

Workshop-2



Santiago Gomez Castro (2226287)

Docente: Javier Alejandro Vergara Zorrilla

Universidad Autónoma de Occidente

Facultad de Ingeniería

Santiago de Cali

2024

¿Sobre qué trata este trabajo?

El trabajo llamado WorkShop-2 nos entrega 2 dataset en formato csv, el primer dataset se encuentran todas las canciones de 125 géneros musicales diferentes junto al nombre del artista, álbum donde aparece la canción, a continuación una explicación sobre las columnas y su contenido:

- **track_id:** El código único que Spotify le da a cada canción.
- **artists:** Los nombres de los artistas que interpretan la canción. Si hay más de uno, sus nombres están separados por un punto y coma (;).
- **album_name:** El nombre del álbum donde aparece la canción.
- **track_name:** El nombre de la canción.
- **popularity:** La popularidad de la canción, medida del 0 al 100. Cuanto más popular sea, más cerca estará de 100. La popularidad depende de cuántas veces y cuán recientemente ha sido reproducida.
- **duration_ms:** La duración de la canción en milisegundos.
- **explicit:** Indica si la canción tiene contenido explícito (palabras groseras o temas adultos). True significa que sí tiene, y False que no.
- **danceability:** Mide cuán bailable es la canción, con un valor entre 0.0 y 1.0 (1.0 es más bailable).
- **energy:** Mide la intensidad y actividad de la canción, entre 0.0 y 1.0. Canciones rápidas y ruidosas tienen más energía.
- **key:** La tonalidad musical de la canción (por ejemplo, C, D, E, etc.). Si no se detectó ninguna, el valor es -1.
- **loudness:** El volumen general de la canción en decibelios (dB).
- **mode:** Indica si la canción está en tono mayor (1) o menor (0).
- **speechiness:** Mide la cantidad de palabras habladas en la canción. Un valor más cercano a 1.0 indica más palabras habladas.
- **acousticness:** Indica cuán acústica es la canción, con un valor entre 0.0 y 1.0 (1.0 es completamente acústica).
- **instrumentalness:** Predice si la canción es instrumental, es decir, sin voz. Cuanto más cerca de 1.0, más probable que no tenga voces.
- **liveness:** Mide la probabilidad de que la canción haya sido grabada en vivo. Un valor alto (más de 0.8) indica una mayor probabilidad de que sea en vivo.
- **valence:** Mide cuán positiva o feliz suena la canción, entre 0.0 (triste) y 1.0 (feliz).
- **tempo:** El ritmo de la canción en beats por minuto (BPM).
- **time_signature:** El compás de la canción, que indica cuántos tiempos hay en cada medida (por ejemplo, 3/4 o 4/4).
- **track_genre:** El género musical al que pertenece la canción (como pop, rock, jazz, etc.).

El segundo dataset contiene información sobre todas las premiaciones grammys desde el año 1959 hasta el 2019, a continuación una explicación sobre las columnas:

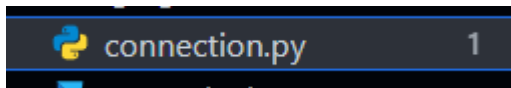
- **year:** El año en el que se celebraron los Grammy.
- **title:** El título del evento de los Grammy (incluye el número de edición y el año).
- **published_at:** La fecha y hora en que se publicó la información.
- **updated_at:** La fecha y hora en que se actualizó por última vez la información.
- **category:** La categoría del premio (por ejemplo, "Record of the Year").
- **nominee:** La canción o proyecto nominado en esa categoría.
- **artist:** El artista que interpreta la canción o está nominado.
- **workers:** Lista de las personas involucradas en la producción del proyecto (productores, ingenieros, etc.).
- **img:** El enlace a la imagen relacionada con el nominado o el premio.
- **winner:** Indica si el nominado fue el ganador en su categoría (True significa que ganó, False que no).

Se realizará todo el proceso de EDA a ambos dataset en donde se analizaran los datos que contienen, realizar modificaciones y organización de la información para realizar análisis profundos y graficación.

Herramientas usadas

- **Python:** Se usó python para la creación de scripts de subida de datos a la base de dato y Dags para el funcionamiento de Airflow
- **Jupyter:** Se emplean Notebooks de jupyter para el EDA de ambos dataset, donde se limpia, transforma la información y a su vez se crean gráficas para entender más sencillamente la información.
- **Ubuntu:** Una máquina virtual Ubuntu para poder ejecutar el proyecto en un ambiente linux ya que Airflow lo requiere para su correcto funcionamiento.
- **Apache-Ariflow:** Gestor de flujos de datos para la modificación, transformación y carga de los datos de forma secuencial.
- **Poetry:** Ambiente de desarrollo en Python para la gestión de las librerías requeridas.
- **Git y Github:** Gestores de versión de código para guardar y compartir el proyecto.
- **PowerBI:** Las visualizaciones principales se crearon usando PowerBI
- **SQLAlchemy:** Esta librería nos permite la conexión a la base de datos para la obtención de los datos y luego actualizarlos después de haber pasado por el EDA.
- **Pandas:** Librería para el análisis de los datos y su manipulación.
- **Dontev:** Libreria para acceder a las credenciales de la base de datos y no exponerlas en el código.
- **PostgreSQL:** Base de datos relacional que nos facilitara el guardado y gestión de los datos.
- **Ngrok:** Se emplea Ngrok para crear un sub túnel para exponer nuestra base de datos local a la máquina Ubuntu.

Subida de datos



Usando un script de python subimos el dataset the_grammy_award.csv a la base de datos antes de proceder al análisis, revisamos en la base de datos para asegurarnos de la subida.

	year	title	published_at	updated_at	category	nominee	artist	workers
1	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Bad Guy	Billie Eilish	Finneas O'Connell, produce
2	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Hey, Ma	Bon Iver	BJ Burton, Brad Cook, Chris
3	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	7 rings	Ariana Grande	Charles Anderson, Tommy I
4	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Hard Place	H.E.R.	Rodney "Darkchild" Jerkins,
5	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Talk	Khalid	Disclosure & Denis Kosiak, i
6	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Old Town Road	Lil Nas X Featuring Billy Ray Cyrus	Andrew "VoxGod" Bolooki,
7	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Truth Hurts	Lizzo	Ricky Reed & Tele, produce
8	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Record Of The Year	Sunflower	Post Malone & Swae Lee	Louis Bell & Carter Lang, pr
9	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	When We All Fall Asleep, Where Do We Go?	Billie Eilish	Finneas O'Connell, produce
10	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	U	Bon Iver	Brad Cook, Chris Messina &
11	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	Norman F***ing Rockwell!	Lana Del Rey	Jack Antonoff & Lana Del R
12	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	thank u, next	Ariana Grande	Tommy Brown, Ilya, Max M
13	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	I Used To Know Her	H.E.R.	David "Swagg R'Celious" Hi
14	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	7	Lil Nas X	Joe Grasso, engineer/mixer;
15	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	Cuz I Love You (Deluxe)	Lizzo	Ricky Reed, producer; Manr
16	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Album Of The Year	Father Of The Bride	Vampire Weekend	Ezra Koenig & Ariel Rechtsh
17	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Bad Guy	[NULL]	Billie Eilish O'Connell & Fin
18	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Always Remember Us This Way	[NULL]	Natalie Hemby, Lady Gaga,
19	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Bring My Flowers Now	[NULL]	Brandi Carlile, Phil Hanserot
20	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Hard Place	[NULL]	Ruby Amanfu, Sam Ashwor
21	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Lover	[NULL]	Taylor Swift, songwriter (Tay
22	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Norman F***ing Rockwell	[NULL]	Jack Antonoff & Lana Del R
23	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Someone You Loved	[NULL]	Tom Barnes, Lewis Capaldi,
24	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Song Of The Year	Truth Hurts	[NULL]	Steven Cheung, Eric Frederi
25	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Best New Artist	Billie Eilish	[NULL]	[NULL]
26	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Best New Artist	Black Pumas	[NULL]	[NULL]
27	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Best New Artist	Lil Nas X	[NULL]	[NULL]
28	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19T05:10:28-07:00	2020-05-19T05:10:28-07:00	Best New Artist	Lizzo	[NULL]	[NULL]

EDA

01_Grammy_EDA

Este NoteBook se encuentra todo el proceso de transformación, limpieza y carga del dataset de grammy_award.

Importación de las librerías que se van a utilizar.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sqlalchemy import create_engine, inspect
from dotenv import load_dotenv
import os
```

Bloque de código que creará la conexión con la base de datos.

```

load_dotenv()

localhost = os.getenv('LOCALHOST')
port = os.getenv('PORT')
nameDB = os.getenv('DB_NAME')
userDB = os.getenv('DB_USER')
passDB = os.getenv('DB_PASS')

try:
    engine = create_engine(f'postgresql+psycopg2://{userDB}:{passDB}@{localhost}:{port}/{nameDB}')
    inspector = inspect(engine)

    connection = engine.connect()
    print("Successfully connected to the database.")

    connection.close()

except Exception as e:
    print(f"Failed to connect to the database: {e}")

```

Successfully connected to the database.

Luego de crear la conexión imprimimos el dataset para observar a primera vista cuáles datos hay y sus tipos.

```

year          int64
title         object
published_at  object
updated_at    object
category      object
nominee       object
artist        object
workers       object
img           object
winner        bool
dtype: object

```

```

year          title          published_at \
0 2019 62nd Annual GRAMMY Awards (2019) 2020-05-19T05:10:28-07:00
1 2019 62nd Annual GRAMMY Awards (2019) 2020-05-19T05:10:28-07:00
2 2019 62nd Annual GRAMMY Awards (2019) 2020-05-19T05:10:28-07:00
3 2019 62nd Annual GRAMMY Awards (2019) 2020-05-19T05:10:28-07:00
4 2019 62nd Annual GRAMMY Awards (2019) 2020-05-19T05:10:28-07:00

          updated_at          category  nominee  artist \
0 2020-05-19T05:10:28-07:00 Record Of The Year Bad Guy Billie Eilish
1 2020-05-19T05:10:28-07:00 Record Of The Year Hey, Ma Bon Iver
2 2020-05-19T05:10:28-07:00 Record Of The Year 7 rings Ariana Grande
3 2020-05-19T05:10:28-07:00 Record Of The Year Hard Place H.E.R.
4 2020-05-19T05:10:28-07:00 Record Of The Year Talk Khalid

          workers \
0 Finneas O'Connell, producer; Rob Kinelski & Fi...
1 BJ Burton, Brad Cook, Chris Messina & Justin V...
2 Charles Anderson, Tommy Brown, Michael Foster ...
3 Rodney "Darkchild" Jerkins, producer; Joseph H...
4 Disclosure & Denis Kosiak, producers; Ingmar C...

          img  winner
0 https://www.grammy.com/sites/com/files/styles/... True
1 https://www.grammy.com/sites/com/files/styles/... True
2 https://www.grammy.com/sites/com/files/styles/... True

```

Primero procedemos en arreglar el tipo de datos de la columnas `published_at` y `updated_at` ya que no era acorde con la naturaleza de los datos, lo pasamos a tipo `Date` que es un tipo más acorde a la información que poseen.

```

year          int64
title         object
published_at  datetime64[ns, UTC]
updated_at    datetime64[ns, UTC]
category      object
nominee       object
artist        object
workers       object
img           object
winner        bool
dtype: object

```

Luego procedemos a borrar la columna de “img” porque no es necesario para los análisis que vamos a realizar.

```
df = df.drop(columns="img")
print(df.info())
```

```
class 'pandas.core.frame.DataFrame'>
RangeIndex: 4810 entries, 0 to 4809
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   year                  4810 non-null  int64
1   title                 4810 non-null  object
2   published_at          4810 non-null  datetime64[ns, UTC]
3   updated_at            4810 non-null  datetime64[ns, UTC]
4   category              4810 non-null  object
5   nominee              4804 non-null  object
6   artist                2970 non-null  object
7   workers               2620 non-null  object
8   winner                4810 non-null  bool
```

Revisamos los valores null que hay en el dataset

```
year          0
title         0
published_at  0
updated_at    0
category      0
nominee       6
artist        1840
workers       2190
winner        0
dtype: int64
```

Como observamos, tenemos un número considerable de valores nulos en la columna de “artists”, pero ya que no podemos rellenar con alguna operación se decide con borrar todos los null de la columna y después de eso cambiar los valores null de la columna de “workers” porque existe la probabilidad de que el artista haya trabajado solo o si trabajo con alguien, y ya que no es muy necesario saber con quien el artista ha trabajado para el análisis.

```

year          0
title         0
published_at  0
updated_at    0
category      0
nominee       0
artist        0
workers       2004
winner        0
dtype: int64

```

```

df['workers'].fillna('Unknown', inplace=True)
print(df.info())

```

Por último actualizamos los datos viejos de la base de datos con los nuevos que se han trabajado.

```

try:
    df.to_sql('grammy_awards', engine, if_exists='replace', index=False)

    print(f"Table 'grammy_awards' updated.")

except Exception as e:
    print(f"Error uploading data: {e}")

finally:
    engine.dispose()

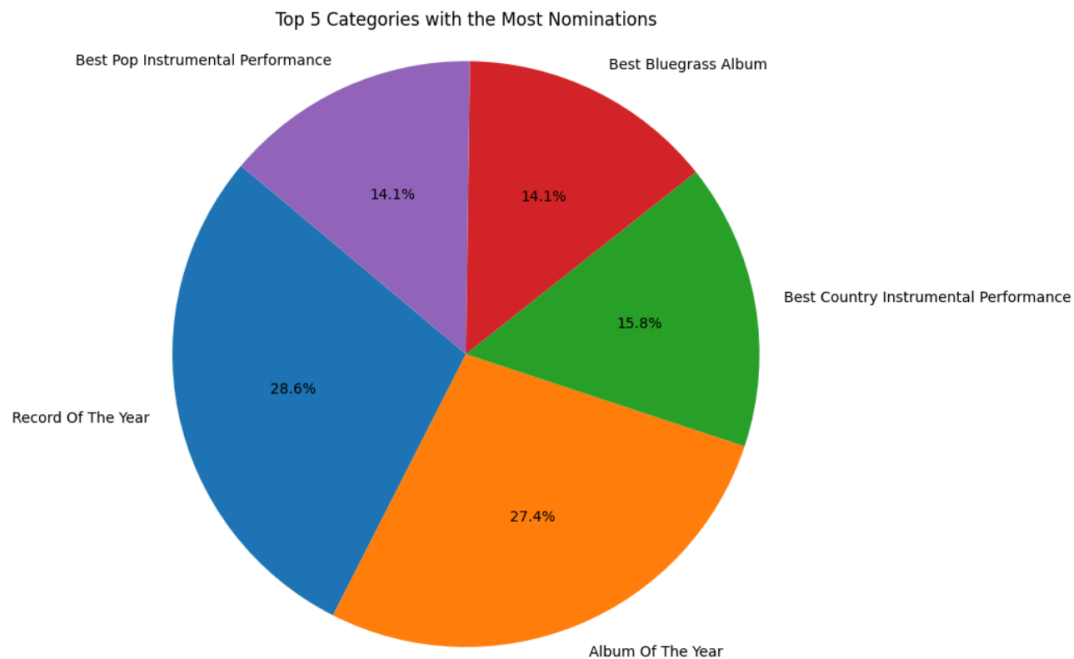
```

Table 'grammy_awards' updated.

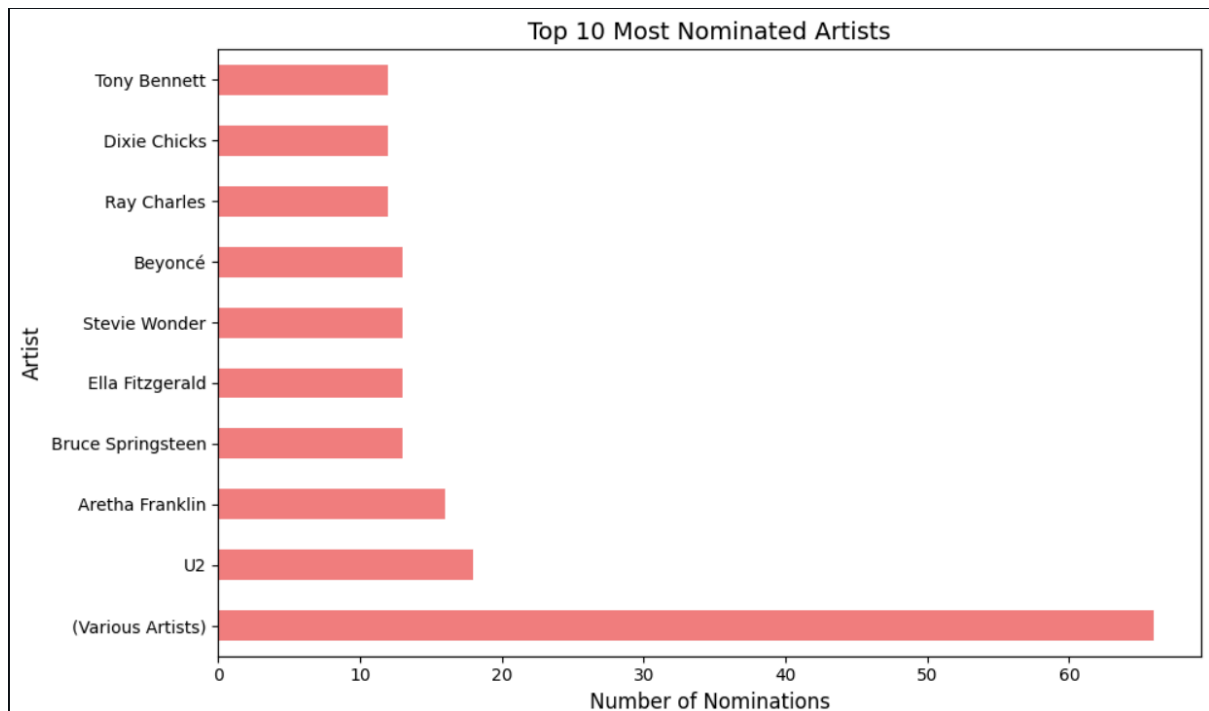
		A2 title	published_at	updated_at	A2 category	A2 nominee	A2 artist	A2 workers
1	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Bad Guy	Billie Eilish	Finneas O'Connell, producer; Rob Kinelski &
2	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Hey, Ma	Bon Iver	BJ Burton, Brad Cook, Chris Messina & Just
3	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	7 rings	Ariana Grande	Charles Anderson, Tommy Brown, Michael
4	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Hard Place	H.E.R.	Rodney "Darkchild" Jerkins, producer; Josep
5	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Talk	Khalid	Discoourse & Denis Kosiak, producers; Igmm
6	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Old Town Road	Lil Nas X Featuring Billy Ray Cyrus	Andrew "VoxGod" Boloski, Jocelyn "Jazzy"
7	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Truth Hurts	Lizzo	Ricky Reed & Tele, producers; Chris Galland
8	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Record Of The Year	Sunflower	Post Malone & Swae Lee	Louis O'Connell, producer; Rob Kinelski &
9	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	When We All Fall Asleep, Where Do We Go?	Billie Eilish	Finneas O'Connell, producer; Rob Kinelski &
10	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	II	Bon Iver	Brad Cook, Chris Messina & Justin Vernon, J
11	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	Norman F***ing Rockwell!	Lana Del Rey	Jack Antonoff & Lana Del Rey, producers; J
12	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	I thank u, next	Ariana Grande	Tommy Brown, Ilya, Max Martin & Victoria
13	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	I Used To Know Her	H.E.R.	David "Swagg R" Celious" Harris, H.E.R., Wal
14	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	7	Lil Nas X	Joe Grasso, engineer/mixer; Montero Lama
15	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	Cuz I Love You (Deluxe)	Lizzo	Ricky Reed, producer; Manny Marroquin &
16	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Album Of The Year	Father Of The Bride	Vampire Weekend	Ezra Koenig & Ariel Rechtshaid, producers;
17	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Solo Performance	Truth Hurts	Lizzo	Unknown
18	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Solo Performance	Spirit	Beyoncé	Unknown
19	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Solo Performance	Bad Guy	Billie Eilish	Unknown
20	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Solo Performance	7 rings	Ariana Grande	Unknown
21	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Solo Performance	You Need To Calm Down	Taylor Swift	Unknown
22	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Duo/Group Performance	Old Town Road	Lil Nas X Featuring Billy Ray Cyrus	Unknown
23	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Duo/Group Performance	Boyfriend	Ariana Grande & Social House	Unknown
24	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Duo/Group Performance	Sucker	Jonas Brothers	Unknown
25	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Duo/Group Performance	Sunflower	Post Malone & Swae Lee	Unknown
26	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Pop Duo/Group Performance	Señorita	Shawn Mendes & Camila Cabello	Unknown
27	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Traditional Pop Vocal Album	Look Now	Eivis Costello & The Imposters	Unknown
28	2019	62nd Annual GRAMMY Awards (2019)	2020-05-19 07:10:28.000 -0500	2020-05-19 07:10:28.000 -0500	Best Traditional Pop Vocal Album	Si	Andrea Bocelli	Unknown

Visualizaciones.

El gráfico de pastel nos muestra las categorías con más nominaciones que se han realizado con el paso del tiempo, desde el año 59 hasta el 2019, observamos que la primera categoría es “album of the year”(álbum del año), de segundo “record of the year”(disco del año) y de tercero “Best Country Instrumental Performance”(Mejor interpretación instrumental country) siendo las 3 categorías que se han nominado en estos aproximadamente 60 años.



La gráfica de barras nos muestra los artistas que fueron los más nominados en los grammy donde el grupo musical U2 son los más nominadas a los grammys seguido por Aretha Franklin y Bruce Springsteen, Ella Fitzgerald, Stevie Wonder y Beyonce como los terceros, esto lo podemos observar en las barras tienen el mismo ancho.



02_Spotify_dataset

Este Notebook se encuentra todo el proceso de transformación, limpieza y carga del dataset de spotify_dataset.

Importación de las librerías

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Accedemos al CSV que se encuentra en la carpeta "/Dirty-Data" e imprimimos para observar los tipos de datos y qué información aparece.

```
df = pd.read_csv('../Dirty-Data/spotify_dataset.csv')

pd.set_option('display.max_rows', 100)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)

df.head(5)
```

Unnamed: 0	track_id	artists	album_name	track_name	popularity	duration_ms	explicit	danceability	energy	key	loudness	mode
0	5SuOikwiRyPMVolQDIUgSV	Gen Hoshino	Comedy	Comedy	73	230666	False	0.676	0.4610	1	-6.746	0
1	4qPND8W1i3p13qLCt0KI3A	Ben Woodward	Ghost (Acoustic)	Ghost - Acoustic	55	149610	False	0.420	0.1660	1	-17.235	1
2	1iJBSr7s7jYXzM8EGcbK5b	Ingrid Michaelson;ZAYN	To Begin Again	To Begin Again	57	210826	False	0.438	0.3590	0	-9.734	1
3	6lfxq3CG4xtTiEg7opyCyx	Kina Grannis	Crazy Rich Asians (Original Motion Picture Soundtrack)	Can't Help Falling In Love	71	201933	False	0.266	0.0596	0	-18.515	1
4	5vjLSffimilP26QG5WcN2K	Chord Overstreet	Hold On	Hold On	82	198853	False	0.618	0.4430	2	-9.681	1

```

RangeIndex: 114000 entries, 0 to 113999
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype  
---  --
0   Unnamed: 0             114000 non-null  int64  
1   track_id               114000 non-null  object  
2   artists                113999 non-null  object  
3   album_name             113999 non-null  object  
4   track_name             113999 non-null  object  
5   popularity             114000 non-null  int64  
6   duration_ms            114000 non-null  int64  
7   explicit               114000 non-null  bool    
8   danceability           114000 non-null  float64 
9   energy                 114000 non-null  float64 
10  key                    114000 non-null  int64  
11  loudness               114000 non-null  float64 
12  mode                   114000 non-null  int64  
13  speechiness            114000 non-null  float64 
14  acousticness           114000 non-null  float64 
15  instrumentalness        114000 non-null  float64 
16  liveness               114000 non-null  float64 
17  valence                114000 non-null  float64 
18  tempo                  114000 non-null  float64 
19  time_signature          114000 non-null  int64  
20  track_genre            114000 non-null  object  
dtypes: bool(1), float64(9), int64(6), object(5)
memory usage: 17.5+ MB
None

```

Revisamos que no haya algún valor repetido

```

1  duplicates = df.duplicated().sum()
   print(duplicates)

0

```

Buscamos si hay valores null

```
df = df.dropna()
null_data = df.isnull().sum()
print(null_data)
```

```
Unnamed: 0      0
track_id        0
artists         0
album_name      0
track_name      0
popularity      0
duration_ms     0
explicit        0
danceability    0
energy          0
key             0
loudness        0
mode            0
speechiness     0
acousticness    0
instrumentalness 0
liveness        0
valence         0
tempo           0
time_signature  0
track_genre     0
dtype: int64
```

Por último borramos la columna “acousticness” por no ser necesaria para los análisis que se va a emplear el dataset.

```
df = df.drop(columns='acousticness')

print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 113999 entries, 0 to 113999
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Unnamed: 0             113999 non-null  int64
1   track_id               113999 non-null  object
2   artists                113999 non-null  object
3   album_name             113999 non-null  object
4   track_name             113999 non-null  object
5   popularity             113999 non-null  int64
6   duration_ms            113999 non-null  int64
7   explicit               113999 non-null  bool
8   danceability           113999 non-null  float64
9   energy                 113999 non-null  float64
10  key                    113999 non-null  int64
11  loudness                113999 non-null  float64
12  mode                   113999 non-null  int64
13  speechiness            113999 non-null  float64
14  instrumentalness        113999 non-null  float64
15  liveness                113999 non-null  float64
16  valence                 113999 non-null  float64
17  tempo                   113999 non-null  float64
18  time_signature          113999 non-null  int64
19  track_genre             113999 non-null  object
dtypes: bool(1), float64(8), int64(6), object(5)
memory usage: 17.5+ MB
None
```

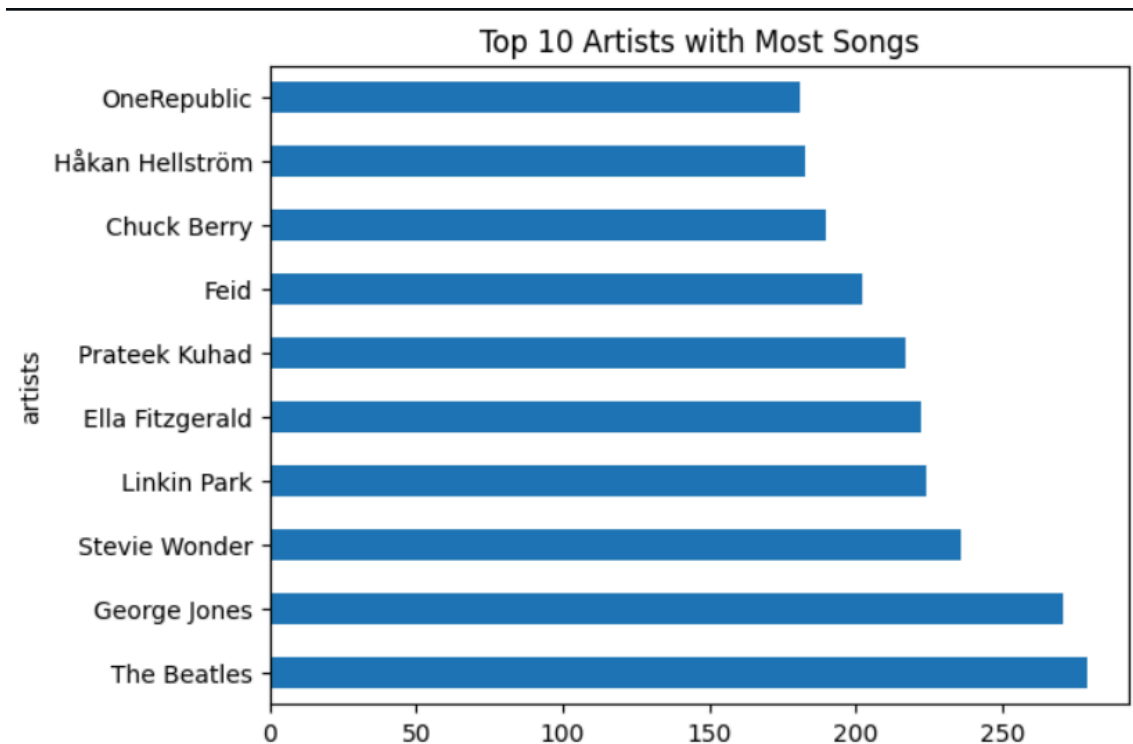
Finalmente guardamos el nuevo CSV en la carpeta “./cleanData”

Finally, we save the data into a CSV file.

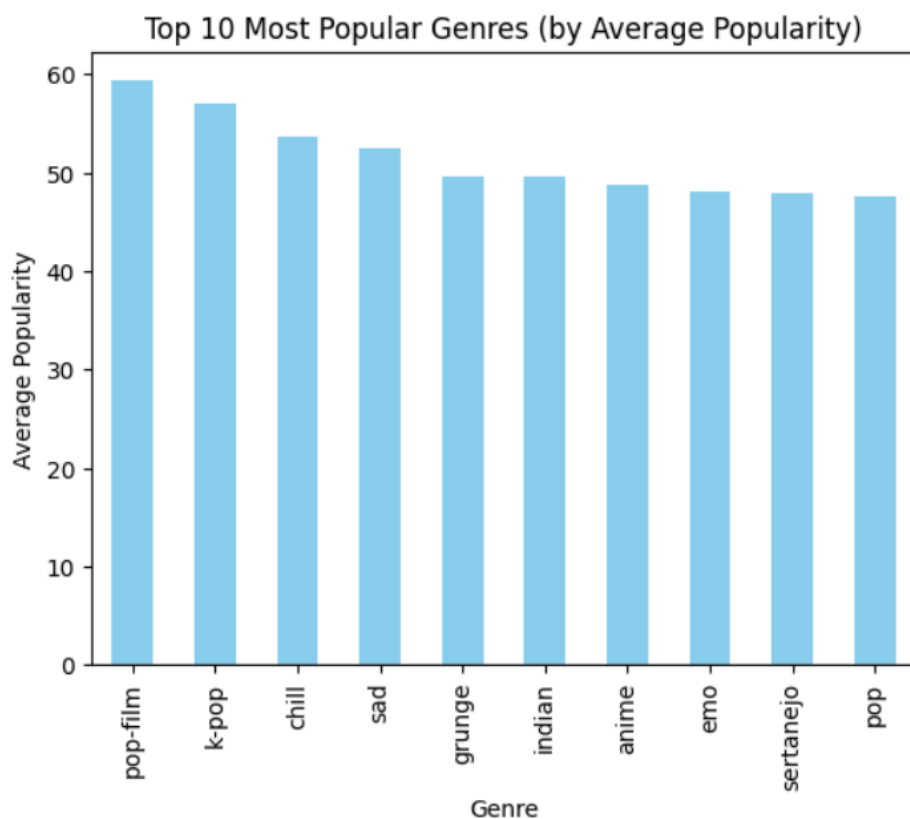
```
df.to_csv('../cleanData/spotify-clean.csv', index=False)
```

Visualizaciones.

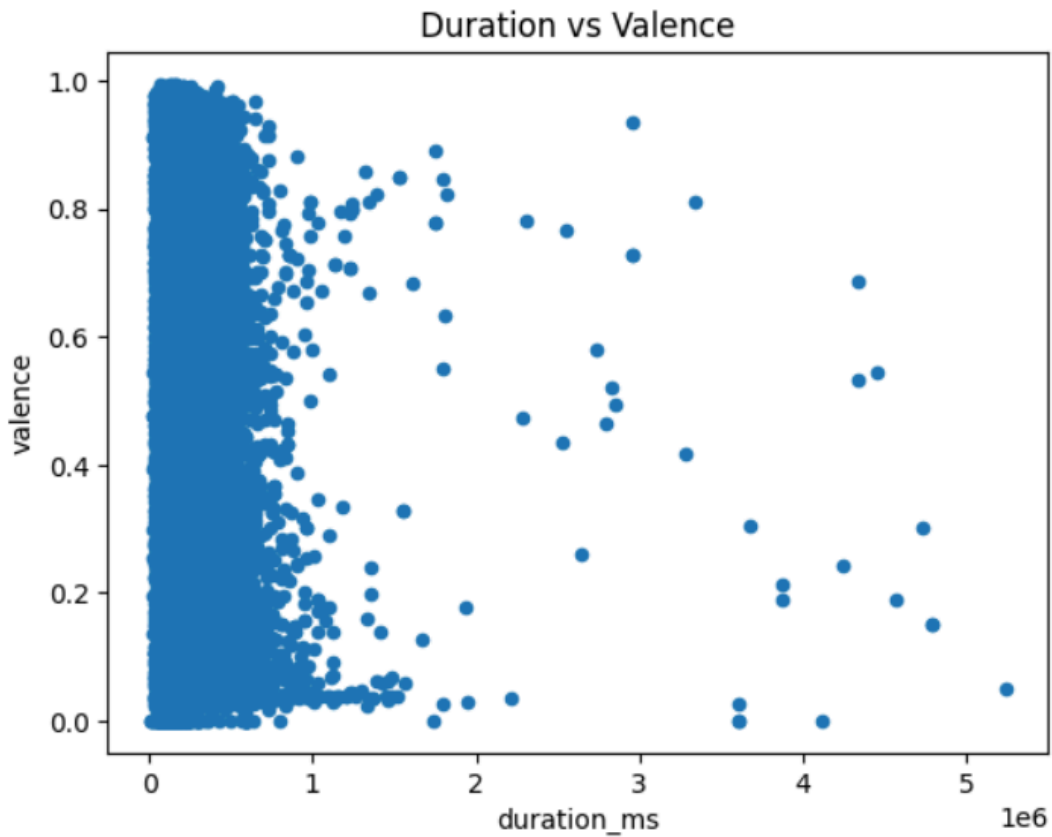
Esta gráfica nos muestra los 10 artistas con más canciones en spotify, la banda británica The Beatles, son la banda con más canciones en spotify seguida por George Jones y Stevie Wonder como artistas con más canciones en spotify.



Esta gráfica muestra la media en popularidad de cada género, con esto obtenemos un top 10 de los géneros más populares encabezado por Pop-film, k-pop y chill.

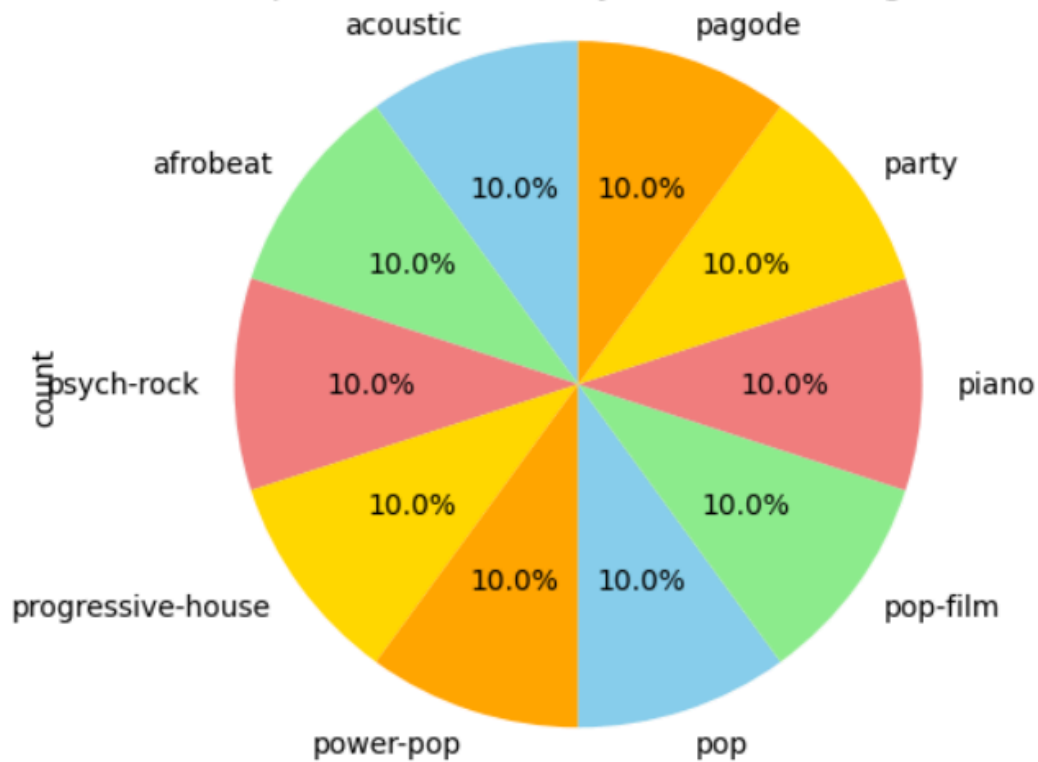


Esta gráfica de dispersión permite observar cómo se distribuye la felicidad de la gente de acuerdo a la duración de una canción, observando que las canciones a menor duración las personas son más felices.



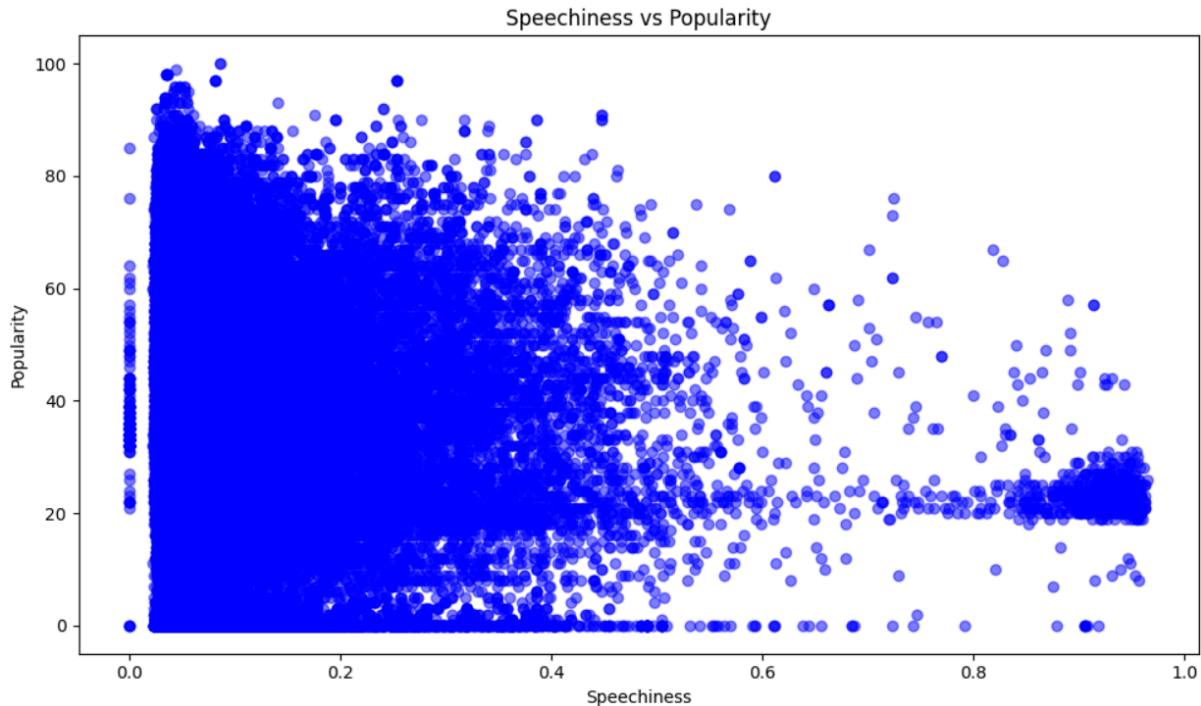
Este gráfico de pastel muestra cual genero posee mas canciones, el problema es que los porcentajes son iguales, mostrando que los géneros poseen las misma cantidad de canciones, entonces se optó por un código que cuente la cantidad de canciones de cada género musical, como resultado esta la imagen de abajo del gráfico de pastel, esta imagen nos muestra que efectivamente todos los géneros tienen la misma cantidad de canciones a excepción de k-pop que posee 999 canciones a diferencia de las 1.000 canciones que tienen los demas generos.

Top 5 Music Genres by Number of Songs



```
track_genre
acoustic      1000
afrobeat      1000
psych-rock    1000
progressive-house 1000
power-pop     1000
...
emo           1000
electronic    1000
electro       1000
world-music   1000
k-pop         999
Name: count, Length: 114, dtype: int64
```


La gráfica de dispersión muestra la popularidad de una canción con la cantidad de palabras que se hablan, podemos concluir que las canciones más populares son aquellas que menos palabras tienen y ya las menos populares son aquellas que tienen muchas más palabras, también podemos ver que a más palabras la popularidad va decayendo.



Merge

La función `merge_all_csv` obtiene los Datasets como formato json para después normalizar a Dataframe, se usa la función **merge** siguiendo los siguientes parámetros: `pd.merge(df1, df2, left_on='nominee', right_on='track_name', how='inner')`.

Donde se usa las columnas "nominee" y "track_name", y se combina los valores que son iguales ya que ambas columnas son el nombre del artista, dejando solo los artistas nominados para los análisis y luego se borra las las columnas ['artists', 'liveness', 'time_signature'], siendo innecesarias para los análisis que se emplearía el Dataset.

```
def merge_all_csv(**kwargs):
    try:
        ti = kwargs["ti"]

        json_data = json.loads(ti.xcom_pull(task_ids="transform_DB"))
        df1 = pd.json_normalize(data=json_data)

        json_data = json.loads(ti.xcom_pull(task_ids="transform_csv"))
        df2 = pd.json_normalize(data=json_data)

        logging.info("CSV files loaded successfully.")

        df_merge = pd.merge(df1, df2, left_on='nominee', right_on='track_name', how='inner')
        logging.info(f"DataFrames merged successfully. Shape: {df_merge.shape}")

        df_merge.drop(columns=['artists', 'liveness', 'time_signature'], inplace=True)
        logging.info("Unnecessary columns dropped successfully.")

        count_duplicated = df_merge['track_id'].duplicated().sum()
        logging.info(f"Number of duplicates in the 'track_id' column before dropping: {count_duplicated}")

        df_merge = df_merge.drop_duplicates(subset=['track_id'], keep='first')

        count_duplicated = df_merge['track_id'].duplicated().sum()
        logging.info(f"Number of duplicates in the 'track_id' column after dropping: {count_duplicated}")

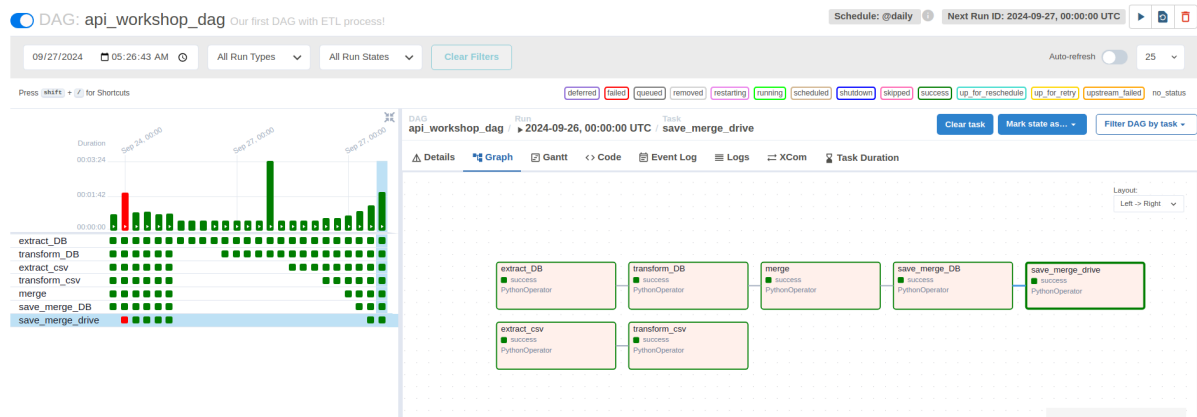
        logging.info("Dataset ready and cleaned.")

        logging.info(f"Final DataFrame: \n{df_merge.head()}")
        return df_merge.to_json(orient='records')

    except Exception as e:
        logging.error(f"Error during dataframe merge: {e}")
```

Airflow

El proceso de Airflow es un proceso ETL completo, teniendo la extracción de los datos de la base de datos y un CSV, la transformación de ambos verificando posibles datos nulos o duplicados y por último la subida del nuevo dataset a la base de datos y en Drive tal como se ve en la imagen de evidencia.



Evidencia:

Nombre	Motivo por el que se te sugiere	Propietario	Ubicación
df_merge.csv	Lo has creado • 0:28	yo	Mi unidad

Conexión a base de datos desde power BI

Base de datos

Base de datos PostgreSQL

4.tcp.ngrok.io:15042;gramy

Nombre de usuario

postgres

Contraseña

••••



Seleccionar en qué nivel hay que aplicar esta configuración

4.tcp.ngrok.io:15042

Atrás

Conectar

Cancelar

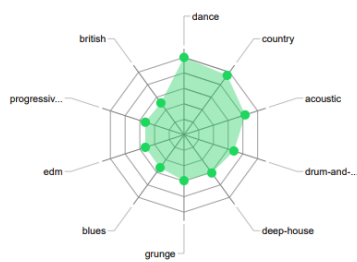
- ☐  public.grammy_awards
- ☐  public.merge_data

Descargamos los datos.

Dashboard

Winner count by track_genre

Axis ● Winner count

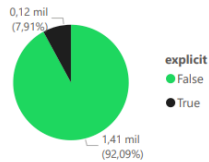


1530

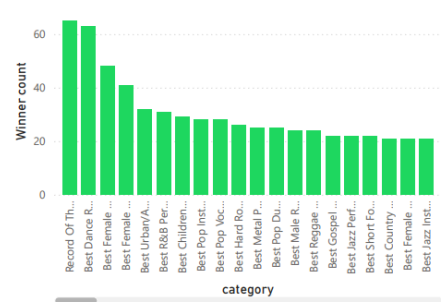
Nominee count



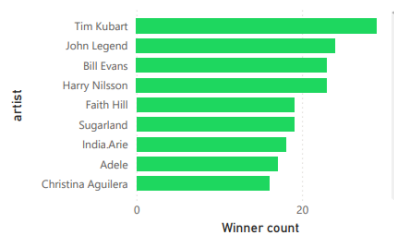
Explicit count per explicit



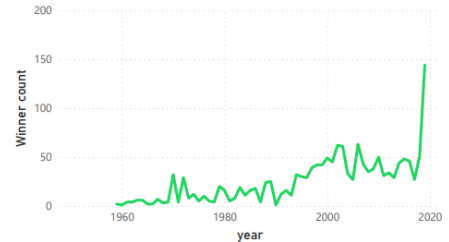
Winner count by category



Winners by artist



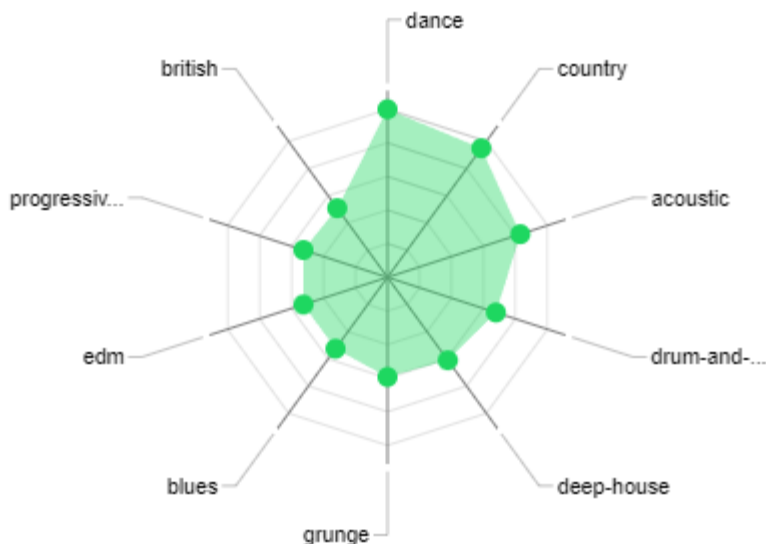
Winner count by year



Este es un gráfico de radar, nos muestra cómo se distribuye los géneros con mas grammys, observamos que empieza con “dance” tocando el borde y luego empieza a alejarse del borde acercándose un poco más al centro, mostrando el top de los géneros con más grammys y la diferencia que los separa.

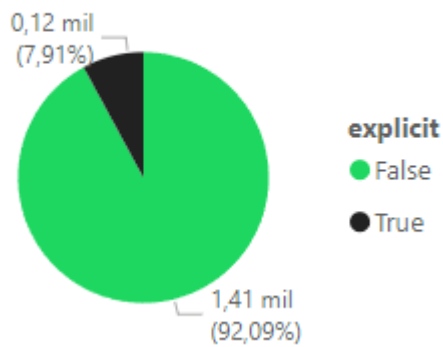
Winner count by track_genre

Axis ● Winner count



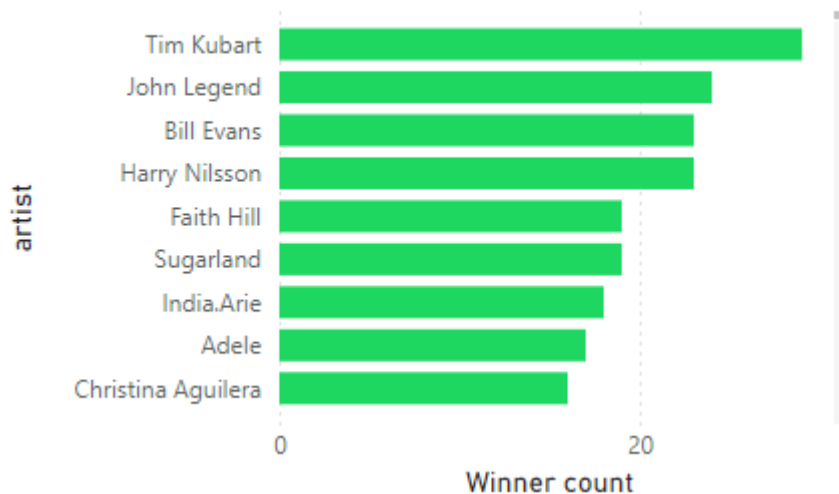
Explicit contiene datos de que si la cancion es grosera o contiene tema de adultos, este grafico de pastel nos muestra el porcentaje de canciones groseras(verde) con las que no son groseras(negro) con grammys, concluyendo que las canciones con tema para adultos tiene gran posibilidad de obtener un grammy.

Explicit count per explicit



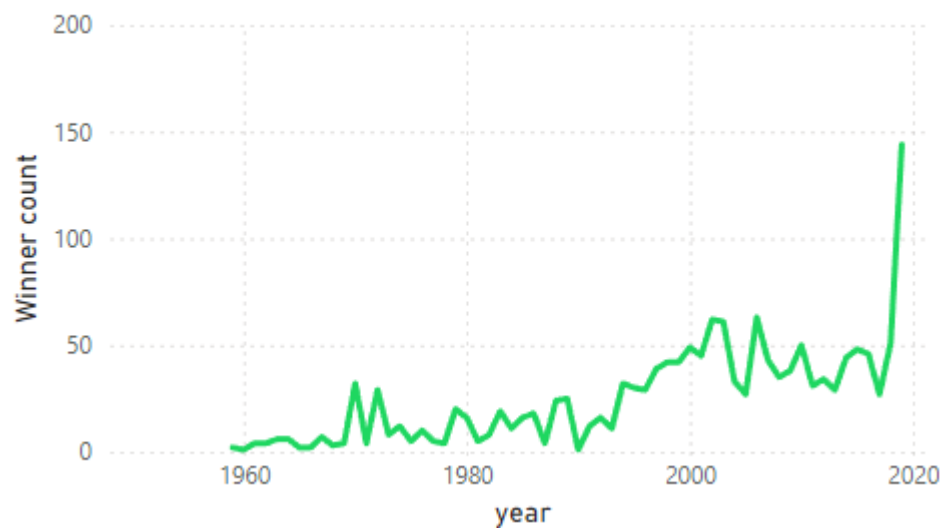
La gráfica muestra los artistas que más grammys han ganado con el paso de los años, siendo Tim Kubart el artista con más grammys seguido por Jhon Legend y de tercero Bill Evans.

Winners by artist



Aquí se muestra los grammys entregados desde los años 50 hasta el 2019 que fue la fecha de corte del dataset, el número de grammys ha ido aumentando desde los años 90 donde se ve una tendencia creciente que se estanca para luego en el año 2019 tener un pico de casi 150 grammys entregados.

Winner count by year



Como última gráfica, tenemos las categorías con más grammys entregados, la primera es disco del año obteniendo más de 60 de grammys, el segundo es la categoría de mejor disco de baile con más de 60 grammys entregados pero por detrás de disco del año, y de tercero está Mejor interpretación vocal country femenina estando por debajo de 50 grammys obtenidos, podemos concluir que las categorías más importantes de los grammy serian las anteriormente mencionadas.

Winner count by category

