What Makes a Song Trend? Cluster Analysis of Musical Attributes for Spotify Top Trending Songs

Zayd Al-Beitawi
California State Polytechnic University, Pomona

Mohammad Salehan
California State Polytechnic University, Pomona

Sonya Zhang
California State Polytechnic University, Pomona

Music streaming services like Spotify have changed the way consumers listen to music. Understanding what attributes make certain songs trendy can help services to create a better customer experience as well as more effective marketing efforts. We performed cluster analysis on Top 100 Trending Spotify Song of 2017 and 2018, using nine musical attributes, including danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. The results show that music structures with high danceability and low instrumentalness increase the popularity of a song and lead them to chart-topping success.

Keywords: cluster analysis, music, Spotify

INTRODUCTION

Music streaming services have revolutionized the way consumers listen to music, not only by lowering the costs but also by providing consumers with an endless library of artists from all genres and musical backgrounds. As of July 2019, Spotify, the leading music streaming service, provides access to over 50 million tracks to 232 million monthly active users, including 108 million paying subscribers. Spotify's payment model structures around a $5 monthly subscription fee that provides a user with unlimited, advertising-free experience. For an additional $5, users receive premium features including offline listening, a mobile app, enhanced sound quality, exclusive content, early album releases, and sound system compatibility.

In recent years, Spotify has allowed users to discover music and create exclusive playlists based on their musical preferences, favorite genres and artists, and even mood. This design has helped in eliminating a potential struggle for users in searching for an extensive database of millions of songs. To optimize such discovery and personalization, streaming services like Spotify not only rely heavily on recommender systems but also on human editors. A deeper understanding of the characteristics and use of playlists and how users create and maintain their playlists can contribute to better recommendations.

As these playlists become more customized based on Spotify's recommendations, certain songs begin to recurrently appear on "Top Song" lists resulting in their trending on the platform. For each song, Spotify provides audio features such as duration, key, and mode. This study intends to investigate whether the success of the trending songs is related to these attributes. The results would allow music streaming services to create better-customized playlists that reduce search time and improve the satisfaction of their users. The findings would also lead to more focused marketing efforts by the artists to attract potential subscribers to their music.

Related Work

Discovery and personalization are a key part of the user experience and critical to the success of the creator and consumer ecosystem in music industry. Both Content-based filtering and Collaborative filtering recommender systems were applied for discovery and personalization by both practitioners and researchers. Data scientists at Spotify had developed Discover Weekly, a personalized playlist which updates weekly and reached 1 billion streams within the first 10 weeks from its release, powered by a scalable factor analysis of Spotify's over two billion user-generated playlists matched to each user's current listening behavior. Others had also generated playlist recommender systems based upon playlist names, social data of musicians, or the Facebook likes of artists and the listening history of songs of a Spotify user. Finally, a survey study finds that track and artist popularity can play a dominant role in the automated playlist generation process. More interestingly, a study shows that very simple popularity-based algorithms can outperform sophisticated algorithms in more general music recommendation scenarios.

Previous studies attempted to classify popular music data with various machine learning algorithms, including decision tree, regression, SVM, Naïve Bayes, and neural network. Most of these studies utilized a more limited and abstract set of musical attributes compared to Spotify's audio features. Only one study used Spotify's audio features to find music popularity; the researchers conducted CART decision tree classification to a dataset containing Indonesia's Daily TOP 200. The songs with streams more than 2 million labeled as popular and the songs with streams less than 2 million labeled as non-popular. The results found five dominated attributes represented the characteristics of popular songs - acousticness, liveness, energy, valence, and key. Songs played with acoustic instruments, medium energy, moderate valence, and high base key are considered as popular songs in Indonesia. In this study, we aim to study the similarities of trendy music in the more influential U.S. market based on Spotify's audio features, using a different machine learning approach – clustering analysis. We hope the results from this study could contribute to discovery and personalization for consumers, as well as to music creation and promotion for creators.

Spotify Audio Features

Using the audio features component of the Spotify API service, users can extract a series of characteristics for each song, such as how acoustic or loud it is. The list of audio features, as well as their data type and definition, are provided by Spotify as displayed in table 1.

TABLE 1
SPOTIFY AUDIO FEATURES

Attribute	Data Type	Definition
Key	integer	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.
Mode	integer	The modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
Time_signature	integer	An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
Danceability	float	Describes how suitable a track is for dancing based on a combination of musical elements, including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable, and 1.0 is most danceable."
Energy	float	A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Loudness	float	An attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud.
Speechiness	float	Detects the presence of spoken words in a track." If the speechiness of a song is above 0.66, it is probably made of spoken words, a score between 0.33 and 0.66 is a song that may contain both music and words (e.g. rap music), and a score below 0.33 means the song does not have any speech.
Acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Instrumentalness	float	Represents the number of vocals in the song. The closer it is to 1.0, the greater likelihood the song contains no vocal content.
Liveness	float	Describes the probability that the song was recorded with a live audience. A value above 0.8 provides a strong likelihood that the track is live.
Valence	float	Describes the musical positiveness conveyed by a track, with a measure from 0.0 to 1.0. Tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry).
Tempo	float	Describes the timing of the music or the speed at which a piece of music is played.
Duration_ms	integer	The duration of the track in milliseconds.

RESEARCH METHODOLOGY

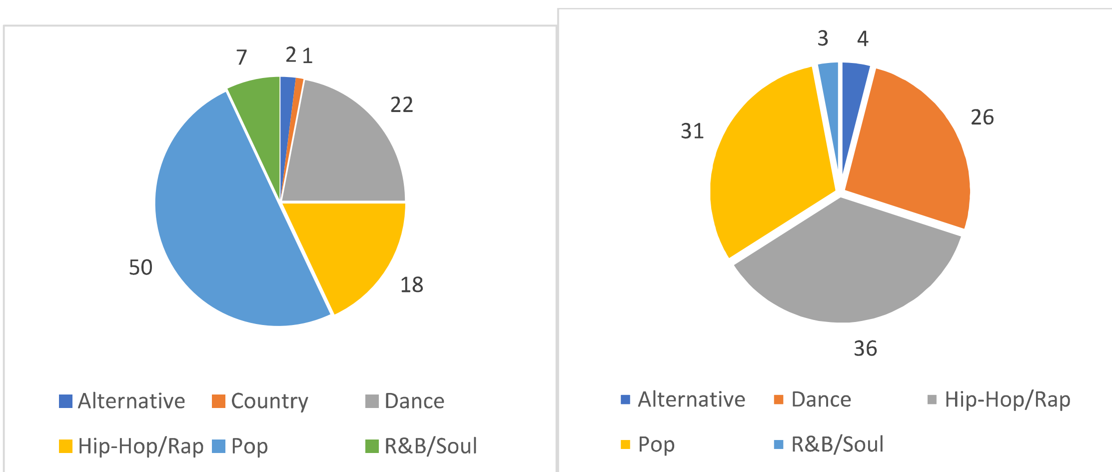
Dataset

At the end of each year, Spotify compiles a variety of lists showcasing the top artists, songs, and albums, and it categorizes some of the lists based on region, streaming platform, and musical genre. To analyze popular musical trends and to understand what leads to their success, we used the "Top 100 Trending Spotify Songs" of years 2017 and 2018 as our primary datasets in this study which is comprised of the top 100 most-streamed tracks of each year on Spotify, including their track ID, song name, artist name, and Spotify audio features. Although we were limited to 200 records, the type of artists and genres featured on the list represent a good variability, with over five genres, as shown in Figure 1.

The descriptive statistics in Table 2 shows that the musical characters of the trending songs are overall consistent across two years. Trending songs in both years are low in speechiness (<0.33), liveness (<0.80), and instrumentalness (average 0.00), meaning songs that top the charts usually do not have speech or vocal in it and they are usually not recorded at live concerts. More importantly we see consistent high danceability, high energy, and low instrumentalness in the trending songs across two years.

The correlations of the musical attributes in Table 3 are mostly consistent with their definitions in Table 1. There is a high correlation (r > 0.70) between loudness and energy, also a moderate correlation (r > 0.30) between Valence and Loudness, but these will not be an issue in this study, which focuses on clustering by measuring distances between records.

FIGURE 1
SPOTIFY TOP 100 SONGS MUSIC GENRES



(Left: 2017, Right: 2018)

Cluster Analysis

We conducted a cluster analysis using k-means clustering to identify groups of trending songs with similar features. K-means clustering involves using "a set of n data points in real d- dimensional space, R^d, and an integer k...to determine a set of k points in R^d...to minimize the mean squared distance from each data point to its nearest center".

Before determining the best value for k, we first cleaned our dataset and rearranged it to filter out unhelpful features. As a result, we removed track ID, song name, and artist name columns, which are all nominal and not suitable in the cluster analysis. After further visualizing the dataset, we decided also to remove the time signature column which had a low variance as it only contained time signatures of 3 and 4, which is challenging to use for more than 2 clusters. We then removed all rows with null values. Once the data was cleaned, we normalized all non-categorical values to make sure all variables have equal importance when the distance is calculated. Lastly, we created dummy variables for categorical columns, which were key and mode. The genre category was not provided by Spotify and was manually collected and included in the dataset by the authors. The genre was excluded from cluster analysis and was saved for comparison with the generated clusters.

Then, we moved on to determine the optimum number of clusters (k) for our k-means clustering model using Python programming language. The goal is to find a set of clusters that contained a significant amount of details without dividing up the dataset into underwhelmingly small clusters or confusingly large clusters. Using the Silhouette method, 2, 4 and 5 seemed to be optimal candidates for the number of clusters as Figure 2 shows. Agglomerative clustering confirmed this view where the most gains was achieved by reducing the number of clusters to 2, 4, and 5 which increased the distance between clusters by 1.66, 1.55, and 1.41 respectively. Figure 3 shows the corresponding Dendrogram. A cluster size k= 2 was too small for analysis so it was discarded. Then, we evaluated k = 5 which generated clusters where 2 of them have a significant overlap. In contrast, the overlap between clusters was not an issue when k = 4. Hence, we selected 4 as the optimal number of clusters and proceeded with k-means clustering.

FIGURE 2
SILHOUETTE CHART

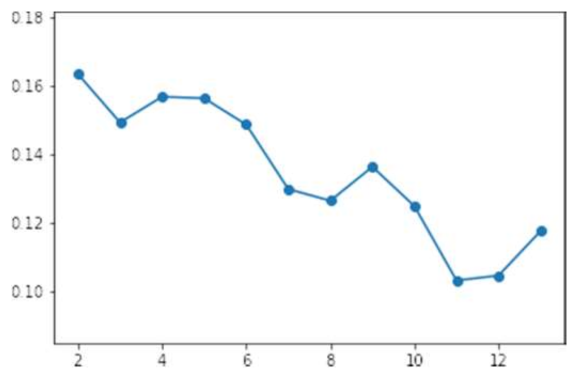
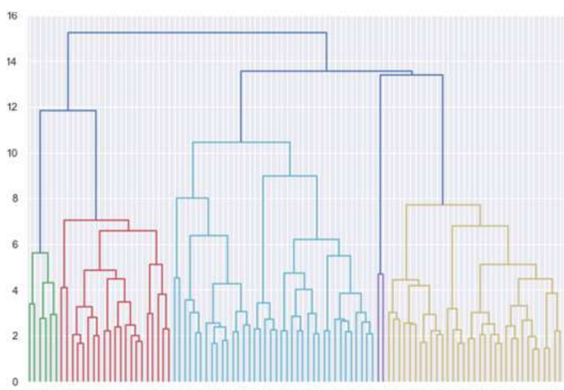


FIGURE 3
DENDOGRAM FOR AGGLOMERATIVE CLUSTERING



We then characterized the clusters and analyzed their patterns to determine if the top trending songs contained specific attributes that directly lead to their success. Using the established clusters, we looked at specific characteristics that result in a higher chance of trending songs. We also wanted to see if each of the clusters matched with a specific music genre, thus potentially providing us with information about the type of musical attributes that make up a specific genre.

TABLE 2
DESCRIPTIVE STATISTICS

Attribute	Year	Mean	SE	Median	SD	Kurtosis	Skewness	Range	Min.	Max.
Danceability	2017	0.70	0.01	0.71	0.13	1.52	-0.89	0.67	0.26	0.93
	2018	0.72	0.01	0.73	0.13	1.63	-0.83	0.71	0.26	0.96
Energy	2017	0.66	0.01	0.67	0.14	-0.83	-0.33	0.59	0.35	0.93
	2018	0.66	0.01	0.68	0.15	-0.19	-0.60	0.61	0.30	0.91
Loudness	2017	-5.65	0.18	-5.44	1.80	1.15	-0.88	9.07	-11.46	-2.40
	2018	-5.68	0.18	-5.57	1.78	-0.20	-0.46	7.73	-10.11	-2.38
Speechiness	2017	0.10	0.01	0.06	0.10	3.51	2.00	0.41	0.02	0.43
	2018	0.12	0.01	0.07	0.10	4.65	2.07	0.51	0.02	0.53
Acousticness	2017	0.17	0.02	0.11	0.17	1.11	1.33	0.69	0.00	0.70
	2018	0.20	0.02	0.11	0.22	1.96	1.62	0.93	0.00	0.93
Instrumentalness	2017	0.00	0.00	0.00	0.03	45.39	6.55	0.21	0.00	0.21
	2018	0.00	0.00	0.00	0.01	97.77	9.84	0.13	0.00	0.13
Liveness	2017	0.15	0.01	0.13	0.08	1.71	1.40	0.40	0.04	0.44
	2018	0.16	0.01	0.12	0.11	4.23	2.01	0.61	0.02	0.64
Valence	2017	0.52	0.02	0.50	0.22	-0.66	0.04	0.88	0.09	0.97
	2018	0.48	0.02	0.47	0.21	-0.67	0.04	0.85	0.08	0.93
Tempo	2017	119.20	2.80	112.47	27.95	0.21	0.88	124.85	75.02	199.86
	2018	119.90	2.88	120.12	28.80	-0.21	0.63	133.14	64.93	198.08

TABLE 3
CORRELATIONS

	Year	Danceability	Energy	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo
Danceability	2017	1.00								
	2018	1.00								
Energy	2017	-0.12	1.00							
	2018	-0.073	1.00							
Loudness	2017	0.04	**0.71**	1.00						
	2018	-0.016	**0.733**	1.00						
Speechiness	2017	0.09	-0.24	-0.46	1.00					
	2018	0.227	-0.074	-0.252	1.00					
Acousticness	2017	0.02	-0.25	-0.14	-0.05	1.00				
	2018	-0.134	-0.421	-0.270	-0.082	1.00				

	Year	Danceability	Energy	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo
Instrumentalness	2017	-0.03	0.10	-0.06	-0.09	-0.07	1.00			
	2018	-0.067	0.094	0.036	-0.070	-0.090	1.00			
Liveness	2017	-0.07	0.13	0.05	-0.03	-0.13	-0.04	1.00		
	2018	-0.039	0.051	0.000	-0.099	-0.150	-0.016	1.00		
Valence	2017	0.38	0.31	**0.42**	-0.13	0.11	-0.07	-0.01	1.00	
	2018	0.414	0.382	**0.408**	-0.051	-0.021	-0.095	-0.043	1.00	
Tempo	2017	-0.31	0.06	-0.13	0.19	-0.24	0.15	0.06	-0.26	1.00
	2018	-0.195	0.062	-0.035	0.103	-0.158	0.178	-0.108	-0.148	1.00

RESULTS AND DISCUSSION

To make sense of the clusters, we drew multiple scatter plots where one dimension was a musical attribute and the other dimension was the generated cluster labels (Figure 4). Among these musical attributes, key and mode did not seem to significantly vary across clusters so we left them out of the analysis. We applied the clustering model developed from year 2017 dataset to year 2018 dataset, to assign each of the 2018 top 100 songs to one of the four clusters from the year 2017. The clustering results for both years are summarized in Table 4 and 5.

For the year 2017 dataset, the largest cluster, Cluster#1, contained 47% of the songs and had the attributes of high danceability (average of 0.74), high loudness (average of -4.50), low instrumentalness (average of 0.00), high valence (average of 0.63), and low tempo (average of 107.33). Songs in this cluster are upbeat, joyful, danceable, and contain few spoken words. The low tempo in combination with the high danceability and low speechiness can also indicate that this cluster contains tracks that are driven by relaxing, redundant beats and rhythms, making them an easy, enjoyable listen.

Cluster#2, the second largest cluster with 27% of the songs, shares the high danceability with Cluster#1, resulting in a majority (72%) of the top trending songs having a danceable music structure. Also notable in Cluster#2 are the following attributes: low energy (average of 0.51), low loudness (average of -7.52), high acousticness (average of 0.26), low liveness (average of 0.13), and low tempo (107.09). Similar to the mellow track structures of Cluster#1, Cluster#2 also shares the laidback, rhythmic structure, with a bigger emphasis on the acoustic instrumentals (high acousticness).

Cluster#4, which is characterized by low danceability (average of 0.56), comparatively high speechiness (average of 0.14), low instrumentalness (average of 0.00), comparatively high liveness (average of 0.17), low valence (average of 0.39), and high tempo (average of 153.97), is comprised of a mix of rap, pop, and dance songs. While energetic through its high tempo, the songs in these clusters follow an unsophisticated structure that may contain redundant lyrics and hooks (high speechiness), and an overly-simplistic instrumental structure (low instrumentalness, low danceability, low valence) that makes an immediate, memorable impression on the audience. Overall, these clusters all consist of an overwhelming majority of Pop and Dance tracks from the trending list (71 out of top 100 songs). As a result, we could draw that the genres of Pop and Dance contain a successful, chart-topping musical structure that is high in loudness and low in speechiness (Figure 5).

TABLE 4
CLUSTER ANALYSIS RESULTS ATTRIBUTE – AVERAGES

Year 2017

Cluster	Average of danceability	Average of energy	Average of loudness	Average of speechiness	Average of acousticness	Average of instrumentalness	Average of liveness	Average of valence	Average of tempo	Count
1	0.74	0.74	-4.50	0.07	0.14	0.00	0.16	0.63	107.33	47
2	0.74	0.51	-7.52	0.13	0.26	0.00	0.13	0.42	107.09	27
3	0.65	0.84	-4.71	0.04	0.08	0.17	0.13	0.52	144.46	2
4	0.56	0.67	-5.90	0.14	0.11	0.00	0.17	0.39	153.97	24
Grand Total	0.70	0.66	-5.65	0.10	0.17	0.00480	0.15	0.52	119.20	**100**

Year 2018

Cluster	Average of danceability	Average of energy	Average of loudness	Average of speechiness	Average of acousticness	Average of instrumentalness	Average of liveness	Average of valence	Average of tempo	Count
1	0.77	0.64	-6.01	0.12	0.22	0.00	0.11	0.56	130.74	46
2	0.72	0.70	-4.71	0.10	0.15	0.00	0.12	0.43	101.64	26
3	0.86	0.55	-7.61	0.52	0.09	0.00	0.09	0.61	105.02	2
4	0.61	0.66	-5.90	0.09	0.21	0.00	0.28	0.41	120.15	26
Grand Total	0.72	0.66	-5.68	0.12	0.20	0.00158	0.16	0.48	119.90	**100**

TABLE 5
CLUSTER ANALYSIS RESULTS – GENRED AND RANKING

Year	Cluster #	Significant Attributes	Genres	Count	Median Ranking
2017	1	High Danceability High Loudness Low Instrumentalness High Valence Low Tempo.	Pop (26) Dance (12) R&B/Soul (2) Hip-Hop/Rap (6) Alternative (1)	47	45
	2	High Danceability Low Energy Low Loudness High Acousticness	Pop (10) Dance (4) Hip-Hop/Rap (8) R&B/Soul (4) Country (1)	27	48
	3	High Energy Low Speechiness Low Acousticness High Instrumentalness	Pop (1) Alternative (1)	2	39.5
	4	Low Danceability Comparatively High Speechiness Low Instrumentalness Comparatively High Liveness Low Valence High Tempo	Pop (13) Dance (6) Hip-Hop/Rap (4) R&B/Soul (1)	24	63
2018	1	Comparatively High Danceability High Acousticness High Instrumentalness Comparatively High Valence High Tempo	Alternative (1) Dance (17) Hip-Hop/Rap (13) Pop (14) R&B/Soul (1)	46	57
	2	High Energy High Loudness Low Valence Low Tempo	Alternative (1) Dance (4) Hip-Hop/Rap (11) Pop (8) R&B/Soul (2)	26	37.5
	3	High Danceability Low Energy Low Loudness High Speechiness Low Acousticness High Valence Low Instrumentalness Low Tempo	Pop (1) Hip-Hop/Rap (1)	2	66.5
	4	Low Danceability Low Speechiness High Acousticness Low Instrumentalness Low Valence	Alternative (2) Dance (5) Hip-Hop/Rap (11) Pop (8)	26	43.5

On the other hand, the smallest cluster, Cluster #3, containing only two songs ranked at 22 and 57, presented a significant attribute - high level of instrumentalness (average of 0.17) unseen in other clusters. As the only one of the four clusters that contained high instrumentalness, this small cluster potentially emphasizes that songs with a sophisticated and varyingly unique musical structure, such as songs in the Alternative genre, while representing a niche market with dedicated consumers, tend to not chart as well as songs with very redundant and easy to follow beat patterns, as well as catchy hooks/phrases, as seen with the more popular trending Pop or Dance genres.

For the 2018 dataset, with 46% of the dataset songs, Cluster#1 is unique for the following attributes: high acousticness (average of 0.22), high instrumentalness (average of 0.0033), high tempo (average of 130), and high time signature (average of 4.0217). The songs in this cluster lean towards the Dance and Pop genres, with a total of 31 combined songs from those two genres. Essentially dance and pop tracks that focus on a unique, memorable instrumental (high instrumentalness) and that also tend to have rapid rhythms and beats (high tempo) have a high chance of achieving charting success.

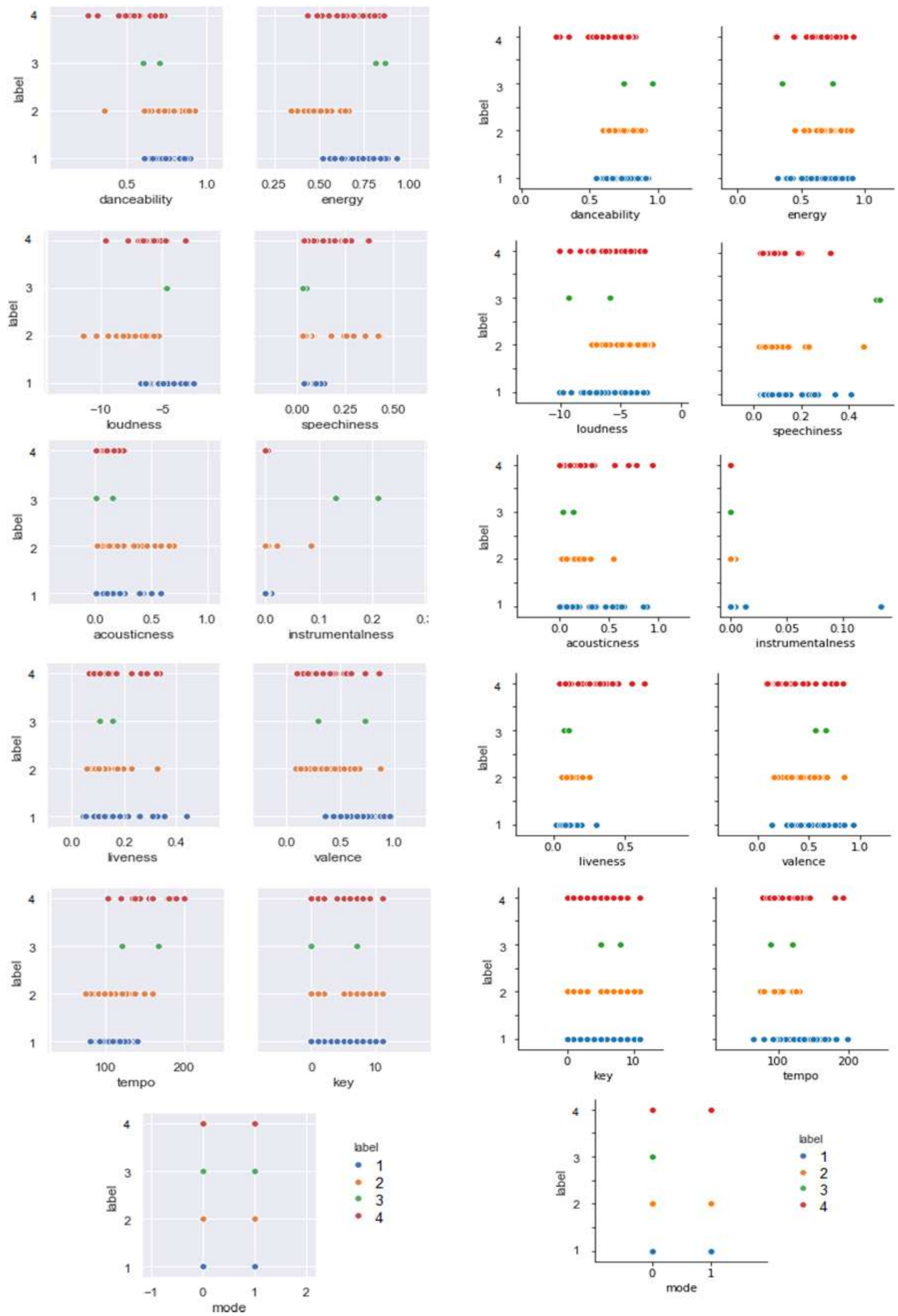
With 26% of the songs, Cluster#2 is unique for the following attributes: high energy (average of 0.6966), low key (average of 5.2308), high loudness (average of -4. 713), low tempo (average of 101.64), and high duration (average of 214 seconds). Unlike the previous cluster which focused on the electronic/dance related tracks of the chart, this cluster mainly categorizes Hip-Hop/Rap and Pop tracks. Cluster#2 is defined by highly energetic tracks that also contain a significant amount of melodies (high mode). The lyrical nor the instrumental content of these tracks are attributed to their success; instead it is the uniqueness of the tracks that are the reason for their placement on the chart, which could be due to a new artistic or creative approach by a well-defined artist who has already had prior placement on the chart. This is most notable with tracks like "All the Stars" by Kendrick Lamar (Ft. SZA), "Pray for Me" by The Weeknd and Kendrick Lamar, and "Back to You" and "Wolves" which are both by Selena Gomez.

Cluster#3 is the smallest of all the clusters, having only 2 songs and the following attributes: high danceability (average of 0.856), low energy (average of 0.5475), high key (average of 6.5), low loudness (average of -7.6125), low mode (average of 0), high speechiness (average of 0.523), low acousticness (average of 0.0885), low instrumentalness (average of 0), low liveness (average of 0.0897), high valence (average of 0.6105), and low duration (average of 182 seconds). Because this cluster has only two songs the overall averages for the described attributes are heavily skewed in comparison to the other clusters' attributes that are more well-defined due to the increased quantity of tracks per cluster. However, it is important to emphasize the only two tracks in this cluster are both Hip-Hip/Rap songs, which means that these results can help define a successful, chart-topping formula for that genre. Specifically, rap songs that focus heavily on lyrical content (high speechiness) and are complemented with a calm and simple instrumental (low loudness, low mode, low acousticness, low instrumentalness), potentially have a higher chance of achieving a few niche spots on a trending list. The high speechiness can also attribute for lyrical content that resonates well with the audience, or are simply very catchy, memorable, or enjoyable. Overall, the two songs in this cluster can either assist in defining a specific formula for rare trending Hip-Hop/Rap songs, or are simply two unique tracks that successfully charted in 2018.

The final cluster, Cluster#4, contains 26% of the songs and is defined by the following: low danceability (average of 0.6096), low speechiness (average of 0.08694), high liveness (average of 0.2799), low valence (average of 0.4089), low time signature (average of 3.923). Similar to Cluster#2, Cluster#4 also consists of mostly Hip-Hop/Rap and Pop songs. In comparison to the high energy attributes of Cluster#2, Cluster#4 nearly contrasts that, with songs that are quite mellow (low danceability) and a lot simpler in both lyrical content (low speechiness) and musical content (low time signature, low valence). With calmer and more slow-paced instrumentals, as well as more redundant lyrics, these songs are simpler in form, yet just as successful. This cluster serves as an opportunity to further define the Hip-Hop/Rap and Pop genres, which combined, account for 67 of the 100 songs. Because a wide range of attributes and versions for those types of songs have achieved chart-topping success, artists can continue to prioritize their unique and creative talents within those genres and still have a high chance of achieving chart-topping success.

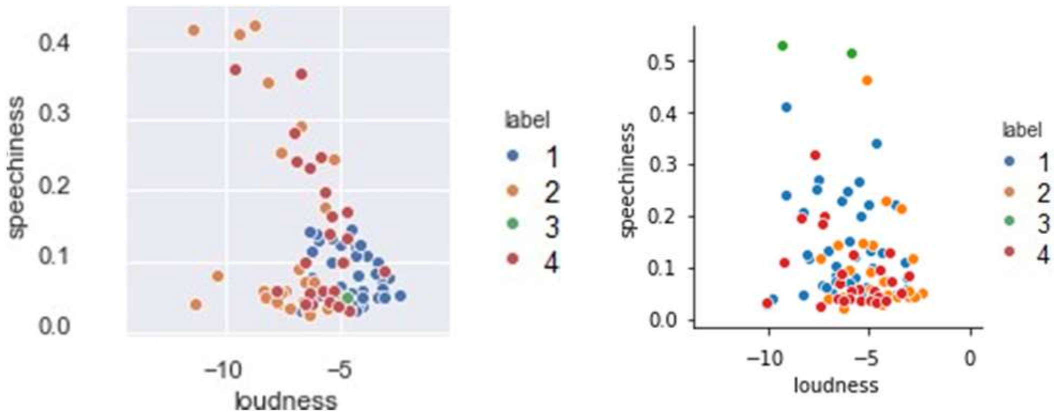
While some musical characteristics of the trending songs had changed in year 2018, including increased acousticness and decreased valence, other characteristics including danceability, energy, and loudness remained the same. The most trending music genres also changed from 2017's Pop and Dance to 2018's Hip-Hop Rap, but these three genres share the common strong rhythms and simple forms of music that the masses enjoy. Trend changes across time are natural due to a variety of factors including the constantly evolving musical landscape and overall tastes of listeners. Consistent with the descriptive statistics, trending songs in both years are low in speechiness (<0.33), liveness (<0.80), and instrumentalness (average 0.00). More importantly, the clustering analysis shows songs that are high in danceability and rank higher in the chart than the others.

FIGURE 4
DISTRIBUTION OF CLUSTERS ON DIFFERENT SONG ATTRIBUTE



(Left: 2017, Right: 2018)

FIGURE 5
SPEECHINESS VS LOUDNESS



(Left: 2017, Right: 2018)

CONCLUSION AND FUTURE WORK

The intention behind conducting a cluster analysis in this study was to automatically characterize trendy music based on the musical attributes defined by Spotify. We found clusters that not only vary in size but also contain a variety of significant attributes in each cluster. The completeness and homogeneity scores between clusters and genres were equal to 7.18% and 8.26% for year 2017 and equal to 4.13% and 4.78% for year 2018 respectively. These low scores indicate little overlap between genres and our clusters. This approach challenges the traditional music genres and provides new insight into how music can be automatically classified into different trending categories based on musical attributes and potentially provide better recommendations. The most popular songs tended to be the more exciting, radio-friendly songs that we all hear on our commute to work or while shopping at a supermarket. These songs follow a formulaic, pop-friendly sound, with a danceable music structure that tends to put the audience in a good mood.

While our results have shown us that certain musical attributes and song genres lead to long-term chart-topping success, we noticed the potential to further expand upon the reasoning behind the immediate success of a song. Specifically, we would like to incorporate the artist's name as part of the analysis. With regards to the 2018 dataset, the artists Post Malone, XXXTentacion, and Drake had a combined total of 16 songs on the Top 100 list. By devising a binary column that would emphasize the star power of a potential artist (0 = new artist, 1 = established, popular artist), we can further incorporate a clearly impactful attribute into the cluster analysis model. This would refine our results even further and help establish an even clearer and more distinguishable set of clusters. While subconsciously a listener may continuously enjoy a song for the more technical attributes associated with it, they may initially listen to a song based on who created it and whether they truly value that artist. As a result, it is important that we include that artist attribute in future research.

As future work, we will try to optimize our model and results with larger sample size, perform time series analysis and forecasting, and also explore additional attribute of trendy music across genre, culture, time, and whether those vary across different audience segments (e.g., age, location, social-economical class, etc.). In the long run, we will create a recommender agent that provides better discovery and personalization for both consumers and creators based on musical attributes and the clusters automatically generated from popular songs in the past.

REFERENCES

Bonnin, G., & Jannach, D. (2015, January). Automated Generation of Music Playlists: Survey and Experiments. *ACM Computing Surveys (CSUR) Surveys*, 47(2).

Chiru, C., & Popescu, O. (2017). Automatically Determining the Popularity of a Song in Lecture Notes. *Artificial Intelligence and Bioinformatics*, pp. 488-491.

Fields, B. (2011). *Contextualize Your Listening: The Playlist as Recommendation Engine*. A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of London. Retrieved from https://www.researchgate.net/profile/Benjamin_Fields2/publication/266271406_Contextualize_Your_Listening_The_Playlist_as_Recommendation_Engine/links/561e2d5108aef097132b32bc.pdf

Germain, A., & Chakareski, J. (2013, September 30-October 2). Spotify Me: Facebook- assisted automatic playlist generation. *Proceedings of the 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)* (pp. 25-28). Pula, Italy.

Herremans, D., Martens, D., & Sorensen, K. (2014, July). Dance Hit Song Prediction. *Journal of Music Research*, 43(3), 291-302.

Jacobson, K., Murali, V., Newett, E., Whitman, B., & Yon, R. (2016, September 15-19). Music Personalization at Spotify. *RecSys '16 Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 373-373). Boston, Massachusetts, USA.

Kanungo, T. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), 881- 892.

McFee, B., Bertin-Mahieux, T., Ellis, D., & Lanckriet, G. (2012, April 16–20). The Million Song Dataset Challenge. *Proceedings of AdMIRe'12 Workshop* (pp. 909-916). Lyon, France.

Pichl, M., Zangerle, E., & Specht, G. (2015, November 11-17). *Towards a Context-Aware Music Recommendation Approach: What is Hidden in the Playlist Name?* IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA.

Pichl, M., Zangerle, E., & Specht, G. (2016, December 11-13). Understanding playlist creation on music streaming platforms. In *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM)* (pp. 475-480). San Jose, California.

Rajyashree, R., Anand, A., Yash, S., & Mahajan, H. (2018). *Predicting Hit Music using MIDI features and Machine Learning.* The 3rd International Conference on Communication and Electronics Systems (ICCES).

Ramadhina Febirautami, L., Surjandari, I., & Laoh, E. (2018, July 20-22). Determining Characteristics of Popular Local Songs in Indonesia's Music Market. *Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE)*. Zhengzhou, China.

Shmueli, G. (2017). *Data mining for business analytics: concepts, techniques, and applications*. John Wiley & Sons.

Spotify. (2019). Retrieved from https://en.wikipedia.org/wiki/Spotify

Spotify. (n.d.). *Get Audio Features for a Track*. Retrieved from https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Teague, E. J. (2012). Saving the Spotify revolution: Recalibrating the power imbalance in digital copyright. *Case W. Res. JL Tech. & Internet*, 4, 207.