1. Yearly Top 100
   1. Regression ranking和 feature
   2. Text mining: lyrics title
   3. Which feature people like? (e.g. keys)
   4. At a glance:
      1. X: singer, Y: # of songs, bar chart
2. Daily Top 100
   1. Time series:
      1. Feature 变化 (e.g. 一年内的歌的dancability的变化)
   2. # of days in top 100 (top 20 songs)
   3. # of days in top 100 (top 20 singers)
   4. 动态变化图: x: 播放量, y: 歌, z: 时间
3. Pick singers
   1. Top 1 Singer:
      1. Popularity of his songs by time in the year (x: time, y: popularity, color: song)
   2. Compare two singers:
      1. Radar chart of song features
4. Pick songs
5. Trend
6. Region:
   1. Map showing features in different countries (1 map 1 feature)

We are using open-source data from Spotify, and collecting the features of each song using its API. We get the list of the songs to focus on by scraping data from [https://spotifycharts.com/regional](spotifychart) which contains the daily top 200 songs.

We also managed to get the information of top songs by each country using the crawler. However, since the data for some countries are missing, we are still discussing about the ways to handle it, such as focusing only on the continents with enough data.

Also, due to the size of the dataset (there are too many songs right now!), it would be impossible to carefully analyze every song on the list. We plan to focus on the top songs and the most popular singers.

```{r}

library(GGally)

ggpairs(df,columns= c(1,5,6,8,10,11,13,14,15,16))

summary(lm(ranking~danceability+energy+loudness+speechiness+acousticness+instrumentalness+liveness+valence+tempo, data = df))

```

```{r}

library(ggplot2)

df <- readr::read\_csv('top2018.csv')

df <- tibble::rowid\_to\_column(df, var = 'ranking' )

tb <- data.frame(table(df[,c("artists")]))

tb\_sort <- tb[which(tb$Freq>1),]

ggplot(data=tb\_sort, aes(x=reorder(Var1, -Freq), y=Freq)) +

geom\_bar(stat="identity", fill = "lightblue") +

xlab("Artists")+

ylab("Number of songs in Top 100")+

ggtitle("Hot Singers In 2018")+

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

```

lm:

To verify the result of the above plot, we then ran a linear regression with ranking as the dependent variable and the other features as independent variables. From the summary of the model, we can see that “Liveness” is indeed significant to the ranking of a song, with p-value smaller than 0.05. This has confirmed what we found from the above plot that there is a significant correlation between a song’s liveness with its ranking in 2018.

Histogram:

Just as mentioned before, we would like to reduce the scope of the singers we analyze. Thus, we only focus on those who have more than 2 hot songs in Top 100. As shown in the barchart, Post Malone and XXXTentacion both rank top on the hottest singer chart each with 6 songs. They are very popular rappers, which shows global enthusiasm towards rap in 2018. Also, Ed sheeran and Ariana Grande are still popular in 2018. We could focus on these singers for further investigation.

Useful link:

<https://www.kaggle.com/cihanoklap/top-songs-on-spotify-what-makes-them-popular/data>

(Ed on top100 daily trend)

<https://www.kaggle.com/fangya/spotify-2017-popular-song-analysis>

(nothing more)

<https://www.kaggle.com/caicell/spotify-chart-trend-seasonal-arima/notebook>

Trend for top 1 。。。。