EDA Steam

Memoria

Juan Casas Lopez

Obtención de datos

Obtuve los datos de un dataset disponible en kaggle: (https://www.kaggle.com/datasets/fronkongames/steam-games-dataset).

El dataset fue a su vez obtenido a partir de la API pública de Steam. Los datos, a fecha de elaboración del EDA, fueron actualizados por última vez el 12 de enero de 2024.

El dataset está disponible tanto en formato CSV como JSON, para la elaboración del EDA utilizo el CSV.

Librerías utilizadas

Empleo numpy, pandas, seaborn, matplotlib, wordcloud y datetime.

Dataset

El dataset contiene 40 columnas y 85103 entradas. De ellas me quedo con 27 columnas para continuar con la elaboración del EDA, y reseteo índices en orden de AppID.

A continuación, creo columnas nuevas interesantes para utilizar más adelante en el análisis.

El dataset incluye además de juegos datos de programas también ofrecidos a través de Steam, para filtrarlas creo una columna nueva asignando 1 a aquellas filas que en el campo de categorías incluya la palabra “Software” y elimino dichas filas.

La columna “Estimated owners” incluye valores de rangos de propietarios, p.ej. 500 000 – 1 000 000. Creo una nueva columna “Owners” con el valor medio del rango.

Creo una columna “Ratio” que es el ratio entre críticas positivas y negativas de cada juego.

Creo una columna “OS3” donde un valor 1 indica soporte para los 3 sistemas operativos suportados (Windows, macOS, Linux)

Creo una columna “Free” con aquellos juegos cuyo precio sea 0,00 $

Transformo los valores de la columna “Release Date” a datetime, en origen se encuentran en múltiples formatos de fecha diferentes.

Creo una vista para poder analizar por separado a los juegos más populares del total del conjunto, separo por número de propietarios superior a 500 000.

Análisis de idiomas soportados

La columna “Supported languages” incluye una lista con las localizaciones disponibles para cada uno de los juegos, con 11 176 valores diferentes.

Algunos de los idiomas tienen más de una localización, como por ejemplo Español – España y Español – Latinoamérica.

Creo un wordcloud a partir de los datos y en él se puede observar que los idiomas más comunes parecen ser inglés, chino, español, francés, alemán y ruso.

Para comprobar el valor exacto de juegos que incluyen soporte para un idioma creo una columna de valores 1 y 0 en función de si el nombre del idioma aparece al menos una vez en la lista de localizaciones disponibles.

Con ello puedo observar que el 95,6% de los juegos incluyen soporte para inglés, el 25,1% para chino (al menos una localización sea en chino tradicional o simplificado), el 22,2% para francés y el 21,7% para español.

Sistemas operativos

En base a las columnas con datos booleanos sobre soporte de cada uno de los 3 sistemas operativos donde está disponible Steam (Windows, macOS y Linux), puedo observar que el 100% de los juegos incluyen soporte para Windows, un 20,2% para macOS y un 13,8% para Linux.

Analizando los juegos con más de 500 000 propietarios aumenta el porcentaje con soporte para los otros sistemas operativos, subiendo a un 35,3% para macOS y un 25,9% para Linux.

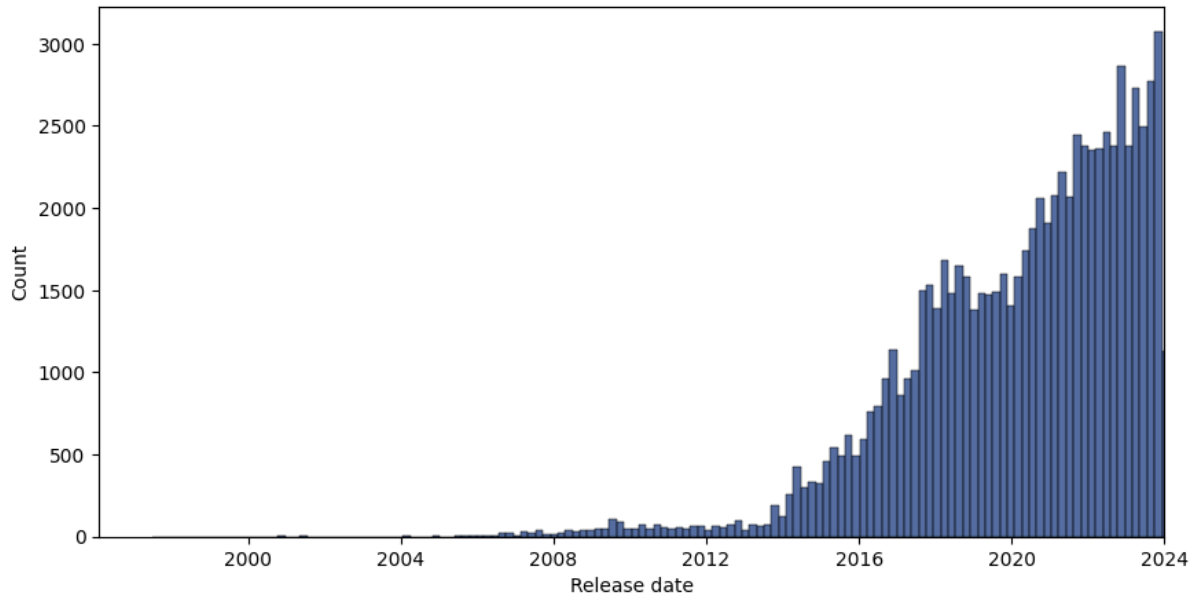
Juegos más populares

En base a los datos de número de propietarios estimado los juegos más populares son:

- Con entre 100 y 200 millones:
  - DOTA 2
- Con entre 50 y 100 millones (en ningún orden en particular):
  - Team Fortress 2
  - Counter-Strike: Global Offensive
  - PUBG: BATTLEGROUNDS
  - New World

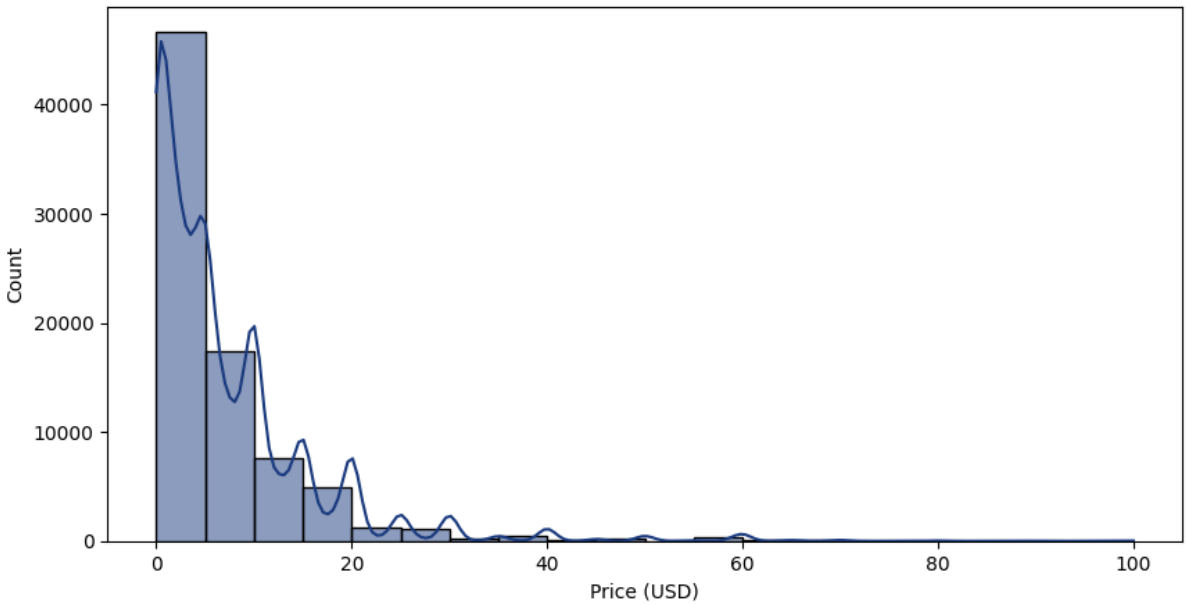
Fecha de lanzamiento

Analizo la distribución de los juegos en base a su fecha de lanzamiento. Donde se puede apreciar el fuerte crecimiento de la plataforma en los últimos años

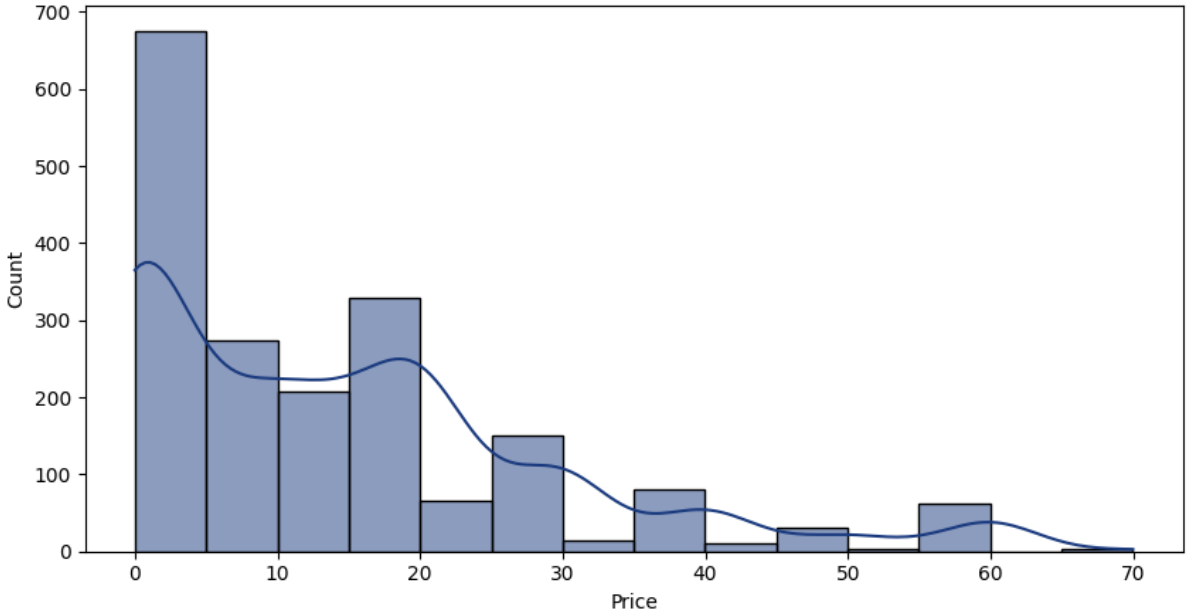


Precio

Analizo la distribución de los juegos en base a su precio en el mercado estadounidense. Se puede apreciar cómo hay una agrupación de los precios al final de cada bloque de 5 $, efecto de los precios terminados en 4.99 y 9.99.



Analizando por separados los juegos con número estimado de propietarios superior a 500 000, la distribución es similar a la general:



En cuanto al porcentaje de los juegos que son gratuitos, el porcentaje general es del 15,7% subiendo notablemente al 24,8% entre aquellos con un número estimado de propietarios superior a 500 000.

Categorías y géneros

En las columnas de categorías y de géneros se encuentran separadas con comas todas las categorías y géneros aplicables a cada uno de los juegos.

Creo dos wordclouds cada uno para representar cada una de las dos columnas.

Desarrolladores y Distribuidoras

Busco las distribuidoras y los desarrolladores más populares entre los juegos con más de 500 000 propietarios estimados

Análisis de columnas numéricas

Creo una vista con las columnas con objetos de tipo entero, booleano y datetime. En una matriz de correlación no se encuentra correlación lineal entre las diferentes columnas, más allá de la correlación debida a la popularidad de un juego; es decir, los juegos con más propietarios estimados tienen a su vez más críticas, una cifra mayor de pico de jugadores concurrentes, etc.

Análisis de hipótesis

"Los juegos multijugador retienen a los jugadores durante más horas a lo largo del tiempo"

Para comprobar la hipótesis planteada creo dos columnas que representan la aparición de la expresión "Multi-player" y "Single-player", respectivamente, en la columna "Categories".

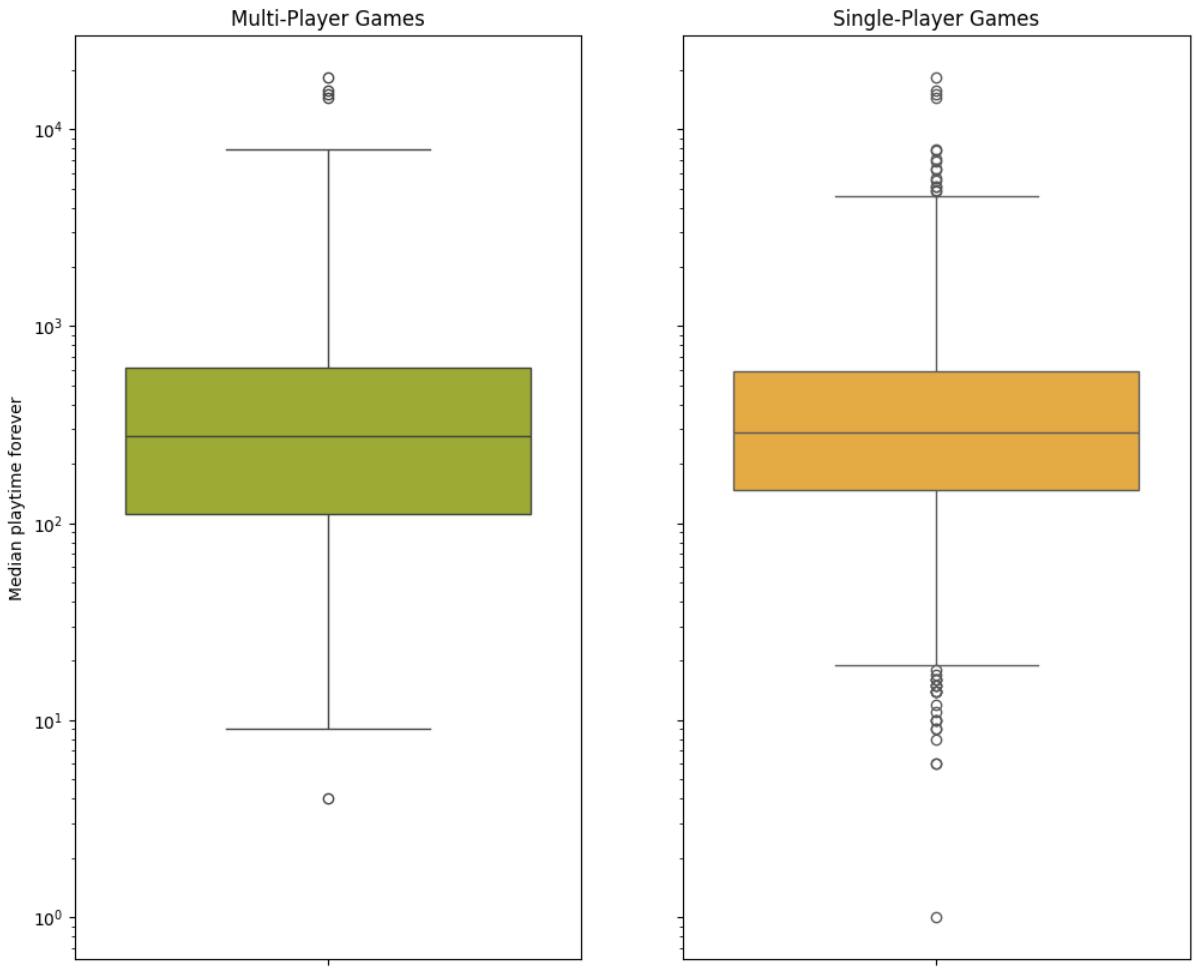
Comparo entonces los valores de "Median playtime forever" en cada una de las dos vistas. La diferencia entre ambas es mínima y no representa una diferencia notable entre los dos subgrupos, negando por tanto la hipótesis planteada.

Juegos multijugador:

- P50: 278
- Media: 617
- Desviación típica: 1278

Juegos de un jugador:

- P50: 287
- Media: 572
- Desviación típica: 1091


Multi-Player Games
Single-Player Games
Median playtime forever
10^4
10^3
10^2
10^1
10^0