March 30, 2022

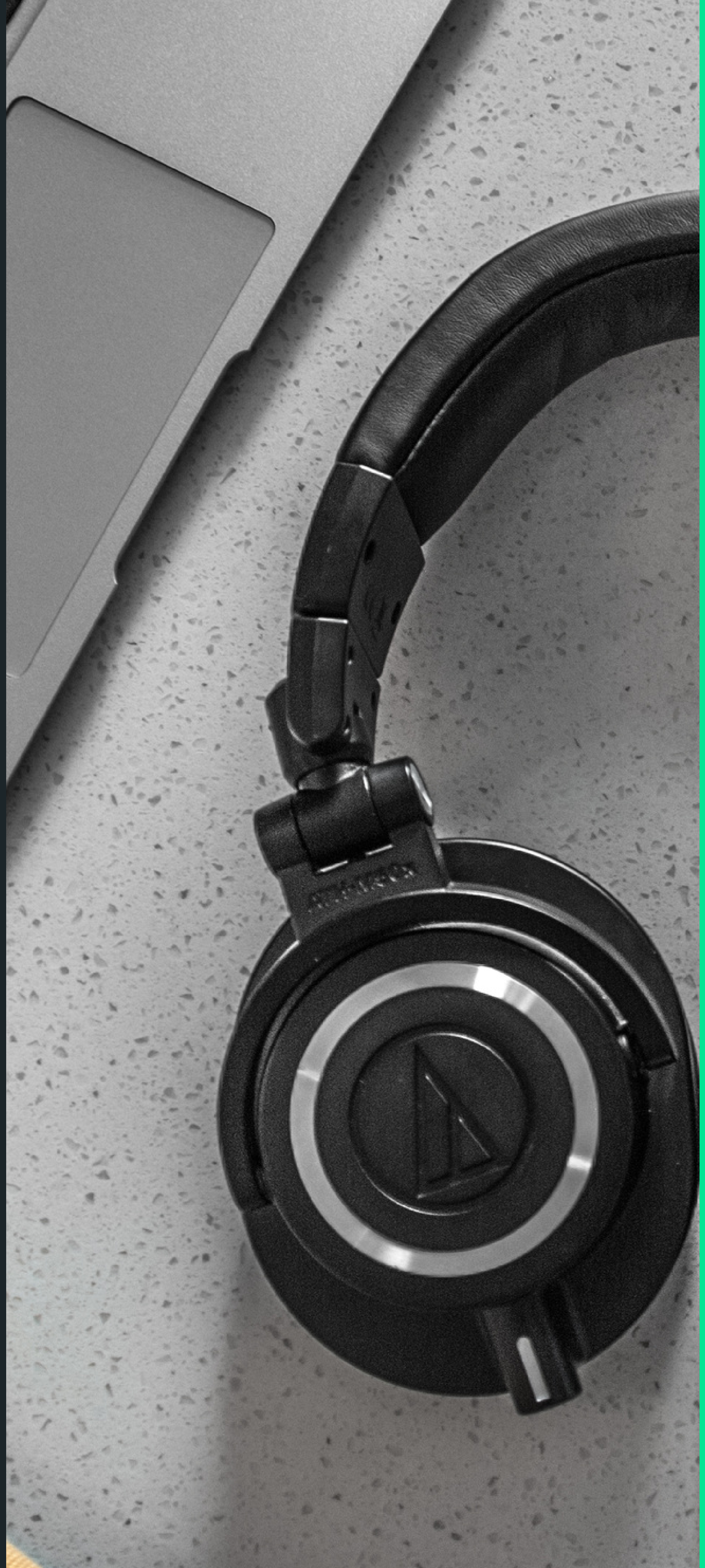
SPOTIFY: SONG ANALYSIS

SPOTIFY: SONG ANALYSIS
SPOTIFY: SONG ANALYSIS
SPOTIFY: SONG ANALYSIS

BY: SYDNEY TRAN

Project Statement

Project Statement
Project Statement
Project Statement



Utilizing Spotify's Top 200 Weekly (Global) charts from 2020 and 2021, can we identify what makes a song popular based on specific features of a song?

Model performance will be determined by the RMSE and R2 score. The success of the model will be measured by an increase of at least 10% from the baseline score.

Project Roadmap

DATA CLEANING & EDA

- Genres
- Outliers

PREPROCESSING

- Polynomial Features
- One-Hot Encoding Categorical Features

MODELING

- Linear Regression/LASSO
- Decision Trees
- Random Forest

RMSE

Measure of error (in terms of popularity score 0-100)

R2

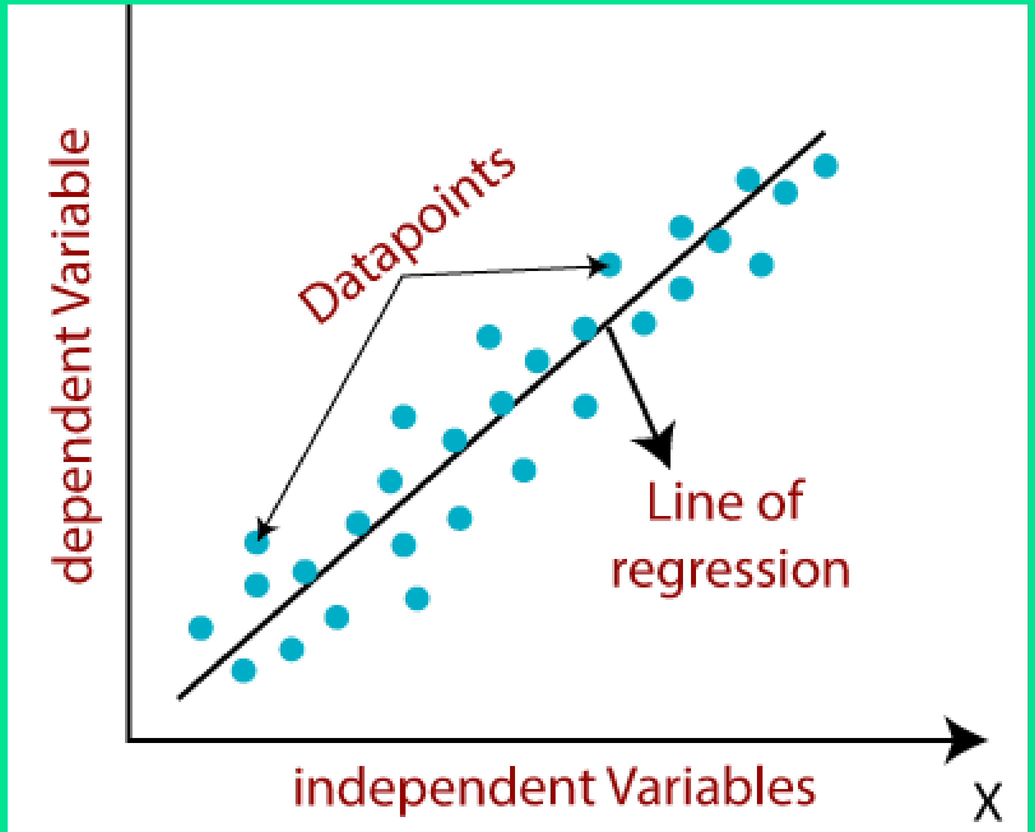
"Statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable"

POLYNOMIAL FEATURES

Features that are created by raising existing features to an exponent

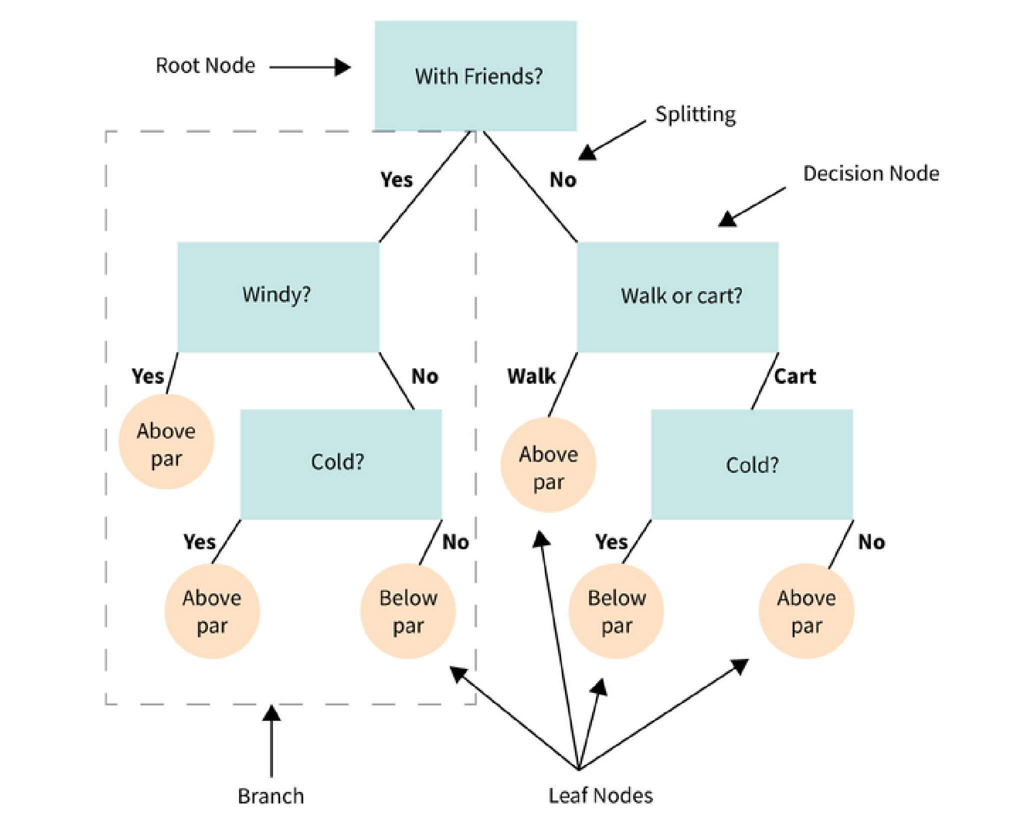
ONE-HOT ENCODING

Converting categorical data into binary representation



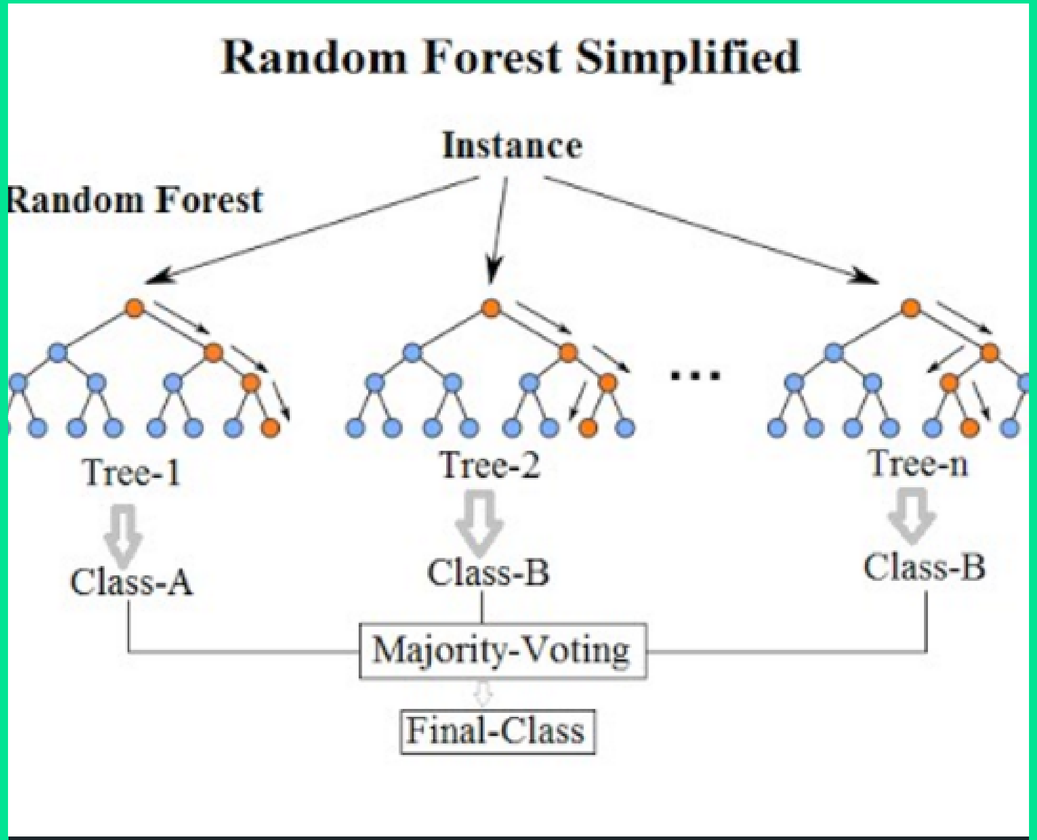
LINEAR REGRESSION

Establishing a relationship between the dependent and independent variable



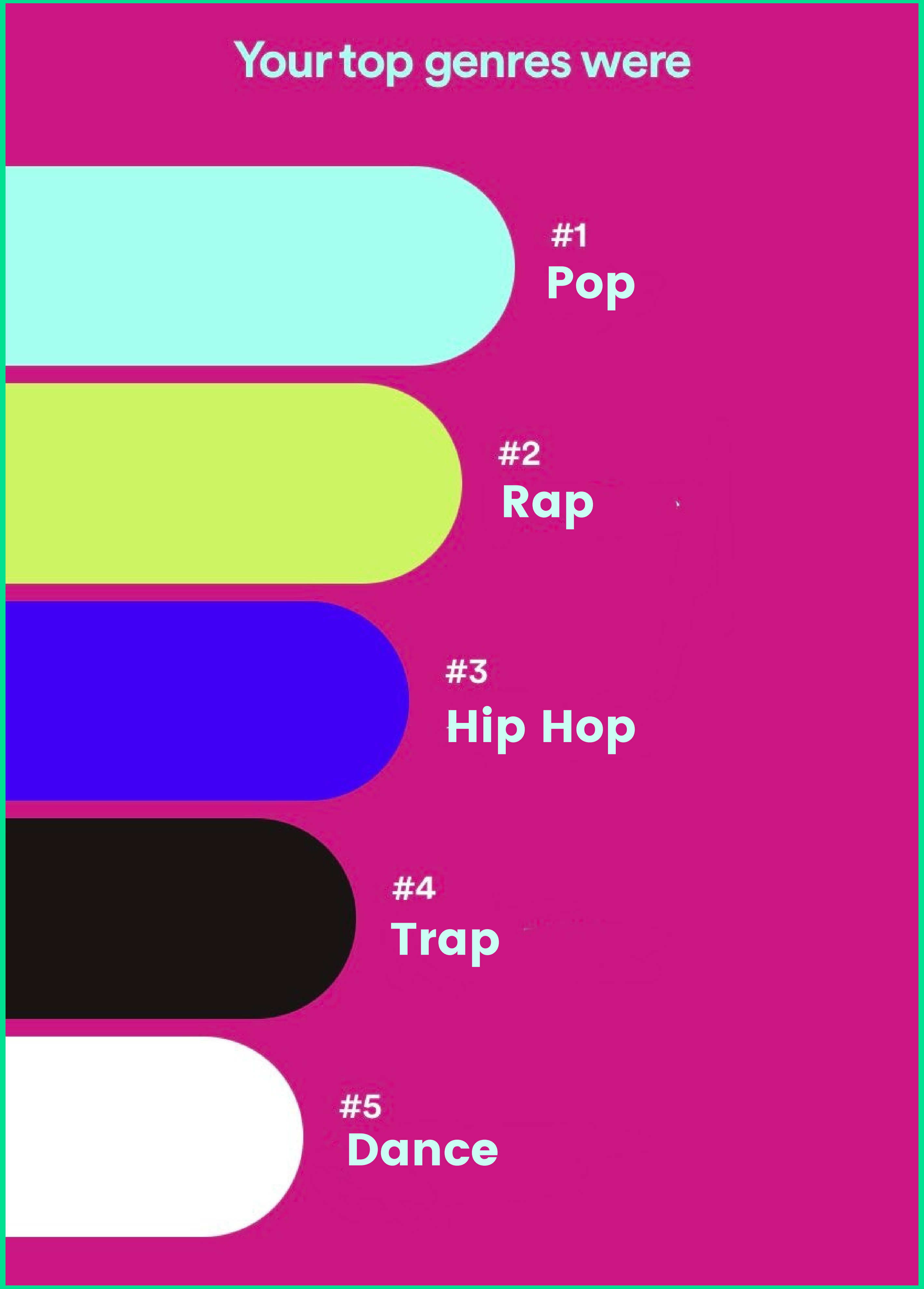
DECISION TREES

Takes a dataset, finds rules based on the X data and splits data into smaller datasets



RANDOM FOREST

A number of decision trees on various subsamples of datasets



Your top genres were

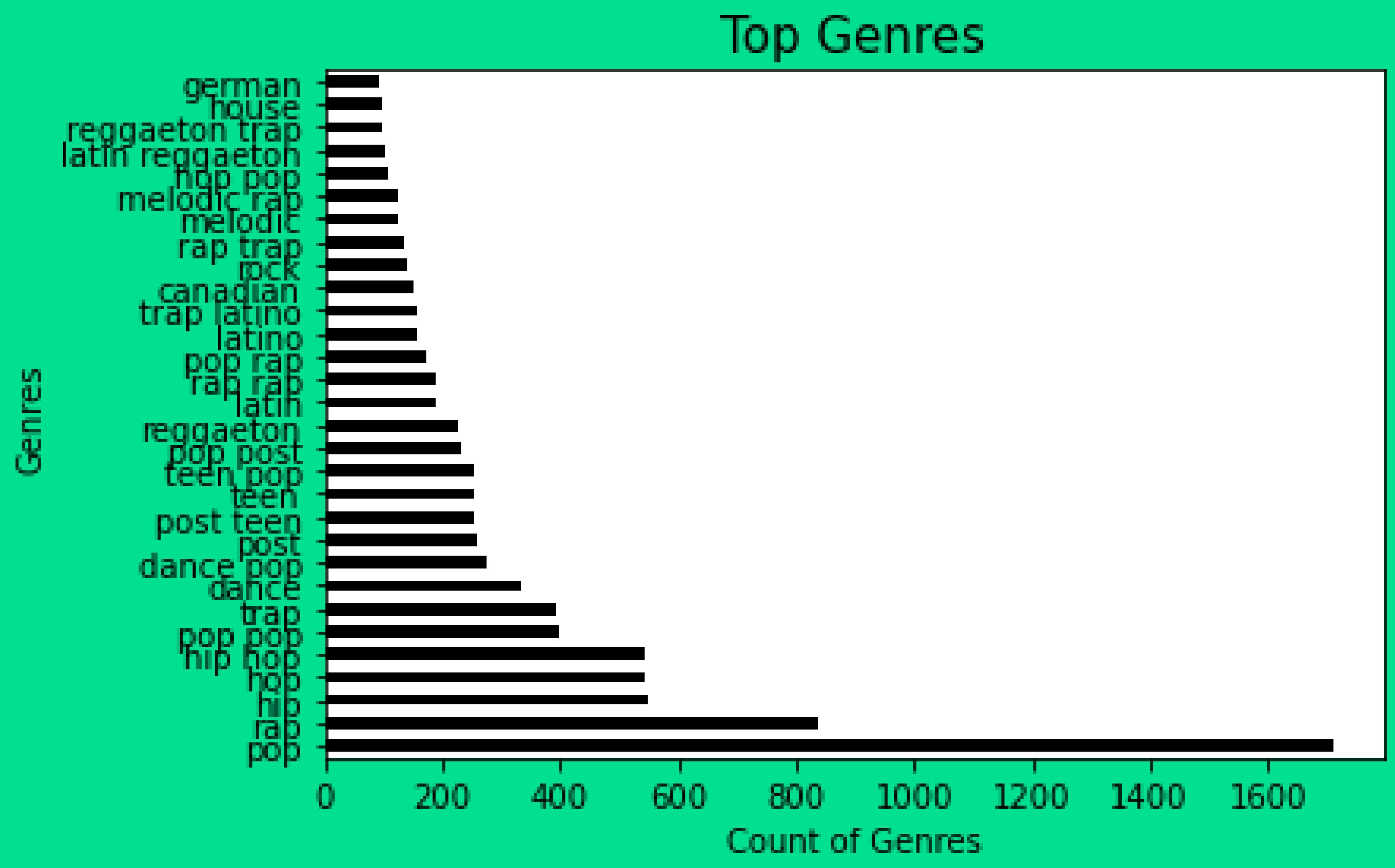
#1 Pop

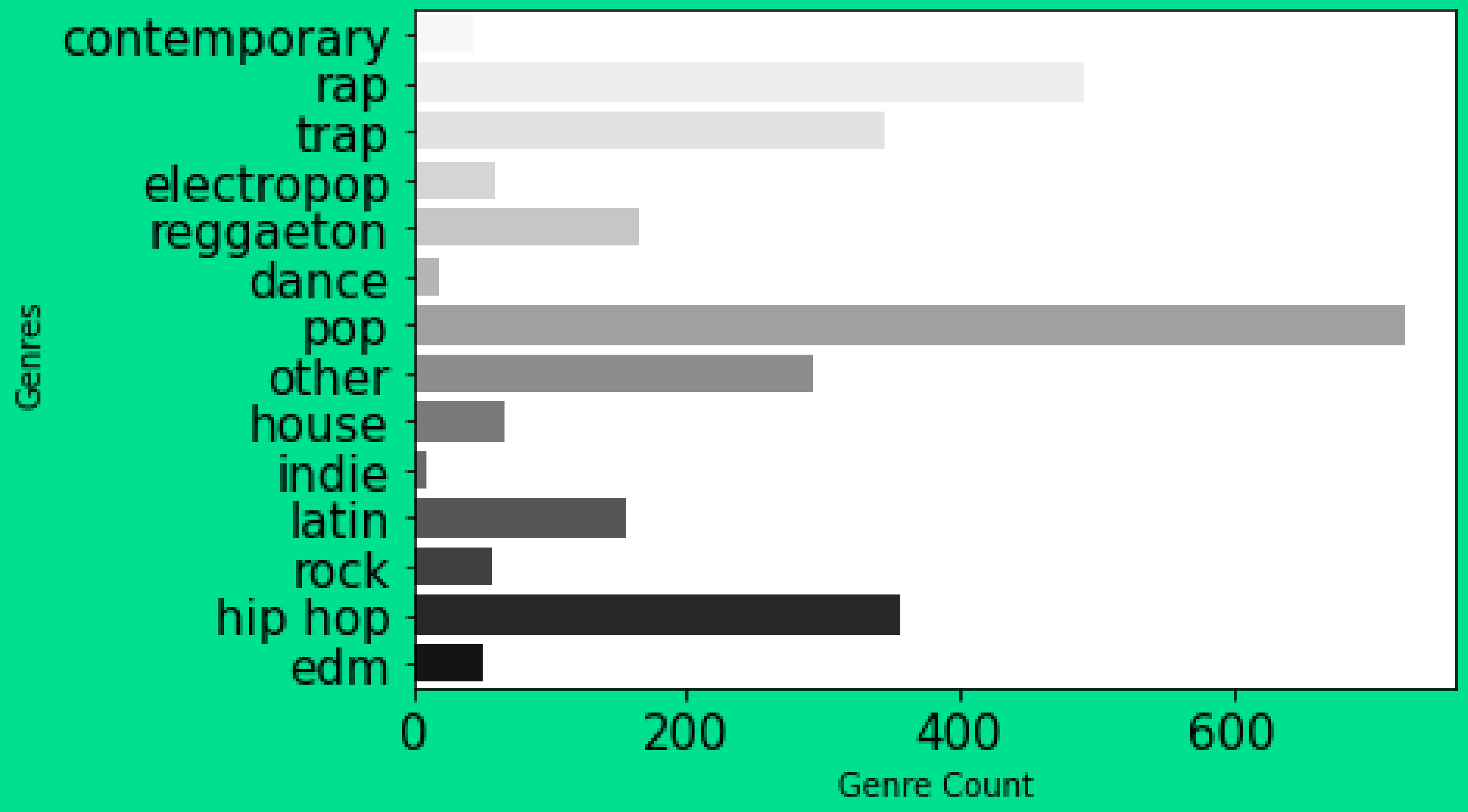
#2 Rap

#3 Hip Hop

#4 Trap

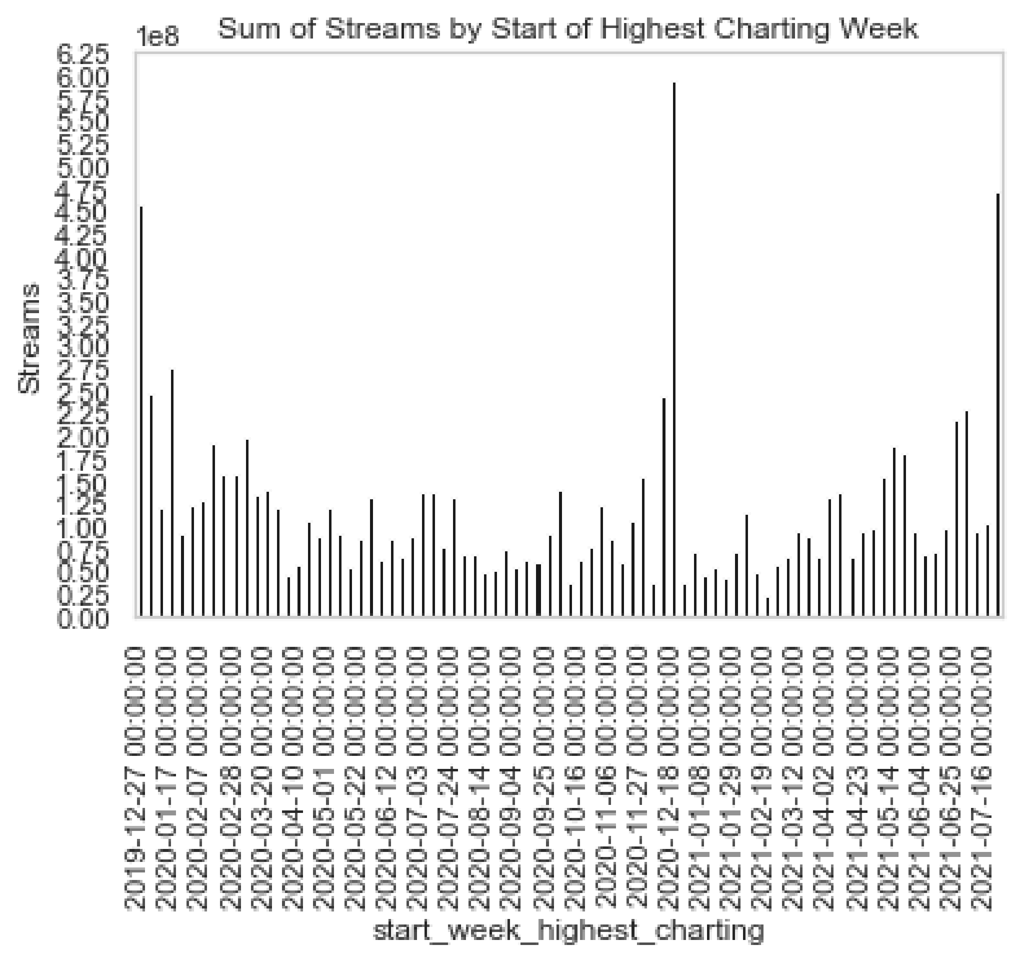
#5 Dance





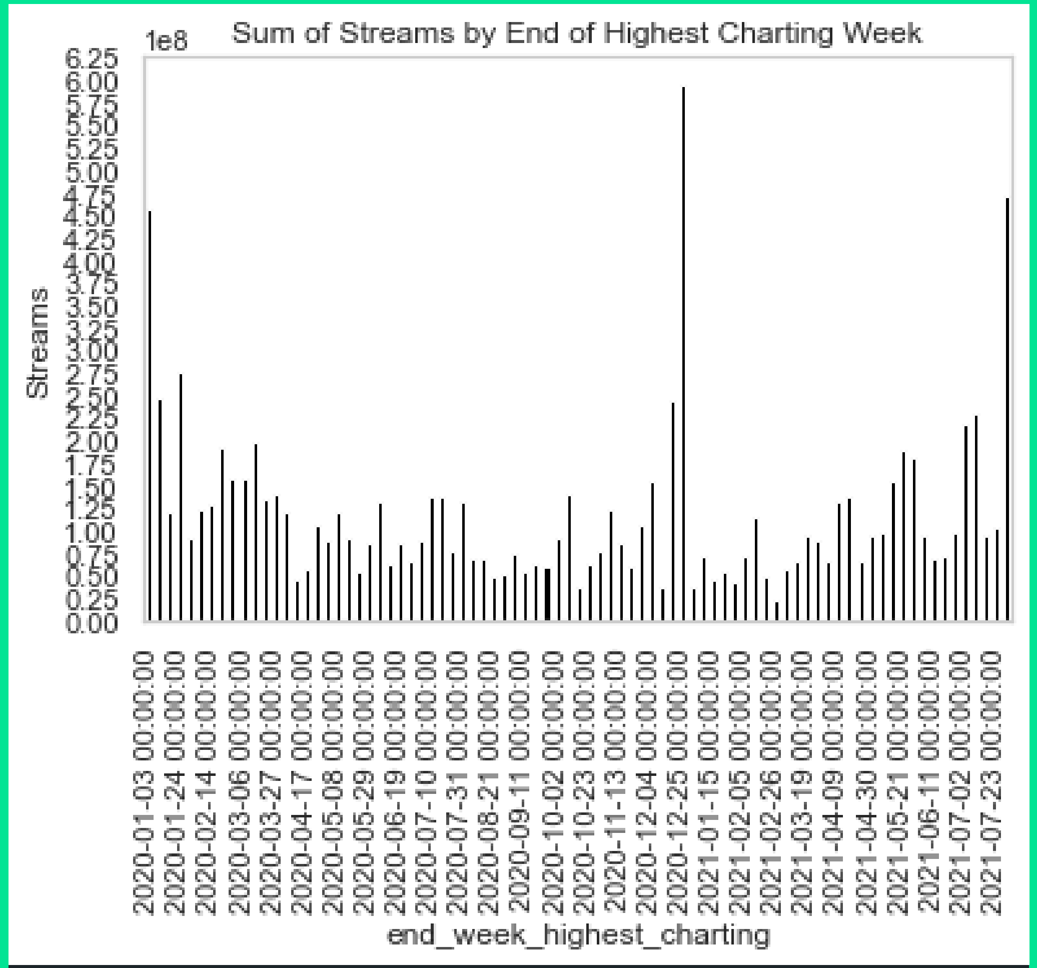
Lists of 335 various genres to 14 larger "family" genres

Christmas Songs: Outliers



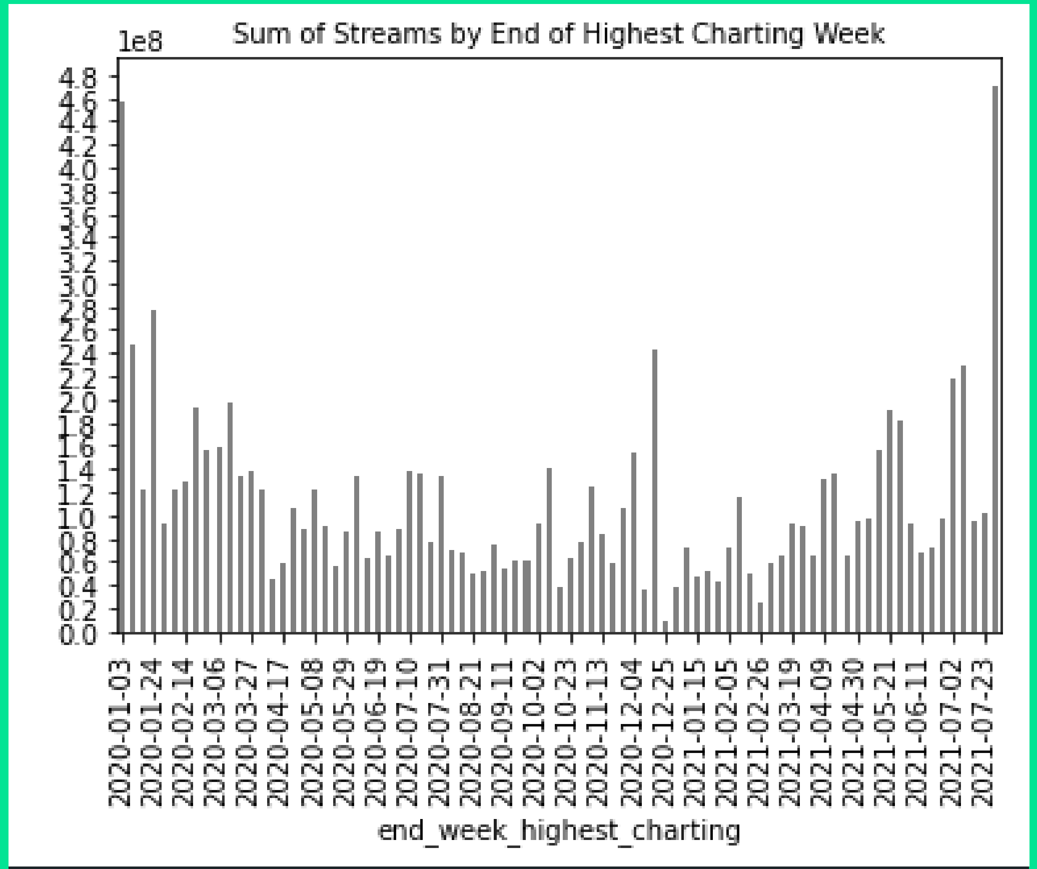
START OF HIGHEST CHARTING WEEK

2021-12-18



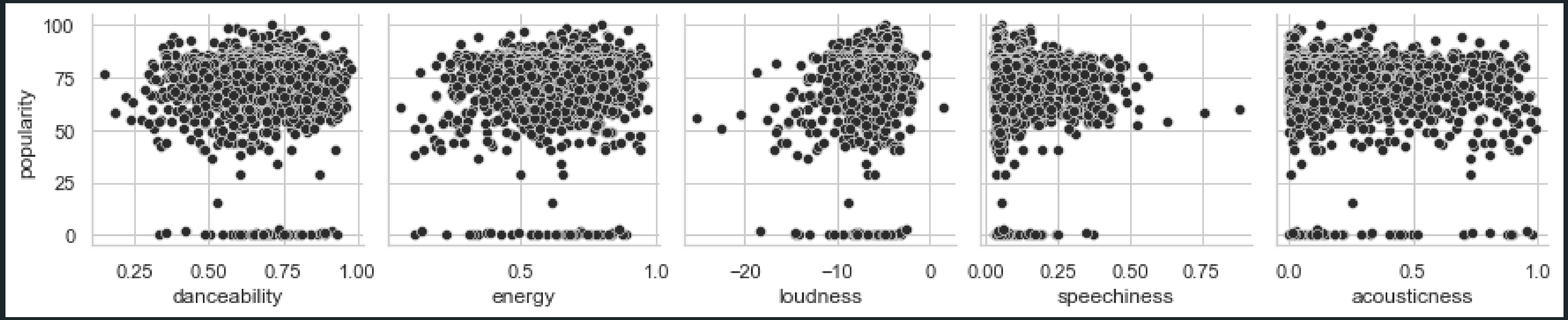
END OF HIGHEST CHARTING WEEK

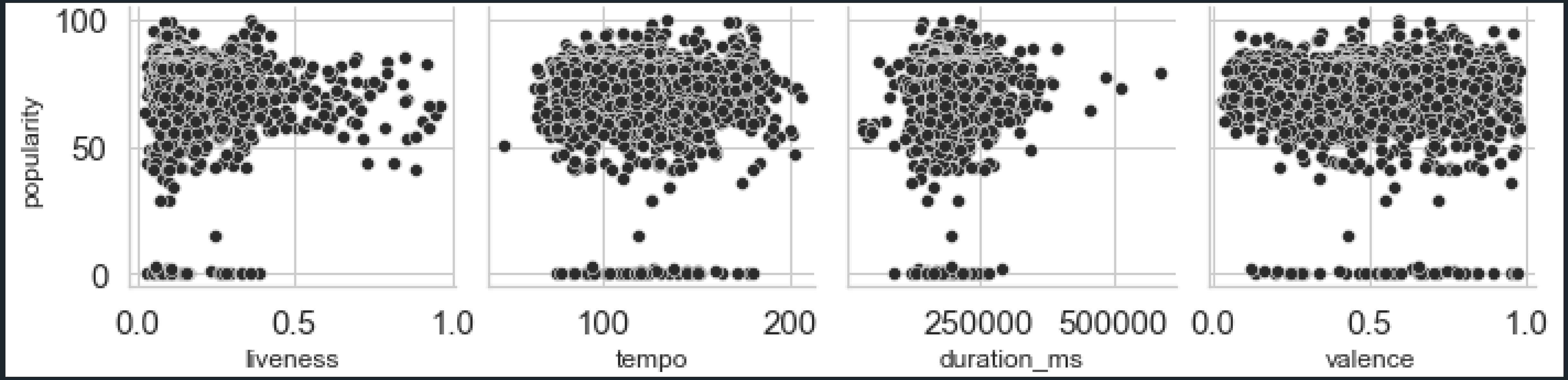
2021-12-25



OUTLIERS REMOVED

CORRELATIONS: POPULARITY

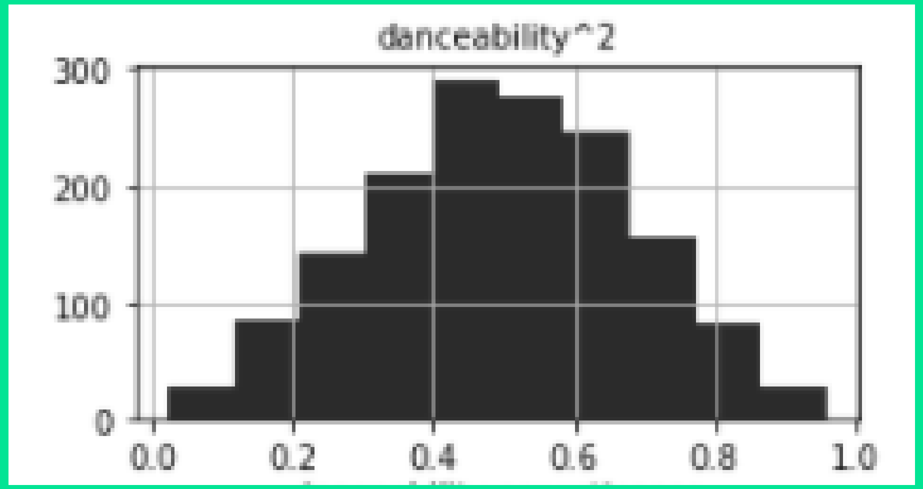




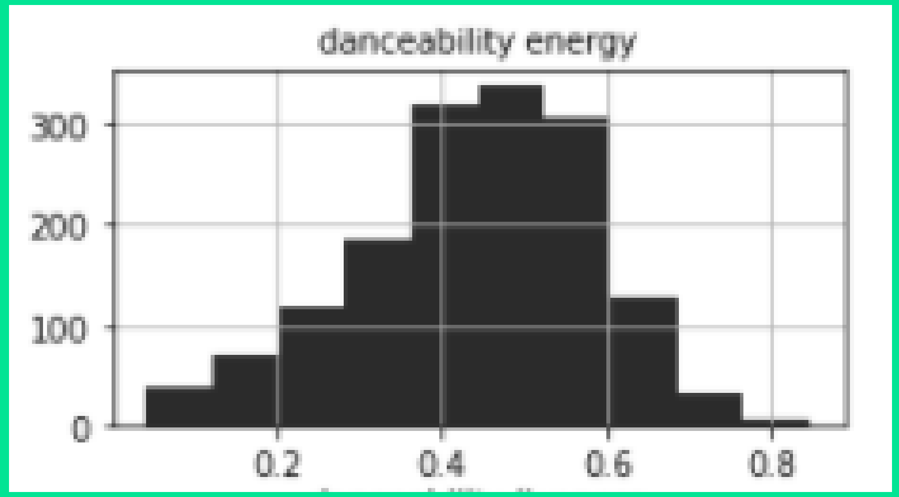
No clear correlations between popularity and numerical features

Polynomial Features

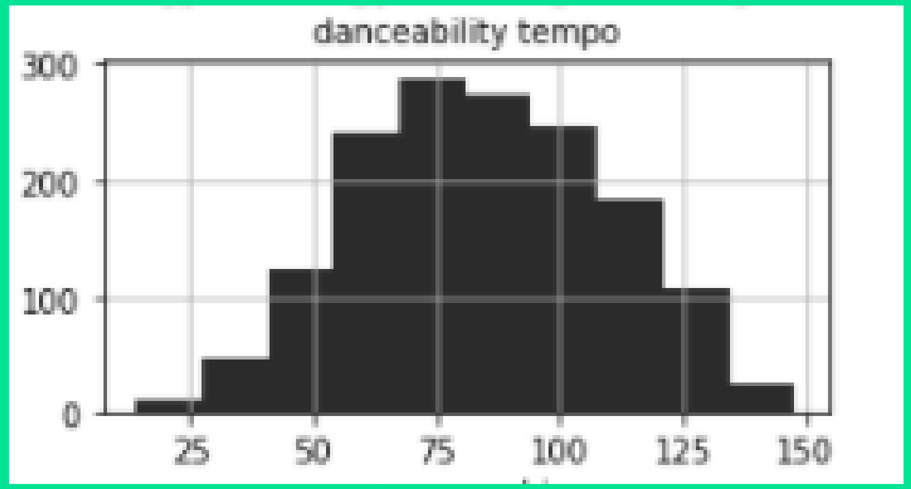
DANCEABILITY^2



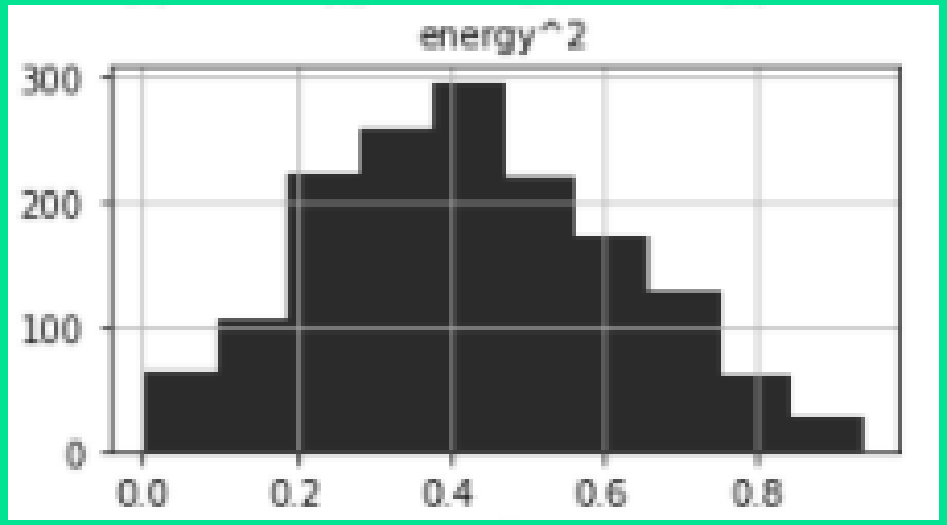
DANCEABILITY ENERGY



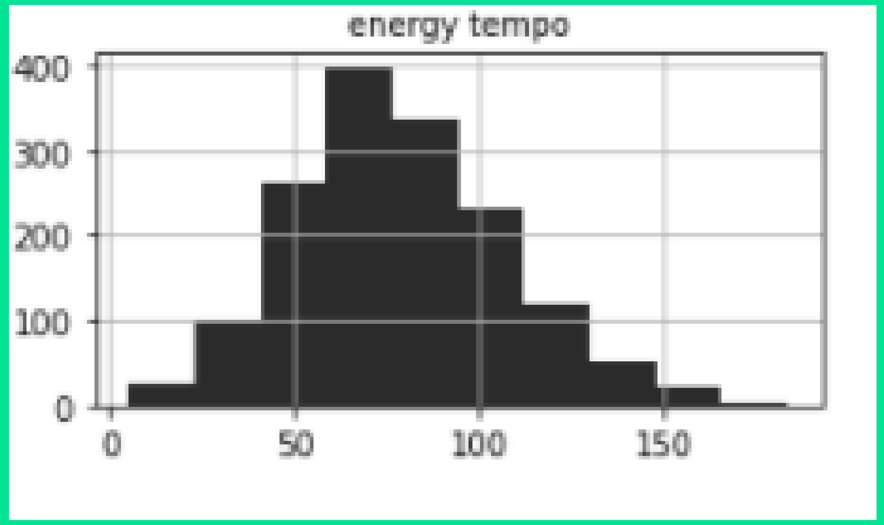
DANCEABILITY TEMPO



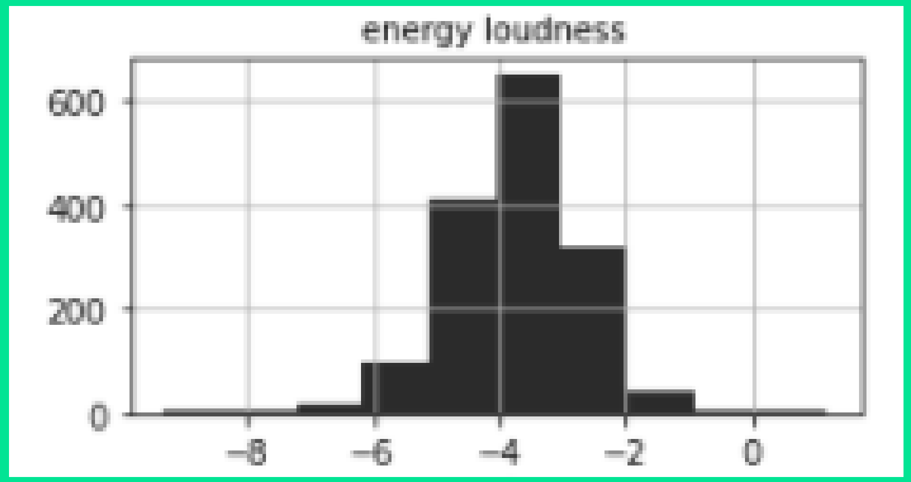
ENERGY^2

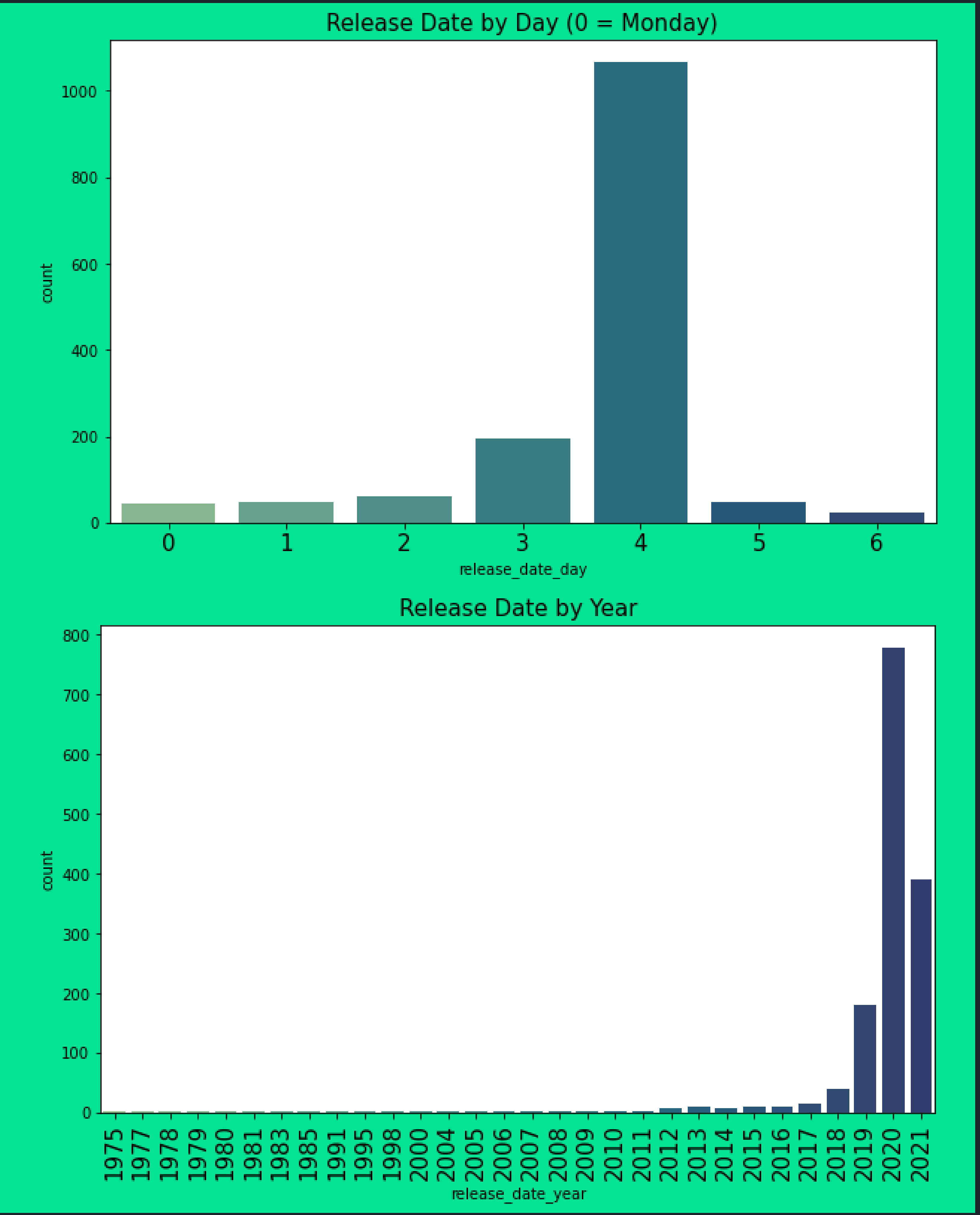


ENERGY TEMPO



ENERGY LOUDNESS



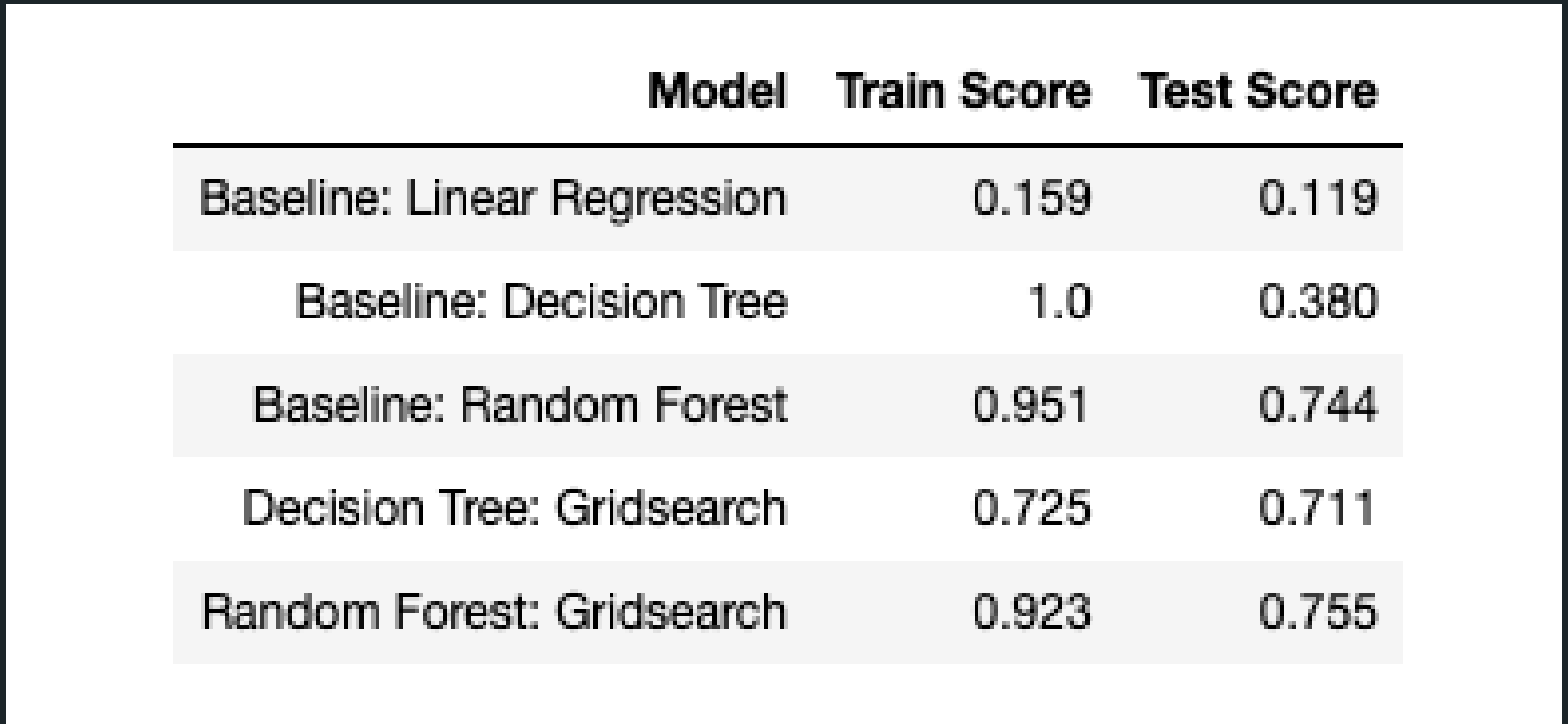


Release Date

CREATED 2 NEW FEATURES: DAY & YEAR

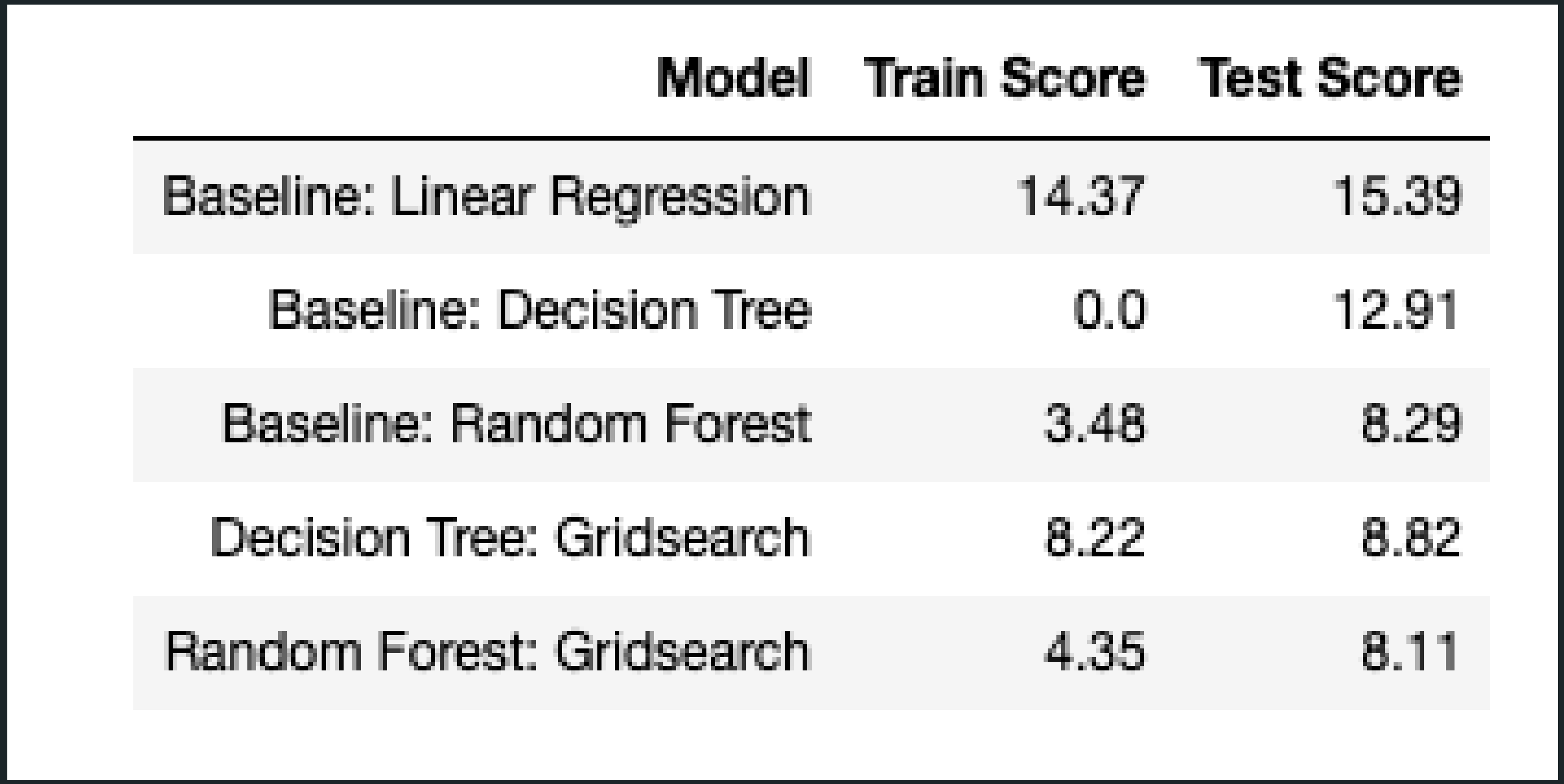
Baseline Models

R2



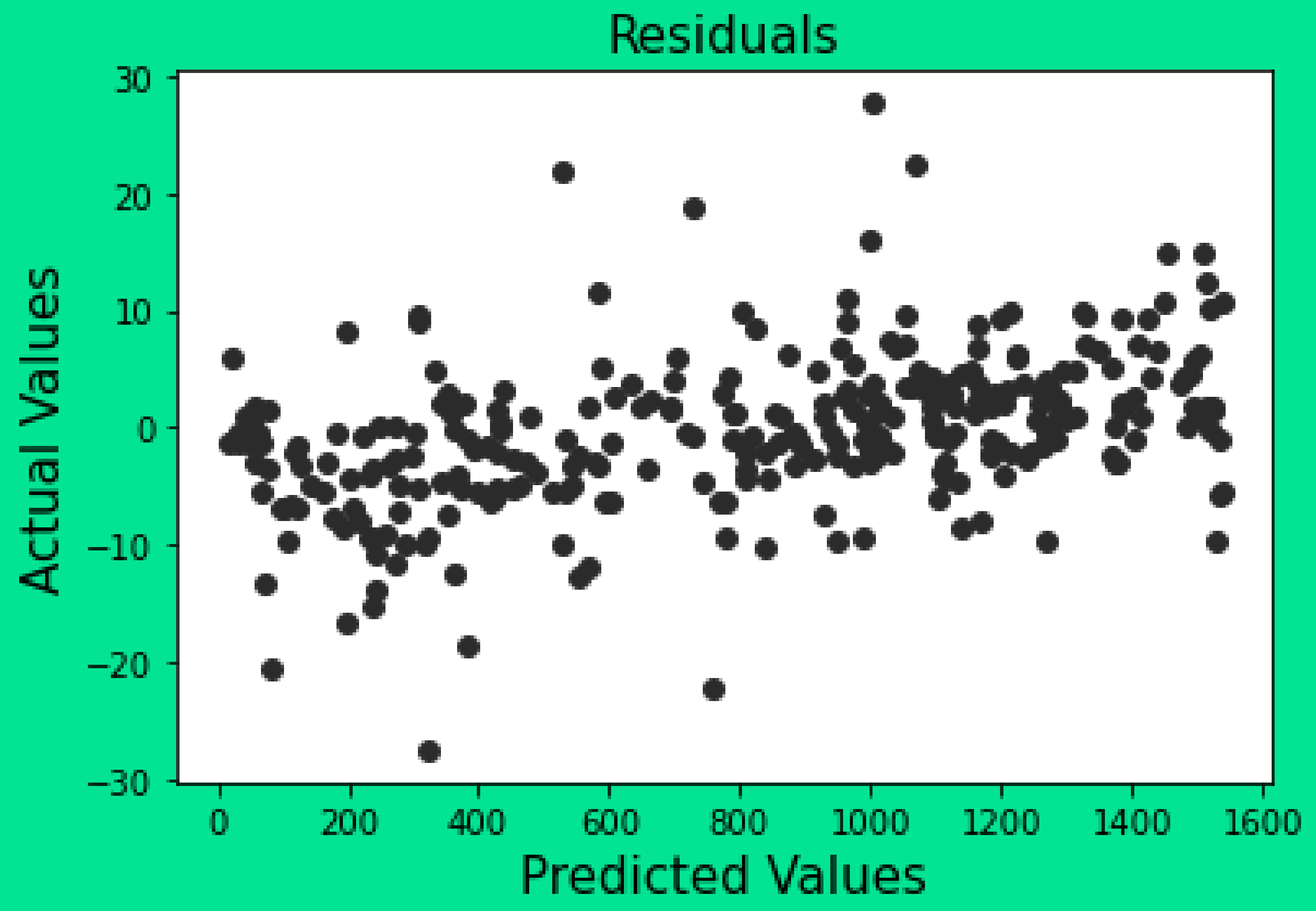
Model	Train Score	Test Score
Baseline: Linear Regression	0.159	0.119
Baseline: Decision Tree	1.0	0.380
Baseline: Random Forest	0.951	0.744
Decision Tree: Gridsearch	0.725	0.711
Random Forest: Gridsearch	0.923	0.755

RMSE



Model	Train Score	Test Score
Baseline: Linear Regression	14.37	15.39
Baseline: Decision Tree	0.0	12.91
Baseline: Random Forest	3.48	8.29
Decision Tree: Gridsearch	8.22	8.82
Random Forest: Gridsearch	4.35	8.11

polynomial features
polynomial features
polynomial features
polynomial features



RANDOM FOREST - POLYNOMIAL FEATURES ADDED

R2 Train Score: 0.962
R2 Test Score: 0.794

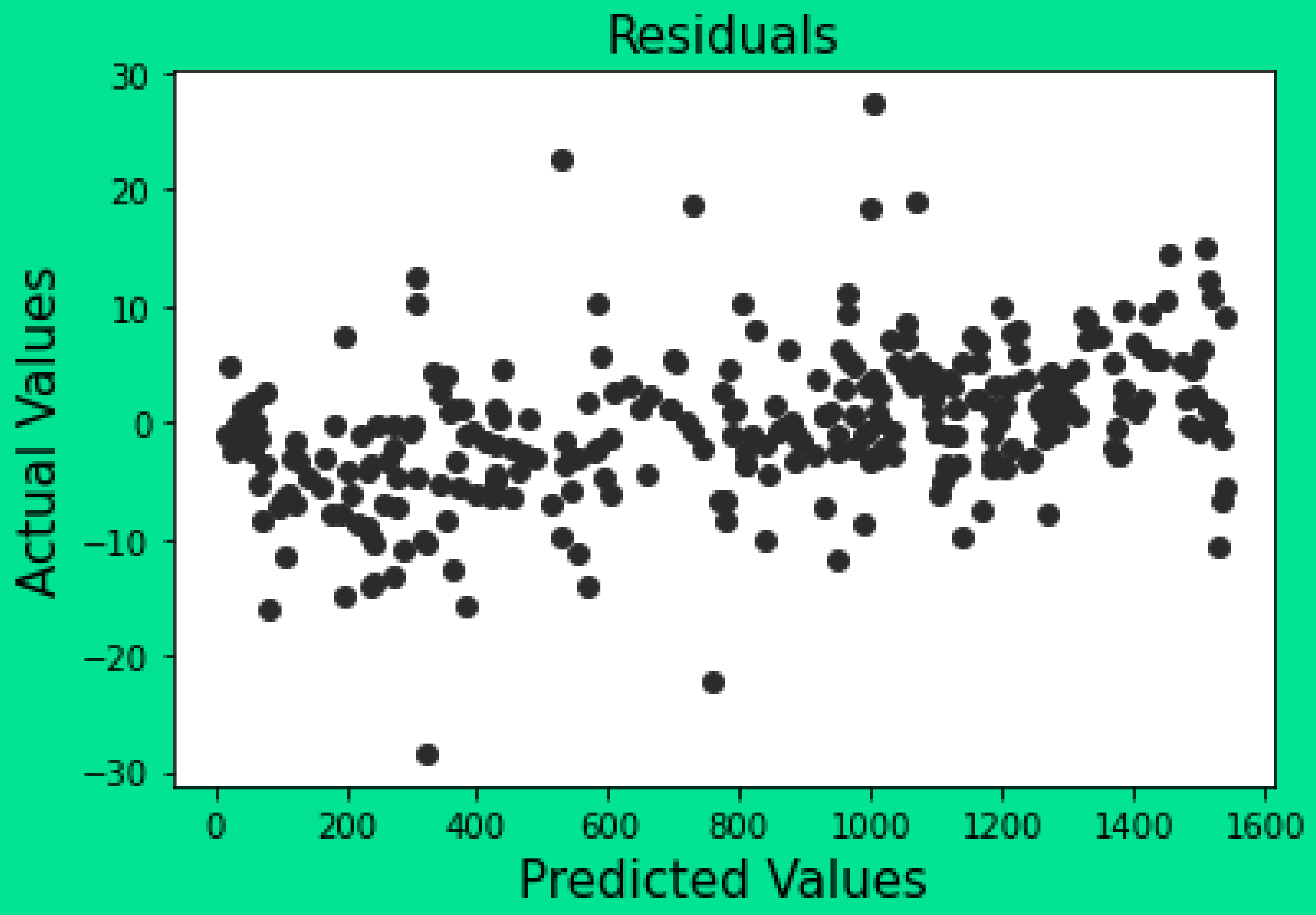
RMSE Train Score: 2.91
RMSE Test Score: 6.66

chord

chord

chord

chord



RANDOM FOREST - CHORD ADDED

R2 Train Score: 0.962

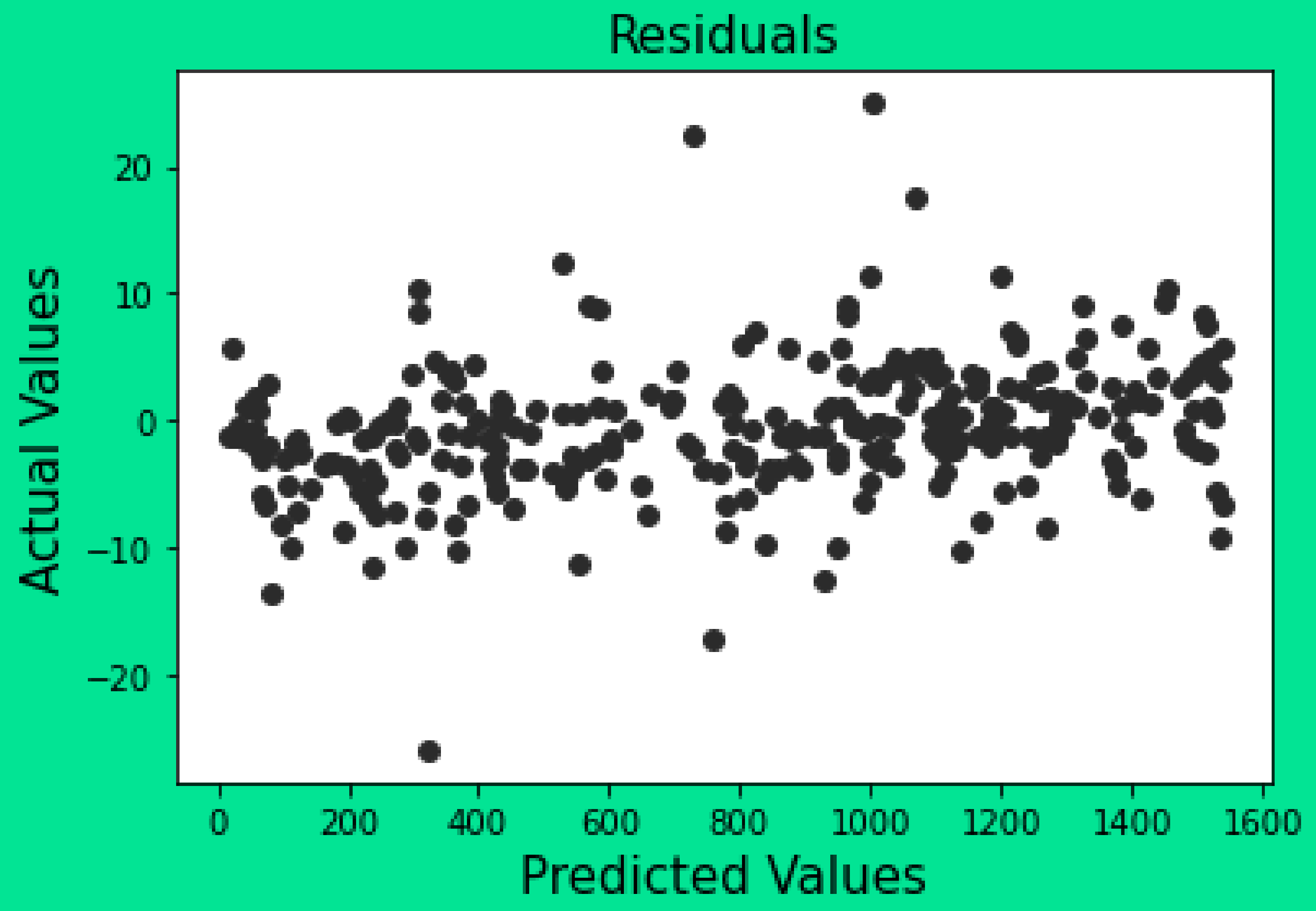
R2 Test Score: 0.799

RMSE Train Score: 2.93

RMSE Test Score: 6.58

release date & year

release date & year
release date & year
release date & year



RANDOM FOREST - RELEASE DATE & YEAR ADDED

R2 Train Score: 0.972

R2 Test Score: 0.868

RMSE Train Score: 2.49

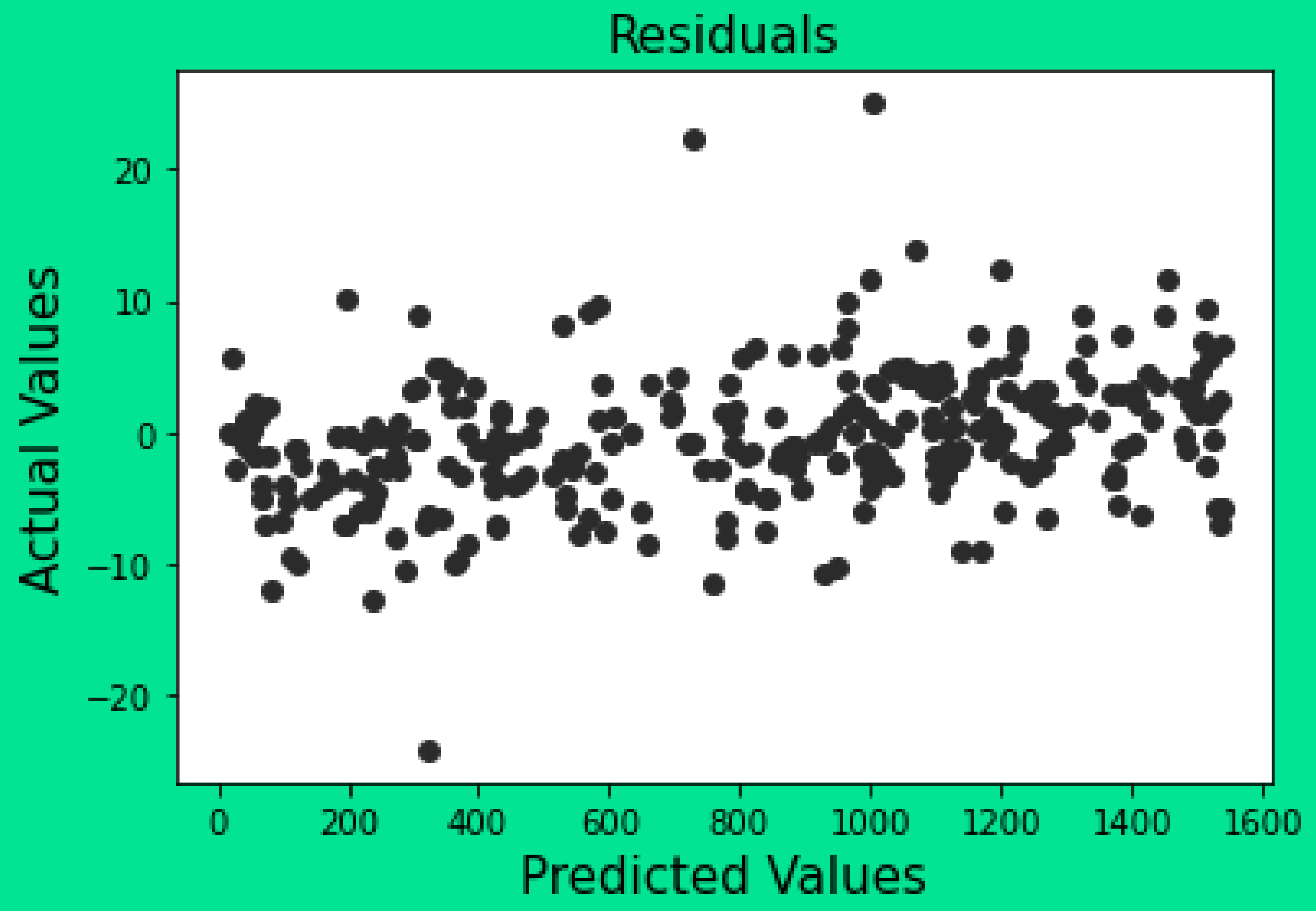
RMSE Test Score: 5.32

LASSO

LASSO

LASSO

LASSO



RANDOM FOREST - AFTER LR/LASSO

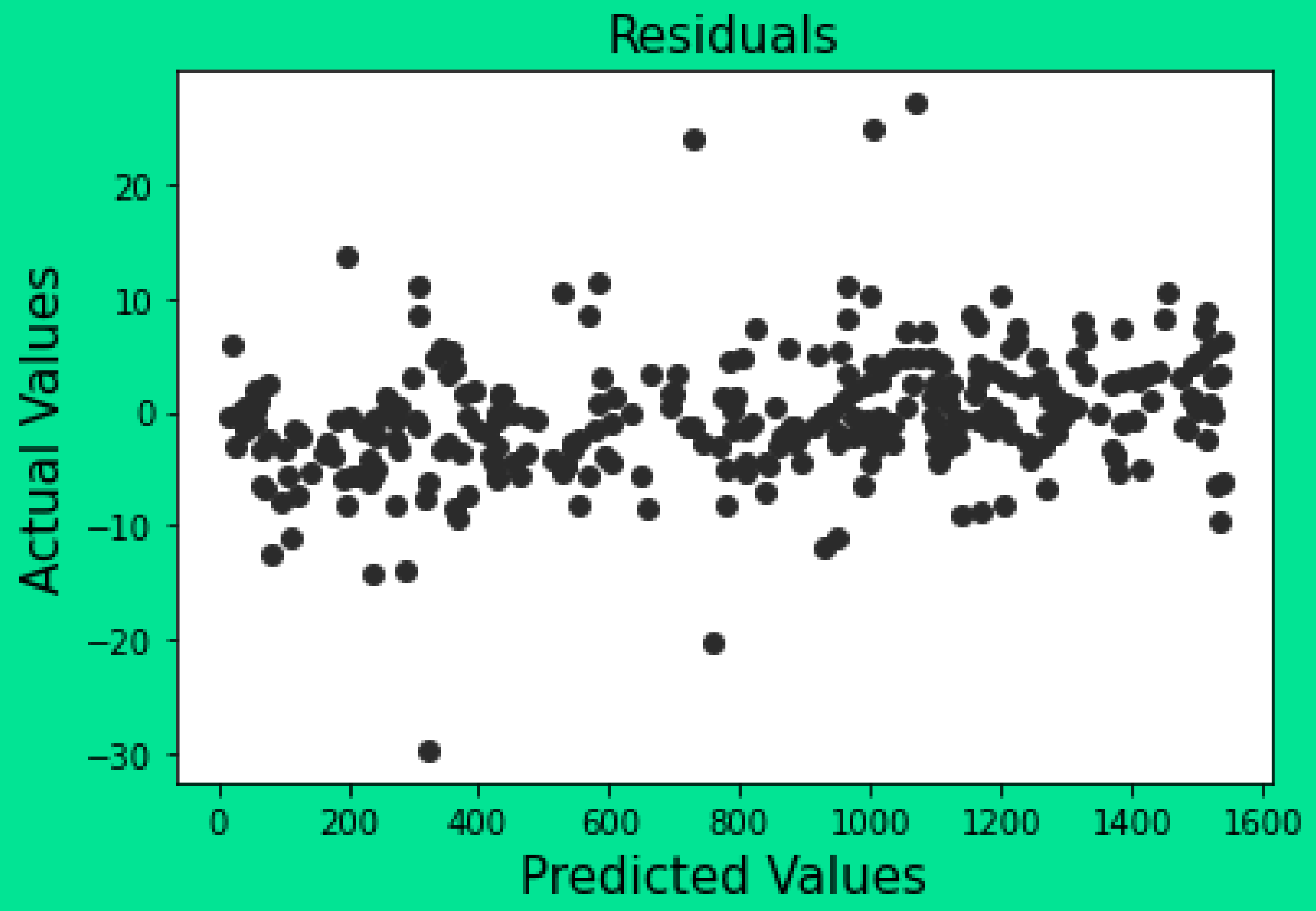
R2 Train Score: 0.972

R2 Test Score: 0.872

RMSE Train Score: 2.48

RMSE Test Score: 5.25

random forest gridsearch
random forest gridsearch
random forest gridsearch
random forest gridsearch



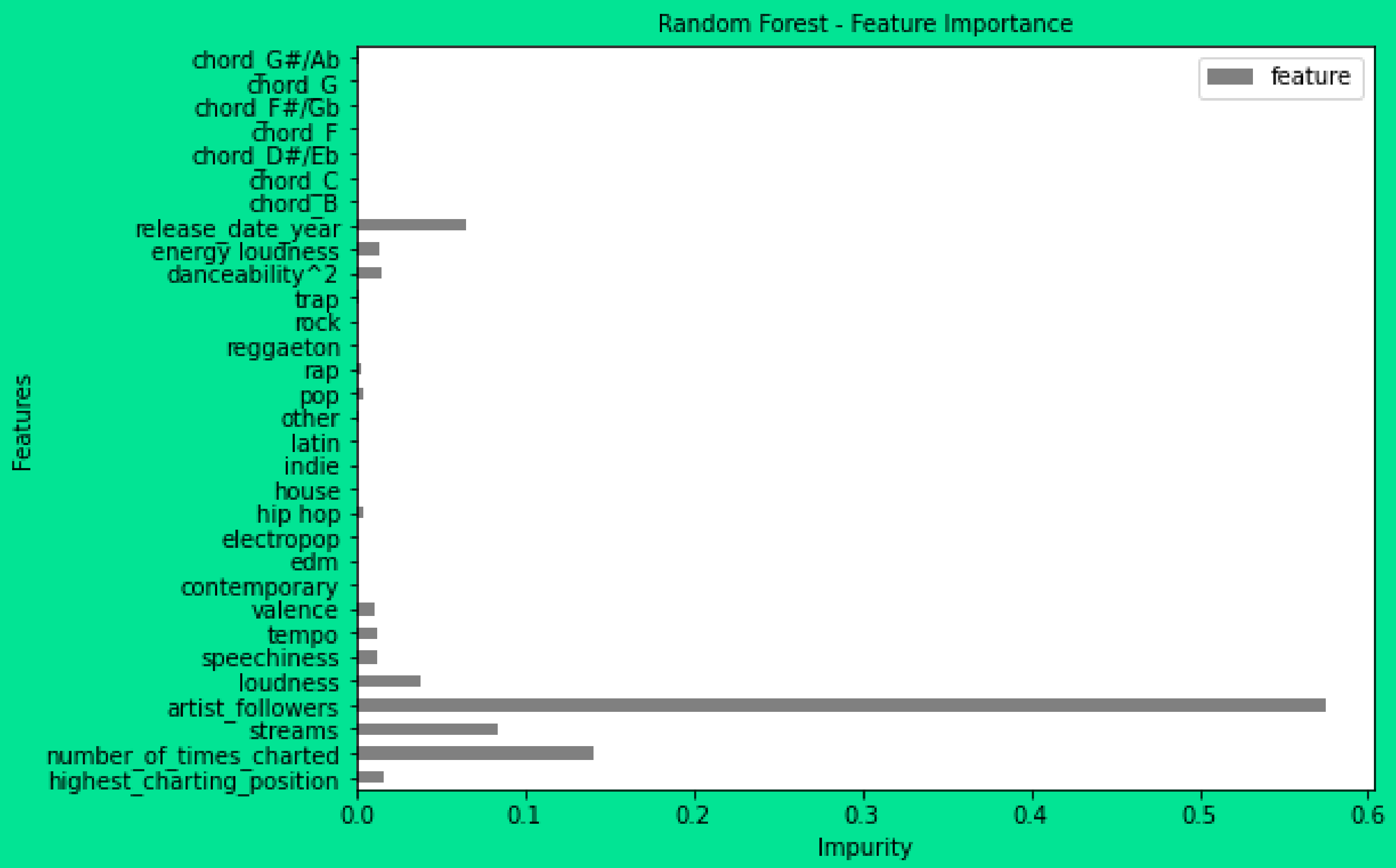
RANDOM FOREST - WITH GRIDSEARCH

R2 Train Score: 0.903

R2 Test Score: 0.848

RMSE Train Score: 4.65

RMSE Test Score: 5.71



Feature Importance

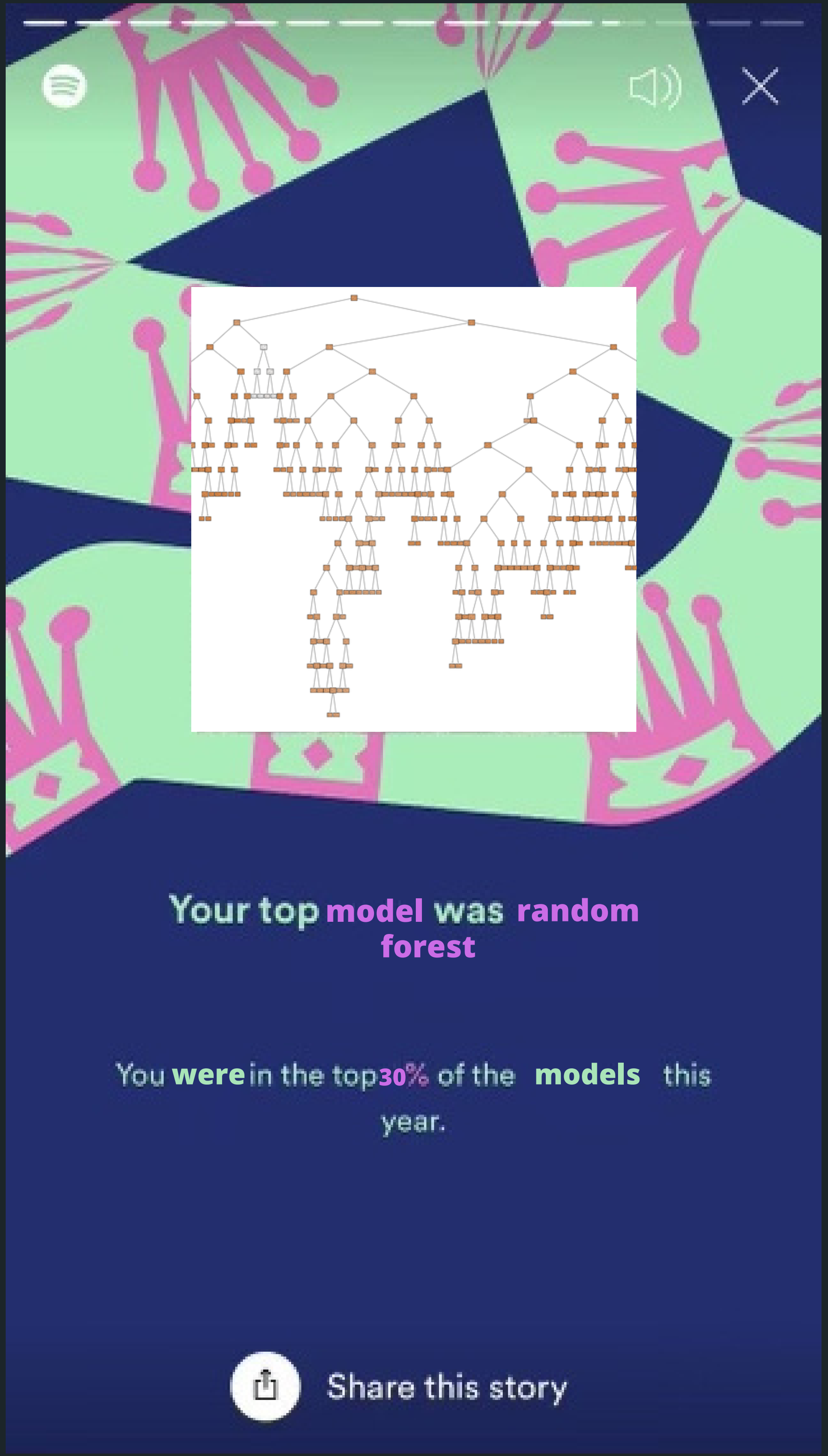
● ● ●

RANDOM FOREST

- artist followers
- number of times charted
- streams
- release date (year)

Conclusion

You trained 10 different models, but things got pretty serious with one...



Your top model was random forest

You were in the top30% of the models this year.

Share this story

Improvement of ~14% in R2 score and ~30% in RMSE

Feature Importance: artist followers, streams, release date

Recommendation: focus on artist following, release date, but also specific audio features

Sources

- https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20(R2),variables%20in%20a%20regression%20model.
- https://static.javatpoint.com/tutorial/machine-learning/images/linear-regression-in-machine-learning.png
- https://www.mastersindatascience.org/wp-content/uploads/tree-graphic.jpg
- https://upload.wikimedia.org/wikipedia/commons/7/76/Random_forest_diagram_complete.png
- https://imgflip.com/meme/290748647/Spotify-Wrapped
- https://www.reddit.com/r/outerwilds/comments/r6hydf/spotify_wrapped_2021_guess_who_my_top_artist_was/