CONTENT

INTRODUCTION

Music is a universal language that can express emotions, convey messages, and inspire people. Music can be a subject of analysis, as it has various aspects such as melody, harmony, rhythm, lyrics, and genre. Analyzing music can help us understand its meaning, structure, and influence on society and culture. Music analysis can also reveal the connections between music and other aspects of world . For these reasons, music analysis can be regarded as a valuable and rewarding activity that can enrich our knowledge and experience of music.

Data scrawling

Dataset requirements:

- Legal.
- Variety.
- Quality.
- Easy as fast to get.

Exploratory Data

- Shape of dataset

  3229 rows and 16 columns

- Meanings of each columns:

  The table in the next slide.

- Datatype of each columns:

  float64(9), int64(3), object(4)

Column Name	Description
name	The title of the track.
album	The album to which the track belongs.
artist	The artist or artists who performed the track.
release_date	The date when the track was released.
length	The duration of the track in milliseconds.
popularity	The popularity of the track, measured on a scale from 0 to 100, where 100 is the most popular.
danceability	How suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. The value ranges from 0 to 1, with 1 being the most danceable.
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.
energy	A perceptual measure of intensity and activity of a track, typically energetic tracks feel fast, loud, and noisy. The value ranges from 0.0 to 1.0.
instrumentalness`	Predicts whether a track contains no vocals. The value ranges from 0.0 to 1.0, with 1.0 representing a higher likelihood that the track is instrumental.
liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
loudness	The overall loudness of the track in decibels (dB). The values typically range between -60 and 0 dB.
speechiness	Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audiobook, poetry), the closer to 1.0 the attribute value.
tempo	The overall estimated tempo of a track in beats per minute (BPM).
time_signature	The 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).

PREPROCESSING

An important stage for all models

Tasks to do:

- **Are there any duplicate?**

  Yes, there are 437 rows of duplicates. Because a song can be included in many playlists. -> remove them.

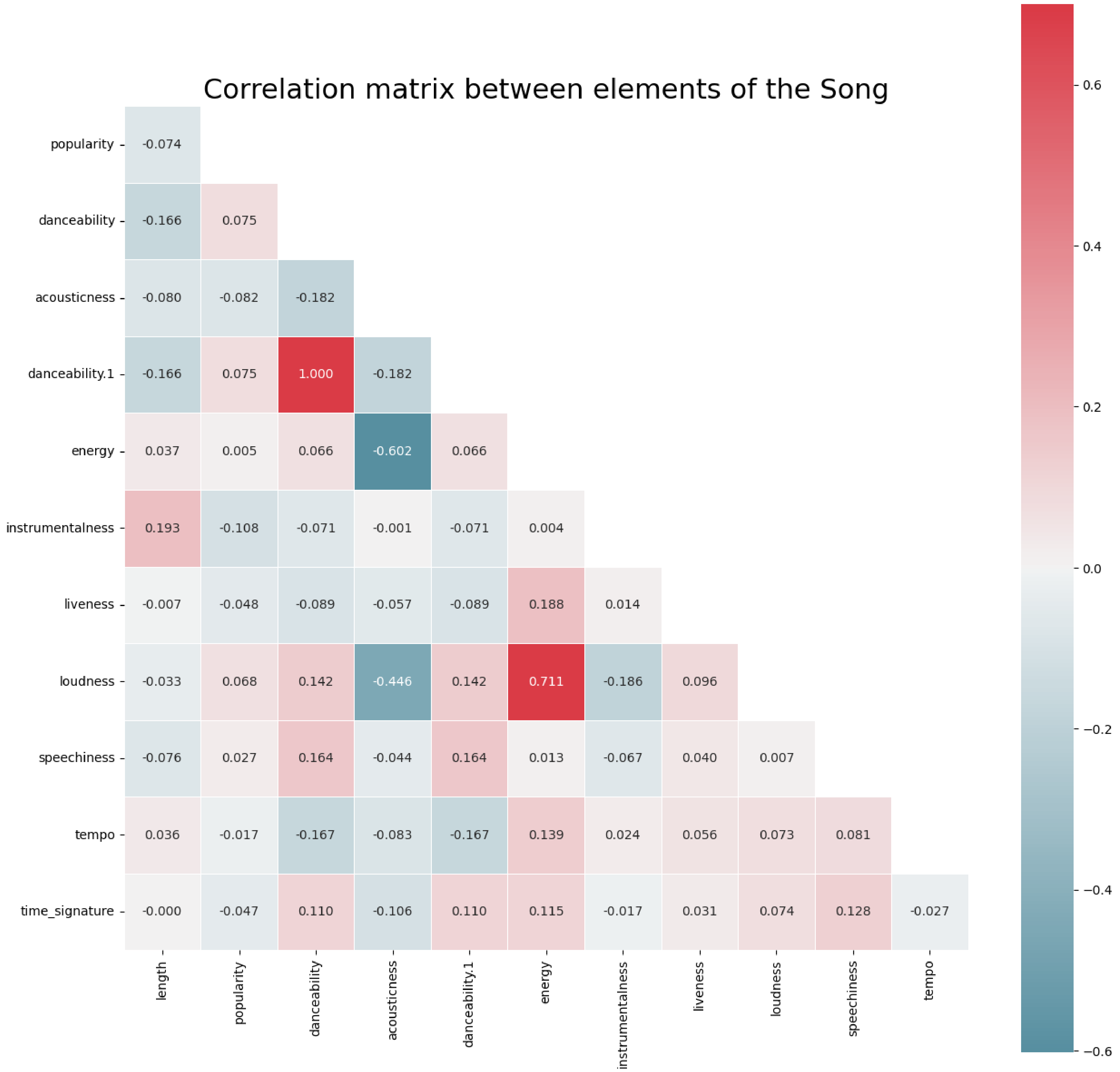
- **Nan handling and datatype converting**

  Due to no NULL value, and datatype is proper so this activity is free.

- **Analyze correlation between features**

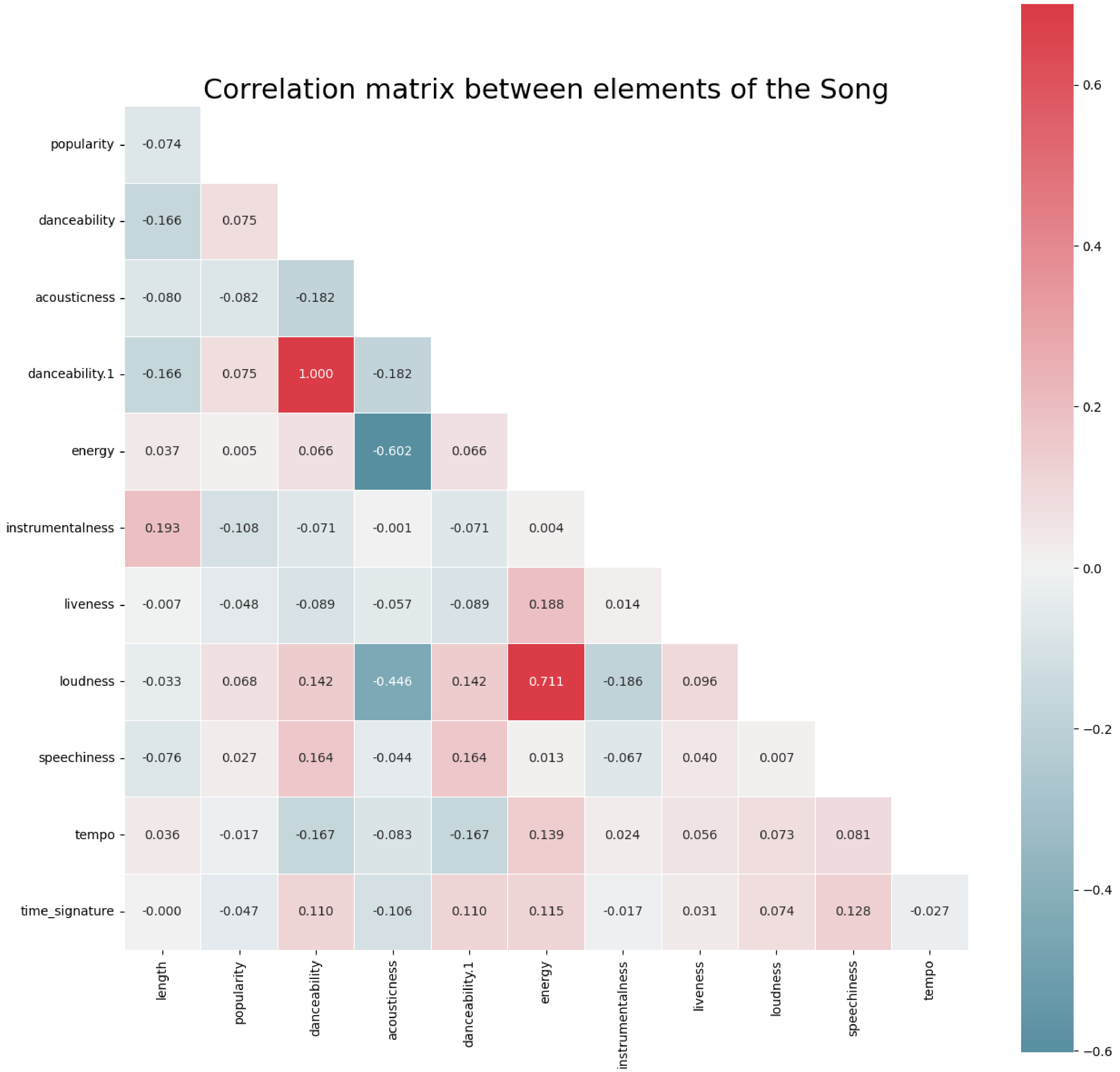
RESULT

- **energy** and **loudness** have the highest correlation (0.711)
- **energy** - **acouticness** have the lowest correlation (-0.602)
- moderate negative correlation between **acouticness** and **loudly** (- 0.446)



RESULT

- The other correlation seems to quite faint to be brought under consideration.
- I found that two fields **danceability.1** and **danceability** have abnormal relationship. Let's find what happened?



RESULT

```
res = df['danceability.1'] != df.danceability
res.sum()
```

```
0
```

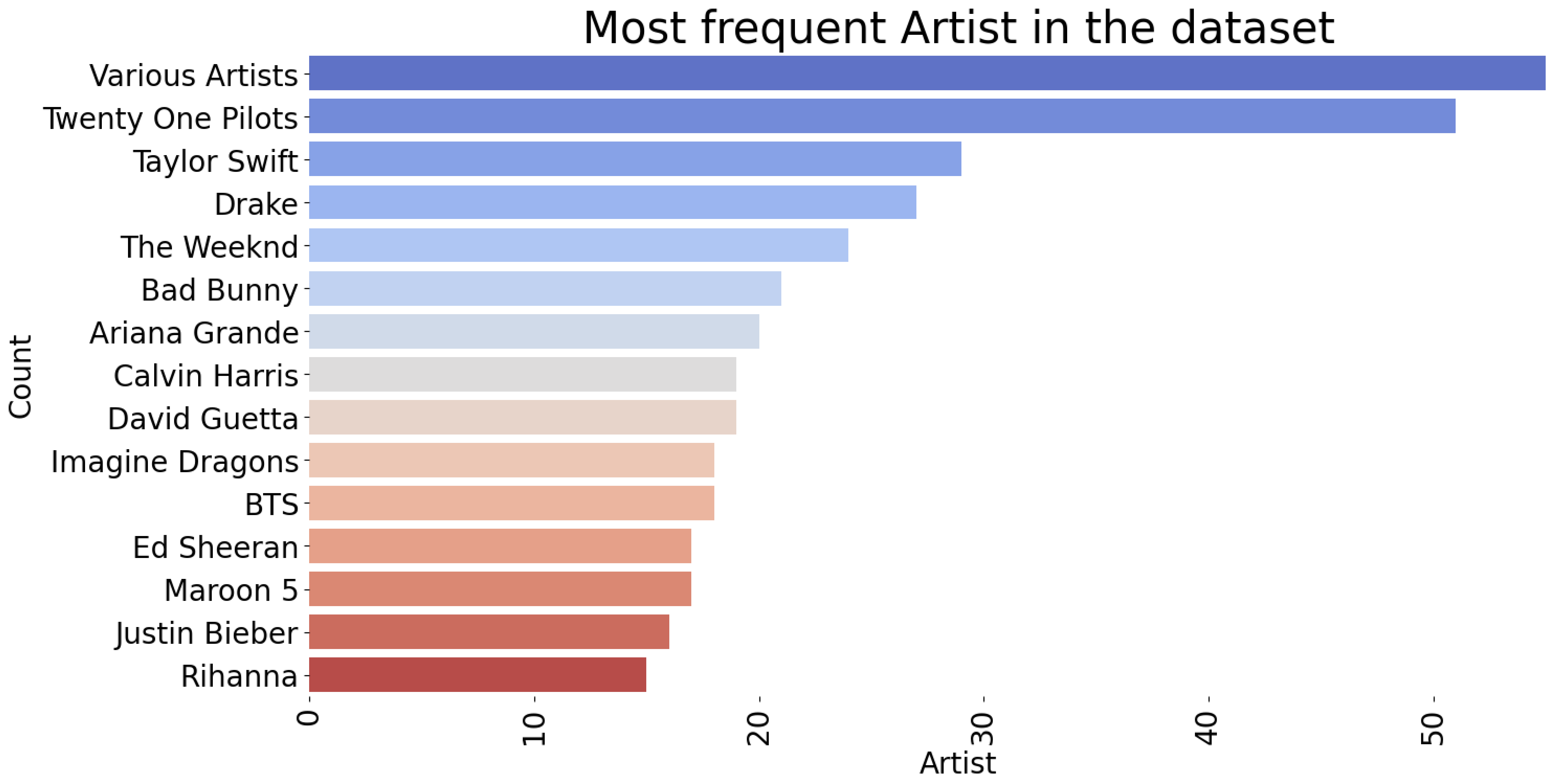
RESULT : Due to no diffence between the 2 columns. I will drop column ['danceability.1']

```
df = df.drop(columns = ['danceability.1'],axis = 1)
df.head(5)
```

After preprocessing stage, all the data set is save into ‘clean_data.csv’

Questions:

The data set is collected mainly from many Top 100 songs in the years 2021,2022,2023,... Let see how frequently the artist appeared in this dataset?


Most frequent Artist in the dataset
Various Artists
Twenty One Pilots
Taylor Swift
Drake
The Weeknd
Bad Bunny
Ariana Grande
Calvin Harris
David Guetta
Imagine Dragons
BTS
Ed Sheeran
Maroon 5
Justin Bieber
Rihanna
Count
0
10
20
30
40
50
Artist

Analysis

Twenty One Pilots is the most popular artist in the dataset, with a count of nearly 50 songs.

This could mean that the dataset was collected from a source that favors this artist, such as a fan playlist. Alternatively, it could mean that Twenty One Pilots has a large and loyal fan base that listens to their songs frequently.

Analysis

Rihanna is the least frequent artist in the Top 15, with a count of nearly 15

This could mean that the dataset was collected from a source that does not favour this artist much. Alternatively, it could mean that she has a low demand or a declining popularity among the listeners of the dataset. The second hypothesis is more likely to be true, because we all know that It's a long time since her last song has been released.

Analysis

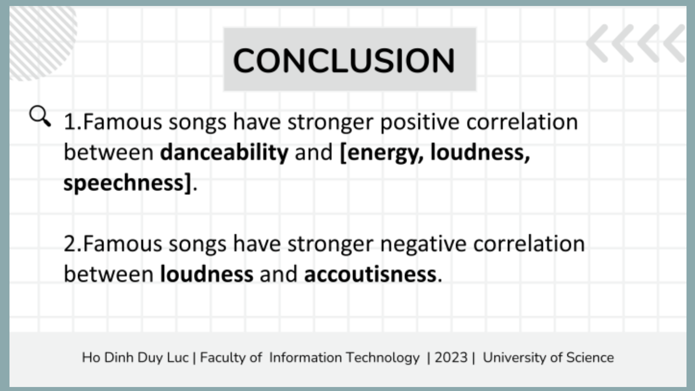
- There is a gap between the top four artists (Twenty One Pilots, Taylor Swift, Drake and The Weeknd) and the rest of the artists.

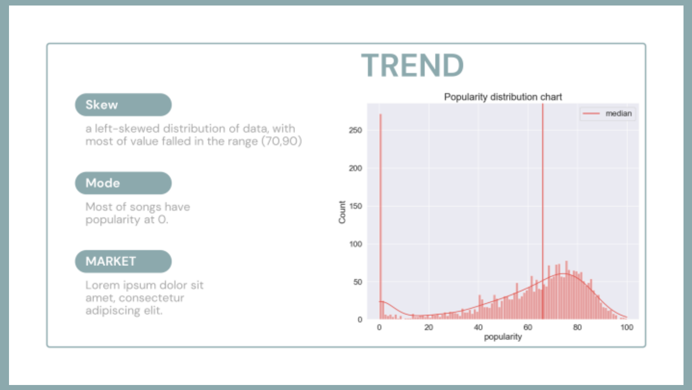
  This could mean that the dataset has a skewed distribution, where a few artists dominate the majority of the songs. This could also reflect the current trends and preferences of the music industry and the listeners.

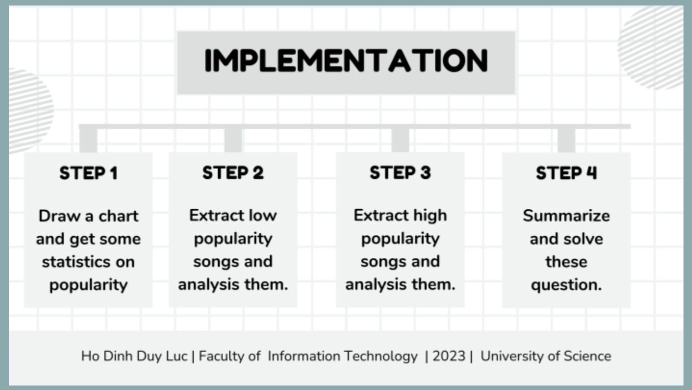
02.

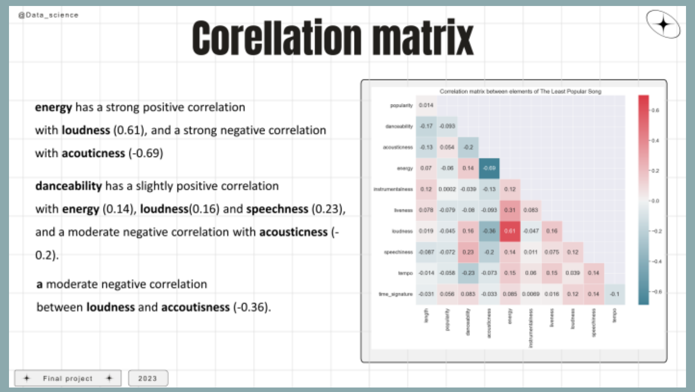
How to become popular?

Deeper insights.









IMPLEMENTATION

STEP 1

Draw a chart and get some statistics on popularity

STEP 2

Extract low popularity songs and analysis them.

STEP 3

Extract high popularity songs and analysis them.

STEP 4

Summarize and solve these question.

TREND

SKEW

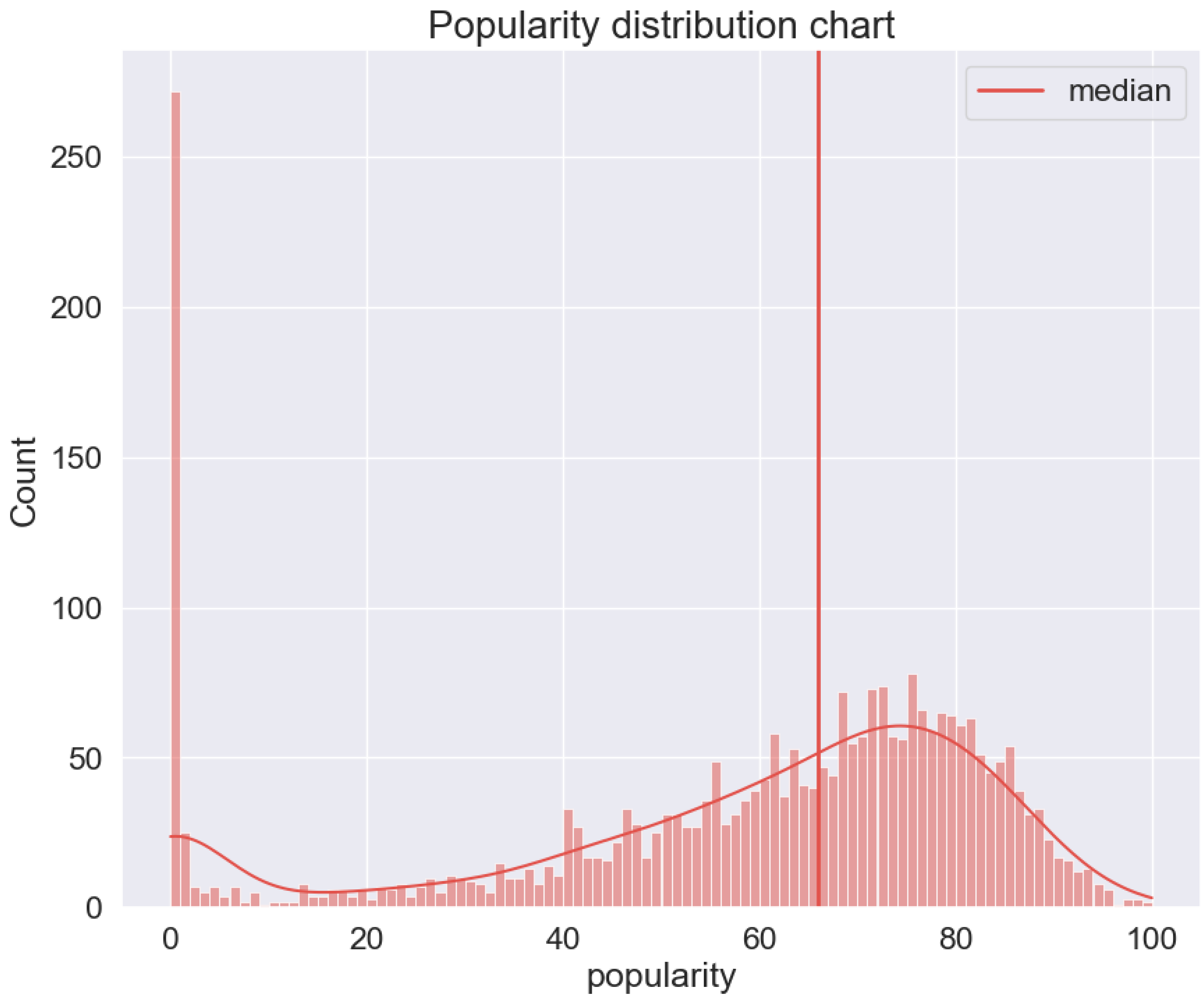
A left-skewed distribution of data, with most of value falled in the range (70,90)

MODE

Most of songs have popularity at 0.

MEDIAN

Median value recorded at 66.0

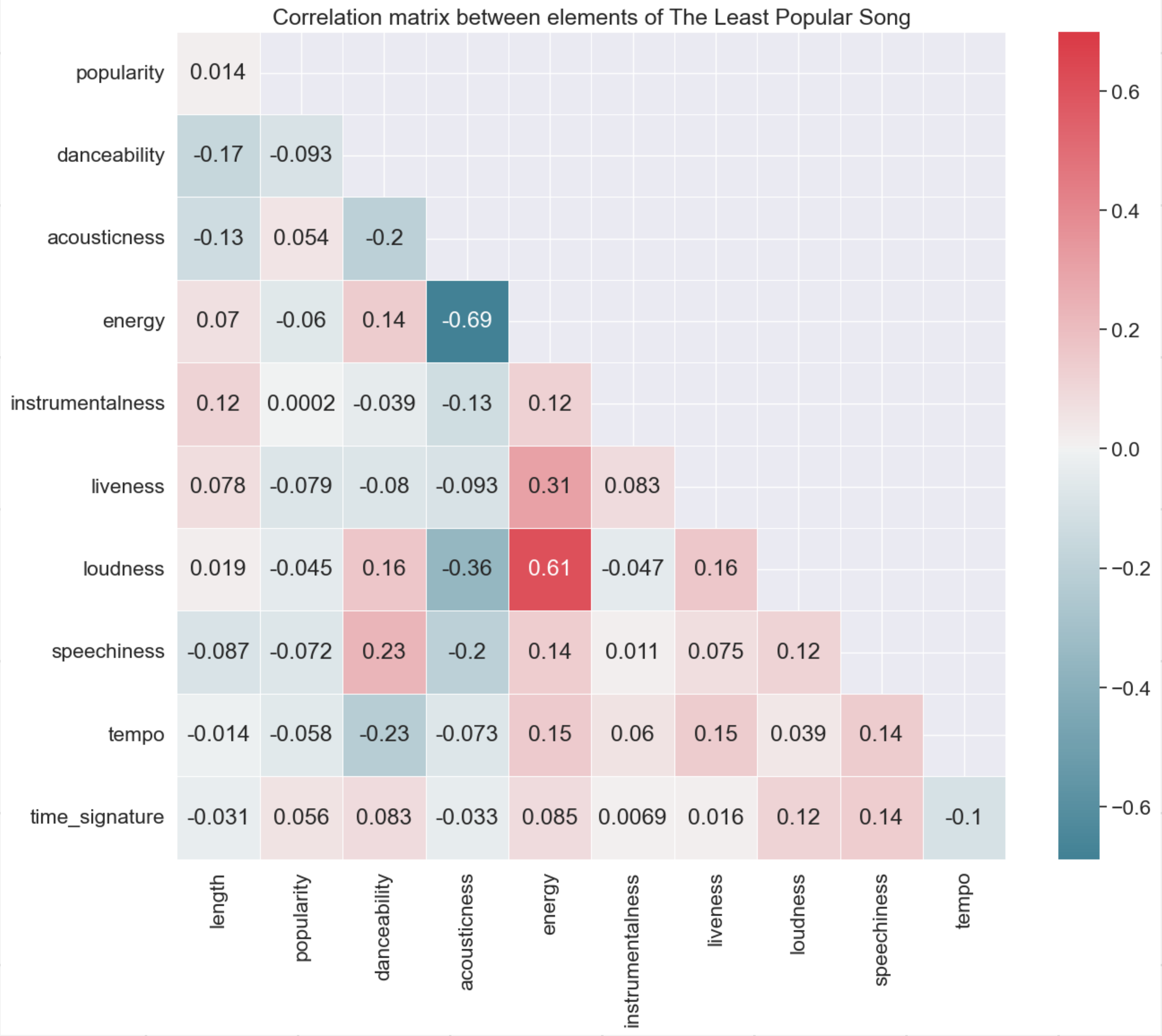


Corellation matrix

energy has a strong positive correlation with **loudness** (0.61), and a strong negative correlation with **acouticness** (-0.69)

danceability has a slightly positive correlation with **energy** (0.14), **loudness**(0.16) and **speechness** (0.23), and a moderate negative correlation with **acousticness** (-0.2).

a moderate negative correlation between **loudness** and **accoutisness** (-0.36).

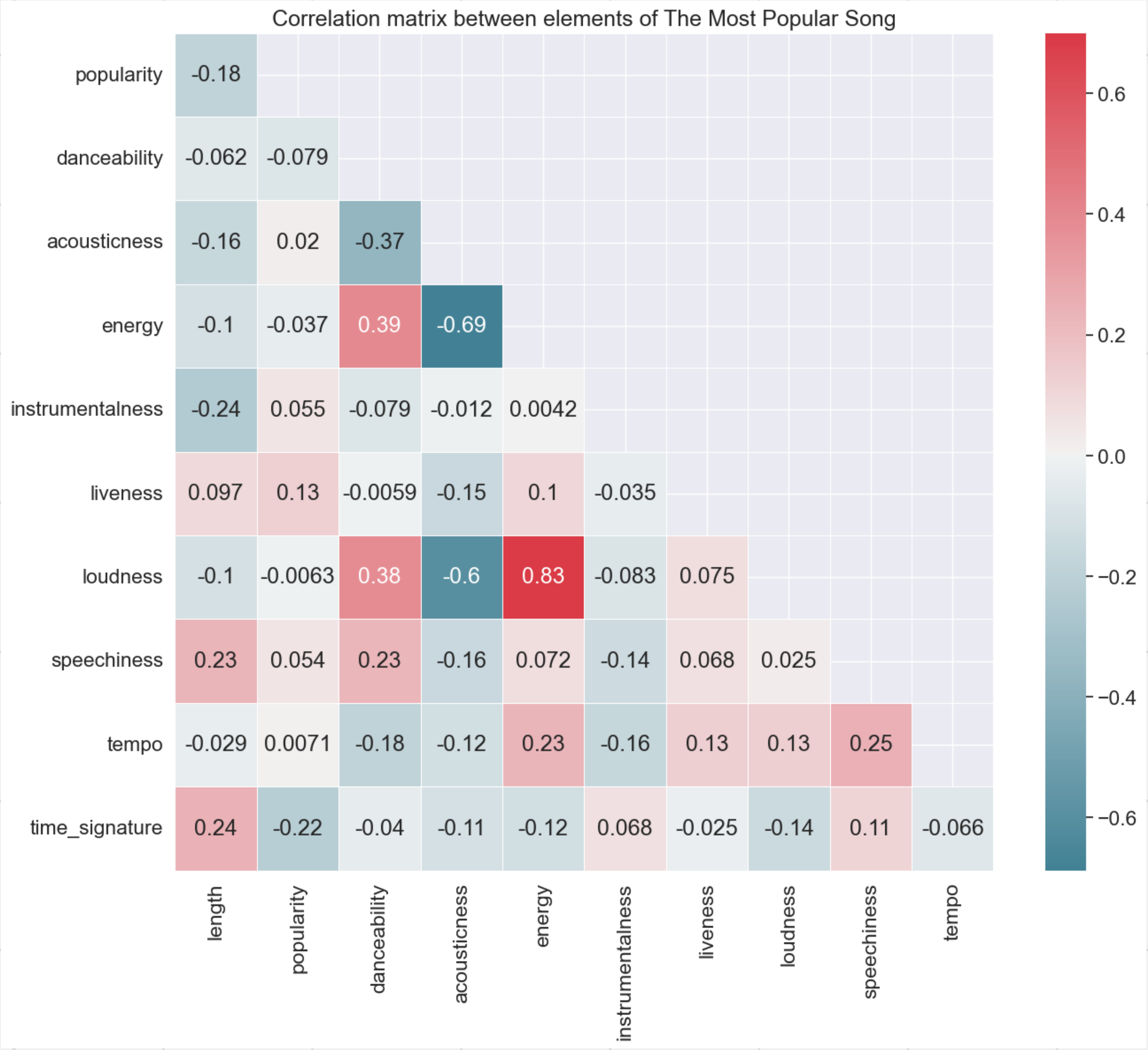


Corellation matrix

energy has a strong positive correlation with **loudness** (0.83), and a strong negative correlation with **acousticness** (-0.69)

danceability has a moderate positive correlation with **energy** (0.39), **loudness**(0.38) and **speechness** (0.23), and a moderate negative correlation with **acousticness** (-0.34).

a significant negative correlation between **loudness** and **accoutisness** (-0.6)



CONCLUSION

1.Famous songs have stronger positive correlation between **danceability** and **[energy, loudness, speechness]**.

2.Famous songs have stronger negative correlation between **loudness** and **accoutisness**.

03.

The most contributer

Deeper insights.

IMPLEMENTATION

STEP 1

Extract data from dataset

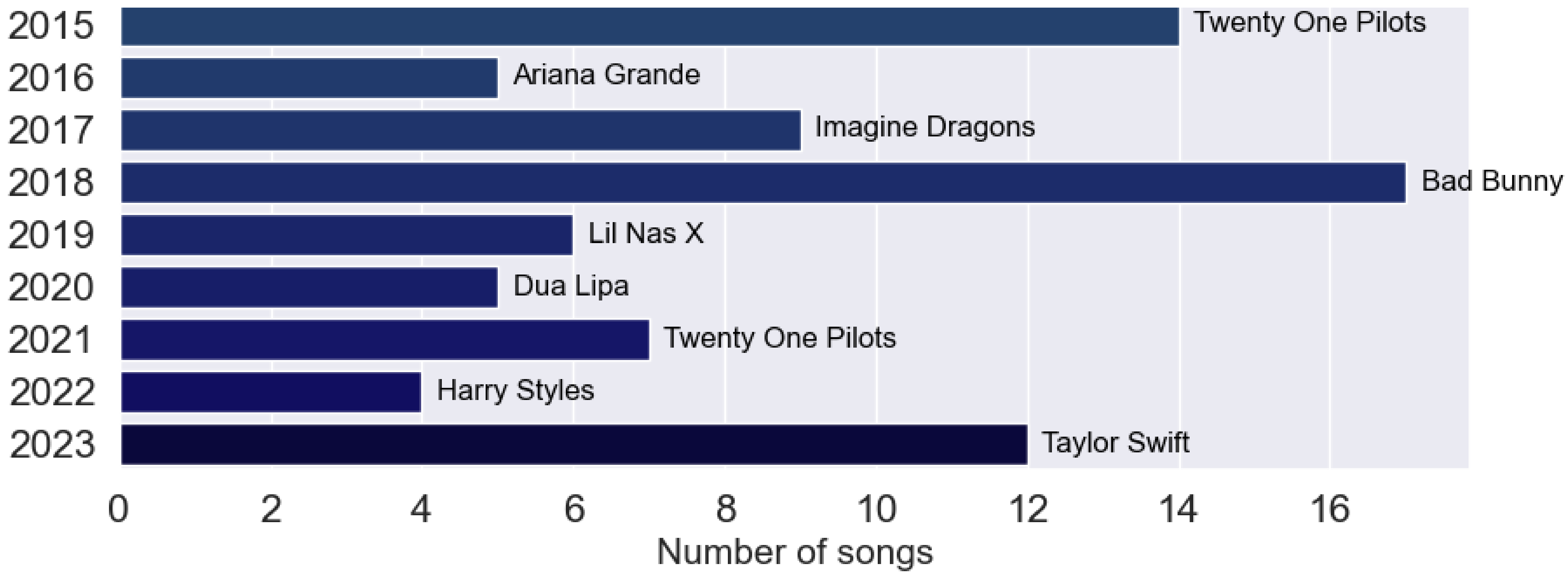
STEP 2

Create a matrix

STEP 3

Group data and visualize result.

Artist with most song from 2015 - 2023



04.

WHICH MONTH IS BEST FOR A NEW SONG RELEASE?

Deeper insights.

IMPLEMENTATION

STEP 1

Filter song which have high popularity.

STEP 2

Extract month data

STEP 3

Group data and visualize result.

TREND

BEST

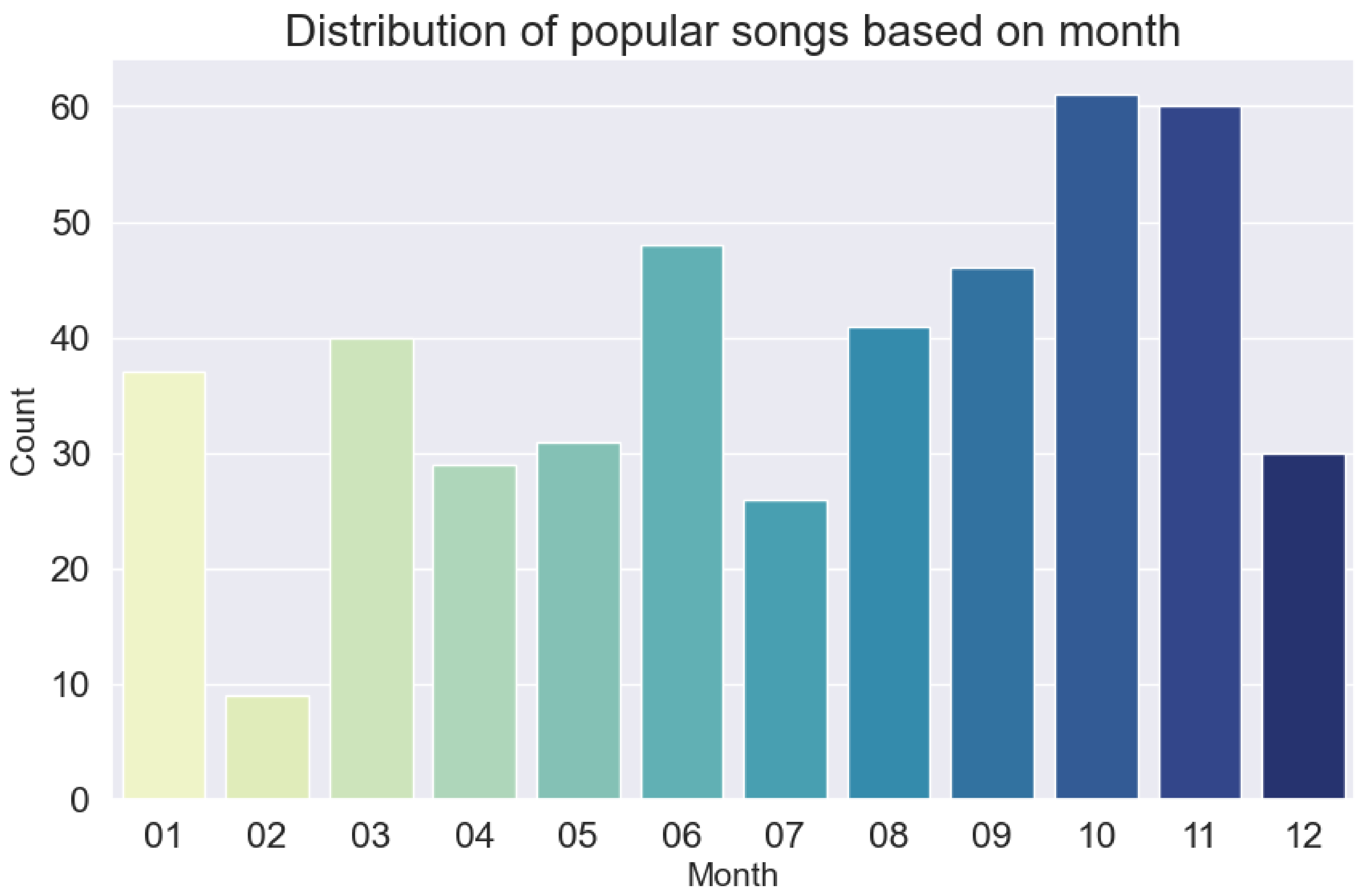
October, November

WORST

February

OPTINAL CHOICE

January, March, June, August, September.



Analysis

- The popularity of songs is influenced by seasonal factors, such as holidays, weather, and mood.

  For example, in the end of year may have more popular songs because of Christmas songs, winter songs, or songs that reflect the end of the year. January and February may have fewer popular songs because of the post-holiday slump, cold weather, or songs that are too upbeat for the winter blues.

Analysis

- The popularity of songs is also affected by the release dates of new albums, singles, or music videos.
- The popularity of songs is not evenly distributed across the year, but rather follows a cyclical pattern, with peaks and troughs.

05.

Find a familiar rhythm. How to do that?

Deeper insights.

IMPLEMENTATION

Cosine similarity

$$\text{Similarity}(p,q) = \cos\theta = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}}\sqrt{\sum_{i=1}^{n} q_i^{\,2}}}$$

RESULT

TEST

```
id_find = 184
recommend_id = get_recommendations(id_find, cosine_sim)
recommend_id =list(map(lambda id: id + 10000, recommend_id))
origin_song = id_find + 10000
recommend_id.insert(0,origin_song)
recommend_id
```

```
[10184, 10058, 10073, 10479, 10575, 10724, 10989, 11748, 12276, 12539, 12786]
```

RESULT

```
song_id.query('id in @recommend_id')
```

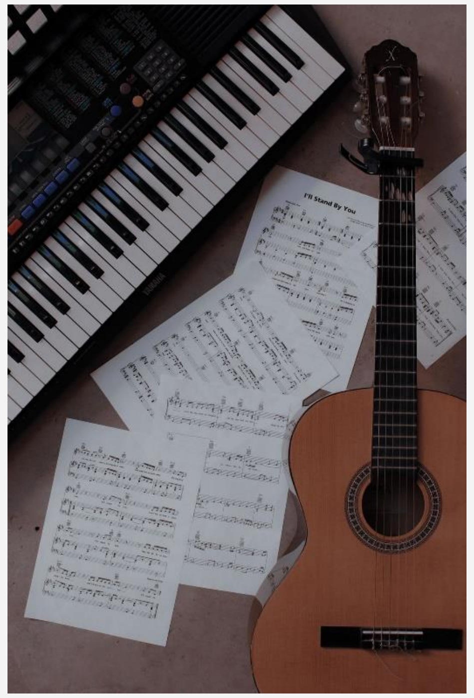
	id	name	album	artist	release_date
58	10058	Money	Money	Cardi B	2018-10-23
73	10073	Ladbroke Grove	AJ Tracey	AJ Tracey	2019-02-08
184	10184	JIKJIN	THE SECOND STEP : CHAPTER ONE	TREASURE	2022-02-15
479	10479	Dime Si Te Acuerdas	Dime Si Te Acuerdas	Bad Bunny	2018-02-22
575	10575	Love Lies (with Normani)	Love Lies (with Normani)	Khalid	2018-02-14
724	10724	Aristocrate	En esprit	Heuss L'enfoiré	2019-01-25
989	10989	INDUSTRY BABY (feat. Jack Harlow)	INDUSTRY BABY (feat. Jack Harlow)	Lil Nas X	2021-07-23
1748	11748	INDUSTRY BABY (feat. Jack Harlow)	MONTERO	Lil Nas X	2021-09-17
2276	12276	Questions	Questions	Lost Frequencies	2022-06-03
2539	12539	Drive (feat. Wes Nelson)	Drive (feat. Wes Nelson)	Clean Bandit	2021-07-30
2786	12786	Let You Down	Perception	NF	2017-10-06

Final project – Introduction to Data Science
Ho Dinh Duy Luc

University of Science
Faculty of Information Technology

Model

06



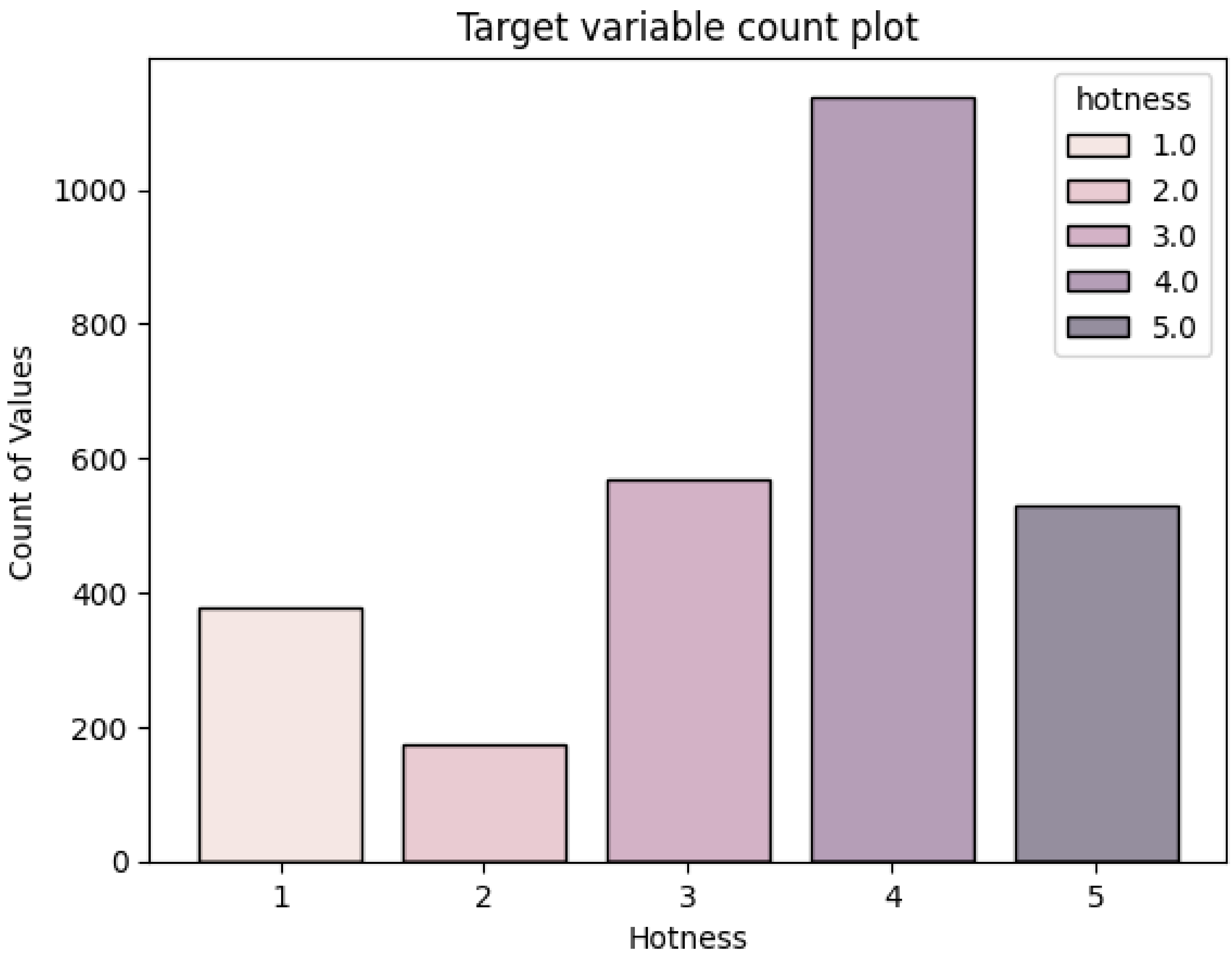
Forecasting a song's popularity is a challenging yet crucial task with extensive applications in the music industry. It involves analyzing factors like musical composition, artist popularity, and cultural trends. Accurate predictions aid record labels, streaming platforms, and artists in optimizing marketing strategies and enhancing the overall music consumption experience. Successfully navigating this dynamic landscape requires a harmonious blend of artistic expression and data-driven insights.

PRESENTATIONS ARE COMMUNICATION TOOLS THAT CAN BE USED AS DEMONSTRATIONS, LECTURES, SPEECHES, REPORTS, AND MORE.

PREPROCESSING

An important stage for all models

✿ Create new columns for labels and stars



Model

- Random Forest
- Naïve Bayes
- MLP Classifier

MODEL 1: RANDOM FOREST CLASSIFIER

```
confusion_matrix(y_test, y_pred)
```

```
array([[104,   1,   1,   1,   0],
       [  0,  37,   0,   0,   0],
       [  0,   1, 144,   0,   0],
       [  0,   0,   0, 283,   4],
       [  0,   0,   0,   0, 122]], dtype=int64)
```

```
print("Accuracy score of Random forest model:" , accuracy_score(y_test, y_pred))
```

```
Accuracy score of Random forest model: 0.9885386819484241
```

MODEL 1: RANDOM FOREST CLASSIFIER

	precision	recall	f1-score	support
1.0	0.97	1.00	0.99	104
2.0	1.00	0.95	0.97	39
3.0	0.99	0.99	0.99	145
4.0	0.99	1.00	0.99	284
5.0	1.00	0.97	0.98	126
accuracy			0.99	698
macro avg	0.99	0.98	0.99	698
weighted avg	0.99	0.99	0.99	698

MODEL 2: Naïve Bayes Classifier

```
confusion_matrix(y_test, NBC_scaled_predict)
```

```
array([[107,   0,   0,   0,   0],
       [  0,  37,   0,   0,   0],
       [  0,   0, 145,   0,   0],
       [  0,   0,   0, 287,   0],
       [  0,   0,   0,   0, 122]], dtype=int64)
```

```
print("Accuracy score of Naive Bayes with scaled test input:" , accuracy_score(y_test, NBC_scaled_predict))
```

```
Accuracy score of Naive Bayes with scaled test input: 1.0
```

MODEL 2: Naïve Bayes Classifier

	precision	recall	f1-score	support
1.0	1.00	1.00	1.00	107
2.0	1.00	1.00	1.00	37
3.0	1.00	1.00	1.00	145
4.0	1.00	1.00	1.00	287
5.0	1.00	1.00	1.00	122
accuracy			1.00	698
macro avg	1.00	1.00	1.00	698
weighted avg	1.00	1.00	1.00	698

MODEL 3: Multi-Layer perceptron Classifier

```
MLPC_model.score(X_test_scaled, y_test)
```

```
0.9469914040114613
```

```
confusion_matrix(y_test, MLPC_scaled_predict)
```

```
array([[106,   1,   0,   0,   0],
       [ 25,   5,   7,   0,   0],
       [  0,   0, 141,   4,   0],
       [  0,   0,   0, 287,   0],
       [  0,   0,   0,   0, 122]], dtype=int64)
```

MODEL 3: Multi-Layer perceptron Classifier

	precision	recall	f1-score	support
1.0	0.99	0.81	0.89	131
2.0	0.14	0.83	0.23	6
3.0	0.97	0.95	0.96	148
4.0	1.00	0.99	0.99	291
5.0	1.00	1.00	1.00	122
accuracy			0.95	698
macro avg	0.82	0.92	0.82	698
weighted avg	0.98	0.95	0.96	698

REFERENCES

- **REFERENCES 1**

  Introduction to Data Science and Programming for Data Science slides and labs.

- **REFERENCES 2**

  Python’s library documents: Pandas, Sklearn, numpy, ...e.t.c

THANK YOU

Presentation by Ho Dinh Duy Luc