



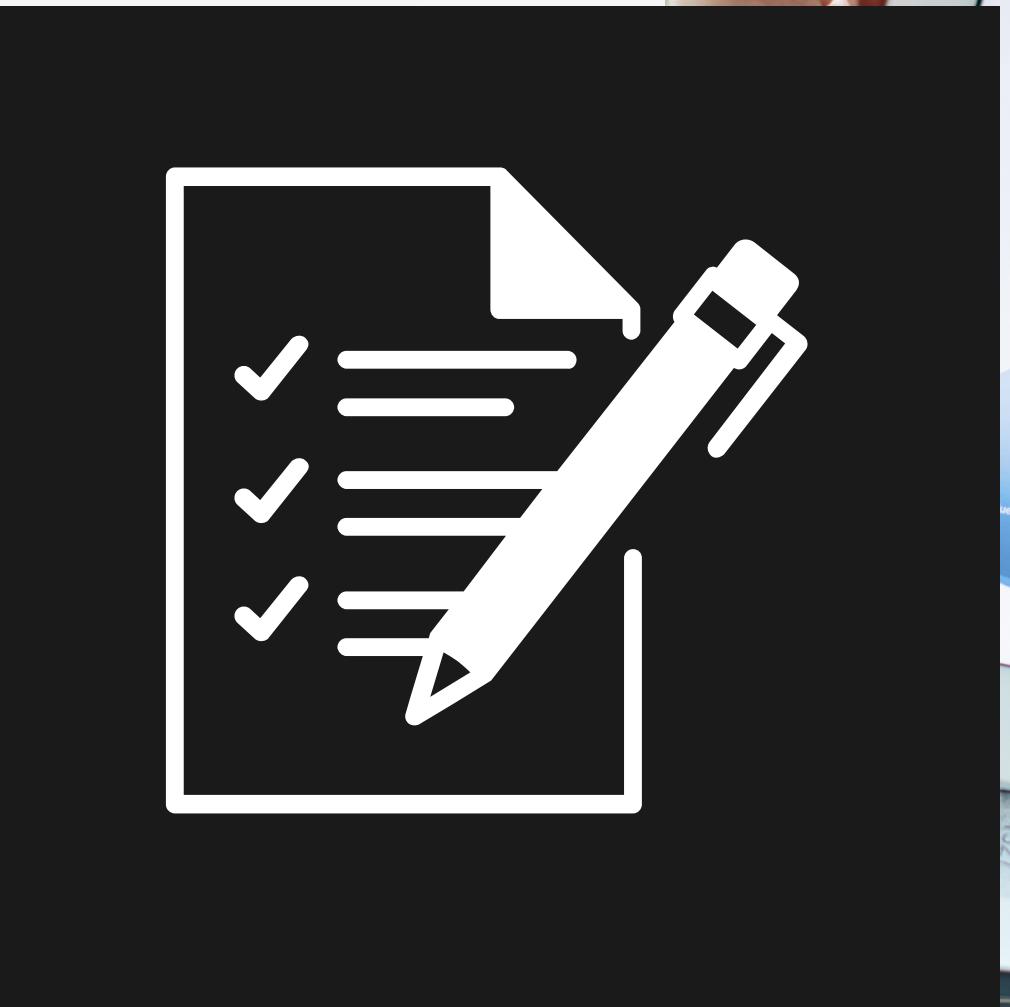
SISTEMA DE RECOMENDACIÓN DE SPOTIFY

CONTENIDO

| | | | |
|------------|-----------------------------|--------------|--------------------------|
| 03 | OBJETIVOS ESPECIFICO | 9-10 | ANÁLISIS DE VARIABLES |
| 04 | OBJETIVOS GENERALES | 11-19 | GRÁFICAS DE LOS DATOS |
| 05 | DESCRIPCIÓN DE LOS DATOS | 20 | HALLAZGOS |
| 6 | TABLA DE VARIABLES | 21 | CONCLUSIONES |
| 7-8 | LIMPIEZA DE DATOS | 22 | DESPEDIDA |

OBJETIVO GENERAL

- Proponer un sistema de recomendación efectivo y preciso que le ofrezca recomendaciones valiosas a los usuarios de Spotify por medio de Machine Learning.



OBJETIVOS

Objetivo n° 1

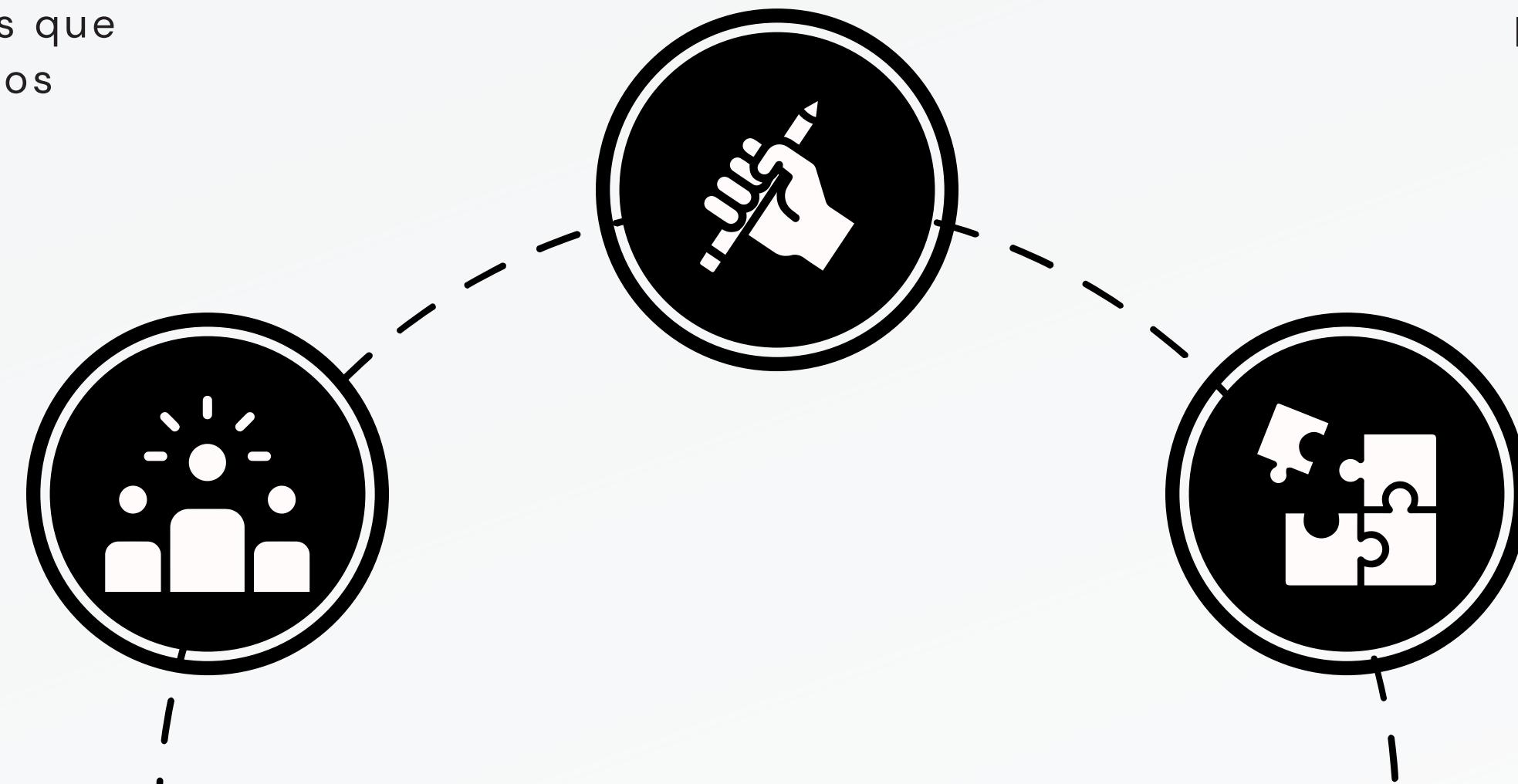
Reconocer patrones en los gustos musicales que ayuden a predecir la relevancia de las recomendaciones que se le hagan a los usuarios

Objetivo n° 2

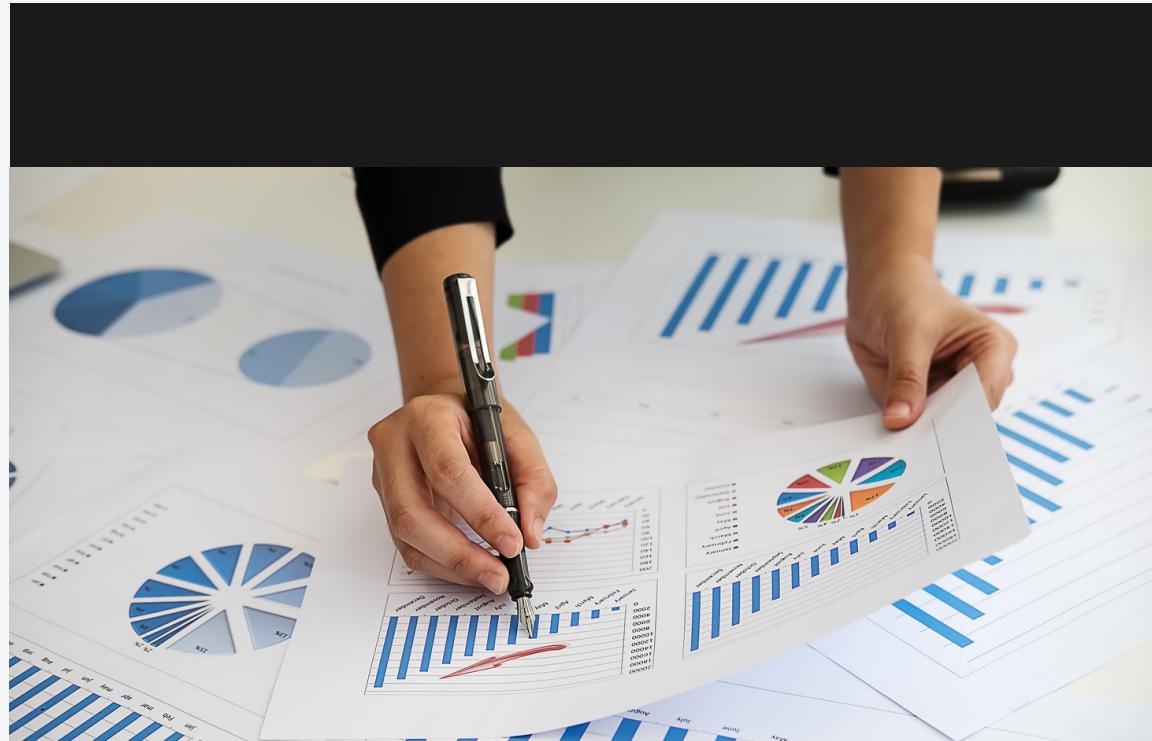
Identificar qué variables representan mejor los gustos musicales de los usuarios.

Objetivo n° 3

Desarrollar un modelo de manera iterativa para encontrar la mejor combinación de parámetros para el sistema de recomendación



DESCRIPCIÓN DE LOS DATOS



- El conjunto de datos consiste en aproximadamente 10,000 playlists públicas de Spotify que han sido creadas entre enero 2010 y noviembre 2017. Cada playlist contiene datos como; el nombre de la playlist y la lista de canciones, la lista de canciones contiene información específica referente a cada una de las canciones, a continuación se profundizará más acerca de qué significa cada uno de los campos del conjunto de datos.

TABLA DE VARIABLES

| Nombre | Descripción | Tipo |
|--------------|---|-----------------------|
| Pid | Identificador de la playlist | Categórica nominal |
| Name | Nombre de la playlist, para algunas playlists este campo es omitido. | Categórica nominal |
| num_holdouts | Número de canciones que han sido omitidas para esta playlist. | Cuantitativa discreta |
| tracks | Una colección que contiene las canciones de cada playlist | Cuantitativa discreta |
| pos | Posición de la canción en la playlist. | Cuantitativa discreta |
| track_name | Nombre de la canción. | Categórica nominal |
| track_uri | El URI de la canción en Spotify. | Categórica nominal |
| artist_name | Nombre del álbum de la canción. | Categórica nominal |
| artist_uri | El URI del artista en Spotify. | Categórica nominal |
| album_name | Nombre del álbum en donde esta la música. | Categórica nominal |
| album_uri | El URI del álbum en Spotify. | Categórica nominal |
| duration_ms | La duración de la canción en milisegundos. | Cuantitativa continua |
| num_samples | Número de canciones en total incluidas en la playlist | Cuantitativa discreta |
| num_tracks | Número de canciones en total de la playlist, incluyendo num_holdouts. | Cuantitativa discreta |



LIMPIEZA DE DATOS

01

CARGA DE DATOS

Para iniciar, primero se carga el archivo JSON utilizando la función 'json.load()', lo que permite que los datos sean accesibles y manejables en el código. A continuación, se realiza una tarea significativa al convertir estos datos en un formato tabular mediante la creación de un DataFrame de Pandas, que se nombra 'df'.

02

DATOS RELEVANTES

Iterando a través de cada fila del DataFrame 'df' utilizando un bucle 'for', se extraen datos esenciales de cada pista musical, como su posición en la lista, el nombre del artista, el URI de la pista, entre otros. Además se identifican las listas vacías almacenando sus índices.

03

ELIMINACIÓN DE FILAS

En esta parte se procede a eliminar de manera estratégica las filas que corresponden a las listas de reproducción que se identificaron previamente como vacías. Al hacerlo, asegura que el DataFrame solo contenga listas de reproducción con contenido válido y significativo. Esta acción es esencial para garantizar que los datos estén en un estado óptimo para análisis posteriores, donde se busca obtener información precisa y valiosa que nos ayude con nuestro modelo.

VARIABLES NUMERICAS

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|----------|--------|----------|----------|-----|--------|--------|--------|---------|
| holdouts | 9000 | 74.51 | 42.09 | 5 | 43 | 67 | 94 | 225 |
| tracks | 9000 | 105.73 | 65.05 | 10 | 51 | 87 | 163 | 250 |
| samples | 9000 | 31.22 | 37.61 | 1 | 5 | 10 | 25 | 100 |
| 100 | 281000 | 23601.69 | 63151.59 | 0 | 200026 | 224574 | 256013 | 9158194 |

VARIABLES CATEGORICAS

| Type | Artist | Track | Album |
|--------------|----------------|---------------------------|---------------------------|
| Country | Drake | Closer | Views |
| Rap | Kanye West | Roses | Coloring Book |
| Oldies | Kendrick Lamar | Ride | Stoney |
| Workout | Rihanna | Broccoli (ft. Lil Yachty) | More Life |
| Rock | The Weeknd | Ignition – remix | The Life Of Pablo |
| Throwback | Eminem | Gold Digger | Beauty Behind The Madness |
| Throwbacks | Luke Bryan | Forever | Greatest Hits |
| Party | J. Cole | No Role Modelz | 2014 Forest Hills Drive |
| Jams | Chris Brown | Home | Original Album Classics |
| Classic Rock | Future | Let Me Love You | Good Kid, M.A.A.D City |
| 148 | 4877 | 335 | 1126 |
| 97 | 2592 | 241 | 806 |
| 75 | 1902 | 229 | 762 |
| 74 | 1734 | 226 | 743 |
| 72 | 1644 | 222 | 713 |
| 59 | 1494 | 215 | 701 |
| 54 | 1486 | 213 | 660 |
| 54 | 1472 | 211 | 657 |
| 43 | 1418 | 203 | 656 |
| 40 | 1308 | 203 | 632 |

GRÁFICO DE DISPERSIÓN DE DURACIÓN DE CANCIONES

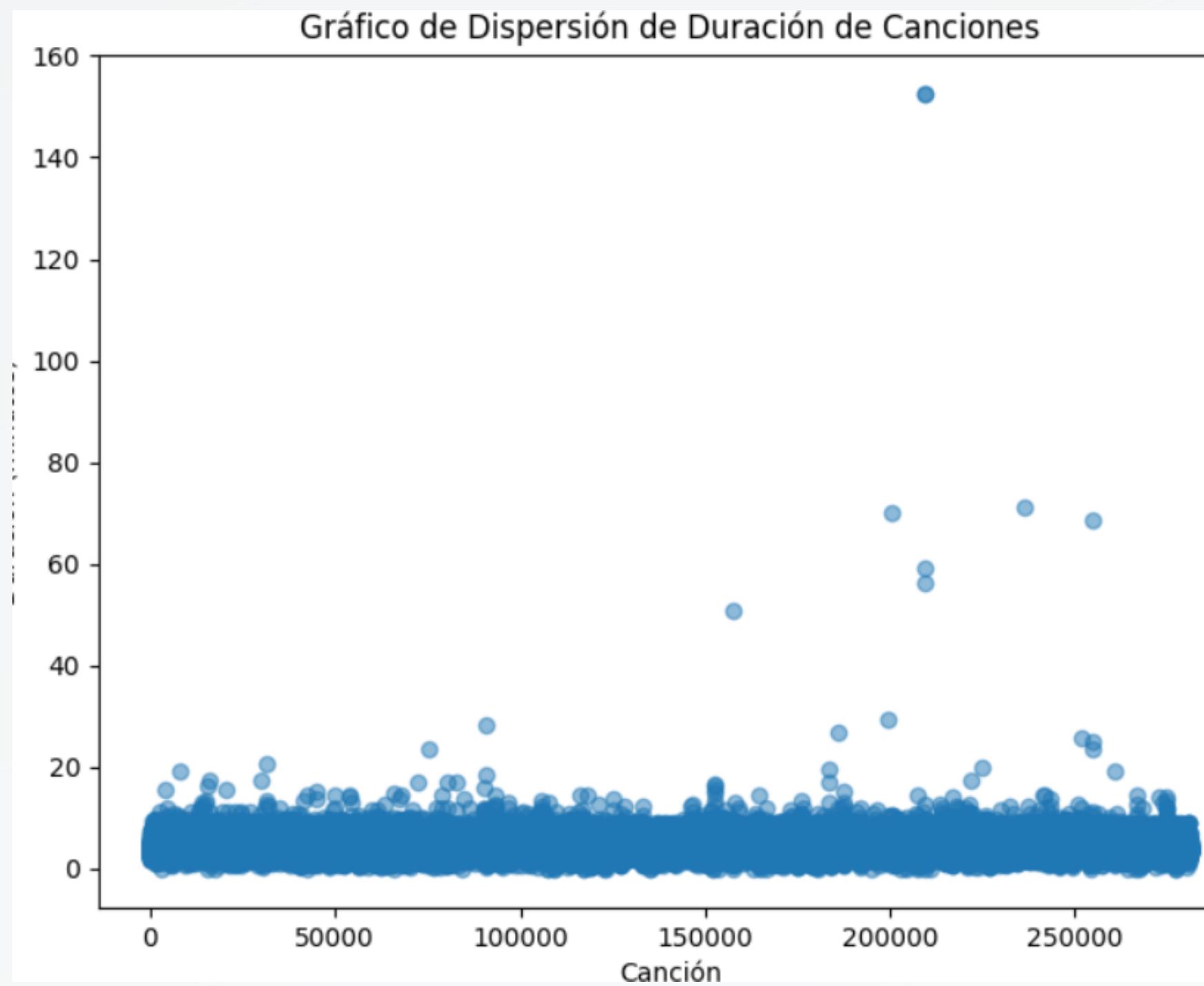


GRÁFICO DE DISPERSIÓN DE NÚMERO DE MUESTRAS

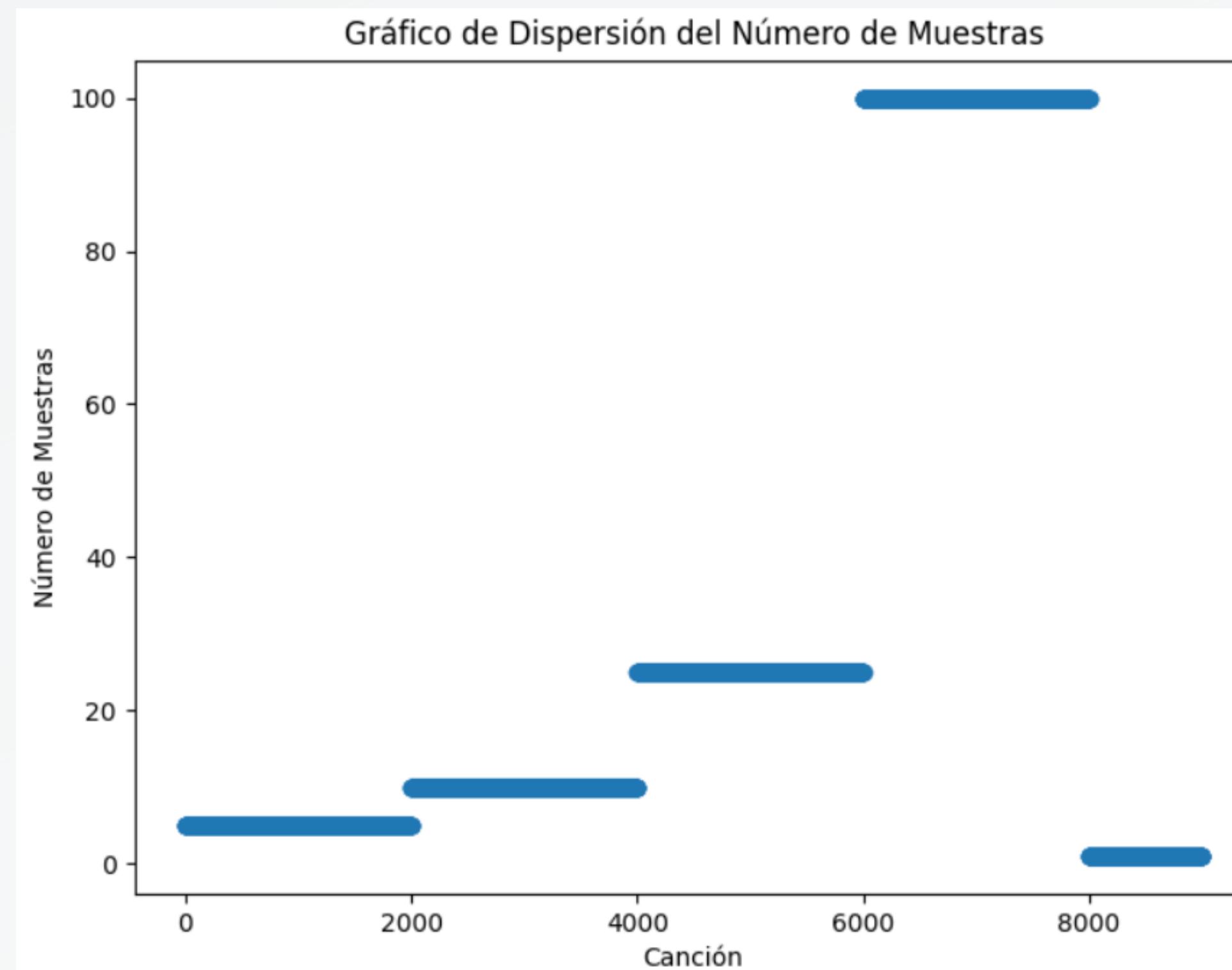


GRÁFICO DE DISPERSIÓN DE NÚMERO DE CANCIONES

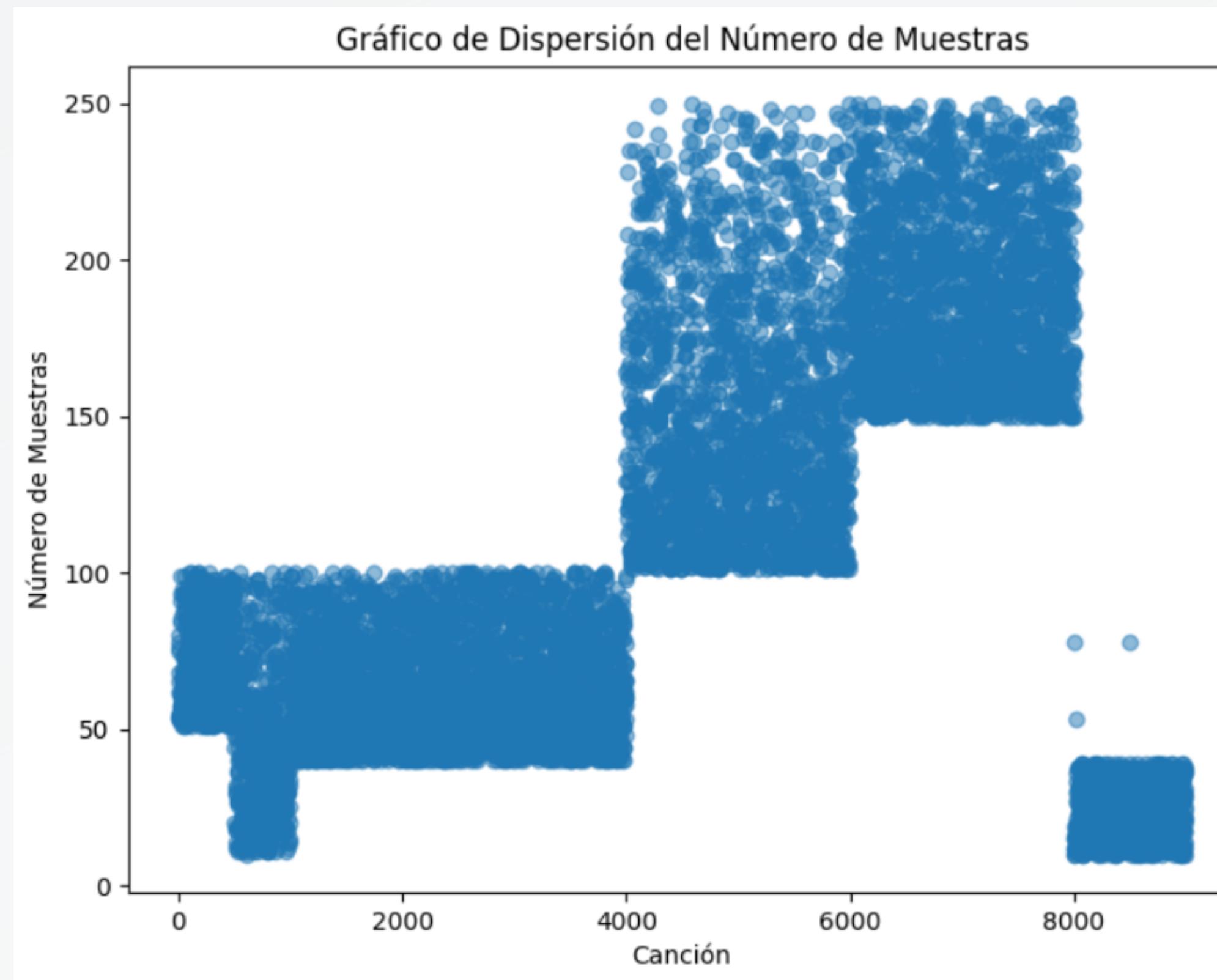
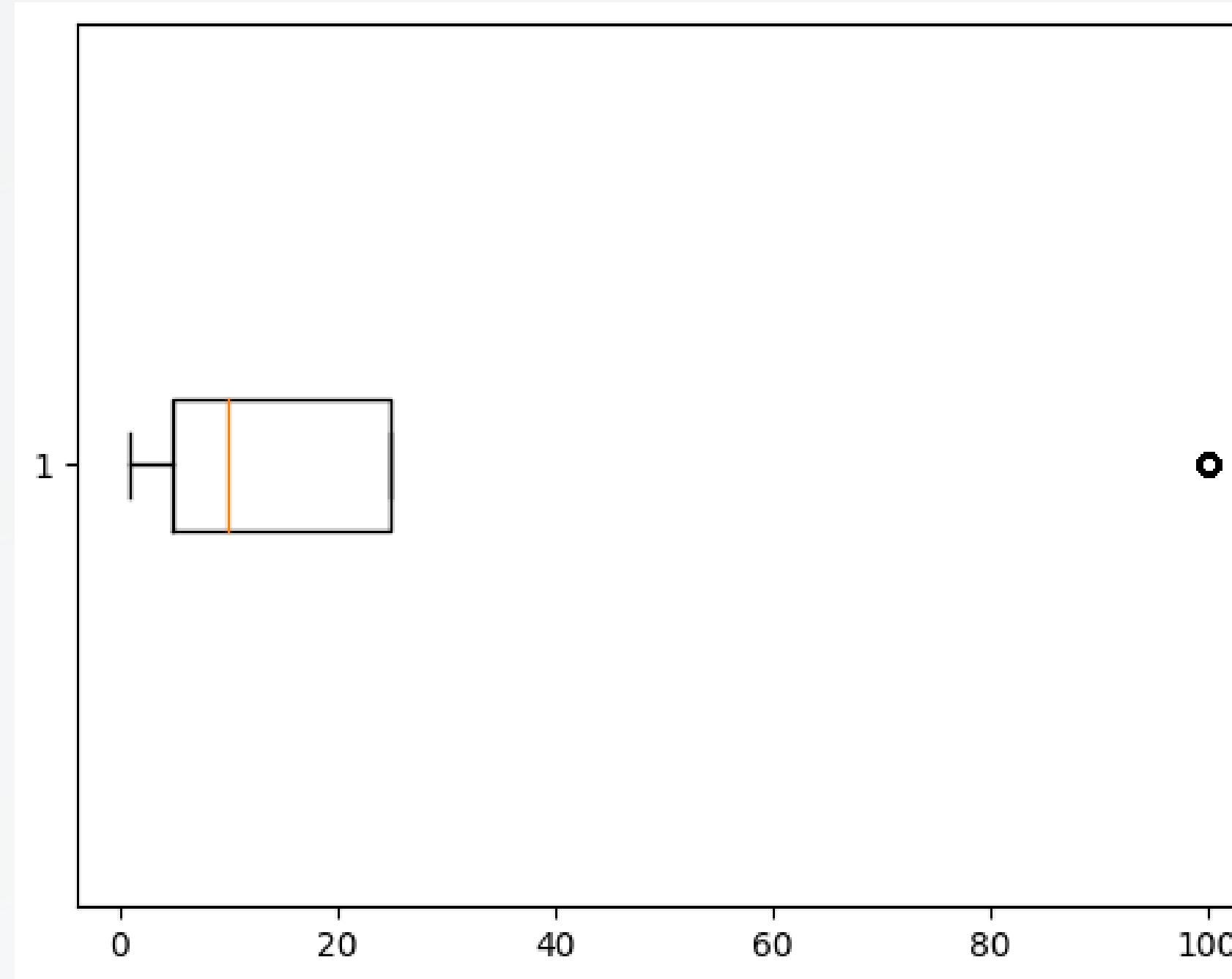


DIAGRAMA DE CAJA Y BIGOTES DE MUESTRAS



Cuartil 1 (Q1)

- 0.925

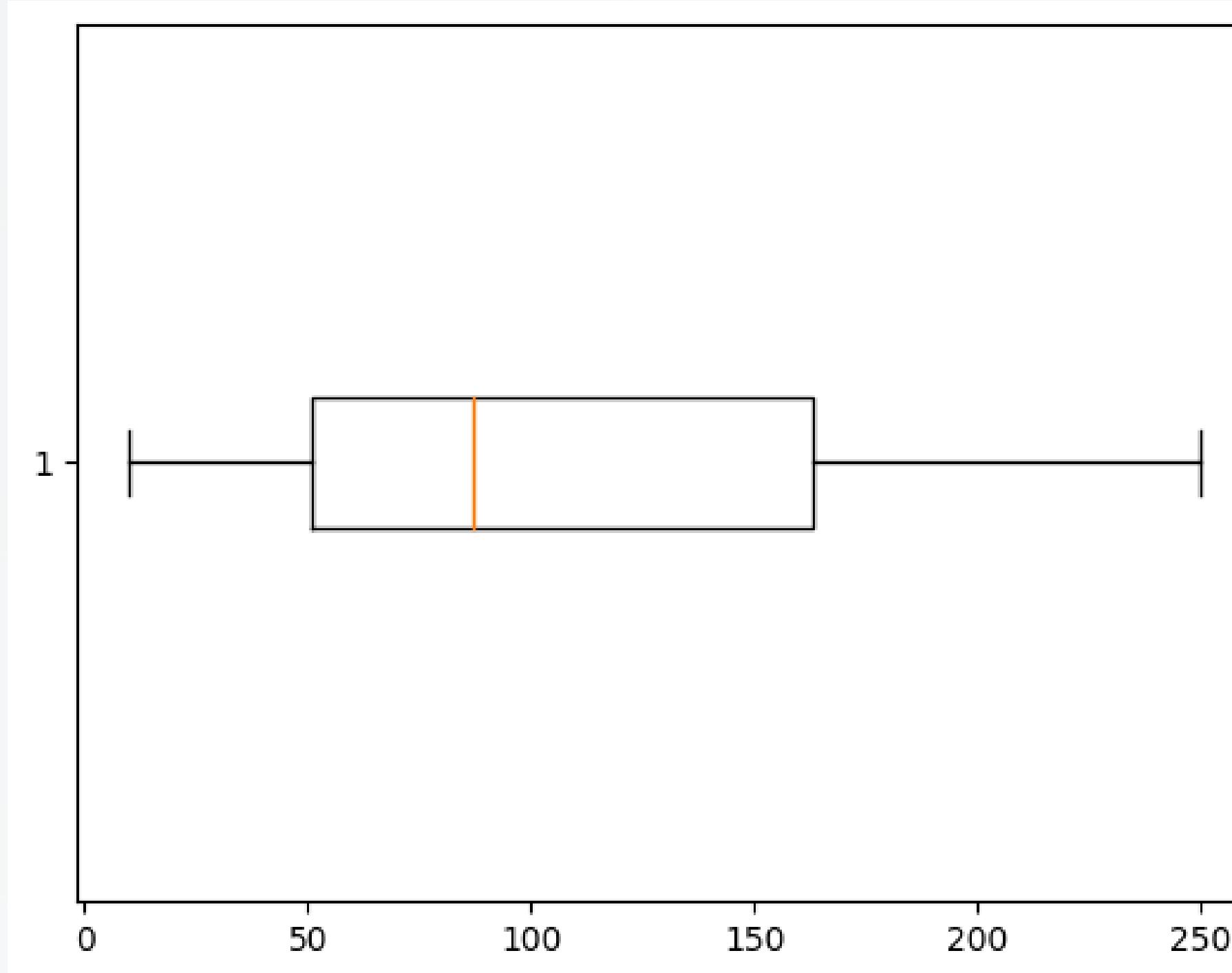
Mediana (Q2)

- 0.925

Cuartil 3 (Q3)

- 1.075

DIAGRAMA DE CAJA Y BIGOTES DE TRACKS



Cuartil 1 (Q1)

- 0.925

Mediana (Q2)

- 0.925

Cuartil 3 (Q3)

- 1.075

GRÁFICO DE BARRAS TOP 10 ARTISTAS CON MÁS CANCIONES EN EL DATASET

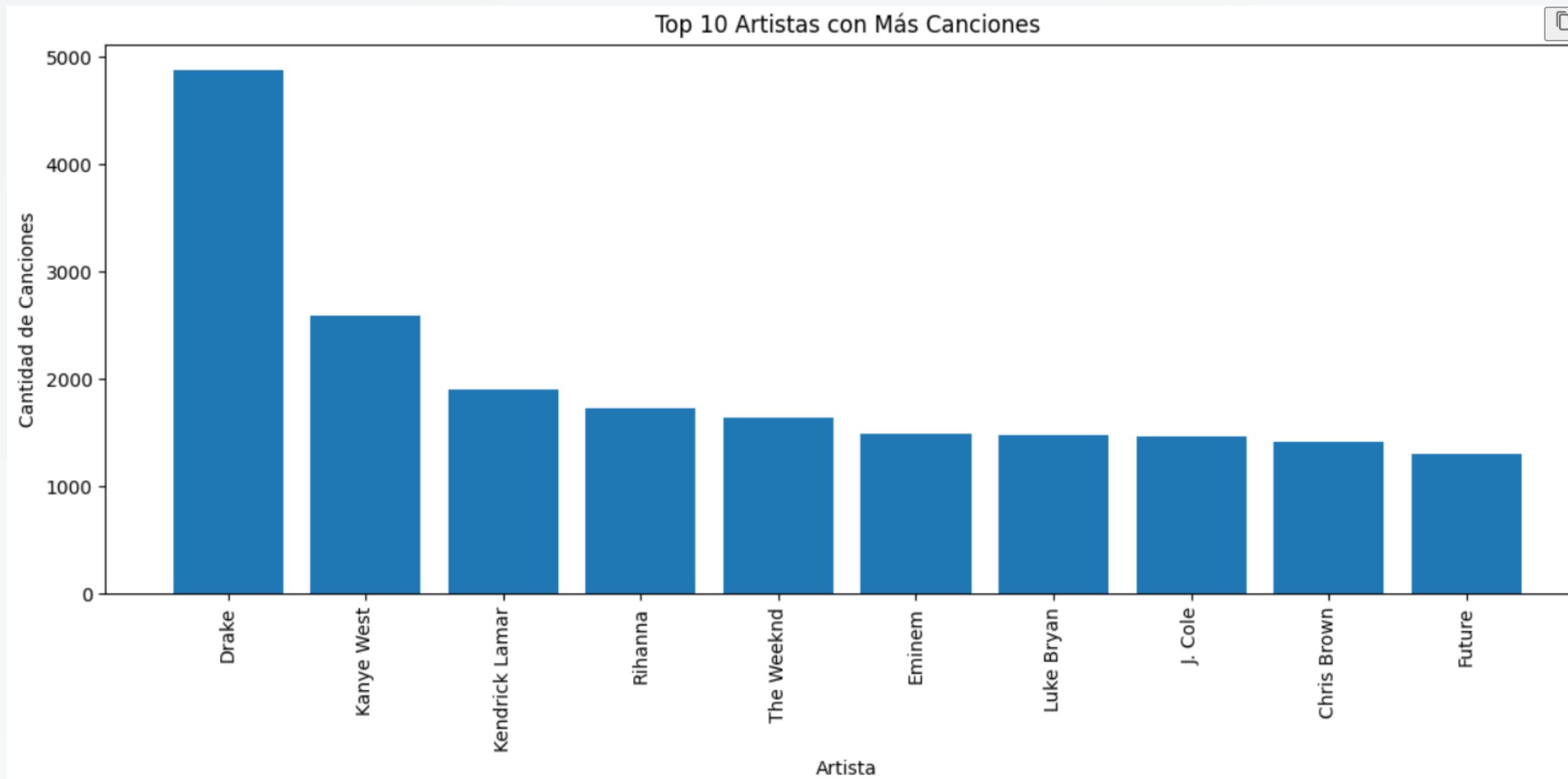


GRÁFICO DE BARRAS TOP 10 ARTISTAS CON MÁS ÁLBUMES

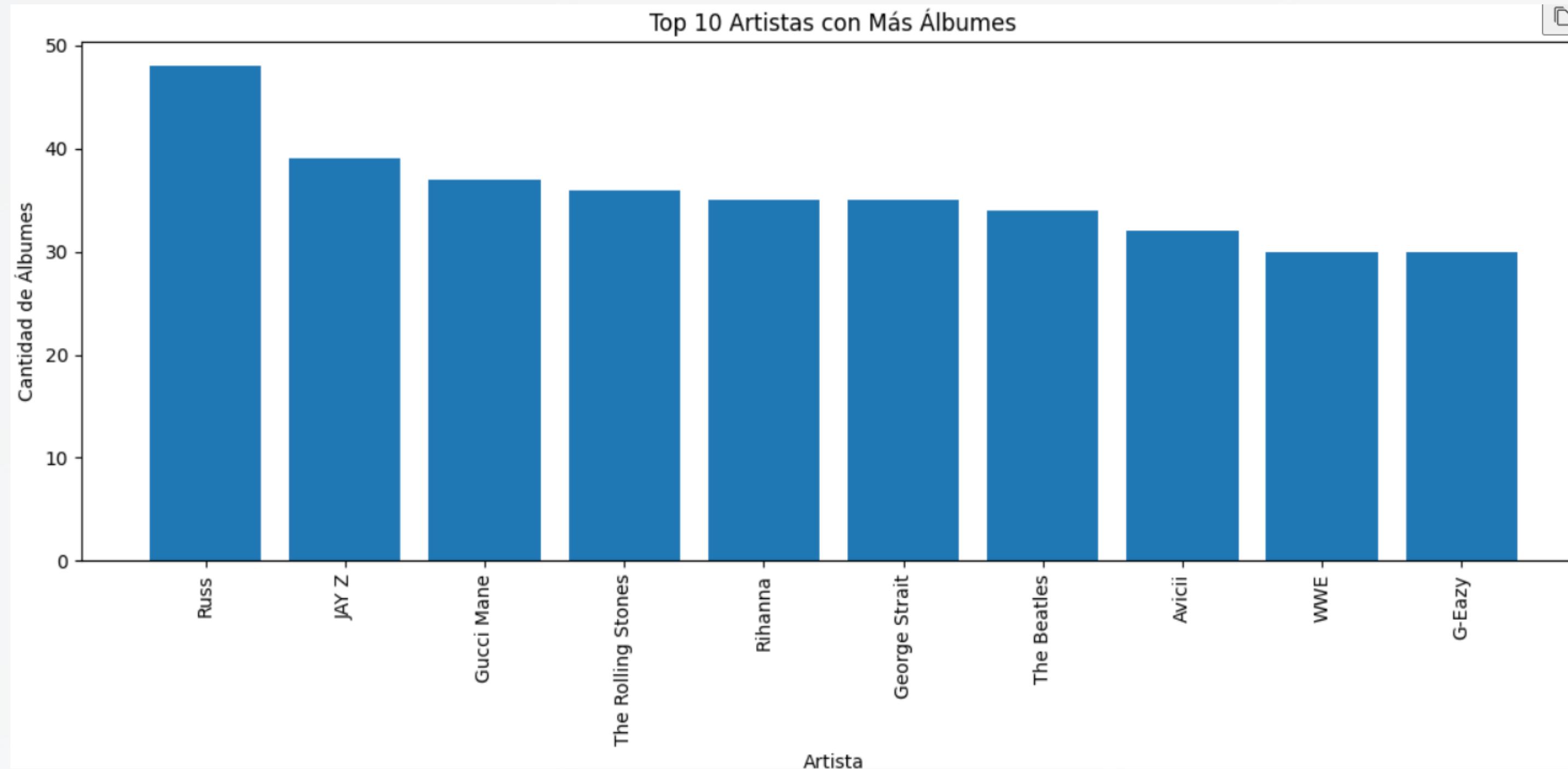
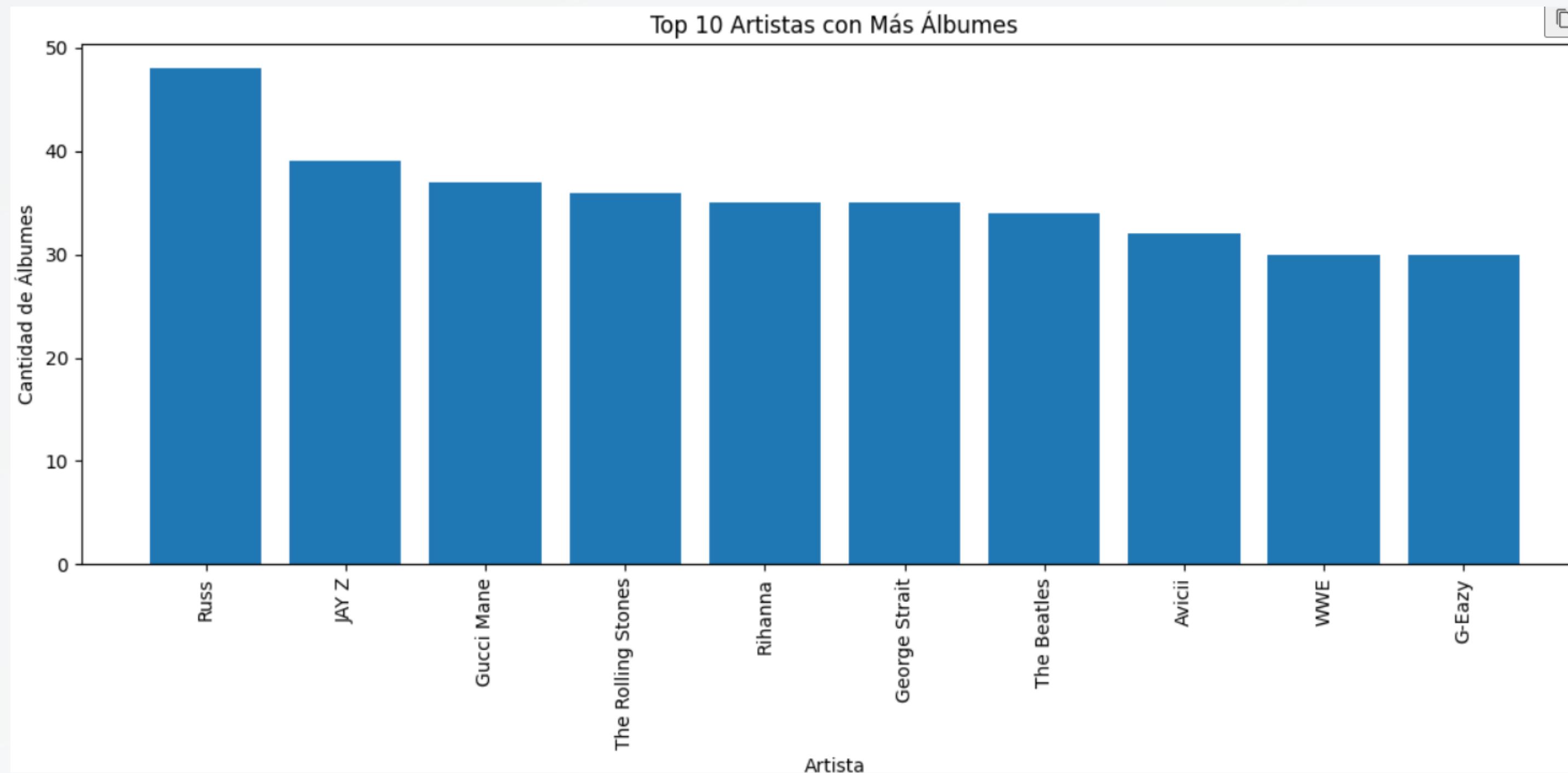
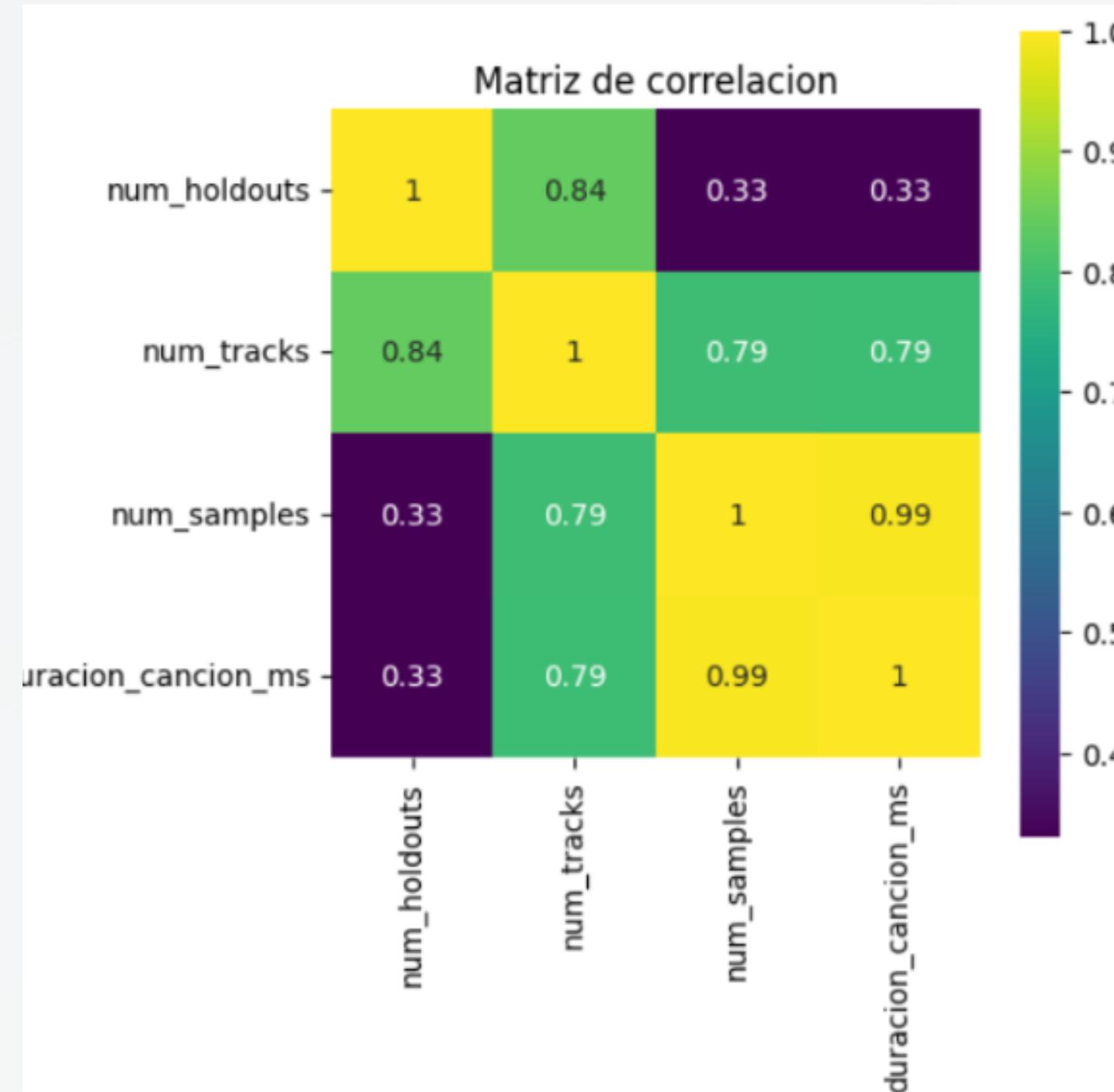


GRÁFICO DE BARRAS TOP 10 CANCIONES QUE MÁS APARECEN EN LAS PLAYLIST



CRUCE DE VARIABLES



| Variable numérica: num_holdouts | | | | | |
|--|--------------|--------|----------|--------------|--|
| | sum_sq | df | F | PR(>F) | |
| name | 9.233319e+06 | 4225.0 | 1.155197 | 0.000017 | |
| Residual | 5.247850e+06 | 2774.0 | NaN | NaN | |
| Variable numérica: num_tracks | | | | | |
| | sum_sq | df | F | PR(>F) | |
| name | 2.208698e+07 | 4225.0 | 1.340201 | 2.851953e-17 | |
| Residual | 1.082047e+07 | 2774.0 | NaN | NaN | |
| Variable numérica: num_samples | | | | | |
| | sum_sq | df | F | PR(>F) | |
| name | 7.340573e+06 | 4225.0 | 1.227161 | 2.126903e-09 | |
| Residual | 3.927427e+06 | 2774.0 | NaN | NaN | |
| Variable numérica: duracion_cancion_ms | | | | | |
| | sum_sq | df | F | PR(>F) | |
| name | 3.981124e+17 | 4225.0 | 1.197451 | 1.135024e-07 | |
| Residual | 2.182870e+17 | 2774.0 | NaN | NaN | |

HALLAZGOS

Hallazgo n° 1

El análisis reveló relaciones entre variables numéricas, la influencia de la variable categórica "name" en algunas de estas variables, y estadísticas descriptivas importantes para el desarrollo del modelo y la toma de decisiones de arquitectura.

Hallazgo n° 2

Se encontró una correlación positiva entre 'num_holdouts' y 'num_tracks', indicando que las listas de reproducción con más 'holdouts' tienden a tener más canciones, lo que sugiere una relación entre diversidad y longitud de las listas.

Hallazgo n° 3

El test ANOVA demostró que la variable 'name' tenía un efecto significativo en las variables numéricas relacionadas con muestras y duración de canciones, lo que la hace importante para futuras predicciones con el modelo. Además, se identificaron los artistas y canciones más influyentes en el conjunto de datos.

CONCLUSIONES



La variable 'name' influye de manera significativamente en todas las variables numéricas evaluadas en el test de ANOVA.



La popularidad de los artistas puede llegar a tener una correlación importante al momento de realizar predicciones de acuerdo a los gustos de cada persona dados ciertos parámetros.



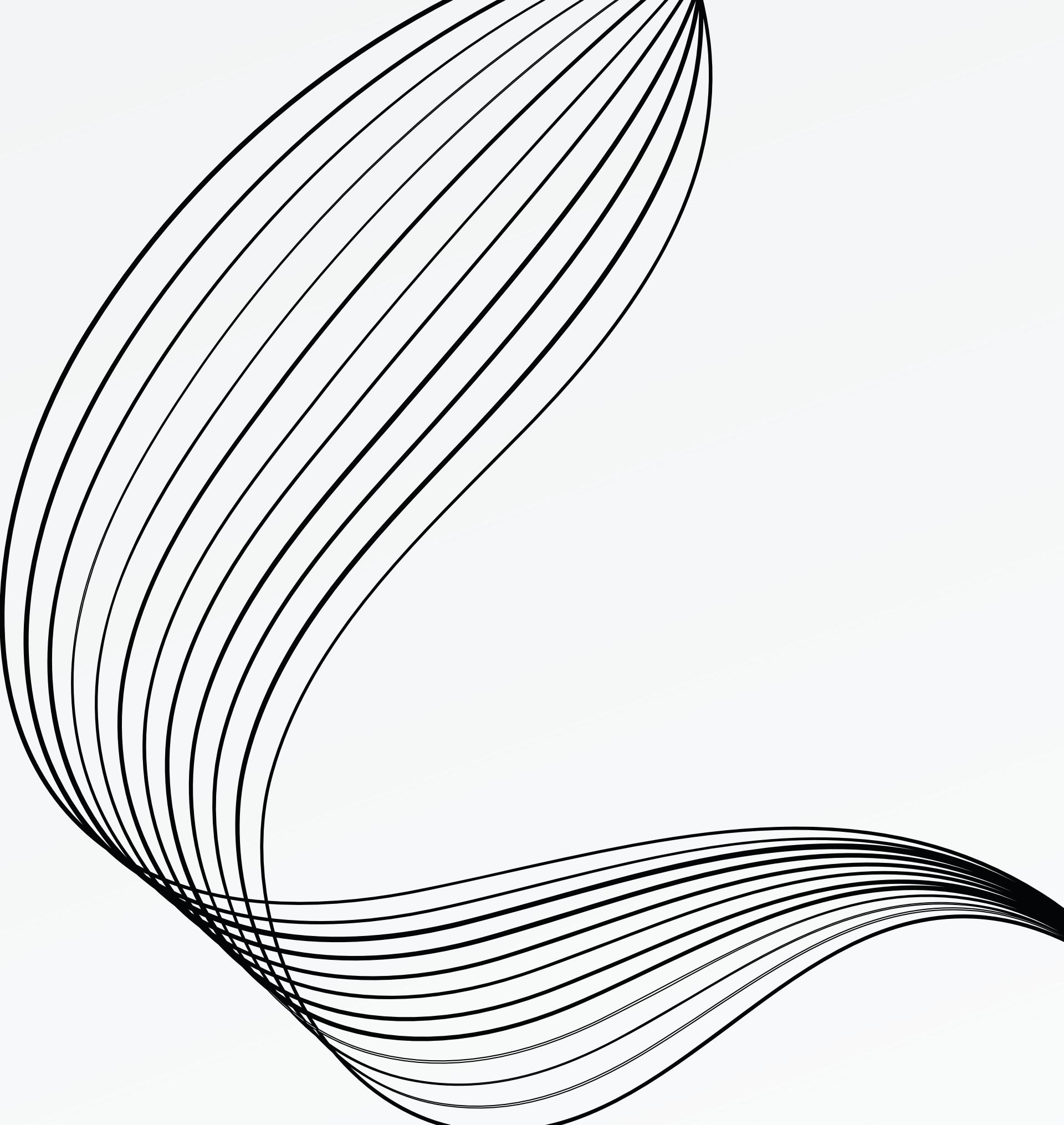
Las relaciones encontradas para la cantidad de canciones y duración de las canciones pueden llegar tener una correlación con el género o gusto del usuario.

REFERENCIAS

Spotify, AIcrowd (2020) Spotify
Million Playlist Dataset Challenge.
Extraído de AIcrowd.

C.W, Chen, P. Lamere, M. Schedl, H.
Zamani. (2018) Recsys Challenge
2018: Automatic Music Playlist
Continuation, in Proceedings of the
12th ACM Conference on
Recommender Systems.

C.W, Chen, P. Lamere, M. Schedl, H.
Zamani. (2018) An Analysis of
Approaches Taken in the ACM
RecSys Challenge 2018 for Automatic
Music Playlist Continuation.



MUCHAS
GRACIAS

