Taller 1 grupal

Preparación de datos y análisis exploratorio de canciones y artistas de Spotify.

Joshep Andersson Blanco Reres [a,c] , Laura Estrada Saldarriaga [a,c]

Fabián Camilo Peña[b,c]

a Estudiante de Maestría en Inteligencia Artificial.
b Profesor, Departamento de Ingeniería de Sistemas
c Pontificia Universidad Javeriana, Bogotá, Colombia

Facultad de Ingeniería

Gestión de datos.

Taller 1 – Primer Semestre 2023

Contenido

1. Notebook.

El notebook se alimenta directamente de la base de datos compartida mediante la api de Google drive. Se podrá consultar y ejecutar sin importar el correo Gmail que se use.

https://colab.research.google.com/drive/1l8G72g2qj85NmDgBla0LA9INOlJTW72R?usp=sharing

2. Diccionario de Datos.

Campo	Descripción	Tipo
Acousticness	Medida de confianza de 0.0 a 1.0 de si la pista es acústica. 1.0 representa una alta confianza, la pista es acústica.	Numérica
Danceability	Describe cuán adecuada es una pista para bailar basada en una combinación de elementos musicales que incluyen tempo, estabilidad de ritmo, fuerza de ritmo y regularidad general. Un valor de 0.0 es menos bailable y 1.0 es más bailable.	Numérica
Duration_ms	la duración de la pista en milisegundos.	Numérica
Duration_min	numérica, la duración de la pista en minutos.	Numérica
Energy	La energía es una medida de 0.0 a 1.0 y representa una medida perceptiva de	Numérica

	intensidad y actividad. Por lo general, las pistas enérgicas se sienten rápidas, fuertes y ruidosas. Por ejemplo, el death metal tiene alta energía, mientras que un preludio de Bach obtiene un puntaje bajo en la escala. Las características perceptivas que contribuyen a este atributo incluyen el rango dinámico, el volumen percibido, el timbre, la tasa de inicio y la entropía general.	
Explicit	Ya sea que la pista tenga o no letras explícitas (true = sí lo hace; falso = no, no lo hace o desconocido).	categórico
Id	la identificación de Spotify para la pista.	Ordinal
Instrumentalness	Predice si una pista no contiene voces. Los sonidos "OOH" y "AAH" se tratan como instrumentales en este contexto. Las pistas de rap o palabras habladas son claramente "vocales". Cuanto más cerca sea el valor de instrumentalidad a 1.0, mayor es la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0.5 están destinados a representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1.0.	Numérica
Key	La clave general estimada de la pista. Integers mapean los lanzamientos utilizando notación de clase de tono estándar. P.ej. 0 = C, 1 = C#/db, 2 = D, y así sucesivamente. Si no se detectó ninguna clave, el valor es -1.	Numérica
Liveness	numérica, detecta la presencia de una audiencia en la grabación. Los valores de vida más altos representan una mayor probabilidad de que la pista se haya realizado en vivo. Un valor superior a 0.8 proporciona una fuerte probabilidad de que la pista esté en vivo.	Numérica
Loudness	Overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.	Numérica
Mode	El modo, indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico.	Numérica

	Mayor está representado por 1 y menor es 0.	
Popularity	Lla popularidad de una pista es un valor entre 0 y 100, siendo 100 el más popular. La popularidad se calcula por el algoritmo y se basa, en su mayor parte, en el número total de obras de teatro que ha tenido la pista y cómo son las obras recientes.	Numérica
Release_date	fecha La canción fue lanzada	Fecha
Speechiness	El habla detecta la presencia de palabras habladas en una pista. Cuanto más exclusivamente hablan en la grabación (por ejemplo, programa de entrevistas, audiolibro, poesía), más cerca de 1.0 el valor del atributo. Los valores superiores a 0.66 describen pistas que probablemente se hagan completamente de palabras habladas. Los valores entre 0.33 y 0.66 describen pistas que pueden contener música y discurso, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores por debajo de 0.33 probablemente representan música y otras pistas sin voz.	Numérico
Tempo	Estimado general de una pista en ritmos por minuto (BPM). En la terminología musical, el tempo es la velocidad o el ritmo de una pieza dada y deriva directamente de la duración promedio del latido.	Numérica
Valence	Mide de 0.0 a 1.0 que describe la positividad musical transmitida por una pista. Las pistas con alta valencia suenan más positivas (por ejemplo, felices, alegres, eufóricas), mientras que las pistas con baja valencia suenan más negativas (por ejemplo, triste, deprimida, enojada).	Numérica
Year	año se lanzó la canción	Fecha.

Segundo dataset. Artistas

id	Id Spotify	Numérica
followers	Número de seguidores del artista	Numérica
genres	Géneros cantados por el artista	Fecha.
name	Nombre del artista	Objeto.
popularity	Índice de popularidad del artista.	Numérica

3. Análisis de calidad:

Columnas que pasaron por el proceso de limpieza de datos:

Base artists_mod: artists object, id_artists object, release_date object
tracks_mod: genres object, name object

Completitud

El principal error de la base de datos es completitud, se encontraron un total de 38.606 registros con al menos un valor perdido en su registro, esto impacta el 40% de los registros.

Precisión

Las bases de datos no incluían el género por canción, a pesar de que un artista pueda identificarse con un género común, hay artistas versátiles que fluctúan entre géneros, la información deja de ser precisa. Indica si los datos no logran representar correctamente las características del género en cuanto al consumo de música por parte de los oyentes.

Disponibilidad

La base de datos es estática y no puede ser accedida para determinar información en tiempo real, los datos resultarán obsoletos a largo plazo.

Conformidad

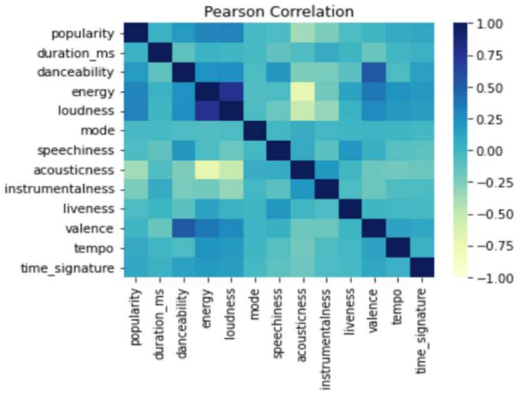
Hay datos en campos de fecha como release_date en la que el formato de fecha varía, hay datos que solo incluían el año por ejemplo y fueron completado en el proceso de limpieza de los datos, eliminando campos como "Year 2022". Los datos no siguen un conjunto de reglas explícitas o estándares para la captura y publicación, lo que dificulta el análisis y requiere un proceso previo de calidad.

Aparentemente no hay indicadores de falencias en consistencia.

4. Análisis exploratorio de datos (EDA):

El índice de popularidad depende de la información actual, es decir, una canción de los años 1980 pudo haber sido muy popular en su momento y la data que sustentaría su éxito sería la de discos vendidos, pero Spotify carece de esta información por lo cual no refleja con precisión estas canciones. Cruzar la información con listas billboard, puede arrojar data interesante y nos ayudaría a entender mejor los precedentes de la creciente popularidad de música latina y predecir futuras tendencias en el mercado de la música, y predecir el siguiente "Despacito" de Luis Fonsi.

Pearson entre canciones más populares:



Análisis de variables:

- Las canciones más populares son aquellas que recaen en la categoría: energy y loudness.
- La categoría acousticness no se relaciona con las categorías: energy and loudness. La música acústica carece de actividad y actividad (energy), además de no tener altos decibeles durante la canción (loudness).
- Las canciones más bailables (danceability) son altamente relacionadas con ser positivas (valence).

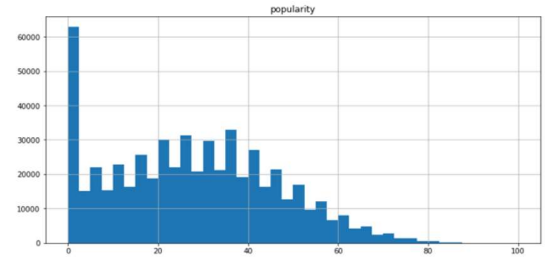
Análisis exploratorio de datos (EDA) por canciones.

Pontificia Universidad JAVERIANA
Bogotá

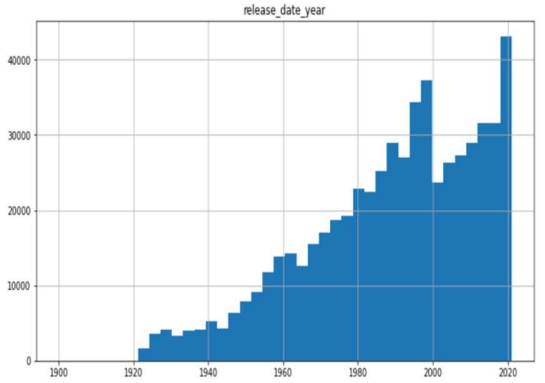
Facultad de Ingeniería

Gestión de datos.

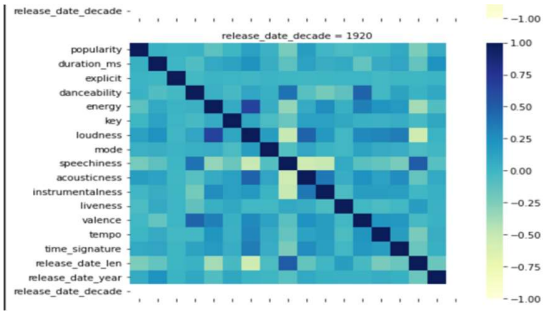
Taller 1 – Primer Semestre 2023



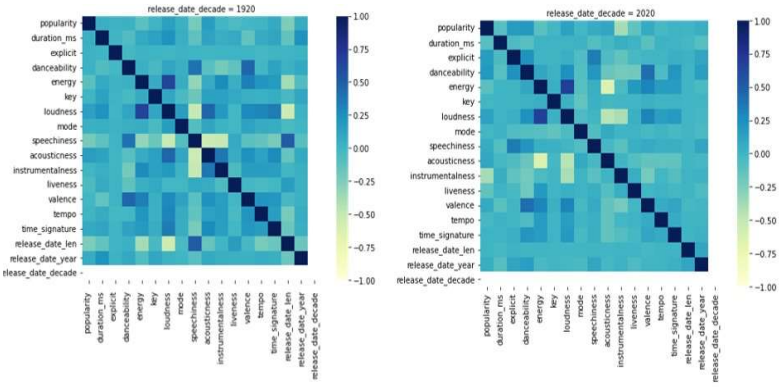
1. La gran mayoría de las canciones no alcanzan popularidad, las más populares se encuentran en la parte más baja de la gráfica. La mayor probabilidad es que una canción no alcance popularidad dentro de los usuarios de la plataforma.



2. Es notable el crecimiento de canciones lanzadas con los años, es común que el inicio de la década sea ostensiblemente menor que el final. El pico de lanzamientos durante 2020 puede haber sido fuertemente impactado por la aparición del COVID, siendo el ingreso por reproducciones el único adquirido por los artistas. Las restricciones para eventos masivos afectaron en gran medida la difusión de música en vivo.



3. Las canciones más populares de todas las décadas se caracterizan para tener una alta relación con las categorías de energy and loudness (revisar workbook .)



4. Se pueden apreciar aspectos interesantes y correlaciones que han aparecido o desaparecido a lo largo de los años.
Canciones con bastante letra también eran bailables (años 1920 a 1950 fue así) sin embargo empieza a

desaparecer esa tendencia en los años posteriores.
Adicionalmente, las canciones con más letras se relacionan en 2020 con unas canciones más explícitas, algo que no pasaba en el siglo anterior.

Análisis exploratorio de datos (EDA) por artistas.

1. *Crecimiento de la música latina y el kpop*

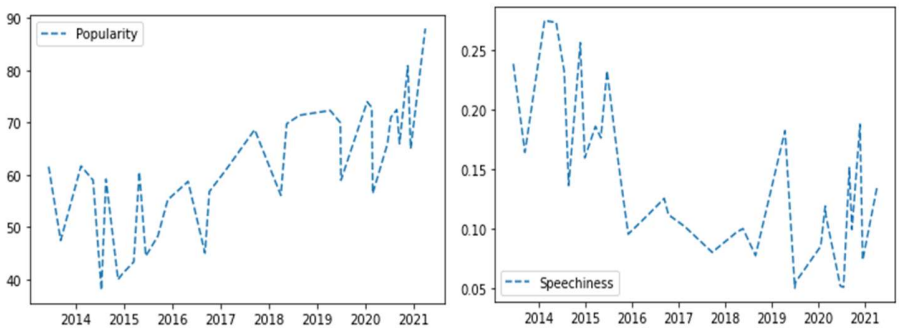
genres	name	popularity
['canadian pop', 'pop', 'post-teen pop']	Justin Bieber	100
['pop', 'post-teen pop']	Taylor Swift	98
['canadian hip hop', 'canadian pop', 'hip hop'...	Drake	98
['latin', 'reggaeton', 'trap latino']	Bad Bunny	98
['k-pop', 'k-pop boy group']	BTS	96
canadian contemporary r&b', 'canadian pop', ...	The Weeknd	96
['chicago rap', 'melodic rap']	Juice WRLD	96
['trap latino']	Myke Towers	95
['dance pop', 'pop', 'uk pop']	Dua Lipa	95
['latin', 'reggaeton', 'reggaeton colombiano',...	J Balvin	95

El crecimiento de la música latina, junto con BTS es muy importante a nivel global.
Aun así, se destacan artistas de la "ola anterior" como Justin Bieber y Taylor Swift que son líderes en popularidad.

2. *¿Qué diferencia a BTS?*

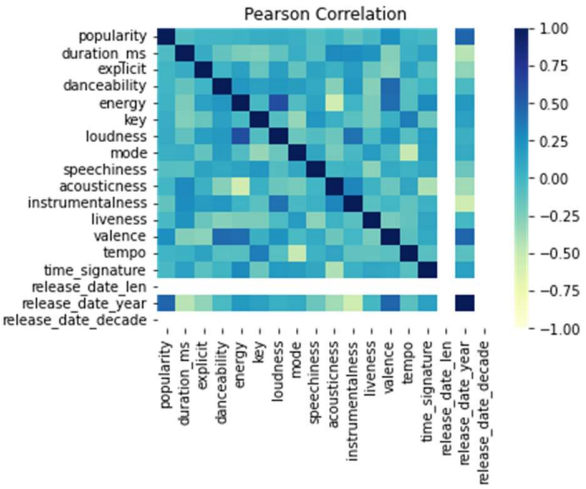
BTS particularmente destaca en canciones que **no son explicitas**, con una alta duración, un sonido muy enérgico (**energic**) y veloz (**tempo**).

Es uno de los grupos con presencia de público en sus canciones más pop (**liveness-speechiness**) , algo que rompe la tendencia estándar de ésta década. (**Ver comparativo entre décadas.)**



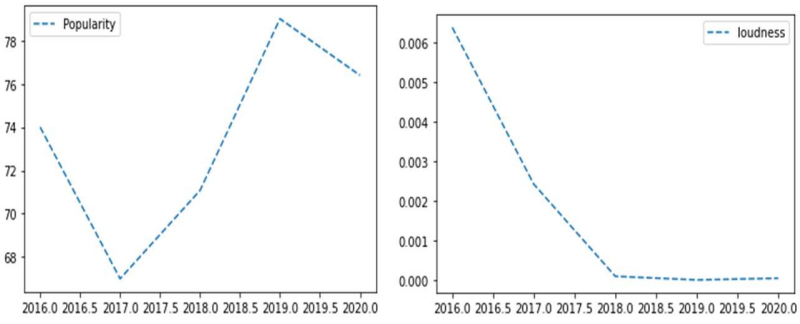
Si vemos la popularidad del artista, estaban en su pico más alto a nivel internacional, su popularidad coincide con su etapa más pop de su carrera, al ser un grupo que empezó un teniendo mucho rap.

3. ¿Qué diferencia a Bad Bunny?



Bad Bunny tiene sonidos fuertes (Alto loudness) combinadas con sonidos instrumentales, es decir con alta presencia de coros.

Tienen en común con BTS que son canciones para bailar, pero no cuentan con una fuerte presencia de público en sus canciones, ni tienen tanta duración (muy apegados al estándar musical).



Si vemos la popularidad del artista, estaba en su pico más alto en 2020, su popularidad coincide con su etapa de mayor trap urbano del artista, que cambió el reggaetón enérgico de sus predecesores, por uno más oscuro y calmado, como lo muestra su último álbum.

Pontificia Universidad JAVERIANA
Bogotá

Facultad de Ingeniería

Gestión de datos.

Taller 1 – Primer Semestre 2023