

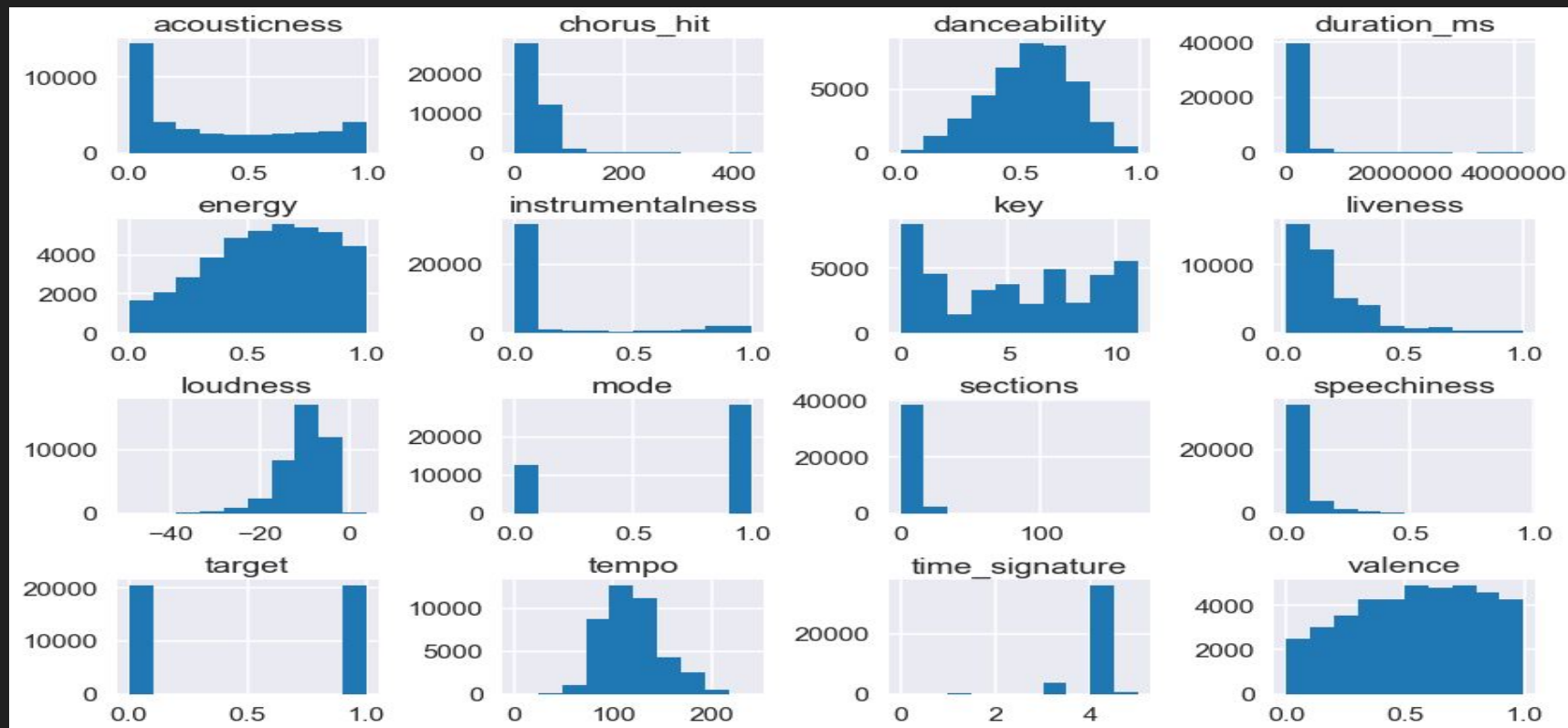
Spotify Hit Prediction and Analysis

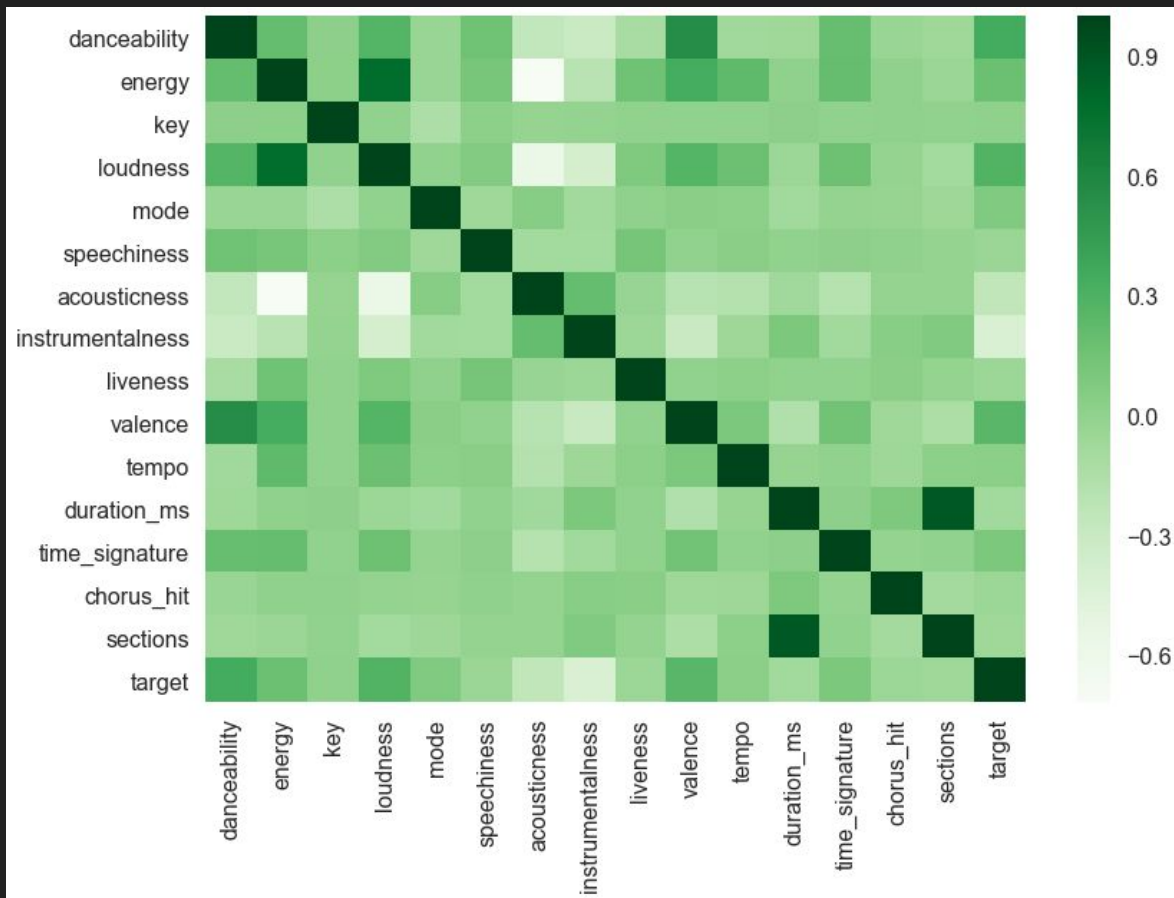
Investigating trends in hit songs and evaluating hit potential

Examining the Data

The dataset from this project is sourced from Kaggle and contains audio features of over 6000 songs from the 1960's through the 2010's. First we will examine the distribution of each feature.

Feature Distribution





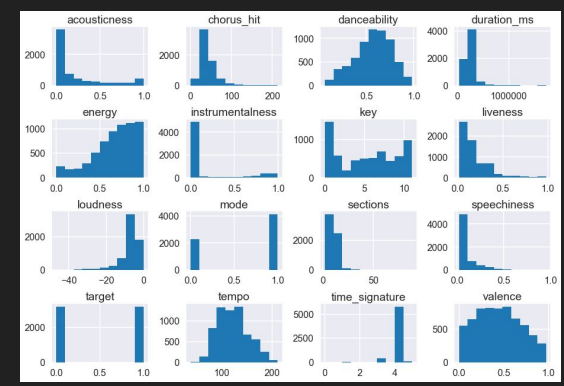
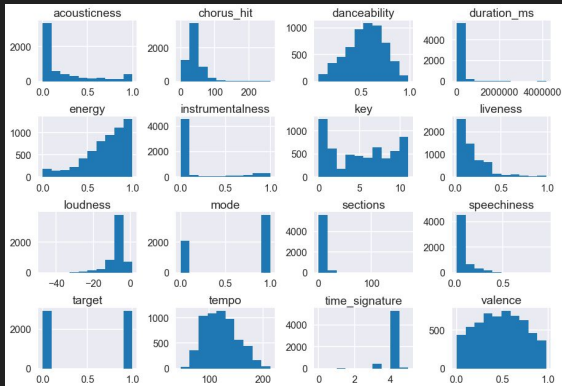
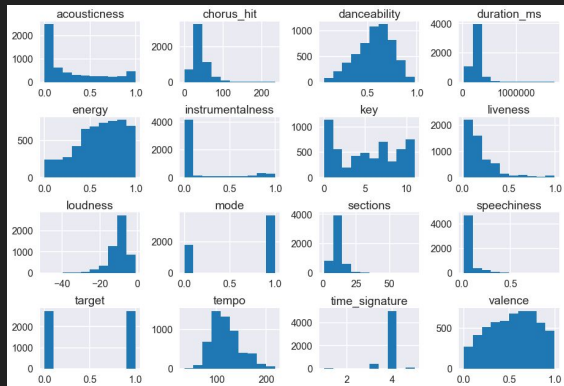
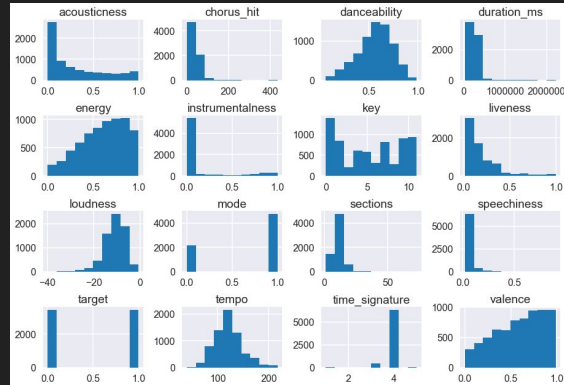
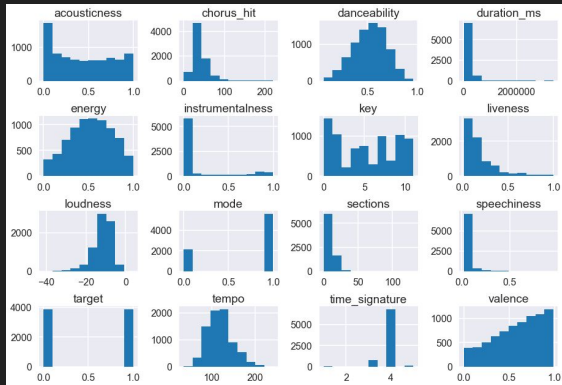
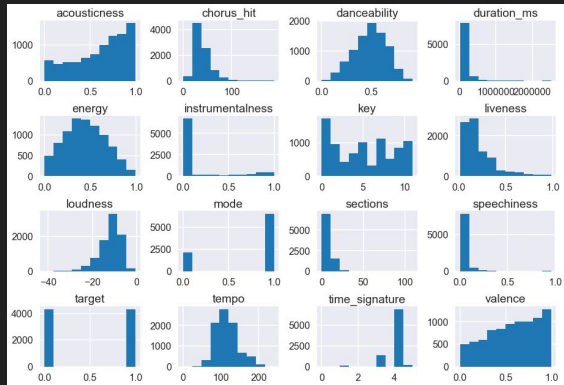
Heatmap analysis

The heatmap shows that the following features are most strongly correlated with the target variable:

- Danceability
- Energy
- Loudness
- Valence

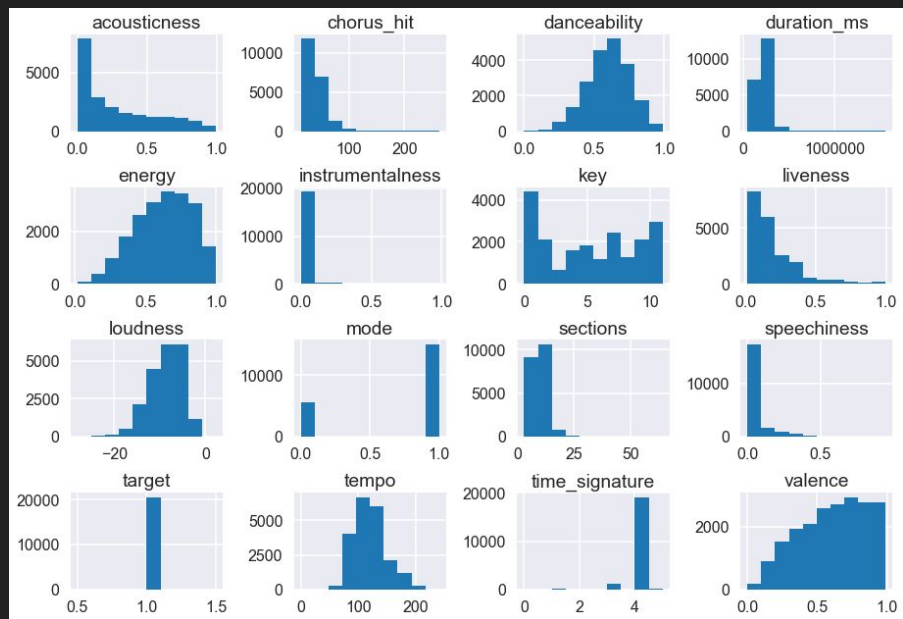
We can also see that Instrumentalness and Acousticness display negative correlation with the target

Analyzing features by decade

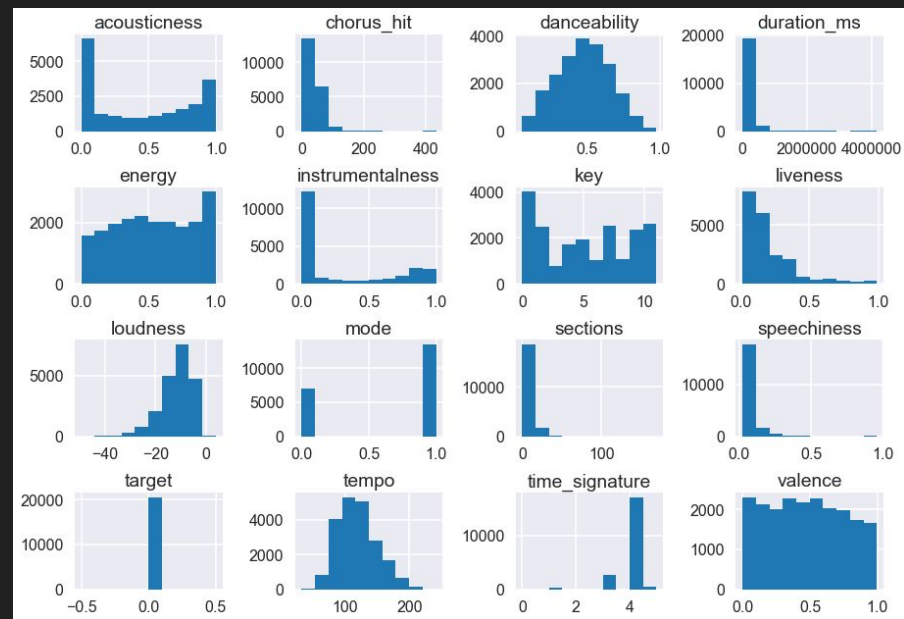


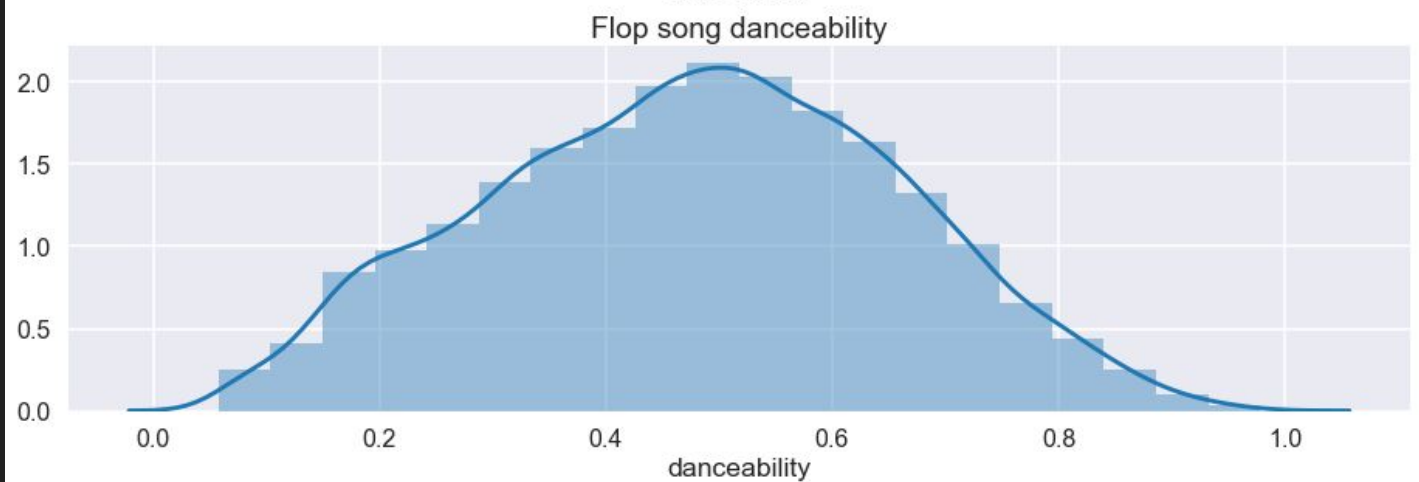
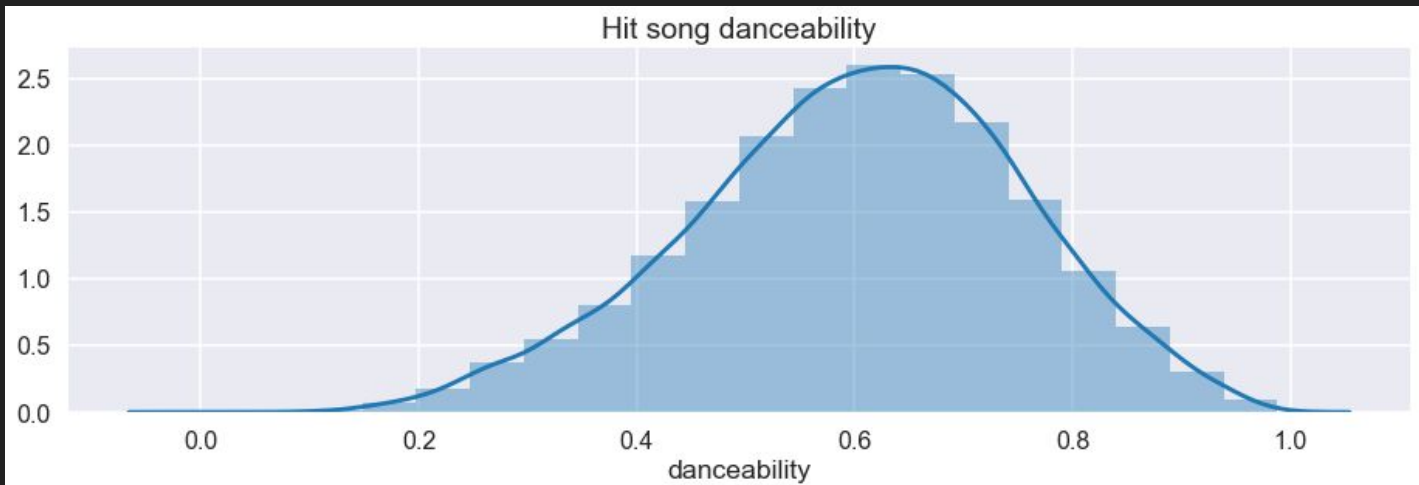
Difference between hit and flop songs

Hit

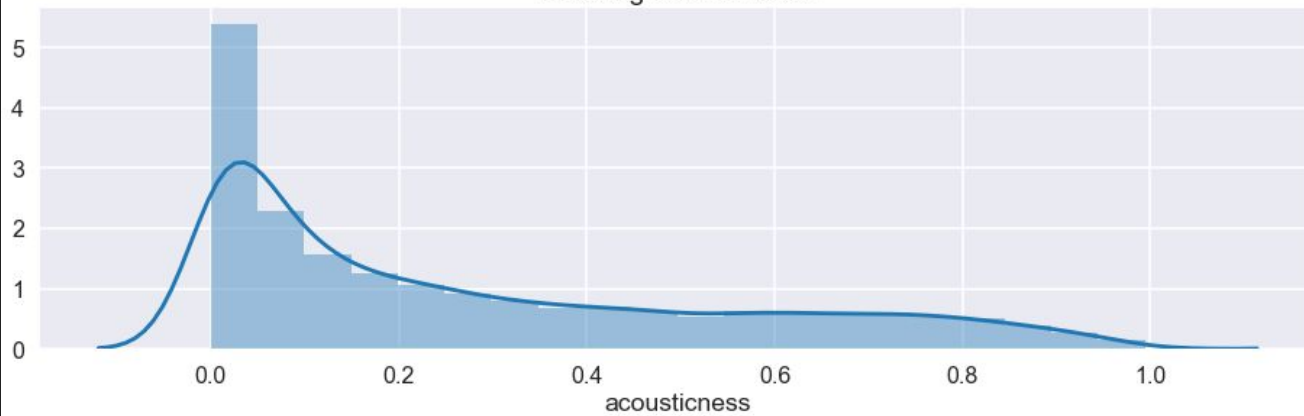


Flop

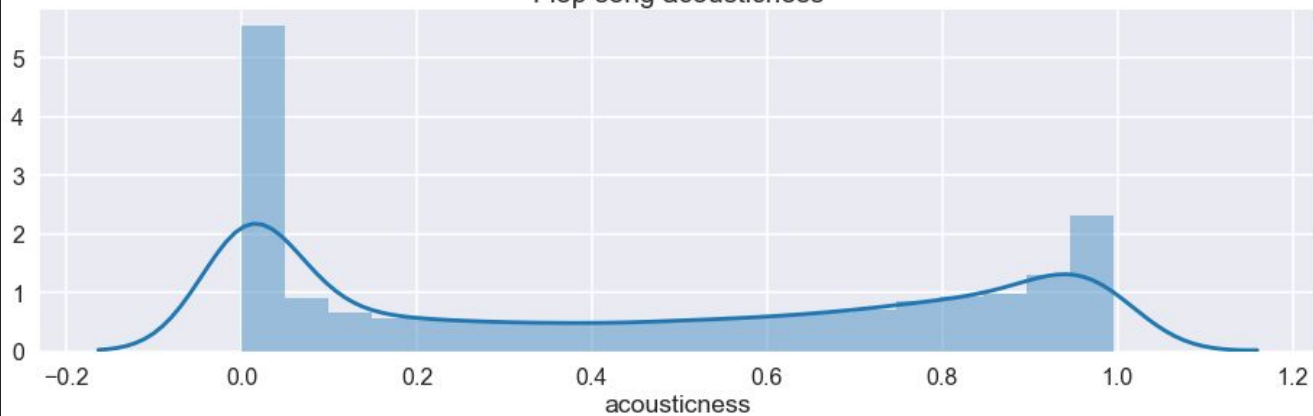


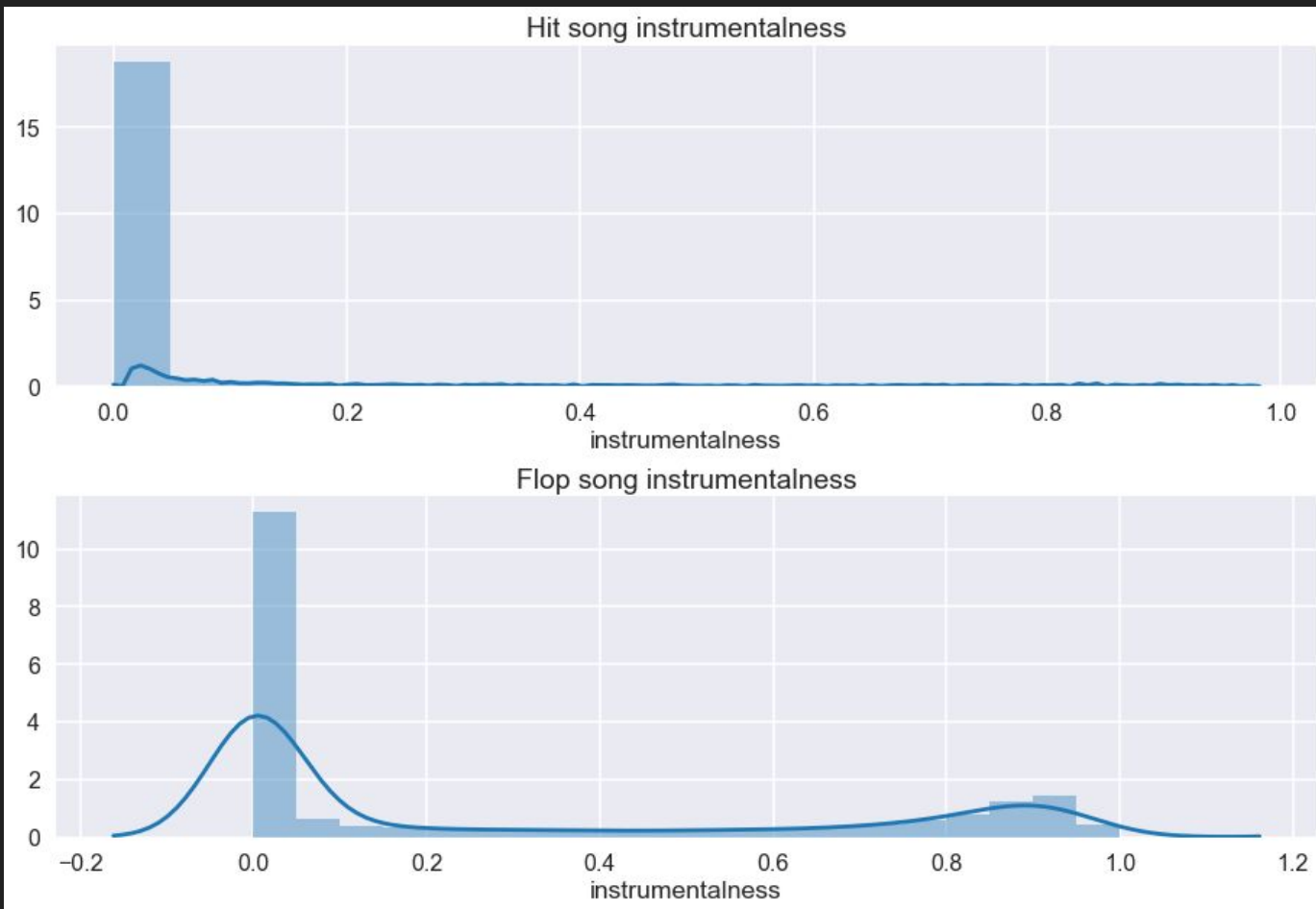


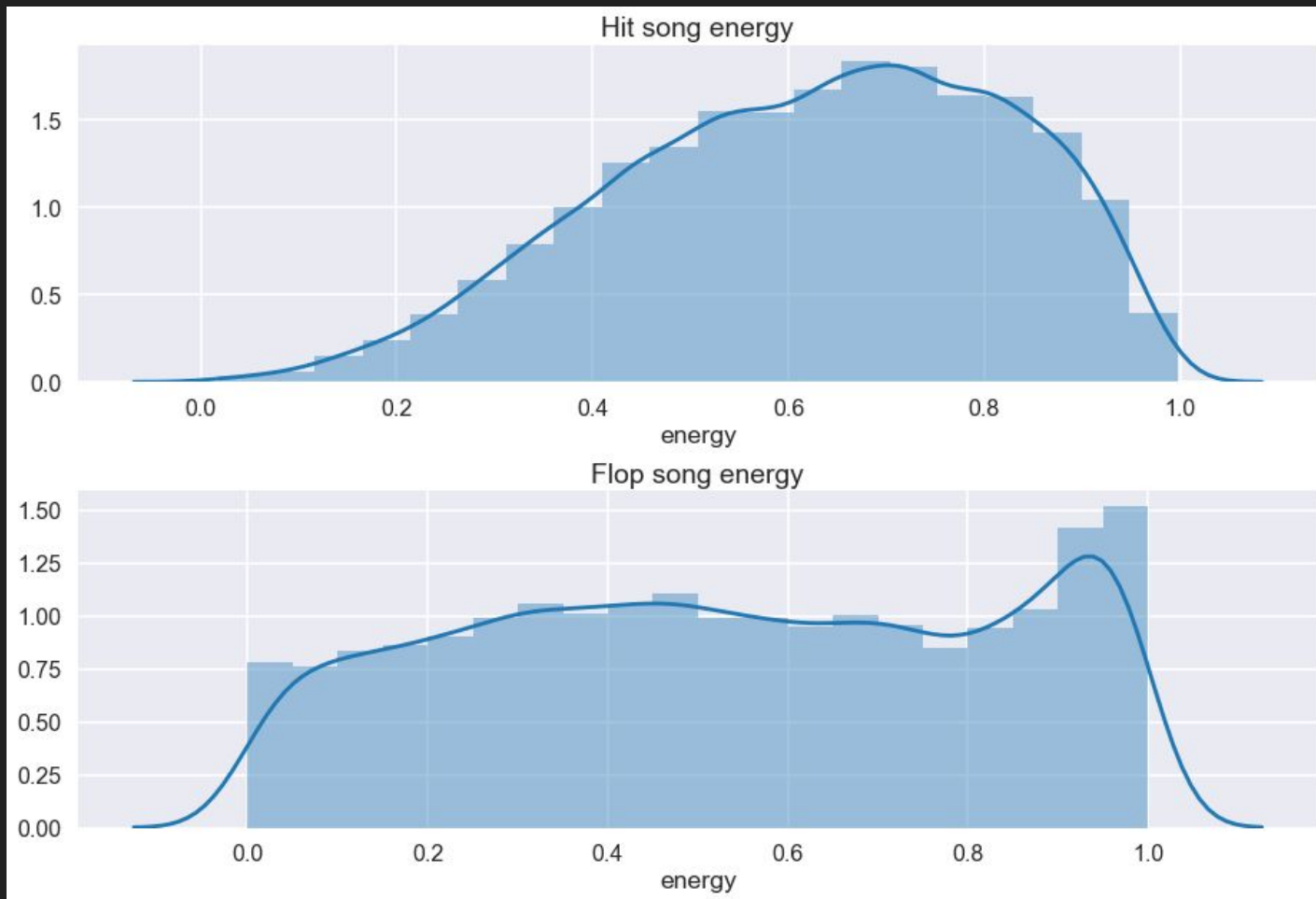
Hit song acousticness



Flop song acousticness







Machine Learning: Random Forest Feature Importances

danceability : 0.1961608604247969
energy : 0.10186251245010265
key : 0.0
loudness : 0.10804753662129706
mode : 0.0012185736458913317
speechiness : 0.02482497234444852
acousticness : 0.1508940885169172
instrumentalness : 0.2650231879735413
liveness : 0.0017610270658901823
valence : 0.06123975608998724
tempo : 0.001843649491325074
duration_ms : 0.04616232259887296
time_signature : 0.007187765634488112
chorus_hit : 0.0013198312574551716
sections : 0.032453915884986216

Random Forest Accuracy

Initial Accuracy: 73%

Tuned Accuracy: 78%

Tuned Parameters:

'bootstrap': True,
'max_depth': 100,
'max_features': 3,
'min_samples_leaf': 3,
'min_samples_split': 8,
'n_estimators': 1000

XGBoost Feature Importances

danceability : 0.10043808
energy : 0.0479939
key : 0.020840691
loudness : 0.04576259
mode : 0.07882381
speechiness : 0.05622998
acousticness : 0.11864295
instrumentalness : 0.25530356
liveness : 0.02467498
valence : 0.043647353
tempo : 0.027719742
duration_ms : 0.061845098
time_signature : 0.05478216
chorus_hit : 0.02297637
sections : 0.040318668

XGBoost Final Parameters

```
'colsample_bytree': 0.5,  
'eta': 0.05,  
'gamma': 0.3,  
'max_depth': 3,  
'min_child_weight': 5
```

XGBoost Accuracy

Accuracy: 78%

F1 Score: 0.79

ROC AUC Score: 0.77

Conclusions

The most important musical features for a song's popularity are:

- Danceability
- Instrumentalness
- Acousticness

Ultimately a song's success is due to many factors outside of the musical features of the song, but this model can predict a song's musical hit potential with a high degree of accuracy.