Analyzing and Modeling Top 1000 Songs on Spotify

DSCI 550: Data Science at Scale

Team #02:
Ying Yang (Leader)
Yixin Qu
Zhiqian Li
Wenjing Huang

Project Idea

Dataset we got from Kaggle **"Spotify Top Songs Streamed in 2023".**

Analysis:

First objective : To identify common music features and patterns among top 1000 streamed songs.

Second objective:

- Apply regression models to predict future songs based on selected music features.
- Employ classification models to predict if future songs will become super hit by assessing if their streaming volumes rank in the top 25 of a 1000-song dataset.

Description of Dataset

Description :

This dataset contains a comprehensive list of the most famous songs of 2023 as listed on Spotify. The dataset offers a wealth of features beyond what is typically available in similar datasets. It provides insights into each song's attributes, popularity, and presence on various music platforms. The dataset includes information such as **track name, artist(s) name, release date, Spotify playlists and charts, streaming statistics, Apple Music presence, Deezer presence, Shazam charts, and various audio features**.

- Kaggle **"Spotify Top Songs Streamed in 2023"**
- Size: 48 KB in size with 1,000 rows with 24 attributes.
  - Track details: track name, artist name, artist count, and release year
  - Platforms: Spotify
  - Crucial audio features: bpm, key, mode, and danceability percentage
- Why the dataset is appropriate
  - Attributes
  - Breadth: nearly 1,000 top-tier
  - "A leading streaming platform, which makes it an official and reliable resource" (Castillo et al., 2023)

Current Progress | Data Cleaning

- Import the dataset
  - First import and read the dataset into Google Colab

```
[1] import pandas as pd
    import matplotlib.pyplot as plt
    import numpy as np
    import seaborn as sns
```

```
[2] data = pd.read_csv("./sample_data/spotify-2023.csv", encoding='latin-1')
    data.tail(5)
```

	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_cha
948	My Mind & Me	Selena Gomez	1	2022	11	3	953	
949	Bigger Than The Whole Sky	Taylor Swift	1	2022	10	21	1180	
950	A Veces (feat. Feid)	Feid, Paulo Londra	2	2022	11	3	573	
951	En La De Ella	Feid, Sech, Jhayco	3	2022	10	20	1320	
952	Alone	Burna Boy	1	2022	11	4	782	

5 rows × 24 columns

Current Progress | Data Cleaning

- Missing Value

```
[4] data.isna().sum().sum()

145
```

```
[5] data.isna().sum()

track_name              0
artist(s)_name          0
artist_count            0
released_year           0
released_month          0
released_day            0
in_spotify_playlists    0
in_spotify_charts       0
streams                 0
in_apple_playlists      0
in_apple_charts         0
in_deezer_playlists     0
in_deezer_charts        0
in_shazam_charts       50
bpm                     0
key                    95
mode                    0
danceability_%          0
valence_%               0
energy_%                0
acousticness_%          0
instrumentalness_%      0
liveness_%              0
speechiness_%           0
dtype: int64
```

```
[7] #delete in_shazam_charts
    data.drop(columns=['in_shazam_charts'],inplace=True)
```

```
[8] data['key']=data['key'].fillna('unknown')
```

Then we got

```
[9] data.isna().sum()

track_name              0
artist(s)_name          0
artist_count            0
released_year           0
released_month          0
released_day            0
in_spotify_playlists    0
in_spotify_charts       0
streams                 0
in_apple_playlists      0
in_apple_charts         0
in_deezer_playlists     0
in_deezer_charts        0
bpm                     0
key                     0
mode                    0
danceability_%          0
valence_%               0
energy_%                0
acousticness_%          0
instrumentalness_%      0
liveness_%              0
speechiness_%           0
dtype: int64
```

```
[10] data.shape

(953, 23)
```

Current Progress | Data Cleaning

- Dealing with “Streams”
  - while preprocessing the data, we can’t process the “Streams”

```
[14] print(data.dtypes)
```

```
track_name              object
artist(s)_name          object
released_year            int64
released_month           int64
released_day             int64
streams                 object
bpm                      int64
key                     object
mode                    object
danceability_%           int64
valence_%                int64
energy_%                 int64
acousticness_%           int64
instrumentalness_%       int64
liveness_%               int64
speechiness_%            int64
dtype: object
```

Find why the streams object

```
[15] data["streams"] = data["streams"].astype(str)
     df_non_numeric = data[pd.to_numeric(data["streams"], errors="coerce").isna()]

     print("Sample of non-numeric values in the 'streams' column:")
     print(df_non_numeric[["track_name", "streams"]])
```

```
Sample of non-numeric values in the 'streams' column:
                                track_name  \
574  Love Grows (Where My Rosemary Goes)

                                                   streams
574  BPM110KeyAModeMajorDanceability53Valence75Ener...
```

Delete the 574 row

```
[16] data = data.drop(574)
     data['streams'] = data['streams'].astype(int)
     print(data['streams'].dtype)
```

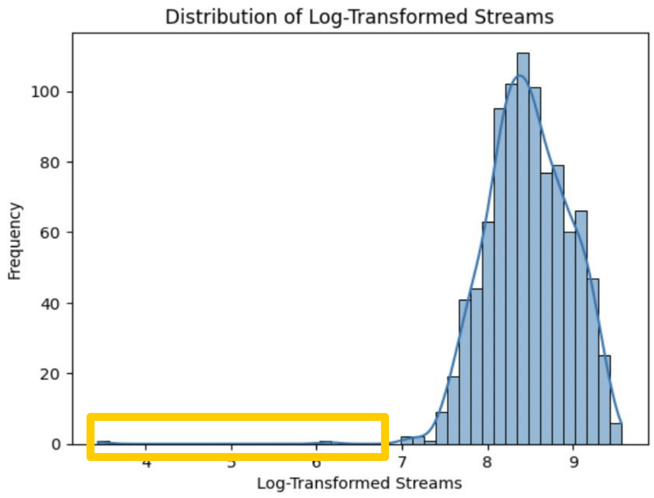
```
int64
```

Current Progress | Data Cleaning

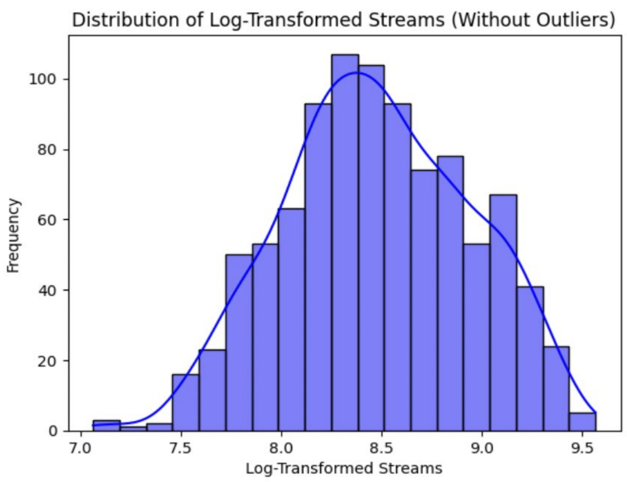
- Dealing with Outlier
  - Seaborn & Matplotlib to draw the Distribution Histograms

```
sns.histplot(data['streams_log'], kde=True)
plt.title('Distribution of Log-Transformed Streams')
plt.xlabel('Log-Transformed Streams')
plt.ylabel('Frequency')
plt.show()
```

```
data_cleaned = data[~data.index.isin(outliers.index)]
# Plotting the distribution of log-transformed streams without outliers
sns.histplot(data_cleaned['streams_log'], kde=True, color="blue", edgecolor='black')
plt.title('Distribution of Log-Transformed Streams (Without Outliers)')
plt.xlabel('Log-Transformed Streams')
plt.ylabel('Frequency')
plt.show()
```



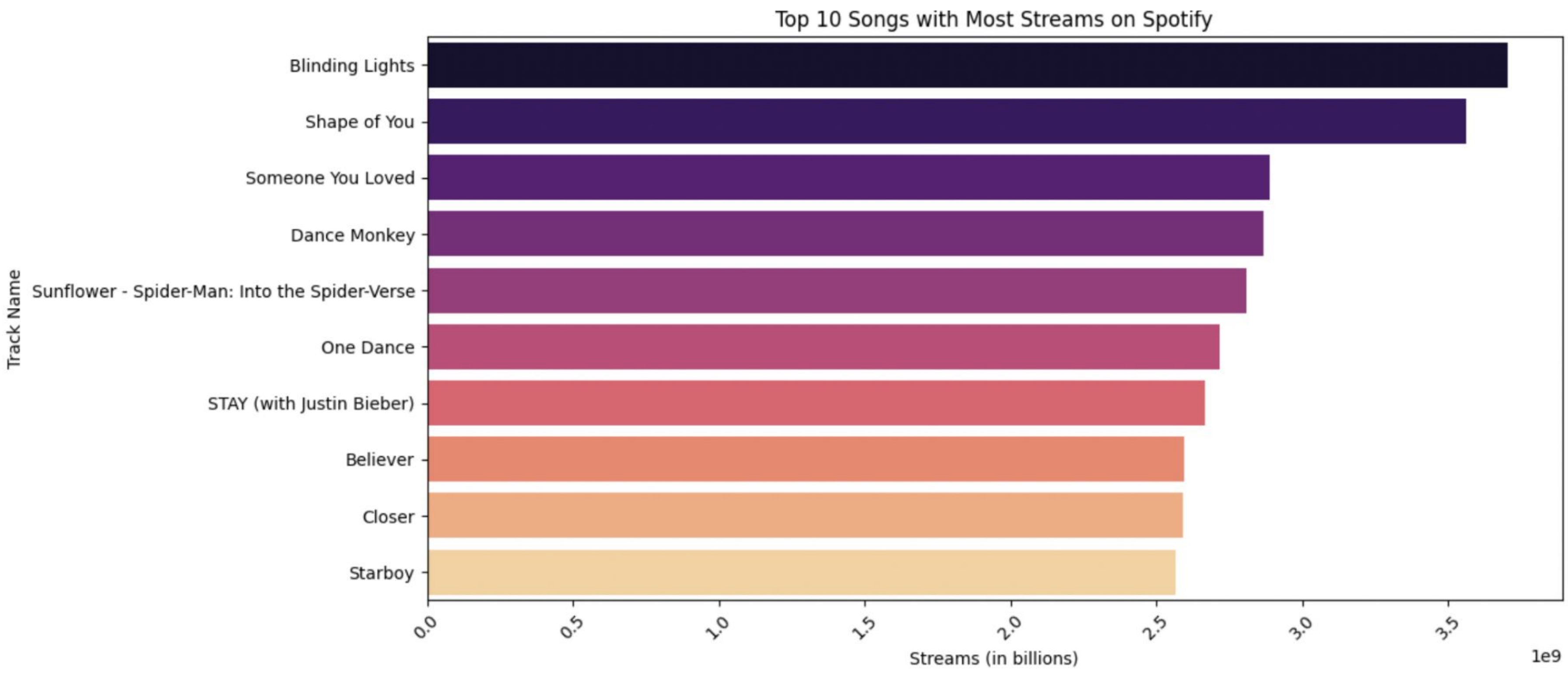
With Outlier



Without Outlier

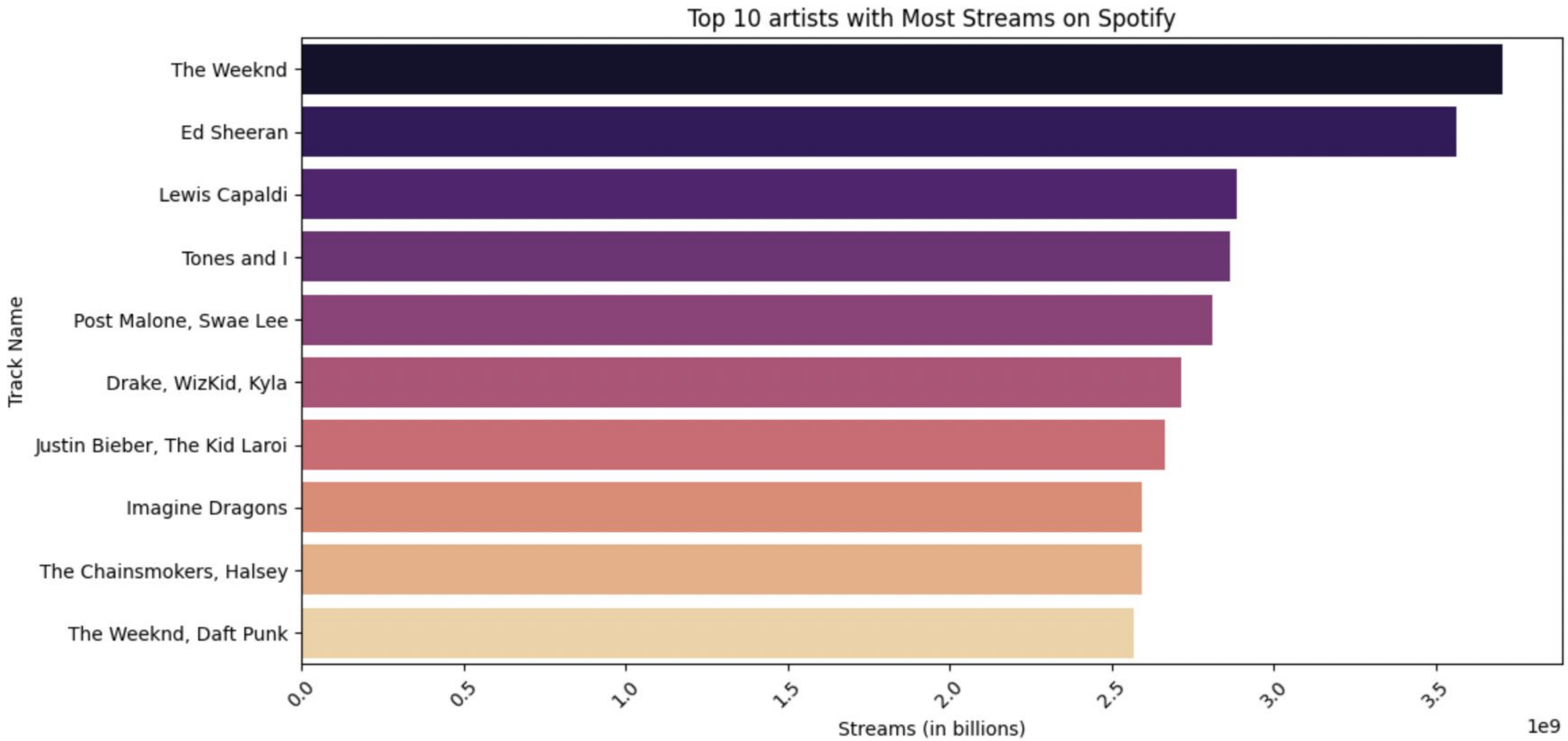
Current Progress | EDA Analysis (Visualizations)

Top 10 songs with the highest number of streams on Spotify



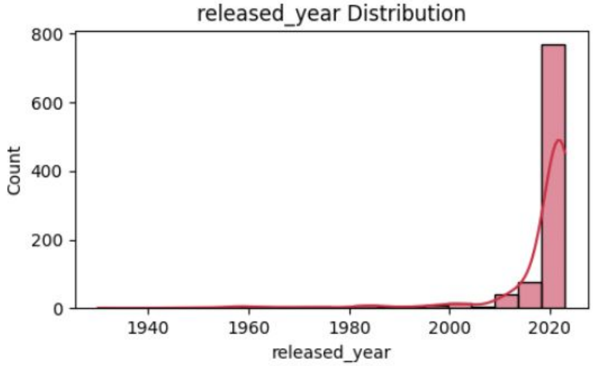
Current Progress | EDA Analysis (Visualizations)

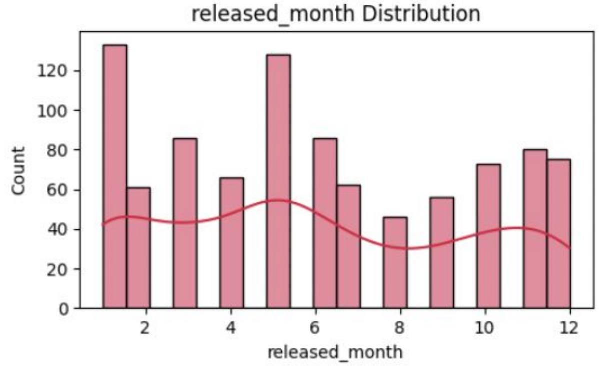
Top 10 artists with the highest number of streams on Spotify

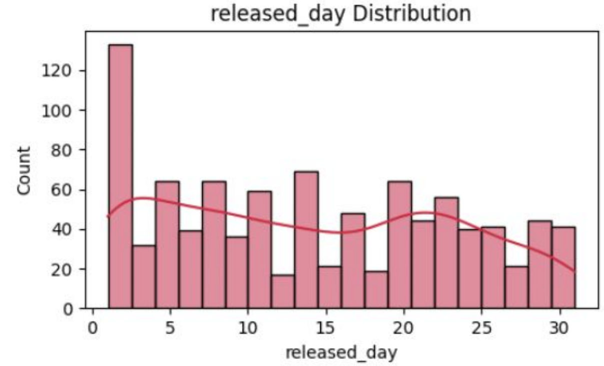


Current Progress | EDA Analysis (Visualizations)

Univariate Analysis for the released time



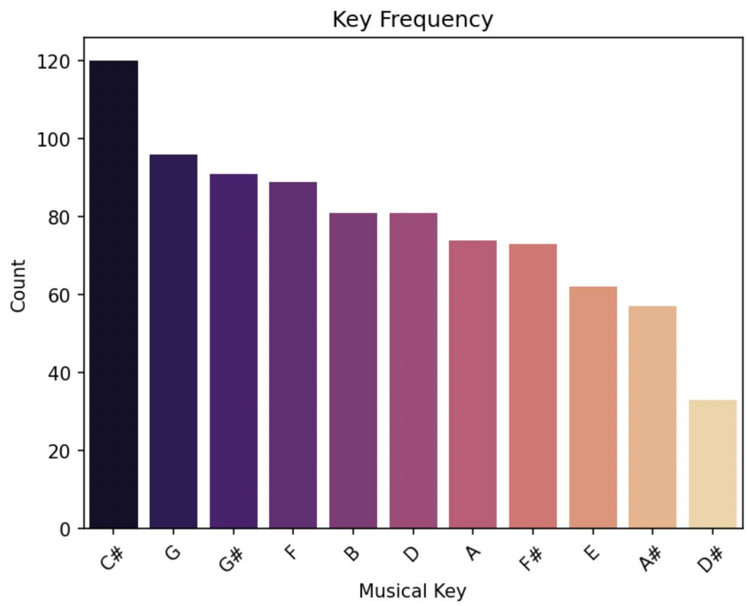


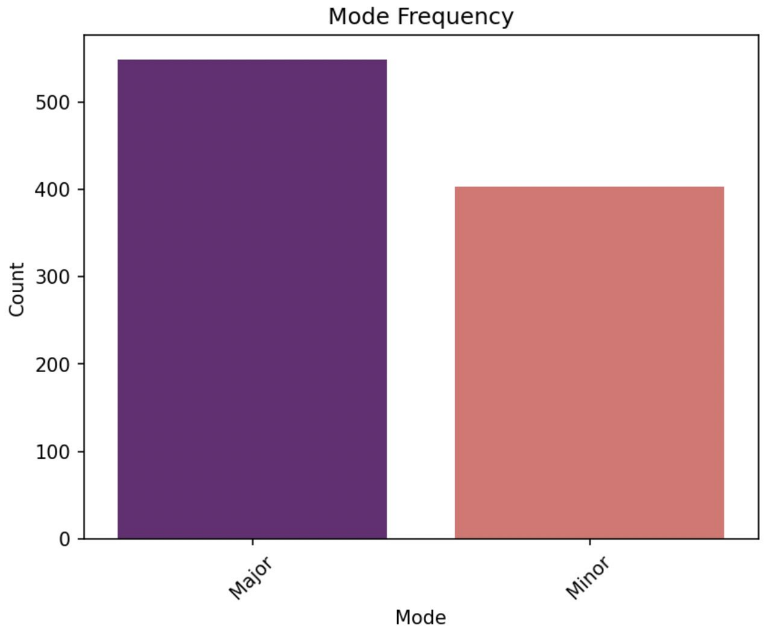


- Most of top-streamed songs have been released in recent years
- Songs released in January and May, as well as those released at the beginning of each month are more frequently found among the top 1000 high-streamed songs.
  - January: Spotify users actively search for new songs to start the new year
  - May: Great season to travel and have festivals to promote new songs

Current Progress | EDA Analysis (Visualizations)

Univariate Analysis for key and mode



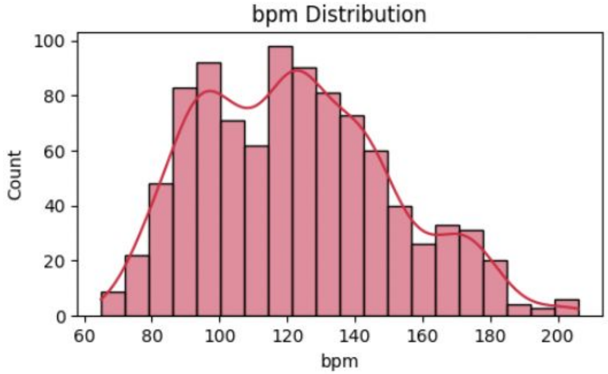


The Key of C# and the Major mode are the most frequently used among the top 1000 most-streamed songs

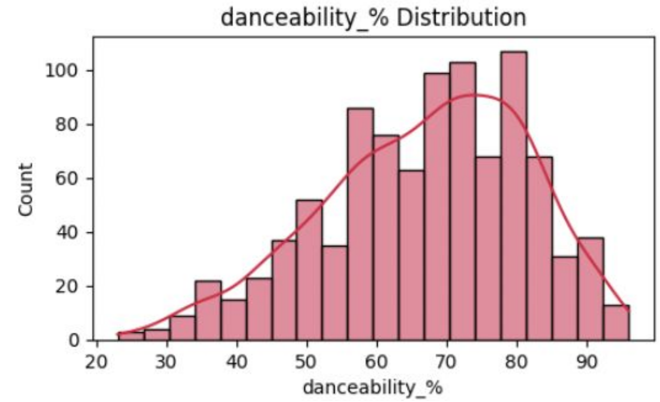
Current Progress | EDA Analysis (Visualizations)

Univariate Analysis for the music features

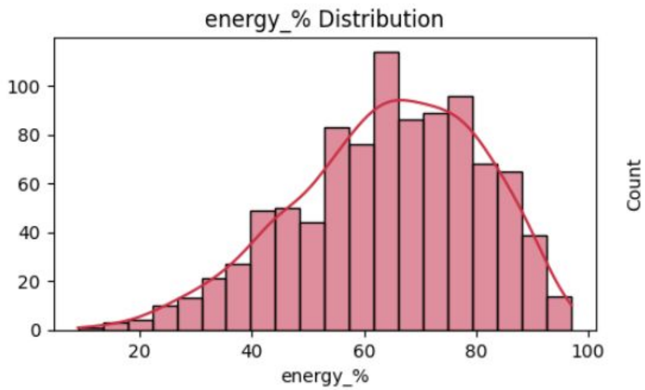
- The majority of the most-streamed songs typically feature a moderately fast BPM, relatively high levels of danceability and energy, along with a broad distribution of valence.



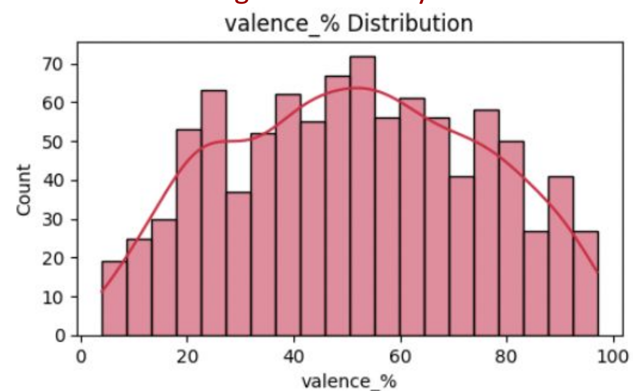
Average BPM is 122



Average danceability % is 67%



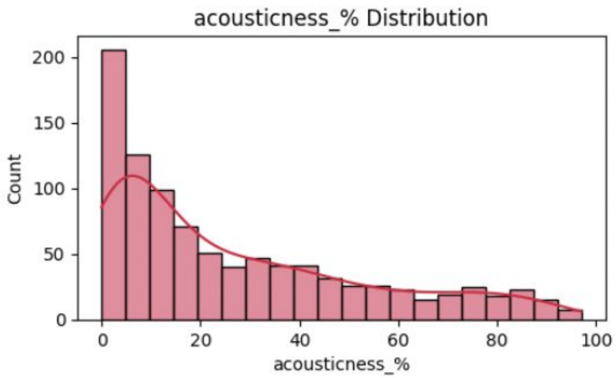
Average energy % is 64%

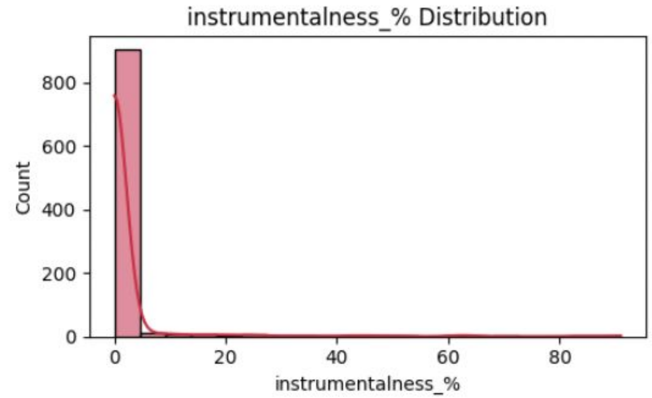


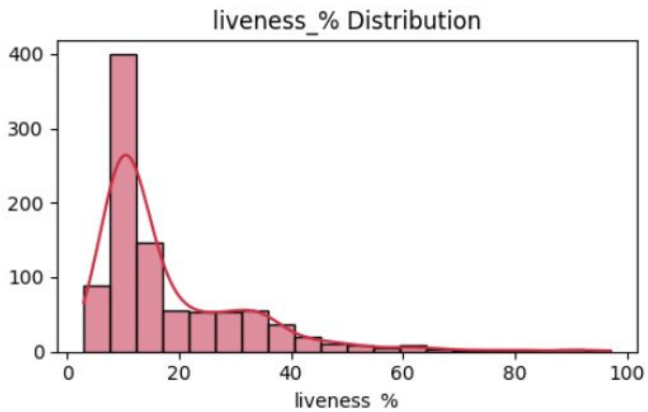
Average valence % is 51%

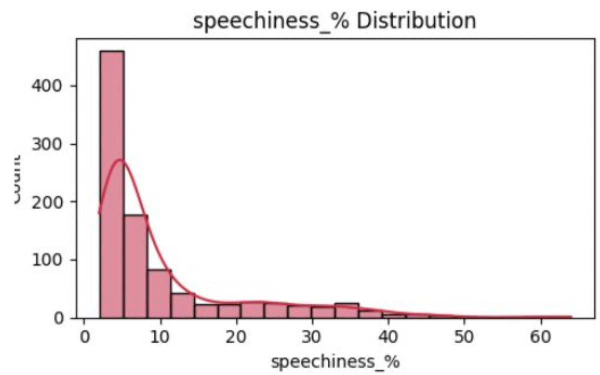
Current Progress | EDA Analysis (Visualizations)

Univariate Analysis for the music features





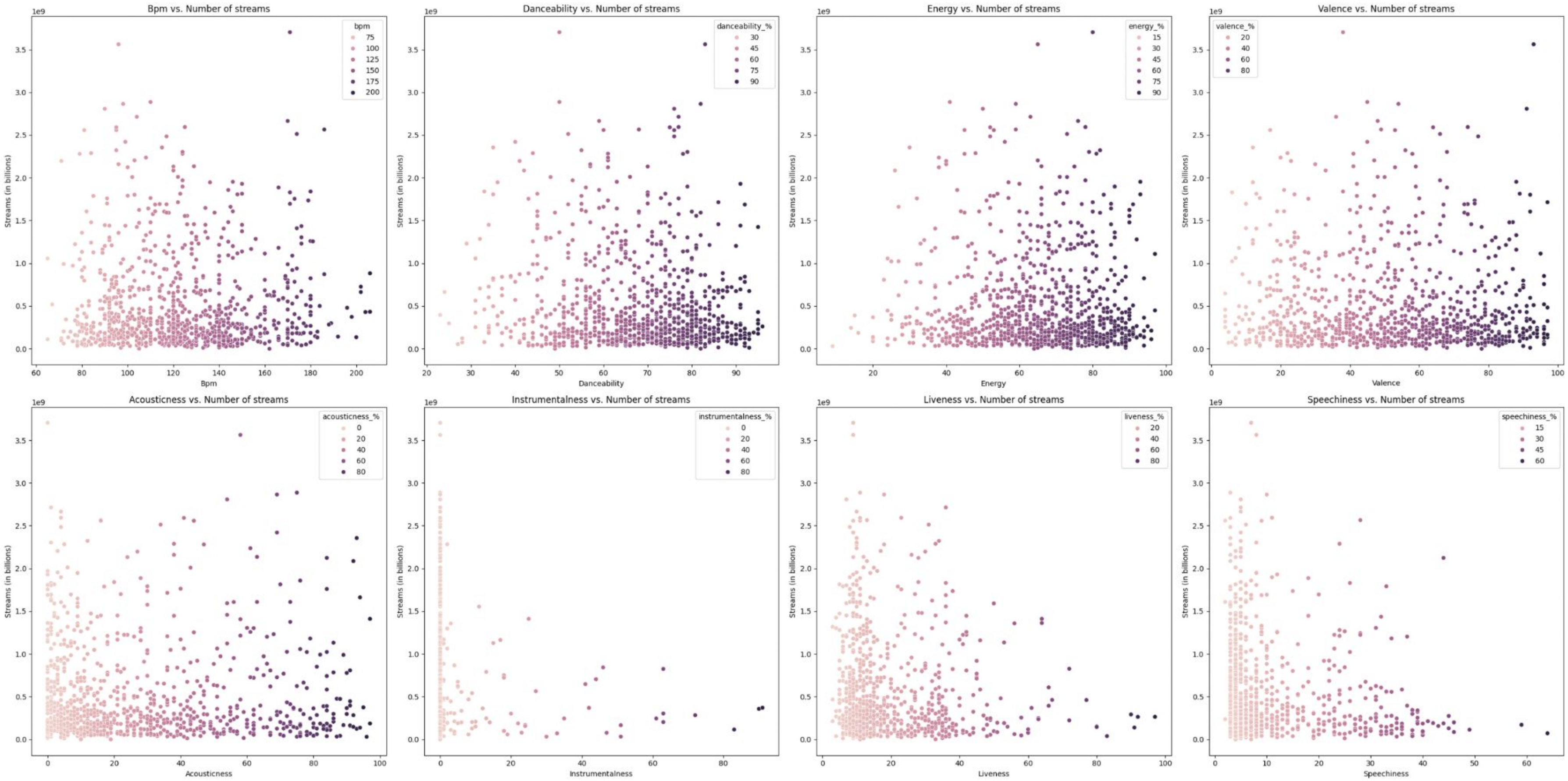




The majority of the most-streamed songs have a low percentage of acousticness, instrumentalness, liveness, and speechiness in their musical characteristics.

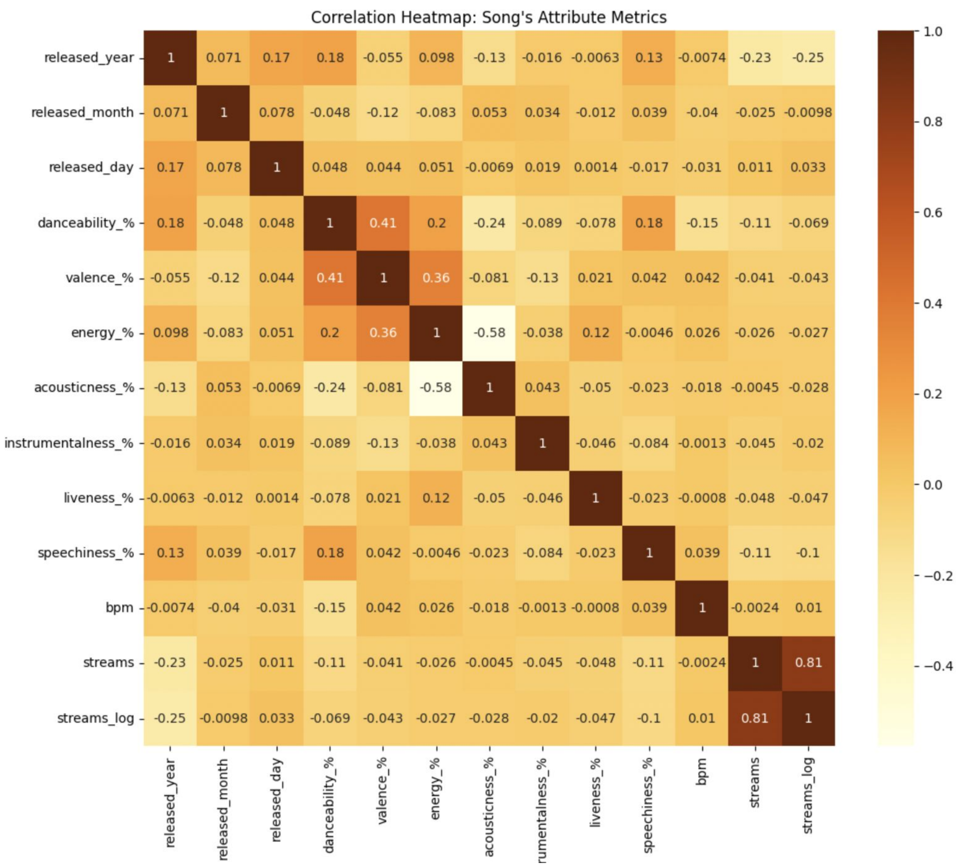
Current Progress | EDA Analysis (Visualizations)

Bi-variate Analysis for the music features with streams (show similar patterns as Univariate Analysis)



Current Progress | EDA Analysis (Visualizations)

Bi-variate Analysis for the music features with streams



Highly Correlated Value

- Danceability and valence are positive correlated
- Energy and valence are positive correlated

Current Progress | Linear Models

The linear regression model's performance, both with a single feature ("speechiness_%") and a combination of features ("speechiness_%" and "liveness_%"), demonstrated limited predictive power.

The decision tree model with a single feature ("speechiness_%") gave the best predictive result, with a mean squared error of 0.188 and R^2 Score of 0.028.

```
Decision Tree - Mean Squared Error: 0.18841488483452115
Decision Tree - R^2 Score: 0.02781008646462057б
```

Future Step

- In the following weeks, we will identify and test several new variables from other datasets as potential predictor for “streams_log” variable.

- In addition, we will categorize songs as “hit” and “not hit” based on their streams and develop classification models to predict if the given song will become hit.

Reference

- International Federation of the Phonographic Industry. (2023). *IFPI Global Music Report 2023.* IFPI. https://globalmusicreport.ifpi.org/

- Araujo, C. V. S., Cristo, M. A. P., & Giusti, R. (2020). A model for predicting music popularity on streaming platforms. *Revista de Informatica Teorica e Aplicada*, *27*(4), 108–117. https://doi.org/10.22456/2175-2745.107021

Thank you!

DSCI 550: Data Science at Scale

Team #02:
Ying Yang (Leader)
Yixin Qu
Zhiqian Li
Wenjing Huang