Decoding Playlist Success

Sarang Nirwan
Cristina Racoviță
Smriti Dangi
Bogdan Bîndilă

1

Broadside
Come & Go

Good Charlotte
The River

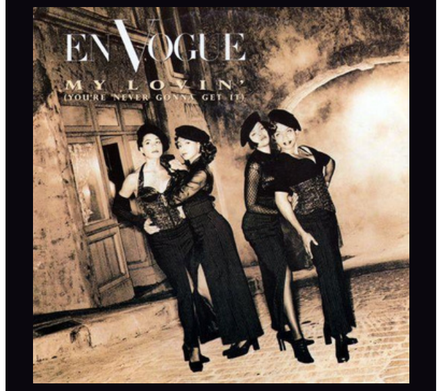
blink-182
Bored To Death

Good Charlotte
The Anthem

Boys Like Girls
Love Drunk

2

Bruno Mars
That's What I Like



En Vogue
My Lovin'

I.O.I
Whatta Man

Michel Jackson
The Way You
Make Me Feel

Sugarhill Gang
Apache

Research Question

Can we predict the **number of followers** of a playlist only using **audio features** and **genres** of the contained songs?

Data Sources

Playlist

- **collaborative**: string
- **duration_ms**: long
- **num_albums**: long
- **num_edits**: long
- **num_followers**: long
- **num_tracks**: long
- **pid**: long
- **tracks**: array
  - **album_name:** string
  - **artist_uri:** string
  - **track_name:** string
  - **track_uri:** string

Spotify Million Playlist Dataset

Spotify API

Artist

- **id**: string
- **name**: string
- **uri**: string
- **genre**: array
  - **element:** string

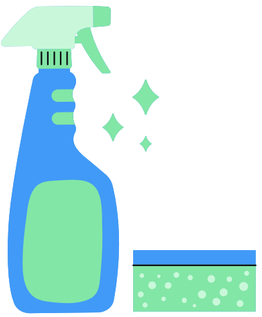
Audio Feature

- **acousticness**: double
- **danceability**: double
- **duration_ms**: double
- **energy**: double
- **valence**: double
- **instrumentalness**: double
- **key**: double
- **liveness**: double
- **loudness**: double
- **mode**: double
- **speechiness**: double
- **tempo**: double
- **time_signature**: double
- **uri**: string

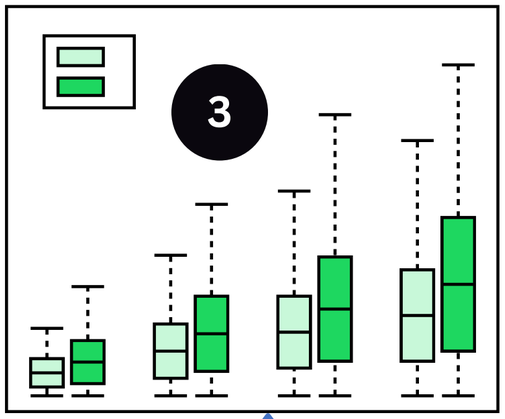
Overview

Spotify API + Spotify Million Playlist Dataset

1 Data Collection

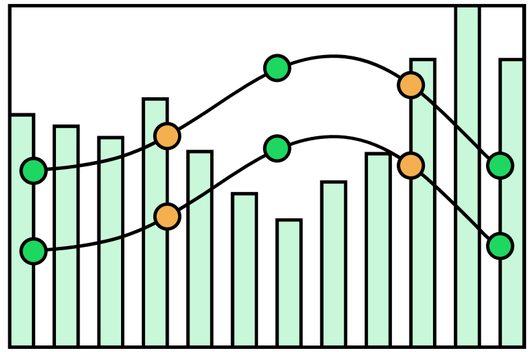


2 Data Cleaning

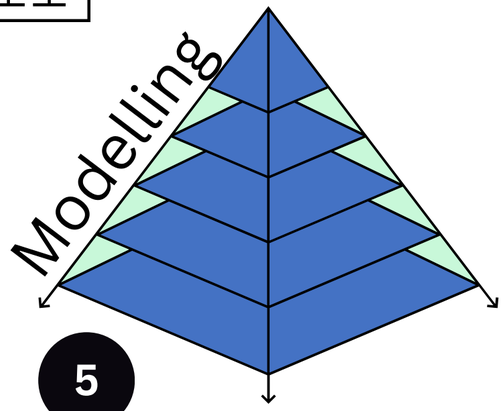


3 Exploratory Data Analysis

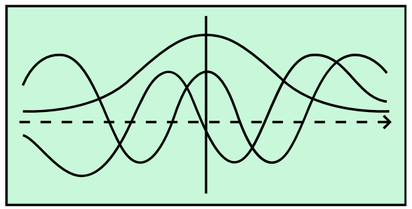
4 Feature Creation



Data Preprocessing



5



Prediction & Analysis

6

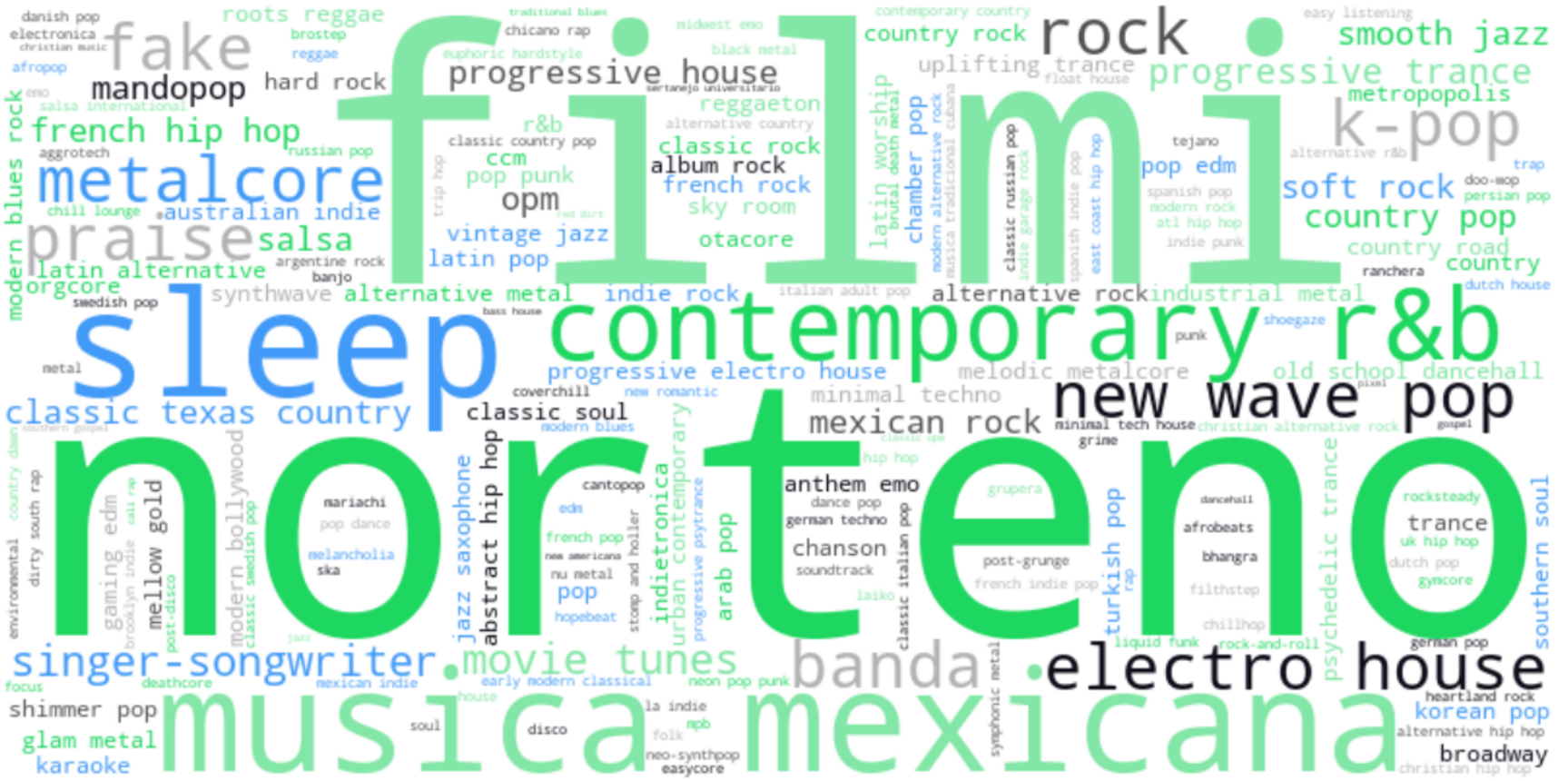
Data Questions

Artist Genre:
What are the **most popular genres**?
How many genres are the artists generally associated with?

Playlists:
What is a overall **composition** of the playlists?
What is the **average duration** of a playlist?
How often are the playlist **modified**?

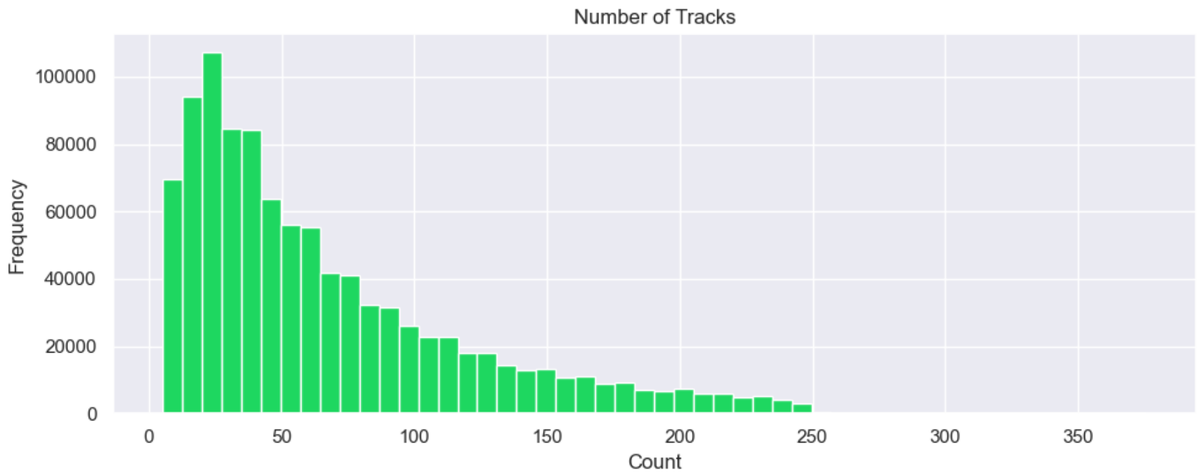
Audio:
What are the **most popular audio features**?

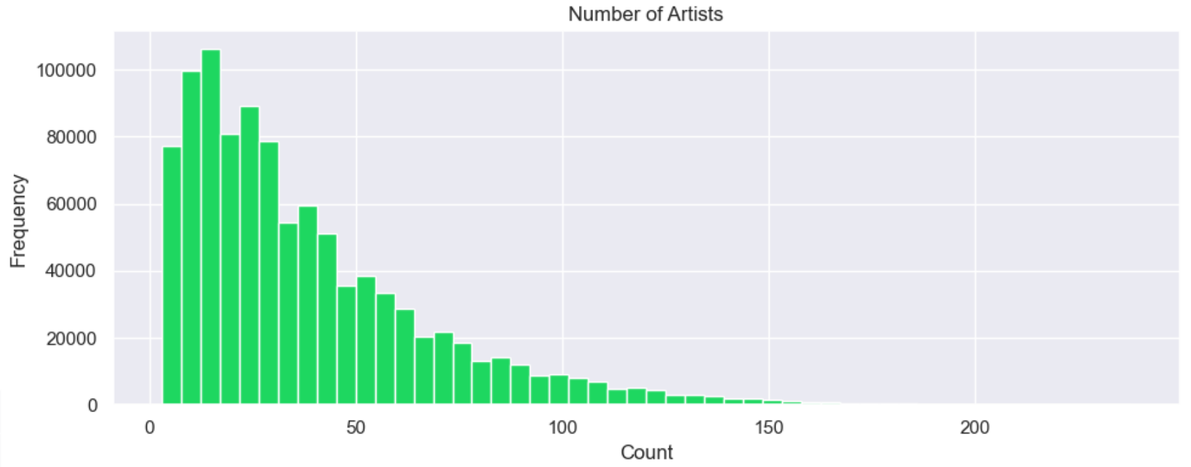
Artists/Genre Analysis

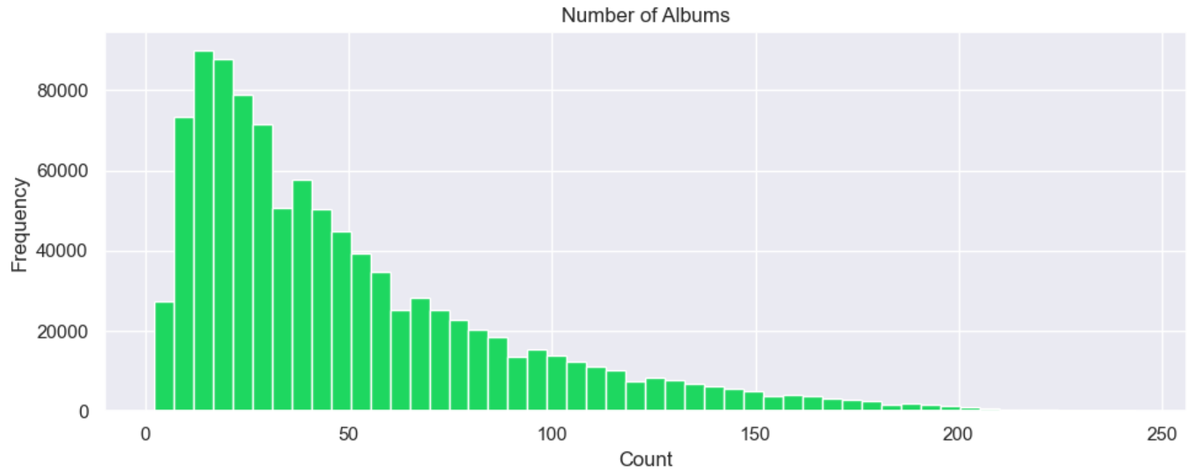


^Word cloud of most popular genres overall

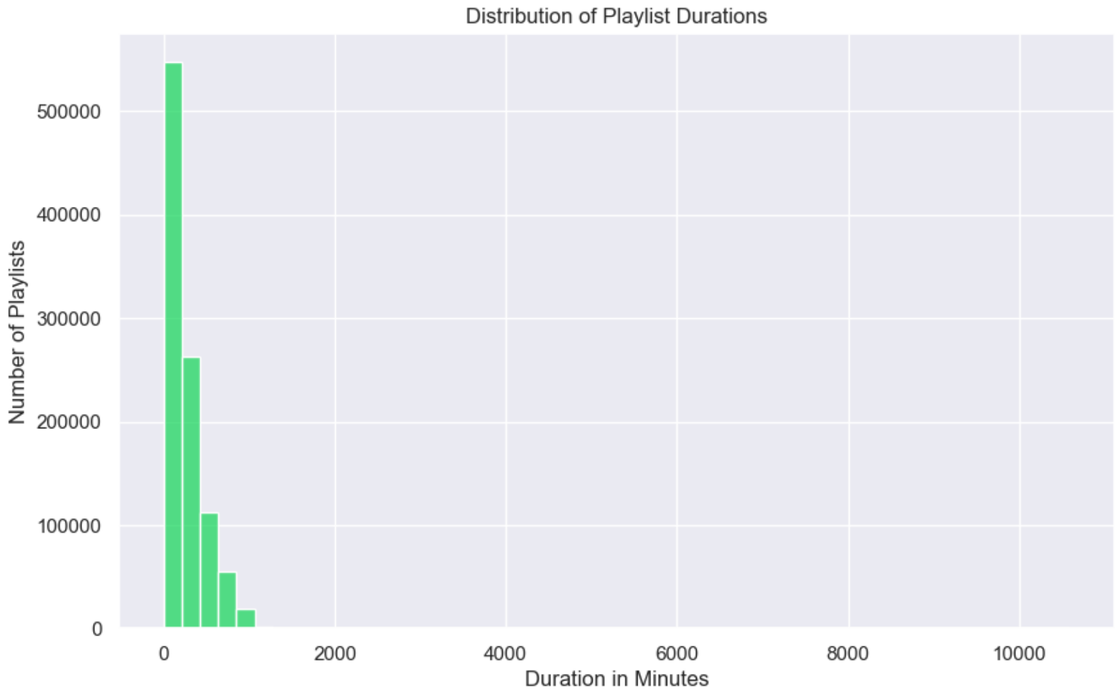
Playlist Composition Analysis





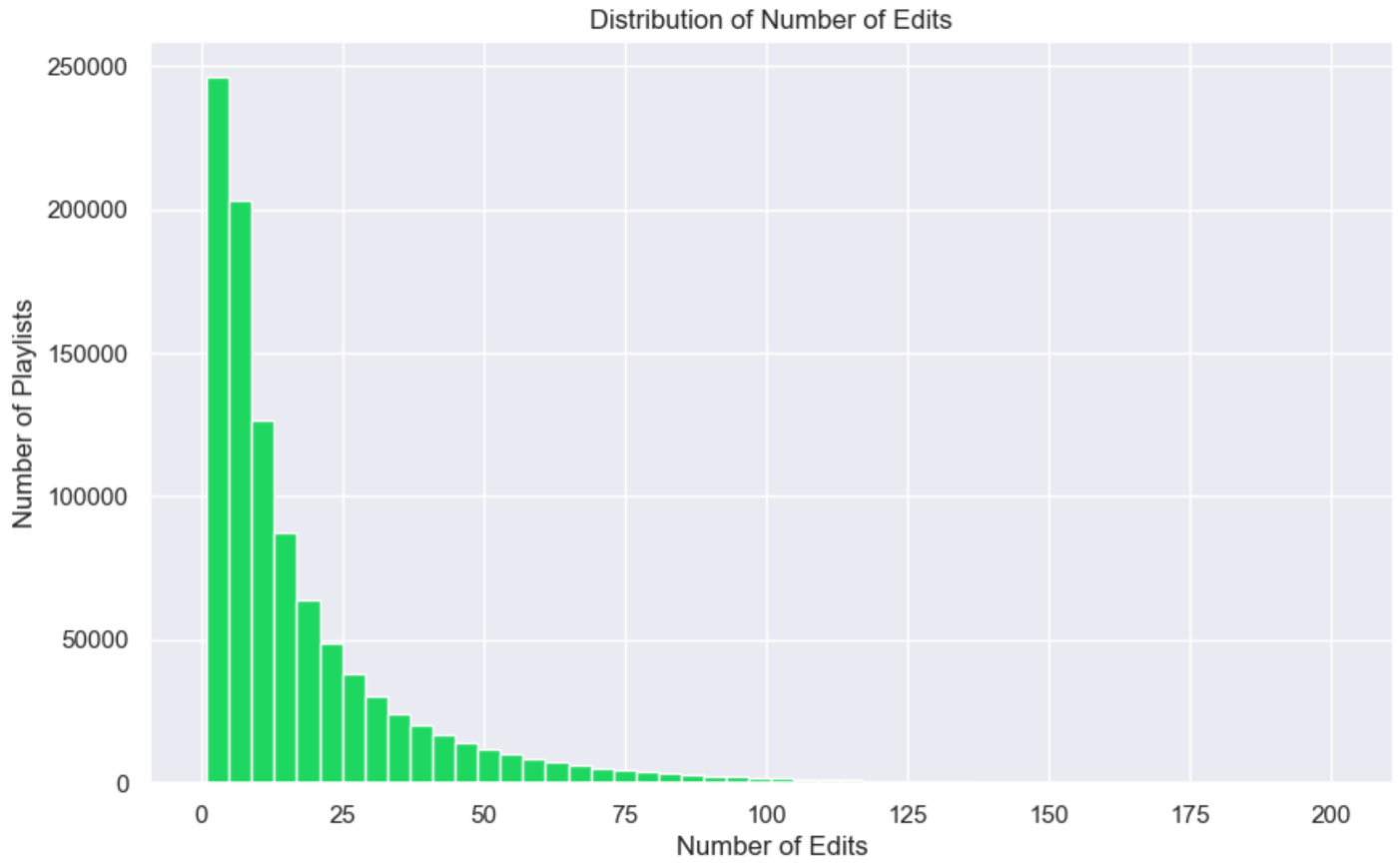


Playlist Duration Analysis

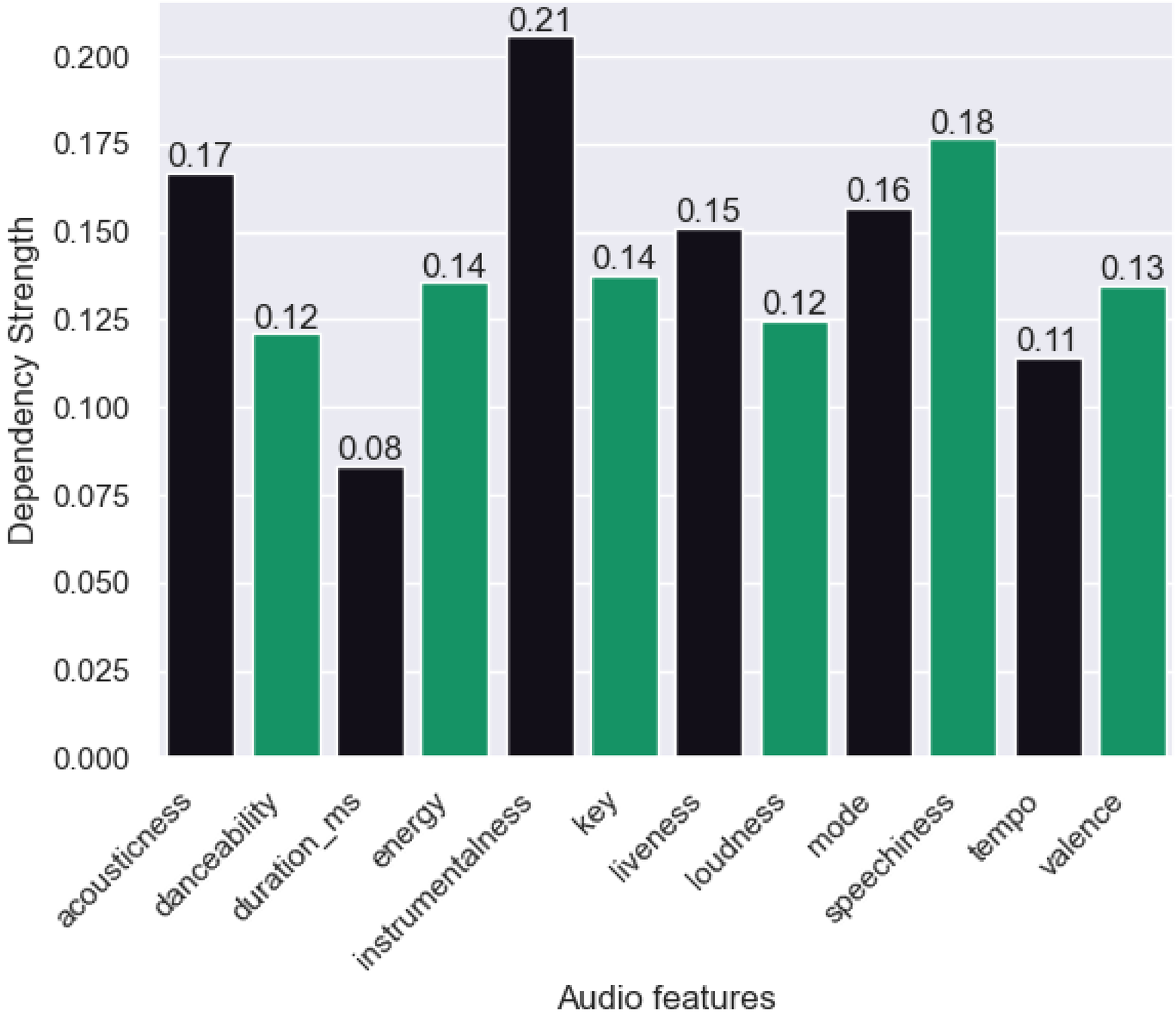


statistic	duration (hours, minutes, seconds)
mean	(4, 20, 0)
std	(3, 34, 0)
min	(0, 2, 0)
25%	(1, 39, 0)
50%	(3, 10, 0)
75%	(5, 57, 0)
max	(176, 25, 0)

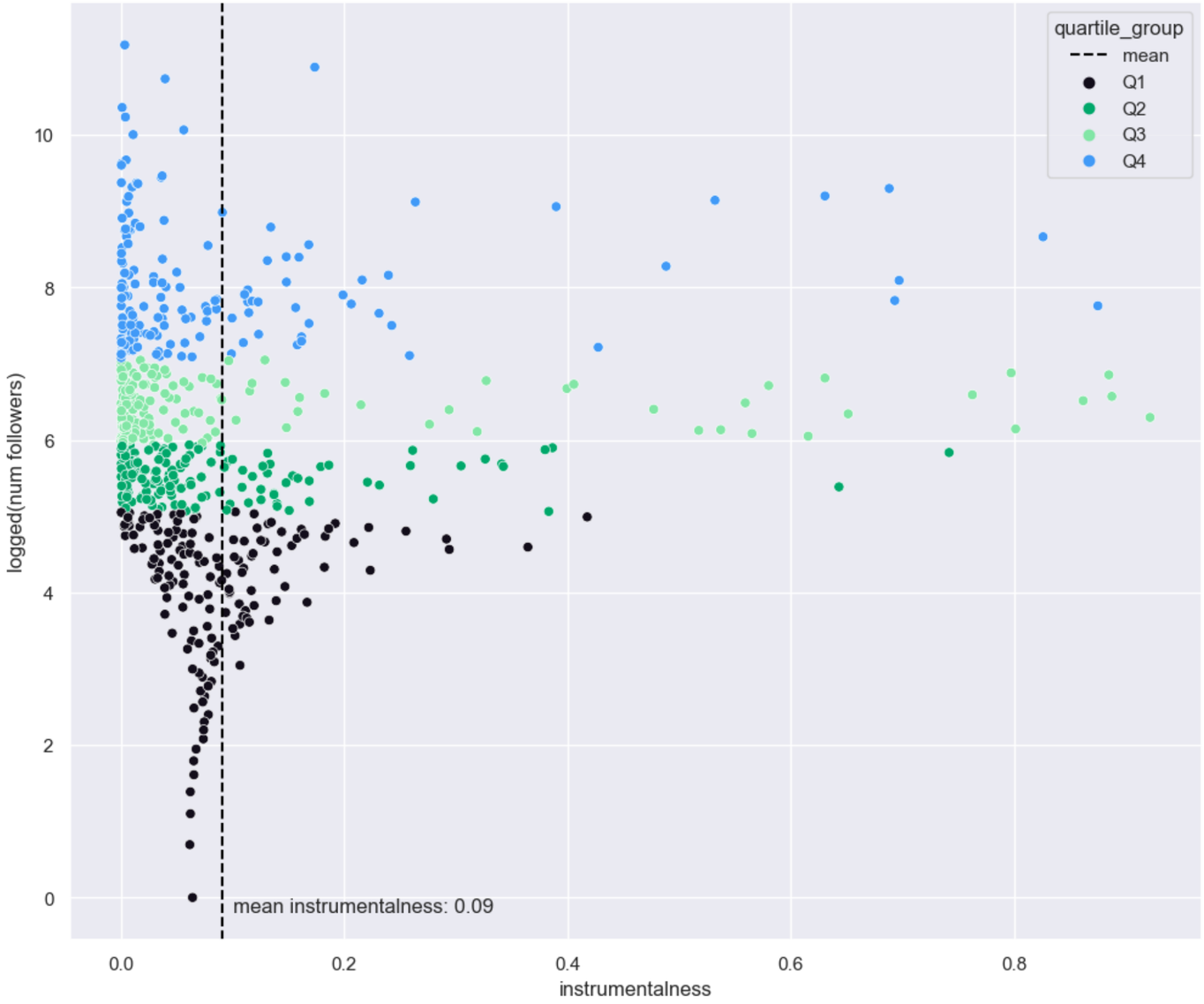
Playlist Revision Analysis



Mutual information regression scores



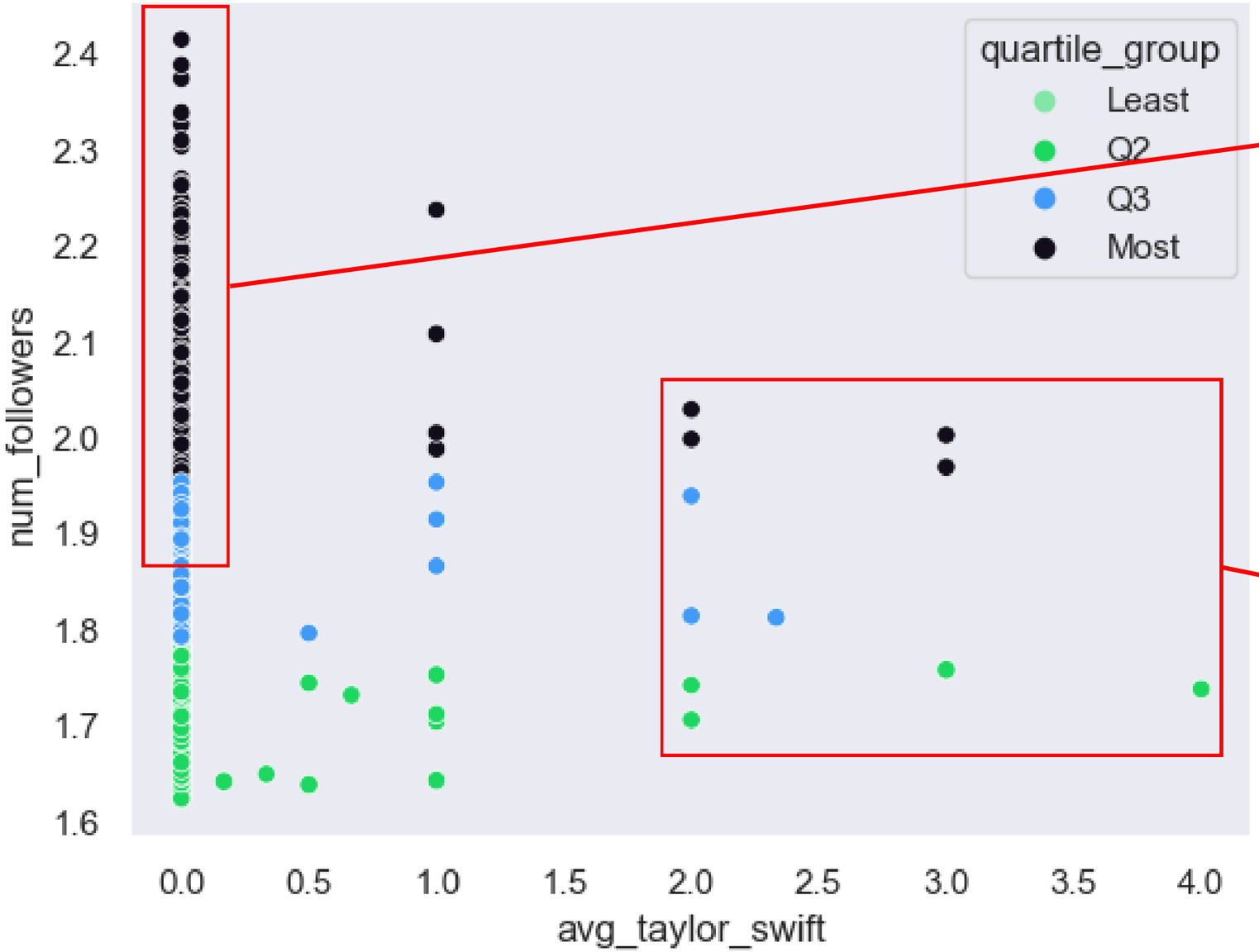
Instrumentalness feature



Hypothesis: Clustering of features if playlists are grouped based on num followers

Result: Little clustering

“Taylor Swift” feature



Lower avg. Taylor Swift count
mostly higher followers count

Higher avg. Taylor Swift count
mostly lower followers count

Data preprocessing

Distribution of the number of followers is right-skewed

followers #	1	2	3	4-100	101 - 1000	1001 - 71643
observations %	75.42	14.96	4.69	4.84	0.064	0.0176

aggregated data

fix the distribution

Remove 0.001% of large outliers
Undersample playlists with one follower
Apply log transformationto remove skewness

prepare data

Split data into train-val (80%) and test (20%)
Standardize features

Modeling

baseline

always predict the average follower count of train-val playlists

hyperparameter optimization

Random Forest Regressor + **3-fold cross validation**

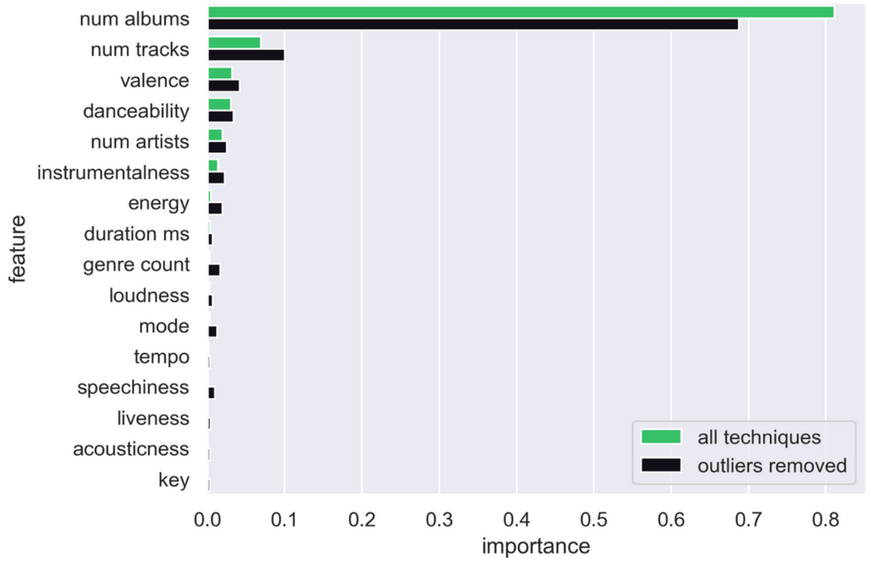
search space - 90 combinations

number of trees
percentage of playlists used to train a tree
minimum playlists number per tree node
number of features sampled to train a tree

get best model instance

Results

Feature importance given by model



Audio features don’t impact the popularity

Metrics in relation to data preprocessing

Conclusion

Audio Features do not matter too much when it comes to the **playlist popularity**

Popular Artist = Popular Playlist

High-level aggregated features have no **predictive power**