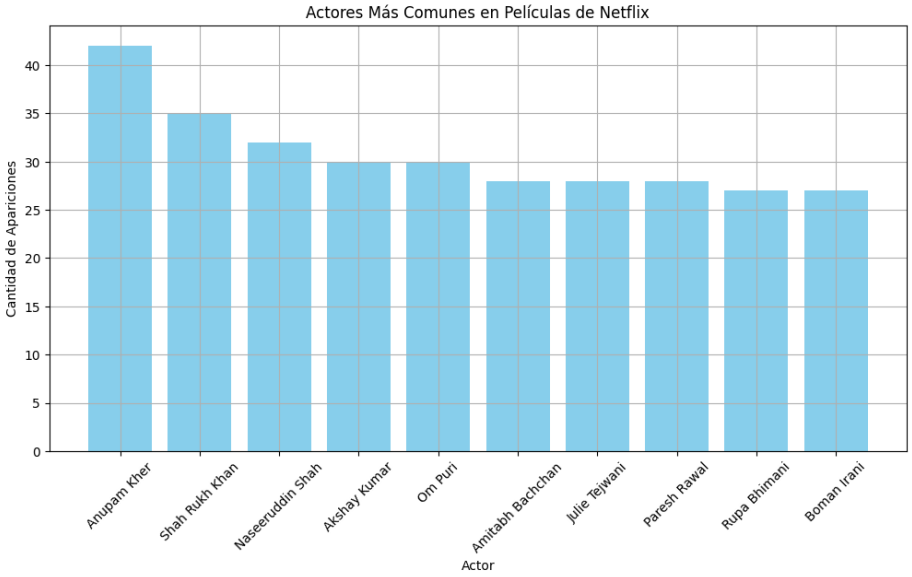

Figure 12: Actores Más Comunes en Películas

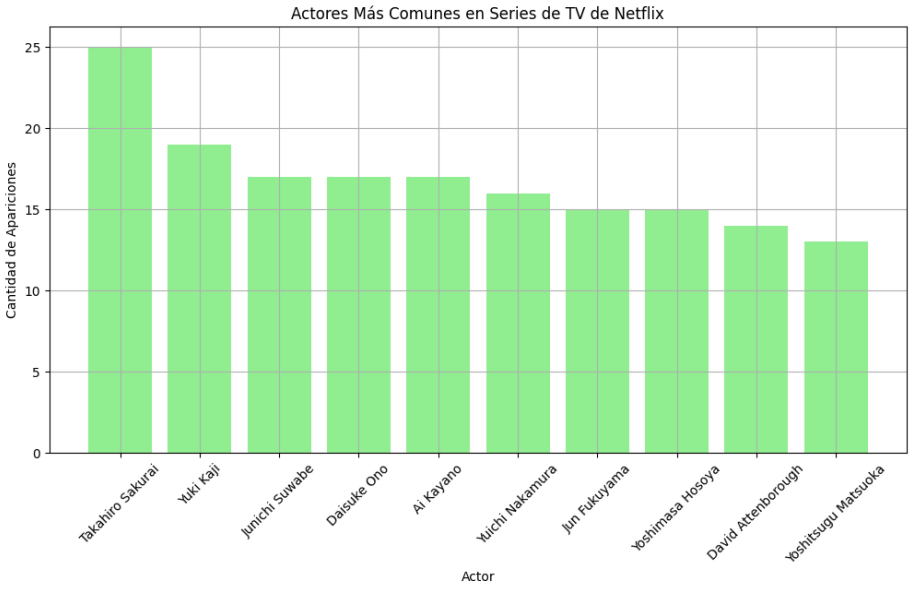

Figure 13: Actores Más Comunes en Series de TV

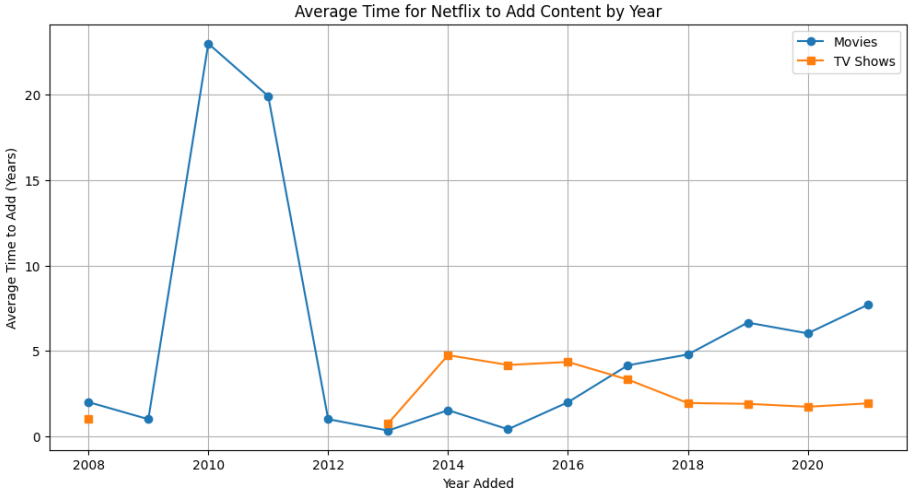

Figure 14: Tiempo promedio anual en añadir el contenido a la plataforma desde su estreno

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

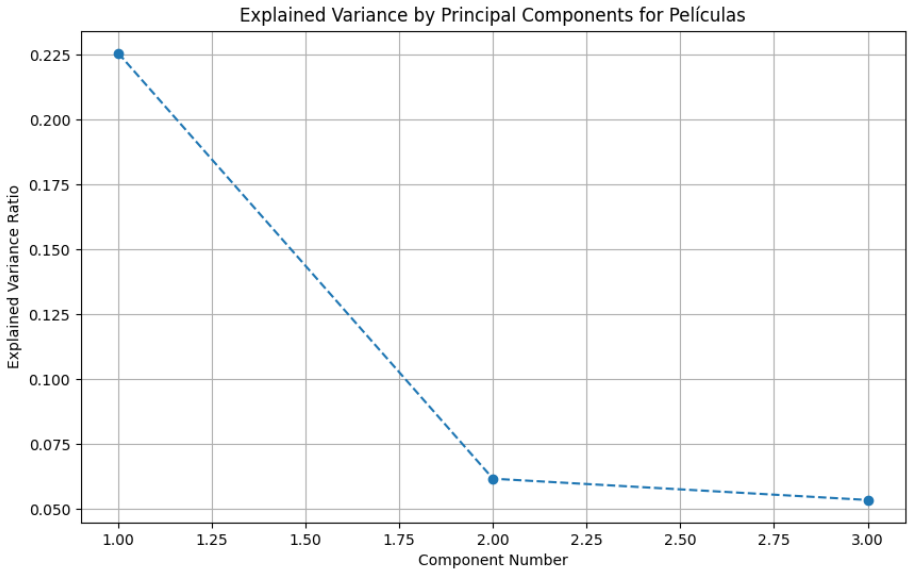

Figure 15: Varianza explicada por componentes principales (Películas)

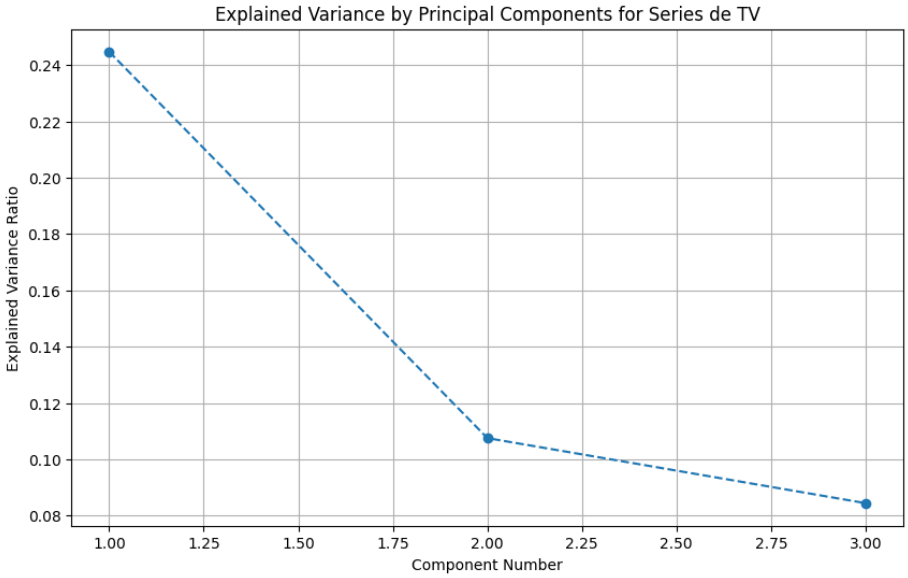

Figure 16: Varianza explicada por componentes principales (Series)

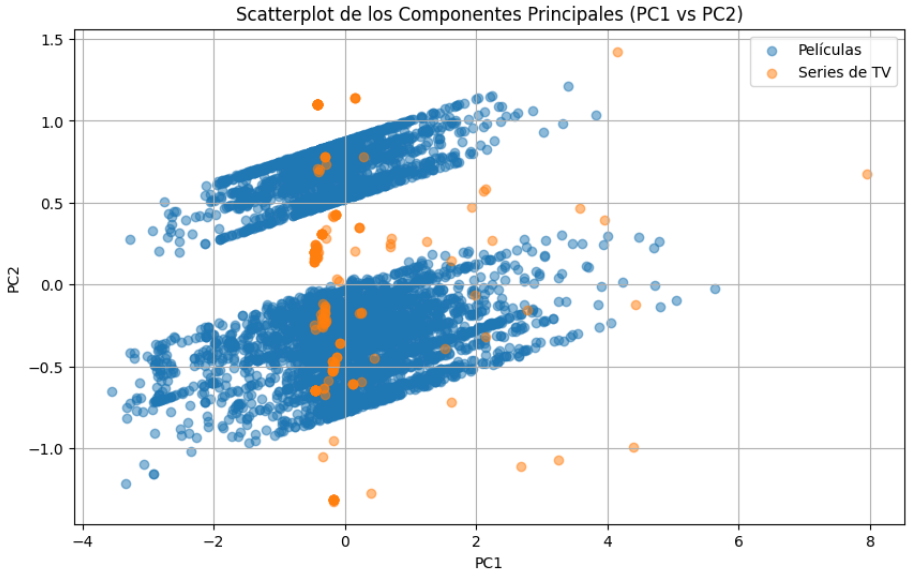

Figure 17: Gráfico de puntos para los componentes principales

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.

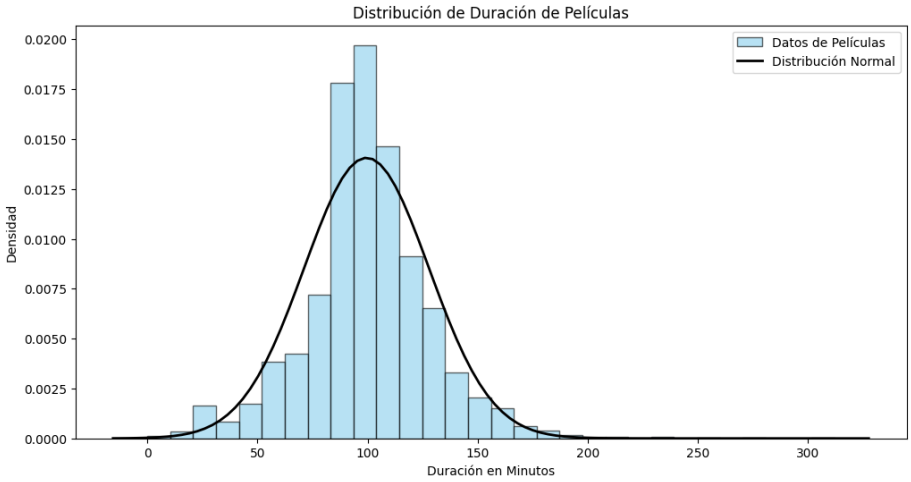

Figure 18: Test de normalidad (Peliculas)

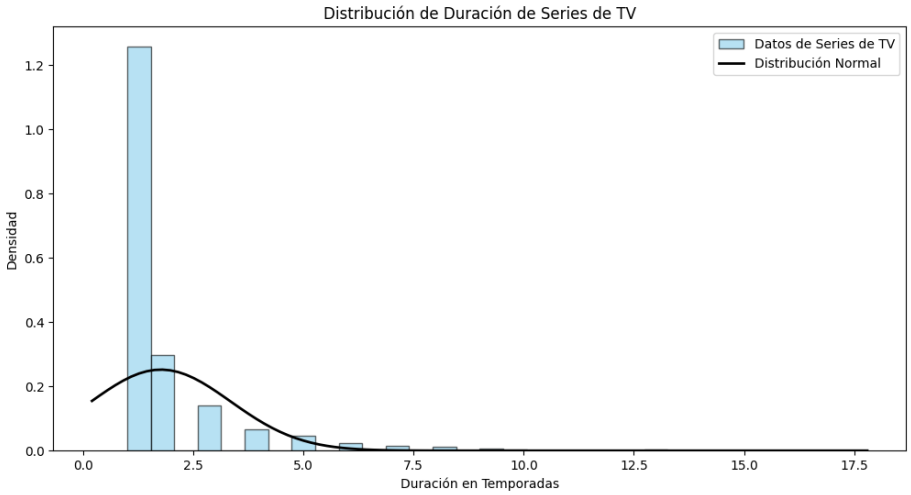

Figure 19: Test de normalidad (Series)

5 Formulación de Hipótesis

5.1 Hipótesis Nula y Alternativa

- Hipótesis nula H0 y alternativa H1.
- Justificación de las hipótesis basada en el EDA y PCA.

5.2 Pruebas Estadísticas Seleccionadas

Descripción de las pruebas a utilizar (t-test, ANOVA, chi-cuadrado, etc.) y su relevancia.

6 Análisis de Correlación

6.1 Matriz de Correlación

- Cálculo de la matriz de correlación.
- Visualización con heatmap.

6.2 Interpretación de Correlaciones

- Identificación de correlaciones fuertes.
- Discusión sobre multicolinealidad y su impacto en el análisis.

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.