



# UN VIAJE A TRAVÉS DE LOS DATOS DE LA INDUSTRIA DEL ENTRETENIMIENTO.

---

## ANÁLISIS DE DATOS DE PELÍCULAS Y SERIES

**Presentado por:**  
Cola Geronimo

# Índice

<b>Introducción.....</b>	<b>1</b>
Contexto de la Industria.....	1
Justificación.....	1
Preguntas de Investigación.....	1
Hipótesis.....	2
Resumen.....	2
<b>Datos.....</b>	<b>2</b>
Base de Datos Principal - movies_metadata.csv.....	2
Base de Datos Secundaria - tv_shows.csv.....	3
<b>Metodología.....</b>	<b>3</b>
Fases del Proyecto:.....	4
1. Confección del entorno de trabajo.....	4
2. Carga de Datos.....	4
3. Limpieza y Normalización de Datos.....	5
4. Análisis Exploratorio de Datos y Visualización.....	5
5. Hipótesis y Modelado.....	6
Limitaciones Metodológicas.....	7
<b>Resultados.....</b>	<b>7</b>
Primera Hipótesis.....	7
Gráfico de Dispersion.....	7
Segunda Hipótesis.....	8
Gráfico de Dispersion.....	8
Gráfico de Puntos.....	8
Gráfico de Barros.....	9
Tercera Hipótesis.....	9
Mapa Coroplético.....	9
Cuarta Hipótesis.....	10
Mapa Coroplético.....	10
<b>Conclusiones.....</b>	<b>10</b>
Primera Hipótesis.....	10
Segunda Hipótesis.....	11
Tercera Hipótesis.....	12
Cuarta Hipótesis.....	12
Recomendaciones Estratégicas y Prácticas.....	13

# Introducción

## Contexto de la Industria

En la actualidad, la globalización y transformación digital ha reestructurado toda la estructura del entretenimiento a nivel mundial.

La industria del entretenimiento es dominada por las grandes productoras y las grandes plataformas de streaming, las que junto a la democratización del acceso al contenido y la segmentación de las audiencias han dado lugar al exceso de producción audiovisual, situando en un elevado nivel de importancia el desarrollo e implementación de sistemas de recomendación precisos y estratégicos.

## Justificación

Esta investigación surge de la necesidad de entender el momento por el que está pasando la industria del entretenimiento, en días donde la saturación del contenido dificulta el descubrimiento de contenido de calidad y donde las plataformas disputan firmemente la atención del usuario, es fundamental identificar patrones reales de éxito basados en datos concretos.

## Preguntas de Investigación

¿Qué factores predicen la calidad de una producción?

¿Cómo ha evolucionado la calidad del contenido televisivo en la era del streaming?

¿Qué patrones geográficos caracterizan la producción de contenido de alta calidad?

## Hipótesis

Se plantearon las siguientes hipótesis:

Hipótesis	Enunciado
H1	No existe una relación relevante entre el presupuesto de una película y su calificación promedio (rating).
H2	La calidad de las producciones (medida por el <i>rating</i> ) ha disminuido significativamente a lo largo del tiempo.
H3	La producción audiovisual está altamente concentrada geográficamente.
H4	El liderazgo en volumen de producción no garantiza liderazgo en calidad.

## Resumen

Este estudio representa un análisis sobre la industria audiovisual contemporánea, analizando más de 44.000 producciones entre películas y series de televisión. El análisis busca responder preguntas sobre los factores, variables y circunstancias que determinan el éxito en la industria.

En otras palabras, el objetivo del análisis es evaluar y cotejar la calidad, las tendencias temporales y diversos factores entre las películas y las series de televisión.

## Datos

Para este análisis, se preseleccionaron y manejaron dos bases de datos principales, si bien estas pertenecen a dos formatos audiovisuales distintos, dan lugar a la comparación directa entre estos dos formatos.

### Base de Datos Principal - [movies\\_metadata.csv](#)

Este archivo contiene metadatos de las 45.000 películas incluidas en el conjunto de datos completo de MovieLens. Este conjunto incluye películas estrenadas hasta julio de 2017. Los datos disponibles son: reparto, equipo técnico, palabras clave de la

trama, presupuesto, recaudación, carteles, fechas de estreno, idiomas, productoras, países, número de votos en TMDB y promedio de votos.

Las principales variables analizadas son:

- Presupuesto de producción (budget).
- Calificación promedio de usuarios (vote\_average).
- Duración en minutos (runtime).
- Fecha de lanzamiento (release\_date).
- Géneros cinematográficos (genres).
- El lenguaje original de las películas (original\_languages).

## Base de Datos Secundaria - [tv\\_shows.csv](#)

Los datos recopilados comprenden una lista exhaustiva de programas de televisión disponibles en diversas plataformas de streaming. Contiene información de series, incluyendo el rating de IMDb, la duración promedio de los episodios y el país de origen.

Las principales variables analizadas son:

- Calificación IMDb (IMDb, formato numérico estandarizado).
- Calificación Rotten Tomatoes (Rotten Tomatoes, formato estandarizado).
- Año de lanzamiento (Year).
- Distribución por plataforma (Netflix, Hulu, Prime Video, Disney+).

Las bases debieron pasar por un proceso de limpieza, de normalización para estandarizar formatos, manejar valores nulos (sobre todo en las columnas de "ratings" y "duración"), asegurando la compatibilidad de las variables a la hora de compararlas.

## Metodología

Se diseñó y llevó a cabo una secuencia de pasos que implicó: la recolección de múltiples fuentes, la limpieza y depuración de datos, el análisis exploratorio de la información y la confirmación o refutación de hipótesis, empleando herramientas y librerías de visualización de datos, detalladas en la siguiente sección:

## Fases del Proyecto:

### 1. Confección del entorno de trabajo

En primer lugar, fue necesario acondicionar

- A. Google Colaboratory: es el entorno de desarrollo donde se llevará a cabo el análisis, funcionando sobre el lenguaje de programación Python.
- B. Google Drive: se emplea el almacenamiento en la nube de Google para alojar las bases de datos y acceder a estas por este medio.
- C. Librerías:

#### **Análisis de Datos**

- a. NumPy.
- b. Pandas.

#### **Visualización**

- c. Matplotlib.pyplot:.
- d. Seaborn.

#### **Geoespaciales y de Mapeo**

- e. Geopandas.
- f. Cartopy.
  - i. cartopy.feature.
  - ii. cartopy.crs.
- g. Contextily.

### 2. Carga de Datos

- A. **Fuentes:** se importaron dos conjuntos de datos en colaboración con Google Drive, directamente al entorno de trabajo Google Colaboratory: **movies\_metadata.csv** (para películas) y **tv\_shows.csv** (para series).
- B. **Identificación de variables principales:** se identificaron las variables fundamentales para la comparación: budget, vote\_average, IMDb\_rating, curtime, episode\_run\_time, el país de origen, etc.

### 3. Limpieza y Normalización de Datos

Esta fase fue fundamental para asegurar la comparabilidad de los dos conjuntos de datos, compuesta de:

- A. **Filtrado de valores nulos:** se eliminaron filas donde el rating era nulo, ya que son valores principales para medir la calidad.
- B. **Conversión de tipos:** fue necesario convertir las columnas de rating a formato numérico y las columnas de tiempo a formato de fecha y hora para el análisis temporal.
- C. **Normalización de ratings:** se normaliza la escala de los ratings, asegurando que ambos conjuntos representan la calidad en el mismo formato.
- D. **Información geográfica:** se requirió procesar las cadenas de texto de países y géneros para obtener valores únicos y medibles, trascendentales para el mapa coroplético.

### 4. Análisis Exploratorio de Datos y Visualización

Se llevó a cabo un análisis exploratorio de los datos para obtener una interpretación inicial y verificar la distribución de los datos:

- A. **Análisis de medidas estadísticas:** se calcularon los promedios de rating (Series: 7.09, Películas: 6.00), medias, medianas y desviaciones estándar, esenciales para la investigación.
- B. **Distribución de ratings:** se generaron histogramas y gráficos de densidad para observar la frecuencia y las tendencias de la variable en ambos conjuntos. También se visualizó la relación Presupuesto vs. Rating y otros hallazgos significativos.
- C. **Análisis temporal:** se calcularon los promedios de ratings por año y el total de series por año, para luego generar los gráficos relacionados con la evolución de las series a lo largo del tiempo y la tendencia en su calidad. Se segmentó por períodos temporales.
- D. **Análisis Geográfico:** se registró el número total de producciones por país y se calculó el rating promedio para cada uno de estos países.

## 5. Hipótesis y Modelado

A. Primera Hipótesis (Presupuesto vs. Rating): para esta hipótesis se calculó el coeficiente de Correlación de Pearson entre el presupuesto y el rating, buscando demostrar o no una relación relevante. Para esto se dejó afuera:

- a. Películas con presupuesto menores a \$10.000 dólares.
- b. Películas con ratings menores a 1.

Se realizó un gráfico de dispersión entre el Presupuesto y el Rating, junto a una línea para representar la tendencia.

B. Segunda Hipótesis (Tendencia Temporal): para este caso fue necesario agrupar ambos conjuntos de datos por medio del Año de Estreno y emplear el rating promedio de la suma de series por año. Para verificar la hipótesis se representó:

- a. En un gráfico de barras se plasmó el rating promedio de todas las series del conjunto de datos.
- b. Con un gráfico de torta se demostró la distribución de la calidad de las series
- c. A partir de los ratings promedios de la suma de series por año, se creó un gráfico de líneas buscando representar la evolución de la calidad de las series desde el año 2000 hasta el 2021.
- d. Con un diagrama de dispersión se demostró la tendencia de la calidad de las series a lo largo de 2000-2021.

C. Tercera Hipótesis (Concentración Geográfica): fue necesario identificar los 15 países con mayor producción cinematográfica, para luego:

- a. Presentar un gráfico de barras sobre los países con dicha producción.
- b. Generar un mapa coroplético para demostrar la densidad en cuanto a producción cinematográfica se refiere.

D. Cuarta Hipótesis (Volumen vs. Calidad): fue necesario calcular el rating promedio de los 15 países y calcular el rating promedio para cada uno, para luego:

- a. Presentar un gráfico de barras sobre los países con mayor rating promedio.

- b. Generar un mapa coroplético para visualizar la distribución geográfica de la calidad promedio por país.

## Limitaciones Metodológicas

- Cobertura Geográfica: existe un notable sesgo hacia producciones en inglés
- Disponibilidad de Datos: ciertos valores de importantes variables están incompletas en algunos registros.
- Periodo Temporal: los conjuntos de datos no cuentan con registros en los últimos años (2022-2025)
- Sesgo hacia producciones en plataformas de streaming principales

## Resultados

### Primera Hipótesis

No existe una relación significante entre el presupuesto y la calidad de la producción, películas debajo presupuesto superan en calidad a varias producciones costosas, demostrando que los factores creativos y artísticos, como el guion, la dirección y la actuación, predominan sobre las grandes inversiones.

- Coeficiente de correlación de Pearson:  $r = 0.074$
- Muestra analizada: 8.377 películas con datos confiables
- Correlación muy débil.

### Gráfico de Dispersion

El gráfico muestra la relación entre el Presupuesto (USD) de las películas en el eje X y su Rating en el eje Y.

- La nube de puntos es amplia y dispersa, especialmente en el lado izquierdo (presupuestos bajos). Hay una concentración de puntos en el rango de presupuesto bajo a medio-bajo (cercano a 0 USD), cubriendo la mayoría del espectro de ratings (aproximadamente de 2.0 a 10.0, aunque la densidad es mayor entre 4.0 y 8.0). Es una correlación muy débil o casi nula
- La línea de tendencia de color rojo, es casi horizontal, pero muestra una pendiente ligeramente positiva.
- Rating de Películas por Presupuesto:

- ◆ Bajo Presupuesto (50 M\$ USD): Muestra la mayor variabilidad, cubriendo todo el espectro de calidad (desde ratings muy bajos hasta los más altos, 10.0).
- ◆ Alto Presupuesto (200 M\$ USD): Estos títulos se concentran en el rango de rating medio a alto (5.0 - 7.5) y evitan consistentemente los ratings más bajos.

## Segunda Hipótesis

La calidad de las series ha experimentado un deterioro sostenido y significativo desde el año 2000, coincidiendo temporalmente con la explosión de plataformas de streaming y la saturación del mercado.

- Correlación temporal:  $r = -0.113$  (tendencia negativa).
- Cambio promedio anual: -0.0269 puntos de rating.
- Período analizado: 2000-2021.
- Promedio general de 7.06.
- Series analizadas: 4,021 producciones.

## Gráfico de Dispersión

El gráfico muestra la tendencia de la calidad a través de los últimos 20 años.

- ➔ La línea de tendencia (roja punteada) tiene una pendiente negativa. Confirmando la disminución en el rating.
- ➔ Hay una diferencia de 0.56 puntos entre el mejor año (2000, 7.35) y el peor año (2021, 6.79), representando una caída del 7.6% con respecto al rating máximo del período.

## Gráfico de Puntos

Este gráfico refleja la evolución de los ratings promedio de las series por año de estreno.

- ➔ La línea roja de rating promedio varía fuertemente entre 6.8 y 7.4.
- ➔ Las series se mantienen por encima del umbral de "Calidad Media (6.0+)" y se acercan y superan el "Promedio General (7.06)" en varios años.
- ➔ A partir de 2010, el Rating se mantiene más estable alrededor del promedio general.

→ Análisis por décadas:

- ◆ Series clásicas ( $\leq$ 2010): rating promedió 7.09
- ◆ Series modernas ( $>$ 2010): rating promedió 6.98
- ◆ Diferencia: -0.11 puntos.

## Gráfico de Barras

Este gráfico refleja la cantidad de series que se produjeron durante las últimas dos décadas.

- Pre-2010: la producción anual era relativamente baja y estable, moviéndose de 26 a 106 series. En este período, el rating promedio era más volátil, pero alcanzaba los puntos más altos.
- Post-2010; impulsado por el auge de las plataformas de streaming (Netflix, Amazon, etc.), la cantidad de series se dispara, pasando de 133 en 2011 a un pico de 467 en 2019.

## Tercera Hipótesis

Se puede confirmar con un alto grado de certeza que la producción audiovisual está altamente concentrada geográficamente, con un dominio absoluto de Estados Unidos, seguido por un grupo selecto de países de Europa Occidental.

- Dataset analizado: 38,868 películas con datos geográficos
- Países identificados: 158 países únicos
  - La producción significativa se concentra en solo 15 países.
    - 1. Estados Unidos: 19,501 películas (50.2% del total)
    - 2. Reino Unido: 3,855 películas (9.9% del total)
    - 3. Francia: 3,570 películas (9.2% del total)

## Mapa Coroplético

El mapa representa la producción cinematográfica mundial por país, demostrando la concentración

- Estados Unidos está coloreado con el rojo más intenso, lo que indica la mayor categoría de producción (más de 17.500 películas), dominando por alto margen.

- La producción en Europa muestra países con distintos tonos de amarillo, representando una producción media.
- Muchos países en América del Sur, África y parte de Asia están sombreados de color gris, representando una industria significativamente menor o nula.

## Cuarta Hipótesis

Esta hipótesis puede confirmarse, si bien EE.UU es el país con mayor producción, en un top de 10 países productores de los datasets, Estados Unidos no participaría.

- 1º Lugar en Calidad: Corea del Sur con un Rating Promedio de 6.53.
- 2º Lugar en Calidad: Japón con un Rating Promedio de 6.46.
- 3º Lugar en Calidad: Hong Kong con un Rating Promedio de 6.30.
- Los países que lideran en calidad tienen un volumen de producción significativamente menor.

## Mapa Coroplético

El mapa representa el rating promedio del total de producciones audiovisuales de cada país, demostrando claramente, que no es el mayor productor quien lidera en calidad.

- Corea del Sur y Japón en tonos verde oscuro/verde claro (la categoría más alta de rating)
- EE. UU. aparece en rojo representando ratings bajos en la escala de calidad (5.86).

## Conclusiones

### Primera Hipótesis

La inversión no define el éxito, siendo el rating altamente independiente del presupuesto. El presupuesto es irrelevante para predecir la calidad percibida por el usuario. Las grandes productoras deberían tomar decisiones priorizando, el mérito artístico, la historia y el talento creativo, sobre consideraciones de presupuesto, ya que estos son factores mucho más críticos para la aprobación de usuario.

Factores involucrados:

- Saturación de contenido por competencia entre plataformas.
- Priorización de volumen sobre calidad.
- Fragmentación de audiencias.
- Cambio en estándares de producción.

Sin embargo, podemos determinar que las producciones con muy altos presupuestos, tienen una tendencia a evitar los peores ratings.

## Segunda Hipótesis

Si bien el rating promedio anual de las series muestra una tendencia descendente en los últimos 20 años, esta tendencia no se traduce como un colapso total en la calidad de las series. Podemos decir que el rating promedio se ha mantenido en gran parte estable en las últimas dos décadas, rondando por los 7.0 puntos.

El promedio general de 7.06 sirve como referencia, si bien el rating final está por debajo de este promedio, la mayoría del periodo analizado se mantiene cerca de este número, demostrando una calidad promedio robusta a pesar de la tendencia.

Por otra parte, no debemos dejar de lado el aumento en la cantidad de series producidas por años, ya que esto tiene un efecto directo en la dilución del rating promedio:

- Cuando se producen anualmente aproximadamente 50 series, la competencia es intensa y los recursos se concentran en asegurar que cada proyecto sea un éxito crítico, elevando el promedio.
- Al producir más de 250 series por año, la necesidad de llenar catálogos y mantener a los suscriptores activos suele superar la prioridad de la excelencia crítica. Se producen más series de "relleno" o de nicho, que aunque no son malas, tienen ratings más cercanos a 6.8 - 7.0 y no llevan el promedio hacia arriba.

La tendencia a la baja en el rating promedio es el resultado natural de la tendencia al alza en la cantidad de series. La industria prioriza la cantidad para satisfacer la demanda del streaming, lo que inevitablemente arrastra ligeramente a la baja la métrica de calidad promedio.

## Tercera Hipótesis

La producción global está monopolizada por Estados Unidos, que por sí solo supera la producción combinada de todas las demás regiones analizadas.

Sin embargo, debemos admitir un sesgo geográfico muy significativo hacia producciones anglófonas, limitando la representación global del análisis.

América del Norte y Europa suman más del 80% de la producción total analizada, representando un sesgo geográfico masivo.

Factores involucrados:

- El dataset representa principalmente industria anglófona
- Hegemonía Cultural: dominio estadounidense en narrativas globales.
- Existe una difícil competencia para las industrias emergentes.

## Cuarta Hipótesis

El liderazgo en volumen de producción presenta correlación inversa con el liderazgo en calidad. Corea del Sur, Japón y Francia demuestran que la excelencia cualitativa se logra con producción selectiva y enfocada, no masiva.

A mayor volumen de producción, tendencia a menor calidad promedio.

Factores involucrados:

- Modelo estadounidense prioriza volumen, sacrificando calidad promedio
- Aún existen industrias cinematográficas que mantienen su excelencia con una producción moderada.
- Saturación de producción.

Sin embargo, cabe recalcar que Estados Unidos es el mayor mercado de entretenimiento del mundo por ingresos, pero su "rating promedio" puede variar significativamente según el tipo de contenido, la audiencia y las métricas de medición. Es posible que otros países, como los que tienen una fuerte industria televisiva local o con mayor consumo de medios per cápita, tengan promedios de rating más altos.

Además, debemos aclarar que el rating promedio, es una medida que puede variar entre países debido a factores, como la diversidad de la oferta de contenido, los hábitos de consumo cultural, la fragmentación de las audiencias, la disponibilidad de plataformas de transmisión y las diferencias en los métodos de medición.

## Recomendaciones Estratégicas y Prácticas

Basado en la validación de las hipótesis, se proponen las siguientes acciones para optimizar un sistema de recomendación:

1. Sistema de ponderación dinámica: priorizar el formato Serie, dada su mayor calidad promedio.
  - a. Priorizar la calidad en series: ante la tendencia a la baja, se recomienda que el sistema penalice las series más recientes que no cumplan con un umbral de rating alto.
2. Diversificación de Origen: fomentar y recomendar contenido de países con bajo volumen de producción, pero altos ratings de calidad, para evitar la fatiga por el contenido proveniente de los países tradicionales.

El análisis ofrece a la industria herramientas para:

- Orientar decisiones de inversión en producción audiovisual.
- Optimizar estrategias de contenido para plataformas de streaming.
- Identificar oportunidades de mercado en regiones subrepresentadas.
- Optimizar asignación de recursos presupuestarios.