Memoria

El objeto de este EDA consiste en averiguar cuales son y como se relacionan entre sí las características que han influido en las canciones de mayor éxito de un año precovid como el 2019 y otro en el que mucha gente se vio afectado por los efectos de la pandemia como fue el 2020.

Para ello vamos a bucear entre los datasets que nos ofrece Kaggle sobre la app Spotify. Desafortunadamente no he podido incluir un DataSet de 2018, que también iba a ser objeto de estudio, porque que le faltaba la columna de género musical, lo que suponía prescindir de muchas comparaciones relevantes.

En primer lugar, voy a moldearlos y configurarlos de modo que presenten una información homogénea y de interés para el análisis a realizar, por lo que es necesario:

- Renombrar las columnas, que, a pesar de presentar la misma información en los csv, aparecen las mismas nombradas de forma diferente.
- Escalar las unidades de medida para su comparación.
- Eliminar columnas que no muestran información de utilidad.
- Añadir la columna “date”
- Modificar la columna “posición” del año 2020

En este primer punto he tenido dificultad para cada uno de los pasos ya que no sale todo a la primera. He tenido que pelear con el código y preguntar a compañeros para resolver errores.

Para unir la columna “date”, que es algo sencillo y básico, llegué a construir una función con un bucle que me creara un nuevo DataFrame de una sola columna con valores 2019 o 2020 para después unirlo los DataSets (en el Notebook de pruebas dejo constancia). Un disparate de principiante.

El siguiente paso es unirlos, para ello necesito que las categorías de los valores de las columnas sean iguales. En la tabla de 2019 los valores numéricos venían en formato int64 y en 2020 en formato float64.

A nivel de código y con objeto de trabajar y mostrar los conocimientos adquiridos, he creado una función en un archivo .py que se llama desde Notebook principal del Eda.

Algo interesante a destacar ha sido construir un nuevo DataFrame para agrupar los géneros con menos presencia de cada año. Así las gráficas de tartas se ven más limpias y es más sencillo de seguir el hilo de la presentación.

Previamente cuando me dispuse a empezar a crear gráficas no se me pintaban ya que no había indicado *%matplotlib inline* esto ha supuesto horas de agobio y tensión, afortunadamente di con la solución y pude continuar.

Respecto a las gráficas, he usado para su representación las librerías Seaborn y Plotly. El motivo de la selección de estas librerías es que se ajustan a las necesidades de la información extraída de los DataFrame trabajados. Descarté de plano usar Matplolib básico ya que las ilustraciones son menos atractivas visualmente. No he usado Tableau ya que quería que se quedase reflejado el gráfico resultante del código en el Main_Notebook. Por último, respecto a Folium y Streamlit son herramientas que no he precisado usar para la presentación realizada.

Como conclusión

- La muestra de datos debe ser suficientemente representativa. A mayor tamaño más precisión.
- Si las variables a tratar no son numéricas pueden estar sesgadas por el criterio de quien determina los parámetros por las que se rigen. Ejemplo: la bailabilidad o las variantes de los géneros.
- Como hemos visto existen factores externos que influyen en los resultados.
- Es un EDA con fines académicos que me ha servido para:
- Aplicar conocimiento aprendido
- Comprender el campo del estudio
- Aunque difícilmente pueda tomar decisiones informadas de la información obtenida.