Proyecto Final

Alumno: Francesco Centarti Maestu

Comisión: 46240

Tutor: Joaquín Armesto

Abstract

En este proyecto elegí un dataset relativo a videojuegos, dado que soy fanático desde niño y es una actividad que sostuve hasta la actualidad, siendo uno de mis principales pasatiempos y descansos de mis actividades laborales. Además me encantaría en algún momento de mi vida poder trabajar en cuestiones relacionadas a los videojuegos.

Mi meta es poder cumplimentar con las consignas y objetivos del presente curso de Data Science con un tema como el que elegí que me apasiona y obtener resultados que tal vez pueda acercarme a mi objetivo laboral de desempeñarme laboralmente con datos y videojuegos.

Este proyecto se basa datos de la industria gamer a nivel mundial, que es el sector económico involucrado en el desarrollo, la distribución, la mercadotecnia, la venta de videojuegos y del hardware asociado. Engloba a docenas de disciplinas de trabajo y emplea a miles de personas alrededor del mundo. Se estiman ingresos en 2023 en la industria del gaming de 400,000 millones de dólares, lo que superaría el Producto Interno Bruto de países como Dinamarca, Colombia, Chile o Finlandia, además superando ingresos generados por la música o el cine.

Contexto Comercial:

Trabajamos como una empresa especializada en datos que se desempeña en la industria de los videojuegos, llamada PlayData Solutions, que se encarga de colaborar con sus clientes en la toma de decisiones en el mercado.

Al contratar los servicios de esta empresa, se encargará -por ejemplo- de identificar tendencias del mercado, perfiles de jugadores y oportunidades de nicho, utilizando modelos para predecir ventas, optimizar su inventario y estrategia de lanzamiento, definir su audiencia y diseñar campañas de marketing efectivas. Todo a través de visualizaciones atractivas y fáciles de entender para comunicar los resultados de manera efectiva.

Objetivo inicial:

Obtener información que permitan recomendar o dirigir la inversión económica en un género de videojuegos que sea rentable a futuro.

Hipótesis iniciales:

- Existe una relación significativa entre el género de un juego y su éxito;
- Los géneros de los juegos podrían influir en su éxito de una manera que pueda ser predecida por características específicas de cada género;
- Existe una relación entre la frecuencia y las ventas, es decir, los géneros con más juegos son los que más se venden;
- El género con más ventas, a su vez contiene el o los juegos más exitosos;
- Ciertas características como plataforma, editora y año de lanzamiento pueden influir de manera significativa en las ventas de un juego;

Fuente:

https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat

Se descargó archivo CSV que fue renombrado como “ventasvideoj.csv” e importado desde Google Drive.

Contexto analítico:

Luego de la importación del dataset y librerías que serán utilizadas, se presenta una introducción a los datos de forma preliminar, con información básica relativa a tamaño, tipo de datos y valores faltantes.

Introducción a la base de datos

Tamaño y tipos de datos

```
[ ] df.shape

    (16719, 16)
```

Conclusión:

- Cantidad total de **16719 filas** en el dataset.
- Cantidad total **16 columnas** en el dataset.

```
[ ] df.size

    267504
```

```
[ ] df.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 16719 entries, 0 to 16718
    Data columns (total 16 columns):
     #   Column           Non-Null Count  Dtype
    ---  ------           --------------  -----
     0   Name             16717 non-null  object
     1   Platform         16719 non-null  object
     2   Year_of_Release  16450 non-null  float64
     3   Genre            16717 non-null  object
     4   Publisher        16665 non-null  object
     5   NA_Sales         16719 non-null  float64
     6   EU_Sales         16719 non-null  float64
     7   JP_Sales         16719 non-null  float64
     8   Other_Sales      16719 non-null  float64
     9   Global_Sales     16719 non-null  float64
     10  Critic_Score     8137 non-null   float64
     11  Critic_Count     8137 non-null   float64
     12  User_Score       10015 non-null  object
     13  User_Count       7590 non-null   float64
     14  Developer        10096 non-null  object
     15  Rating           9950 non-null   object
    dtypes: float64(9), object(7)
    memory usage: 2.0+ MB
```

Conclusión:

- La base de datos contiene **9 columnas numéricas** y **7 categóricas**.

A continuación podemos observar a modo de introducción:

```
[ ] df
```

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer	Rating
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	322.0	Nintendo	E
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN	NaN	NaN	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	8.3	709.0	Nintendo	E
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	192.0	Nintendo	E
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16714	Samurai Warriors: Sanada Maru	PS3	2016.0	Action	Tecmo Koei	0.00	0.00	0.01	0.00	0.01	NaN	NaN	NaN	NaN	NaN	NaN
16715	LMA Manager 2007	X360	2006.0	Sports	Codemasters	0.00	0.01	0.00	0.00	0.01	NaN	NaN	NaN	NaN	NaN	NaN
16716	Haitaka no Psychedelica	PSV	2016.0	Adventure	Idea Factory	0.00	0.00	0.01	0.00	0.01	NaN	NaN	NaN	NaN	NaN	NaN
16717	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01	NaN	NaN	NaN	NaN	NaN	NaN
16718	Winning Post 8 2016	PSV	2016.0	Simulation	Tecmo Koei	0.00	0.00	0.01	0.00	0.01	NaN	NaN	NaN	NaN	NaN	NaN

16719 rows × 16 columns

Descripción de columnas y variables disponibles inicialmente en el dataset

Name: Nombre del videojuego.

Platform: La plataforma o consola en la que se lanzó el juego (por ejemplo, PlayStation, Xbox, PC, etc.).

Year_of_Release: El año en que se lanzó el juego.

Genre: El género del juego (por ejemplo, acción, aventura, deportes, etc.).

Publisher: La empresa editora del juego.

NA_Sales: Las ventas totales del juego en América del Norte.

EU_Sales: Las ventas totales del juego en Europa.

JP_Sales: Las ventas totales del juego en Japón.

Other_Sales: Las ventas totales del juego en otras regiones.

Global_Sales: Las ventas totales globales del juego. Esta es la variable objetivo que podríamos utilizar en análisis y modelado predictivo.

Critic_Score: La puntuación promedio otorgada por críticos de videojuegos.

Critic_Count: El número de críticas de videojuegos utilizadas para calcular la puntuación promedio de los críticos.

User_Score: La puntuación promedio dada por usuarios de videojuegos.

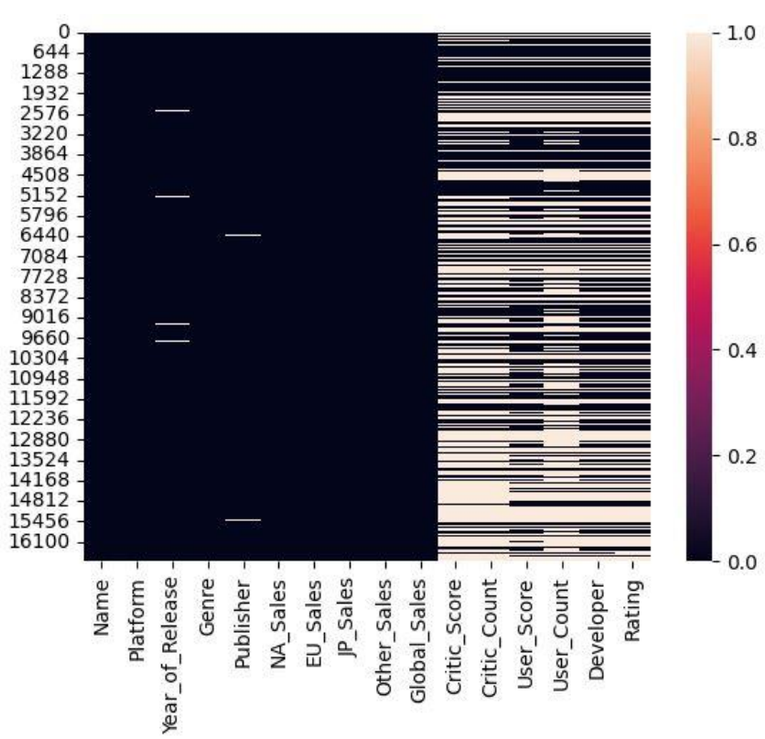
User_Count: El número de usuarios que calificaron el juego y se utilizó para calcular la puntuación promedio de usuarios.

Developer: El nombre de la empresa o estudio de desarrollo del juego.

Rating: La clasificación por edad del juego (por ejemplo, "E" para todos, "M" para maduros, etc.).

Sales (todas las columnas) = En millones de unidades

Respecto a los valores faltantes, esta es la información que obtenemos del mapa de calor:



Se realiza trabajo de data wrangling para las columnas relevantes, eliminando duplicados y clasificado las ventas en diferentes categorías basadas en cuartiles. Esto proporciona una base sólida y limpia para continuar con el análisis y la modelización del proyecto.

Data Wrangling

Introducción:

- Tenemos algunos valores faltantes en el dataset, en las columnas Name (2),Year_of_Release (269), Genre (2) Publisher (54), Critic_Score (8582), Critic_Count (8582), User_Score (6704), User_Count (9129) Developer (6623), Rating (6769)
- En primer lugar, en los casos como Name y Genre con valores nulos muy bajo (2), eliminamos las filas con nulos en estas columnas.
- Para el año de lanzamiento imputamos los valores nulos en la columna con la mediana y con la moda en la columna Publisher.
- Para el caso de las columnas Critic_Score, Critic_Count, User_Score, User_Count, Developer y Rating, dada la gran cantidad de valores nulos y que considero que no son esenciales para mi análisis opto por eliminarlas.

Tratamiento

[ ] *4 celdas ocultas*

A continuación analizamos si contamos con valores duplicados en nuestro dataset

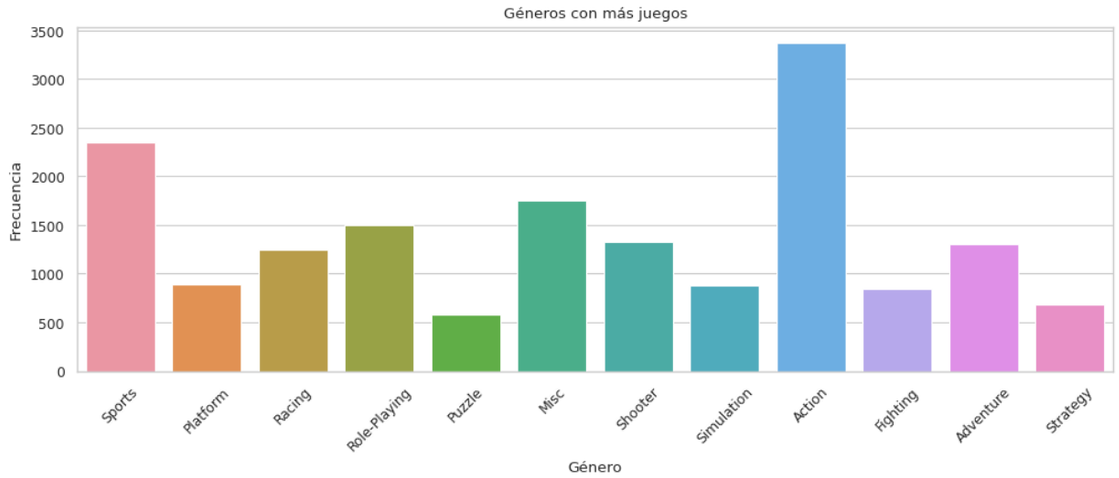
[ ] *5 celdas ocultas*

Detección de Colinealidad

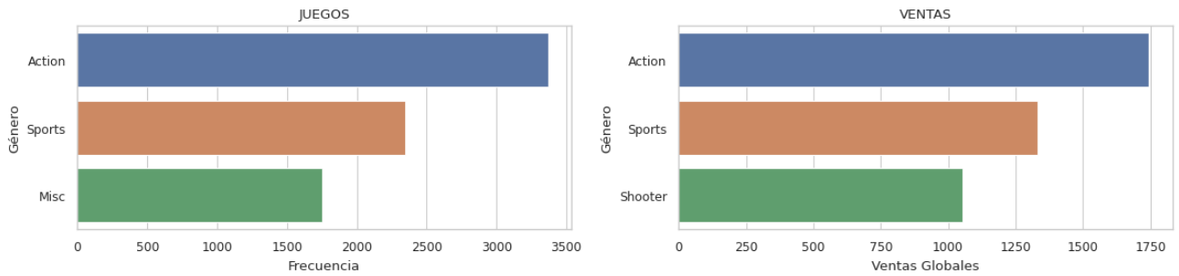
[ ] *8 celdas ocultas*

Análisis exploratorio de datos (EDA), centrado y relacionado al objetivo del trabajo relativo a intentar predecir en qué género de videojuegos invertir, mediante análisis bivariados y multivariados.

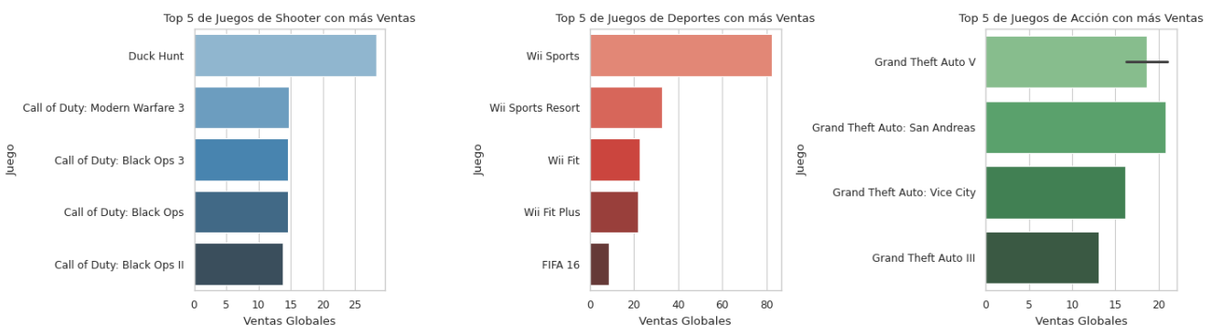
¿Cuáles son los géneros más populares?



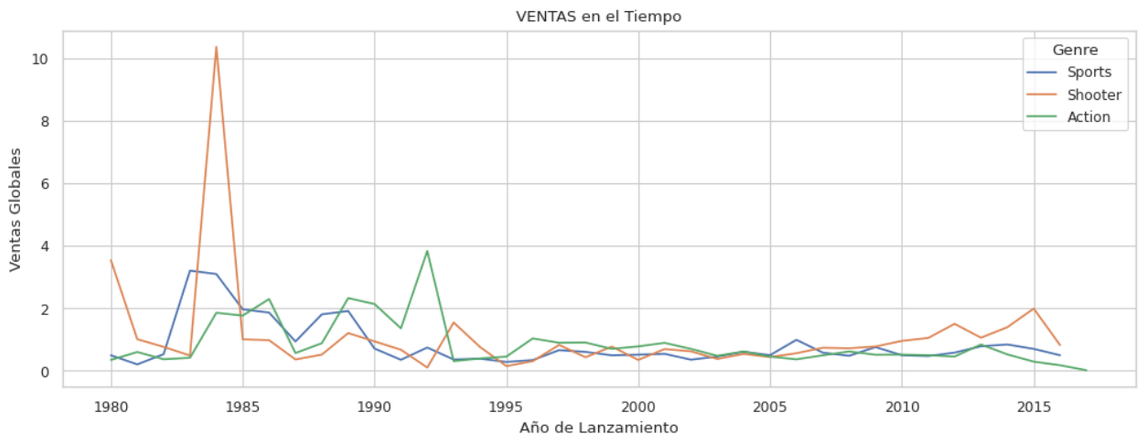
¿Son los géneros con más juegos los que más se venden?



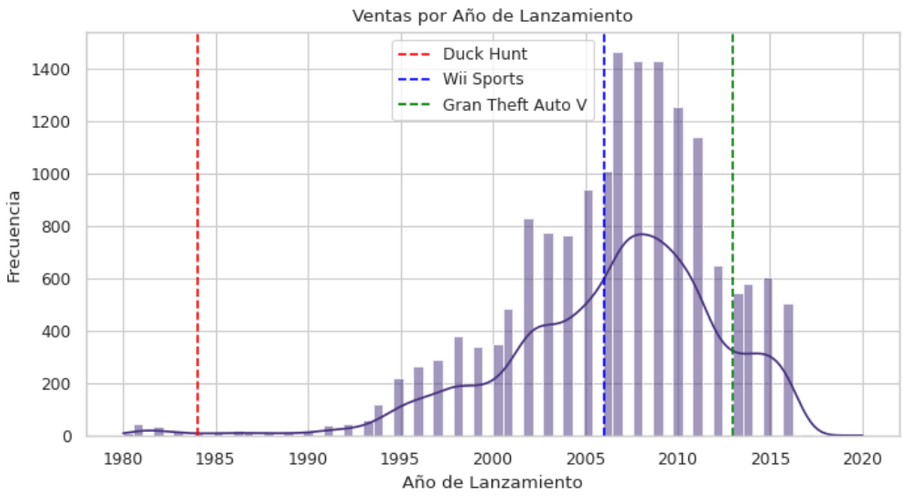
¿Cuáles son los juegos más exitosos?



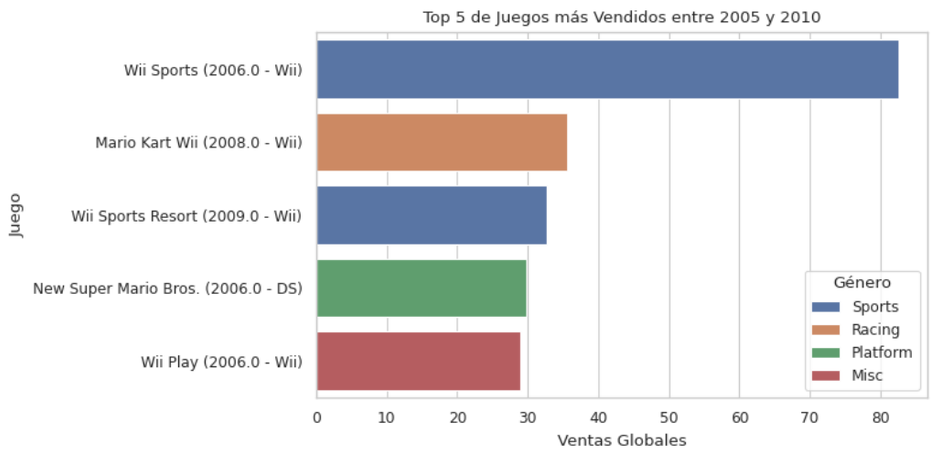
¿Cómo fueron las tendencias de ventas en el tiempo de los 3 géneros más vendidos?



¿Cómo han sido las ventas globales a lo largo del tiempo? ¿han sido un hito en la historia los lanzamientos de estos 3 juegos?



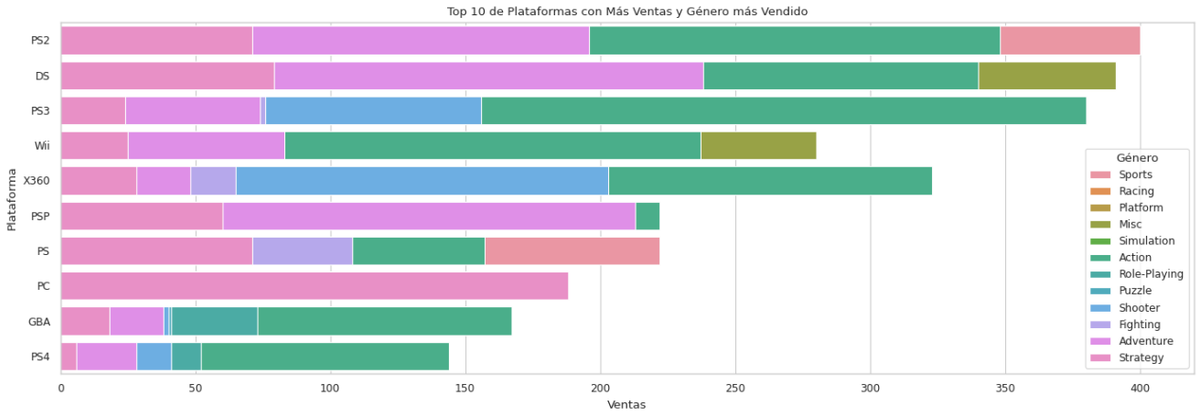
¿Qué sucede en ese pico desde 2005 a 2010?



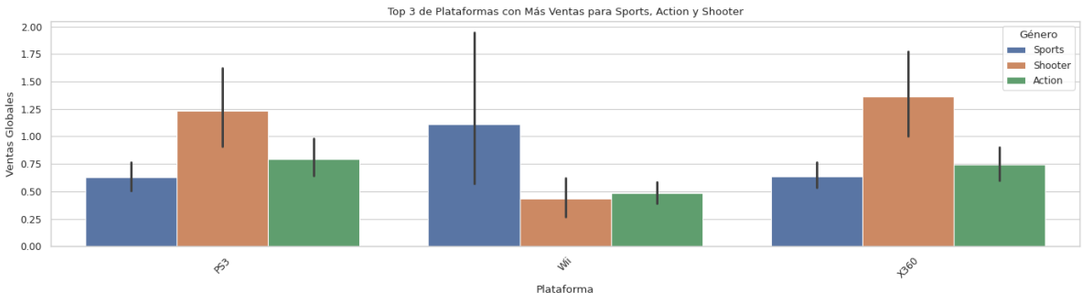
** En el período observado de mayor cantidad de ventas a nivel global, también vemos la influencia de los dos géneros más vendidos. Nintendo Wii se ubica, en el rango de tiempo observado que va desde 2005 a 2010 como una de las plataformas con más ventas, siendo además el título Wii Sports el juego más vendido.*

"El 23 de septiembre de 2009, Nintendo anunció su primer recorte de precio de la consola. En Japón, el precio disminuyó de 25 000 a 20 000 JPY a partir del 1 de octubre de ese año. En los Estados Unidos a partir del 27 de septiembre su precio se redujo 50 USD, al pasar de 249,99 a 199 USD. Finalmente en Europa, el precio de Wii bajó de 249 a 199 EUR. Tras el recorte en los precios, en diciembre de 2009, Nintendo vendió más de 3 000 000 consolas en Estados Unidos, mucho más que las 2 140 000 de unidades en el año anterior. Además estableció un récord mensual en dicho país y acabó con una racha de 9 meses de ventas declinantes. Este éxito también se le atribuyó al lanzamiento de juegos como New Super Mario Bros Wii. Al término de ese mes, pasó a ser la consola casera con mayores ventas producida por Nintendo en toda su historia...". FUENTE: Wikipedia https://es.wikipedia.org/wiki/Wii

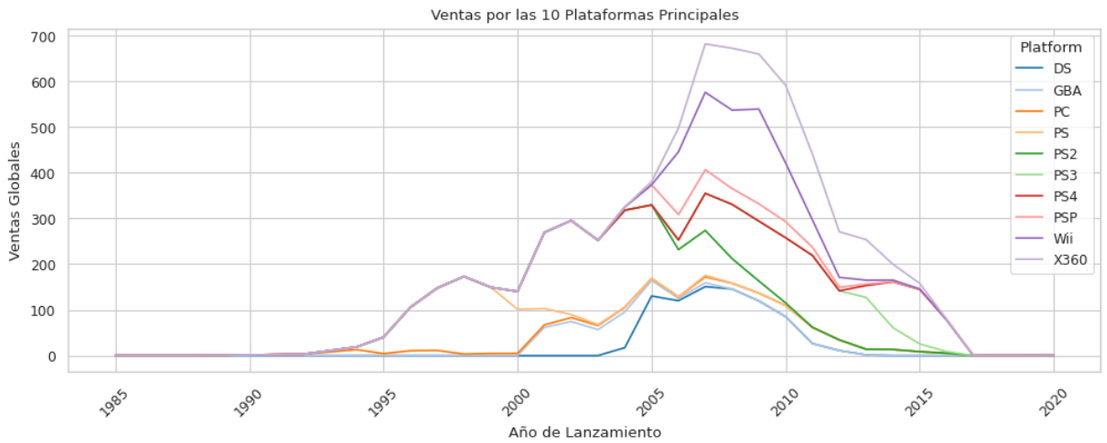
¿Qué géneros son más vendidos por las principales plataformas?



¿Cuáles son las plataformas con más ventas de los géneros más buscados por el público?



¿Cómo han sido las ventas por género en el tiempo?



Seguidamente, nos centramos en dos modelos predictivos, un modelo de clasificación para predecir el éxito de un juego basado en el género, y un modelo de regresión para predecir las ventas globales de un juego en función de su género, plataforma, editora y año de lanzamiento.

Para esto fueron preparados datos, codificadas las variables, divididos los datos en conjuntos de entrenamiento y prueba, seleccionadas las características relevantes, para luego construir y evaluar los modelos.

A lo largo del curso, fuimos incorporando nuevos conceptos y conocimientos tanto teóricos como prácticos, que me permitieron intentar perfeccionar mi proyecto, desde observaciones de resultados y análisis iniciales hasta trabajos de clustering y validación cruzada de modelos como puede observarse en el colab.

Modelo de Clasificación

Objetivo: predecir el éxito de un juego basado en el género.

Descripción Inicial y definición de éxito: Comenzamos con estadísticas descriptivas de las ventas globales para entender la distribución de los datos.

```
[ ] # Mostrar estadísticas descriptivas de las ventas globales
    sales_stats = df['Global_Sales'].describe()
    print(sales_stats)
```

```
count    16713.000000
mean         0.533481
std          1.548122
min          0.010000
25%          0.060000
50%          0.170000
75%          0.470000
max         82.530000
Name: Global_Sales, dtype: float64
```

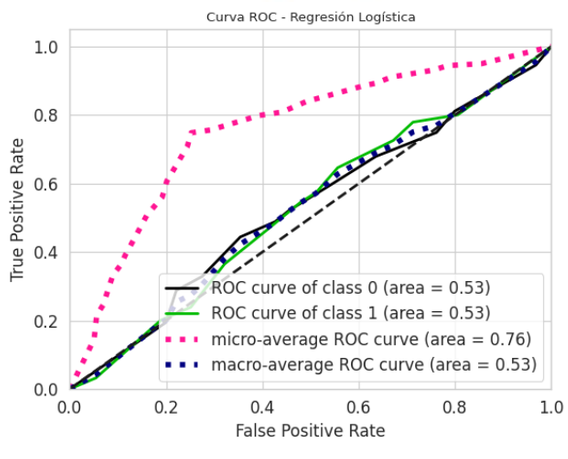
** hay una variación significativa en las ventas globales, con una gran cantidad de juegos que tienen ventas relativamente bajas, pero algunos juegos han alcanzado ventas muy altas*

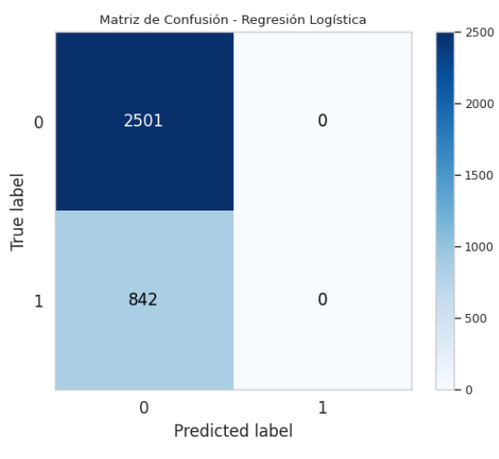
Se crea una nueva columna llamada 'Exito' basada en un umbral definido en el tercer cuartil de las ventas globales para clasificar juegos como exitosos.

Modelo de Regresión Logística: Entrenamiento de un modelo de Regresión Logística utilizando la variable categórica 'Genre' como característica. Se obtiene una precisión del modelo del 75%.

```
# Calcular la precisión del modelo
accuracy = accuracy_score(y_test, y_pred)
print(f"Precisión del modelo de Regresión Logística: {accuracy:.2f}")
```

```
Precisión del modelo de Regresión Logística: 0.75
```





Modelo de Random Forest: Se obteniene una precisión similar al modelo de Regresión Logística. Sin embargo, la matriz de confusión muestra que el modelo tiene dificultades para predecir la clase positiva (éxito), ya que clasifica todas las instancias en la clase negativa.

```
# Imprimir métricas de evaluación
print(f"Precisión del modelo de Random Forest: {accuracy}")
print("Matriz de Confusión:")
print(conf_matrix)
```

```
Precisión del modelo de Random Forest: 0.7481304217768472
Matriz de Confusión:
[[2501    0]
 [ 842    0]]
```

** el modelo acertó en clasificar 2501 instancias como negativas (clase 0, juegos no exitosos), pero tuvo dificultades en clasificar instancias como positivas (clase 1, juegos exitosos), ya que todas las 842 instancias fueron clasificadas incorrectamente. Esto sugiere que el modelo tiene dificultades para clasificar la clase positiva*

Modelo SVM Lineal: Con resultados similares a los modelos anteriores pero clasifica todas las instancias en una sola clase, lo que indica problemas con el desequilibrio de clases (clase 0 teniendo muchas más muestras que la clase 1).

```
[ ]  clase_count = pd.Series(y).value_counts()

     print(clase_count)
```

```
0    12510
1     4203
Name: Exito, dtype: int64
```

Regresión Logística para desequilibrio de Clases: La matriz de confusión y el reporte de clasificación revelan que, aunque la precisión global disminuye, el modelo ahora es capaz de predecir mejor la clase 1 (éxito).

```
Matriz de Confusión:
[[1222 1279]
 [ 359  483]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.77      0.49      0.60      2501
           1       0.27      0.57      0.37       842

    accuracy                           0.51      3343
   macro avg       0.52      0.53      0.48      3343
weighted avg       0.65      0.51      0.54      3343
```

** La precisión del modelo es buena para la clase 0 pero baja para la clase 1. El recall, en cambio, es alto para la clase 1 pero bajo para la clase 0. Aunque la precisión global disminuye, el modelo ahora es capaz de predecir mejor la clase 1 (éxito)*

Previo a la validación cruzada, sería conveniente la inclusión de características relevantes para mejorar la capacidad predictiva de los modelos.

Validación Cruzada: Los resultados muestran que los modelos se comportan de manera similar en diferentes conjuntos de datos de prueba. Se aplica la técnica de SMOTE para abordar el desequilibrio de clases para luego entrenar un modelo de Random Forest con los datos balanceados. Utilizando GridSearchCV se exploran diferentes combinaciones de hiperparámetros para el modelo de Random Forest. Luego, se evalúa el mejor modelo en el conjunto de prueba.

```
Mejores hiperparámetros: {'max_depth': None, 'n_estimators': 300}
Precisión del modelo en el conjunto de prueba: 0.9979060723900688
Recall del modelo en el conjunto de prueba: 0.9916864608076009
Precisión del modelo en el conjunto de prueba: 1.0
F1-score del modelo en el conjunto de prueba: 0.9958258795468097
AUC-ROC del modelo en el conjunto de prueba: 0.9958432304038005
Matriz de Confusión:
[[2501    0]
 [   7  835]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2501
           1       1.00      0.99      1.00       842

    accuracy                           1.00      3343
   macro avg       1.00      1.00      1.00      3343
weighted avg       1.00      1.00      1.00      3343
```

** Estos resultados indican un rendimiento excepcional del modelo de clasificación en la predicción de instancias positivas y negativas, con una precisión general del 99.79%. La capacidad para identificar instancias positivas (recall) también es muy alta (99.17%)*

Modelo de Regresión

Objetivo: desarrollar un modelo de regresión capaz de predecir las ventas globales de videojuegos en función de variables como género, plataforma, editora y año de lanzamiento.

Modelo de regresión lineal: mostró un rendimiento insatisfactorio, con un R^2 muy bajo, indicando una baja capacidad para explicar la variabilidad de los datos.

```
# Evaluar el modelo
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Error Cuadrático Medio (MSE): {mse}")
print(f"Coeficiente de Determinación (R²): {r2}")
```

```
Error Cuadrático Medio (MSE): 4.123720329664874
Coeficiente de Determinación (R²): 0.0017536633956982683
```

One-Hot Encoding: La aplicación a variables categóricas resultó en un rendimiento deficiente, con un MSE extremadamente alto y un R^2 negativo, sugiriendo una inadecuada adaptación del modelo a la codificación.

```
print(f"Error Cuadrático Medio (MSE): {mse}")
print(f"Coeficiente de Determinación (R²): {r2}")
```

```
Error Cuadrático Medio (MSE): 3.5949030782433972e+16
Coeficiente de Determinación (R²): -1.74274655322022e+16
```

Modelo de Bosque Aleatorio: mostró un rendimiento mejor que la regresión lineal, pero aún se percibió como insuficiente para una predicción precisa de las ventas globales.

```
print(f"Error Cuadrático Medio (MSE) - Bosque Aleatorio: {mse_rf}")
print(f"Coeficiente de Determinación (R²) - Bosque Aleatorio: {r2_rf}")
```

```
Error Cuadrático Medio (MSE) - Bosque Aleatorio: 1.8275104249850749
Coeficiente de Determinación (R²) - Bosque Aleatorio: 0.11405469777699451
```

Optimización de Hiperparámetros para Random Forest: resultó en una mejora significativa en la puntuación del modelo.

```
# Definir el modelo con los mejores parámetros encontrados
best_rf_model = RandomForestRegressor(
    max_depth=None,
    max_features='sqrt',
    min_samples_leaf=1,
    min_samples_split=5,
    n_estimators=200
)


X = df[['Genre', 'Platform', 'Publisher', 'Year_of_Release']]  # Features
y = df['Global_Sales']  # Target

best_rf_model.fit(X, y)
```

```
RandomForestRegressor(max_features='sqrt', min_samples_split=5,
                      n_estimators=200)
```

Sin embargo, el rendimiento aún no alcanzó niveles altos, sugiriendo la posibilidad de que las características actuales no sean suficientes para predecir con precisión las ventas globales.

```
print(f"Error Cuadrático Medio (MSE): {mse}")
print(f"Coeficiente de Determinación (R²): {r2}")
```

```
Error Cuadrático Medio (MSE): 3.8737545430293343
Coeficiente de Determinación (R²): 0.06226393345215919
```

Conclusiones finales y elección de modelo

La precisión del 75%, tanto la Regresión Logística como el Bosque Aleatorio en el modelo de clasificación ha demostrado que son efectivos para clasificar juegos según su éxito.

En contraste, el Modelo de Regresión, a pesar de mejoras mediante la optimización de hiperparámetros, no alcanzó métricas de rendimiento satisfactorias.

La elección del modelo debería alinearse en primer lugar con los objetivos del proyecto, por lo que en un primer momento me hubiera inclinado por el modelo de regresión, que de hecho a lo largo del proyecto fue el primero en el que trabajé.

Sin embargo, me gustó el enfoque que pudo buscarse con el modelo de clasificación y quedé mucho más conforme con sus resultados donde para clasificar juegos según su éxito ha alcanzado una precisión alta.

En el caso del modelo de regresión, me resulta muy interesante el enfoque de predecir las ventas globales de juegos pero el rendimiento debería mejorarse. Considero que sería interesante y de mucha utilidad que en cualquiera de los dos casos se puedan explorar y sumar nuevas características para mejorar la capacidad predictiva de los modelos.