Data Card del Dataset de Google Play Apps (Después de preparación de los datos con transformaciones básicas)

Kristian Restrepo, Sebastian Restrepo

2 de octubre de 2025

1. Descripción General

- **Nombre del dataset:** Google Play Store Apps (Procesado).
- **Propósito:** Dataset con transformaciones básicas para un mejor manejo en un posterior paso de modelado.
- **Fuente:** Kaggle
- **Periodo:** Dataset estático con transformaciones aplicadas en octubre 2025.
- **Estado:** Listo para modelado tras limpieza exhaustiva y feature engineering.

2. Composición Final

- **Registros originales:** 10,841 apps.
- **Registros después de limpieza:** 10,357 apps (eliminación de 483 duplicados y 1 valor inválido).
- **Variables finales:** 38 columnas (14 originales + 24 nuevas variables derivadas).
- **Memoria:** Incremento de 5.98 MB a 8.36 MB debido a feature engineering.
- **Calidad:** 100 % de valores válidos excepto Rating (variable objetivo).

Variables principales (post-procesamiento)

Variable	Tipo	Descripción	Estado
Variables originales (limpias)			
Rating	Numérica	Calificación (1–5)	1,465 faltantes (14.1 %)
Reviews	Numérica	Número de reseñas	Completo
Size	Numérica (MB)	Tamaño imputado	Completo + flag
Price	Numérica (USD)	Precio imputado	Completo + flag
Type	Categórica	Free/Paid corregido	Completo
Variables transformadas			
Reviews_log	Numérica	log1p(Reviews)	Estabilizada
Installs_log	Numérica	log1p(Installs)	Estabilizada
Size_log	Numérica	log1p(Size)	Estabilizada
Variables categorizadas (bins)			
Price_bin	Categórica	5 niveles de precio	Completo
Size_bin	Categórica	5 niveles de tamaño	Completo
Installs_bin	Categórica	7 niveles de popularidad	Completo
Variables binarias (indicadores)			
is_free	Binaria	Aplicación gratuita	92.6 % = 1
is_large_app	Binaria	Tamaño ¿50MB	12.7 % = 1
has_high_installs	Binaria	Instalaciones ¿1M	31.3 % = 1
is_top_category	Binaria	FAMILY o GAME	31.2 % = 1
is_everyone_rated	Binaria	Content Rating = Everyone	80.9 % = 1
Variables derivadas (feature engineering)			
review_rate	Numérica	Reviews/Installs	Engagement
Genres Main	Categórica	Primer género extraído	33 únicas
days_since_update	Numérica	Días desde actualización	Media: 1,539 días
update_recency	Categórica	6 bins temporales	Completo
popularity_score	Numérica	Score compuesto (0-100)	Media: 1.15

3. Transformaciones Aplicadas

3.1 Limpieza de datos

- **Duplicados eliminados:** 483 filas (4.46 %).
- **Valores inválidos:** 1 registro con Rating = 19.0 eliminado.
- **Validación de consistencia:** Type vs Price corregido automáticamente.

3.2 Imputación estratégica

- **Size:** Mediana por Category × Type + flag size_missing (1,696 casos).
- **Content Rating:** Moda por Category + flag (1 caso).
- **Price:** 0 si Free, mediana por Category si Paid + flag (1 caso).

- **Versiones Android/Current:** Moda por Category + flags (3 y 8 casos respectivamente).
- **Rating:** NO imputado (variable objetivo).

3.3 Transformaciones logarítmicas (cifras exactas del output)

- **Reviews_log:** De media 405,905 (mediana 1,680) a media 7.25 (mediana 7.43).
- **Installs_log:** De media 14,157,759 (mediana 100,000) a media 11.10 (mediana 11.51).
- **Size_log:** De media 20.18 MB (mediana 12.00) a media 2.58 (mediana 2.56).

4. Distribuciones Post-Transformación (cifras exactas)

- **Rating:** Media 4.19, mediana 4.30 (outliers inválidos eliminados).
- **Type:** Free (9,592 - 92.6 %), Paid (765 - 7.4 %).
- **Content Rating:** Everyone (8,382 - 80.9 %), Teen (1,146 - 11.1 %), Mature 17+ (447 - 4.3 %), Everyone 10+ (377 - 3.6 %), Adults only 18+ (3), Unrated (2).
- **Category (Top 10):** FAMILY (1,943 - 18.8 %), GAME (1,121 - 10.8 %), TOOLS (843 - 8.1 %), BUSINESS (427 - 4.1 %), MEDICAL (408 - 3.9 %).

5. Variables Derivadas Clave (estadísticas exactas)

- **review_rate:** Media 0.04, mediana 0.02 (engagement).
- **Genres Main:** Géneros únicos extraídos del primer elemento.
- **update_recency:** Distribución de actualización por bins temporales.
- **popularity_score:** Media 1.15, mediana 0.01 (altamente sesgado hacia apps poco populares).

6. Correlaciones Mantenidas

- **Installs_log vs Reviews_log:** Correlación fuerte preservada tras transformación logarítmica.
- **Size_log vs Installs_log:** Relación moderada estabilizada.
- **Price vs popularidad:** Correlación negativa confirmada tras limpieza.
- **Rating:** Sigue siendo independiente de otras variables (desafío para modelado).

7. Flags de Calidad (casos exactos)

- **size_missing:** 1,696 casos marcados.
- **content_rating_missing:** 1 caso marcado.
- **android_ver_missing:** 3 casos marcados.
- **current_ver_missing:** 8 casos marcados.
- **price_missing:** 1 caso marcado.
- Estos flags preservan información sobre patrones de ausencia para el modelado.

8. Estado para Modelado (cifras finales exactas)

- **Registros listos:** 8,892 apps con Rating válido (85.9 % del dataset limpio).
- **Variables predictoras:** 37 disponibles (excluyendo Rating).
- **Distribuciones:** Estabilizadas mediante transformaciones logarítmicas.
- **Outliers:** Preservados como información legítima, transformados para reducir impacto.
- **Consistencia:** 100 % validada entre variables relacionadas.

9. Mejoras Implementadas

- Eliminación de sesgos por duplicados y valores inválidos.
- Imputación inteligente que preserva patrones por categoría.
- Transformaciones que estabilizan distribuciones asimétricas.
- Feature engineering que captura patrones complejos (engagement, recencia, popularidad).
- Creación de variables binarias para modelos lineales.
- Categorización de variables continuas para análisis no paramétrico.

10. Limitaciones Residuales

- 14.1 % de apps sin Rating (eliminadas para modelado).
- Sesgo hacia apps gratuitas persiste (modelo de mercado real).
- Rating sigue siendo difícil de predecir por baja correlación con predictores.
- Variables temporales limitadas (solo fecha de última actualización).
- Ausencia de información sobre desarrolladores o descripciones textuales.

11. Usos Recomendados Post-Procesamiento

- **Modelado supervisado:** Predicción de Rating usando algoritmos lineales y no lineales.
- **Clasificación:** Apps de alto vs bajo rating usando variables binarias.
- **Clustering:** Segmentación de apps por características transformadas.
- **Análisis de importancia:** Identificación de factores clave para el éxito.
- **Benchmark:** Comparación de técnicas de preprocesamiento y feature engineering.

12. Consideraciones Éticas

- Datos públicos, sin información personal sensible.
- Transformaciones documentadas y reproducibles.
- Sesgos de mercado preservados intencionalmente (reflejo de realidad).
- Uso recomendado para investigación, educación y desarrollo de prototipos.
- No apto para decisiones comerciales críticas sin validación adicional.