

# 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**

AppID	Name	Release date	Estimated owners	Pe
-------	------	--------------	------------------	----

Por ejemplo dado que el nombre estaba en la columna ApplD los demas datos

	Release date
2567970	0 - 20000
2669610	0 - 20000
2571900	0 - 20000
2485520	0 - 0
2173350	0 - 20000
...	...
2673370	0 - 0
2741410	0 - 20000
2743870	0 - 20000
2739140	0 - 20000
2725770	0 - 20000

Columnas eliminadas: ['Movies', 'Screenshots', 'Notes', 'Recommendations', 'Achievements', 'Negative', 'Positive', 'Metacritic score']

Dimensiones originales: (111452, 42)

Dimensiones actuales: (111452, 25)

## Otros Ejemplos

### II. Qué descubrieron al analizar los datos:

Algunas columnas como “Estimated owners” tienen valores como “0 - 20000” esto es un rango expresado como texto, y no puede ser tratado como valor numérico directo.

- “Release date” es object (texto), pero sería útil transformarlo a tipo datetime para hacer análisis temporal.
- “Reviews” y “Notes” son object, pero están vacías o tienen muy poca información en muchos registros.
- “Tags”, “Categories”, “Genres”, “Screenshots” están como texto plano, pero representan listas o conjuntos múltiples separados por comas

```

Unique Words in Tags column:
['1980s', '1990s', '2.5D', '2D', '2D Fighter', '2D Platformer', '360 Vi

Total unique tags: 453

Unique Words in Categories column:
['Captions available', 'Co-op', 'Commentary available', 'Cross-Platform

Total unique categories: 43

```

Los elementos dentro de la columna Tags and Categories tenían comas dentro de sus elementos por lo que podría generar confusión

Missing values per column:

Name	6
About the game	6483
Reviews	100828
Website	64994
Support url	60693
Support email	19025
Metacritic url	107447
Score rank	111408
Notes	93003
Developers	6475
Publishers	6778
Categories	7566
Genres	6440
Tags	37423
Screenshots	4192
Movies	9620

dtype: int64

Number of duplicate rows: 6

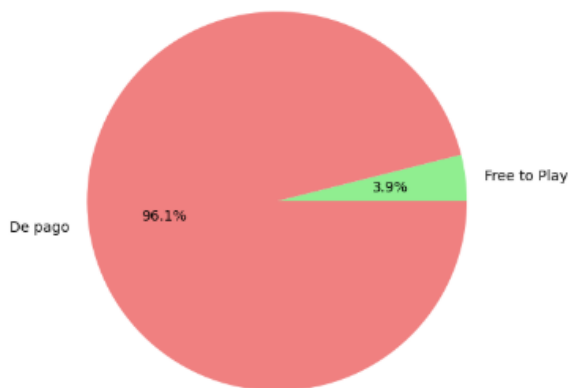
Existen filas que no contienen datos de algunas columnas

La mayoría de tipos de datos eran Object como fecha de salida o dueños estimados del juego

Name	object
Release date	object
Estimated owners	object
Peak CCU	int64
Required age	int64
Price	float64
DiscountDLC count	int64
N/A	int64
About the game	object
Supported languages	object
Full audio languages	object
Reviews	object
Header image	object
Website	object
Support url	object
Support email	object
Windows	bool
Mac	bool
Linux	bool
Metacritic score	int64
Metacritic url	object
User score	int64
Positive	int64
Negative	int64

### III. Que reflejan los patrones de tendencia

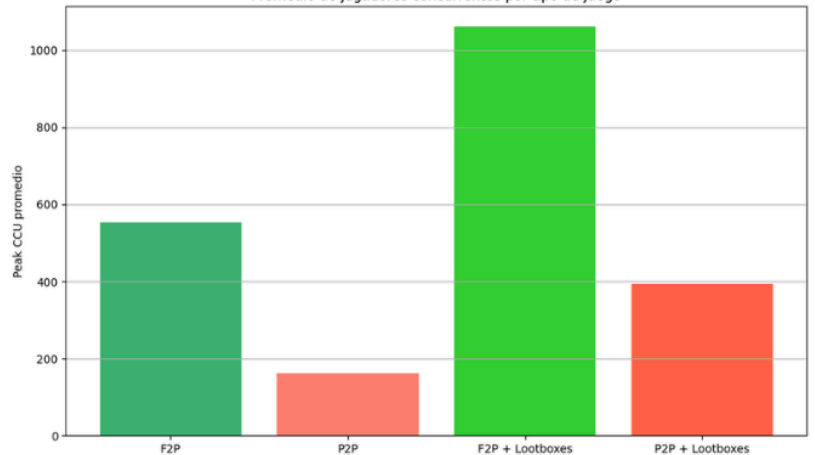
Distribución de juegos Free to Play vs de pago



Existen una cantidad muy baja de juegos Gratis a comparación de los juegos de pago.

Sin embargo, estos juegos gratis suelen retener una cantidad de jugadores mayor a la de los juegos P2P

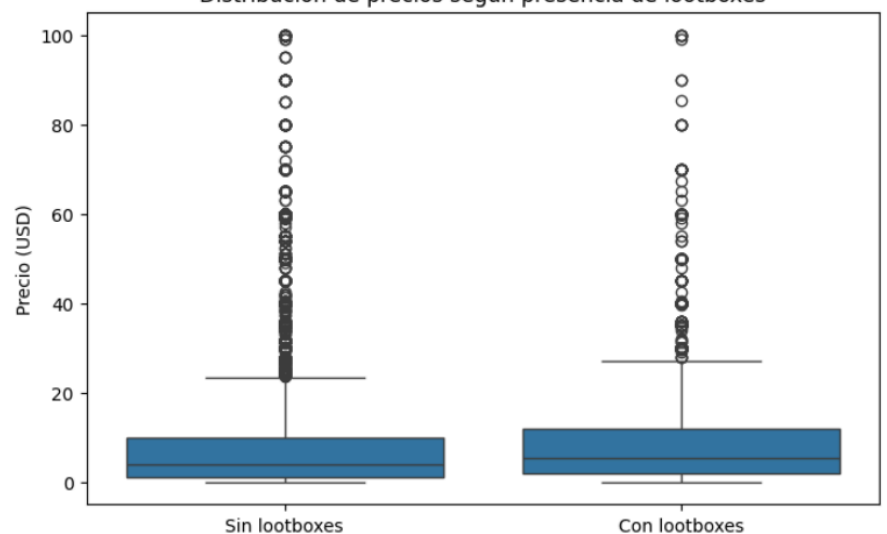
Promedio de jugadores concurrentes por tipo de juego



#### - Outliers:

Esto sugiere que algunos juegos implementan un modelo de “doble monetización”: cobran por entrar (precio base) y además incluyen lootboxes.

Distribución de precios según presencia de lootboxes





#### 4. Cómo es afectado el comportamiento del jugador frente al tiempo de juego y reseñas de otros.

- Las reseñas (positivas/negativas) y las recomendaciones están fuertemente vinculadas.
- El precio no influye en la valoración ni en el tiempo de juego.
- Los logros (Achievements) son irrelevantes para otras métricas en este dataset.
- El tiempo de juego tiene una leve relación con reseñas positivas, pero no es determinante.

