# Spotify Hit Songs Analytics & Prediction

Shun-Ting (Trista) Wang

# Introduction

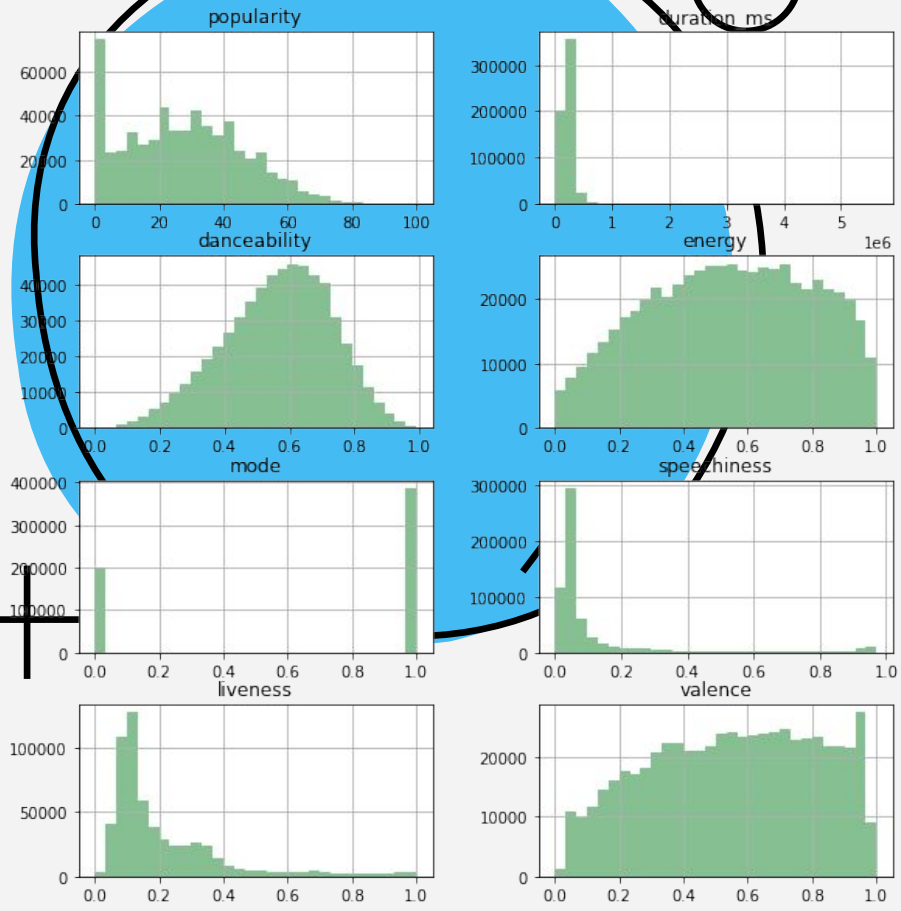
The goal of analyzing hit songs on Spotify is to identify the factors that contribute to a song's success. Spotify calculates a song's "popularity" metric based on its recent streams, rather than just its total number of streams. By using data to predict the popularity of songs in recent years, we can help music industry professionals, such as artists, producers, and record labels, understand what elements to prioritize when creating and promoting new music.

# I will talk about...

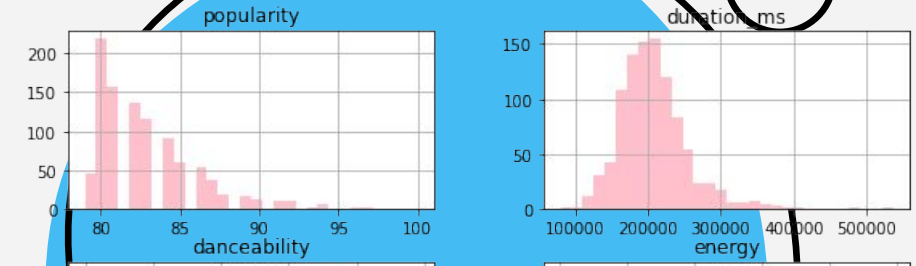
- Distribution of the song features
- The common feature for the top artist's hit song
- The importance of song feature to make a song popular
- The audio characteristics change over the years
- The correlation between the song features
- Predictive model for popular song

# Distribution of audio features

# The whole dataset

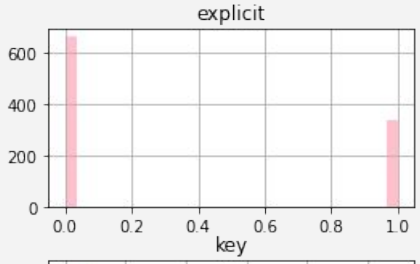


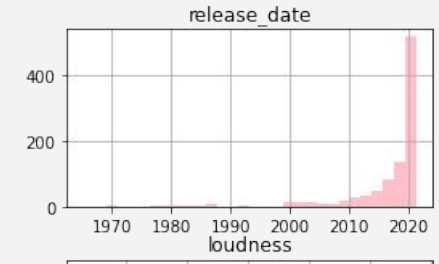
# Top 1000 songs

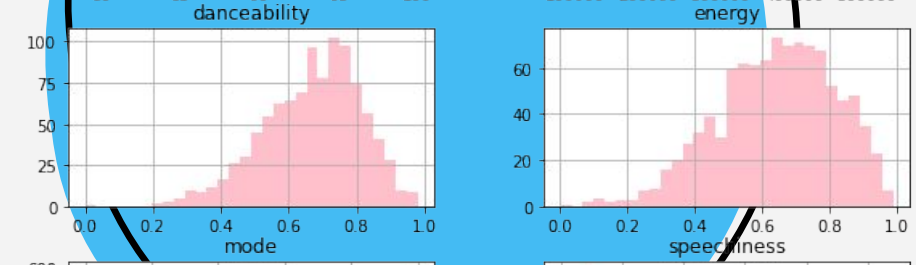






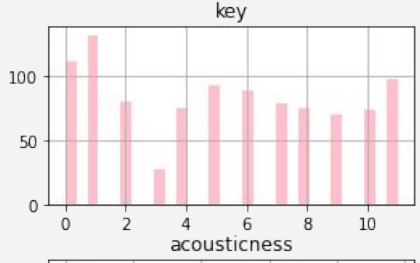


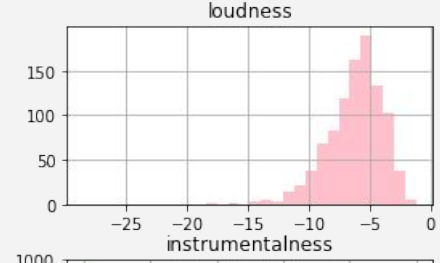



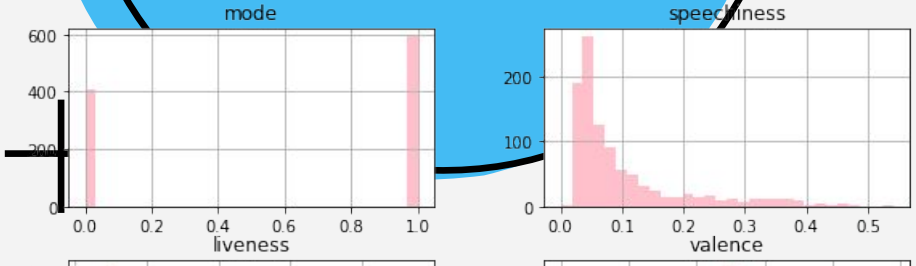







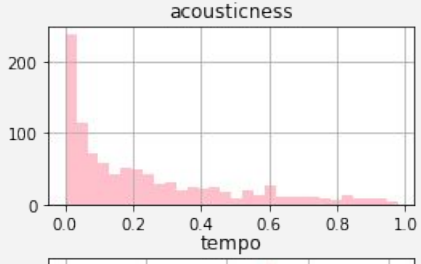


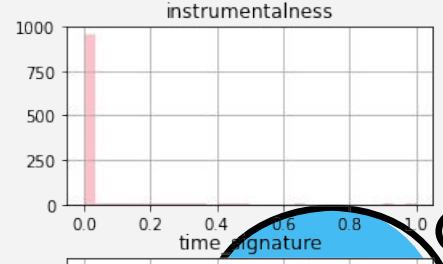



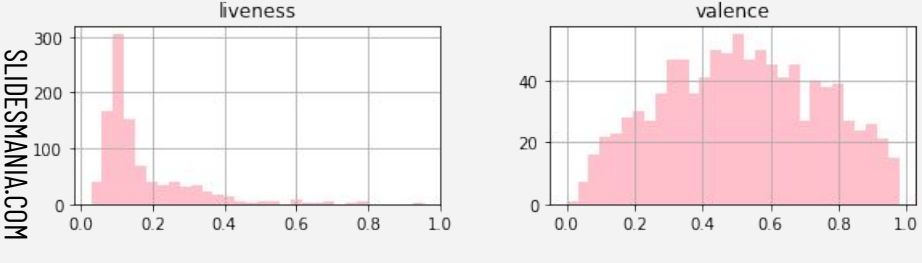







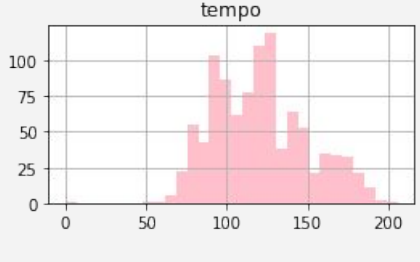


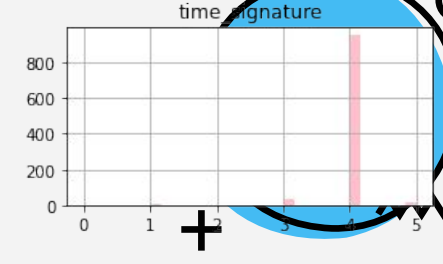









# Features of top 1000 songs

- Top songs are released in recent years.
- Top songs’ duration is shorter.
- Top songs are more explicit, danceable, energetic, louder.
- Top songs are slower, less acoustic, and instrumental.
- The distribution of the audio features for the general data and top songs is identical.

# The common feature for the top 3 artists' hit song

```
top_artists = top_songs['artists'].value_counts().head(20)
top_artists
```

```
Justin Bieber      12
Billie Eilish      11
Juice WRLD         10
Ariana Grande      10
Taylor Swift        9
The Weeknd          9
Bruno Mars          8
Harry Styles        8
Bad Bunny           8
XXXTENTACION        8
Ed Sheeran          8
Imagine Dragons     7
Dua Lipa            7
BTS                 7
Coldplay            6
One Direction       6
Lewis Capaldi       6
Pop Smoke           6
Maroon 5            6
Post Malone         5
```

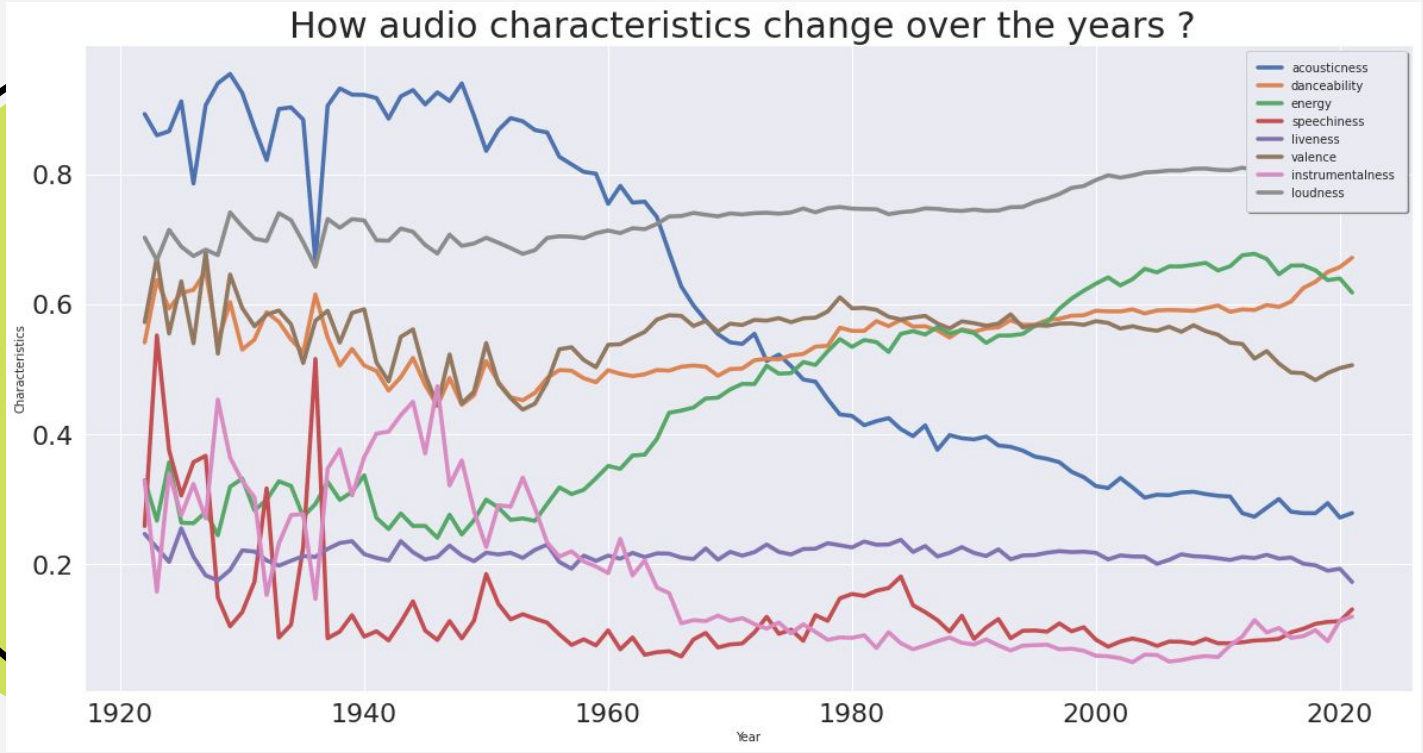
# Top artist's song

- In the top 1000 songs:
  - Justin Bieber has 12 songs
  - Billie Eilish has 11 songs
  - Juice WRLD has 10 songs.

- The acousticness of the top three artists is higher than the average for the top 1000 songs.

- After comparing the features of the songs of three specific artists to the overall song data, I found that the mean danceability of these artists' songs is higher than the average.

# The audio characteristics change over the years




- Acousticness and instrumentalness have been declined.
- Energy, loudness and danceability have been increased.
- As of 2021, the most important features for song popularity are loudness, followed by danceability, energy, and valence.
- Valence has maintained for decades.

# The importance and correlation of song features

```
corr_matrix = df.corr()
corr_matrix["popularity"].sort_values(ascending=False)
```

```
popularity          1.000000
year                0.590801
loudness            0.327002
energy              0.302179
explicit            0.211749
danceability        0.186879
time_signature      0.086713
tempo               0.071224
duration_ms         0.027638
key                 0.015306
valence             0.004560
mode               -0.033652
speechiness        -0.047415
liveness           -0.048736
instrumentalness   -0.236403
acousticness       -0.370724
```

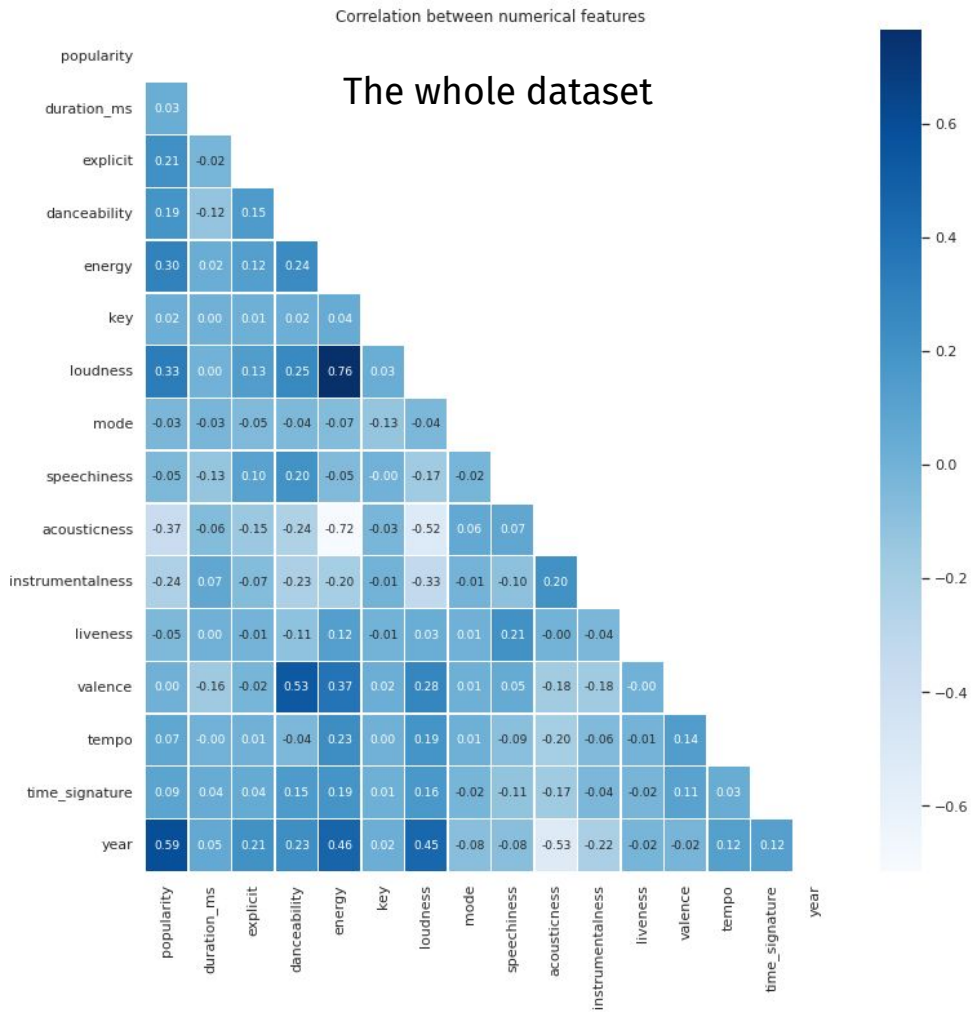
The whole dataset

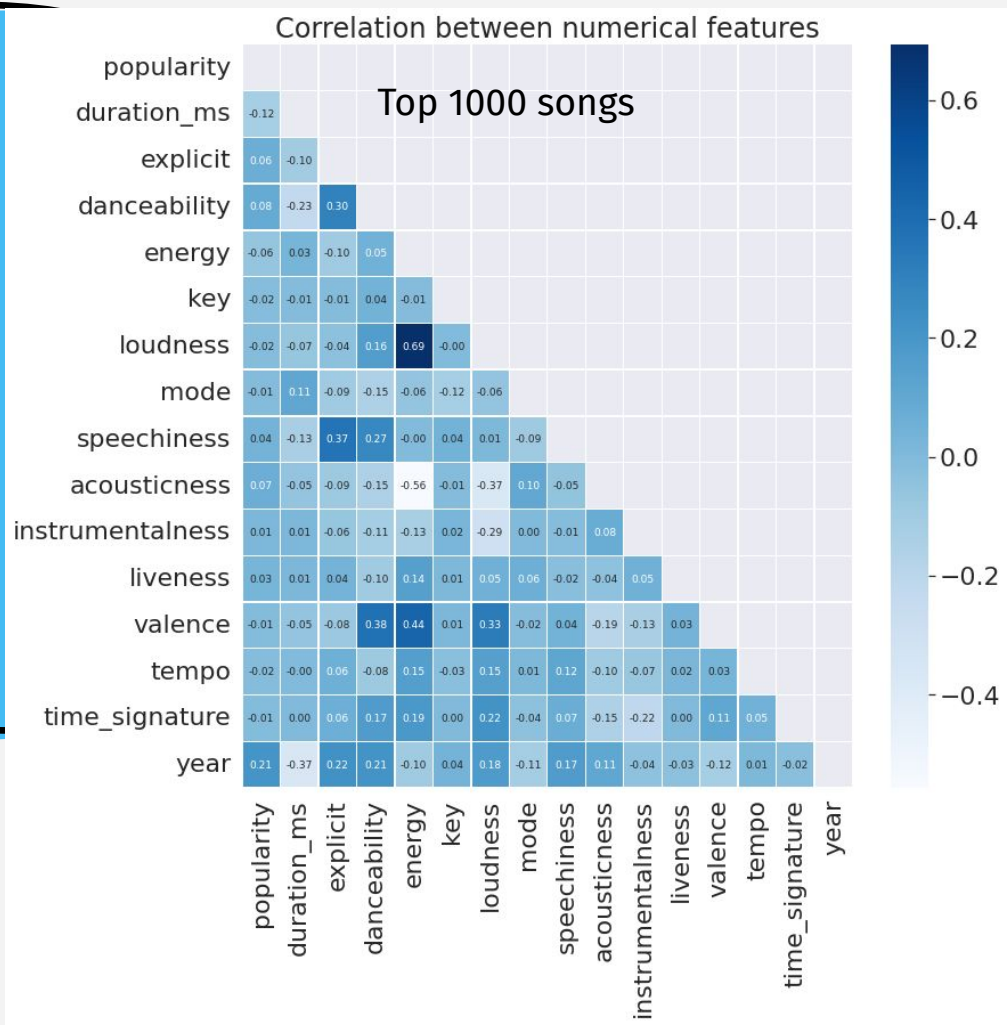
```
corr_matrix2 = top_songs.corr()
corr_matrix2["popularity"].sort_values(ascending=False)
```

```
popularity          1.000000
year                0.207013
danceability        0.084594
acousticness        0.067642
explicit            0.064873
speechiness         0.035651
liveness            0.029827
instrumentalness    0.013224
valence            -0.005754
time_signature     -0.007879
mode               -0.009817
key                -0.015195
tempo              -0.016290
loudness           -0.016726
energy             -0.062493
duration_ms        -0.119533
```

Top 1000 songs

# The whole dataset





# Song Features

- As expected, popularity is highly correlated with the year released. As the Spotify algorithm decides how popular a song is, it generates the "popularity" metric by how recent the streams of the song are, not just the total number of streams the song has received.
- danceability plays as the most important factor in influencing a song's popularity with a 0.084 correlation ratio.
- Acousticness, explicit, and speechiness are also relatively highly correlated with popularity.

# Song Features

- Acousticness is only correlated with popularity and year but not with other audio features.
- Explicit is highly correlated with speechiness, danceability, and year.
- Speechiness is highly correlated with energy, danceability, and year.
- Valence and danceability are highly correlated.

# Song Features

- The correlation of audio feature between top songs and general dataset is quite different. Therefore, we can infer that there are certain features that make a song popular.
- From this data, we can infer that an artist with a high energy and more danceable song has the best chance of gaining the most popularity.
- Acoustic songs also have a good chance to be popular.
- In recent years, the hip-hop genre has become more popular. Hip-hop songs tend to be more speechy and have more explicit lyrics, which may contribute to their popularity. These features may make a song more likely to be successful.

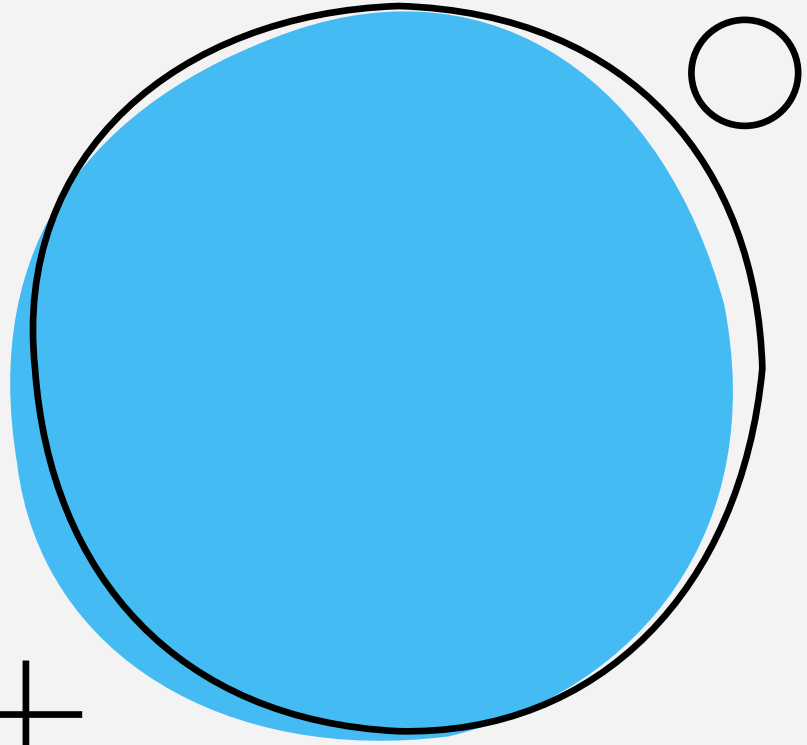
# Predictive Model

# Preprocess data for machine learning

- Split 20% data as testing data
- Created a highly popular column for the songs which the popularity is more than 50. Use this feature as target to predict the song popularity.
- Created a pipeline to normalize data and build machine learning model

# Predictive model

- Logistic Regression
- Random Forest
- XGBoost



# Logistic Regression

```
TEST RESULTS:

Classification Report:
              precision    recall  f1-score   support

           0     0.9417    0.6683    0.7818    102141
           1     0.2444    0.7218    0.3651     15179

    accuracy                         0.6752    117320
   macro avg     0.5930    0.6950    0.5735    117320
weighted avg     0.8515    0.6752    0.7279    117320


ROC AUC Score: 0.6950441701965839
```

Accuracy: 67.52%

| | Attribute | Importance |
|---|---|---|
| 1 | explicit | 1.377579 |
| 5 | loudness | 0.761548 |
| 2 | danceability | 0.548141 |
| 12 | tempo | 0.088612 |
| 13 | time_signature | 0.024828 |
| 6 | mode | 0.018649 |
| 4 | key | 0.010556 |
| 9 | instrumentalness | -0.001266 |
| 0 | duration | -0.024052 |
| 7 | speechiness | -0.076989 |
| 10 | liveness | -0.100813 |
| 3 | energy | -0.266279 |
| 8 | acousticness | -0.603936 |
| 11 | valence | -0.724806 |

# Random Forest

```
TEST RESULTS:

Classification Report:
              precision    recall  f1-score   support

           0     0.8972    0.9668    0.9307    102141
           1     0.5327    0.2547    0.3446     15179

    accuracy                         0.8747    117320
   macro avg     0.7150    0.6107    0.6377    117320
weighted avg     0.8501    0.8747    0.8549    117320


ROC AUC Score: 0.6107473900452489
```

Accuracy: 87.47%

| | Column | Feature importance |
|---|---|---|
| 1 | explicit | 0.388820 |
| 5 | loudness | 0.226532 |
| 8 | acousticness | 0.063170 |
| 11 | valence | 0.048904 |
| 7 | speechiness | 0.044044 |
| 2 | danceability | 0.037344 |
| 0 | duration | 0.036990 |
| 3 | energy | 0.031452 |
| 9 | instrumentalness | 0.031162 |
| 10 | liveness | 0.028648 |
| 6 | mode | 0.019880 |
| 12 | tempo | 0.016101 |
| 13 | time_signature | 0.015658 |
| 4 | key | 0.011294 |

# XGBOOST

```
TEST RESULTS:

Classification Report:
              precision    recall  f1-score   support

           0     0.9473    0.7018    0.8063    102141
           1     0.2687    0.7373    0.3939     15179

    accuracy                         0.7064    117320
   macro avg     0.6080    0.7195    0.6001    117320
weighted avg     0.8595    0.7064    0.7529    117320


ROC AUC Score: 0.7195413768139607
```

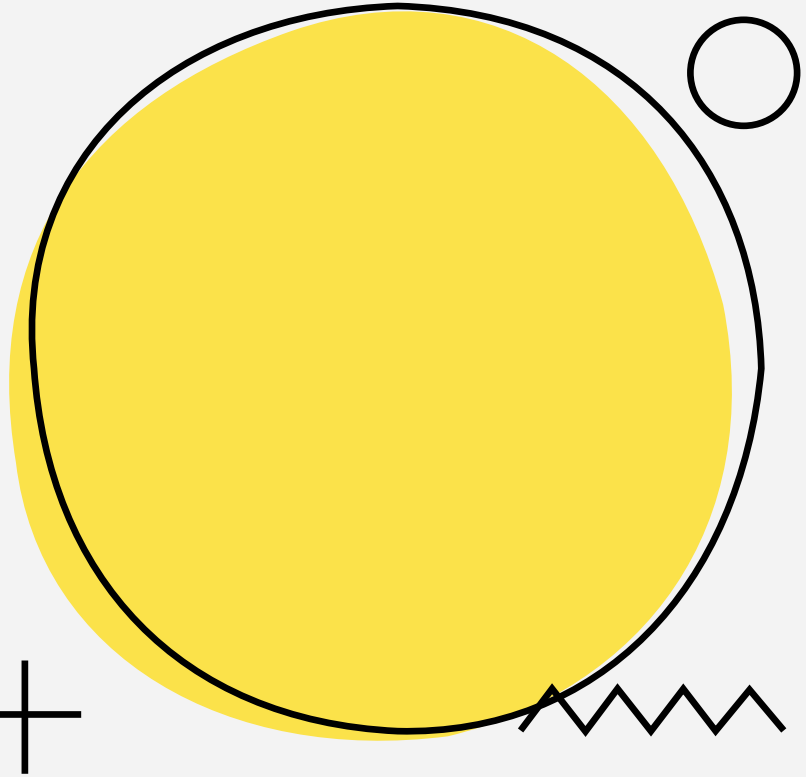
Accuracy: 70.64%

| | Column | Feature importance |
|---|---|---|
| 1 | explicit | 0.388820 |
| 5 | loudness | 0.226532 |
| 8 | acousticness | 0.063170 |
| 11 | valence | 0.048904 |
| 7 | speechiness | 0.044044 |
| 2 | danceability | 0.037344 |
| 0 | duration | 0.036990 |
| 3 | energy | 0.031452 |
| 9 | instrumentalness | 0.031162 |
| 10 | liveness | 0.028648 |
| 6 | mode | 0.019880 |
| 12 | tempo | 0.016101 |
| 13 | time_signature | 0.015658 |
| 4 | key | 0.011294 |

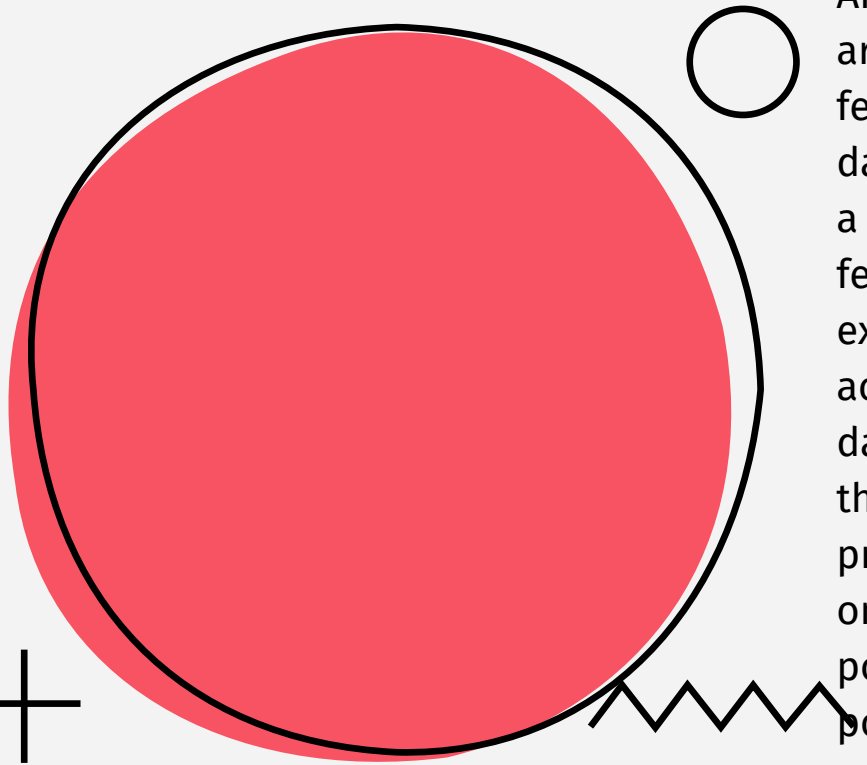
# Conclusion

I used audio features to predict the popularity of songs in a dataset containing approximately 600,000 songs. I defined songs with a popularity score greater than 50 as the target for prediction and used a random forest model, which had the highest accuracy of 87.47%.

Logistic Regression model gave the result of 67.52% accuracy which was the worst result. Random Forest has 87.47% accuracy which was the best result I have. XGBoost has 70.64% accuracy which was a moderate result.



# Conclusion

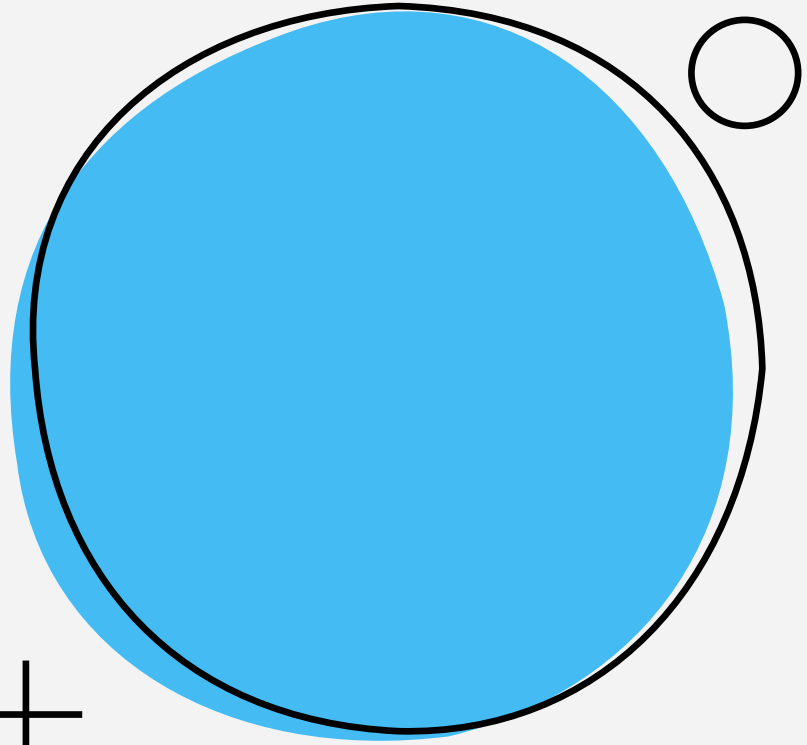


After analyzing the songs of the top three artists, I found that the most common feature among their songs is higher danceability. According to the best model, a random forest, the most important feature for predicting song popularity is explicitness, followed by loudness, acousticness, valence, speechiness, and danceability. While the overall accuracy of the predictive model was 84%, the precision for predicting popular songs was only 53%. This precision rate could potentially be improved with more data on popular songs.

# Conclusion

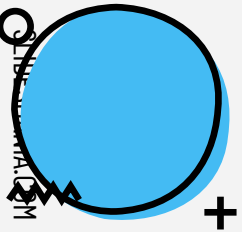
Based on the results, we can see that the audio feature changed from year to year. As time goes by, the popularity will most certainly decrease.

While audio features are certainly important in understanding the characteristics of popular songs, there are many other factors that can influence a song's success, such as the artist's popularity, the marketing efforts behind the song, the lyrics of the song, and so on.



# Further study

1. Incorporating additional data sources
2. Developing more advanced machine learning techniques
3. Examining trends in specific genres or markets
4. Improving the accuracy of predictions
5. Applying the desired song to three predictive models and using a voting system to decide if it is likely to be a hit or not




Thank
you!
SLIDESMANIA.COM