**MEMORIA**

---

Felipe Alvim
Data Science Bootcamp - The Bridge Madrid
Análisis EDA de ventas de videojuegos (1980-2016)

---

## Introducción

Este conjunto de datos ofrece información sobre las ventas de videojuegos según plataforma, editora y género para los títulos más populares a nivel mundial.
Con más de 16,000 registros de ventas entre 1980 y 2016, este análisis busca descubrir relaciones clave entre estas variables y proporcionar insights sobre las tendencias de una industria que mueve millones de dólares cada año.

### Objetivo

Este estudio tiene como objetivo analizar cuáles plataformas lideran las ventas globales y por región, identificar los géneros que predominan en las preferencias de los jugadores, y observar cómo estos factores han cambiado a lo largo del tiempo.
Este análisis ayudará a comprender la evolución de la industria de los videojuegos y a identificar tendencias clave que podrían ayudar a los desarrolladores y editores a alcanzar el éxito en el futuro.

### Hipótesis

1. ¿Cuáles son los juegos, géneros, plataformas y editoriales más populares en términos de cantidad de publicaciones?
2. Correlación de datos
3. ¿Cuáles son las cifras de ventas por región y a nivel mundial de los videojuegos?
4. ¿Cómo se distribuye el consumo de videojuegos a nivel mundial? ¿Qué regiones

representan el mayor porcentaje de las ventas globales y por qué?
5. ¿Cuáles son los videojuegos más vendidos por región y a nivel mundial?
6. ¿Cuáles son las plataformas de videojuegos más populares en términos de ventas por región y a nivel mundial?
7. ¿Cuáles son los géneros de videojuegos más vendidos por región y a nivel mundial?
8. ¿Cuáles son las editoriales de videojuegos con mayor volumen de ventas por región y a nivel mundial?
9. ¿Cómo ha evolucionado la popularidad de los diferentes géneros de videojuegos a lo largo de los años?
10. ¿Existe alguna correlación entre las preferencias de los jugadores en diferentes regiones y los factores demográficos, culturales o económicos?
11. ¿Cómo se comparan las ventas de videojuegos por género, plataforma y editorial en diferentes regiones?

Librerías utilizadas:

La siguiente lista contiene todas las librerías utilizadas para realizar este análisis:

- Pandas
- Numpy
- Matplotlib.pyplot
- Scipy.stats
- Seaborn
- Os
- Sys
- IPython.display

Dataset objeto de estudio:

El conjunto de datos utilizado para este estudio se descargó del sitio web Kaggle (https://www.kaggle.com/) en formato csv.

**Análisis de los datos**

Analisis inicial

Justo después de cargar los datos, se empieza con una primera exploración obtenida a través de las funciones:
info()
describe()
head()
tail()
shape()
isna().sum()

Después de ejecutar dichas funciones se observan los siguientes datos.

Son 10 columnas al total, de las cuales son divididas del siguiente modo:

- 6 columnas numéricas
    - Year - El ano en cuestión en formato 'yyyy'
    - NA_Sales - Ventas en la región de Norteamérica (en millones USD)
    - EU_Sales - Ventas en la región de Europa (en millones USD)
    - JP_Sales - Ventas en la región de Japón (en millones USD)
    - Other_Sales - Ventas acumuladas de regiones no seleccionadas previamente (en millones USD)
    - Global_Sales - Ventas acumuladas de todo el mundo (en millones USD)

- 4 Columnas categóricas
    - Name - Título del videojuego
    - Platform - Título de la consola
    - Genre - Título de género del juego
    - Publisher - Título de la empresa editorial que lanza el videojuego

Se identificó la presencia de valores nulos en el conjunto de datos que requerían ser tratados. En concreto, se observaron los siguientes valores faltantes:
Columna Year: 271 (1.63% del total de los datos)
Columna Publisher 58 (0.34% del total de los datos)

Dado el bajo porcentaje de valores nulos, se optó por completar el conjunto de datos de manera precisa, evitando eliminar registros o insertar valores predichos basados en los datos existentes. Para ello, los valores faltantes se obtuvieron mediante investigaciones en sitios especializados como VGChartz, donde se encuentran datos completos sobre videojuegos. Posteriormente, estos valores se incorporaron al conjunto de datos utilizando diccionarios que contenían la información faltante y un bucle for para realizar las actualizaciones, como se ilustra en el siguiente fragmento de código:

```
for n in df.index:
    if n in game_publisher.keys():
        df.at[n, 'Publisher'] = game_publisher[n]
```

En el código de arriba itera entre todos los índices del dataset (df), se el valor ‘n’ del índice está presente en las claves del diccionario creado, entonces se cambia el valor nulo del dataset por el valor relativo a dicha clave del diccionario, insertando el valor correcto.

A través de la función duplicated() se encuentra un dato duplicado en el dataset, que fué eliminado utilizando la siguiente función:

```
df[df.duplicated(keep=False)]
```

Tal función establece el parámetro keep en False para establecer como verdadera la primera aparición en True, manteniéndolo y eliminando el dato duplicado.

Luego se salva el archivo con un nuevo nombre, manteniendo el archivo original por si fuera necesario realizar alguna consulta posterior, a partir de este momento se empieza

el análisis descriptivo utilizando el archivo “limpio”.

Análisis exploratorio de los datos:

1. Análisis Univariado:

    - Se emplearon gráficos de conteo (countplots) para analizar la distribución de las variables categóricas, proporcionando una visión general de las frecuencias de cada categoría.

2. Análisis Bivariado:

    - Se exploraron las relaciones entre las variables numéricas y categóricas mediante mapas de calor de correlación (heatmaps).
    - Se examinaron las relaciones entre las variables categóricas utilizando diversos gráficos de barras y comparaciones.

3. Visualizaciones e Insights:

    - Se crearon clasificaciones de los 10 principales en categorías clave utilizando gráficos de barras, incluyendo:
        - Ventas por Región: Se visualizaron las 10 principales ventas por región, el acumulado global y comparaciones entre regiones.
        - Ventas por Juego: Se destacaron los 10 juegos más vendidos a nivel global, segmentados por región mediante gráficos de barras.
        - Ventas por Plataforma: Se mostraron las plataformas con mejor desempeño por región, globalmente y a lo largo de diferentes años mediante gráficos de barras.
        - Ventas por Género: Se ilustraron las principales ventas por género en cada región, por década y como tendencias a lo largo del tiempo utilizando gráficos de barras y líneas.
        - Ventas por Editorial: Se presentaron los principales publicadores

por región mediante gráficos de barras y un mapa de árbol (treemap) para resumir el dominio global.

- Se analizaron las tendencias de ventas globales a lo largo del tiempo, identificando el año con las mayores ventas acumuladas mediante gráficos de líneas y la función idxmax para determinar el año con más ventas en toda la serie temporal.
- Se generaron gráficos comparativos para evaluar las ventas por géneros, plataformas y publicadores en todas las regiones, todo dentro de un único gráfico para facilitar la comparación.
- Se creó un gráfico circular para ilustrar la contribución relativa de cada región a las ventas globales.

Bibliografía:

https://www.kaggle.com/

https://www.vgchartz.com/