



# DealData

Análisis exploratorio y técnicas estadísticas del  
DataSet comercio electrónico Amazon

**Materias:** Ciencia de Datos II y Estadística y Exploración de Datos **Grupo N° 6**

**Integrantes:** Melania Ligorria, Miguel Rojas, Carlos Direní, Nicolás Allende, Emanuel Guaraz, Lucas Ryser, Juan Clavijo y Guadalupe Mendoza

**Profesores:** Nahuel Pratta y Marcos Ugarte



# Información general y Limpieza

El DataFrame inicial contiene 1465 filas y 16 columnas, todas interpretadas como texto.

## Eliminación de nulos

Se eliminan celdas y valores nulos sin alterar el DataFrame original.

## Validación de duplicados

Se verifica la ausencia de valores duplicados en el dataset.

# Limpieza y preparación

Columnas como discounted\_price, actual\_price, rating y rating\_count se limpian de símbolos y se convierten a tipos numéricos.

# 1º Técnica: análisis de Correlación y Regresión Lineal

## Objetivo: relación entre calificaciones y cantidad

Estudiar la relación entre la cantidad de calificaciones (`rating_count`) y la calificación promedio (`rating`) de los productos. La hipótesis es que a mayor cantidad de calificaciones, el valor promedio debería estabilizarse o reflejarse mejor el valor promedio.

### Análisis Exploratorio Inicial

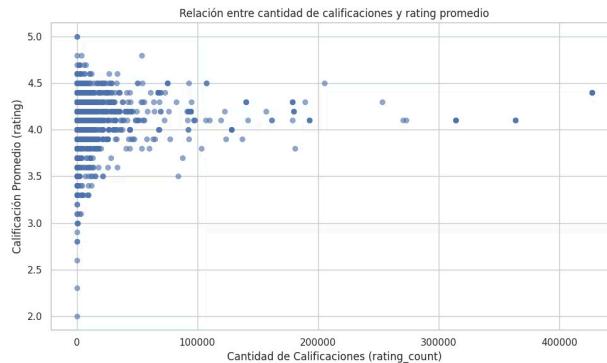
Conocer la cantidad de registros, promedios y distribución es crucial antes de calcular correlaciones.

### Conclusión Preliminar

El dataset contiene 1463 productos. Los `ratings` se concentran entre 4 y 4.3 ( $\sigma \approx 0.29$ ), mientras que `rating_count` varía enormemente ( $\sigma \approx 42.753$ ), anticipando outliers.

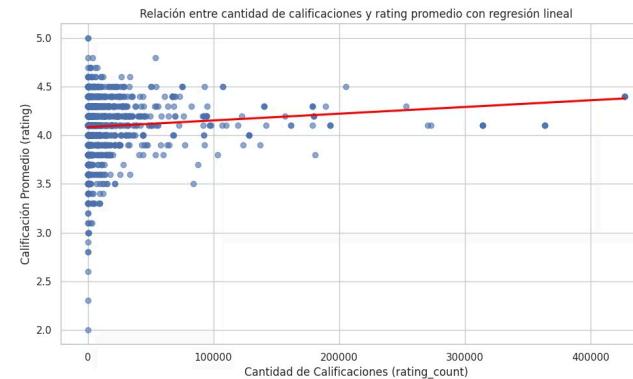
# Visualización de la relación

Un scatterplot revela que la mayoría de los productos se agrupan con bajo rating\_count y un rating entre 4 y 4.5.



## Dispersión de Puntos

La mayoría de los productos tienen menos de 20.000 calificaciones, con ratings concentrados.



## Pendiente de Regresión

La línea de regresión muestra una pendiente positiva muy leve, indicando una relación débil.

## Outliers

Productos con más de 100.000 calificaciones influyen, pero la relación general es débil.

**Conclusión:** La relación positiva es muy débil, sugiriendo la necesidad de coeficientes no paramétricos como Spearman o Kendall.

# Verificación de supuestos para Pearson

## a) Normalidad de las variables

Para aplicar el coeficiente de correlación de Pearson, es recomendable que las variables tengan una distribución aproximadamente normal. Evaluamos esto mediante histogramas, Q-Q plots y el test de Shapiro-Wilk.

### Interpretación anticipada

Si los p-valores < 0.05, se rechaza la normalidad → Pearson no es totalmente adecuado.

### Normalidad de las variables

#### Interpretación

Resultados del test de Shapiro-Wilk

**rating:** estadístico = 0.925, p-valor ≈ 2.24e-26

**rating\_count:** estadístico = 0.414, p-valor ≈ 4.34e-56

Los p-valores de Shapiro-Wilk (rating: 2.24e-26,

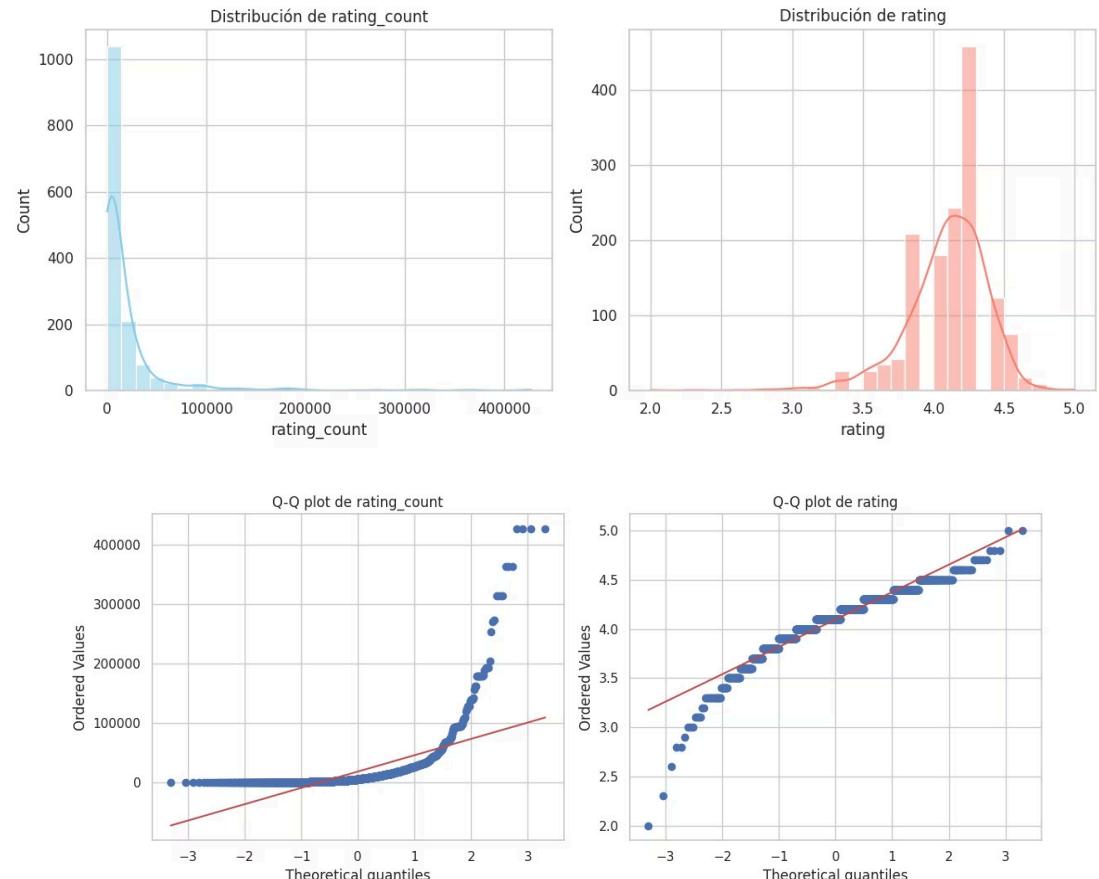
rating\_count: 4.34e-56) son < 0.05, rechazando la normalidad.

#### Interpretación

Ambos p-valores son mucho menores que 0.05, lo que indica que debemos rechazar la hipótesis de normalidad.

Los histogramas muestran que rating\_count tiene una distribución fuertemente sesgada hacia la derecha (muchos productos con pocas calificaciones y pocos productos con muchísimas calificaciones).

rating también presenta ligera desviación de la normalidad, concentrándose entre 4 y 4.5.



### Conclusión

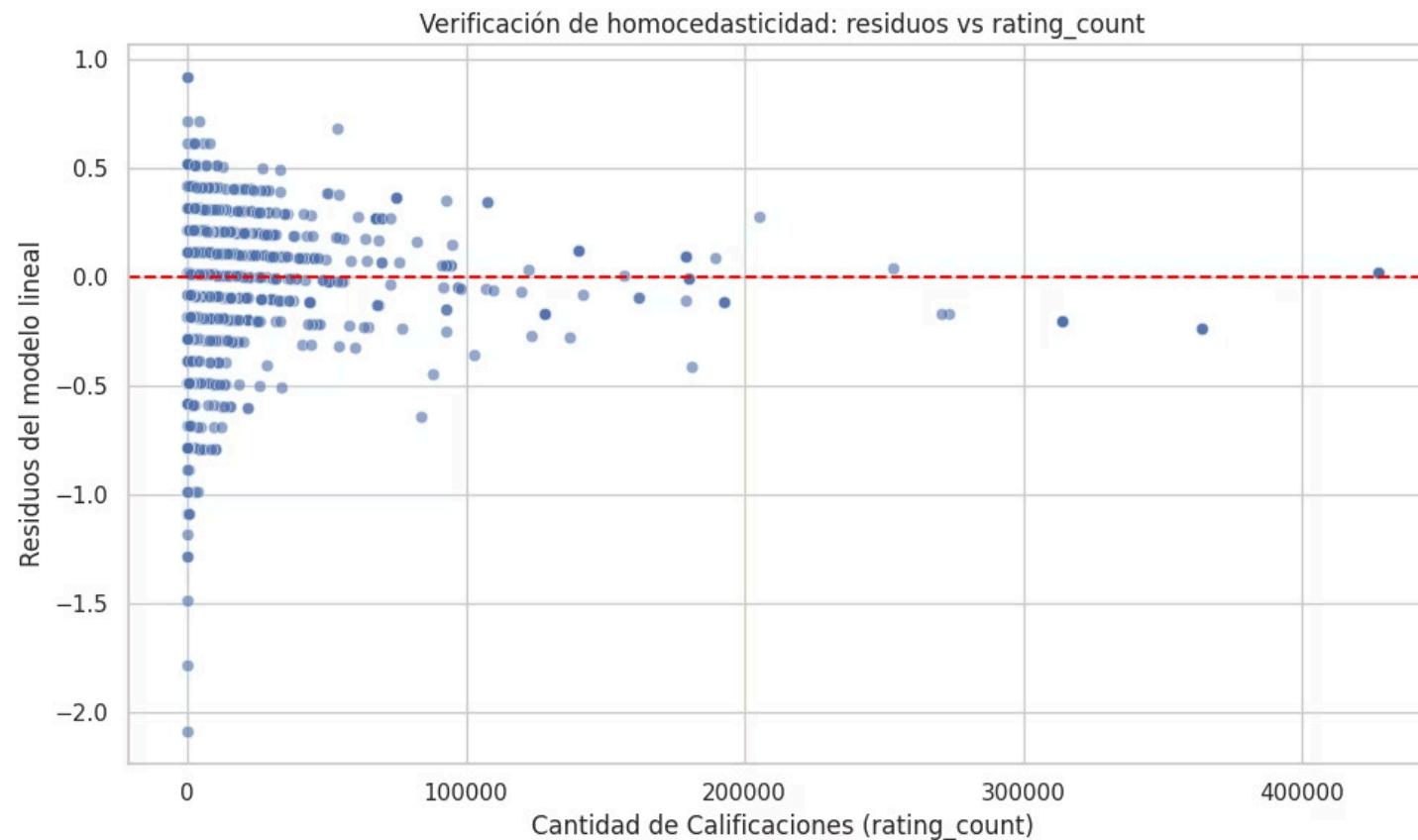
Los datos no cumplen el supuesto de normalidad requerido para Pearson. Por lo tanto, este coeficiente no es totalmente adecuado para describir la relación entre rating\_count y rating.

Se recomienda considerar coeficientes no paramétricos como Spearman o Kendall, que no dependen de la normalidad de las variables.

## b) Homocedasticidad

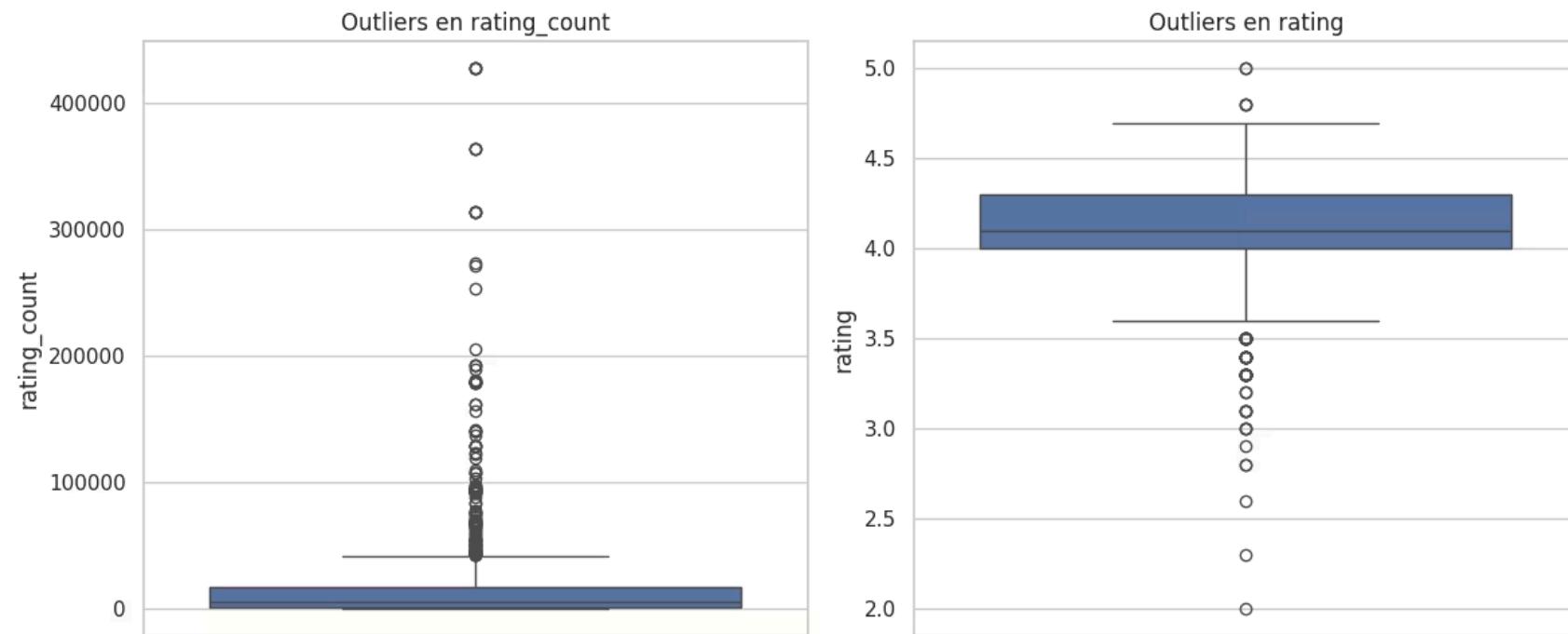
Aunque no hay un patrón de "embudo" evidente, la alta dispersión de `rating_count` sugiere que la homocedasticidad no se cumple perfectamente.

**Conclusión:** esto refuerza la recomendación de usar métodos no paramétricos.



## c) Outliers y transformación Logarítmica

El `rating_count` presenta una alta concentración de datos en valores bajos y outliers muy grandes, mientras que el `rating` se concentra entre 4.0 y 4.5.



### ⓘ Transformación Logarítmica

Para mitigar el efecto de los outliers en `rating_count`, se aplica una transformación logarítmica ( $\log(1 + \text{rating\_count})$ ).

Esto permite una interpretación más fiable de la relación, ya que la distribución sesgada puede afectar los coeficientes de correlación.

# Coeficientes de Correlación y Regresión Lineal

En esta sección ajustamos un **modelo de regresión lineal simple** para estudiar cómo la cantidad de calificaciones influye en el rating promedio (rating).

Usamos la variable transformada `rating_count_log` ( $\log(1 + \text{rating\_count})$ ) para reducir el efecto de outliers y estabilizar la relación.

## Coeficientes de Correlación

- **Pendiente**: **0.236** (p-value  $\approx 7e-20$ ) → positiva, débil, pero estadísticamente significativa.
- **Spearman**: **0.181** (p-value  $\approx 3.4e-12$ ) → positiva, débil, refleja la relación por rangos.
- **Kendall**: **0.129** (p-value  $\approx 2.8e-12$ ) → positiva, débil, consistente con Spearman.

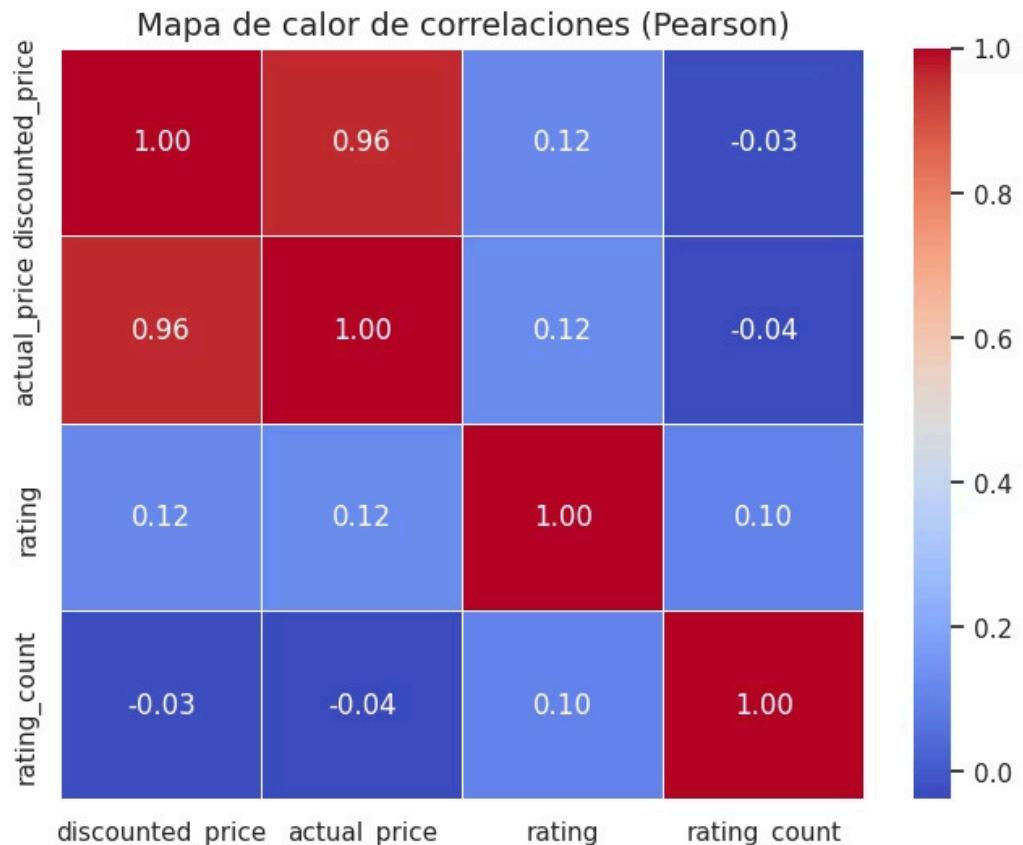
## Conclusión final del análisis de correlación

Relación positiva, débil, pero estadísticamente significativa. Los productos más populares tienden a tener un rating ligeramente mayor.

- La fuerza de la asociación es débil, lo que indica que la popularidad de un producto no garantiza un rating alto.
- Los coeficientes no paramétricos (Spearman y Kendall) son más robustos frente a la distribución sesgada y confirman la misma tendencia.
- La transformación logarítmica permitió reducir la influencia de los outliers extremos, haciendo los resultados más interpretables y confiables.

**Conclusión:** la relación es débil; la popularidad no garantiza un rating alto. Este hallazgo respalda la observación previa: aunque algunos productos muy populares tienen ratings altos, la mayoría de los ratings se mantiene en un rango estrecho ( $\approx 4 - 4.5$ ) independientemente de la cantidad de calificaciones.

# Mapa de calor de correlaciones



## Principales hallazgos:

- discounted\_price y actual\_price → **correlación muy alta (0.96).**
- rating vs rating\_count → **positiva pero débil ( $\approx 0.10 - 0.12$ ).**
- Precios y número de reseñas casi no influyen en el rating.

## Conclusión:

Los precios están fuertemente relacionados entre sí, pero **no explican de forma significativa la valoración de los usuarios.** La relación entre popularidad y rating existe, pero es débil.

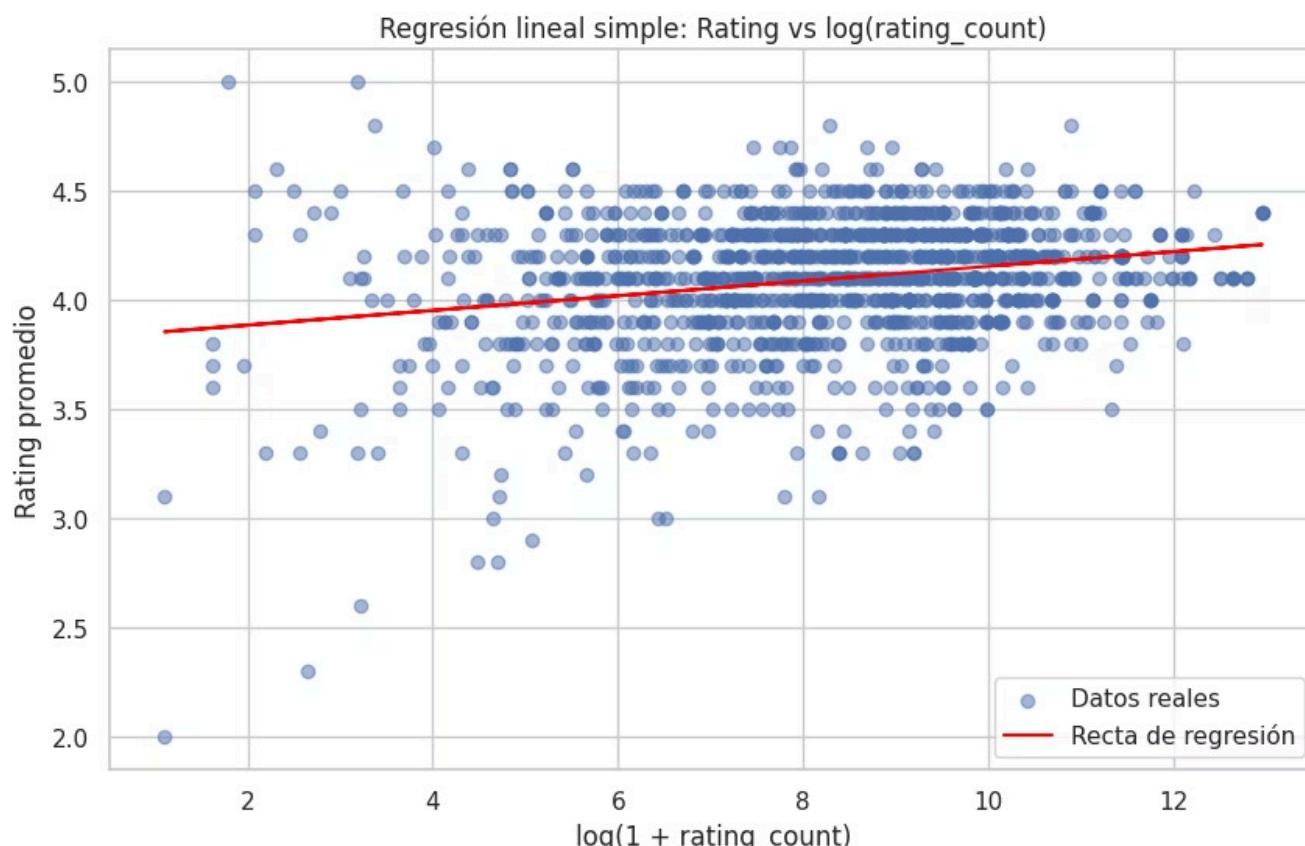
# Regresión lineal simple: prediciendo Rating a partir de la cantidad de calificaciones (log-transformada)

## Coeficientes del modelo

- **Pendiente (b):** 0.0336 → indica que, por cada incremento unitario en el logaritmo del número de calificaciones, el rating promedio aumenta en ~0.034 puntos.
- **Intercepto (a):** 3.8175 → valor de rating estimado cuando el número de calificaciones es cero (punto de inicio de la recta).
- **R<sup>2</sup> = 0.055** → el modelo explica solo el 5.5% de la variabilidad del rating.
- **RMSE = 0.281** → error medio moderado: el modelo predice con un desvío promedio de 0.28 respecto a los ratings reales.

## Conclusión final del análisis de regresión lineal

- Existe una **relación positiva pero débil** entre la cantidad de calificaciones y el rating promedio.
- El valor de **R<sup>2</sup> bajo** muestra que esta variable, por sí sola, **no predice bien el rating**.
- El modelo confirma la **tendencia general**: productos con más calificaciones tienden a tener ratings ligeramente más altos.
- Sin embargo, la **popularidad no asegura un buen rating**, ya que hay muchos otros factores que influyen.
- La transformación logarítmica de *rating\_count* permitió estabilizar la relación y reducir el efecto de los outliers.



# 2º Técnica: ANOVA rating por categoría de producto

ANOVA (Analysis of Variance) compara las medias de rating entre distintas categorías de productos para identificar diferencias significativas.

## Objetivo

Determinar si existen diferencias significativas en el rating promedio entre las categorías de productos.

## Interpretación anticipada

Si p-value < 0.05, hay diferencias significativas en el rating entre categorías.

### ☐ Resumen de categorías

El dataset tiene 211 categorías. Las 10 categorías con mayor rating promedio se sitúan entre 4.5 y 4.6.

#### 1. Revisamos cuántas categorías distintas hay y hacemos un pequeño resumen de sus ratings promedio

```
❶ # Número de categorías
num_categorías = df_anova['category'].nunique()
print(f"El dataset tiene {num_categorías} categorías distintas.")

❷ # Promedio y mediana de rating por categoría
resumen_categoria = df_anova.groupby('category')[['rating']].agg(['mean', 'median']).sort_values('mean', ascending=False)
print(resumen_categoria.head(10)) # Mostramos las 10 primeras
```

El dataset tiene 211 categorías distintas.

category	mean	median
Computers&Accessories Tablets	4.6	4.6
Computers&Accessories NetworkingDevices Network...	4.5	4.5
Electronics Cameras&Photography Accessories Film	4.5	4.5
Computers&Accessories Components Memory	4.5	4.5
Electronics HomeAudio MediaStreamingDevices Str...	4.5	4.5
OfficeProducts OfficeElectronics Calculators Basic	4.5	4.5
HomeImprovement Electrical CordManagement	4.5	4.5
Home&Kitchen Kitchen&HomeAppliances Coffee,Tea&...	4.5	4.5
Home&Kitchen Kitchen&HomeAppliances SmallKitche...	4.5	4.5
Electronics PowerAccessories SurgeProtectors	4.5	4.5

# Resultados del modelo ANOVA

El modelo ANOVA de una vía evalúa las diferencias en el rating entre categorías.

1

## F-statistic

**2.77**: Indica variación entre las medias de las categorías.

2

## p-value

**1.24e-27**: Mucho menor que 0.05, rechazando la hipótesis nula.

3

## Conclusión

Hay diferencias significativas en el rating promedio entre al menos algunas categorías.

Esto confirma que la categoría de producto influye en el rating promedio.

## 2. Ajustamos el modelo ANOVA (una vía) para evaluar si hay diferencias significativas en el rating entre categorías.

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
import seaborn as sns
import matplotlib.pyplot as plt

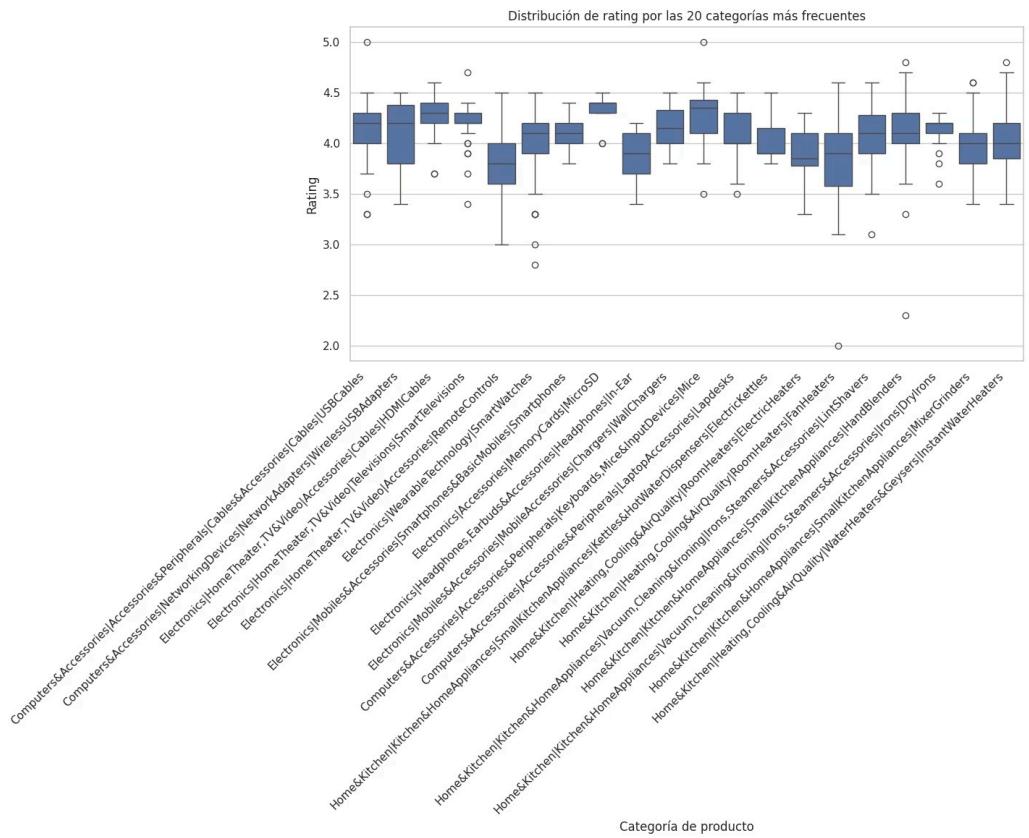
## Modelo ANOVA
model = ols('rating ~ C(category)', data=df_anova).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

	sum_sq	df	F	PR(>F)
C(category)	38.828028	210.0	2.766259	1.242121e-27
Residual	83.616212	1251.0	NaN	NaN

# Visualizamos los resultados con un boxplot de las 20 categorías más frecuentes para que sea más legible

```
# Selección de las 20 categorías más frecuentes
top_categories = df_anova['category'].value_counts().nlargest(20).index
df_top = df_anova[df_anova['category'].isin(top_categories)]  
  
# Resumen numérico para estas 20 categorías
resumen_top = df_top.groupby('category')['rating'].agg(['count', 'mean', 'median']).sort_values('mean', ascending=False)
print(resumen_top)  
  
# Boxplot
plt.figure(figsize=(12, 6))
sns.boxplot(x='category', y='rating', data=df_top)
plt.xticks(rotation=45, ha='right')
plt.title('Distribución de rating por las 20 categorías más frecuentes')
plt.xlabel('Categoría de producto')
plt.ylabel('Rating')
plt.show()
```

category	count	mean	median
Electronics Accessories MemoryCards MicroSD	13	4.330769	4.40
Computers&Accessories Accessories&Peripherals K...	24	4.287500	4.35
Electronics HomeTheater, TV&Video Accessories Ca...	24	4.254167	4.30
Electronics HomeTheater, TV&Video Televisions Sm...	63	4.289524	4.20
Computers&Accessories Accessories&Peripherals C...	231	4.153247	4.20
Electronics Mobiles&Accessories MobileAccessori...	16	4.143750	4.15
Home&Kitchen Kitchen&HomeAppliances Vacuum, Clea...	24	4.129167	4.20
Electronics Mobiles&Accessories Smartphones&Bas...	68	4.100000	4.10
Computers&Accessories NetworkingDevices Network...	18	4.094444	4.20
Computers&Accessories Accessories&Peripherals L...	14	4.064286	4.00
Home&Kitchen Kitchen&HomeAppliances SmallKitche...	19	4.057895	4.10
Home&Kitchen Heating, Cooling&AirQuality WaterHe...	23	4.052174	4.00
Home&Kitchen Kitchen&HomeAppliances Vacuum, Clea...	22	4.036364	4.10
Home&Kitchen Kitchen&HomeAppliances SmallKitche...	19	4.031579	3.90
Electronics WearableTechnology SmartWatches	76	4.025000	4.10
Home&Kitchen Kitchen&HomeAppliances SmallKitche...	27	4.011111	4.00
Electronics Headphones, Earbuds&Accessories Head...	52	3.898077	3.90
Home&Kitchen Heating, Cooling&AirQuality RoomHea...	28	3.890000	3.85
Home&Kitchen Heating, Cooling&AirQuality RoomHea...	20	3.810000	3.90
Electronics HomeTheater, TV&Video Accessories Re...	49	3.800000	3.80



## Conclusión final del análisis de ANOVA de una vía

### Interpretación gráfica y numérica:

- Las categorías con mayor promedio de rating incluyen Electronics|Accessories|MemoryCards|MicroSD (mean ≈ 4.33) y otras relacionadas con accesorios electrónicos.
- Algunas categorías tienen ratings más bajos, alrededor de 3.8–3.9, lo que indica que no todos los tipos de productos son igualmente bien valorados.
- El **F-statistic ≈ 2.77** y el **p-value ≈ 1.24e-27** mostraron que al menos una categoría difiere significativamente de las demás en promedio.
- El **boxplot** confirma visualmente la variabilidad entre categorías, mostrando medianas, rango intercuartílico y posibles outliers.
- Las diferencias en altura de las cajas y posición de medianas reflejan claramente la dispersión y diferencias significativas entre categorías.

## Conclusión final técnica ANOVA

El análisis ANOVA nos permite **detectar diferencias significativas en los ratings según la categoría**.

Combinando **resumen numérico + boxplot**, obtenemos tanto precisión como visualización clara.

Este análisis sirve para **identificar categorías que destacan o necesitan atención** y puede guiar decisiones comerciales o de producto.

# Conclusión final del análisis

El análisis de datos revela insights clave sobre la relación entre la popularidad del producto y su calificación, así como la influencia de la categoría.

## Correlación débil

Existe una relación positiva débil entre `rating_count` y `rating`, pero la popularidad no garantiza un rating alto.

## ANOVA significativo

Las categorías de productos tienen un impacto significativo en el rating promedio, con algunas destacando por sus altas calificaciones.

## Decisiones estratégicas

Estos hallazgos pueden guiar decisiones comerciales y de producto, identificando categorías a potenciar o mejorar.

**Próximos pasos:** considerar métodos robustos o no paramétricos para análisis futuros y combinar variables para mejorar modelos predictivos.