Data Analysis: Spotify

21-03-2022

Proyecto Personal

Creado por: Torres, Leandro Joel

Nombre del logotipo

Introducción

En el presente proyecto tiene como finalidad analizar los datos de un dataset de las canciones de la plataforma de Spotify, mediante una limpieza de datos con Excel y la creación de dashboard con Power BI

Dataset: https://www.kaggle.com/datasets/ektanegi/spotifydata-19212020

Mirando y limpiando datos

Antes que nada si abrimos la pagina donde se encuentra el dataset, nos brinda una descripción de los datos, que nos es de gran ayuda!!

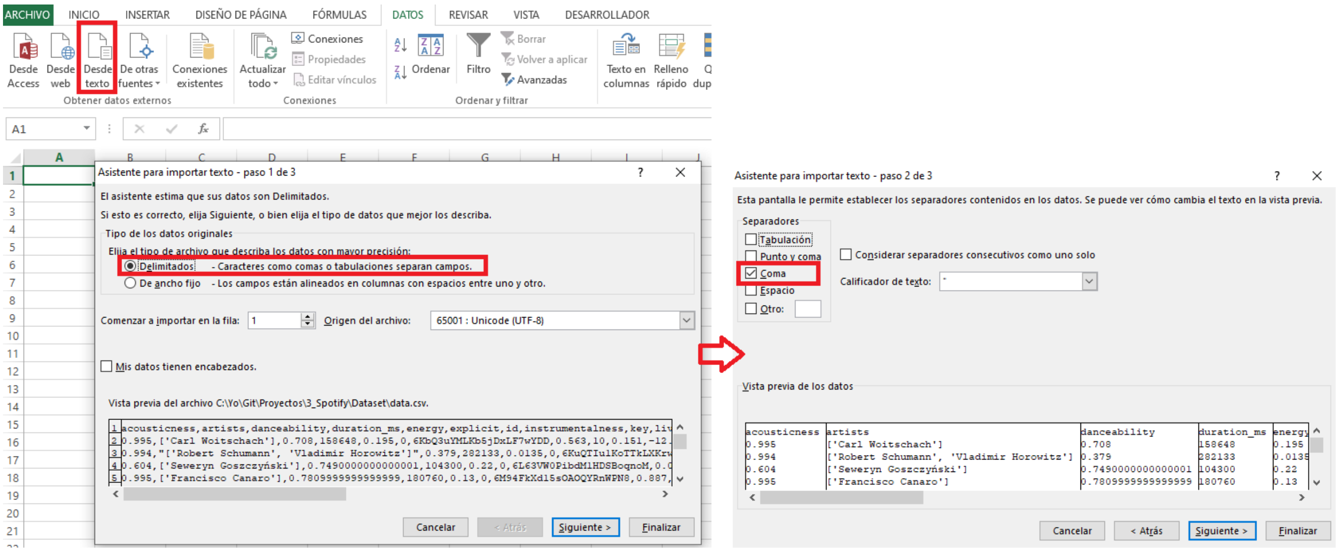
Description

The "data.csv" file contains more than 160.000 songs collected from Spotify Web API. The dataset is from Spotify and contains 169k songs from the year 1921 to year 2020. Each year got
top 100 songs.

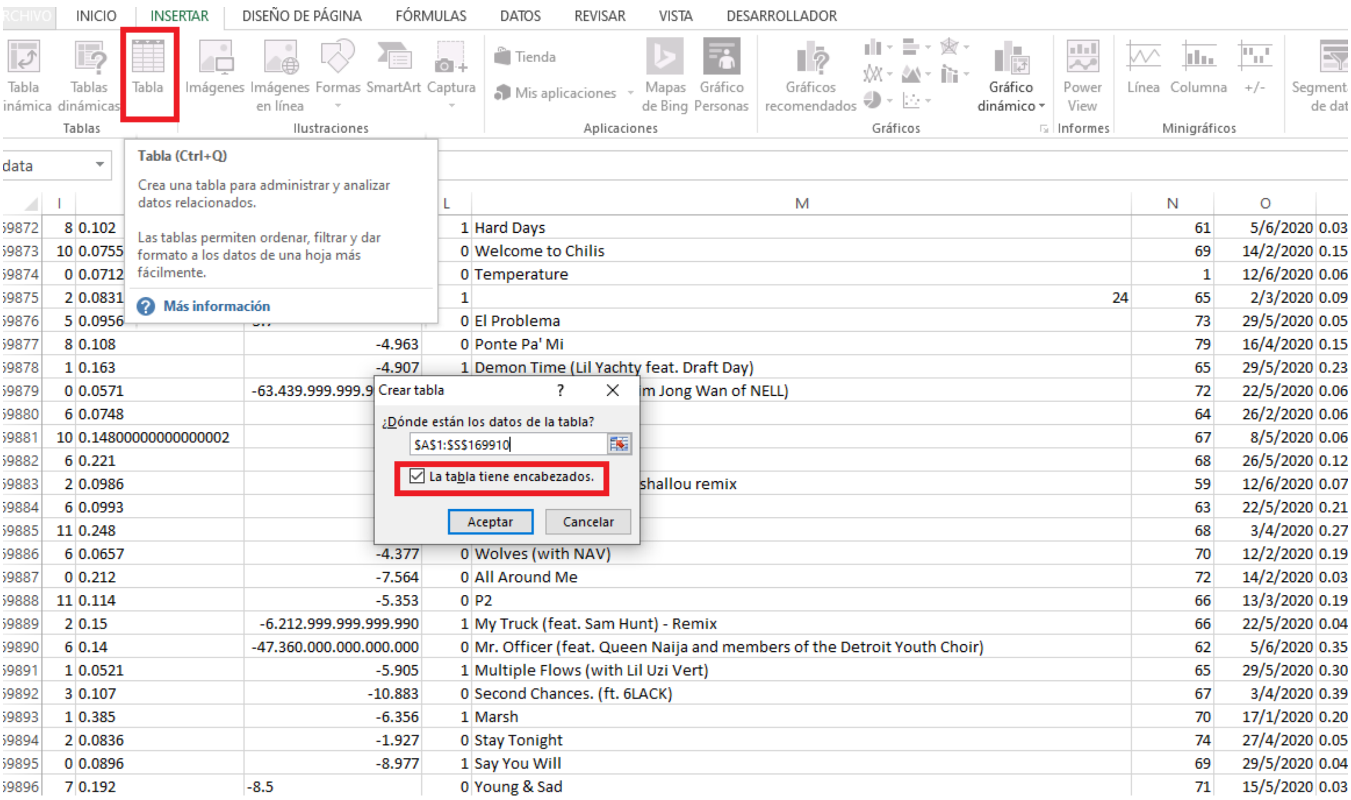
Primary:

- id (Id of track generated by Spotify)
  Numerical:
- acousticness (Ranges from 0 to 1)
- danceability (Ranges from 0 to 1)
- energy (Ranges from 0 to 1)
- duration_ms (Integer typically ranging from 200k to 300k)
- instrumentalness (Ranges from 0 to 1)
- valence (Ranges from 0 to 1)
- popularity (Ranges from 0 to 100)
- tempo (Float typically ranging from 50 to 150)
- liveness (Ranges from 0 to 1)
- loudness (Float typically ranging from -60 to 0)
- speechiness (Ranges from 0 to 1)
- year (Ranges from 1921 to 2020)
  Dummy:
- mode (0 = Minor, 1 = Major)
- explicit (0 = No explicit content, 1 = Explicit content)
  Categorical:
- key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on...)
- artists (List of artists mentioned)
- release_date (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)
- name (Name of the song)

Notamos que si abrimos en un bloc de notas, los datos están separados por comas, entonces ya viste esto, abrimos Excel, vamos a la pestaña de DATOS y luego seleccionamos la delimitacion por comas , con ello ya importamos el dataset.



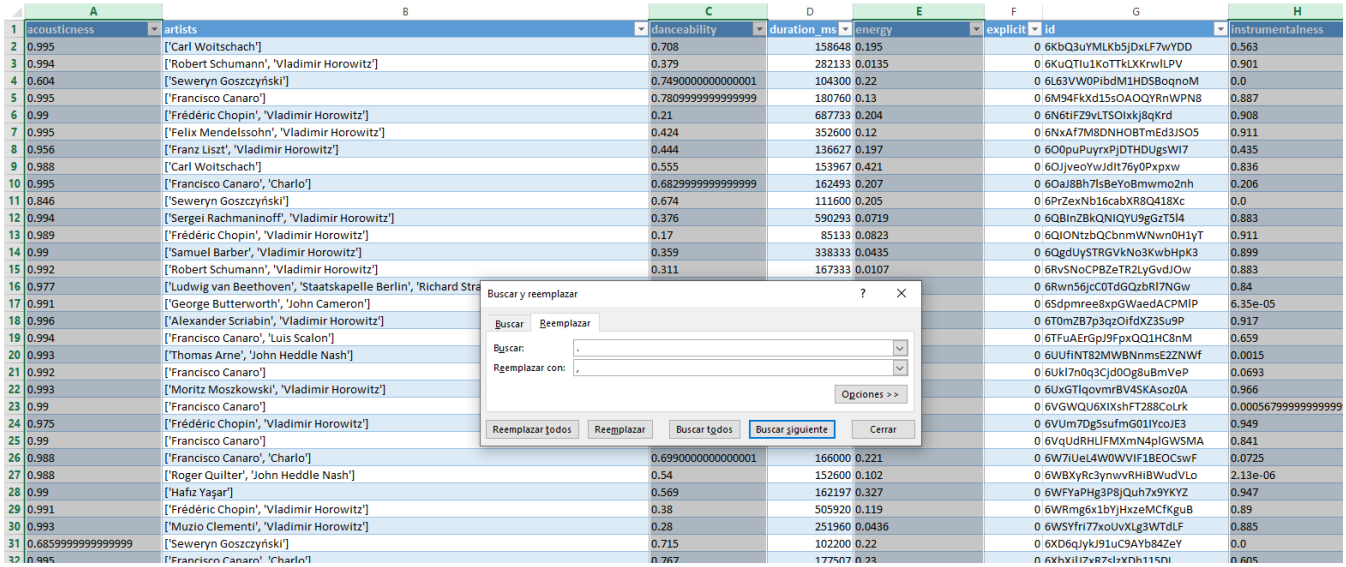
Una vez importado los datos vamos a crear una tabla para un mejor manejo de los mismos. Entonces vamos a la pestaña INSERTAR y seleccionamos tabla (Nota: si bien excel lo selecciona automaticamente, como puede haber filas con valores nulos, esto puede fallar ,así que es mejor seleccionar manualmente los datos, usando la tecla Shift , y la tecla Fin)



(Realizado esto, guardemos el libro de Excel en el directorio que queramos!!)

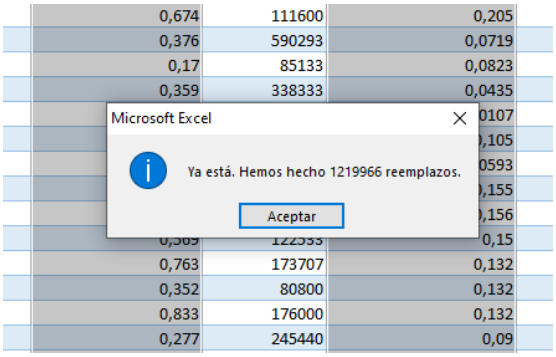
Lo primero que notamos es que hay datos numericos que se encuentran a la izquierda de cada celda, esto en excel significa que no lo está tratando como numeros sino como texto. Entonces tenemos que transformar los puntos por comas ,así tener decimales y no textos , para ello solo presionamos Ctrl+B, para buscar puntos y reemplazar por comas en todas las columnas numericas

Entonces seleccionamos las columnas y luego reemplazamos en todo



	A	B	C	D	E	F	G	H
1	acousticness	artists	danceability	duration_ms	energy	explicit	id	instrumentalness
2	0.995	['Carl Woitschach']	0.708	158648	0.195	0	6KbQ3uYMLKb5jDxLF7wYDD	0.563
3	0.994	['Robert Schumann', 'Vladimir Horowitz']	0.379	282133	0.0135	0	6KuQTIu1KoTTkLXKrwlLPV	0.901
4	0.604	['Seweryn Goszczyński']	0.7490000000000001	104300	0.22	0	6L63VW0PibdM1HDSBoqnoM	0.0
5	0.995	['Francisco Canaro']	0.7809999999999999	180760	0.13	0	6M94FkXd15sOAOQYRnWPN8	0.887
6	0.99	['Frédéric Chopin', 'Vladimir Horowitz']	0.21	687733	0.204	0	6N6tiFZ9vLTSOIxkj8qKrd	0.908
7	0.995	['Felix Mendelssohn', 'Vladimir Horowitz']	0.424	352600	0.12	0	6NxAf7M8DNHOBTmEd3JSO5	0.911
8	0.956	['Franz Liszt', 'Vladimir Horowitz']	0.444	136627	0.197	0	6O0puPuyrxPjDTHDUgsWI7	0.435
9	0.988	['Carl Woitschach']	0.555	153967	0.421	0	6OJjveoYwJdIt76y0Pxpxw	0.836
10	0.995	['Francisco Canaro', 'Charlo']	0.6829999999999999	162493	0.207	0	6OaJ8Bh7lsBeYoBmwmo2nh	0.206
11	0.846	['Seweryn Goszczyński']	0.674	111600	0.205	0	6PrZexNb16cabXR8Q418Xc	0.0
12	0.994	['Sergei Rachmaninoff', 'Vladimir Horowitz']	0.376	590293	0.0719	0	6QBInZBkQNIQYU9gGzT5l4	0.883
13	0.989	['Frédéric Chopin', 'Vladimir Horowitz']	0.17	85133	0.0823	0	6QIONtzbQCbnmWNwn0H1yT	0.911
14	0.99	['Samuel Barber', 'Vladimir Horowitz']	0.359	338333	0.0435	0	6QgdUySTRGVkNo3KwbHpK3	0.899
15	0.992	['Robert Schumann', 'Vladimir Horowitz']	0.311	167333	0.0107	0	6RvSNoCPBZeTR2LyGvdJOw	0.883
16	0.977	['Ludwig van Beethoven', 'Staatskapelle Berlin', 'Richard Stra				0	6Rwn56jcC0TdGQzbRl7NGw	0.84
17	0.991	['George Butterworth', 'John Cameron']				0	6Sdpmree8xpGWaedACPMlP	6.35e-05
18	0.996	['Alexander Scriabin', 'Vladimir Horowitz']				0	6T0mZB7p3qzOifdXZ3Su9P	0.917
19	0.994	['Francisco Canaro', 'Luis Scalon']				0	6TFuAErGpJ9FpxQQ1HC8nM	0.659
20	0.993	['Thomas Arne', 'John Heddle Nash']				0	6UUfiNT82MWBNnmsE2ZNWf	0.0015
21	0.992	['Francisco Canaro']				0	6Ukl7n0q3Cjd0Og8uBmVeP	0.0693
22	0.993	['Moritz Moszkowski', 'Vladimir Horowitz']				0	6UxGTlqovmrBV4SKAsoz0A	0.966
23	0.99	['Francisco Canaro']				0	6VGWQU6XIXshFT288CoLrk	0.00056799999999999
24	0.975	['Frédéric Chopin', 'Vladimir Horowitz']				0	6VUm7Dg5sufmG01IYcoJE3	0.949
25	0.99	['Francisco Canaro']				0	6VqUdRHLlFMXmN4plGWSMA	0.841
26	0.988	['Francisco Canaro', 'Charlo']	0.6990000000000001	166000	0.221	0	6W7iUeL4W0WVIF1BEOCswF	0.0725
27	0.988	['Roger Quilter', 'John Heddle Nash']	0.54	152600	0.102	0	6WBXyRc3ynwvRHiBWudVLo	2.13e-06
28	0.99	['Hafız Yaşar']	0.569	162197	0.327	0	6WFYaPHg3P8jQuh7x9YKYZ	0.947
29	0.991	['Frédéric Chopin', 'Vladimir Horowitz']	0.38	505920	0.119	0	6WRmg6x1bYjHxzeMCfKguB	0.89
30	0.993	['Muzio Clementi', 'Vladimir Horowitz']	0.28	251960	0.0436	0	6WSYfrI77xoUvXLg3WTdLF	0.885
31	0.6859999999999999	['Seweryn Goszczyński']	0.715	102200	0.22	0	6XD6qJykJ91uC9AYb84ZeY	0.0
32	0.995	['Francisco Canaro', 'Charlo']	0.767	177507	0.23	0	6XhXiUZx87slzXDh115DL	0.605

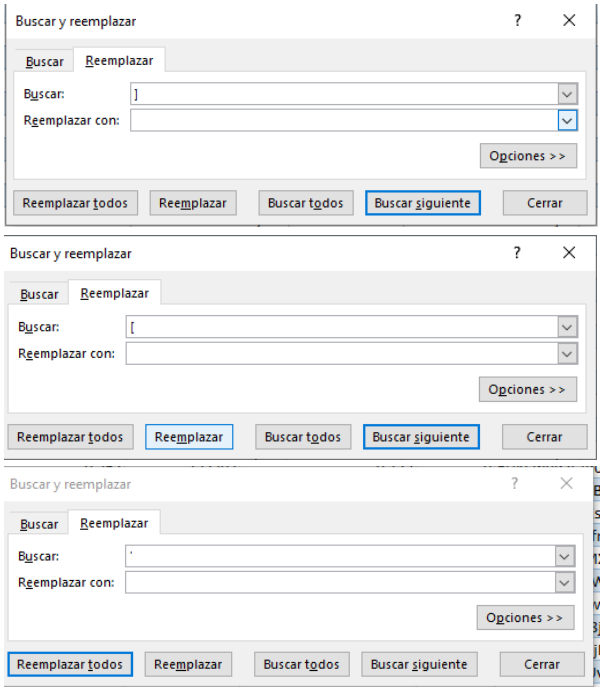
Esto puede tardar unos minutos porque son muchos datos



Por otro lado si miramos la columna artist

artists
['Carl Woitschach']
['Robert Schumann', 'Vladimir Horowitz']
['Seweryn Goszczyński']
['Francisco Canaro']
['Frédéric Chopin', 'Vladimir Horowitz']
['Felix Mendelssohn', 'Vladimir Horowitz']
['Franz Liszt', 'Vladimir Horowitz']
['Carl Woitschach']
['Francisco Canaro', 'Charlo']
['Seweryn Goszczyński']
['Sergei Rachmaninoff', 'Vladimir Horowitz']
['Frédéric Chopin', 'Vladimir Horowitz']
['Samuel Barber', 'Vladimir Horowitz']
['Robert Schumann', 'Vladimir Horowitz']
['Ludwig van Beethoven', 'Staatskapelle Berlin', 'Richard Strauss']
['George Butterworth', 'John Cameron']
['Alexander Scriabin', 'Vladimir Horowitz']
['Francisco Canaro', 'Luis Scalon']
['Thomas Arne', 'John Heddle Nash']
['Francisco Canaro']
['Moritz Moszkowski', 'Vladimir Horowitz']
['Francisco Canaro']
['Frédéric Chopin', 'Vladimir Horowitz']
['Francisco Canaro']
['Francisco Canaro', 'Charlo']

Vamos a borrar los corchetes usando tres veces buscar y reemplazar (Ctrl+B) ,uno para el corchete izquierdo ,otro para el derecho y otro para las comillas, seleccionando la columna de artist, y haciendo click en “Reemplazar todos”



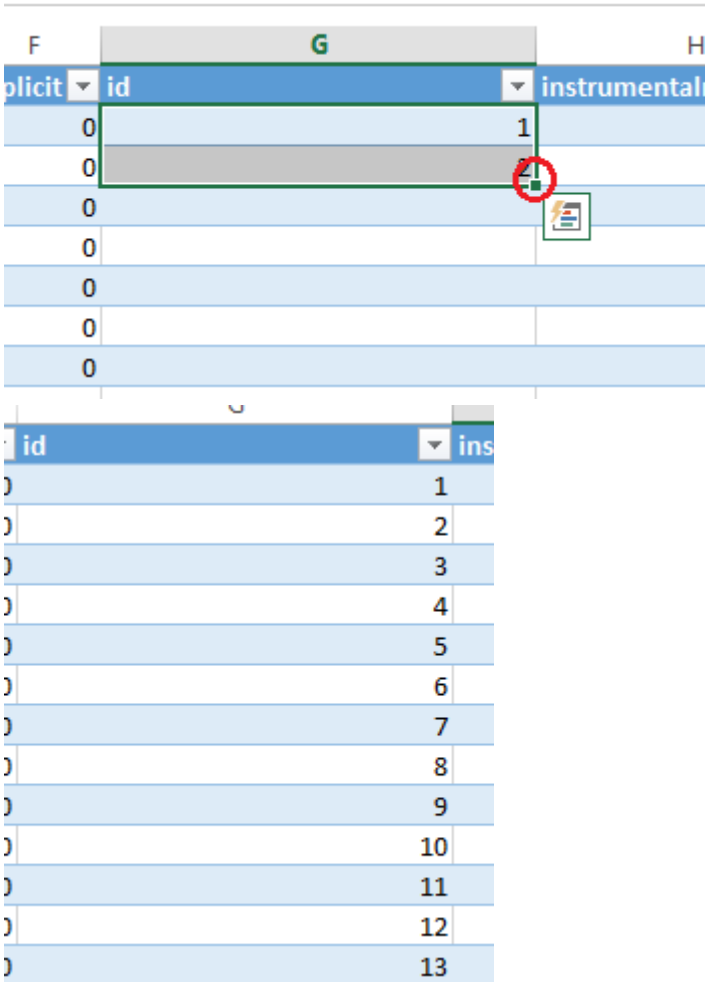
Deberia quedar así

artists
Carl Woitschach
Robert Schumann, Vladimir Horowitz
Seweryn Goszczyński
Francisco Canaro
Frédéric Chopin, Vladimir Horowitz
Felix Mendelssohn, Vladimir Horowitz
Franz Liszt, Vladimir Horowitz
Carl Woitschach
Francisco Canaro, Charlo
Seweryn Goszczyński
Sergei Rachmaninoff, Vladimir Horowitz
Frédéric Chopin, Vladimir Horowitz
Samuel Barber, Vladimir Horowitz
Robert Schumann, Vladimir Horowitz
Ludwig van Beethoven, Staatskapelle Berlin,
George Butterworth, John Cameron
Alexander Scriabin, Vladimir Horowitz
Francisco Canaro, Luis Scalon

Por otro lado la columna id tiene muchos valores que nos no intersa, podemos realizar en numeración común , por lo que borramos los datos

licit	id	instrume
0	6KbQ3uYMLKb5jDxLF7wYDD	
0	6KuQTIu1KoTTkLXKrwlLPV	
0	6L63VW0PibdM1HDSBoqnoM	
0	6M94FkXd15sOAOQYRnWPN8	
0	6N6tiFZ9vLTSOIxkj8qKrd	
0	6NxAf7M8DNHOBTmEd3JSO5	
0	6O0puPuyrxPjDTHDUgsWI7	
0	6OJjveoYwJdIt76y0Pxpxw	
0	6OaJ8Bh7lsBeYoBmwmo2nh	
0	6PrZexNb16cabXR8Q418Xc	
0	6QBInZBkQNIQYU9gGzT5l4	

Una vez que borramos todo (seleccionando los datos y borrando con Supr) , escribimos 1 y 2 , y luego ya podemos arrastrar para todas las filas, solo tenemos que seleccionar los estos dos números y luego hacemos doble click en el cuadrado de la esquina infierior derecha

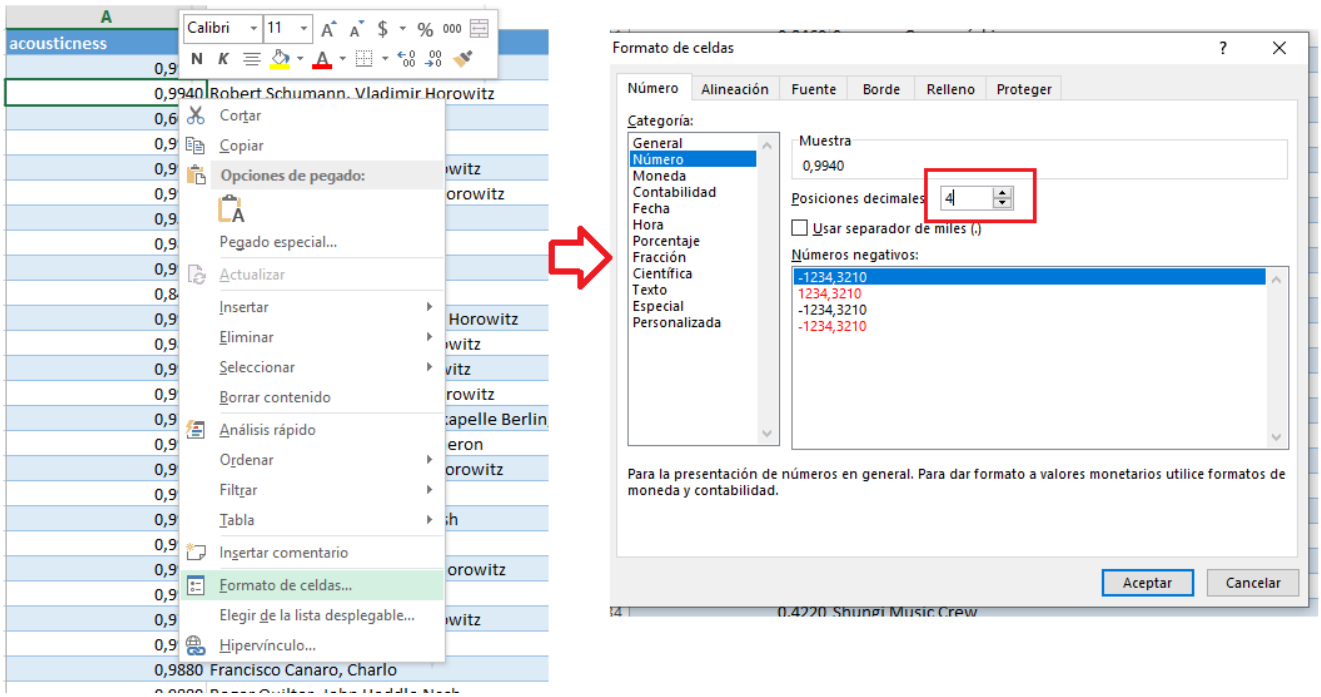


Observamos la siguiente columna

	N	O	P	Q	R	S
	popularity	release_date	speechiness	tempo	valence	year
	0	1928	0,0506	118.469	0,779	1928
	0	1928	0,0462	8.397.200.000.000.000	0,0767	1928
	0	1928	0,929	107.177	0,88	1928
	0	25/9/1928	0,0926	108.003	0,72	1928
	1	1928	0,0424	62.149	0,0693	1928
	0	1928	0,0593	63.521	0,266	1928
	0	1928	0,04	80.495	0,305	1928
	0	1928	0,0474	123,31	0,857	1928
	0	3/10/1928	0,127	119.833	0,493	1928
	0	1928	0,954	81.249	0,759	1928
	0	1928	0,0352	141,39	0,0393	1928
	1	1928	0,0317	8.598.899.999.999.990	0,346	1928
	0	1928	0,0424	96.645	0,042	1928
	0	1928	0,0556	78,98	0,216	1928
	0	1/1/1928	0,0716	80.204	0,406	1928
	0	1928	0,051	79.831	0,169	1928
	0	1928	0,036	66.947	0,0488	1928
	0	16/9/1928	0,157	117.167	0,849	1928
	0	1928	0,0474	76,93	0,596	1928
	0	17/9/1928	0,0886	111.679	0,832	1928
	0	1928	0,0381	74.737	0,661	1928
	0	20/9/1928	0,295	121.779	0,568	1928
	0	1928	0,0316	105.031	0,168	1928
	0	16/9/1928	0,0712	115.212	0,933	1928
	0	20/9/1928	0,115	117.899	0,782	1928

Notamos que release_date tiene la fecha de salida igual que la columna year, pero la misma tiene muchos datos de meses y días que están incompletos, por lo que borraremos dicha columna seleccionandola y apretando “supr”, entonces solo nos quedamos con la columna year

Por útlimo las columnas
Acousticness – danceability – energy – instrumentalness – liveness – speechiness – valence
Modificaremos el formato a Numerico con 4 decimales ya que quiero tener 2 decimales de porcentaje , por ejemplo 0,9876 = 98,76%
Entonces, seleccionamos estás columnas y hacemos click en formato de celda



Dimension Key y mode

Esto es opcional pero si leemos en la descripción de las columnas en la pagina web, notamos que la variable key contiene la tonalidad de la canción, en mi caso personal sé música, por lo que podriamos crear la tonalidad para ser más específico

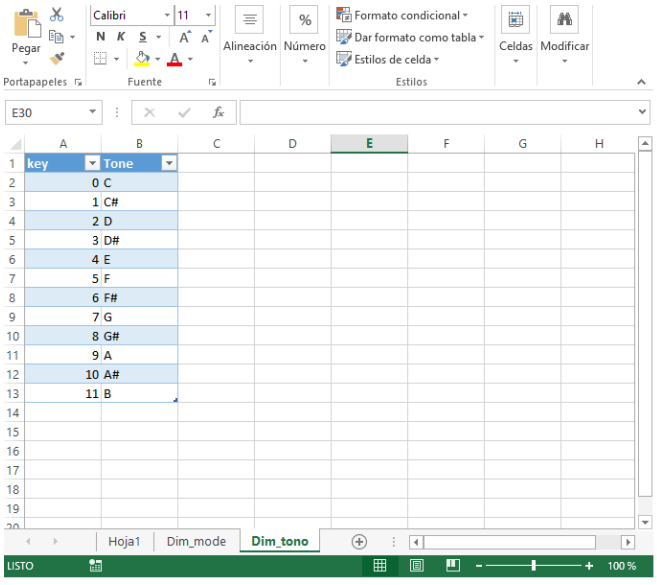
Key	Tonalidad
0	C
1	C#
2	D
3	D#
4	E
5	F
6	F#
7	G
8	G#
9	A
10	A#
11	B

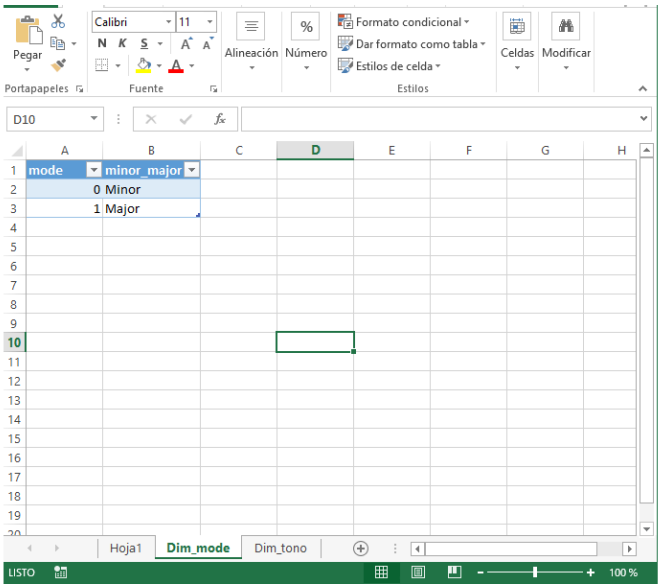
Donde “#” significa una tonalidad sostenida.

Luego en la columna mode es si es una tonalidad mayor o menor

Mode	Menor_mayor
0	Minor
1	Major

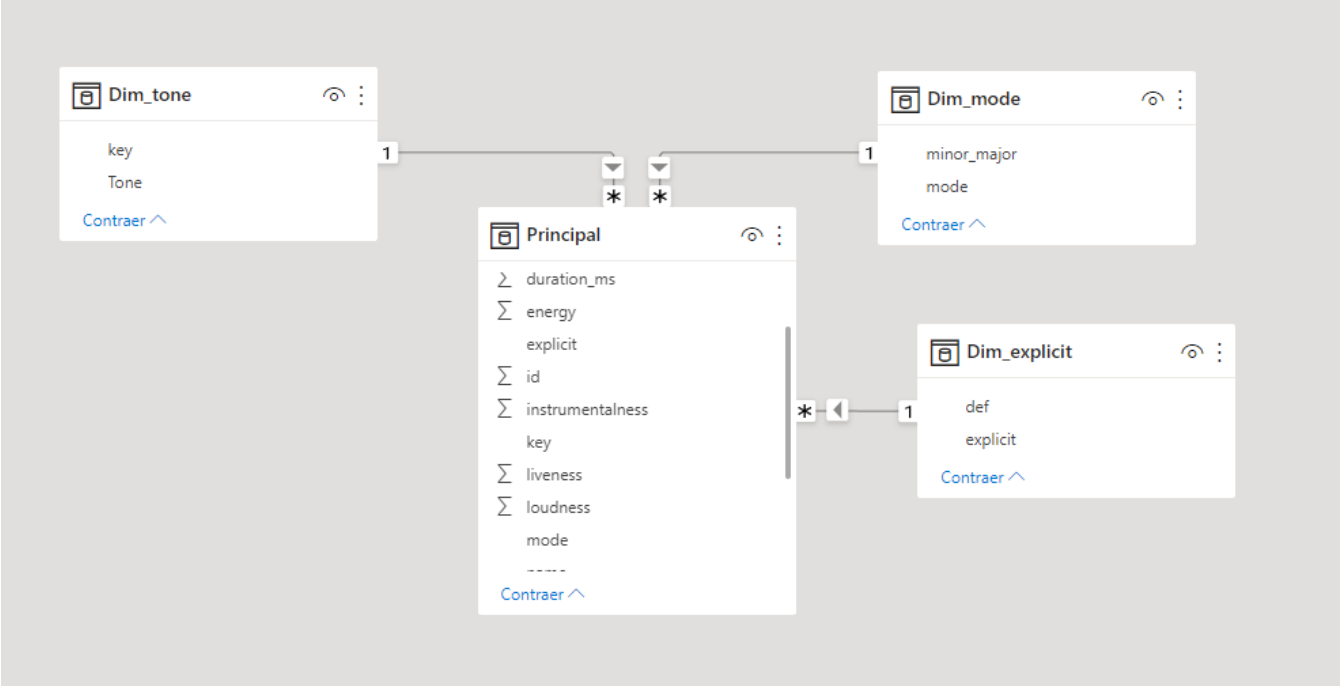
Por ello entonces realizamos dos nuevas hojas de calculo más



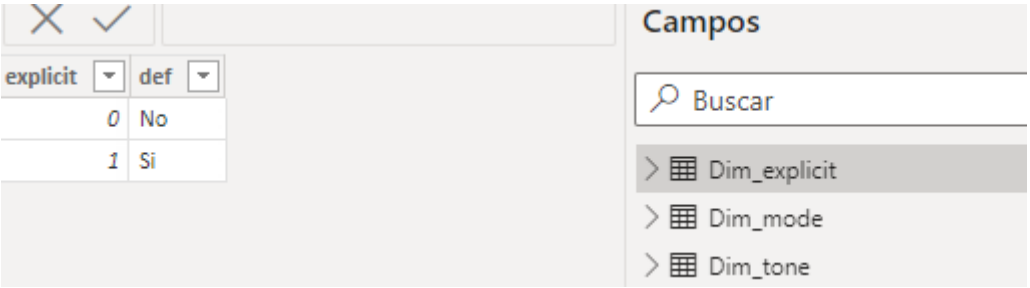


Dashboard (Power BI)

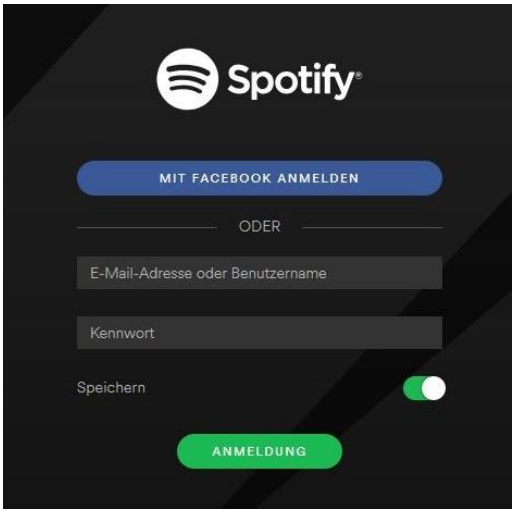
Abrimos el Excel en el power BI y debería crear el siguiente modelo relacional automatico, sino lo realizamos manualmente, con “key” de “Dim_tone” y “key” de “Principal” con cardinalidad uno a varios. Lo mismo con “mode”



Por otro lado noté a última hora que me falta una tabla dimensión de la columna explicit, la podemos crear desde Power BI o desde el Excel .

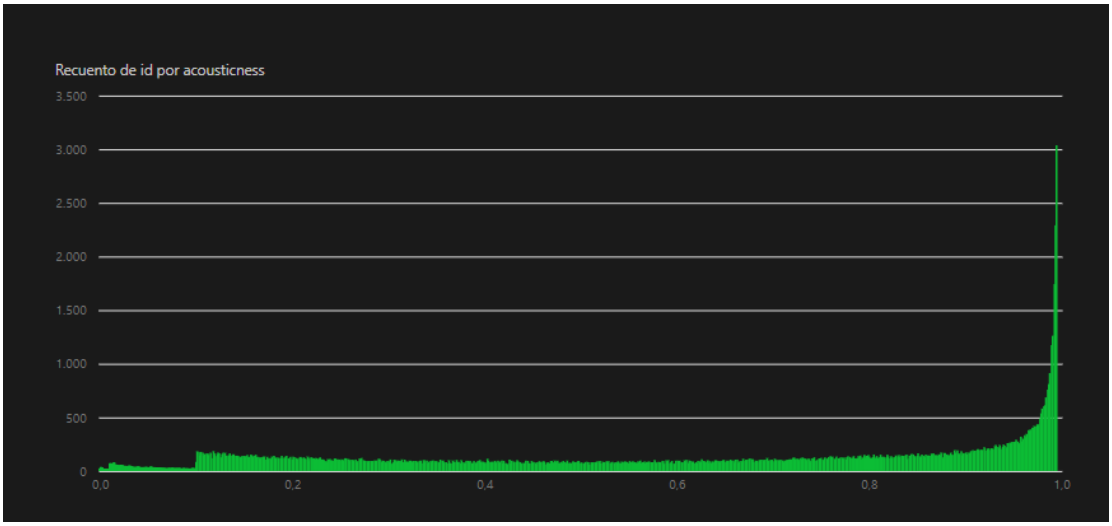


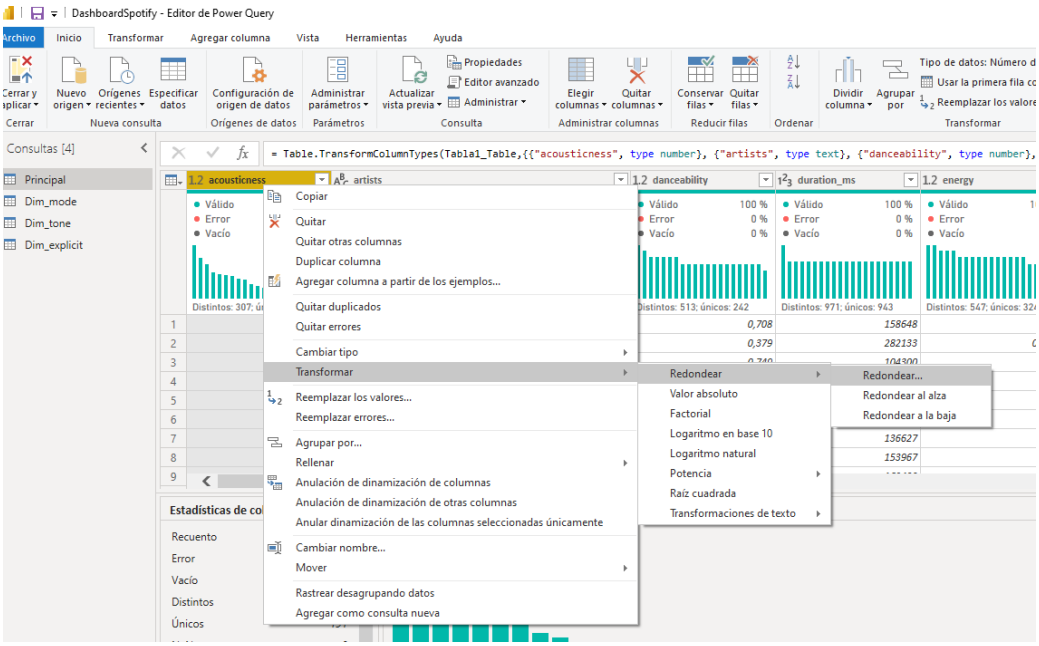
Para crear el dashboard usaré de guía el siguiente boceto web.

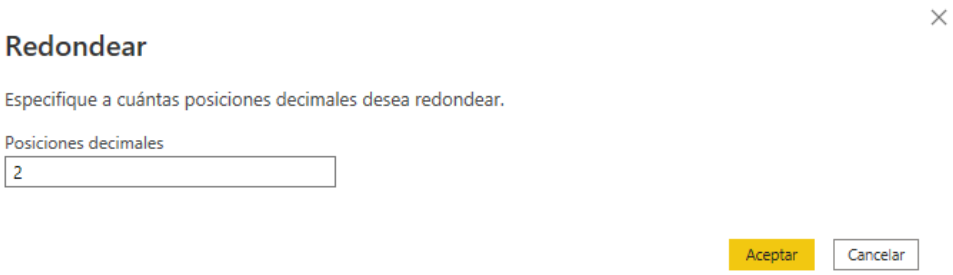


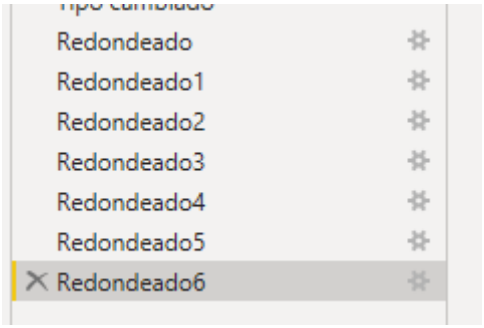
Otra limpieza de datos

Realizando el dashboard, noté uno de los errores en la cantidad de intervalos en las columnas con escala de 0 a 1. Por ejemplo 0,166661 es diferente a 0,166662, creando otra barra más , entonces vamos a redondear los números al segundo decimal. Para limpiar los datos utilizaremos Power Query.



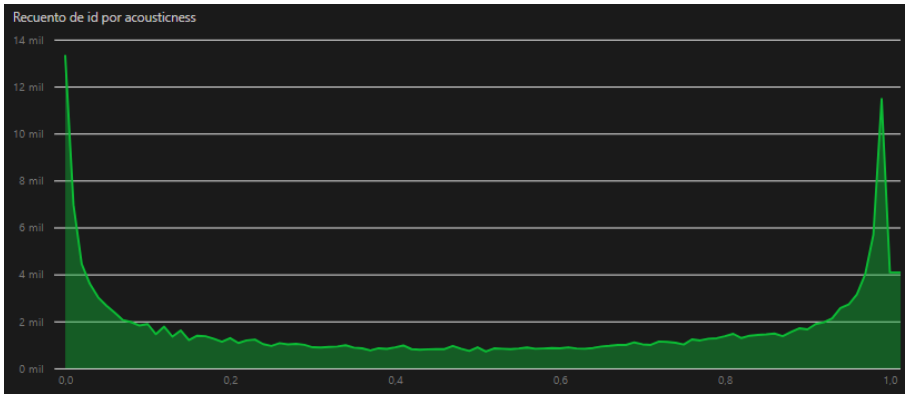






En total 6 redondeos.

Realizado esto, la gráfica ha cambiado a algo sin tantos intervalos



NOTA: Hay muchos valores outliers, o sea que salen del rango del 0 a 1 ,esto se puede filtrar desde las propiedades de los gráficos en Power BI.

Solapas

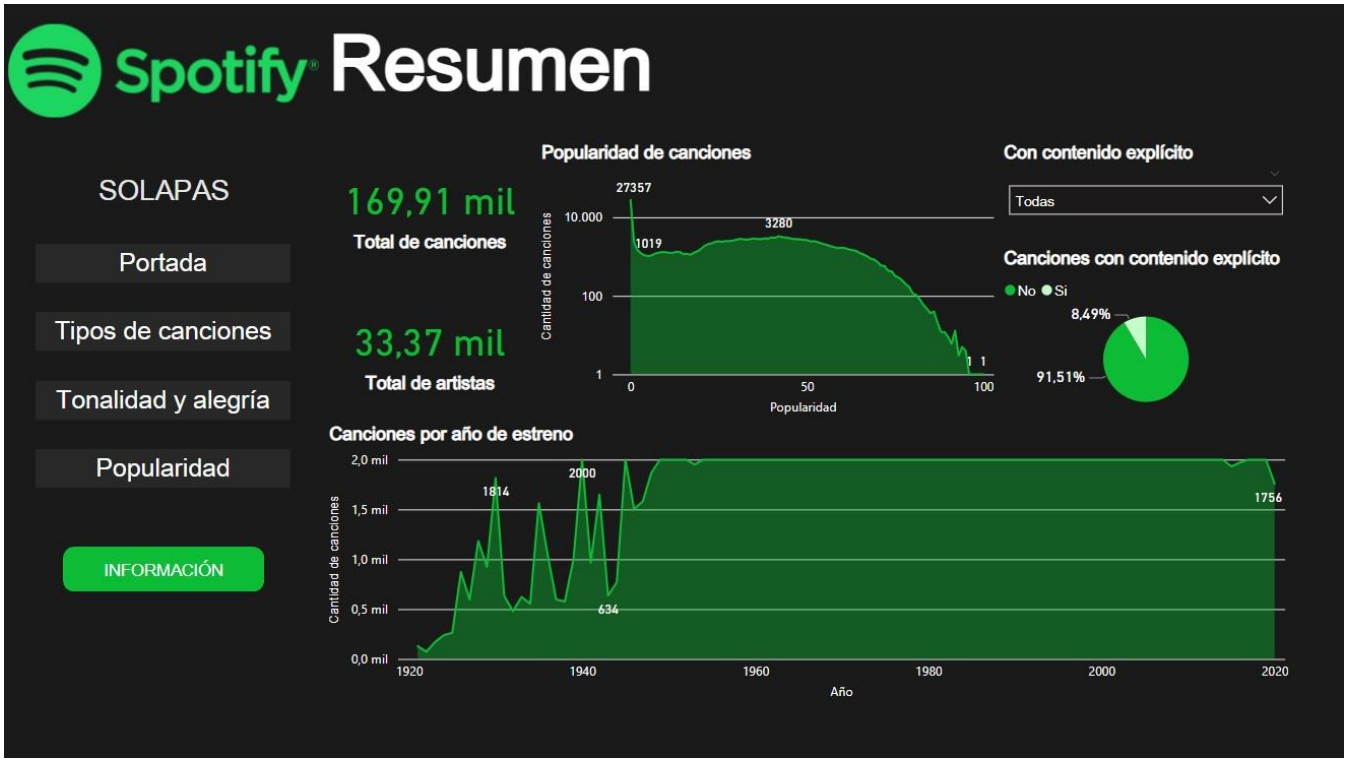
El mismo pueden observarlo desde Power BI para poder mirar con mayor detalle todos los filtros y gráficas.

Portada

Es una portada con un imagen png para el logo de Spotify, con una botonera para moverse por todo el dashboard, tiene info de mi nombre y última actualización

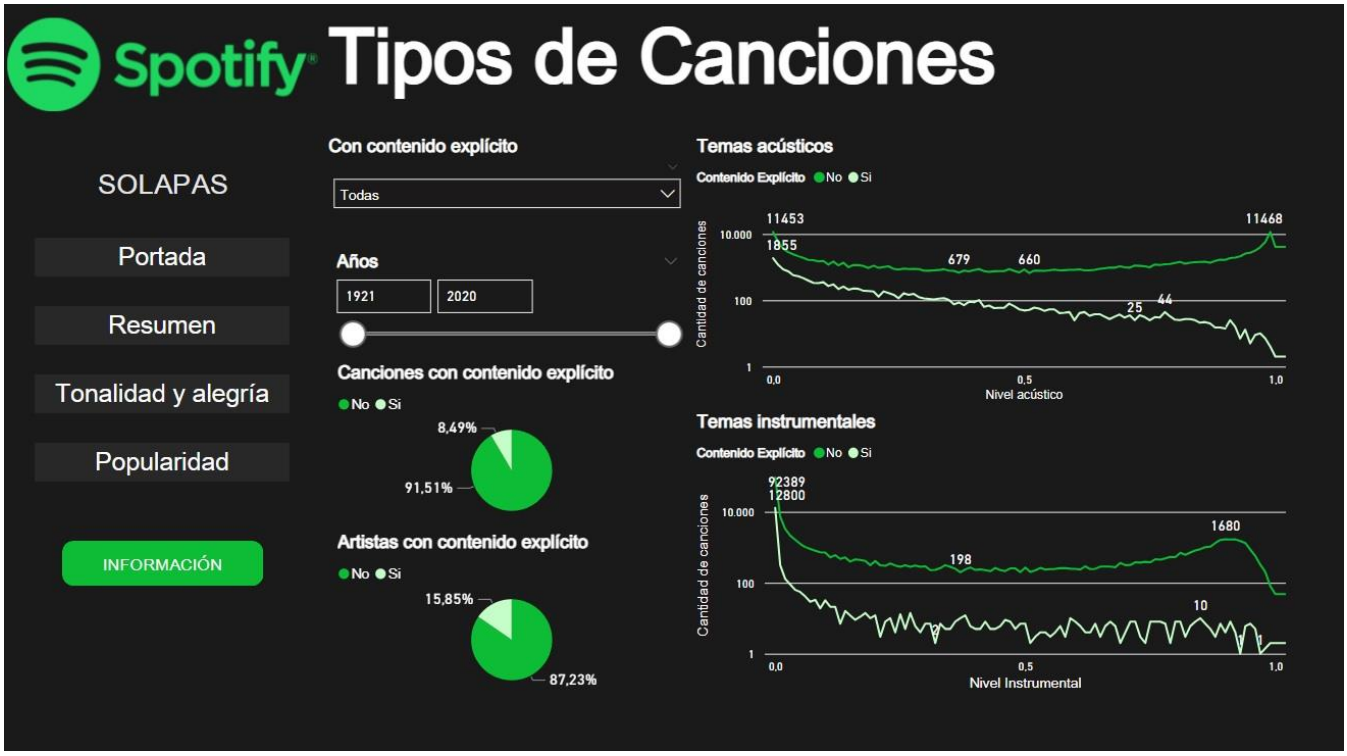
Resumen

Un breve resumen de la cantidad de datos que tenemos en el dataset



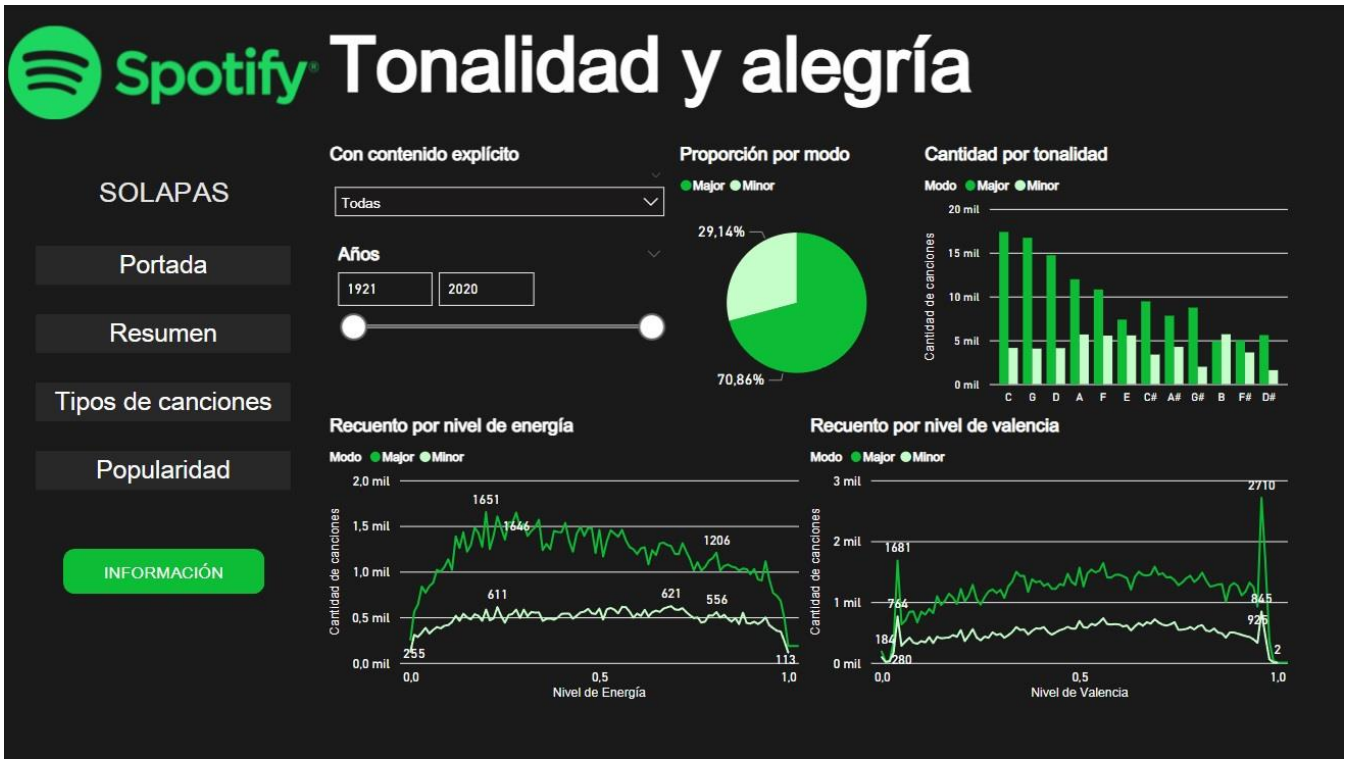
Tipos de canciones

Esta solapa tiene como denotar los temas acústicos y instrumentales en nuestro dataset. Filtrándolo por contenido explícito.



Tonalidad y alegría

Esta solapa muestra como son los niveles de valencia (alegría) y la tonalidad en nomenclatura americana



Popularidad

Este último dashboard tiene como fín , ver como varía la popularidad al pasar de los años

Conclusiones

Realizado este proyecto podemos observar muchos detalles, como por ejemplo la cantidad de contenido explícito fue subiendo con los años, también que sí la canciones es acústica o instrumental lo más probable es que no tenga contenido explícito, lo que más me interesó fue que los niveles de valencia (cualidad que define si una canción es alegre) y de energía (define si la canción es frenética) son altos en tonos mayores y que también existen más canciones en la tonalidad de C (en escala americana es DO), las canciones más populares son las actuales.