REPORTE FINAL ANALISIS DE DATOS

Introduccion a la Ciencia de Datos

Profesor
Jaime Alejandro Romero Sierra

Rey David Rojas Huerta
11/25/2024

Introducción del proyecto

Breve descripción de los objetivos del proyecto.

El objetivo principal del proyecto es analizar las características musicales y auditivas que influyen en la popularidad de una canción. A través de este análisis, el objetivo es identificar patrones que optimicen la creación, promoción y selección de canciones para listas de reproducción populares. En última instancia, el proyecto proporcionará conocimientos basados en datos que podrán utilizar artistas, productores y fabricantes.

Justificación y Contexto

En una industria musical cada vez más competitiva y basada en datos, la popularidad de una canción es un factor clave en su éxito comercial. Sin embargo, no siempre está claro qué factores contribuyen al éxito de una canción. Características como el género musical, el tono, la energía e incluso la estructura de los acordes juegan un papel clave en su aceptación por parte del público. Es importante abordar este tema porque

- Para artistas y productores: brinde orientación sobre qué elementos priorizar al hacer música para aumentar sus posibilidades de éxito.

- Para compañías discográficas: brindar análisis para sustentar decisiones estratégicas respecto de los géneros y estilos de música que se deben promover.

Este análisis puede transformar la forma en que se crean y promueven las canciones, alineando estrategias con los gustos y preferencias del público.

Fuentes de Datos

En este proyecto se empleó una base de datos de aproximadamente **45,195 canciones**, que contiene información detallada sobre las características musicales y auditivas de cada una. Esta base de datos fue obtenida de Kaggle, específicamente de la sección de musica. A continuación, se resumen sus principales características:

- **Origen de los Datos**: Los datos fueron recolectados de plataformas digitales que brinda todas las herramientas y recursos más importantes para progresar al máximo en data science.
- **Cantidad de Datos**: El dataset cuenta con registros de **45,195 canciones**, ofreciendo un panorama amplio y representativo de diversos géneros y estilos musicales.

- **Características Principales**:
    - **Categóricas**: Artista, nombre de la canción, género musical, tonalidad (key), modo (mode).
    - **Numéricas**: Popularidad, duración de la canción (duration_ms), energía, acústica (acousticness), bailabilidad (danceability), instrumentalidad (instrumentalness), vivacidad (liveness), volumen (loudness), hablabilidad (speechiness), tempo y valencia (valence).
    -

Metodología

Proceso de limpieza de datos: Este proceso se encuentra en el siguiente enlace

Análisis Exploratorio de Datos (EDA)

1. Descripción General de los Datos

Visión General

- **Resumen del Dataset:**
    - Número total de registros: 45,195 canciones.
    - Número total de columnas: 17 variables.
    - Sin valores nulos: Todas las columnas contienen datos completos (non-null).
    - Tipos de datos presentes: 6 categóricos y 11 numéricos.

Tipos de Variables

- **Variables Categóricas:**
    - artist_name: Nombre del artista o banda.
    - track_name: Nombre de la canción.
    - key: Tonalidad de la canción (e.g., Do mayor, Re menor).

- mode: Tipo de modo musical (mayor o menor).
- music_genre: Género musical.
- obtained_date: Fecha de obtención del dato (puede considerarse categórica si no se analiza como fecha).

- **Variables Numéricas:**
  - popularity: Popularidad de la canción.
  - acousticness: Nivel de acústica de la canción.
  - danceability: Qué tan bailable es la canción.
  - duration_ms: Duración en milisegundos.
  - energy: Nivel de energía de la canción.
  - instrumentalness: Probabilidad de que la canción sea instrumental.
  - liveness: Presencia de elementos en vivo.
  - loudness: Volumen medio en decibelios.
  - speechiness: Cantidad de palabras habladas en la canción.
  - tempo: Velocidad de la canción en BPM.
  - valence: Qué tan positiva o alegre es la canción.

Resumen Estadístico

	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence
count	45195.000000	45195.000000	45195.000000	4.519500e+04	45195.000000	45195.000000	45195.000000	45195.000000	45195.000000	45195.000000	45195.000000
mean	44.316753	0.305900	0.558362	2.216374e+05	0.600142	0.180900	0.194171	-9.037802	0.091843	119.880067	0.456941
std	15.171509	0.332173	0.173866	1.254725e+05	0.257952	0.316391	0.157867	6.017529	0.099471	28.374292	0.241051
min	0.000000	0.000001	0.059600	-1.000000e+00	0.000792	0.000000	0.009670	-47.046000	0.022300	34.347000	0.000000
25%	34.000000	0.023000	0.451000	1.777680e+05	0.448000	0.000000	0.098500	-10.585500	0.036600	97.953000	0.270000
50%	44.890299	0.169000	0.561083	2.236170e+05	0.625000	0.000315	0.132000	-7.227285	0.051100	119.957332	0.456337
75%	55.000000	0.520000	0.679000	2.657235e+05	0.807000	0.181246	0.233000	-5.279000	0.093250	137.653000	0.638000
max	96.000000	0.996000	0.980000	4.830606e+06	0.999000	0.996000	1.000000	1.949000	0.942000	220.041000	0.992000

Frecuencia de Categorías - key

G 11.476933%

C 10.925987%

C# 10.760040%

D 10.565328%

A 9.549729%

F 8.746543%

B 7.615887%

E 7.613674%

A# 6.704281%

G# 6.609138%

F# 6.210864%

D# 3.221595%

Frecuencia de Categorías - mode

Major 64.159752%

Minor 35.840248%

Frecuencia de Categorías - music_genre

Electronic 10.045359%

Rock 10.036508%

Jazz 10.018807%

Alternative 10.014382%

Blues 10.007744%

Classical 10.003319%

Hip-Hop 9.976767%

Rap 9.974555%

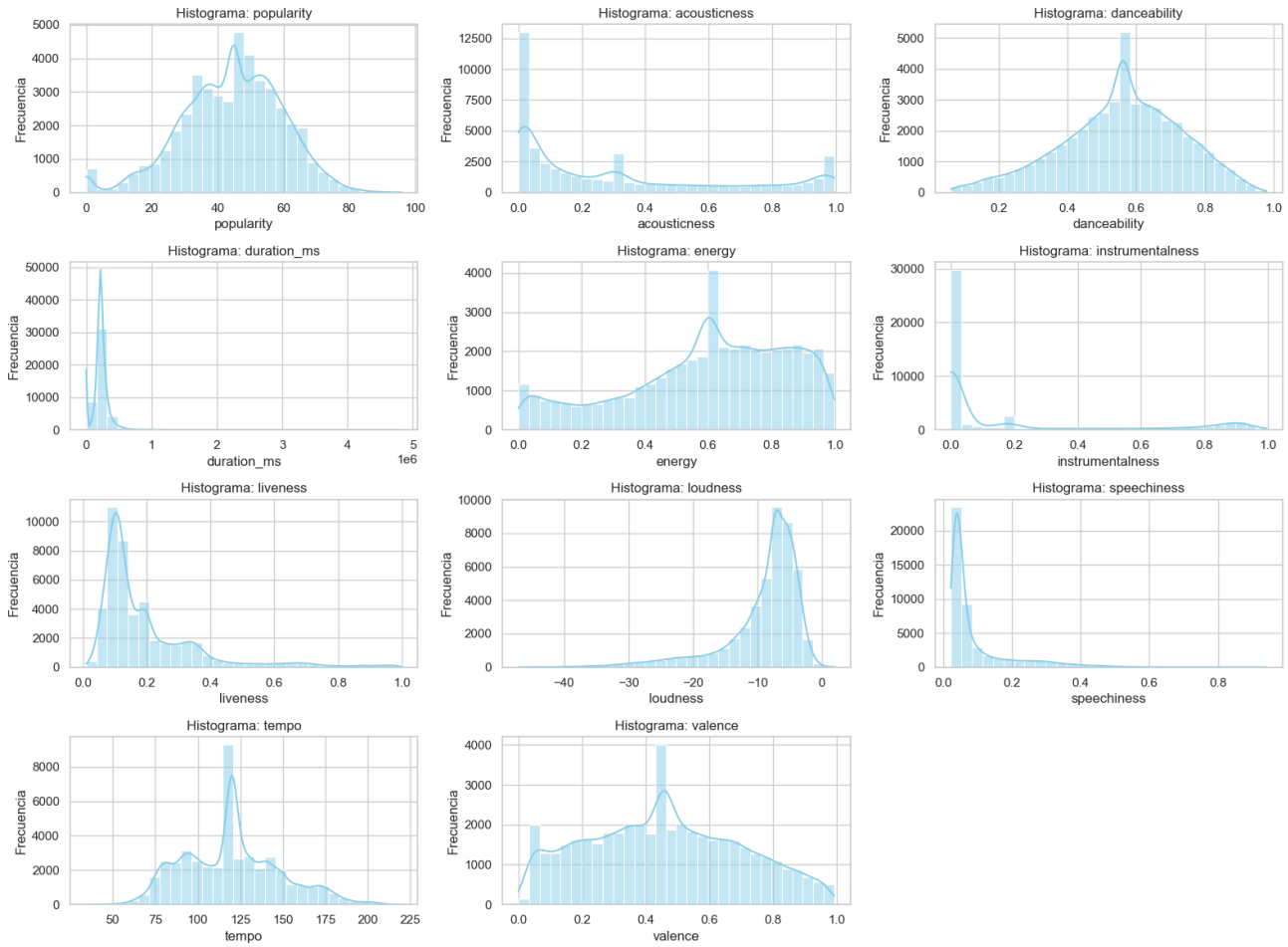
Country 9.972342%

Anime 9.950216%

2. Visualización y Distribución de Variables Individuales

Se puede apreciar una distribución normal y baja cantidad de datos atípicos promedio entre las gráficas, sin embargo, la duración, instrumentalness, liveness, loudness, speechiness y tempo cuentan con una gran cantidad de valores atípicos.

En el caso de la duración, en el análisis estadístico observamos un valor negativo, lo cual sería un error en el caso de tempo por sus cualidades no debería de variar demasiado, en cuanto a las otras mencionadas cada una cuenta con cierto sentido, por ejemplo liveness hace referencia a la presencia de audiencia durante la grabación, lo cual es absolutamente variable según el género de la canción, speechiness por su lado son la cantidad de 'palabras' por canción que igual según el género serán muchas, pocas o inclusive ninguna en cada canción.



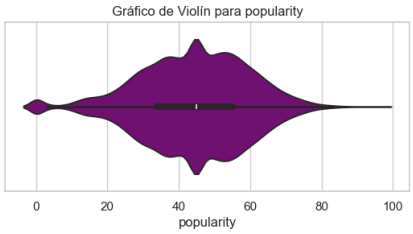


Gráfico de Violín para popularity
popularity

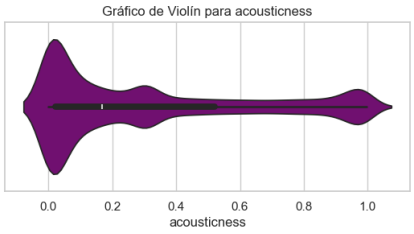


Gráfico de Violín para acousticness
acousticness

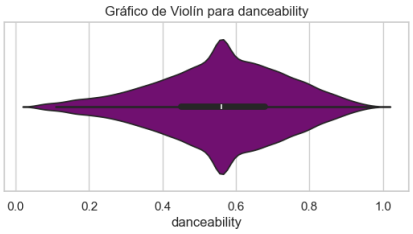


Gráfico de Violín para danceability
danceability

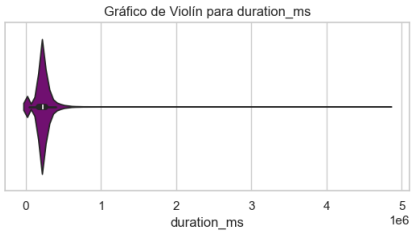


Gráfico de Violín para duration_ms
duration_ms
1e6

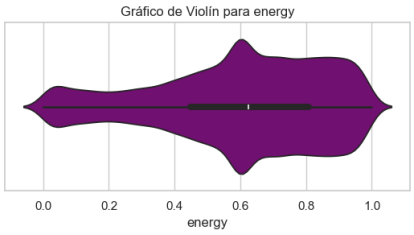


Gráfico de Violín para energy
energy

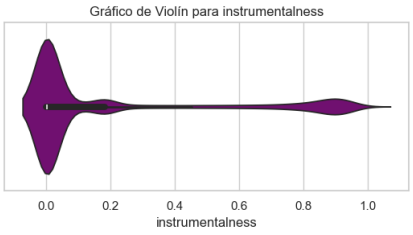


Gráfico de Violín para instrumentalness
instrumentalness

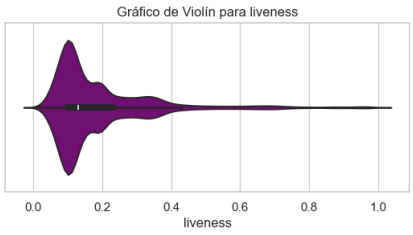


Gráfico de Violín para liveness
liveness

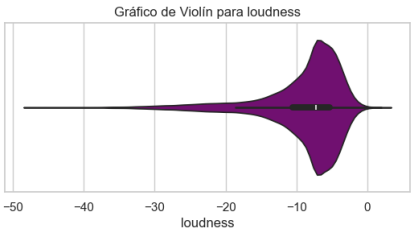


Gráfico de Violín para loudness
loudness

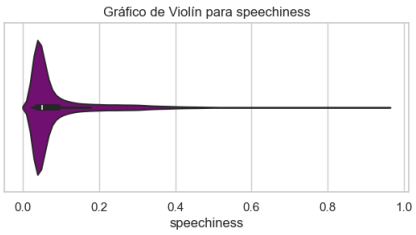


Gráfico de Violín para speechiness
speechiness

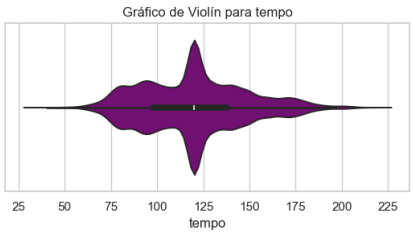


Gráfico de Violín para tempo
tempo

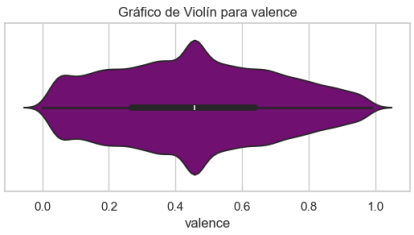
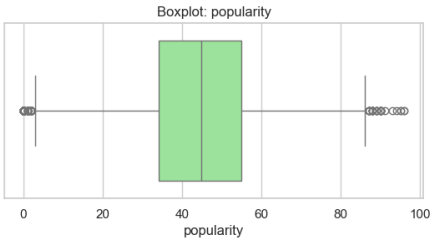
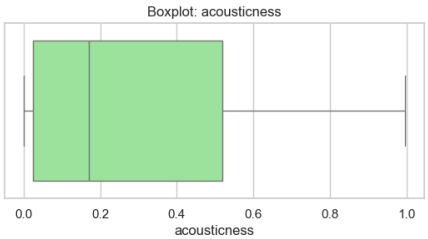
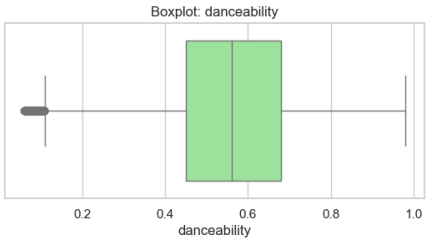
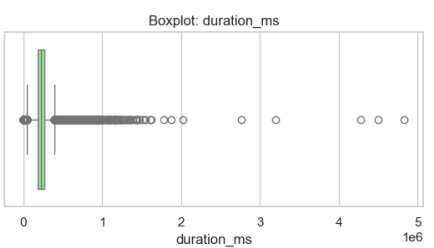


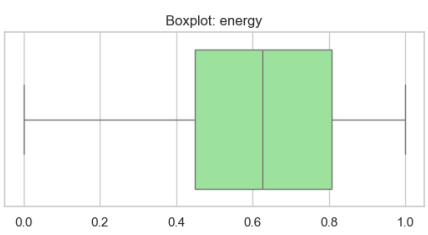
Gráfico de Violín para valence
valence

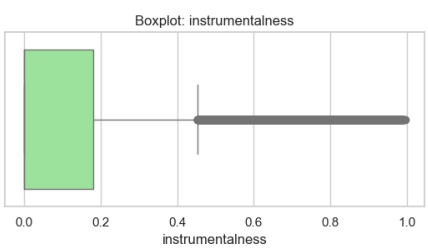

Boxplot: popularity
popularity

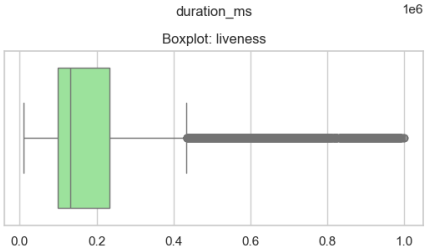

Boxplot: acousticness
acousticness

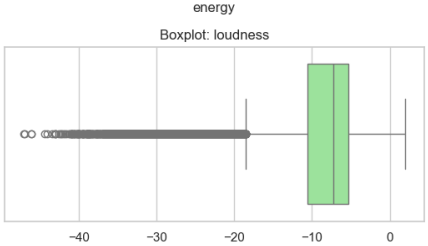

Boxplot: danceability
danceability

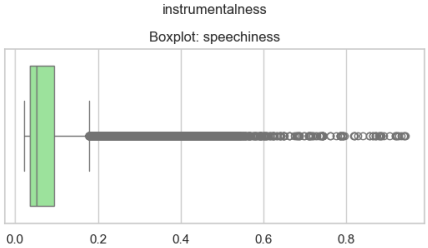

Boxplot: duration_ms
duration_ms
1e6

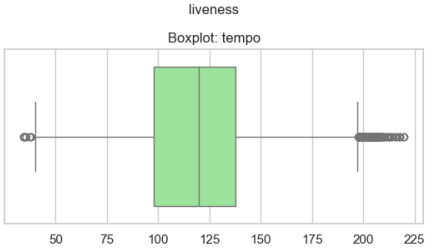

Boxplot: energy
energy

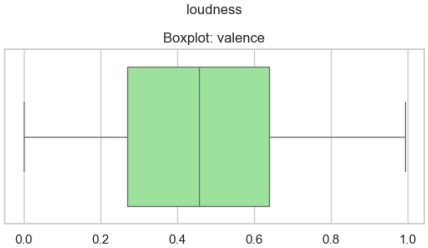

Boxplot: instrumentalness
instrumentalness

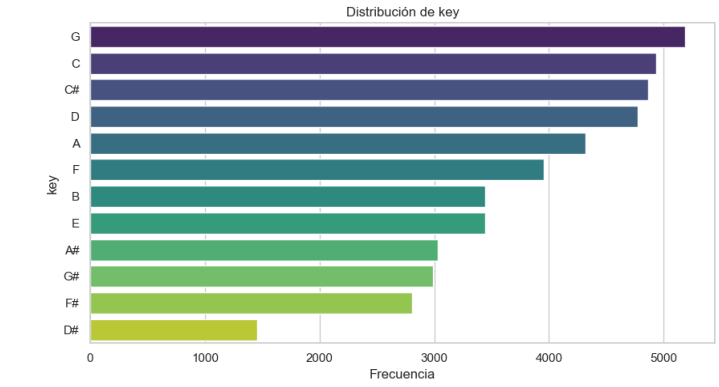

Boxplot: liveness
liveness

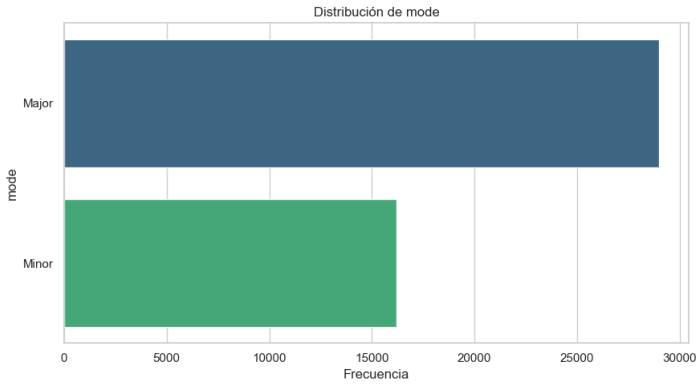

Boxplot: loudness
loudness

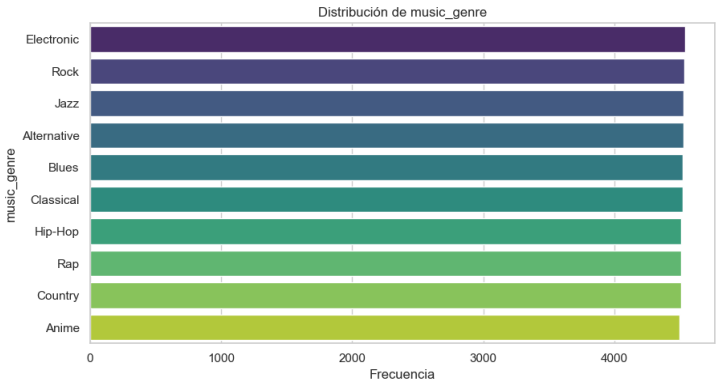

Boxplot: speechiness
speechiness


Boxplot: tempo
tempo


Boxplot: valence
valence







Como se observa en las gráficas de barras, la frecuencia de key es adecuada para cada una, a excepción de las canciones con que se encuentran en D# la cual es mucho más baja que las demás.

Para el mode las canciones predominantes son en una escala Mayor sobre la escala Menor.

Para music_gente en este caso se encuentran distribuido equitativamente.

3.Correlación entre Variables y Análisis de Hallazgos

Relación entre Variables Numéricas y Popularidad

- **energy**: Hay una **correlación positiva moderada** de **0.21** entre la energy y la popularity. Esto indica que, en general, las canciones con más energía tienden a ser más populares. Esta relación es consistente con la intuición de que las canciones más enérgicas, como las de géneros populares, pueden generar más engagement y ser más atractivas para un público amplio.
- **danceability**: También observamos una **correlación positiva moderada** de **0.34** entre danceability y popularity. Las canciones más fáciles de bailar tienden a ser más populares, lo cual puede ser un indicador de que los oyentes prefieren

canciones con ritmos pegajosos y contagiosos, un atributo común en géneros como el pop y el reggaetón.

- **loudness**: La loudness tiene una fuerte **correlación positiva** de **0.30** con popularity. Esto sugiere que las canciones más fuertes, en términos de volumen percibido, tienden a tener mayor popularidad. Esto es consistente con las características de producción de canciones populares, que a menudo priorizan una mezcla más intensa.

Variables con Baja Correlación con Popularidad

- **instrumentalness**: La instrumentalness tiene una **correlación negativa moderada** de **-0.35** con popularity. Esto sugiere que las canciones predominantemente instrumentales (sin voz) tienen menos popularidad, lo cual es consistente con la prevalencia de canciones con vocales prominentes en géneros populares.

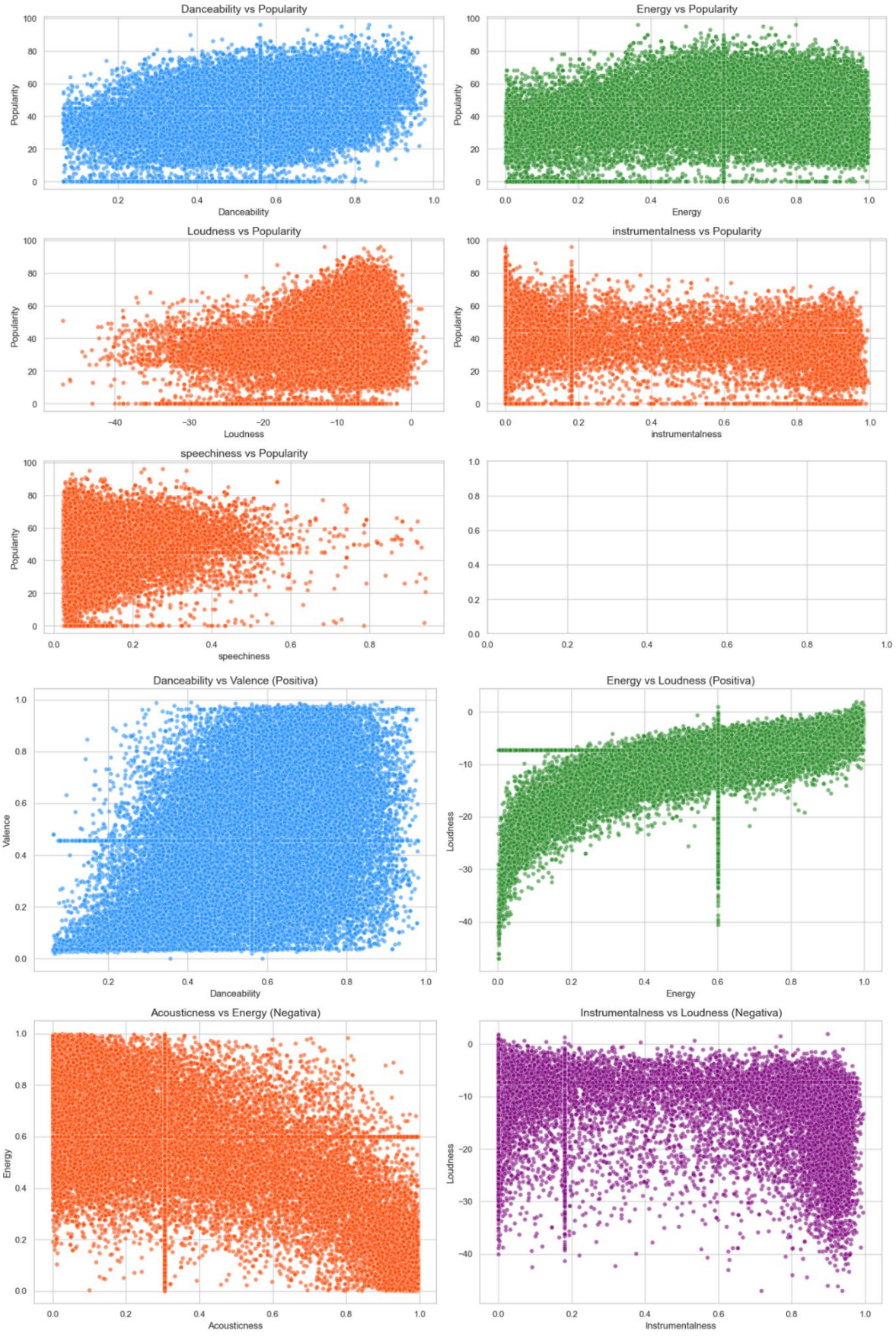
Relaciones Interesantes entre Otras Variables

- **energy y loudness**: Tienen una **correlación muy fuerte** de **0.80**, lo que indica que las canciones con más energía tienden a ser más ruidosas, lo cual es coherente con la producción de canciones populares que buscan captar la atención del oyente.
- **danceability y valence**: También existe una **correlación moderada positiva** de **0.41** entre estas dos variables, lo que implica que las canciones más bailables tienden a tener un tono emocional más positivo. Esto es típico en géneros que favorecen la alegría y el disfrute, como el pop y la música electrónica.
- **Acousticness y energy**: Una **correlación modera negativa de -0.75**, lo que implica que canciones más acústicas tienen a ser menos enérgicas. Por lo tanto, aunque no todas las canciones acústicas carecen de energía, las propiedades de las canciones más acústicas factores tiende a hacer que se perciban como más relajadas y menos energéticas.

Implicaciones para el Modelo

- Las variables energy, danceability, y loudness tienen correlaciones relativamente fuertes con la popularidad, lo que las convierte en buenas candidatas para incluir en el modelo de predicción.
- **acousticness** y **instrumentalness** podrían ser variables que requieren un tratamiento especial, ya que su relación negativa con la popularidad podría reflejar

la predominancia de géneros más electrónicos y con vocales prominentes en las canciones populares.



Canciones con mayor danceability y energy tienden a tener más popularidad, canciones con un loudness cercano a 0 suelen tener mayor popularidad, canciones con menor intrumentalness y speechiness tienden a tener mayor popularidad, aunque muy posiblemente varíe con el género, loudness y acousticness tienden a tener una relación positiva y negativa con energy respectivamente.

4.Análisis de Valores Atípicos (Outliers)

Se opto por métodos estadísticos como valores que superan los 1.5x el rango intercuartil en boxplots para identificar los Ouliers y estos fueron los resultados:

Variable: popularity

Cantidad de Outliers: 691

Rango: 2.50 a 86.50

Variable: danceability

Cantidad de Outliers: 337

Rango: 0.11 a 1.02

Variable: duration_ms

Cantidad de Outliers: 6582

Rango: 45834.75 a 397656.75

Variable: instrumentalness

Cantidad de Outliers: 8548

Rango: -0.27 a 0.45

Variable: liveness

Cantidad de Outliers: 3123

Rango: -0.10 a 0.43

Variable: loudness

Cantidad de Outliers: 3769

Rango: -18.55 a 2.68

Variable: speechiness

Cantidad de Outliers: 6476

Rango: -0.05 a 0.18

Variable: tempo

Cantidad de Outliers: 286

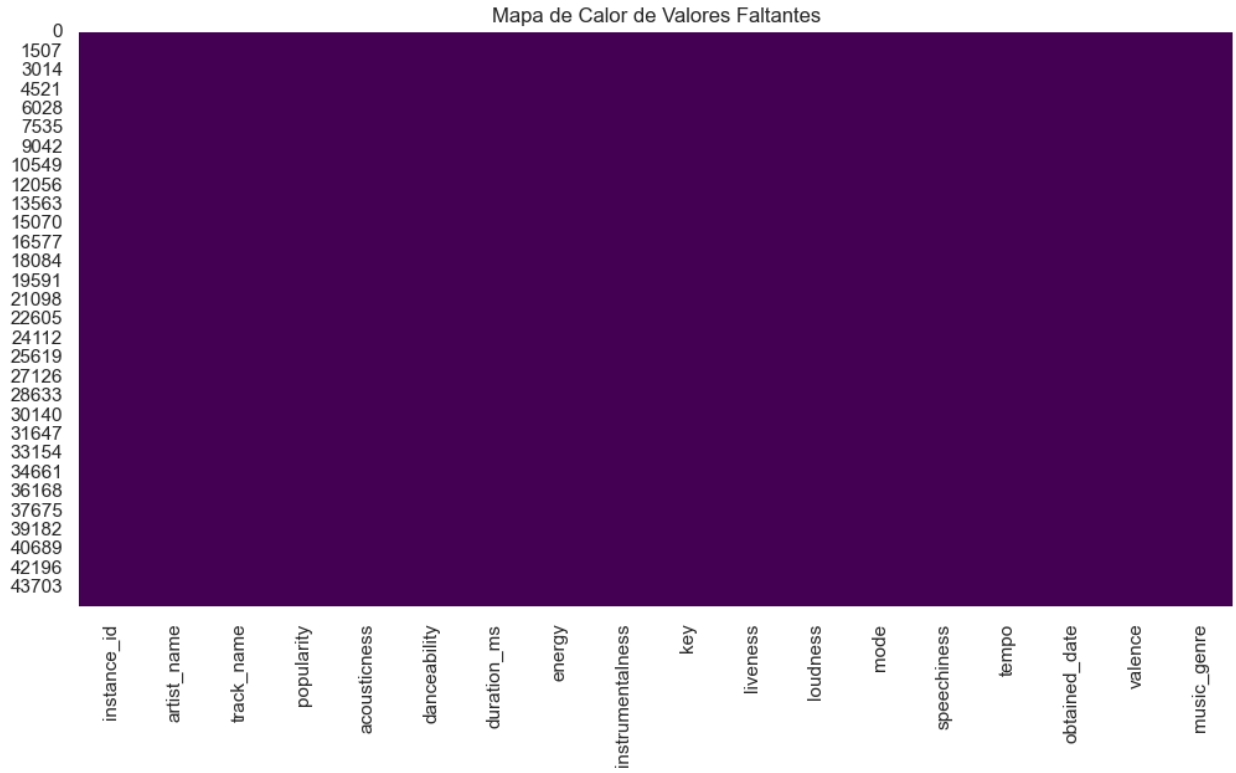
Rango: 38.40 a 197.20

Tratamiento de Outliers:

Solo se eliminaron los Outliers de las variables duration_ms. Tempo y danceability, ya que, como se mencionó anteriormente ciertas características auditivas tienden a tener un gran volumen de Outliers que al observarse en graficas de parejas de variables llegan a tener sentido.

5.Análisis de Valores Faltantes

No se cuenta con valores faltantes.



6. Relación entre Variables Categóricas y Numéricas

En cuanto a key no existe mucha variación entre las variables numéricas, a excepción de la clave de D# que suele tener una mínima variación respecto a las otras, muy posiblemente debido a que está presente en menor cantidad entre todas las canciones.

En cuanto a mode no existe variación notable entre las características de las canciones en tonalidad mayor o menor.

En music_genre se encuentra una gran variación importante entre los géneros, a continuación, se presentarán:

El género Anime y Classical tiende a ser menos popular.

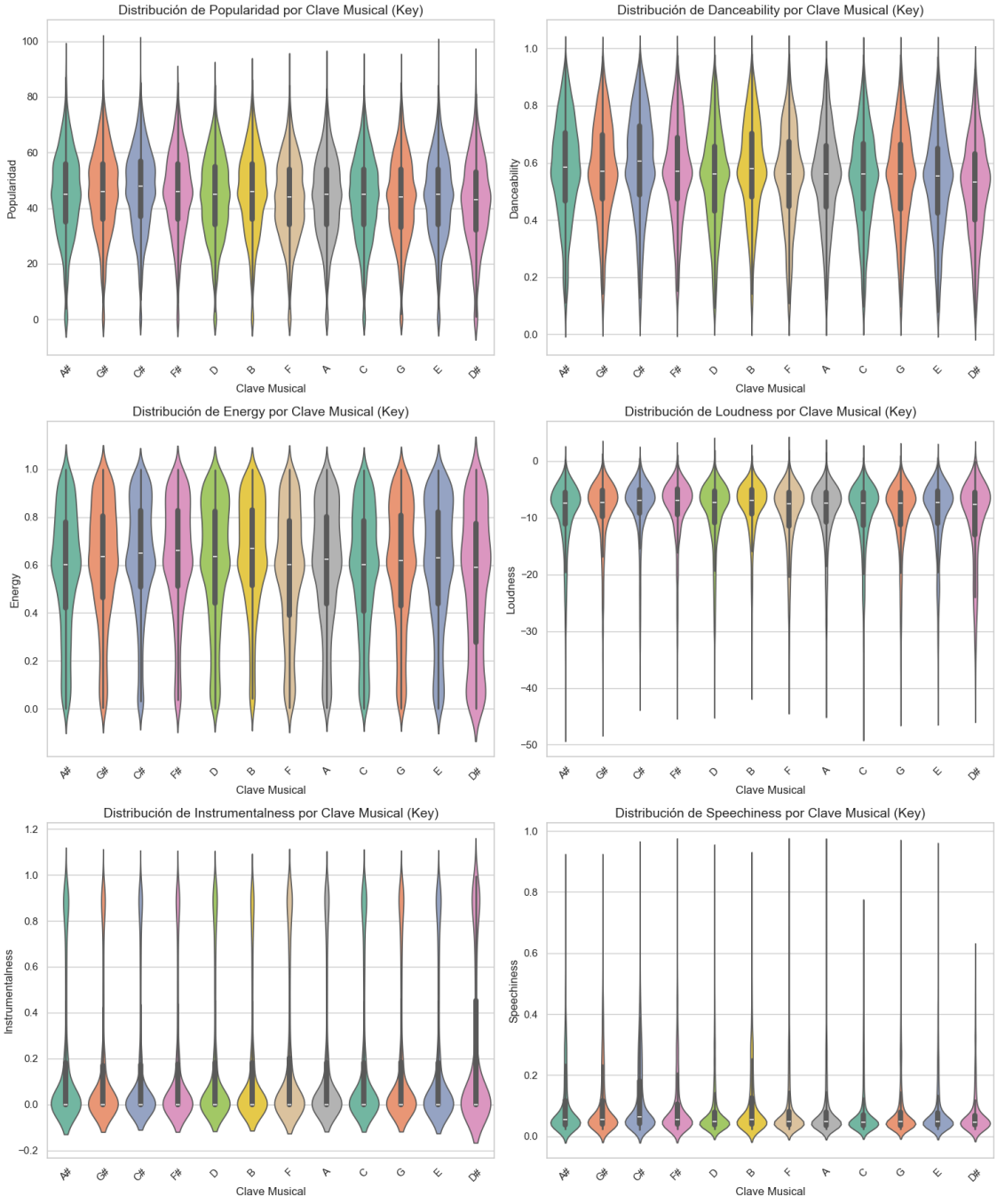
El género Anime, Jazz y Alternative tienen una menor danceability respecto a las otras, y Classical una danceabilidad muy baja.

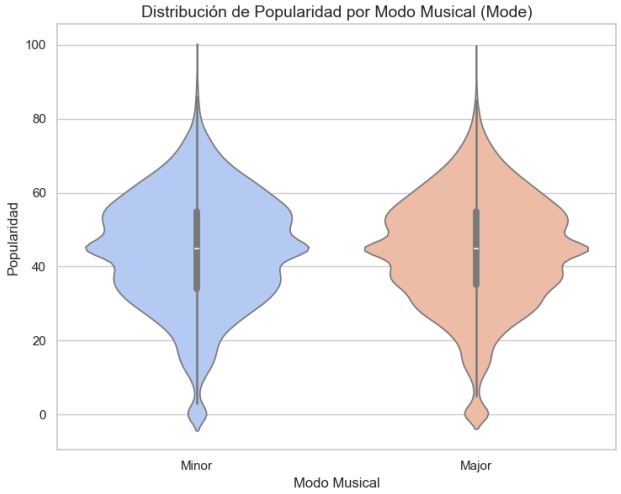
El género Jazz y Blues tienen una menor energía respecto a las otras, y Classical una energía muy baja.

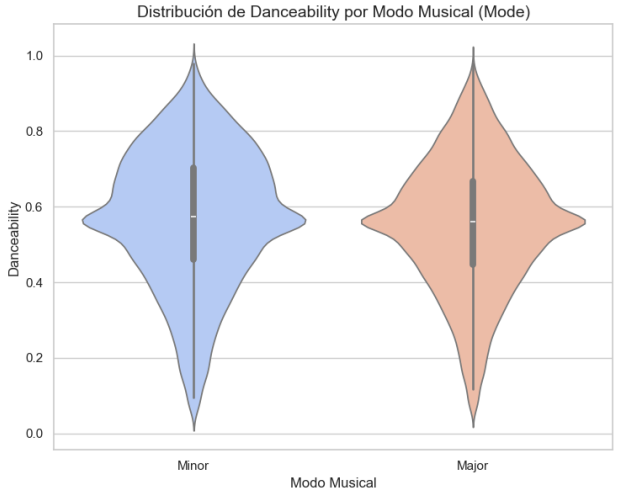
El loudness en el género Classical tiene a ser muy negativo.

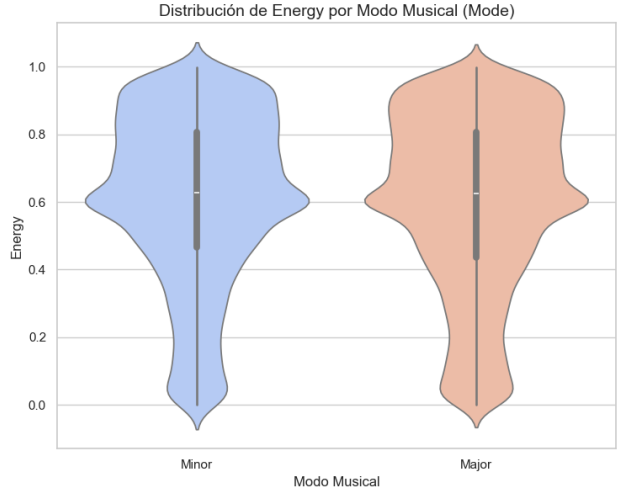
El instrumentalness en los géneros Electronic, Anime, Jazz y sobre todo Classical suele en gran medida mayor a los otros géneros.

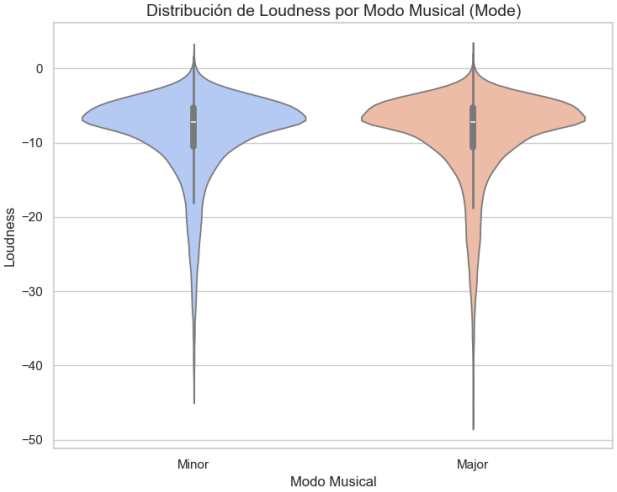
El speechiness en los géneros Rap e Hip-Hop es mayor a todos los demás géneros.

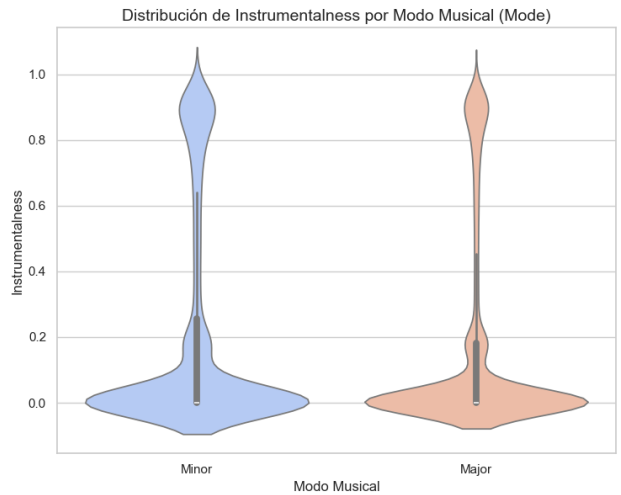


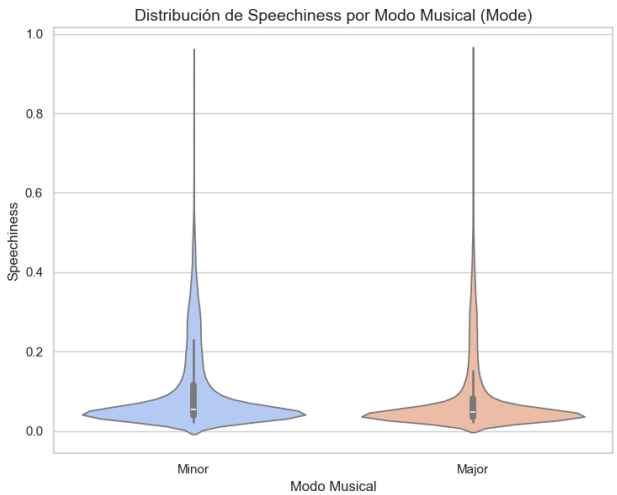

Distribución de Popularidad por Modo Musical (Mode)
Popularidad
Modo Musical
Minor
Major

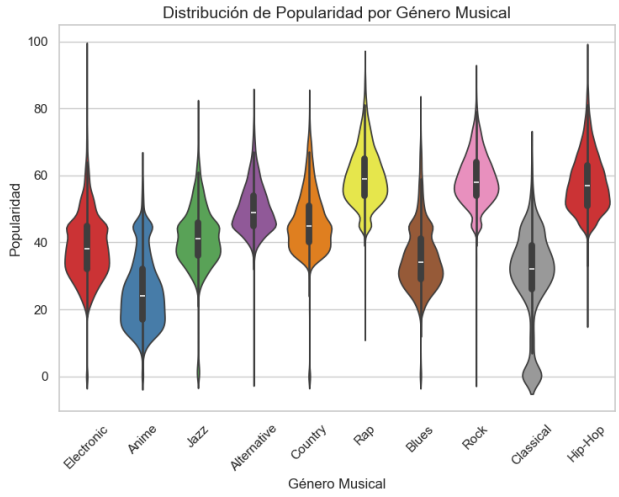

Distribución de Danceability por Modo Musical (Mode)
Danceability
Modo Musical
Minor
Major

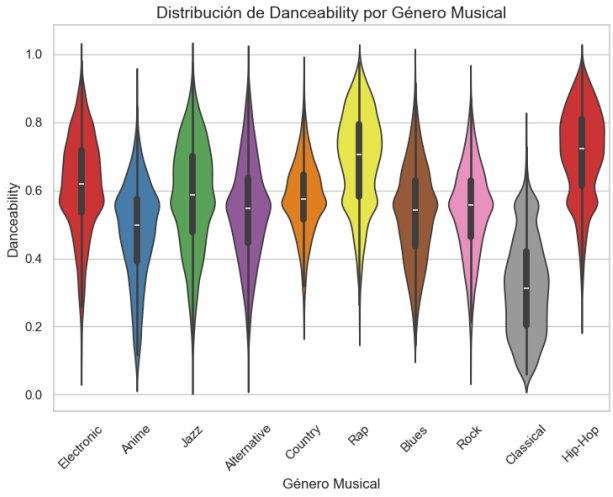

Distribución de Energy por Modo Musical (Mode)
Energy
Modo Musical
Minor
Major

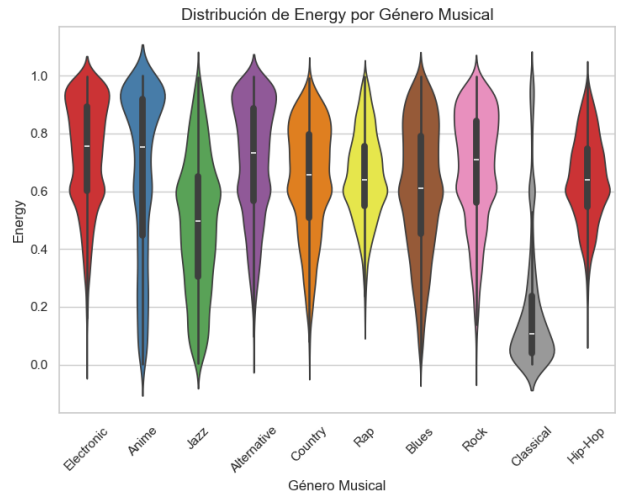

Distribución de Loudness por Modo Musical (Mode)
Loudness
Modo Musical
Minor
Major

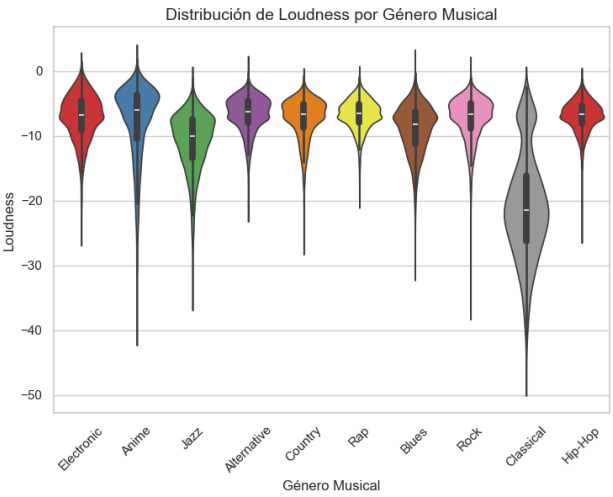

Distribución de Instrumentalness por Modo Musical (Mode)
Instrumentalness
Modo Musical
Minor
Major

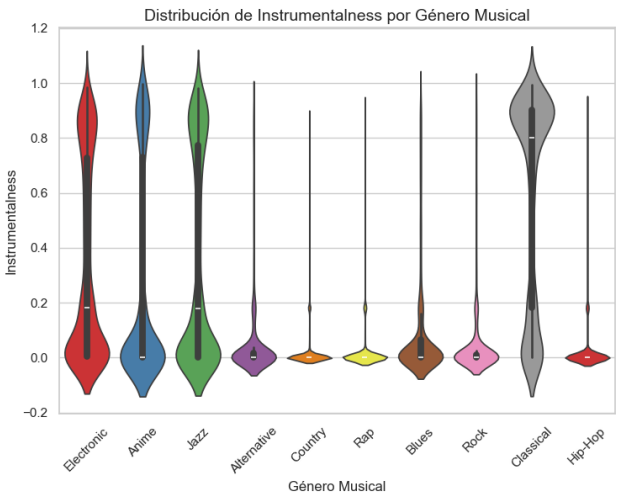

Distribución de Speechiness por Modo Musical (Mode)
Speechiness
Modo Musical
Minor
Major

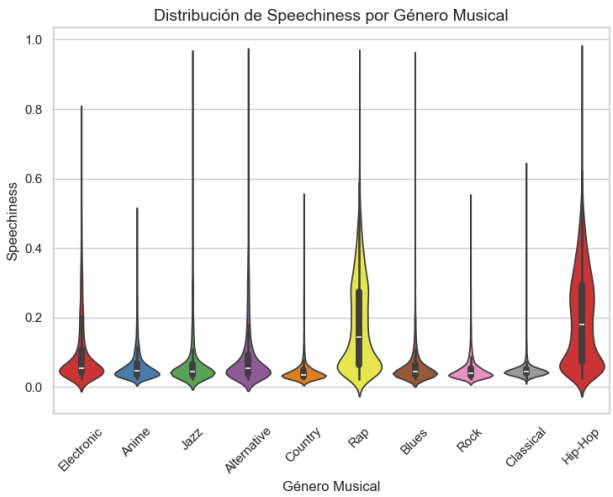

Distribución de Popularidad por Género Musical
Popularidad
Género Musical
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop


Distribución de Danceability por Género Musical
Danceability
Género Musical
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop


Distribución de Energy por Género Musical
Energy
Género Musical
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop


Distribución de Loudness por Género Musical
Loudness
Género Musical
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop


Distribución de Instrumentalness por Género Musical
Instrumentalness
Género Musical
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop


Distribución de Speechiness por Género Musical
Speechiness
Género Musical
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop

7. Observaciones y Hallazgos Importantes

Variable Objetivo: Popularidad (popularity)

La variable que se busca predecir es la **popularidad de la canción**, que está influenciada principalmente por las siguientes variables:

- **Energy**: Correlación positiva moderada (0.21), canciones más energéticas tienden a ser más populares.
- **Danceability**: Correlación positiva moderada (0.34), las canciones más bailables tienden a ser más populares.
- **Loudness**: Correlación fuerte (0.30), canciones más ruidosas tienen mayor popularidad.
- **Instrumentalness**: Correlación negativa (-0.35), las canciones instrumentales son menos populares.

Hallazgos Clave

- **Energy y Loudness** tienen una correlación muy fuerte (0.80), indicando que las canciones energéticas suelen ser más ruidosas.
- **Danceability y Valence** tienen correlación positiva (0.41), sugiriendo que canciones bailables tienden a ser alegres.
- **Anomalías**: Géneros como **Anime** y **Classical** tienen menor popularidad y características auditivas distintas, como baja energía y danceability.

Anomalías y Patrones Inesperados:

- En el análisis de las variables **duration_ms**, **liveness**, **speechiness**, **tempo**, y **instrumentalness**, se observa una gran cantidad de outliers. Esto se puede interpretar como que ciertos géneros, como el jazz o el clásico, tienden a tener características únicas que no se ajustan a la norma de la mayoría de las canciones populares.
- La variable **D#** en la tonalidad (key) tiene una frecuencia mucho más baja que otras tonalidades, lo que puede indicar que ciertos tonos son menos utilizados en canciones populares.

Implicaciones de Género:

- Los géneros como **Anime**, **Classical**, y **Jazz** muestran una **baja popularidad**, lo que podría deberse a la menor accesibilidad de estos géneros para el público general.
- **Rap** y **Hip-Hop** tienen una alta **speechiness**, lo que es consistente con las características de estos géneros, que a menudo incluyen letras rápidas y complejas.

Implicaciones para el Modelo

- **Variables clave**: **Energy**, **Danceability**, **Loudness**, y **Valence** son las más relevantes para el modelo.
- **Tratamiento de Outliers**: Algunos outliers son válidos, especialmente para géneros específicos, y deben manejarse adecuadamente.
- **Modelo multivariante**: Modelos como **Random Forest** o **Árbol de decisiones** serían adecuados para capturar las interacciones entre variables, ya que muchas están correlacionadas.

Modelo de Machine Learning

Modelo elegido: **Random Forest Regressor**
Tipo de modelo: **Regresión**

El modelo **Random Forest** es un algoritmo de aprendizaje automático supervisado que utiliza un conjunto de árboles de decisión para realizar predicciones. En lugar de depender de un solo árbol de decisión, el Random Forest construye múltiples árboles y combina sus predicciones para obtener un resultado más preciso y robusto. Este enfoque reduce el riesgo de sobreajuste (overfitting) que es común en los árboles de decisión individuales, y generalmente ofrece un rendimiento mejorado.

Justificación

Elegimos el **Random Forest Regressor** debido a las siguientes razones:

- **Capacidad de manejo de datos complejos**: Dado que nuestro conjunto de datos tiene tanto variables numéricas como categóricas, y debido a la no linealidad de las relaciones entre las características, un **Random Forest** es adecuado ya que puede manejar relaciones no lineales y variables de diferentes tipos.

- **Robustez ante sobreajuste**: Los modelos como el árbol de decisión pueden ser muy sensibles a pequeñas variaciones en los datos de entrenamiento, lo que puede llevar al sobreajuste. El Random Forest, al promediar múltiples árboles, ayuda a mitigar este riesgo.
- **Importancia de características**: Este modelo nos permite entender cuál es la importancia relativa de cada característica (como music_genre y tempo), lo cual es valioso en un análisis exploratorio de datos.
- **Rendimiento general**: Random Forest ha mostrado un buen rendimiento en una amplia variedad de problemas de predicción y es relativamente fácil de implementar y ajustar.

Implementación y Entrenamiento

División de los datos:
Primero, se divide el conjunto de datos en dos partes: **entrenamiento** y **prueba**. Usamos el 80% de los datos para entrenar el modelo y el 20% restante para evaluar su rendimiento.

Métricas usadas para evaluar el modelo:
Para evaluar el modelo de regresión, utilizamos las siguientes métricas:

- **MAE (Error Absoluto Medio)**: Mide la magnitud promedio de los errores entre las predicciones y los valores reales.
- **RMSE (Raíz del Error Cuadrático Medio)**: Es más sensible a los grandes errores y proporciona una medida en las mismas unidades que los datos.
- **R^2 (Coeficiente de Determinación)**: Mide qué tan bien el modelo explica la variabilidad de los datos. Un valor de R^2 cercano a 1 indica que el modelo explica la mayor parte de la variabilidad.
-

Resultados

Métricas de rendimiento:

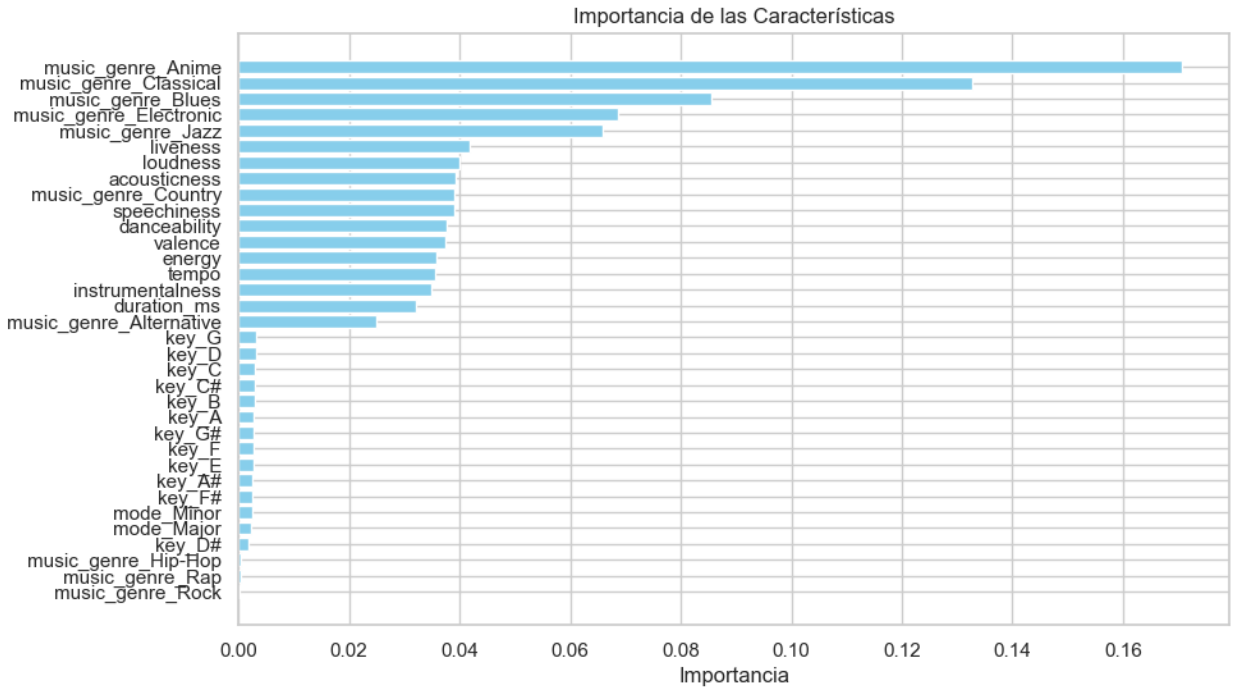
Al evaluar el modelo en el conjunto de prueba, obtenemos los siguientes resultados:

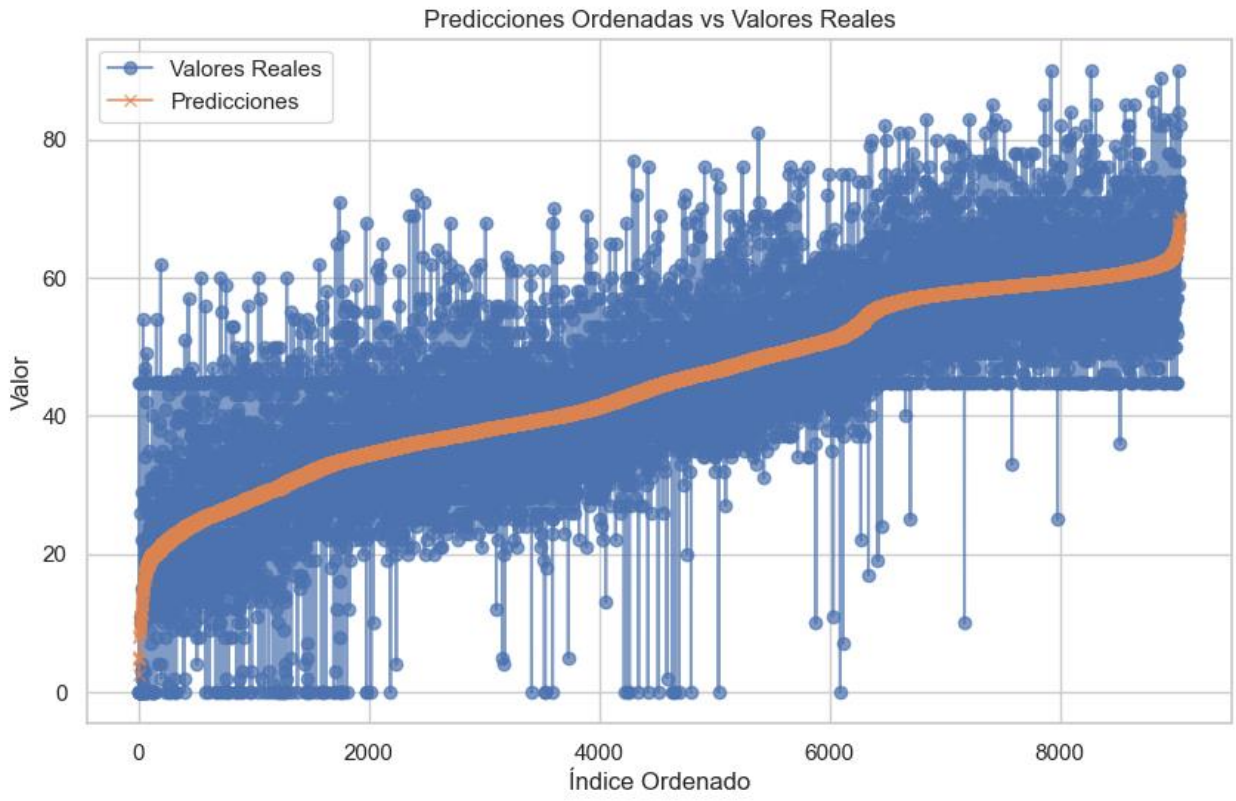
- **MAE (Error Absoluto Medio)**: **7.0162**
- **RMSE (Raíz del Error Cuadrático Medio)**: **9.2546**

- **R^2 (Coeficiente de Determinación): 0.6243**
- **Score promedio de la validación cruzada: 117.4901**

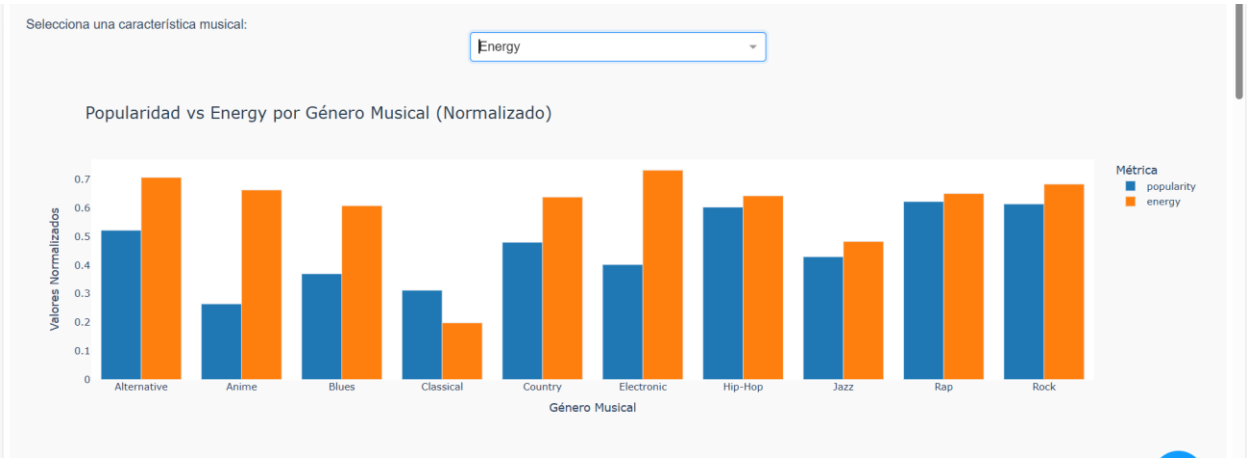
Interpretación de las métricas:

- **MAE de 7.0162**: El modelo tiene un error promedio de aproximadamente 7 unidades al predecir la popularidad de las canciones. Esto indica que las predicciones son razonablemente precisas.
- **RMSE de 9.2546**: La raíz del error cuadrático es de aproximadamente 9, lo que sugiere que algunas predicciones están algo alejadas de los valores reales, pero aún dentro de un margen razonable para este tipo de problema.
- **R^2 de 0.6243**: El modelo explica aproximadamente el 62.43% de la variabilidad en los datos. Aunque no es un valor extremadamente alto, indica que el modelo tiene una capacidad moderada para predecir la popularidad.
- **Validación cruzada**: El **score promedio de la validación cruzada** de 117.4901 muestra que el modelo mantiene un rendimiento consistente en diferentes subconjuntos de los datos.

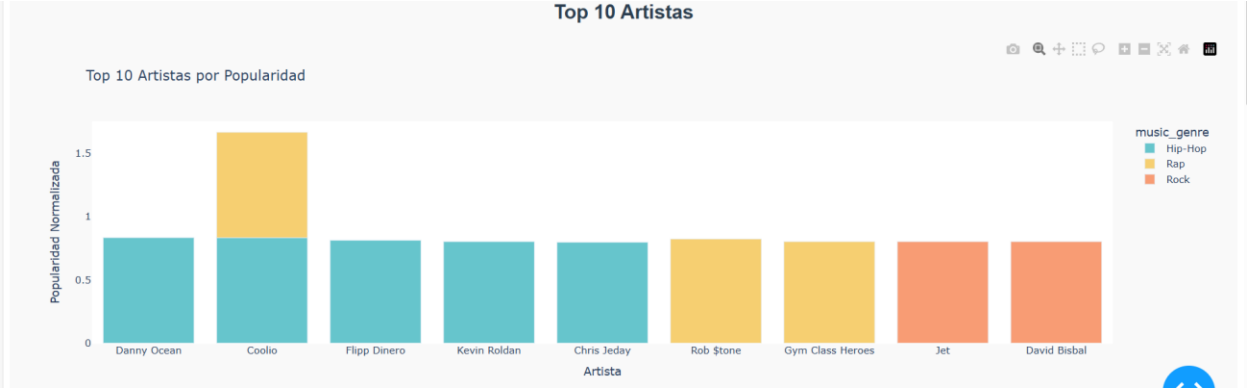




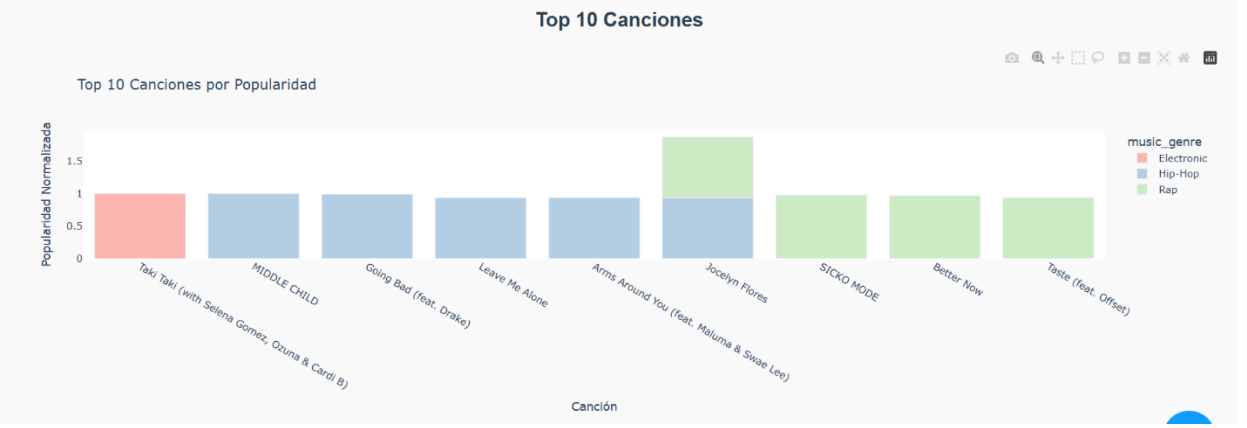
Dashboard



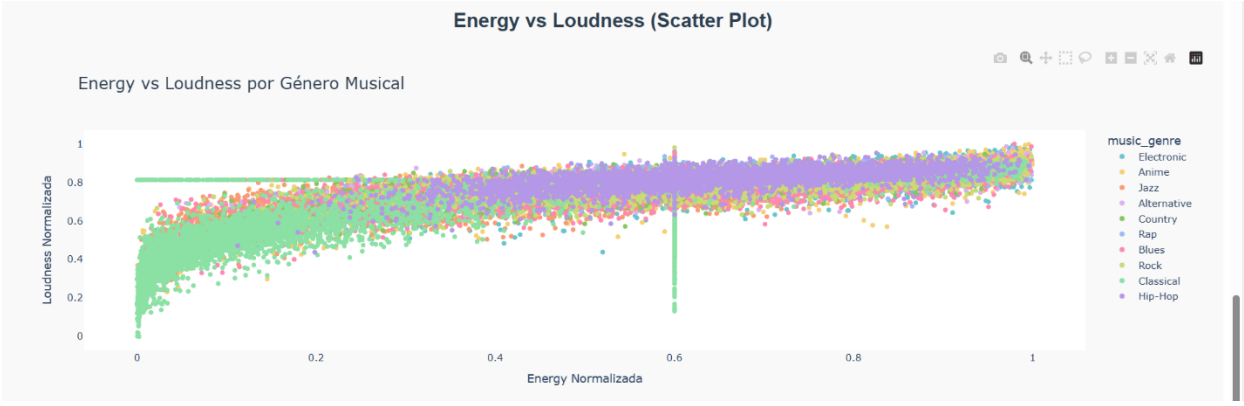
Este primer dashboard permite al usuario observar la popularidad y una característica musical respecto a cada genero dentro de mi dataframe, de este modo los usuarios a la vez pueden ver la relación de la característica respecto a la popularidad, sin necesidad de usar gráficos en exceso.



Este segundo dashboard permite a los usuarios visualizar los top 10 artistas con mayor promedio de popularidad, a la vez que el genero que mayoritariamente representan estos artistas, inclusive aquellos que llegan a representar dos géneros musicales a la vez.



En este tercer dashboard observamos las top 10 canciones con mayor popularidad, igualmente que en el caso de los artistas se presenta mediante colores el genero de la canción, igualmente existiendo canciones con doble género musical.



Por ultimo estos dos dashboard permiten a los usuarios observar la correlaciones mas importantes, la cual es respecto a la variable energy, es un manera óptima de obsérvala entendible para la gran mayoría de usuarios.

Uso y beneficios:

Los dashboard son una gran manera de permitir a los usuarios interactuar con el entorno de las gráficas, una de las partes mas importantes de un análisis y lo mejor es que son bastantes amigables con los usuarios generando un entorno de fácil entendimiento, y para los mas expertos entender el problema o objetivo con solo verlas.

Conclusiones y Futuras Líneas de Trabajo

Conclusiones

- **Resumen de los hallazgos principales:**

    1. Se identificaron las características musicales que influyen significativamente en la **popularidad** de las canciones. Por ejemplo, variables como instrumentalness, music_genre y energy demostraron una fuerte correlación con niveles altos de popularidad.
    2. El análisis categórico mostró que ciertos géneros musicales, como [inserta géneros destacados], tienden a agrupar canciones más populares, mientras que otros no parecen tener un impacto tan relevante.
    3. La combinación de visualizaciones como los gráficos permitió explorar la relación de cada variable con la popularidad de forma más clara y orientada a usuarios no técnicos.

- **Cumplimiento de los objetivos iniciales:** El proyecto logró cumplir con su principal objetivo: **identificar las características auditivas y musicales que influyen en la popularidad de las canciones**. Los resultados pueden ser utilizados como una base para optimizar estrategias de producción musical y curación de listas de reproducción.

Futuras Líneas de Trabajo

1. **Mejoras a los datos:**
    - Ampliar la base de datos para incluir registros más actuales o datos históricos con rangos de tiempo específicos, lo que podría permitir un análisis de tendencias.
    - Incorporar más características contextuales como la ubicación geográfica de los oyentes, plataformas donde se consumen las canciones y duración de las pistas.
    - Asegurar la calidad de las etiquetas, especialmente en géneros y datos faltantes, ya que los errores en estas áreas pueden introducir sesgos en los resultados.
2. **Optimización del modelo:**
    - Experimentar con modelos más avanzados, como redes neuronales o modelos de clasificación más complejos (*ensemble methods*), para mejorar la precisión en la predicción.
    - Realizar un análisis de *feature importance* más detallado, para evaluar cómo interactúan las características entre sí y si existen relaciones no lineales que el modelo actual no detectó.
    - Implementar técnicas de selección de características para reducir la dimensionalidad y evitar el *overfitting*.
3. **Visualizaciones más efectivas:**
    - Desarrollar dashboards interactivos para facilitar la comprensión y exploración de datos a usuarios finales.
    - Incorporar gráficos avanzados como *facet grids* o análisis de palabras clave para explorar cómo los títulos de las canciones influyen en su popularidad.

- Crear visualizaciones específicas para comparar la popularidad entre diferentes décadas o regiones.

4. **Extensión del análisis:**
    - Incluir un análisis de predicción temporal, explorando cómo las características musicales cambian su relación con la popularidad a lo largo del tiempo.
    - Ampliar el alcance del análisis hacia predicciones más específicas, como identificar *hits* potenciales antes de su lanzamiento.

Referencias

La base de datos fue obtenida de Kaggle:

Prediction of music genre. (2021, 2 noviembre). Kaggle.

https://www.kaggle.com/datasets/vicsuperman/prediction-of-music-genre/data