UNIVERSIDAD NACIONAL SAN AGUSTÍN

FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y SERVICIOS

ESCUELA PROFESIONAL DE CIENCIAS DE LA COMPUTACIÓN

Tópicos en Ciencias de Datos

Alumno:

Apaza Andaluz, Diego Francisco

AREQUIPA

2024

1. **Contexto de mi Dataset (games.csv):**

   Este trabajo se basa en un dataset que contiene información recopilada de múltiples archivos relacionados con videojuegos disponibles en la plataforma Steam no duplicados. Este contiene información detallada sobre un juego: su nombre, fecha de lanzamiento, desarrollador, editor, categorías, géneros, etiquetas, tiempo promedio de juego, disponibilidad en diferentes sistemas operativos, y si incluye mecánicas de monetización como microtransacciones o loot boxes.

2. **Descripción del Dataset:**

   El dataset, una vez unificado y limpiado, **incluye 111,452 registros**, cada uno representando un videojuego único (o una entrada relacionada con un título en particular). Cada fila del dataset representa una instancia individual de un videojuego —es decir, es el nivel más básico de observación (registro). No existen niveles de granularidad geográfica (país, región, ciudad), ni temporal (día, mes, año de eventos), salvo por la fecha de lanzamiento. La información no está etiquetada en el sentido de aprendizaje supervisado, pero sí puede considerarse etiquetada desde una perspectiva analítica: por ejemplo, la columna FreeToPlay puede actuar como una etiqueta binaria para clasificación o segmentación.

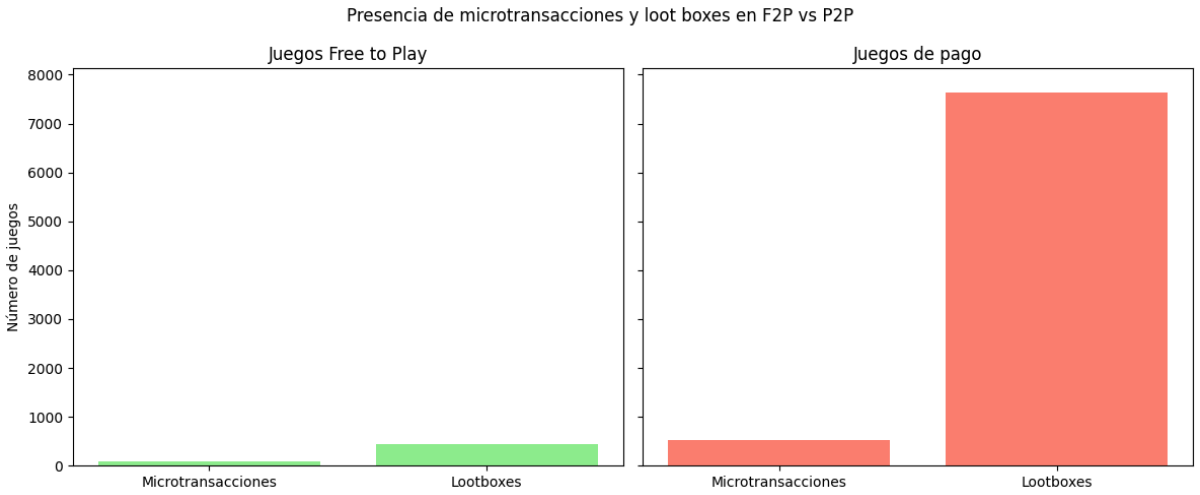
   ```
   Número total de filas: 111452
   ```

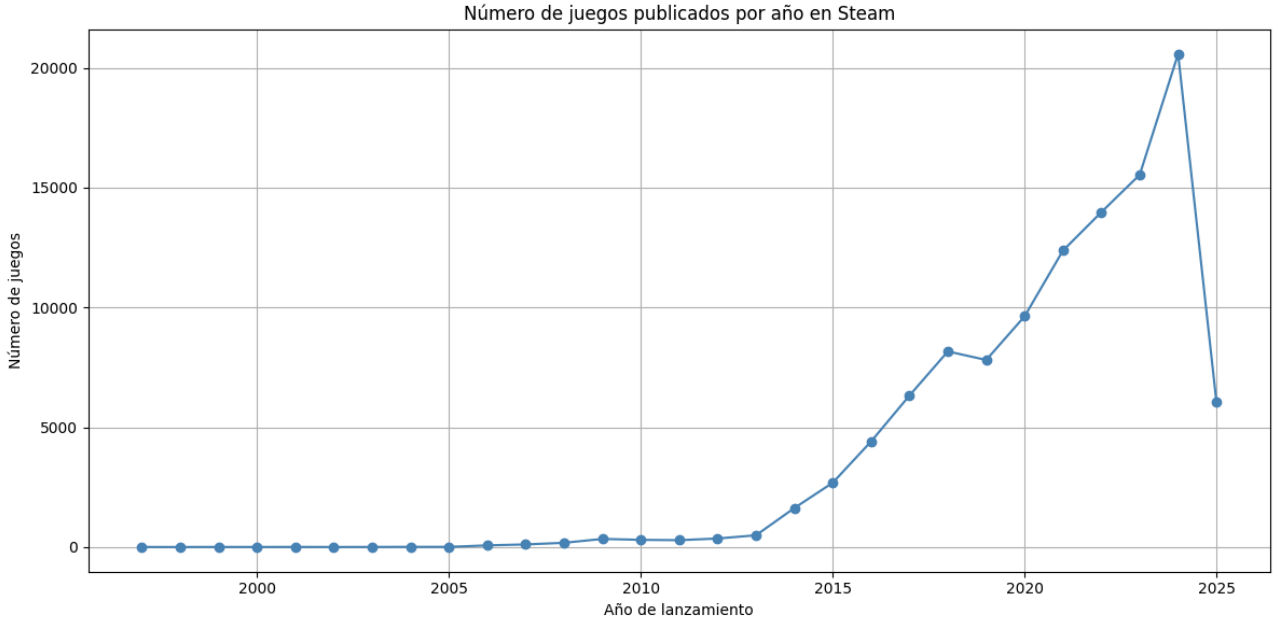
   El tamaño del dataset es adecuado para análisis estadísticos y exploratorios sin comprometer la capacidad computacional (CPU o RAM), además de que se trabajó en la plataforma Google Colab

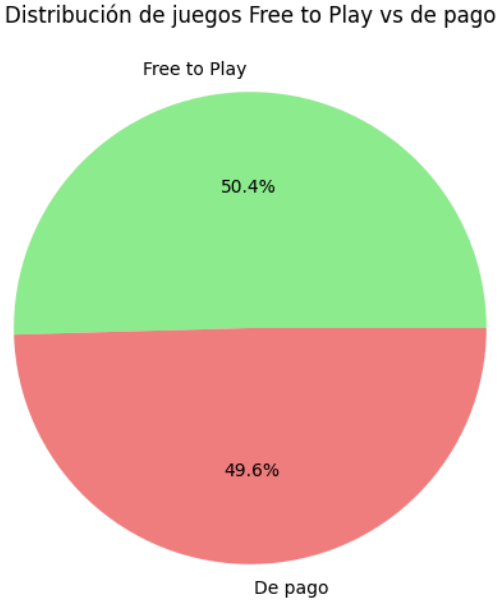
3. **Cual es mi objeto de estudio:**
   Se estudian diferentes títulos de videojuegos publicados en la plataforma Steam. Cada fila representa un videojuego específico como unidad de análisis, conteniendo información técnica, comercial y conductual. Esta información incluye variables como fecha de lanzamiento, desarrollador, precio, presencia de microtransacciones o loot boxes, categorías de juego, tiempo promedio de uso y datos relacionados con la percepción y mecánicas de monetización.

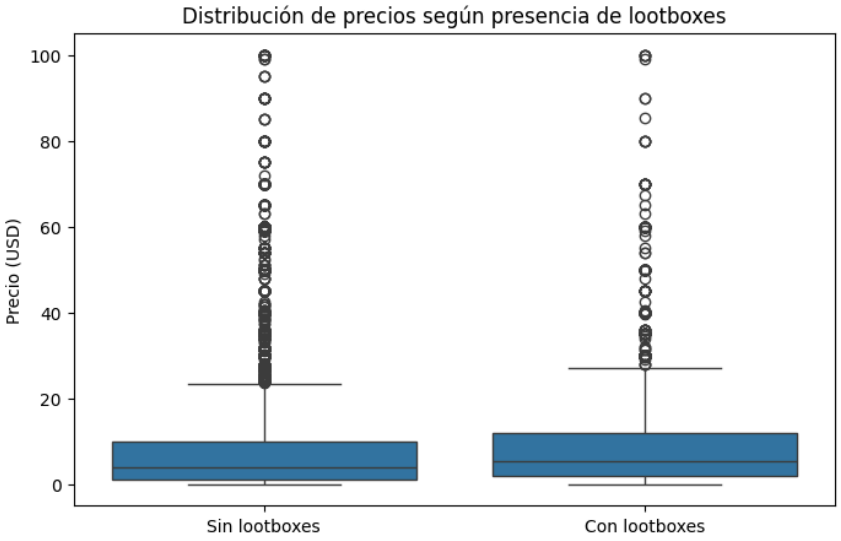
Atributo	Tipo de dato	Significado
Name	object	Nombre del videojuego
Release date	object	Fecha de lanzamiento
Estimated owners	object	Rango estimado de usuarios
Peak CCU	int64	Número máximo de jugadores concurrentes Mínimo 0 | Máximo: 1311366
Required age	int64	Edad mínima recomendada para jugar Mínimo: 0 | Máximo: 21
Price	float64	Precio base del videojuego 0.0 | Máximo: 999.98

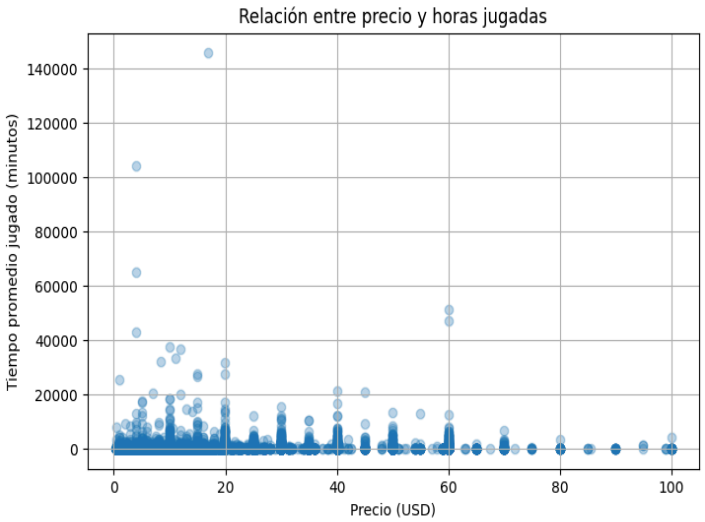
DiscountDLC count	int64	Número de contenidos descargables (DLC) Mínimo: 0 | Máximo: 92
About the game	object	Descripción textual del juego
Supported languages	object	Lista de idiomas en los que el juego está disponible
Windows	bool	Compatible con el sistema operativo Windows
Mac	bool	Compatible con el sistema operativo Mac
Linux	bool	Compatible con el sistema operativo Linux
Score rank	float64	Clasificación del juego
Average playtime forever	int64	Tiempo promedio (en minutos) que los usuarios han jugado el título desde que lo adquirieron.
Average playtime two weeks	int64	Tiempo promedio (en minutos) jugado en las últimas dos semanas por usuarios activos.
Median playtime forever	int64	Tiempo mediano total (en minutos) jugado por los usuarios desde su adquisición
Median playtime two weeks	int64	Tiempo mediano (en minutos) jugado en las últimas dos semanas.
Developers	object	Estudio Desarrollador
Publishers	object	Publicador
Categories	object	Categorías asignadas al juego, como “Single-player”, “Multiplayer”, “Controller support”, etc.
Genres	object	Géneros principales del juego, como Acción, Estrategia, Rol, Aventura, etc.
Tags	object	Etiquetas asignadas por la comunidad o por los desarrolladores que describen el contenido o características.
FreeToPlay	bool	Indica si el juego es gratuito
HasMicrotransactions	bool	Indica si el juego menciona o contiene microtransacciones en su descripción o etiquetas.
HasLootboxKeywords	bool	Indica si el juego incluye referencias a loot boxes u objetos aleatorios como cofres, sobres, gacha, etc.

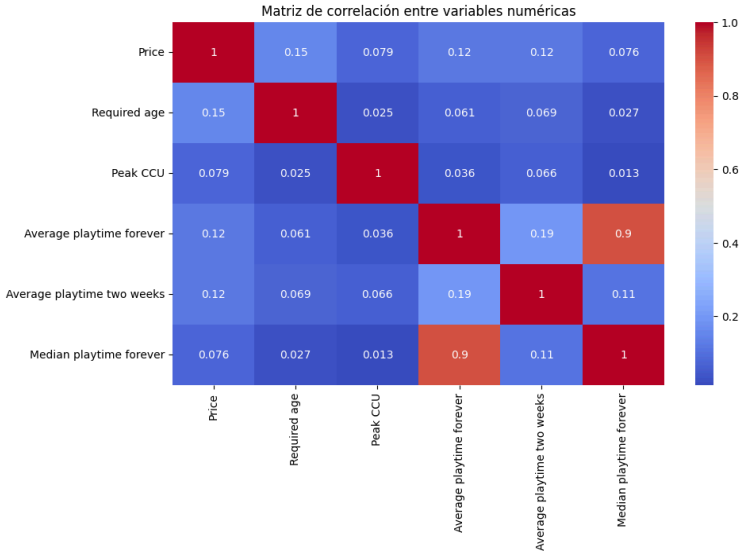

Presencia de microtransacciones y loot boxes en F2P vs P2P
Juegos Free to Play
Juegos de pago
Número de juegos
0
1000
2000
3000
4000
5000
6000
7000
8000
Microtransacciones
Lootboxes
Microtransacciones
Lootboxes


Número de juegos publicados por año en Steam
Número de juegos
0
5000
10000
15000
20000
2000
2005
2010
2015
2020
2025
Año de lanzamiento


Distribución de juegos Free to Play vs de pago
Free to Play
50.4%
49.6%
De pago







Outliers

Peak CCU: máximo >1.3 millones

Playtime: varios juegos con >100,000 minutos

- **¿Podemos eliminarlos? ¿Es importante conservarlos?**

  No necesariamente. Muchos son juegos reales (ej. CS:GO, Dota 2) con miles o millones de usuarios.

  Deben conservarse si se quiere reflejar la realidad de títulos populares.

  Para visualización, se recomienda usar log scale o limitar el eje y.

- **son errores de carga o son reales?**

  Con una revisión manual se sugiere que son juegos conocidos y populares. Se conservarán para el análisis.

En este proyecto no estamos abordando un problema supervisado en el sentido clásico (es decir, sin una variable de salida definida para entrenamiento y predicción). Sin embargo, podríamos considerar la columna HasLootboxKeywords como una “pseudoetiqueta” binaria (True/False) que indica si un juego contiene términos relacionados con loot boxes. Esta variable podría usarse como etiqueta (target) en una posible tarea futura de clasificación.

Al analizar esta columna como salida hipotética, se observa que la distribución está desbalanceada: la mayoría de los juegos no tienen referencias a loot boxes, mientras que solo un subconjunto minoritario sí las tiene. Esto implicaría la necesidad de aplicar técnicas de balanceo (por ejemplo, sobremuestreo o submuestreo) si se desarrollara un modelo supervisado en el futuro.

Respecto a las features (atributos) más importantes, se identifican como potencialmente relevantes:

- Price: importante para analizar el modelo de negocio (free-to-play vs. premium).
- Categories, Tags y Genres: contienen información clave sobre la mecánica y tipo de juego.
- Average playtime forever y Peak CCU: indicadores de popularidad y retención.
- About the game: contiene descripciones ricas en texto que podrían analizarse con NLP.

Otras variables como Score rank, Release date o DiscountDLC count también pueden ser útiles, pero algunas como Developers, Publishers o Supported languages pueden ser descartadas dependiendo del objetivo del análisis, ya que contienen información demasiado específica o dispersa.

No se trata de un problema clásico de series temporales (time series), ya que cada juego es una instancia única en el tiempo. Sin embargo, al observar la variable Release date a nivel de año, sí se puede analizar la evolución temporal de ciertas tendencias, como el aumento de juegos con loot boxes o F2P a partir de 2015. Por tanto, hay una dimensión temporal de interés, pero no es un problema de predicción secuencial.

Conclusión

Este análisis preliminar permitió entender la estructura del dataset, identificar posibles etiquetas y variables explicativas, detectar problemas de calidad y desigualdad de clases, y extraer relaciones útiles entre variables clave de monetización. Esto sienta una base sólida para un análisis más profundo, visualizaciones comparativas o modelos explicativos que puedan apoyar discusiones sobre ética, diseño económico y regulación del gambling en videojuegos.