## **PROYECTO “Video Game Sales and Ratings”**

### **1. Introducción**

Los videojuegos son una de las formas más populares de entretenimiento a nivel mundial. El éxito de un juego depende de diversos aspectos como la plataforma en la que se juega, su género, las opiniones de los críticos y usuarios, así como las ventas en distintas regiones. A veces no es fácil ver las conexiones entre los datos, pero al analizarlos es posible encontrar patrones que puedan explicar por qué algunos videojuegos tienen más éxito que otros en el mercado. Este proyecto analiza datos mundiales de videojuegos para estudiar cómo la plataforma, el género y las calificaciones influyen en las ventas a nivel global y regional.

**2. Objetivo general**

Procesar un conjunto de datos de videojuegos para explorar la relación entre las ventas globales y regionales, las plataformas, los géneros, las puntuaciones de críticos y usuarios, con el fin de identificar patrones que puedan influir en el éxito comercial de los videojuegos.

### **3. Objetivos específicos**

* **Limpiar y organizar** la base de datos obtenida para garantizar que los datos estén listos para su análisis.
* **Analizar la relación** entre las ventas globales y las ventas por región (Norteamérica, Europa, Japón y otras) para determinar si alguna región domina el mercado.
* **Comparar el éxito** de diferentes plataformas (PS4, Xbox, PC, etc.) en las ventas globales y regionales.
* **Evaluar la influencia** de las puntuaciones de críticos y usuarios en las ventas totales de los videojuegos.
* **Explorar las tendencias** de ventas según el género del videojuego identificando los géneros más exitosos en cada región.

### **4. Planteamiento del problema**

La industria de los videojuegos enfrenta un desafío constante: identificar los factores clave que determinan el éxito comercial de un título. Si bien elementos como la plataforma y el género de un juego son considerados determinantes, aspectos adicionales como las calificaciones de críticos, las preferencias de diferentes regiones y el momento del lanzamiento también pueden desempeñar un papel crucial. En un mercado en constante crecimiento, se espera que las ventas de un videojuego puedan ser predichas en función de las tendencias y las preferencias de los consumidores. Sin embargo, existe una desconexión frecuente entre las calificaciones y las ventas: un juego altamente valorado no siempre se traduce en grandes ventas, y los títulos más vendidos no necesariamente obtienen las mejores críticas. Este proyecto busca analizar un conjunto de datos de videojuegos con el fin de descubrir qué características tienen mayor influencia en las ventas globales y regionales, y cómo estas variables están interrelacionadas.

### **5. Preguntas de investigación**

* ¿Cuántos videojuegos tienen información incompleta o nula en términos de ventas o puntuaciones?
* ¿Cuál es el rango de ventas en las diferentes regiones (Norteamérica, Europa)?
* ¿Qué plataformas generan mayores ventas globales?
* ¿Qué géneros de videojuegos son los que mayor se venden?
* ¿Cuál es la distribución de los videojuegos por género y plataforma?
* ¿Qué relación existe entre las puntuaciones de críticos y usuarios y las ventas globales?
* ¿Podemos usar la puntuación de críticos y usuarios para predecir las ventas globales?
* ¿Existen diferencias significativas en las ventas según el año de lanzamiento de los videojuegos?
* ¿Hay juegos con buenas calificaciones que venden poco? ¿Y juegos mal calificados que venden mucho?

### **6. Posible solución**

La solución consistirá en implementar un proceso de limpieza y procesamiento de datos utilizando Python y las bibliotecas de **pandas** y **numpy**. Se abordarán tareas como la eliminación de valores nulos, la normalización de las puntuaciones, y el cálculo de las ventas totales por región. Luego, se realizarán análisis estadísticos para identificar patrones en las ventas según las características de los videojuegos. Visualizaciones como gráficos de barras y de líneas serán empleadas para facilitar la interpretación de los resultados y responder las preguntas planteadas.

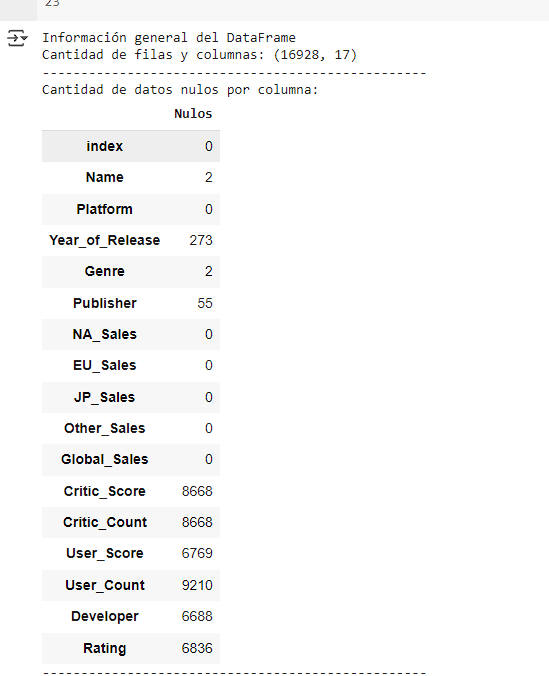
### **7. Pasos y resultados**

1. **¿Cuántos videojuegos tienen información incompleta o nula en términos de ventas o puntuaciones?**

La función isnull() busca los valores nulos, la aplicamos en un principio para la limpieza de datos, nos muestra la cantidad de datos nulos en cada columna.



**Resultados**

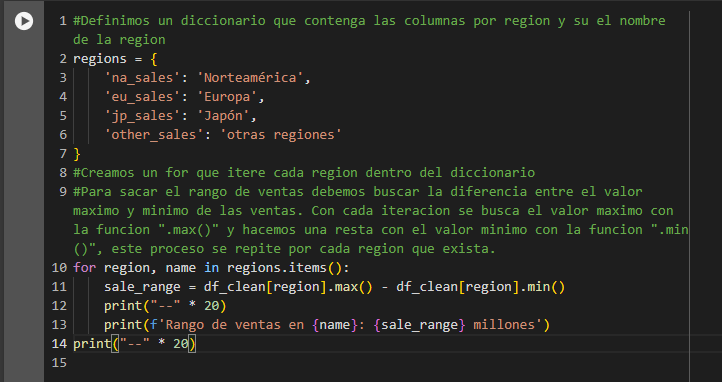
****

El resultado de la ejecución fueron diversos, en ventas no tenemos valores nulos por tanto no se modificará, en el caso de puntuaciones obtenemos bastantes valores NaNs nuestra solución fue llenar estos datos nulos con fillna() reemplazando los valores de las puntuaciones con valores generados aleatoriamente en base al promedio de los datos de la columna correspondiente.

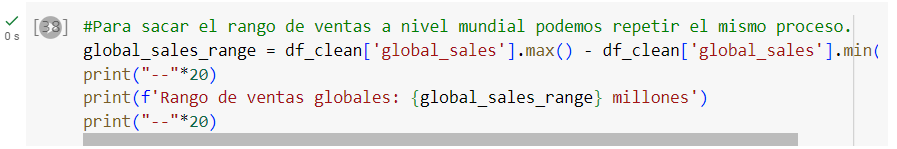
1. **¿Cuál es el rango de ventas en las diferentes regiones (Norteamérica, Europa, Japón)?**

El código calcula el rango de ventas de videojuegos en distintas regiones (Norteamérica, Europa, Japón y otras) mediante la diferencia entre las ventas máximas y mínimas registradas en cada área. Se utilizó un diccionario que asocia las columnas de ventas con nombres reales, se itera sobre cada región, se calcula el rango de ventas y se manda a imprimir..

Por Región:



Global:

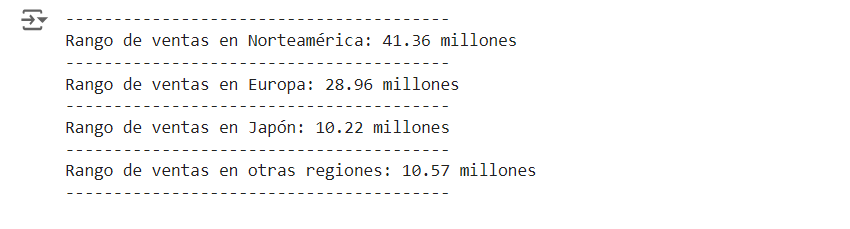


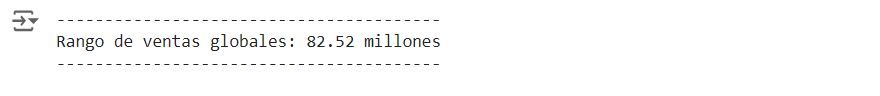
Se utilizó la función `.max()` para encontrar el valor máximo de ventas y la función `.min()` para el valor mínimo.

La diferencia entre estos dos valores se almacenó en variables específicas para cada región.

**Resultados**

Por Región:

****Global:

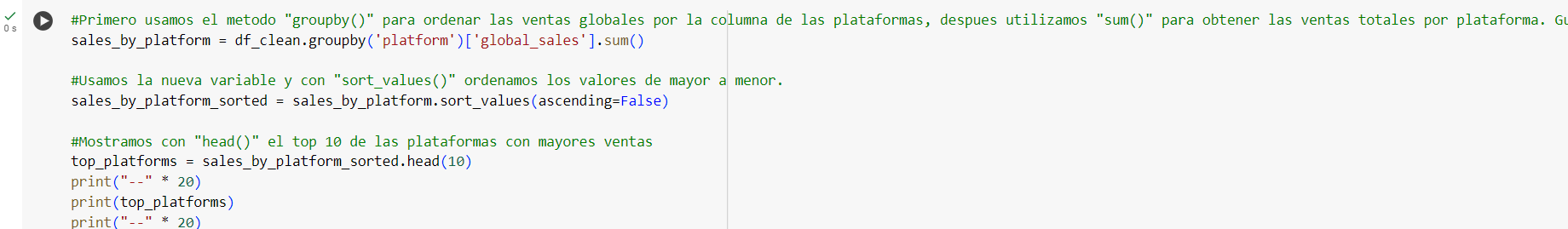


Estos resultados indican que hay variaciones significativas en las ventas de videojuegos entre las diferentes regiones. La cobertura de ventas en Norteamérica y Europa es particularmente alta en comparación con otras regiones, lo que indica que estas son regiones clave en el mercado de los videojuegos.

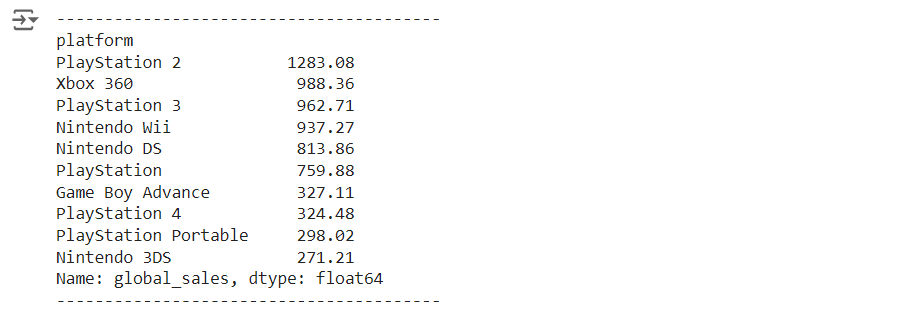
Las ventas globales de 82,52 millones demuestran el potencial y la diversidad del mercado global de videojuegos, aunque también indica que las ventas pueden variar significativamente entre los diferentes títulos de juegos.

1. **¿Qué plataformas generan mayores ventas globales?**

Para agrupar los datos de ventas por plataforma se utilizó groupby(), además de sum() para obtener el total de ventas globales, posteriormente sort\_values(ascending=False) para ordenar las plataformas de mayor a menor en función a lo anterior.



**Resultados**

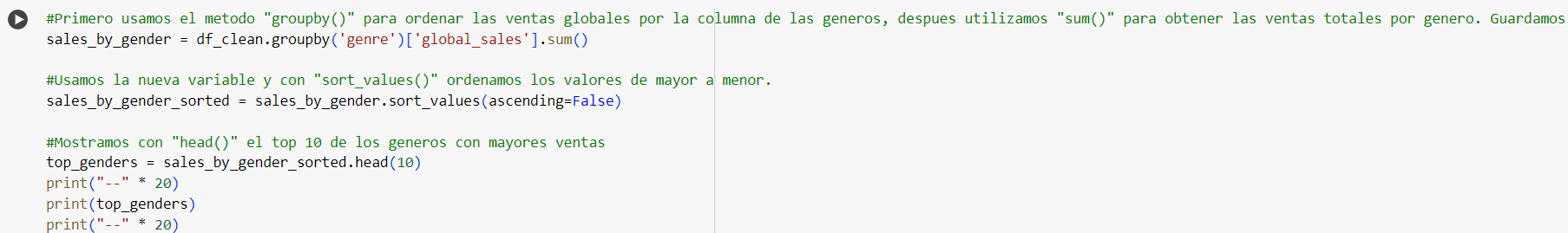
****

PlayStation 2 es la consola con mayores ingresos globales en la historia de los videojuegos, seguida de Xbox 360 y PlayStation 3.

Las consolas de Nintendo como Nintendo Wii y Nintendo DS también se encuentran entre las consolas más vendidas, lo que demuestra su fuerte presencia en el mercado.

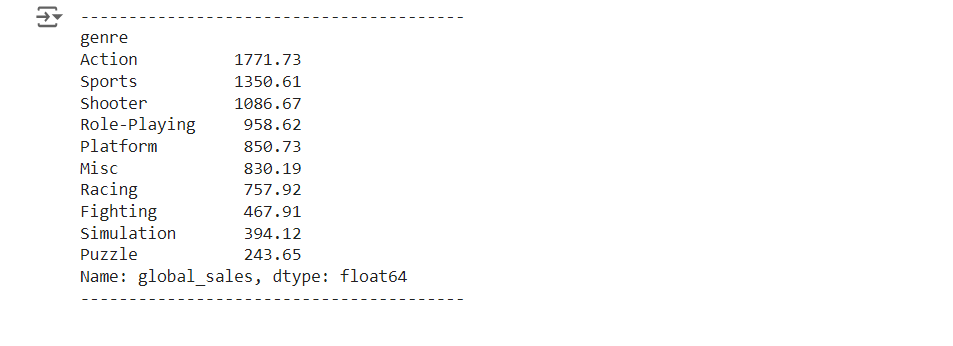
1. **¿Qué géneros de videojuegos son los que mayor se venden?**

Se realizó un análisis de las ventas globales totales de cada género de videojuego. Se agruparon los datos por género y luego se sumaron las ventas globales para cada uno de ellos.



**Resultados**

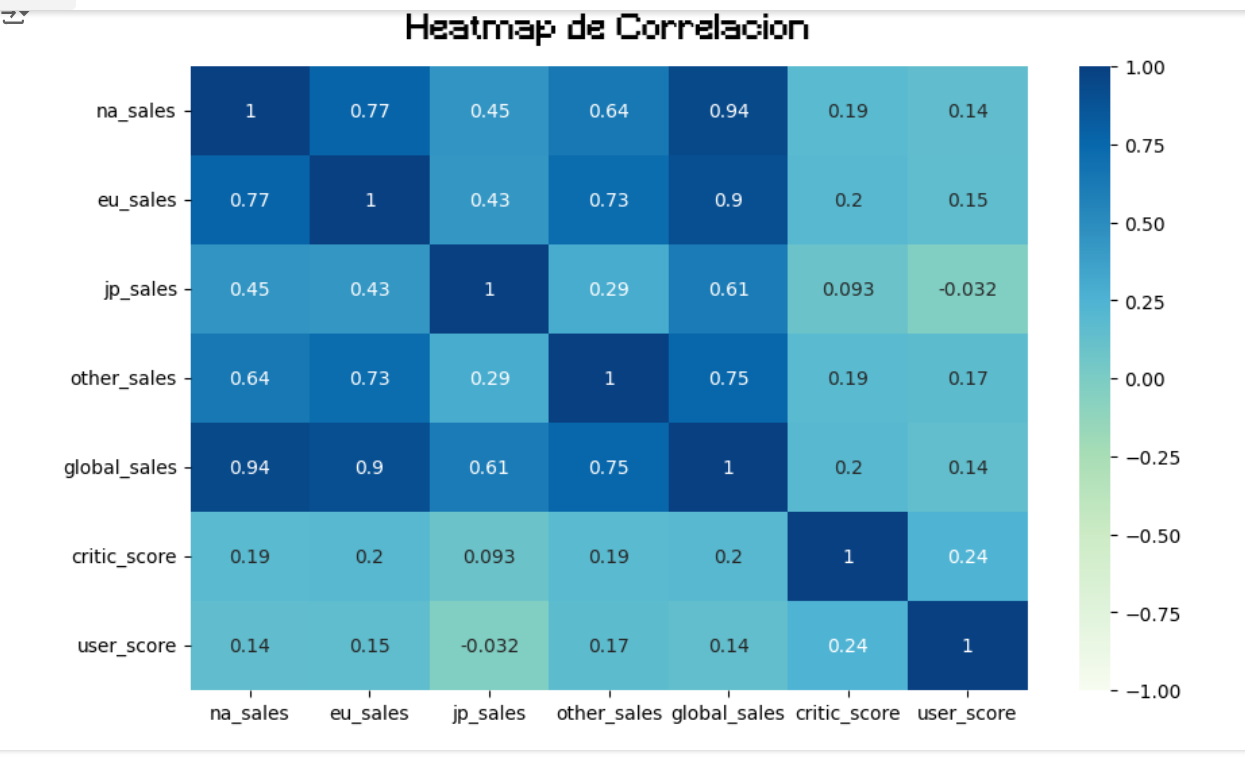
Los géneros más vendidos son Action (1,771.73 millones), Sports (1,350.61 millones) y Shooter (1,086.67 millones), destacando la popularidad de los juegos dinámicos y competitivos a nivel mundial, además ofrecen experiencias intensas y variadas destacan entre su popularidad.



**5.- ¿Qué relación existe entre las puntuaciones de críticos y usuarios y las ventas globales?**

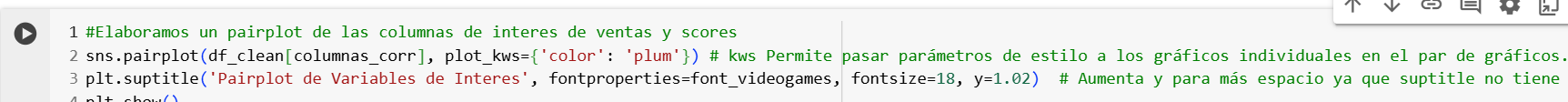
Queremo saber si a partir de las puntuaciones que tienen los videojuegos podemos encontrar una relación directa o positiva donde encontremos que si hay mejores reseñas, haya más compras a nivel global.

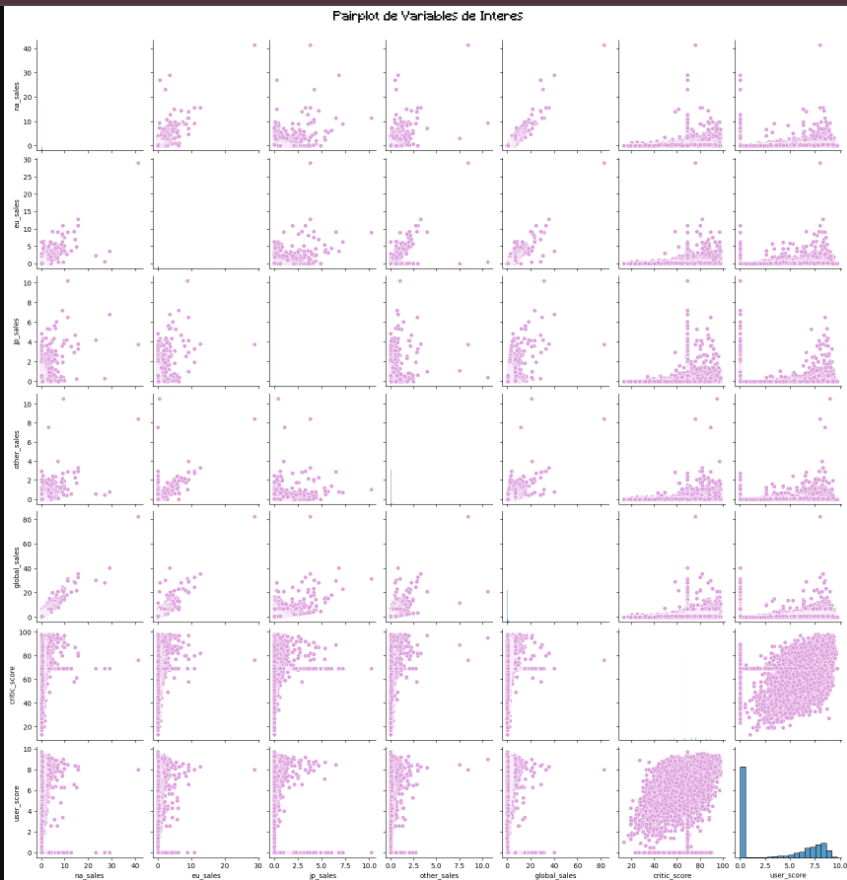




La correlación entre las puntuaciones de la crítica y de los usuarios con las ventas globales es positiva, pero no tan fuerte como la relación entre las diferentes regiones y las ventas globales. Esto sugiere que las reseñas, tanto de la crítica como de los usuarios, influyen en las ventas, pero no son el único factor determinante.

Elaboramos ahora un pairplot de las variables numéricas de interés en nuestro dataframe para poder visualizar posibles correlaciones entre variables:



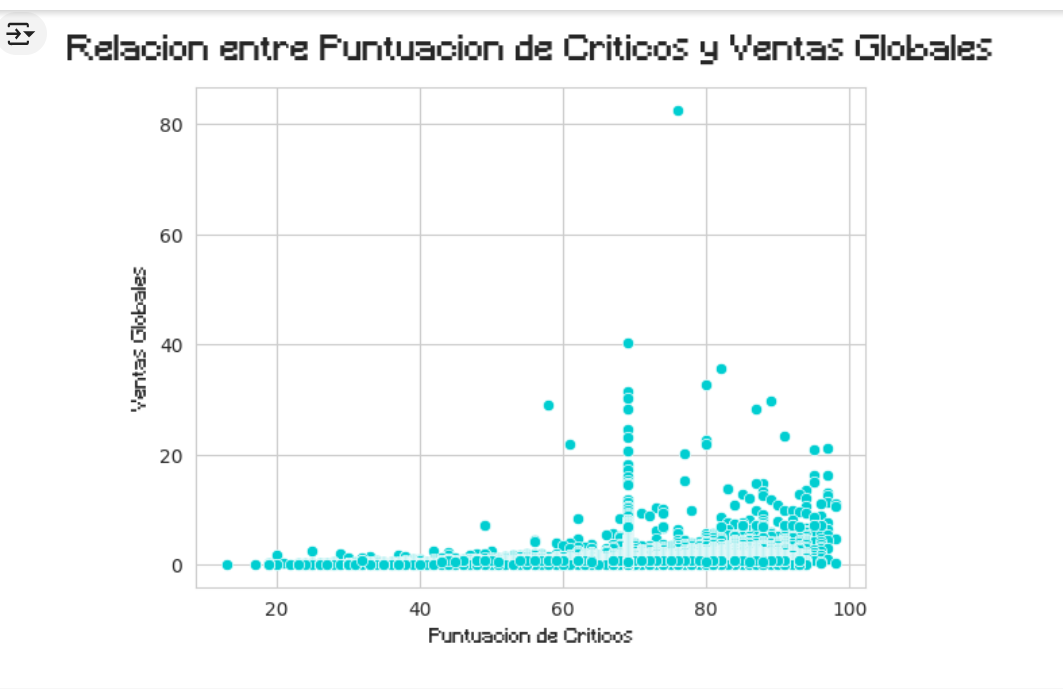


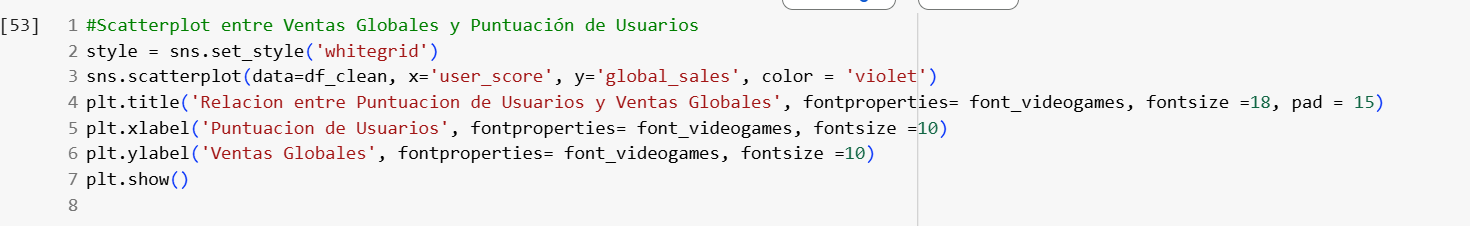
Nuestras observaciones respecto al gráfico son las siguientes:

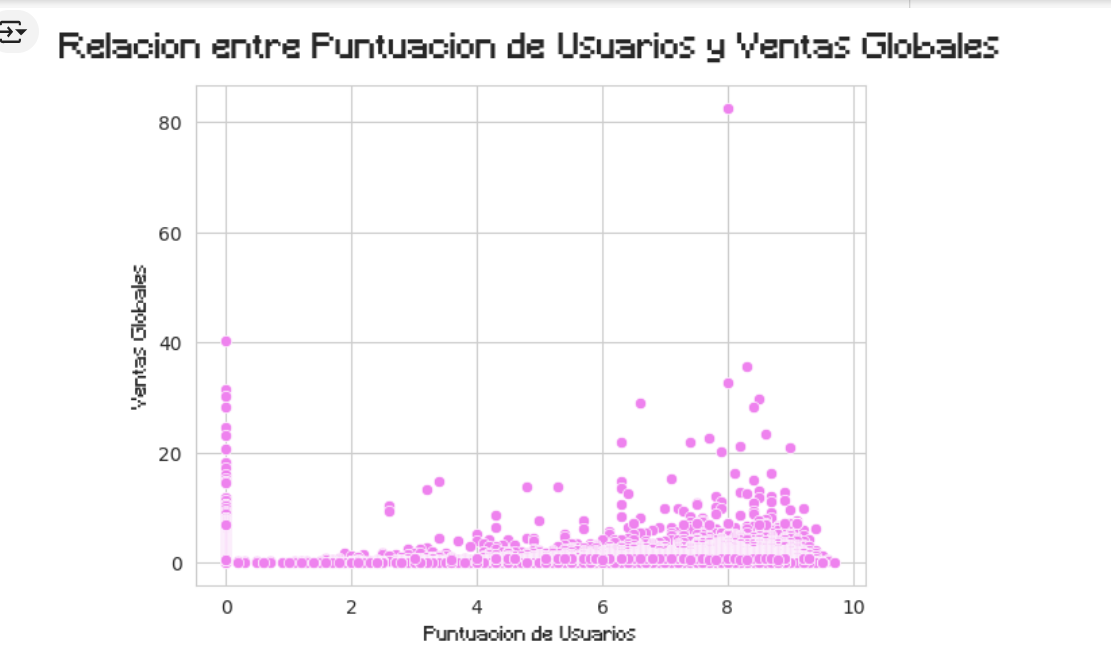
* Ventas por región: Vemos una posible correlación positiva entre las ventas en diferentes regiones (NA, EU, JP, etc.) y las ventas globales. Esto nos indica que si un juego vende bien en una región, es probable que también venda bien en otras.
* Puntuaciones y ventas: Podrías observar una correlación positiva entre las puntuaciones de críticos y usuarios y las ventas globales, aunque esta relación podría ser un poco más débil que la correlación entre las diferentes regiones.
* Observamos también una correlación positiva entre critc score y user score lo que tendría sentido pues si un juego es bueno tendrá buenas reseñas tanto por críticos como por usuarios

Nos enfocaremos en las variables que nos interesan y graficamos sus respectivos scatterplots:







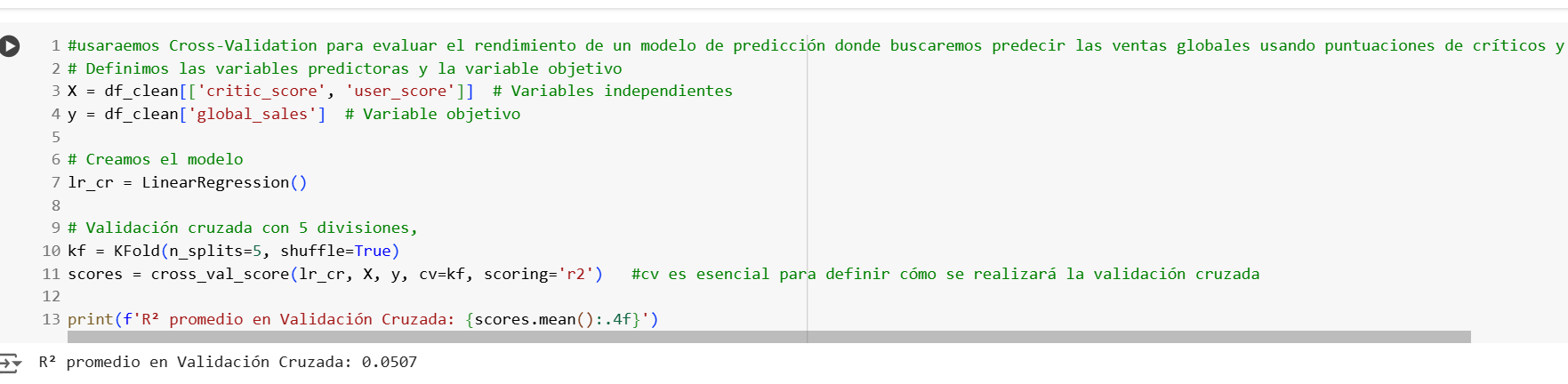


Sobre el primer gráfico tenemos un comportamiento esperado, donde podemos observar que a medida que aumenta el score de las reseñas de críticos aumentan las ventas.

El segundo gráfico muestra una relación interesante entre la puntuación de los usuarios y las ventas globales de los videojuegos. Observamos que algunos juegos con una puntuación de usuarios cercana a 0 tienen ventas sorprendentemente altas. Esto podría explicarse por varios factores, como la falta de reseñas debido a lanzamientos recientes o juegos de nicho, estrategias de marketing agresivas que impulsan las ventas iniciales, o simplemente errores en la recopilación de datos. Sin embargo, es fundamental realizar un análisis más profundo, considerando factores como el género del juego, la plataforma y el tiempo de lanzamiento, para comprender completamente esta relación y obtener conclusiones más sólidas. Al comparar estos datos con otras fuentes de información y analizar tendencias a lo largo del tiempo, podemos obtener una visión más completa de los factores que influyen en el éxito comercial de los videojuegos

**6.- ¿Podemos predecir las ventas globales usando puntuaciones de críticos y usuarios?**

Para poder responder esta pregunta usaremos el cross validation o validación cruzada que nos ayudará a evaluar a través de un modelo de predicción, utilizaremos como variables independientes user y critic score, mientras nuestras global sales es la variable objetivo.

****

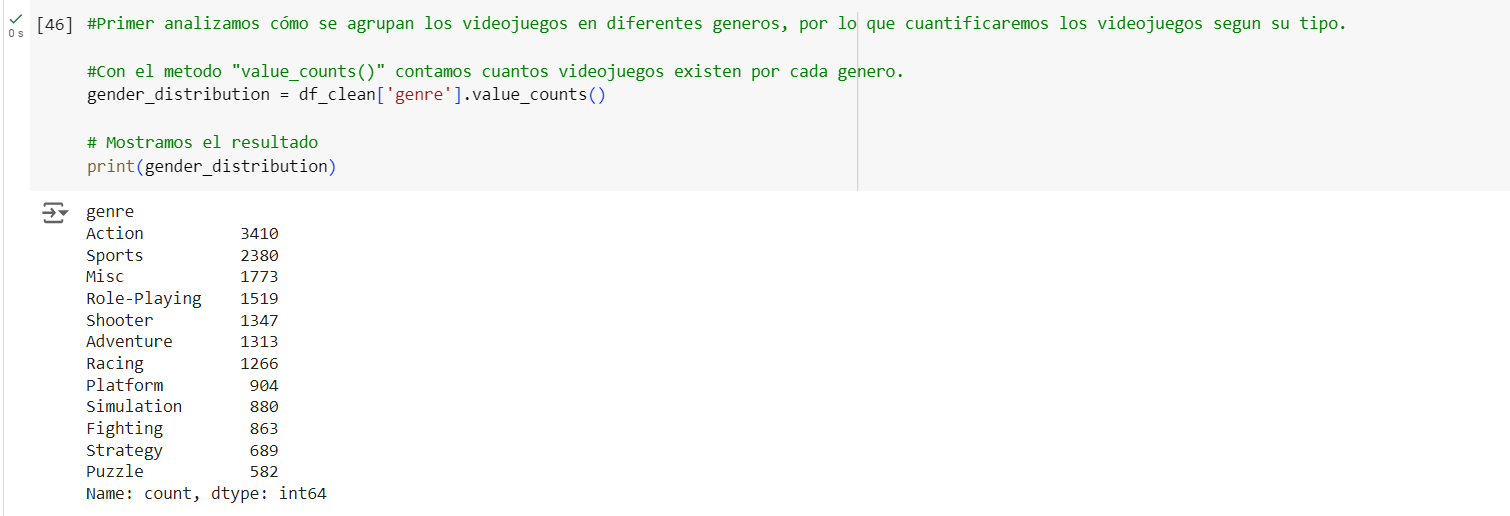
El valor de R² promedio nos indica que nuestras variables predictoras (puntuaciones de críticos y usuarios) explican apenas un 5.44% de la variabilidad en las ventas globales.

Podemos deducir que entonces las puntuaciones de críticos y usuarios, por sí solas, no son suficientes para explicar de manera significativa las ventas globales de los videojuegos. Deben existir otras variables importantes como el género del juego, la plataforma, el presupuesto de marketing, la fecha de lanzamiento, etc., que influyan significativamente en las ventas.

**7.- ¿Cuál es la distribución de los videojuegos por género y plataforma?**

Se realizó un análisis de cómo se agrupan los videojuegos según su género y plataforma.

Iniciamos cuantificando los videojuegos según su tipo.



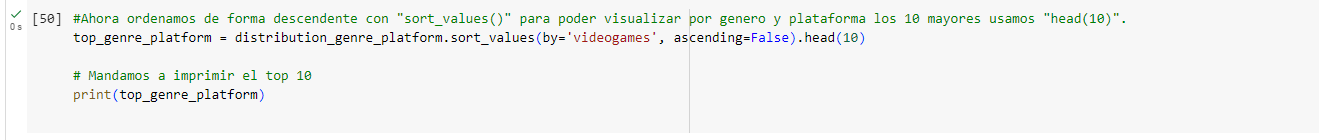
Después por plataforma.



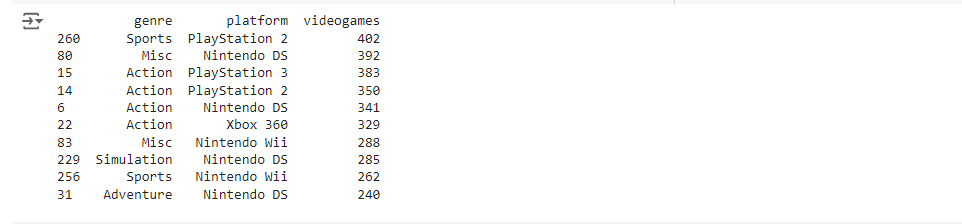
Posteriormente ambos (género y plataforma)



Para finalmente ordenarlos con el top 10, los más relevantes.



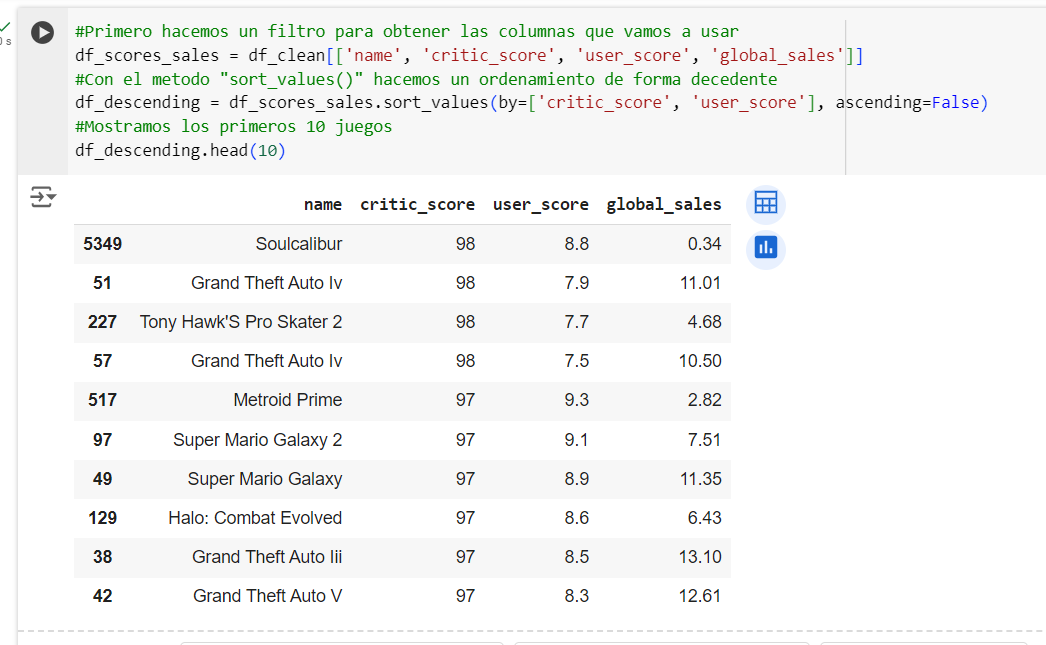
**Resultados**

****

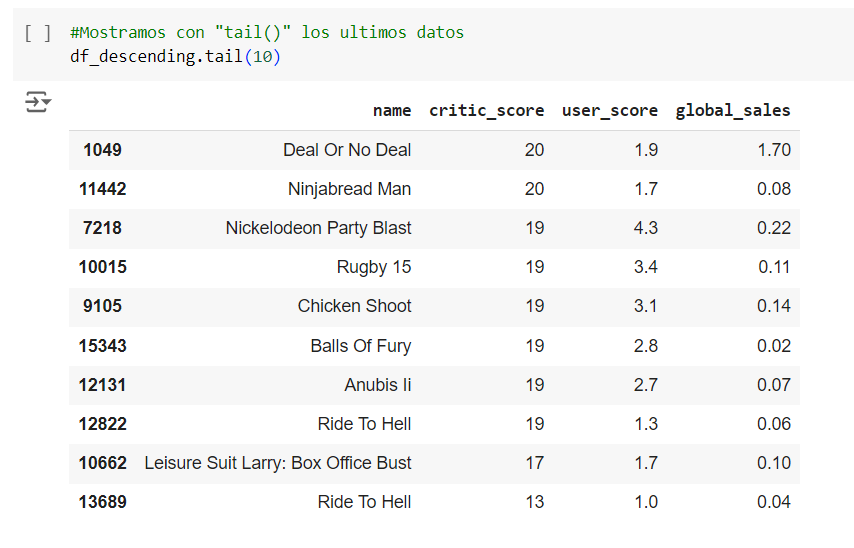
Los resultados muestran que Action (3,410), Sports (2,380), y Misc (1,773) son los géneros con más videojuegos. En términos de plataformas, las más populares son PlayStation 2 (2,188), Nintendo DS (2,164), y PlayStation 3 (1,359). Al combinar género y plataforma, se observa que Sports en PlayStation 2 tiene la mayor cantidad de videojuegos (402), seguido de Misc en Nintendo DS (392) y Action en PlayStation 3 (383).

**8.- ¿Qué relación existe entre las puntuaciones de críticos y usuarios y las ventas globales? ¿Hay juegos con buenas calificaciones que venden poco? ¿Y juegos mal calificados que venden mucho?**

Primero se filtraron las columnas relevantes para el análisis, seleccionando el nombre del juego, las puntuaciones de críticos y usuarios, y las ventas globales. Luego, los juegos se ordenan en forma descendente según las puntuaciones más altas usando la función sort\_values(). Finalmente, se imprimen los 10 juegos mejor calificados para observar si existe una relación entre las altas puntuaciones y las mayores ventas.



Se utiliza la función \*\*tail()\*\* para mostrar los últimos 10 juegos con las peores calificaciones. A pesar de las puntuaciones muy bajas de críticos y usuarios, estos juegos registran algunas ventas, aunque considerablemente menores.



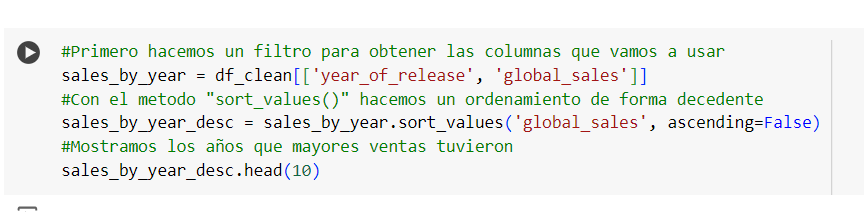
**Resultados**

Juegos como Grand Theft Auto V y Super Mario Galaxy tienen calificaciones excelentes de críticos (97) y usuarios, con ventas que superan los 10 millones de copias, mostrando una relación positiva en estos casos.

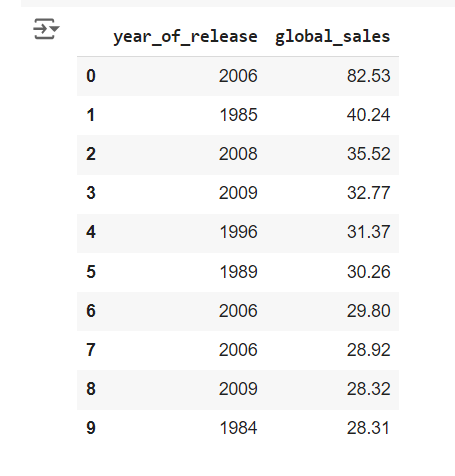
Mientras que juegos como Ride to Hell y Ninjabread Man tienen puntuaciones extremadamente bajas (entre 13 y 20), y las ventas son mínimas, mostrando que en estos casos las malas críticas sí afectan las ventas.

**9.- ¿Hay juegos con buenas calificaciones que venden poco? ¿Y juegos mal calificados que venden mucho?**

Se filtran las columnas relevantes para el análisis, seleccionando el año de lanzamiento y las ventas globales. Luego, se ordenan los datos de forma descendente utilizando sort\_values(). Finalmente, se muestran los 10 años con mayores ventas.



**Resultados**



Los datos muestran que algunos años, como 2006 y 1985, tuvieron ventas significativamente más altas que otros.

### **8.- Consideraciones futuras**

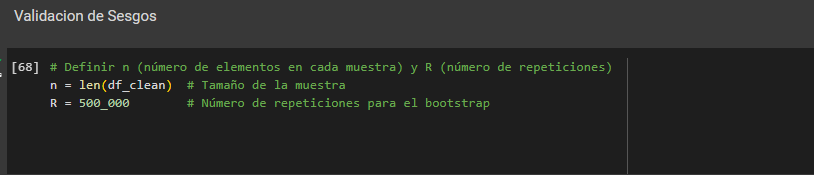
Una vez que los datos estén procesados y limpios, será posible realizar un análisis más profundo. Por ejemplo, se podrían explorar correlaciones avanzadas entre las puntuaciones de los críticos y las ventas globales. Además, este proyecto podría servir como base para crear modelos predictivos que anticipen el éxito comercial de futuros videojuegos basándose en sus características clave. También podría ser interesante analizar el impacto de factores externos como campañas de marketing o lanzamientos simultáneos de consolas.

**8.1. Implementación**

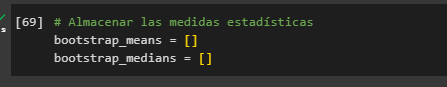
**Bootstrap (validación de sesgos )**

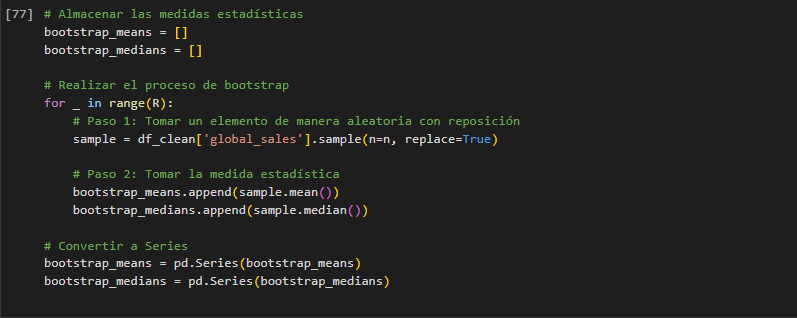
Pasos:

1.Define n (tamaño de la muestra) y R (número de repeticiones):

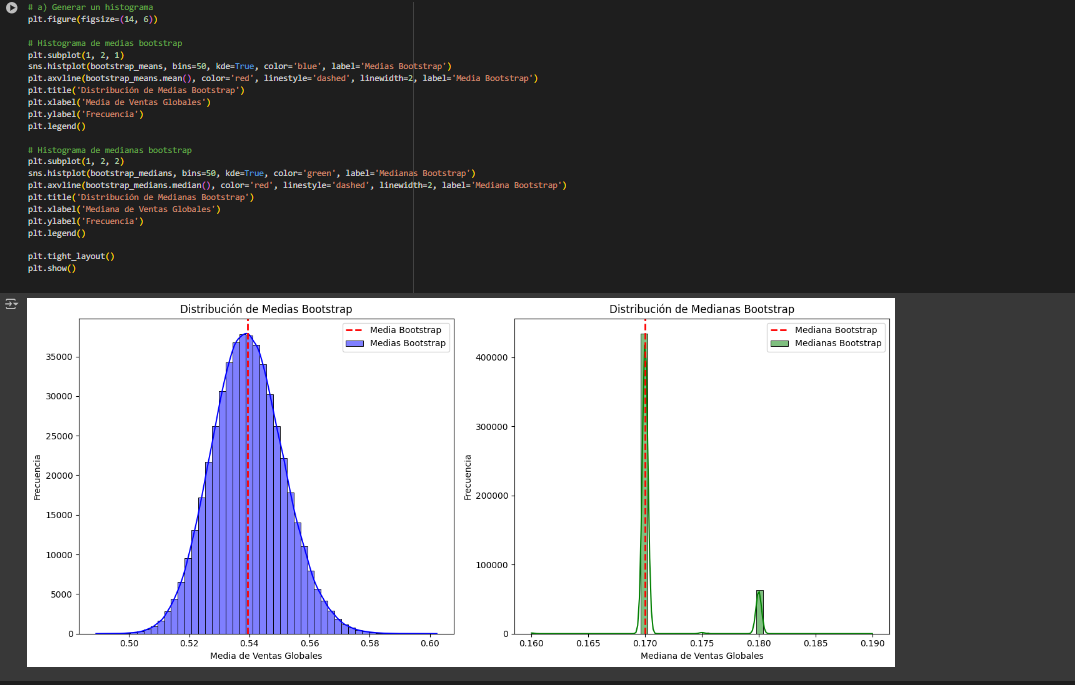


2.Calcular la media y la mediana de cada muestra remuestreada.

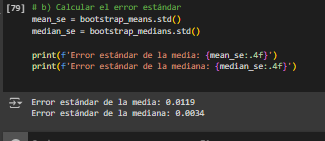




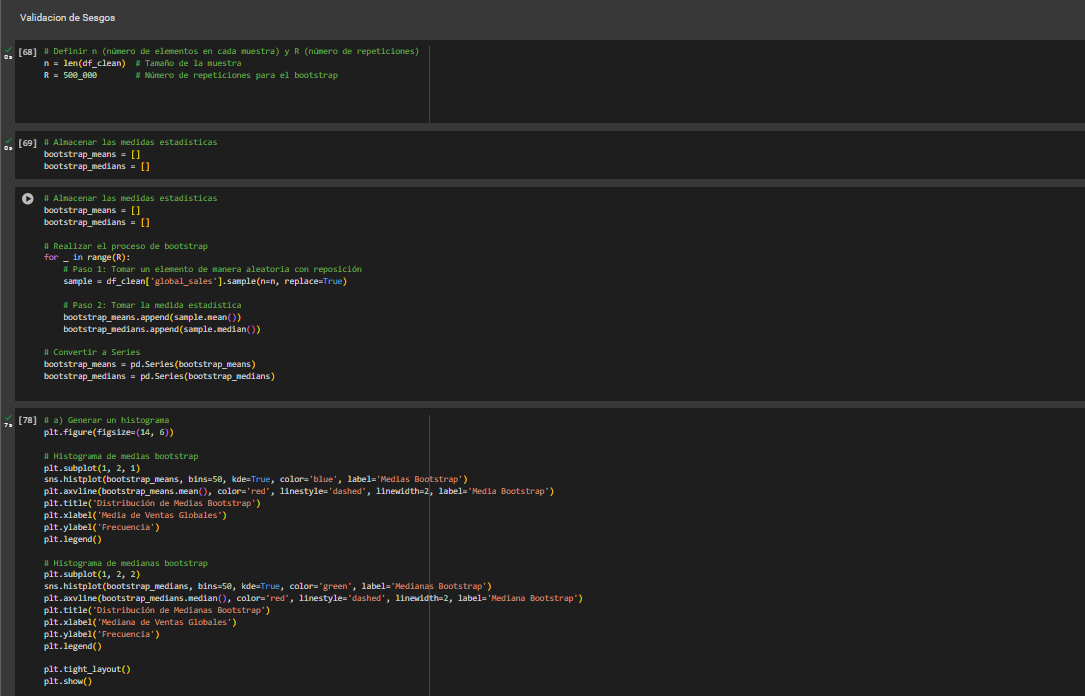
3.Crear histogramas para visualizar la distribución de las medias y medianas calculadas.

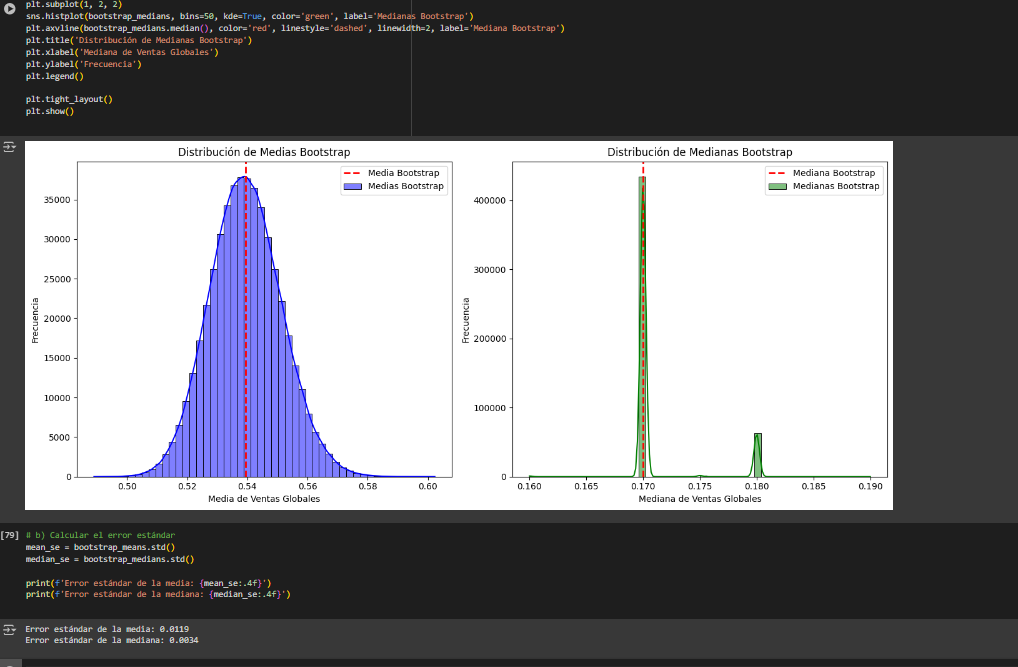


4.Error estándar de las medias y medianas



En el análisis de bootstrap, los errores estándar obtenidos son 0.0119 para la media y 0.0034 para la mediana. Estos valores bajos indican que las estimaciones de la media y la mediana son precisas, con poca variabilidad en torno a los valores originales. La menor variabilidad en la mediana sugiere que es una medida más robusta y menos afectada por valores extremos, lo que refuerza su confiabilidad para representar el centro de los datos.

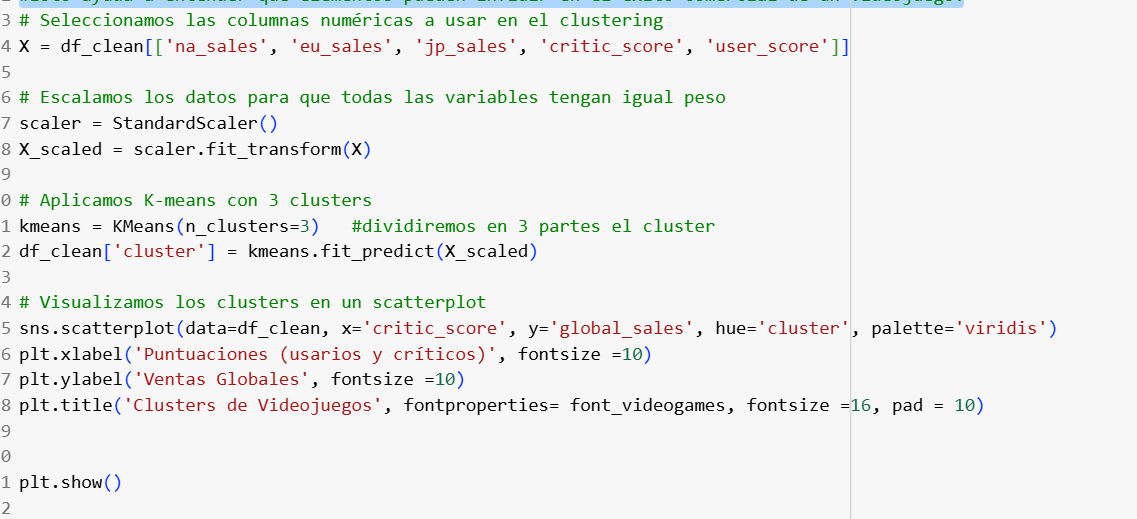


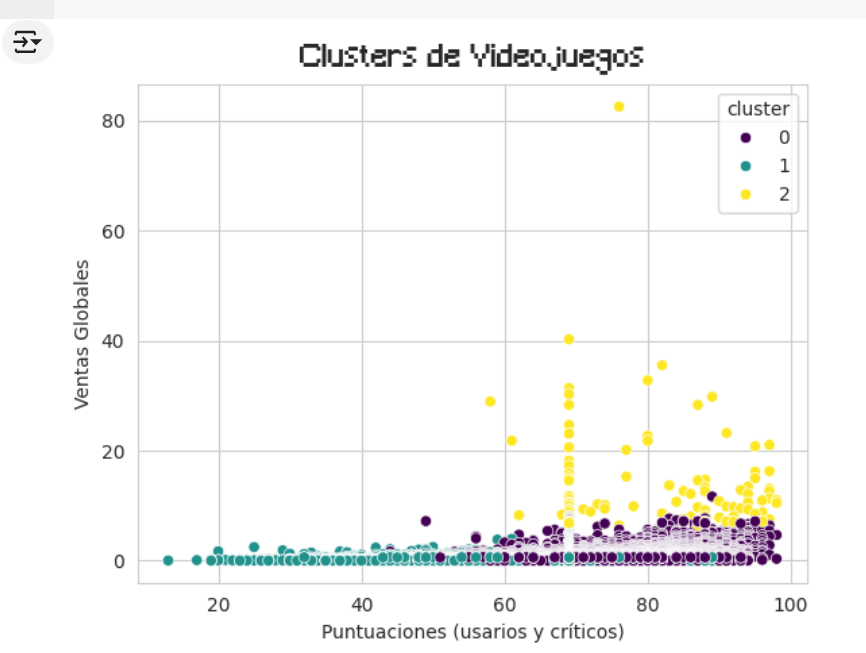


**-Análisis de Clusters de Videojuegos: Relación entre Puntuaciones y Ventas Globales**

El clustering permite identificar patrones en las puntuaciones de los críticos y usuarios en relación con las ventas.

Esto ayuda a entender qué elementos pueden influir en el éxito comercial de un videojuego.





Posibles Interpretaciones:

Basándonos en la distribución de los puntos en el gráfico, podemos hacer algunas inferencias sobre lo que cada cluster podría representar:

* Cluster 1 (color morado) Videojuegos independientes o de nicho: Estos juegos suelen tener puntuaciones de crítica más bajas y ventas globales más modestas. Podrían ser juegos indie, juegos con temáticas muy específicas o juegos que no han recibido una gran promoción. Otra opción son videojuegos que se volvieron populares debido al marketing.
* Cluster 2 (color verde): Videojuegos de gran éxito comercial: Estos juegos suelen tener puntuaciones de crítica moderadas a altas y ventas globales muy altas. Son los juegos que todos conocemos y que suelen estar en las listas de los más vendidos.
* Cluster 3 (color amarillo): Videojuegos con alta puntuación de crítica pero bajas ventas: Este cluster podría incluir juegos que han sido muy bien recibidos por la crítica, pero que no han logrado alcanzar un gran éxito comercial. Podrían ser juegos con mecánicas de juego innovadoras o con temáticas muy específicas que no han conectado con el público general.

--- **Métrica o estadística a utilizar (Pruebas A/B )**

* ¿Hay juegos con buenas calificaciones que venden poco? ¿Y juegos mal calificados que venden mucho?

¿Existen diferencias significativas en las ventas entre videojuegos con puntuaciones altas y bajas de críticos o usuarios?

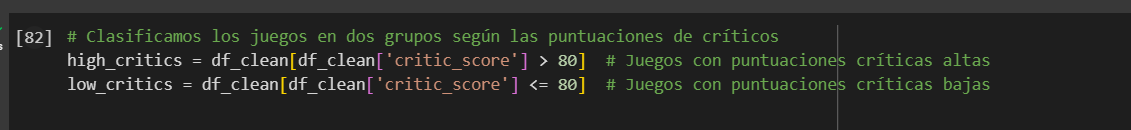
**Métrica**: La métrica que utilizaremos para comparar el comportamiento de los grupos será la **media de ventas globales**. Esto nos permitirá evaluar si existe una diferencia significativa en las ventas entre los videojuegos con puntuaciones altas y bajas.

**Test de hipótesis**

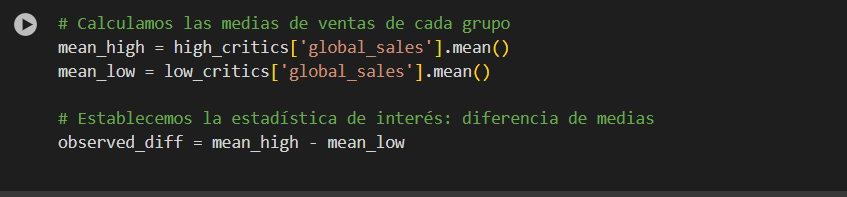
**Hipótesis**:

* **Hipótesis nula (H0)**: No hay diferencia significativa en las ventas medias entre videojuegos con puntuaciones altas y videojuegos con puntuaciones bajas.
* **Hipótesis alternativa (H1)**: Hay una diferencia significativa en las ventas medias entre videojuegos con puntuaciones altas y videojuegos con puntuaciones bajas.

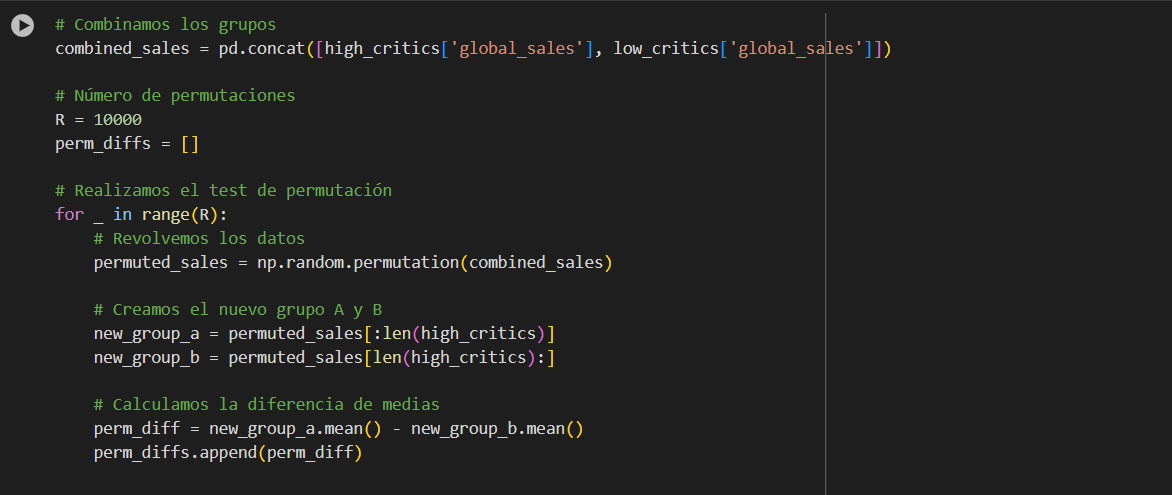
1.Clasificar los juegos en grupos según las puntuaciones de críticos



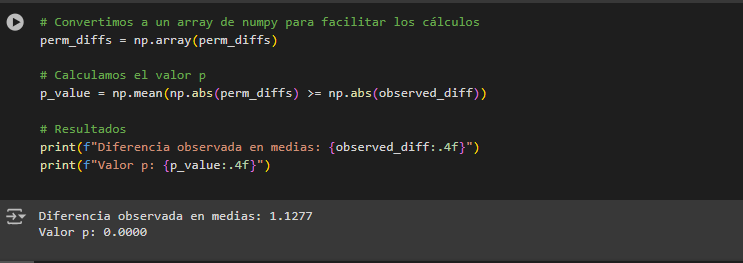
2.Calcular las medias de ventas de cada grupo



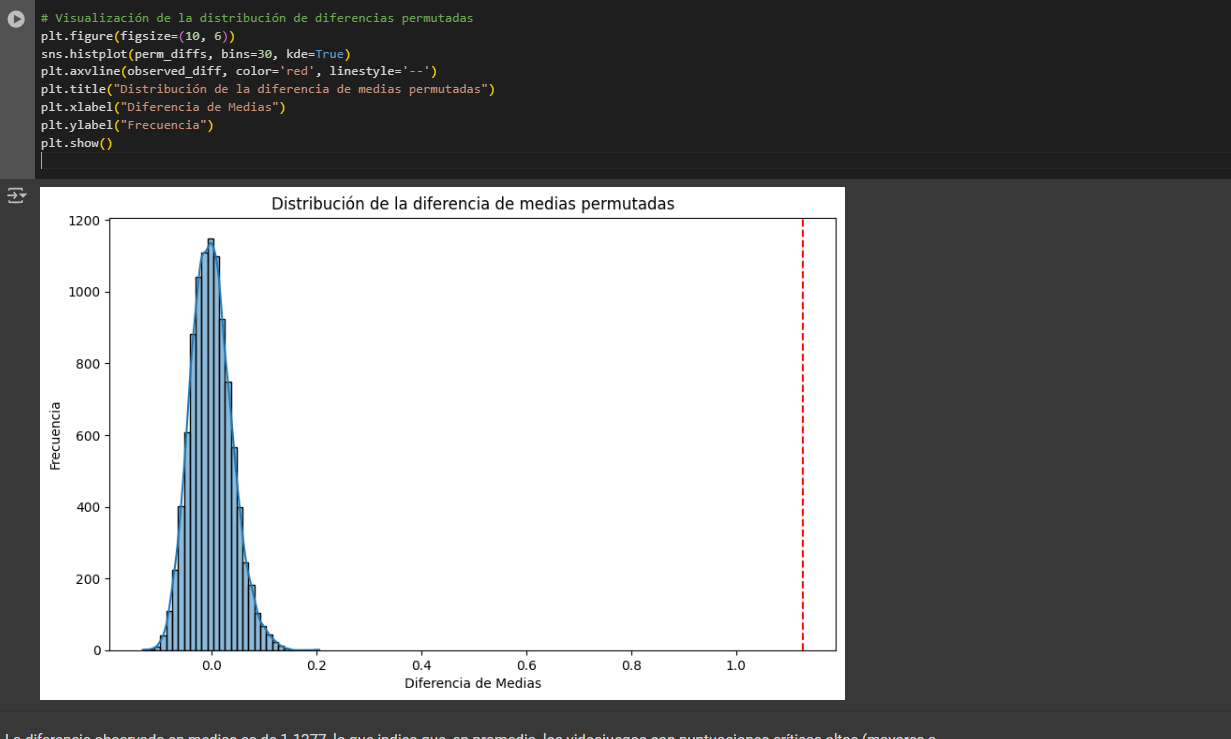
3.Combinación de grupos y realización del test de permutación



4.Cálculo del valor p y resultados



5.Visualización de la distribución de diferencias permutadas

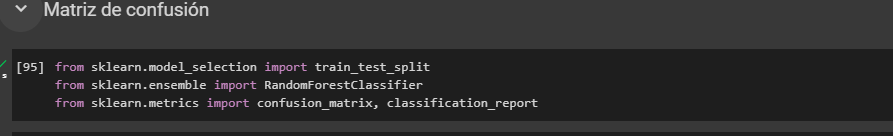


La diferencia observada en medias es de 1.1277, lo que indica que, en promedio, los videojuegos con puntuaciones críticas altas (mayores a 80) venden 1.1277 millones de unidades más que aquellos con puntuaciones críticas bajas (80 o menos). Además, el valor p de 0.0000 sugiere una evidencia muy fuerte en contra de la hipótesis nula, lo que implica que hay una diferencia significativa en las ventas entre estos dos grupos. Los resultados respaldan la idea de que los juegos con mejores críticas tienden a vender más que los mal calificados.

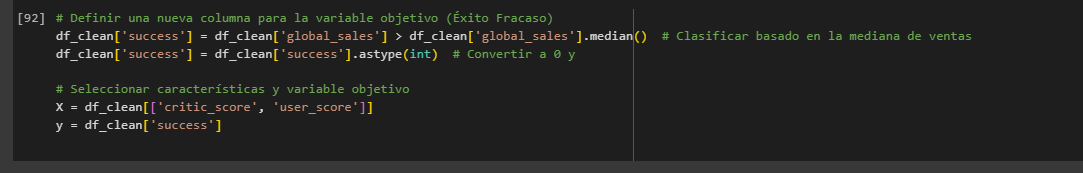
**Rendimiento de un modelo de machine learning**

**Intenta predecir el éxito o fracaso de un videojuego en función de características específicas, como las puntuaciones de críticos y usuarios.**

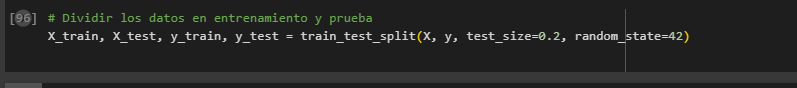
1.Importar librerias faltantes



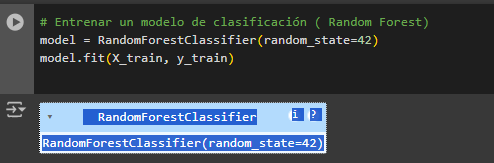
2.Preparar datos



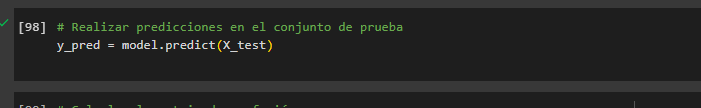
3.Dividir los Datos en Conjuntos de Entrenamiento y Prueba



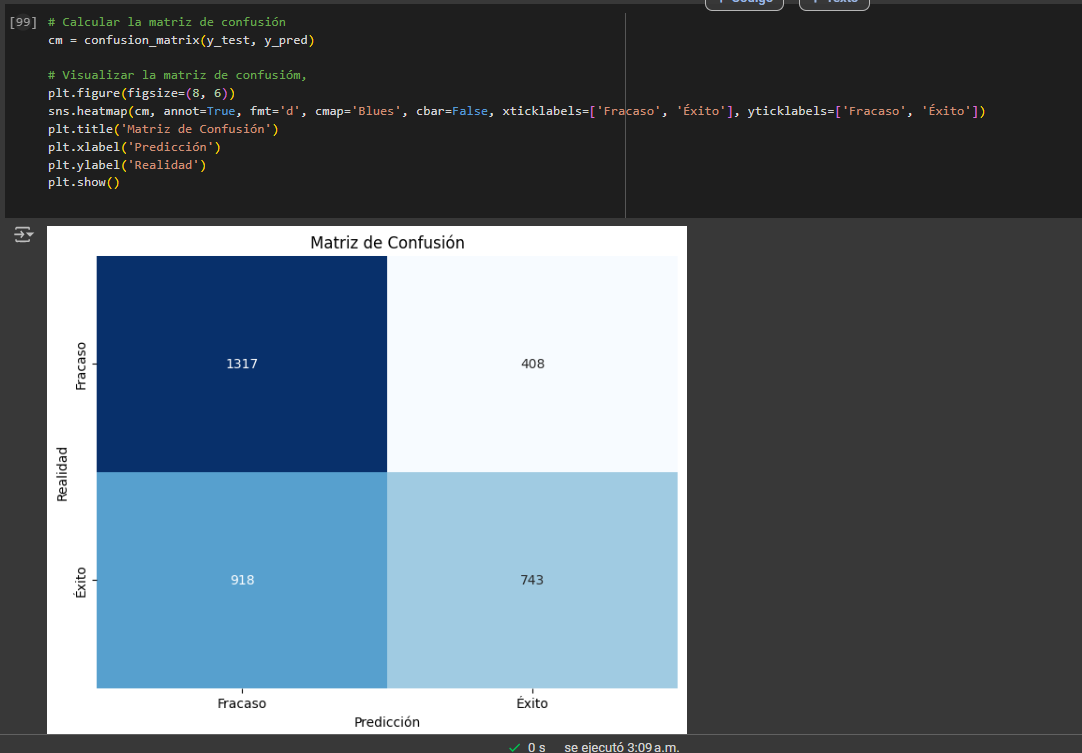
4.Entrenar el Modelo



5.Predicciones



6.Resultados y Matriz de Confusión



### Interpretación de los Valores

1. Verdaderos Positivos (VP):
   * 743: Estos son los casos en los que el modelo predijo correctamente que un videojuego sería un "Éxito". Es decir, 743 videojuegos que realmente fueron exitosos y que el modelo también clasificó como tales.
2. Verdaderos Negativos (VN):
   * 1317: Estos son los casos en los que el modelo predijo correctamente que un videojuego sería un "Fracaso". Así que 1317 videojuegos que realmente fracasaron y que el modelo también clasificó como fracasados.
3. Falsos Positivos (FP):
   * 408: Estos son los casos en los que el modelo predijo incorrectamente que un videojuego sería un "Éxito", pero en realidad fue un "Fracaso". Es decir, 408 videojuegos que el modelo clasificó como éxitos, pero no lo fueron.
4. Falsos Negativos (FN):
   * 918: Estos son los casos en los que el modelo predijo incorrectamente que un videojuego sería un "Fracaso", pero en realidad fue un "Éxito". Es decir, 918 videojuegos que el modelo clasificó como fracasos, pero sí tuvieron éxito.

### **9. Conclusión**

En este análisis, hemos examinado diversos aspectos clave del mercado de videojuegos, revelando patrones sobre las plataformas, géneros y su relación con las ventas y calificaciones.

El procesamiento de datos es esencial para comprender las tendencias en la industria de los videojuegos y su relación con las ventas globales. En este proyecto, nos enfocamos en limpiar y analizar un conjunto de datos globales de videojuegos, permitiendo futuras investigaciones sobre las relaciones entre plataformas, géneros, puntuaciones y ventas. Con una base de datos procesada, será posible avanzar hacia análisis más complejos y construir modelos predictivos que apoyen la toma de decisiones estratégicas en el desarrollo y publicación de videojuegos.

1. Plataformas con mayores ventas globales: Las consolas PlayStation 2, Xbox 360 y PlayStation 3 dominan el mercado, demostrando que la calidad, el acceso a juegos icónicos y una sólida base de jugadores son fundamentales para el éxito comercial. Estas plataformas fueron líderes durante generaciones cruciales en la evolución de los videojuegos.

2. Géneros más vendidos: Los videojuegos de Acción, Deportes y Shooter generan las ventas más altas a nivel global. Esto refleja una clara preferencia por juegos con alta adrenalina, interacción directa y experiencias inmersivas, atrayendo tanto a jugadores casuales como dedicados.

3. Distribución por género y plataforma: Al analizar la distribución de títulos por género y plataforma, se observa que la PlayStation 2 y la Nintendo DS ofrecen una amplia variedad de juegos, especialmente en los géneros de Acción y Deportes. Esto subraya la importancia de contar con un catálogo diverso para mantener la competitividad y satisfacer una amplia gama de preferencias.

4. Relación entre puntuaciones y ventas: Aunque existe una correlación positiva entre altas calificaciones (de críticos y usuarios) y ventas en títulos como Grand Theft Auto V y Super Mario Galaxy, también se encontraron ejemplos de juegos con bajas puntuaciones (como Ride to Hell) que lograron ventas significativas. Esto sugiere que factores como la nostalgia, el marketing y la franquicia a la que pertenecen pueden impulsar las ventas, incluso ante críticas negativas.

En resumen, nuestro análisis no solo responde a preguntas clave sobre los factores que impulsan las ventas, sino también a cómo las preferencias de los jugadores, la atracción de los géneros y las estrategias de las plataformas se combinan para crear fenómenos globales en la industria del videojuego.

### **10. Entregables**

* Código ejecutable en Google Colab o Jupyter Notebook que realice el procesamiento y análisis de datos.
* Documento en Google Docs o Word que describa los pasos y resultados del procesamiento y análisis de datos.
* Presentación en PowerPoint o Google Slides con un resumen de las técnicas aplicadas y los resultados obtenidos en el análisis de los datos.