Data Analysis and Visualization in R

Case Study II

Shabnam Sadegharmaki, Christian Mertes, Julien Gagneur

Case Study II: Top Spotify Tracks of 2017

At the end of each year, Spotify compiles a playlist of the songs streamed most often over the course of that year.

Top Tracks of 2017 included 100 songs. The question is: What do these top songs have in common? Why do people like them?

Dataset I: Top Tracks of 2017

You are going to investigate the dataset, Top Spotify Tracks of 2017, which is available in Kaggle website (https://www.kaggle.com/nadintamer/top-tracks-of-2017/home). You can also download the featuresdf.csv file, uploaded in Moodle.

- The dataset includes:
  - Spotify URI for the song
  - Name of the song
  - Artist(s) of the song
  - Audio features for the song (such as danceability, tempo, key etc.)

A more detailed explanation of the audio features can be found in Kaggle (https://www.kaggle.com/nadintamer/top-tracks-of-2017)

Questions you could investigate

- Look for patterns in the audio features of the songs. Why do people stream these songs the most?
- Try to predict one audio feature based on the others
- See which features correlate the most

You need to support your claims with statistical assessment and prediction models

Dataset II: Spotify's Worldwide Daily Song Ranking

This dataset contains the daily ranking of the 200 most listened songs in 53 countries from 2017 and 2018 by Spotify users. It contains more than 2 million rows, which comprises 6629 artists, 18598 songs for a total count of one hundred five billion streams count.

You can download the data either from Kaggle (https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking) or Moodle. Each row contains a ranking position on a specific day for a song. Which includes:

- Position: Position on charts
- Track Name: Title of song
- Artist: Name of musician or group
- Streams: Number of streams
- URL
- Date
- Region: Country code

Questions you could investigate

- Can you predict what is the rank position or the number of streams a song will have in the future?
- How long does songs "resist" on the top 3, 5, 10, 20 ranking?
- What are the signs of a song that gets into the top rank to stay?
- Do continents share same top ranking artists or songs?
- Are people listening to the very same top ranking songs on countries far away from each other?
- How long time does a top ranking song takes to get into the ranking of neighbor countries?

Case study

- In class working on case study on the 4th February
- deadline for submission is the 5th February at noon (12:00).
- presentations will be given in the last week during exercise sessions.
- support your analysis with statistical testing and prediction models and evaluations.
- submit one solution per group only.
- sign up with your group here: Sign-Up-Sheet (https://docs.google.com/spreadsheets/d/1WgPOHS3ppuSDUzUOLHN4AwukoMqsiiXrU0yfdaCZa34/edit?usp=sharing)
- you can form new groups, you don't have to submit with the same group again.

Warm up exercise

Artists shining in top songs

```
spotify_data <- fread('./extdata/casestudy2/featuresdf.csv')
daily_spotify <- fread('./extdata/casestudy2/spotifys-ranking.csv')

top_artists <- spotify_data %>%
    group_by(artists)  %>%
    summarise(n_apperance = n()) %>%
    filter(n_apperance > 1) %>%
    arrange(desc(n_apperance))

top_artists$artists <- factor(top_artists$artists, levels = top_artists$artists[order(top_artists$n_apperance)]) # in or
```

Warm up exercise

```
ggplot(top_artists, aes(x = artists, y = n_apperance)) +
    geom_bar(stat = "identity",  fill = "tomato2", width = 0.6 ) +
    labs(title = "Top Artists of 2017", x = "Artists", y = "Number of Apperance on the Top 100") +
    theme(plot.title = element_text(size=15,hjust=-.3,face = "bold"), axis.title = element_text(size=12)) +
    geom_text(aes(label=n_apperance), hjust = 2, size = 3, color = 'white') +
    coord_flip()
```