

UNIVERSIDAD AMERICANA

Actividad: Analisis de Datos de Fuentes Diversas  
­­­­­­Asignatura: Inteligencia de Negocios  
Grupo:

Docente:

Emiliano Quintero Sánchez  
Cif: 19014233  
18/09/2024

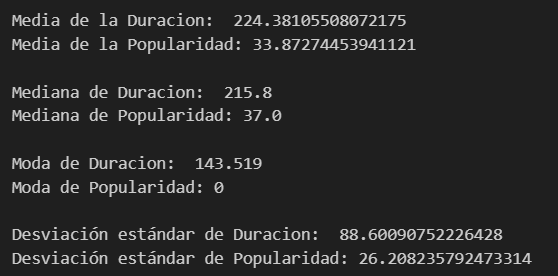
# Recopilacion y Almacenamiento de Datos

Para este ejercicio, decidí usar mis listas de reproducción de Spotify. Estos datos los saque de una página web llamada [Exportify](https://exportify.net). Estos datos exportados desde la cuenta de spotify y convertidos a un csv, el cual almacene en [GitHub](https://github.com/EQuiSan/AnalisisDeFuentesDiversas). La página exporta el siguientes dataset:  
  
Track ID  
Track Name  
Album Name  
Artist Name(s)  
Release Date  
Duration (ms)  
Popularity  
Added By  
Added At  
Genres  
Record LabelDanceability  
Energy  
Key  
Loudness  
Mode  
Speechiness  
Acousticness  
Instrumentalness  
Liveness  
Valence  
Tempo  
Time Signature

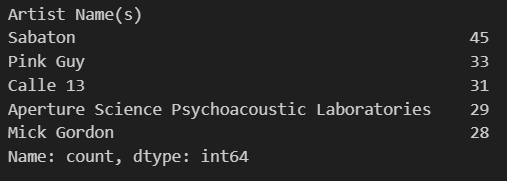
Todos esos datos están guardados en formato CSV con un total de 1053 filas y 23 columnas. Están separados por comas. Inmediatamente se puede notar que algunas columnas tienen datos faltantes que serán llenados.

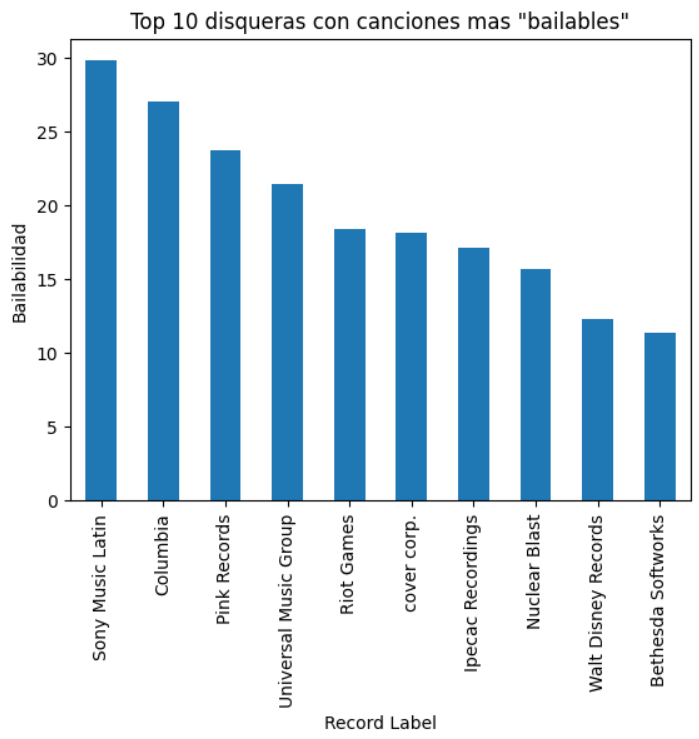
# Análisis Descriptivo

Aquí calcule y presente las estadísticas básicas (media, mediana, moda, desviación estándar) de las variables numerales más relevantes, siendo estas Duration y Popularity.

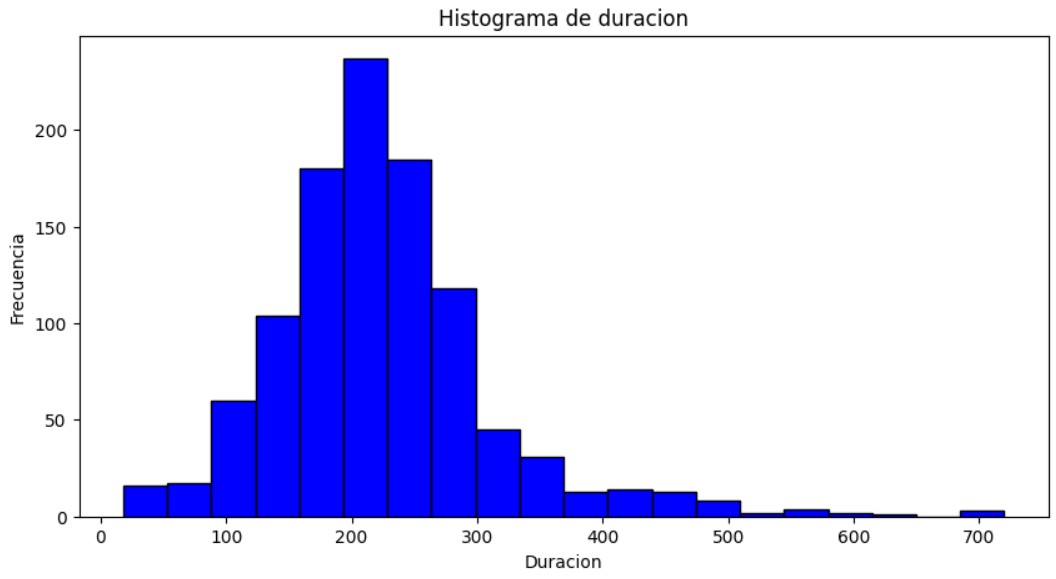


Se creó una tabla de frecuencia de los artistas con más canciones dentro del dataset. En este se observa que el más recurrente es la banda Sabaton, con 45 canciones.



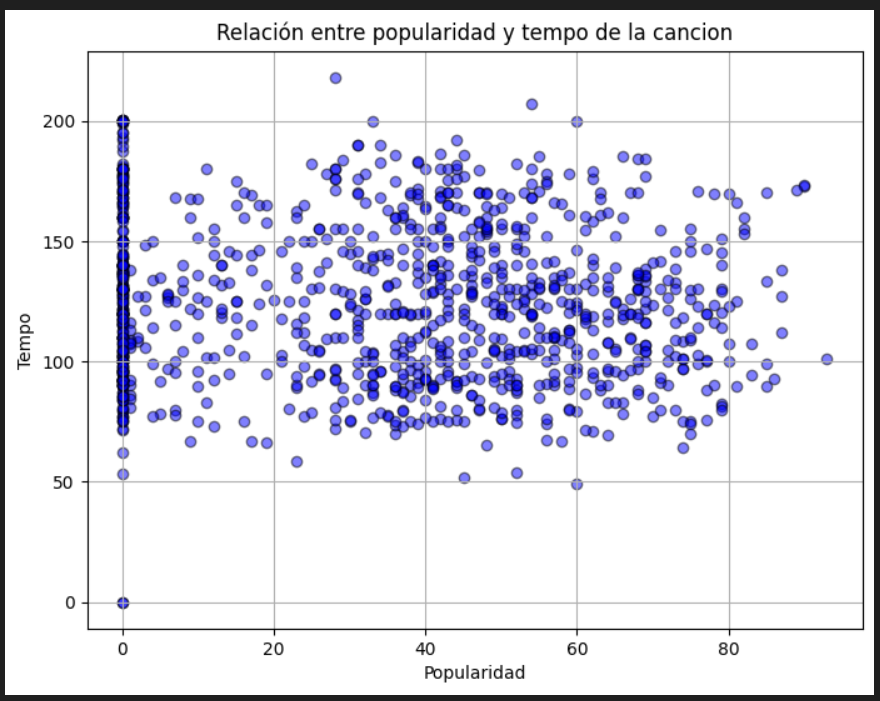
El siguiente grafico de barras representa las 10 disqueras que lanza canciones más “bailables” de mi biblioteca de canciones según las métricas de Spotify. Se puede notar como Sony Music Latin es la que ocupa el puesto más alto, seguido de Columbia y Pink Records  


Este histograma representa las duraciones más frecuentes en segundos de las canciones. Se puede apreciar como la duración promedio anda por los 220-240 segundos, lo cual confirma las métricas que fueron sacadas anteriormente (media de 224.381 segundos). Se puede notar un par de anomalías, con un par de canciones que llegan a los 550 y hasta 700 segundos.

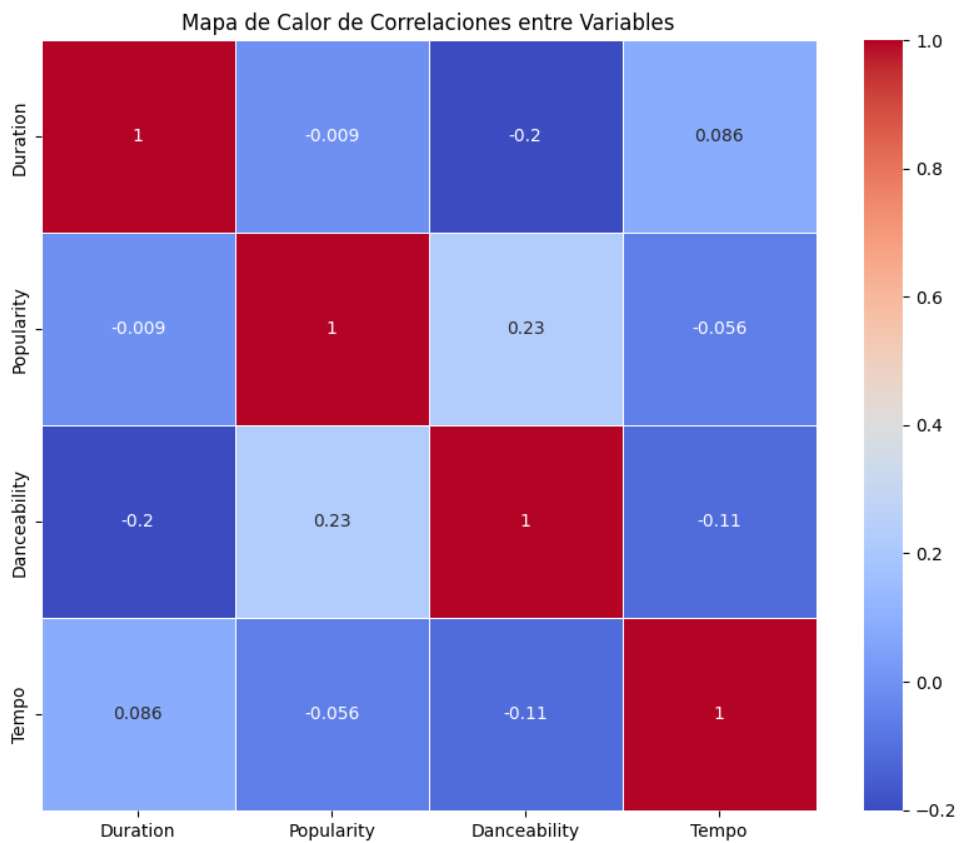


# Exploracion de Datos

Despues de realizar un diagrama de dispersión de datos utilizando la popularidad de las canciones y el tempo de las mismas, se puede notar que el punto con más popularidad tiene un Tempo de poco más de 100. Mientras que en promedio se mantienen entre un 30 y 70 de Popularidad y entre 70 y 200 en Tempo.



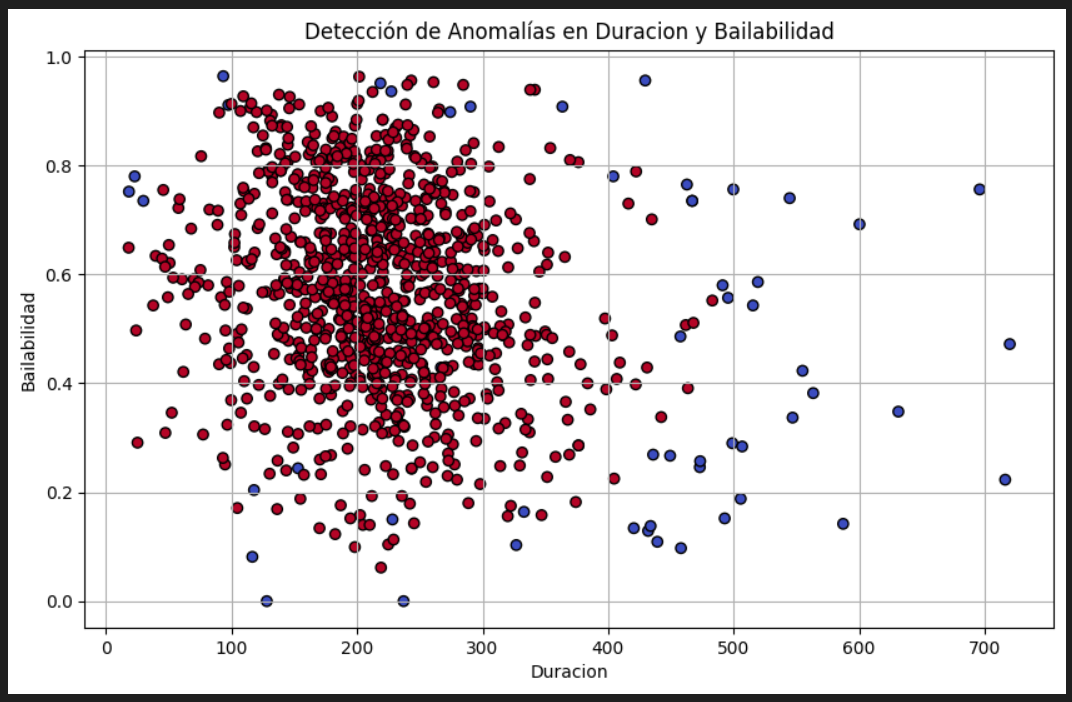
En el mapa de calor siguiente se puede apreciar cómo mientras más se acerca a una correlación de 1, mas rojo se pone el cuadro. Mientras que las correlaciones más débiles están representadas en tonos de azul más oscuros.



Inmediatamente salta a la vista que la correlación más fuerte es de 0.23 entre Bailabilidad y Tempo, lo cual implica que las canciones más populares entre el público, son aquellas cuyos ritmos permiten que sea disfrutable bailarlas según las métricas de Spotify.  
  
Otro dato a notar es que la duración y las Bailabilidad tienen un correlación muy negativa de -0.2. Esto significa que mientras larga sea la canción, menos bailable es de acuerdo al dataset.

# Descubrimiento de Información Relevante

El siguiente gráfico muestra los resultados de la detección de anomalías utilizando el algoritmo Isolation Forest aplicado a la Duracion y la Bailabilidad



Se detectaron 53 anomalías (en azul). Se puede observar como los datos normales están marcados en rojo, y demuestran lo anteriormente mencionado de que canciones muy largas no suelen ser tan bailables. El punto perfecto esta aproximadamente entre los 150 y 260 segundos.

Para concluir, se pueden encontrar muchas relaciones entre diferentes variables encontradas en el dataset. Se encontró que mientras más larga sea una canción, menos popular y bailable es, además de que Tempos muy altos o muy bajos suelen ser muy poco populares. Hay algunas anomalias en las que la duración no afecta tanto la bailabilidad, pero son pocas