DEALDATA
ANÁLISIS EXPLORATORIO Y TÉCNICAS ESTADÍSTICAS DEL DATASET COMERCIO ELECTRÓNICO AMAZON

TIPO DE PROYECTO:
Tecnológico y de investigación.

ESPACIO CURRICULAR/MÓDULO:
Ciencia de Datos II
Estadística y Exploración de Datos II

EJES | UNIDADES CONCEPTUALES:
- **Estadística Descriptiva y Exploratoria (EDA):** limpieza de datos, transformación logarítmica para mitigar sesgo en precios y conteos, y análisis de distribuciones.
- **Inferencia estadística y relaciones:** análisis de correlación (Pearson/Spearman), ANOVA (Análisis de Varianza) y pruebas post-hoc (Test de Tukey) para comparación de medias entre múltiples grupos categóricos.
- **Modelado predictivo y clasificación:** regresión logística binaria (Log-Odds) para predecir una variable dicotómica (is_highly_rated), incluyendo diagnóstico de multicolinealidad (VIF) y evaluación de métricas de desempeño (AUC, Accuracy).

PROBLEMÁTICA | NECESIDAD | CASO:
- **Problemática:** identificar con precisión los **factores de éxito** que determinan si un producto de Amazon es "altamente calificado" (rating ⩾ 4.0). La pregunta clave es: ¿Qué es más importante para una alta calificación, la popularidad (rating_count) o la categoría y el precio?
- **Necesidad:** ofrecer información accionable a nivel comercial y de producto, yendo más allá de la correlación superficial para cuantificar la influencia de cada variable (precio, descuento, reseñas y categoría).
- **Caso:** análisis de un dataset de productos de Amazon (1465 registros) con el objetivo de optimizar la estrategia de calidad y pricing (estrategia de precios).

OBJETIVO GENERAL:
Evaluación y cuantificación la influencia de variables de precio, descuento, popularidad y categoría en la calificación promedio (rating) de productos de Amazon.

OBJETIVOS ESPECÍFICOS:
1. Analizar la fuerza y dirección de la relación entre popularidad (rating_count) y calidad percibida (rating).

2.Determinar si existen diferencias significativas en el rating promedio según la categoría del producto (ANOVA).

3. Construir un modelo de Regresión Logística para predecir la probabilidad de que un producto sea "altamente calificado" (rating ⩾ 4.0).

FUNDAMENTACIÓN | HIPÓTESIS:
- **Fundamentación:** es fundamental aplicar técnicas robustas (ANOVA + Tukey y Regresión Logística) para aislar la influencia de la categoría, un factor categórico que la simple correlación ignora. Se requiere una herramienta predictiva interpretable para cuantificar riesgos y oportunidades.
- **Hipótesis central:** la categoría de producto será el factor más determinante y significativo del rating final, superando la influencia de la popularidad (rating_count) y el precio en la probabilidad de ser un producto "altamente calificado".

ACCIONES | RECURSOS | TIEMPO:
- **Acciones clave:**
  1. Limpieza de datos y transformación logarítmica de variables sesgadas.
  2. Análisis de correlación (Pearson/Spearman).
  3. Ajuste del modelo ANOVA con Test Post-Hoc de Tukey (HSD).
  4. Modelado de regresión logística para predicción binaria (rating ⩾ 4.0).
- **Recursos:** Python, librerías principales: Pandas, Scikit-learn, Statsmodels, Seaborn/Matplotlib.
- **Tiempo**: 10 Semanas / 1 Cuatrimestre.

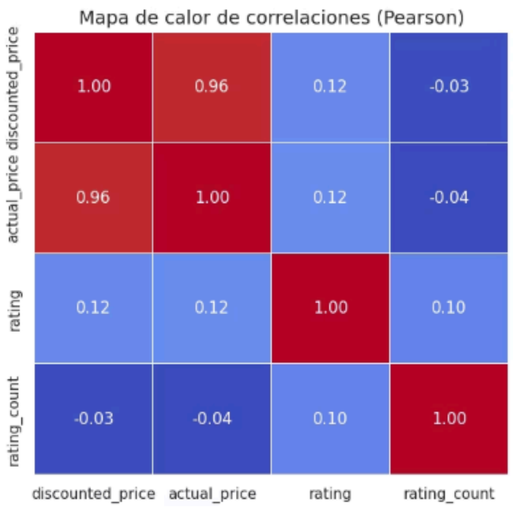
PRODUCTO FINAL | CONCLUSIONES | RESULTADOS ESPERADOS:
Producto final:

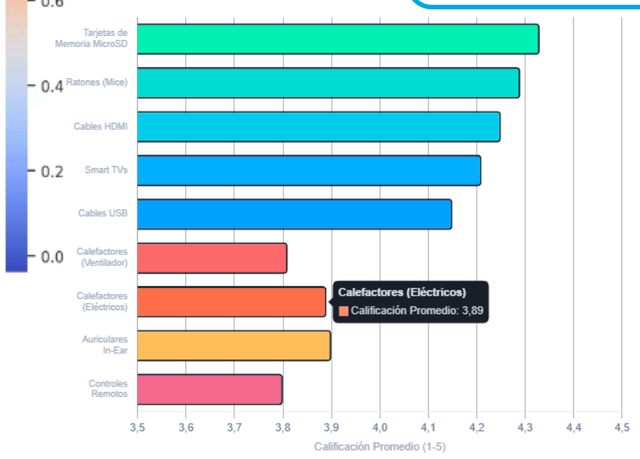
Un modelo interpretable de Regresión Logística que predice la probabilidad de rating alto, y un análisis inferencial sólido que valida la influencia de la categoría.

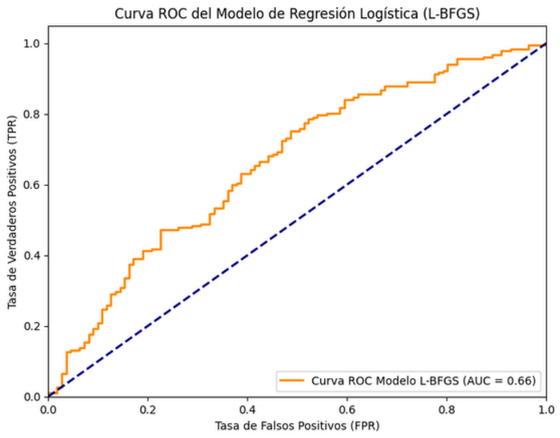
Conclusiones:
- **Correlación:** la popularidad es un **predictor débil (r ≈ 0.236).**
- **Inferencia (ANOVA/Tukey):** la **categoría** es altamente significativa ($p < 0.05$), con diferencias concretas entre pares de categorías (e.g., MicroSD vs. Controles Remotos).
- **Regresión Logística:** el modelo es estadísticamente significativo ($p < 0.05$) AUC = 0.66. Revela que:
  - La probabilidad de rating alto aumenta con las reseñas y precios altos.
  - **Altos descuentos reducen la probabilidad** de obtener un rating alto (coeficiente negativo).

Resultados esperados:

La confirmación de la hipótesis central: el tipo de producto (categoría) es el principal factor de calidad percibida, sugiriendo que la estrategia debe enfocarse en la calidad intrínseca sobre la mera popularidad.







INTEGRANTES: Melania Ligorria, Carlos Direni, Miguel Rojas, Nicolás Allende, Emmanuel Guaraz, Guadalupe Mendoza y Juan Clavijo

DOCENTE: Nahuel Pratta y Marcos Ugarte

TECNICATURA: Tecnicatura Superior en Ciencia de Datos e Intelegencia Artificial

COHORTE: 2024