EDA DE VIDEOJUEGOS

AUTOR: Ángel Cebriá Morales (AngelElPet)

ÍNDICE

INTRODUCCIÓN

- ¿En qué región suele estrenarse los videojuegos antes, en Norte América o en Europa?

A lo largo de esta memoria se intentará resolver a esta pregunta, bajo una hipótesis inicial de que se estrenan más juegos en Norte América que en Europa. Para ello se han buscado tres archivos de tipo Json que tuviesen información relevante sobre los videojuegos estrenados en las plataformas: Xbox, PlayStation y Switch.

1. Esta es la url de PlayStation 'https://api.sampleapis.com/playstation/games'
2. Esta es la url de Xbox 'https://api.sampleapis.com/xbox/games'
3. Esta es la url de Switch 'https://api.sampleapis.com/switch/games'

Una vez descargados y creados los diferentes Json y pasados a DataFrames, estamos listos para empezar el trabajo

Desarrollo de datos

Partiendo de las tres tablas de datos, se puede observar que en la columna de genre, muchos juegos no tienen género, en cambio, tienen escrito una lista vacía.

Por otro lado, cada una de las columnas de las fechas de salida, rdEurope, rdNorthAmerica, rdJapan, rdAustralia, me ha parecido conveniente separar cada una en tres columnas, una para almacenar el dato del día, otra para la del mes y otra para la del año. Además, también conviene revisar sus datos, ya que no todos los datos están escritos en modo fecha, sino que se puede encontrar texto escrito como:

- *Unrealised* que significa que no ha salido
- *TBA*
- *Early acces*
- *Q3 2020*
- *Assorted*
- *Error in Template:Date table sorting: 'Dex' is not a valid month*

En la misma función que genera las columnas anteriores, se comparan las fechas y genera la columna Comp en cada DataFrame que especificará las regiones en las que se lanzó el videojuego.

Añadiremos una nueva columna en cada DataFrame llamada 'plataforma' y en sus observaciones pondremos: playstation, xbox y switch.

Por último, generaremos un DataFrame común concatenando las 3 tablas anteriores y modificaremos la columna de 'id' para que todas sus observaciones sean 1, para poder utilizarlos después.

En conclusión, seguiremos los siguientes pasos:

1. En la columna genre cambiaremos los datos que veamos como una lista vacía por la lista ['Sin Genero']
2. En la columna genre hay datos que vienen con listas de más de un género, por lo que solo nos interesará el primer género.
3. En las columnas, rdEurope, rdNorthAmerica, rdJapan, rdAustralia, cuyos datos ponga *Unrealised*, lo cambiaremos por una fecha del futuro, para después en la lectura de datos podamos aprovecharlos. En este caso pondremos como que saldrán en el año 3000 mes 0 y día 0, cosa que nos facilitará diferenciarlos de las demás observaciones.
4. En las columnas, rdEurope, rdNorthAmerica, rdJapan, rdAustralia, cambiaremos por un numpy.NaN aquellos datos ponga :*TBA, Early acces, Q3 2020, Assorted, Error in Template:Date table sorting: 'Dex' is not a valid month*. Finalmente haremos un `dropna(inplace=True)` en los 3 dataframes.
5. De las columnas, rdEurope, rdNorthAmerica, rdJapan, rdAustralia, obtendremos columnas que guarden el día, el mes y el año de cada una de ellas y una última columna que haya comparado las fechas
6. Creamos la columna 'plataforma' en la que pondrá: playstation, xbox, switch según el DataFrame en el que nos encontremos
7. Concatenaremos los 3 DataFrame anteriores en uno solo y modificaremos la columna 'id' para que sus datos sean 1.
8. Guardar en la carpeta el DataFrame final como el archivo Videojuegos.csv

Desarrollo de gráficos

En el proyecto se va ha utilizar Matplotlib, Seaborn, Plotly y Streamilit, para generar las gráficas.

Se han generado las siguientes graficas:

- Con Seaborn se ha realizado la gráfica de seaborn.countplot de las plataformas y además otro de los géneros principales. En la segunda gráfica solo aparecen aquellos géneros que tengan más de 30 juegos registrados con ellos
- Con Matplotlib, se generan las tablas que clasifican la cantidad de juegos que han salido en cada mes sin tener en cuenta el año y diferenciando entre las regiones. A la función le hemos pedido que se guarde cada gráfica en una imagen, para poder utilizarla después en Streamlit.
- Con Plotly, generamos 2 tipos de gráficas: una es de tipo Pie para ver los datos de la cantidad de juegos que se han estrenado por regiones y otras 4 de barras que muestran la cantidad de juegos que se han estrenado en los diferentes años que hay registros, filtrado por plataformas y regiones.
- Streamlit se ha utilizado para generar 4 gráficas donde se filtran por días la cantidad de juegos que han salido en las diferentes regiones.

Desarrollo de Streamlit

He desarrollado una streamlit.app (https://angelelpet-eda-videojuegos-mainst-zg10vy.streamlit.app/) a modo de presentación del proyecto.

He creado un menú con las siguientes opciones:

1. 'Página principal'

En esta pantalla se podrá ver una explicación del proyecto, algunas imágenes y un desplegable donde estarán los url de donde he obtenido los datos iniciales

2. 'Cargar datos'

En esta pantalla se tendrá que ingresar el archivo src/data/Videojuegos.csv. Después muestra el DataFrame y una gráfica de la cantidad de videojuegos que hay de cada plataforma

3. 'Regiones por Meses'

Muestra 4 desplegables, un por cada región donde se muestra una imagen de la gráfica con la información de la cantidad de juegos estrenados por meses, obtenida del archivo main.py. Cabe resaltar que el mes = 0 mide la cantidad de juegos que no han salido en cada región

4. 'Regiones por Días'

Muestra 4 desplegables, un por cada región donde se muestra una gráfica con la información de la cantidad de juegos estrenados por días. Cabe resaltar que el mes = 0 mide la cantidad de juegos que no han salido en cada región

5. 'Géneros principales'

Muestra una gráfica con aquellos géneros que tienen más de 30 videojuegos

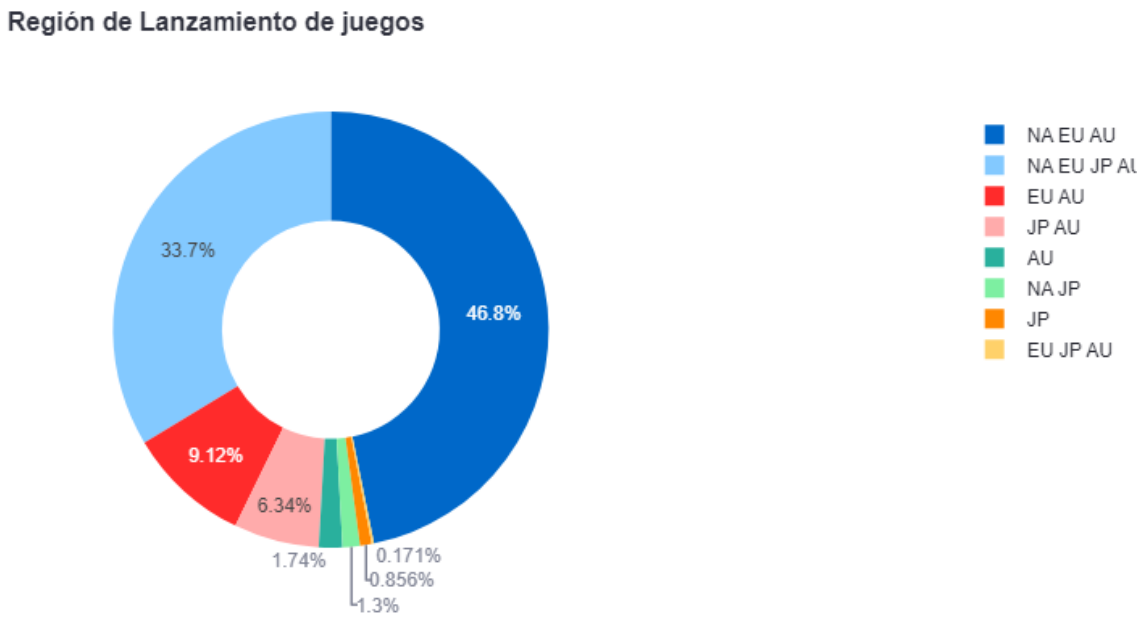
6. 'Conclusión'

Muestra 4 desplegables, uno para cada región, con graficas de barras que muestran la cantidad de juegos que se han estrenado en los diferentes años que hay registros, filtrado por plataformas y regiones.

Además hay una gráfica tipo Pie con los datos de las fechas comparadas que indica la cantidad de juegos que ha salido por regiones y tablas que muestran los resultados filtrados escritas con SQL.

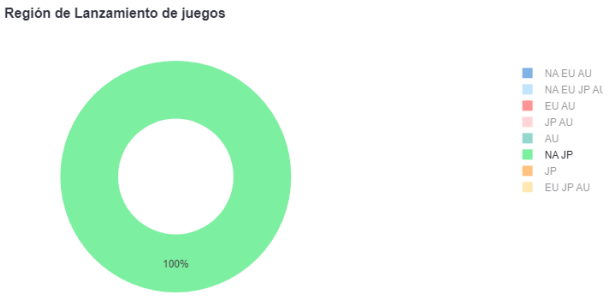
Conclusión

Aquí podemos observar la cantidad de juegos que se estrenaron agrupadas por regiones



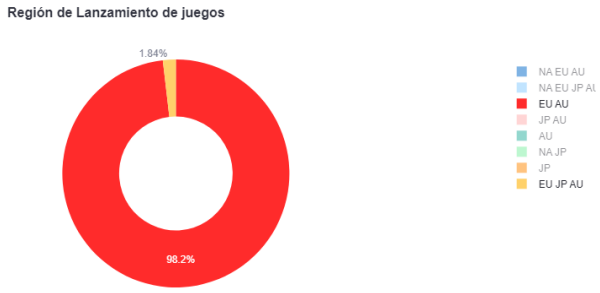
Se puede apreciar fácilmente que solo 53 juegos se estrenaron en Norte América antes que en Europa.

	Comp	Regiones
0	53	NA JP



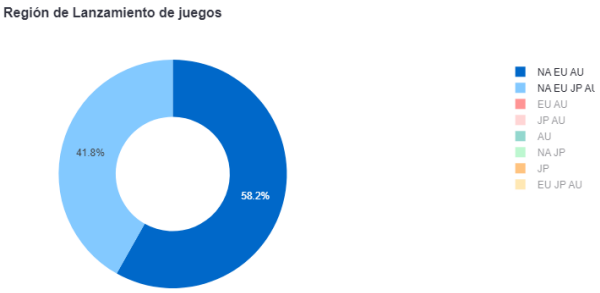
Mientras que un total de 380 videojuegos salieron antes en Europa que en Norte América,

	Comp	Regiones
0	373	EU AU
1	7	EU JP AU



Y como dato adicional, un total de 3200 juegos fueron lanzados a la vez en ambas regiones.

	Comp	Regiones
0	1,914	NA EU AU
1	1,376	NA EU JP AU



Así pues, he demostrado que mi hipótesis inicial era errónea y a pesar de que en la región de Europa hayan salido más videojuegos, realmente destaca mucho más el dato de que en ambas regiones se han estrenado a la vez un total de 3200 videojuegos.

Bibliografía

1. Esta es la url de PlayStation 'https://api.sampleapis.com/playstation/games'
2. Esta es la url de Xbox 'https://api.sampleapis.com/xbox/games'
3. Esta es la url de Switch ‘https://api.sampleapis.com/switch/games'