“I Wish that I Could Be Like the Cool Kids”: an Analysis for Underexposed Indie Artist on Spotify
by Sydney Hu and Stephanie Shaw

Project Objective

The project aims to answer the question that many underexposed indie artists, including Sydney (artist name JINZY), have while releasing their songs to the Spotify platform: how can my songs reach more audience? To answer the question, we first uncover one main method for these artists to gain traction: to be featured on Spotify’s playlist. While there are many influential external factors, such as marketing and promotions, artists are still left with a question unanswered: disregarding the external factors, do their songs have the potential to be featured? In other words, **what are featured indie-pop artists doing *musically? Are they different* compared to underexposed ones?**

Data Scope & Data Collection

We set out to collect data for two groups of songs by artists under the Indie-Pop genre: popular (hereafter referred to as Group 1) and underexposed (Group 2). Group 1 includes songs collected from an un-personalized Spotify-made Indie-pop playlist: “Indie Pop Hits”, which was the top result when searching in the Indie-pop genre. Group 2 is collected from scraping a website called Indie Shuffle. To ensure the validity of the comparison between groups, we set the scope for Group 2 by drawing from the exploratory analysis we had done for Group 1.

Originally, we had set out to find Group 2 within the Spotify platform. However, since songs in Group 2 are meant to be songs not getting featured on Spotify, it became a challenge. In addition, we would not be finding valid data records as it would not be consistent with how consumers usually discover these Indie-pop songs. We investigate other ways to find underexposed artists from external sources. Eventually, we decided to web-scrape the Indie-Pop genre page under a website called Indie Shuffle (https://www.indieshuffle.com/songs/indie-pop/), where dedicated listeners find Indie song recommendations.

To collect data for Group 1, we had utilized Spotify Web API (https://developer.spotify.com/documentation/web-API) and the library Spotipy (https://spotipy.readthedocs.io/en/2.22.1/). We utilized the playlist_tracks function from the Spotipy library using its URI, which returned a dictionary, including information relating to the playlist in addition to the track. We extracted the “track” attribute and put it into a data frame. However, the track attribute does not contain the data of interest – the audio features, so we created a function called create_audio_features that takes a list of tracks as input and utilizes Spotify’s audio_features function to get each track’s audio features. Then, we merge the two data frames using the column ‘URI’ as the key.

Before starting data collection for Group 2, we first ran Group 1’s exploratory analysis and established the reference and baseline for Group 2’s data scope. The two groups will be limited to similar release dates and similar sizes: Group 1 has a sample size of 100. The minimum popularity score to be featured on the Spotify-made playlist will be the upper bound for Group 2.

An external website recommendation presented a problem: the songs might not be on Spotify. Considering the problem and the baseline established by Group 1, we decided to overshoot the sample size and scraped 26 pages, which included 200 songs in total. We used requests and beautiful soup to extract song information, which is contained in the <h5> tag. However, the hierarchy of the website was not well-established, so we resolved to use Regular Expression to extract the artist name and the song name. We stored them in a list of tuples.

Later, we looped through the list while trying Spotipy's search function to query each track's information using the artist and song name and excepting IndexError for tracks not on Spotify. During our process of drawing out song data through Spotify's API, 67 tracks were not found. Additionally, Since Sydney's professional aspiration serves as the underlying motivation of the project, we plan to include her 7 songs within Group 2 after cleaning. After obtaining all the tracks, we repeated similar steps in connecting each track with its audio features.

The raw datasets contain 34 columns, including two columns requiring further processing if needed. The 'album' column, for instance, contains all data about the album in JSON format. Group 1 has 100 entries while Group 2 has 140.

Data Cleaning

Upon further inspection of the raw data, we decided to extract information about the album – name, URI, and release date. All album attributes are inserted as new columns in the data frame. The 'album_release_date' was stored as a PANDAS DateTime object for easier analysis.

After this, we determined the relevant information to analyze for the project's scope. These include the song's URI as the index key, popularity, release date, and the following audio feature attributes: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, duration, liveness, tempo, time signature, and valence.

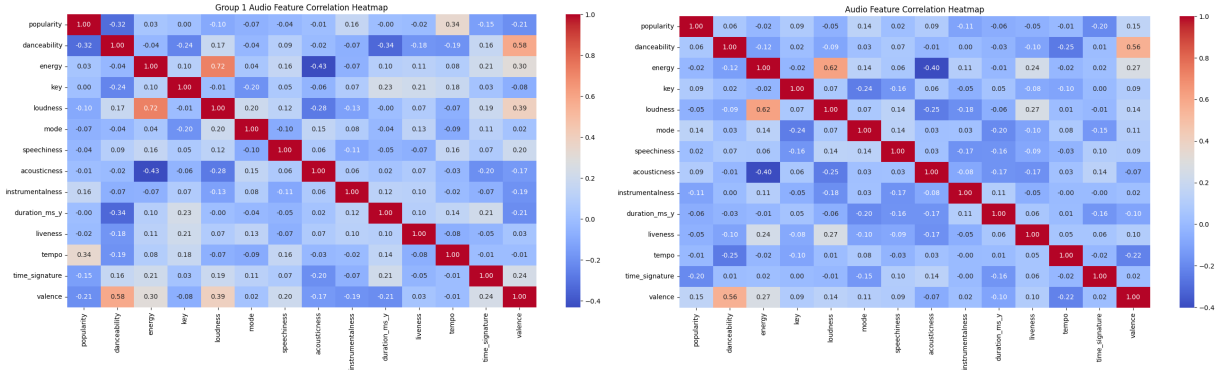
As aforementioned, Group 1 determined Group 2's data scope. The "Indie Pop Hits" playlist included 100 songs. The earliest release date was 2013-01-01. We assumed that the album was made to be up until the present. We also found that the minimum popularity score to be featured on the Spotify-made playlist is 39, which became the upper bound for Group 2. We queried the dataset of 140 songs using two constraints: a popularity score of less than 39 and an album release date later than 2013-01-01, resulting in a sample size of 99.

After cleaning, both datasets contain the column index 'URI' followed by 15 columns. Group 1 has 100 records of songs while Group 2 has 99.

Analysis & Visualization

Before answering our research question: **what featured indie-pop artists are doing differently *musically* compared to underexposed ones,** we first created a scatterplot and regression plot, which did not yield any significant patterns. We decided to run correlation analyses in an attempt to determine if there are key variables that determine a song's popularity.

Correlation Analysis. The correlation analysis reveals insights into the relationships between audio features. The heatmaps display varying degrees of correlation through cool/warm gradients, with higher correlation in darker tones of red (for positive) or blue (for negative).

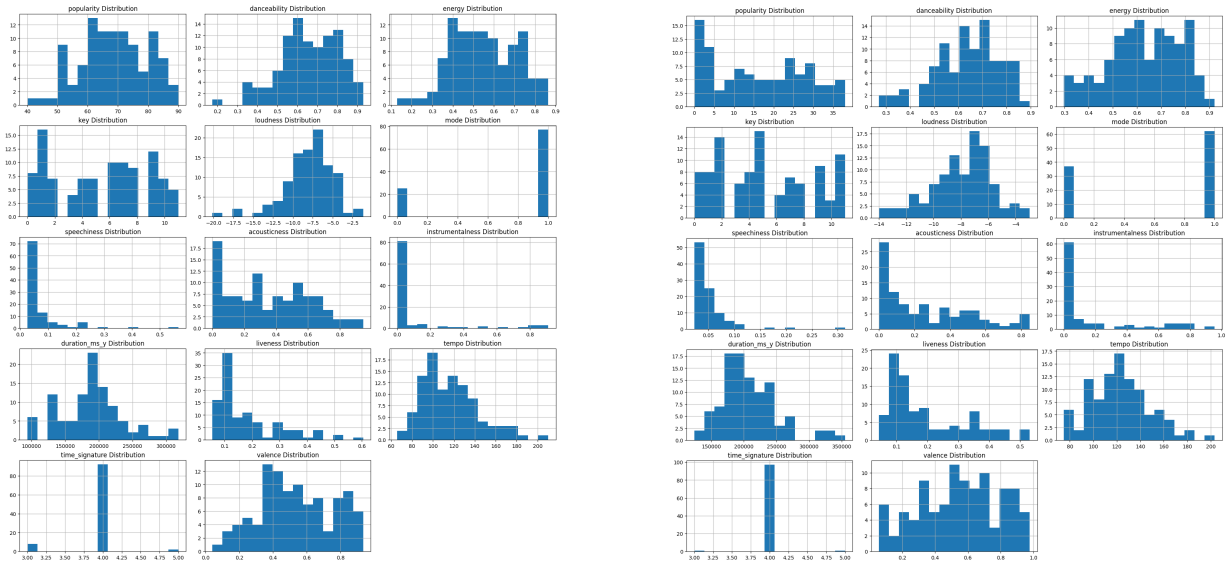


Group 1 (left) and Group 2 (right) Heatmaps for correlation between variables.

Notably, the features with the highest positive correlation were observed between loudness and energy (G1: 0.72; G2: 0.62), danceability and valence (0.58; 0.56), and the highest negative correlation between acousticness and energy (-0.43; -0.40). These statistics reveal a degree of connectedness between qualitative variables, such as danceability, energy, and valence, and more quantifiable attributes such as loudness and acousticness. This finding gives artists a better understanding of Spotify's calculation algorithm, although further experiment is required to establish a causal direction.

Overall, the heat map analysis underscored the difficulty of determining key success variables for popular tracks and the ambiguity Indie artists face when trying to maneuver Spotify's algorithm since none of the audio features strongly correlate with the popularity score.

MANOVA-test. We decided to run a multivariate ANOVA test to check for significant differences between the two groups of songs in the combined set of all the audio features. To check for conditions of normality for further analysis, we created the following histogram panels and the two groups are also similar in sample sizes (G1: n=100, G2: n=99).



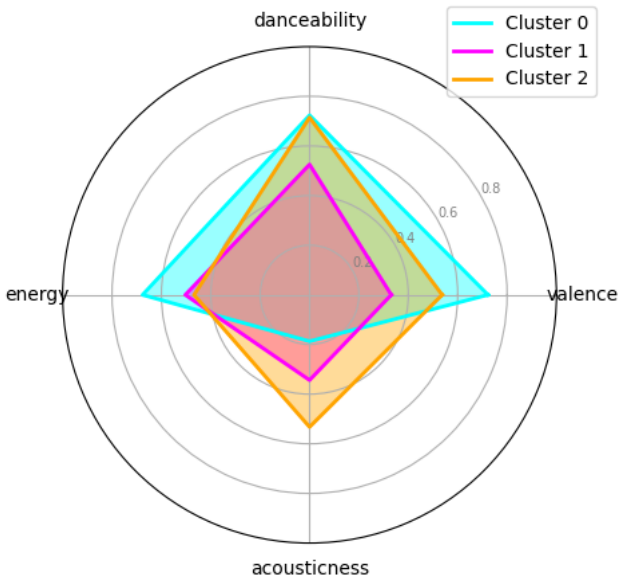
Group 1 (left) and Group 2 (right) attribute histogram panel.

The ANOVA-test analysis conducted on audio features yielded insights into whether there is a statistical difference between featured songs and underexposed songs. The resulting p-value (0.002) indicates a statistically significant difference between the groups. The result suggests that there is not enough evidence and we reject our null hypothesis that the featured and underexposed songs have the same audio features.

Cluster Analysis & Within-Group ANOVA test. We then extracted characteristic featured songs and how underexposed songs compare. To do so, we first found the elbow point, or the optimal number of clusters (k), by detecting the point where the decrease in WCSS starts to slow down. We determined that k = 3 clusters for both Group 1 and Group 2. Then we standardized the data and put records into arrays and clusters using the K-means algorithm, which partitions the data into k clusters relative to the distance to the centroid of the clusters.

After assigning the songs to three clusters, we used ANOVA to determine the most influential factors that distinguished the clusters. Out of the statistically significant audio features, we found the most influential ones include **valence** (F-value: 58.12), **danceability** (30.64), **energy** (27.87), and **acousticness** (22.92). According to Spotify's definition, valence describes the musical positiveness of a track. Danceability is based on "tempo, rhythm stability, beat strength, and overall, overall regularity" and describes how well-suited a track is for dancing. Energy describes a track's intensity and activity: a track with high energy is typically fast, loud, and noisy. Acousticness measures the confidence of whether the track is acoustic.

We created the following radar visualization to show the differences between clusters in Group 1. The axes include the four influential audio features. The clusters are differentiable by colors.



The cluster analysis revealed that the Spotify indie-pop playlist features three clusters of songs:

1. 34 songs that are high in valence(e.g. cheerful, happy), danceability, and energy, while having low acousticness
2. 33 songs that are low in valence(e.g. sad, depressed, angry), low in danceability, low in energy, and medium in acousticness
3. 35 songs that are medium in valence, high in danceability, low in energy, and medium in acousticness.

In addition, we found that the underexposed songs fall under 3 clusters but almost half (49 out of 99) are medium in loudness, high in energy, relatively high in danceability and valance.

Implications

The various analyses comparing featured artists to underexposed Indie-pop artists have revealed insights that can be useful for underexposed Indie artists using Spotify. As the MONOVA test illustrates, the two groups of songs are indeed musically different.

Indie-pop artists who are aiming to be featured can take into account the descriptions outlined in the cluster analysis section as a direction when trying to tap into the same audience or establish their niche and point of difference.

Concluding Remarks

The report focuses on exploring the relationship between in-platform variables, specifically audio features, and popularity. We acknowledge that many more factors go into a song's success on platforms like Spotify. To further help answer the questions underexposed Indie artists like Sydney have when striving for their songs to gain traction, future work should consider other factors, including but not limited to promotional campaigns on Spotify or other platforms (e.g. Instagram). Another direction can be a predictive algorithm that incorporates all factors to determine the probability of a song being featured.