
Benemérita Universidad Autónoma de Puebla

Facultad de ciencias de la computación (FCC)

# Tendencias populares
# En la industria de los videojuegos

Introducción a la ciencia de datos

Responsable: Manzanarez Peña Victor Hugo

Docente: Jaime Alejandro Romero Sierra

# ÍNDICE


## 1. Portada

## 2. Introducción

- 2.1 Objetivo Principal
- 2.2 ¿Por qué es importante este proyecto?
- 2.3 Fuentes de Datos

## 3. Metodología

- 3.1 Descripción de Columnas del Dataset
- 3.2 Proceso de Limpieza de Datos
    - 3.2.1 Revisión de Datos Faltantes
    - 3.2.2 Detección y Manejo de Duplicados
    - 3.2.3 Corrección de Valores Atípicos o Inconsistentes
    - Columna Rank
    - Columna Name
    - Columnas de Ventas
    - Columna Publisher y Genre
    - Columnas Restantes (Year y Platform)
    - 3.2.4 Cambio de Nombres de Columnas

## 4. Análisis Exploratorio de Datos (EDA)

- 4.1 Descripción General de los Datos
- 4.2 Tipos de Variables
- 4.3 Resumen Estadístico
    - 4.3.1 Variables Numéricas
    - 4.3.2 Variables Categóricas
- 4.4 Visualización y Distribución de Variables Individuales
- 4.5 Análisis de Correlación
- 4.6 Análisis de Valores Atípicos (Outliers)
- 4.7 Comparación entre Variables Numéricas y Categóricas
- 4.8 Hallazgos Clave e Implicaciones para el Modelo

## 5. Modelo de Machine Learning

- 5.1 Descripción del Modelo
- 5.2 Justificación de Modelos Seleccionados
- 5.3 Implementación y Entrenamiento
    - 5.3.1 División de Datos
    - 5.3.2 Preparación de Variables (X e y)
    - 5.3.3 Entrenamiento de Modelos
- 5.4 Resultados y Evaluación
    - 5.4.1 Métricas de Regresión Lineal
    - 5.4.2 Métricas de Random Forest
- 5.5 Comparación de Modelos
- 5.6 Conclusiones del Análisis Predictivo

## 6. Dashboard

- 6.1 Visualizaciones Principales
- 6.2 Métricas del Modelo
- 6.3 Insights y Conclusiones

## 7. Conclusiones Generales

# Tendencias populares en la industria de los videojuegos

Saludos cordiales al lector. Hoy día se encuentra redactado en estas páginas el análisis completo del proyecto de introducción a ciencia de datos en el cual se analizan las tendencias populares en la industria de los videojuegos.

## Introducción

El objetivo principal de este proyecto es Identificar que géneros, plataformas y regiones son más rentables en la industria de los videojuegos. Permitiendo a las desarrolladoras tomar decisiones estratégicas en base a datos históricos de ventas y así aumentar el margen de éxito.

## ¿Por qué es importante este proyecto?

La investigación busca resolver diferentes problemáticas a las cuales se enfrentan las desarrolladoras a la hora de crear y lanzar un videojuego tales como lo pueden ser el público objetivo, si el género es el adecuado, plataformas en las cuales lanzar el videojuego, regiones con mayor probabilidad de éxito, entre otras. Con este análisis se busca resolver esas dudas y otorgar un panorama más claro y extenso en base al cual se puedan tomar decisiones que influyan positivamente en las ventas y aceptación del videojuego.

## Fuentes de datos

La única y principal base de en la cual se basa esta investigación es un data set llamado “Video game sales” proveniente de la página “Kaggle” la cual contiene información histórica sobre ventas de videojuegos a nivel global con aproximadamente 16598 filas y 11 columnas (datos previos a la limpieza). La principal característica de esta base es contar únicamente con registros de ventas de videojuegos de al menos 10,000 copias.

# METODOLOGIA

Antes de continuar, se hace una breve mención de las columnas de nuestro data set y que es lo que contienen:

- 1.-Rank: Ranking en base a las ventas totales
- 2.-Name: Nombre del videojuego
- 3.-Platform: Plataforma en la cual se lanzó el videojuego
- 4.-Year: Año en el que se lanzó el videojuego
- 5.-Genre: Genero del videojuego
- 6.-Publisher: editorial del videojuego
- 7.-NA_Sales: Ventas en Norteamérica (en millones)
- 8.-EU_Sales: Ventas en Europa (en millones)
- 9.-JP_Sales: Ventas en Japón (en millones)
- 10.-Other_Sales: Ventas en el resto del mundo (en millones)
- 11.-Global_Sales: Ventas mundiales totales

A continuación se hace un breve pero concisa explicación sobre cómo se limpiaron los datos, tipos de datos ausentes y su manejo y eliminación de duplicados en cada columna de nuestro data set:

**1.- Revisión de datos faltantes**

Lo primero que se comprobó al iniciar la limpieza, fue la cantidad de datos nulos en la base de datos, para lo cual se hizo uso del comando df.isnull().sum(), para posteriormente utilizar el comando df.info() para conocer el tipo de dato de cada columna.

**2.- Detección y manejo de duplicados**

Para la limpieza de esos datos simplemente se utilizó la función drop_duplicates(). Una cosa importante es que si bien logro deshacerse de una cantidad importante de datos duplicados, no logro deshacerse de todos, puesto que algunos de ellos contenían datos nulos (NaN) en algunas de sus columnas, pero esos datos los eliminamos más adelante.

**3.- Corrección de valores atípicos o inconsistentes**

Para este caso se utilizaron diferentes soluciones según la columna que se limpió, a continuación una breve descripción de lo que se hizo con los datos atípicos en cada columna:

**Columna Rank**: Se identificaron los datos atípicos y posteriormente, al no ser un dato crucial para el análisis (lo que significa que el registro no pierde valor en caso de no tener el dato) se reemplazaron por datos NaN para seguidamente rellenarlos con el número 0.

**Columna Name**: Para esta columna se aplicó una estrategia similar (ubicar los datos atípicos y reemplazarlos por datos nulos), sin embargo al ser esta columna de SUMA importancia para el análisis (puesto que sin este dato, el registro queda inutilizado) se tomó la decisión de usar el comando dropna para eliminar todos esos registros.

**Columnas de ventas (NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales):** La limpieza de estas columnas fue un caso particular ya que se pudieron recuperar la mayoría de los datos nulos y atípicos. Al ser estas columnas directamente influenciadas por las demás, se optó por obtener los datos mediante cálculos simples. Lo que se hizo con estas columnas fue lo siguiente: Se eliminaron todos los registros que contenían más de 1 dato nulo en alguna de estas columnas (porque de otra forma, el dato no se puede calcular) y posteriormente se obtuvieron los registros faltantes mediante fórmulas básicas de despeje (A + X = B, se despeja X= B-A)

**Columna Publisher y columna Genre**:

(Se incluirán ambas columnas en una sección ya que se utilizó la misma técnica para ambas.)

Directamente no podemos obtener este dato, sin embargo al haber juegos que se lanzaron varias veces en diferentes plataformas, podemos obtener ambos datos de esos registros. Pero evidentemente no todos los registros contaban con más de un lanzamiento, entonces lo que se hizo fue llenar los datos faltantes en base a los que ya se tenía y dropear a los que no se podían recuperar.

**Columnas restantes (Year y platform)**: Ambas columnas presentaban el mismo problema: los datos contenidos son de suma importancia y no hay

forma de recuperarlos. Por lo tanto aunque en este caso la columna platform representaba menos del 5% de los datos y se podía dropear sin problemas, no sucedía lo mismo con la columna year. Ya que esta columna representaba cerca del 7% de los datos, pero al no haber una manera concreta de poder recuperarlos y además asegurarse de que estos datos fueran correctos, se tomó la decisión de eliminarlos todos (tanto los datos atípicos como los datos nulos).

**Cambio de nombres a columnas:** Si bien esto no fue un paso de limpieza como tal, para mejor manejo y entendimiento de los datos, se tradujeron los nombres de las columnas al español, quedando de la siguiente manera:

- 1.-Rank: Rank
- 2.-Name: Nombre
- 3.-Platform: Plataforma
- 4.-Year: Año
- 5.-Genre: Genero
- 6.-Publisher: Editorial
- 7.-NA_Sales: Ventas_NA
- 8.-EU_Sales: Ventas_EU
- 9.-JP_Sales: Ventas_JP
- 10.-Other_Sales: Ventas_Otras
- 11.-Global_Sales: Ventas_Globales

# Análisis Exploratorio de Datos (EDA)

Para comenzar con nuestro análisis exploratorio de datos, veamos una descripción general de los datos.

El data set cuenta con 14896 registros y 11 variables

Los tipos de datos de cada columna son los siguientes:

- Rank:  Numérica
- Nombre: Categórica
- Plataforma: Categórica
- Año: Numérica
- Género: Categórica
- Editorial:  Categórica
- Ventas_NA:  Decimal
- Ventas_EU:  Decimal
- Ventas_JP: Decimal
- Ventas_Otras: Decimal
- Ventas_Globales: Decimal

A continuación se presenta un resumen estadístico por columna mediante tablas:

Columnas numéricas

| | Rank | Año | Ventas_NA | Ventas_EU | Ventas_JP | Ventas_Otras | Ventas_Globales |
|---|---|---|---|---|---|---|---|
| **Media** | 7823.165951 | 2006.429914 | 0.263392 | 0.146748 | 0.076108 | 0.048136 | 0.534616 |
| **Mediana** | 7777.50 | 2007 | 0.08 | 0.02 | 0.00 | 0.01 | 0.17 |
| **STD** | 5013.547284 | 5.823479 | 0.760085 | 0.453306 | 0.300959 | 0.178331 | 1.416856 |
| **Min** | 0 | 1980 | -0.010000 | -0.010000 | -0.010000 | -0.010000 | -0.010000 |
| **Max** | 16599 | 2020 | 29.08 | 12.88 | 10.22 | 10.57 | 40.24 |

Columnas categóricas

(Se usó una tabla por columna a excepción de las columnas nombre y Editorial, ya que cada una de estas contienen 9845 y 545 datos respectivamente)

**Columna Género**:

| Genero | Frecuencia |
|---|---|
| **Action** | 2985 |
| **Sports** | 2107 |
| **Misc** | 1544 |
| **Role-Playing** | 1346 |
| **Shooter** | 1179 |
| **Racing** | 1152 |
| **Adventure** | 1130 |
| **Platform** | 801 |
| **Simulation** | 770 |
| **Fighting** | 767 |
| **Strategy** | 600 |
| **Puzzle** | 515 |

**Columna Plataforma:**

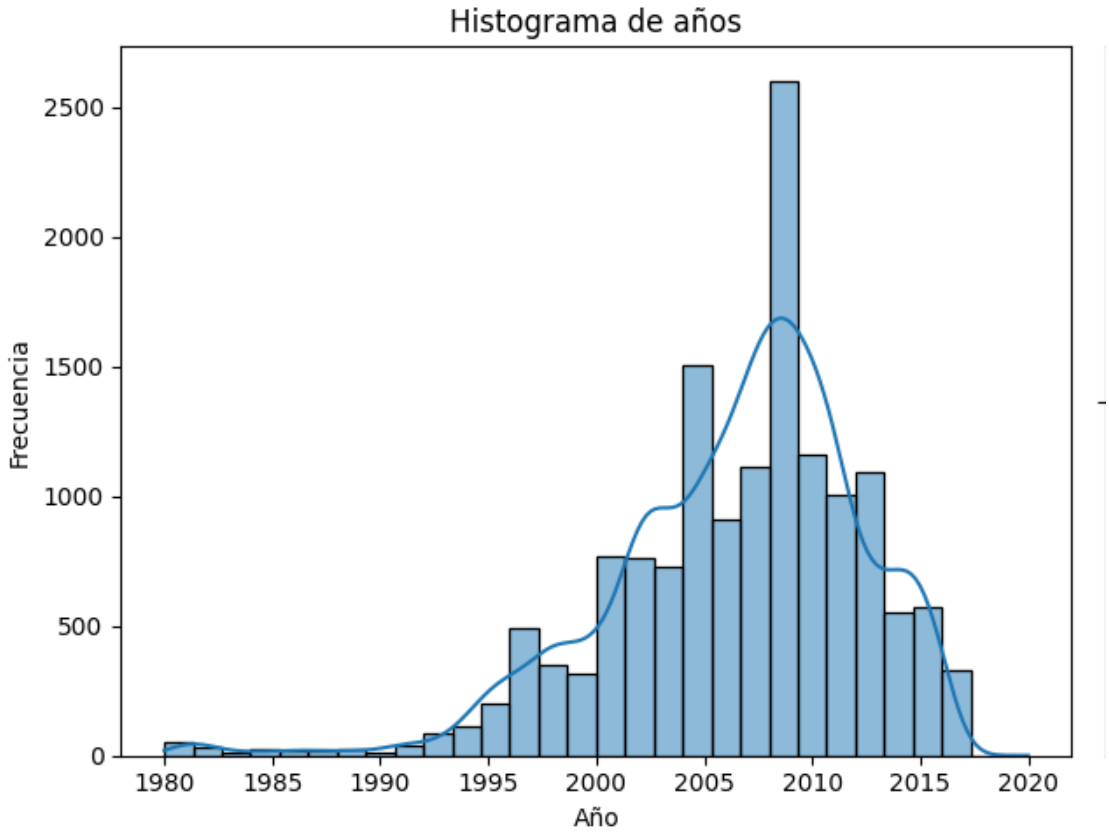
| Plataforma | Frecuencia | Plataforma | Frecuencia | Plataforma | Frecuencia |
|---|---|---|---|---|---|
| **PS2** | 1909 | **GC** | 518 | **GB** | 84 |
| **DS** | 1905 | **3DS** | 473 | **NES** | 83 |
| **PS3** | 1219 | **PSV** | 370 | **DC** | 49 |
| **Wii** | 1171 | **PS4** | 309 | **GEN** | 24 |
| **X360** | 1150 | **N64** | 289 | **NG** | 13 |
| **PS** | 1102 | **SNES** | 218 | **WS** | 6 |
| **PSP** | 1077 | **Xone** | 197 | **SCD** | 5 |
| **PC** | 850 | **SAT** | 164 | **TG16** | 3 |
| **XB** | 749 | **WiiU** | 134 | **3DO** | 2 |
| **GBA** | 715 | **2600** | 107 | **GG** | 1 |

## Visualización y Distribución de Variables Individuales

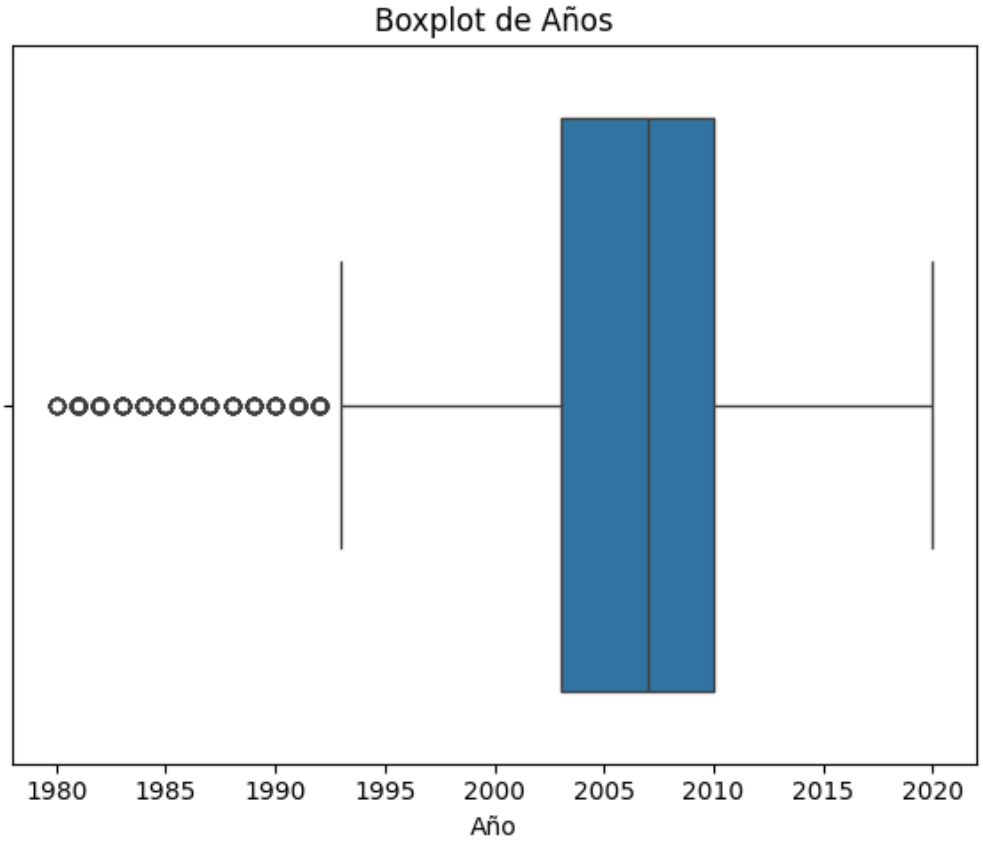
A continuación se presenta un análisis de las variables de forma individual mediante el uso de histogramas y boxplots para las variables numéricas y gráficos de barras para las variables categóricas.

# Variables numéricas

## Columna Año

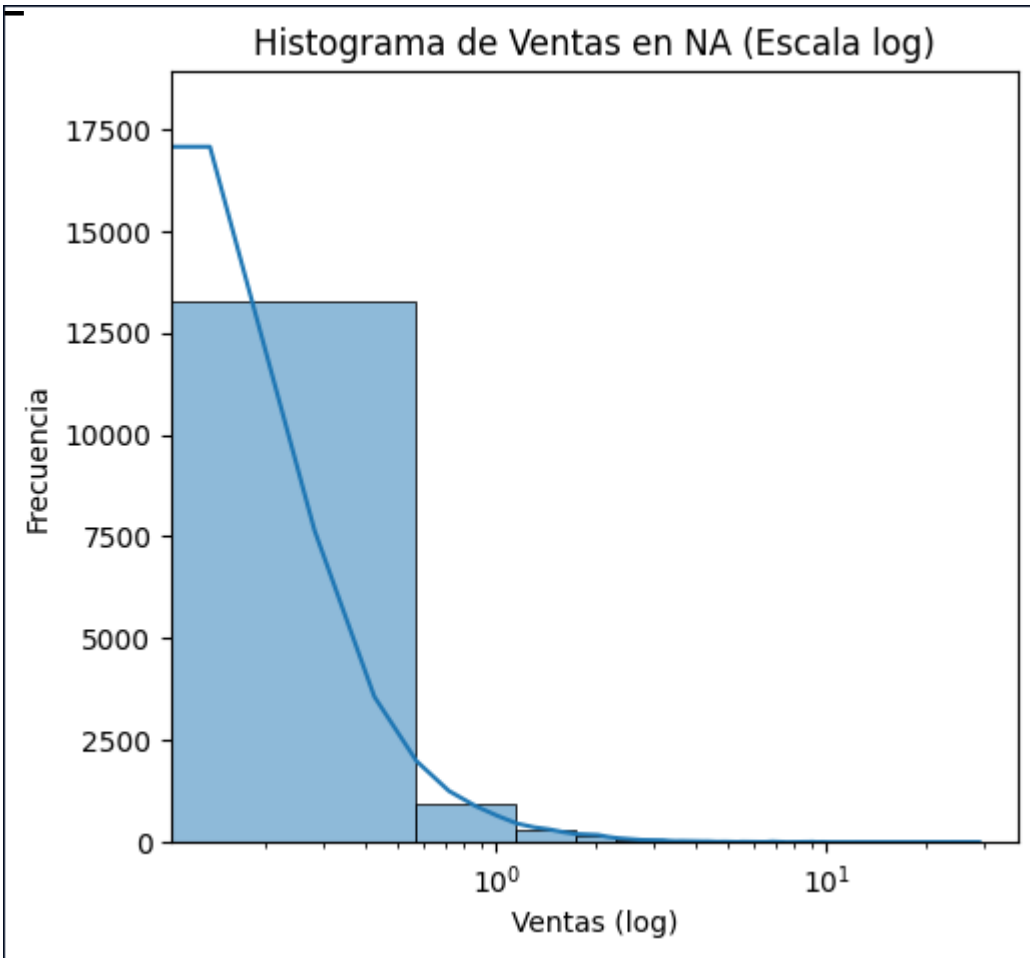



Mediante el siguiente histograma podemos observar una clara observación de lanzamientos entre los años 2000 y 2015, con el pico más alto entre 2008 y 2010 lo cual nos indica que la mayoría de los registros de lanzamientos de nuestro dataset provienen de esta década. La distribución esta sesgada hacia la derecha con menores lanzamientos a años posteriores a 2015 y muchos menos a años anteriores a 1995, lo que nos indica que los lanzamientos en esos años pueden no ser tan representativos.




De igual manera en el siguiente boxplot podemos observar la dispersión de los años e identificar datos atípicos. La mayoría de los lanzamientos se sitúan entre los años 2000 a 2015, sin embargo también se confirma que se cuentan con algunos lanzamientos muy antiguos. El rango intercuartilico está concentrado en años recientes, mostrando que la industria ha estado más activa en los últimos años.

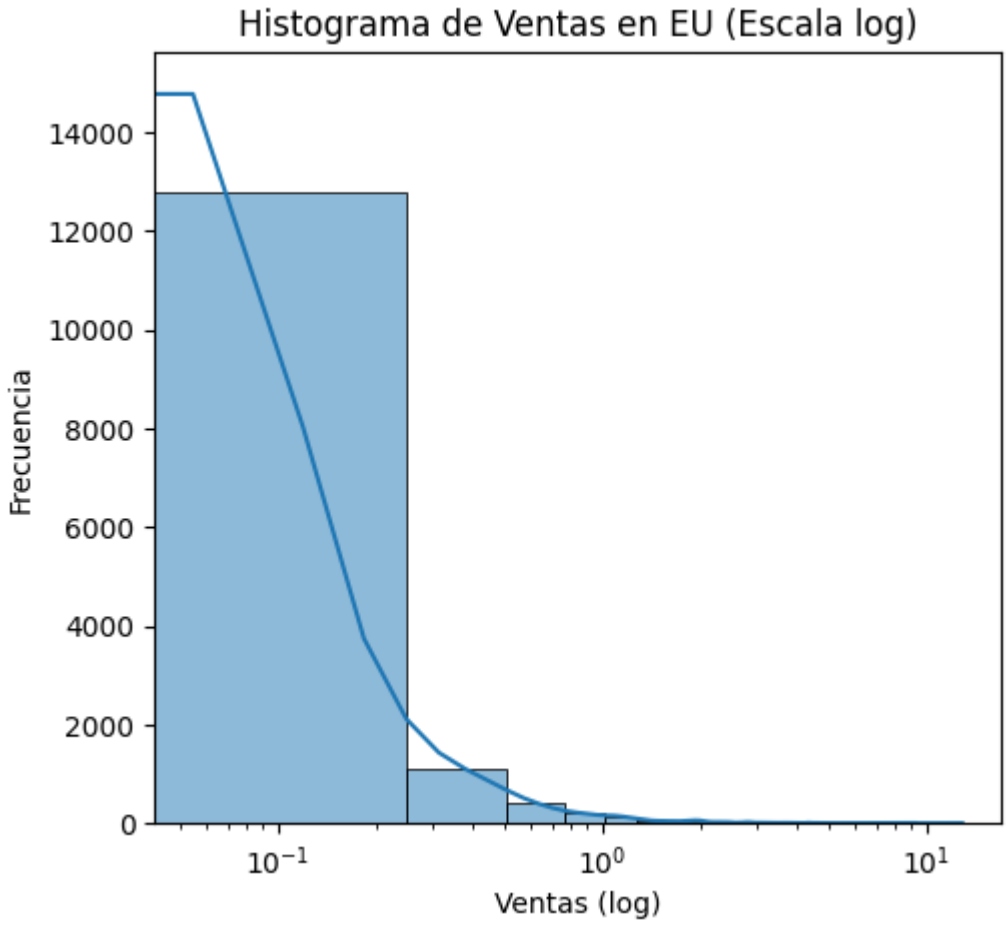
## Columna Ventas_NA




Para la representación de esta variable se utilizó una escala logarítmica y así poder ver todo con mayor claridad. El grafico muestra un muy claro sesgo hacia la derecha, concentrando la mayoría de los datos entre 0 y $10^0$ con muy pocos datos posteriores a estas cifras, lo cual nos indica que la mayoría de videojuegos en Norteamérica venden menos de 1 millón de copias y muy pocos superan esta cifra, dejándolos como valores atípicos.

## Columna Ventas_EU

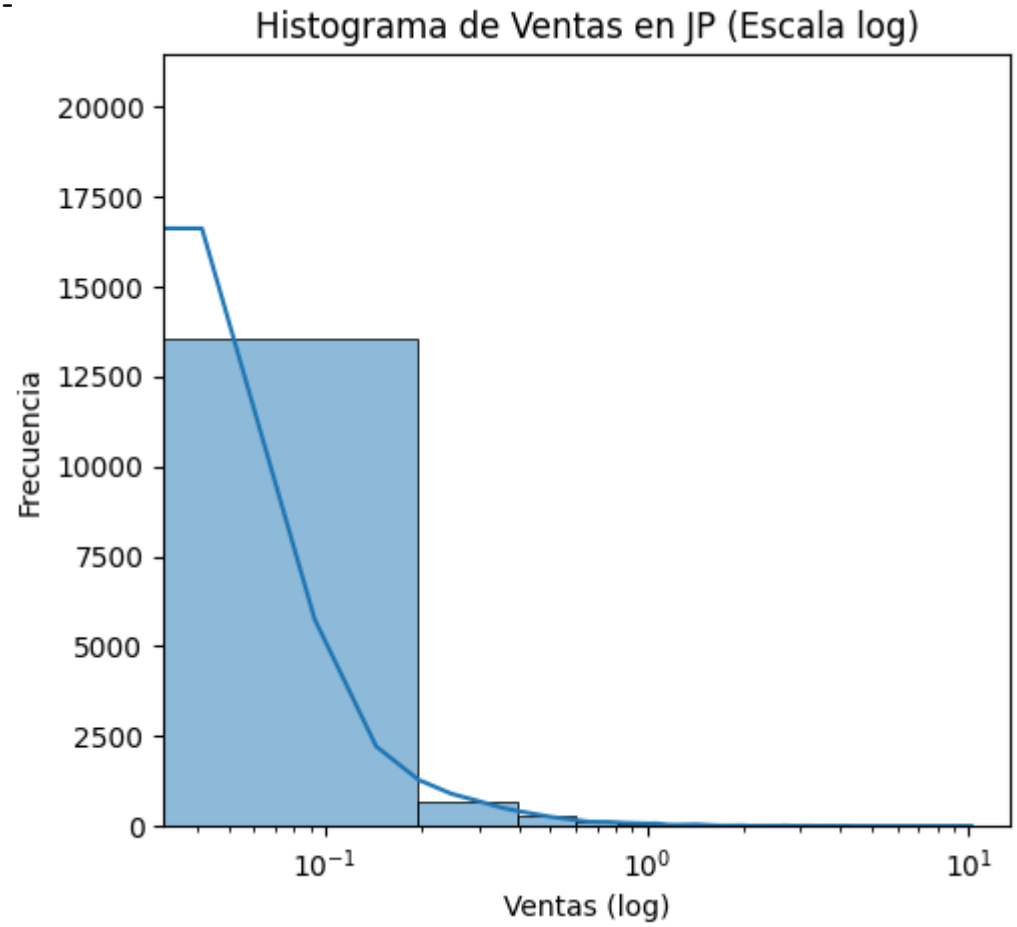
.




Nuevamente para las ventas en EU, se puede observar un sesgo hacia la derecha, sin embargo esta vez la mayoría de los datos se encuentran entre y $10^{-1}$y $10^{0}$, lo cual nos indica que en EU la mayoría de los videojuegos venden entre 100,000 y 1 millón de copias con una mínima cantidad de lanzamientos superando esta cifra, dejándolos de esta forma como datos atípicos. De igual forma el uso de la escala logarítmica permite visualizar mejor la dispersión de los datos.

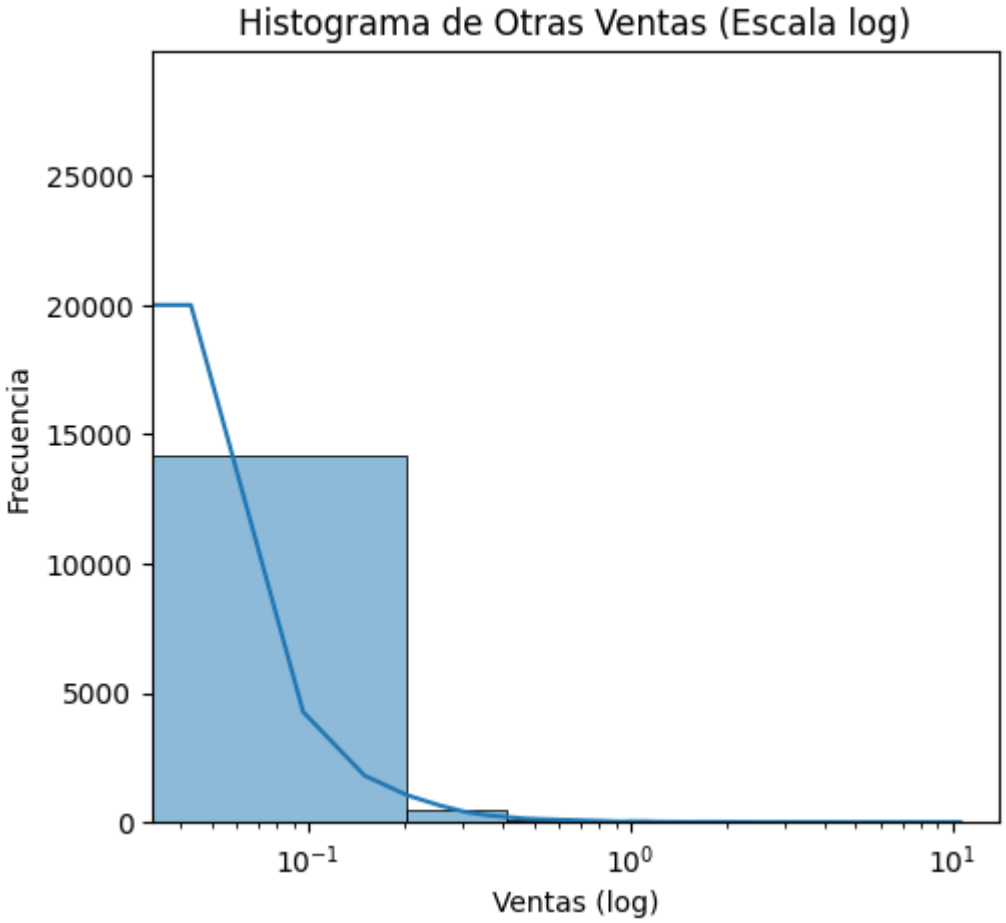
## Columna Ventas_JP

-



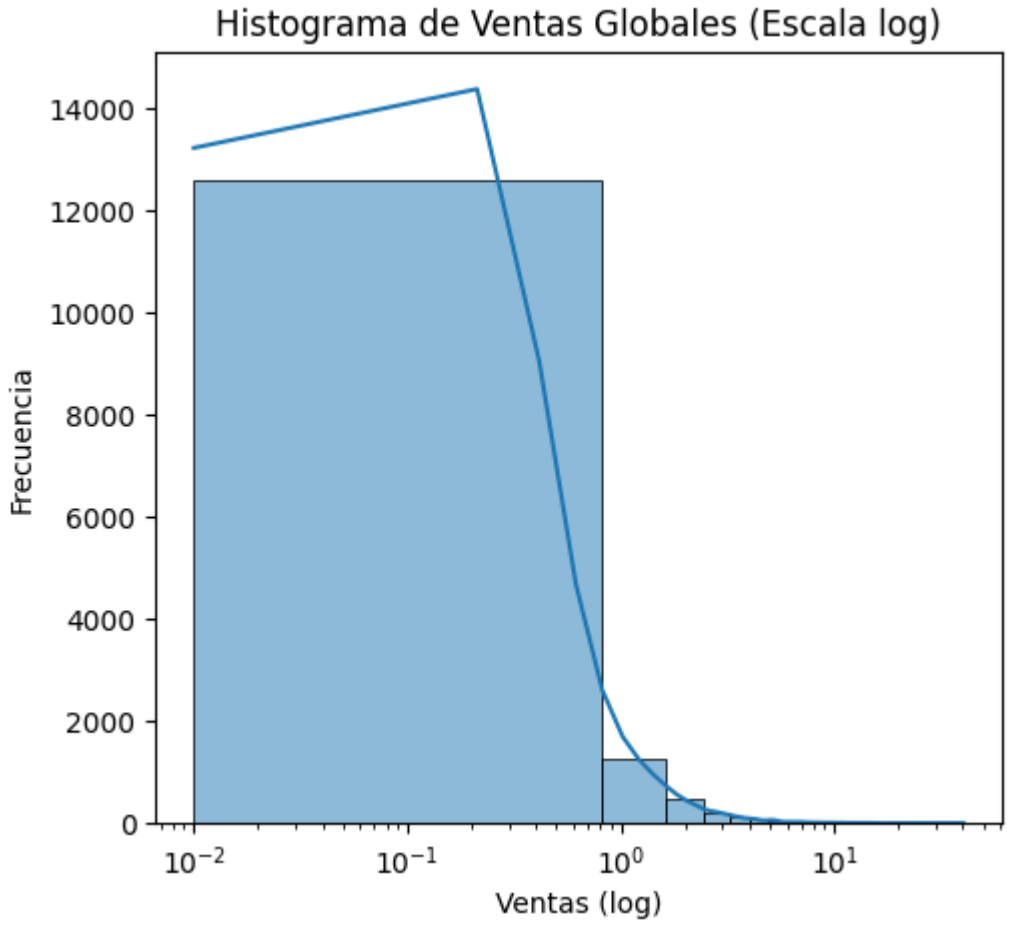

Continuando con la columna de ventas en JP, al igual que en las últimas 2 columnas se observa un sesgo hacia la derecha. Sin embargo esta vez la mayoría de los datos se encuentran entre 0 y $10^{-1}$, lo cual nos indica que la mayoría de los lanzamientos no superan las 100,000 unidades vendidas. De igual forma se empleó la escala logarítmica para poder observar mejor la distribución de los datos.

## Columna Ventas_Otras




La columna Ventas_Otras de la misma forma que sus columnas anteriores, muestra un sesgo a la derecha y similar a la columna recién analizada la mayoría de sus datos se encuentran entre 0 y $10^{-1}$, lo cual nuevamente nos indica que el resto de las ventas de los videojuegos fuera de Norteamérica, Europa y Japón no supera las 100,000 unidades vendidas con muy pequeños datos superando esta cifra y dejándolos como datos atípicos. Nuevamente el uso de la escala logarítmica nos permite ver con más claridad la distribución de los datos

## Columna Ventas_Globales

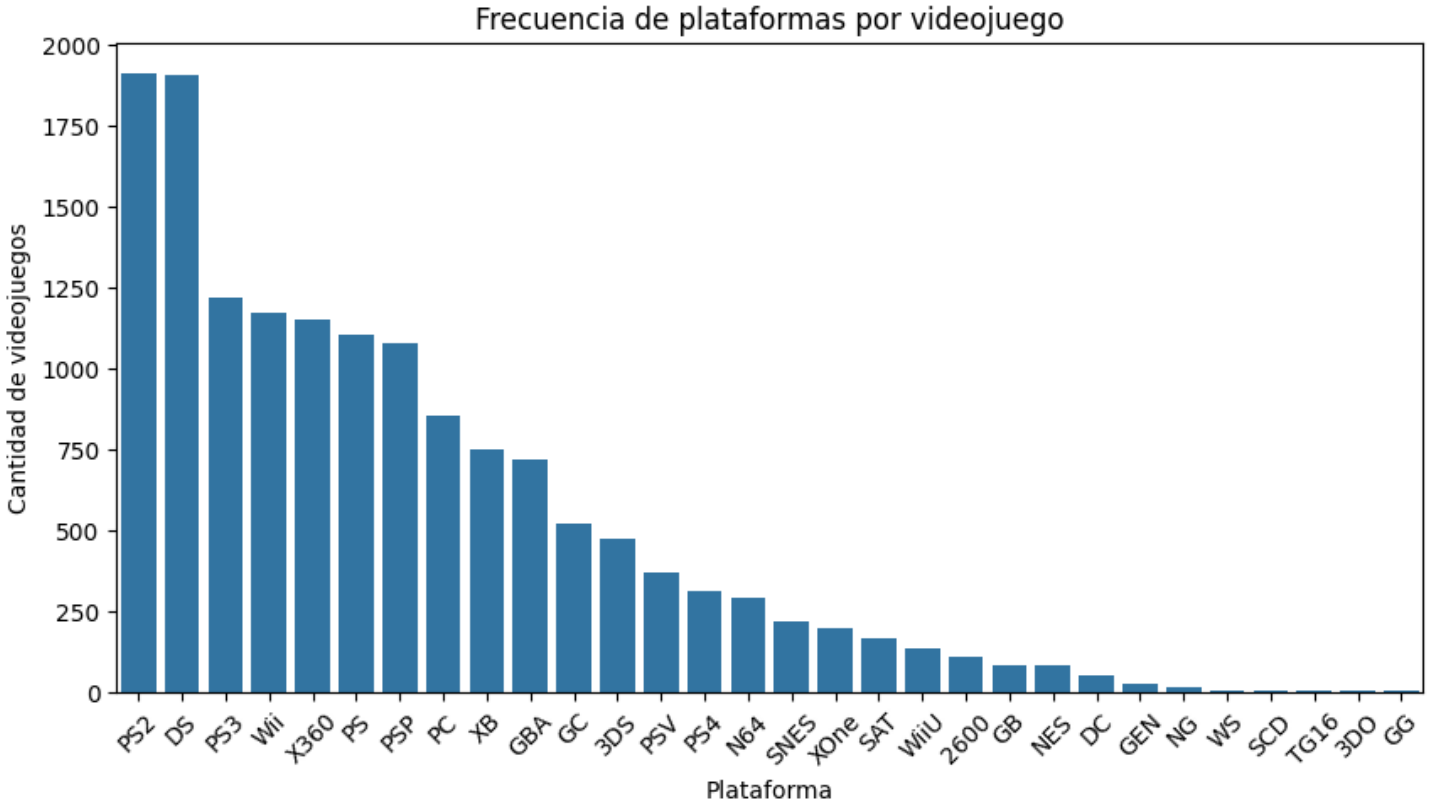



Para finalizar con las columnas numéricas tenemos la variable Ventas_Globales, que de igual manera muestra un sesgo hacia la derecha. Al ser un recuento de las ventas totales de un videojuego en todas las regiones, podemos ver claramente que la mayoría de datos se encuentra entre $10^{-2}$y $10^{0}$, es decir, la mayoría de los lanzamientos alcanzan mínimamente las 10,000 ventas a nivel global, mientras que las demás obtuvieron ventas similares o superiores.

# Variables categóricas

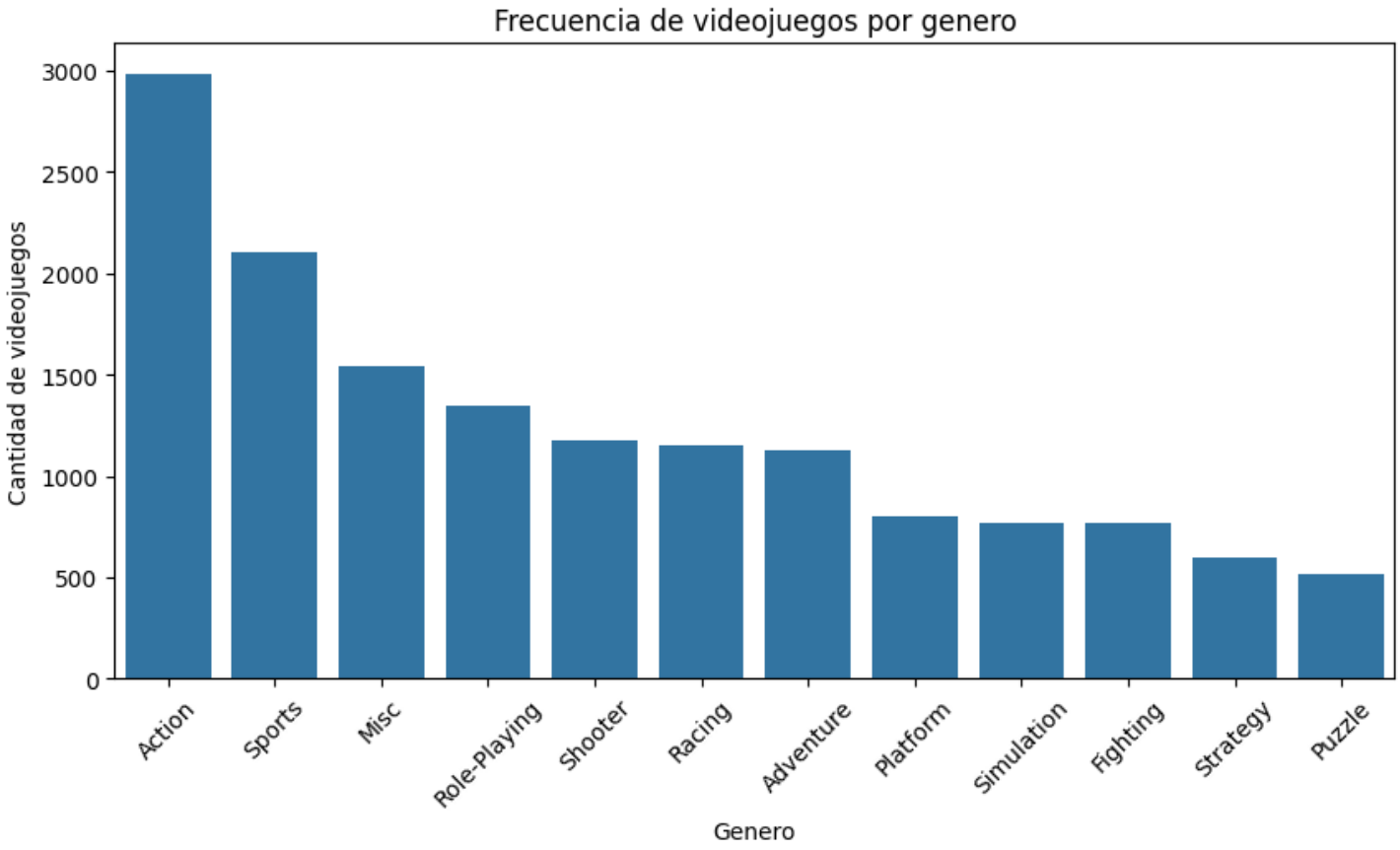
Nuevamente solamente se graficaron las variables Genero y plataforma por la misma razón anterior (las demás variables categóricas cuentan con cantidades de datos enormes mediante las cuales no puede obtenerse nada al ser graficadas).

## Columna Plataforma




Muy bien, en la siguiente grafica podemos observar claramente un superior dominio en cuanto a cantidad de lanzamientos por parte de 2 consolas: la Playstation 2 y el Nintendo DS. Lo cual hace total sentido ya que estas consolas tienen el primer y segundo lugar respectivamente en las consolas más vendidas de la historia. Seguidas posteriormente por la Playstation 3 y el Nintendo Wii y recorriendo asi de esta forma a Xbox hasta la quinta posición, mediante lo cual podemos inferir que las desarrolladoras prefieren lanzar sus videojuegos ya sea en consolas de Playstation o de Nintendo, relegando de esta forma al resto de consolas a posiciones extremadamente marginales (Sega, Atari, etc)

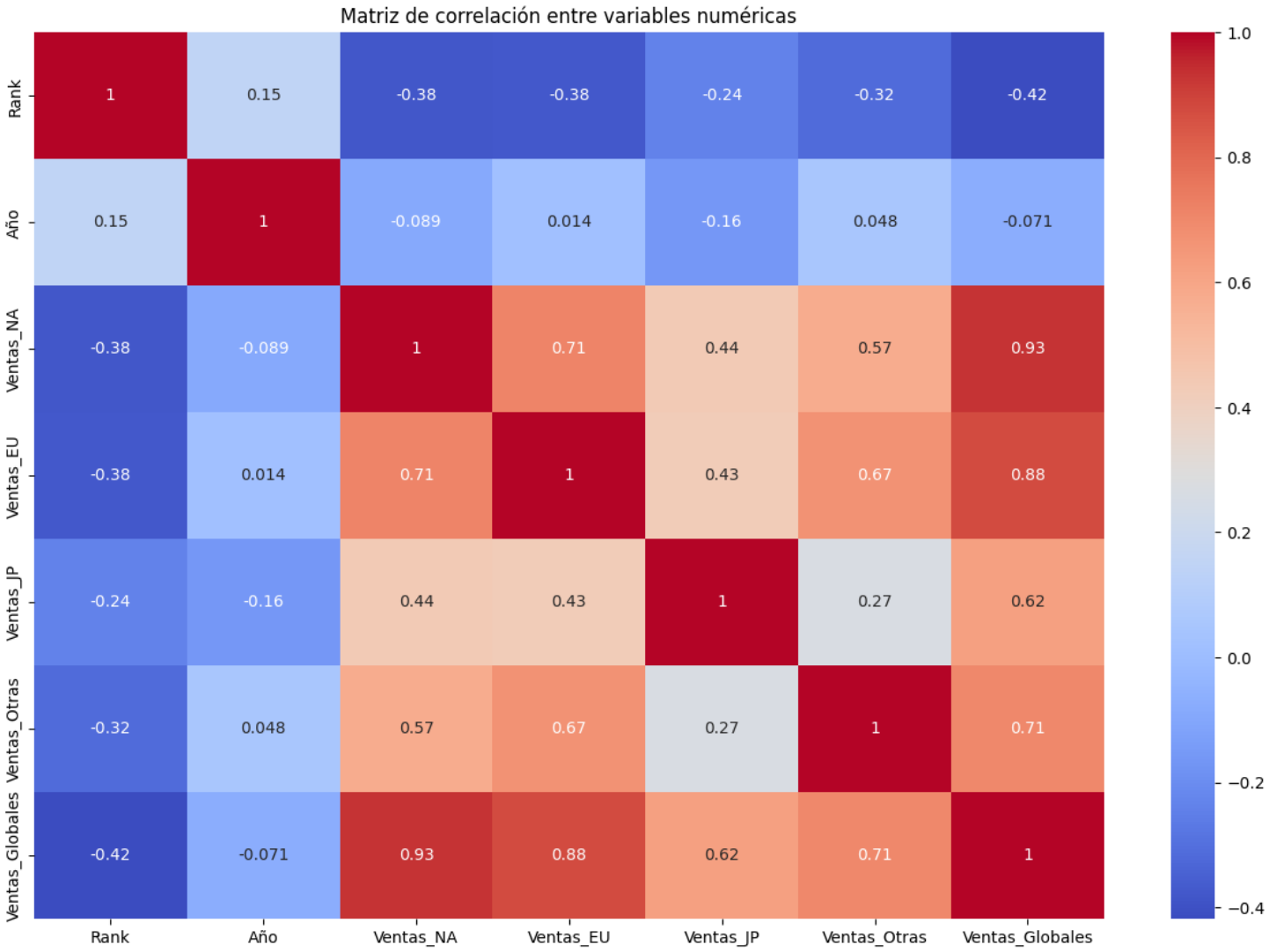
## Columna Género




Para finalizar con los análisis individuales a las variables tenemos la columna género en la cual tenemos a un claro dominante en cuanto a lanzamientos: El género de acción. Seguido por el género de deportes esto los convierte en los 2 géneros con más títulos, infiriendo que logran esta posición debido a la preferencia personal de los clientes por estos 2 roles y convirtiéndolos en los más populares. Dejando así al género strategy y puzzle en los últimos 2 lugares, lo que sugiere una menor popularidad entre los consumidores. Si bien se podría pensar que géneros con mayor presencia podrían ser los más rentables, es importante considerar que la frecuencia de lanzamientos no necesariamente se traduce en mayor éxito comercial. Ya que para inferir rentabilidad sería necesario analizar datos de ventas por género.

## Correlación entre Variables

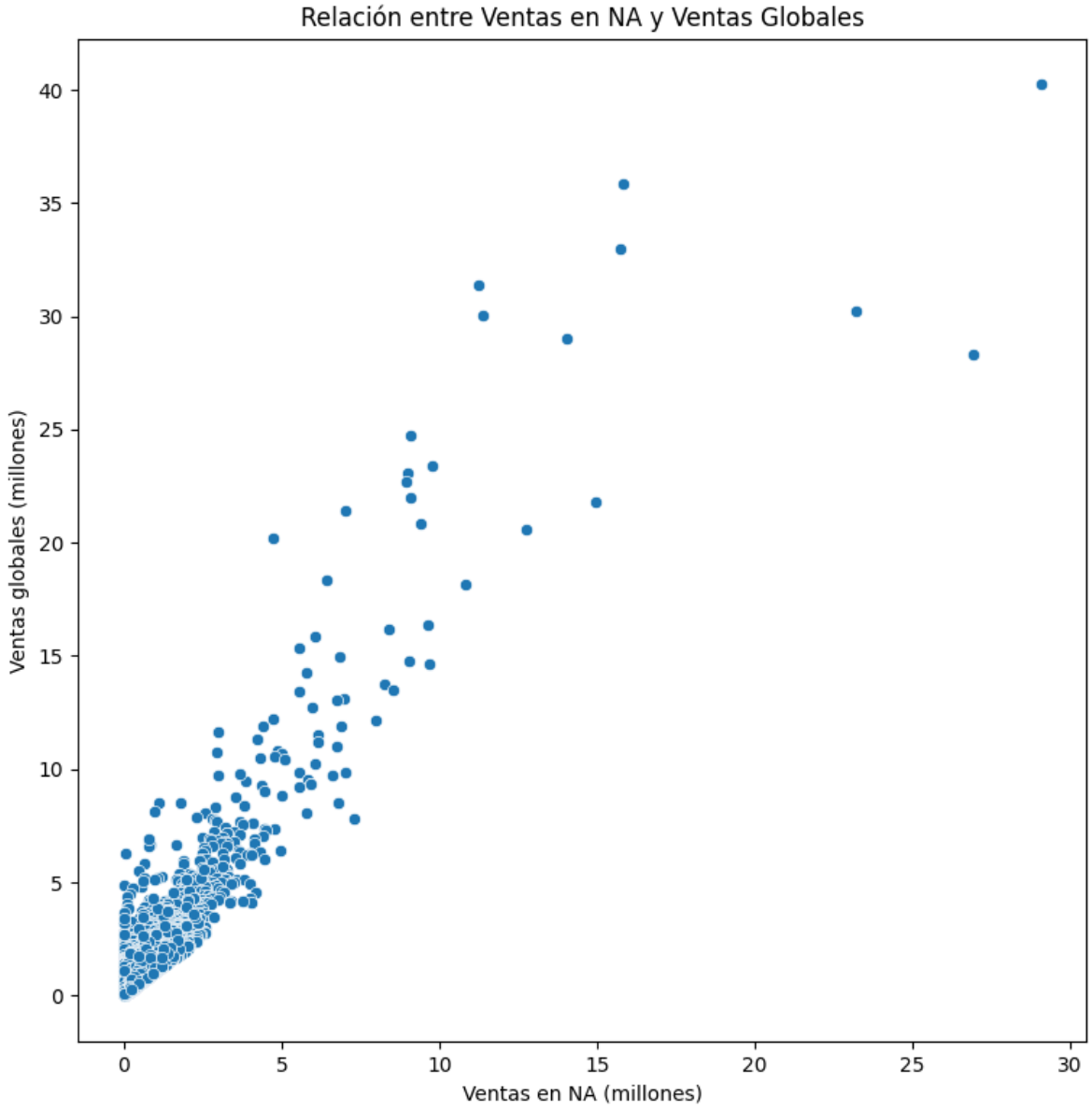
A continuación se muestra la matriz de correlación entre las variables numéricas y un análisis a las correlaciones que pueden ser útiles para el modelo:




Las variables de ventas por región (Norteamérica, Europa, otras) mantienen correlaciones fuertes entre sí y con las ventas globales, lo que sugiere que la mayoría de los videojuegos exitosos tienden a funcionar bien en múltiples mercados. Sin embargo, la correlación baja entre Ventas_JP y Ventas_Otras indica que el mercado japonés y el resto del mundo fuera de NA/EU están menos conectados comercialmente.

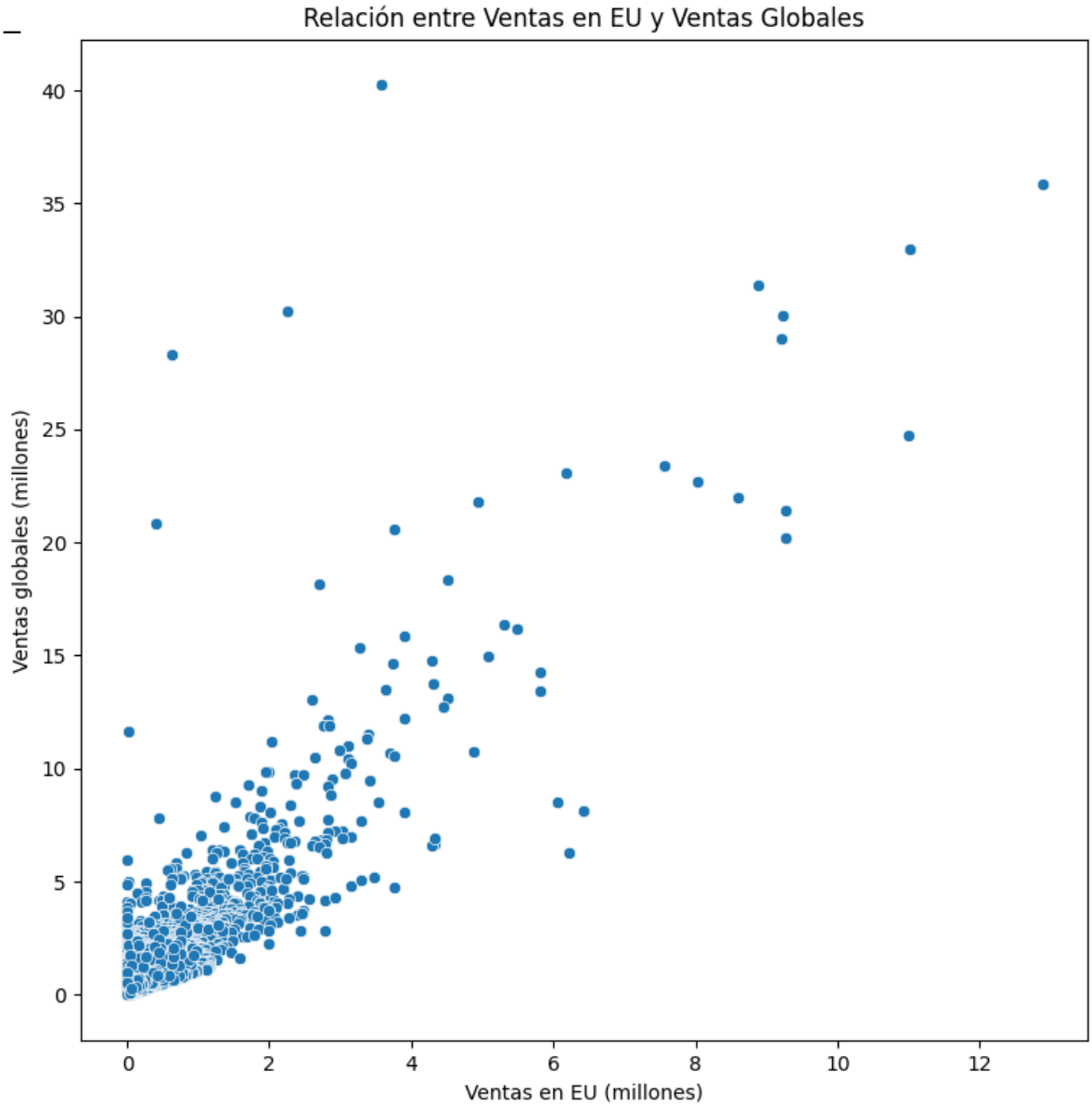
A continuación se muestra cada uno de los ScatterPlots de todas las variables relacionadas y que son útiles para nuestro análisis:

**Ventas NA y ventas Globales**



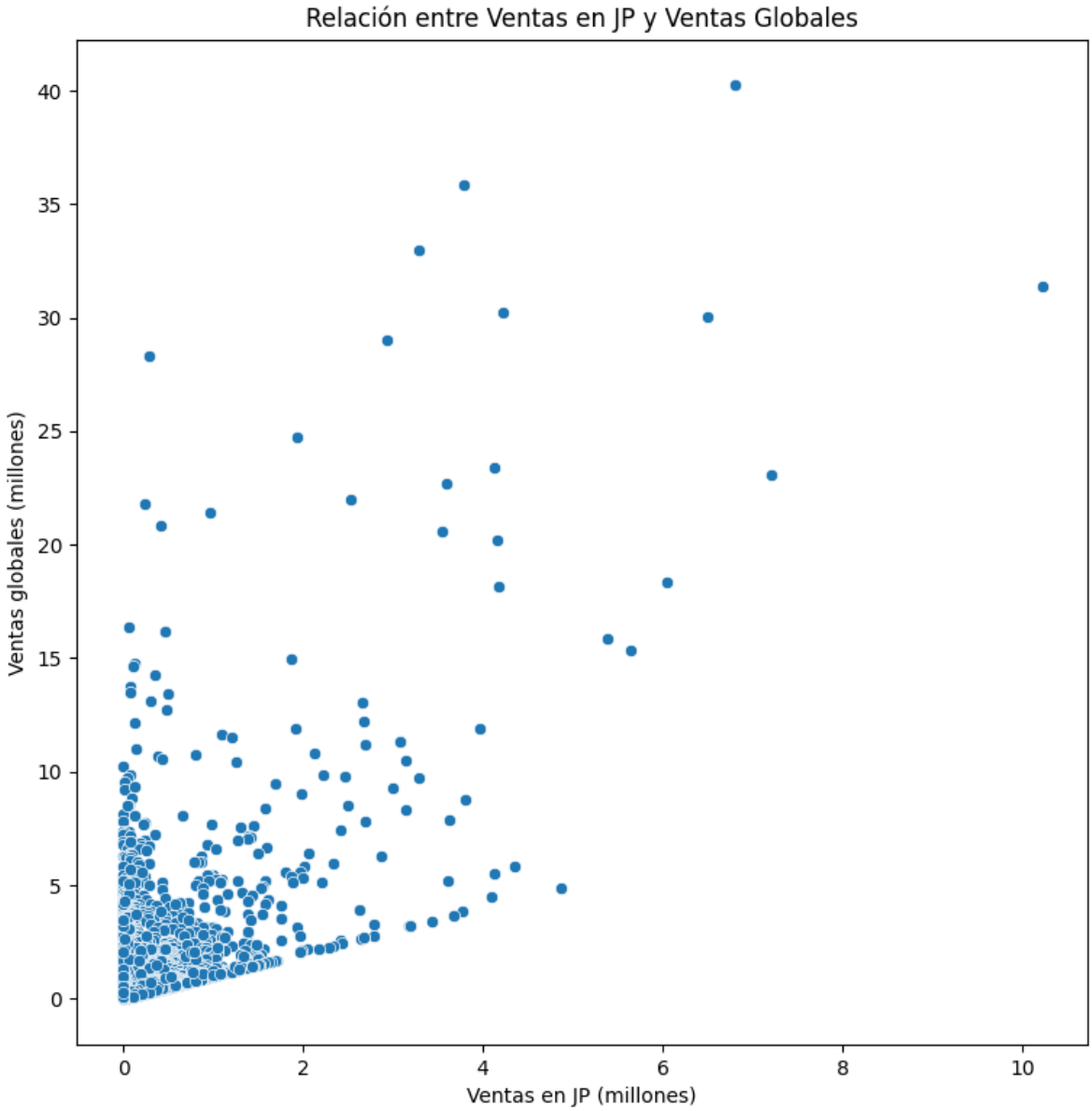

En este scatterplot podemos observar de manera más adecuada la correlación existente entre estas 2 variables, por lo tanto podemos asumir que un lanzamiento con ventas exitosas en NA está directamente relacionado con una mayor probabilidad de éxito global.

**Ventas_EU y Ventas_Globales**



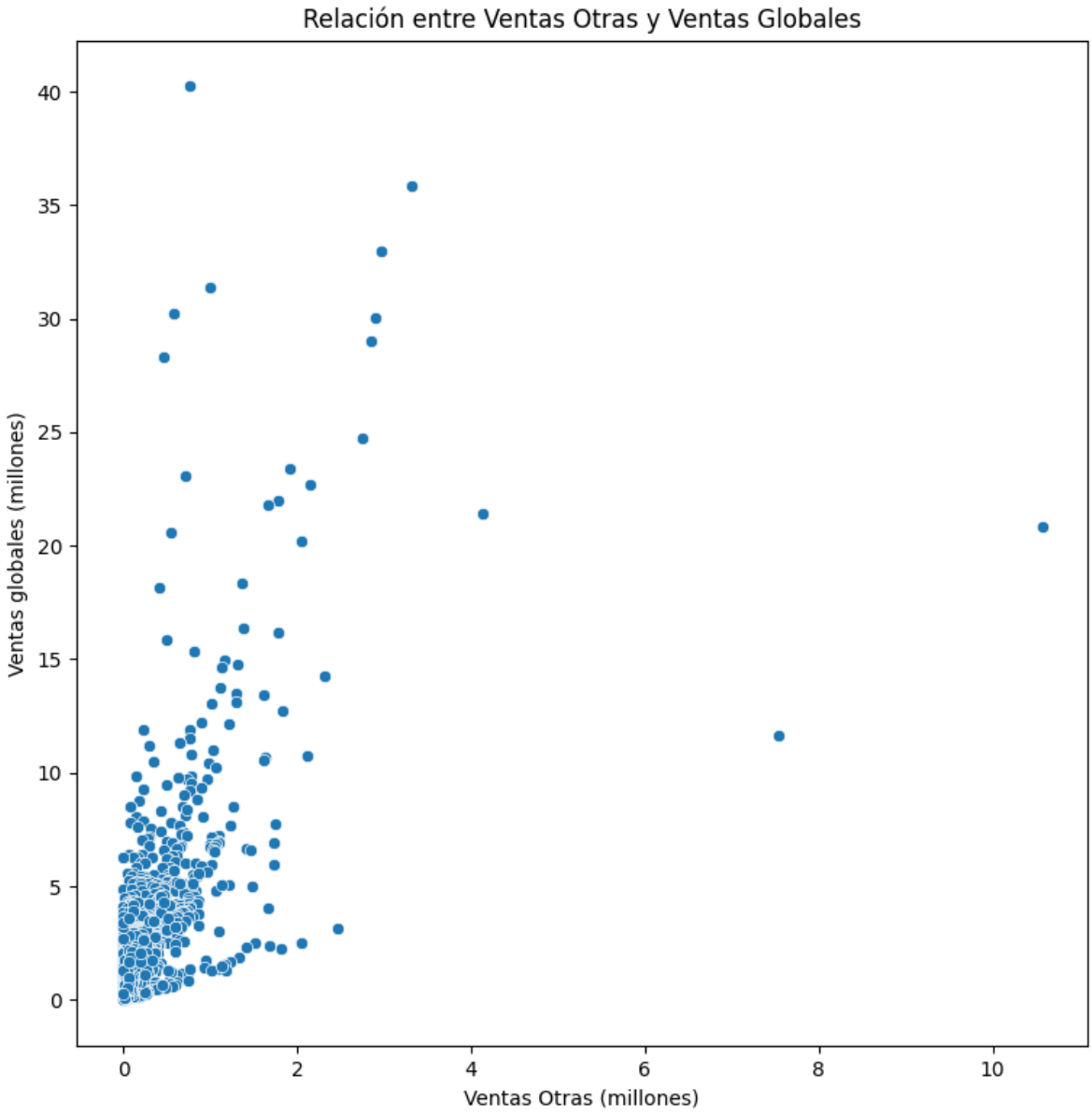

Nuevamente como en el scatterplot anterior, podemos observar una correlación positiva entre las ventas en EU y las ventas globales. El grafico muestra una diagonal ascendente, lo cual indica que una mayor cantidad de ventas en EU influye de gran manera en la probabilidad de éxito global que puede o no tener un título. Sin embargo, también se puede notar una dispersión un poco más pronunciada, lo que puede traducirse en que el éxito global de un título depende de un poco más que del mercado europeo.

**Ventas JP y Ventas Globales**




En este scatter plot podemos observar claramente una relación muy débil que deja en claro una sola cosa: el éxito global está determinado principalmente por el mercado de Norteamérica y Europa. Si bien Japón puede tener algo de presencia en cuanto a la probabilidad de éxito global, esta es muy mínima a comparación de los 2 mercados anteriormente mencionados, posiblemente debido a diferencias culturales, preferencias locales y estilos de juego específicos del mercado japonés.

**Ventas Otras y Ventas_Globales**




Finalmente en la siguiente grafica podemos observar una relación positiva aunque no tan pronunciada como en las regiones anteriores. Lo cual nos ayuda a confirmar lo mencionado anteriormente: La probabilidad de éxito global esta mayormente determinada por el éxito obtenido en el mercado de Norteamérica y Europa, mientras que el mercado japonés y el resto de ventas fuera de estas regiones, si bien influyen mínimamente, NO determinan principalmente la probabilidad de éxito global de un videojuego.

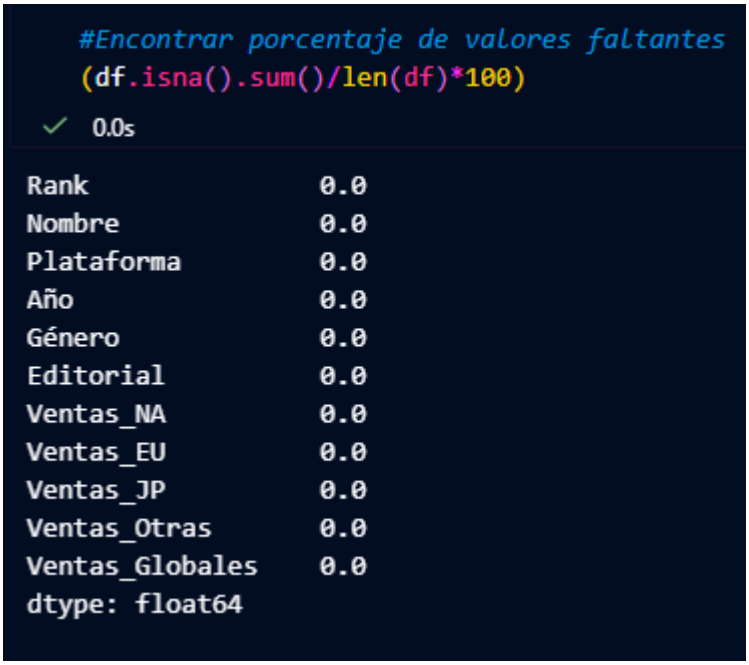
## Análisis de Valores Atípicos (Outliers)

A continuación se utilizó el rango intercuartílico (IQR) para detectar valores extremos en las columnas de ventas y tras un exhaustivo análisis, se tomó la decisión de mantener todos los outliers para cada columna de ventas por región.

Esto debido a que si bien representan valores muy extremos, no son errores, si no casos reales de un éxito superior. Casos de la vida real de videojuegos que alcanzaron un éxito muy grande. Realizar un análisis que no incluya tales casos de éxito podría resultar en una versión parcial e incompleta de la industria.

## Análisis de Valores Faltantes

En este caso, no se encontró ningún porcentaje de valores faltantes en ninguna de las variables.



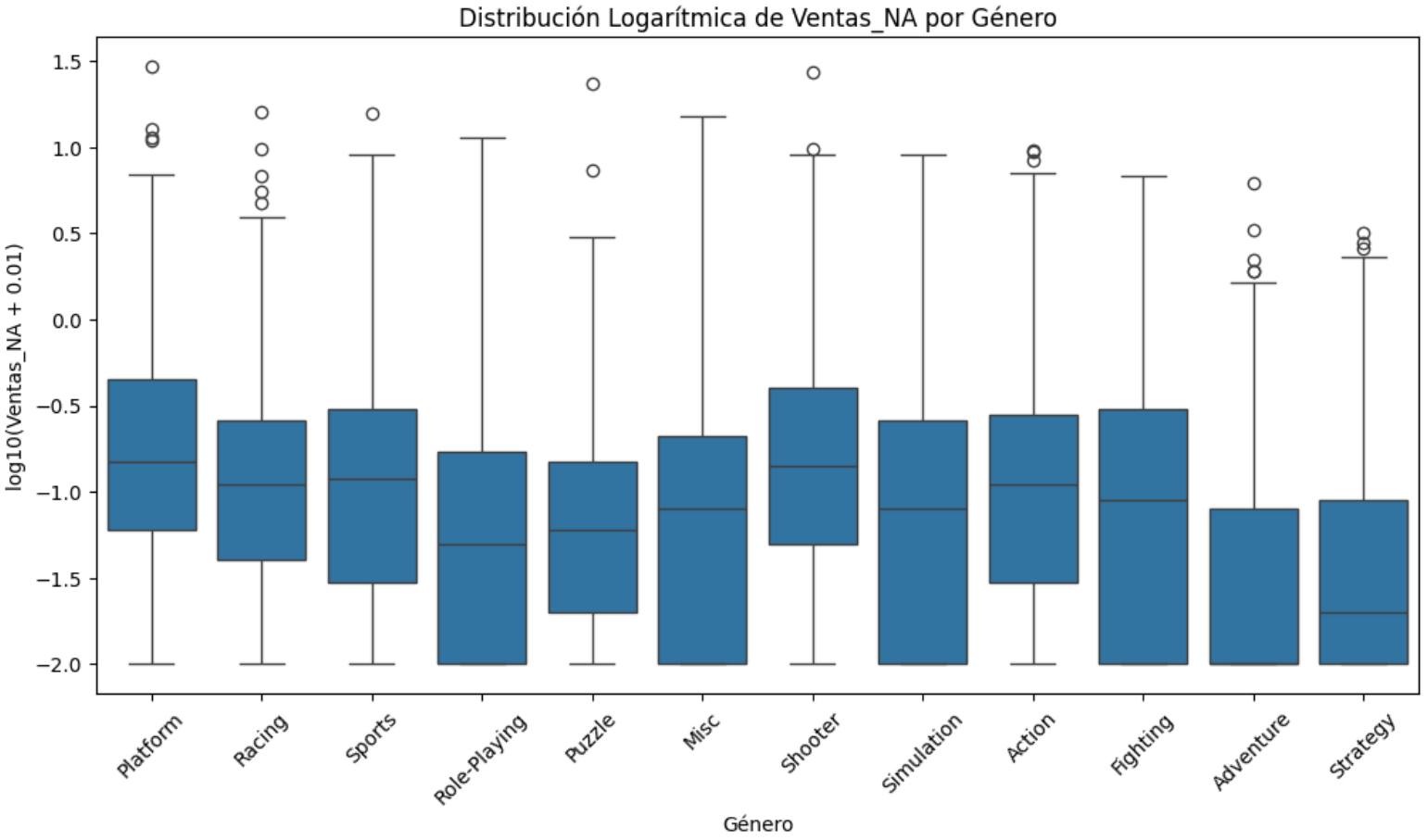
```
#Encontrar porcentaje de valores faltantes
(df.isna().sum()/len(df)*100)
```

✓ 0.0s

```
Rank              0.0
Nombre            0.0
Plataforma        0.0
Año               0.0
Género            0.0
Editorial         0.0
Ventas_NA         0.0
Ventas_EU         0.0
Ventas_JP         0.0
Ventas_Otras      0.0
Ventas_Globales   0.0
dtype: float64
```

## Relación entre variables numéricas y variables categóricas

Muy bien, ahora nos encontramos en una de las fases más interesantes del análisis, ya que aquí finalmente podremos comprobar la relación entre nuestras variables de ventas y nuestras variables categóricas (género y plataforma). A continuación se muestran las gráficas y un breve análisis de cada una de ellas.




(Para un mejor análisis de comparación entre ventas y género, se utilizó una escala logarítmica en el eje Y)

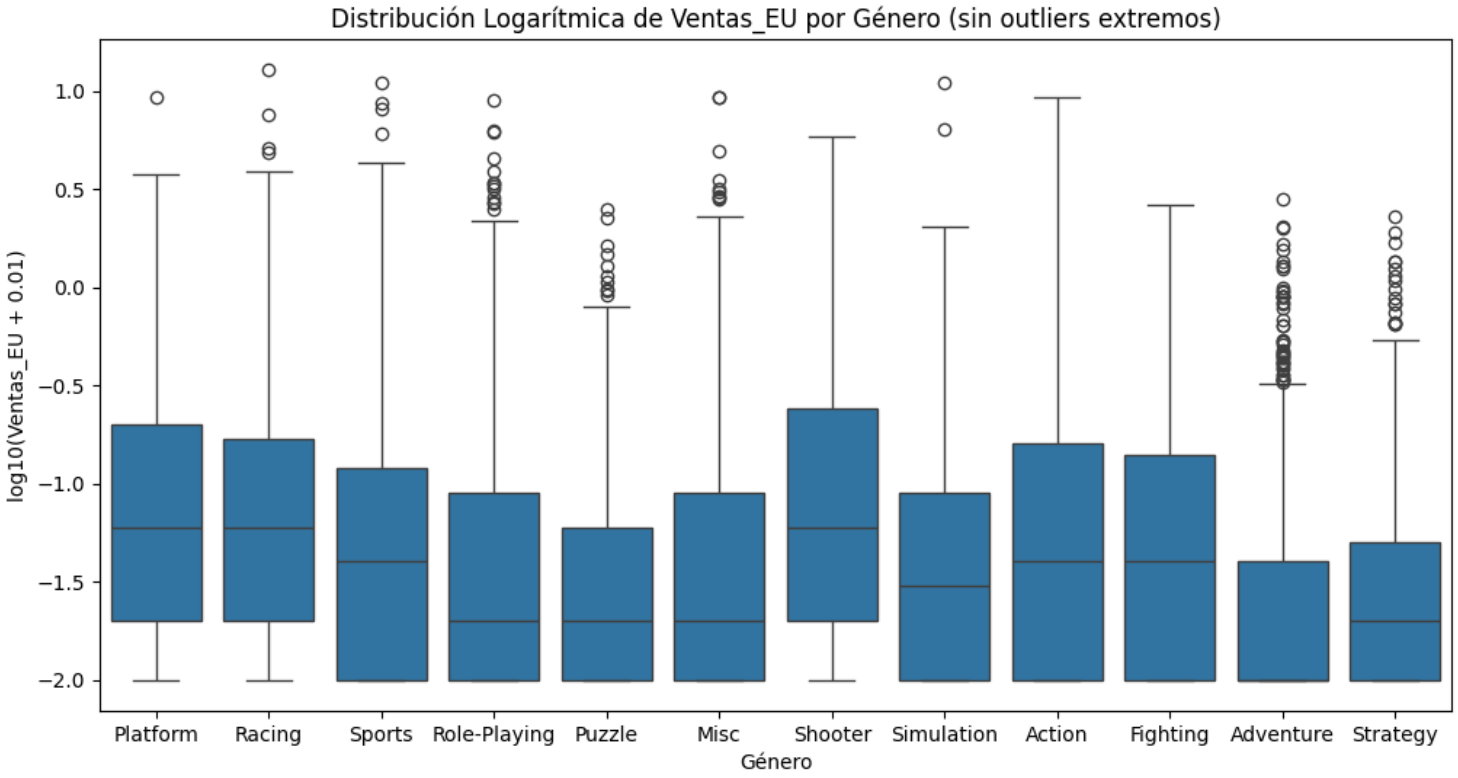
A continuación un análisis de cada uno de los gráficos por columna de ventas:

**Distribución de ventas en NA**

En el grafico podemos observar con más claridad la distribución de las ventas por género. Aquí comprobamos como el género que menos ventas genera en NA es el género de Strategy, esto debido a tener la mediana más baja de ventas de todas, y que si bien también tiene un largo bigote superior, lo cual podría indicarnos distribución, la realidad es que la mayoría de los títulos se

encuentra agrupados en ventas muy bajas. Mientras que por el contrario, el género Shooter se encuentra en la posición número uno en cuanto a ventas, esto debido a su mediana (la más alta) y a una caja posicionada en una posición más alta, lo cual indica que la mayoría de los títulos (si bien hay distribución de ventas) alcanza buenas ventas. Acompañando a la lista de géneros con ventas estables tenemos a los géneros Platform, Racing, Sports y Action, todos con una mediana de ventas, altura de caja similar y al menos un outlier (sin olvidar mencionar un tamaño de bigotes superiores e inferiores parecido, es decir, mucha variabilidad entre sus títulos). Mientras que los géneros Adventure y RPG son los menos rentables al tener la mediana de ventas más bajas después de Strategy y tener la gran mayoría de datos agrupados en una parte baja, lo cual resalta que la mayoría de las ventas son bajas. También es importante mencionar como el género Racing es el que cuenta con más outliers, lo que se traduce en más títulos sobresalientes que los demás generos.

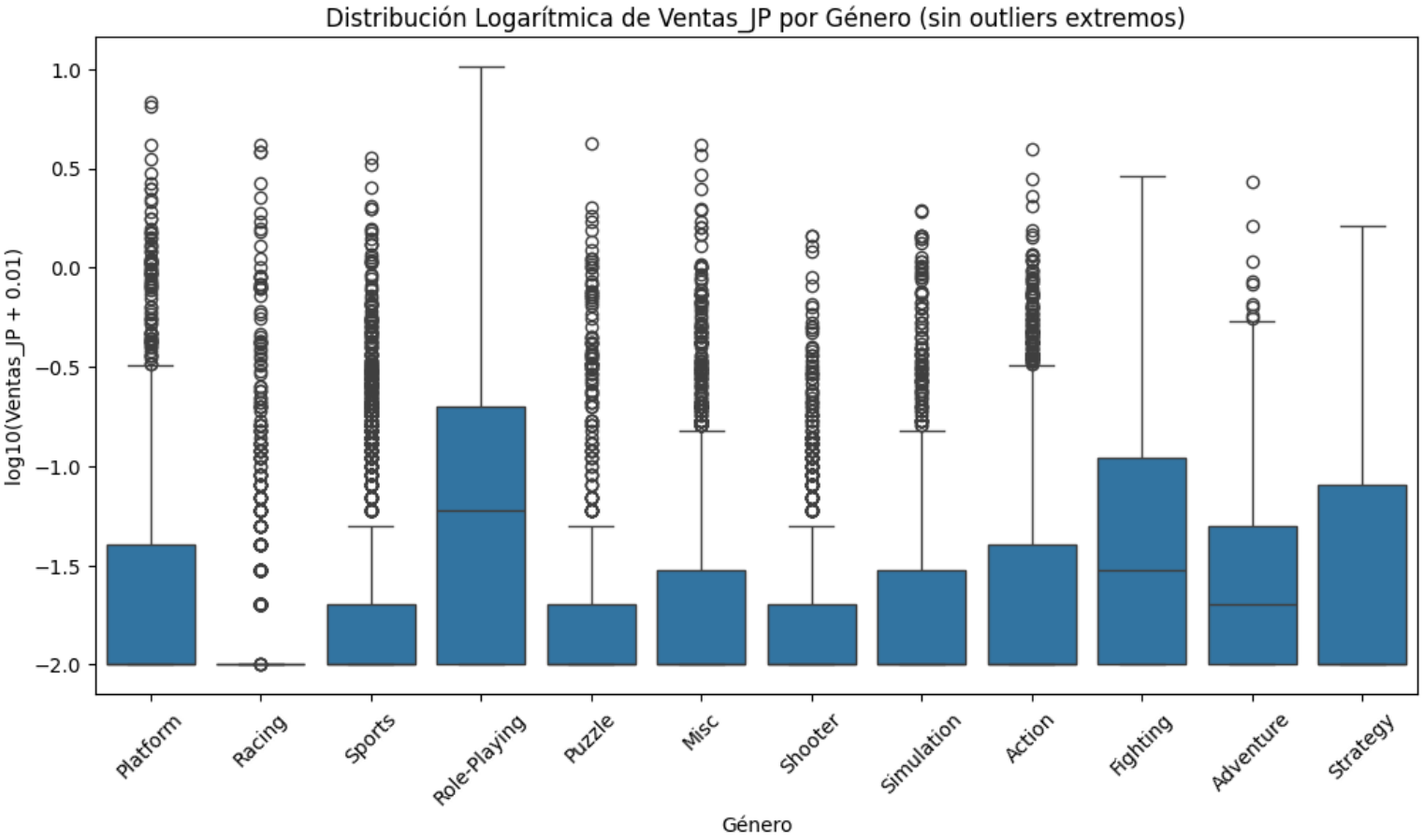
**Distribución de ventas en EU**




(Se omitieron los outliers negativos para una mejor visualización del gráfico)

Empezamos el análisis del grafico mencionando un dato importante: Al igual que en NA, Shooter se alza como el género más rentable, debido a la mediana más alta de entre todas y la posición de caja con mayor altura, con una alta dispersión hacia arriba y poca distribución baja, lo cual indica que es el género con más títulos con buenas, si no excelentes ventas. Aunque no podemos ignorar el hecho de que no cuenta con ningún outlier, lo cual nos dice que si bien es el mejor género en cuanto a ventas, no cuenta con ningún título que sea un éxito extraordinario. Mientras que como peor título esta vez contamos con el género adventure, al contar con la caja más pequeña de todas y por lo tanto, una mediana exageradamente baja. Aunque en contraste, este es el género con mayor cantidad de outliers, lo que significa que si bien casi todos los lanzamientos cuentan con ventas muy bajas, hay muchos que la rompen por completo. En cuanto a los géneros más estables contamos con los géneros Platform y Racing, géneros con un buen tamaño de caja, baja distribución negativa, alta distribución positiva y las segundas medianas más altas después de Shooter. Por otro lado los géneros menos rentables son Strategy y Puzzle, al contar con las cajas más pequeñas de todas, una distribución concentrada negativamente y la misma mediana de ventas, pero también es importante resaltar que aunque sean los 2 peores géneros después de Adventure cuentan con una buena cantidad de outliers, lo que nos dice que si bien en general es un mal género en cuanto ventas, también tiene títulos que alcanzan un gran éxito.

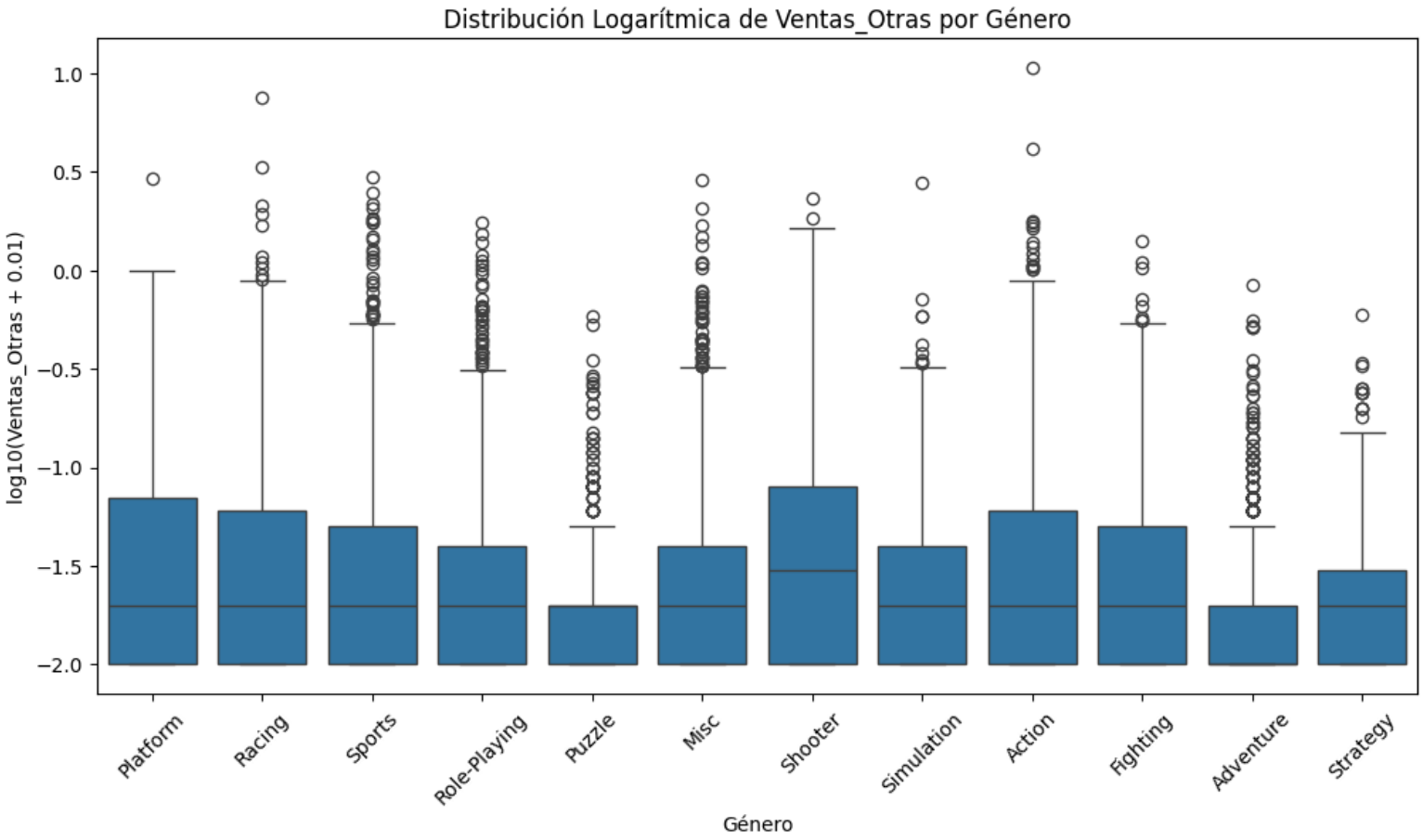
**Distribución de ventas en JP**




Llegamos al mercado japonés y desde ya podemos ver cosas sumamente interesantes. Mientras que en NA y EU, el género RPG se posicionaba como uno de los menos rentables, aquí se redime siendo el mejor género en cuanto a ventas, con una mediana inmensurablemente superior a todas las demás, el mayor tamaño de caja con una gran dispersión positiva, mientras que el segundo genero más rentable es Fighting, otro género que en los anteriores 2 análisis pasaba completamente desapercibido aquí se alza con una gran diferencia como el segundo mejor género, con la segunda mediana y tamaño de caja más alta de todos. Mientras que Racing, un género con buenas ventas en los anteriores análisis, aquí cae como el peor genero de respecto a ventas, aunque eso sí, con una gran cantidad de outliers, lo que puede indicar excepciones altas pero bajísimas ventas generales. Los demás géneros siguen este patrón: bajas ventas generales pero una buena cantidad de excepciones. Aunque eso no puede negar lo obvio: el mercado japonés es totalmente dominado por los generos RPG y Fighting, mientras que la mayoría de otros

géneros es relegado a posiciones con buenos casos excepcionales, pero muy bajas ventas generales.

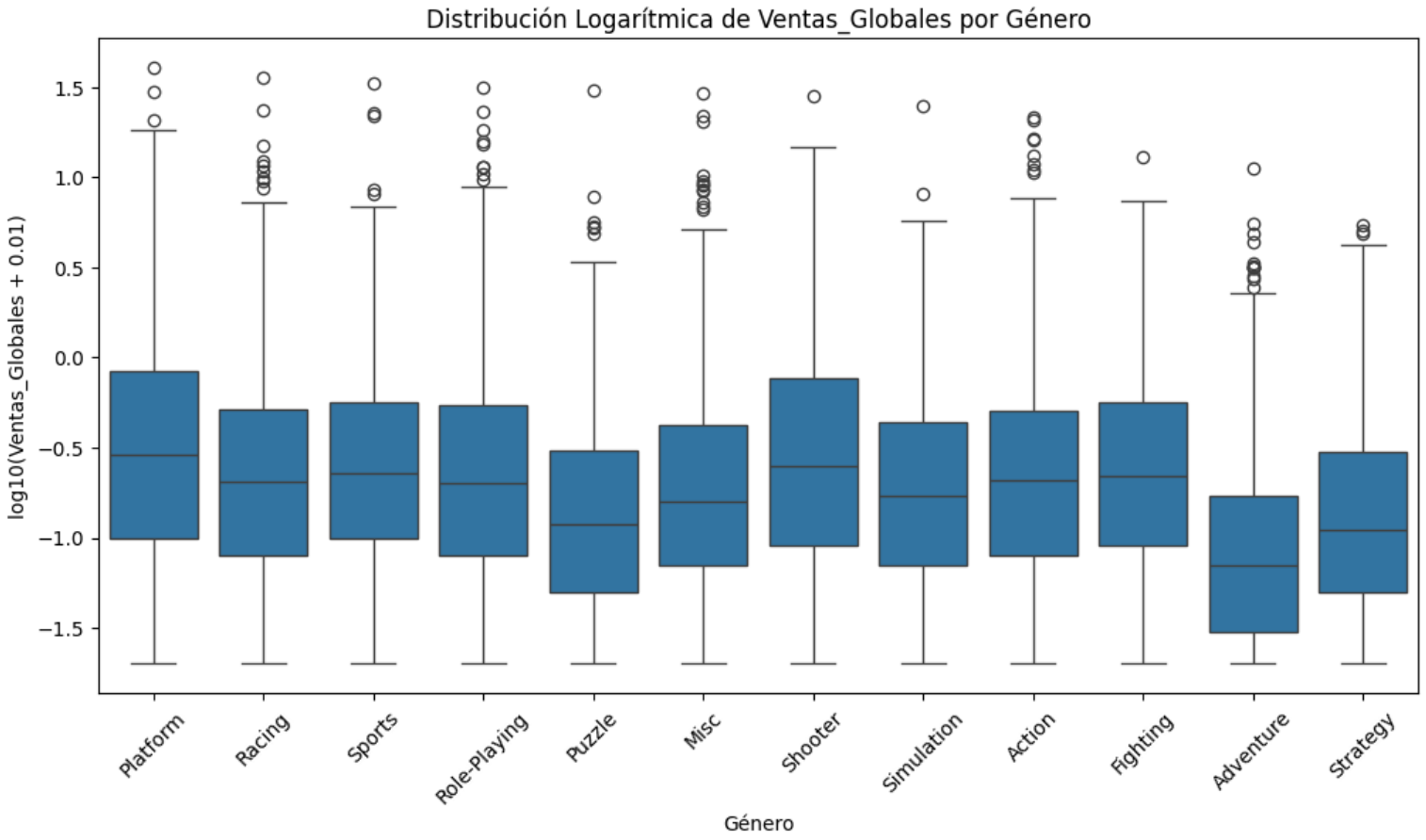
**Distribución de ventas en el resto del mundo (Otras ventas)**




En este grafico podemos observar a primera vista, cajas de tamaños similares con solo algunas más altas o bajas pero todas con una característica: ninguna tiene bigotes negativos, lo cual indica que la mayoría de ventas no son muchas. Nuevamente shooter se alza como el mejor género en cuanto a ventas, esto por tener la caja y la mediana más altas que el resto, pero con una enorme diferencia si lo comparamos con EU o NA. Como los géneros menos rentables esta vez tenemos un empate entre los géneros puzzle y adventure, ambos con el mismo tamaño de caja y la misma distribución aunque con el género adventure teniendo más outliers. A partir de aquí los géneros probablemente más rentables serian platform, racing y sports, los 3 con una diferencia en tamaño de caja mínima pero superior (aunque no en gran medida) a las demás,

eso sí, a pesar de tener un tamaño de caja mayor, la mediana sigue siendo la misma que la de todas las demás, por eso no podemos decir que existe realmente un género “estable” ya que aunque si, efectivamente podemos observar una buena cantidad de outliers en algunos géneros, eso no quita la tendencia general de ventas bajas.

**Distribución de ventas globales por género**




Finalmente tenemos el grafico de ventas globales, en donde nos llevamos una gran sorpresa, ya que aun después de ser el mejor género en cuanto a ventas por región, globalmente el género Platform se alza como el mejor género en cuanto a ventas globales, con una caja y una mediana muy pero muy ligeramente arriba de shooter, marcan la diferencia como genero líder en ventas globales. Nuevamente el género Adventure vuelve a quedar como el género menos rentable esta vez globalmente, con la caja y mediana más bajas de todas, lo cual indica ventas generales bajas. En cuanto al resto de géneros

las ventas se encuentran muy igualadas, con medianas y tamaños de caja bastantes similares, lo cual indica que casi todos los géneros tienen buenas ventas a nivel global, lo cual nos indica lo fuerte y rentable que es la industria de los videojuegos a nivel global.

## Observaciones y hallazgos importantes

En esta fase, la variable objetivo de nuestro modelo de ML será Ventas_Globales. Mientras que las variables más influyentes según nuestra matriz de correlación son las demás columnas de ventas (Ventas_NA, EU, JP y Ventas_Otras) y si bien no mantienen una relación muy fuerte, también se consideraran las variables género y año (en caso de no ser relevantes, se eliminaran del análisis)

A continuación un  resumen de hallazgos claves:

- La probabilidad de éxito global de un lanzamiento está determinada principalmente por su desempeño en los mercados de NA y EU
- Platform se consagra como el género líder en cuanto a ventas globales
- Shooter es el género con mejores ventas en cuanto a ventas por región
- Adventure es el peor género en cuanto a ventas globales y por región
- Los géneros con ventas más estables por región son Sports y Racing
- La frecuencia de un género (lanzamientos) no indica mayor éxito en sus lanzamientos
- El género con más outliers tanto por región como a nivel global es Action
- El mercado japonés esta distanciado por completo de las demás regiones
- El género con menos outliers en NA y EU es Puzzle
- El género con menos outliers en JP es Fighting
- El género con menos outliers fuera de las regiones principales y a nivel global es Strategy

Los outliers negativos fueron conservados en el análisis, ya que representan casos reales de bajas ventas y aportan información relevante sobre los riesgos comerciales del sector.

## Modelo de ML (Machine Learning)

En este análisis nuestra variable objetivo (ventas globales) es numérica, por lo tanto se requieren modelos de regresión y no de clasificación.

En base a lo anterior mencionado, los modelos elegidos son regresión lineal múltiple y random forest ya que, como se mencionó antes, la variable es numérica y además el tamaño del dataset lo permite. Regresión lineal aporta interpretabilidad y facilita explicar qué factores influyen en las ventas, mientras que random forest mejora la precisión prediciendo resultados con variables diversas, interacciones y outliers. Comparar ambos nos permitirá mejorar tanto la explicación de resultados como el desempeño predictivo.

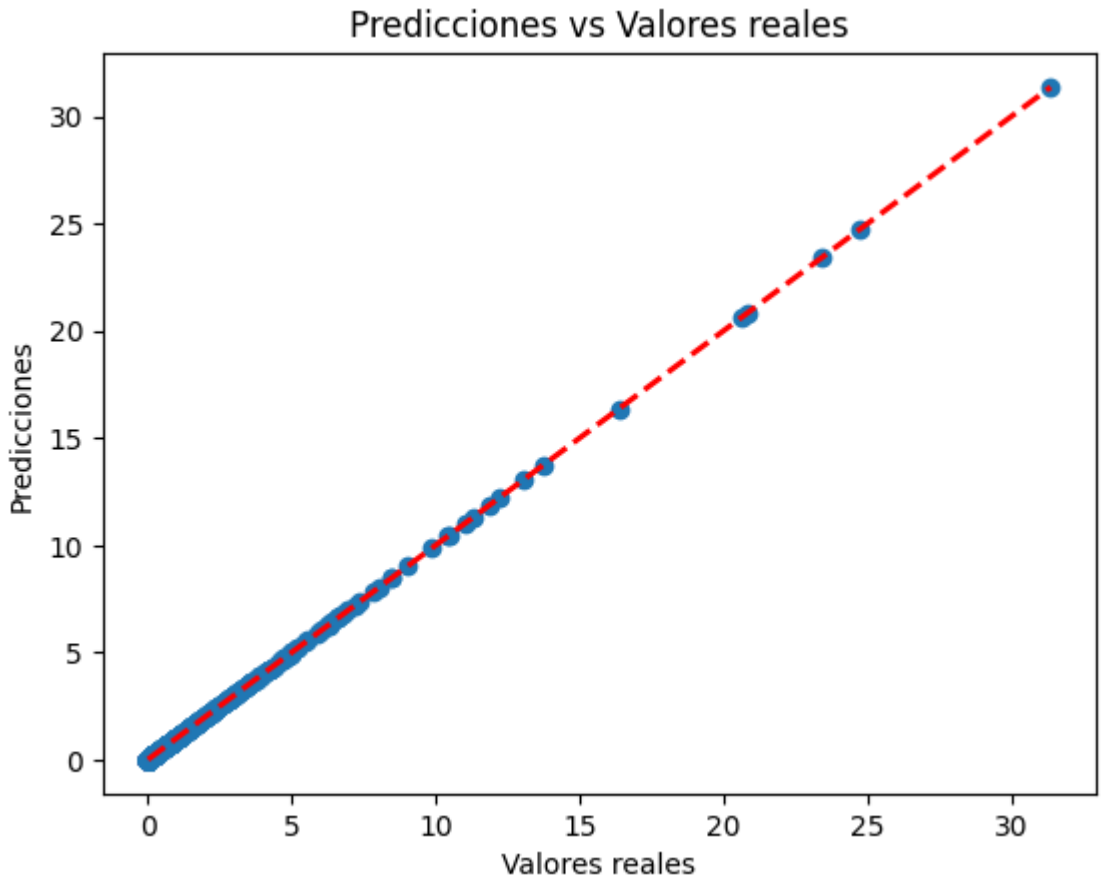
*La implementación y entrenamiento de los modelos se maneja a mayor detalle en el código fuente del análisis*

Como podemos ver en el código fuente del análisis y, el modelo de regresión lineal obtuvo un desempeño casi perfecto ($R^2$: $0.99$ $y$ $RMSE$; $0.0048$), esto debido a que la variable objetivo estaba directamente relacionada con las variables predictoras. De igual forma, nuestro modelo de Random Forest obtuvo muy buenos resultados, lo cual confirma la calidad de los datos en el dataset.

## Visualización de resultados

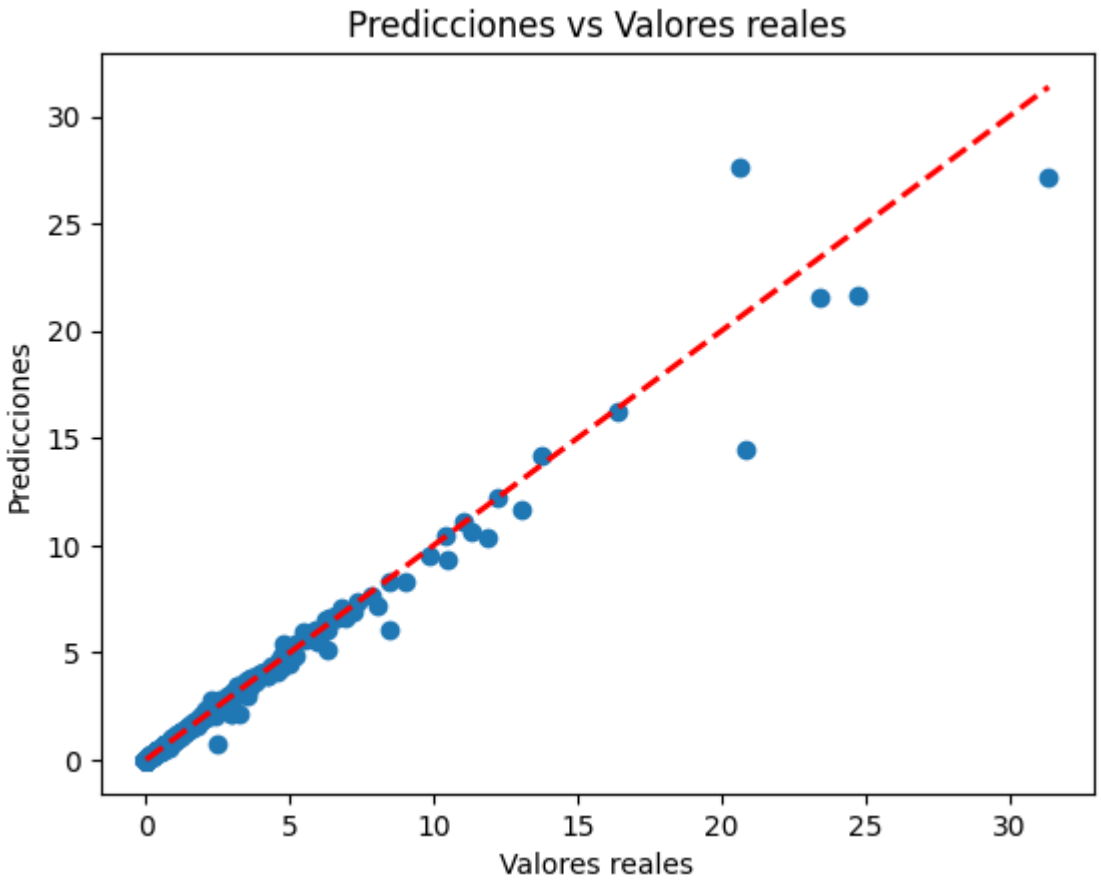
Una vez realizada la implementación, entrenamiento y evaluación de nuestros modelos, procedemos a demostrar los buenos resultados mediante gráficos de dispersión. En estos podemos observar que, mientras más pegados los puntos a nuestra línea central (predicción perfecta), más precisos son los datos obtenidos.

**Gráfico de dispersión de regresión lineal**




Tal y como lo mostraron las métricas, los resultados fueron prácticamente perfectos, alcanzando un $R^2$ de 0.99 (casi 1.0), lo que nos dice que el modelo predice el 99% de los datos, prácticamente perfecto y un RMSE bajísimo, lo cual nos da un margen de error prácticamente nulo. Como ya mencionamos antes, esta perfección podría deberse a la directa relación entre nuestras variables predictoras (ventas por región) y nuestra variable objetivo.

**Gráfico de dispersión de random forest**




Nuevamente corroboramos las buenas métricas obtenidas y mediante este grafico podemos observarlo con mayor claridad. Con un $R^2$ de 0.97, nuestro modelo puede predecir el 97% de los datos mientras que obtuvo un RMSE de 0.22, lo cual también es un gran valor, puesto que RF intenta encontrar patrones un poco más complejos que regresión lineal, por lo tanto le cuesta un poco más adivinar la suma perfecta que el modelo de regresión lineal encuentra de inmediato.

## Conclusión del modelo

Podemos concluir este reporte con que nuestros modelos seleccionados pueden predecir con buena precisión y bajo margen de error. Evidentemente regresión lineal obtuvo un mejor resultado, pero esto es debido a la directa relación entre nuestras variables predictoras y nuestra variable objetivo. Aunque de igual forma no quita que Random Forest obtuvo de igual forma resultados excelentes. Las variables más influyentes desde luego, fueron las de ventas región: “Ventas_EU”, “Ventas_NA”, “Ventas_JP” y “Ventas_Otras”, ya que matemáticamente componen el total de las ventas globales, superando en importancia a otras variables como el género o el año.

# Conclusiones generales

Este análisis nos permitió analizar las ventas del mercado por género, lo cual nos llevó a poder entender un poco más los criterios de éxito global y de éxito por región y como funciona este mercado. Descubrimos datos clave como los géneros líderes en venta en cada región, el género con mejores ventas globales, los peores y mejores géneros en cuanto a ventas tanto regionales como globales, relaciones entre la probabilidad de éxito global y éxito en ciertas regiones, etc. Algunas de las hipótesis que logramos solucionar y que se plantearon en fases previas a este proyecto fueron las siguientes:

**“Los videojuegos de genero Acción y Deportes contienen la mayor parte de las ventas globales.”**

Esta hipótesis fue **falsa**, ya que gracias a nuestro análisis logramos comprobar que si bien efectivamente son los 2 géneros con más lanzamientos, mayor frecuencia de lanzamientos no se traduce en mayores ventas. Eso lo comprobamos ya que los 2 géneros con la mayor parte de las ventas globales son Platform y Shooter respectivamente.

**“La región de Norteamérica representa el mercado más grande en ventas de videojuegos”**

Esta hipótesis fue **cierta**. Junto a EU, NA representa el mercado más importante en cuanto a influencia de probabilidad de éxito global.

Mientras que algunas de las preguntas que pudimos contestar son las siguientes:

***“¿Qué géneros de videojuegos han generado más ventas globales a lo largo del tiempo?”***

Los 3 géneros con mejores ventas a nivel global con el paso del tiempo son:

1. Platform
2. Shooter
3. RPG

***“¿En qué región (Norteamérica, Europa, Japón, Otros) se concentra la mayor cantidad de ventas?”***

Estas se concentran principalmente en la región de Norteamérica.

***“¿Los géneros más populares son los mismos en las diferentes regiones?”***

No. Si bien algunos géneros son bien recibidos de igual forma en todas las regiones, cada región a excepción de una, cuenta con diferentes géneros líderes en ventas:

- Norteamérica: Platform
- Japón: Role-Playing-Game
- Europa y resto del mundo: Shooter

***"¿Existe alguna relación entre las fechas en las que se lanzaron parte de los videojuegos más exitosos?"***

No. La fecha de lanzamiento resulta completamente irrelevante a la hora de influir en la probabilidad de éxito de un juego, ya que mientras el videojuego más exitoso de todos los tiempos se lanzó en 1985, el segundo más exitoso fue lanzado en 2008. Esto está más relacionado a temas de innovación para la época y jugabilidad que con la fecha de lanzamiento.

***"¿Cuáles son los géneros que han decaído más a lo largo de los años?"***

Globalmente, los peores géneros en cuanto a ventas a lo largo del tiempo son:

1. Adventure
2. Strategy
3. Puzzle

Algunas de las posibles mejoras que podría llegar a tener este análisis seria incluir de manera más recurrente a la variable "plataforma", esto para corroborar alguna relación entre la plataforma de lanzamiento y la probabilidad de éxito

Otra variable que podría incluirse seria el Publisher, para detectar si este tiene alguna relación de igual manera con la probabilidad de éxito de un lanzamiento.

## Referencias

Kaggle – Video Game Sales Dataset :
https://www.kaggle.com/datasets/gregorut/videogamesales