Drifting Upstream

Predicting song popularity from music features

The Problem

- Regression Records music label
- Extra promotional budget for Q3 and Q4 of 2024
- Hundreds of songs in catalog, must choose 3
- **Predict song popularity from music features**

Data Wrangling & EDA Pt. 1

945 Songs

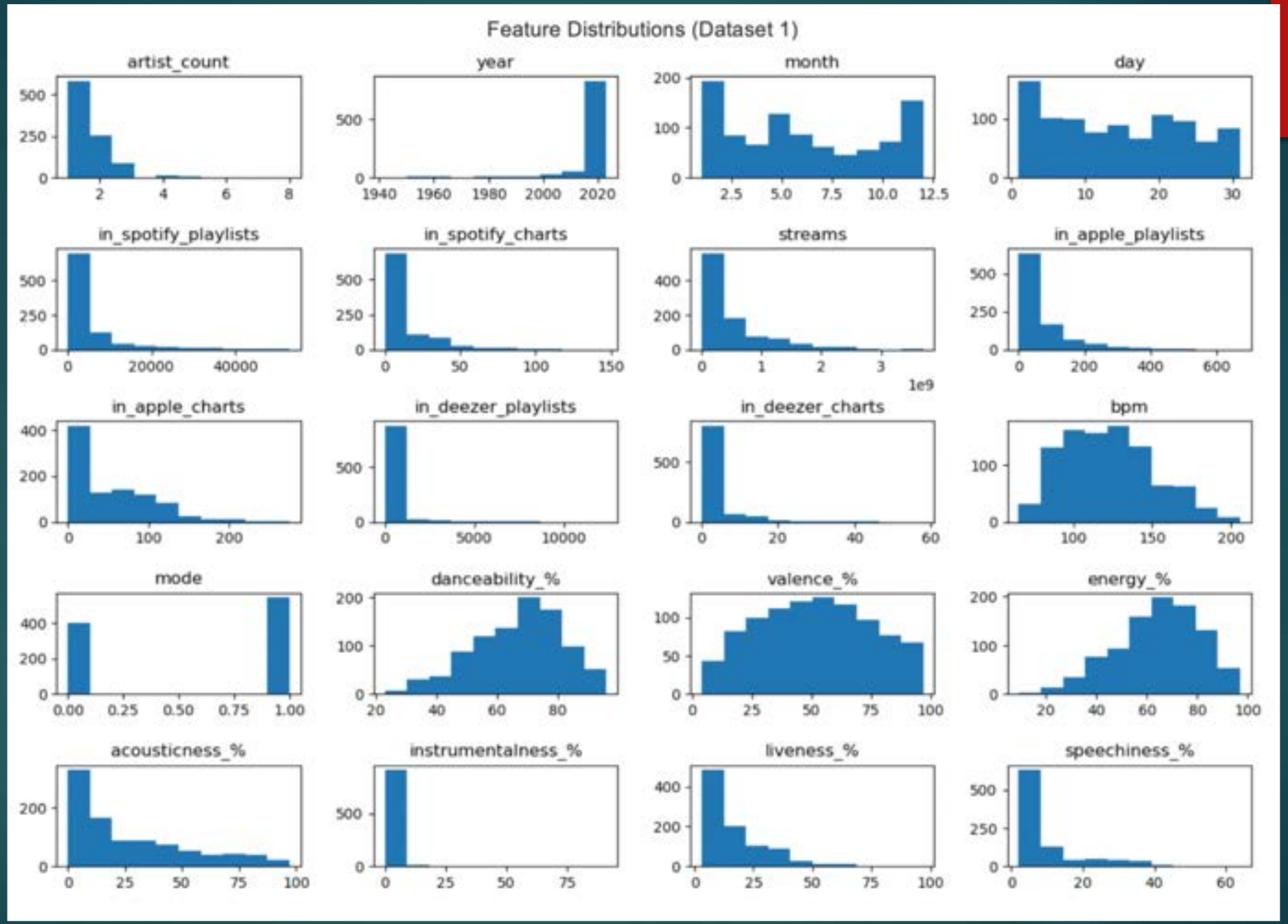
Target Variable
Total # Streams

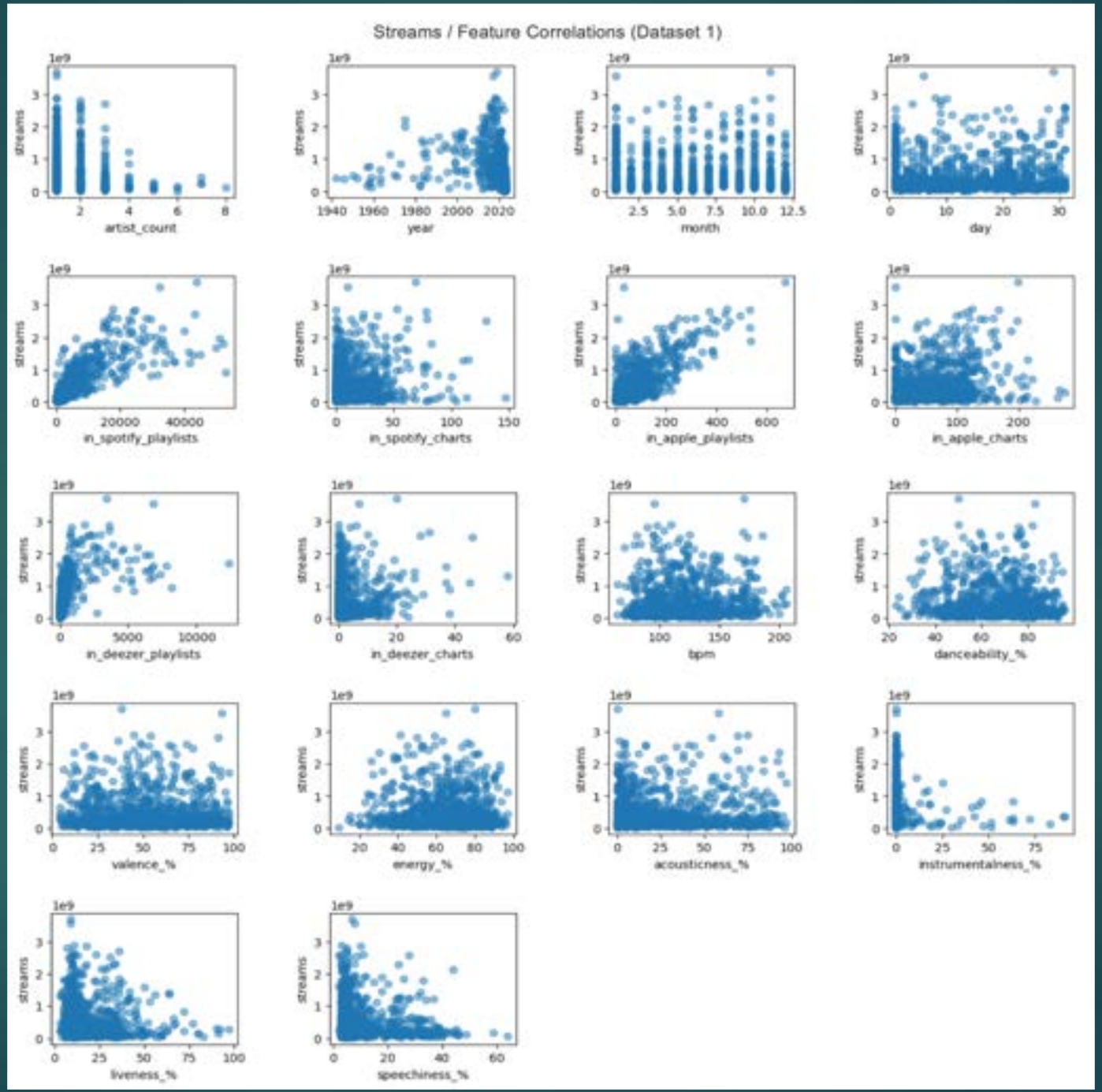
Median ~290 million
Min ~1.4 million
Max ~3.7 billion

Data Wrangling & EDA Pt. 1

Features

- Tempo
- Danceability
- Valence
- Energy
- Acousticness
- Instrumentalness
- Liveness
- Speechiness
- Mode


Feature Distributions (Dataset 1)
artist_count
year
month
day
in_spotify_playlists
in_spotify_charts
streams
1e9
in_apple_playlists
in_apple_charts
in_deezer_playlists
in_deezer_charts
bpm
mode
danceability_%
valence_%
energy_%
acousticness_%
instrumentalness_%
liveness_%
speechiness_%


Streams / Feature Correlations (Dataset 1)
streams
1e9
artist_count
year
month
day
in_spotify_playlists
in_spotify_charts
in_apple_playlists
in_apple_charts
in_deezer_playlists
in_deezer_charts
bpm
danceability_%
valence_%
energy_%
acousticness_%
instrumentalness_%
liveness_%
speechiness_%

Modeling Pt. 1

Random Forest Regressor

- Train / Test MAE 335.5M / 332.4M

Extra Trees Regressor

- Train / Test MAE 0 / 340M

Gradient Boosting Regressor

- Train / Test MAE 310M / 348M

(70 / 30 train test split)

Modeling Pt. 1

Random Forest Regressor

PowerTransformer()

- Train / Test MAE 335.5M / 332.2M

90 / 10 Split

- Train / Test MAE 332M / 268M

Remove top 5% outliers (80 / 20)

- Train / Test MAE 247M / 270M

Modeling Pt. 1

Random Forest Regressor

Hyperparameters

- n_estimators = 200
- max_features = 6
- max_depth = 30
- min_samples_split = 7
- min_samples_leaf = 11
- random_state = 42

Feature Importance

1. **Tempo (0.166281)**
2. **Danceability (0.161662)**
3. **Acousticness (0.155589)**
4. **Valence (0.146104)**
5. Speechiness (0.119351)
6. Liveness (0.115062)
7. Energy (0.109871)
8. Mode (0.023337)
9. Instrumentalness (0.002743)

Data Wrangling & EDA Pt. 2

32,833 Songs

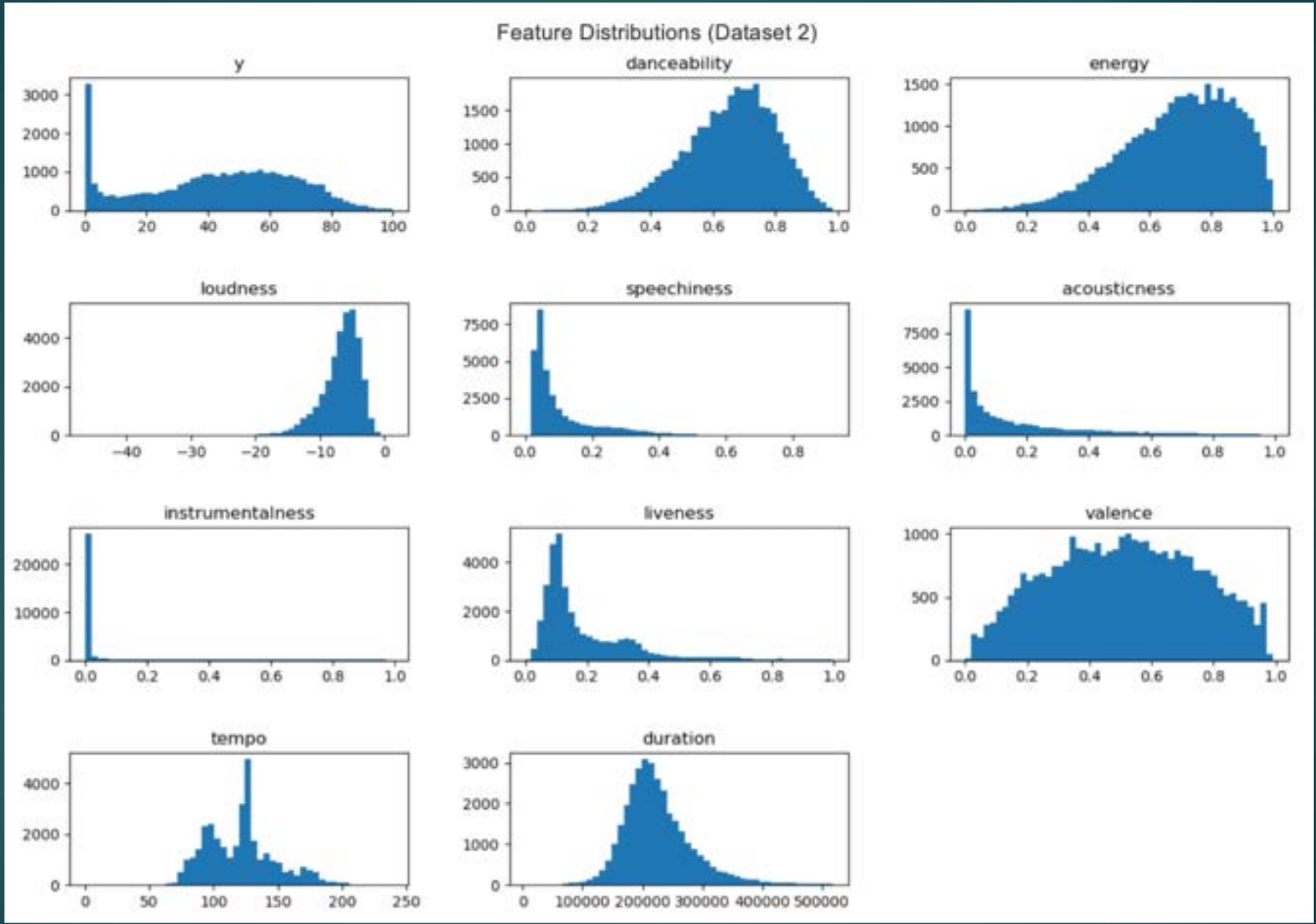
Target Variable
Popularity Score (1-100)

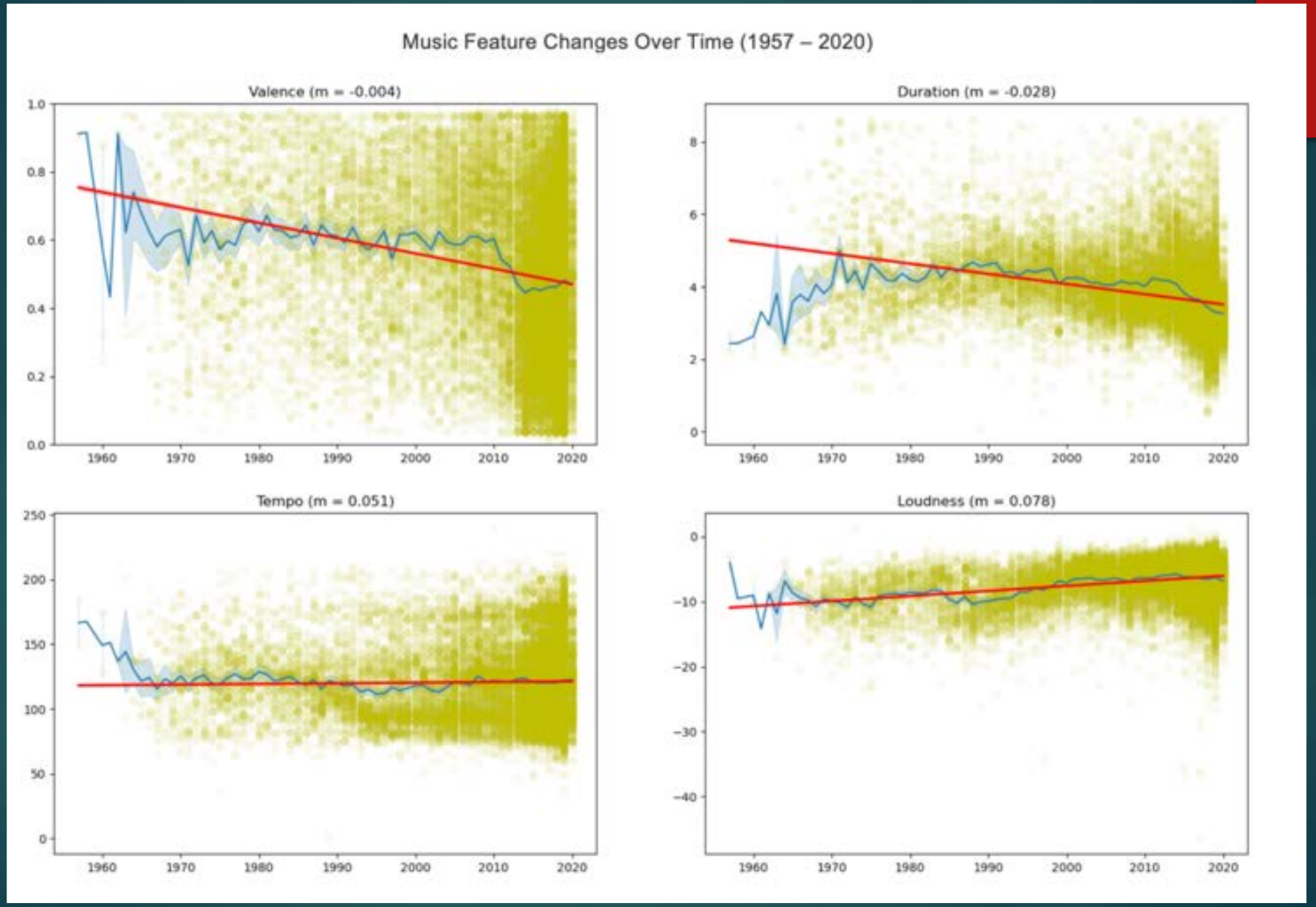
Mean 42.48
Median 45

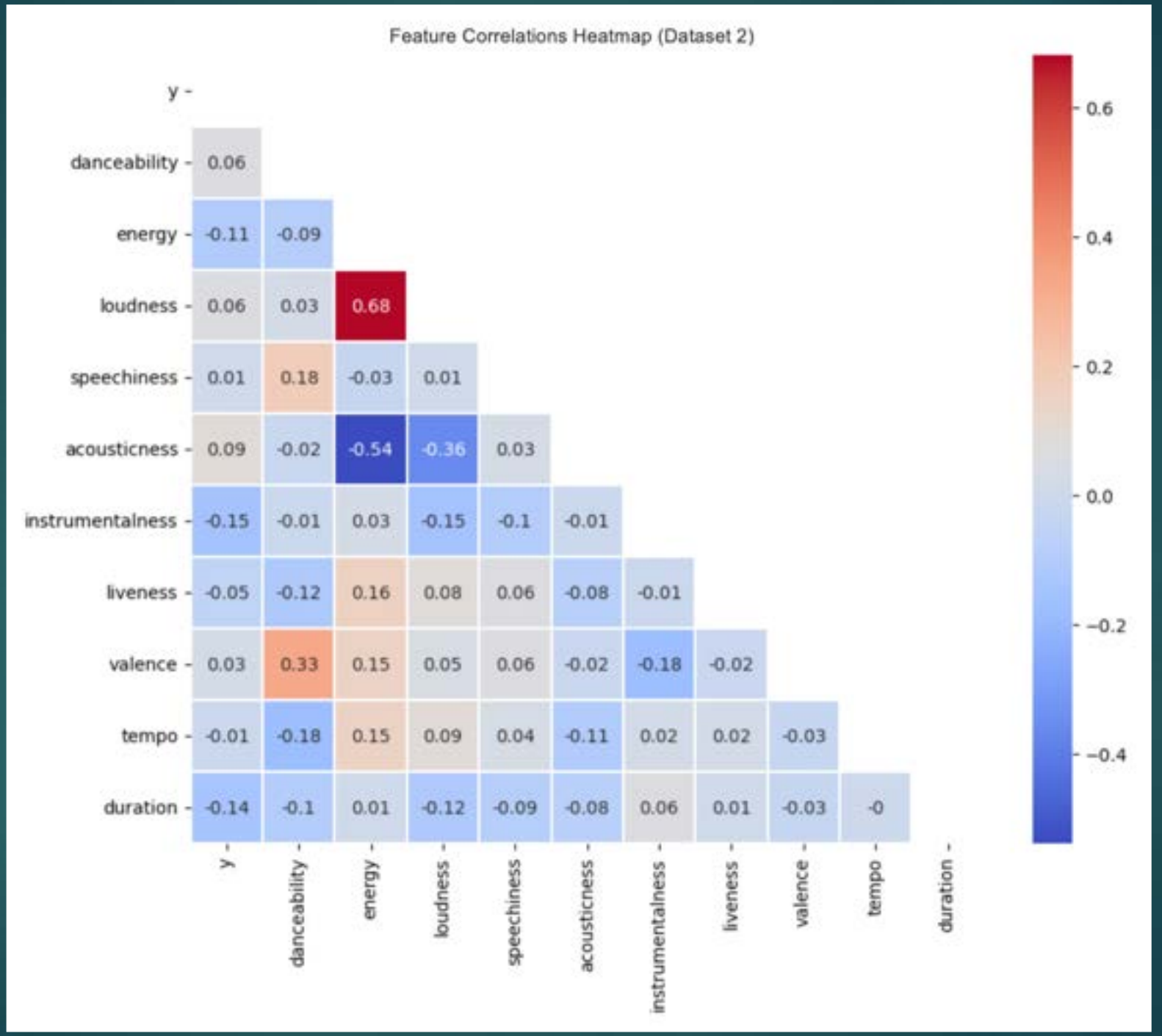
Data Wrangling & EDA Pt. 2

Features

- Tempo
- Danceability
- Valence
- Energy
- Acousticness
- Instrumentalness
- Liveness
- Speechiness
- **Duration**
- **Loudness**


Feature Distributions (Dataset 2)
y
danceability
energy
loudness
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration


Music Feature Changes Over Time (1957 – 2020)
Valence (m = -0.004)
Duration (m = -0.028)
Tempo (m = 0.051)
Loudness (m = 0.078)
1960
1970
1980
1990
2000
2010
2020


Feature Correlations Heatmap (Dataset 2)
y
danceability
energy
loudness
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration
0.06
-0.11 -0.09
0.06 0.03 0.68
0.01 0.18 -0.03 0.01
0.09 -0.02 -0.54 -0.36 0.03
-0.15 -0.01 0.03 -0.15 -0.1 -0.01
-0.05 -0.12 0.16 0.08 0.06 -0.08 -0.01
0.03 0.33 0.15 0.05 0.06 -0.02 -0.18 -0.02
-0.01 -0.18 0.15 0.09 0.04 -0.11 0.02 0.02 -0.03
-0.14 -0.1 0.01 -0.12 -0.09 -0.08 0.06 0.01 -0.03 -0
0.6
0.4
0.2
0.0
−0.2
−0.4

Modeling Pt. 2

Random Forest Regressor
- Standard Scaler
- CV / Test MAE 16.86 / 16.31

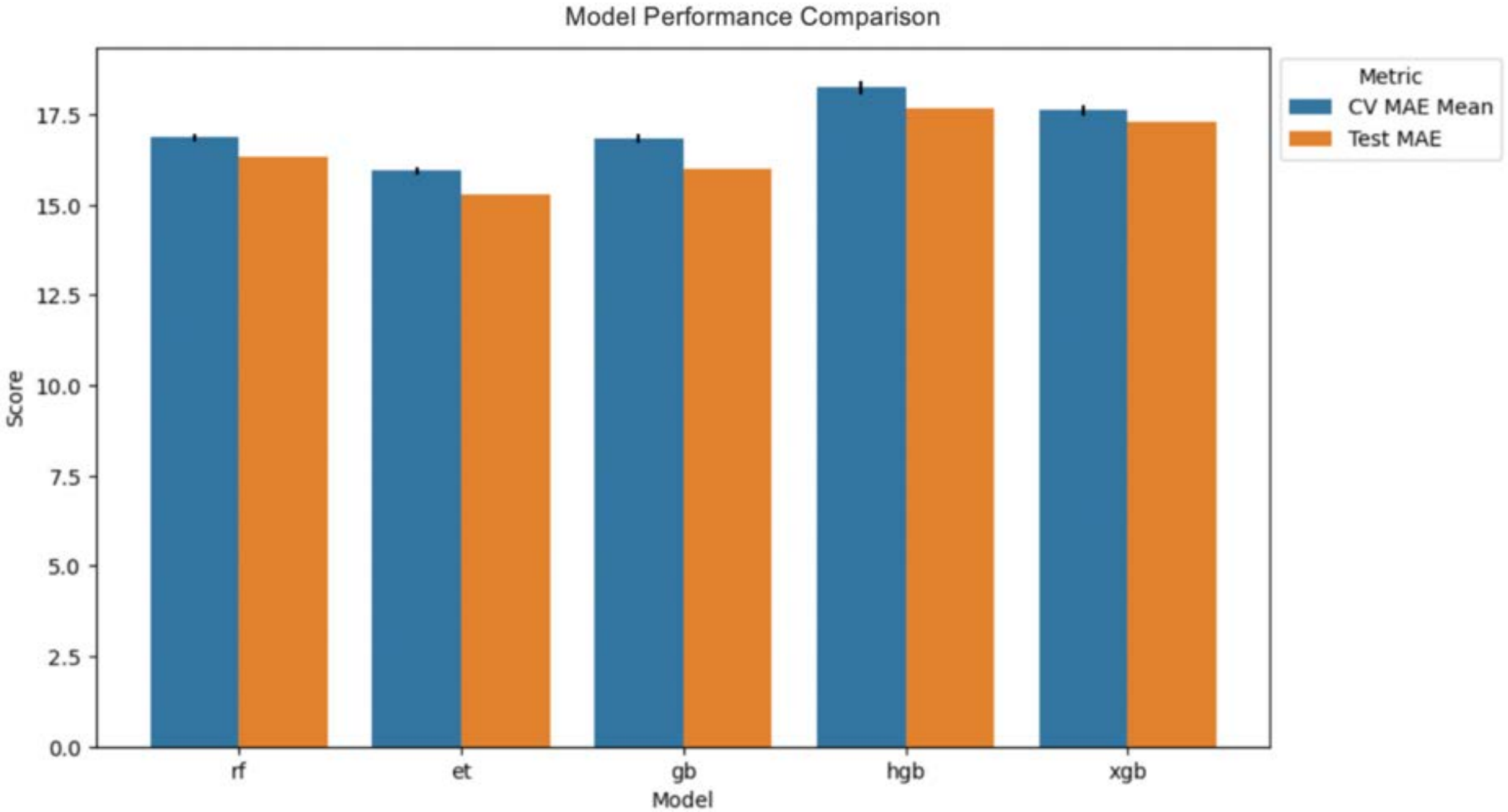
Extra Trees Regressor
- Power Transformer
- CV / Test MAE 15.94 / 15.30

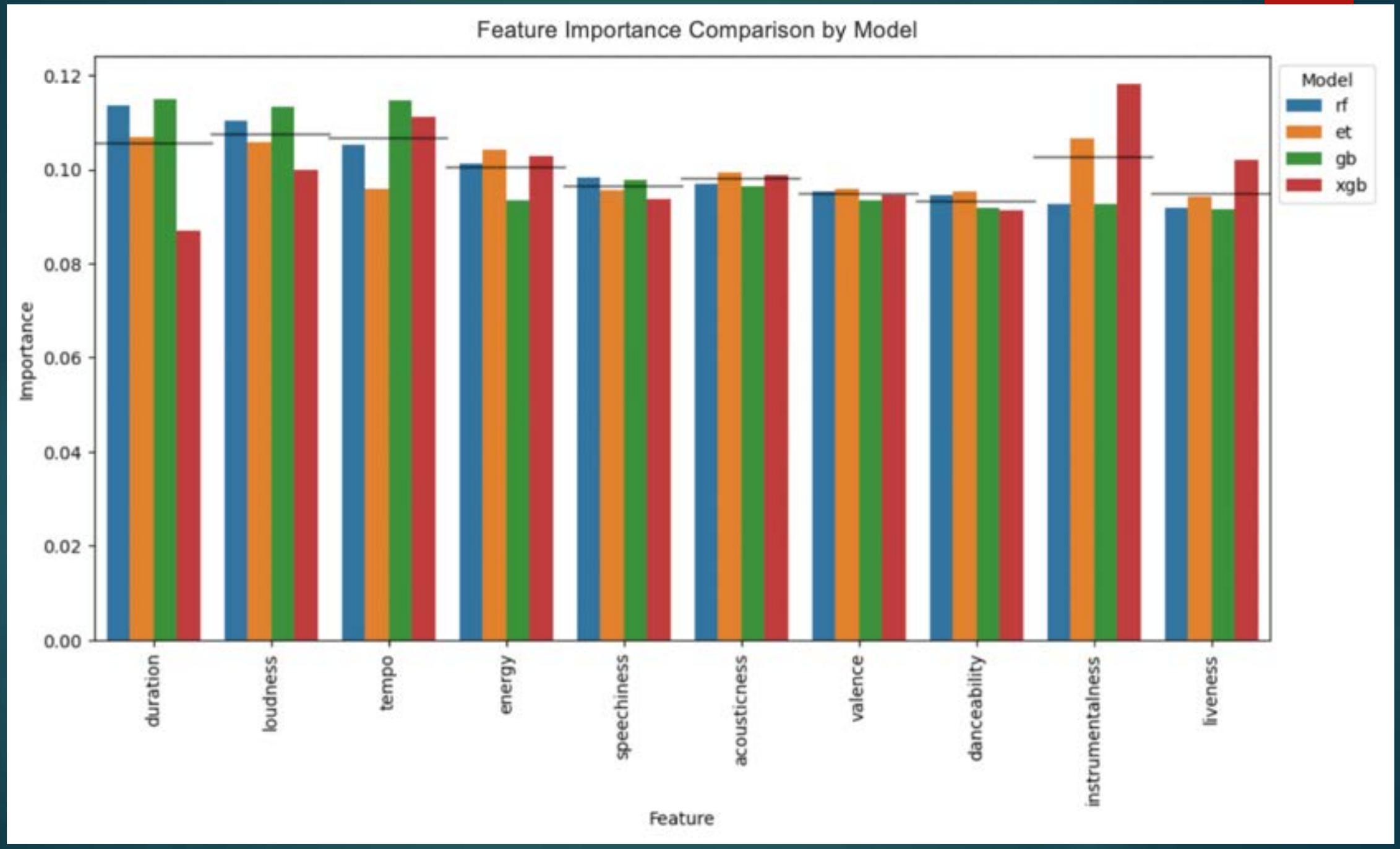
Gradient Boosting Regressor
- Min Max Scaler
- CV / Test MAE 16.83 / 16.00

Hist Gradient Boosting Regressor
- Robust Scaler
- CV / Test MAE 18.24 / 17.68

XGBoost Regressor
- Power Transformer
- CV / Test MAE 17.62 / 17.29

(80 / 20 train test split)


Model Performance Comparison
Metric
CV MAE Mean
Test MAE
Score
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
rf
et
gb
hgb
xgb
Model


Feature Importance Comparison by Model
Importance
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Model
rf
et
gb
xgb
duration
loudness
tempo
energy
speechiness
acousticness
valence
danceability
instrumentalness
liveness
Feature

Modeling Pt. 2

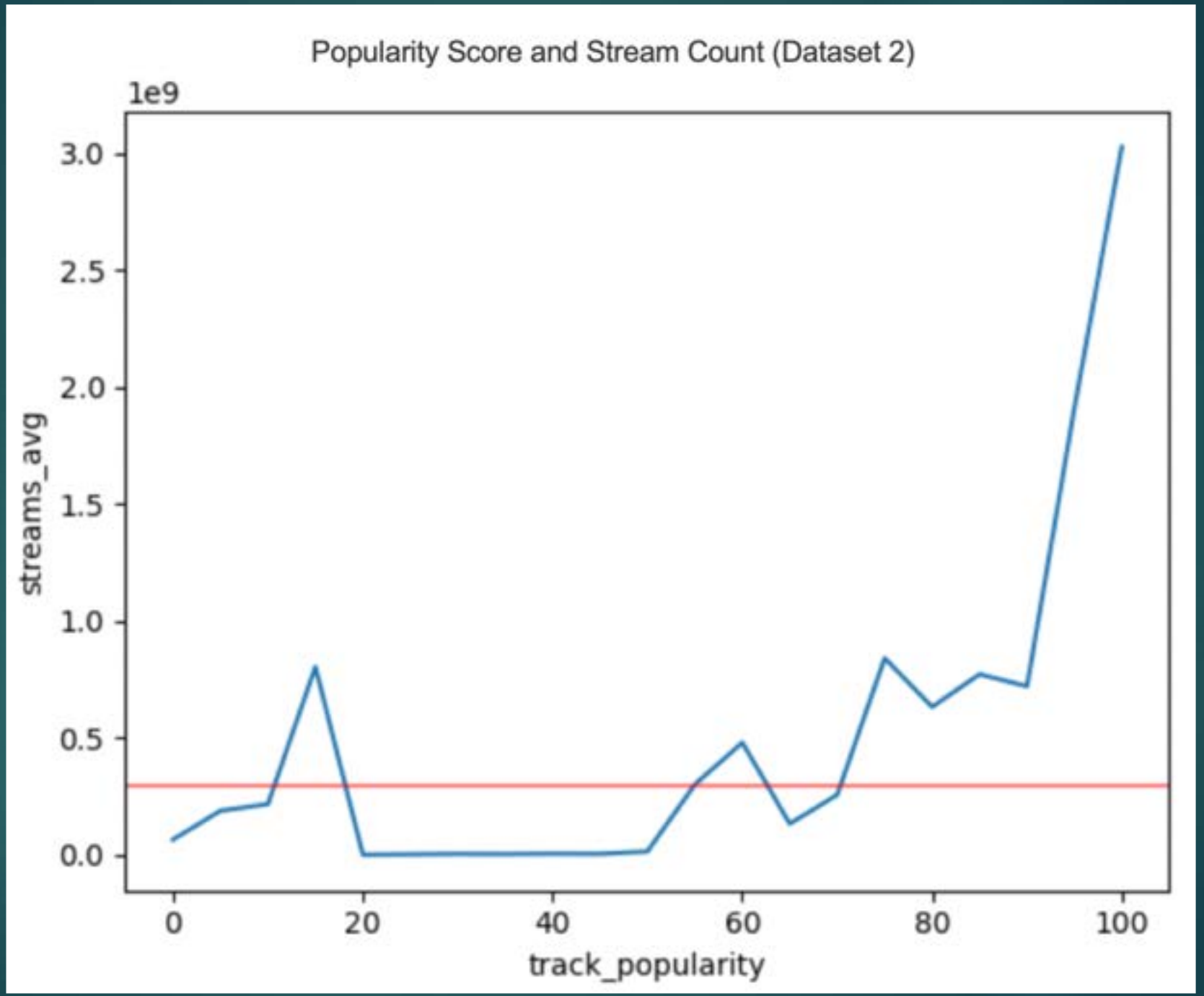
Extra Trees Regressor

Hyperparameters

- n_estimators = 1458
- max_depth = 46
- random_state = 42

Feature Importance

1. **Duration (0.106840)**
2. **Instrumentalness (0.106628)**
3. **Loudness (0.105952)**
4. **Energy (0.104153)**
5. Acousticness (0.099422)
6. Tempo (0.095905)
7. Valence (0.095804)
8. Speechiness (0.095739)
9. Danceability (0.095383)
10. Liveness (0.094174)


Popularity Score and Stream Count (Dataset 2)
1e9
3.0
2.5
2.0
1.5
1.0
0.5
0.0
streams_avg
0
20
40
60
80
100
track_popularity

Conclusion

Streams / Popularity Score

Random Forest Regressor (Set 1)

- Test MAE **268M**
- Tempo, Danceability, Acousticness, Valence

Extra Trees Regressor (Set 2)

- Test MAE **15.30**
- Duration, Instrumentalness, Loudness, Energy

Future Work

- Use current models for baseline predictions
- Seek dataset with thorough observations and features
- Quantify success metric