Recomendación de canciones basada en listas de reproducción de servicios de música online.

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

Aqui viene el abstract

Table of Contents

# 1 Introducción

[Spotiy](https://spotify.com) es un servicio en linea de musica, con mas de 140 millones de usuarios activos alrededor del mundo y mas de 20 millones de canciones. Ademas de lo anterior el servicio cuenta con mas de 2 billones de listas de reproduccion a diciembre de 2018. A pesar de eso una de sus funcionalidad principal

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# 2 Descripcion del Conjunto de Datos

Se obtiene acceso al conjunto de datos principal [Million Playlist Dataset](https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge) publicado por Spotify en la plataforma [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.

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()  
  
  
#load spotify metadata info  
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)) {  
 cat("---")  
   
 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()  
  
#extract\_id <- function(l) {  
# parts <- strsplit(l, ":")  
# return(parts[[1]][3])  
#}  
  
#spotify\_tracks\_ids <- sapply(spotify\_track\_uris, extract\_id)  
  
  
  
#ids <- paste(spotify\_tracks\_id\_by\_100[1], collapse = NULL)  
#tra <- left\_join(x = tracks, y = egresos\_clientes, by = c("id" = "id2")) %>% na.omit()  
  
# "xyz,yz"  
  
#tr <- unique(tr$track\_name)  
  
# Crearemos dos archivos para sus respectivos analisis estadistico. El primer archivo es la lista de listas de reproduccion, mientras que el segundo es la lista de canciones encontradas en las listas de reproduccion.  
#for (f in files[-1]) {  
# i <- fromJSON(f)  
   
# # Join the playlists to the original d dataframe  
# j = i$playlists %>% select(c("name", "collaborative", "pid", "modified\_at", "num\_tracks", "num\_albums", "num\_followers"))  
# d <- rbind(d, j)  
   
 # Join tracks to the tr dataframe  
# jtr <- do.call("rbind", t$playlists[]$tracks) %>% select(-c("pos"))  
# tr <- rbind(tr, jtr)  
#}  
  
#View(tr)  
#tr <- t$playlists$tracks %>% mutate(id = n)  
  
#str(t)  
# d <- t %>% select(c("playlists"))  
# View(d)  
#View(t$playlists)  
  
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)  
  
# Loading   
  
#t$playlists$tracks[1]  
#as.data.frame(unlist(t$playlists$tracks))  
#unlist(t$playlists$tracks)  
  
#mpd <- lapply(files, fromJSON)  
#df <- rbind\_pages(mdp)

A partir de la muestra se construyen los siguientes conjuntos de datos, con information que se obtiene a traves de la integracion de la API de Spotify.

## 2.1 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. |

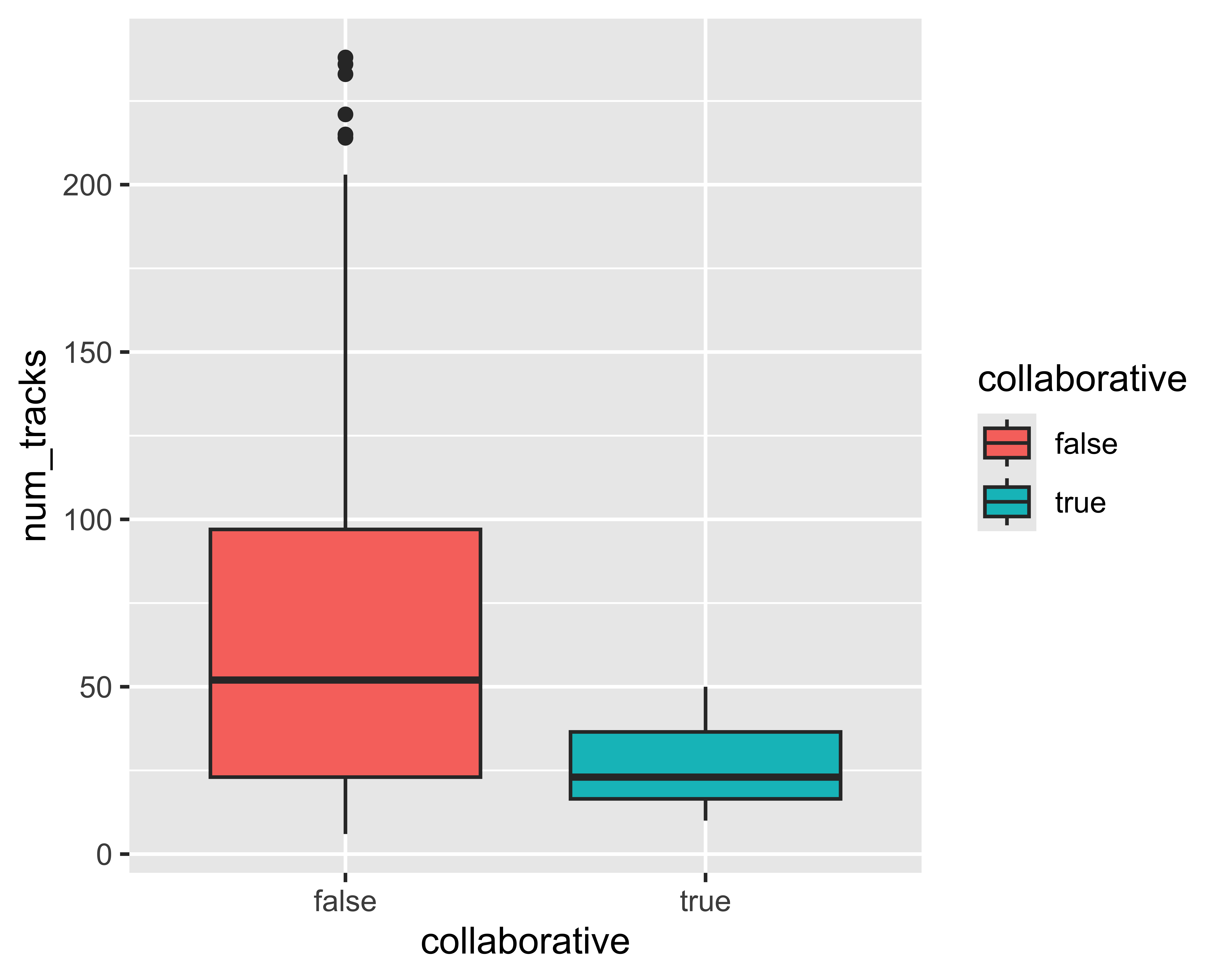
### 2.1.1 Canciones

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 del conjunnto de datos

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()

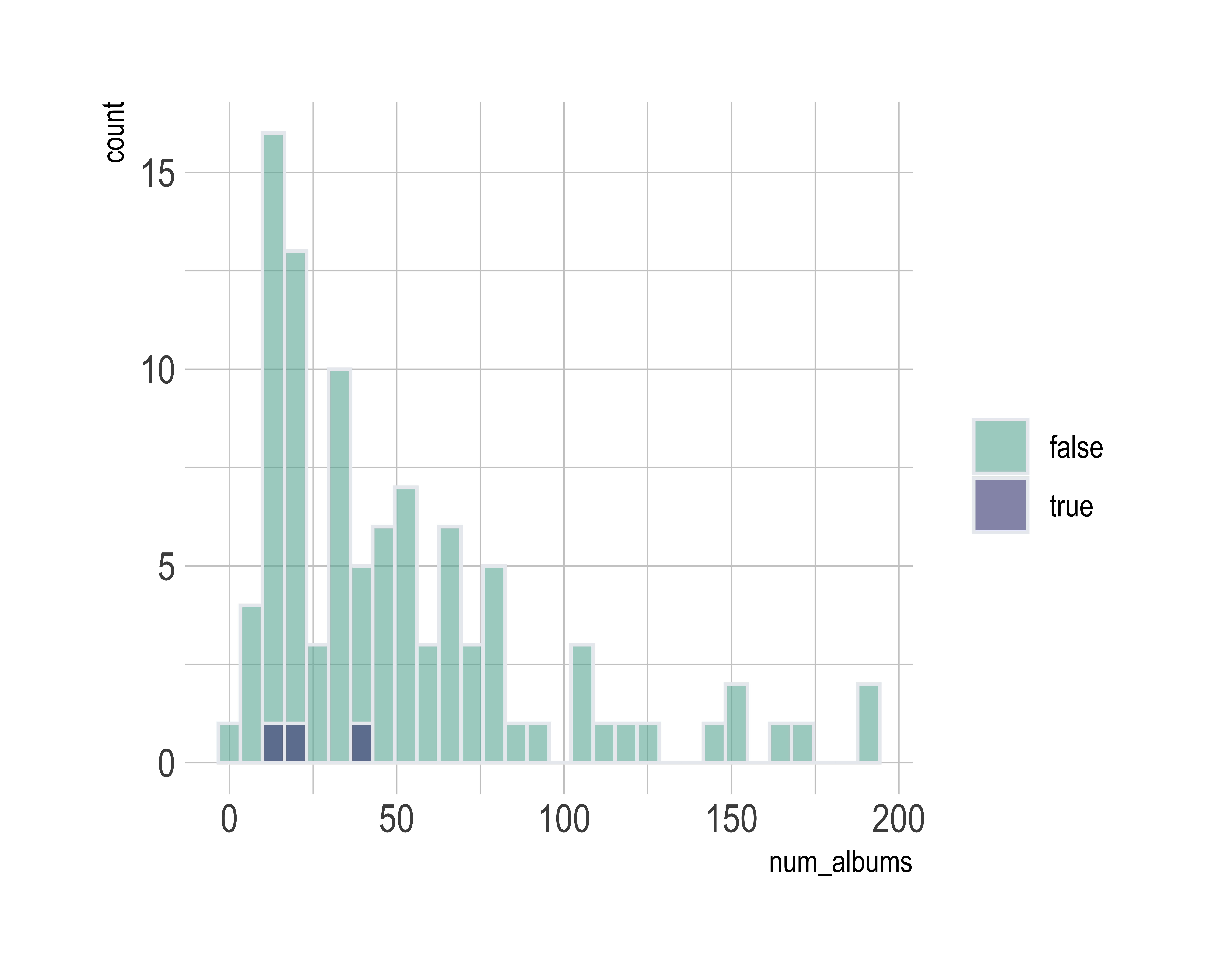


#table <- playlists %>% group\_by(collaborative) %>% summarise(mean\_num\_tracks=mean(num\_tracks), mean\_num\_albums=mean(num\_albums), sd\_num\_tracks=sd(num\_tracks), sd\_num\_albums=sd(num\_albums), total\_tracks=sum(num\_tracks), total\_albums=sum(num\_albums))  
  
#flextable(table) %>% theme\_vanilla()  
  
# 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)  
#set\_  
#sumarizor\_table <- set\_header\_labels(summary\_table, values = list(  
# collaborative: "Public/Private"  
#))  
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 |

# 3 Histograma de frecuencia por Numero de Canciones

library(tidyverse)  
library(hrbrthemes)  
library(viridis)  
library(magrittr)  
library(ggplot2)  
library(dplyr)  
  
playlists <- read.csv("data/playlists.csv")  
  
playlists %>% ggplot( aes(x=num\_albums, fill=collaborative)) + geom\_histogram( color="#e9ecef", alpha=0.6, position = 'identity') + scale\_fill\_manual(values=c("#69b3a2", "#404080")) + theme\_ipsum() + labs(fill="")

 ## El Artista Mas Escuchado

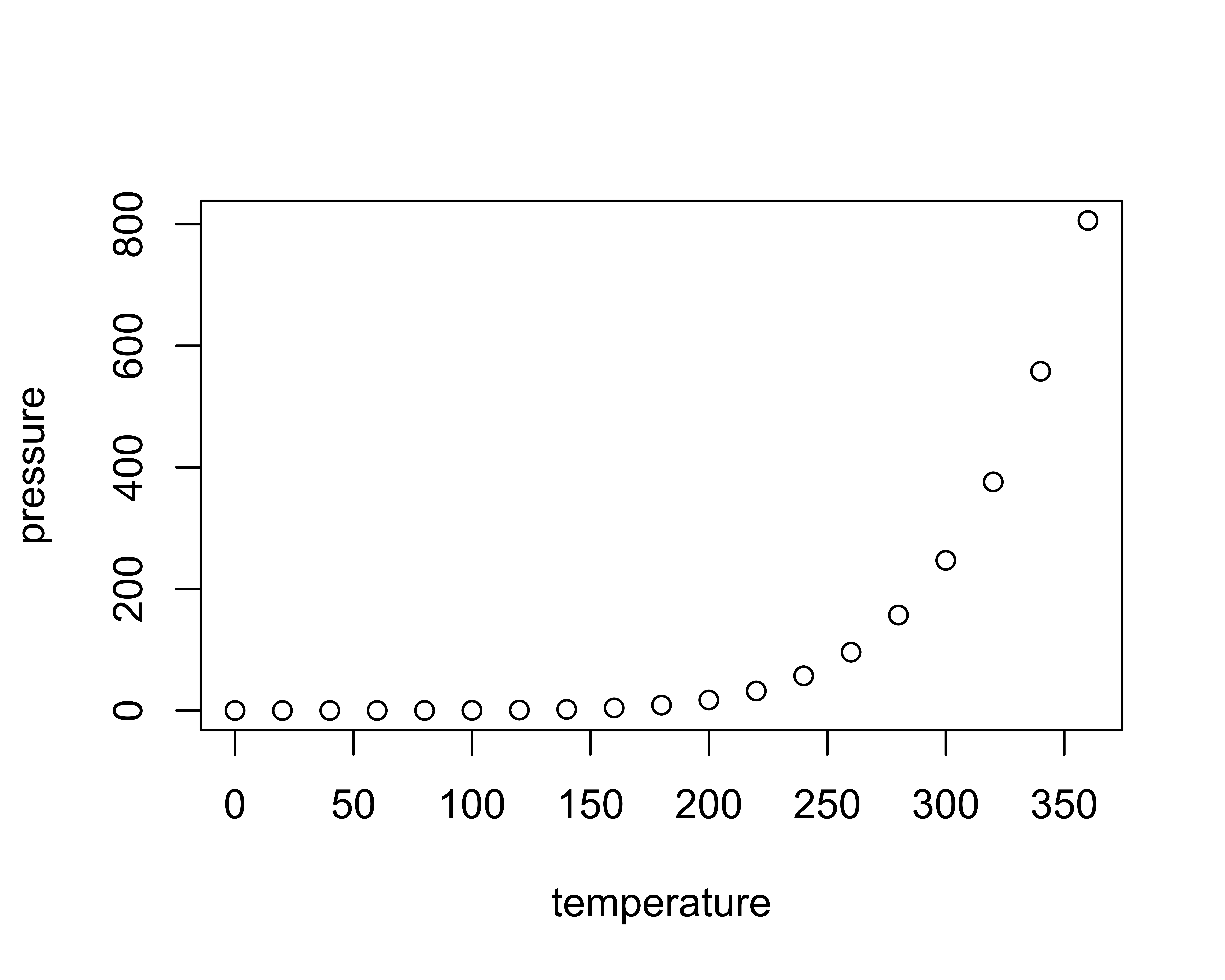
library(wordcloud2)  
library(readr)  
library(magrittr)  
  
dtracks <- read\_csv("data/tracks.csv")  
wc\_tracks <- dtracks %>% group\_by(artist\_name) %>% summarise(n = n\_distinct(track\_uri))  
  
wordcloud2(data = wc\_tracks, size = 2)

# 4 La cancion mas escuchada

library(wordcloud2)  
library(readr)  
library(magrittr)  
  
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)

## 4.1 Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

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)