Selección automática de la siguiente canción, basada en las listas de reproducción del servicio de música en línea: Spotify.

Javier Eduardo Jaimes Velasquez[[1]](#footnote-20)

Paula Nathalia Pineda Ortiz[[2]](#footnote-22)

Brandon Valencia Murillo[[3]](#footnote-24)

Johan Alejandro Cifuentes Gonzalez[[4]](#footnote-26)

Melvin Damar Pineda Cañon[[5]](#footnote-28)

2024-06-07

Abstract

Las plataformas de música en línea deben emplear una forma eficiente y eficaz que resuelve uno de los problemas más críticos, recomendación de música automática. En este documento se explora uno de tantos métodos expuestos en la actualidad para resolver la tarea. Se espera poder evaluar su funcionamiento respecto a otros y también su efectividad a la hora de resolver el problema.

Table of Contents

# 1 Introducción

Las canciones se volvieron parte de cotidianidad, y por ellos los servicios de música en línea juegan un papel importante en la vida de cada ciudadano. Los consumidores de música ya no necesitan un dispositivo de propósito dedicado, como el radio. Ahora la música está disponible en una computadora, o en las manos a través de un smartphone.

Las plataformas de música en línea abundan, de acuerdo con *“Wikipedia List of Online Music Databases” (n.d.)*, la variedad es extensa. Por ende, cada proveedor debe esforzarse en crear una experiencia única en un mercado abundado de opciones.

Dichas plataformas están compuestas de muchas o tal vez cientos de funcionalidades, pero una en ella es clave: la recomendación automática de canciones. De ella depende la permanencia de muchos de sus clientes. La capacidad intrínseca del sistema de proveer de forma automática la siguiente canción es clave a la hora de hacer clientes felices. Como se menciona en *Pichl, Zangerle, and Specht (2016)*: *“Un entendimiento profundo de los atributos de las listas de reproducción y como los usuarios crean y mantienen dichas listas puede contribuir de manera natural a mejores recomendaciones y personalizaciones”*.

En este documento se espera poder responder la pregunta: ¿cuál es el método de selección de música automática que mejor contribuye a una mejor experiencia de los usuarios de la plataforma de música en la línea con foco en Spotify?

En el artículo *Lu et al. (2016)*, se menciona diferentes enfoques que resuelven nuestra pregunta. Para el desarrollo de la pregunta en este documento se decidido elegir el enfoque basado en contenido. Si bien este método no es el más usado en sistemas de recomendación se espera que al menos pueda plantear una forma eficaz de resolver la pregunta.

# 2 Analisis Exploratorio de Datos

## 2.1 Conjunto de Datos

Se obtuvo acceso al conjunto de datos [Million Playlist Dataset](https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge) publicado por Spotify en la plataforma de [AI Crowd](https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge). A partir de la fuente original tomamos como referencia una muestra de 1000, en las que aleatoriamente se escogieron 100.

Para una idea incial del conjunto de base se encuentra: datasets/spotify\_millions\_playlist/mpd.slice.0-999.json.

library(jsonlite)  
library(dplyr)  
library(magrittr)  
library(jsonlite)  
library(tidyr)  
library(spotifyr)  
  
Sys.setenv(SPOTIFY\_CLIENT\_ID = "cf91febd27b84f08baa8c5053f9d8f2e")  
Sys.setenv(SPOTIFY\_CLIENT\_SECRET = "f16c1e86e1d8462384a5def0f0fddab4")  
  
# Archivo original de spotify  
t <- fromJSON("datasets/spotify\_millions\_playlist/mpd.slice.0-999.json")  
  
# list of playlist  
d <- t$playlists %>% select(c("name", "collaborative", "pid", "modified\_at", "num\_tracks", "num\_albums", "num\_followers", "tracks"))  
  
# Se toma como muestra 100 lista de reproducciones.  
set.seed(100)  
srow <- sample(1:nrow(d), 100)  
p <- d[srow, ]  
tr <- do.call("rbind", p$tracks) %>% select(-c("pos")) %>% distinct()  
  
  
# Obtención de datos a través del API de Spotify  
access\_token <- get\_spotify\_access\_token()  
  
spotify\_track\_uris <- as.list(tr$track\_uri)   
spotify\_tracks\_ids <- lapply(spotify\_track\_uris, function(x) strsplit(x, ":")[[1]][3])  
spotify\_tracks\_id\_by\_100 <- split(spotify\_tracks\_ids, ceiling(seq\_along(spotify\_tracks\_ids) / 100))  
  
audio\_features\_by\_ids <- get\_track\_audio\_features(gsub(" ", "", toString(spotify\_tracks\_id\_by\_100[1][[1]])))  
  
for (i in 2:length(spotify\_tracks\_id\_by\_100)) {  
 audio\_features\_by\_ids <- rbind(audio\_features\_by\_ids, get\_track\_audio\_features(gsub(" ", "", toString(spotify\_tracks\_id\_by\_100[i][[1]]))))  
 Sys.sleep(1)  
}  
  
trf <- left\_join(x = tr, y = audio\_features\_by\_ids, by = c("track\_uri" = "uri")) %>% na.omit()  
  
pj <- toJSON(p, auto\_unbox = TRUE)  
write\_json(pj, path = "data/playlist.json")  
  
pf <- p %>% select(-c("tracks"))  
  
write.csv(pf, "data/playlists.csv", row.names = FALSE, quote = FALSE)  
write.csv(trf, "data/tracks.csv", row.names = FALSE, quote = FALSE)

A partir de la muestra se construyen los siguientes conjuntos de datos, con information obtenida a traves del [API de Spotify](https://developers.spotify.com).

### 2.1.1 Lista de Reproducciones

La lista de reproduccion tiene la siguiente estructura:

| Nombre del campo | Tipo del Campo | Descripción |
| --- | --- | --- |
| Name | Caracteres | Nombre de la lista de reproduccion |
| Collaborative | Logico | Describe si la lista de reproduccion es publica o privada. |
| pid | Entero | Identificador unico de la lista de reproduccion |
| modified\_at | Entero | Tiempo |
| num\_tracks | Entero | Numero de canciones en la lista de reproduccion. |
| num\_albums | Entero | Numero de albumes en la lista de reproduccion |
| num\_followers | Entero | Numero de seguidores de la lista de reproduccion. |

### 2.1.2 Canciones

Cada lista de reproduccion contiene un numero de canciones con la siguiente estructura.

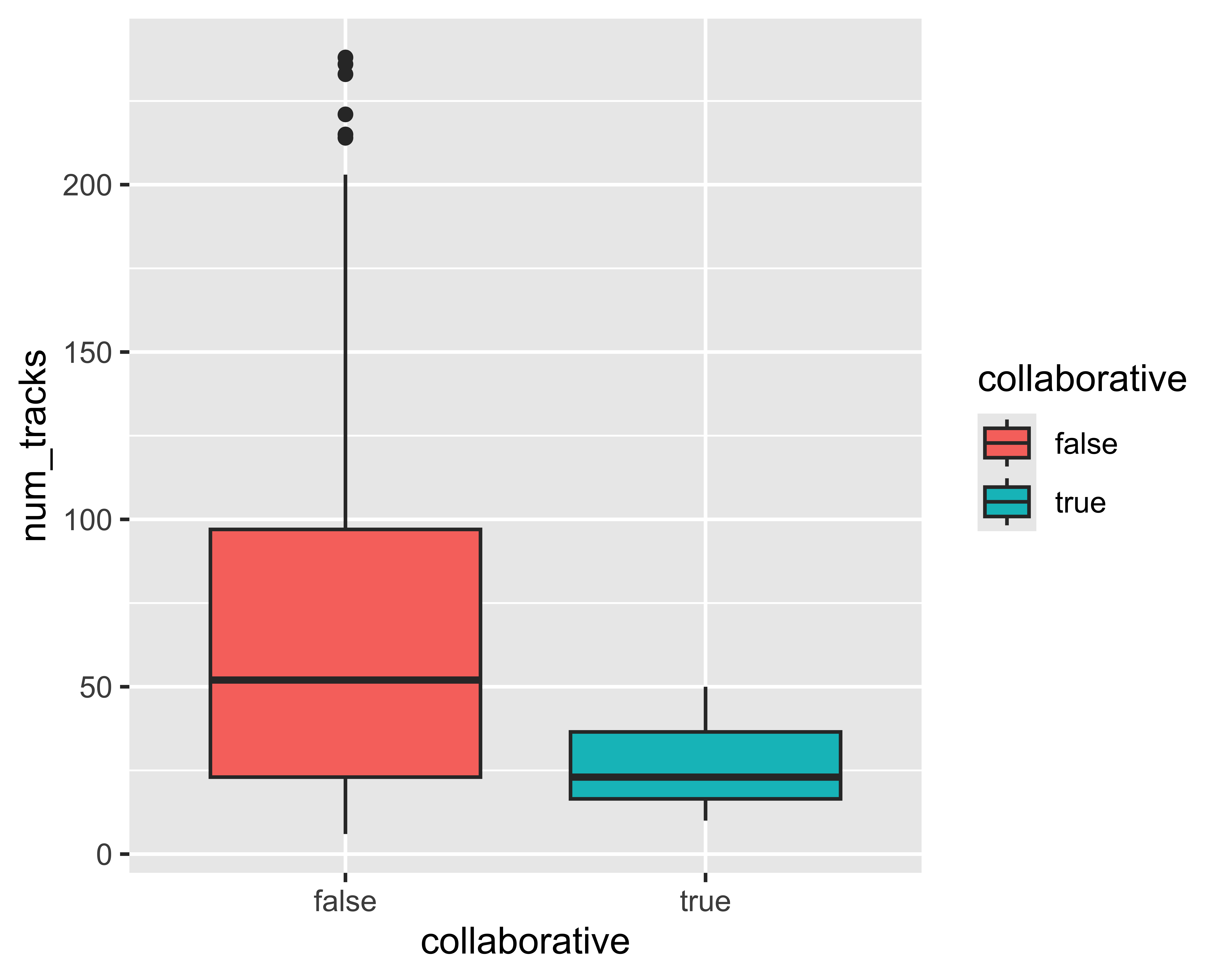
| Nombre del campo | Tipo del Campo | Descripción |
| --- | --- | --- |
| artist\_name | Caracteres | Nombre del artista |
| track\_uri | Caracteres | Identificador unico de cancion |
| artist\_uri | Caracteres | Identificador unico de artista |
| track\_name | Caracteres | Nombre de la cancion |
| album\_uri | Caracteres | Identificador unico del album |
| duration\_ms.x | Caracteres | Duracion de la cancion en ms. |
| album\_name | Caracteres | Nombre del album |
| danceablity | Real | Describe el atributo de baile de una cancion. |
| energy | Real | Describe en terminos de intensidad y actividad una cancion. |
| loudness | Real | Describe el volumen de la cancion. |
| speechiness | Real | Describe la presencia de palabras en una cancion |
| acousticness | Real | Describe si una cancion es acustica. |
| instrumentalness | Real | Describe si una cancion no contiene voces. |
| liveness | Real | Describe si hay presencia de audiencia en una cancion. |
| tempo | Real | Describe el actual BPM de una cancion. |

## 2.2 Estadisticas descriptivas

A continuacion se presentan los primeros resultados de nuestro analisis exploratorio inicial.

### 2.2.1 Tabla Lista de Reproducciones.

library(tidyverse)  
library(hrbrthemes)  
library(viridis)  
library(magrittr)  
library(ggplot2)  
library(dplyr)  
library(flextable)  
  
playlists <- read.csv("data/playlists.csv")  
  
playlists\_by\_group <- playlists %>% group\_by(collaborative) %>% summarise(mnum\_tracks=mean(num\_tracks))  
#playlists\_by\_group %>% ggplot( aes(x=collaborative, y=mnum\_tracks, fill=collaborative)) + geom\_boxplot()  
playlists %>% ggplot( aes(x=collaborative, y=num\_tracks, fill=collaborative)) + geom\_boxplot()

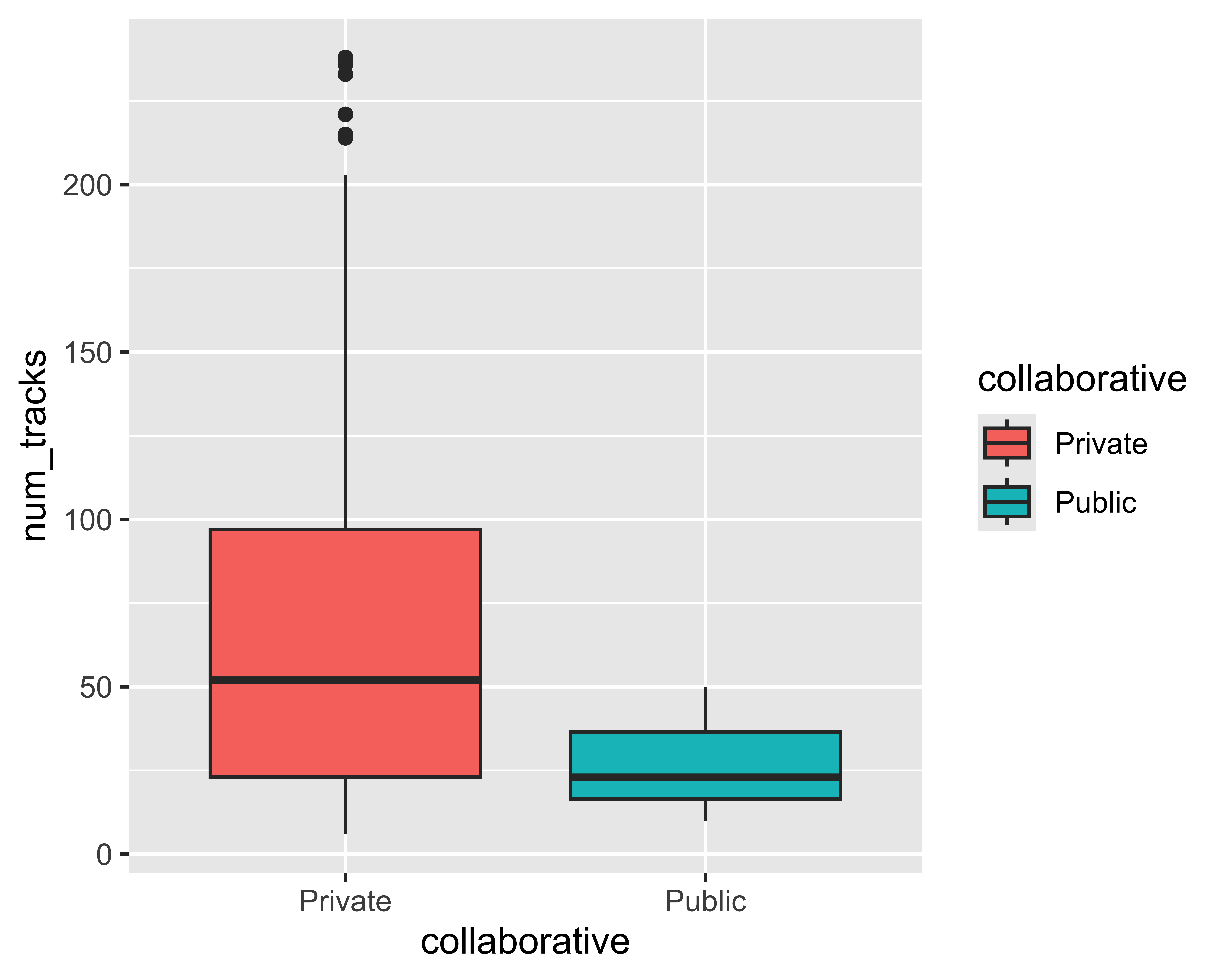


# Se clona la lista de canciones (playlist) para propositos de visualizacion  
data\_summary <- playlists[, c("collaborative", "num\_tracks", "num\_albums", "num\_followers")]  
data\_summary$collaborative <- ifelse(data\_summary$collaborative, "Public", "Private")  
summary <- data\_summary %>% summarizor(by = c("collaborative"), overall\_label = "Total")   
summary\_table <- summary %>% as\_flextable(spread\_first\_col = TRUE)  
summary\_table

|  |  | Private (N=97) |  | Public (N=3) |  | Total (N=100) |
| --- | --- | --- | --- | --- | --- | --- |
| num\_tracks | | | | | | |
| Mean (SD) |  | 69.3 (59.7) |  | 27.7 (20.4) |  | 68.1 (59.3) |
| Median (IQR) |  | 52.0 (74.0) |  | 23.0 (20.0) |  | 49.5 (70.2) |
| Range |  | 6.0 - 238.0 |  | 10.0 - 50.0 |  | 6.0 - 238.0 |
| num\_albums | | | | | | |
| Mean (SD) |  | 50.8 (43.3) |  | 24.3 (16.3) |  | 50.0 (42.9) |
| Median (IQR) |  | 41.0 (51.0) |  | 21.0 (16.0) |  | 39.5 (51.0) |
| Range |  | 3.0 - 194.0 |  | 10.0 - 42.0 |  | 3.0 - 194.0 |
| num\_followers | | | | | | |
| Mean (SD) |  | 1.5 (1.5) |  | 1.0 (0.0) |  | 1.5 (1.5) |
| Median (IQR) |  | 1.0 (0.0) |  | 1.0 (0.0) |  | 1.0 (0.0) |
| Range |  | 1.0 - 11.0 |  | 1.0 - 1.0 |  | 1.0 - 11.0 |

### 2.2.2 Distribucion de Listas de Reproduccion

library(tidyverse)  
library(hrbrthemes)  
library(viridis)  
library(magrittr)  
library(ggplot2)  
library(dplyr)  
library(flextable)  
  
playlists <- read.csv("data/playlists.csv")  
  
#p\_bygroup <- playlists %>% group\_by(collaborative) %>% summarise(mnum\_tracks=mean(num\_tracks))  
#playlists\_by\_group %>% ggplot( aes(x=collaborative, y=mnum\_tracks, fill=collaborative)) + geom\_boxplot()  
  
p\_bygroup <- playlists[, c("collaborative", "num\_tracks")]  
p\_bygroup$collaborative <- ifelse(p\_bygroup$collaborative, "Public", "Private")  
p\_bygroup %>% ggplot( aes(x=collaborative, y=num\_tracks, fill=collaborative)) + geom\_boxplot()



p\_bygroup

collaborative num\_tracks  
1 Private 83  
2 Private 52  
3 Private 161  
4 Private 53  
5 Private 203  
6 Private 56  
7 Private 162  
8 Private 62  
9 Private 97  
10 Private 36  
11 Private 165  
12 Private 18  
13 Private 233  
14 Private 42  
15 Private 16  
16 Private 171  
17 Private 49  
18 Private 25  
19 Private 140  
20 Private 75  
21 Private 113  
22 Private 29  
23 Private 45  
24 Private 46  
25 Private 16  
26 Private 37  
27 Private 100  
28 Private 30  
29 Private 8  
30 Private 85  
31 Private 10  
32 Private 238  
33 Private 26  
34 Private 23  
35 Private 12  
36 Private 55  
37 Private 88  
38 Private 17  
39 Private 58  
40 Private 16  
41 Private 23  
42 Private 17  
43 Private 59  
44 Private 22  
45 Private 14  
46 Private 42  
47 Private 114  
48 Private 72  
49 Private 17  
50 Private 103  
51 Private 104  
52 Private 31  
53 Private 68  
54 Private 74  
55 Private 122  
56 Private 36  
57 Private 35  
58 Private 28  
59 Private 48  
60 Private 52  
61 Private 14  
62 Public 23  
63 Private 26  
64 Private 88  
65 Private 164  
66 Private 14  
67 Private 236  
68 Private 12  
69 Private 15  
70 Private 23  
71 Private 85  
72 Private 121  
73 Private 44  
74 Private 6  
75 Private 12  
76 Private 41  
77 Private 78  
78 Private 11  
79 Private 92  
80 Private 10  
81 Private 134  
82 Private 103  
83 Private 7  
84 Private 214  
85 Private 104  
86 Public 10  
87 Private 55  
88 Private 49  
89 Private 24  
90 Private 18  
91 Private 55  
92 Private 215  
93 Private 44  
94 Public 50  
95 Private 81  
96 Private 221  
97 Private 109  
98 Private 61  
99 Private 59  
100 Private 47

## 2.3 El Artista Mas Escuchado

library(wordcloud2)  
library(readr)  
library(magrittr)  
library(dplyr)  
  
dtracks <- read\_csv("data/tracks.csv")  
wc\_tracks <- dtracks %>% group\_by(artist\_name) %>% summarise(n = n\_distinct(track\_uri))  
  
wordcloud2(data = wc\_tracks, size = 2)

# 3 La cancion mas escuchada

library(wordcloud2)  
library(readr)  
library(magrittr)  
library(dplyr)  
  
dtracks <- read\_csv("data/tracks.csv")  
wc\_tracks <- dtracks %>% group\_by(album\_name) %>% summarise(n = n\_distinct(track\_uri))  
  
wordcloud2(data = wc\_tracks, size = 2)

## 3.1 Conclusiones

Se puede apreciar información relevante sobre el análisis exploratorio. La mayoría de las listas de reproducción son privadas, en comparación con las listas de reproducción pública. Esto refuerza la idea inicial de que al crear un modelo basado en listas de reproducción (Contenido generado por usuario) podría llevar a aumentar la satisfacción en la experiencia de los usuarios de las plataformas de música en línea. Los artistas y álbumes más escuchados podrían utilizarse como características del modelo de recomendación de música automática. En buena medida reflejan la relación que existe entre los usuarios dueños de sus listas de reproducción y sus preferencias musicales.

Lu, Y., Z. Qiao, P. Zhang, and L. Guo. 2016. “Ranking-Based Music Recommendation in Online Music Radios.” In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, 614–19. Los Alamitos, CA, USA: IEEE Computer Society. <https://doi.org/10.1109/DSC.2016.102>.

Pichl, M., E. Zangerle, and G. Specht. 2016. “Understanding Playlist Creation on Music Streaming Platforms.” In *2016 IEEE International Symposium on Multimedia (ISM)*, 475–80. Los Alamitos, CA, USA: IEEE Computer Society. <https://doi.org/10.1109/ISM.2016.0107>.

“Wikipedia List of Online Music Databases.” n.d. <https://en.wikipedia.org/wiki/List_of_online_music_databases>.

1. Politécnico Grancolombiano, [jajaimes4@poligran.edu.co](mailto:jajaimes4@poligran.edu.co) [↑](#footnote-ref-20)
2. Politécnico Grancolombiano, [papineda1@poligran.edu.co](mailto:papineda1@poligran.edu.co) [↑](#footnote-ref-22)
3. Politécnico Grancolombiano, [brvalencia6@poligran.edu.co](mailto:brvalencia6@poligran.edu.co) [↑](#footnote-ref-24)
4. Politécnico Grancolombiano, [jcifuntes@poligran.edu.co](mailto:jcifuntes@poligran.edu.co) [↑](#footnote-ref-26)
5. Politécnico Grancolombiano, [mdpineda1@poligran.edu.co](mailto:mdpineda1@poligran.edu.co) [↑](#footnote-ref-28)