AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

SI 618: DATA MANIPULATION & ANALYSIS

Exploratory Data Analysis Project Report - 2

Fall 2022

MOTIVATION

There has been a revolution in the music listening market thanks to Spotify, which has dominated it. A staggering 125 million people subscribe to Spotify, and that doesn't even include those who use the app for free. Having a musical ear is an activity we all take part in everyday, so it's imperative that musicians and artists look at any trends and directions that listeners take in order to compete in the competitive environment. Based on Spotify data, we wish to create a regression analysis of attributes of songs that correlate with popularity. Using our data will allow others to consider the different factors that may affect the number of streams that an artist receives.

As part of the project, we intend to analyze different characteristics that might affect a song's popularity during a specific period of time. Due to rapid change in music tastes, it becomes crucial for new and established artists alike to have an in-depth understanding of what their audience digs.

Moreover, the project analyzed the top fifty artists of all time as well as popular artists during different decades. Additionally, it aims to classify songs into different categories (or clusters) based on how similar they are.

The three questions that the project aims to answer are:

1. Can the popularity of a song be attributed to certain characteristics? What changes have taken place in these characteristics over time?
2. In the history of music, which artists were most popular? From one decade to the next, who were the most liked or most favored artists?
3. Is it possible to separate song features into different clusters based on their characteristics? The clusters represent different categories, so what do they tell us?

THE DATASET

Data is plotted as 170653 rows divided into 13 numerical columns and 6 categorical columns. The dataset contains more than 160k songs gathered from Spotify Web API. A time element is also included in the dataset via the year and date when the song was released.

URL: Spotify-Data 1921-2020 | Kaggle

The numerical columns in the dataset are:

- **duration_ms**: The duration of the track in milliseconds.

- **acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **danceability**: Danceability 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**: Energy is 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.
- **instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- **liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **loudness**: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
- **speechiness**: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. 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**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **popularity**: The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time.
- **year**: The release year of track, (Ranges from 1921 to 2020)
- **id**: The Spotify ID for the track.

Similarly, the categorical columns in the dataset are:

- **key**: 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. If no key was detected, the value is -1.
- **release_date:** Date of release mostly in yyyy-mm-dd format, however precision of date may vary
- **name**: The title of the track
- **mode**: The binary value representing whether the track starts with a major (1) chord progression or a minor (0)
- **explicit**: The binary value whether the track contains explicit content or not, (0 = No explicit content, 1 = Explicit content)

DATA CLEANING & PREPROCESSING

The dataset used in the project was in .csv format. In order to analyze the dataset, it was simply downloaded from the source and uploaded into the project's Jupyter notebook using the pandas library.

At first, the data set had to be cleaned, which required checking for null values in the data. The data did not contain any null values, however, there was redundant information present in the dataset. Duplicate rows existed in the dataset, which was not interpreted because of unique values of the *'id'* attribute. After removing the *'id'* and '*release_date'* attributes, we were able to find duplicate rows, which were simply removed from our analyses.

Moreover, the 'artists' attribute contained multiple names of artists in the form of list. Therefore, the symbols were removed and only names were assigned using a comma. Additionally, another column *'decades'* was created to better analyze the data over a span of ten years.

Data set limitations include a limited number of nearly 2000 songs from each year. Songs in the older decades have a lower popularity metric (ranging 0-100) because they were played less on Spotify. Songs from the present decade have a higher popularity metric.

Task 1: Can the popularity of a song be attributed to certain characteristics? What changes have taken place in these characteristics over time?

Using the cleaned data, a correlation heatmap was plotted for the different numerical variables. This was done using the seaborn library. However, this correlation heatmap consisted the overall correlation of characteristics, but we know that these may very well vary according to different time period with changes in the musical industry or in people's choices and preferences. Therefore, a more rigorous analyses is necessary which considers the temporal data as well.

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

Before diving deeper into the analyses, a data-frame consisting of the correlation of various features such as *'duration_min', 'acousticness', 'danceability', 'energy', 'explicit', 'instrumentalness', 'key', 'liveness', 'loudness', 'mode', 'speechiness', 'tempo'* and *'valence'* was created with respect to the 'popularity' attribute.

This data-frame was then manipulated using *pivot_table()* function to convert rows into columns, which would help in visualization. However, due to the density of the data, a single line-plot visualizing the variation between the correlation of popularity with different characteristics was not interpretable. Therefore, it was decided to use a violin plot to account for the variation in correlation of popularity with individual features over time.

Finally, some features that were highly correlated with popularity were visualized using line plots and their inferences were noted down.

Task 2: In the history of music, which artists were most popular? From one decade to the next, who were the most liked or most favored artists?

This task required the data to be manipulated further, since the attribute 'artists' had the name of multiple contributors to one song in the same row, it had to be changed. The function *explode()* was used to break down the rows which contained multiple artists into different rows such that each row now contained only one artist. Finding the popular music artists of all time required the data to be grouped by the attribute *'artists'* which was further used to calculate the total sum of popularity for each artist. This data was visualized using a bar plot for the top fifty artists of all time.

Furthermore, for each decade a new data frame was created, which contained the popularity of each artist for that particular decade. These data-frames were used to create word clouds which represent the popularity of different artists of the decade. The artists which were most popular were represented in the word cloud by a larger and bolder font, whereas artists with lower popularity were represented using smaller font sizes.

Task 3: Is it possible to separate song features into different clusters based on their characteristics? The clusters represent different categories, so what do they tell us?

For this task, a new data-frame was created which consisted only of the characteristic attributes such as *'duration_min', 'acousticness', 'danceability', 'energy', 'explicit', 'instrumentalness', 'key', 'liveness', 'loudness', 'mode', 'speechiness', 'tempo'* and *'valence'*. The attributes *'duration_min'*, *'loudness'* and *'tempo'* were scaled using *MinMaxScaler()* because their high values could have affect our clusters. Moreover, Principal Component Analysis (PCA) was performed on the data which was followed by K-Means Clustering method which was used to find the optimal number of clusters that could be created.

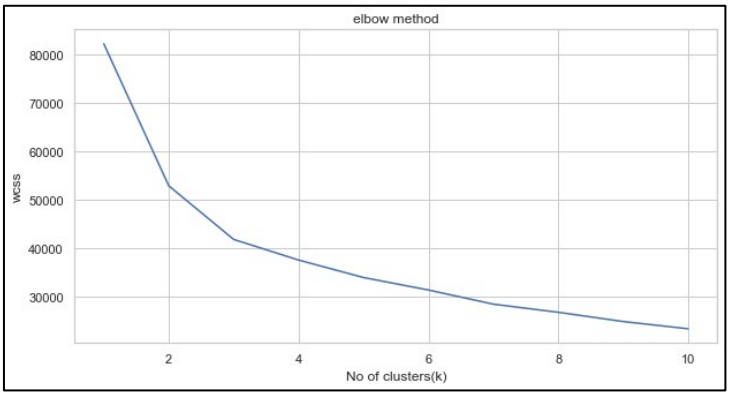

Fig. 1: Analyzing optimum number of clusters

The optimal number of clusters were calculated using the Elbow Method, a total of three clusters were

selected for our analyses. After applying the K-Means algorithm, the newly created cluster labels were then merged with the original data-frame to label the rows to their assigned clusters. Furthermore, a scatterplot was created to visualize the different clusters and a pivot table was created to analyze different characteristics of the points in these clusters.

ANALYSIS & RESULTS

Analysis was performed using libraries such as matplotlib, seaborn and plotly. The information was visualized using various techniques such as violin-plots, scatter-plots, heatmaps, line-plots and word clouds. The inferences taken out from the analyses are as follows,

Task 1: Can the popularity of a song be attributed to certain characteristics? What changes have taken place in these characteristics over time?

The main focus of this task is to analyze different characteristics that a song has with respect to popularity of the song. A heatmap was used to perform correlational analysis between different characteristics that a song could have. Moreover, since the heatmap collectively shows the correlational values for the entire time period from 1920s till 2020s, we used violin plots to analyze correlation between different characteristics and popularity over the decades.

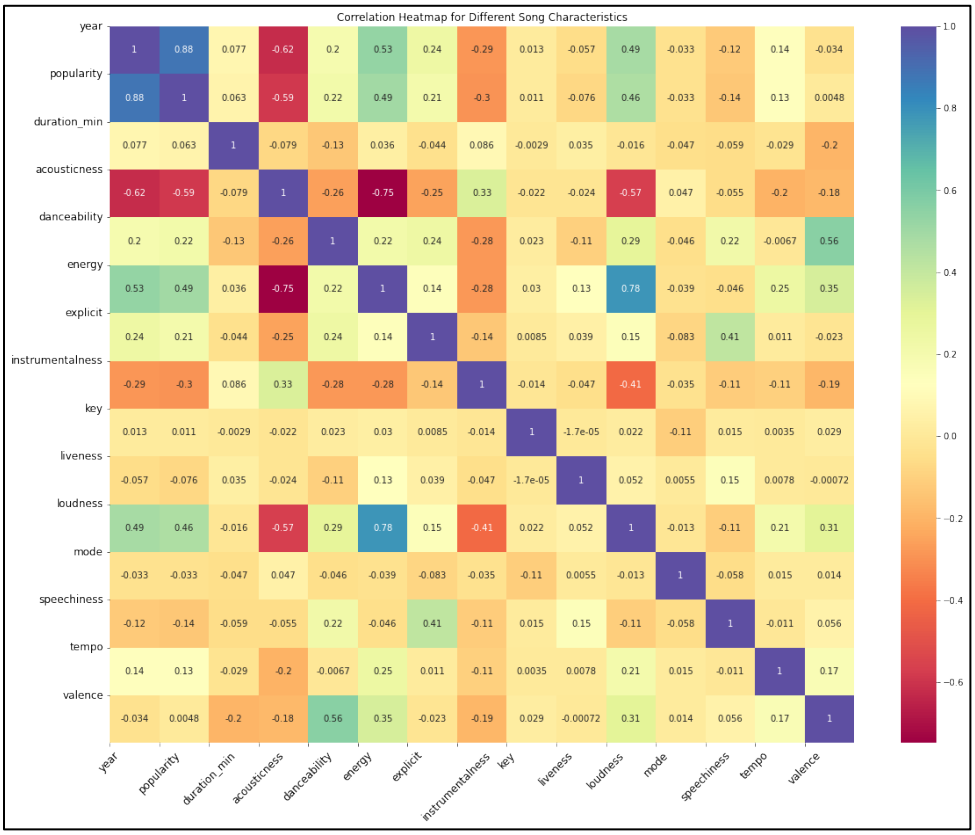

Fig. 2: Correlation Heatmap

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

Inferences:

In line with expectations, popularity increases with the release year. When Spotify computes a song's popularity, it not only considers the total number of streams the song has received, but also the frequency of its streams.

An effective correlation ratio of 0.49 indicates that energy plays a role in determining whether a song is popular. Despite the energetic nature of many popular songs, they may not be suited to dancing, as evidenced by the low correlation between dancing and the songs. It is still possible for a song with low energy to be popular despite its low energy.

Acousticness is the least correlated with popularity, with a score of -0.59, which is understandable since popular songs these days contain electric instruments or remixes. Compared to the vast majority of popular songs, you rarely hear an orchestra or pure acoustics playing them.

Moreover, we find that,

- There is a strong correlation between loudness and energy
- There is a negative correlation between acousticness and energy, loudness, and year.
- There is a strong correlation between valence and danceability.

According to this data, an artist with a high-energy song with electric instruments is more likely to be popular.

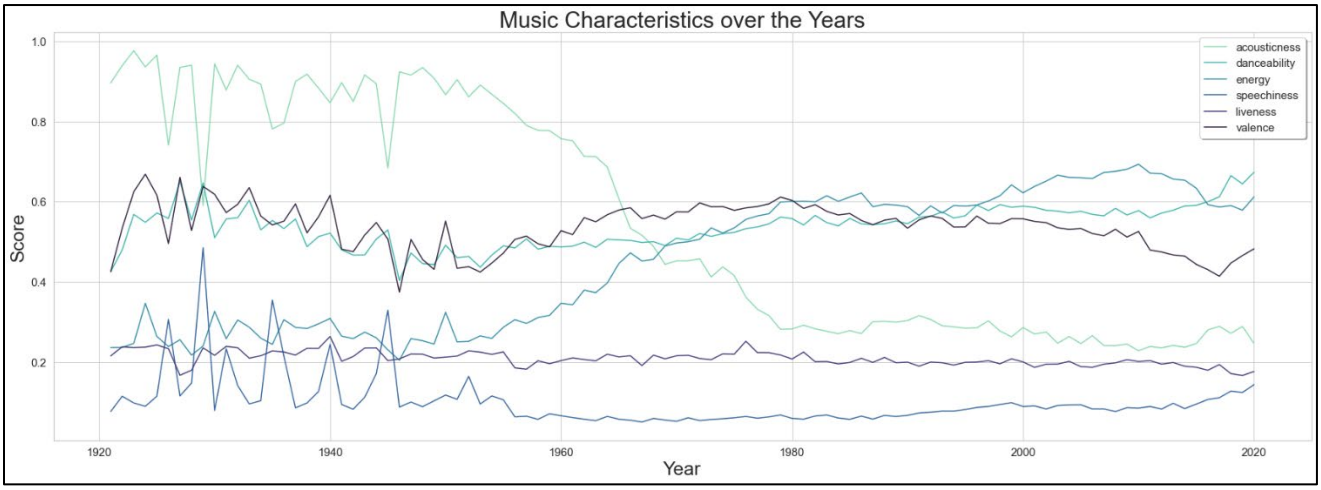

Fig. 3: Different characteristics over time

Inferences:

Prior to the 1960s, main stream music was highly acoustic. Music became more energetic after 1960, and danceability increased. With increasing energy, the level of acousticness fell. According to our hypothesis, this can be attributed to the rise of two specific music genres: hip-hop and electronic dance music (EDM). With the popularity of EDM and hip-hop, these songs focus on energy and hype, resulting in less acoustic content, as well as more energy, but fewer acoustic elements.

Danceability doesn't increase as energy increases, meaning people find both energetic music and acoustic music equally enjoyable.

Moreover, some other observations are,

- After 1960, the music market with a uniform score of speechiness seemed to have a steadily decreasing degree of speechiness before 1960. As a result of low mean scores for music with high speechiness since 1960, it's plausible that people didn't like it.
- Since the turn of the century, mean danceability has virtually remained the same. Due to the changes in acousticness and energy over the past decade, it is probably not correlated to these attributes.
- Inferring from this result, energy is probably inversely proportional to acousticness. It goes without saying that as energy increases, acousticness decreases.

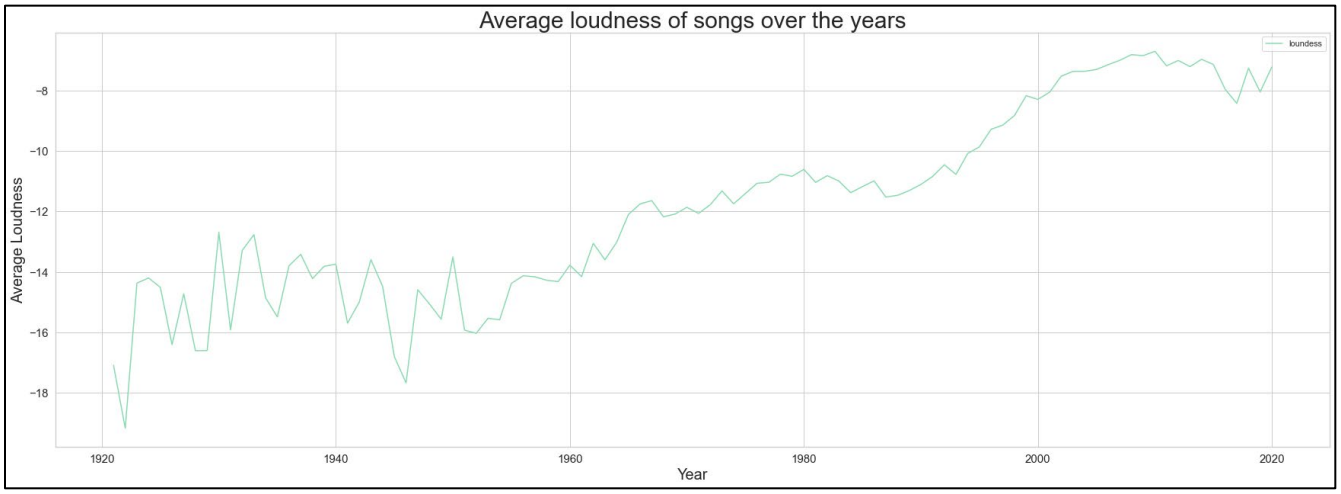

Fig. 4: Loudness over the years

Inferences:

In recent decades, music has become louder on the whole, as you can see in the graph. As people's tastes change, this can also reflect their musical preferences. Especially in the rap industry, people started to prefer music that was "loud" and had "heavy bass" because of the growth of EDM and hip-hop music. A possible reason for this could be an increase in loudness due to the development of recording technology, where adding more layers or more audio effects usually results in a louder sound.

Task 2: In the history of music, which artists were most popular? From one decade to the next, who were the most liked or most favored artists?

Bar plots and word clouds were used as visualization techniques for this task. As the musical preferences of the audience changes, it is natural for new artists to gain increasing popularity over a period of time. Libraries such as seaborn, wordcloud and collections were used in this task.

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

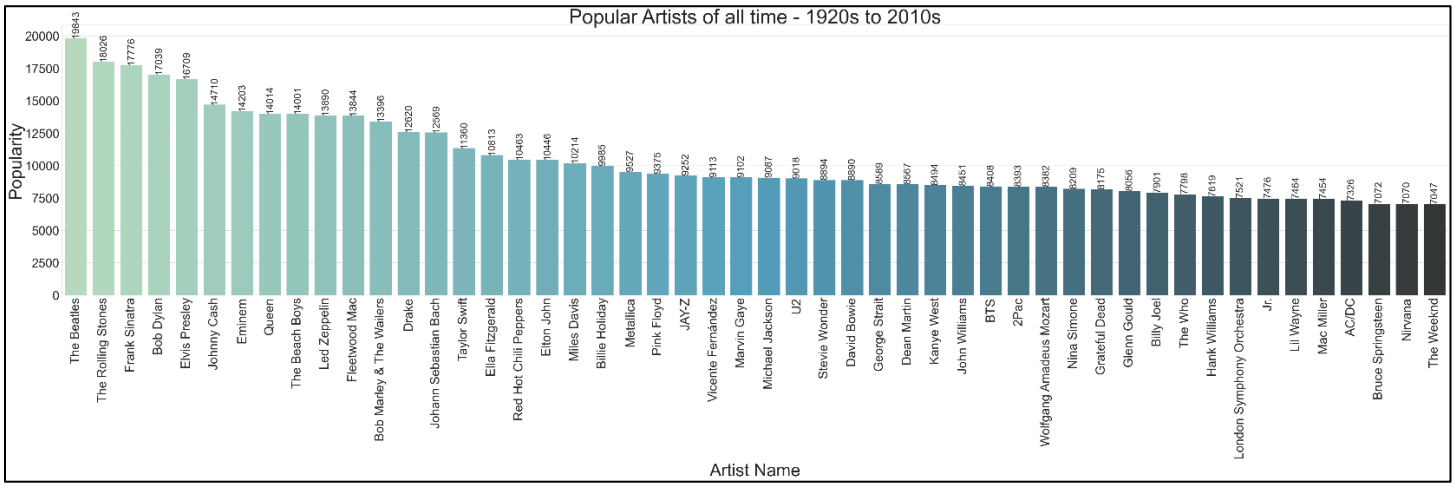

Fig. 5: Top 50 Artists & Bands of all time

Inferences:

The above bar plot depicts top fifty artists and bands of all time, and we can see that The Beatles leads the popularity chart, followed by The Rolling Stones and Frank Sinatra.

From the word clouds, we can see that,

- During the 1920's Louis Armstrong and his different orchestra's were dominating the popularity charts. Moreover, artists such as Frederic Chopin, Vladimir Horowitz, Fats Waller, Tommy Dorsey were also popular.
- In the 1930's artists such as Billie Holiday, Teddy Wilson, Robert Johnson were more prominent.
- Eric Sattie, Doris Day, Artur Rubinstein and Frank Sinatra were a few prominent artists, pianists and composers from the 1940's.
- 1950's saw the rise of king Elvis, Miles Davis, Glenn Gould and Ella Fitzgerald. Frank Sinatra still remained one of the popular artists.
- During the 1960's bands such as The Bach Boys, The Beatles and The Rolling Stones gained popularity. Moreover, the famous Bob Dylan came into the picture during this period.
- The 1970's was flooded with artists and bands such as Queen, Bob Marley, Billy Joel, Led Zeppelin, Eagles and Elton John.
- Musical bands such as Metallica, The Smiths and U2 gained popularity in the 1980's along with artists such as Johann Sebastian Bach and Prince.
- 1990's saw the rise of Nirvana, 2Pac, Sublime, Green Day and The Notorious B.I.G.
- During the 2000's Eminem, Kanye West, Taylor Swift, JAY-Z, John Mayer and Red Hot Chilli Peppers dominated the charts.
- During the 2010's bands such as BTS, The Weeknd and One Direction gained popularity. Moreover, artists such as Drake, Mac Miller and Lana Del Rey also dominated the charts.

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

Fig. 6: Word Cloud depicting popular artists over the decades

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

Task 3: Is it possible to separate song features into different clusters based on their characteristics? The clusters represent different categories, so what do they tell us?

Scatter plots were used to differentiate between the clusters visually. Using the elbow method, we were able conclude that the optimum number of clusters can be three. After performing K-Means clustering on the features, the scatter plot obtained was as follows,

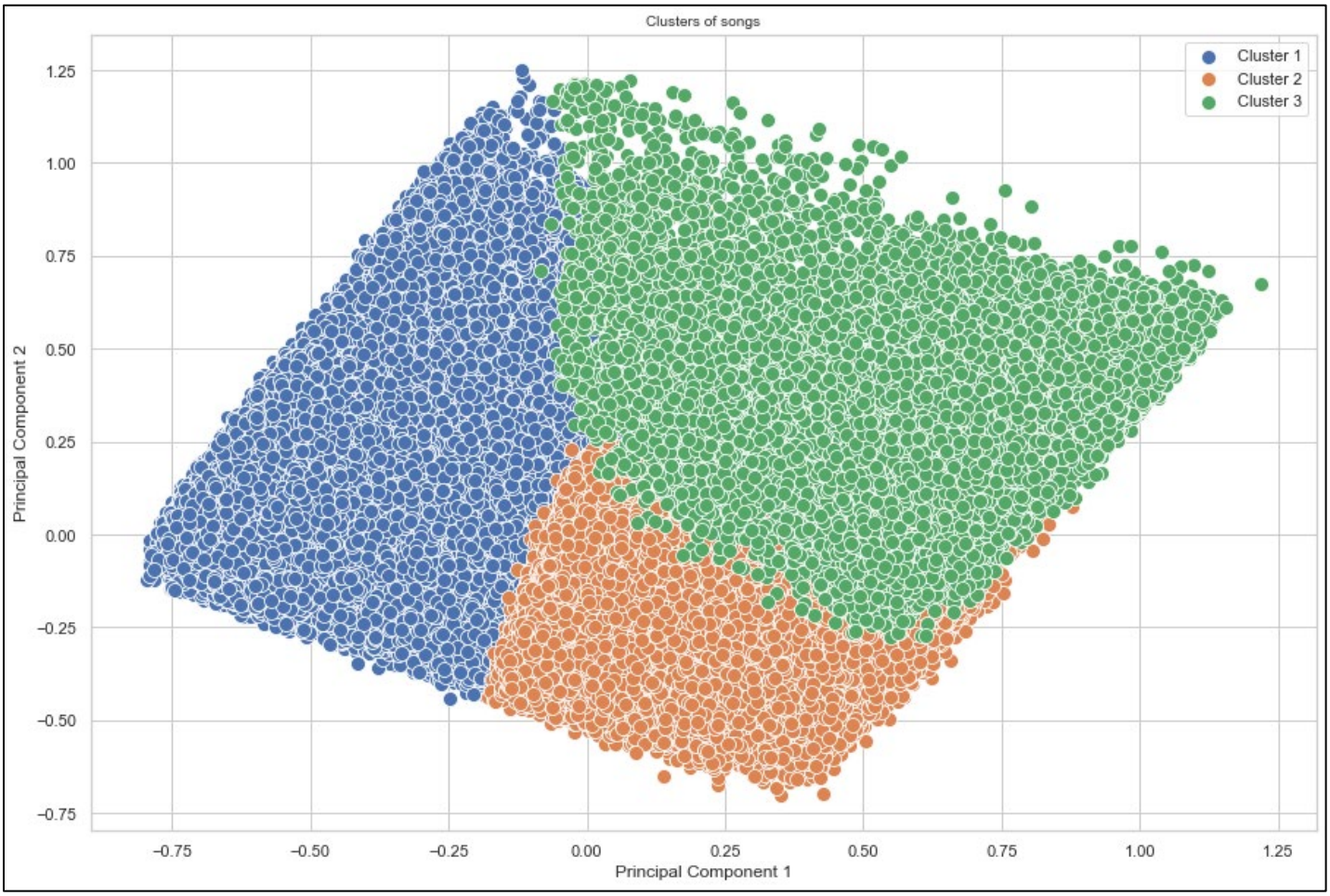

Fig. 7: Clusters from K-Means Clustering

cluster	acousticness	danceability	duration_min	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year
0	0.149306	0.590877	3.984819	0.690461	0.152174	0.061093	5.281173	0.210044	-8.145097	0.687838	43.336087	0.099491	122.971435	0.599896	1991.898168
1	0.794206	0.516951	3.553933	0.317673	0.028691	0.030269	5.153379	0.211735	-13.139477	0.746948	22.633480	0.100983	112.970112	0.498186	1965.548865
2	0.879209	0.419237	4.186852	0.248790	0.002373	0.814395	5.044685	0.183343	-17.504190	0.682902	15.245747	0.059703	107.098789	0.396694	1958.158227

Fig. 8: Properties of different clusters

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

Inferences:

From the above clusters, we were able to analyze the following:

- **Cluster 0**: These songs have a high loudness, energy, and tempo, while having a low acousticity and instrumentality. They have the highest valence of all clusters. They are certainly the most dynamic cluster. They may be in the EDM, hip-hop, country, or country music genres that mainly use electrical elements.
- **Cluster 1:** Due to the liveness and speechiness indexes being so high, it appears that the songs in this cluster were most likely played live. This cluster also exhibits high acoustic qualities. There is a possibility that this is pop music.
- **Cluster 2**: This is a list of songs with the highest number of acousticness as well as instrumentals. This cluster has one of the slowest music styles that we have seen so far. It is possible for these songs to be lullabies as well as natural songs.

Songs in Cluster 0:

	name	artists
8052	Honsool	Agust D
8053	Happy Does	Kenny Chesney
8055	Como Llora	Juanfran
8056	2020 Riots: How Many Times	Trey Songz
8059	If You're Too Shy (Let Me Know)	The 1975
8060	Shimmy	Aminé
8061	Money Talk (feat. YoungBoy Never Broke Again)	Rich The Kid, YoungBoy Never Broke Again
8062	Hold Me Close (feat. Ella Henderson)	Sam Feldt, Ella Henderson
8066	No Judgement	Niall Horan
8067	Wetty	Fivio Foreign

Fig. 10: Songs belonging to cluster 0

Songs in Cluster 1:

	name	artists
8051	homecoming queen?	Kelsea Ballerini
8054	Sad Forever	Lauv
8057	midnight love	girl in red
8058	Copy Cat (feat. Tierra Whack)	Melanie Martinez, Tierra Whack
8063	I miss you, I'm sorry	Gracie Abrams
8064	Still Softish	Josh Richards, Bryce Hall
8065	327	Westside Gunn, Joey Bada$$, Tyler, The Creator...
8073	That's Tuff (feat. Quavo)	Rich The Kid, Quavo
8099	River Of Tears	Alessia Cara
8104	Girls Need Love	Summer Walker

Fig. 9: Songs belonging to cluster 1

Songs in Cluster 2:

	name	artists
16141	In Peace	Donato Manna
16159	Onthou	Ever So Blue
16209	Heavy Rain & Gentle Thunder	Epic Soundscapes
32110	Fracture	Stephan Moccio
32117	Sagittabondo	Maura Bellucci
32138	Heavenly Harps	Matooma
32146	Isla De Flores	Berlioz
32164	Continuo	Mauro Cangemi
32179	Parts Unknown	Night Poets
32192	Buoyant	Keira Barton

Fig. 11: Songs belonging to cluster 2

AKSHAT PORWAL (akshatp)
akshatp@umich.edu
December 8, 2022

REFERENCES

- Web API Reference | Spotify for Developers
- Plotly Python Graphing Library
- Generating Word Cloud in Python - GeeksforGeeks
- ddhartma/Spotify-dataset-analysis-160kTracks-1921-2020: A song popularity study of 160k Spotify tracks between 1921-2020 via descriptive stats and Linear Regression (github.com)
- Spotify Wrapped: Data Visualization and Machine Learning on Your Top Songs | by Adam Reevesman | Towards Data Science
- Is my Spotify music boring? An analysis involving music, data, and machine learning | by Juan De Dios Santos | Towards Data Science
- What Makes a Song Likeable?. Analyzing Spotify's Top Tracks Of 2017… | by Ashrith | Towards Data Science