Curso: Data Science
Comisión: 39960
Fecha de entrega: 3 de mayo de 2023
Nombre: Catalina Miranda

Desafío N°9: Data Wrangling

ABSTRACT

La temática que se desarrollará en el presente proyecto final se relaciona con el mundo de la música, más específicamente, con el ranking de la mundialmente reconocida revista Billboard. Desde el año 1958, cada semana, la institución da a conocer los "Hot 100", es decir, las 100 canciones más aclamadas del momento. La clasificación no solo está basada en las ventas físicas y digitales de los Estados Unidos, sino también en las reproducciones por radio y por plataformas de *streaming* a nivel global.

La base de datos seleccionada agrupa información relacionada con las canciones y los artistas que lograron ingresar al Hot 100 del ranking de la revista desde el año 1958 hasta el 2021. Por cada semana, se muestran la posición y los títulos de las 100 canciones de la clasificación junto con sus correspondientes los cantautores o grupos musicales. Además, el conjunto de datos incluye información técnica de las canciones presentes en el ranking, como la valencia, el tempo, el modo y la acusticidad, entre otros. Mediante la utilización de APIs se incluirán los rankings del año 2022.

A partir de la información histórica, se identificarán los artistas con mayor trayectoria y las canciones más exitosas. Por otro lado, en base a los datos técnicos de las canciones del ranking, se podrá establecer si existen patrones musicales que hacen que una canción sea más escuchada que otra. En un contexto comercial, dicha información resulta de interés para los productores musicales y compañías discográficas, porque les permite conocer las características que deberían tener las nuevas melodías para convertirse en "*hits*" mundiales.

PROBLEMA Y CONTEXTO COMERCIAL

Una discográfica está teniendo problemas con los artistas que contrata. Las canciones que producen no llegan a posicionarse dentro de las canciones más escuchadas, y las recaudaciones de la compañía, junto con su reputación, están cayendo drásticamente. Mientras tanto, otras empresas del ámbito musical están creciendo notablemente. Por tal motivo, la discográfica desea conocer cuáles son las características de las canciones más exitosas, para poder predecir cuáles van a ser parte del "Hot 100" y cuáles no. El objetivo principal es implementar un modelo de Machine Learning que indique las probabilidades de que la canción se posicione entre las mejores del mundo.

ANALISIS EXPLORATORIO DE DATOS (EDA)

En esta sección se realizará el análisis exploratorio de datos, conocido por sus siglas en inglés como EDA. Este proceso tiene como objetivo examinar la base de datos seleccionada para encontrar patrones, descubrir anomalías y probar hipótesis usando medidas estadísticas.

A la hora de llevar a cabo el EDA, se siguieron los siguientes pasos recomendados por el libro *"Hands-on exploratory data analysis with Python"* de Suresh Kumar Mukhiya y Usman Ahmed.

1. Definición del problema;
2. Recolección de datos;
3. Carga de datos en el formato deseado;
4. Procesamiento de datos;
5. Limpieza de datos;
6. Análisis de datos. Es el EDA propiamente dicho, que incluye la aplicación de conceptos de estadística descriptiva para conocer más a fondo los datos.

A continuación, se detallará cada una de las etapas mencionadas previamente.

1. DEFINICIÓN DEL PROBLEMA

Teniendo en cuenta el contenido de la base de datos seleccionada, se formularon las siguientes preguntas de interés, que se abordarán a lo largo del desarrollo del proyecto.

- ¿Las canciones más exitosas, es decir, aquellas que ocuparon mejores posiciones durante más semanas en el ranking, tienen características en común? De ser así, ¿esas características fueron cambiando con el paso de los años?

- ¿Existe un género predominante dentro del top 10 del ranking? ¿Existe alguna relación entre el género musical y las características de las canciones?

Las respuestas a dichas preguntas permitirán conocer las variables claves a incluir en el modelo.

2. RECOLECCIÓN DE DATOS

El conjunto de datos está compuesto por dos archivos en formato csv: el primero llamado "Top 100" que contiene 12 columnas y 327896 filas, y el segundo denominado "Audio" que posee 19 columnas y 29504 registros. Ambos archivos están relacionados por el campo SongID, que permite identificar de manera unívoca a cada una de las canciones que aparece en el set de datos.

A continuación, se especifican los atributos de cada archivo.

"Top 100" indica las canciones que ingresaron al top 100 del Billboard Chart entre agosto de 1958 y mayo de 2021. Los campos son los siguientes:

1. Url: link que permite ingresar al sitio web de Billboard y visualizar el ranking de la semana correspondiente.
2. Week position: posición de la canción en el top 100 de la semana correspondiente
3. Song: nombre de la canción
4. Performer: cantante/grupo musical
5. SongID: clave primara que relaciona las tablas
6. Instance: indica la cantidad de veces que ha aparecido la canción en el top 100
7. Previous week position: indica la posición en el ranking la semana anterior a la analizada
8. Peak position: indica la máxima posición alcanzada por la canción en el ranking.
9. Weeks on chart: indica la cantidad de semanas de la canción en el ranking.
10. Day: día de la semana en la que se publicó el ranking
11. Month: mes en el que se publicó el ranking
12. Year: año correspondiente al que se publicó el ranking

"Audio" especifica los atributos musicales de las canciones que ingresaron al ranking.

1. SongID: clave primara que relaciona las tablas
2. Spotify_genre: género según la clasificación de Spotify
3. Spotify_track_ID: código que identifica de forma unívoca a la canción en Spotify
4. Spotify_track_preview_url: link que permite acceder a 30 segundos de la canción.
5. Spotify_track_duration_ms: duración de la canción en milisegundos.
6. Spotify_track_explicit: "Explicit" es una clasificación de Spotify que significa que la canción usa lenguaje grosero, violento u ofensivo. Se indica como verdadero o falso.

7. Spotify_track_album: álbum al que pertenece la canción
8. Danceability: describe qué tan bailable es la canción, en función de elementos como el tempo y la estabilidad del ritmo. 0 significa que es muy poco bailable y 1 significa que es 100% bailable.
9. Energy: es una medida que varía entre 0 y 1 y que representa un porcentaje de intensidad de actividad. Normalmente, una canción enérgica se suele rápida, fuerte y ruidosa.
10. Key: es el tono, las notas o la escala de la canción que forma la base. Las 12 claves varían entre 0 y 11.
11. Loudness: es la calidad de la canción, desde -60 a 0 dB promediada a lo largo de toda su duración. Cuanto mayor es el valor, más alta es la canción.
12. Mode: las canciones se clasifican en menores (0) o mayores (1).
13. Speechiness: detecta la presencia de palabras en la canción. 1 significa que el 100% del tiempo se dicen palabras.
14. Acousticness: o acusticidad, que es la propiedad de lo que suena agradablemente, 1 significa que la canción es acústica, 0 significa que no es nada acústica.
15. Instrumentalness: indica si existe contenido vocal. 1 significa que no hay contenido vocal. La saqué
16. Liveness: detecta la presencia de audiencia durante el grabado de la canción. Cuanto más cerca de uno se encuentra el valor, mayor probabilidad de que se haya grabado en vivo.
17. Valence: es una medida que varía entre 0 y 1, y que describe la positividad musical transmitida por una pista. Las pistas con una valencia alta suenan más positivas (felices, alegres, eufóricas, etc), mientras que las pistas con una valencia baja suenan más negativas (tristes, deprimidas, enfadadas, etc).
18. Tempo: es la velocidad de la canción y se mide en BMP (*Beats por minuto*).
19. Spotify_track_popularity: mide la popularidad de la canción en Spotify. Los valores varían entre 0 y 100, indicando qué tan popular es el artista en relación a todos los demás artistas de la plataforma.

Link de descarga del conjunto de datos seleccionado:

https://www.kaggle.com/datasets/heemalichaudhari/billboard-hot-weekly-charts

Al conjunto de datos seleccionado se le anexará información descargada desde APIs.

En primer lugar, se descargarán los Billboard charts faltantes del año 2021 y se agregará al análisis el año 2022 completo. Para ello, se usó el API billboard.py, disponible en Python para acceder a los datos del ranking Hot100. Luego de su instalación se especificaron las semanas del ranking deseadas. Dado que se trata de una cantidad interesante de semanas, se generó una serie con la fecha inicial

(mayo de 2021), y cada una de las fechas posteriores se calcularon sumando 7 días. Los datos obtenidos se agregaron a los archivos planos, para luego poder ser anexados al modelo relacional planteado para el presente proyecto final.

Por otro lado, al extender la cantidad de cuadros analizados, se encontraron nuevas canciones que no poseen datos de sus atributos musicales. Para completar la tabla Audio, y minimizar la cantidad de datos ausentes, los atributos de las canciones se buscaron mediante el API web de Spotify, disponible para todos sus usuarios. Este proceso también fue útil para completar los datos que originalmente faltaban en el archivo Audio. A partir del nombre de la canción y del artista es posible identificar el URI o *track*_id, que son identificadores unívocos de la canción y que permiten extraer los atributos buscados. El URI contiene al track_id, por lo cual se decide trabajar con él.

3. OBTENCIÓN DE DATOS EN PYTHON

Los archivos fueron cargados en Google Colab como Pandas Dataframes, y se identificaron con el mismo nombre de los archivos planos (Top100 y Audio). Los datos descargados de APIs se obtuvieron como archivos planos, llamados downloads_h100 y Audio2, que luego se cargaron en Python.

4. PROCESAMIENTO DE DATOS

Incluye las etapas de estructuración y ordenamiento de los datos para que su posterior análisis sea más sencillo.

Dado que no todas las columnas de los dataframes serán de utilidad para el análisis, aquellas que no aportaban información relevante o que no pudieran ser correctamente interpretadas fueron eliminadas.

En el caso del DataFrames Top100, se eliminaron los atributos "*Instance*", porque no se encontró explicado claramente el criterio a la hora de calcularlo, y "URL", porque no tiene utilidad durante el análisis.

Del DataFrames "Audio" se eliminaron cinco columnas. En primer lugar, se suprimió *"Spotify_track_id*", código que identifica a la canción en Spotify, pero que carece de utilidad porque ya se cuenta con un código identificador (SongID). La segunda eliminada fue *"Spotify_track_preview_url*", un link que permite escuchar los primeros 30 segundos de la canción y que no aporta ningún dato relevante en el tipo de análisis que se quiere realizar. En tercer lugar, se eliminó el campo "Instrumentalness" porque provee la misma información que el campo *"Speechiness*", y se optó por conservar este segundo. Luego se eliminó la columna *"Liveness*" que no

se considera de interés para el análisis a realizar y finalmente, se quitó el atributo *"Spotify_track_popularity"* porque no tiene relación con el Billboard Chart.

La información contenida en los DataFrames obtenidos de APIs se concatenó a los DataFrames de la base de datos. Audio2 se agregó a Audio, y h100 a Top100.

Luego de este procedimiento se obtuvieron dos DataFrames, uno con 10 columnas y el otro con 14.

Si bien luego de eliminar los campos inservibles se obtienen tablas mucho más útiles, es necesario ordenar los datos aún más. En base a los datos disponibles, se propone un esquema relacional como el de la figura 1.

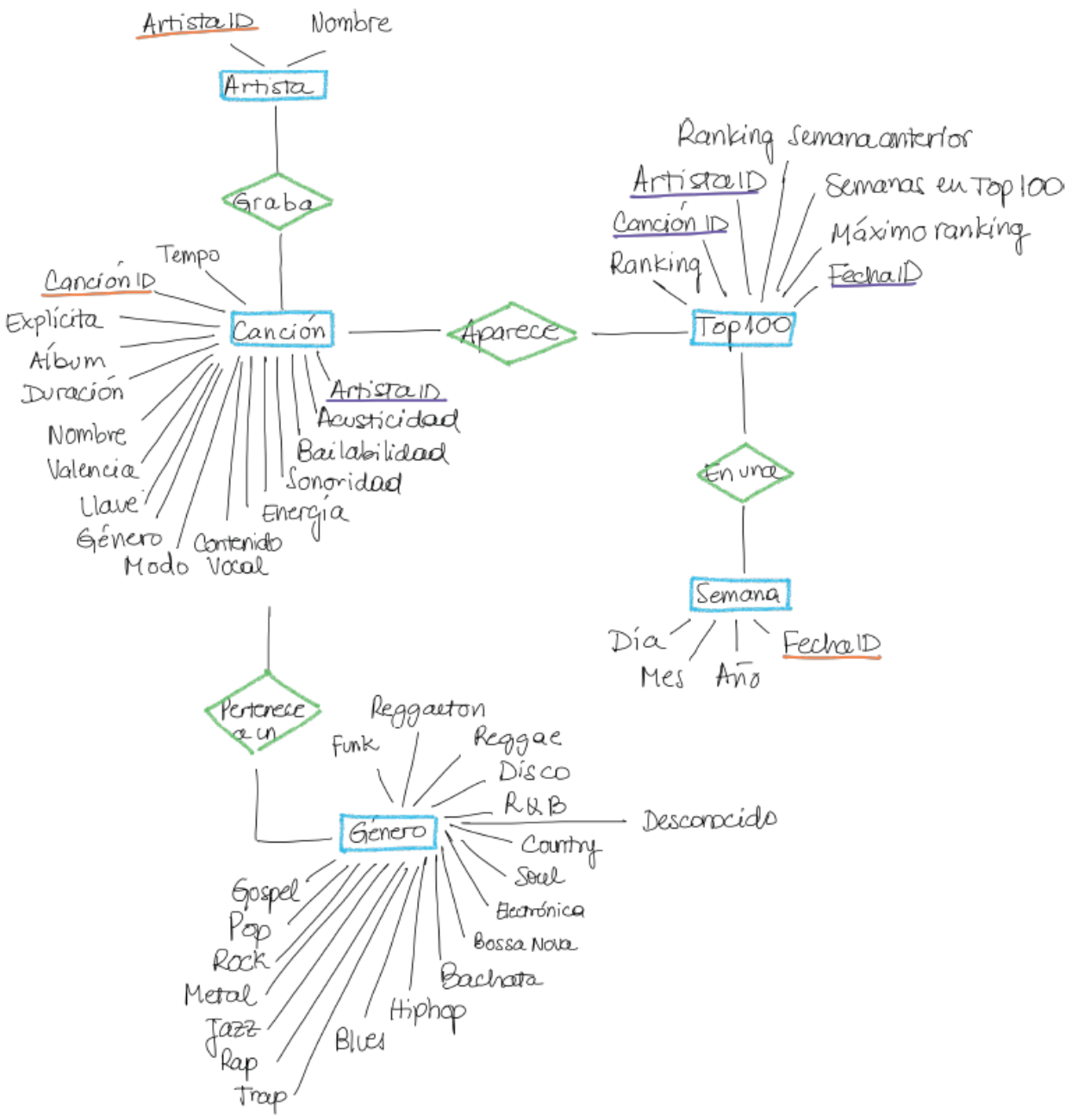

Figura 1. Diagrama de entidad relación de la base de datos seleccionada

Se decidió reordenar a todos los atributos en cinco tablas diferentes. Los campos que están subrayados en naranja son claves primarias, mientras que los subrayados en violeta son claves foráneas.

La primera se llama Artista, y contiene la información relevante del cantante, como su nombre y su identificador primario "Artista ID", que permite relacionarlo con las demás tablas. Además, esta tabla puede usarse durante el análisis de datos para almacenar información importante como la cantidad de

canciones que alguna vez obtuvo en el ranking, su máximo ranking, su posición más recurrente (en la que estuvo más veces), la cantidad de apariciones en el ranking, la cantidad de semanas consecutivas, la cantidad de semanas en el ranking, la cantidad de semanas en el ranking con la misma canción, la cantidad de álbumes, la fecha de su primera y última aparición en el ranking, etc. De esta manera, la tabla Artista es una especie de "perfil" del cantante, donde se ven resumidos sus logros principales.

La tabla Canción especifica todos los atributos técnicos de las canciones que aparecen en el "Hot100". Posee como clave primaria "Canción ID" y como clave foránea a "Artista ID".

La tabla Top100 resume todas las apariciones en el ranking, junto con otros indicadores específicos, como el ranking de la semana inmediatamente anterior y el máximo ranking alcanzado por la canción. Posee como claves foráneas a "Canción ID", "Artista ID" y "Fecha ID", que es la clave primaria de la cuarta y última tabla.

La tabla Fecha posee campos que permiten identificar a la semana del ranking. Contiene cinco campos. Los tres del día, mes y año, del archivo Top100, y dos nuevos: la semanaID, que la relaciona con la tabla Top100, y la fecha, que es una concatenación del día, mes y año, y que cobra importancia a la hora de visualizar los datos. Al separar la información de la semana estudiada en un nuevo dataframe, se observa claramente que en total se cuenta con datos de 3279 semanas, siendo la primera de ellas la de 1958-08-02 y la última 2021-05-29. Dado que los datos de días y meses de un solo dígito estaban almacenados con un solo carácter, se agregó un cero delante de los mismos para evitar confusiones al concatenar la fecha. De esta manera, se obtiene la fecha en formato YYYY-MM-DD.

Finalmente, la tabla Género indica los géneros musicales a los que pertenecen las canciones del ranking.

Estas cinco tablas constituyen cinco DataFrames en Python, relacionados por las claves primarias Artista ID, Canción ID y Fecha ID. Para relacionar los datos de un DataFrames a otro se utilizó la función *merge*().

Las tablas 1 a 5 enumeran detalladamente las variables que contiene cada DataFrames.

Tabla 1. DataFrames Artista

Columna	Tipo de variable	Tipo de dato	Descripción
Artista ID	Identificador	int64	Clave primaria que permite identificar de forma unívoca al artista. No puede haber dos iguales.
Nombre	Categórica	string	Nombre del artista

Tabla 2. DataFrames Canción

Columna	Tipo de variable	Tipo de dato	Descripción
Canción ID	Identificador	string	Clave primaria que permite identificar de forma unívoca a la canción. No puede haber dos iguales
Duración	Cuantitativa continua	float64	Duración de la canción en milisegundos
Álbum	Cualitativa	string	Álbum al que pertenece la canción
Género	Cualitativa	string	Género musical de la canción
Nombre	Cualitativa	string	Nombre de la canción
Explicit	Cualitativa	string	Indica la presencia de palabras obscenas u ofensivas en la canción.
Artista ID	Identificador	int64	Clave primaria que permite identificar de forma unívoca al artista. No puede haber dos iguales.
Valencia	Cuantitativa continua entre 0 y 1	float64	Mide la positividad musical de la canción. Las melodías con Valencia alta suenan más positivas, felices, y eufóricas, mientras que las canciones con baja Valencia suenan negativas, tristes, o enojadas.
Tempo	Cuantitativa continua	float64	Es la velocidad o ritmo de la melodía, y deriva directamente de la duración promedio del pulso. Se mide en beats por minuto (BPM). Si los BPM son bajos el tempo es lento.
Energía	Cuantitativa continua entre 0 y 1	float64	Representa un porcentaje de intensidad de actividad. Es una medida perceptiva de intensidad y actividad. Generalmente, las canciones energéticas se sienten rápidas, fuertes y ruidosas.
Modo	Cualitativa discreta (0 o 1)	float64	Indica la modalidad (mayor o menor) de una canción, es decir, el tipo de escala a partir del cual deriva el contenido musical. 1 significa mayor y 0 menor.
Clave	Cuantitativa discreta entre 0 y 11.	float64	Es el tono, las notas o la escala de la canción que forma la base. Las 12 claves varían entre 0 y 11.
Acusticidad	Cuantitativa continua entre 0 y 1	float64	Es la confianza en que la canción sea acústica. 1 significa que muy probablemente lo sea.
Contenido Vocal	Cuantitativa continua entre 0 y 1	float64	Detecta la presencia de palabras en la canción. Cuanto más cerca de 1 se encuentra el contenido vocal, mayor contenido puramente vocal posee la grabación. Valores entre 0.33 y 0.66 describen canciones que contienen tanto palabras como melodías. Menos

			de 0.33 indica que probablemente solo haya melodía.
Bailabilidad	Cuantitativa continua entre 0 y 1	float64	Describe qué tan apta es una canción para bailarla basada en la combinación de elementos musicales, incluyendo el tempo, la estabilidad del ritmo, la fuerza del pulso y la regularidad en general. 1 significa que es muy bailable.
Sonoridad	Cuantitativa	float64	Mide los niveles de audio en función de la forma en que los humanos perciben el sonido. Se mide en decibeles (dB). Los valores típicos oscilan entre -60 y 0 dB. Cuanto mayor es la amplitud, más fuerte e intenso es el sonido.

Tabla 3. Dataframe Top100

Columna	Tipo de variable	Tipo de dato	Descripción
Ranking	Cualitativa	int64	Indica la posición en el ranking de canción en la semana correspondiente
Canción ID	Identificador	string	Clave primaria que permite identificar de forma unívoca a la canción. No puede haber dos iguales.
Artista ID	Identificador	int64	Clave primaria que permite identificar de forma unívoca al artista. No puede haber dos iguales.
Ranking semana anterior	Cualitativa	float64	Indica el ranking de la canción en la semana inmediatamente anterior.
Máximo ranking	Cualitativa	int64	Indica la máxima posición alcanzada por la canción
Semanas en Hot100	Cuantitativa discreta	int64	Indica la cantidad de semanas de la canción en el Billboard chart.
Fecha	Identificador	datetime64	Identifica de forma unívoca la fecha de publicación del ranking

Tabla 4. DataFrames Semana

Columna	Tipo de variable	Tipo de dato	Descripción
Día	Cualitativa	int64	Día en que se publicó el ranking semanal.
Mes	Cualitativa	int64	Mes en el que se publicó el ranking semanal.
Año	Cualitativa	int64	Año en el que se publicó el ranking semanal.
Fecha	Identificador	datetime64	Identifica de forma unívoca la fecha de publicación del ranking. Es la concatenación de los tres campos anteriores.

Tabla 5. DataFrames Género

Columna	Tipo de variable	Tipo de dato	Descripción
Canción ID	Identificador	string	Clave primaria que permite identificar de forma unívoca a la canción. No puede haber dos iguales.

Pop	Cualitativa	boolean	Indica si la canción pertenece al género pop
Rock	Cualitativa	boolean	Indica si la canción pertenece al género rock
Música Clásica	Cualitativa	boolean	Indica si la canción pertenece al género música clásica
Jazz	Cualitativa	boolean	Indica si la canción pertenece al género jazz
Blues	Cualitativa	boolean	Indica si la canción pertenece al género blues
Hip Hop	Cualitativa	boolean	Indica si la canción pertenece al género Hip Hop
Rap	Cualitativa	boolean	Indica si la canción pertenece al género rap
Trap	Cualitativa	boolean	Indica si la canción pertenece al género trap
Electrónica	Cualitativa	boolean	Indica si la canción pertenece al género electrónica
Country	Cualitativa	boolean	Indica si la canción pertenece al género country
Soul	Cualitativa	boolean	Indica si la canción pertenece al género soul
Funk	Cualitativa	boolean	Indica si la canción pertenece al género funk
Disco	Cualitativa	boolean	Indica si la canción pertenece al género disco
Reggae	Cualitativa	boolean	Indica si la canción pertenece al género reggae
Reggaetón	Cualitativa	boolean	Indica si la canción pertenece al género reggaetón
Gospel	Cualitativa	boolean	Indica si la canción pertenece al género góspel
Metal	Cualitativa	boolean	Indica si la canción pertenece al género metal
R&B	Cualitativa	boolean	Indica si la canción pertenece al género R&B
Bachata	Cualitativa	boolean	Indica si la canción pertenece al género bachata
Bossa Nova	Cualitativa	boolean	Indica si la canción pertenece al género bossa nova
Desconocido	Cualitativa	boolean	Indica si la canción pertenece al género desconocido

Para evitar volver a correr todo el código de procesamiento de datos, los cinco DataFrames se exportaron en formato Excel.

5. LIMPIEZA DE DATOS

Consiste en la eliminación de *outliers* y en el tratamiento de datos ausentes.

DATOS AUSENTES

Tabla Artista

No se encontraron valores nulos en ninguna de las columnas.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10494 entries, 284 to 322636
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Artista     10494 non-null  string
 1   Artista ID  10494 non-null  int64
dtypes: int64(1), string(1)
memory usage: 246.0 KB
```

Figura 2. Resumen de la tabla Artista

Tabla Canción

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30432 entries, 0 to 30431
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Canción ID       30432 non-null  string
 1   Track ID         25354 non-null  object
 2   Duración         25354 non-null  float64
 3   Bailabilidad     25292 non-null  float64
 4   Energía          25292 non-null  float64
 5   Clave            25292 non-null  float64
 6   Sonoridad        25292 non-null  float64
 7   Modo             25292 non-null  float64
 8   Contenido Vocal  25292 non-null  float64
 9   Acusticidad      25292 non-null  float64
 10  Valencia         25292 non-null  float64
 11  Tempo            25292 non-null  float64
 12  Nombre           30432 non-null  string
 13  Artista ID       30432 non-null  int64
dtypes: float64(10), int64(1), object(1), string(2)
memory usage: 3.3+ MB
```

Figura 3. Resumen de la tabla Canción

Como puede verse, hay muchos valores ausentes en esta tabla. Esto se debe a que en la base de datos original muchas de las canciones no tenían completos sus atributos. Mediante la API de Spotify se buscarán completar los campos vacíos.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30432 entries, 0 to 30431
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Canción ID       30432 non-null  object
 1   Track ID         30328 non-null  object
 2   Duración         30328 non-null  float64
 3   Bailabilidad     30262 non-null  float64
 4   Energía          30262 non-null  float64
 5   Clave            30262 non-null  float64
 6   Sonoridad        30262 non-null  float64
 7   Modo             30262 non-null  float64
 8   Contenido Vocal  30262 non-null  float64
 9   Acusticidad      30262 non-null  float64
 10  Valencia         30262 non-null  float64
 11  Tempo            30262 non-null  float64
 12  Nombre           30432 non-null  object
 13  Artista ID       30432 non-null  int64
dtypes: float64(10), int64(1), object(3)
memory usage: 3.5+ MB
```

Figura 4.

Como puede verse en la figura 4, mediante la API de Spotify se lograron rellenar muchos de los registros vacíos. Los registros que, aún luego de la utilización de la API, quedaron sin rellenar se buscaron manualmente en Spotify. Sin embargo, todos ellos resultaron inexistentes en la aplicación, y por lo tanto se decidió que dichos registros no pueden completarse y deberá adoptarse una estrategia de tratamiento de nulos. En otros casos, puede observarse que se encontró el ID de Spotify, pero la aplicación no registraba los atributos técnicos de la canción. En resumen:

- De las 30432 canciones, solamente 104 no se consiguieron en la aplicación.
- De las 30328 que sí se encontraron, 66 no contenían datos de atributos musicales.

En el caso de las variables cuantitativas, los valores ausentes se rellenarán luego del estudio de los valores atípicos.

Tabla Top100

La figura 5 muestra un resumen de la tabla Top100. Se ve que únicamente en la columna Ranking semana anterior existen valores nulos. Dado que el primer Billboard Chart se publicó en agosto de 1958, para la semana 1 no existen registros del ranking de la semana anterior porque se trata de la primera aparición. Por tal motivo, los registros del 1 al 100 se rellenaron con "1ra vez".

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 336395 entries, 0 to 336394
Data columns (total 7 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   Ranking                  336395 non-null  int64
 1   Canción ID               336395 non-null  object
 2   Ranking semana anterior  304441 non-null  float64
 3   Máximo Ranking           336395 non-null  int64
 4   Semanas en top100        336395 non-null  int64
 5   Fecha                    336395 non-null  datetime64[ns]
 6   Artista ID               336395 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(4), object(1)
memory usage: 18.0+ MB
```

Figura 5. Resumen del dataframe Top100

En el caso de los demás registros, es probable que el hecho de que no esté haya un ranking de la semana anterior se deba a que se trata de la primera aparición de la canción en el ranking. Para verificar que esto sea así se utiliza la columna "Semanas en hot 100" que indica la cantidad de semanas de la canción en el ranking. Si este valor es igual a 1, entonces se trata de la primera aparición de la canción, y se reemplaza el dato ausente por el número 0, que indica que esa es la primera aparición de la canción. Se decidió usar una clave numérica para categorizar a las canciones que aparecen por primera vez para poder mantener el tipo de datos numérico (float64) de la columna.

Después de realizar estos reemplazos quedan 2586 valores ausentes. En estos casos se debe tener en cuenta que la columna "Semanas en hot 100" no indica necesariamente que las semanas sean consecutivas. Puede suceder que la canción haya aparecido en el ranking varias semanas atrás, por lo cual no es la primera vez en el ranking, pero tampoco se cuenta con un dato. Esto ocurre así porque la columna "Ranking semana anterior" únicamente tiene en cuenta el ranking de la semana inmediatamente anterior. En estos casos el valor se reemplazó por 101, que indica que la semana anterior la canción no estuvo dentro del Top100. Luego de este procedimiento no existen valores ausentes (figura 6).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 336395 entries, 0 to 336394
Data columns (total 7 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Ranking                 336395 non-null  int64
 1   Canción ID              336395 non-null  string
 2   Ranking semana anterior 336395 non-null  float64
 3   Máximo Ranking          336395 non-null  int64
 4   Semanas en top100       336395 non-null  int64
 5   Fecha                   336395 non-null  datetime64[ns]
 6   Artista ID              336395 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(4), string(1)
memory usage: 20.5 MB
```

Figura 6. Resumen de la tabla Top100 luego del reemplazo de valores ausentes

Tabla Semana

La figura 7 muestra los campos de la tabla Semana, junto con sus tipos de datos y la cantidad de registros no nulos. Como puede observarse, no existen valores ausentes. Los valores de día, mes y año están almacenados como int64, y la fecha como datetime64. Como puede verse en la figura 7, se cuenta con datos de 3362 Billboard Charts.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3362 entries, 1 to 3362
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Día     3362 non-null   int64
 1   Mes     3362 non-null   int64
 2   Año     3362 non-null   int64
 3   Fecha   3362 non-null   datetime64[ns]
dtypes: datetime64[ns](1), int64(3)
memory usage: 131.3 KB
```

Figura 7. Resumen de la tabla Semana

Tabla Género

No se encontraron valores ausentes en la tabla Género

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30578 entries, 17 to 1
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Canción ID      30578 non-null  string
 1   Pop             30578 non-null  boolean
 2   Rock            30578 non-null  boolean
 3   Música Clásica  30578 non-null  boolean
 4   Jazz            30578 non-null  boolean
 5   Blues           30578 non-null  boolean
 6   Hiphop          30578 non-null  boolean
 7   Rap             30578 non-null  boolean
 8   Trap            30578 non-null  boolean
 9   Electrónica     30578 non-null  boolean
 10  Country         30578 non-null  boolean
 11  Soul            30578 non-null  boolean
 12  Funk            30578 non-null  boolean
 13  Disco           30578 non-null  boolean
 14  Reggae          30578 non-null  boolean
 15  Reggaetón       30578 non-null  boolean
 16  Gospel          30578 non-null  boolean
 17  Metal           30578 non-null  boolean
 18  R&B             30578 non-null  boolean
 19  Bachata         30578 non-null  boolean
 20  Bossa Nova      30578 non-null  boolean
 21  Desconocido     30578 non-null  boolean
```

Figura 8. Resumen de la tabla Género

OUTLIERS

Tabla Semana

Dado que el campo fecha se obtuvo como una concatenación de las columnas día, mes y año, únicamente se verificarán los valores de dichos campos. Debe cumplirse que los valores del atributo día se encuentren entre 1 y 31, que los del mes estén entre 1 y 12, y que los de año estén entre 1958 y 2021. Cabe destacar que estos campos no tienen interés estadístico, por lo cual la verificación de datos atípicos solo se realiza para verificar que todos los datos estén correctamente registrados y sean coherentes. Las figuras 9-11 muestran las distribuciones de las variables de la tabla Semana.

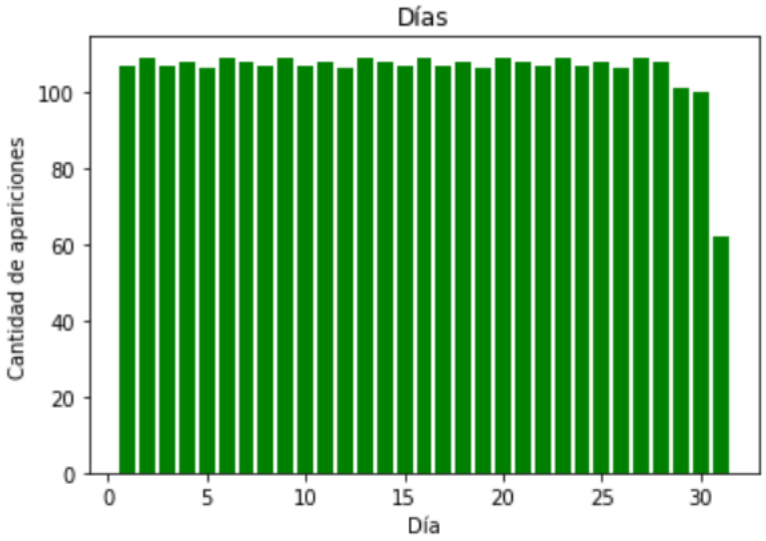

Figura 9. Valores del campo Día

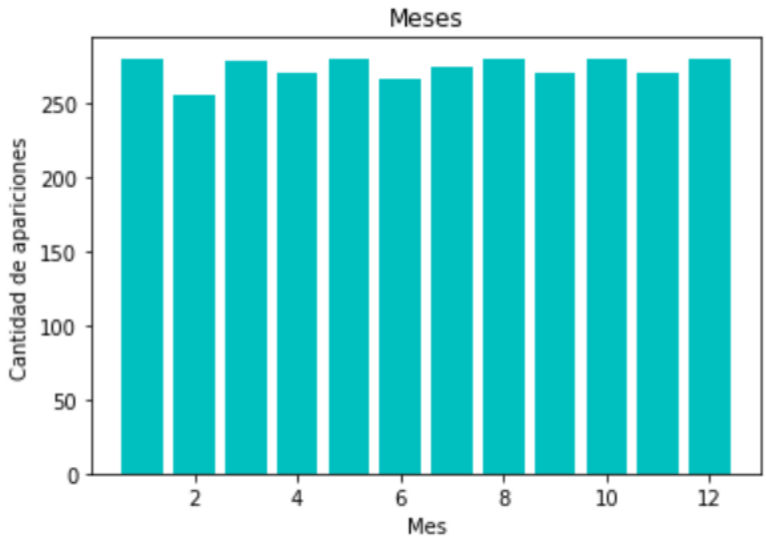

Figura 10. Valores del campo Mes

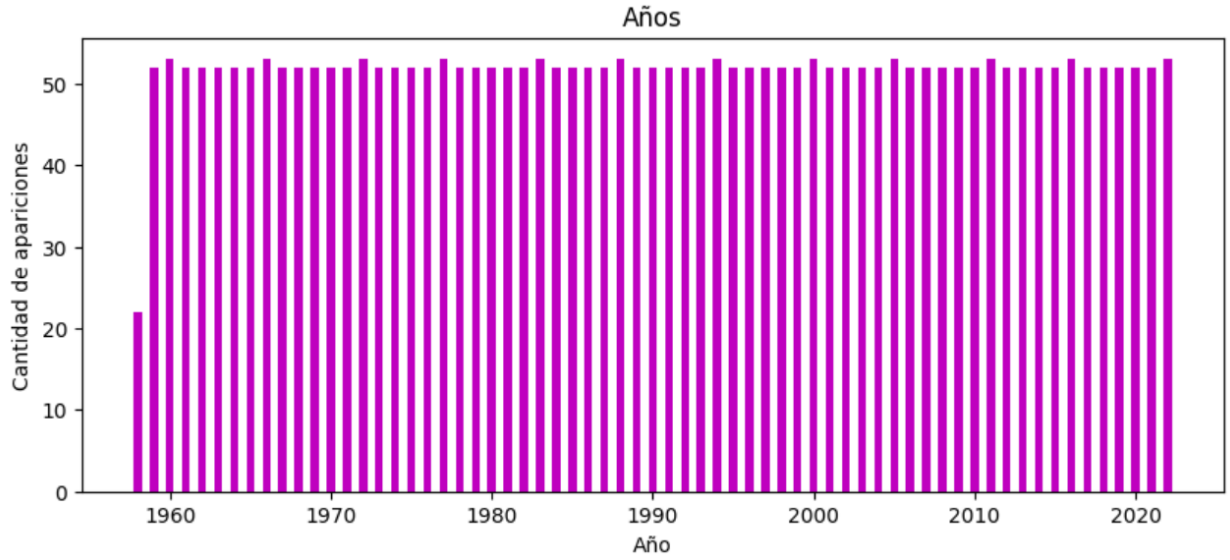

Figura 11. Valores del campo Año

Tabla Canción

De los 16 campos que posee la tabla, 2 son identificadores (Canción ID y Artista ID). De los 14 restantes, 6 son cualitativos (Nombre, Álbum, Explícito, Género, Modo y Clave) y 8 son cuantitativos. Estos últimos son los que indican los parámetros técnicos que caracterizan a la canción.

A continuación, se analizan los campos cuantitativos.

Columna duración

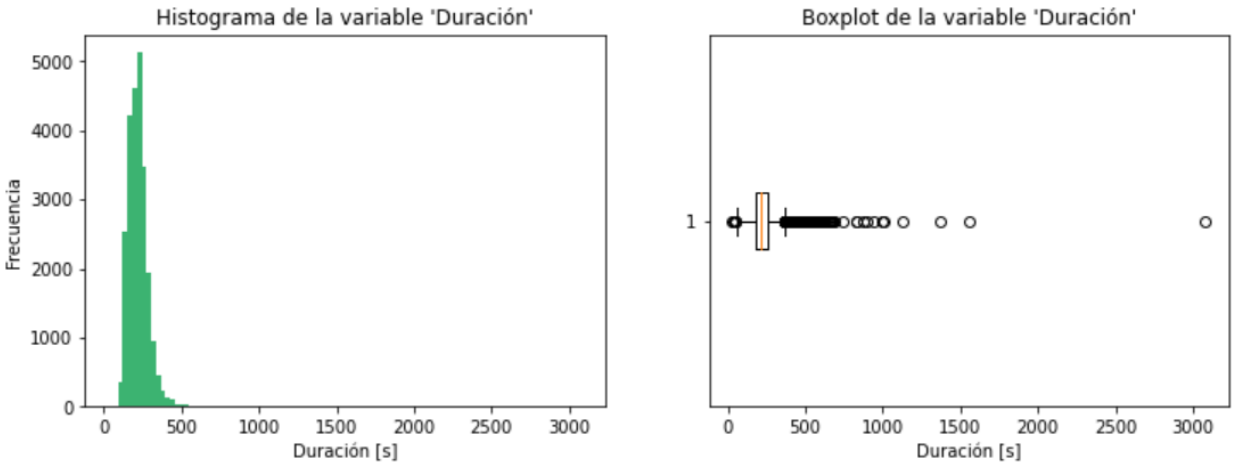

Figura 12. Distribución de la variable Duración

Los gráficos ignoran la presencia de valores ausentes gracias al uso de dropna().

Tabla 5. Medidas de resumen de la columna duración

Media	Mediana	Moda	Q1	Q3	Inlier superior	Inlier inferior

220.78	215	160 y 205.73	175.05	253.38	370.88	57.55

El histograma indica la presencia de posibles valores atípicos, y el diagrama de caja y bigote lo confirma. La mayoría de las canciones tienen una duración entre 0 y 500 segundos, que son aproximadamente 8 minutos. Sin embargo, existen muchos registros con valores mayores a 500 segundos. Estos valores son extremadamente grandes para ser canciones, por lo que es imprescindible modificarlos. Como estrategia se decide reemplazarlos por la media de la distribución. La misma estrategia se aplica a los datos ausentes.

Columnas Valencia, Energía, Acusticidad, Contenido Vocal y Bailabilidad.

Los campos cuantitativos que varían entre 0 y 1 se analizan en conjunto. La figura 13 muestra los histogramas de las 5 variables. Se observa que las distribuciones tienen formas muy diferentes.

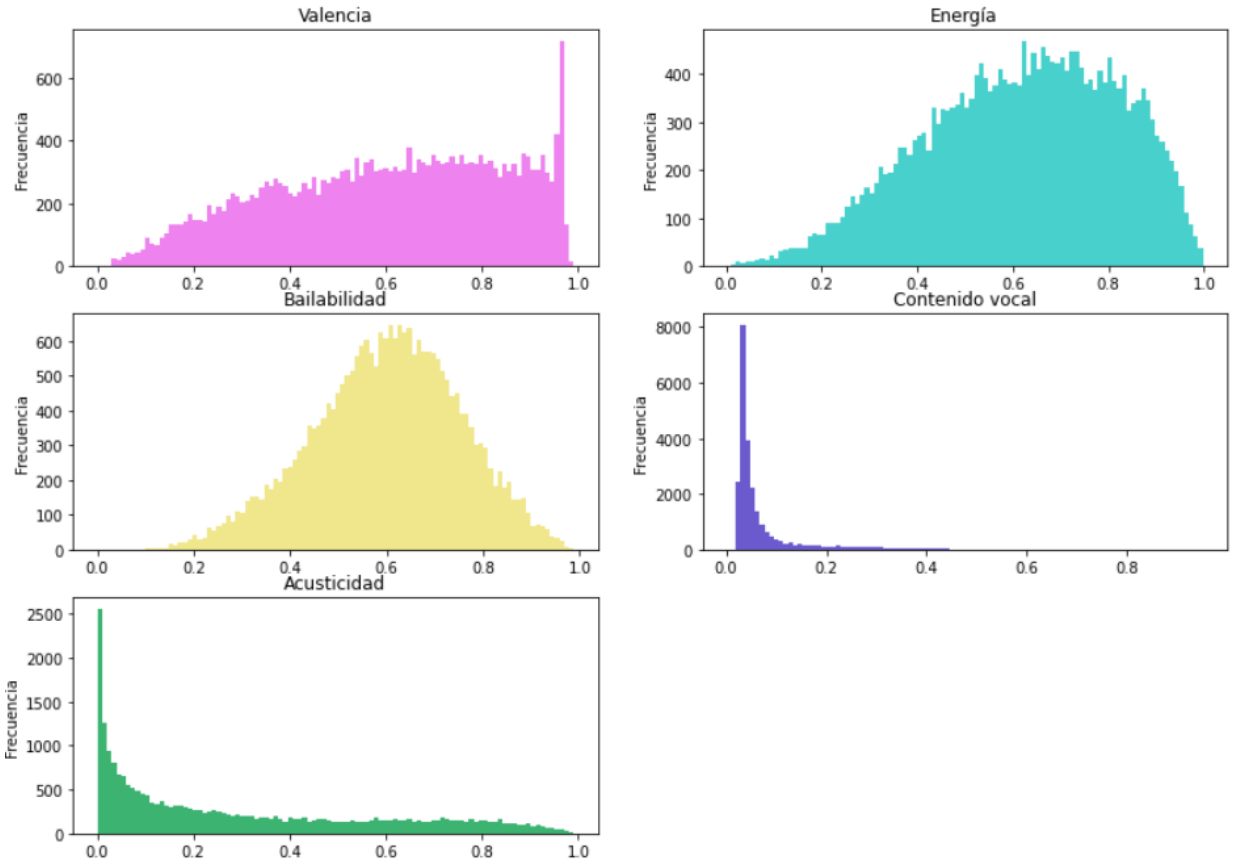

Figura 13. Distribuciones de variables cuantitativas

En la figura 13 se complementa el histograma con los diagramas de caja.

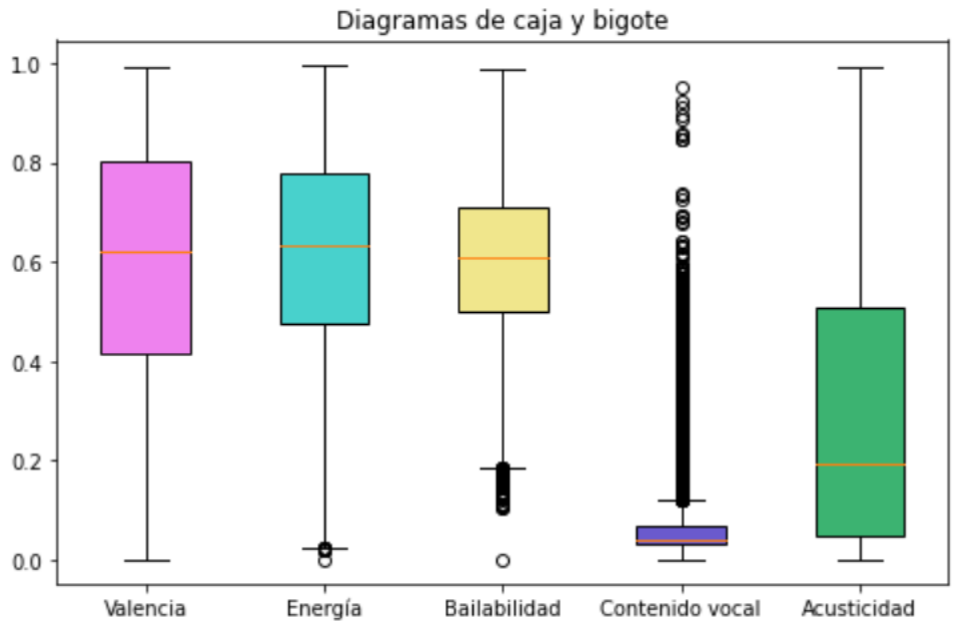

Figura 14. Diagramas de caja y bigote

En el caso de la valencia y la acusticidad no se encuentran valores atípicos. La figura 14 muestra que la valencia tiene una distribución que podría considerarse normal con gran varianza, mientras que la acusticidad está lejos de ser normal. Se encuentran muchos valores cercanos a 0, que indican que la melodía no es acústica.

Con respecto a la energía, se podría decir que hay un solo valor atípico por debajo del Inlier inferior.

En el caso de la Bailabilidad, también se observan *outliers* por debajo del *inlier* inferior. La distribución de esta variable es normal, por lo que se deben eliminar esos valores atípicos.

Finalmente, el contenido vocal tiene una distribución sesgada hacia la derecha.

Columna Tempo

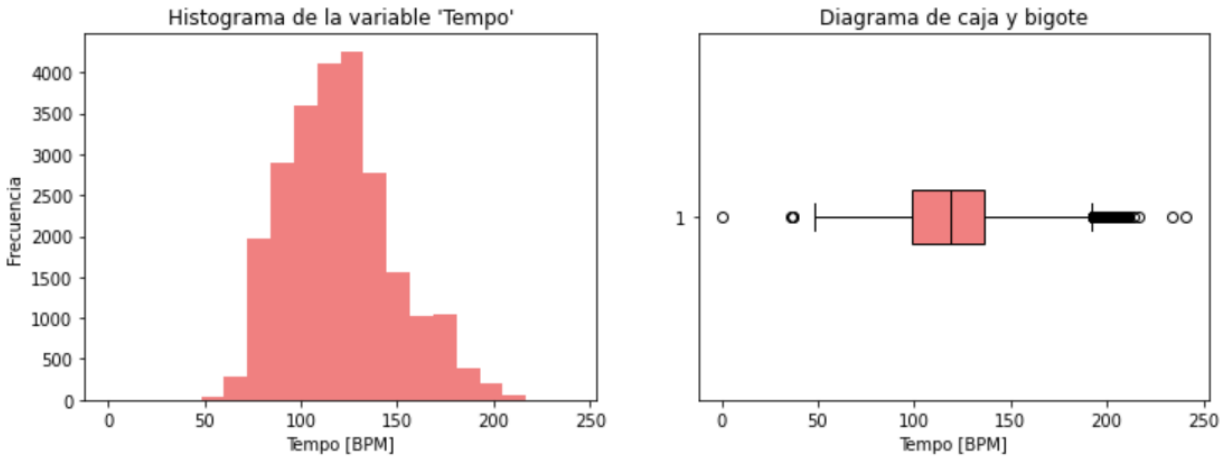

Figura 15. Distribución del tempo

Se verifica la existencia de *outliers* tanto por debajo del inlier inferior, como por encima del *inlier* superior. Se reemplazan por la media de la distribución.

Columna Sonoridad

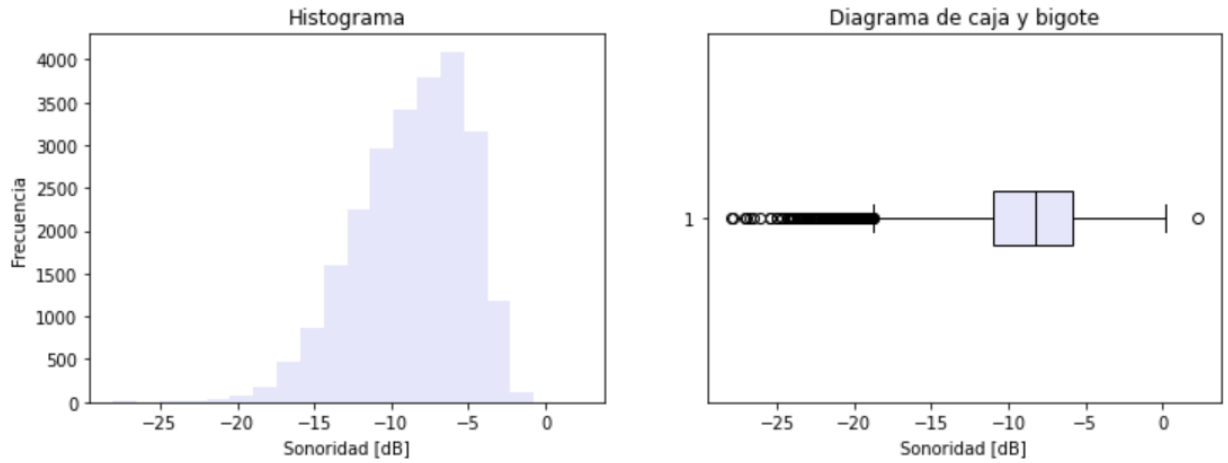

Figura 16. Distribución de la sonoridad

Se eliminan los *outliers* de la distribución.

Campos Clave y Modo

Se realiza un recuento de los registros de cada categoría con el único fin de verificar la inexistencia de valores incorrectos.

Tabla Top100

Posee 7 campos, de los cuales 3 son identificadores. Entre los 4 restantes, hay 3 variables cualitativas (“Ranking”, “Ranking semana anterior” y “Máximo ranking”) y una cuantitativa (“Semanas en Hot

100"). Las posiciones en el ranking se consideran variables cualitativas porque no tiene sentido sumarlos, restarlos ni promediarlos.

La variable cuantitativa "Semanas en Hot 100" se analiza a continuación. Al ordenar los valores se encuentra que la mínima cantidad de semanas registradas es 1, mientras que la máxima es 87.

En la figura 17 se observa un histograma con 19 clases. El número de clases fue seleccionado siguiendo la regla de Sturges ($K = 1 + 3.3\log(N)$, siendo N la cantidad de observaciones). Como puede verse, la distribución está muy lejos de ser normal.

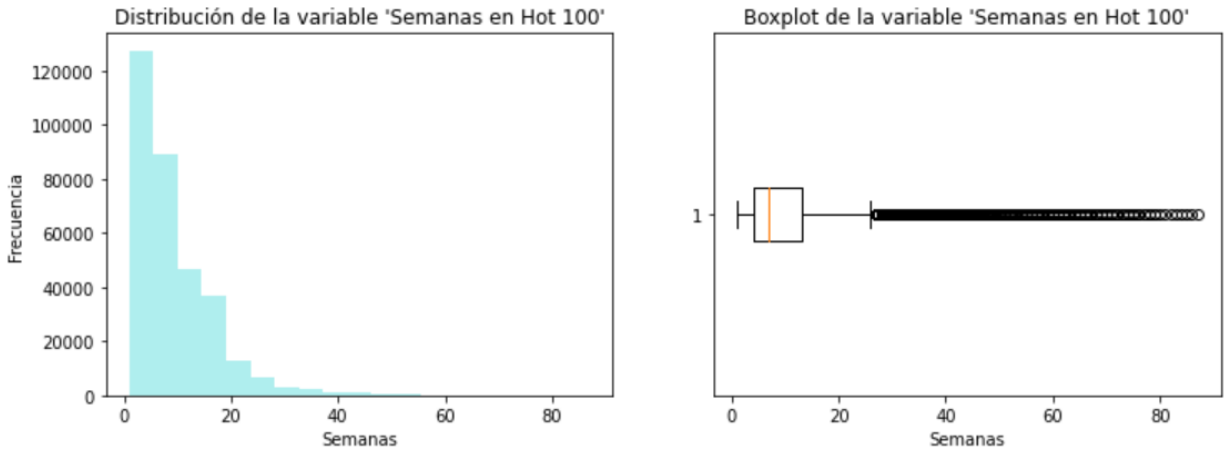

Figura 17. Distribución de la variable "Semanas en Hot 100"

El histograma indica que existen muchos artistas que han aparecido pocas veces en el ranking, y muy pocos artistas que hayan acumulado más de 40 semanas en el ranking en su carrera.

Dada la gran cantidad de valores que se encuentran más allá del tercer cuartil, no se considerarán atípicos dichos valores.

Las medidas de tendencia central que derivan del histograma se encuentran en la tabla 6.

Tabla 6. Medidas de tendencia central de la variable "Semanas en Hot 100"

Media	Mediana	Moda
9.15	7	0 y 1

Como en toda distribución sesgada hacia la derecha se verifica que la media es mayor a la mediana, la cual a su vez es mayor a la moda. Además, se comprueba que la distribución es bimodal. Usando el método *skew*() se obtuvo que el sesgo en la distribución es 1.8, por lo cual puede afirmarse que la distribución es altamente no uniforme.

En el caso de las variables cualitativas se realizó un procedimiento similar al de los días, meses y años porque el único objetivo es verificar que no haya errores en los valores almacenados. Las posiciones

del ranking deben variar entre 1 y 100, excepto en el caso del campo "ranking semana anterior" que incluye también al 0 y al 101.

Las figuras 18 a 20 muestran las distribuciones de las variables cualitativas de la tabla Top100. En todos los casos se verifica la inexistencia de valores incoherentes.

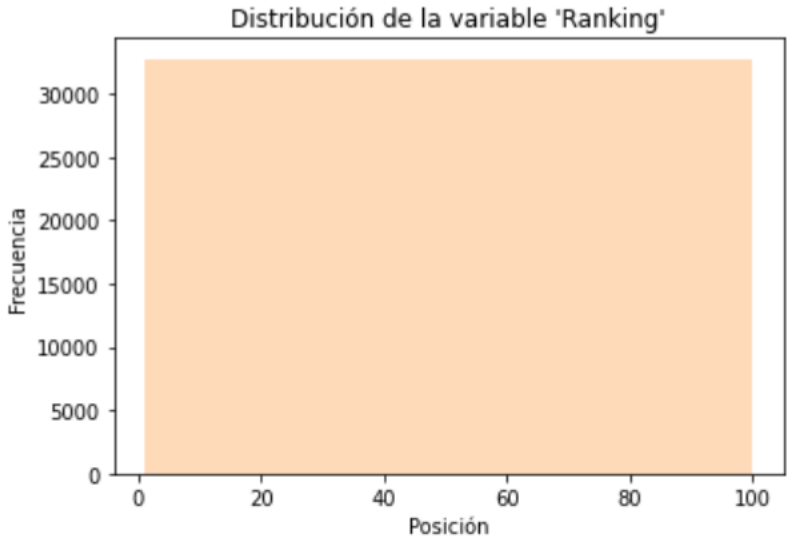

Figura 18. Distribución de la variable "Ranking".

Dado que por cada semana se muestran las 100 canciones, la distribución de la variable "Ranking" es totalmente uniforme.

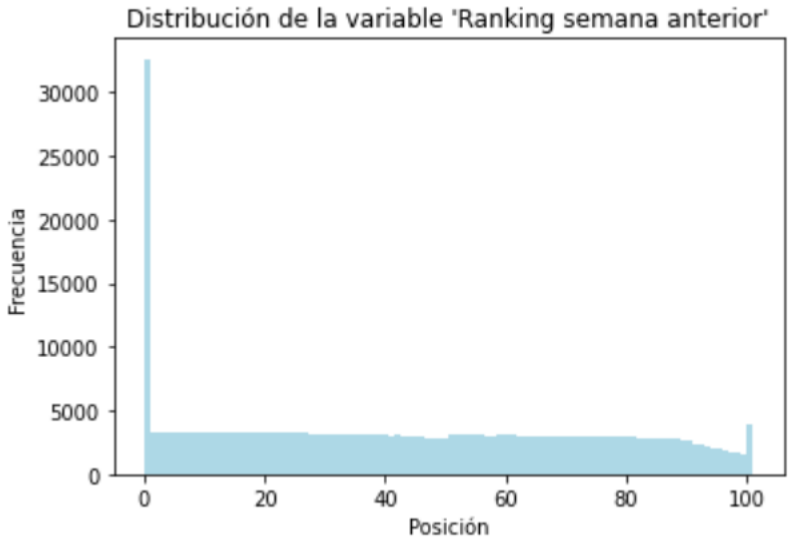

Figura 19. Distribución de la variable "Ranking semana anterior"

Se observa que los valores más recurrentes son el 0 y el 101, casualmente aquellos valores establecidos arbitrariamente para indicar las canciones que aparecen por primera vez en el ranking, o

que vuelven a aparecer después de más de una semana, respectivamente. Dejando de lado estos valores, se verifica que las posiciones del ranking entre 1 y 100 tienen una distribución uniforme.

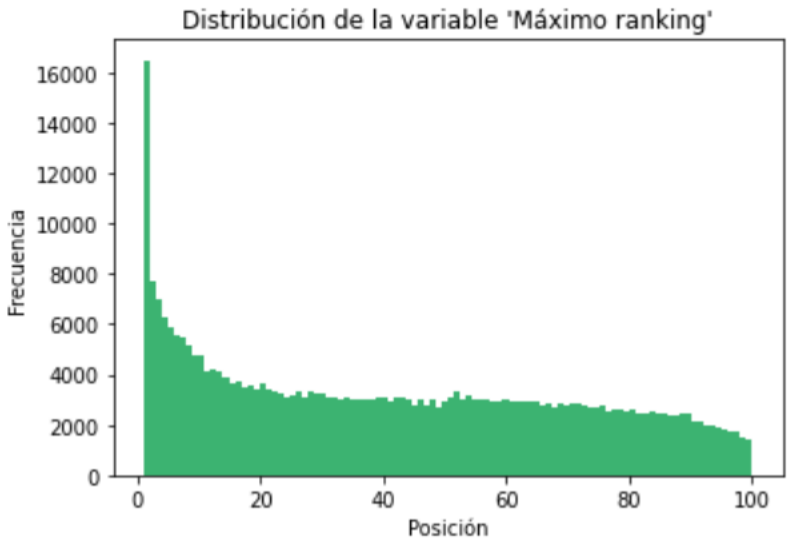

Figura 20. Máximo ranking

En cuanto a la variable "Máximo ranking", se observa que el valor más recurrente es el número 1, lo cual indica que la mayoría de las canciones del ranking alguna vez pasan por la posición número 1.

Tabla Artista

Las columnas del dataframe no tienen interés estadístico. Son los nombres y los identificadores de los artistas. Durante el análisis esta tabla se usará para almacenar nuevas variables, las cuales sí serán de interés estadístico.

Tabla Género

Las columnas del DataFrames tampoco tienen interés estadístico.

6. ANÁLISIS DE DATOS

CORRELACION ENTRE VARIABLES

Se evaluó la relación entre las 8 variables cuantitativas de la tabla canción usando la función *scatter_matrix* de pandas. Los resultados se observan en la figura 21.

Se podría decir que existe una especie de correlación débil entre la sonoridad y la energía. Se observa que, si hay valores grandes de energía, también hay valores altos de sonoridad. Teniendo en cuenta el significado de cada una de las variables, tiene sentido que haya una correlación positiva entre ellas.

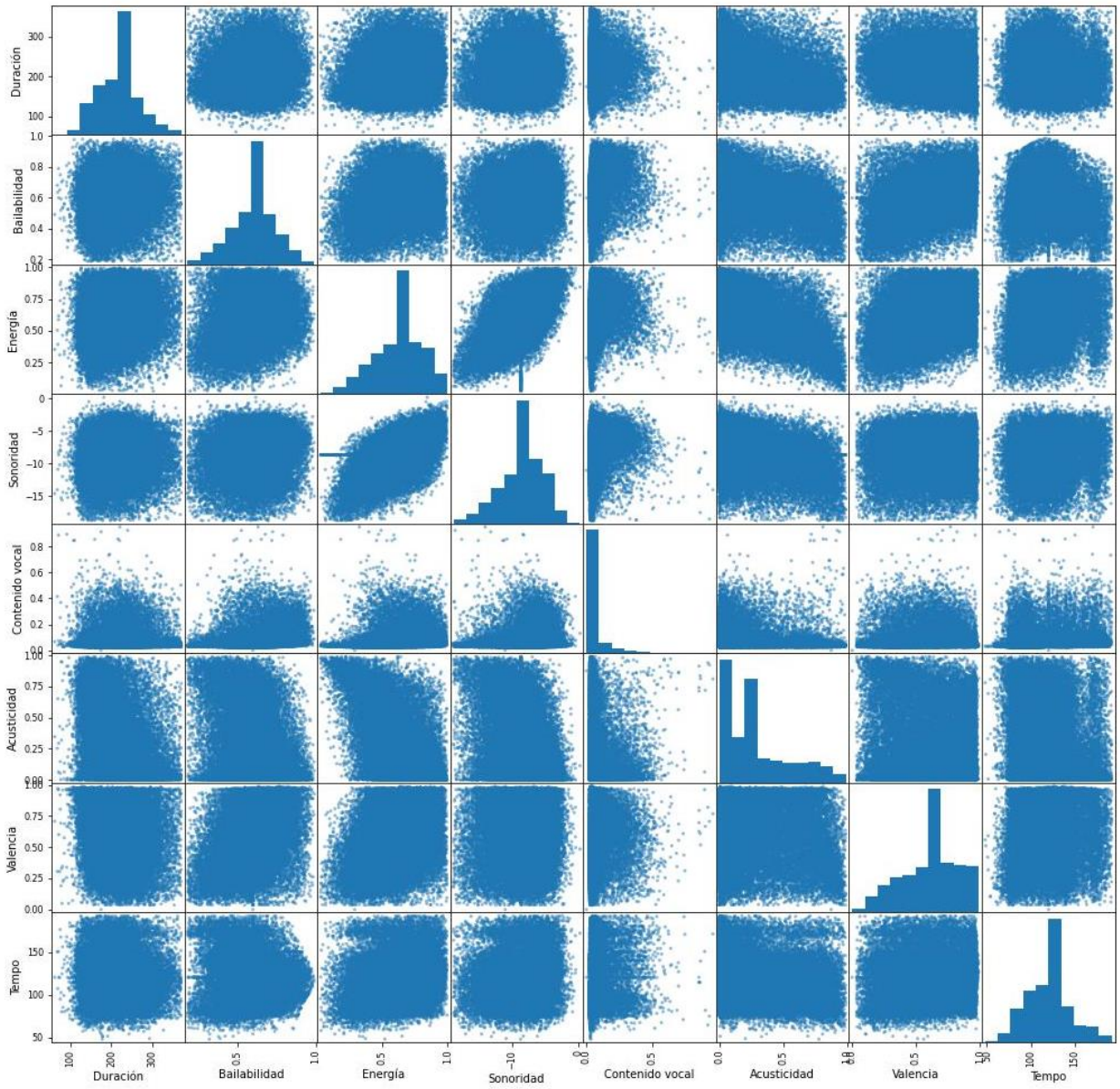

Figura 21. Correlación entre variables

Ahora se relacionan estas variables con las dos variables cualitativas de la tabla.

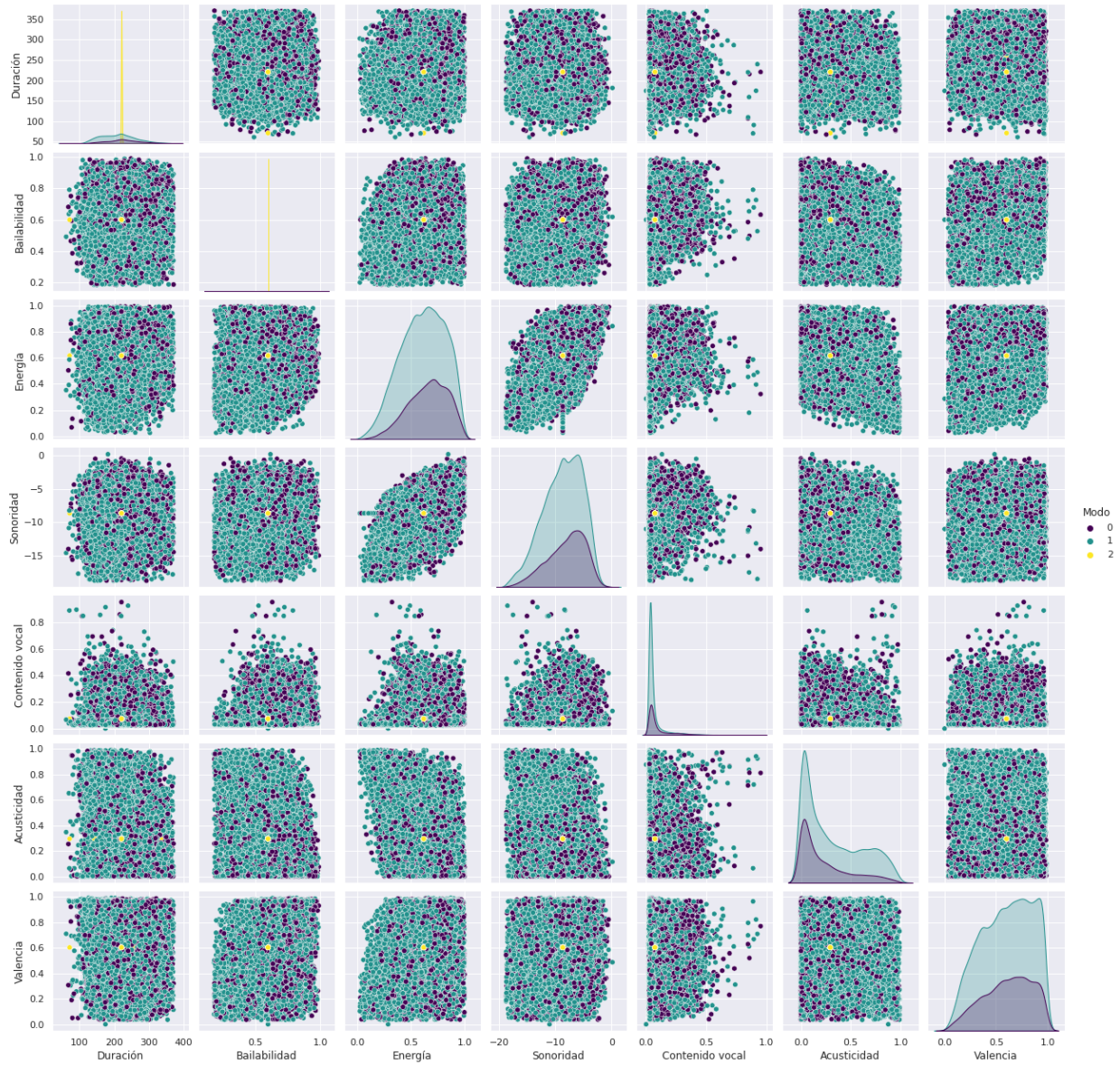

Figura 22. Correlación entre variables

En la figura 22 se agregó la relación con la variable cualitativa modo. En todos los histogramas de las variables consigo mismas se observa que cuando el modo el igual a 0 todas las variables toman valores más pequeños. En los demás casos, es imposible sacar conclusiones porque los valores se encuentran muy dispersos.

EDA

Las dos preguntas a responder durante el análisis exploratorio son las siguientes:

- **Pregunta 1**: ¿Las canciones más exitosas, es decir, aquellas que más tiempo y mejores posiciones ocuparon en el ranking, tienen características en común? De ser así, ¿esas características fueron cambiando con el paso de los años?
- **Pregunta 2**: ¿Existe un género predominante dentro del top 10 del ranking? ¿Existe alguna relación entre el género y las propiedades musicales?

Indicadores generales

Teniendo en cuenta que por cada semana hay 100 registros, uno correspondiente a cada posición del ranking, en el período de tiempo analizado existieron 3279 semanas. La primera semana registrada en el dataset es la del 2 de agosto de 1958, y la última es la del 29 de mayo de 2021. Sabiendo que por cada semana pueden aparecer hasta 100 canciones diferentes, pudieron aparecer como máximo 327900 canciones diferentes. Sin embargo, a lo largo de este tiempo aparecieron solamente 29389 canciones diferentes, lo cual indica que las canciones se repiten con mucha frecuencia dentro del ranking. Con respecto a los artistas, aparecieron 10061, lo cual indica que existen artistas con más de una canción en el ranking.

En la figura 23 se observan los artistas con mayor cantidad de semanas en el ranking. Elton John lidera el top 10, seguido por Kenny Chesney y Tim McGraw.

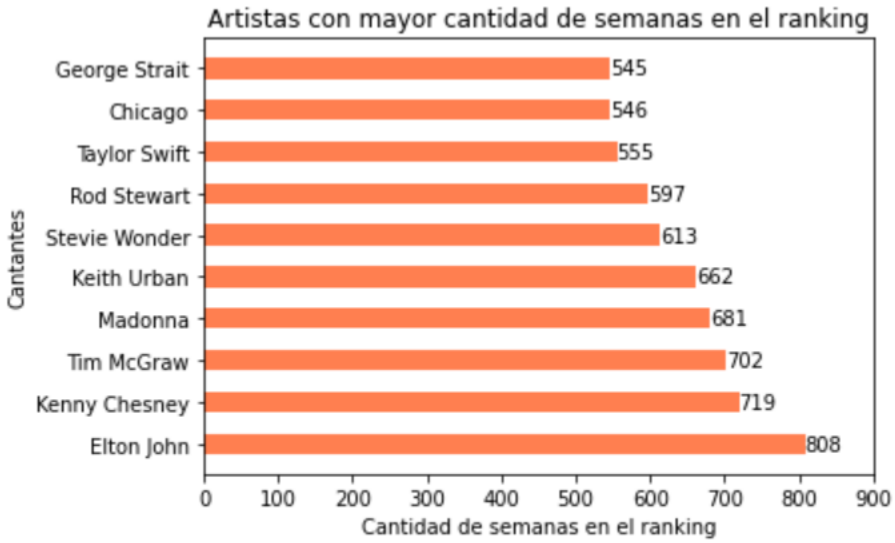

Figura 23. Artistas con mayor cantidad de semanas en el ranking

Pregunta 1

Antes de responder a la pregunta número 1 es necesario identificar a las canciones más exitosas. Para ello, se plantearon tres criterios diferentes, los cuales se muestran en la tabla 7.

Tabla 7. Criterios para identificar a las canciones más exitosas

Criterio 1	Canciones con más semanas en el ranking
Criterio 2	Canciones con más semanas en el top 10
Criterio 3	Canciones con más semanas en la posición número 1

De cada uno de estos criterios se va a establecer un orden, que mostrará todas las canciones, desde las más reproducidas a las menos conocidas. Usando las funciones de pandas *head* y *tail*, se podrán conocer los atributos de estas canciones, y compararlos para saber si existen diferencias entre ellos.

Las variables cuantitativas se normalizarán visualizarán en un gráfico de coordenadas paralelas.

Criterio 1

Se analizarán canciones en función de la época, y se seleccionarán las primeras y últimas 10 canciones.

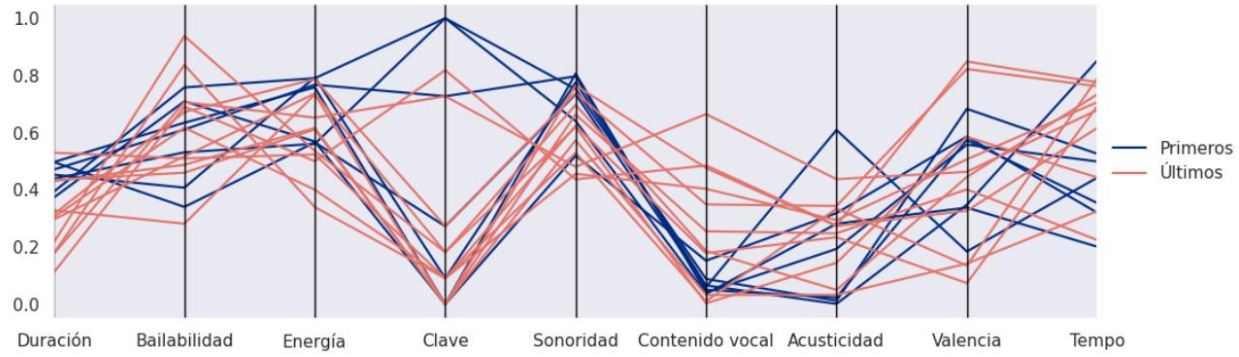

Figura 24. Canciones del último año

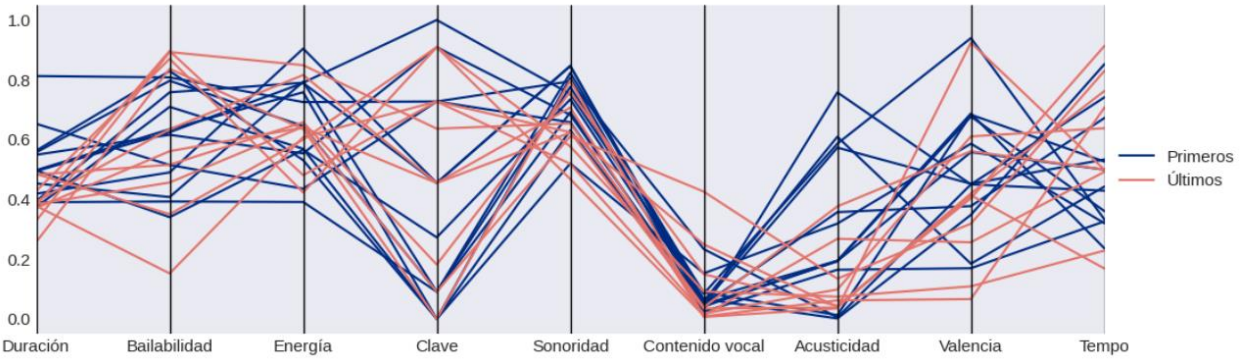

Figura 25. Canciones de los últimos cinco años

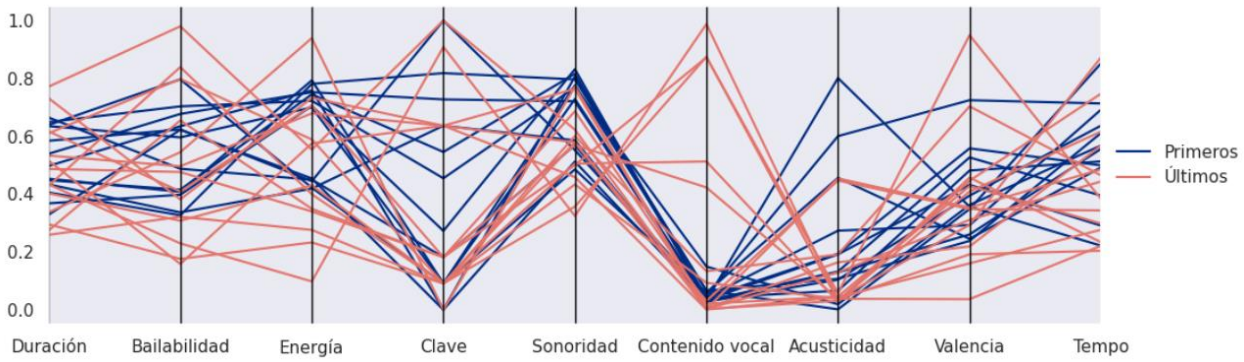

Figura 26. Canciones de todos los tiempos

En los tres casos se observa que los valores que toma la variable clave, son muy amplios. Tanto en las primeras como en las últimas canciones varía entre 0 y 100%, y no se puede establecer un patrón que distinga al tipo de canción. Algo similar sucede con el tempo y la sonoridad, pero en un rango de valores más pequeños, entre 20 y 80% aproximadamente.

Con respecto a la bailabilidad, las canciones primeras del ranking se concentran en valores entre 40 y 80%, mientras que las últimas toman valores más amplios (entre 20 y 90%). El contenido vocal parece ser una variable clave para distinguir a las canciones por su ranking, porque las canciones que se encuentran primeras toman valores pequeños, menores al 20%, mientras que las últimas suelen tener contenidos acústicos mucho mayores. En el caso de la energía, también podría decirse que las canciones exitosas se concentran alrededor de 50 y 80%, mientras que las últimas suelen tener valores menores.

La valencia podría ser otra variable de referencia, porque las primeras canciones del ranking se concentran en un rango más pequeño con respecto a las últimas del ranking (30-70% vs 10-90%).

La duración no pareciera ser una variable relevante porque los valores oscilan bastante.

Criterio 2

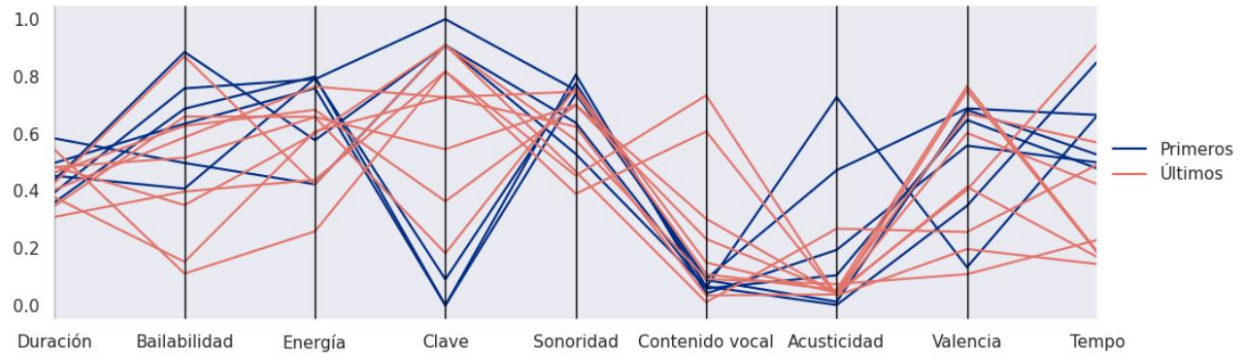

Figura 27. Canciones del último año

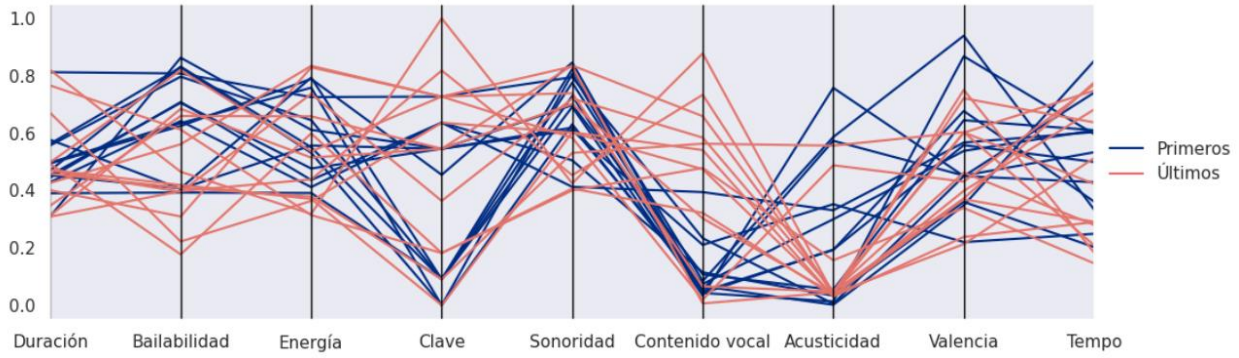

Figura 28. Canciones de los últimos cinco años

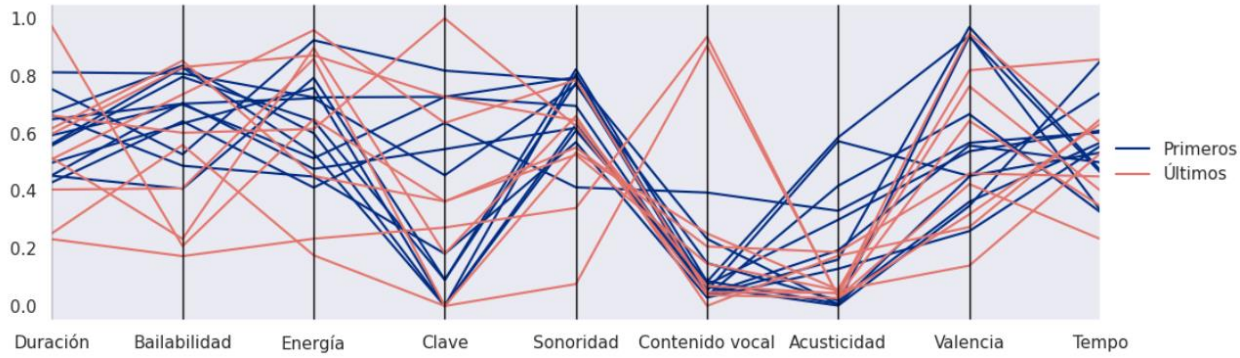

Figura 29. Canciones de todos los tiempos

Los patrones relevantes del criterio 1 parecen repetirse. El contenido vocal de las canciones exitosas sigue siendo bajo, y la bailabilidad sigue siendo mayor al 40%. En este caso aparecen canciones poco escuchadas con contenido de sonoridad bajo, mientras que las canciones más escuchadas siguen mostrando valores de sonoridad mayores al 50%.

En el caso de la energía, continúa el patrón de que las canciones exitosas se concentran alrededor de 50 y 80%, mientras que las últimas suelen tener valores menores.

Criterio 3

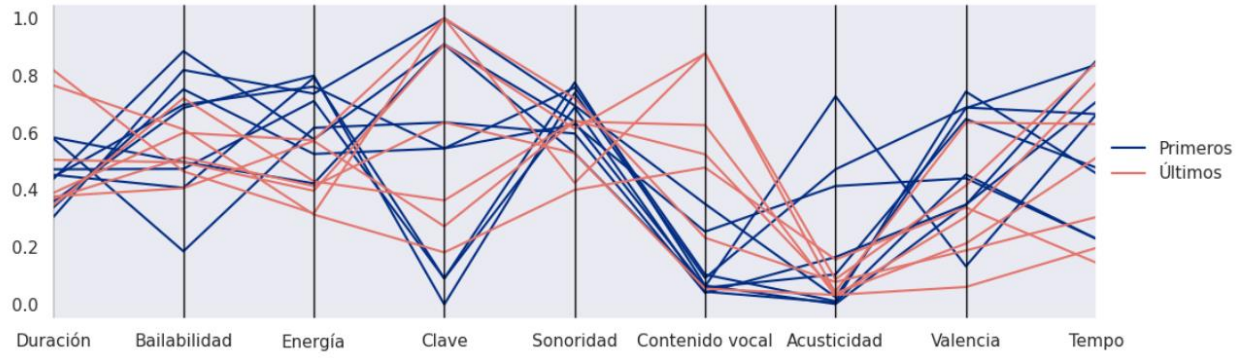

Figura 30. Canciones del último año

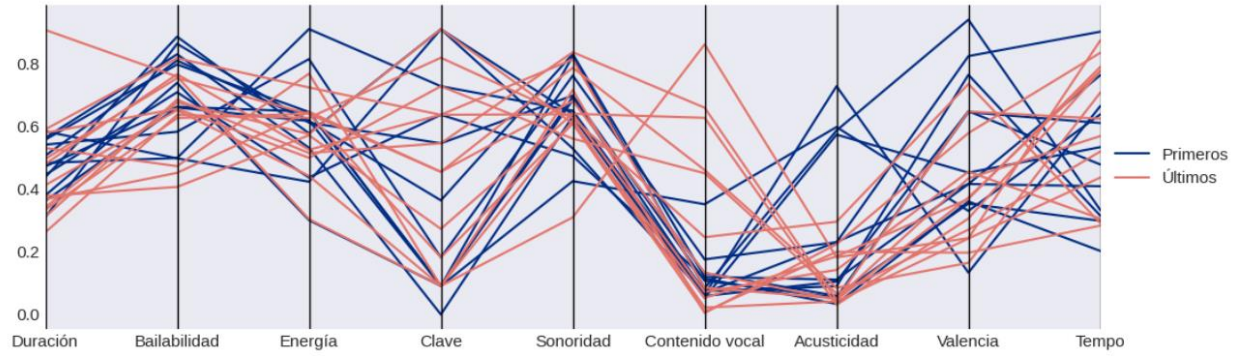

Figura 31. Canciones de los últimos cinco años

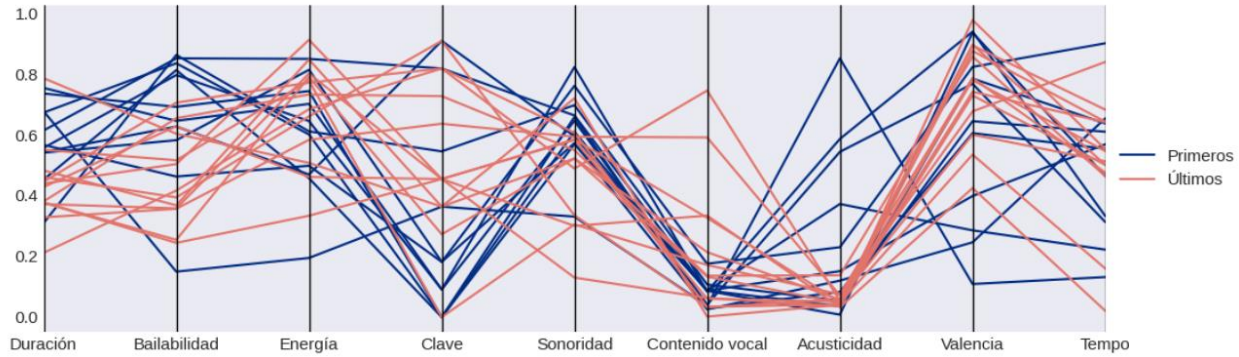

Figura 32. Canciones de todos los tiempos

Finalmente, el tercer criterio repite las tendencias encontradas anteriormente en las variables de sonoridad y contenido vocal, pero con excepciones en la bailabilidad y la energía. En estos dos casos, los rangos de aparición de canciones primeras del ranking se expanden. La valencia toma valores muy amplios.

Luego del análisis podría decirse que las dos variables claves son la sonoridad y el contenido vocal.

Pregunta 2

En la figura 33 se observa la frecuencia con la que aparecen los géneros en todo en ranking, mientras que la figura 34 solamente muestra los géneros predominantes en el top 10.

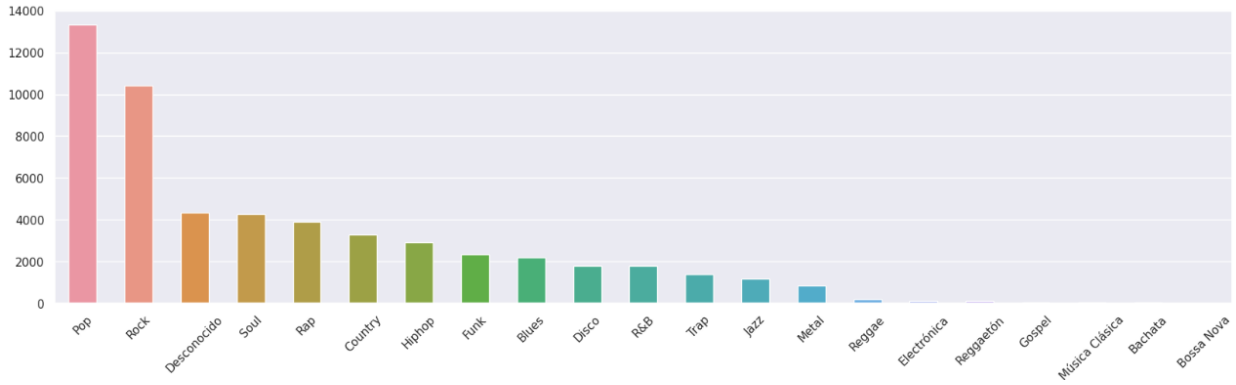

Figura 33.

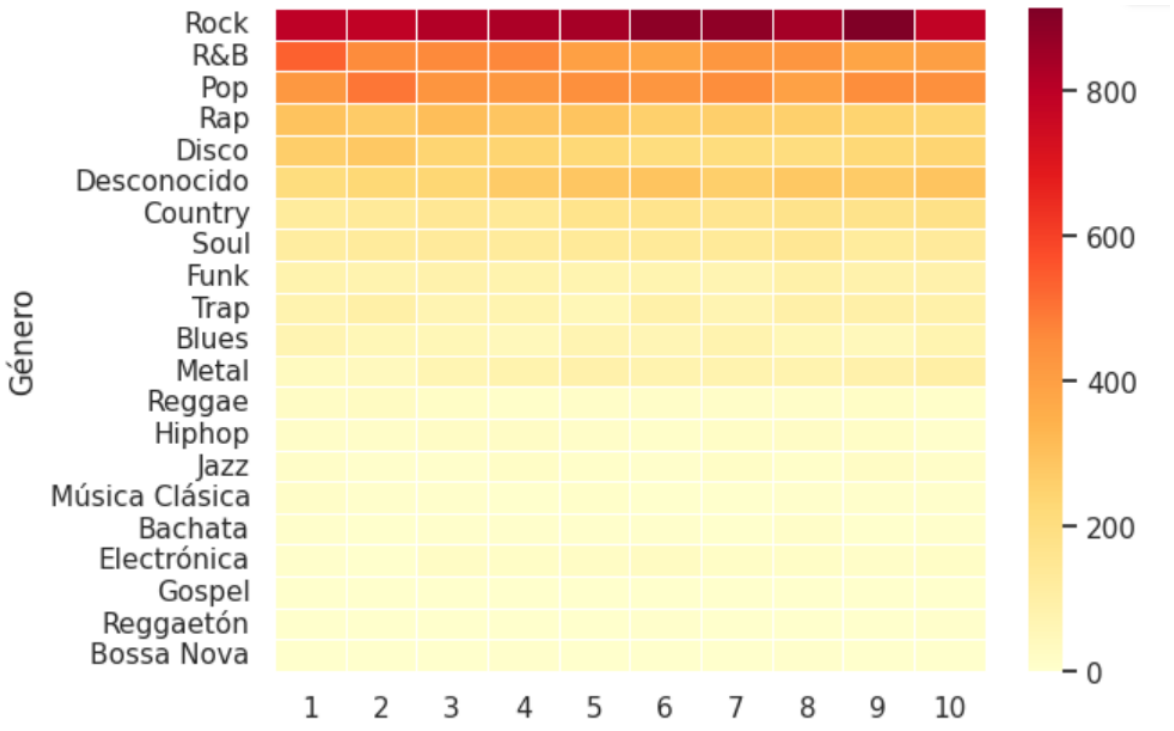

Figura 34. Géneros predominantes en el top 10

La figura 33 indica que el género predominante es el rock, seguido por el R&B y el pop.

Finalmente, en la figura 34 se observa la relación entre los atributos musicales de las canciones y su género. Se ve que no se pueden distinguir patrones claros.

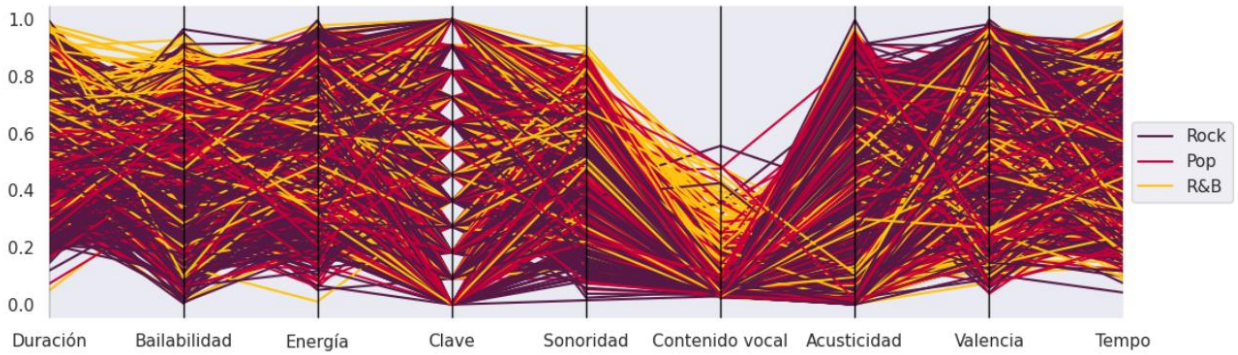

Figura 35. Diagrama de coordenadas paralelas mostrando la relación entre el género musical y los atributos cuantitativos