INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

Tecnicatura Superior en
Ciencia de Datos e Inteligencia Artificial

DealData
Exploración de datos de consumo digital

Proyecto ABP
Materias: Ciencia de Datos II y Estadísticas y Exploración de Datos II

Integrantes:

- Carlos Direni
- Nicolás Allende
- Melania Ligorria
- Guadalupe Mendoza
- Miguel Rojas Medina
- Emanuel Guaraz
- Lucas Ryser
- Juan Clavijo

Docentes Guía:

- Nahuel Pratta
- Marcos Ugarte

AÑO: 2025

Dirección General de EDUCACIÓN TÉCNICA Y FORMACIÓN PROFESIONAL
Ministerio de EDUCACIÓN
CÓRDOBA Seguimos haciendo

ISPC
INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

Evidencia nº 2

Integrantes del equipo:

Nombre de los integrantes	Usuario GitHub
Melania Ligorria	https://github.com/mel-ligorria
Miguel Rojas	https://github.com/Migueerm
Guadalupe Mendoza	https://github.com/Guadamendoza
Carlos Direni	https://github.com/Cdireni1
Nicolás Allende Olmedo	https://github.com/AllendeNicolas
Emanuel Guaraz	https://github.com/JEmanuelG
Lucas Ryser	https://github.com/lucasryser6
Juan Clavijo	https://github.com/juancla001

Link Repositorio Github: 🔗https://github.com/mel-ligorria/DealData

Acerca del Data Set:

Este conjunto de datos contiene información sobre más de 1000 calificaciones y reseñas de productos de Amazon, según detalles que aparecen en el sitio web oficial de la empresa.
El Dataset, posee un considerable número de registros, y una amplia cantidad de variables de interés, que podemos utilizar para realizar distintos tipos de análisis sobre los productos más solicitados de ésta tienda reconocida a nivel mundial.

Link Notebook:
https://colab.research.google.com/drive/1FUs-T7F-fdEd8b0vWGsv8vkRHw6KxL3h?usp=sharing

Link presentación: https://gamma.app/docs/DealData-a58pu2m8zqh7bne

Informe de Análisis de Datos

Dataset de productos de Amazon

Introducción al dataset

Este trabajo se basa en un conjunto de datos de **1465 registros y 16 variables**, que contienen información sobre precios, calificaciones y categorías de productos de Amazon.

- Todas las columnas aparecen inicialmente como tipo object (texto).
- Se identificó la necesidad de limpieza, ya que varias columnas contienen símbolos como ₹, %, comas que impiden su análisis numérico.

Análisis de correlación

Objetivo: Estudiar la relación entre cantidad de calificaciones (**rating_count**) y la calificación promedio (**rating**).

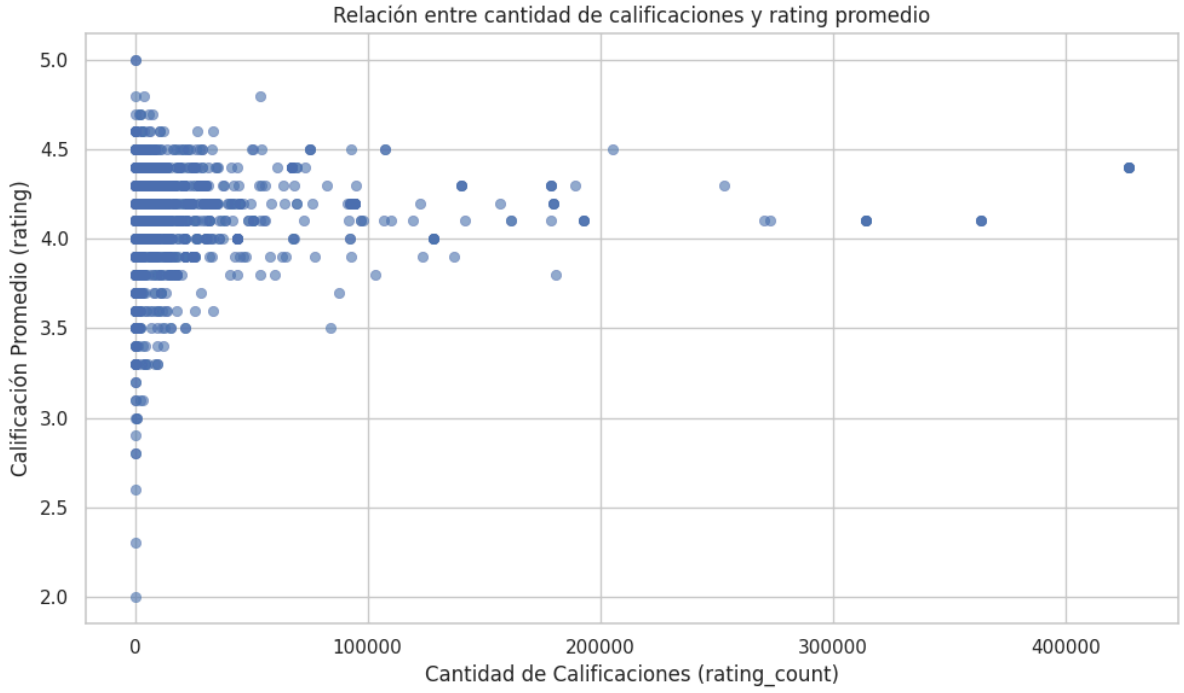
1. Exploración inicial

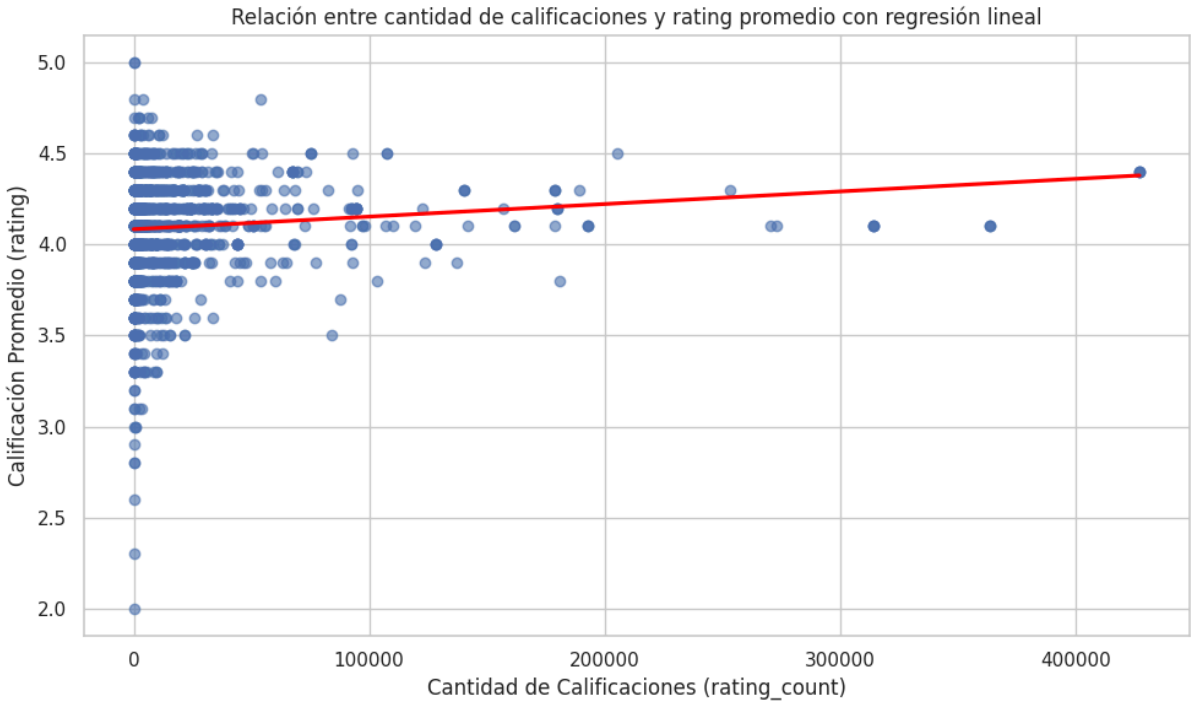
El dataset contiene **1463 productos válidos** tras la limpieza.

rating: Está concentrado entre 4.0 y 4.3 con una variabilidad de 0.29 ($\sigma \approx 0.29$).
rating_count: Con valores desde 2 hasta más de 400.000 reseñas y una variabilidad de 42.753 ($\sigma \approx 42.753$), se anticipa la presencia de **outliers**.

Esta exploración inicial confirmó que los datos no siguen una distribución normal y que la variable rating_count es altamente sesgada.

2. Visualización de la relación





Interpretación de los gráficos

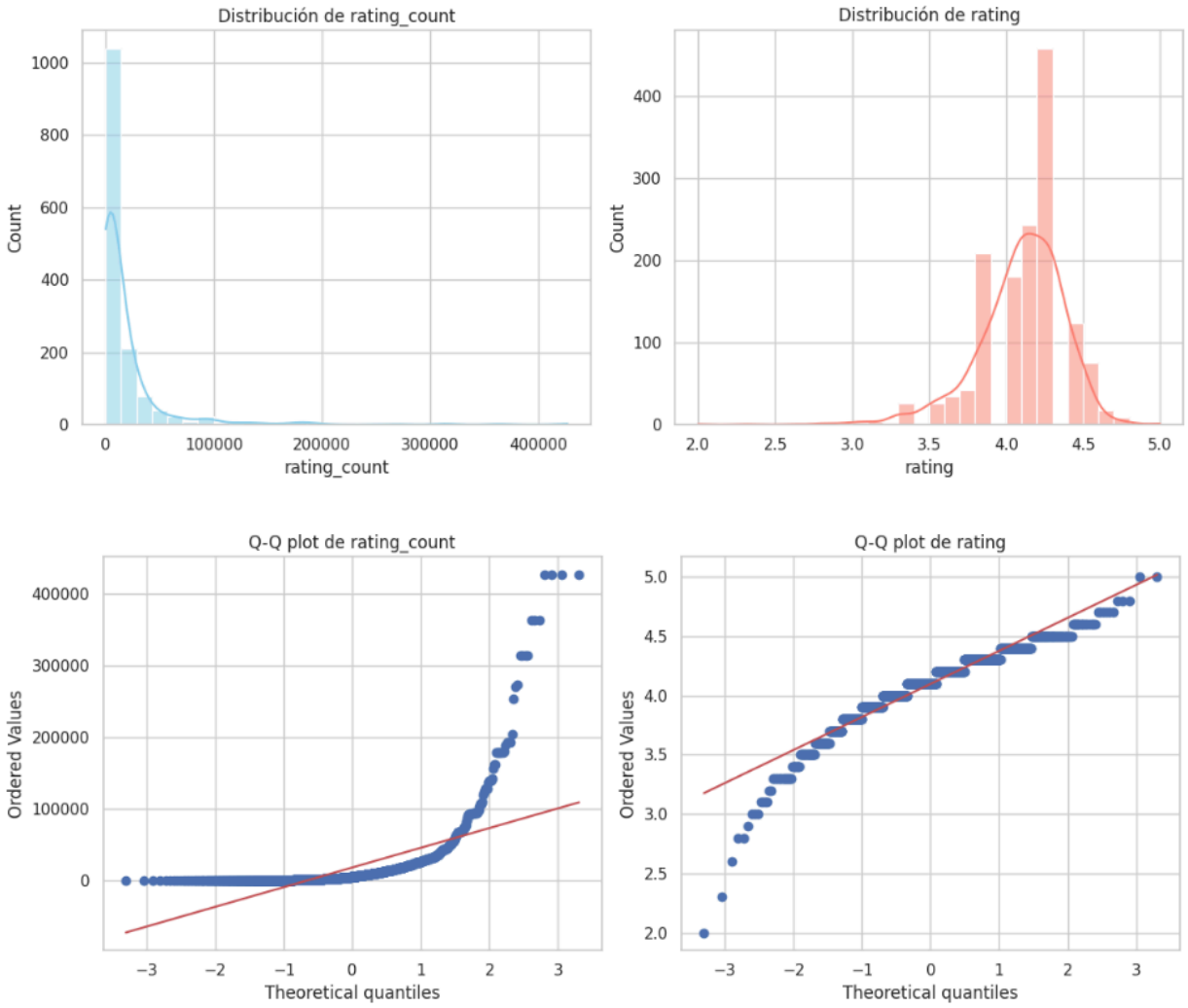
- La mayoría de los productos se agrupan con rating_count bajo y un rating concentrado entre 4 y 4.5.
- **La línea muestra una pendiente positiva muy leve**.
-Aunque la relación es positiva, la dispersión es muy grande y la pendiente casi horizontal. Eso significa que la **fuerza de la relación lineal es débil**.

Conclusión

El gráfico muestra que a mayor cantidad de calificaciones (**rating_count**), la calificación promedio (**rating**) tiende a aumentar levemente, pero la relación es muy débil.

Esto anticipa que el coeficiente de correlación de Pearson será bajo, y sugiere que es mejor aplicar coeficientes no paramétricos como Spearman o Kendall, más robustos frente a outliers y relaciones no lineales.

3. Verificación de supuestos para Pearson



3.1 Normalidad de las variables

Interpretación

Resultados del test de Shapiro-Wilk

- **rating**

  Estadístico = 0.925, p-valor ≈ 2.24e-26

- **rating_acount:**

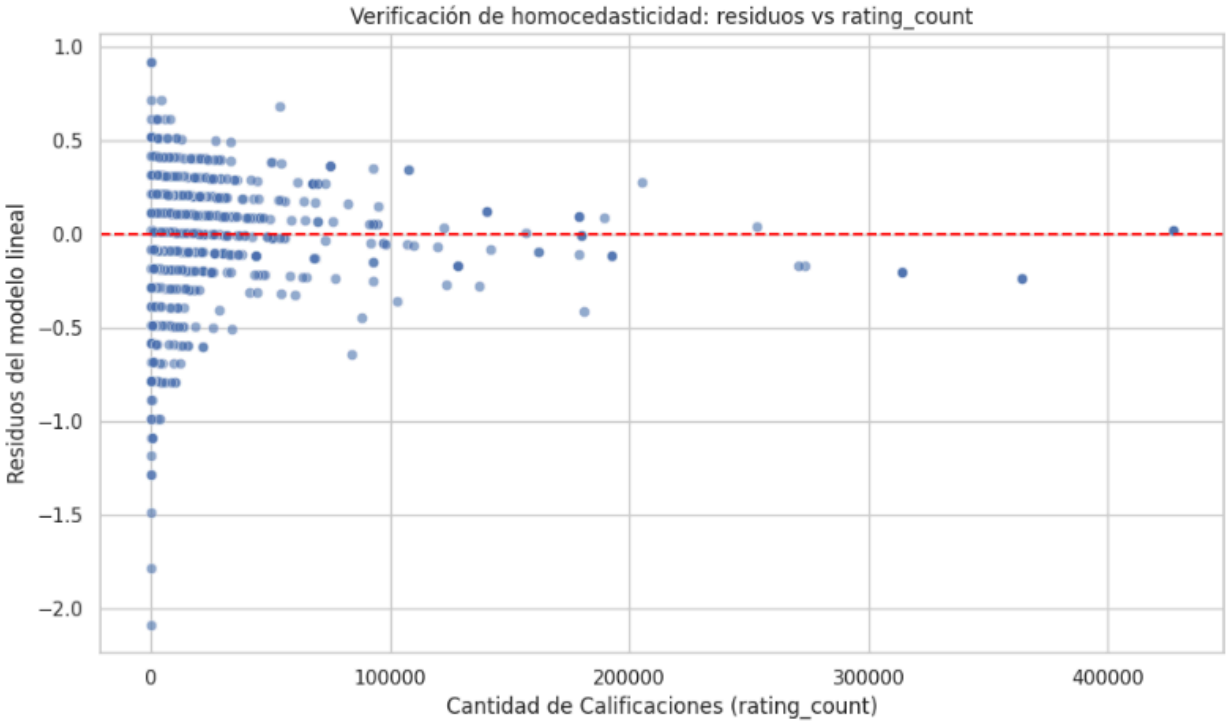
  Estadístico = 0.414, p-valor ≈ 4.34e-56

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.
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.

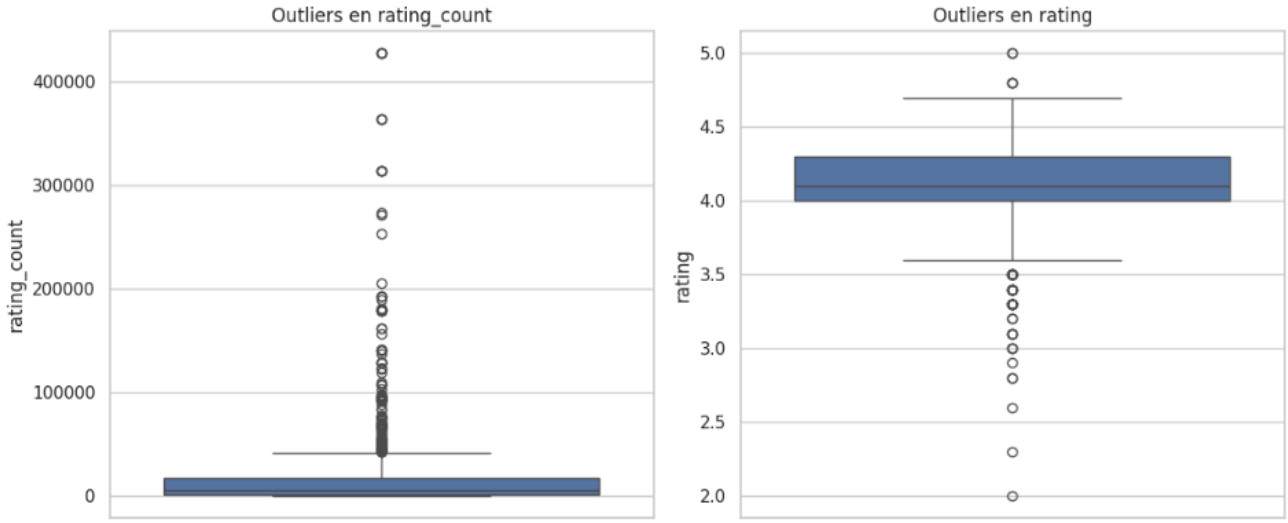
3.2 Homocedasticidad



Conclusión
El gráfico no muestra patrón de embudo evidente, pero dado que **rating_count** tiene muchos outliers y alta dispersión, la homocedasticidad no se cumple perfectamente.

Esto refuerza que el uso de Pearson puede no ser confiable y que métodos no paramétricos (Spearman/Kendall) son más adecuados.

3.3 Outliers



Interpretación de los gráficos

Rating_count (izquierda)
- Se observa una alta concentración de datos con valores bajos
- Existen outliers muy grandes
Esto refleja una distribución altamente sesgada a la derecha

Rating (derecha)
- La distribución de calificaciones está concentrada entre 4.0 y 4.5, con la mediana cercana a 4.1.
- Existen algunos outliers hacia abajo y algunos hacia arriba.

Esto muestra que en general los usuarios tienden a calificar alto, lo que también es común en este tipo de datasets.

Conclusiones preliminares en relación con la hipótesis

rating_count tiene una gran variabilidad y no está normalmente distribuido, lo que habrá que tener en cuenta al calcular correlaciones (posiblemente aplicar una transformación logarítmica).

rating es más estable y muestra menos dispersión, la mayoría de productos tienen calificaciones promedio bastante altas.

Esto respalda la hipótesis, los productos con pocas calificaciones pueden mostrar valores extremos (outliers en el rating).
A medida que el número de calificaciones aumenta, el promedio tiende a concentrarse en el rango 4.0–4.5, lo que sugiere mayor estabilidad estadística.

En nuestro dataset, **rating_count** tiene una distribución altamente sesgada a la derecha. La mayoría de los productos tienen pocas calificaciones, mientras que algunos tienen cientos de miles. Esto puede afectar los coeficientes de correlación, especialmente Pearson.
Para mitigar el efecto de los outliers y poder interpretar mejor la relación, **aplicamos una transformación logarítmica.**

4. Cálculo de coeficientes de correlación

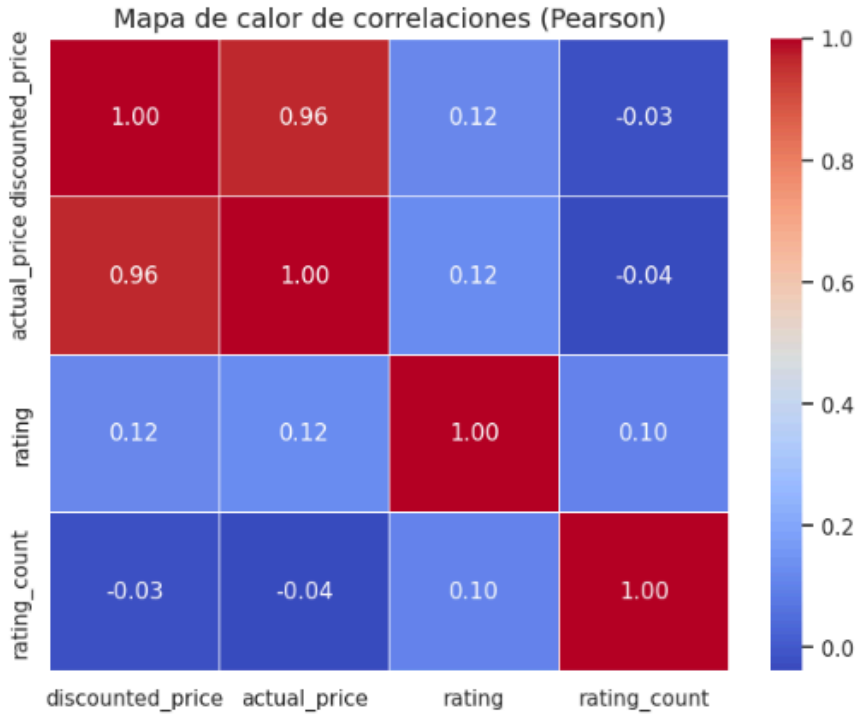
Conclusión final del análisis de correlación

Resultados obtenidos:
- Pearson: 0.236 (p-value ≈ 7e-20) → positiva, débil, pero estadísticamente significativa.
- Spearman: 0.181 (p-value ≈ 3.4e-12) → positiva, débil, refleja la relación por rangos.
- Kendall: 0.129 (p-value ≈ 2.8e-12) → positiva, débil, consistente con Spearman.

Interpretación:
- Existe una relación positiva entre la cantidad de calificaciones y el rating promedio.
- 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.

Conclusión práctica:
La relación existe, pero es débil.
Para análisis posteriores o predicciones, es recomendable considerar métodos robustos o no paramétricos si se trabaja con **rating_count** sin transformar.
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 (≈ 4 – 4.5) independientemente de la cantidad de calificaciones.

5. Mapa de calor de correlaciones



Interpretación del mapa de calor de correlaciones

Se observa una correlación muy alta y positiva (0.96) entre **discounted_price** y **actual_price**.

La relación entre **rating** y **rating_count** es positiva pero débil (≈0.10–0.12).

No hay una relación fuerte entre las variables de precio (**discounted_price, actual_price**) y el **rating**. Los coeficientes (≈0.12) muestran que el precio no determina significativamente la calificación.

Entre **rating_count** y precios la relación es prácticamente nula (−0.03 a −0.04), lo que significa que la cantidad de reseñas no depende del valor monetario del producto.

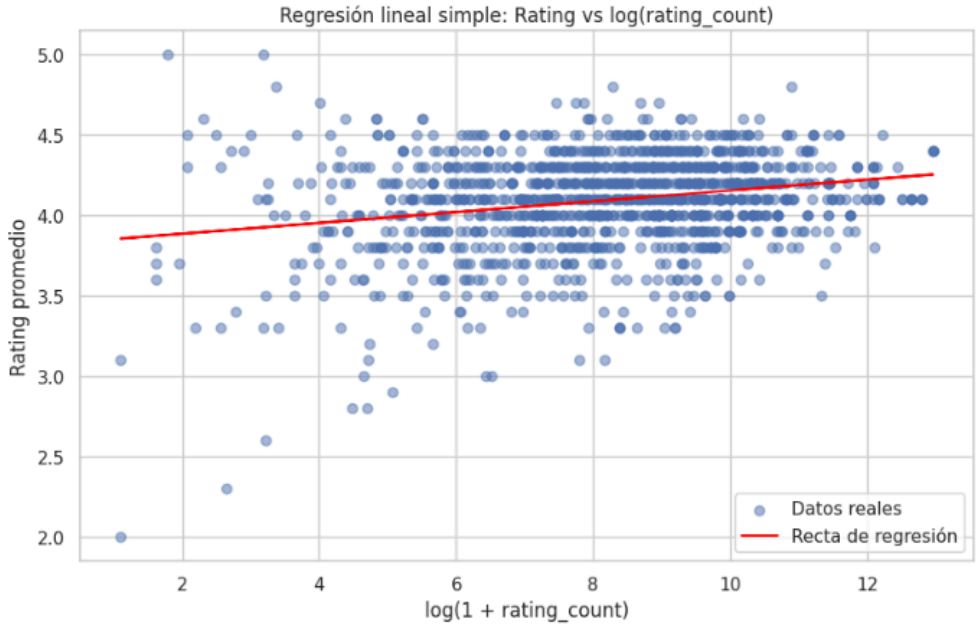
Conclusión

El mapa de calor confirma lo que vimos con los coeficientes de correlación individuales:

- Existe una relación lineal muy fuerte entre precios, pero eso no influye directamente en la percepción de calidad de los usuarios (rating).
- La relación entre popularidad (cantidad de calificaciones) y rating es positiva pero débil, lo que implica que tener más reseñas no garantiza una mejor valoración.

En general, los precios y el número de reseñas no son predictores fuertes del rating, lo cual justifica aplicar técnicas adicionales como regresión lineal y ANOVA para analizar mejor los factores que pueden incidir en la valoración de los productos.

Regresión Lineal Simple: prediciendo Rating a partir de la Cantidad de Calificaciones (log-transformada)



Interpretación de los datos

Pendiente (b = 0.0336): Efecto positivo débil.

Intercepto (a = 3.8175): Rating estimado con pocas reseñas.

R^2 = 0.055: El modelo explica solo el 5.5% de la variabilidad del rating.

RMSE = 0.281: Error moderado.

Conclusión final del análisis de regresión lineal simple

La regresión lineal muestra una **relación positiva débil**: a mayor cantidad de calificaciones, el rating promedio tiende a subir ligeramente.

El R^2 bajo confirma que **esta variable por sí sola no predice perfectamente** el rating.

Este análisis sirve para **visualizar tendencias generales** y puede combinarse con otras variables para mejorar el modelo predictivo.

ANOVA (una vía): rating vs categoría de producto

1. Revisamos cuántas categorías distintas hay y sus ratings promedio
Interpretación y conclusión

- El dataset cuenta con 211 categorías, lo que indica una gran diversidad de productos.
- Las 10 categorías con mayor rating promedio tienen ratings entre 4.5 y 4.6, mostrando que ciertos tipos de productos tienden a recibir calificaciones más altas.
- La mediana es muy cercana al promedio, lo que indica que los ratings están relativamente concentrados y no hay demasiada dispersión en estas categorías top.

Este análisis preliminar permite identificar categorías destacadas y proporciona contexto antes de aplicar ANOVA, mostrando dónde podrían existir diferencias significativas entre grupos.

2. Ajustamos el modelo ANOVA (una vía)

Evaluar si existen diferencias significativas en el rating entre categorías.

Resultados clave:

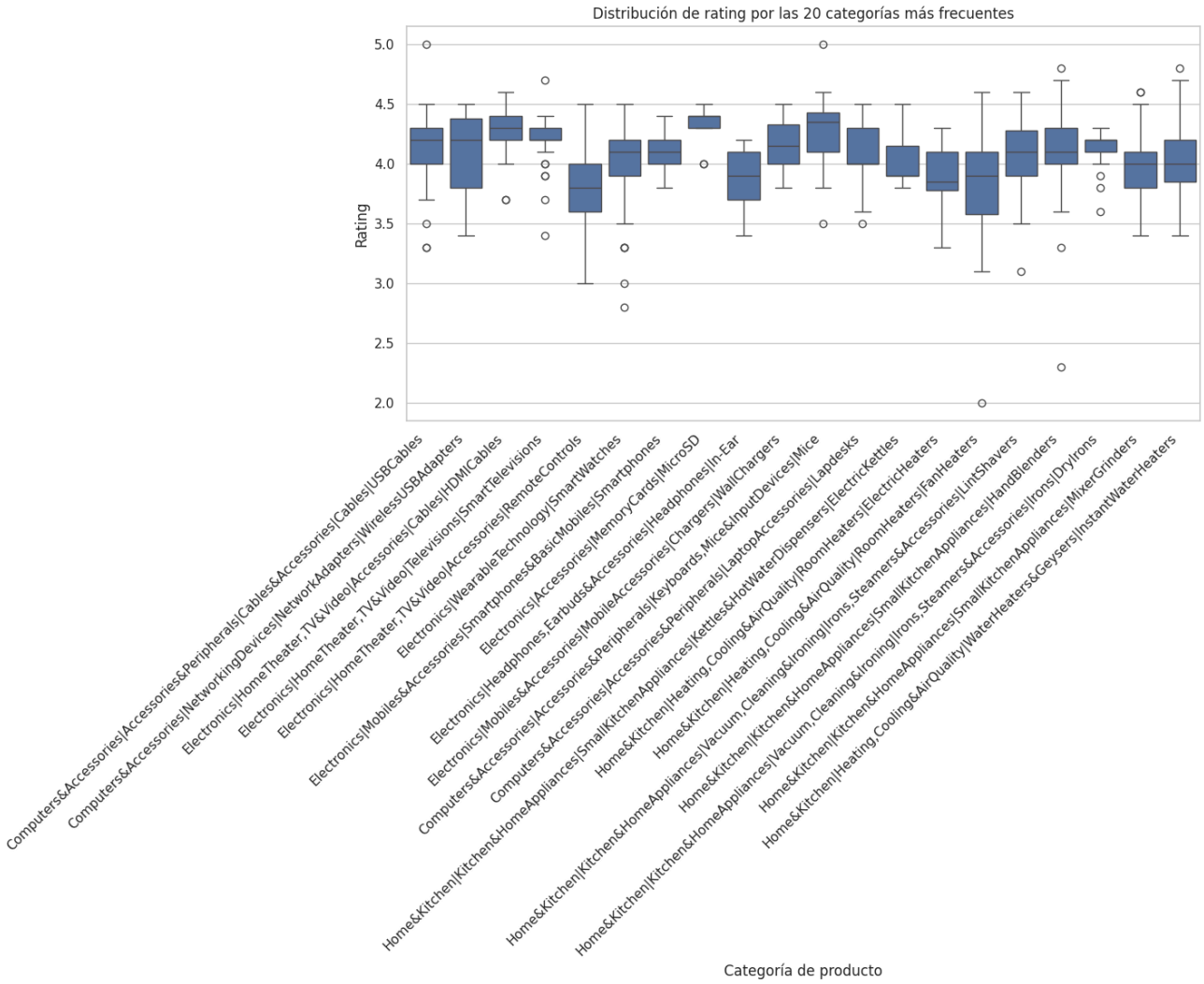
- **F-statistic ≈ 2.77:** indica que existe variación entre las medias de las categorías; no todas las medias son iguales.
- **p-value ≈ 1.24 e-27:** mucho menor que 0.05, por lo que rechazamos la hipótesis nula de igualdad de medias entre todas las categorías.
- **Residual:** representa la variación dentro de cada categoría, que no se explica por el factor categoría.

Conclusión:

Hay diferencias significativas en el rating promedio entre al menos algunas categorías.
Aunque tenemos 211 categorías, este análisis confirma que la categoría de producto influye en el rating promedio.

3. Visualizamos los resultados

Boxplot de las 20 categorías más frecuentes para que sea más legible



Conclusión final del análisis de ANOVA de una vía

Interpretación gráfica y numérica:

- Las categorías con mayor promedio de rating incluyen: Electronics, Accessories, Memory Cards, 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 general del trabajo

1. La **limpieza de datos** fue clave para transformar variables y permitir el análisis correcto.
2. El **análisis de correlación** muestra que la relación entre popularidad (reseñas) y calificación es positiva pero débil.
3. La **regresión lineal simple** confirma que el rating no depende fuertemente del número de reseñas.
4. El **ANOVA** evidencia que la categoría del producto sí influye en las calificaciones, siendo un factor más determinante que el precio o la cantidad de reseñas.

En conjunto, este trabajo demuestra la utilidad de combinar **técnicas estadísticas (correlación, regresión, ANOVA)** con **visualizaciones gráficas** para interpretar datos complejos y guiar la toma de decisiones.