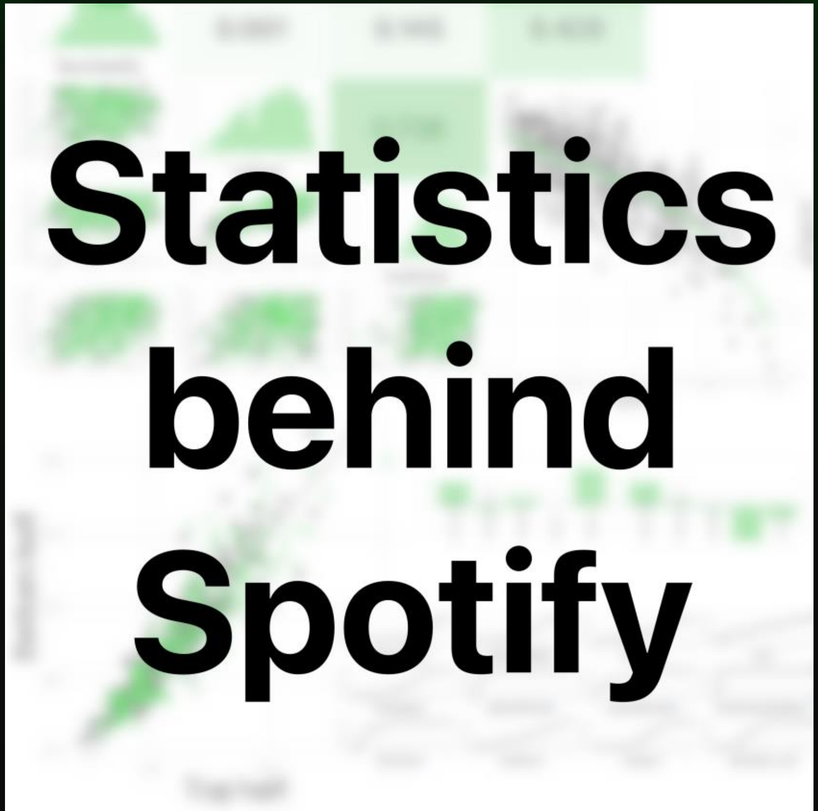
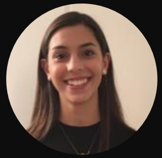
Spotify




Andrea Cioffi


Michele Di Sabato


Francesco Pascuzzi


Chiara Schembri

Statistics behind Spotify

A nonparametric approach to music

NPS Project • 2021-2022

Spotify

Home

Dataset

Goals

EDA

Models

Artist-based

Song-based

Conclusions

DATASET DESCRIPTION

TECHNICAL FEATURES

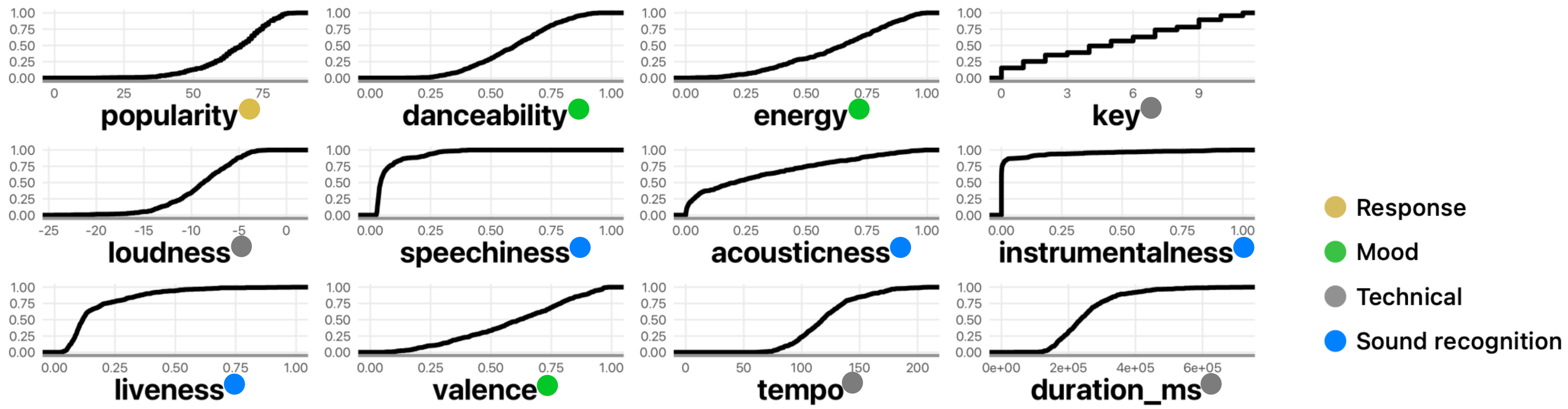
- Key **(C)**
- Mode **(C)**
- Tempo **(Beats Per Minute)**
- Duration **(Milliseconds)**
- Loudness **(Decibels)**

MOOD

- Danceability **(F)**
- Energy **(F)**
- Valence **(F)**

SOUND RECOGNITION

- Speechiness **(F)**
- Instrumentalness **(F)**
- Liveness **(F)**
- Acousticness **(F)**



(**C**) : categorical variable (**F**) : float variable, between 0 and 1 (**I**) : integer variable

Spotify

GOALS

Our goal is to help new and upcoming artists to broad their audience and get more visibility. This analysis could also help both the Spotify platform and record labels to find new talents.

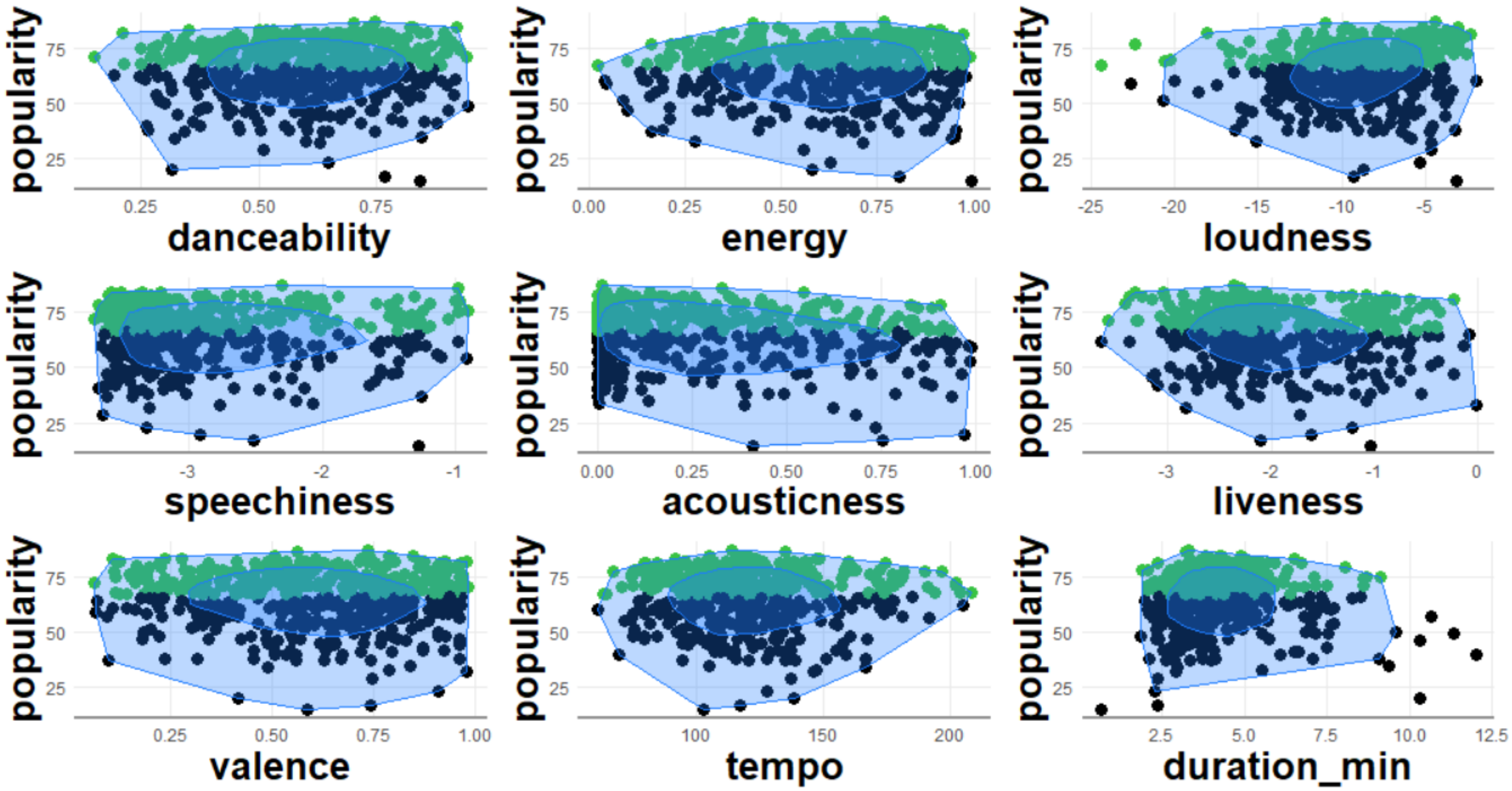
Which features of a song should an artist emphasize to get to the top?

Is it possible to support content decision makers with data-driven insights?
(as Netflix is already doing)

Spotify

- Home
- Dataset
- Goals
- **EDA**
- Models
  - Artist-based
  - Song-based
- Conclusions

EXPLORATORY DATA ANALYSIS

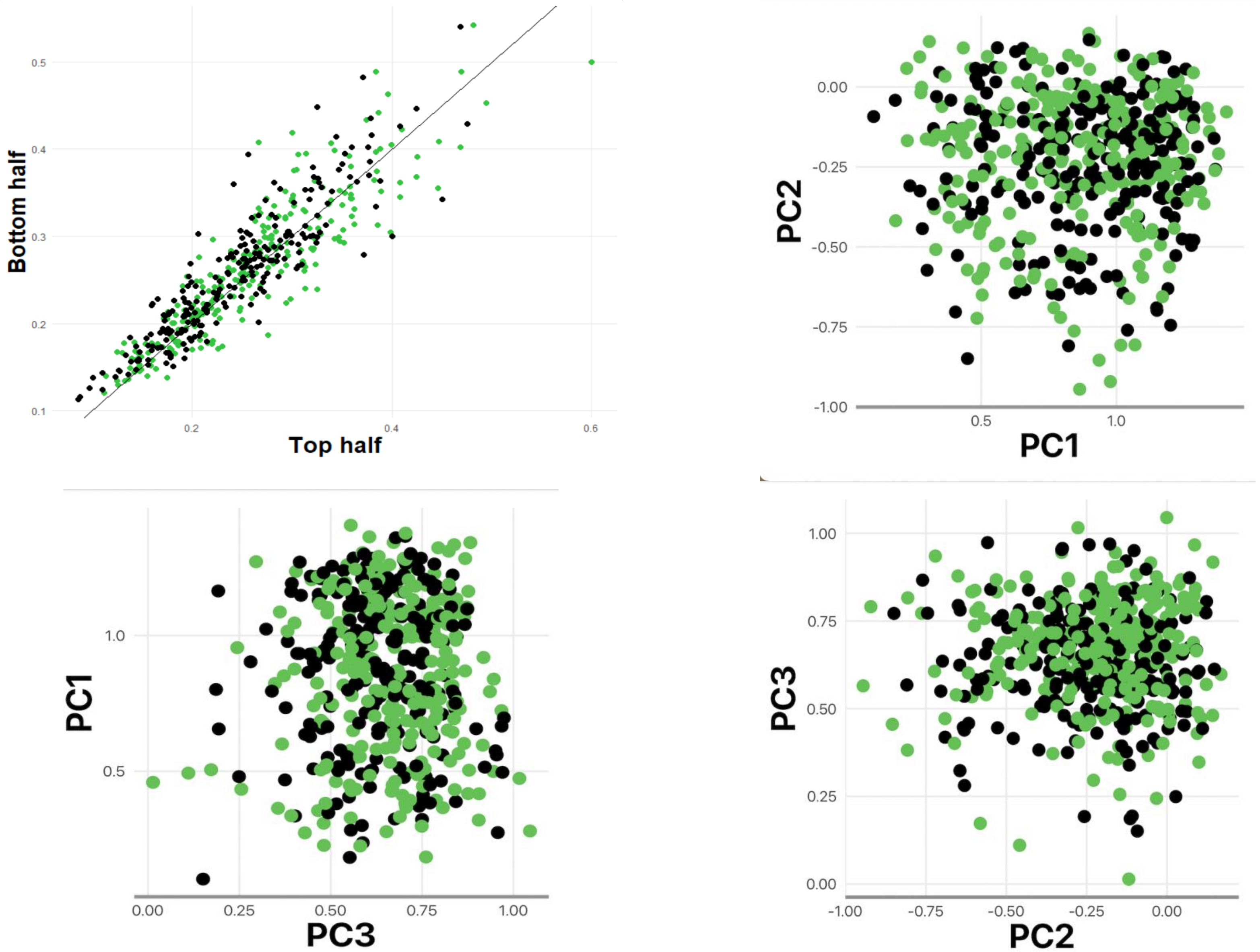


Popularity Bagplot

- most popular
- least popular

Spotify

- Home
- Dataset
- Goals
- **EDA**
- Models
  - Artist-based
  - Song-based
- Conclusions



DD-Plot of the distributions of continuous features & Top 3 Principal Components

● most popular
● least popular

TWO APPROACHES

ARTIST – BASED

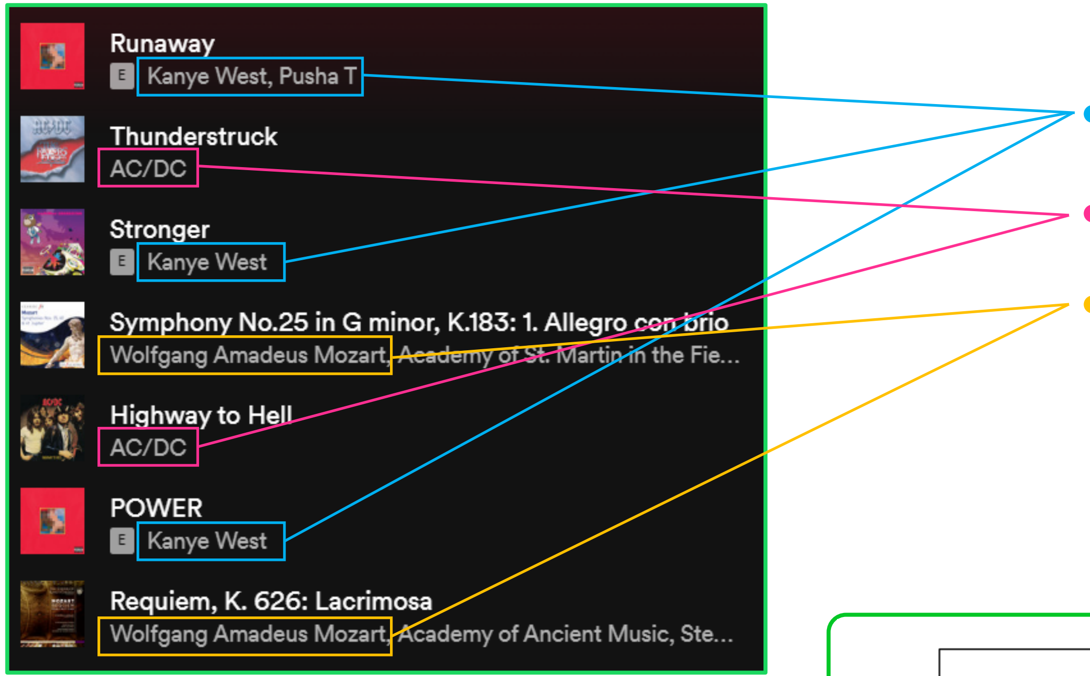
CLUSTERING
GAM + MIXED EFFECTS

SONG – BASED

DIFFERENCE IN DIFFERENCES (DiD)
GAM

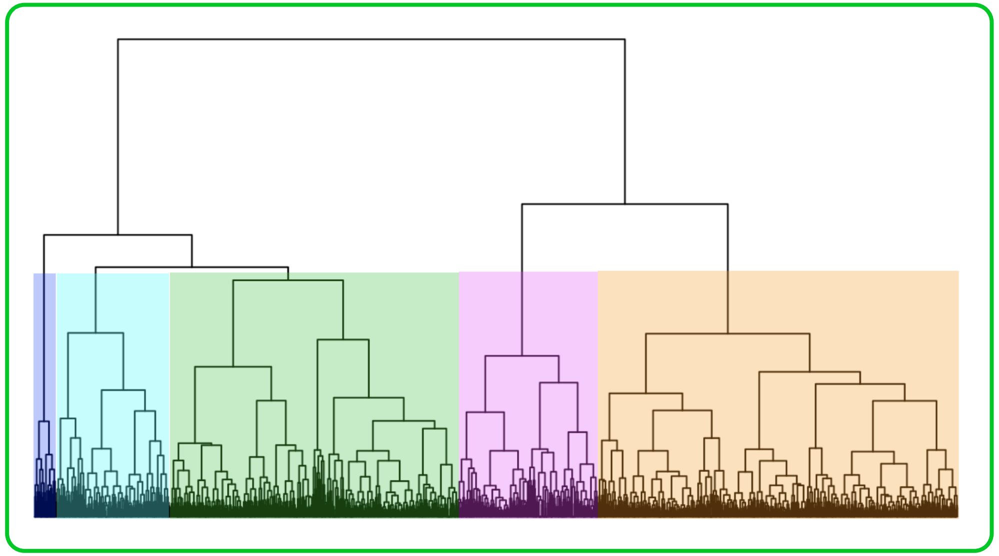
Spotify
Home
Dataset
Goals
EDA
Models
Artist-based
Song-based
Conclusions

ARTIST-BASED MODEL



- Average features for Kanye West
- Average features for AC/DC
- Average features for Mozart

CLUSTERING ARTISTS BASED ON THEIR AVERAGE FEATURES



Dendrogram of Hierarchical clustering with Ward's linkage and euclidean distance

Spotify

Home

Dataset

Goals

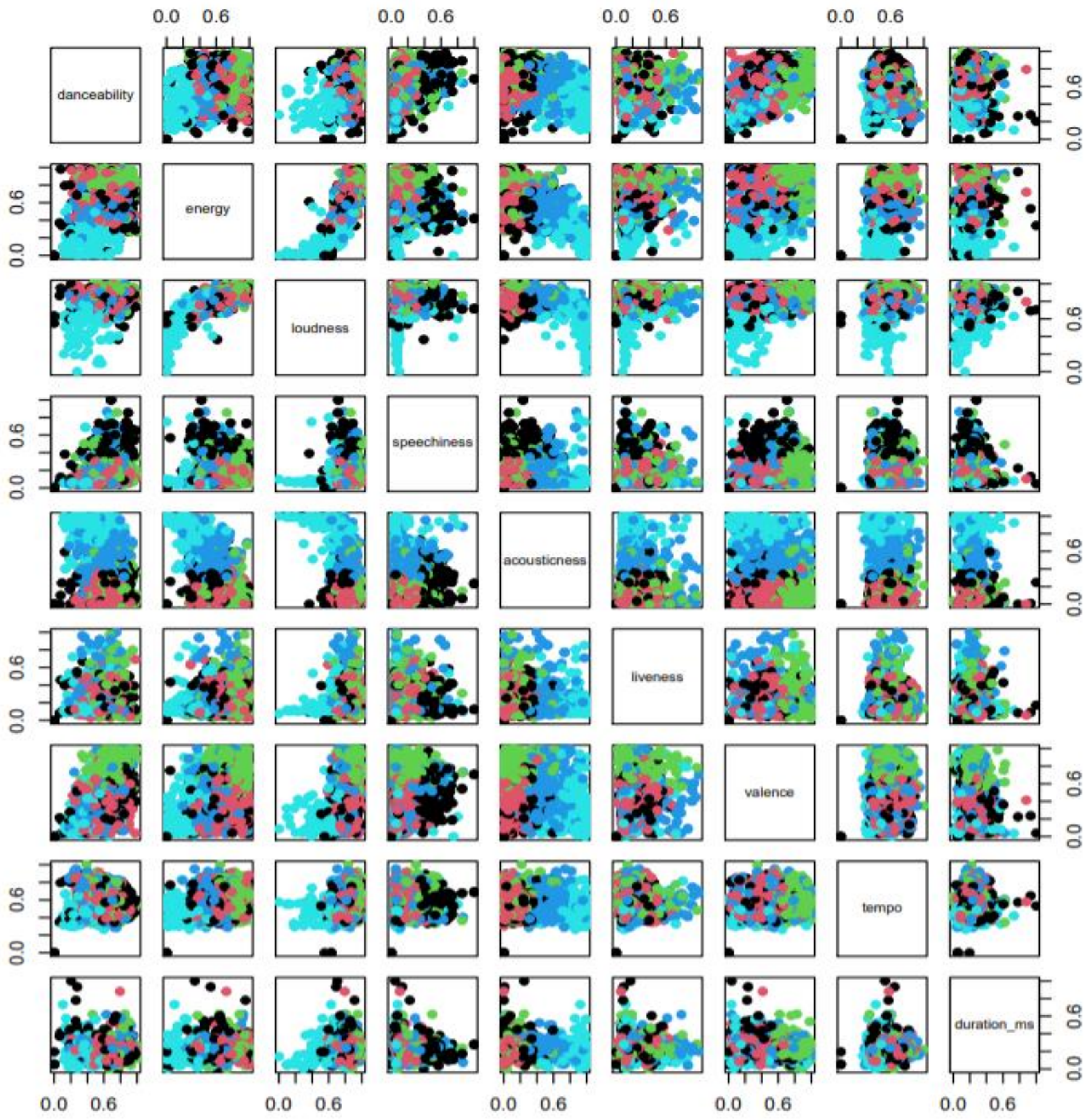
EDA

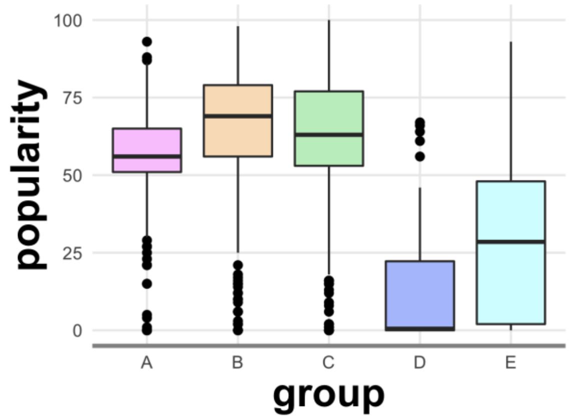
Models

Artist-based

Song-based

Conclusions





Pairplot and Popularity Boxplot by groups

ARTIST-BASED MODEL

$$\boxed{y_i = f(x_{1i}) + f(x_{2i}) + f(x_{3i}) + x_{4i} + x_{5i} + \varepsilon_i}$$

$y_i := popularity$

$x_{1i} := excess\ popularity$

$x_{2i} := general\ popularity$

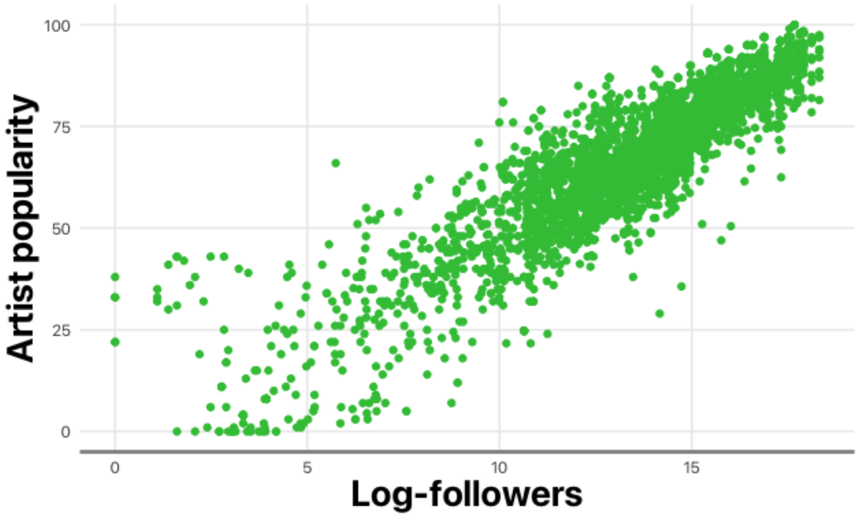
$x_{3i} := duration\ (min)$

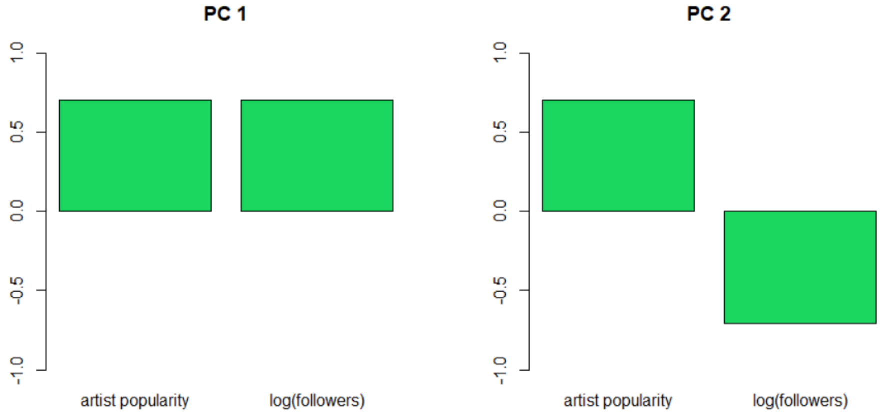
$x_{4i} := groups\ (derived\ by\ our\ cluster)$

$x_{5i} := featuring\ (1\ if\ multiple\ artists)$

Derived from a PCA on **artist popularity** and **followers (log)**, both scaled:

- ***excess popularity*** : difference between popularity and followers
- ***general popularity*** : sum of the two

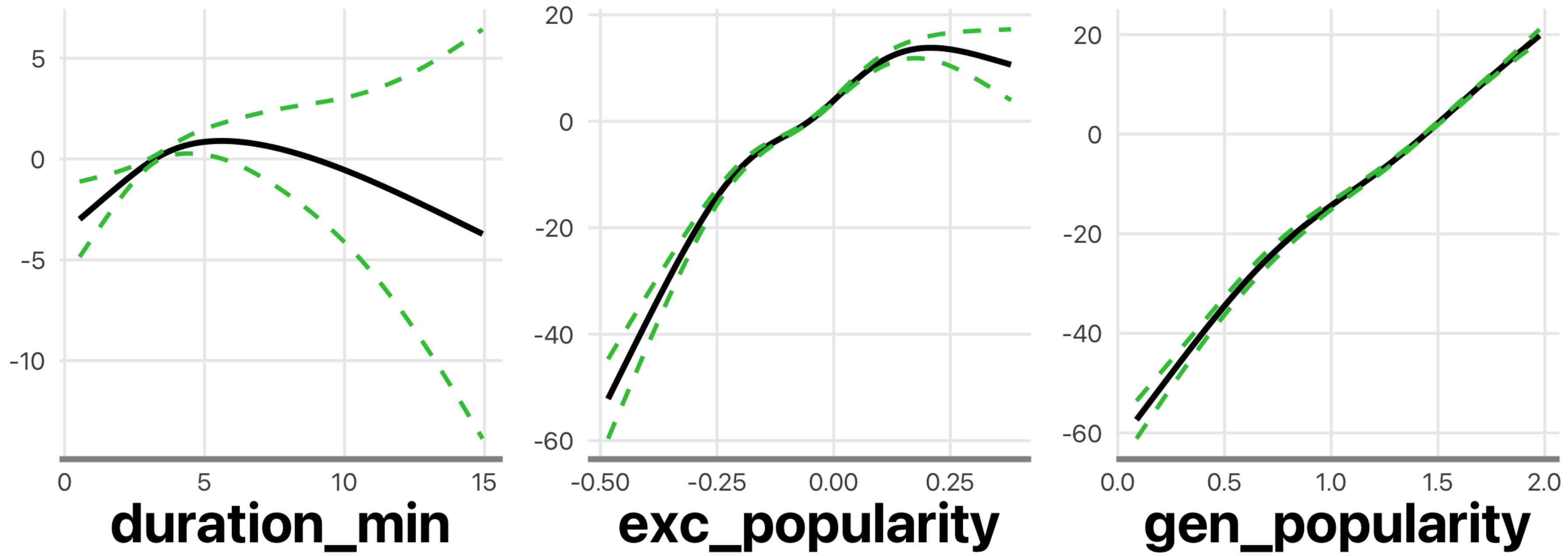




Spotify

- Home
- Dataset
- Goals
- EDA
- **Models**
  - **Artist-based**
  - Song-based
- Conclusions

ARTIST-BASED GAM



GOODNESS OF FIT

- R^2 = 71.5%
- MAE on test set = 7.3

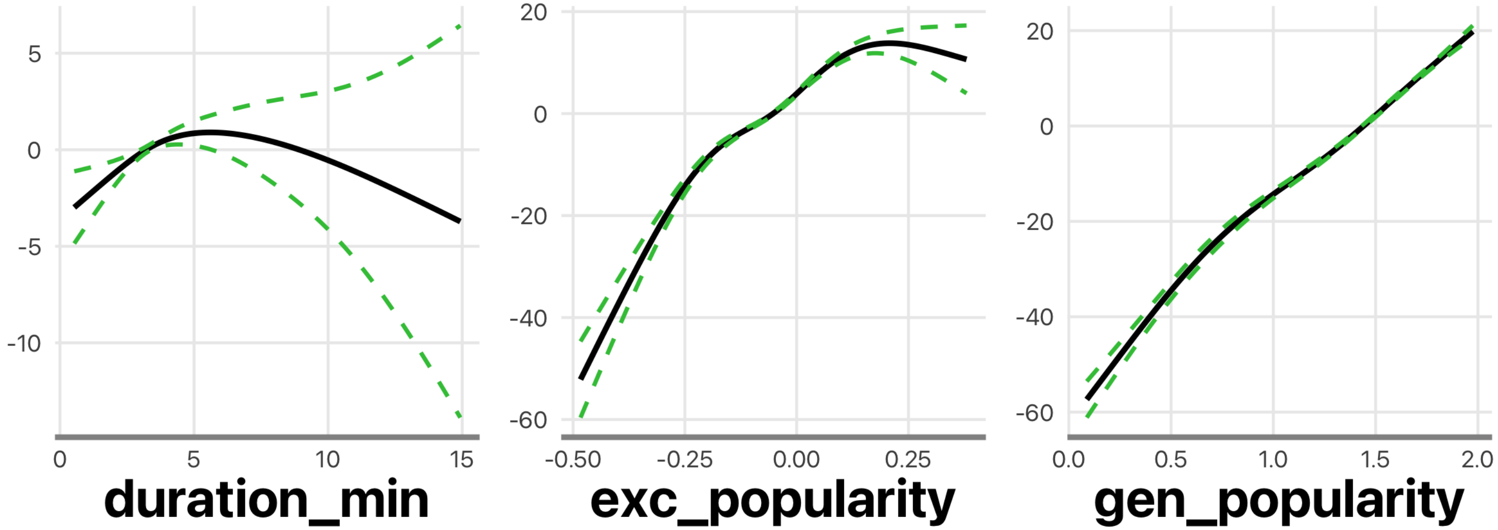
FEATURE	PERM P-VAL
Duration	0.017
General Pop.	<2e-16
Excess Pop.	<2e-16
Groups	<2e-16
Featuring	0.003

Spotify
Home
Dataset
Goals
EDA
Models
Artist-based
Song-based
Conclusions

MIXED RANDOM EFFECTS

$$y_{ij} = f(x_{1ij}) + f(x_{2ij}) + f(x_{3ij}) + f(x_{4ij}) + \alpha_j + \varepsilon_{ij} \qquad \forall i = 1, \dots, n_j$$

$$\left\{\begin{array}{ll}
x_{1ij} & \coloneqq \textit{duration of song } \boldsymbol{i} \textit{ in group } \boldsymbol{j} \textit{ (in minutes)} \\
x_{2ij} & \coloneqq \textit{excess popularity of song } \boldsymbol{i} \textit{ in group } \boldsymbol{j} \\
x_{3ij} & \coloneqq \textit{general popularity of song } \boldsymbol{i} \textit{ ingroup } \boldsymbol{j} \\
\alpha_j & \coloneqq \textit{group specific random intercept} \sim \mathcal{N}(0, \sigma^2_{groups}) \\
\varepsilon_{ij} & \coloneqq \textit{gaussian error} \sim \mathcal{N}(0, \sigma^2)
\end{array}\right.$$



Spotify

Home

Dataset

Goals

EDA

Models

- Artist-based
- **Song-based**

Conclusions

SONG-BASED MODEL

$$y_i = f(x_{1i}) + f(x_{2i}) + f(x_{3i}) + f(x_{4i}) + \varepsilon_i$$

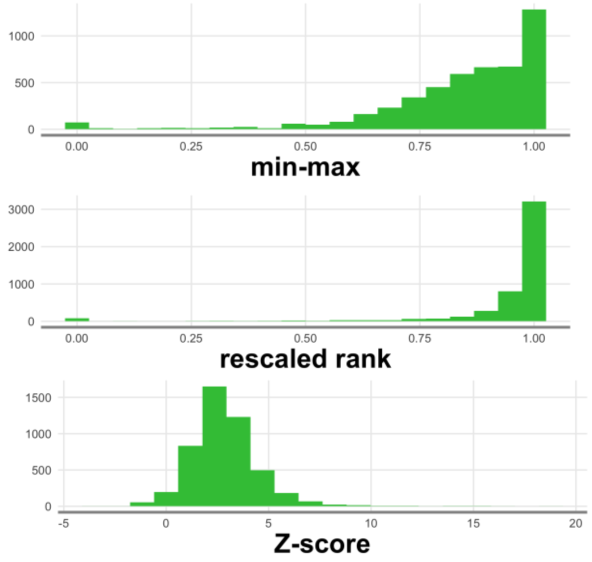
$x_{1i} \coloneqq energy$

$x_{2i} \coloneqq duration\ (min)$

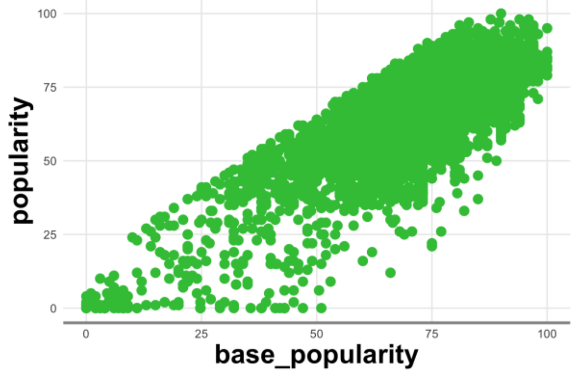
$x_{3i} \coloneqq danceability$

$x_{4i} \coloneqq valence$

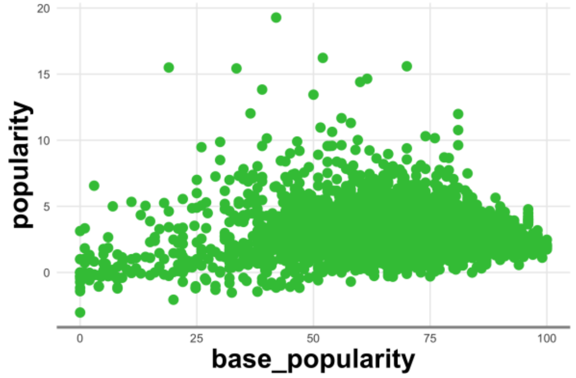
$$y_i \coloneqq difference\ in\ popularity$$



Original popularity



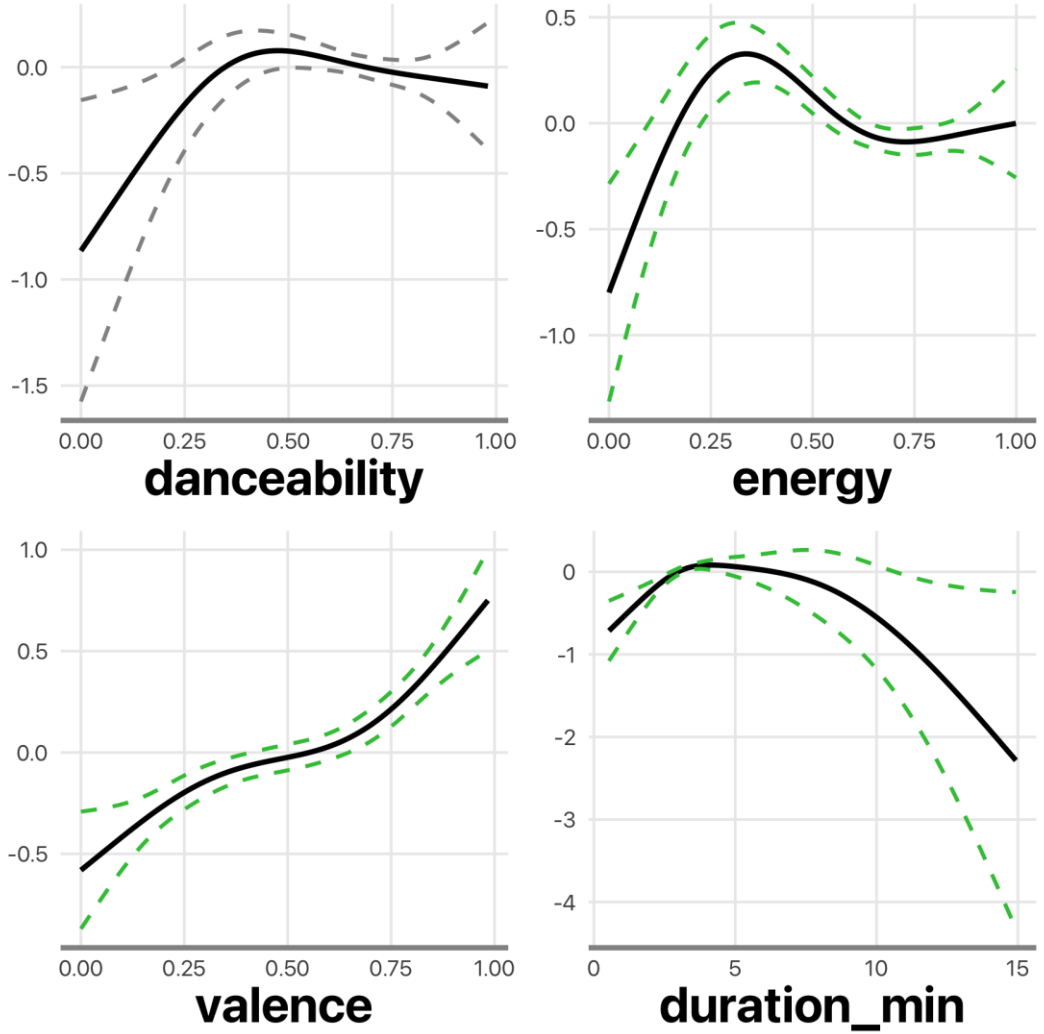
Normalized popularity



Transformations are considered on the whole discography of each artist

- Home
- Dataset
- Goals
- EDA
- **Models**
  - Artist-based
  - **Song-based**
- Conclusions

SONG-BASED GAM



Generalized Additive Model with **standardized** popularity

GOODNESS OF FIT

- $R^2 = 3.6\%$
- MAE on test set = 1.17

CONCLUSIONS

OUR ANALYSIS POINTED OUT :

- POPULARITY OF A SONG IS STRONGLY RELATED TO THE ARTIST
- THERE ARE SOME SIGNIFICANT FEATURES, BUT THEY ARE NOT SUFFICIENT TO PREDICT THE POPULARITY
- A NETFLIX APPROACH FOR PREDICTING THE POPULARITY OF A SONG MIGHT BE UNFEASIBLE

Spotify

Home

Dataset

Goals

EDA

Models

Artist-based

Song-based

Conclusions

REFERENCES

- Spotify API
  https://developer.spotify.com/documentation/web-api/
- Supporting content decision makers with machine learning
  Netflix Technology Blog, link
- A Nonstochastic Interpretation of Reported Significance Levels (1983)
  Freedman D., Lane D., Journal of Business & Economic Statistics, 1:4, 292-298
- Practical variable selection for generalized additive models (2011)
  Marra G., Wood S., Computational Statistics & Data Analysis, 55(7), 2372-2387
- Mgcv package documentation
  https://cran.r-project.org/web/packages/mgcv/mgcv.pdf#page=201
- Introduction to linear mixed effect models, UCLA
  https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/