# Data Card del Dataset de Google Play Apps (Antes de preparación de los datos)

Kristian Restrepo, Sebastian Restrepo

2 de octubre de 2025

## 1. Descripción General

- **Nombre del dataset:** Google Play Store Apps.
- **Propósito:** Exploración y análisis de características de apps móviles para comprender distribución de ratings, precios, reseñas y descargas.
- **Fuente:** Kaggle.
- **Periodo:** Dataset estático sin actualización en tiempo real (snapshot único, sin series temporales).
- **Nota:** El dataset se emplea con fines exploratorios y de prácticas de ML, no para producción.

## 2. Composición

- **Registros:** 10,841 apps.
- **Variables:** 14 columnas principales (originalmente 13, 1 agregada para entendimiento de los datos).
- **Tipos:** texto, categóricas, numéricas.
- Se detectaron **duplicados** que afectan la calidad y deben ser depurados en etapas posteriores.

### Principales variables

| Variable | Tipo | Descripción | Valores Faltantes |
|---|---|---|---|
| App | Categórica | Nombre de la aplicación | 0 |
| Category | Categórica | Categoría oficial | 0 |
| Rating | Numérica | Calificación (1–5) | 1,474 (13.6 %) |
| Reviews | Numérica | Número de reseñas | 0 |
| Size | Numérica (MB) | Tamaño de la app | 1,696 (15.6 %) |
| Installs | Categórica | Descargas (rangos) | 0 |
| Type | Categórica | Free/Paid | 1 |
| Price | Numérica (USD) | Precio de la app | 1 |
| Content Rating | Categórica | Clasificación de edad | 1 |
| Genres | Categórica | Géneros | 0 |
| Last Updated | Texto | Fecha última actualización | 0 |
| Current Ver | Texto | Versión actual | 8 (0.07 %) |
| Android Ver | Texto | Versión mínima requerida | 3 (0.03 %) |
| Installs Numeric | Numérica | Descargas convertidas a número | 1 |

# 3. Preprocesamiento mínimo para mejor comprensión de los datos

- Conversión de variables `object` a numéricas (`Size`, `Price`, `Installs Numeric`).
- Creación de `Installs Numeric` a partir de rangos categóricos.
- Detección y conteo de **outliers**, aunque aún no se han eliminado.
- No se realizó imputación de valores faltantes en esta etapa.

# 4. Valores Faltantes

- **Críticos:** Rating (13.6 %), Size (15.6 %).
- **Menores:** Current Ver (0.07 %), Android Ver (0.03 %), Type, Price, Content Rating, Installs Numeric (cada uno con 1 caso).
- Estos faltantes pueden introducir **sesgo** si no se imputan correctamente, en especial en Rating y Size.

# 5. Distribución y Estadísticos

- **Rating:** media 4.19, mediana 4.3; ligera asimetría, presencia de outliers.
- **Reviews:** media 444,111, mediana 2,094; **fuertemente sesgada**, con valores extremos muy altos.
- **Size:** media 21.5 MB, mediana 13 MB; distribución sesgada hacia apps pequeñas.

- **Price:** mayoría de apps gratuitas, precios extremos hasta 400 USD; **distribución muy desbalanceada**.
- **Installs Numeric:** media 15.4M, mediana 100K; distribución **muy asimétrica y con cola larga**.

# 6. Variables Categóricas

- **Content Rating:** Everyone (79.2 %), Teen (11.6 %), Mature 17+ (4.9 %), Everyone 10+ (4.2 %). Fuerte **desbalance** hacia "Everyone".
- **Type:** Free (93.1 %), Paid (6.9 %). Desbalance hacia apps gratuitas.
- **Installs (rangos):** principales: 1M+ (16.8 %), 10M+ (13.4 %), 100K+ (12.3 %).

# 7. Correlaciones

- En general, las correlaciones lineales (Pearson) son débiles, mientras que las correlaciones monótonas (Spearman) muestran relaciones más fuertes, lo que indica que varias variables se relacionan de forma no estrictamente lineal.
- La relación más importante se da entre **Installs Numeric** y **Reviews**, con Pearson 0.64 (moderada-fuerte) y Spearman 0.97 (muy fuerte). Esto confirma que el número de reseñas crece junto con las instalaciones, reflejando popularidad.
- **Size** mantiene una correlación baja con **Reviews** (0.24 Pearson, 0.37 Spearman) y con **Installs Numeric** (0.35 Spearman), lo que sugiere una influencia secundaria del tamaño de la app en la adopción.
- **Price** muestra correlaciones negativas con popularidad: -0.17 con **Reviews** y -0.24 con **Installs Numeric**. Esto indica que las aplicaciones de pago tienden a ser menos descargadas y recibir menos reseñas.
- **Rating** permanece prácticamente independiente de las demás variables, sin correlaciones relevantes, lo que limita su uso como variable objetivo directamente explicada por las demás.

# 8. Outliers y Anomalías

- **Métodos usados:** IQR, Z-Score, Isolation Forest.
- **Totales:**
  - IQR: 3,489 (32.2 %).
  - Z-Score: 654 (6 %).
  - Isolation Forest: 1,084 (10 %).
- **Más afectados:** Reviews (1,925), Installs Numeric (828), Price (800).

- Nota: los outliers corresponden en gran parte a apps muy exitosas (descargas masivas) o precios atípicos. Se recomienda **no eliminarlos directamente**, ya que pueden contener información valiosa.

# 9. Limitaciones

- Sesgo hacia apps gratuitas y clasificación "Everyone".
- Alta proporción de outliers que pueden distorsionar modelos predictivos.
- Valores faltantes significativos en Rating y Size.
- Dataset sin información temporal longitudinal (descargas acumuladas, no series de tiempo).
- Gran cantidad de datos duplicados presentes.
- Distribuciones muy asimétricas que dificultan modelos lineales sin transformaciones.

# 10. Usos Recomendados

- Análisis de mercado de apps móviles.
- Modelos de predicción de rating o descargas (con imputación adecuada).
- Benchmark de **preprocesamiento, imputación y manejo de outliers**.
- Ejemplo didáctico para **EDA y prácticas de ML inicial**.
- Estudios sobre impacto de precio y reseñas en popularidad.

# 11. Consideraciones Éticas

- Datos públicos, sin información personal sensible.
- Dataset no garantiza representatividad global del mercado.
- Uso recomendado solo para investigación y enseñanza.