



ESTUDIO DE LAS RECOMENDACIONES DE STEAM

GRUPO 3

ADOLFO ARENAS P.
ALEJANDRO MORI A.
IGNACIO HUMIRE S.
LEONARDO RIKHARDSSON
MARIO BENAVENTE C.

¿QUÉ ES STEAM?



¿QUÉ ES STEAM?

ES UNA PLATAFORMA DE
DISTRIBUCIÓN DIGITAL DE
VIDEOJUEGOS, SOFTWARE Y
CONTENIDO MULTIMEDIA CREADA
POR VALVE CORPORATION



MOTIVACIÓN Y OBJETIVOS

¿CÓMO PUEDO HACER QUE MI PROYECTO SE VENDA EN UN MERCADO TAN SATURADO?

MOTIVACIÓN Y OBJETIVOS

¿CÓMO PUEDO HACER QUE MI PROYECTO SE VENDA EN UN MERCADO TAN SATURADO?



MOTIVACIÓN Y OBJETIVOS

¿CÓMO PUEDO HACER QUE MI PROYECTO SE VENDA EN UN MERCADO TAN SATURADO?



- Proporcionar información valiosa que permita a los desarrolladores de videojuegos tomar decisiones que aumenten sus probabilidades de éxito de ventas al lanzar su juego en la plataforma Steam.

PREGUNTAS

1. ¿Qué parámetros influyen más en la cantidad de ventas totales?

PREGUNTAS

- 1. ¿Qué parámetros influyen más en la cantidad de ventas totales?**
- 2. ¿Se puede predecir la cantidad de juegos con género "X" que habrán en un año determinado?**

PREGUNTAS

- 1. ¿Qué parámetros influyen más en la cantidad de ventas totales?**
- 2. ¿Se puede predecir la cantidad de juegos con género "X" que habrán en un año determinado?**
- 3. ¿Es posible identificar grupos de juegos (precios, número de *reviews* similares) que ayuden a entender las características comunes de juegos exitosos o populares?**

EXPLORACIÓN DE DATOS

Géneros por Año

Nube de Palabras de Géneros para el Año 1997

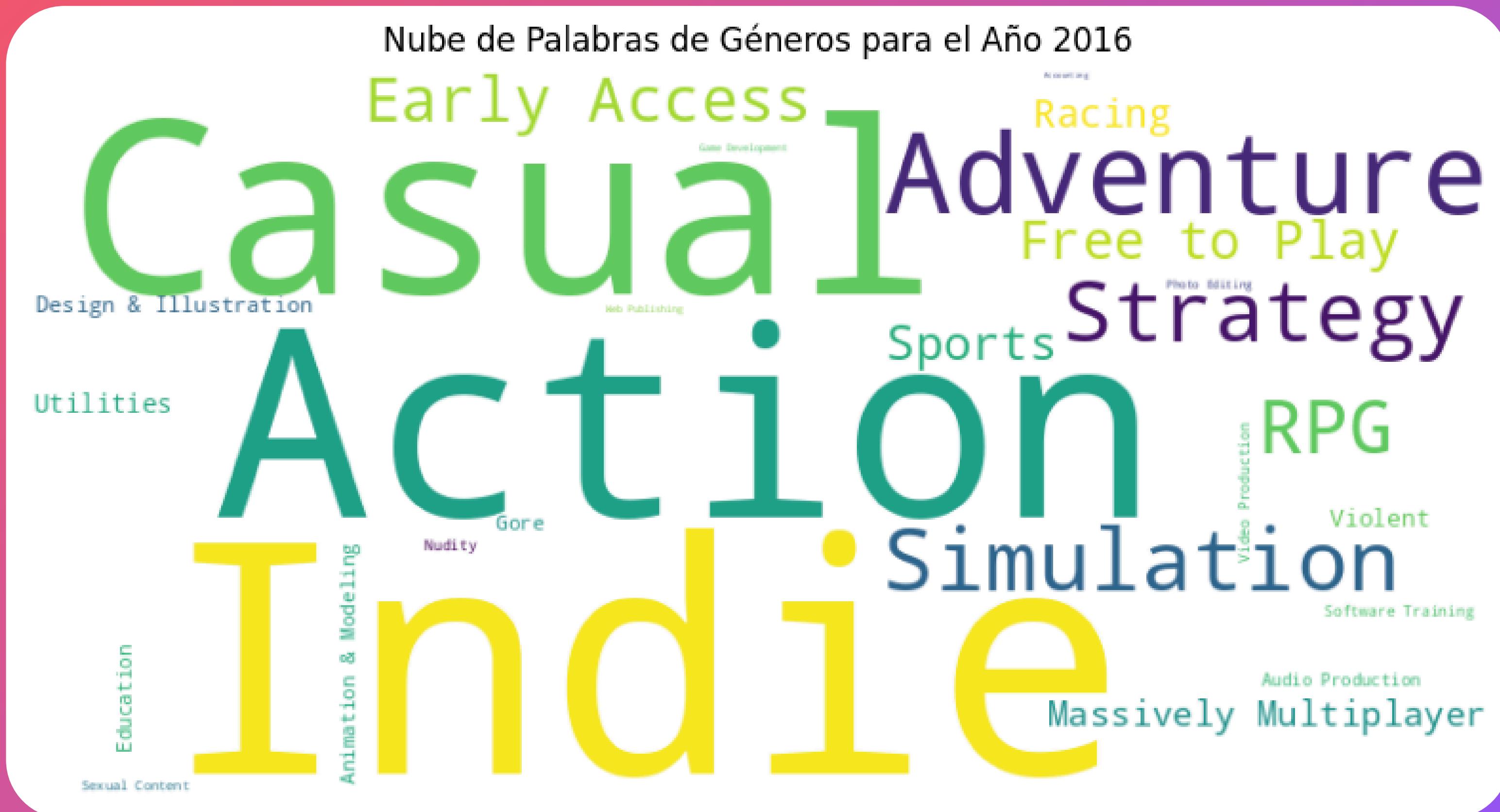


Géneros por Año

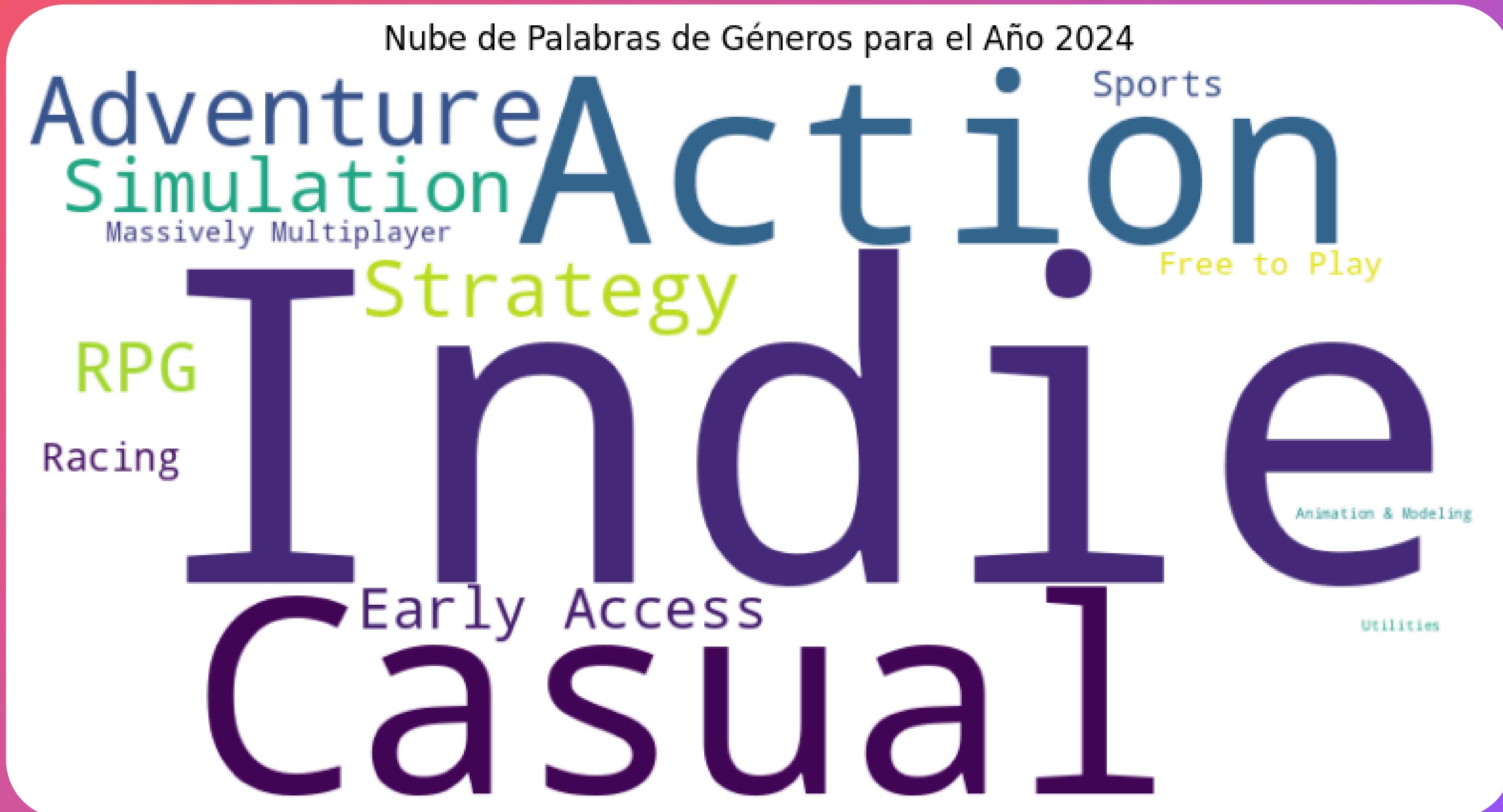
Nube de Palabras de Géneros para el Año 2007



Géneros por Año

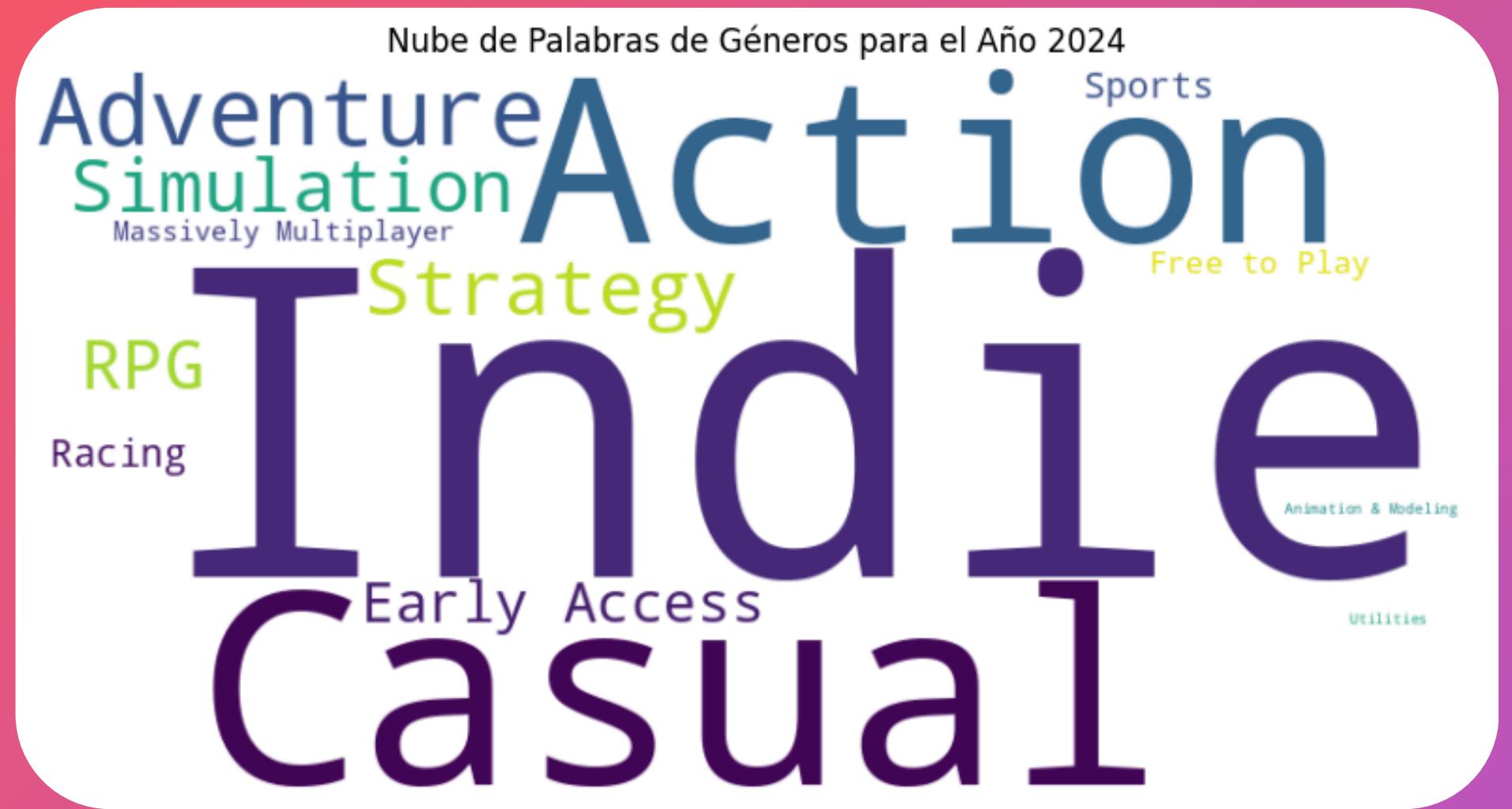


Géneros por Año

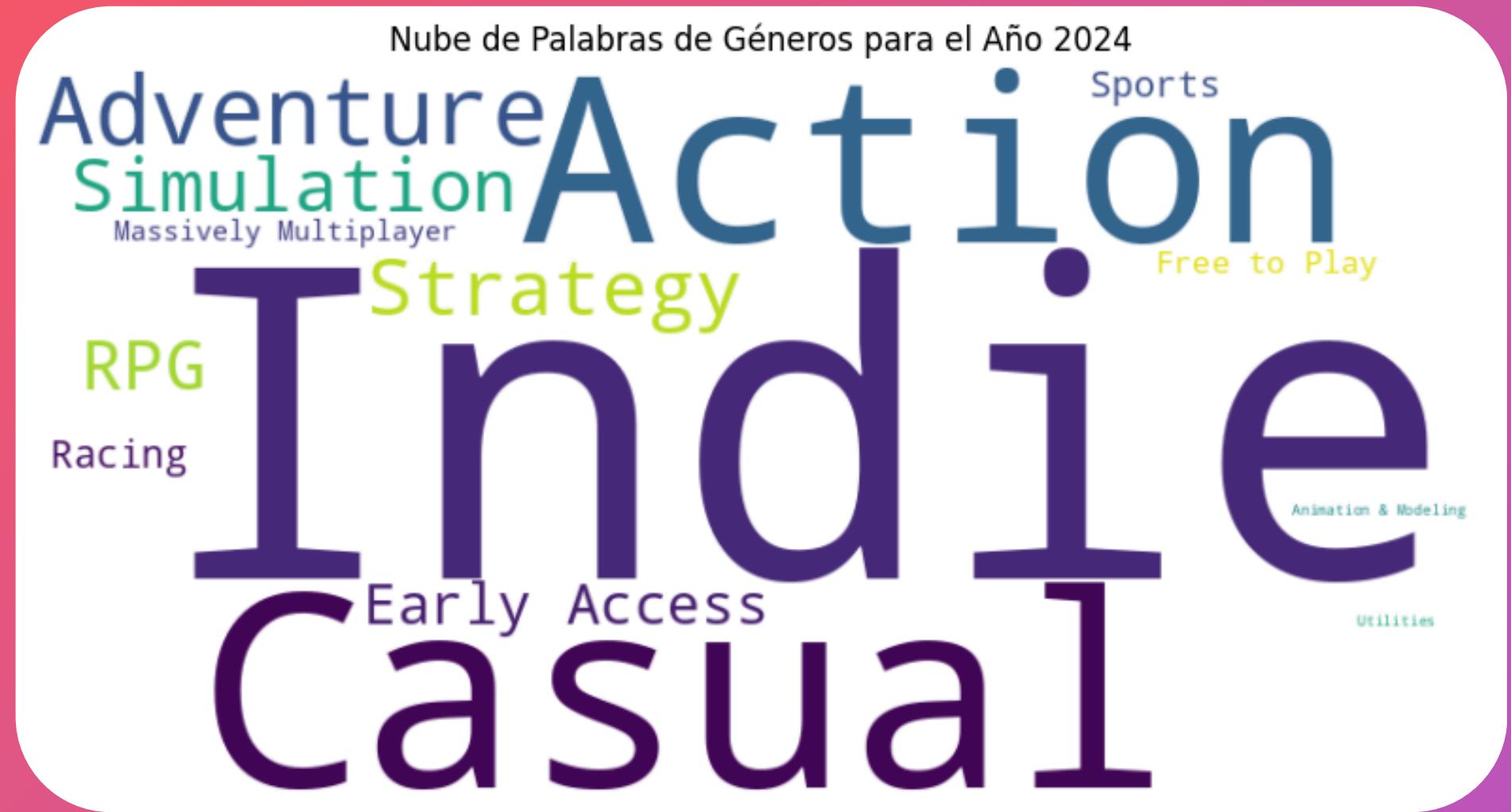


CONCLUSIONES DATOS

CONCLUSIONES DATOS

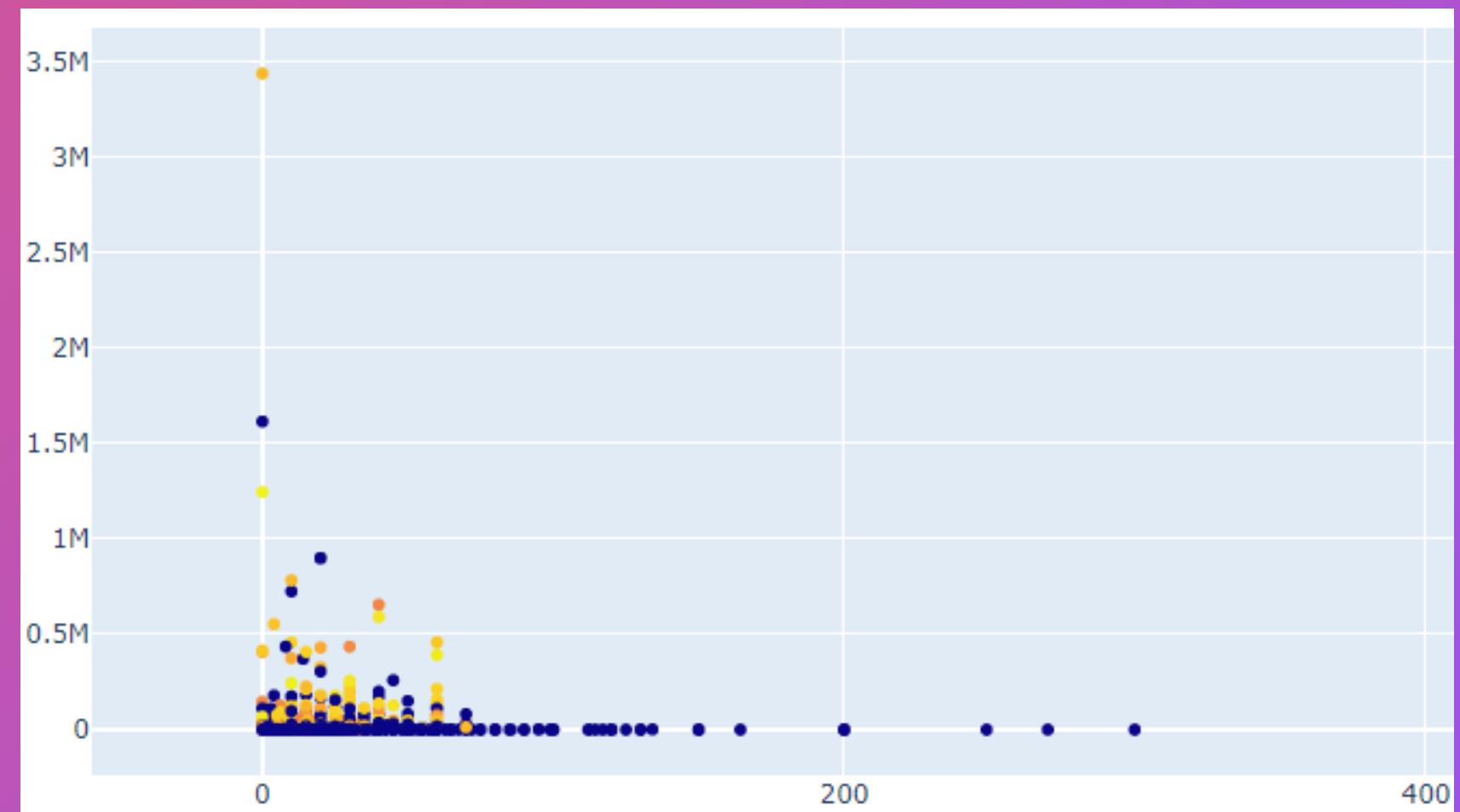


CONCLUSIONES DATOS



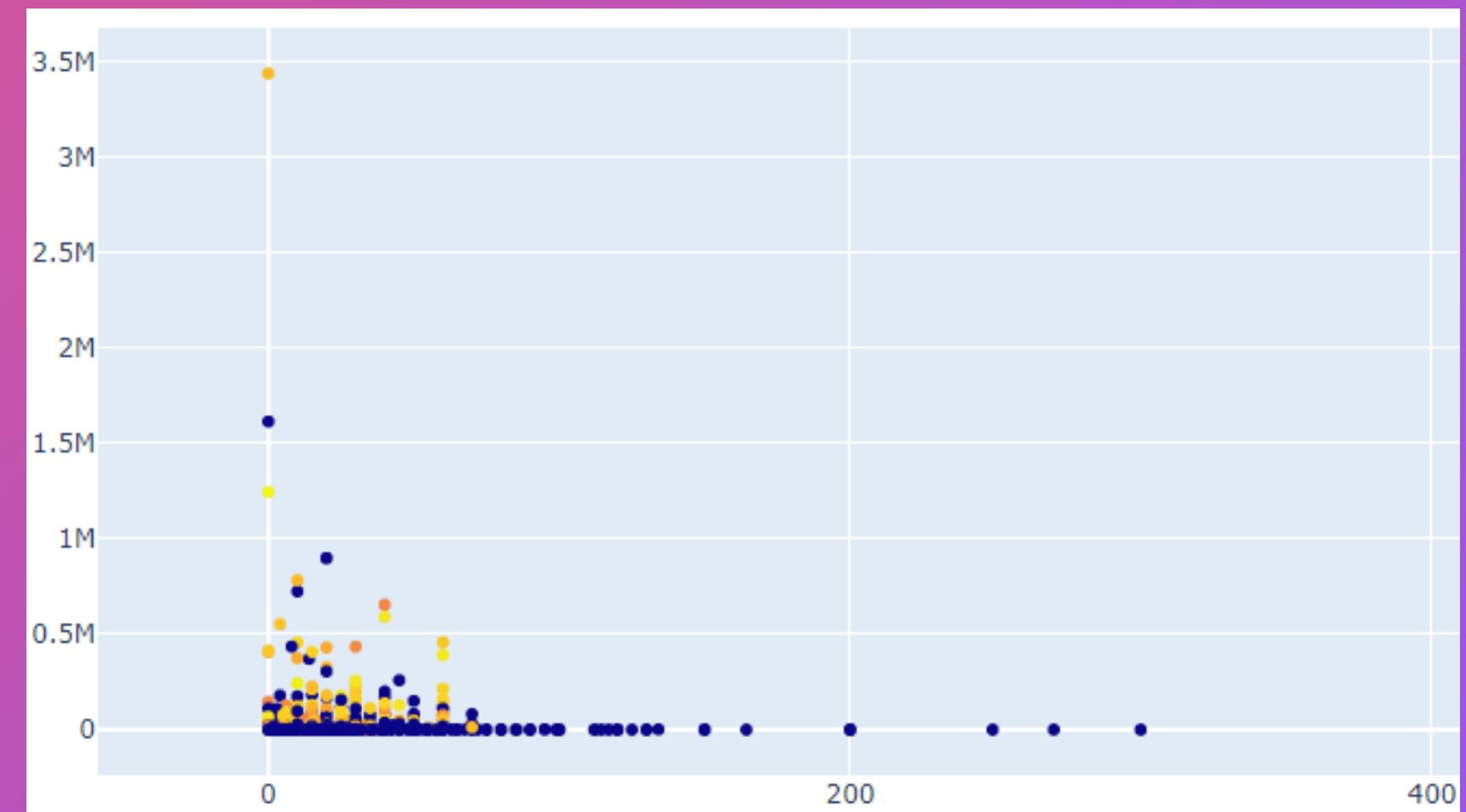
Riqueza, popularidad,
géneros.

CONCLUSIONES DATOS



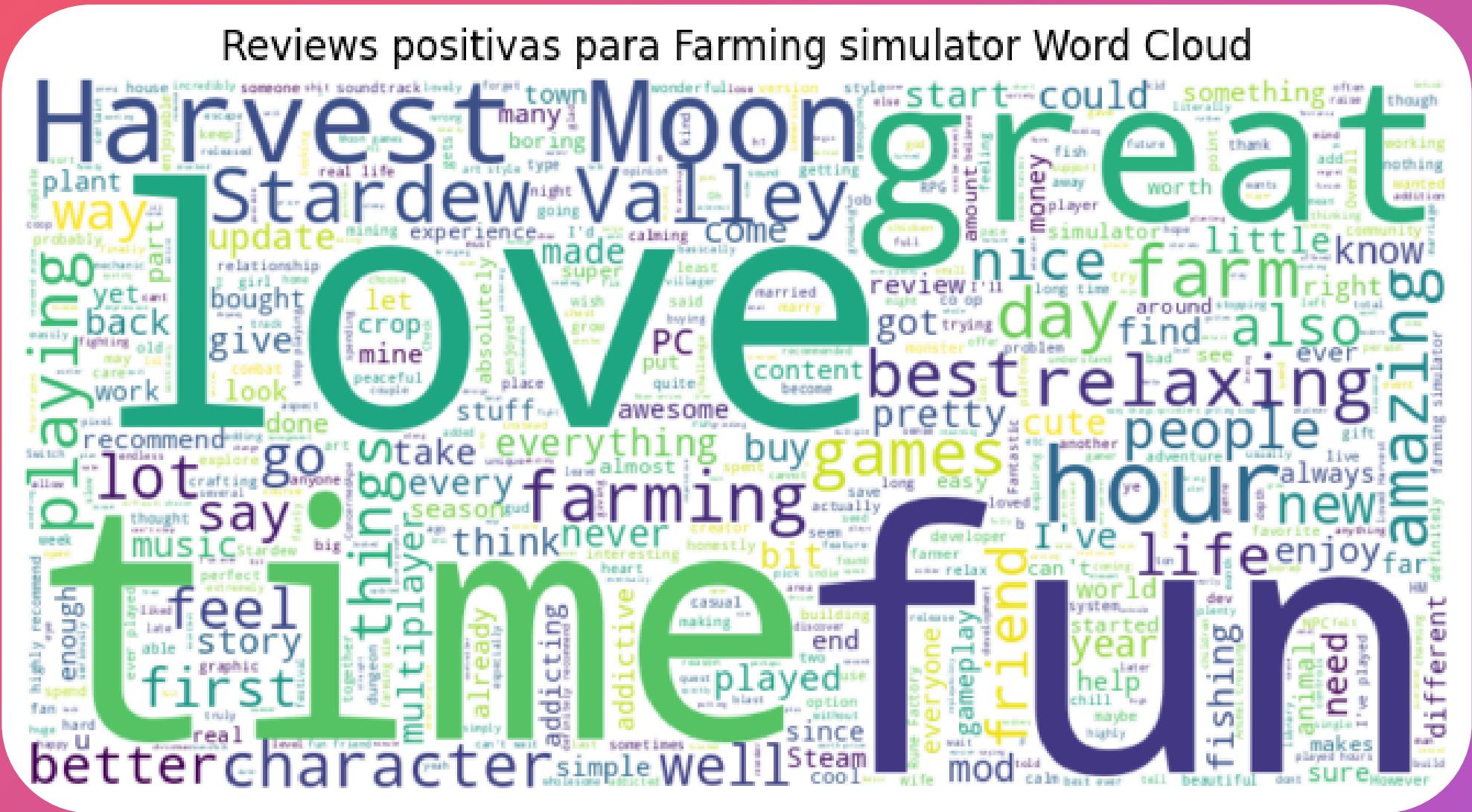
CONCLUSIONES DATOS

Preprocesamiento, sample



CONCLUSIONES DATOS

Reviews positivas para Farming simulator Word Cloud



CONCLUSIONES DATOS

Reviews positivas para Farming simulator Word Cloud



Datos no relevantes para los experimentos

EXPERIMENTOS Y SUS RESULTADOS

PREGUNTA 1

¿Qué parámetros influyen más en la cantidad de ventas totales?

- Uso de clasificadores para predecir rango de ventas.
- Clase [0, 20000] dominante.
- 30% de los datos destinados al entrenamiento.
- Medir desempeño con: *Presicion*, *Recall* y *F1-score*.

Predicción de la cantidad de dueños basándose en el precio y la puntuación en Metacritic

Clasificador	Precisión Promedio	Recall Promedio	F1-score Promedio
Base Dummy	0.453	0.453	0.453
<u>Decision Tree</u>	<u>0.604</u>	<u>0.704</u>	<u>0.639</u>
Gaussian Naive Bayes	0.572	0.514	0.500
KNN	0.605	0.605	0.635

Predicción de la cantidad de dueños basándose en el precio y las reseñas positivas

Clasificador	Precisión Promedio	Recall Promedio	F1-score Promedio
Base Dummy	0.453	0.453	0.453
<u>Decision Tree</u>	<u>0.749</u>	<u>0.781</u>	<u>0.757</u>
Gaussian Naive Bayes	0.262	0.244	0.157
KNN	0.736	0.774	0.748

Predicción de la cantidad de dueños basándose en la fecha de lanzamiento y el número de recomendaciones

Clasificador	Precisión Promedio	Recall Promedio	F1-score Promedio
Base Dummy	0.453	0.453	0.453
<u>Decision Tree</u>	<u>0.616</u>	<u>0.682</u>	<u>0.605</u>
Gaussian Naive Bayes	0.121	0.201	0.091
KNN*	0.616	0.670	0.629

Predicción de la cantidad de dueños basándose en reseñas positivas y negativas

Clasificador	Precisión Promedio	Recall Promedio	F1-score Promedio
Base Dummy	0.453	0.453	0.453
<u>Decision Tree</u>	<u>0.702</u>	<u>0.703</u>	<u>0.687</u>
Gaussian Naive Bayes	0.103	0.225	0.115
KNN	0.697	0.699	0.685

**¿Qué parámetros influyen más en
la cantidad de ventas totales?**

**¿Qué parámetros influyen más en
la cantidad de ventas totales?**

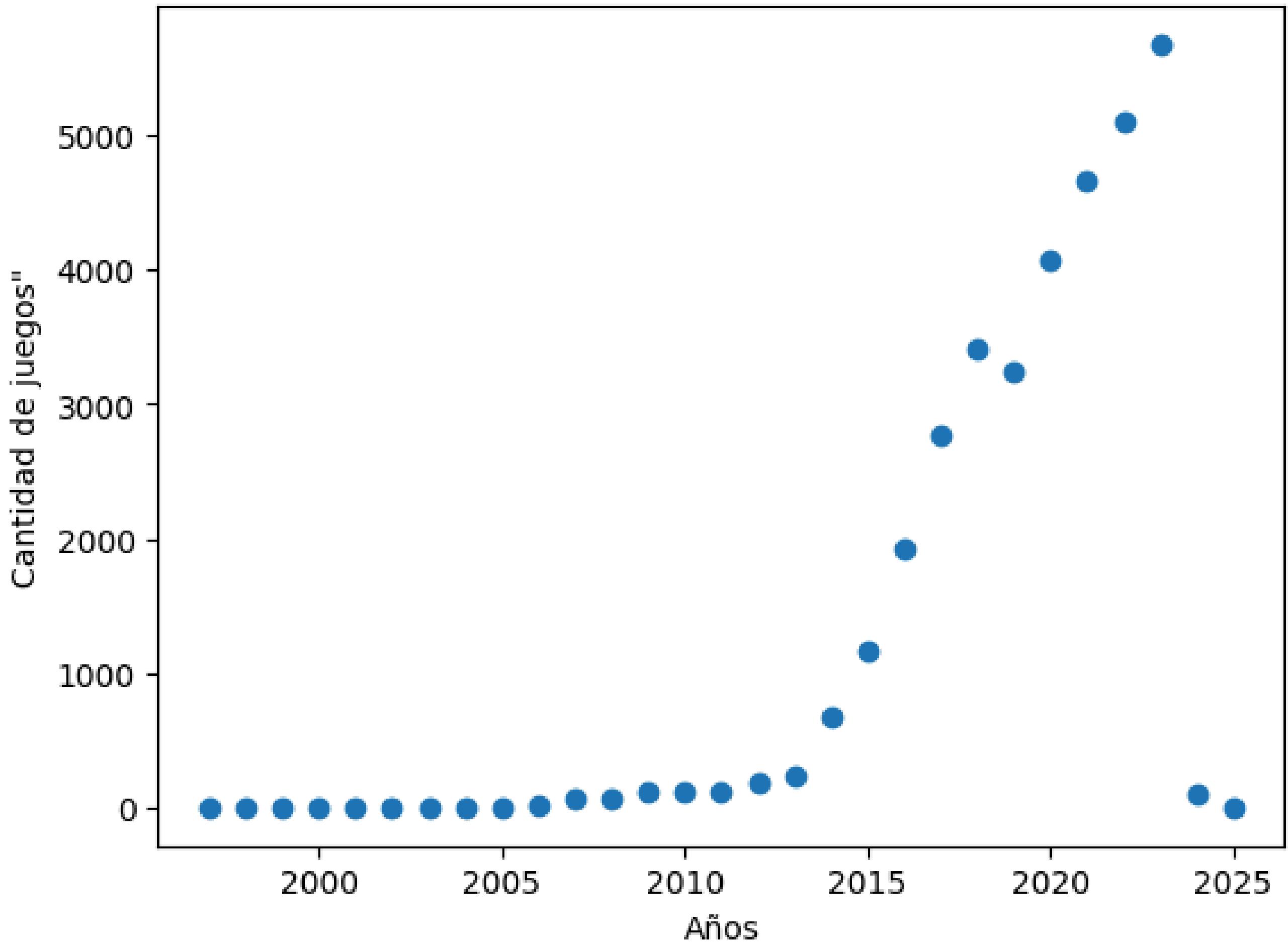
**Los atributos más
significativos son el precio
del juego y la cantidad de
reseñas positivas.**

PREGUNTA 2

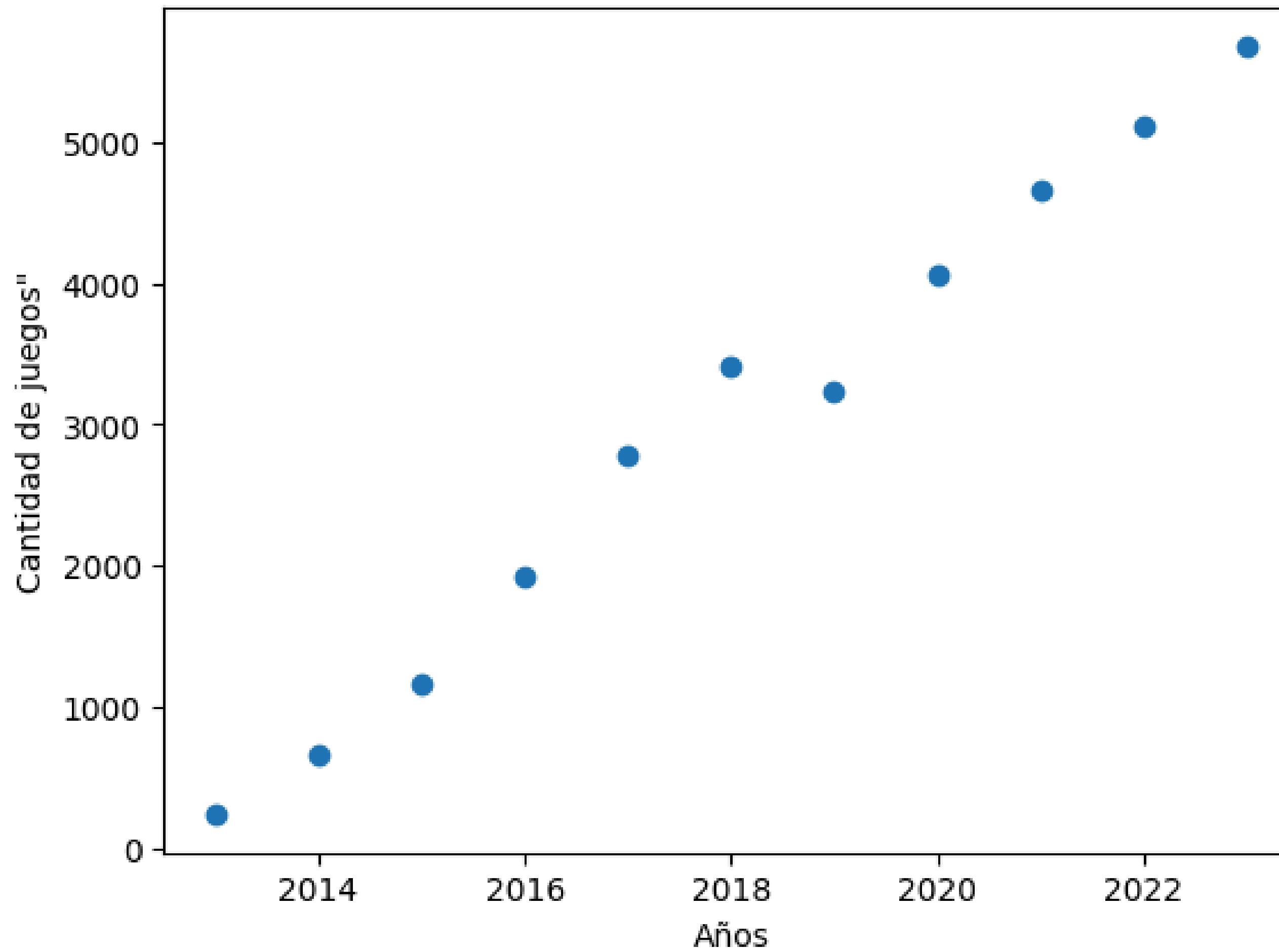
¿Es posible estimar cuántos juegos de un género específico se lanzarán en un año determinado?

- Para preprocessar, extraemos el año de ‘*release date*’.
- Agrupar los videojuegos lanzados por año para hacer la regresión lineal.
- Evaluar si se debe restringir rango de años.
- Experimentar con regresión exponencial para ver si se ajusta mejor.

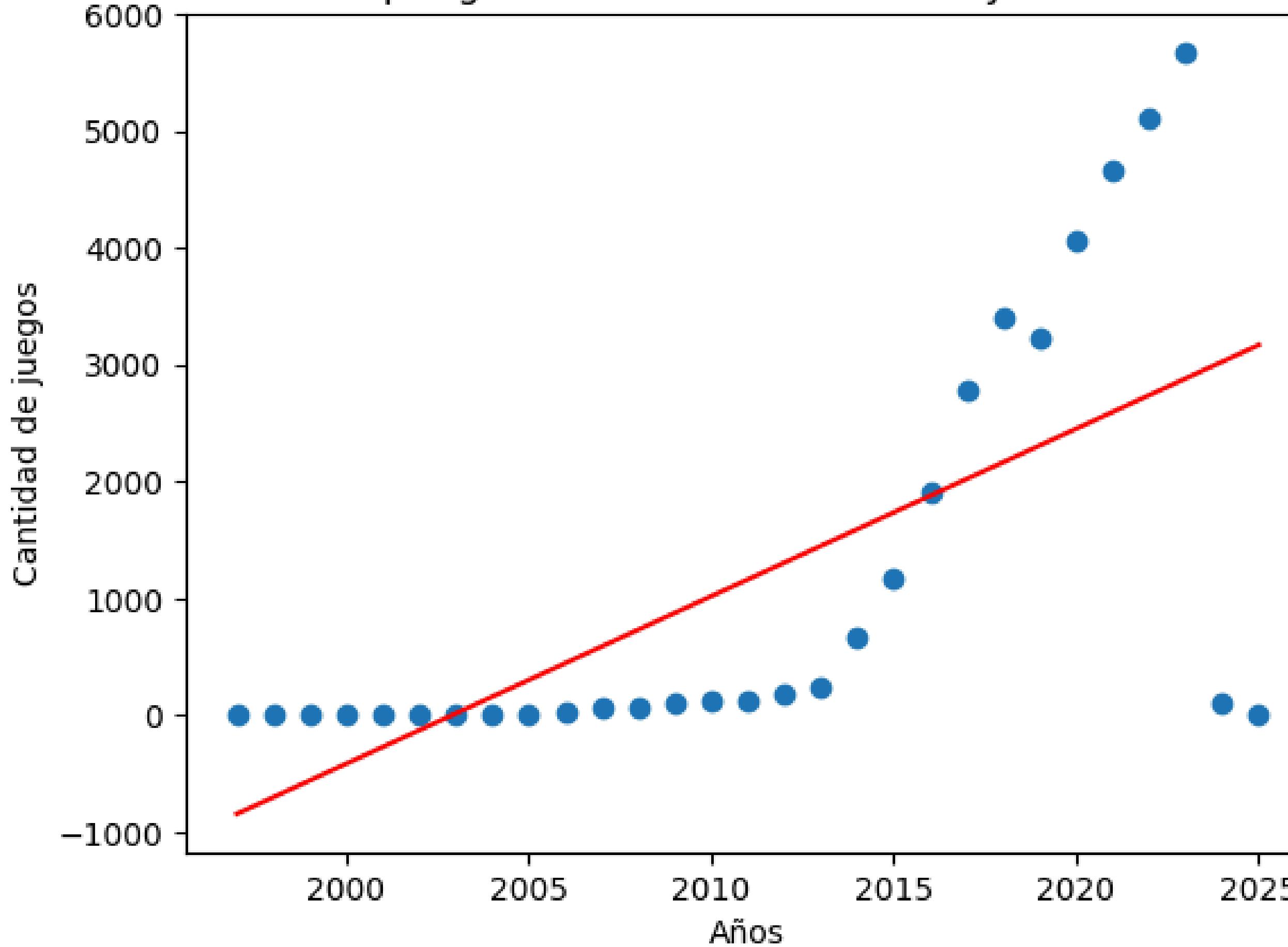
Cantidad de juegos por género "Action" a través de los años



Cantidad de juegos por género "Action" entre 2013 y 2023

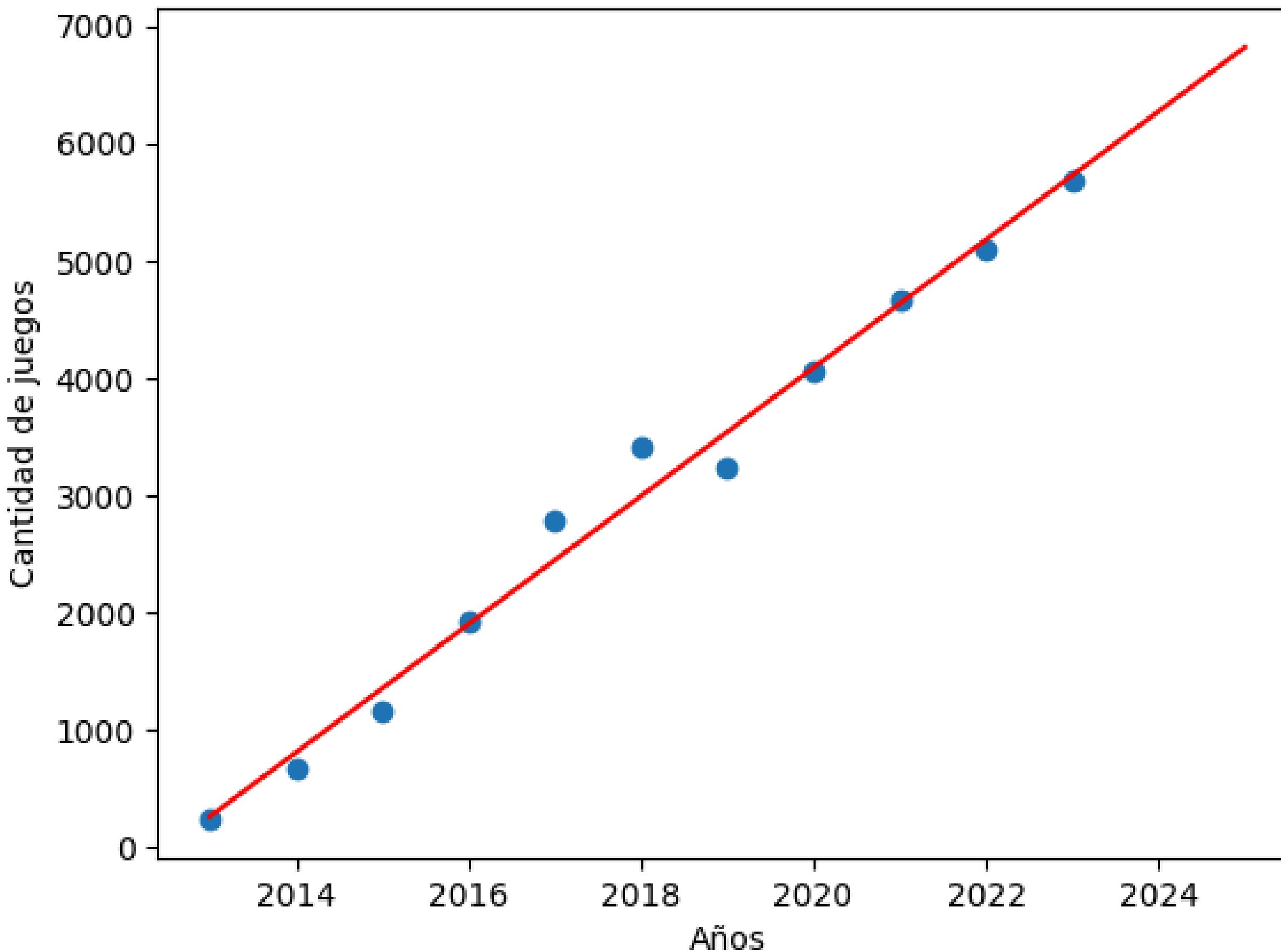


Regresión lineal para la cantidad de juegos por género "Action" entre 1997 y 2025



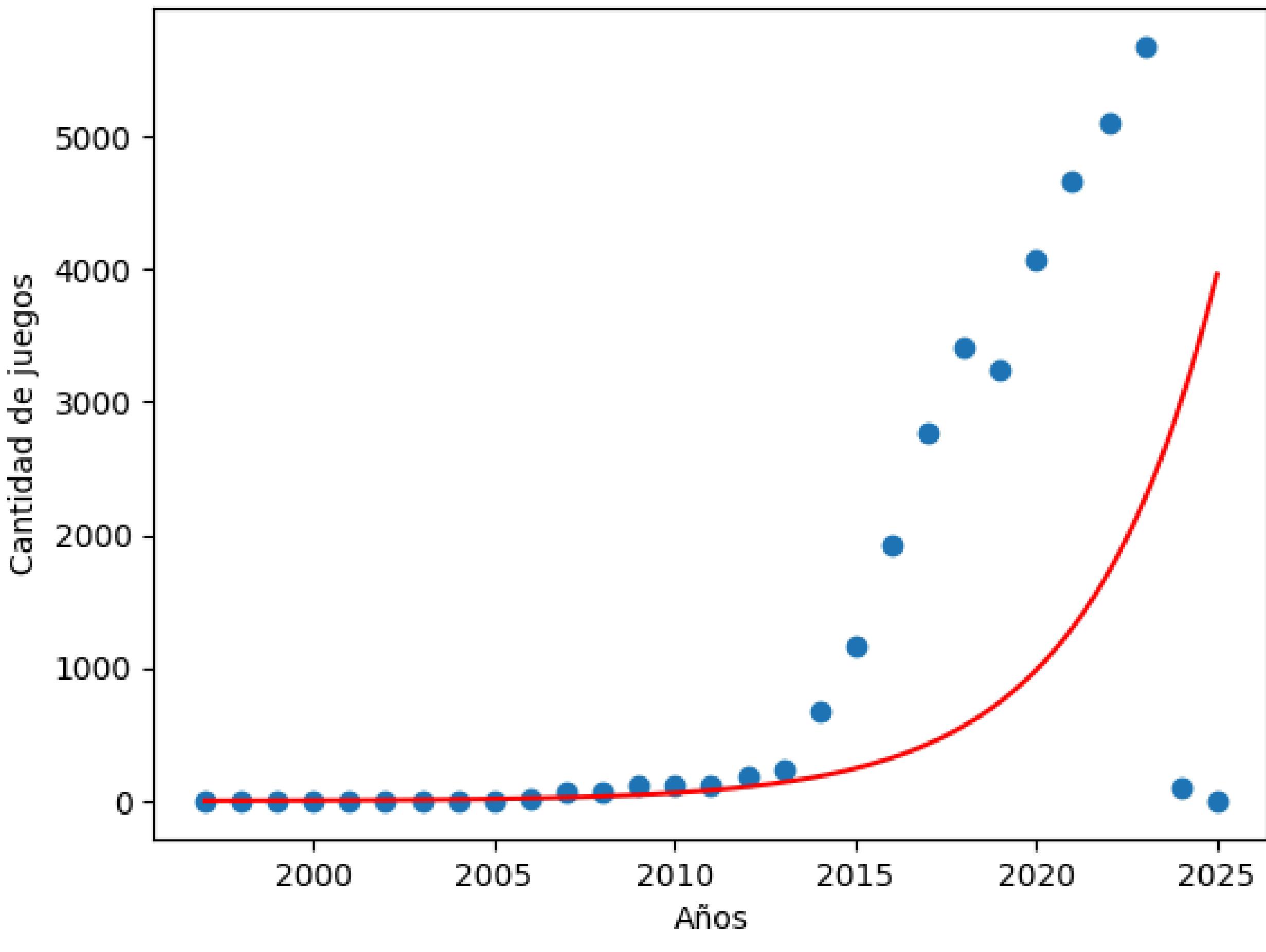
$R^2=0.450$

Regresión lineal para la cantidad de juegos por género "Action" entre 2013 y 2025



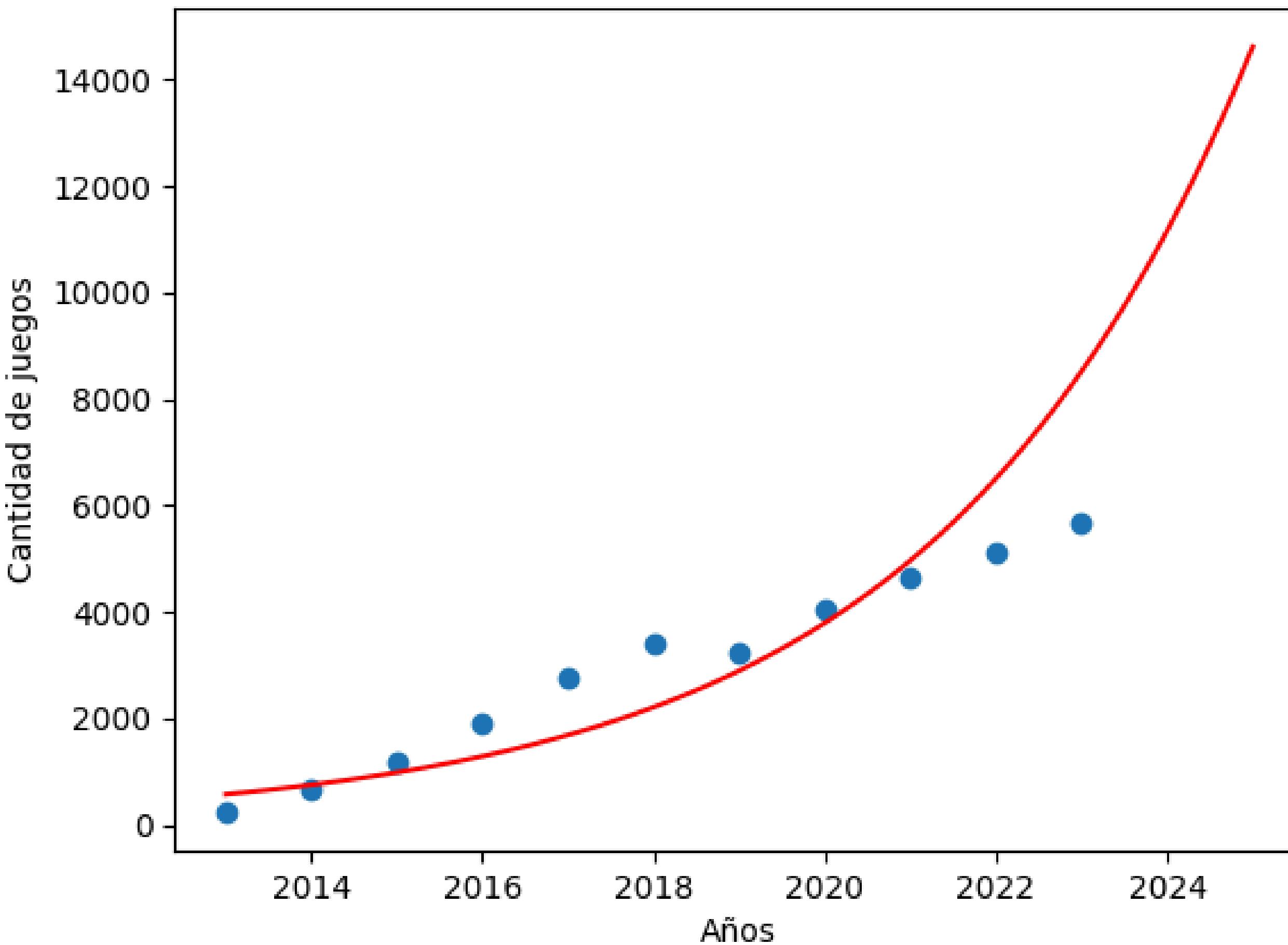
$R^2=0.986$

Regresión exponencial para la cantidad de juegos por género "Action" entre 1997 y 2025



$R^2=0.025$

Regresión exponencial para la cantidad de juegos por género "Action" entre 2013 y 2025



$R^2=0.594$

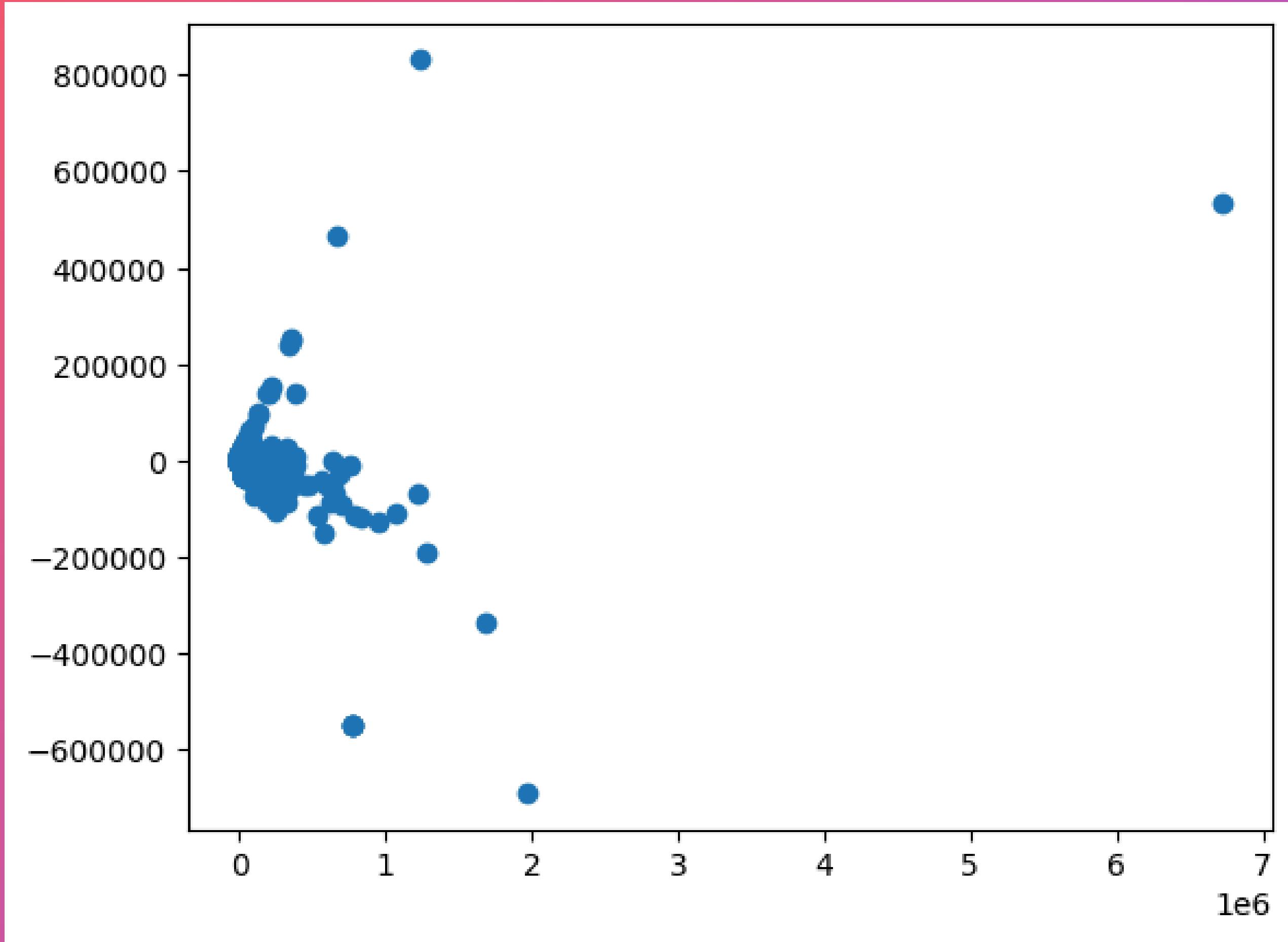
	Valor estimado para 2025	Error absoluto medio	Coeficiente de determinación
Regresión lineal	<u>6820.518</u>	<u>144.181</u>	<u>0.986</u>
Regresión exponencial	14614.965	789.846	0.594

PREGUNTA 3

¿Es posible identificar grupos de juegos (según precio o número de reviews positivas, por ejemplo) que ayuden a entender las características comunes de juegos exitosos o populares?

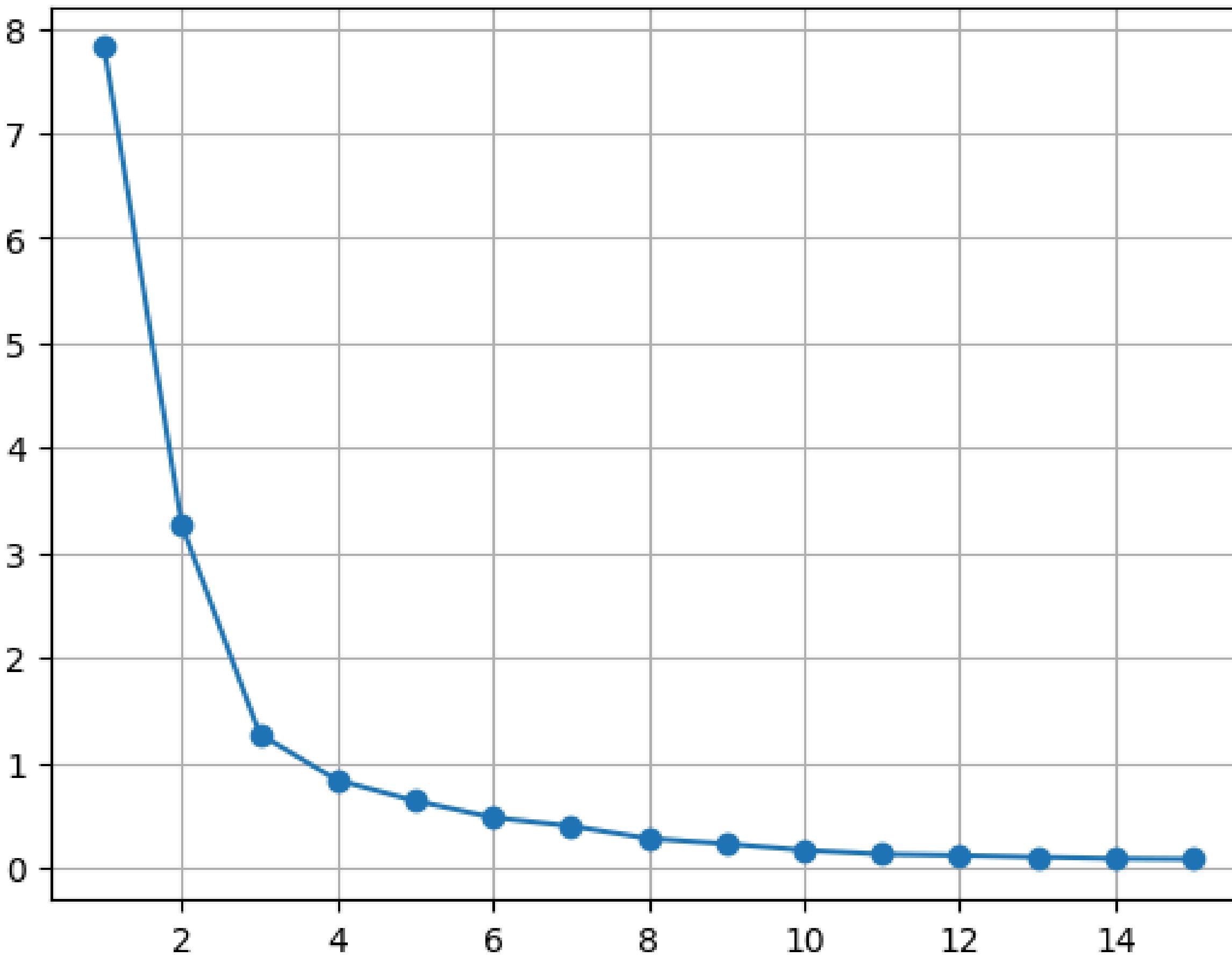
- Para el dataset, usamos los atributos que influyen a las ventas: '*price*' (precio del juego), '*positive*' (cantidad de reseñas positivas), '*negative*' (cantidad de reseñas negativas) y '*recommendations*' (recomendaciones totales).

- Aplicamos *PCA* a los datos, para obtener una visualización de los datos en dimensión reducida.
- Buscamos con el método del codo la cantidad de Clusters para el problema, y aplicamos algún clasificador para encontrar dichos Clusters: en este caso se usará *K-means*.

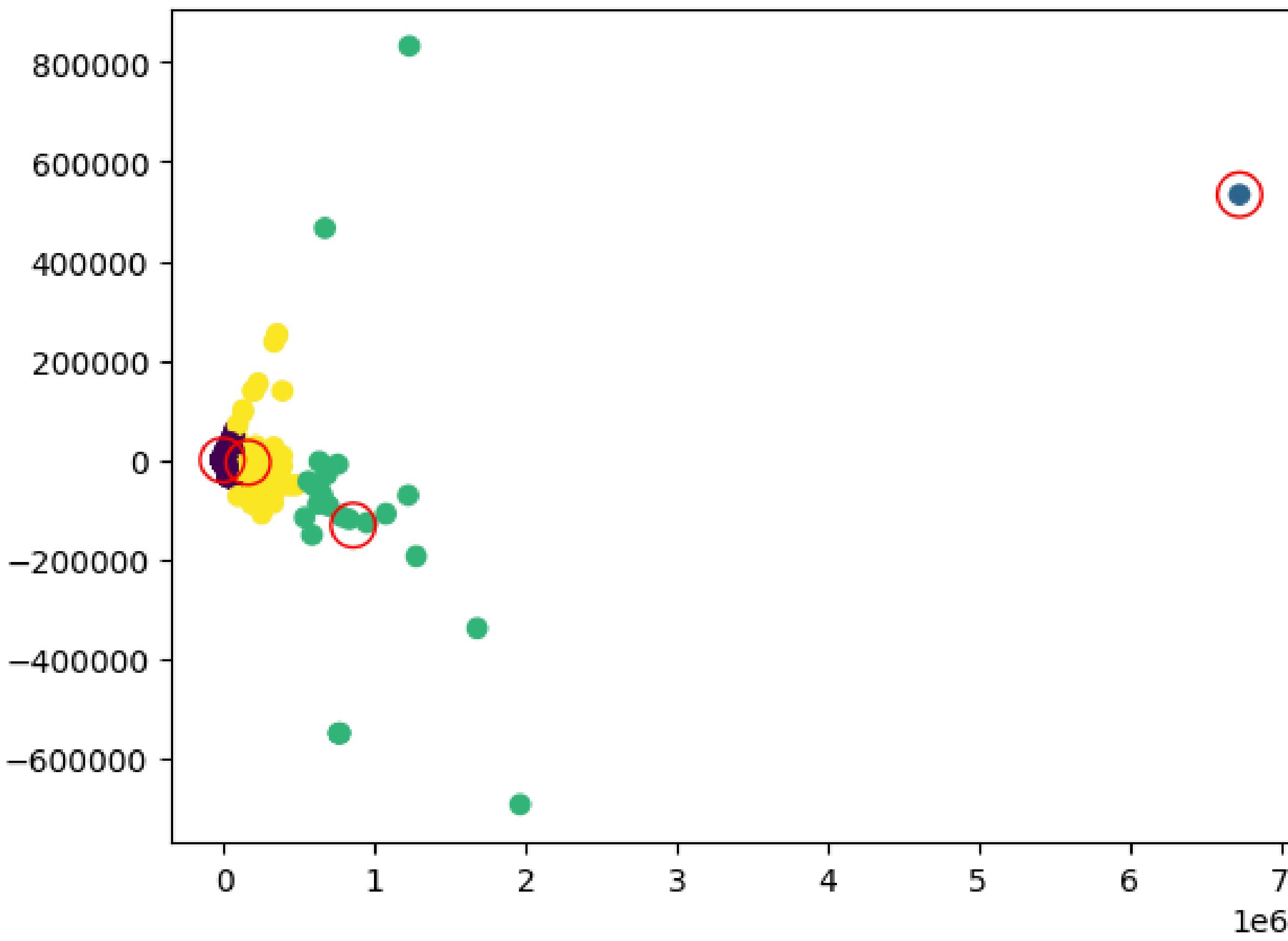


1e13

Metodo del codo, desde 1 hasta 15 clusters



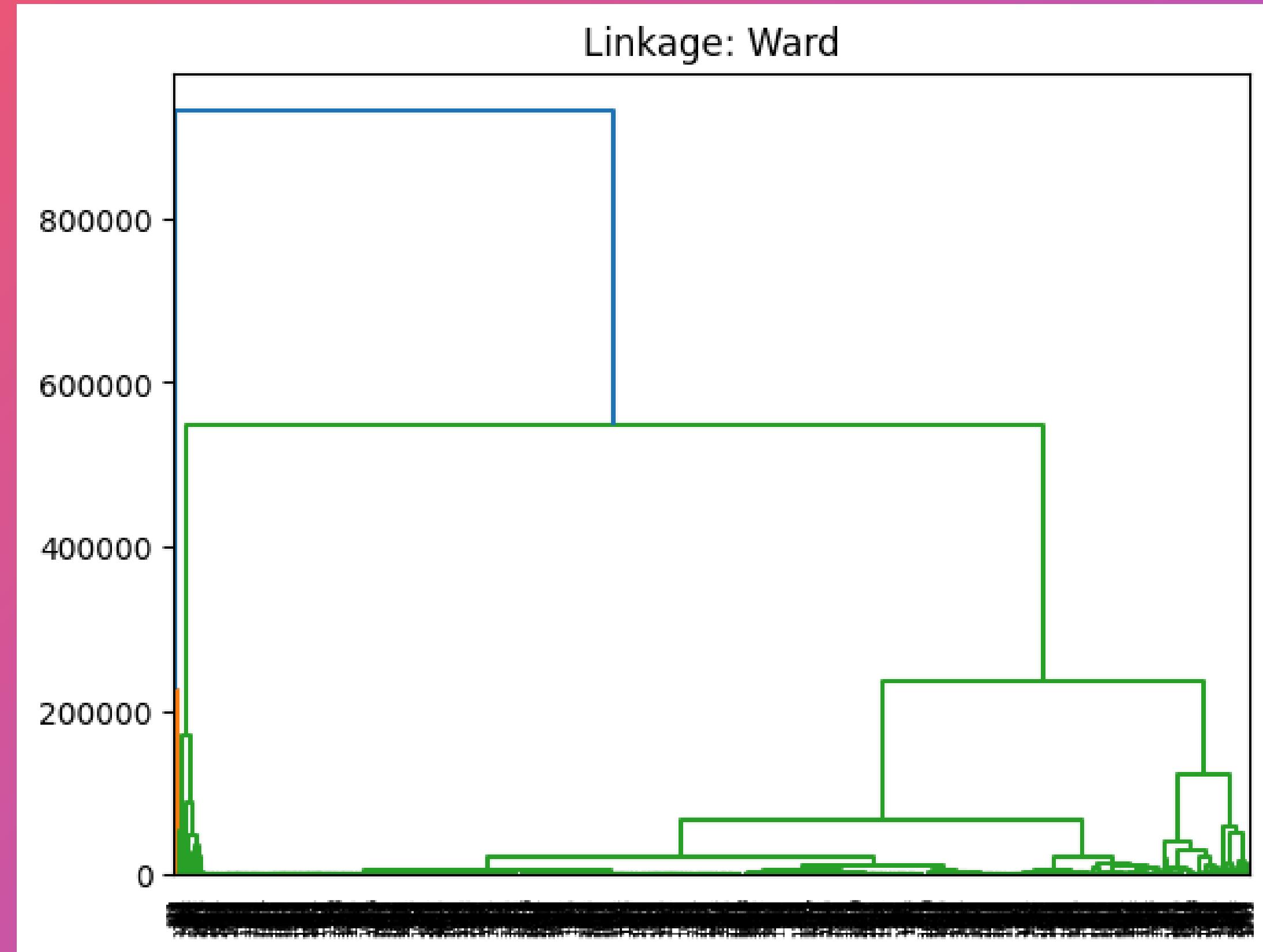
K-Means



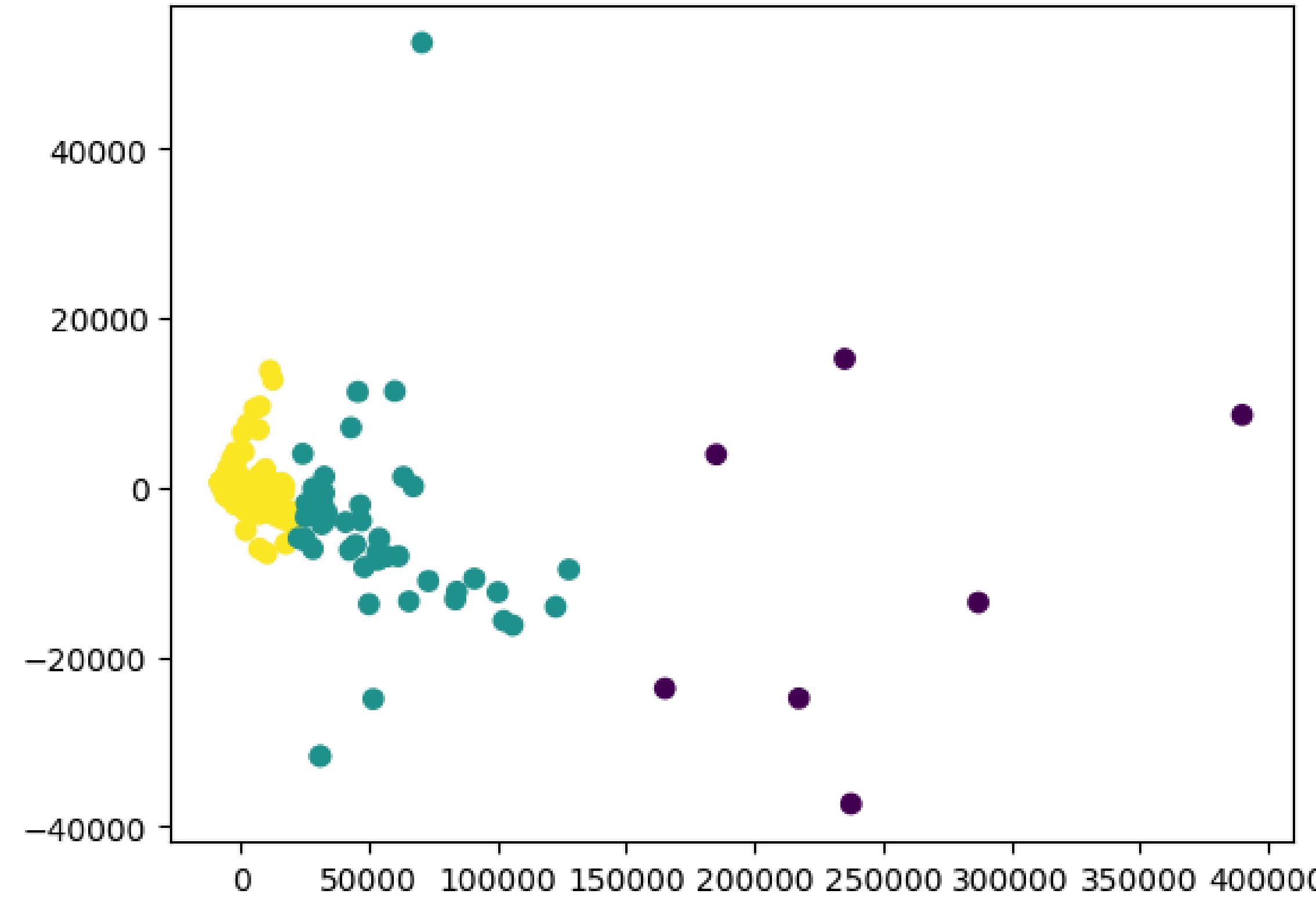
¿Qué podemos ver en estos *clusters*?

**Por la cantidad enorme de datos, se hizo un
sample del 10% para poder realizar
Agglomerative Clustering y *DBSCAN*.**

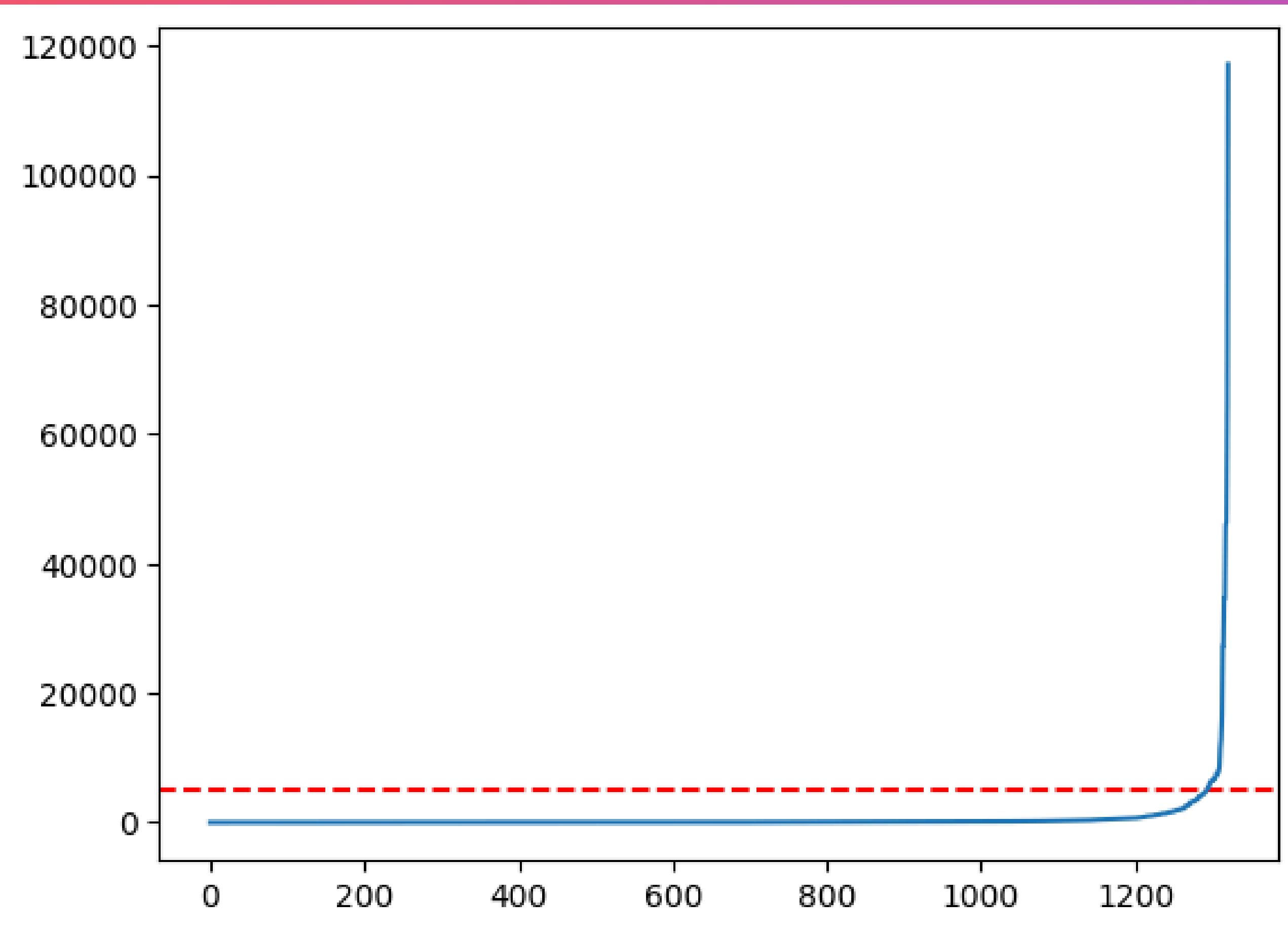
Linkage del sample



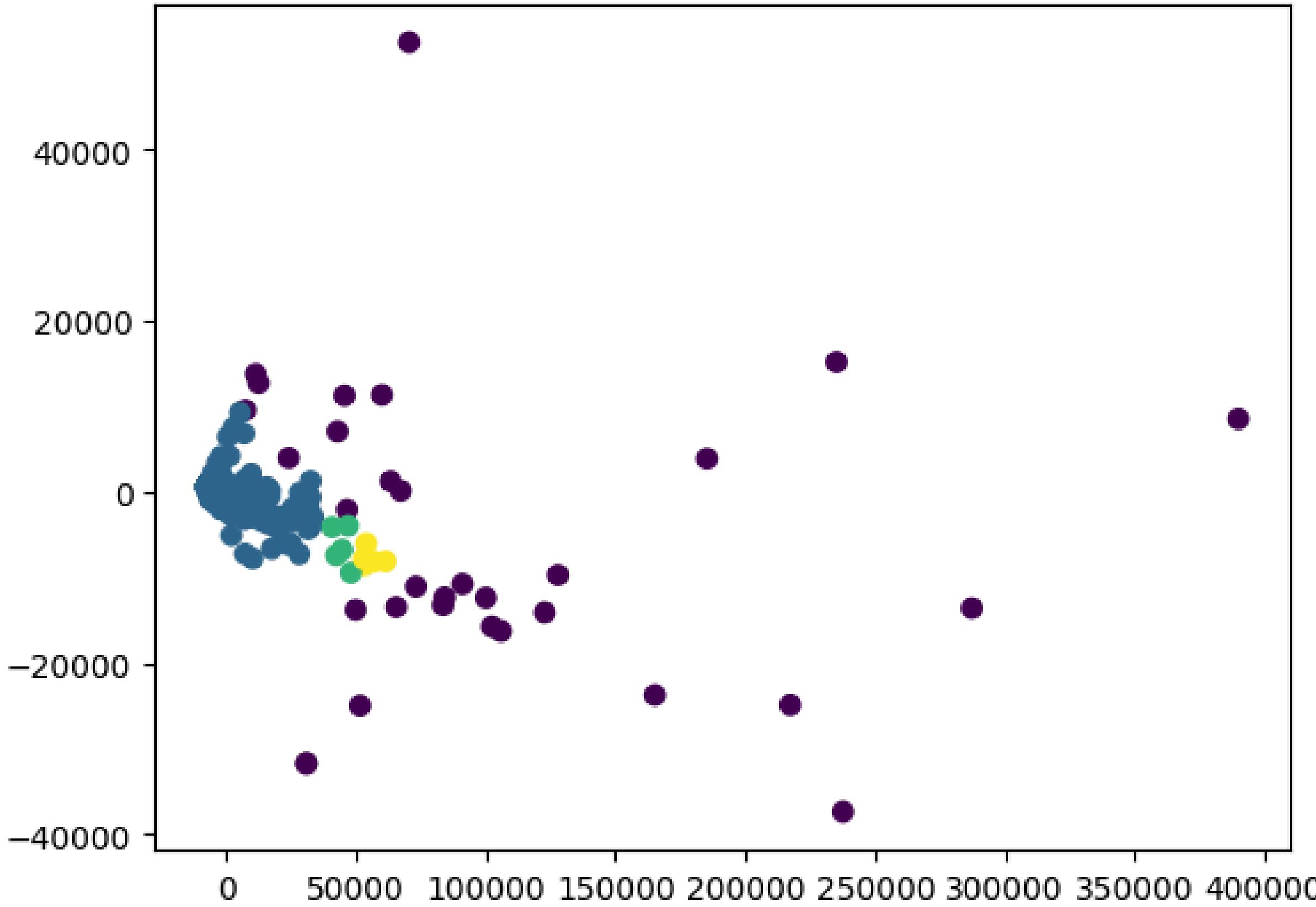
Hierarchical: ward, 3 clusters



**Luego, determinamos la rodilla para encontrar
un valor óptimo de *eps* para *DBSCAN*.**



DBSCAN: $\text{eps}=5000$, $\text{min_samples}=5$



Finalmente, evaluamos los *clusters* con el coeficiente de *Silhouette* para cada uno de los métodos usados.

Silhouette Score

X K-Means 4

0.955

X ward all

0.919

X DBSCAN

0.906

Analisamos los clusters K-means

Cluster: Emerging Hits

			name	main_genre	price	positive	negative	\
8874	Counter-Strike: Global Offensive			Action	0.0	5764420	766677	
recommendations								
8874		3441592						

Cluster: Niche Favorites

			name	main_genre	price	positive	negative	\
241		Garry's Mod		Indie	9.99	822326	29004	
559	Tom Clancy's Rainbow Six® Siege			Action	19.99	312232	64137	
829	Tom Clancy's Rainbow Six® Siege			Action	19.99	312816	64201	
1148	ARK: Survival Evolved			Action	29.99	461567	98701	
1572	Cyberpunk 2077			RPG	59.99	391643	129925	

recommendations

241	725462
559	899435
829	899455
1148	435328
1572	458744

Cluster: Top Sellers

		name	main_genre	price	positive	negative	\
10		Far Cry® 5	Action	59.99	100620	25286	
11		Forza Horizon 4	Racing	59.99	122539	15095	
19		Oxygen Not Included	Indie	24.99	82902	3014	
129		Apex Legends™	Action	0.00	415524	66608	
149		American Truck Simulator	Indie	19.99	104521	3859	

recommendations

10	114588
11	126316
19	80467
129	1000
149	87888

Cluster: Underperformers

		name	main_genre	price	positive	\
0		WARSAW	Indie	23.99	589	
1		Alien Breed 3: Descent	Action	9.99	349	
2		Hero of the Kingdom II	Adventure	7.99	2046	
3		Aerofly FS 2 Flight Simulator	Action	37.49	1490	
4		Kanjozoku Game レーサー	Massively Multiplayer	5.99	392	

negative recommendations

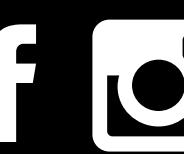
0	212	427
1	134	285
2	120	1615
3	408	1831
4	57	493

EL FUTURO

- Usar clasificadores para predecir precios acorde a las características de un juego.
- Encontrar alguna otra metodología para responder a la primera pregunta.
- Abordar el tema desde la perspectiva de un consumidor.



ELE



/ The L Group

