Universidad del Valle de Guatemala
Facultad de Ingeniería

PROYECTO No. 2

Spotify Million Playlist Dataset Challenge

Situación Problemática: Describe la situación problemática que da lugar al problema.

Un estudio hecho por Digital Music Alliance, en 2018, concluyó que el 54% de los usuarios de plataformas de streaming de música están reemplazando los álbumes por las playlists en sus hábitos. Pero los usuarios no solo escuchan las playlists, sino que también las crean. Los usuarios de Spotify han creado y compartido más de 4 billones de playlists.

Los usuarios crean una playlist por diversas razones: algunas playlists agrupan la música de manera categórica. Ya sea por género, artista, año o incluso país, por estado de ánimo, tema u ocasión. Incluso algunas playlists están hechas para dar un mensaje específico, de amor, celebración, bienvenida, etc…

Al comprender esta relación tan profunda entre la música y las personas, se puede hacer una relación congruente entre canciones y poder generar las playlist indicadas para los usuarios. Al aprender más sobre la naturaleza de las listas de reproducción, también podemos sugerir otras pistas que un oyente disfrutará en el contexto de una playlist. Esto puede facilitar la creación de playlist y, en última instancia, ayudar a las personas a encontrar más música que aman.

Problema científico: Se enuncia el problema científico que se desprende de la situación planteada. Se comprende bien cuál es el problema.

El problema proviene de la pregunta ¿Cómo podemos facilitar la creación de playlist a los usuario, mientras se mantiene la congruencia entre las canciones que contiene esta misma?
La solución a este problema es generar un modelo el cual pueda entrenarse a base de diversas playlist las cuales tienen una congruencia entre las canciones o algunas que no la tengan. El modelo implementará algoritmos de nearest neighbor y correlación de Pearson. Nearest neighbor determina según los datos proporcionados un patrón de gustos y preferencias y utiliza los datos de un vecino cercano con características similares al inicial y partiendo de estos datos genera las recomendaciones.
Por otra parte, el algoritmo de correlación de Pearson es un algoritmo de similaridad que recolecta los datos de preferencias de los usuarios y determina un peso de similitud para estimar la relación que existe entre dos usuarios y crear recomendaciones de contenido en base a dicha similaridad.

Objetivos: Se plantean los objetivos a cumplir para darle solución al problema planteado. Se enuncia al menos un objetivo general y 2 específicos. Los objetivos deben ser medibles y alcanzables durante la investigación

Objetivo General

Generar un sistema de recomendación de canciones para una playlist en base a las canciones que ya contiene la playlist y así facilitar dicha creación para el usuario.

Objetivos específicos

- Generar un sistema de recomendación utilizando el mejor de los algoritmos nearest neighbor y correlación de pearson.
- Por medio de métricas tales como R-precision y NDCG se podrá determinar la efectividad de cada algoritmo y poder llegar a una conclusión certera.
- Medir la congruencia entre canciones por medio de los atributos que ofrecen en la data (eg. artista, año, género).
- Crear una playlist de manera aleatoria ingresando solo el género que al usuario le gustaría escuchar.

Descripción de los datos: Se describen los datos, tanto las variables y observaciones como las operaciones de limpieza que se le hicieron si fueron necesarias.

Operaciones de limpieza necesarias:

La data entregada viene en formato JSON, este formato no es compatible con las librerías de análisis de datos en Python. Fue necesario realizar un programa que realice la conversión de todos los JSON's a CSV para poder utilizar las herramientas necesarias para analizar los datos.

Variables:

Info:

- slice: el rango de cortes que hay en este archivo en particular, como 0-999
- version: la versión actual del MPD (que debería ser v1)
- description: una descripción del MPD
- license: información de licencia del MPD
- generated_on: fecha de generación del corte.

Playlist:

- pid: playlist id

- name: name of playlist
- description: string opcional, descripcion de playlist.
- modified_at: segundos - marca de tiempo (en segundos desde la época) cuando esta lista de reproducción se actualizó por última vez. Los tiempos se redondean a la medianoche GMT de la fecha en que se actualizó la lista de reproducción por última vez.
- num_artists:el número total de artistas únicos para las pistas en la lista de reproducción.
- num_albums: el número de álbumes únicos para las pistas en la lista de reproducción
- num_tracks: el número de pistas en la lista de reproducción
- num_followers:el número de seguidores que tenía esta lista de reproducción en el momento en que se creó el MPD.
- num_edits:el número de sesiones de edición separadas. Las pistas añadidas en una ventana de dos horas se consideran añadidas en una única sesión de edición.
- duration_ms: duración total de todas las canciones en milisegundos.
- collaborative: verdadero si es playlist colaborativa, falso si no lo es.
- tracks: una matriz de información sobre cada pista en la lista de reproducción. Cada elemento de la matriz es un diccionario con los siguientes campos:
    - track_name: nombre
    - track_uri: URI de spotify
    - album_name: nombre del álbum
    - album_uri: URI de spotify del álbum
    - artist_name: nombre del artista primario
    - artist_uri: URi de spotify del artista primario
    - duration_ms: duración en milisegundos
    - pos: posición de la canción en la playlist.

El conjunto de datos contiene PID de 0 a 999 999 (1 millón de listas de reproducción), con M pistas únicas

Para replicar el conjunto de prueba de la competencia, eliminamos algunas listas de reproducción de las listas de reproducción originales de modo que:

- Todas las pistas en el conjunto de desafíos aparecen en el MPD
- Todas las pistas reservadas aparecen en el MPD

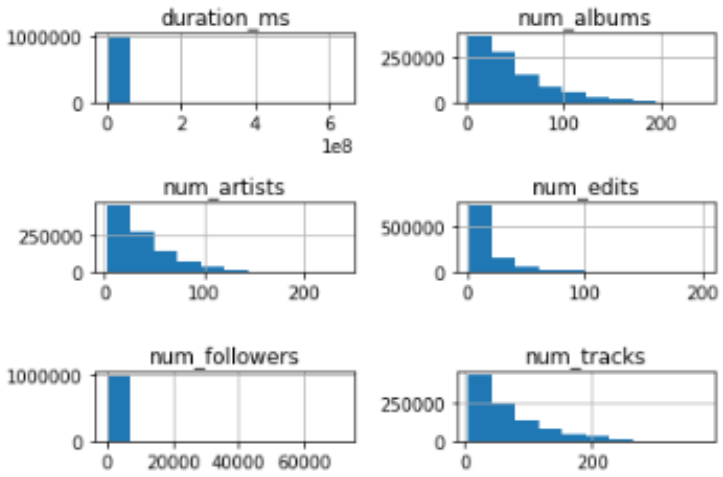
El conjunto de prueba contiene 10 desafíos diferentes, cada desafío contiene 1000 listas de reproducción muestreadas de MPD:

- ❖ Predecir pistas para una lista de reproducción dado su título y las primeras 5 pistas
- ❖ Predecir pistas para una lista de reproducción dado su título y las primeras 10 pistas
- ❖ Predecir pistas para una lista de reproducción dado su título y las primeras 25 pistas
- ❖ Predecir pistas para una lista de reproducción dado su título y 25 pistas aleatorias
- ❖ Predecir pistas para una lista de reproducción dado su título y las primeras 50 pistas
- ❖ Predecir pistas para una lista de reproducción dado su título y 50 pistas aleatorias
- ❖ Predecir pistas para una lista de reproducción dado su título y las primeras 100 pistas
- ❖ Predecir pistas para una lista de reproducción dado su título y 100 pistas aleatorias
- ❖ Predecir pistas para una lista de reproducción dado su título y las primeras 200 pistas
- ❖ Predecir pistas para una lista de reproducción dado su título y 200 pistas aleatorias

Número de:

- ➔ Listas de reproducción: 1000000
- ➔ Pistas: 66346428
- ➔ Pistas únicas: 2262292
- ➔ Álbumes únicos: 734684
- ➔ Títulos únicos: 92944

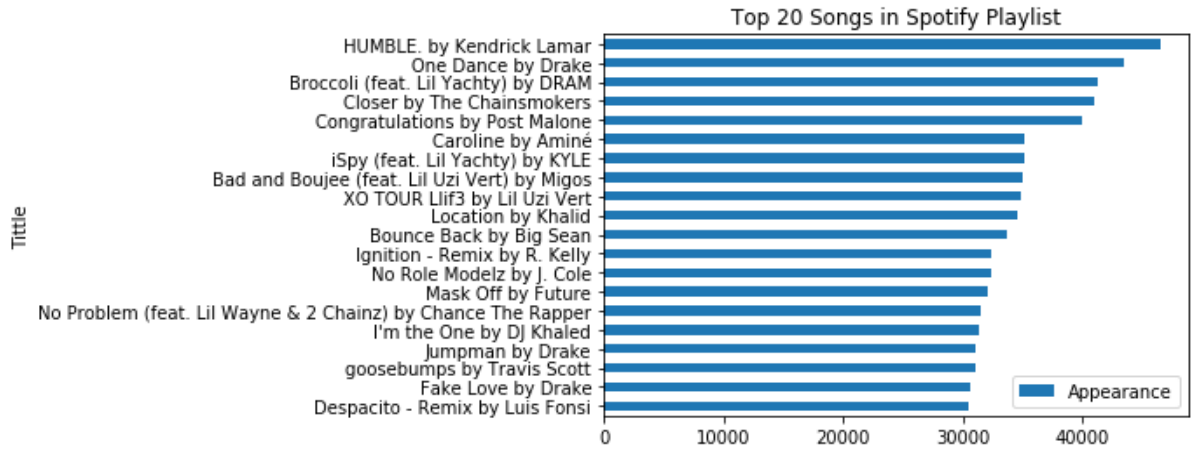
Distribución de: duración de la lista de reproducción, número de álbumes/lista de reproducción, número de artista/lista de reproducción, número de ediciones/lista de reproducción, número de seguidores/lista de reproducción, número de pistas/lista de reproducción.



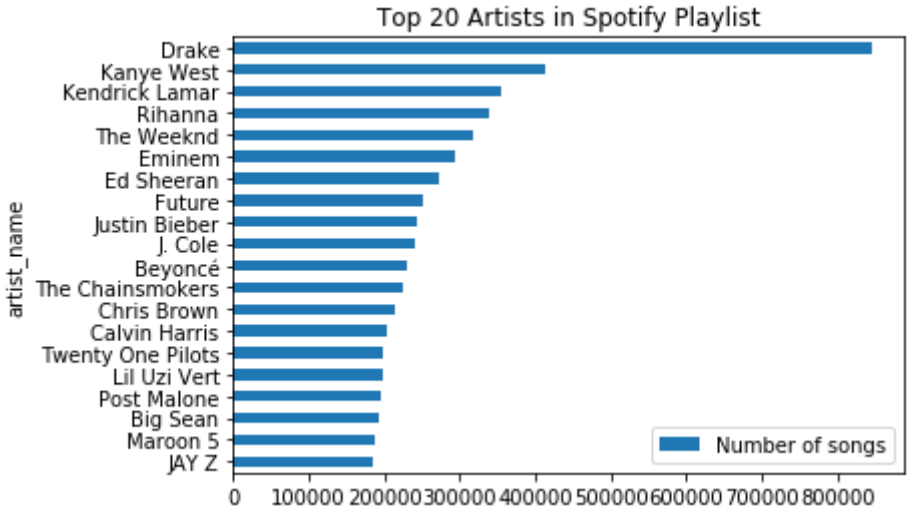
Como podemos ver, todas las distribuciones están sesgadas a la izquierda, lo que significa que si estamos buscando un valor promedio, debemos elegir "Mediana", no "Media".

- ➢ Mediana de duración de la lista de reproducción: 11422438.0
- ➢ Mediana del número de álbumes en cada lista de reproducción: 37.0
- ➢ Mediana del número de artistas en cada lista de reproducción: 29,0
- ➢ Mediana del número de ediciones en cada lista de reproducción: 29,0
- ➢ Mediana del número de seguidores en cada lista de reproducción: 1.0
- ➢ Mediana del número de pistas en cada lista de reproducción: 49,0

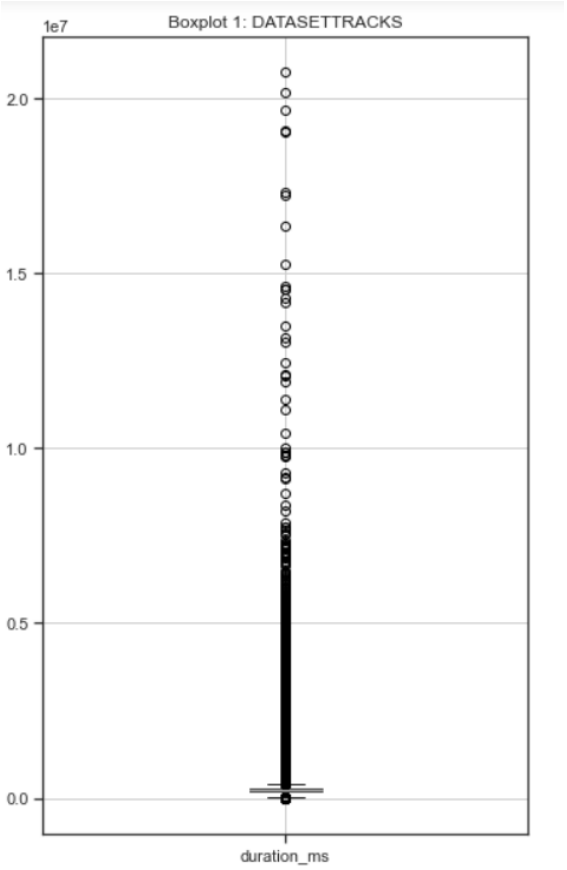
Las 20 mejores canciones en las listas de reproducción de Sporify

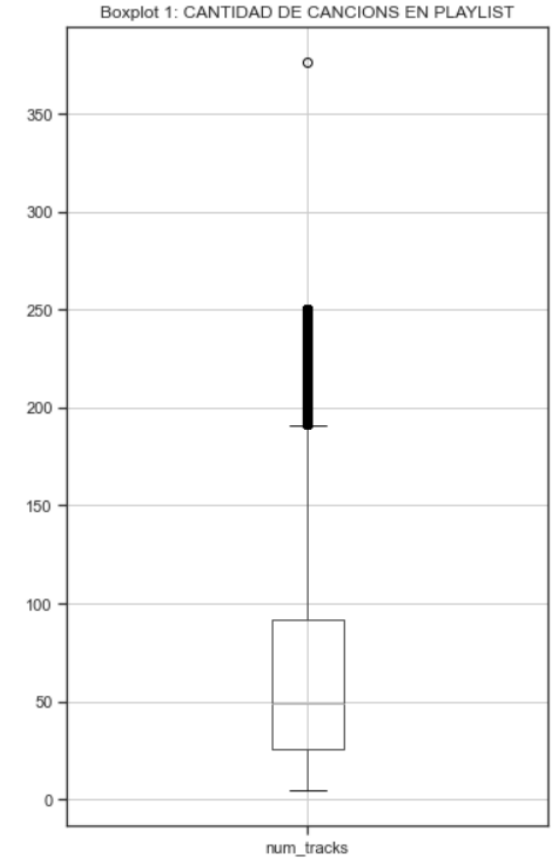


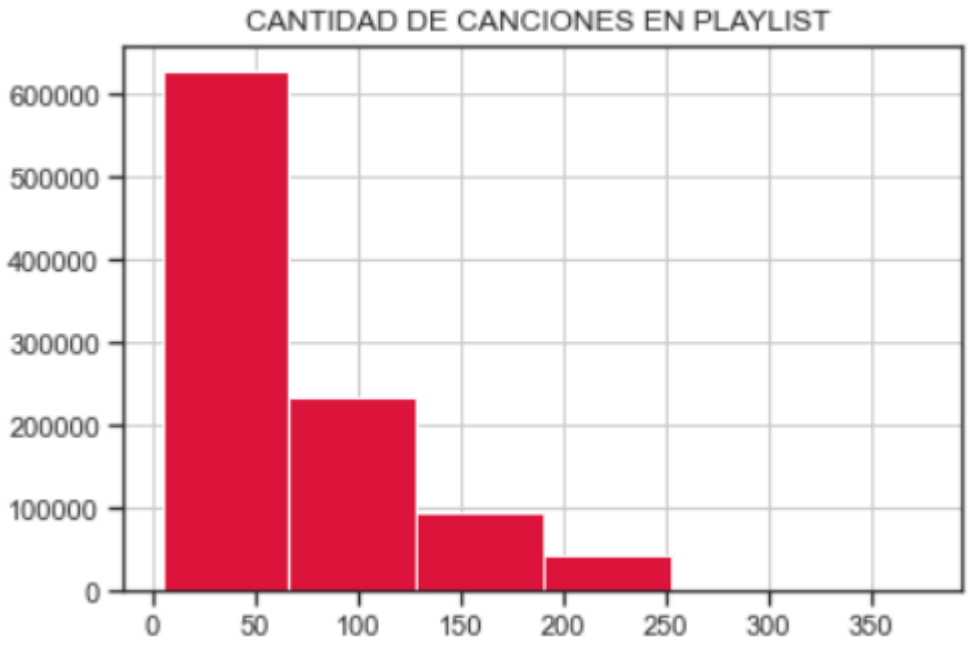
Los 20 mejores artistas en las listas de reproducción de Spotify

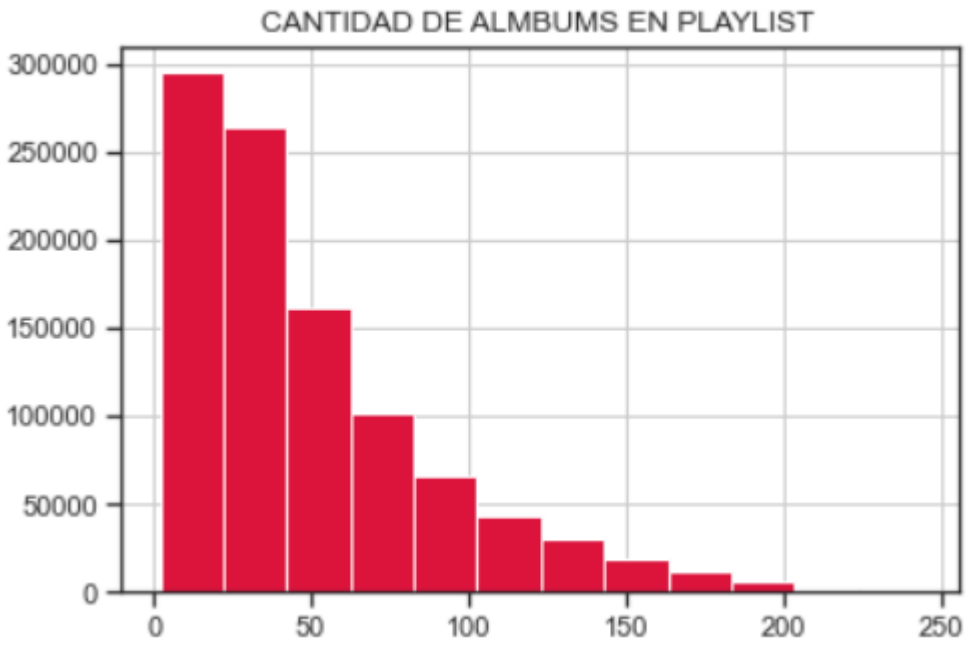


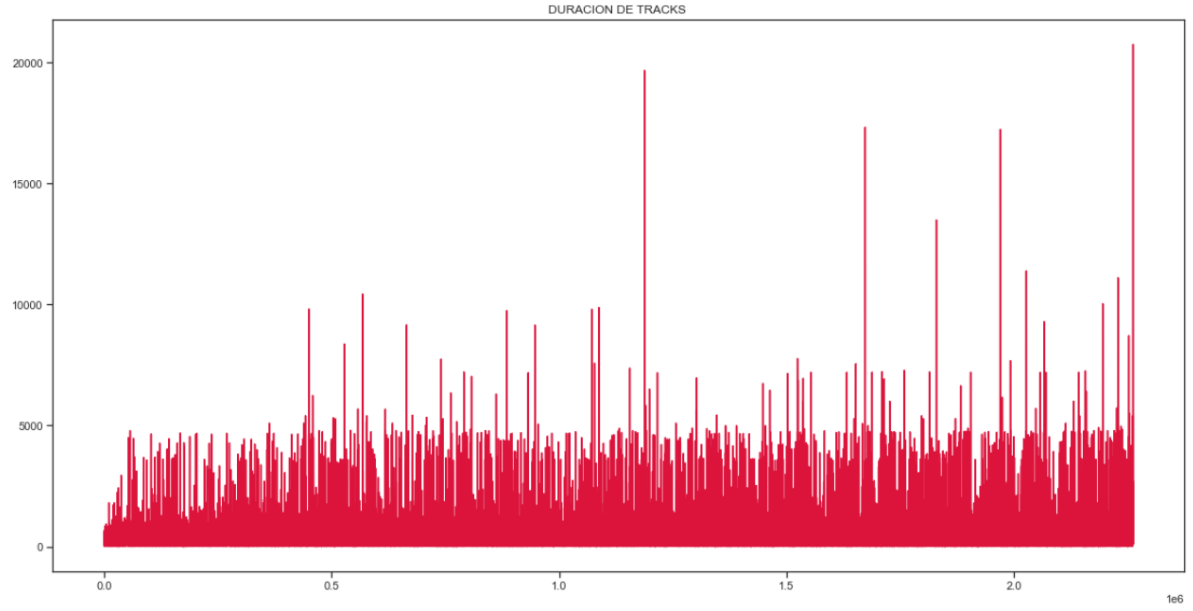
- Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
- Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos


Boxplot 1: DATASETTRACKS
1e7
2.0
1.5
1.0
0.5
0.0
duration_ms


Boxplot 1: CANTIDAD DE CANCIONS EN PLAYLIST
350
300
250
200
150
100
50
0
num_tracks


CANTIDAD DE CANCIONES EN PLAYLIST
600000
500000
400000
300000
200000
100000
0
0
50
100
150
200
250
300
350


CANTIDAD DE ALMBUMS EN PLAYLIST
300000
250000
200000
150000
100000
50000
0
0
50
100
150
200
250



Media duracion de track en segundos

```
statistics.median(DATASETTRACKSDIVS)
```

```
2.2528000000000006
```

Inicialmente se puede observar que la distribución no es simétrica para ninguna de las dos boxplots. Podemos observar que la mayoría de playlist se encuentran en el rango de 0 a 50 canciones, lo mismo se puede decir de la cantidad de álbumes en la playlist. Esto nos puede dar una pista de cómo están distribuidas

las playlists de los usuarios en términos de álbumes y canciones. En el caso de los edits la mayoría se encuentra en el rango aproximado de 0 a 20. Para la duración de las canciones podemos observar que la mayoría duran aproximadamente entre 2.25 minutos y 1.2 minutos, aunque se debe de ajustar la gráfica para mostrar valores mayores a 2.25 minutos. Podemos observar que la media calculada muestra el valor mencionado previamente (2.25 minutos). Lo cual puede ayudar a la hora de la recomendación, tal vez evitar canciones demasiado cortas o largas es una buena idea. Igualmente al observar la cantidad de álbumes por playlist, tal vez se podría evitar recomendar canciones que sean del mismo álbum.