Song Popularity Prediction

Anton Reut s24382

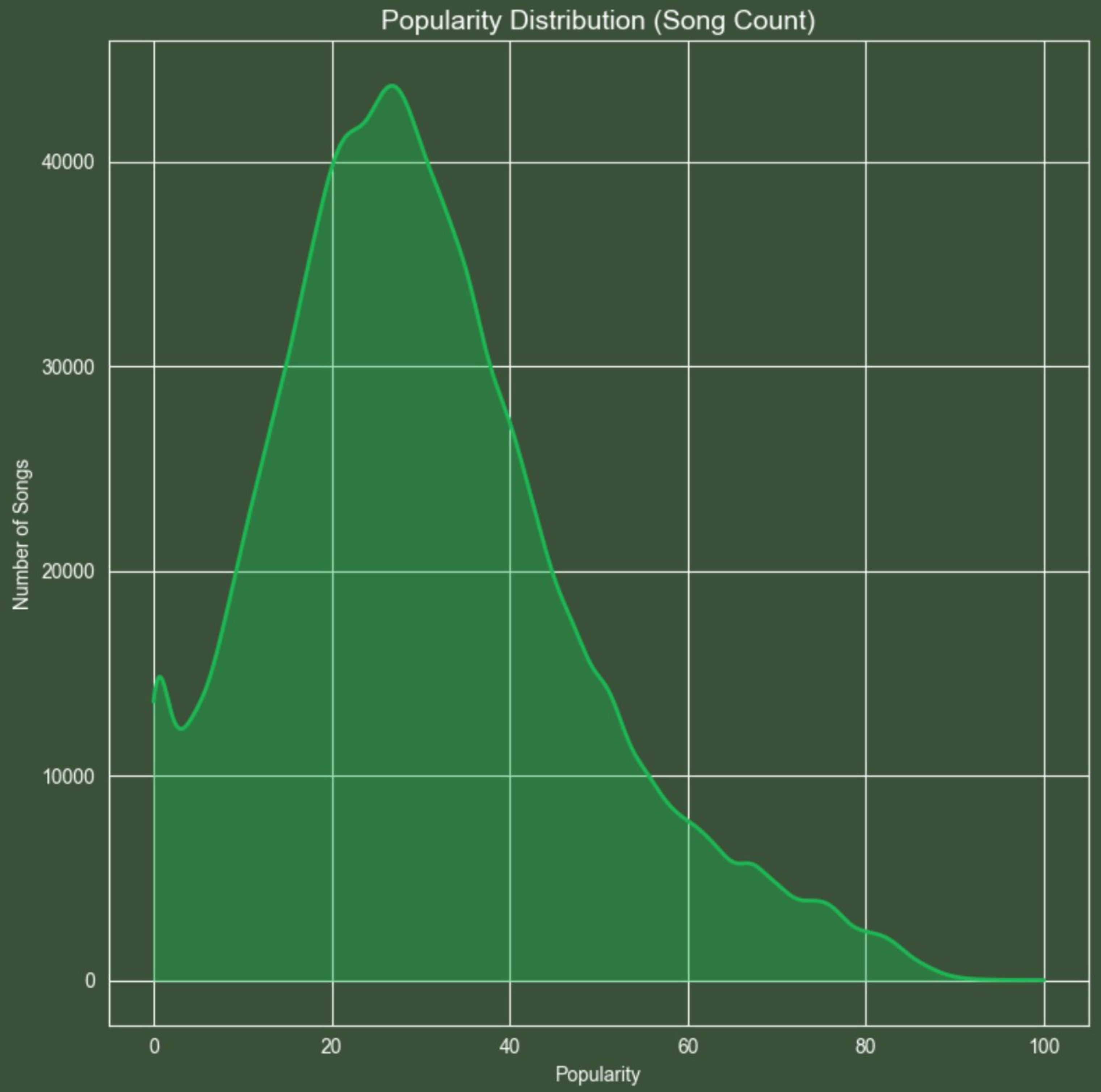
June 11 - 2025

Project EWD

Song Popularity Prediction
Anton Reut , s24382

Agenda

Objective



Agenda

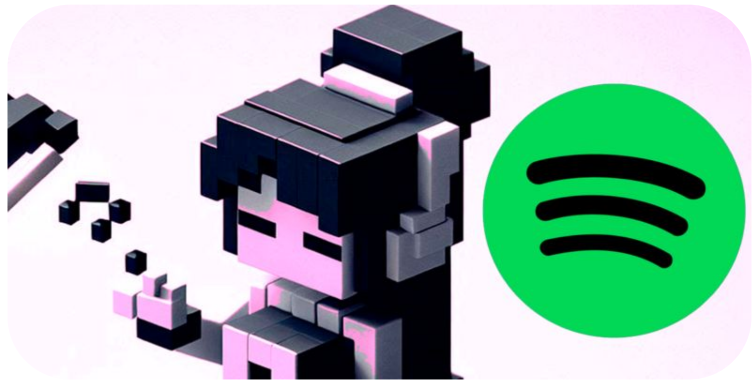
Goals

- Analyze **Emotional** & **Musical** Correlates of Song **Popularity**
- **Predict** Song Popularity
- Examine the **Evolution** of Music Over the Last **50 Years**

Challenges

- Popularity is Multifactorial
- Emotion Is Hard to Quantify
- Genres identifying and cross products
- Collinearity Between Features
- Ensuring Model Generalization

Data



Summary

- 3 files
  - {i} .json — 2
  - .csv — 1
- 39 columns
  - # Integer — 18
  - A String — 15
  - # Decimal — 3
  - Other — 3

🎧 500K+ Spotify Songs with Lyrics, Emotions & More

A Dataset for Music Recommendation and Emotion Analysis (500K+ Tracks)

final_milliondataset_BERT_500K_revised.json (1.64 GB)

This dataset was part of the **Top 200** projects in the **NVIDIA Llama-Index** Contest, supporting the Abracadabra project — a Retrieval-Augmented Generation (RAG) system for intelligent playlist creation using LLMs.

Over 30 features including:

- Popularity, Energy, Danceability, Speechiness, Tempo, Loudness, Key
- Acousticness, Instrumentalness, Time Signature
- Contextual tags (e.g., Good for Party, Relaxation, Study, Exercise, Driving, etc.)

3 similar songs per track (with artist, title, and similarity score)

[Link to the dataset]

Exploration

Good for Party = 0.34corr

Good for Social Gatherings = 0.19corr

Unfortunately the basic data does any columns that have correlation >0.20

Develop the Dataset further to cover more data (that currently are not numeric)

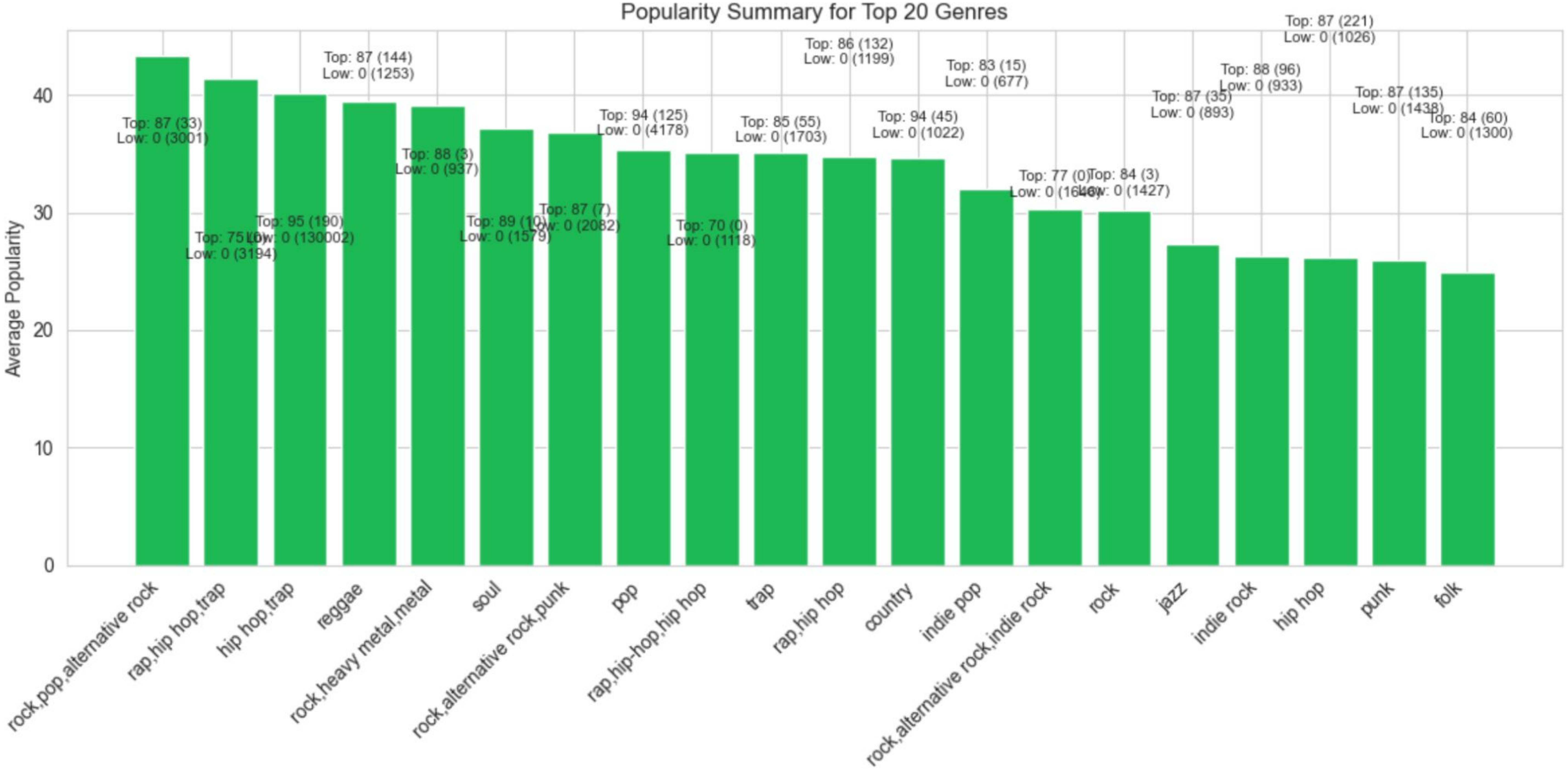
Feature Correlation Matrix

	Tempo	Loudness (db)	Popularity	Energy	Danceability	Positiveness	Speechiness	Liveness	Acousticness	Instrumentalness	Good for Party	Good for Work/Study	Good for Relaxation/Meditation	Good for Exercise	Good for Running	Good for Yoga/Stretching	Good for Driving	Good for Social Gatherings	Good for Morning Routine
Tempo	1.00	0.13	0.02	0.18	-0.11	0.04	0.01	0.01	-0.16	0.01	0.17	-0.32	-0.27	0.35	0.31	-0.22	-0.18	-0.02	-0.04
Loudness (db)	0.13	1.00	0.08	0.75	0.08	0.16	0.06	0.09	-0.55	-0.24	0.13	-0.37	-0.30	0.23	0.15	-0.31	-0.05	0.01	0.01
Popularity	0.02	0.08	1.00	0.05	0.07	0.07	-0.11	-0.05	-0.06	-0.03	0.34	-0.03	-0.02	0.05	0.03	-0.02	0.02	0.19	0.03
Energy	0.18	0.75	0.05	1.00	-0.02	0.26	0.11	0.18	-0.68	-0.06	0.15	-0.45	-0.33	0.31	0.20	-0.31	-0.12	-0.01	-0.02
Danceability	-0.11	0.08	0.07	-0.02	1.00	0.42	0.25	-0.12	-0.11	-0.17	0.11	-0.16	-0.16	0.22	0.07	-0.15	0.19	0.09	0.20
Positiveness	0.04	0.16	0.07	0.26	0.42	1.00	0.10	0.02	-0.13	-0.14	0.13	-0.17	-0.15	0.22	0.14	-0.14	0.03	0.04	0.05
Speechiness	0.01	0.06	-0.11	0.11	0.25	0.10	1.00	0.11	-0.09	-0.14	0.00	-0.16	-0.10	0.06	0.08	-0.09	0.03	-0.02	-0.04
Liveness	0.01	0.09	-0.05	0.18	-0.12	0.02	0.11	1.00	-0.07	-0.03	0.00	-0.06	-0.04	0.02	0.02	-0.04	-0.05	-0.02	-0.04
Acousticness	-0.16	-0.55	-0.06	-0.68	-0.11	-0.13	-0.09	-0.07	1.00	0.06	-0.11	0.50	0.35	-0.21	-0.11	0.31	-0.10	-0.02	-0.10
Instrumentalness	0.01	-0.24	-0.03	-0.06	-0.17	-0.14	-0.14	-0.03	0.06	1.00	-0.03	0.07	0.08	-0.04	-0.03	0.09	-0.03	-0.01	-0.02
Good for Party	0.17	0.13	0.34	0.15	0.11	0.13	0.00	0.00	-0.11	-0.03	1.00	-0.07	-0.04	0.49	0.26	-0.04	-0.06	0.06	-0.00
Good for Work/Study	-0.32	-0.37	-0.03	-0.45	-0.16	-0.17	-0.16	-0.06	0.50	0.07	-0.07	1.00	0.62	-0.14	-0.07	0.51	-0.07	-0.03	-0.08
Good for Relaxation/Meditation	-0.27	-0.30	-0.02	-0.33	-0.16	-0.15	-0.10	-0.04	0.35	0.08	-0.04	0.62	1.00	-0.09	-0.04	0.83	-0.04	-0.02	-0.05
Good for Exercise	0.35	0.23	0.05	0.31	0.22	0.22	0.06	0.02	-0.21	-0.04	0.49	-0.14	-0.09	1.00	0.50	-0.07	-0.11	-0.00	-0.01
Good for Running	0.31	0.15	0.03	0.20	0.07	0.14	0.08	0.02	-0.11	-0.03	0.26	-0.07	-0.04	0.50	1.00	-0.04	-0.06	-0.02	-0.06
Good for Yoga/Stretching	-0.22	-0.31	-0.02	-0.31	-0.15	-0.14	-0.09	-0.04	0.31	0.09	-0.04	0.51	0.83	-0.07	-0.04	1.00	-0.04	-0.01	-0.04
Good for Driving	-0.18	-0.05	0.02	-0.12	0.19	0.03	0.03	-0.05	-0.10	-0.03	-0.06	-0.07	-0.04	-0.11	-0.06	-0.04	1.00	0.06	0.18
Good for Social Gatherings	-0.02	0.01	0.19	-0.01	0.09	0.04	-0.02	-0.02	-0.02	-0.01	0.06	-0.03	-0.02	-0.00	-0.02	-0.01	0.06	1.00	0.32
Good for Morning Routine	-0.04	0.01	0.03	-0.02	0.20	0.05	-0.04	-0.04	-0.10	-0.02	-0.00	-0.08	-0.05	-0.01	-0.06	-0.04	0.18	0.32	1.00

Agenda

Exploration

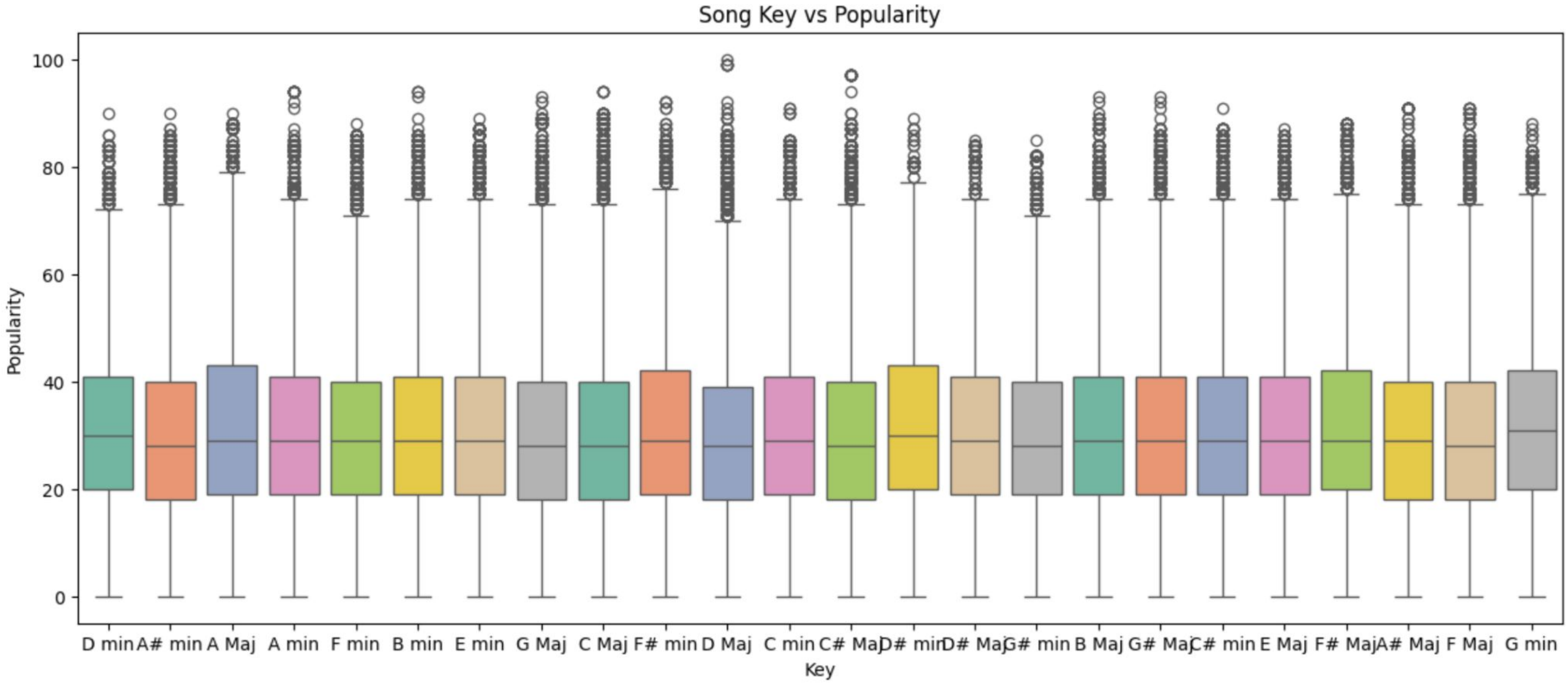
Genre should be a good determination point in **predicting song popularity**



Agenda

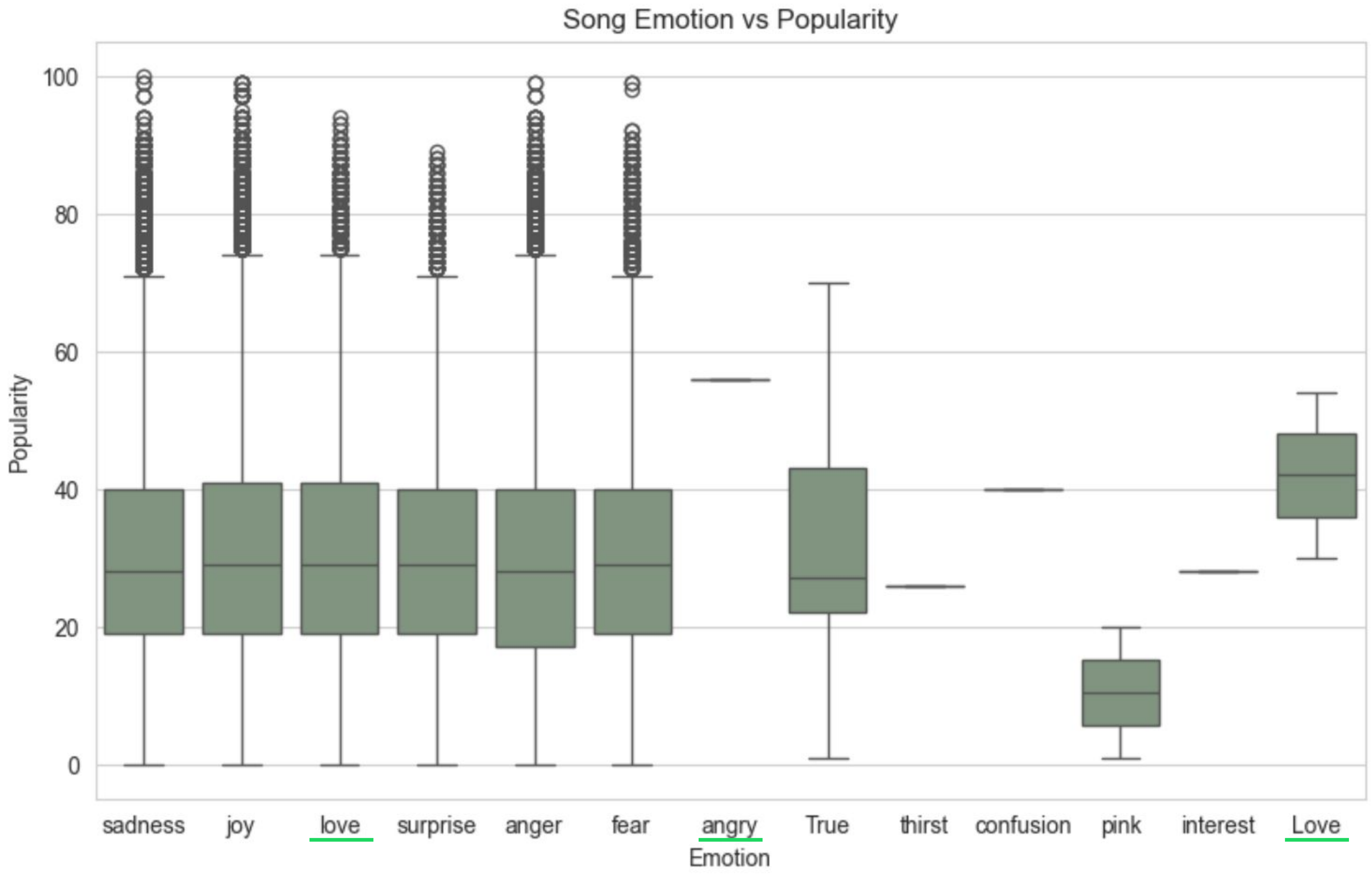
Exploration

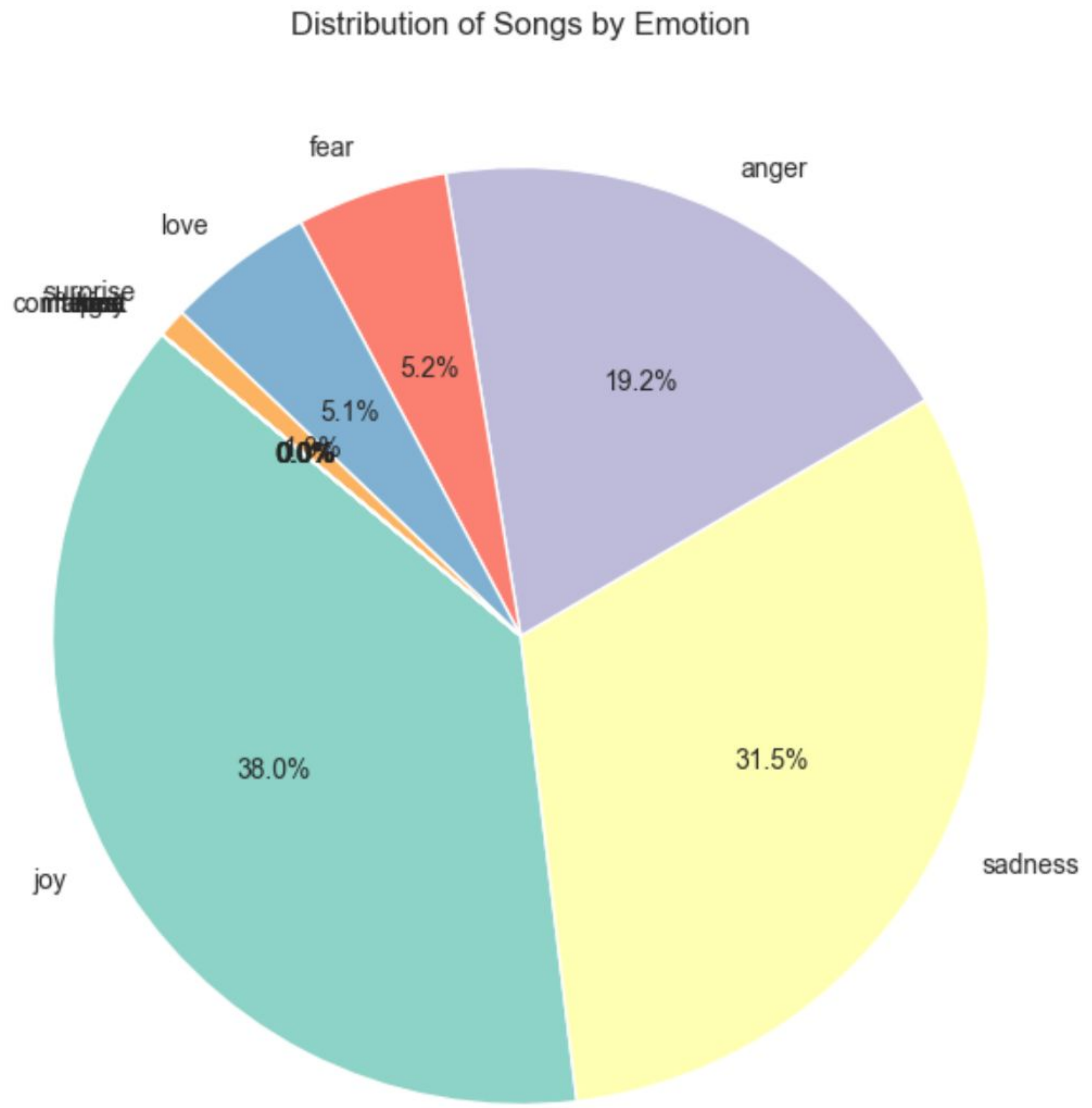
Song key does not give strong insights on song popularity



Agenda

Exploration



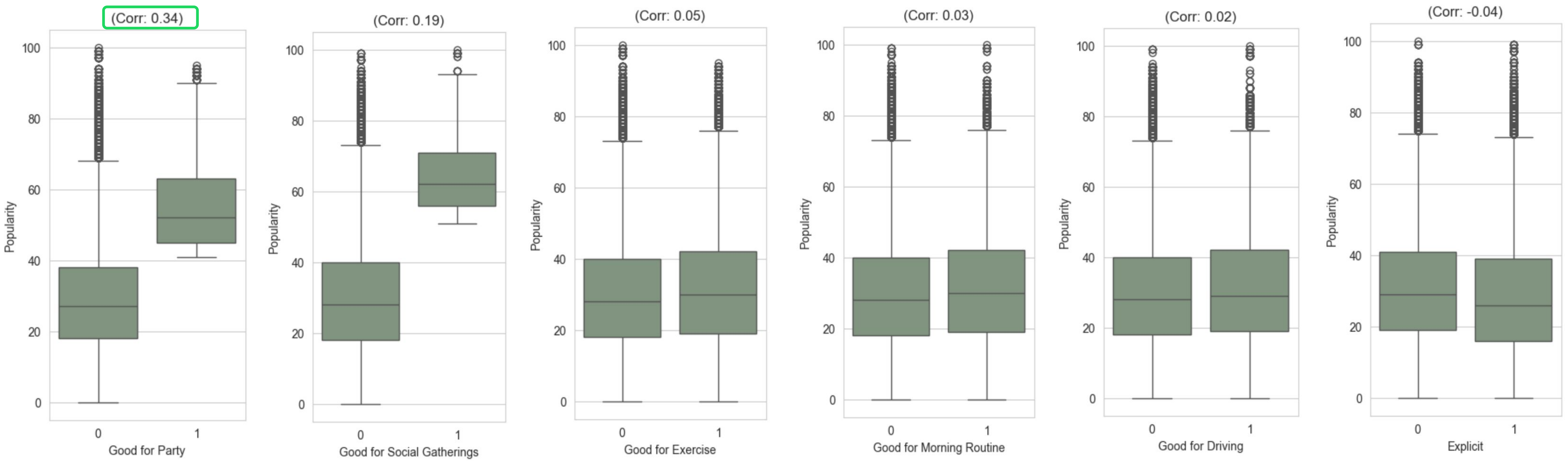


Agenda

Exploration

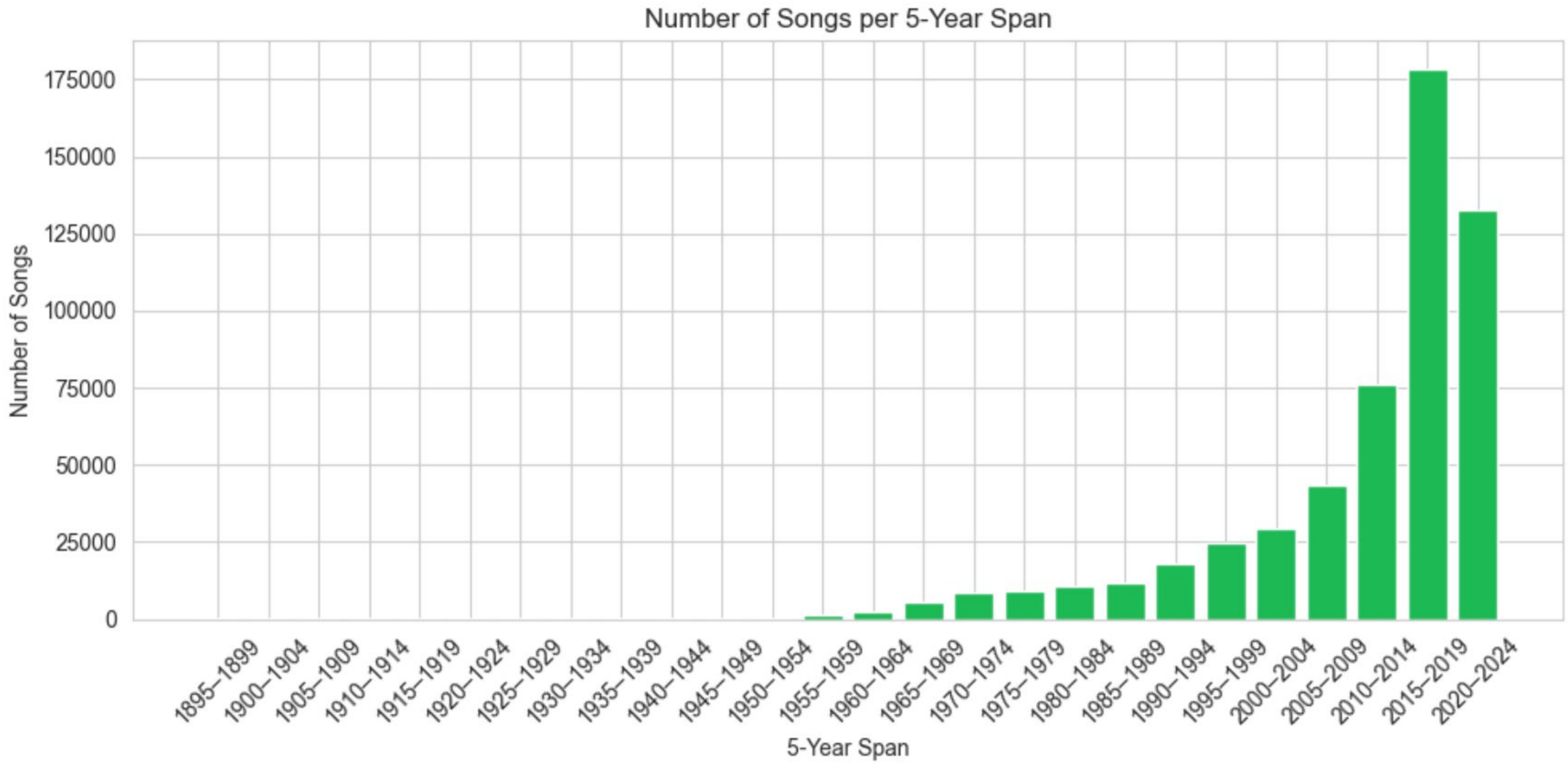
Song Popularity Prediction
Anton Reut , s24382

Boolean features with their correlation on Popularity



Agenda

Music Evolution

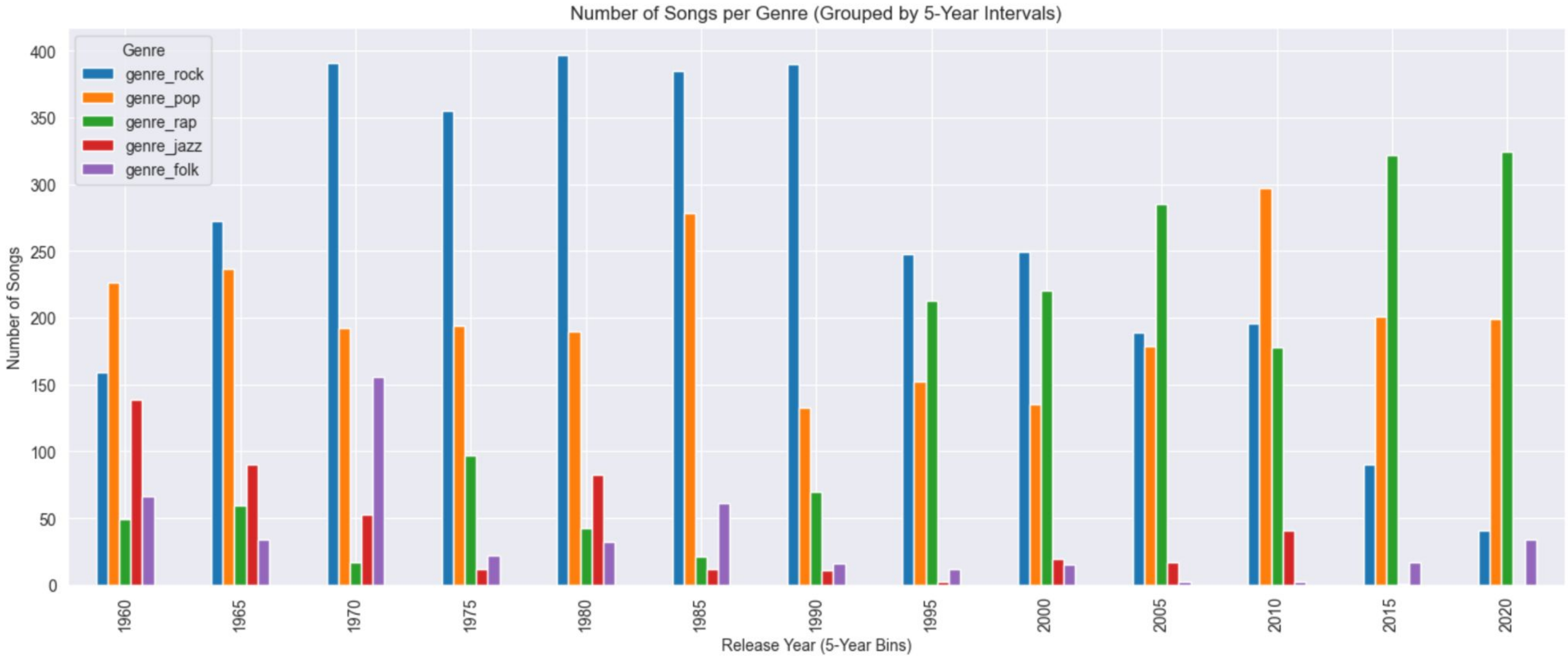


When exploring song evolution we have to account that current dataset has more recent songs.

Agenda

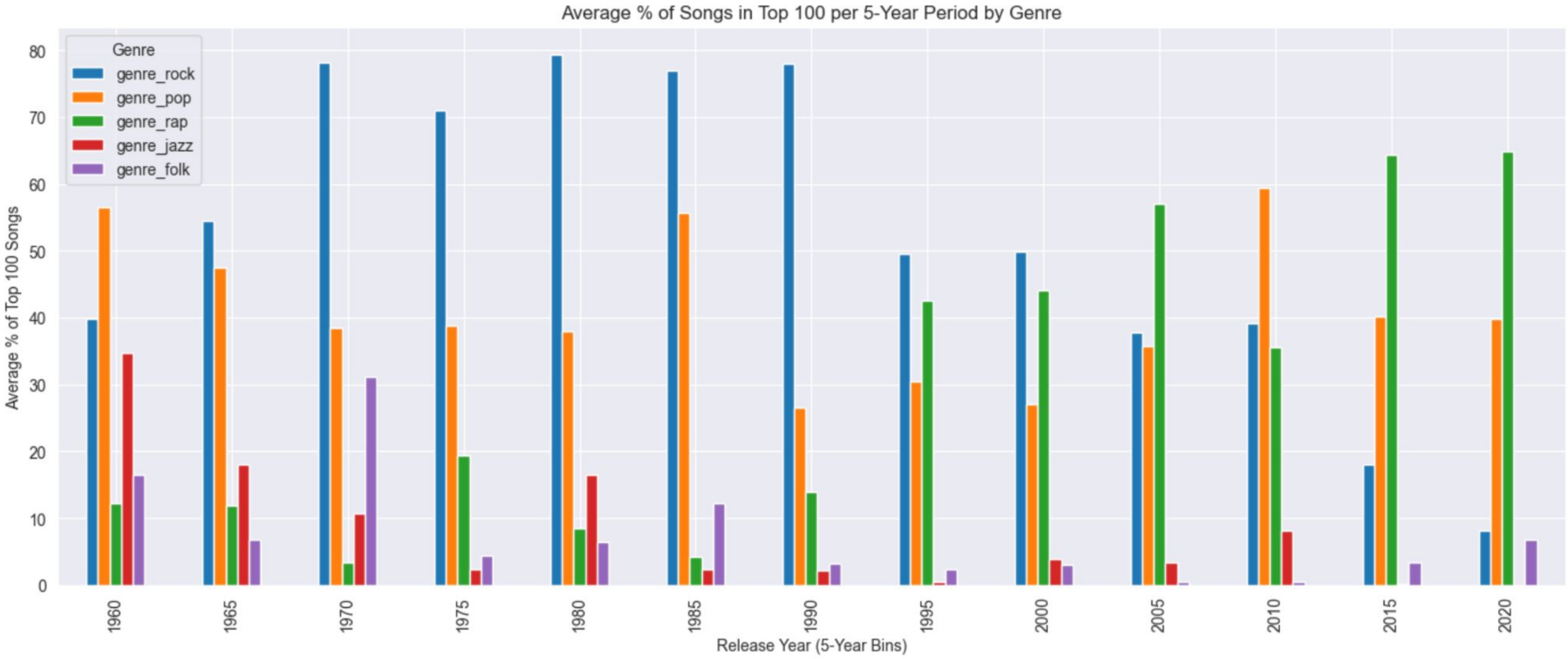
Music Evolution

Bear in mind one song can be in multiple genres (ex. “rock, pop, indie rock”)



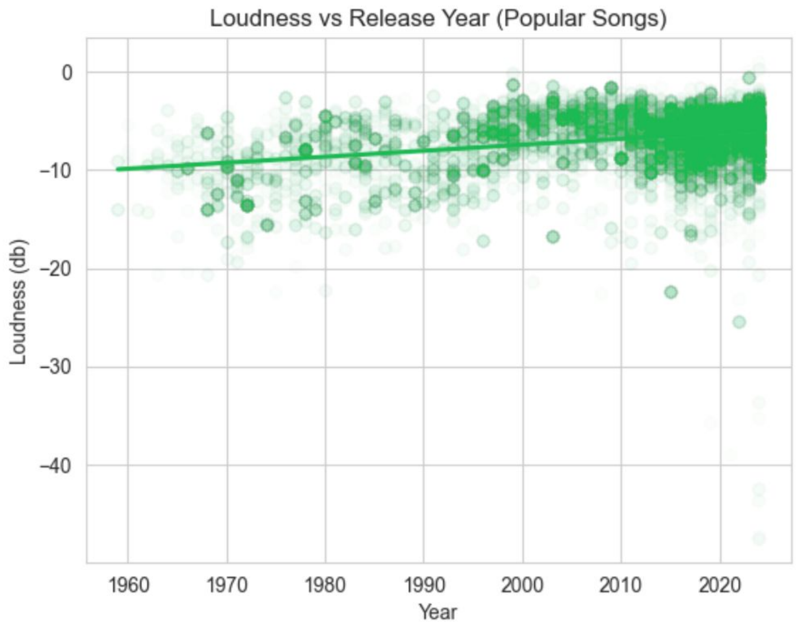
Music Evolution

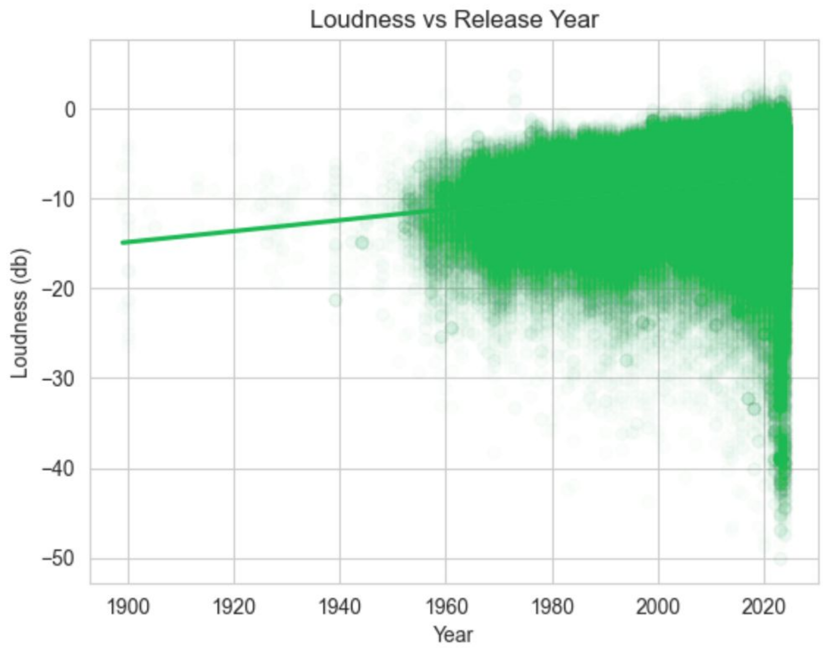
Bear in mind one song can be in multiple genres (ex. “rock, pop, indie rock”)



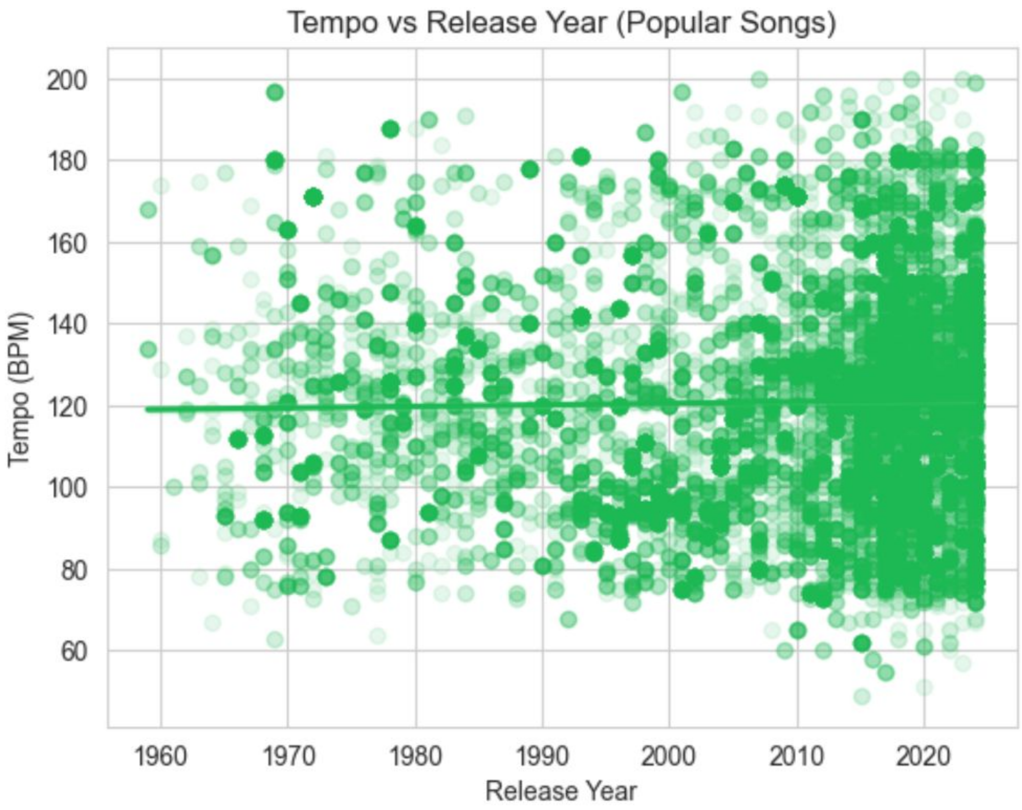
Agenda

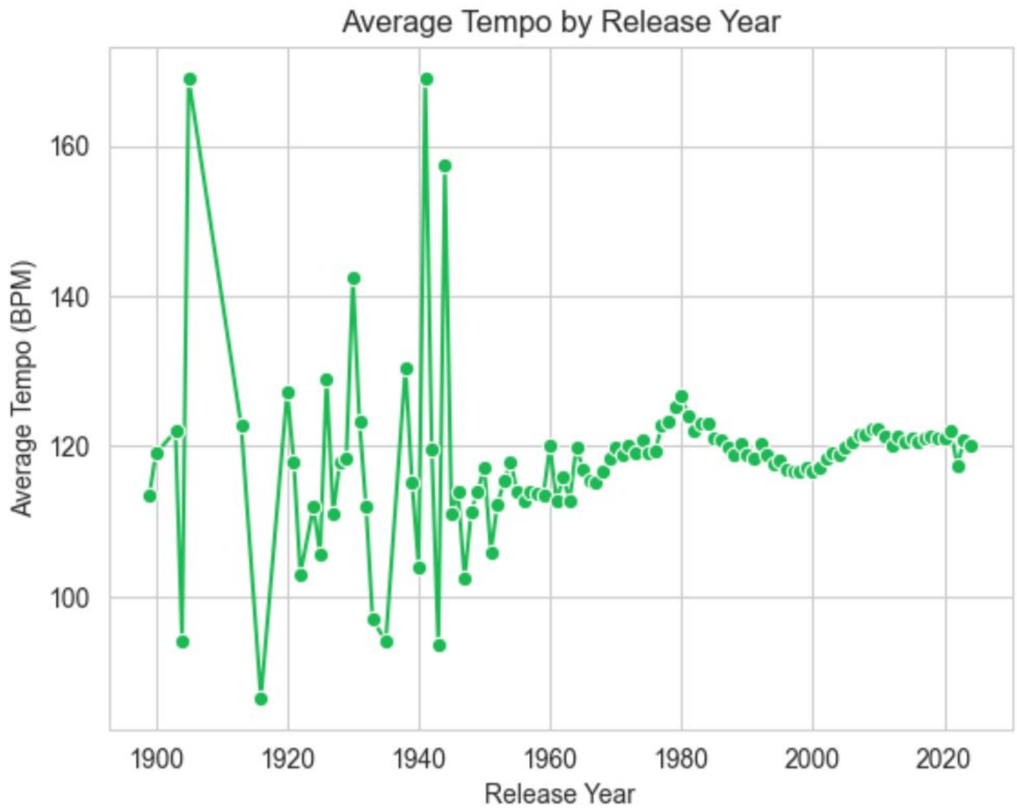
Music Evolution

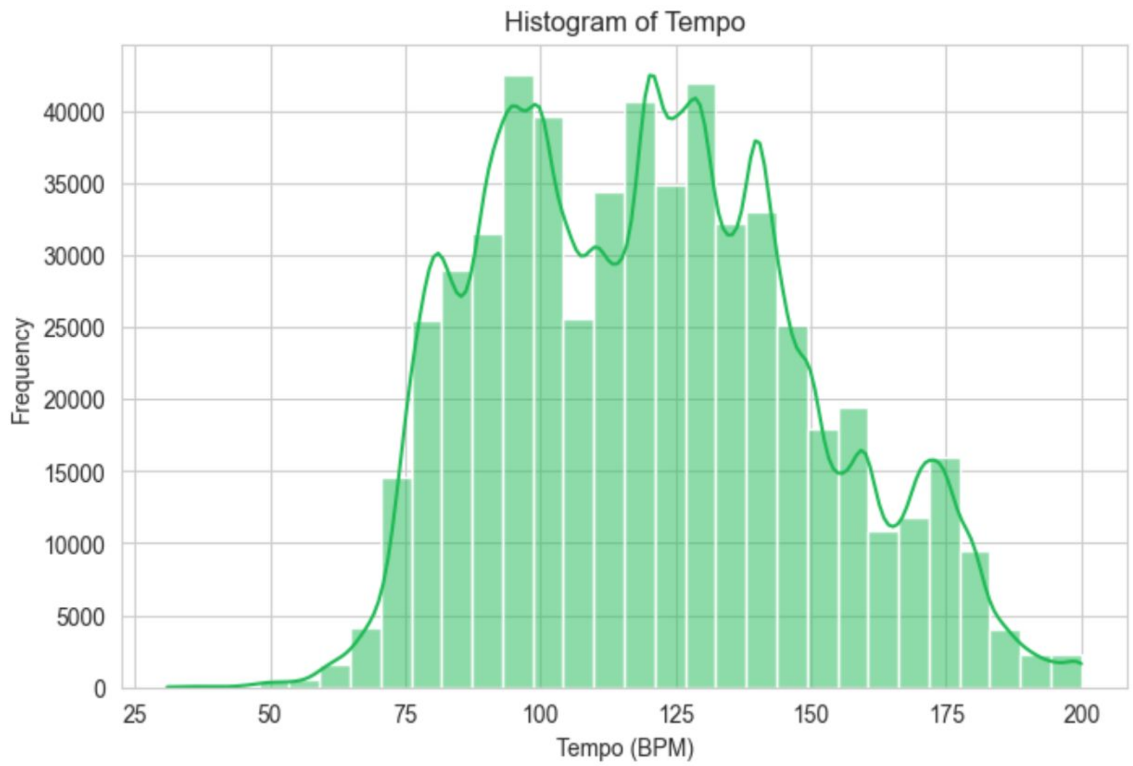




- Loudness is clearly increasing over the years
- Tempo is slowly increasing, but the average stays around 120 BPM
- Average tempo high variance before 1960 is related to the number of songs in dataset per year

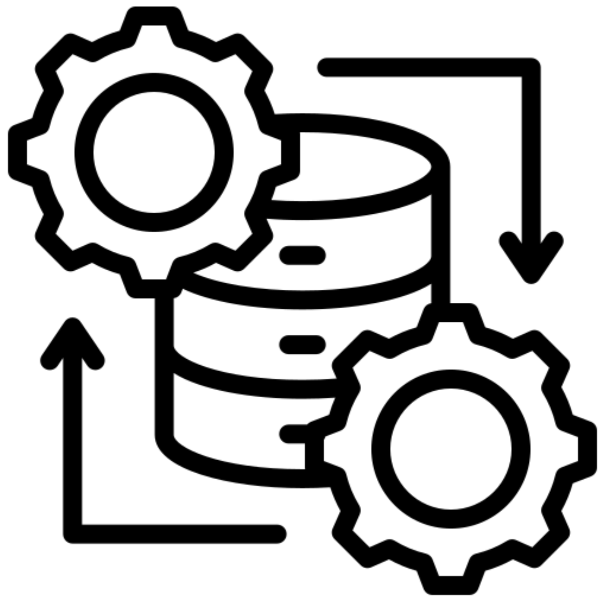






Agenda

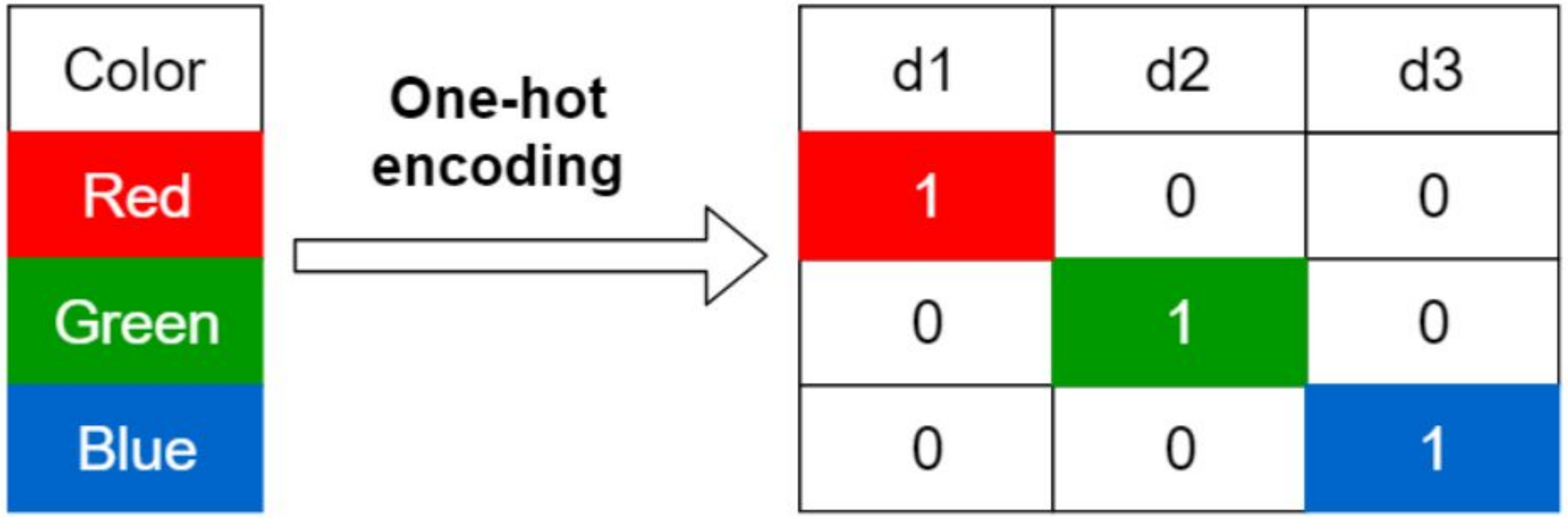
Preparation



1. Drop columns ['Artist(s)', 'song', 'Album','Similar Songs', 'ISRC']
2. Finding and fixing outliers, like emotion: ‘love’ and ‘Love’
3. Length from string into integer '1:23' → 83 (seconds)
4. Loudness from string into float '-6.5db' → -6.5
5. Date from string into Year, Month, Age '29th April 2013' → Year:2013, Month:4, Age:12
6. Dropping NA columns

Agenda

Preparation



String 15

Because of using Regressors One-hot encoding needed to be performed:

1. **Emotion** from string into **Binary** (emotion_sadness (0:1), emotion_love(0:1)
2. **Explicit** from string into **Binary**
3. **Key** from string into **Binary**
4. New column **Major**, that depending on Key (maj or min) in **Binary**
5. Rhythm signature from string into **Binary** `'1/4' → ¼:1, ¾:0`

Agenda

Preparation

Genre is one of the most **important column** in the dataset, so it needed to be transformed carefully.
Yet **big dimensionality** can also be a **problem.**

```
Genre:
'pop,indie rock,
alternative '
'cloud rap'
    83 types
```

Aggregating data

```
Grouped Genre:
'pop, rock'
'rap, others'
```

One-hot encoding

genre_others	genre_pop	genre_rap
0	0	1
0	0	1
0	0	1

8 columns

Agenda

Preparation

TF IDF vectorising words

chorus	0.060455
yeah	0.057695
pre	0.046918
ooh	0.046350
outro	0.044201
uh	0.043697
oh	0.040049
intro	0.038568
verse	0.032214
bridge	0.025563

1st iteration

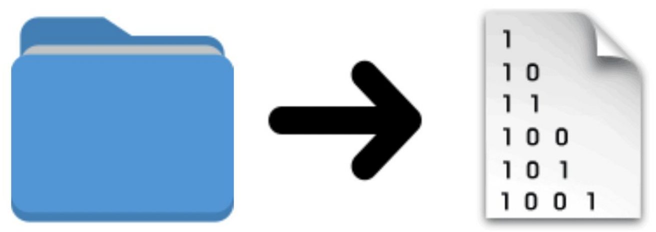
Text replacing

yeah	0.058190
ooh	0.046422
uh	0.043672
oh	0.040589
ah	0.023297
eyes	-0.017430
life	-0.018260
people	-0.018535
left	-0.020057
man	-0.020153
hook	-0.052610

2nd iteration

One-hot encoding

top 3 on each side



Preparation

1. Language detection using `langdetect` library

Cell execution finished in 1h 35m
View cell

2. Is English encoding

```
Starting language detection...
Language detection complete.

Distribution of detected languages:
Language
en    549005
tl       439
so       379
id       334
cy       229
```

Language	English
en	1
en	1
en	1
en	1
en	1

Modeling and Evaluation

Main parameters for evaluation

- R2 - helps understand how well your model captures the underlying patterns in data, beyond just prediction error
- MAE - gives a clear sense of the typical error size
- MSE - is commonly used in regression tasks because it is differentiable, is Sensitive to outliers

Pipeline

1. Get balanced probe
2. Train different models
3. Train with preprocessing Scaling
4. Train with Cross Validation

Agenda

Modeling and Evaluation

First iterations of models before genre grouping, TF IDF and dataset balancing

	Model	MAE	MSE	R2
0	LinearRegression	12.579228244054006	257.7972771095773	0.2255186672733096
1	RandomForest	8.775657394047617	138.08533911189792	0.5851604071832663
2	XGBoost	10.430037498474121	178.22996520996094	0.46455687284469604

After genre grouping, TF IDF and dataset balancing

	Model	MAE	MSE	R2
0	LinearRegression	16.423171085242686	391.90933906203753	0.3024577228075598
1	RandomForest	9.029301224899598	168.52480157511044	0.7000500826659337
2	XGBoost	11.398643493652344	221.57508850097656	0.6056281924247742

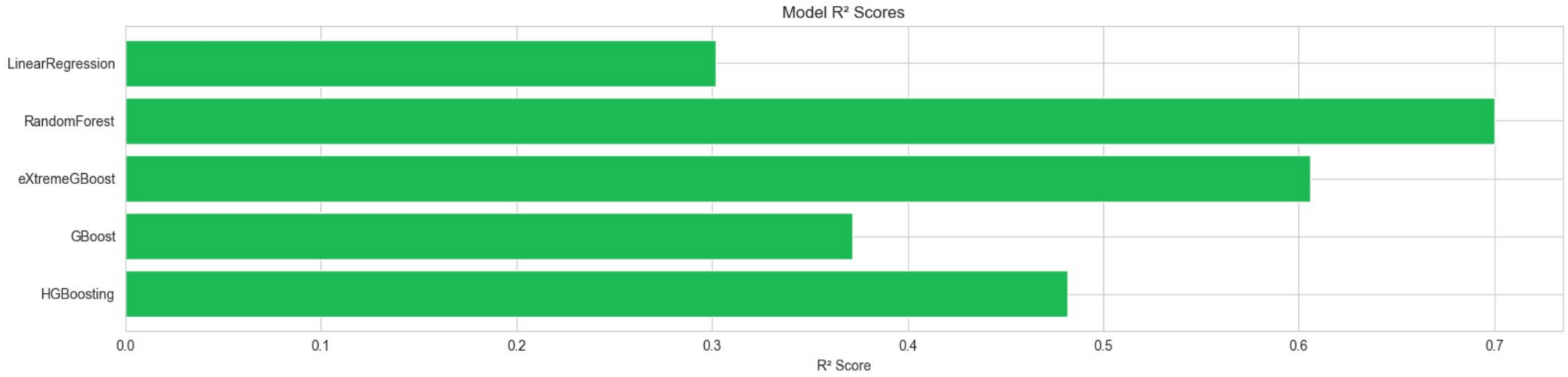
After Scaling Features

	Model	MAE	MSE	R2
0	LinearRegression	16.423171085242725	391.9093390620383	0.30245772280755834
1	RandomForest	9.024775120481927	168.3887767187517	0.7002921873534269
2	eXtremeGBoost	11.398643493652344	221.57508850097656	0.6056281924247742
3	GBoost	15.178872903759737	352.96343441357345	0.3717758336768415
4	HGBoosting	13.162735272293684	291.16213112852716	0.4817732681092164

Agenda

Modeling and Evaluation

After Standard Scaling



After RandomizedSearchCV

```
R²: 0.6989
MSE: 169.1511
MAE: 9.1237
```

```
rf = RandomForestRegressor(random_state=42)
search = RandomizedSearchCV(
    rf,
    param_grid,
    n_iter=5,  # Lowered from 20
    cv=4,       # Lowered from 5 if applicable
    scoring='neg_mean_squared_error'
)
✓ [14] 26m 53s
```

```
Feature importances (sorted):
Good for Party: 0.1188
Loudness: 0.0838
Danceability: 0.0600
Length: 0.0573
Good for Exercise: 0.0555
Energy: 0.0466
Word count: 0.0464
Good for Social Gatherings: 0.0453
Positiveness: 0.0438
```

Results are worse due to Overfitting to Validation Data and launching it on full dataset can take more than 9 hours of fitting.
Data still has some noise due to human nature and emotions.

Agenda

Conclusions

Critical and Important Features

1) **Genre**
   a) Popular genres (Pop, Rap, Rock) correlate strongly with popularity.
   b) Rare genres (Jazz, Classical) often indicate unpopularity.
   c) Recommendation: Use one-hot encoding for genres if rare subgenres are informative, but group them, to avoid dimensionality is an issue.
2) **Audio Features**
   a) High energy, danceability, and loudness predict popularity.
   b) Low acousticness and instrumentalness are also linked to popularity.
3) Songs **become popular at parties** as large crowds listen to them.
   a) Good for Parties
   b) Good for Social Gatherings
4) **Song Time Signature**
   a) Most people like rather simple 4/4 time signature then else

Agenda

Conclusions

Music popularity is still too **hard to precisely** measure and predict due to its emotional nature, yet it can be at least **estimated** based on the **song details**.

Agenda

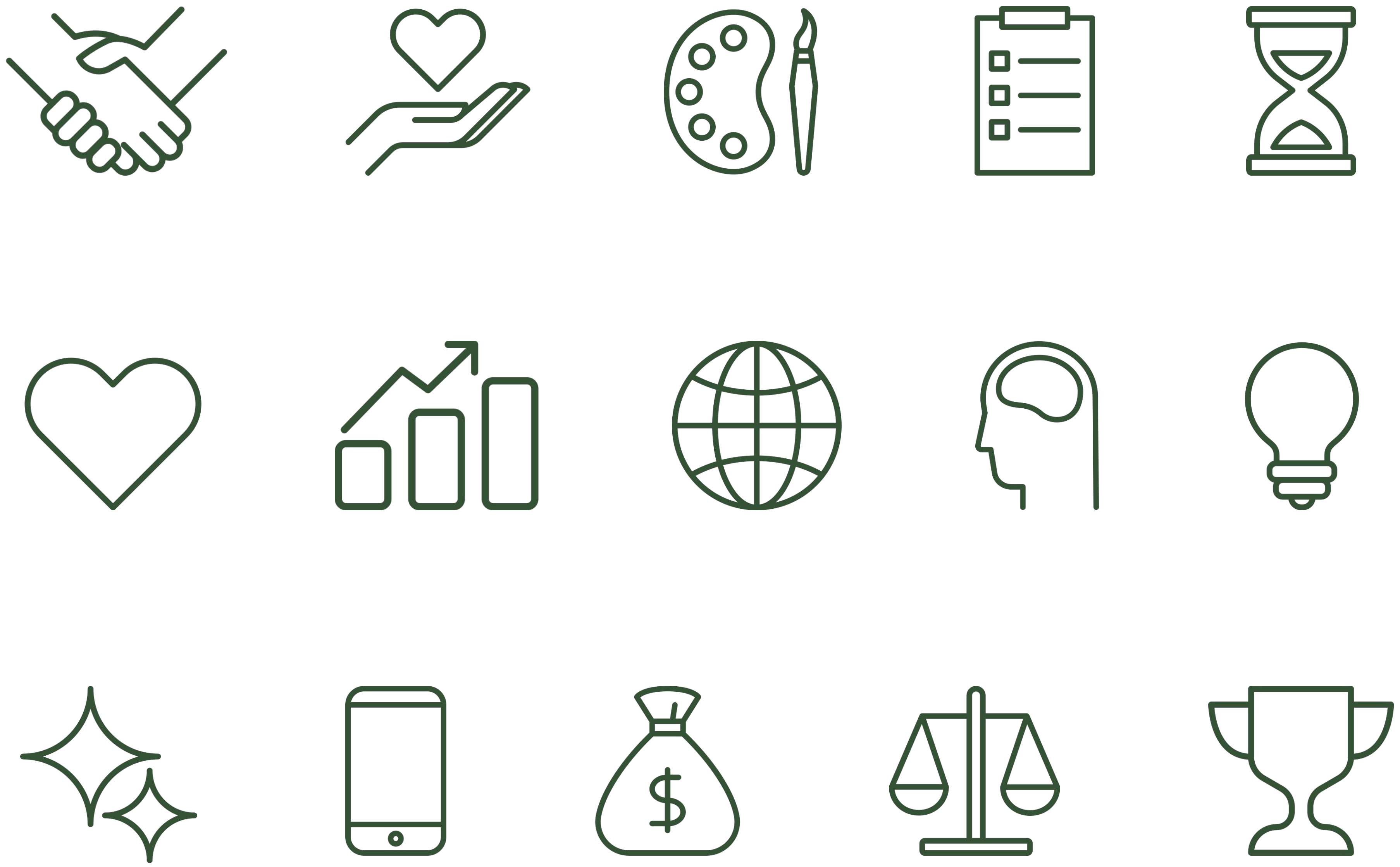
Thank you for your attention

Conclusions

Music popularity is still too **hard to precisely** measure and predict due to its emotional nature, yet it can be at least **estimated** based on the **song details**.

Agenda

Elementy graficzne



Użyj tych elementów w swojej prezentacji Canva. Udanego projektowania! Pamiętaj, aby usunąć lub ukryć tę stronę, zanim wyświetlisz prezentację.