

main

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.5      v purrr  0.3.4
## v tibble  3.1.6      v dplyr  1.0.7
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.1.1      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

Narative:

What musical characteristics of an electronic song make it popular? This is the question I aim to answer in my project.

By characteristics of a song, I mean the basic features in the context of music theory such as tempo, key, time signature, etc. But I also mean more complex features in the context of how humans interpret music such as, loudness, energy, danceability, etc.

By popularity of a song, I mean how many times that song was listened to compared to other songs.

For music producers, an accurate answer to this question would be very valuable as it would give them evidence for selecting target characteristics to maximize a songs popularity. It would also give them target characteristics for a certain type of song. For example, if a music producer wanted to create a high energy and very danceable song, they would have evidence as to which key, time signature, tempo, etc. they would need to maximize the songs popularity.

My response variable would be the popularity of a song based on its musical characteristics. However, my dataset only contains time-series measures of popularity. That is, my dataset only contains information of how popular a song is currently but I am more interested in how popular a song is independent of when it was uploaded. The songs popularity in the dataset is also dependent on the popularity of the artists from the assumption that more popular artists have bigger platforms to promote their song. I would like my response variable to also be independent of the popularity of the artists. Therefor I need to construct a custom response variable.

I am not certain what my response variable will be specifically. But here are some ideas I have. My response variable will be independent from the upload date of a track and independent from the number of followers the songs artists have. I am looking for ways to construct a response variable that matches these conditions.

My hypothesis is that a songs popularity is dependedent on its musical characteristics.

Data:

This dataset is retrieveed from Spotify's API. I will attach the python script I used to create the dataset. Please note, the python script is very messy right now. It contains some functions that I ended up not using and other useless blocks of code I used for testing. I plan to clean it up and comment it later.

Script: <https://colab.research.google.com/drive/1VipujOuut-qT1iVkMw8lASIqxjEGDv?usp=sharing>

The dataset was created by selecting playlists within the Electronic Music Genre, then calling the API to retrieve track information for every track within every playlist. Then for every track I called the API again to get information that was not included in the first call. This is information like artist information and upload date. I then parsed everything into a dictionary and wrote it to the CSV file.

I still need to do a bit more cleaning of the dataset. I need to remove some duplicates. I also need to remove old tracks that have been republished in new albums. These old tracks have a popularity near 0 because only their republished version is being listened to by users. The old track and the republished version are otherwise identical but the republished version has the accurate popularity.

I have been working on cleaning and creating the dataset for some time and while I wanted to have a complete polished dataset for this proposal, I am running out of time so I thought I would just submit what I have.

For the dataset, I wanted a tracks audio features (the characteristics described above) and metrics to help me build my response variable (current track popularity, time uploaded, track artists follower counts)

```
data <- read.csv('data/spotify_tracks.csv')
```

```
glimpse(data)
```

```
## Rows: 2,121
## Columns: 25
## $ track.name      <chr> "Move Your Body", "Say Nothing (feat. MAY-A)", "You've~
## $ artists.names   <chr> ["'Ã-wnboss', 'Sevek'"], ["'Flume', 'MAY-A'"], ["'Dis~
## $ artists.ids      <chr> ["'37czgDRfGMvgRiUKHvnnhj', '0a0IluXr131XqrXFwFCFGT'"]~
## $ track.popularity <int> 79, 69, 72, 85, 79, 51, 83, 52, 76, 75, 86, 63, 69, 6~
## $ artist.ids       <chr> ["'37czgDRfGMvgRiUKHvnnhj', '0a0IluXr131XqrXFwFCFGT'"]~
## $ track.id         <chr> "6GomT970rC0kKAyyrwJeZi", "424Uwmm1kNW07Ty1nOhSpl", "~
## $ release.date     <chr> "2021-10-29", "2022-02-02", "2022-01-28", "2021-11-19~
## $ danceability     <dbl> 0.848, 0.478, 0.658, 0.308, 0.788, 0.774, 0.601, 0.73~
## $ energy           <dbl> 0.821, 0.822, 0.908, 0.861, 0.945, 0.784, 0.787, 0.90~
## $ key              <int> 2, 2, 4, 11, 9, 10, 0, 10, 9, 11, 0, 10, 3, 11, 10, 7~
## $ loudness         <dbl> -5.408, -1.961, -8.071, -4.112, -5.091, -4.549, -6.17~
## $ mode             <int> 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,~
## $ speechiness      <dbl> 0.0527, 0.0642, 0.1860, 0.1720, 0.0599, 0.0351, 0.031~
## $ acousticness     <dbl> 0.016900, 0.101000, 0.033500, 0.012000, 0.223000, 0.1~
## $ instrumentalness <dbl> 4.03e-04, 6.60e-05, 1.98e-02, 1.03e-03, 2.97e-06, 1.0~
## $ liveness         <dbl> 0.0962, 0.1200, 0.2210, 0.2770, 0.1150, 0.2930, 0.142~
## $ valence          <dbl> 0.2490, 0.3090, 0.4500, 0.0386, 0.4660, 0.5330, 0.525~
## $ tempo            <dbl> 125.051, 130.058, 123.920, 171.966, 128.036, 127.042,~
## $ type             <chr> "audio_features", "audio_features", "audio_features",~
## $ id               <chr> "6GomT970rC0kKAyyrwJeZi", "424Uwmm1kNW07Ty1nOhSpl", "~
## $ duration_ms      <int> 157445, 232959, 253787, 158774, 168053, 177689, 17516~
## $ time_signature   <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4,~
## $ follow           <int> 33201, 2257049, 7669295, 125502, 3198289, 1251567, 18~
## $ pop              <int> 130, 135, 158, 146, 150, 153, 85, 140, 74, 297, 168, ~
## $ num.artist       <int> 2, 2, 2, 2, 2, 2, 1, 2, 1, 4, 2, 2, 4, 1, 1, 2, 1, 3,~
```