SPOT-POP

Classification and Prediction of Spotify Songs Popularity

Ali Asghari Adib, Melika Shahhosseini

ERDOS Data Science Bootcamp May 2024 Project

OUTLINE

- Overview
- Dataset
- Exploratory Data Analysis (EDA)
    - General EDA
    - Data Cleaning
    - Removing outliers
    - Classification
- Model Selection and Results

OVERVIEW

The main objective of this project is to develop a predictive model that can classify the popularity of Spotify tracks based on their audio features. By analyzing a dataset containing various attributes of Spotify tracks, we aim to identify key factors that contribute to a track's popularity and create a reliable predictive system.

Stackholders:

- Music Producers and Artists: Gain insights into the factors contributing to track popularity.
- Marketing and Promotion Teams: Utilize predictions to target potential hits for marketing campaigns.

DATASET

The dataset consists of more than 100,000 Spotify tracks spanning 125 different genres, with each track described by a range of audio features and metadata.

Feature Name	Description	Feature Name	Description
track_id	The Spotify ID for the track	explicit	Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)
artists	The artists' names who performed the track. If there is more than one artist, they are separated by a ;	danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements
album_name	The album name in which the track appears	energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale
track_name	Name of the track	key	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1
popularity	The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.	loudness	The overall loudness of a track in decibels (dB)
duration_ms	The track length in milliseconds	mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0
speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.	valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
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	tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration
instrumentalness	Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content	time_signature	An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4.
liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.	track_genre	The genre in which the track belongs

PRELIMINARY EDA AND CLEANING

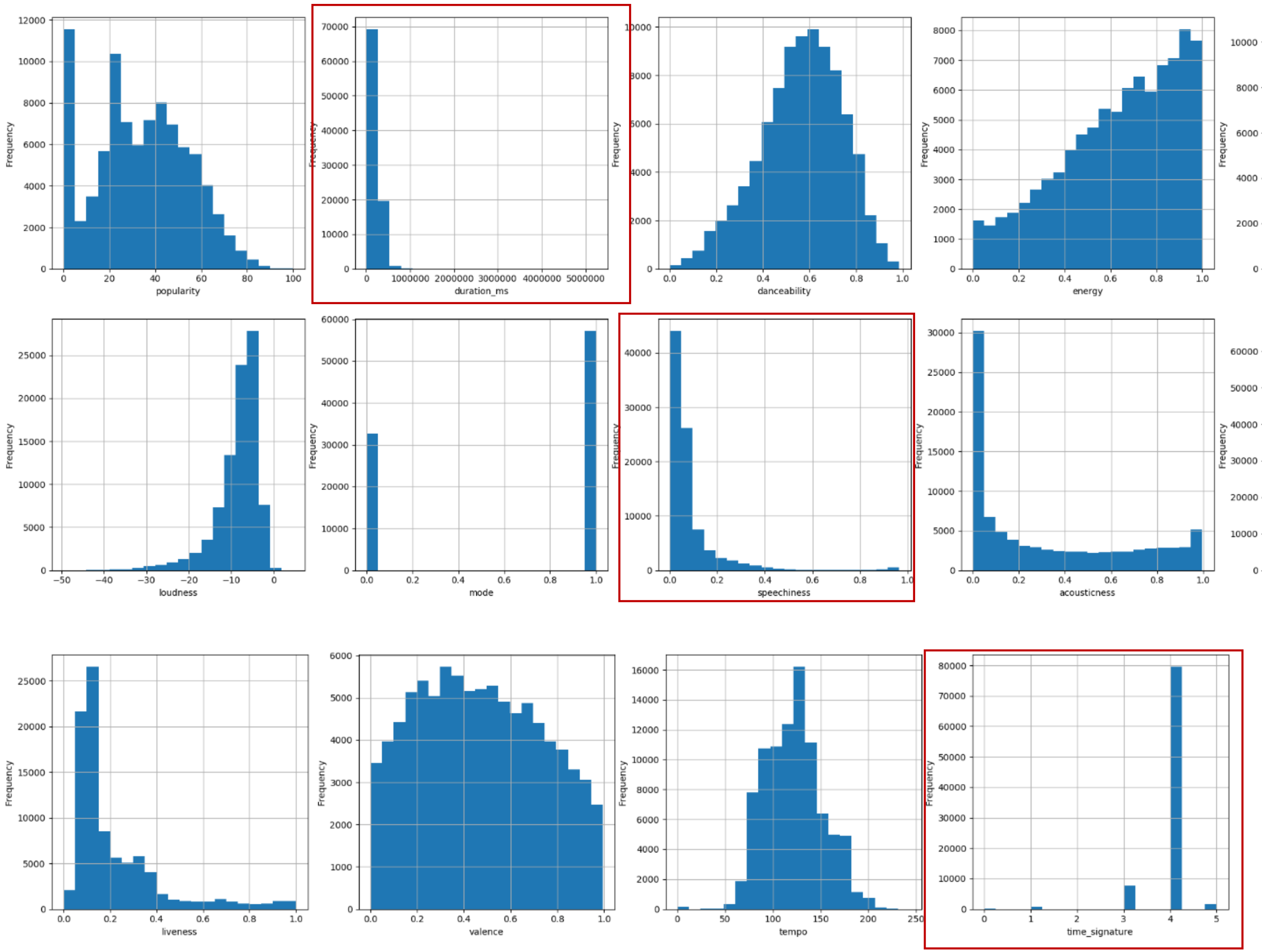
```
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        114000 non-null  int64
 1   track_id          114000 non-null  object
 2   artists           113999 non-null  object
 3   album_name        113999 non-null  object
 4   track_name        113999 non-null  object
 5   popularity        114000 non-null  int64
 6   duration_ms       114000 non-null  int64
 7   explicit          114000 non-null  bool
 8   danceability      114000 non-null  float64
 9   energy            114000 non-null  float64
 10  key               114000 non-null  int64
 11  loudness          114000 non-null  float64
 12  mode              114000 non-null  int64
 13  speechiness       114000 non-null  float64
 14  acousticness      114000 non-null  float64
 15  instrumentalness  114000 non-null  float64
 16  liveness          114000 non-null  float64
 17  valence           114000 non-null  float64
 18  tempo             114000 non-null  float64
 19  time_signature    114000 non-null  int64
 20  track_genre       114000 non-null  object
dtypes: bool(1), float64(9), int64(6), object(5)
```

- **Null Values**: One null value was detected and removed from the dataset.
- **Duplicates**: No duplicated rows were found initially. However, after further EDA, it was learned that there are 24259 duplicated tracks in the data that have been assigned different Genres.

#	track_id	artists	album_name	track_name	track_genre
2002	2K7xn816oNHJZ0aVqdQsha	The Neighbourhood	Hard To Imagine The Neighbourhood Ever Changing	Softcore	alt-rock
3002	2K7xn816oNHJZ0aVqdQsha	The Neighbourhood	Hard To Imagine The Neighbourhood Ever Changing	Softcore	alternative
91105	2K7xn816oNHJZ0aVqdQsha	The Neighbourhood	Hard To Imagine The Neighbourhood Ever Changing	Softcore	rock

- Duplicated tracks were removed from the dataset (keep last). And the new dataset shape was (89741, 21).

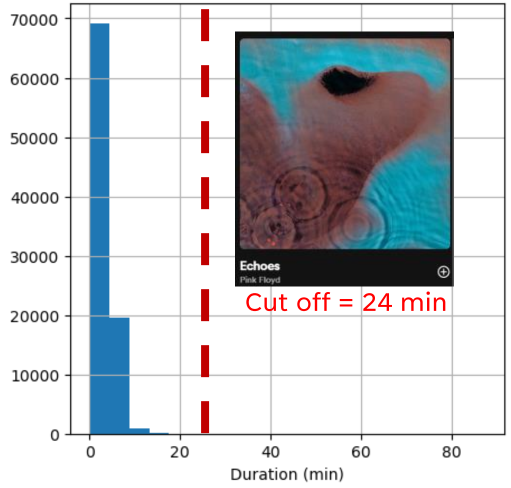
DEEPER EDA AND CLEANING

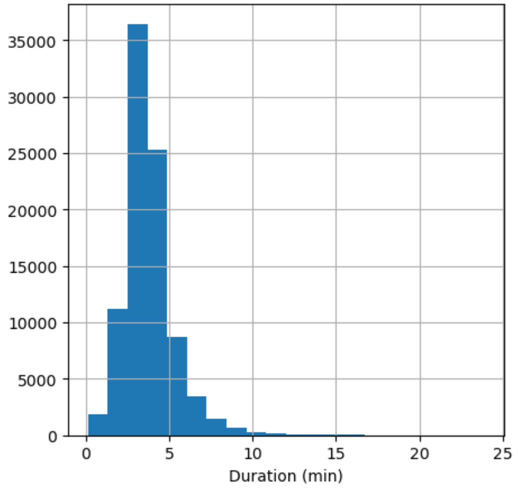


DEEPER EDA AND CLEANING

Duration:

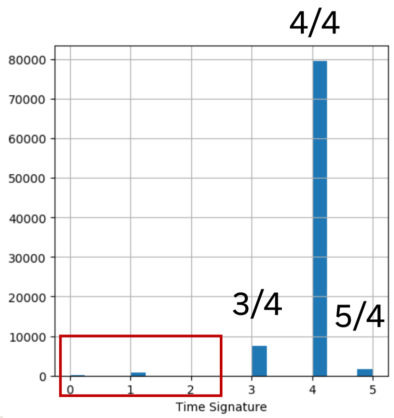
We chose Echoes from pinkfloyd as a representative long song and therefore chose its duration as a cutoff. Entries with duration longer than 24 minutes are dropped from the set.





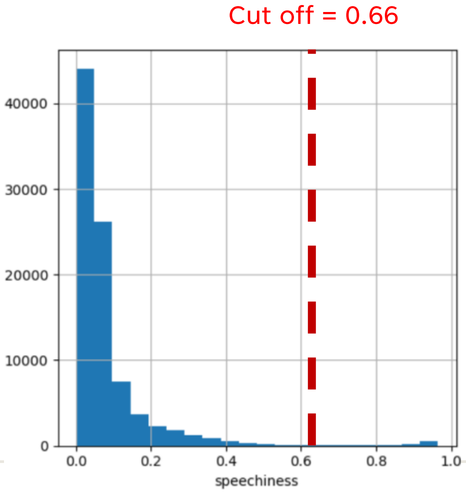
Time Signature:

Based on the Data description, the time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4. Therefore, values of 0, 1 and 2 are meaningless and need to be removed



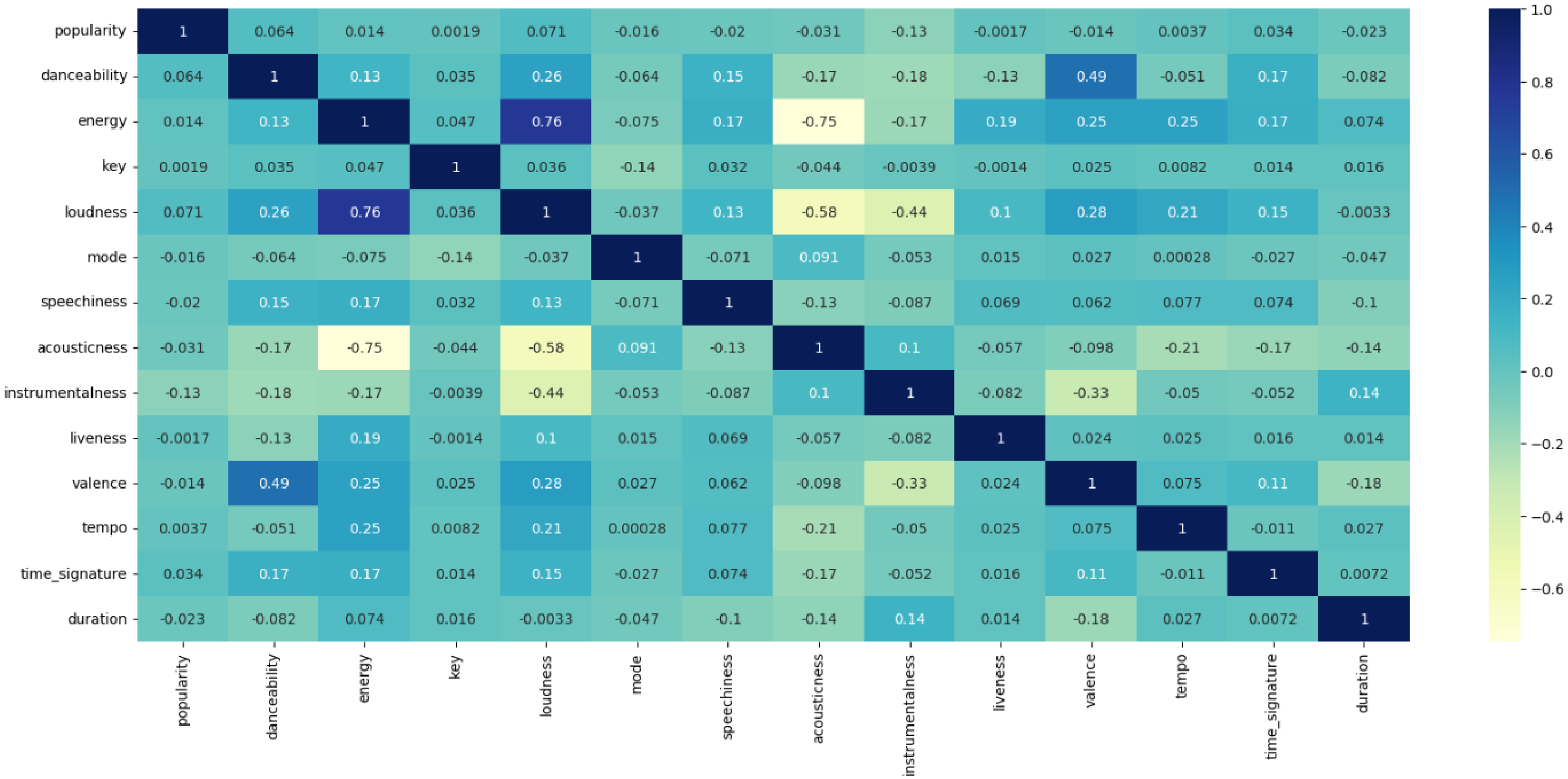
Speechiness:

Based on the Data description, the speechiness greater than 0.66 does not contain any musical instruments and is therefore dropped.



Data shape after these cleanings is (87873, 19).

CORRELATION MATRIX



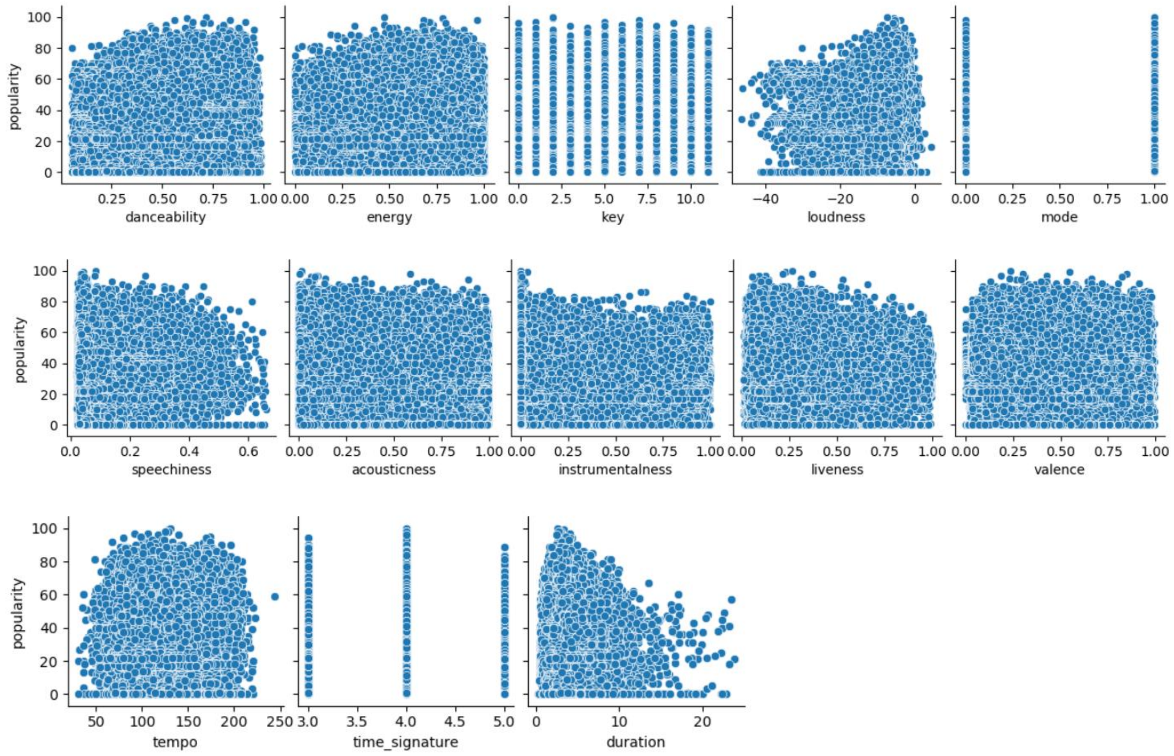
FEATURE ENGINEERING

Artists do not tend to change their genre or name to gain popularity, so out of all the categorical values we only keep 'explicit'ness which seems to slightly have an affect on popularity.

We encode this Boolean entry to a 0 and 1 int type array for modeling purposes.

Since very little correlation was observed between any of the features and popularity and our attempts with linear regression models were not successful, we decided to categorize popularity as follow and predict whether song entries will fall into "Popular" or "Not Popular" categories.

The categorized probability is replaced with the numeric probability in the dataset.



Popularity	Categorized Popularity
<50	Not Popular
>=50	Popular

```
decision
Not Popular    66850
Popular        21023
```

MODEL CHOICE AND DATA PREPARATION

We are going to use and compare three classification models to predict songs popularity based on the features: **Logistic Regression**, **KNN** and **Random Forest**

DATA Preparation:

1. Input to the model is the cleaned dataset with the 'Popularity Decision' column dropped. 'Popularity Decision' is the output of the model.
2. We are using sklearn library and we are splitting our data with the stratify option to maintain a similar value count between our splits in different categories with the test size being 30%.
3. StandardScaler from sklearn reprocessing tools is imported and use to scale the input data.

```
y_train.value_counts(normalize=True)
```

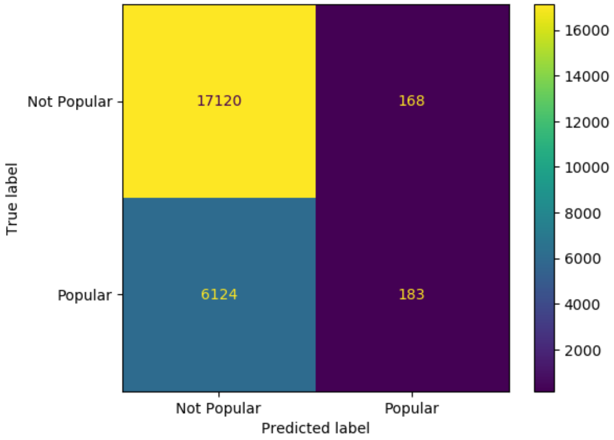
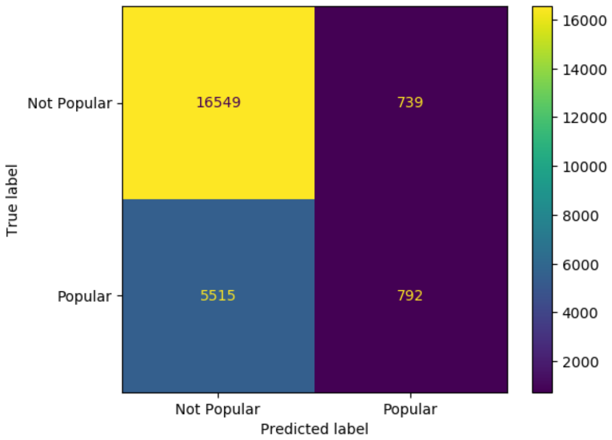
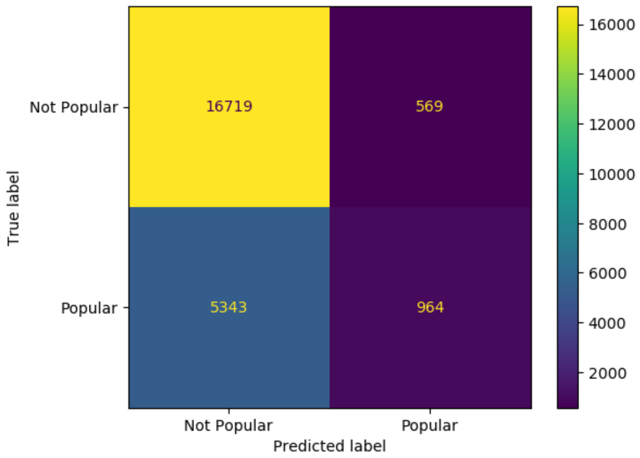
```
decision
Not Popular    0.732699
Popular        0.267301
Name: proportion, dtype: float64
```

```
y_test.value_counts(normalize=True)
```

```
decision
Not Popular    0.732698
Popular        0.267302
Name: proportion, dtype: float64
```

MODEL RESULTS AND CONCLUSIONS

- While the models have performed well on determining “Not Popular” songs, the model is under performing on predicting “Popular” songs, which can be attributed to the smaller portion of popular data.
- Future work includes better feature selection schemes and working with GridSearchCV to find the best fit for each model.

	LogR	KNN (K=30)	RF
Accuracy	73.33%	73.49%	74.94%
Conf Matrix			
Metrics	precision recall f1-score support Not Popular 0.74 0.99 0.84 17288 Popular 0.52 0.03 0.05 6307	precision recall f1-score support Not Popular 0.75 0.96 0.84 17288 Popular 0.52 0.13 0.20 6307	precision recall f1-score support Not Popular 0.76 0.97 0.85 17288 Popular 0.63 0.15 0.25 6307

THANK YOU

Spot-POP Team

asghariadib.1@osu.edu
shahhosseini.2@osu.edu