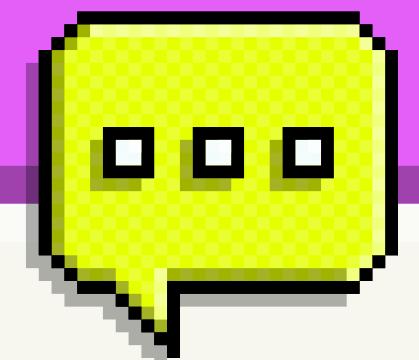
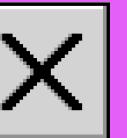
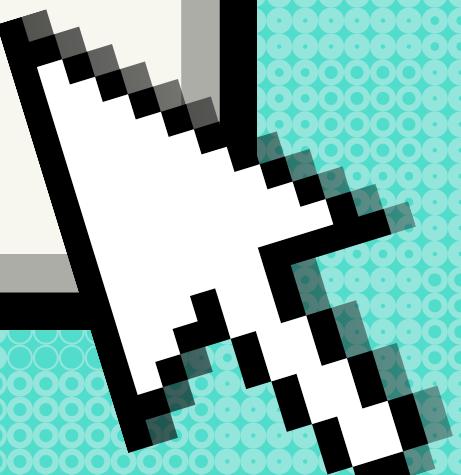


Equipo 30



VIDEOJUEGOS



Proyecto final

índice

01. Introducción
02. Objetivos
03. Planteamiento del problema
04. Resultados
05. Conclusiones

¡Comenzamos!

Introducción

Análisis de Videojuegos: Factores Clave del Éxito

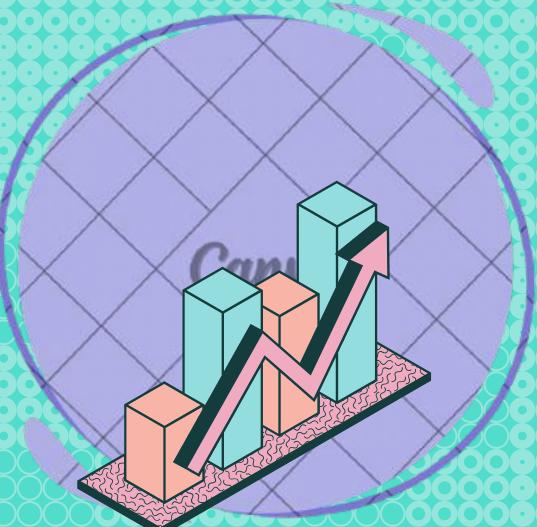
Popularidad mundial de los videojuegos como entretenimiento.



Impacto en las ventas globales y regionales.



Factores del éxito: plataforma, género, críticas y ventas

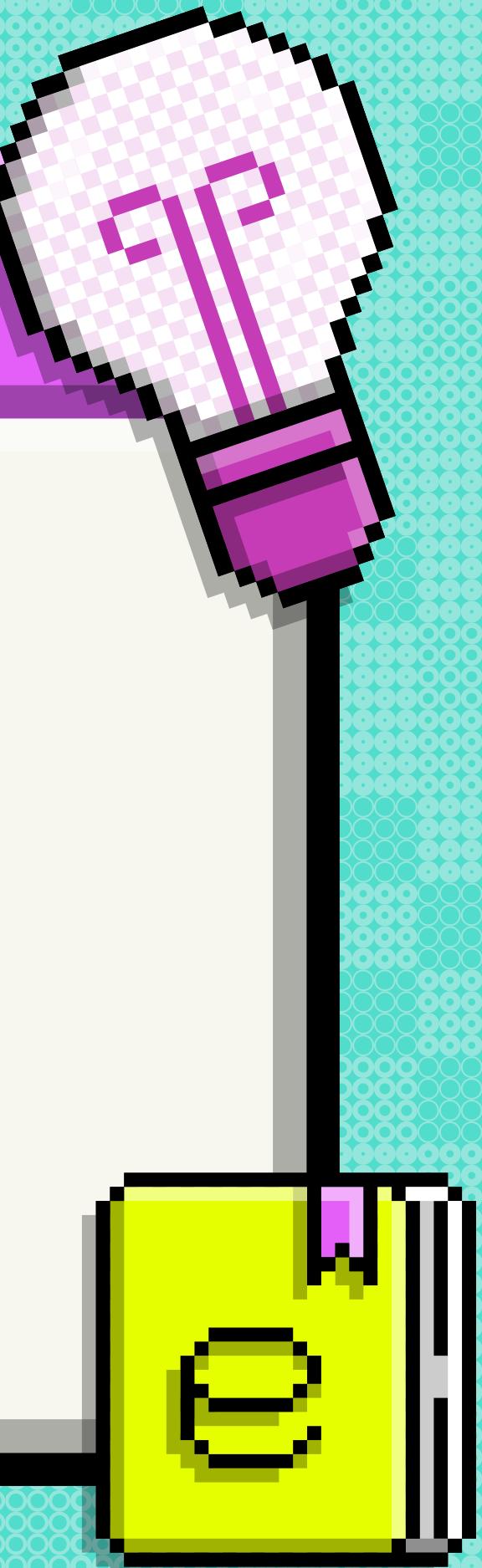
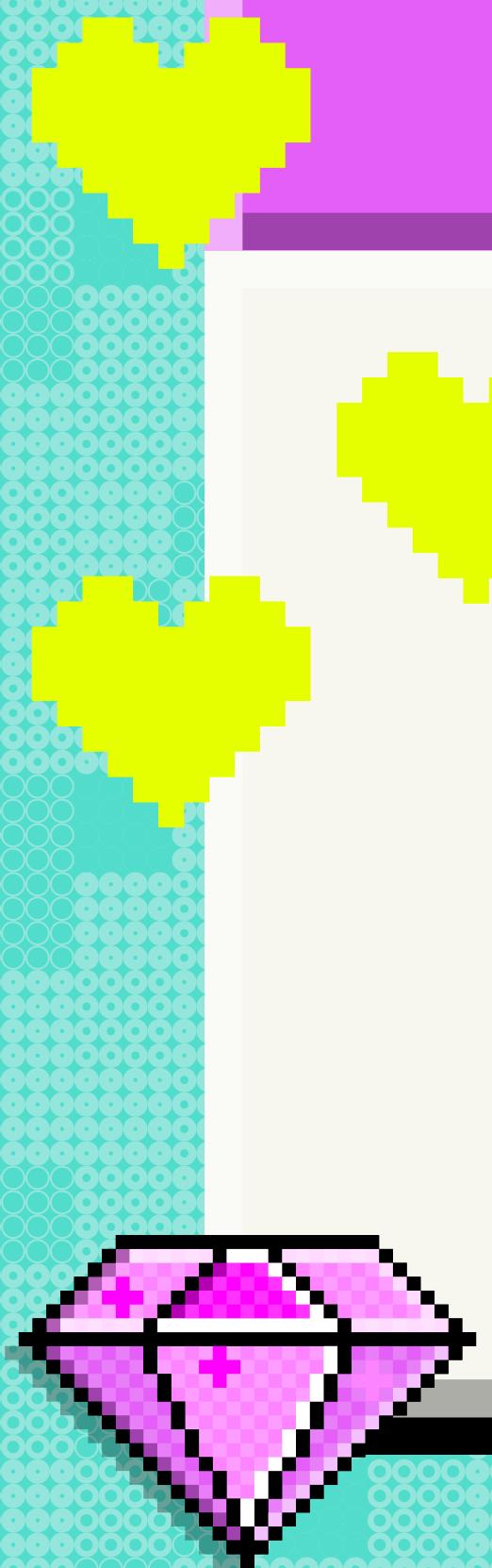


Estudio de datos globales: plataforma, género y calificaciones.



Patrones que explican el éxito de ciertos videojuegos.

Objetivos



Generales

- Analizar la relación entre ventas, plataformas, géneros y puntuaciones para identificar los factores que influyen en el éxito comercial de los videojuegos.

Específicos

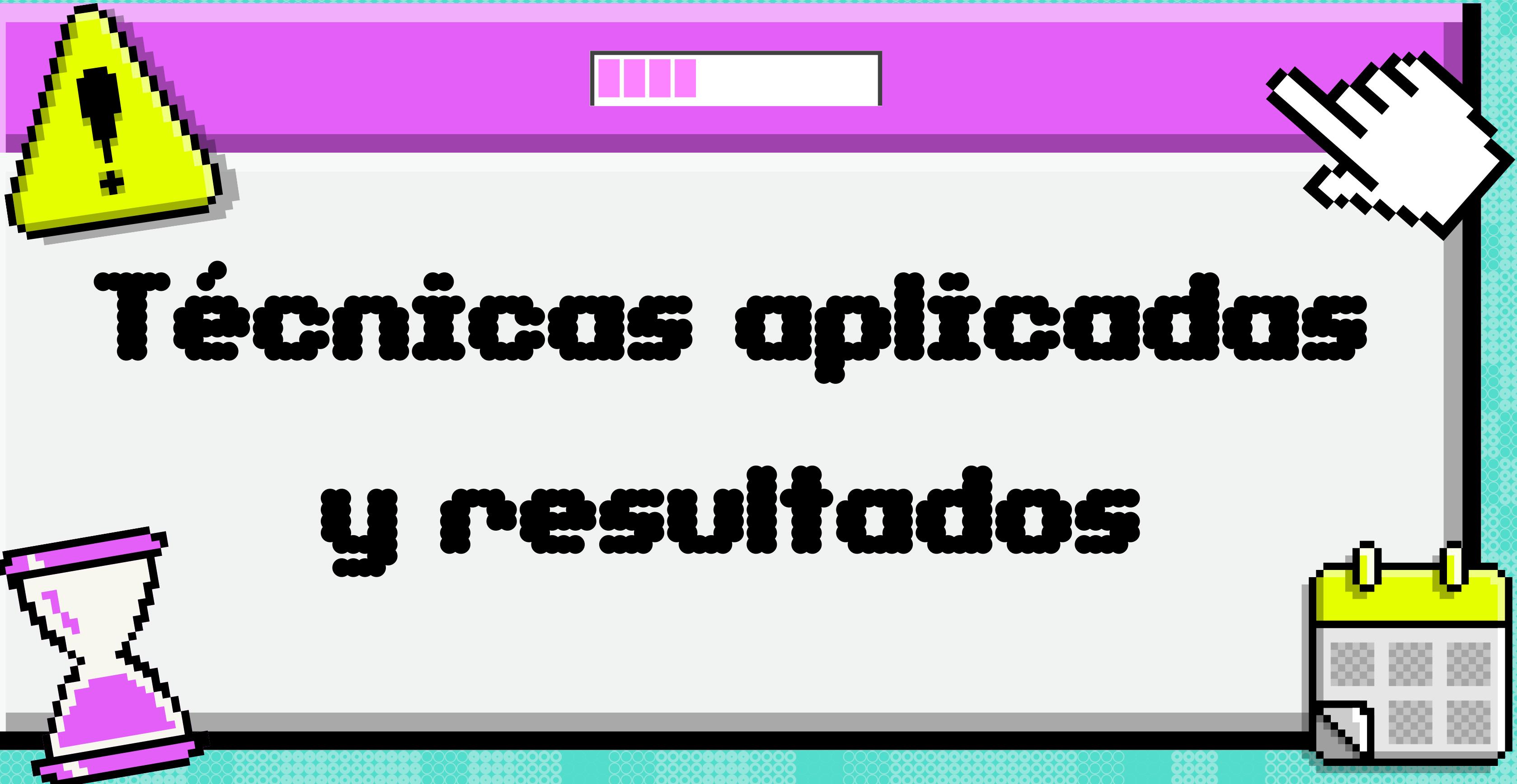
- Preparar los datos de ventas, plataformas, géneros y puntuaciones.
- Estudiar la relación entre ventas globales y regionales.
- Comparar el éxito de diferentes plataformas en ventas.
- Evaluar el impacto de las puntuaciones en las ventas.
- Identificar tendencias de ventas por género y año de lanzamiento.

Planteamiento del problema

La industria de los videojuegos busca identificar qué factores determinan el éxito comercial.

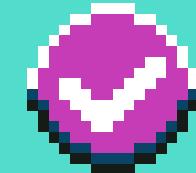
Aunque la plataforma y el género son claves, otros elementos como las calificaciones, preferencias regionales y el año de lanzamiento también influyen.

Existe una desconexión entre críticas y ventas, ya que los juegos más vendidos no siempre son los mejor valorados. Este proyecto analizará datos para descubrir qué características impactan más en las ventas globales y regionales.

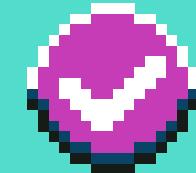


Técnicos aplicados
y resultados

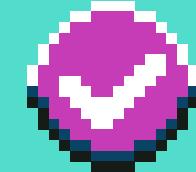
Limpieza de la BD



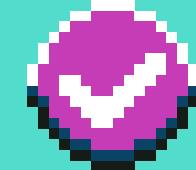
01. Importar bibliotecas



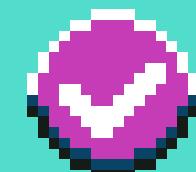
02. Cargar datos (csv, api)



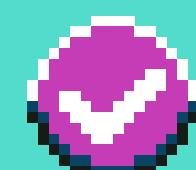
03. Hacer la exploración



04. Limpiar los datos



05. Transformar los datos



06. Dar formato a la BD



#1. Bibliotecas

Import requests / pandas / numpy / pprint / display

#2. Datos

df = pd.read_csv('ruta')

response = requests.get(url)

#3. Exploración

df.head / df.shape / df.isnull().sum() / df.nunique() / df.dtypes /
df.duplicated().sum() / df.columns

#4. Limpiar Datos

df.dropna(axis=1, how='all').reset_index(drop=True)

df.drop_duplicates()

df.drop(columns=['columna'])

#5. Transformar

.astype('Int64')

.apply(pd.to_numeric, errors='coerce')

.fillna() #fechas aleatoria en rango, promedios, 0

.str.title/upper/replace/strip()

#6. Formato

.rename #columnas a través de diccionario

df['platform'].map(mapping) #aplicar diccionario con + descripción

.reset_index(drop=True)

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

El código calcula el rango de ventas de videojuegos en varias regiones [Norteamérica, Europa, Japón y otras] tomando la diferencia entre las ventas máximas y mínimas de cada área. Utiliza un diccionario para asociar las columnas de ventas con sus nombres y luego imprime los resultados por región.

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.

```
1 #Definimos un diccionario que contenga las columnas por region y su el nombre de la region
2 regions = {
3     'na_sales': 'Norteamérica',
4     'eu_sales': 'Europa',
5     'jp_sales': 'Japón',
6     'other_sales': 'otras regiones'
7 }
8 #Creamos un for que itere cada region dentro del diccionario
9 #Para sacar el rango de ventas debemos buscar la diferencia entre el valor maximo y minimo de las ventas. Con cada iteracion se busca el valor maximo con la funcion ".max()" y hacemos una resta con el valor minimo con la funcion ".min()", este proceso se repite por cada region que exista.
10 for region, name in regions.items():
11     sale_range = df_clean[region].max() - df_clean[region].min()
12     print("—" * 20)
13     print(f'Rango de ventas en {name}: {sale_range} millones')
14 print("—" * 20)
15
```

```
Rango de ventas en Norteamérica: 41.36 millones
-----
Rango de ventas en Europa: 28.96 millones
-----
Rango de ventas en Japón: 10.22 millones
-----
Rango de ventas en otras regiones: 10.57 millones
-----
```

```
[ ] 1 #Para sacar el rango de ventas a nivel mundial podemos repetir el mismo proceso.
2 global_sales_range = df_clean['global_sales'].max() - df_clean['global_sales'].min()
3 print("—" * 20)
4 print(f'Rango de ventas globales: {global_sales_range} millones')
5 print("—" * 20)
```

```
Rango de ventas globales: 82.52 millones
-----
```

¿Qué plataformas generan mayores ventas globales?

```
[44] #Primero usamos el metodo "groupby()" para ordenar las ventas globales por la columna de las plataformas  
sales_by_platform = df_clean.groupby('platform')['global_sales'].sum()  
  
#Usamos la nueva variable y con "sort_values()" ordenamos los valores de mayor a menor.  
sales_by_platform_sorted = sales_by_platform.sort_values(ascending=False)  
  
#Mostramos con "head()" el top 10 de las plataformas con mayores ventas  
top_platforms = sales_by_platform_sorted.head(10)  
print("—" * 20)  
print(top_platforms)  
print("—" * 20)
```

```
-----  
platform  
PlayStation 2      1283.08  
Xbox 360          988.36  
PlayStation 3      962.71  
Nintendo Wii       937.27  
Nintendo DS        813.86  
PlayStation         759.88  
Game Boy Advance   327.11  
PlayStation 4       324.48  
PlayStation Portable 298.02  
Nintendo 3DS       271.21  
Name: global_sales, dtype: float64  
-----
```

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.

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

¿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.

Los géneros más vendidos son Action, Sports y Shooter. Destacando la popularidad de los juegos dinámicos y competitivos a nivel mundial, además ofrecen experiencias intensas y variadas destacan entre su popularidad.

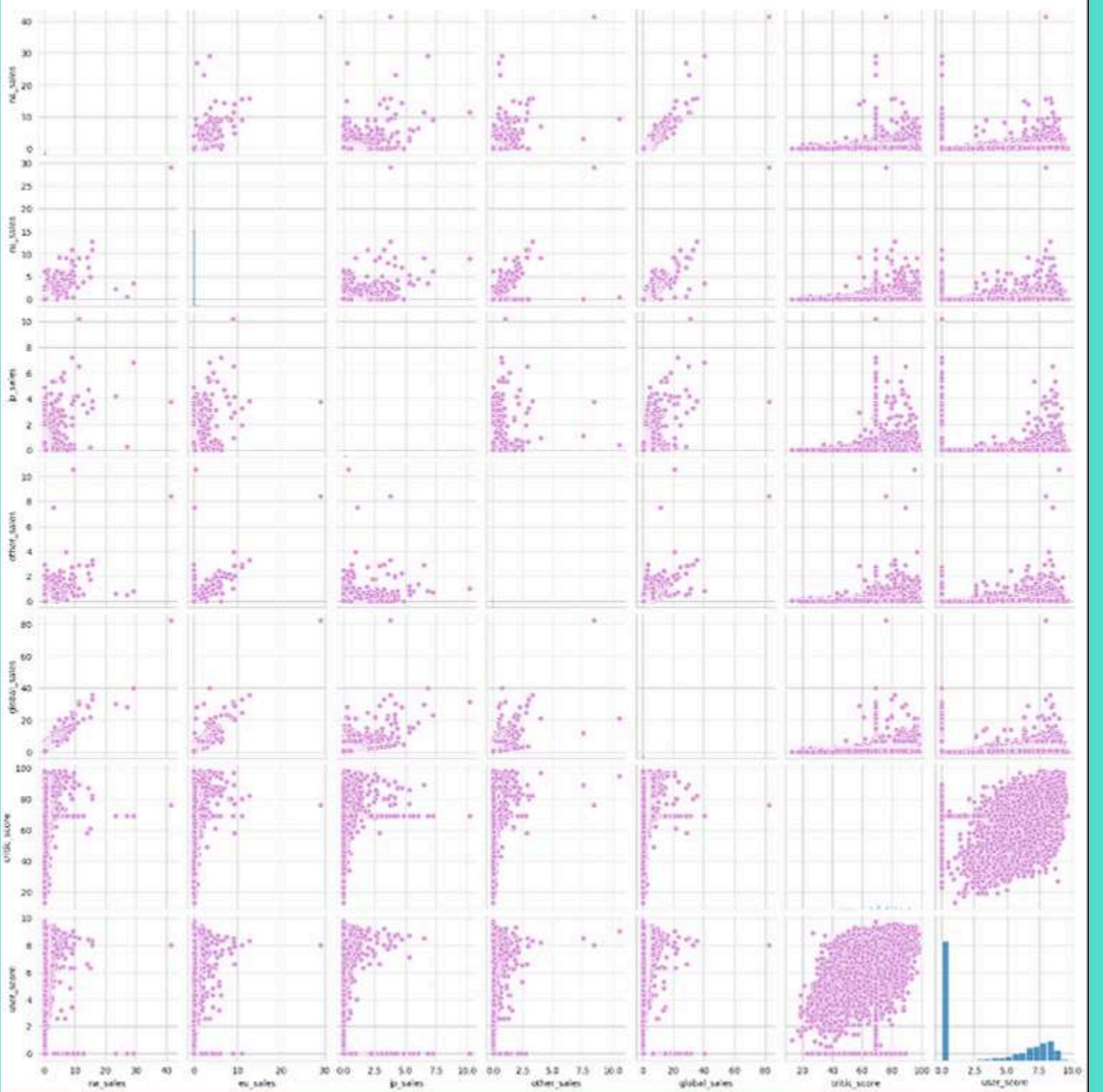
```
[1]: primero usamos el metodo "groupby()" para ordenar las ventas globales por la columna de sales_by_gender = df_clean.groupby('genre')['global_sales'].sum()
```

```
ahusmos la nueva variable y con "sort_values()" ordenamos los valores de mayor a menor, sales_by_gender_sorted = sales_by_gender.sort_values(ascending=False)
```

```
mostramos con "head()" el top 10 de los géneros con mayores ventas  
top_genders = sales_by_gender_sorted.head(10)  
print("... * 20")  
print(top_genders)  
print("... * 20")
```

```
...  
genre  
Action      1771.73  
Sports       1350.61  
Shooter      1086.67  
Role-Playing 958.62  
Platform     850.73  
Misc         830.19  
Racing        757.92  
Fighting      467.91  
Simulation    394.12  
Puzzle        243.65  
Name: global_sales, dtype: float64
```

Pairplot de Variables de Interés



PAIRPLOT DE LAS VARIABLES DE INTERÉS

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.

Leve correlación positiva entre las puntuaciones de críticos y usuarios y las ventas globales, esta relación es un poco más débil que la correlación entre las diferentes regiones.

Observamos también una correlación positiva entre critic score y user score lo que tendría sentido pues si un juego es bueno tendrá buenas reseñas tanto por críticos como por usuarios.

¿Qué relación existe entre las puntuaciones de críticos y usuarios y las ventas globales?

Elegimos columnas numéricas de interés:

'na_sales', 'eu_sales', 'jp_sales', 'other_sales',
'global_sales', 'critic_score', 'user_score'

Generamos una matriz de correlación y la gráfiamos en un heatmap:

```
sns.heatmap[matriz_correlacion, annot=True,  
cmap='GnBu', vmin=-1, vmax=1]
```

Scatterplots de ventas globales y puntuaciones:

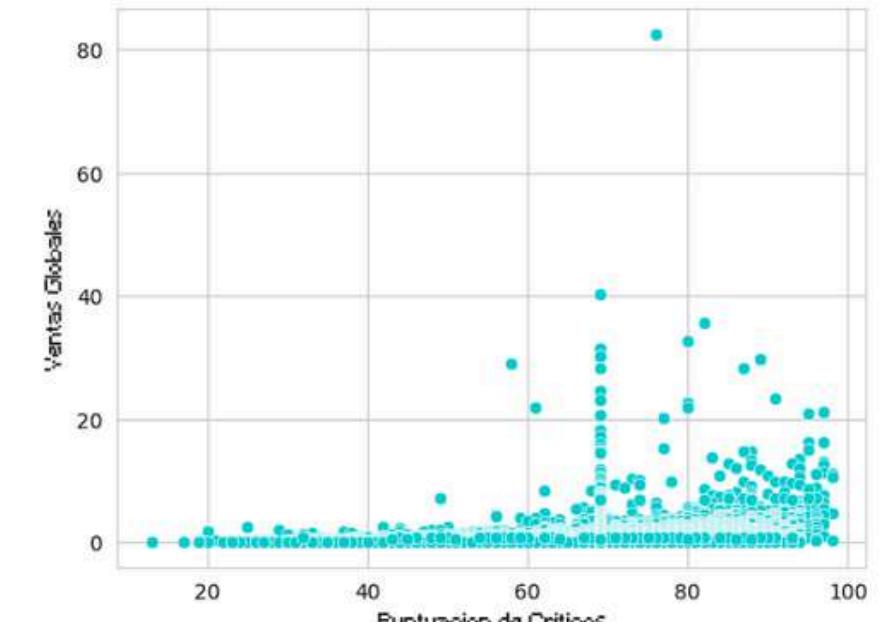
```
sns.scatterplot[data=df_clean, x='critic_score',  
y='global_sales', color = 'darkturquoise']
```

```
sns.scatterplot[data=df_clean, x='user_score',  
y='global_sales', color = 'violet']
```

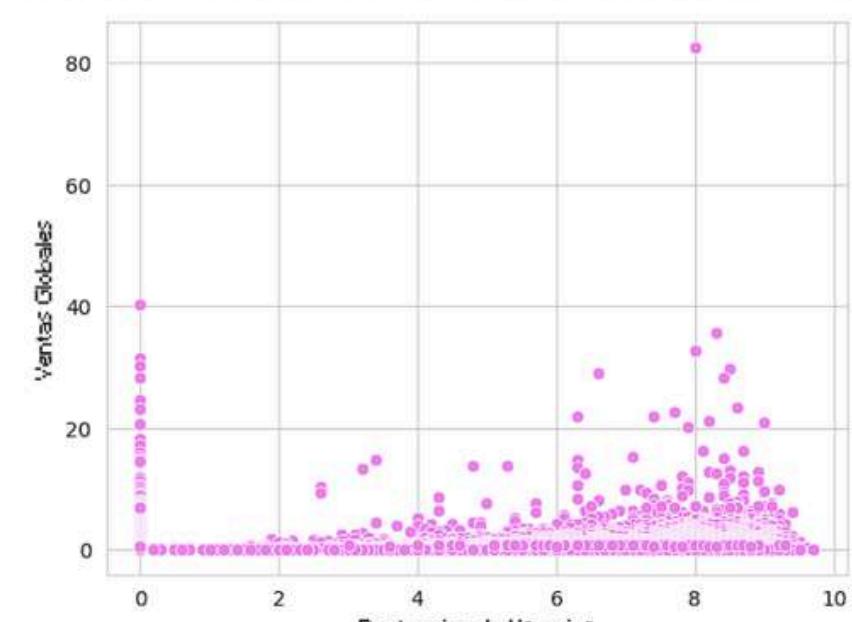
Heatmap de Correlación



Relación entre Puntuación de Críticos y Ventas Globales



Relación entre Puntuación de Usuarios y Ventas Globales



¿Podemos predecir las ventas globales usando puntuaciones de críticos y usuarios?

```
1 #Cross-Validation
2 #Lo usaremos para evaluar el rendimiento de un modelo de predicción
3 # Definimos las variables predictoras y la variable objetivo
4 X = df_clean[['critic_score', 'user_score']] # Variables Independientes
5 y = df_clean['global_sales'] # Variable objetivo
6
7 # Creamos el modelo
8 lr_cr = LinearRegression()
9
10 # Validación cruzada con 5 divisiones,
11 kf = KFold(n_splits=5, shuffle=True)
12 scores = cross_val_score(lr_cr, X, y, cv=kf, scoring='r2') #cv = cantidad de divisiones
13
14 print(f'R2 promedio en Validación Cruzada: {scores.mean():.4f}')
```

R² promedio en Validación Cruzada: 0.0544

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.

¿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 (H_0): No hay diferencia significativa en las ventas medias entre videojuegos con puntuaciones altas y videojuegos con puntuaciones bajas.
- Hipótesis alternativa (H_1): Hay una diferencia significativa en las ventas medias entre videojuegos con puntuaciones altas y videojuegos con puntuaciones bajas.

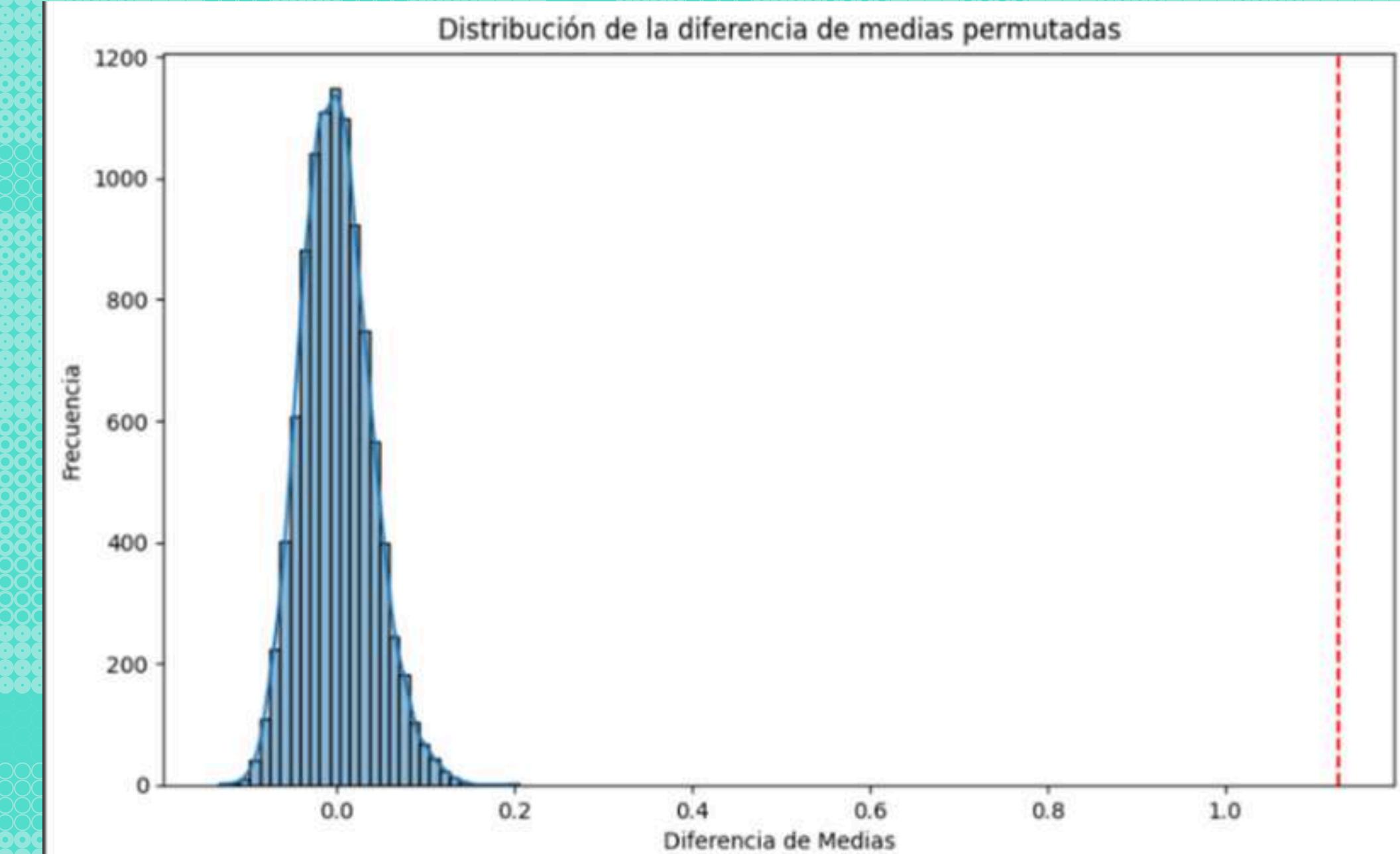
¿Existen diferencias significativas en las ventas entre videojuegos con puntuaciones altas y bajas de críticos o usuarios?

Test de hipótesis

Hipótesis:

Hipótesis nula (H_0): No hay diferencia significativa en las ventas medias entre videojuegos con puntuaciones altas y videojuegos con puntuaciones bajas.

Hipótesis alternativa (H_1): Hay una diferencia significativa en las ventas medias entre videojuegos con puntuaciones altas y videojuegos con puntuaciones bajas.



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

Pasos

- Combina los resultados de ambos grupos en un mismo conjunto de datos
- Revuelve los datos
- Usando muestreo aleatorio sin reposición, construye un nuevo grupo A del mismo tamaño que el original.
- El resto de los datos conforman nuestro nuevo grupo B.

Cuantifica la métrica o estadística que calculaste con los grupos originales y guarda el resultado.

```
▶ # Convertimos a un array de numpy para facilitar los cálculos  
perm_diffs = np.array(perm_diffs)  
  
# Calculamos el valor p  
p_value = np.mean(np.abs(perm_diffs) >= np.abs(observed_diff))  
  
# Resultados:  
print(f"Diferencia observada en medias: {observed_diff:.4f}")  
print(f"Valor p: {p_value:.4f}")
```

Diferencia observada en medias: 1.1277
Valor p: 0.0000

- Repite los pasos 1-5 R veces para obtener una distribución de la estadística de interés.

Bootstrap (validación de sesgos)

```
Validación de Sesgos

[68] # Definir n (número de elementos en cada muestra) y R (número de repeticiones)
n = len(df_clean) # Tamaño de la muestra
R = 500_000 # Número de repeticiones para el bootstrap

[69] # Almacenar las medidas estadísticas
bootstrap_means = []
bootstrap_medians = []

# Almacenar las medidas estadísticas
bootstrap_means = []
bootstrap_medians = []

# Realizar el proceso de bootstrap
for _ in range(R):
    # Paso 1: Tomar un elemento de manera aleatoria con reposición
    sample = df_clean['global_sales'].sample(n=n, replace=True)

    # Paso 2: Tomar la medida estadística
    bootstrap_means.append(sample.mean())
    bootstrap_medians.append(sample.median())

# Convertir a Series
bootstrap_means = pd.Series(bootstrap_means)
bootstrap_medians = pd.Series(bootstrap_medians)

[78] # a) Generar un Histograma
plt.figure(figsize=(14, 6))

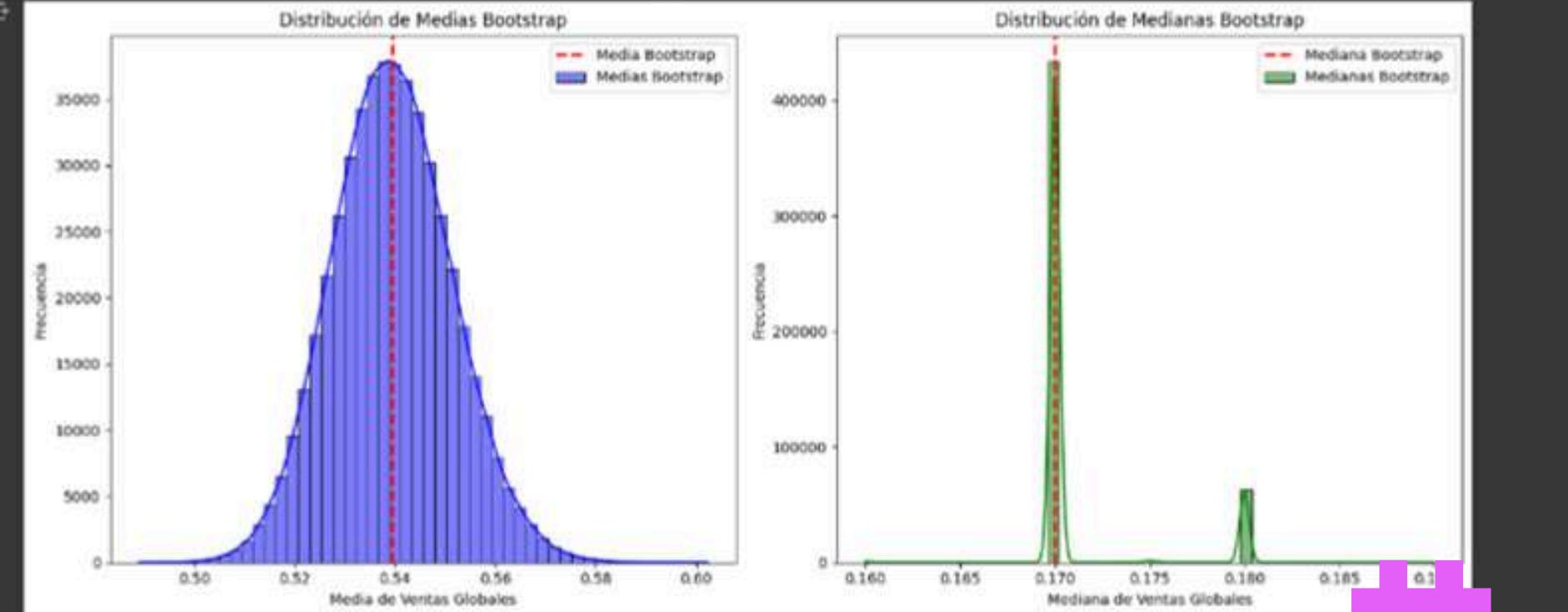
# Histograma de medias bootstrap
plt.subplot(1, 2, 1)
sns.histplot(bootstrap_means, bins=50, kde=True, color='blue', label='Medias Bootstrap')
plt.axvline(bootstrap_means.mean(), color='red', linestyle='dashed', linewidth=2, label='Media Bootstrap')
plt.title('Distribución de Medias Bootstrap')
plt.xlabel('Media de Ventas Globales')
plt.ylabel('Frecuencia')
plt.legend()

# Histograma de medianas bootstrap
plt.subplot(1, 2, 2)
sns.histplot(bootstrap_medians, bins=50, kde=True, color='green', label='Medianas Bootstrap')
plt.axvline(bootstrap_medians.median(), color='red', linestyle='dashed', linewidth=2, label='Mediana Bootstrap')
plt.title('Distribución de Medianas Bootstrap')
plt.xlabel('Mediana de Ventas Globales')
plt.ylabel('Frecuencia')
plt.legend()

plt.tight_layout()
plt.show()
```

```
plt.subplot(1, 2, 3)
sns.histplot(bootstrap_medians, bins=50, kde=True, color='green', label='Medianas Bootstrap')
plt.axvline(bootstrap_medians.median(), color='red', linestyle='dashed', linewidth=2, label='Mediana Bootstrap')
plt.title('Distribución de Medianas Bootstrap')
plt.xlabel('Mediana de Ventas Globales')
plt.ylabel('Frecuencia')
plt.legend()

plt.tight_layout()
plt.show()
```



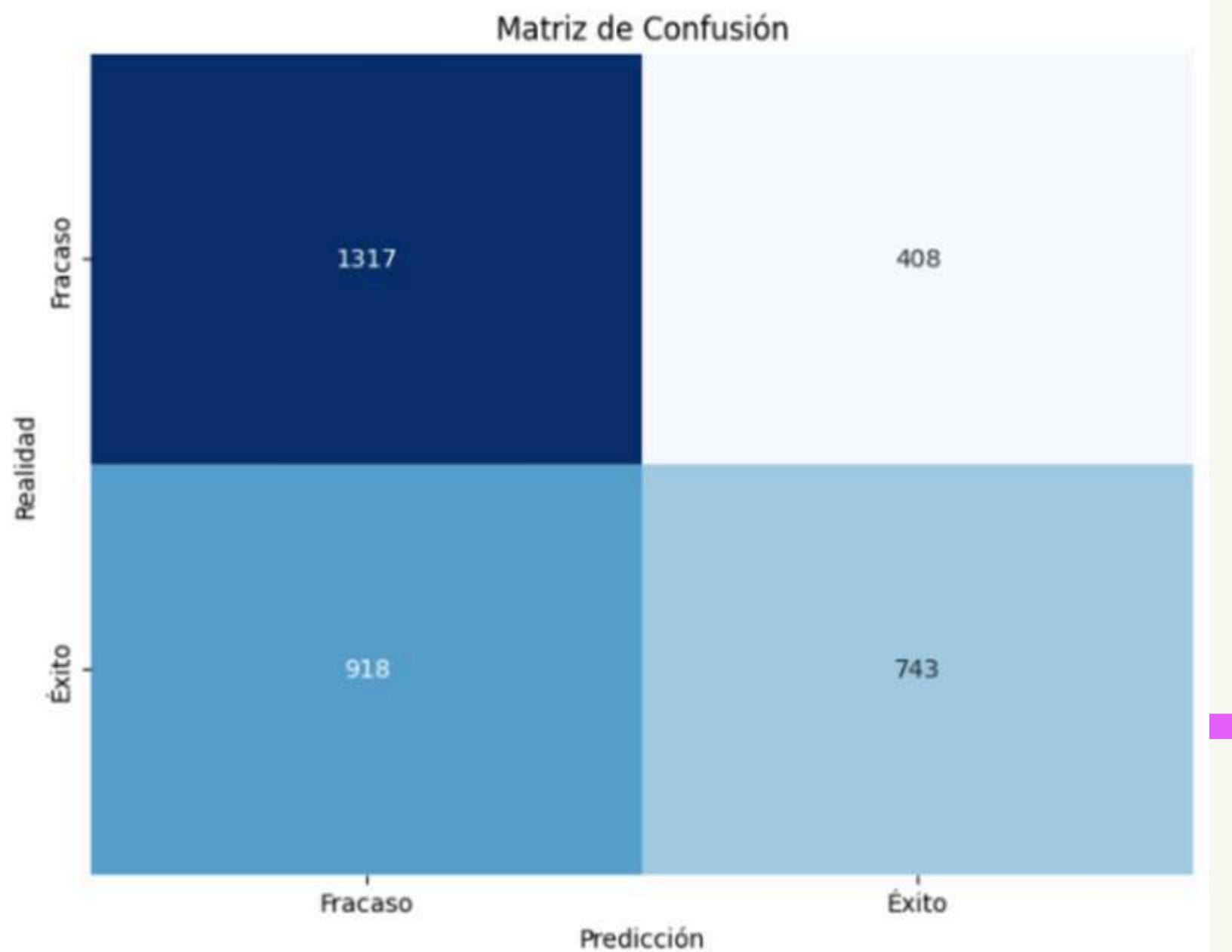
```
[79] # b) Calcular el error estándar
mean_se = bootstrap_means.std()
median_se = bootstrap_medians.std()

print(f'Error estándar de la media: {mean_se:.4f}')
print(f'Error estándar de la mediana: {median_se:.4f}')

>Error estándar de la media: 0.0119
>Error estándar de la mediana: 0.0034
```

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.



Bootstrap (validación de sesgos)

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.



1. Rendimiento de ventas por plataforma

Las plataformas de videojuegos varían en su éxito global.

PlayStation 2
Xbox 360
PlayStation 3

Refleja la preferencia por estas consolas a nivel mundial.



2. Popularidad de géneros de videojuegos

Tipo de contenido que atrae más a los jugadores.

Los géneros de videojuegos que más venden son Acción, Deportes y Shooters.

RESULTADOS

3. Distribución por género y plataforma

Concentración significativa de juegos en estas consolas.

La mayoría de los videojuegos se agrupan en géneros como Acción y Deportes, distribuyéndose principalmente en plataformas como PlayStation 2 y Nintendo DS.



4. Relación entre puntuaciones y ventas

La relación entre las puntuaciones de críticos y usuarios y las ventas es débil, lo que indica que las ventas no dependen principalmente de la crítica, sino de otros factores como el marketing o la popularidad de la plataforma.



Conclusiones

- Plataformas más vendidas: PlayStation 2, Xbox 360 y PlayStation 3 lideran en ventas globales.
- Géneros más exitosos: Acción, deportes y shooters son los géneros con mayores ventas.
- Distribución de videojuegos: PlayStation 2 y Nintendo DS tienen la mayor cantidad de títulos en varios géneros clave.
- Puntuaciones y ventas: Juegos con buenas críticas suelen vender más, pero hay excepciones donde títulos mal calificados logran ventas significativas.

MUCHAS GRACIAS

