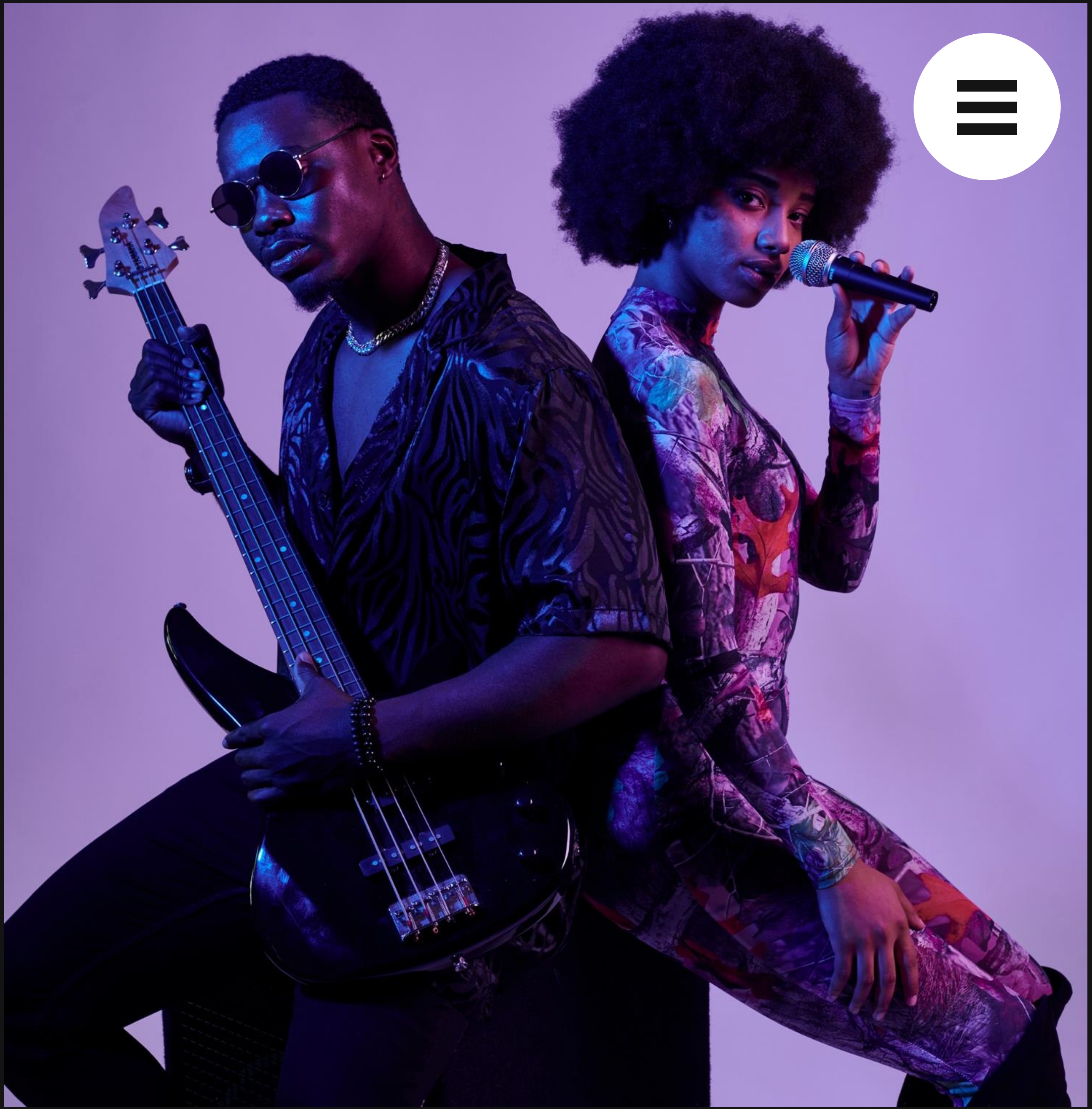
PREDICTING SONG POPULARITY

Jonah Lee

Rajiha Mehdi

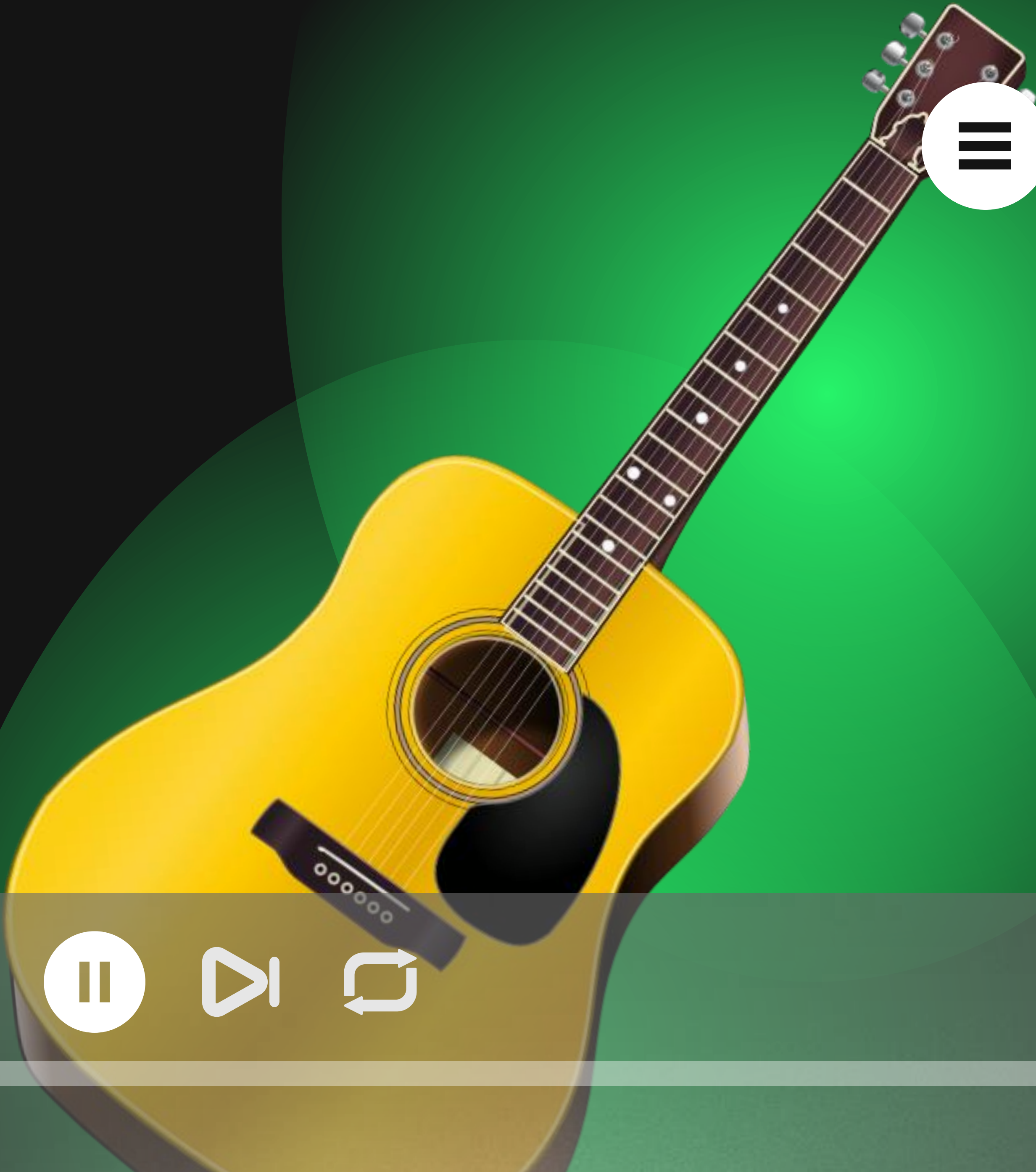
Clement Orcibal

Henry Tang



AGENDA

- Introduction
- Dataset & Preprocessing
- Insights
- Predictive Model
- Evaluation of Model
- Next Steps



Executive Summary

OBJECTIVE

DEVELOP GUIDELINES FOR MUSIC CREATORS & PROMOTORS + PREDICT POPULARITY FOR NEW SONGS

DELIVERABLES

1. INSIGHTS ON FEATURES THAT DRIVE POPULARITY
2. GENRE-SPECIFIC INSIGHTS FOR TOP GENRES
3. PREDICTIVE MODEL TO PREDICT POPULARITY

ACTIONABLE INSIGHTS

UNIVERSAL TRENDS

- **Top 5 Drivers**: Accousticness, instrumentalness, duration, loudness, energy
- Lower acousticness universally appeals to audiences
- Higher energy and moderate loudness positively correlate with popularity
- Preference for shorter duration – experiment with shorter duration to align with streaming trends

GENRE-SPECIFIC

- Genre with highest popularity: pop-film, k-pop, chill
- Pop-Film: Energy & duration are the most critical factors, suggesting a preference for upbeat and medium-length tracks.
- K-Pop: Explicit content, loudness, & danceability most critical. Explicit content: strong negative impact.
- Chill: Speechiness and liveness dominate, indicating an appeal for conversational, live-like qualities.

PREDICTIVE MODEL CREATED

LASSO USED TO FEATURE SELECT, AND THEN RANDOM FOREST USED FOR PREDICTING POPULARITY

Final Output: R^2 : 44.76% | MSE: 208.97

Model is effective for general trend analysis & actionable insights for feature optimization. Room to improve predictive power

Key Stakeholders

WHO?	WHY?
Record Labels/Producers/ Studios	Gain insights to produce music that aligns with evolving listener preferences, maximizing hit potential
Marketing & Advertising Agencies	Optimize promotional efforts by focusing on songs with higher chances of commercial success
Streaming Platforms	Improve recommendation systems by predicting future hits and enhancing user engagement
Artists	Understand what features make a song popular, guiding creative decisions toward higher chart potential

INTRODUCTION

PROJECT AIM

KEY STAKEHOLDERS

Background:
The Evolution of Music and the Digital Age

Sonically

Changes in sound and emerging genres can be traced back to shifts in culture and identity.

Technically

Shift from analog – vinyl records, CD's, cassette tapes, to radio, and now to the digital streaming era, which has made listening to songs more accessible than ever before.

Democratized Music Industry

The Opportunity

Easier than ever for artists to accumulate revenue through streams as opposed to direct sales of digital or physical copies of their music.

The Challenge

Oversaturation of the music market and the difficulty for artists to stand out considering how much content is getting released.

PROJECT AIM

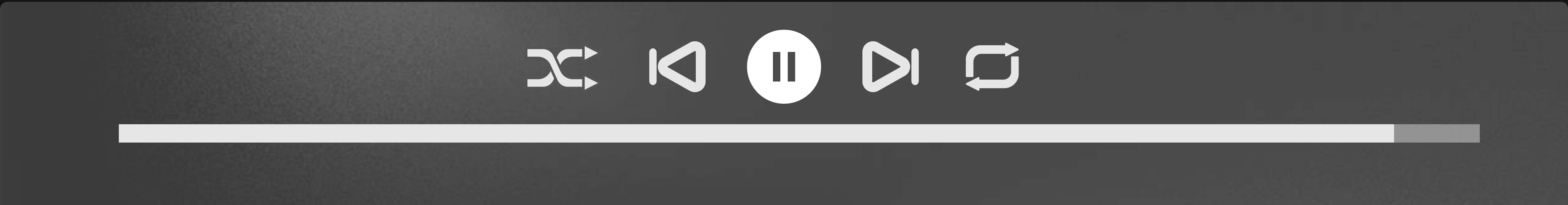
TWO-PRONGED SUPPORT FOR
MUSIC CREATORS, STUDIOS, & MARKETERS

#1
What Makes a Song Popular?
Identify key features that determine popularity of a song

hello@reallygreatsite.com
www.reallygreatsite.com

#2
Will 'this' be popular?
Predict likelihood of a song being popular

123 Anywhere St., Any City, ST 12345



Unique Value Proposition

	Commercial Focus	Song Popularity Prediction
Current/ Market Gap	• User-centric with limited use for marketing/commercial teams	• Based on user behavior • No pre-release insight
Our Project	• Focuses on marketing • Predicts Hit Potential • Can be used by artists to improve song	• Predicts based on features • Before user-data

Stakeholder Benefits

Our analysis will benefit three major groups within the music industry:

1. Artists and Record Labels

Optimized music production: knowing what features are associated with popularity can aid in guiding the creative process for artists

Resource allocation: prioritizing the enhancement of certain song features over others to save on unnecessary production costs

Selection of collaborators: understanding what features make a song popular can help artists choose certain producers or other collaborators known for a specific 'sound' to enhance its appeal

Stakeholder Benefits

2. Streaming Platforms

Enhancing algorithms:
By prioritizing songs with popular attributes, listener satisfaction is expected to increase

Improved identification of trends:
Can use feature selection as a baseline for understanding if the musical landscape is shifting in a new direction

3. Advertisers

Strategic music selection: Advertisers can choose songs expected to be popular as background tracks and sign partnerships with these artists

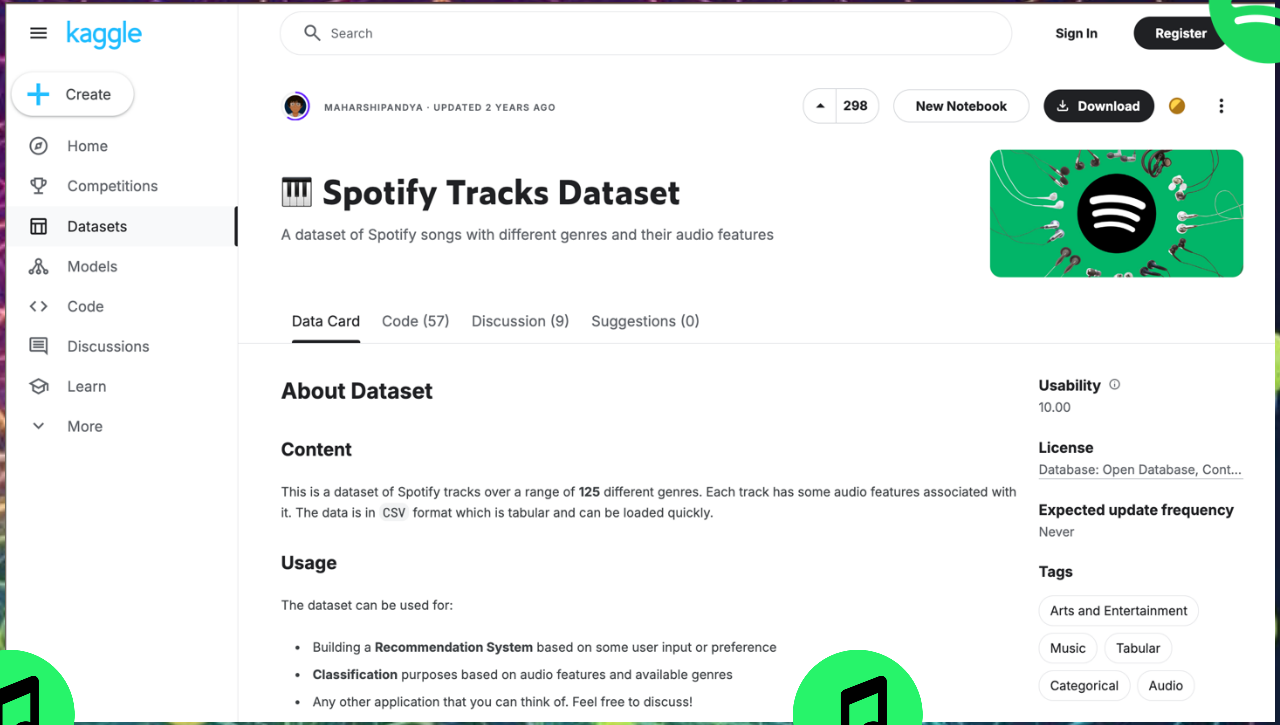
Brand association:
Brands affiliated with popular songs and high listener engagement can enhance the effectiveness of ad placements within streaming services

Goal: To leverage data analytics to uncover what makes a song successful in today's landscape.

DATASET

Dataset Overview

Total Tracks: 114000
Unique Artists: 31437
Unique Albums: 46589
Total Genres: 114



Properties

Popularity	Int64
Duration_ms	Int64
Explicit	Bool
Danceability	Float64
Energy	Float64
Key	Int64
Loudness	Float64
Mode	Int64
Speechiness	Float64
Acousticness	Float64
Instrumentalness	Float64
Liveness	Float64
Valence	Float64
Tempo	Float64
Time_signature	Object
Track_genre	Object

Data pre-processing

Outliers

Outliers in popularity show songs that are either very successful or very unpopular. Here, the outliers are highly popular songs with a score of 100. The same song appears in different genres, pointing to possible data issues. However, since these entries are valid, keeping them ensures accurate insights for playlists or marketing. Using a wider range (3 * IQR) found no outliers, showing these scores are not unusual.

Null values

The dataset was made of very few null values.

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	popularity	81344 non-null	int64
1	duration_ms	81344 non-null	int64
2	explicit	81344 non-null	bool
3	danceability	81344 non-null	float64
4	energy	81344 non-null	float64
5	key	81344 non-null	int64
6	loudness	81344 non-null	float64
7	mode	81344 non-null	int64
8	speechiness	81344 non-null	float64
9	acousticness	81344 non-null	float64
10	instrumentalness	81344 non-null	float64
11	liveness	81344 non-null	float64
12	valence	81344 non-null	float64
13	tempo	81344 non-null	float64
14	time_signature	81344 non-null	int64
15	track_genre	81344 non-null	object

Genres consolidation

We map the correlation between features and popularity by genre, highlights that explicit (strong language, mature themes) song were popular for k-pop genre.

Exploratory results

	popularity	duration_ms	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
count	81344.000000	8.134400e+04	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000	81344.000000
mean	34.635966	2.314117e+05	0.559275	0.635025	5.285922	-8.593940	0.632339	0.088992	0.329670	0.184731	0.219721	0.463280	122.145034	3.896968
std	19.438777	1.164945e+05	0.177746	0.258639	3.557612	5.304765	0.482171	0.116628	0.339961	0.331591	0.198271	0.263383	30.128881	0.456396
min	0.000000	0.000000e+00	0.000000	0.000000	0.000000	-49.531000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	21.000000	1.738710e+05	0.446000	0.455000	2.000000	-10.451250	0.000000	0.036100	0.015900	0.000000	0.098500	0.241000	99.408000	4.000000
50%	35.000000	2.152040e+05	0.573000	0.678000	5.000000	-7.262000	1.000000	0.049100	0.190000	0.000089	0.133000	0.449000	122.030000	4.000000
75%	49.000000	2.673460e+05	0.690000	0.857000	8.000000	-5.140000	1.000000	0.087000	0.629000	0.153000	0.283000	0.676000	140.128250	4.000000
max	100.000000	5.237295e+06	0.985000	1.000000	11.000000	4.532000	1.000000	0.965000	0.996000	1.000000	1.000000	0.995000	243.372000	5.000000

Potential feature importance

duration_ms and loudness have a higher variance so they could have a greater impact in predicting popularity

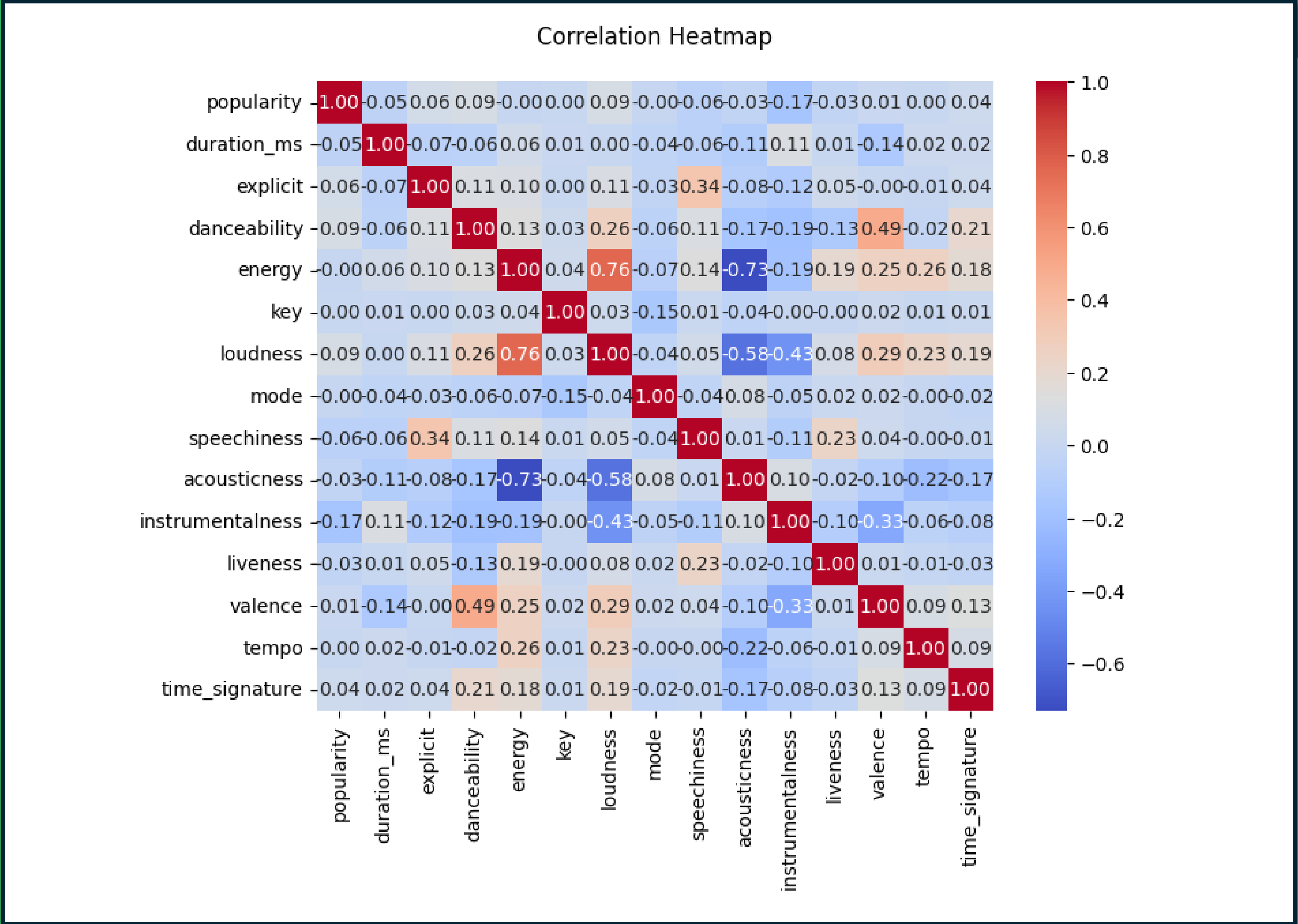
Highlights categorical variables

Key and mode are categorical and numeric

Skewness

Instrumentalness and speechiness have a mean way lower than their max value suggesting a lot of ‘0’ in the data

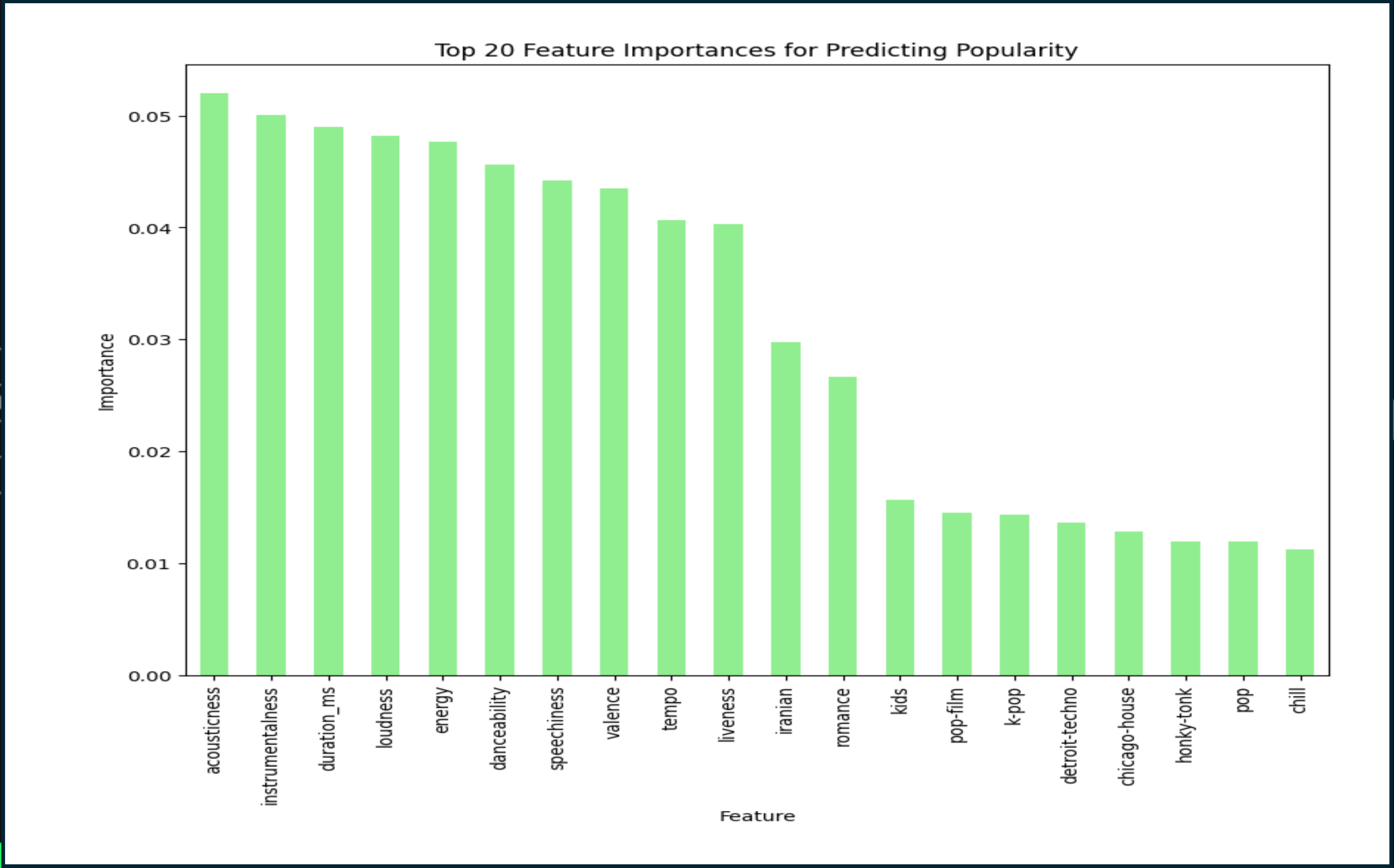
Correlation - Heatmap



INSIGHTS

TOP FEATURES
& POPULARITY CORRELATION

Feature Importance



Derivable Insights

Accoustic Elements: Less is better for a more popular track

Our model identified that lesser amounts of acousticness was the most important feature in determining how popular it would be.

Oftentimes, songs that are more associated with acoustics tend to be softer and simpler.

Small negative correlation (-0.032) suggests that it may be a complement to other features

Derivable Insights

Instrumentals and Popularity

Considering its strong negative correlation (-0.1745), it appears as though songs that include good-sounding vocals are important in making a popular hit.

However, this should be a well-balanced aspect of the song, as having higher amounts of speechiness was found to slightly negatively correlate (-0.0649) with popularity.

Song Duration

Another key takeaway is that shorter songs are found to be slightly more popular – this reflects a current trend in the industry where quick and catchier tracks tend to perform better in the streaming era.

This has parallels and implications with platforms such as TikTok, Instagram Reels, and Youtube Shorts which favour shorter audio snippets that hook the viewer.

Derivable Insights

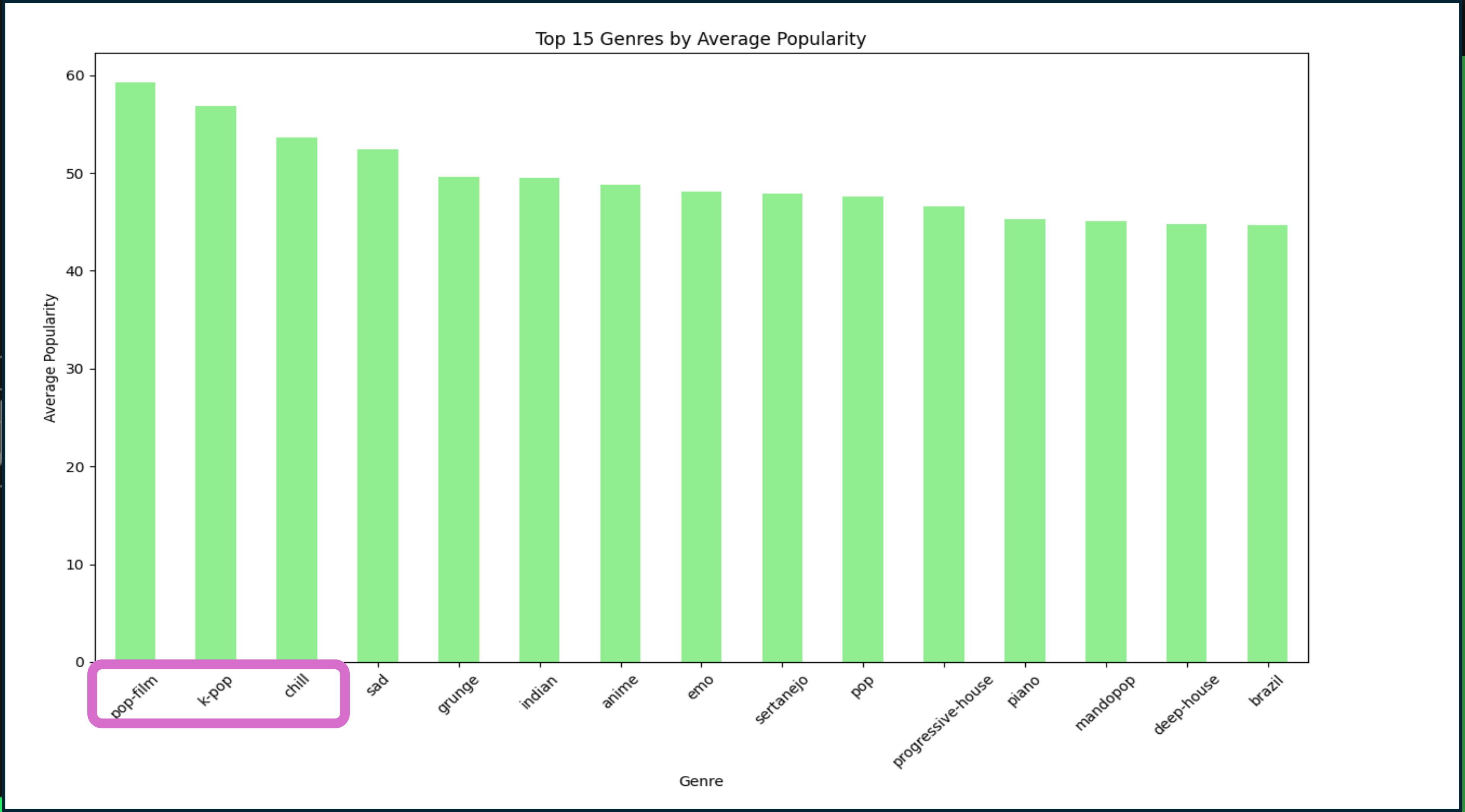
The Feel of the Song: Danceability, Loudness, Valence, and Tempo

Considering all of these factors are positively correlated with song popularity, it is crucial for artists seeking to make a popular song to focus their efforts on making songs that are more "positive" sounding and get people moving to the beat.

INSIGHTS

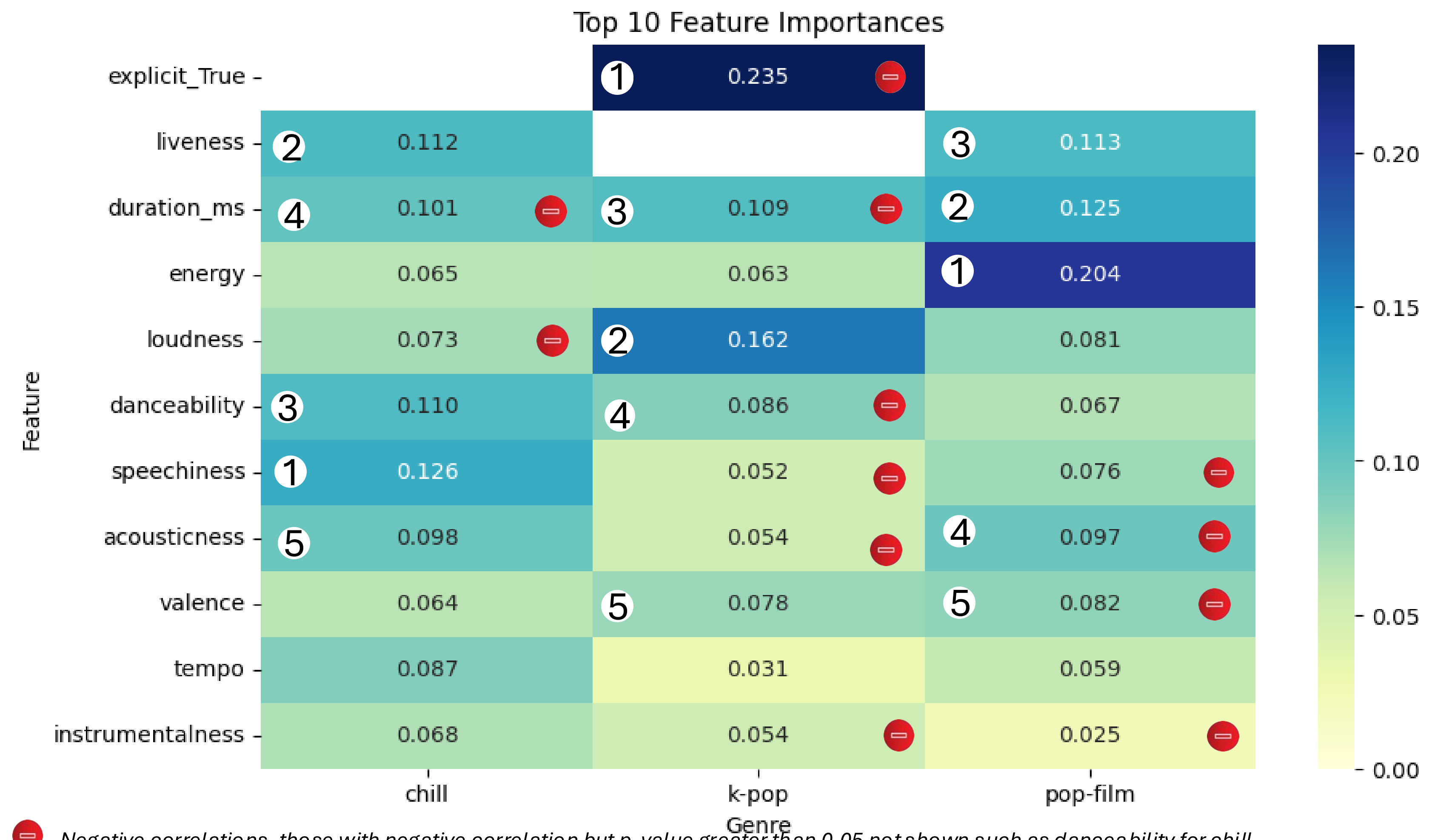
A LOOK AT TOP 3 GENRES

Genre-Wise Average Popularity



Next Steps:
Deep Dive in to top 3 genres

Top 3 Genres: Features Importance Comparison



Negative correlations, those with negative correlation but p-value greater than 0.05 not shown such as danceability for chill

High Negative Impact of Explict Content on K-Pop

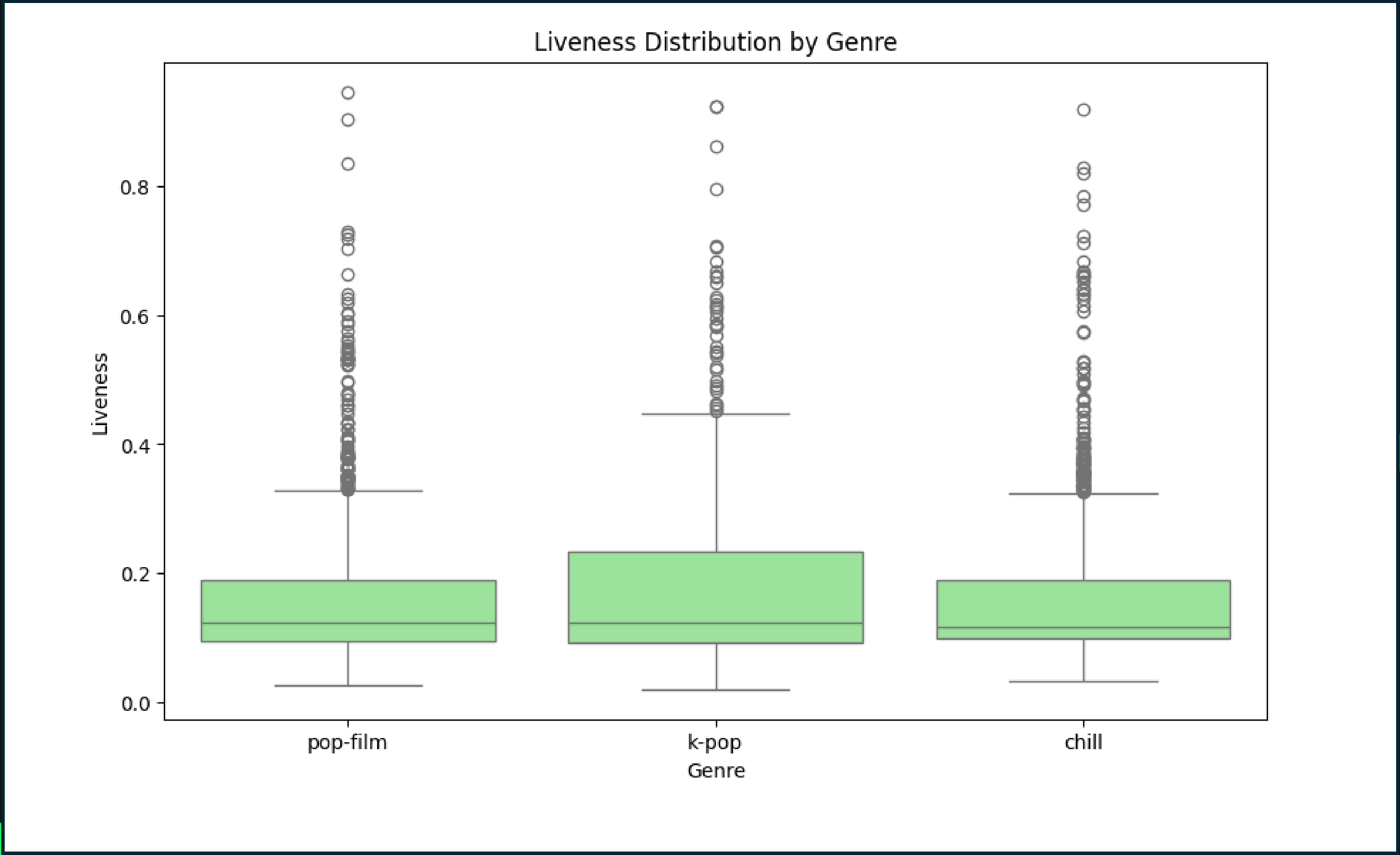
Duration plays an important role across genres

Energy x Pop: Most important – preference for upbeat/energetic songs

Loudness: 2nd most important in K-Pop
As expected, negative impact for chill songs

Speechiness x Chill: Most important
Indicating preference for spoken word/conversational style

Top 3 Genres: Liveness Comparison



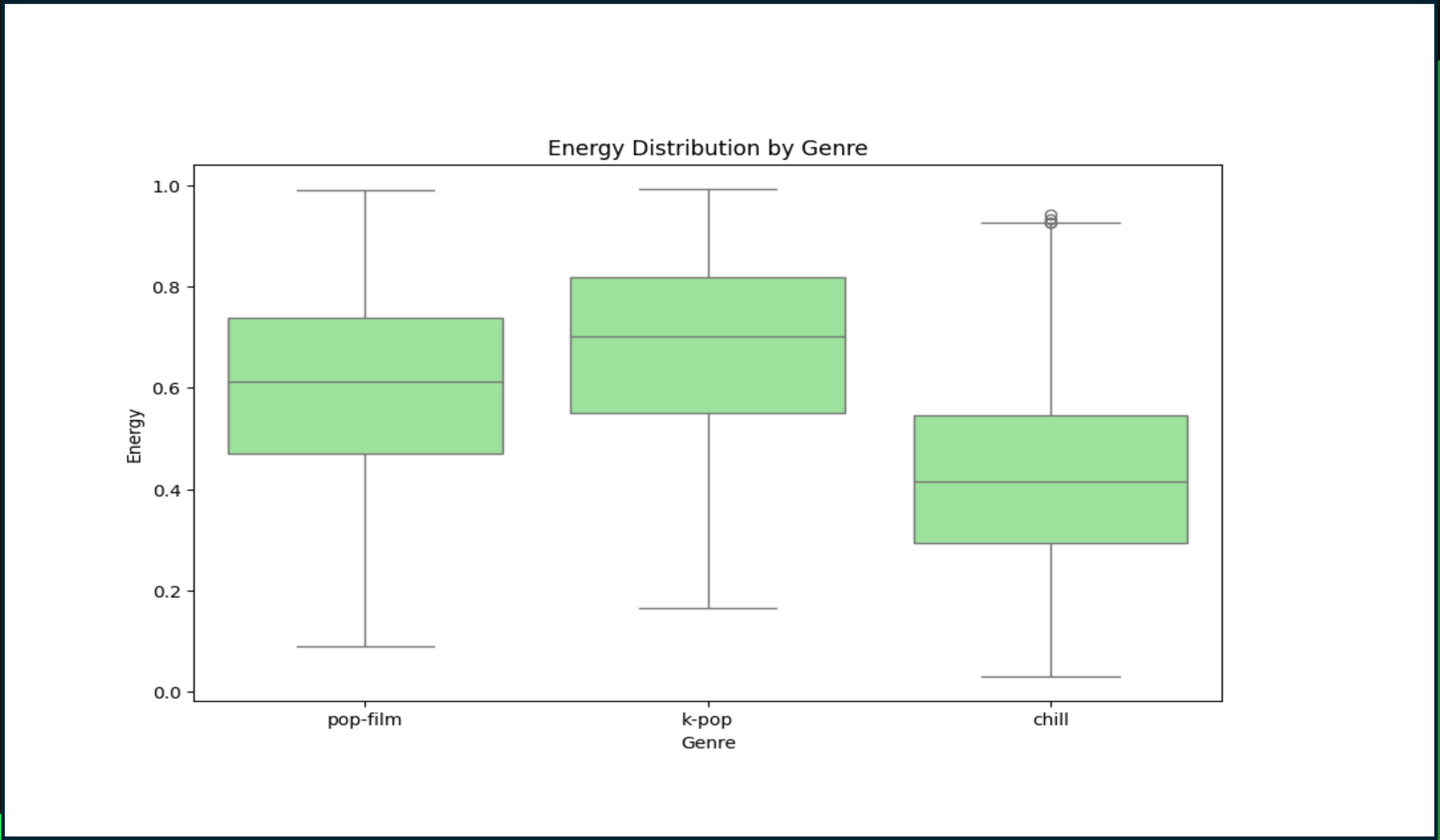
INSIGHTS:

Genre Similarity: comparable median, preference for lower liveliness

Higher Diversity in K-Pop: suitable for tracks with studio & live-like experiences

Chill: Tightest distribution - stronger preference for subdued studio-like recordings

Top 3 Genres: Energy Comparison

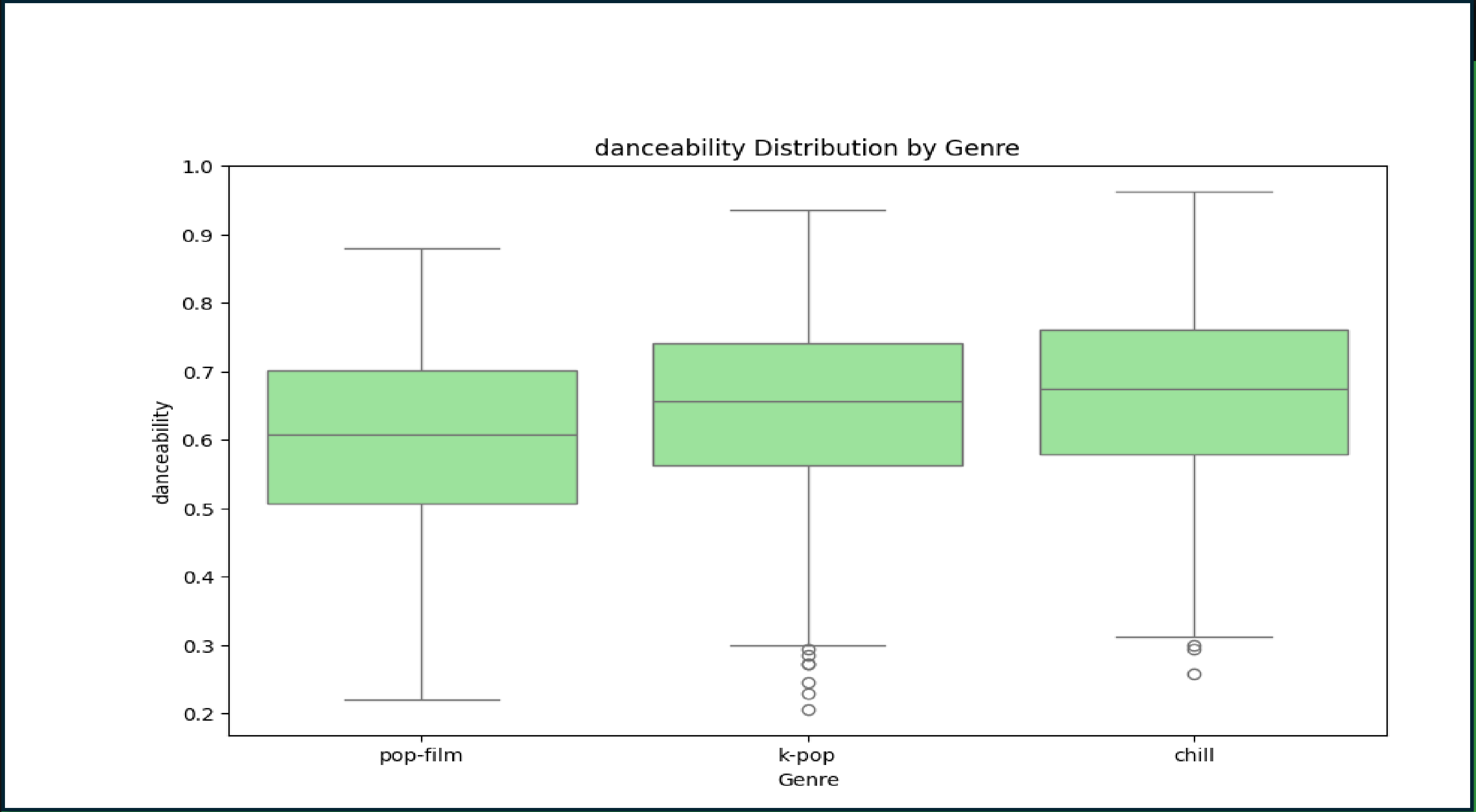


INSIGHTS:

K-Pop: Higher energy with wide range. highlight energetic aspects

Chill: Lower energy level, likely targeting 'calmer'/chill moments

Top 3 Genres: Danceability Comparison



INSIGHTS:

Chill: Interestingly highest median, maybe rhythimic in spite of being 'chill'

K-Pop: Versatile in preference for danceablity, but general preference for danceability seen

Pop-Film: Lower danceability

- Likely to augment tempo of on-screen scenes

PREDICTIVE MODEL

Predictive Model Overview

<u>OBJECTIVE:</u> Develop a model to predict song popularity to support creators/marketers

Data Splitting

- Train Set: 64%
- Validation: 16%
- Test: 20%

Feature Selection

- LASSO (139 Features selected - by shrinking coefficients of less important features to zero.)

Select Predictive Model

- Random Forest chosen

Hyperparameter Tuning

- bootstrap: True
- max_depth: None
- max_features: 'sqrt'
- min_samples_split : 5
- n_estimators: 100

Feature Importance

- Extracted key features
- Top 5 Fatures: 25% importance [Acoustic, Instrumental, duration, loudness, energy]

Model Performance

- MSE = 208.97
- R^2 = 44.76%

Model Development Rationale

Lasso

Why?

- To select only a subset of the original features to train the random forest to speed up the training process.
- Performs feature selection by shrinking coefficients of less important features to zero.

Outcome

- Retained a subset of features for further analysis.
- 139 Features Selected out of 141

Random Forest

Why Random Forest?

- Robust with irrelevant features.
- Easy to report feature importance metrics.

Outcome

MSE = 208.97

R^2 = 44.76%

Hyperparameter Tuning & Cross Validation

Why?

- To identify the optimal configuration of hyperparameters for improving model performance.
- To ensure the model is robust and generalizes well across different subsets of data.

Outcome:

- Best Parameters:
    - bootstrap: True
    - max_depth: None
    - max_features: 'sqrt'
    - min_samples_split: 5
    - n_estimators: 100
- Lowest MSE from cross-validation: **215.99**

Model Performance

R^2
44.76%

MSE
208.97

SAMPLE PREDICTION

	Actual Popularity	Predicted Popularity
12070	35	31.313409
11608	45	43.597055
11703	41	42.304948
15844	35	36.732331
8576	39	24.963353

MODEL EVALUATION

KEY METRICS:

- R^2 shows moderate predictive ability ie 44.76% variability in song popularity explained by model. High MSE indicates low predictive power, further improvement required

POTENTIAL APPLICATIONS/STRENGTHS:

•Identifies general trends and relationships between song features and popularity.
•Provides actionable insights for feature optimization (e.g., energy, duration)

LIMITATION/NEXT STEPS

•Moderate accuracy, particularly with predictiveness
•Further feature engineering to be explored to enhance model

CONCLUSION

•Model is a good starting point for understanding key features, and providing insights
•While not perfect, it is valuable for understanding broad patterns and supporting creative and promotional strategies.

Feature Optimization Recommendation
Top Songs vs Sample Song

- **Average Values**
  - Calculated averages of selected features for the **top 10% popular songs**.
  - Helps in understanding the characteristics of highly popular songs.
- **Correlation with Popularity**
  - Examined **positive** and **negative correlations** between features and popularity.
  - To see how to improve the popularity of the song.
- **Sample Song**
  - The sample song was selected from the test set. The popularity prediction and the selected features are reported
  - The predicted popularity is below 50. To improve the song, we can decrease duration&energy&speechness, increase danceability&valence

```
Average Values for Top 10% Songs by Popularity:
acousticness            0.287050
instrumentalness        0.084164
duration_ms        217894.444193
loudness               -7.878983
energy                  0.632058
danceability            0.591220
speechiness             0.080050
valence                 0.476535
tempo                 120.405115
liveness                0.180531
iranian                 0.000000
romance                 0.000000
dtype: float64

Correlation with Popularity (positive or negative):
acousticness: -0.0322 (negative)
instrumentalness: -0.1745 (negative)
duration_ms: -0.0533 (negative)
loudness: 0.0902 (positive)
energy: -0.0009 (negative)
danceability: 0.0873 (positive)
speechiness: -0.0649 (negative)
valence: 0.0113 (positive)
tempo: 0.0019 (positive)
liveness: -0.0290 (negative)
iranian: -0.1818 (negative)
romance: -0.1645 (negative)
```

Sampled Song's Important Features and Popularity:

	acousticness	instrumentalness	duration_ms	loudness	energy	danceability	speechiness	valence	tempo	liveness	iranian	romance	Actual Popularity	Predicted Popularity
387	0.00321	0.000051	237586	-2.828	0.977	0.541	0.157	0.263	134.938	0.0796	0	0	74	46.130758

Conclusion/Key Outcomes

Guidelines for Stakeholders

- Shorter song durations
- Less accousticness
- More danceability

Are key in creating a popular song

Genre Specific Guidelines

- Chill has highest danceability median
- K-pop has highest energy
- Across genres, there is a preference for lower liveness

Predictive Model

Feature Selection: LASSO for initial feature filtering

Prediction Model: Random Forest to predict popularity

Feature Importance: Extracted key contributing features

Insights: Analyzed relationships between key features and popularity.

Feature Comparison for Predicted Song

The ability to look at a song and gain a popularity prediction based on existing data is insightful for artists looking to optimize their track

Appendix

Project Scope

Music has always existed as a space that promotes connectivity. It has the unique ability to amplify and evoke emotion, and has evolved over time in a number of perspectives.

Sonically, changes in sound and new emerging genres can be traced back to shifts in culture and identity such as the rhythm and blues music of the 1940s, the hippie movement of the 1960's, or the mainstreaming of rap and hip hop in the early 80s.

From a more technological perspective, we have seen a shift from analog – vinyl records, CD's, cassette tapes, to radio, and now to the digital streaming era, which has made listening to songs more accessible than ever before.

Project Scope

The ease of access to music and information in this current age has presented artists both new and old with unseen challenges and opportunities.

It is now easier than ever for artists to release songs and accumulate revenue through streams as opposed to making money through touring or purchasing either digital or physical copies of their music.

However, with this, a key challenge that emerges is the oversaturation of the music market and the difficulty for artists to stand out considering how much content is getting released. Although the democratization of music creation and distribution has made this creative outlet much more accessible to all artists, it comes at the expense of less potential visibility and a successful career.

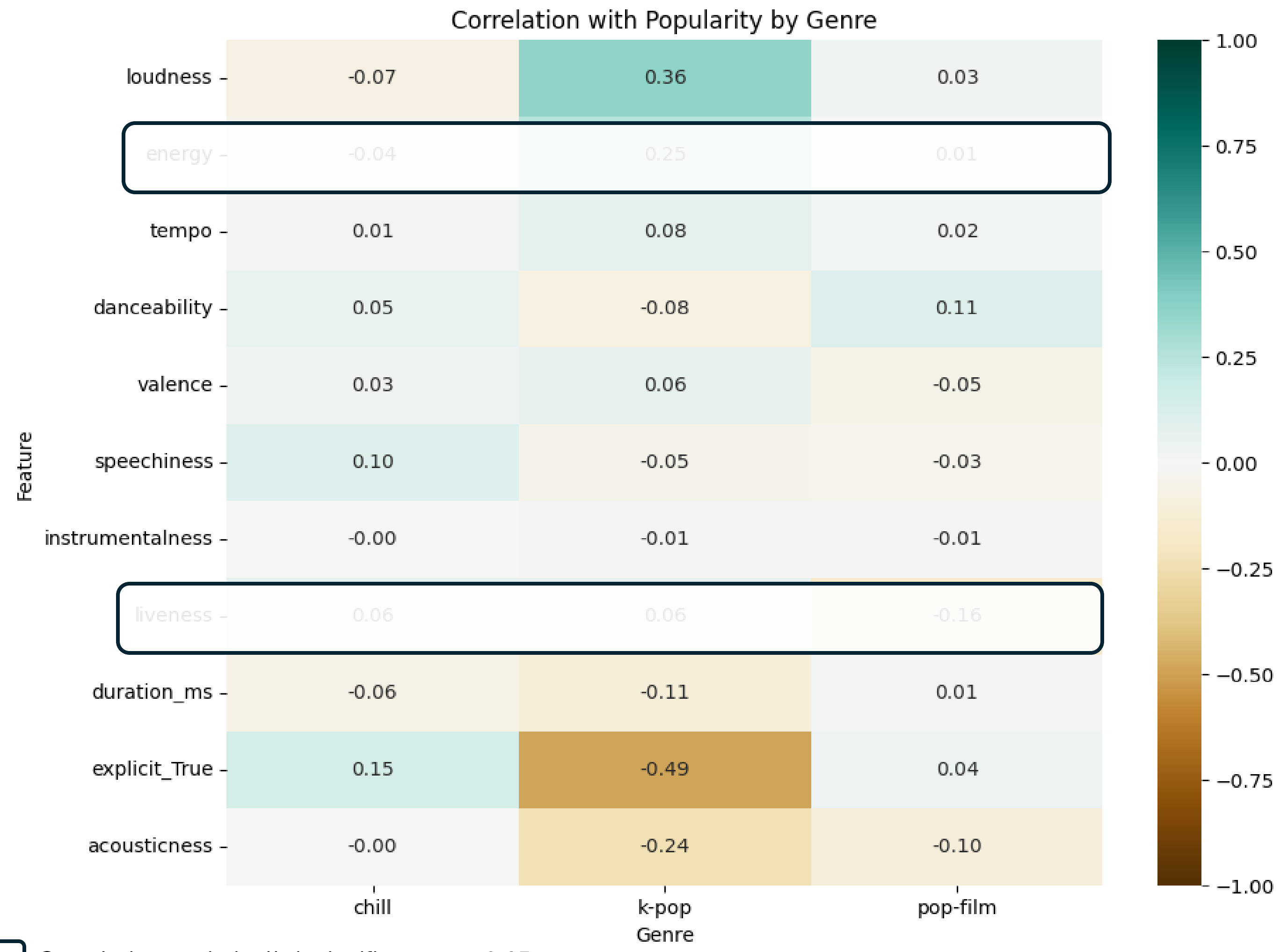
Current Research in Spotify Data

Considering the availability and scope of Spotify data, there exists a growing body of research that has looked at what makes current songs popular.

- Songs that have been found to sound similar to prior popular hits have been found to be less likely to succeed, and that there is an optimal level of differentiation that can predict if a song may rise to the top of the charts (Askin and Mauskapf, 2017)
- A cluster analysis performed on the Top 100 Trending Spotify Songs of 2017 and 2018 found that songs that had high ‘danceability’ scores and low ‘instrumentalness’ increased the popularity of a song (Al-Beitawi, Salehan, & Zhang, 2020)
- A research paper by Nijkamp (2018) on the relationship between song data and popularity based on streams utilized a regression model and found that lack of lyrics are negatively related to stream count as was song duration, while features such as danceability and speechiness were found to be positively related to stream count.

A unique perspective that our research will provide will be how the definition of what makes a song 'popular' shifts over the years and how this influences the likelihood of future songs reaching top charts.

Top 3 Genres: Top Features Correlation Heatmap



Correlation statistically insignificant – p > 0.05

Positive & Negative relationship indicated within random forest heatmap. Low correlations but good for understanding trends/initial relationship

Will need more complex modeling

Sources

Askin, N., & Mauskapf, M. (2017). What makes popular culture popular? Product features and optimal differentiation in music. American Sociological Review, 82(5), 910–944. DOI: https://doi.org/10.1177/0003122417728662

Al-Beitawi, Z., Salehan, M., & Zhang, S. (2020b). What makes a song trend? Cluster analysis ofmusical attributes for Spotify Top trending songs. Journal of Marketing Development and Competitiveness, 14(3), 79–91.

Nijkamp, R. (2018). Prediction of product success: Explaining song popularity by audio features from Spotify data.

Pandya, Maharshi. "Spotify Tracks Dataset." Kaggle, 2021, https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/data.