IDS 561 Analytics for Big Data

Recommendation System and Trend Analysis on Spotify

A detailed analysis of music marketing through Big Data

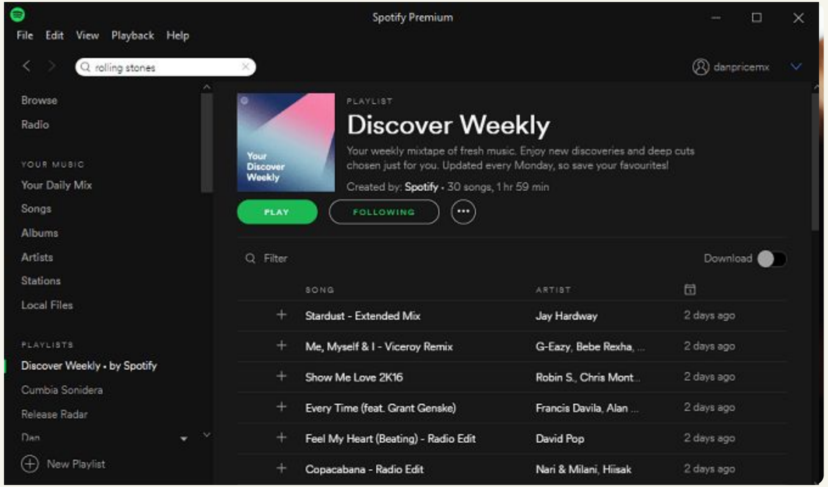
Amin Abbasi/Matthew Yuan/Chia-Hsuan Lin/Bella Chen

What Is The Target?

- **Problem Statement: Managing music data overload for personalized experiences.**

- **Key Goal: Predicting and enhancing music recommendations.**

- **Relevance: Optimizing user engagement and revenue in the online media industry.**



Motivation

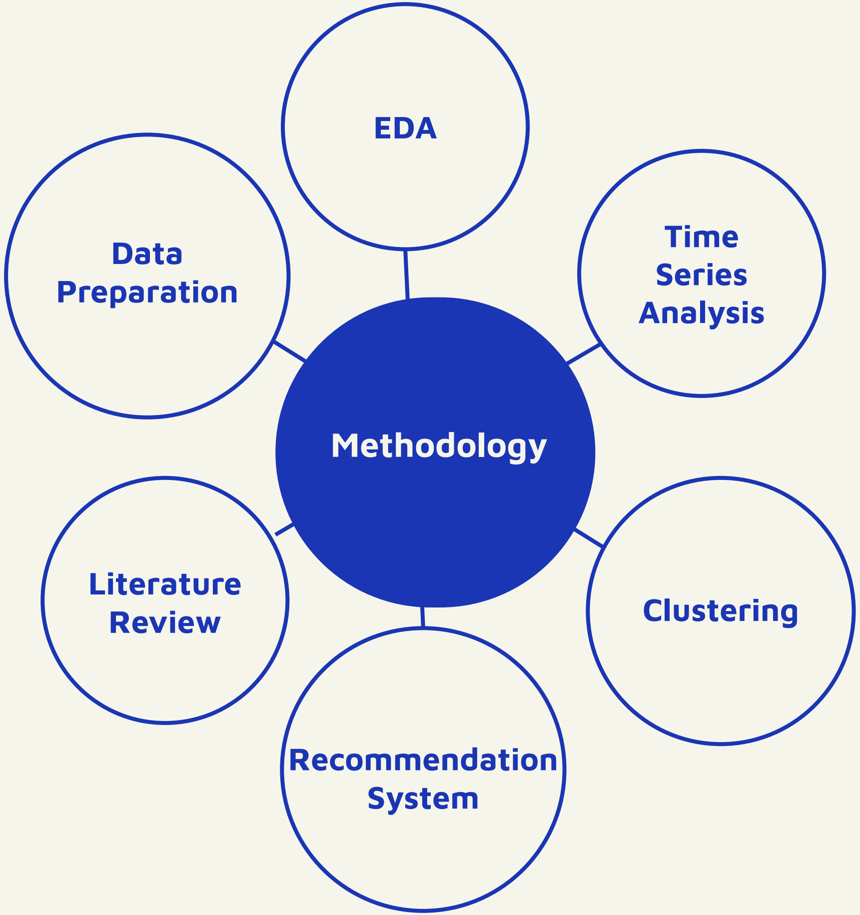
- **Challenge**: Vast datasets exceed listener capacity.
- **Objective**: Create personalized, efficient recommendation systems.
- **Impact**: Enhanced user satisfaction and business ROI.

Methodology Overview

Steps: Data Preparation, EDA, Time Series Analysis, Clustering, Recommender System.

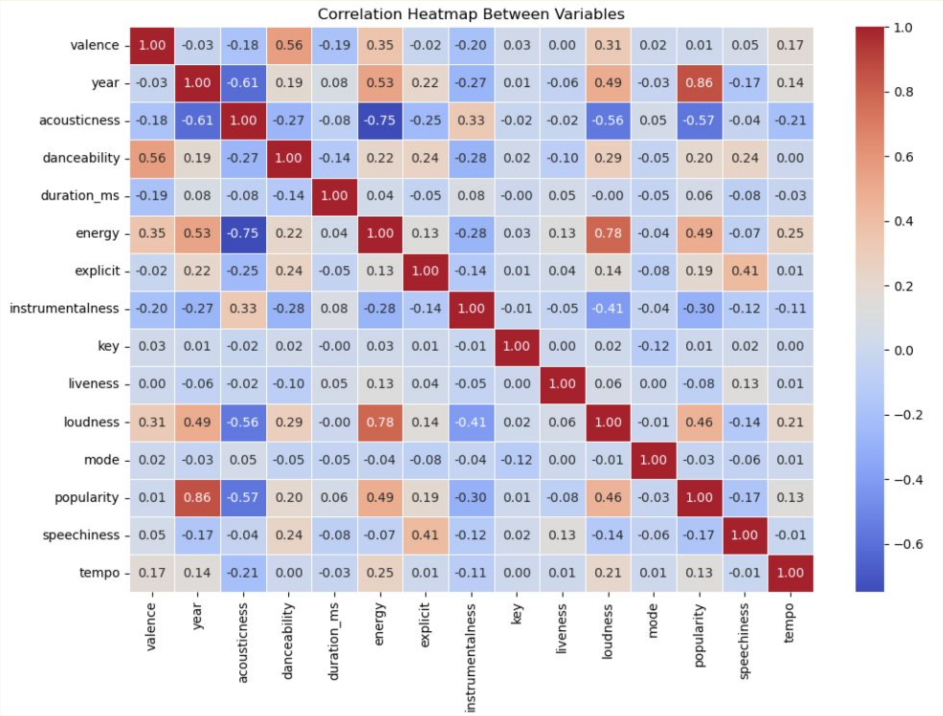
Datasets Used:
We have collected out data from GitHub Platform

- data.csv (170,653 entries, 19 columns)
- data_by_year.csv (100 entries, 14 columns)
- data_by_genre.csv (2972 entries, 14 columns)



Exploratory Data Analysis (EDA)

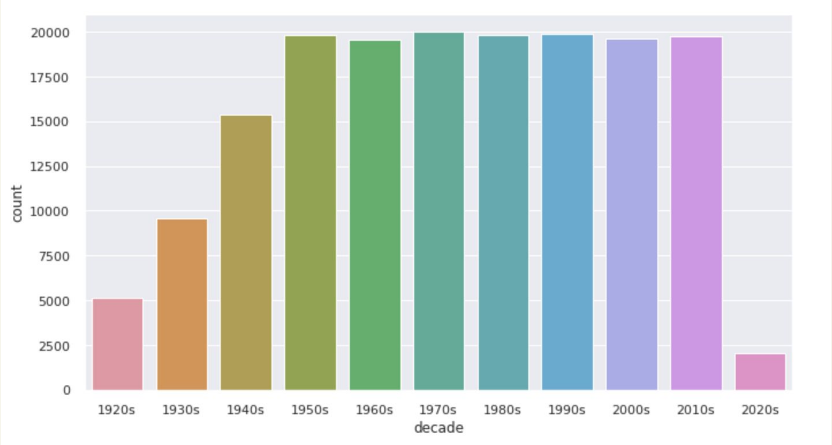
- The correlation heatmap shows the relationships between numerical variables in the Spotify dataset
- 1 indicates a perfect positive correlation



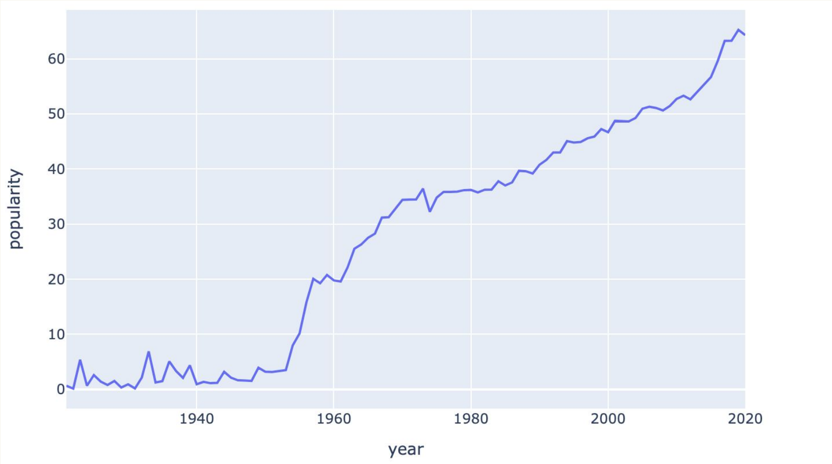
Correlation Matrix between variables

Exploratory Data Analysis (EDA)

- Music over time: Using the data grouped by year, we can understand how the overall sound of music has changed from 1920 to 2020



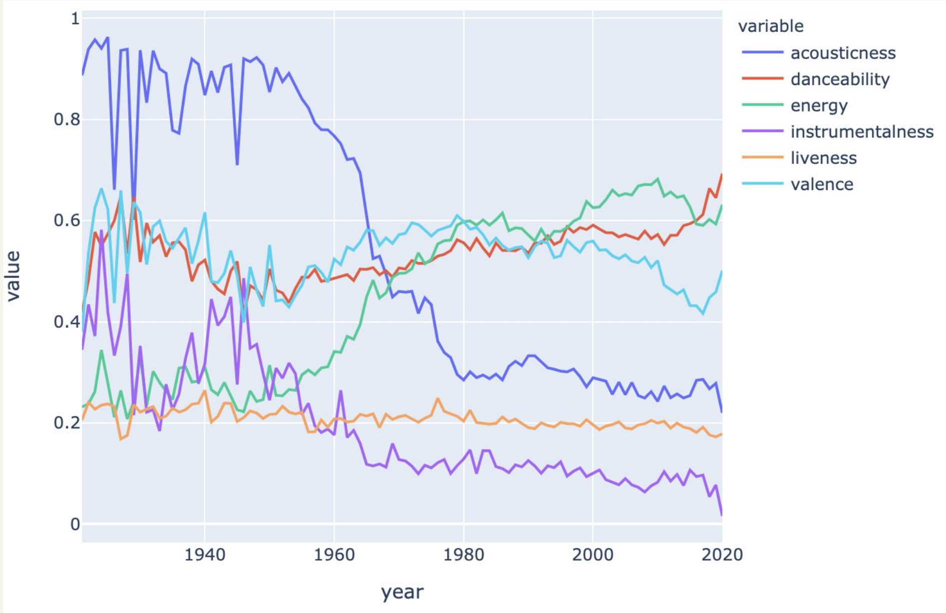
Number of songs per decade



Popularity trends over years

Exploratory Data Analysis (EDA)

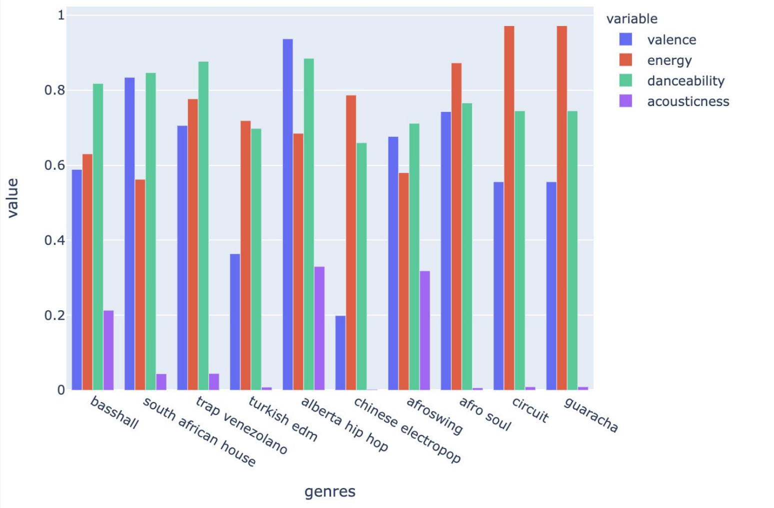
- The graph indicates the trend over years in each variable, acousticness and instrumentaines show a downward trend year by year.
- On the other hand, energy shows a upward trend year by year.



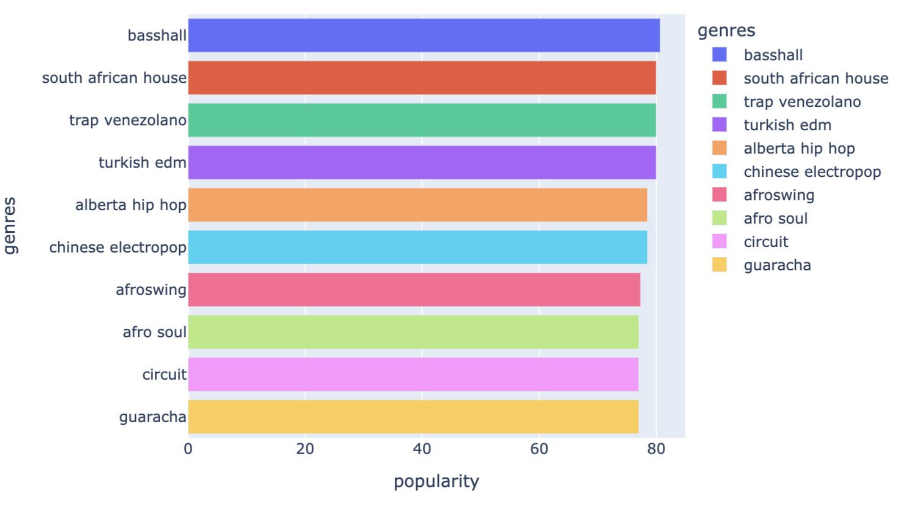
Variable trends over years

Exploratory Data Analysis (EDA)

- **Characteristics of different genres: compare different genres and understand their unique differences in sound.**

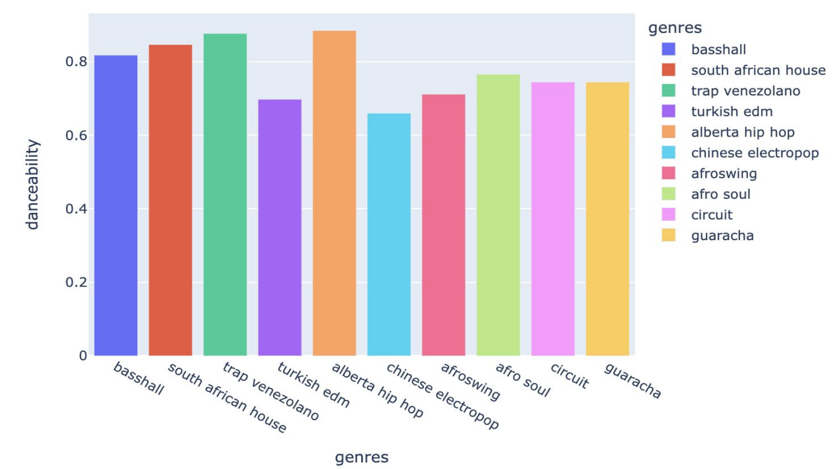


Characteristics of different genres

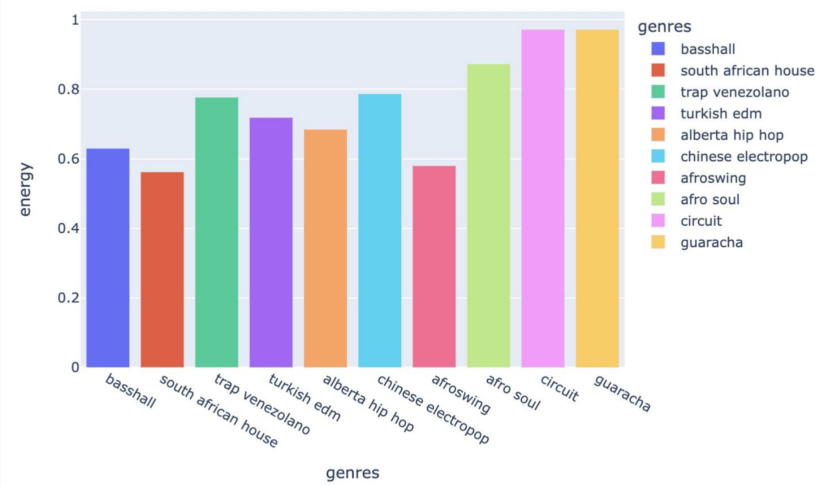


Top genres by popularity

Exploratory Data Analysis (EDA)

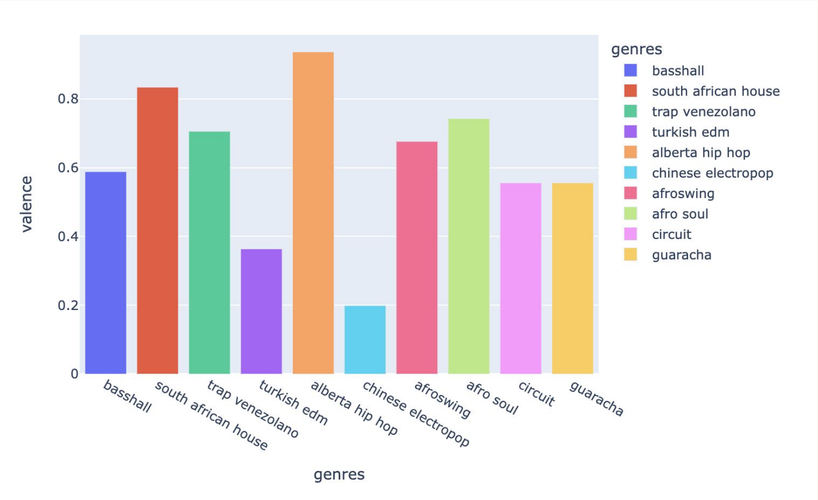


Danceability distribution for top 10 popular genres

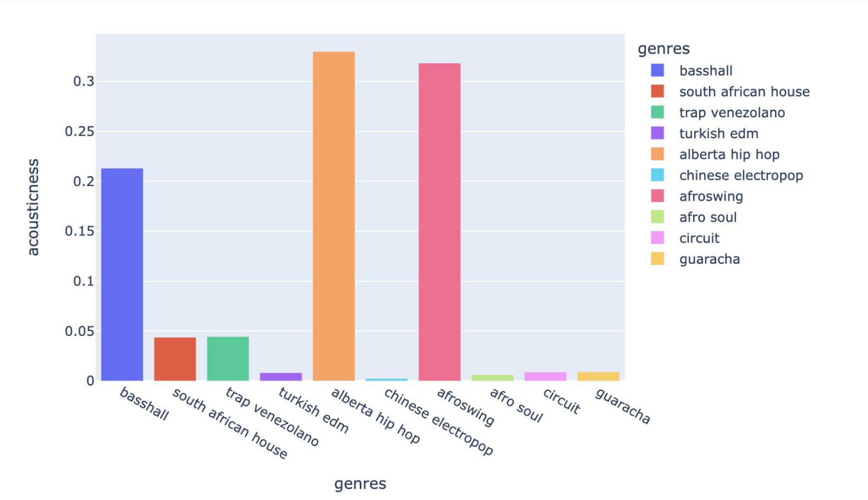


Energy distribution for top 10 popular genres

Exploratory Data Analysis (EDA)

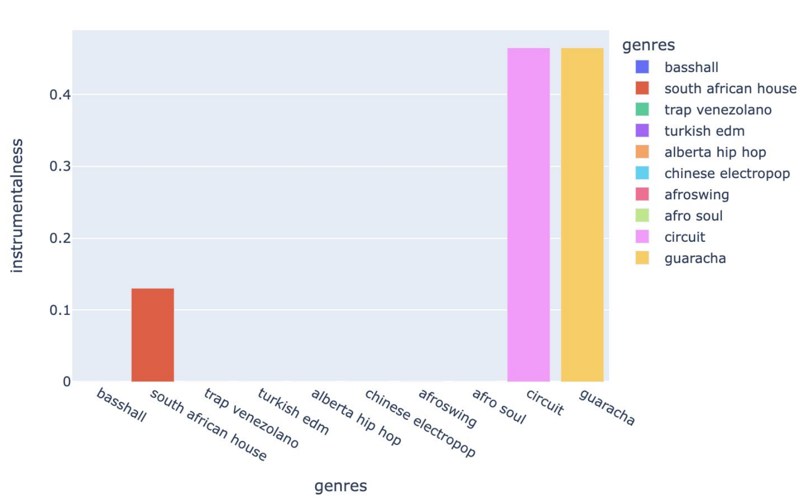


Valence distribution for top 10 popular genres

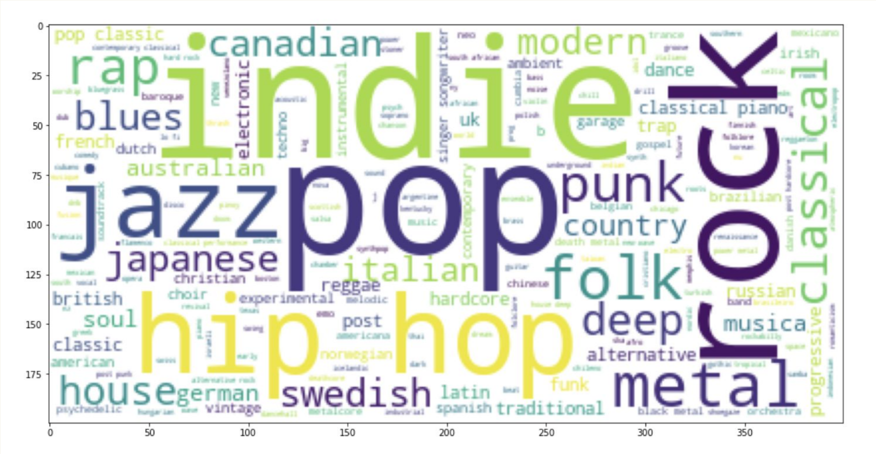


Acousticness distribution for top 10 popular genres

Exploratory Data Analysis (EDA)



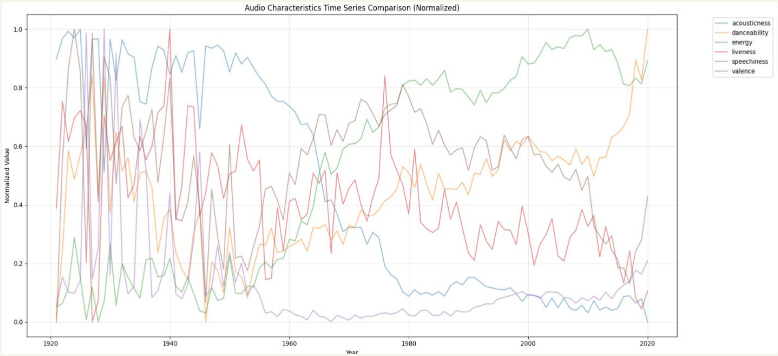
Instrumentalness distribution for top 10 popular genres

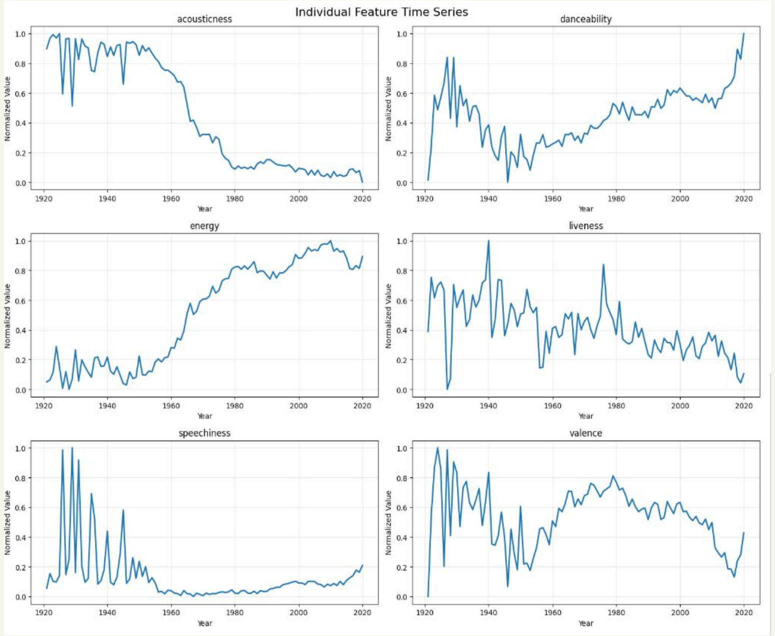


Genre Word Cloud

Time Series Analysis

- **Moving Average Method is applied in our model to uncover interesting patterns and trends within the data.**
- **Trends in Music Features Over Time**
- These trends highlight changes in musical styles, listener preferences, and production techniques over time.
- **Focus on Modern Music Characteristics**: Recent trends show a preference for high energy and danceable music, reflecting increased consumer demand for upbeat and lively tracks.





Time Series Analysis

Seasonality Analysis:

Feature	Seasonal Strength	Pattern
acousticness	0.060	Weak
danceability	0.122	Moderate
duration_ms	0.149	Moderate
energy	0.040	Weak
instrumentalness	0.084	Weak
liveness	0.221	Moderate
loudness	0.076	Weak
speechiness	0.220	Moderate
tempo	0.113	Moderate
valence	0.145	Moderate
popularity	0.038	Weak

- Seasonality Analysis of music
- Forecasting Analysis (Use Model Error Metrics Comparison ARIMA、Prophet、Exponential Smoothing Model)

Model Error Metrics Comparison Across Features:

MSE Comparison:

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
ARIMA	0.0004	0.0026	5.010809e+08	0.0022	0.0007	0.0002	0.3087	0.0006	5.0675	0.0048	33.1339
Prophet	2.2941	0.6965	7.470343e+10	0.9915	0.4443	0.0285	80.6431	0.0246	30696.2446	3.1604	1863.7915
Exponential Smoothing	0.0022	0.0022	5.835300e+08	0.0074	0.0030	0.0001	1.7321	0.0020	13.1469	0.0045	31.4390

RMSE Comparison:

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
ARIMA	0.0188	0.0511	22384.8368	0.0469	0.0263	0.0127	0.5556	0.0244	2.2511	0.0693	5.7562
Prophet	1.5146	0.8346	273319.2783	0.9957	0.6665	0.1687	8.9801	0.1567	175.2034	1.7778	43.1717
Exponential Smoothing	0.0466	0.0468	24156.3658	0.0859	0.0545	0.0109	1.3161	0.0452	3.6259	0.0672	5.6071

MAE Comparison:

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
ARIMA	0.0149	0.0349	16296.6344	0.0385	0.0223	0.0105	0.4311	0.0173	1.7380	0.0622	4.4363
Prophet	1.2484	0.6554	220434.8015	0.7732	0.4939	0.1261	6.6892	0.1441	135.5109	1.3982	35.1966
Exponential Smoothing	0.0359	0.0344	18419.5836	0.0731	0.0466	0.0097	1.1116	0.0378	2.7797	0.0618	4.6300

MAPE Comparison:

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
ARIMA	5.7530	5.4262	7.8215	6.3027	47.0175	5.7198	5.6244	14.7428	1.449	13.8851	7.2139
Prophet	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Exponential Smoothing	13.3466	5.4250	8.7037	11.9119	58.7935	5.2095	14.6300	35.3128	2.319	13.7228	7.6255

R2 Comparison:

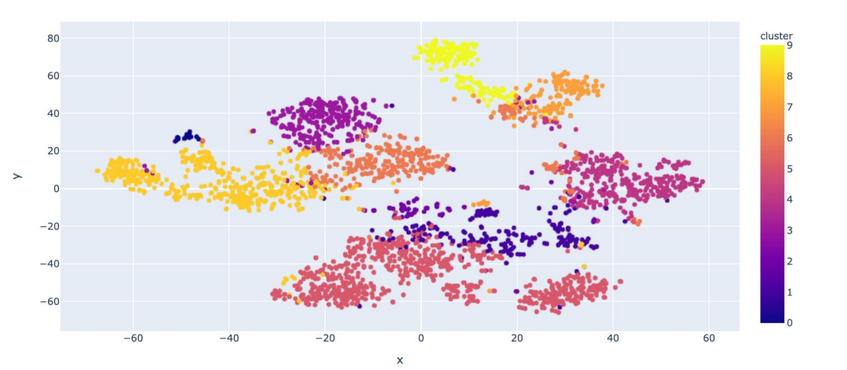
	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
ARIMA	-0.0569	-0.5067	-0.6772	-1.4029	-0.2075	-0.4911	-0.3939	-0.6348	-0.2947	-4.1259	-0.3507
Prophet	-6885.6043	-400.2150	-249.0423	-1082.8301	-773.3792	-261.2447	-363.1708	-66.4845	-7841.3448	-3370.5429	-74.9789
Exponential Smoothing	-5.5224	-0.2639	-0.9532	-7.0732	-4.1832	-0.0989	-6.8221	-4.6236	-2.3588	-3.8162	-0.2816

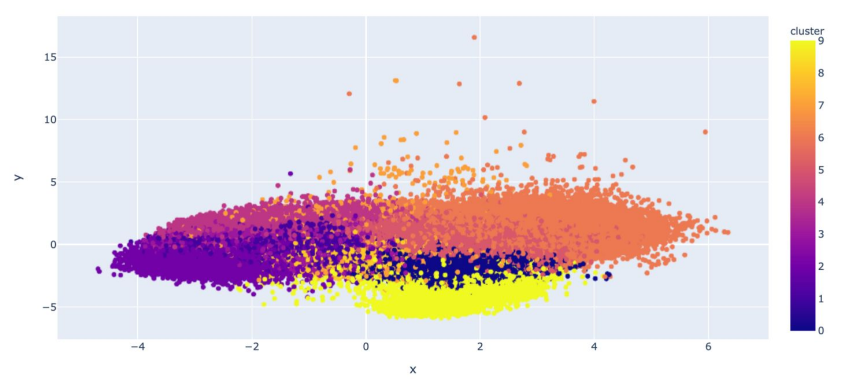
Clustering with K-Means

- **Goal**:
  - Group songs or genres into clusters based on their features (like energy, tempo, danceability).
  - identify patterns or similarities between songs

- **Process**:
  - PySpark and scikit-learn libraries
  - the Silhouette score, which shows how well-separated the clusters are.
  -
  - A specific number of clusters (e.g., 5 or 10) was finalized based on trial and error and performance metrics.
- **Result**:

Songs in the same cluster share similar characteristics (e.g., similar tempo or energy levels).This grouping helps recommend similar songs to users based on their preferences.





Recommendation System

```
recommend(100)
```

	prediction	rating	count
8602	0	5	1017.0
25406	1	5	953.0
24271	1	5	817.0
18762	0	5	528.0
7380	1	5	526.0

- Based on the K mean clustering, clusters of genres are categorized.
- The project used attribute-based recommendations, meaning the system compares a user's chosen song attributes (like genre or energy) with those of other songs.
- The matrix shows the top recommendations made by the recommender system given high rating and playcount.
- With which the Attributes like ratings and playcount are taken into consideration giving out the track ids

Results

- Outcome: Highly accurate recommendations that align with user preferences.
- Validation: Compared recommended genres with user listening history.
- Insights: Trends in genre popularity and evolving characteristics like energy and acousticness

By combining time series analysis, clustering, and our recommendation model, we delivered accurate song suggestions aligned with user preferences, validated by listening history. Our analysis also uncovered trends in genre popularity and shifts in song characteristics like energy and acousticness, showcasing how Big Data drives innovation on platforms like Spotify.

Thank You!