Who am I?

Product Manager

FinTech

Part Time Student MSCA

#Rstats fanatic

@JorgeArgueta

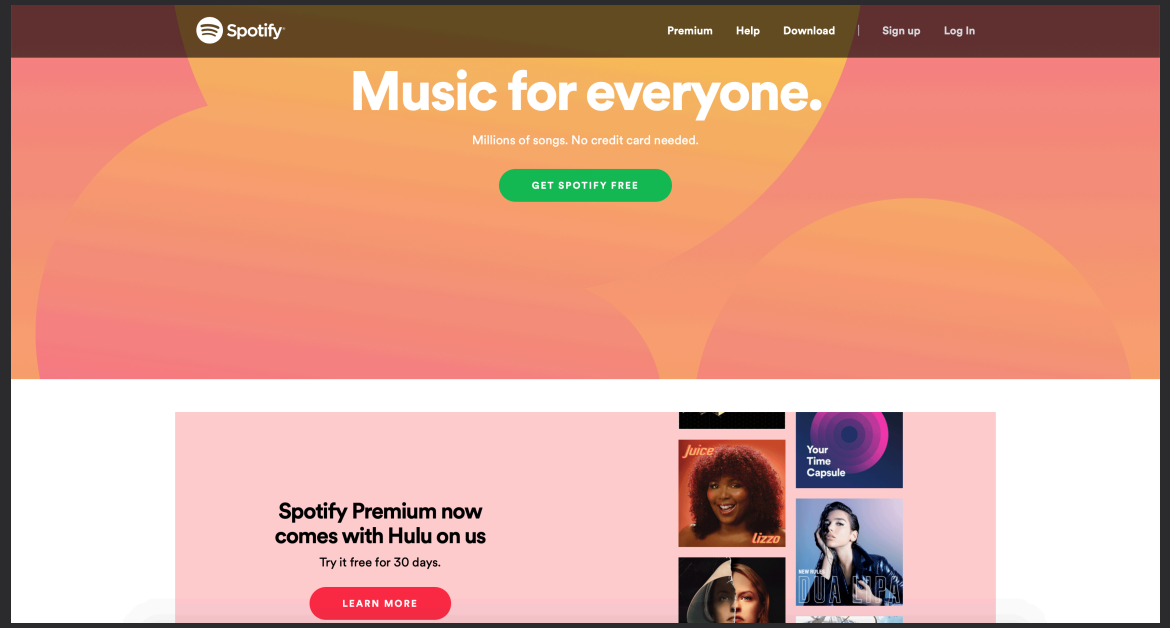
WEB SCRAPPING WWW.SPOTIFY.COM

LYRICS **SENTIMENT ANALYSIS**

CLUSTERING TOP 100 SONGS USA

by Jorge Argueta

What is Spotify?



Daily Top 200 Songs

2018 USA

Web Scraping

WEB SCRAPING

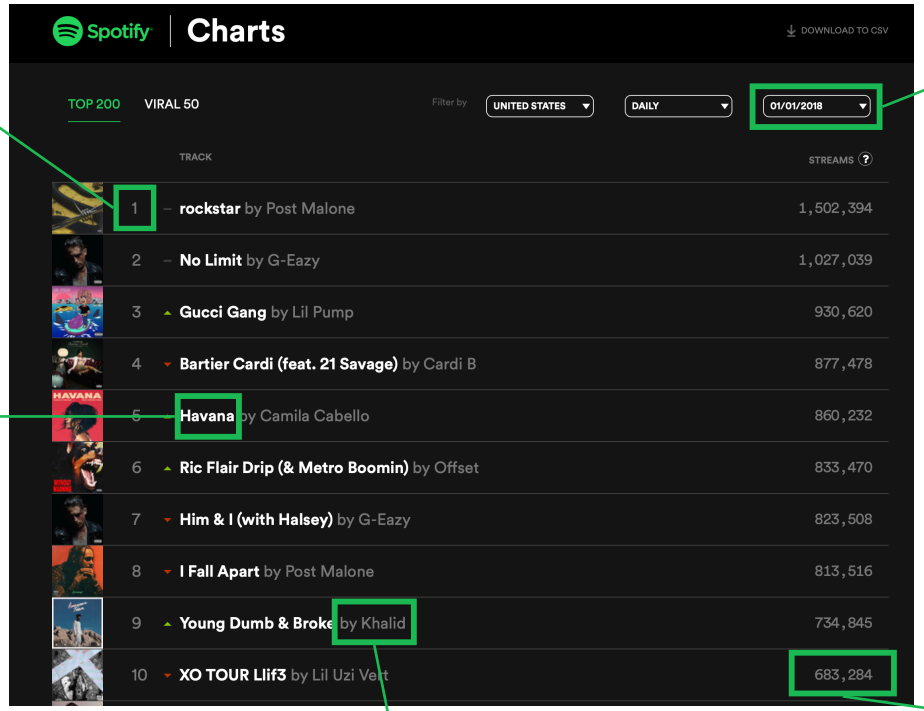
rvest::html_nodes extract pieces out of HTML documents

RANKING OF SONGS

html_nodes('.chart-table-position')

DATE

html_nodes('.responsive-select+... ')



NAMES OF SONGS

html_nodes('strong')

TOTAL DAILY STREAMS

html_nodes('.td.chart-table-streams')

NAMES OF THE ARTIST

html_nodes('.chart-table-track span')

WEB SCRAPING

rvest::html_nodes extract pieces out of HTML documents

```{r}
SpotifyScrape <- function(x){
  page <- x
  rank <- page %>%
    read_html() %>% #Reads an HTML page
    html_nodes('.chart-table-position') %>% #RVEST.PKG: extract pieces out of HTML docs. using XPath & css selectors.
    html_text() %>% #RVEST.PKG:Extract attributes, text and tag name from html
    as.data.frame()
  track <- page %>%
    read_html() %>%
    html_nodes('strong') %>%
    html_text() %>%
    as.data.frame()
  artist <- page %>%
    read_html() %>%
    html_nodes('.chart-table-track span') %>%
    html_text() %>%
    as.data.frame()
  streams <- page %>%
    read_html() %>%
    html_nodes('td.chart-table-streams') %>%
    html_text() %>%
    as.data.frame()
  dates <- page %>%
    read_html() %>%
    html_nodes('.responsive-select~ .responsive-select+ .responsive-select .responsive-select-value') %>%
    html_text() %>%
    as.data.frame()

#combine, name, and make it a tibble
  chart <- cbind(rank, track, artist, streams, dates) #Combine R Objects by Columns
  names(chart) <- c("Rank", "Track", "Artist", "Streams", "Date") #Functions to get or set the names of an object
  chart <- as.tibble(chart)#TIBBLE.PKG:turns an existing object into a so-called tibble
 return(chart) #Final tibble 5 columns & (200 rows * 365 days) = 73,000
}
```

SCRAPPING SUMMARY

365 URLS * (200 ROWS * 5 COLMS)

dim(spotify)

[1] 73000 5

WEB SCRAPING

rvest::html_nodes extract pieces out of HTML documents

```{r}
head(spotify, n = 10)
```

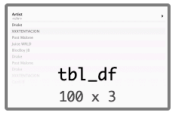
Rank <fctr>	Track <chr>	Artist <chr>	Streams <dbl>	Date <date>	WeekDay <ord>	Month <ord>
1	rockstar	Post Malone	1502394	2018-01-01	Mon	Jan
2	No Limit	G-Eazy	1027039	2018-01-01	Mon	Jan
3	Gucci Gang	Lil Pump	930620	2018-01-01	Mon	Jan
4	Bartier Cardi (feat. 21 Savage)	Cardi B	877478	2018-01-01	Mon	Jan
5	Havana	Camila Cabello	860232	2018-01-01	Mon	Jan
6	Ric Flair Drip (& Metro Boomin)	Offset	833470	2018-01-01	Mon	Jan
7	Him & I (with Halsey)	G-Eazy	823508	2018-01-01	Mon	Jan
8	I Fall Apart	Post Malone	813516	2018-01-01	Mon	Jan
9	Young Dumb & Broke	Khalid	734845	2018-01-01	Mon	Jan
10	XO TOUR Llif3	Lil Uzi Vert	683284	2018-01-01	Mon	Jan

1–10 of 10 rows

```
[1] 73000     7
```

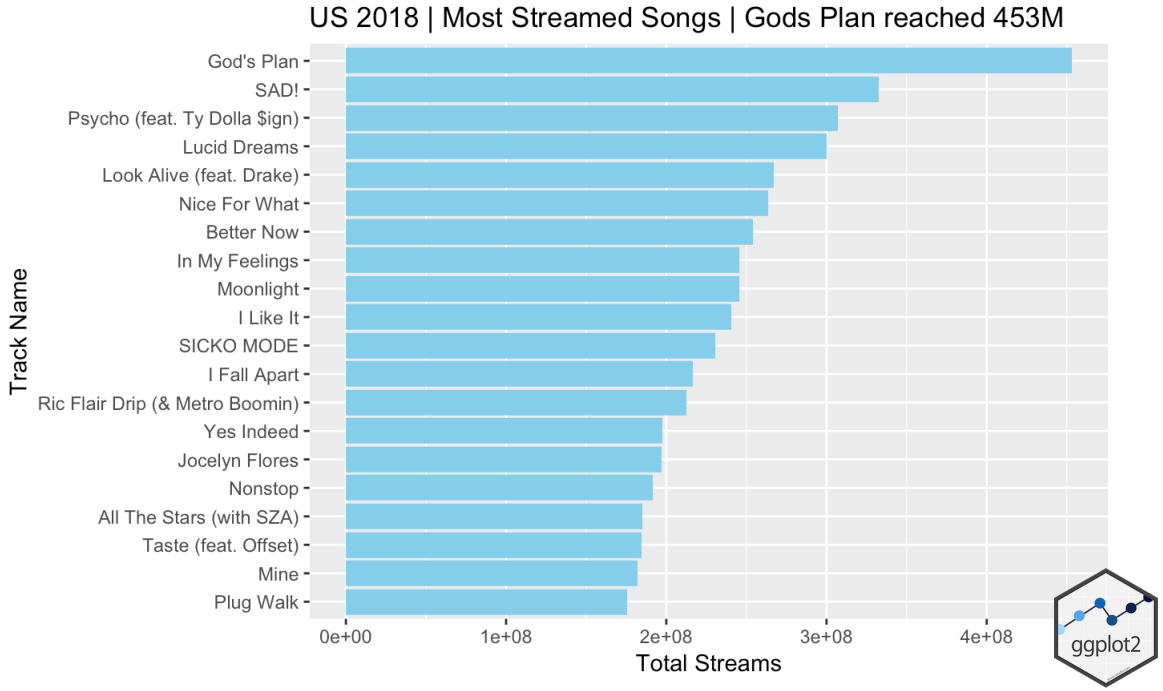
dplyr::group_by group by one or more variables

```
#Group by track and sum Total Streams
by_streams2 <- spotify %>%
  group_by(Track) %>%
  summarise(TotalStreams = sum(Streams)) %>%
  arrange(desc(TotalStreams)) %>%
  top_n(100)

#Create a df with unique tracks and artists
spotify2 <- spotify %>%
  select(Track, Artist) %>%
  distinct(Track, Artist)

#Left join to prep our data and get the lyrics
top100songs <- left_join(by_streams2, spotify2, by = "Track") %>%
  arrange(desc(TotalStreams)) %>%
  select(Artist, Track, TotalStreams) %>%
  filter (! duplicated(TotalStreams)) %>%
  print()
```



Artist <chr>	Track <chr>	TotalStreams <dbl>
Drake	God's Plan	453226629
XXXTENTACION	SAD!	332633597
Post Malone	Psycho (feat. Ty Dolla $ign)	306877012
Juice WRLD	Lucid Dreams	299907223
BlocBoy JB	Look Alive (feat. Drake)	266861797
Drake	Nice For What	263455062
Post Malone	Better Now	254098207
Drake	In My Feelings	245715031
XXXTENTACION	Moonlight	245626527
Cardi B	I Like It	240430007

1–10 of 100 rows



Sentiment Analysis

GENIUS

genius::genius_lyrics retrieves song lyrics from Genius.com

```
#1st get the song lyrics
song1<-genius_lyrics(artist=top100songsv2$Artist[1],song = top100songsv2$Track[1])
```

track_title <chr>	line <int>	lyric <chr>
God's Plan	1	And they wishin' and wishin' and wishin' and wishin'
God's Plan	2	They wishin' on me, yuh
God's Plan	3	I been movin' calm, don't start no trouble with me
God's Plan	4	Tryna keep it peaceful is a struggle for me
God's Plan	5	Don't pull up at 6 AM to cuddle with me
God's Plan	6	You know how I like it when you lovin' on me
God's Plan	7	I don't wanna die for them to miss me
God's Plan	8	Yes, I see the things that they wishin' on me
God's Plan	9	Hope I got some brothers that outlive me
God's Plan	10	They gon' tell the story, shit was different with me

1–10 of 20 rows

tidytext::unnest_tokens split a column into tokens

```
song1 %>%
 unnest_tokens(word, lyric) %>%
 select(word) %>%
  print()
```

word <chr>
and
they
wishin
and
wishin
and
wishin
and
wishin
they

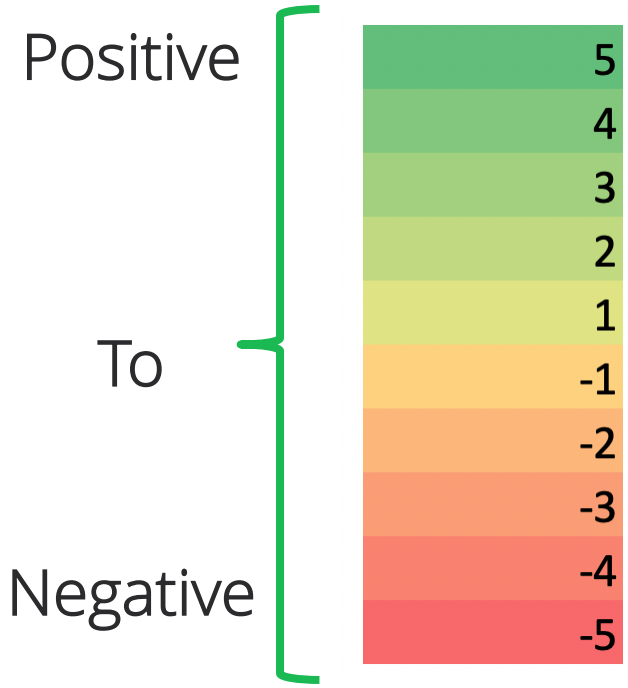
1–10 of 347 rows

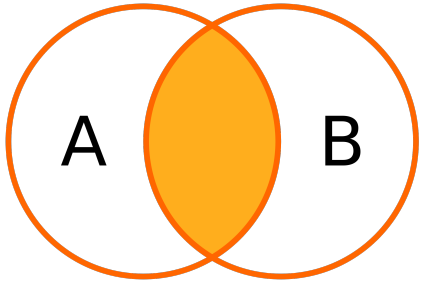
TRANSFORMATION
FROM 20 ROWS
TO 347 ROWS

1 word : 1 row

tidytext::get_sentiments get a tidy data frame of a single sentiment lexicon

LEXICON “afinn” (N = 2,476)






```
song1 %>%
 unnest_tokens(word, lyric) %>%
 select(word) %>%
 inner_join(lexicon)
```

R Console

tbl_df
33 x 2

word <chr>	score <int>
calm	2
no	-1
trouble	-2
peaceful	2
struggle	-2
like	2
die	-3
miss	-2
yes	1
hope	2

1–10 of 33 rows

tidytext::get_sentiments get a tidy data frame of a single sentiment lexicon

```r
```{r}
sentiments <- sapply(
  #X = 1:5
  X = 1:nrow(top100songsv2)
  , FUN = function(row_num, topSongTBL){

      sentiment <- NA
      tryCatch({
          lyricTBL <- genius::genius_lyrics(
            artist = topSongTBL[["Artist"]][row_num]
            , song = topSongTBL[["Track"]][row_num]
          )

        sentiment <- lyricTBL %>%
          unnest_tokens(word, lyric) %>%
          select(word) %>%
          inner_join(lexicon) %>%
          summarise(score = sum(score))

        sentiment <- sentiment[[1]]

      }, error = function(e){
          print(paste0("Failed for song name: ", topSongTBL[["Track"]][row_num]))
      })

      return(sentiment)
  }
  , topSongTBL = top100songsv2
)
print(as.data.frame(sentiments))
```
```

Artist <chr>	Track <chr>	TotalStreams <dbl>	Sentiment <dbl>
Drake	Gods Plan	453226629	-24
XXXTENTACION	SAD!	332633597	-40
Post Malone	Psycho	306877012	-18
Juice WRLD	Lucid Dreams	299907223	-36
BlocBoy JB	Look Alive	266861797	-82
Drake	Nice For What	263455062	-14
Post Malone	Better Now	254098207	39
Drake	In My Feelings	245715031	-43
XXXTENTACION	Moonlight	245626527	-50
Cardi B	I Like It	240430007	123

1–10 of 100 rows 3 4 5 6 … 10 Next

Hierarchical Clustering

dplyr::mutate create or transform variables

```{r}
sentiment1 <- sentiment %>%
  na.omit() %>%                                  # listwise deletion of missing
  mutate(Track_Artist = paste0(Track, " by ", Artist),
         TotalStreamsScld = scale(TotalStreams),  # standardize variables
         SentimentScld = scale(Sentiment)) %>%    # standardize variables
  select(Track_Artist,TotalStreams, Sentiment, TotalStreamsScld, SentimentScld ) %>%
  print()
```

Track_Artist <chr>	TotalStreams <dbl>	Sentiment <dbl>	TotalStreamsScld <dbl>	SentimentScld <dbl>
Gods Plan by Drake	453226629	-24	4.764429981	-0.139971820
SAD! by XXXTENTACION	332633597	-40	2.904255314	-0.376693759
Psycho by Post Malone	306877012	-18	2.506954192	-0.051201093
Lucid Dreams by Juice WRLD	299907223	-36	2.399443628	-0.317513274
Look Alive by BlocBoy JB	266861797	-82	1.889710492	-0.998088848
Nice For What by Drake	263455062	-14	1.837160838	0.007979391
Better Now by Post Malone	254098207	39	1.692829243	0.792120813
In My Feelings by Drake	245715031	-43	1.563516866	-0.421079122
Moonlight by XXXTENTACION	245626527	-50	1.562151672	-0.524644971
I Like It by Cardi B	240430007	123	1.481994181	2.034910990

1–10 of 89 rows

CLUSTERING

stats::hclust

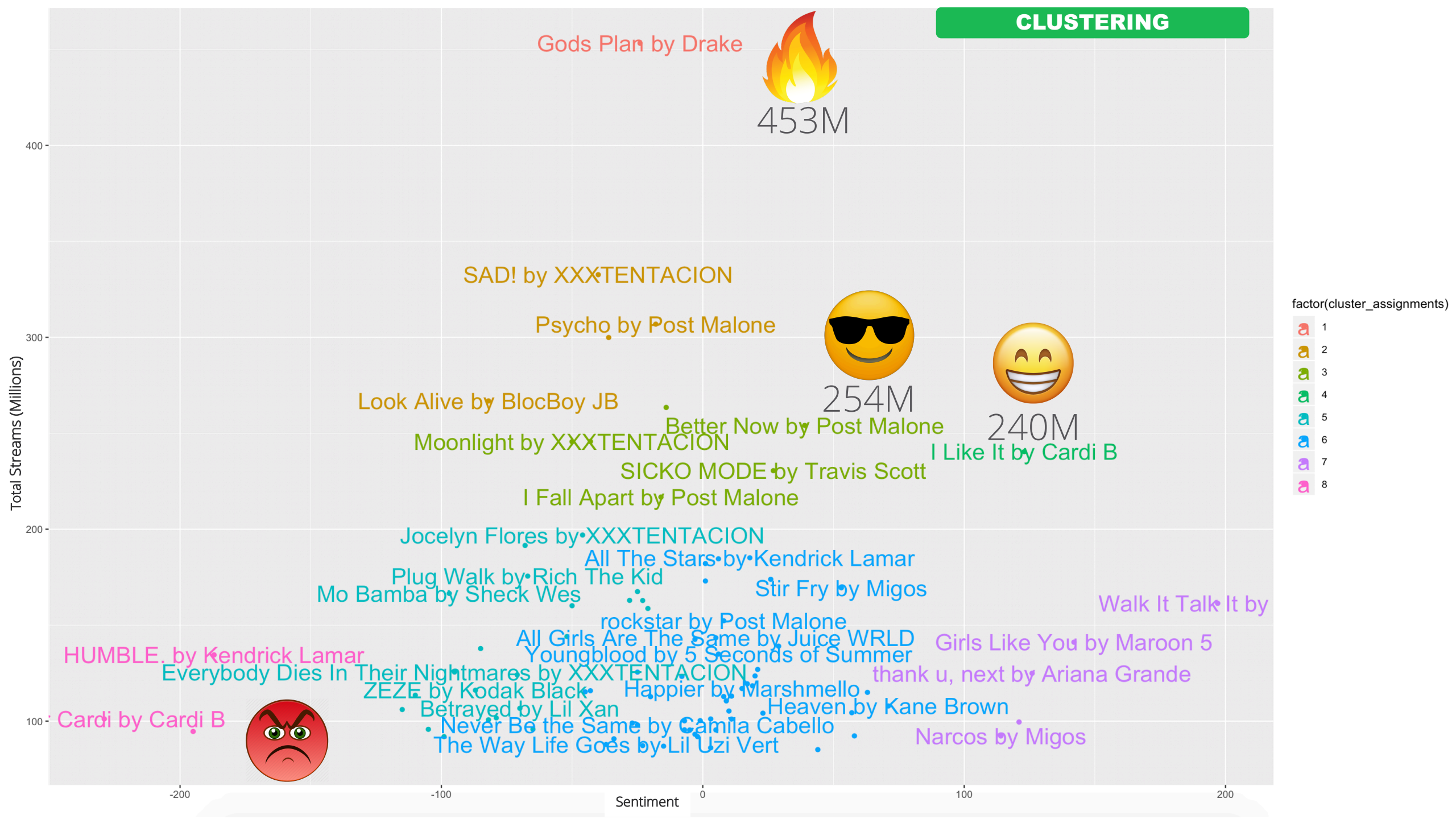
Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.

```{r}
dist_songs <- dist(sentiment2, method = "euclidean")
hc_songs <- hclust(dist_songs, method = "complete")
cluster_assignments <- cutree(hc_songs, k = 8)
cluster_assignments2 <- as.data.frame(cluster_assignments)

sentiment3 <- cbind(sentiment1, cluster_assignments2)
```

Artist <chr>	Track <chr>	TotalStreams <dbl>	Sentiment <dbl>	Track_Artist <chr>	TotalStreamsScld <dbl>	SentimentScld <dbl>	cluster_assignments <int>
Drake	Gods Plan	453226629	-24	Gods Plan by Drake	4.764429981	-0.139971820	1
XXXTENTACION	SAD!	332633597	-40	SAD! by XXXTENTACION	2.904255314	-0.376693759	2
Post Malone	Psycho	306877012	-18	Psycho by Post Malone	2.506954192	-0.051201093	2
Juice WRLD	Lucid Dreams	299907223	-36	Lucid Dreams by Juice WRLD	2.399443628	-0.317513274	2
BlocBoy JB	Look Alive	266861797	-82	Look Alive by BlocBoy JB	1.889710492	-0.998088848	2
Drake	Nice For What	263455062	-14	Nice For What by Drake	1.837160838	0.007979391	3
Post Malone	Better Now	254098207	39	Better Now by Post Malone	1.692829243	0.792120813	3
Drake	In My Feelings	245715031	-43	In My Feelings by Drake	1.563516866	-0.421079122	3
XXXTENTACION	Moonlight	245626527	-50	Moonlight by XXXTENTACION	1.562151672	-0.524644971	3
Cardi B	I Like It	240430007	123	I Like It by Cardi B	1.481994181	2.034910990	4

1–10 of 89 rows

Previous 1 2 3 4 5 6 ... 9 Next


CLUSTERING
Gods Plan by Drake
453M
SAD! by XXXTENTACION
Psycho by Post Malone
254M
240M
Look Alive by BlocBoy JB
Better Now by Post Malone
Moonlight by XXXTENTACION
I Like It by Cardi B
SICKO MODE by Travis Scott
I Fall Apart by Post Malone
Jocelyn Flores by XXXTENTACION
All The Stars by Kendrick Lamar
Plug Walk by Rich The Kid
Mo Bamba by Sheck Wes
Stir Fry by Migos
Walk It Talk It by
rockstar by Post Malone
All Girls Are The Same by Juice WRLD
Girls Like You by Maroon 5
HUMBLE. by Kendrick Lamar
Youngblood by 5 Seconds of Summer
Everybody Dies In Their Nightmares by XXXTENTACION
thank u, next by Ariana Grande
ZEZE by Kodak Black
Happier by Marshmello
Betrayed by Lil Xan
Heaven by Kane Brown
Cardi by Cardi B
Never Be the Same by Camila Cabello
Narcos by Migos
The Way Life Goes by Lil Uzi Vert
Total Streams (Millions)
400
300
200
100
-200
-100
0
100
200
Sentiment
factor(cluster_assignments)
1
2
3
4
5
6
7
8

Where can you find me or my code?

https://github.com/argdata/talks

Thanks!!!