SPOTIFY TOP HITS ANALYSIS

USING POWERBI AND TABLEAU

OUR TEAM

Bence Danko
Id: 015179996

Saumya Varshney
Id: 017417283

Sreenidhi Hayagreevan
Id: 018195489

Victor Dumaslan
Id: 014407289

BUSINESS PROBLEM

OVERVIEW:

The music industry underwent significant changes over decades, driven by the rise of streaming platforms. Understanding what makes a song popular can help artists, producers, and marketers optimize for success.

KEY OBJECTIVE:

To analyze trends and attributes that contribute to the popularity of tracks on Spotify.

BUSINESS QUESTIONS

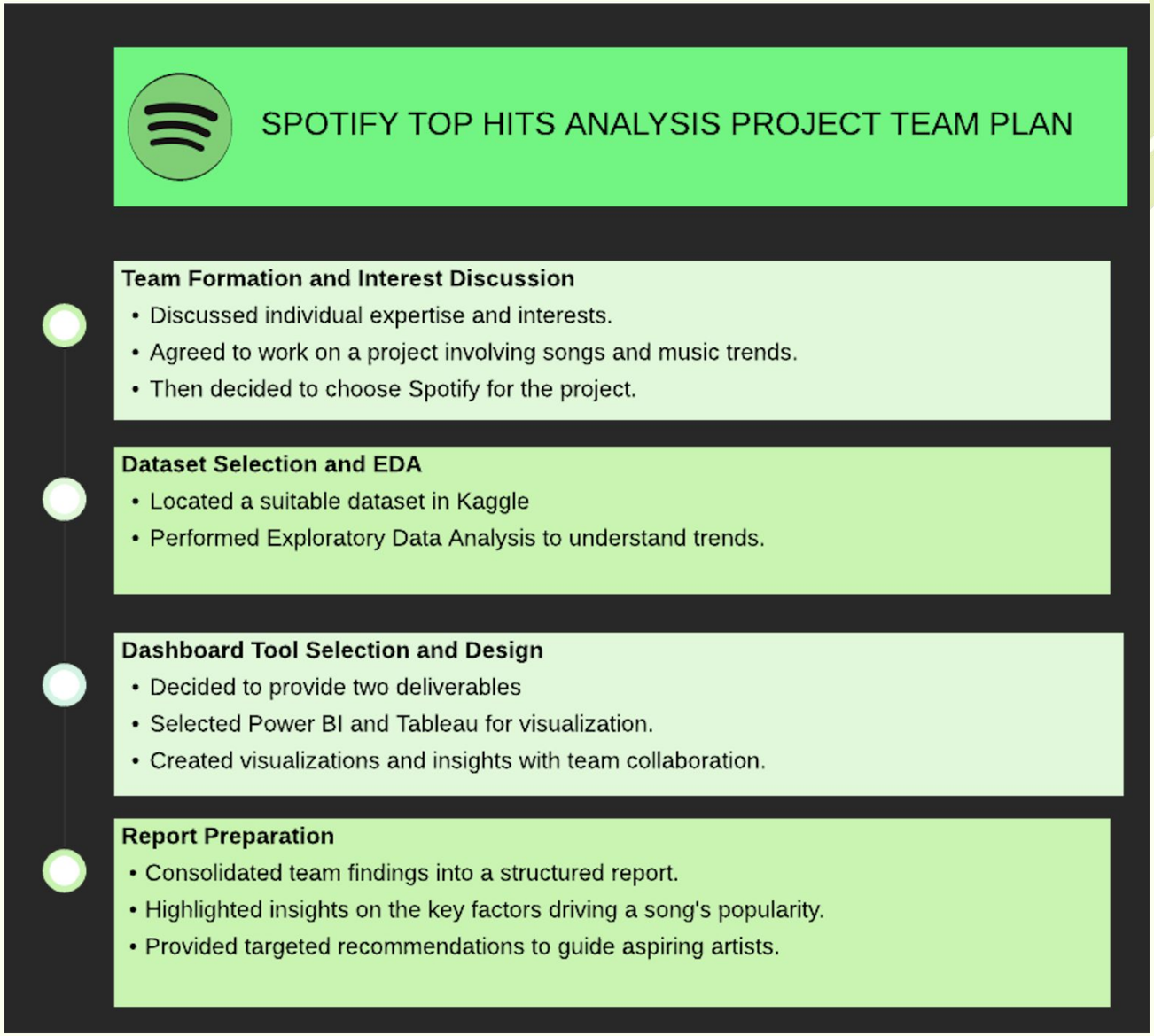
01 What features of a song (e.g., danceability, energy, tempo) contribute most to its popularity?

02 How has the popularity of genres evolved over the years?

03 Which artists and production styles have dominated the streaming era?

04 What are the geographical trends in top hits streaming?

TEAM PLAN


SPOTIFY TOP HITS ANALYSIS PROJECT TEAM PLAN
Team Formation and Interest Discussion
• Discussed individual expertise and interests.
• Agreed to work on a project involving songs and music trends.
• Then decided to choose Spotify for the project.
Dataset Selection and EDA
• Located a suitable dataset in Kaggle
• Performed Exploratory Data Analysis to understand trends.
Dashboard Tool Selection and Design
• Decided to provide two deliverables
• Selected Power BI and Tableau for visualization.
• Created visualizations and insights with team collaboration.
Report Preparation
• Consolidated team findings into a structured report.
• Highlighted insights on the key factors driving a song's popularity.
• Provided targeted recommendations to guide aspiring artists.

TEAM CONTRIBUTION

- EDA
- Tableau Dashboard Design
- Report Preparation

Bence Danko

- EDA
- Power BI Dashboard Design
- Report Preparation

Saumya Varshney

- Power BI Dashboard Design
- Report Preparation
- Project ideas and dataset identification

Sreenidhi Hayagreevan

- Tableau Dashboard Design
- Report Preparation

Victor Dumaslan

DATASET DESCRIPTION

01

Source:

- Kaggle – Spotify Top Hits Dataset (2000-2020)
- Daily Streaming Dataset (2017-2019)

02

Features:

- Track/Song Metadata: Song title,lyrics artist name, release year
- Audio Features: Danceability, energy, loudness, tempo, etc.
- Popularity Metrics: Spotify popularity score, Streams

03

Size:

- Contains records of [2000 number of songs] across two decades
- Joined with daily streaming counts 2017-2019 [2,423,033 records of daily streaming counts]

Dataset Before EDA

	artist	song	Country	Streams	Date
1258	blink-182	All The Small Things	Ireland	4297	3/30/2017
59421	blink-182	All The Small Things	Ireland	4589	7/20/2017
122034	blink-182	All The Small Things	Ireland	5225	8/23/2018
127564	Britney Spears	Toxic	United Kingdom	61937	5/3/2018
150516	Britney Spears	Toxic	Denmark	8872	8/7/2018

	artist	song	duration_ms	explicit	year	popularity	danceability
0	Britney Spears	Oops!...I Did It Again	211160	False	2000	77	0.751
1	blink-182	All The Small Things	167066	False	1999	79	0.434
2	blink-182	All The Small Things	167066	False	1999	79	0.434
3	Faith Hill	Breathe	250546	False	1999	66	0.529
4	Faith Hill	Breathe	250546	False	1999	66	0.529

1. Songs duration is in milliseconds.
2. Genres are comma separated.
3. Geolocation details with streaming details.

danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	genre
0.751	0.834	1	-5.444	0	0.0437	0.3000	0.000018	0.3550	0.894	95.053	pop
0.434	0.897	0	-4.918	1	0.0488	0.0103	0.000000	0.6120	0.684	148.726	rock, pop
0.529	0.496	7	-9.007	1	0.0290	0.1730	0.000000	0.2510	0.278	136.859	pop, country
0.551	0.913	0	-4.063	0	0.0466	0.0263	0.000013	0.3470	0.544	119.992	rock, metal

EXPLORATORY DATA ANALYSIS (EDA)

Data Cleaning

1. Checking the null values
2. Analysing the dataset information

EDA

1. Converting the song duration from millisec to minutes
2. Consolidating Artist name for word cloud
3. Exploding the Genre feature for better visualization
4. Join datasets on Artist/Song name for geolocation columns

Feature Extraction

1. Grouping years into decade
2. Categorizing Popularity into High, Medium & Low

Data Cleaning

1. Checking the null and empty values in the dataset for all the columns.
2. Analysis of dataset information and distribution.

```
[5]  1 songs.isnull().sum()
```

```
[6]  1 songs.isna().sum()
```

	0
artist	0
song	0
duration_ms	0
explicit	0
year	0
popularity	0
danceability	0
energy	0
key	0
loudness	0
mode	0
speechiness	0
acousticness	0
instrumentalness	0
liveness	0
valence	0
tempo	0
genre	0
duration_mins	0
decade	0
popularity_category	0

dtype: int64

```
songs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3710 entries, 0 to 3709
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   artist                3710 non-null   object
 1   song                  3710 non-null   object
 2   duration_ms           3710 non-null   int64
 3   explicit              3710 non-null   bool
 4   year                  3710 non-null   int64
 5   popularity            3710 non-null   int64
 6   danceability          3710 non-null   float64
 7   energy                3710 non-null   float64
 8   key                   3710 non-null   int64
 9   loudness              3710 non-null   float64
 10  mode                  3710 non-null   int64
 11  speechiness           3710 non-null   float64
 12  acousticness          3710 non-null   float64
 13  instrumentalness      3710 non-null   float64
 14  liveness              3710 non-null   float64
 15  valence               3710 non-null   float64
 16  tempo                 3710 non-null   float64
 17  genre                 3710 non-null   object
 18  duration_mins         3710 non-null   float64
 19  decade                3710 non-null   int64
 20  popularity_category   3710 non-null   object
dtypes: bool(1), float64(10), int64(6), object(4)
memory usage: 583.4+ KB
```

```
songs.describe()
```

	duration_ms	year	popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_mins	decade
count	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000	3710.000000
mean	229648.378976	2009.338005	59.683827	0.672044	0.718555	5.416712	-5.512056	0.549326	0.106184	0.126632	0.014218	0.180976	0.554179	119.732831	3.827473	2004.862534
std	37558.152257	5.786443	21.025246	0.138686	0.151506	3.627869	1.926589	0.497628	0.095457	0.168379	0.084218	0.141350	0.219317	26.841252	0.625969	5.388140
min	113000.000000	1998.000000	0.000000	0.129000	0.054900	0.000000	-20.514000	0.000000	0.023200	0.000019	0.000000	0.021500	0.038100	60.019000	1.883333	1990.000000
25%	205360.000000	2004.000000	56.000000	0.585000	0.619000	2.000000	-6.489500	0.000000	0.040700	0.014300	0.000000	0.086925	0.394000	98.523000	3.422667	2000.000000
50%	224886.500000	2010.000000	65.000000	0.681000	0.734000	6.000000	-5.293000	1.000000	0.063200	0.055700	0.000000	0.124000	0.561500	120.028000	3.748108	2010.000000
75%	249533.000000	2014.000000	73.000000	0.767000	0.837000	8.000000	-4.165250	1.000000	0.136750	0.176000	0.000061	0.240000	0.731000	133.040250	4.158883	2010.000000
max	484146.000000	2020.000000	89.000000	0.975000	0.999000	11.000000	-0.276000	1.000000	0.576000	0.976000	0.985000	0.853000	0.973000	210.851000	8.069100	2020.000000

EDA

```
# explode genres
songs_genre_exploded = songs.assign(genre=songs['genre'].str.split(', ')).explode('genre')
```

```
# Process the daily_df in chunks
for chunk in pd.read_csv('data/Spotify_Daily_Streaming_Cleaned.csv', chunksize=chunk_size):
    # Perform the join operation
    merged_chunk = pd.merge(chunk, normalize_df, left_on=['Track Name', 'Artist'], right_on=['song', 'artist'], how='inner')
    # Append the result to the list
    results.append(merged_chunk)

# Concatenate all the results into a single DataFrame
final_df = pd.concat(results, ignore_index=True)
```

1. Genres are exploded based on comma for better analysis.
2. Daily streaming data is merged for geolocation streaming analysis

Feature Extraction

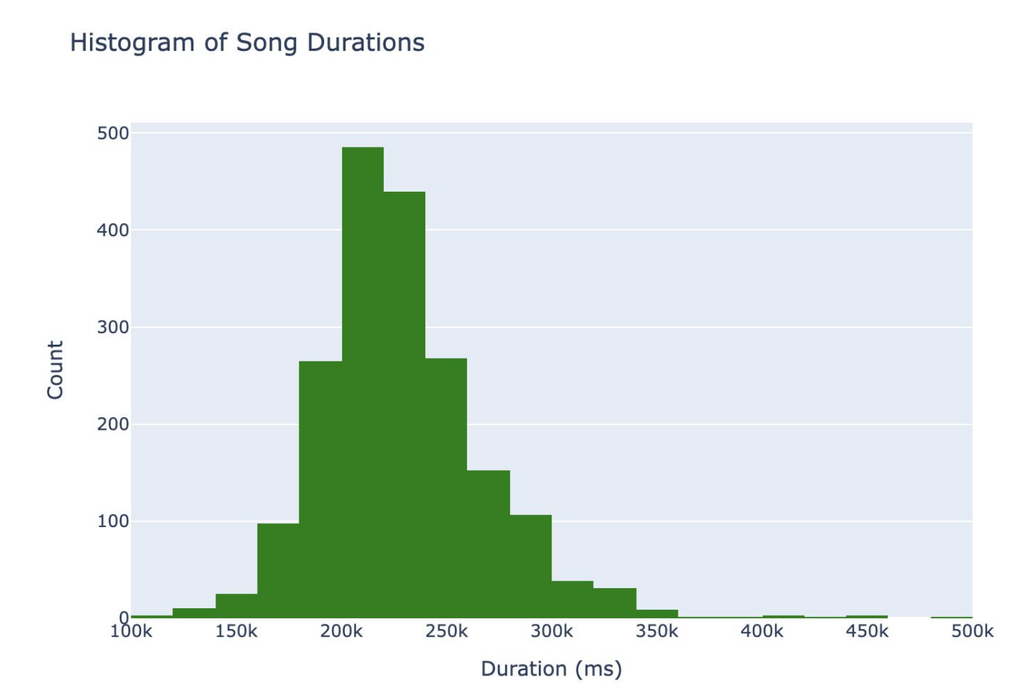
```
[7]  1 # Converting song duration from millisec to mins
     2 songs['duration_mins'] = songs['duration_ms'] / (1000 * 60)
```

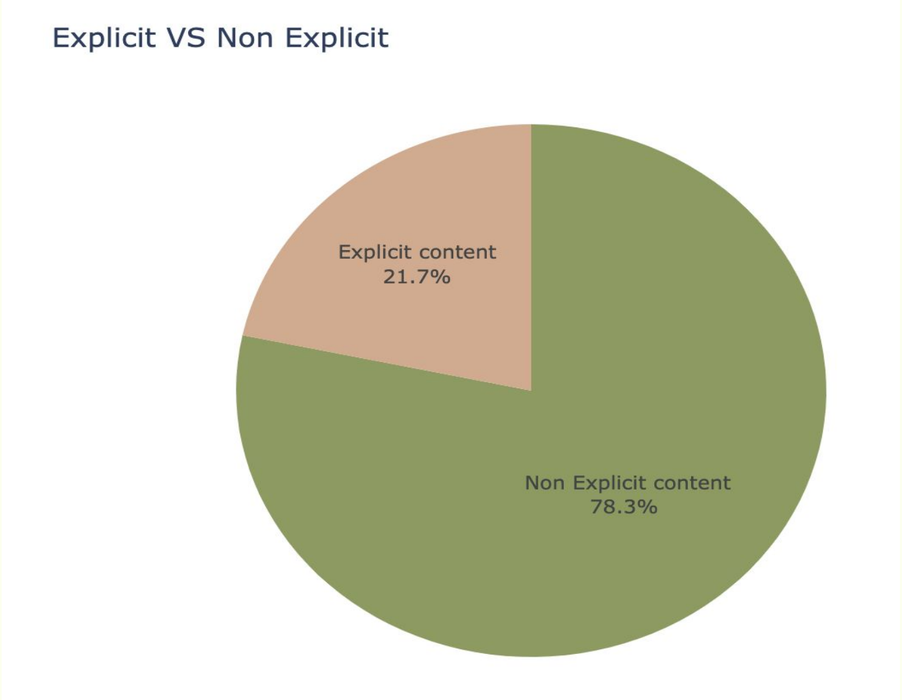
```
[8]  1 # Group Years into Decades
     2 songs['decade'] = (songs['year'] // 10) * 10
```

```
[9]  1 # 3. Categorize `popularity` into High, Medium, Low
     2 songs['popularity_category'] = pd.cut(songs['popularity'], bins=[0, 49, 79, 100],
     3                                       labels=['Low', 'Medium', 'High'], include_lowest=True)
     4
```

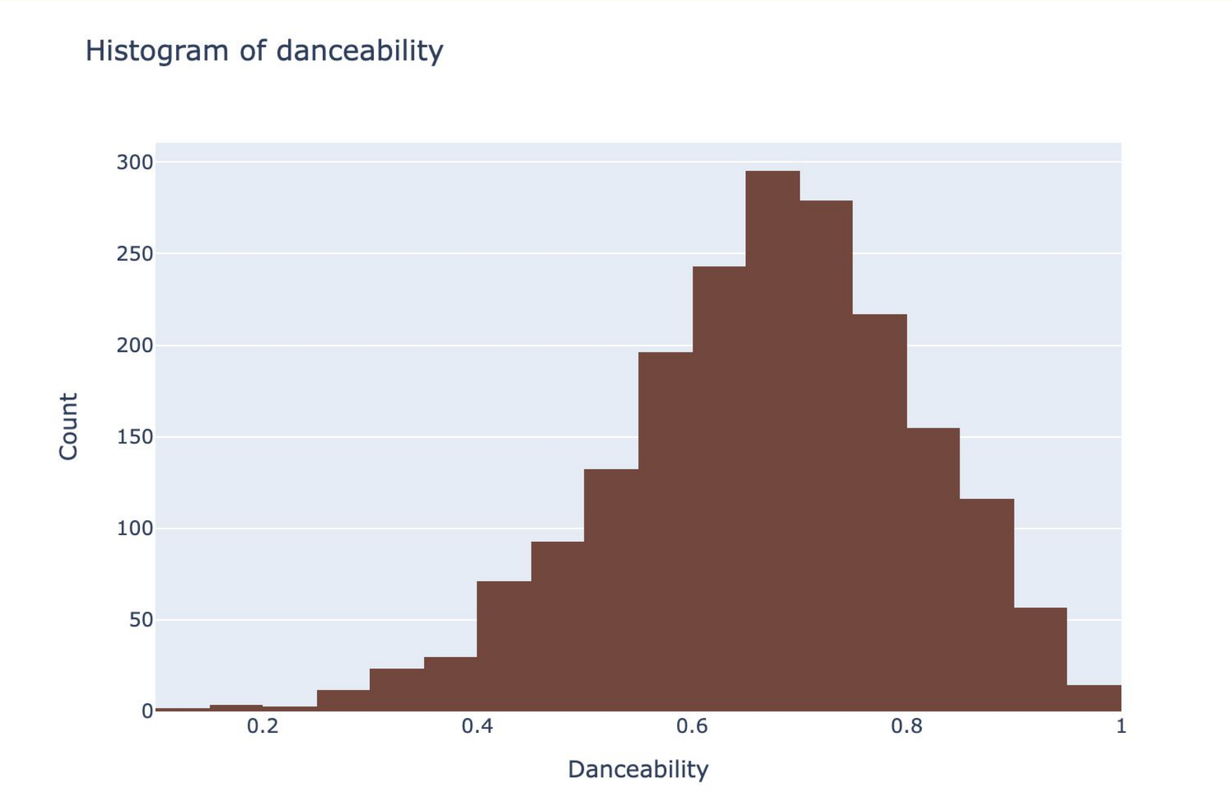
1. Converting the song duration from millisec to minutes.
2. Grouping the years into decade for better analysis based on decade.
3. Categorizing the popularity scores.

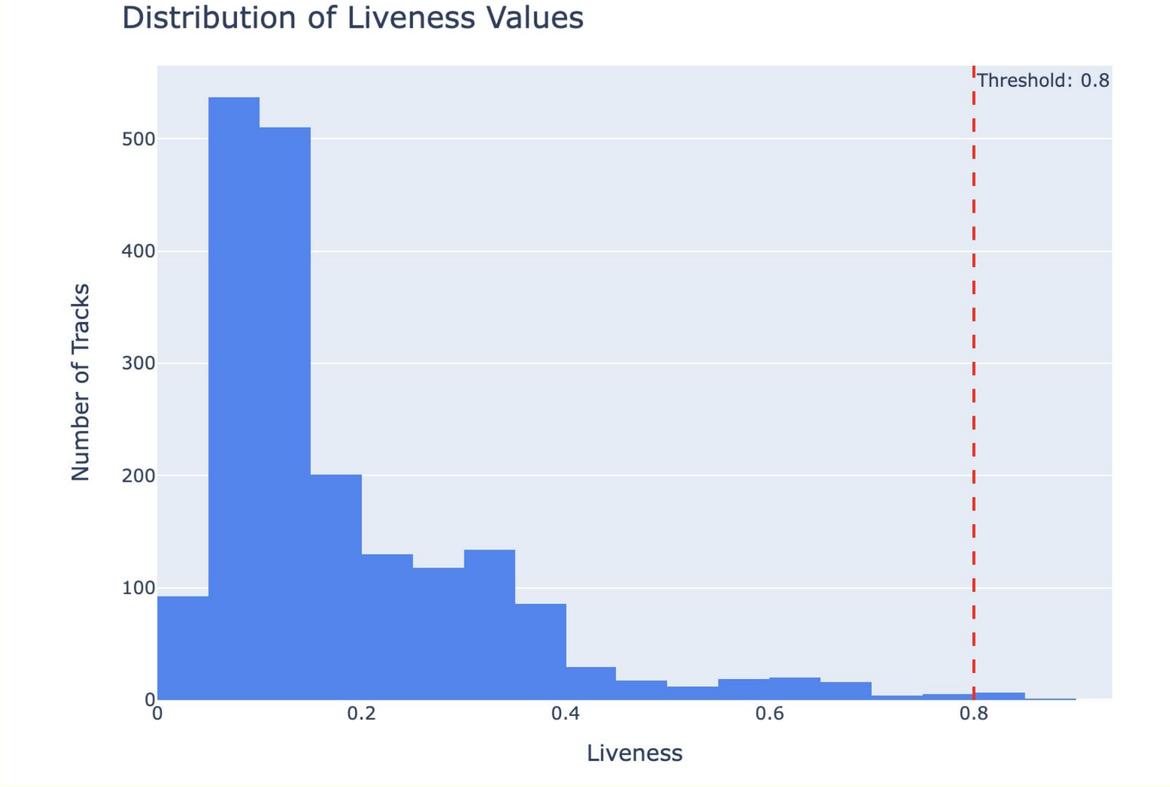
Data Distribution



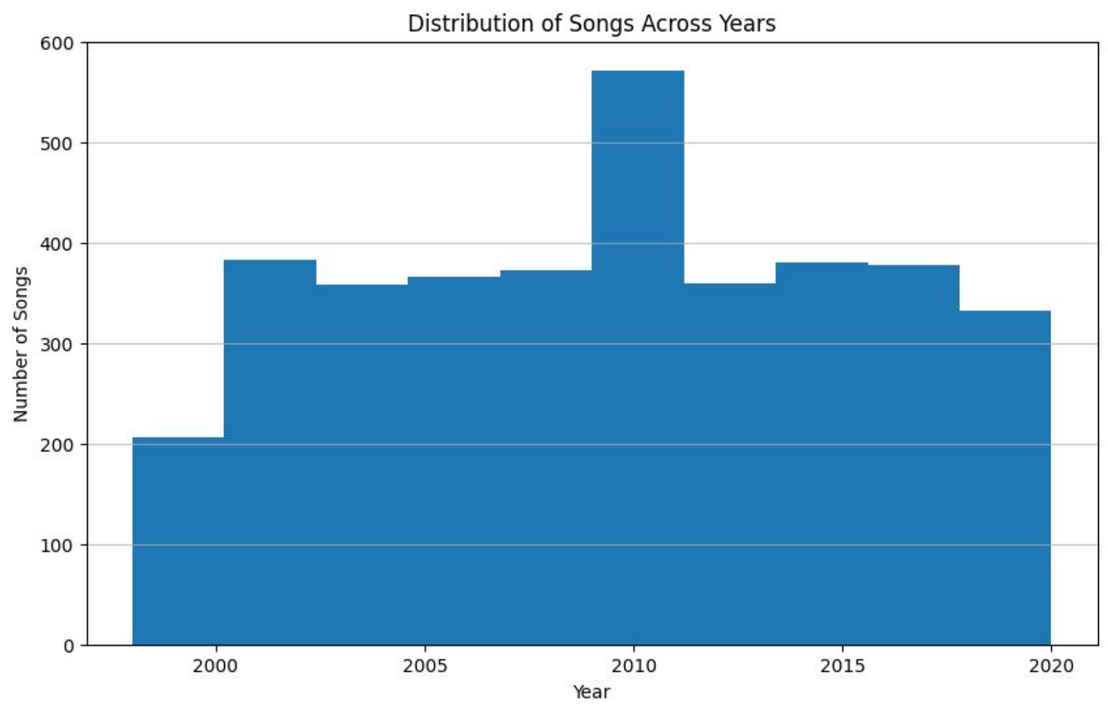


Data Distribution



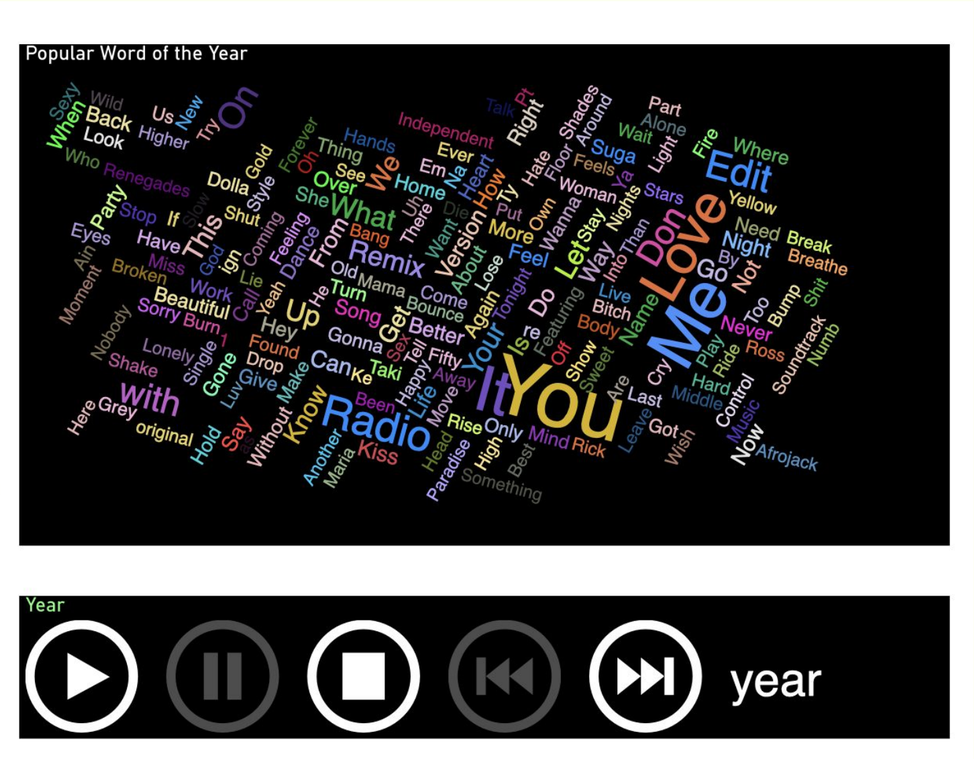


Data Distribution

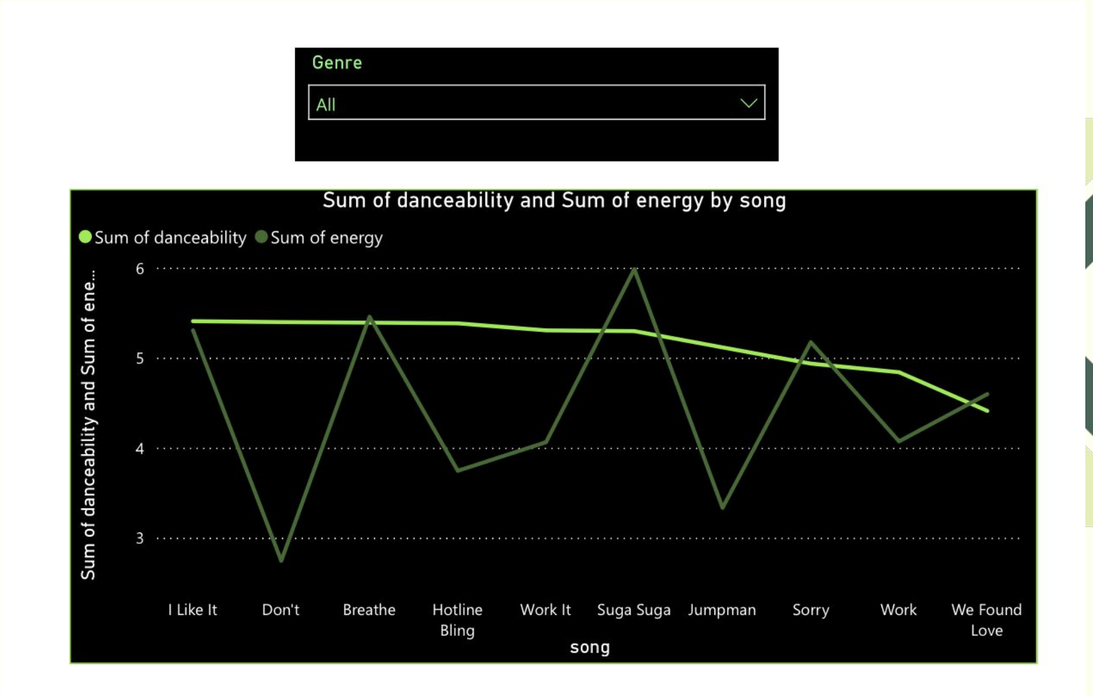


PowerBI Charts

Word cloud showing popular words in the songs over the years



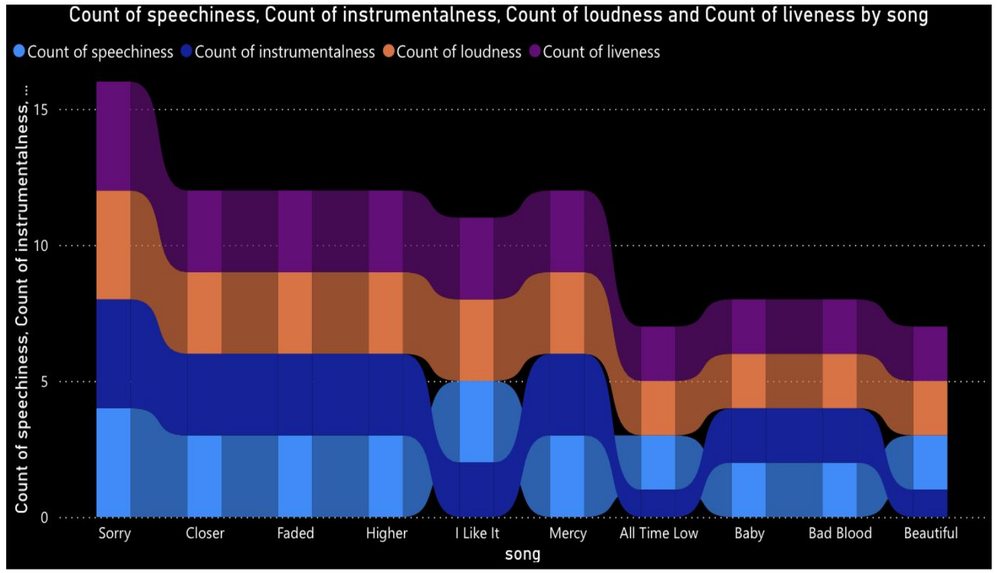
Genre-wise Top 10 songs and its Danceability Vs Energy



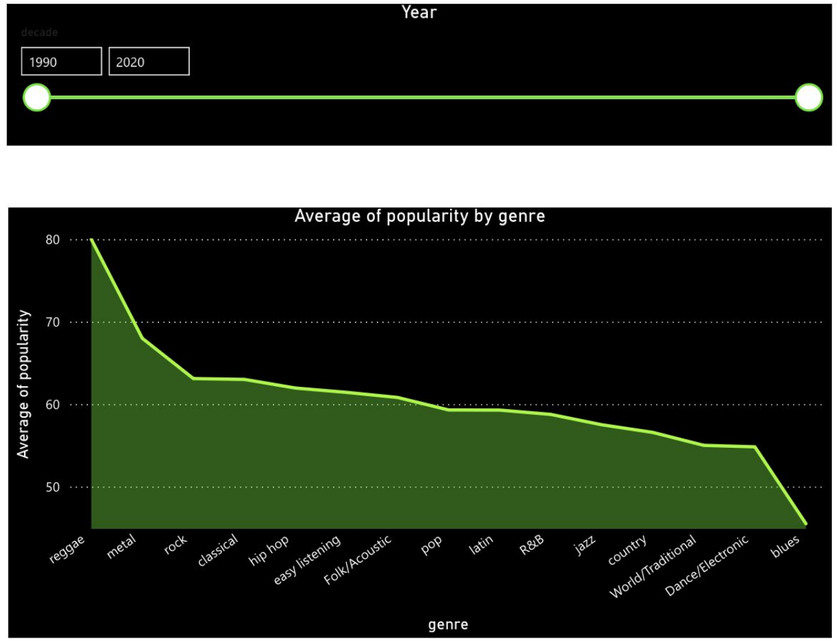
Insights: The word cloud reveals recurring lyrical themes like "You," "Love," and "Me." The line chart shows variations in danceability and energy among top 10 songs, balancing or emphasizing attributes uniquely.

PowerBI Charts

Speech factor, Instrument usage, loudness and liveliness of Top 10 Popular Songs



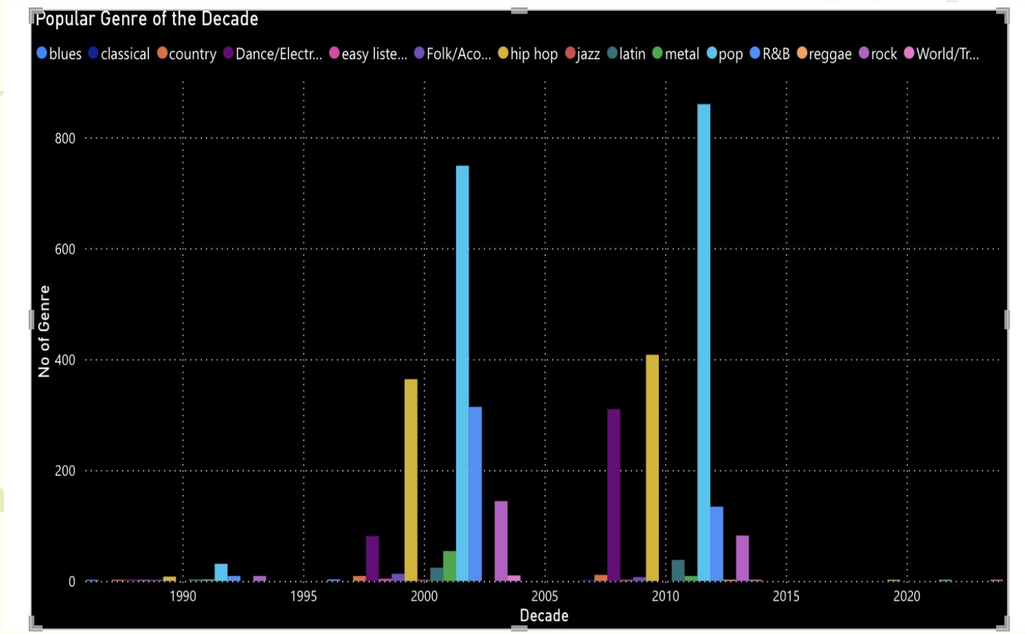
Average Popularity of Genre within a particular period of time



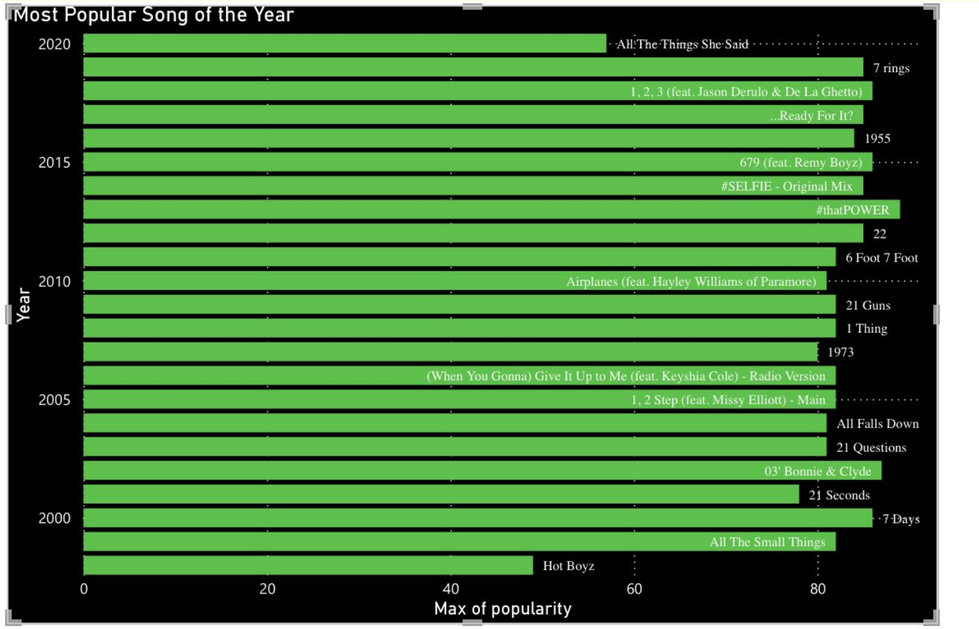
Insights: The first chart shows varying acoustic profiles (speechiness, instrumentalness, loudness, liveness) in top 10 songs. The second highlights genre popularity over time, with reggae and metal ranking higher compared to others during specific periods.

PowerBI Charts

Popular Genre of the Decade



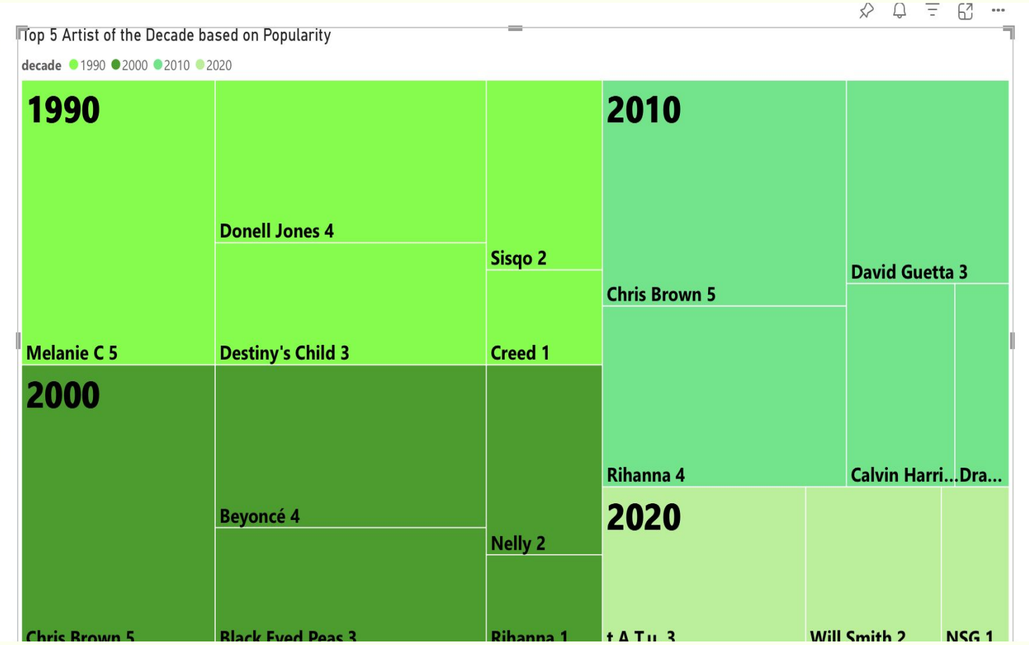
Most popular song of the year



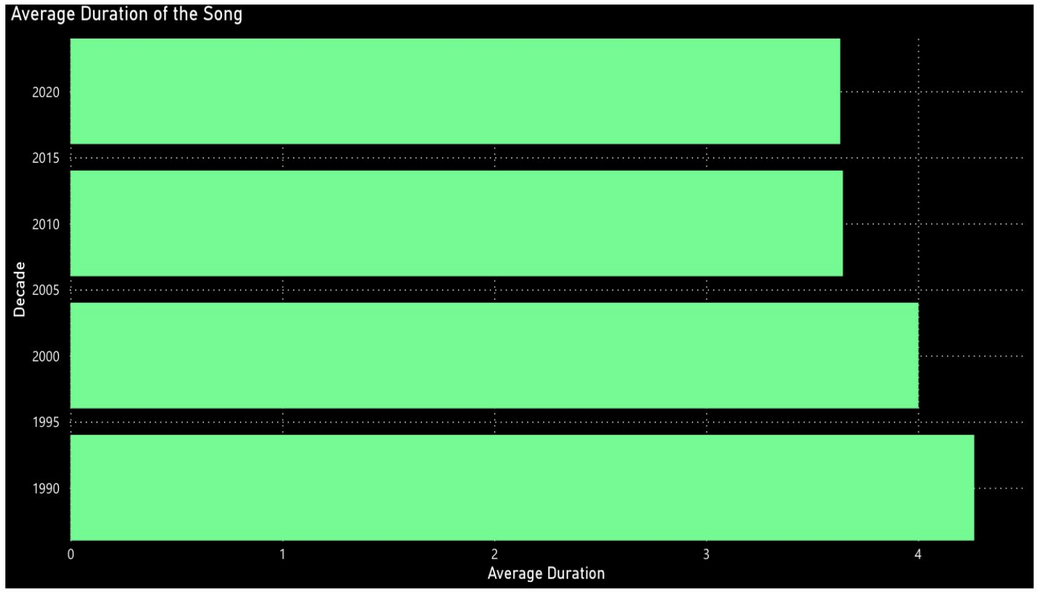
Insights: Pop and hip-hop were the dominant genres in the 2000s and 2010s. The chart also highlights yearly hit songs, showcasing shifts in musical preferences over time.

PowerBI Charts

Top 5 Artist of the decade based on popularity



Average duration of song every decade and its impact

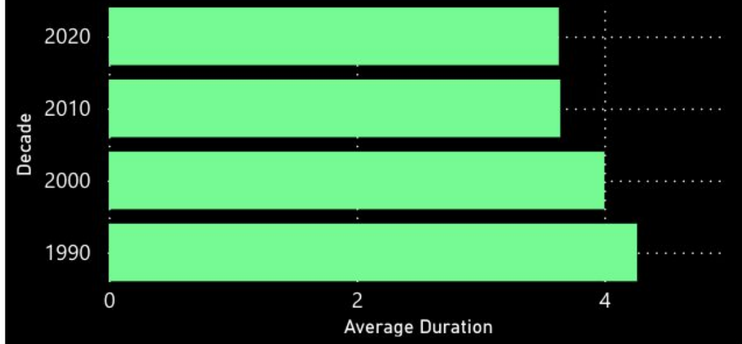


Insights: The charts highlight top artists by decade, like Beyoncé (2000s) and Chris Brown (2010s), and show a gradual increase in average song duration, reflecting evolving preferences and production styles.

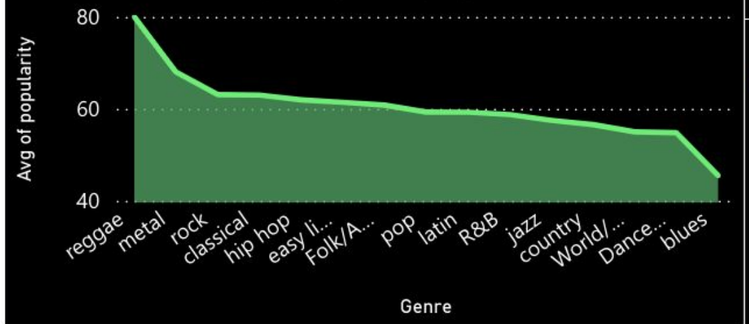
POWERBI DASHBOARD

Spotify Top Hits Analysis

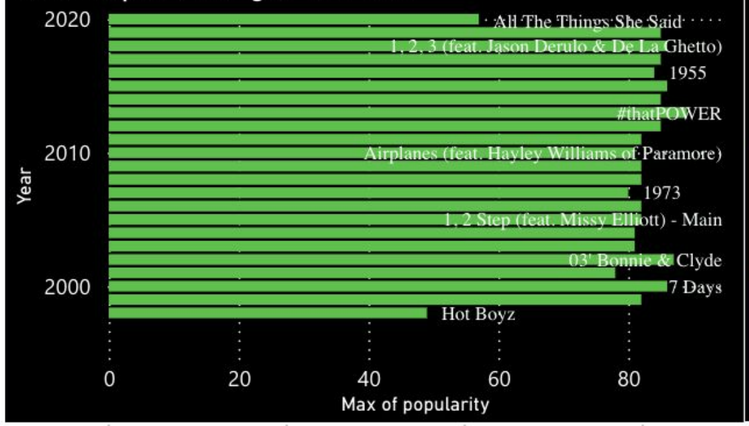
Average Duration of the Song



Average Popularity by Genre

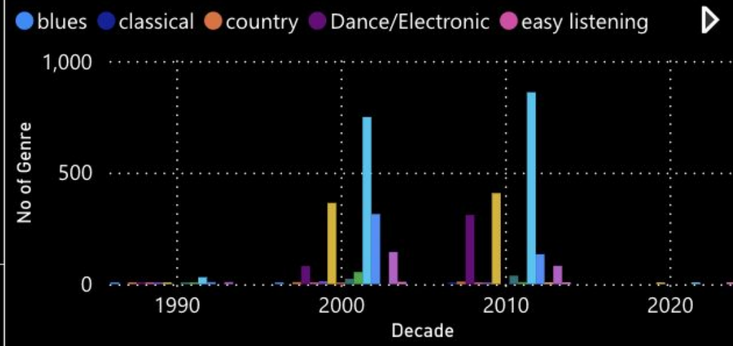


Most Popular Song of the Year

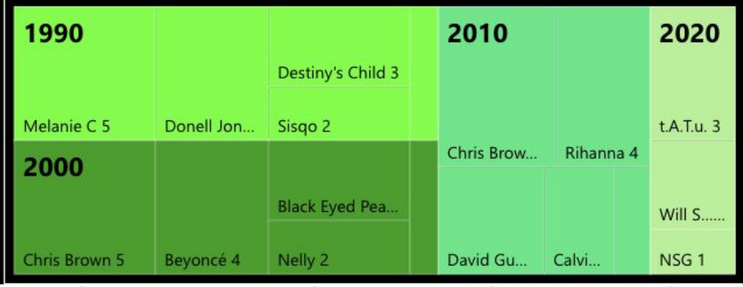


Popular Word of the Year

Popular Genre of the Decade



Top 5 Artist of the Decade based on Popularity



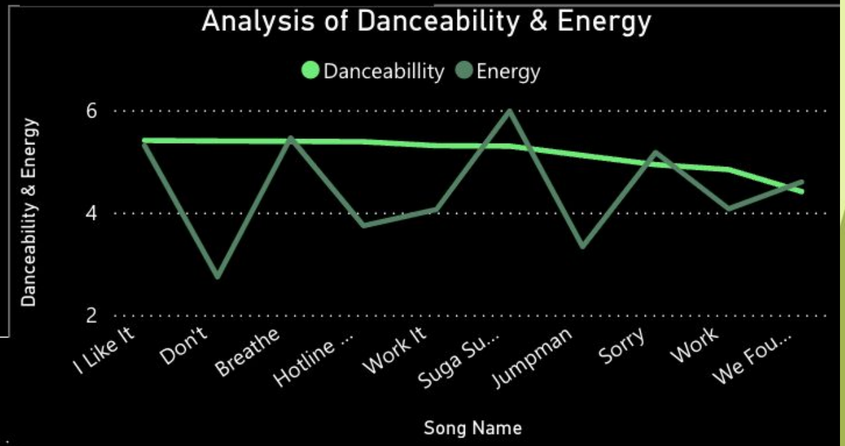
Year

1990 — 2020

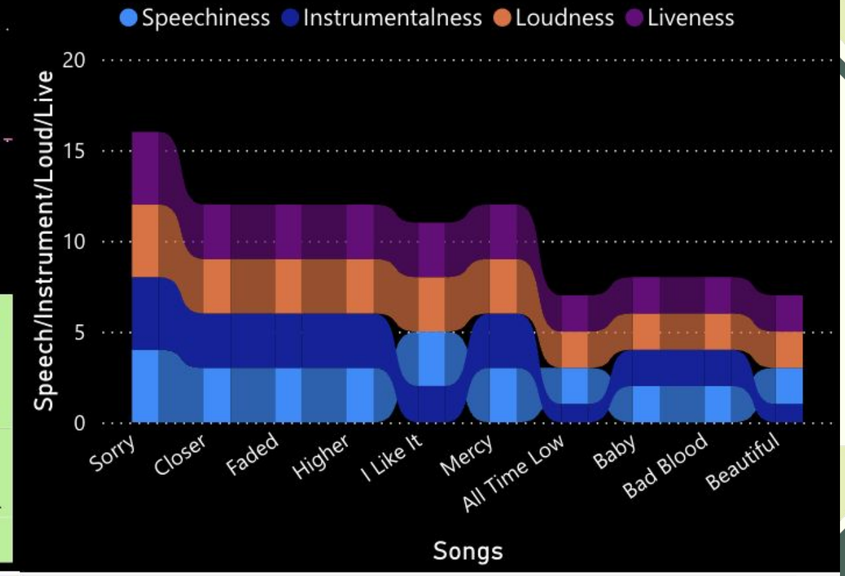
Genre

All

Analysis of Danceability & Energy



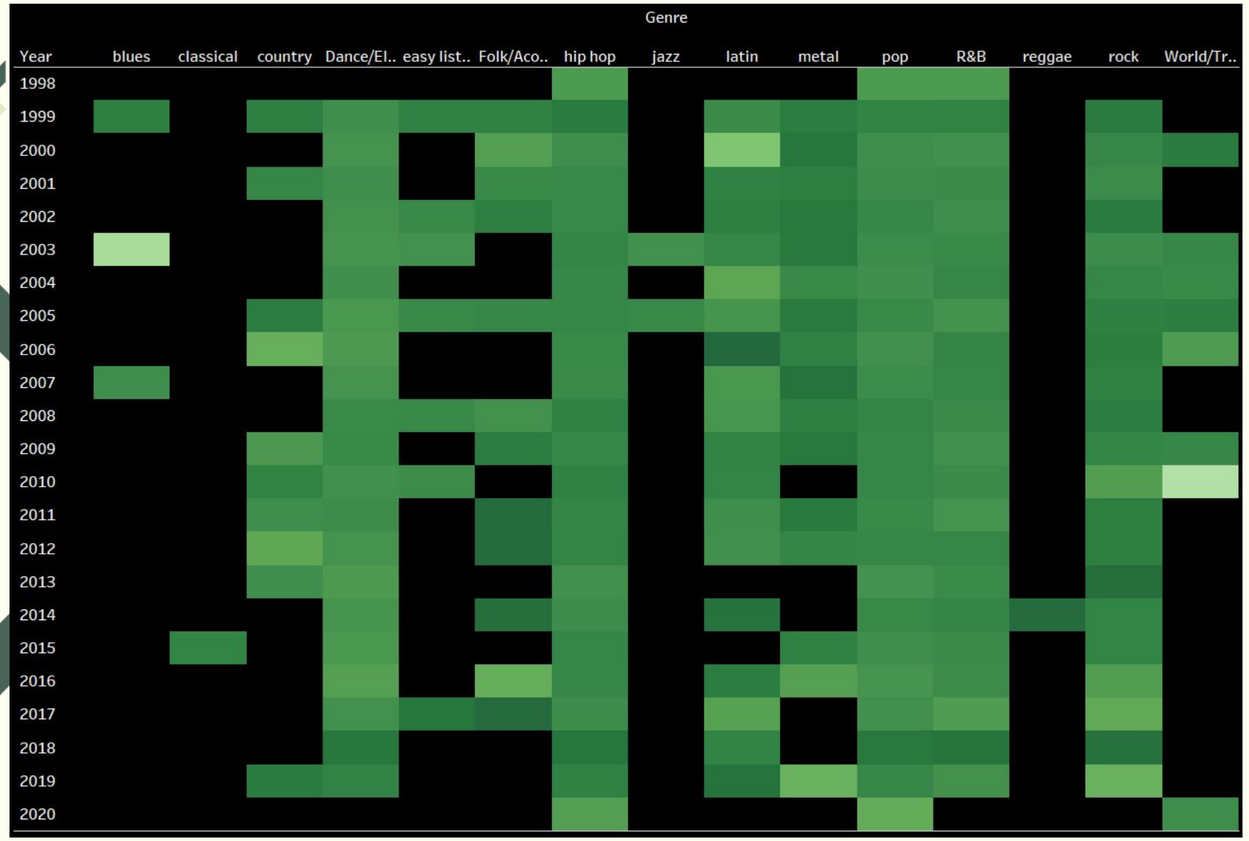
Analysis of Speechiness, Instrumentalness Loudness & Liveness



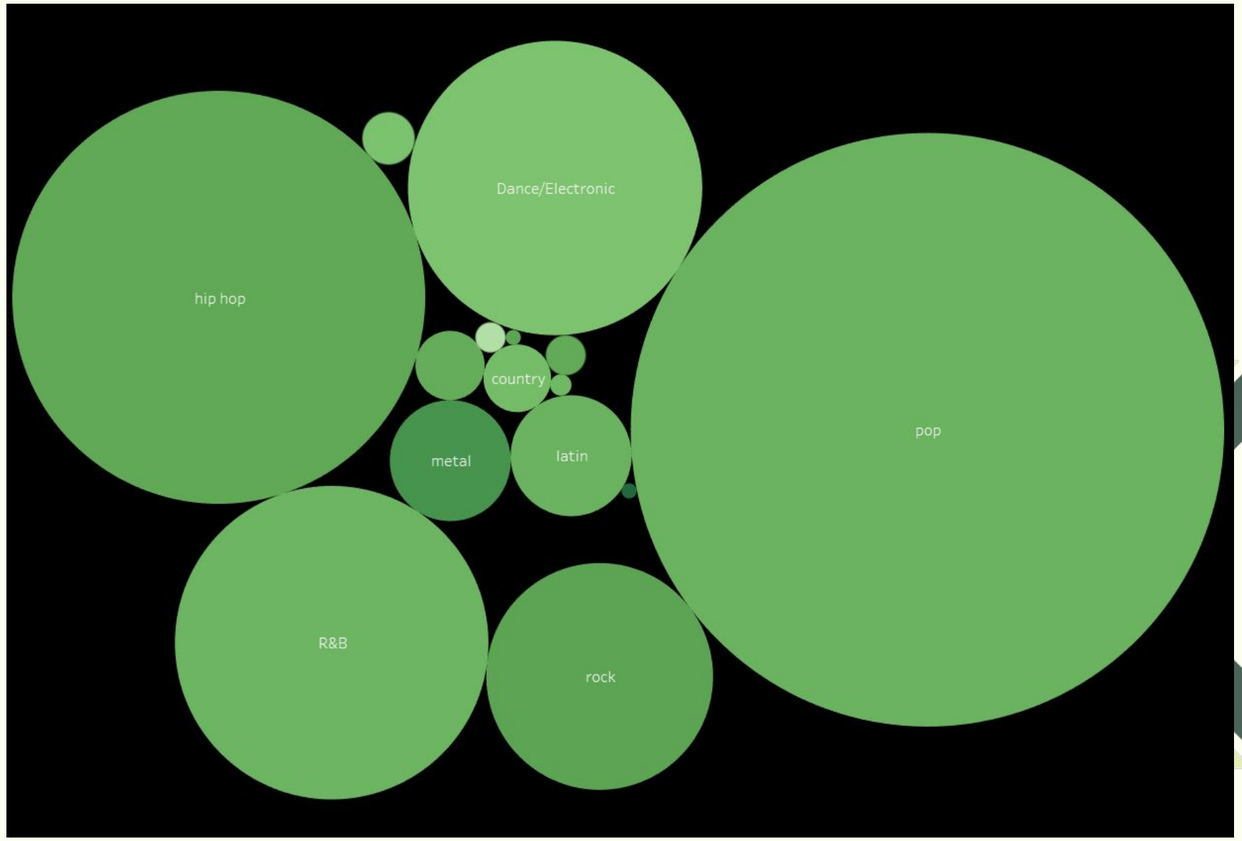
PowerBI Dashboard Demo

PowerBI Link

Tableau Charts



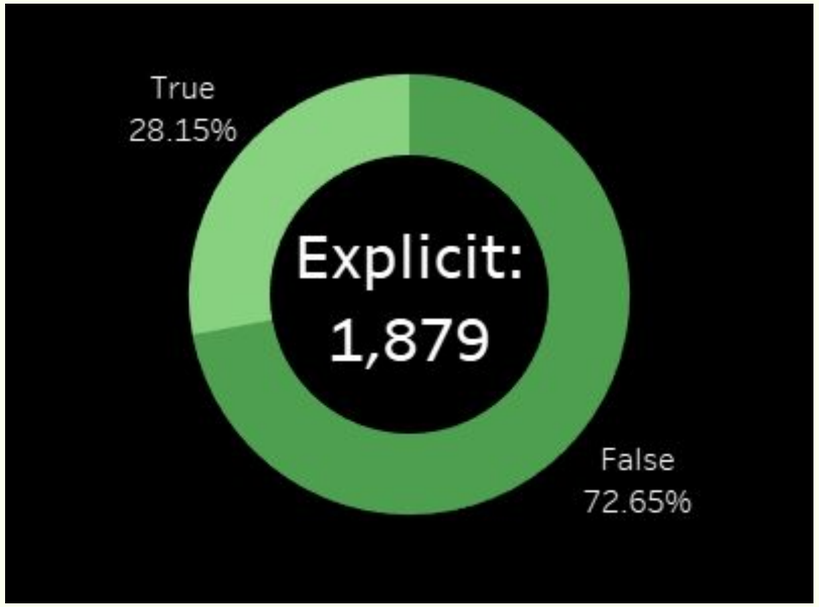
Genre Popularity Over Time



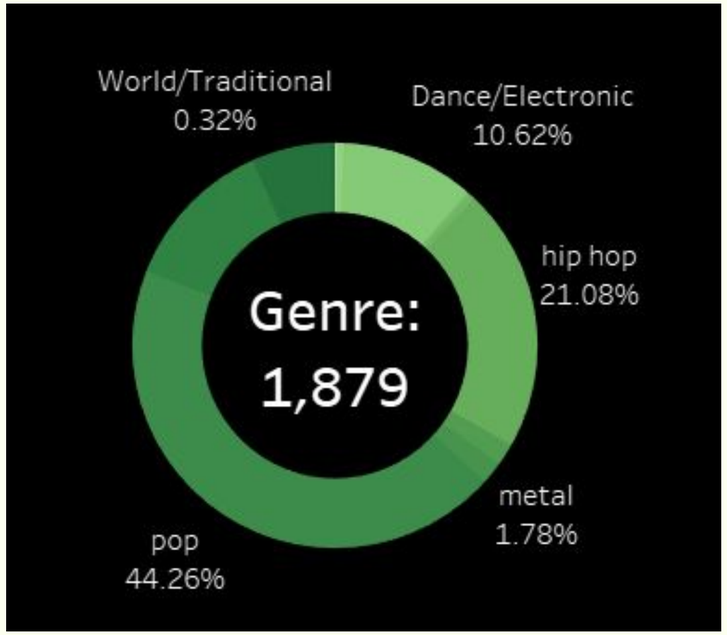
Genres by Song Count and Popularity

Insights: Pop and Hip Hop dominate top hits by year and song count

Tableau Charts



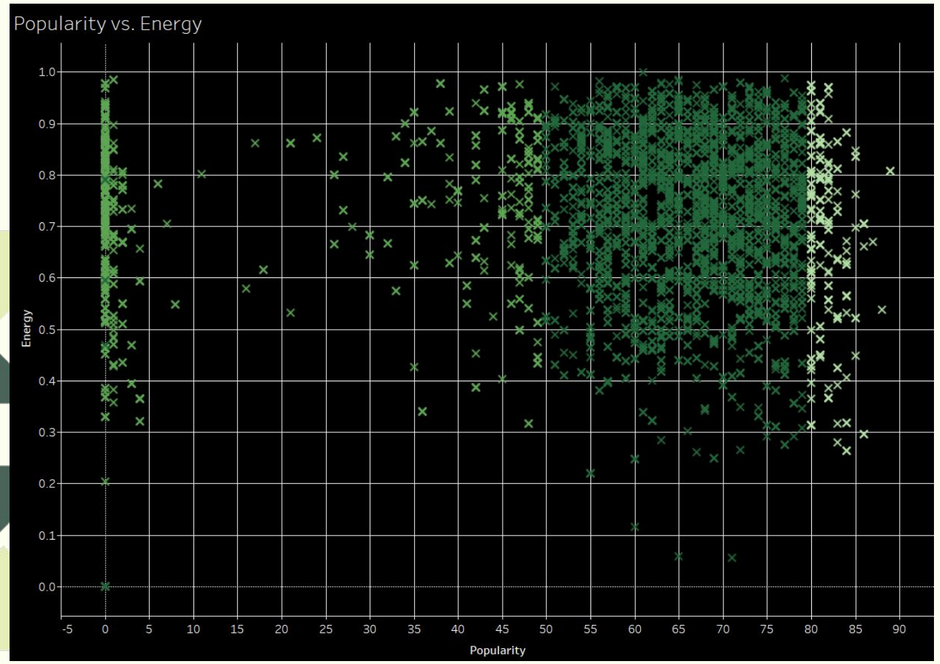
Explicit vs Non-Explicit Songs



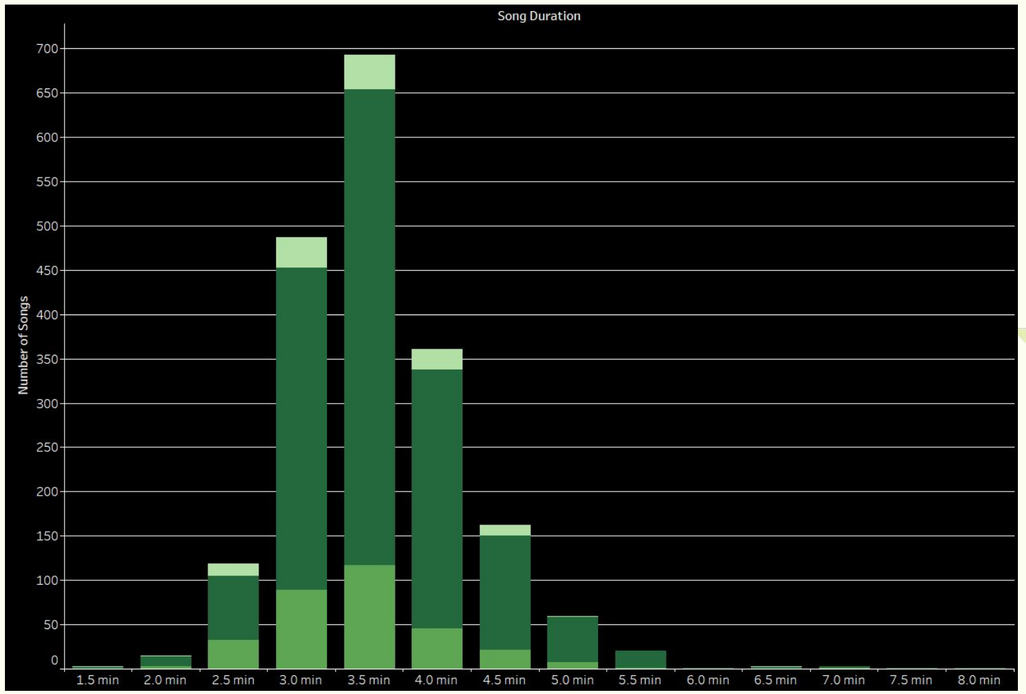
Song Count based on Genre

Insights: A significant portion of top hits may contain explicit contents, but majority do not

Tableau Charts



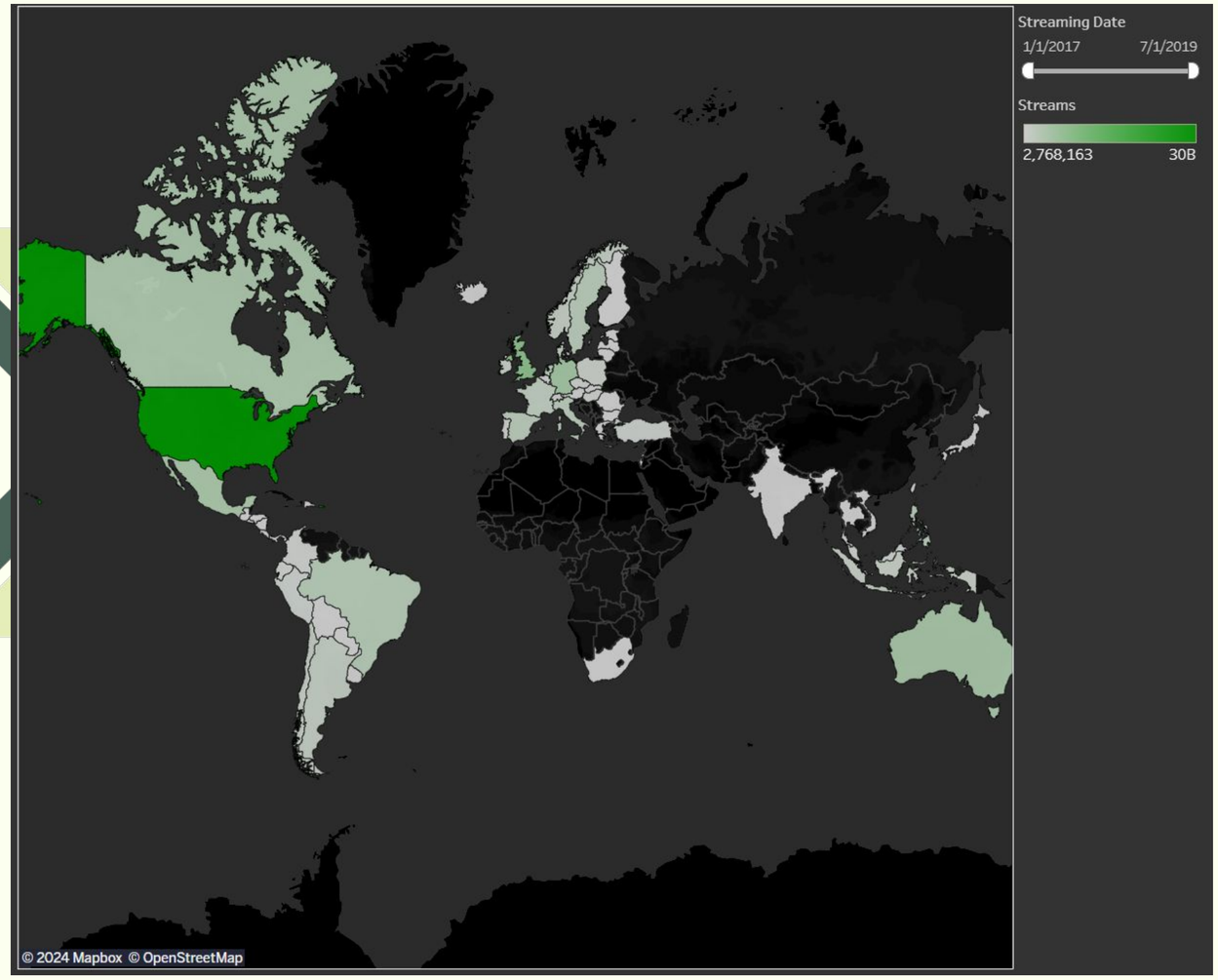
Popularity vs ‘Features’ Scatter Plot



Distribution of Song Durations by Popularity

Insights: Majority of popular songs feature high energy, danceability, and tempo. Most are within the duration of 3 to 4 minutes

Tableau Charts



Total Streams by Country
2017-2019

Insights: Majority of top hits stream in the United States, but international streaming is increasing (Asia)

TABLEAU DASHBOARD

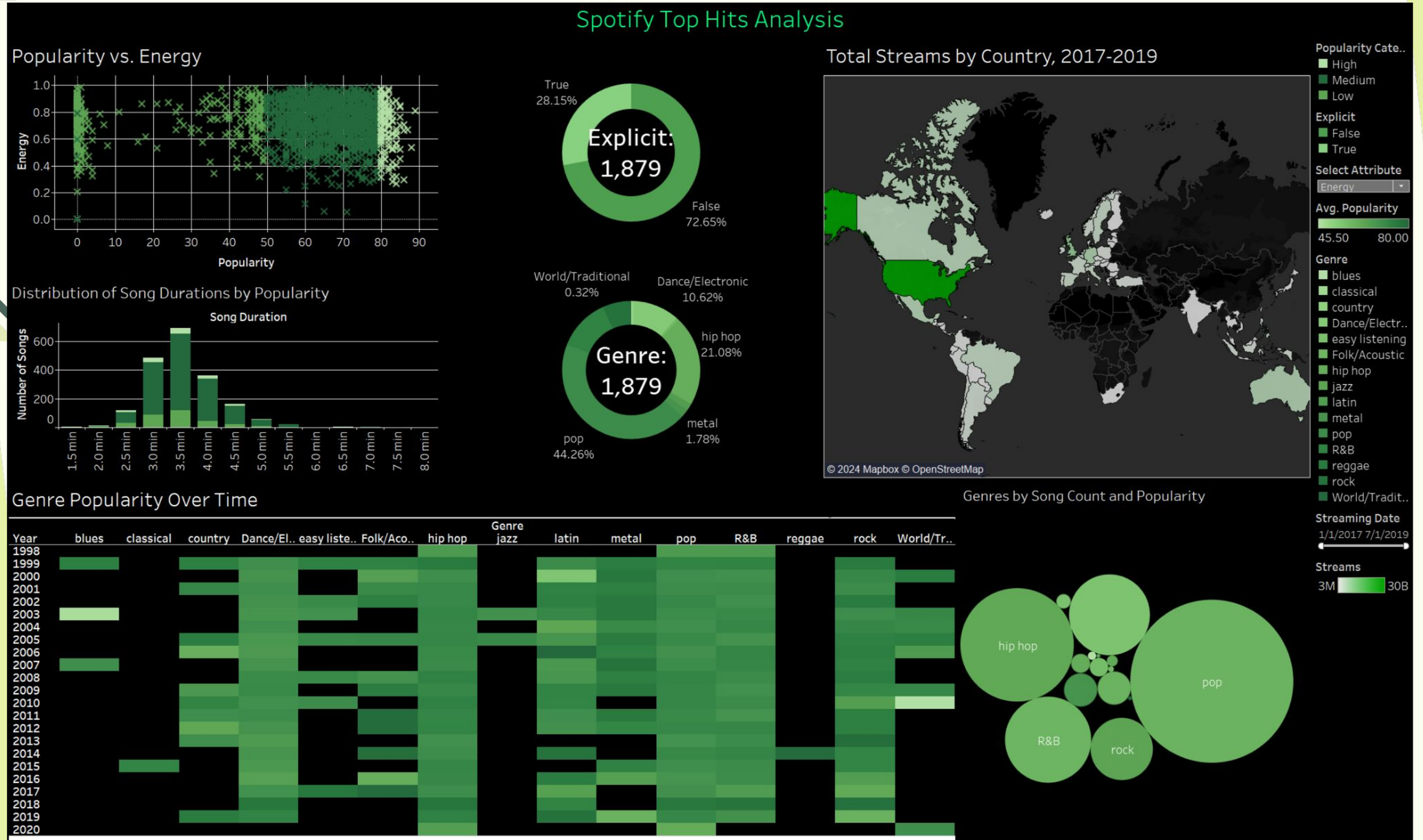


Tableau Dashboard Demo

Tableau Dashboard Link

Tableau Story Link

Recommendations/Next Steps

Recommendations:

- Focus on creating songs with high energy, danceability, and tempo based on current trends
- Explore dominating genres, like pop and hip hop, and blend styles for broader appeal
- Target a song duration of 3 to 4 minutes for maximum appeal
- Tailor marketing strategies by targeting regions with higher and growing streaming counts (Asia)

Next Steps:

- Expanding analysis to include data beyond 2019 for updated trends
- Leveraging machine learning for predictive analysis on song popularity

Conclusion

Conclusion:

- Identified features like liveness, energy, and tempo that contribute to a song’s popularity
- Observed shifts in genre popularity across years
- Highlighted geographical streaming trends for top hits with Spotify

THANK YOU