Popularity Prediction of Spotify Tracks

Can we predict track popularity with relevant attribute information?

Context

Spotify has changed how the music industry uses data.

Streaming services are the top way people listen to music and the number of streaming services is increasing.

For customer retention and attraction, it’s important that people are hearing the tracks they want.

Goal

Can we use our collected data to predict track popularity? We can push predicted popular tracks to playlists and use those tracks in marketing to attract customers, feeding on people’s desire to be up-to-date on pop culture.

Data

Dataset from Kaggle, contained Spotify tracks and attributes from 1921-2020.

There are several attributes Spotify uses to quantify tracks.

- Acousticness
- Danceability
- Energy
- Instrumentalness
- Liveness
- Loudness
- Speechiness
- Tempo
- Valence
- Popularity

Data Wrangling

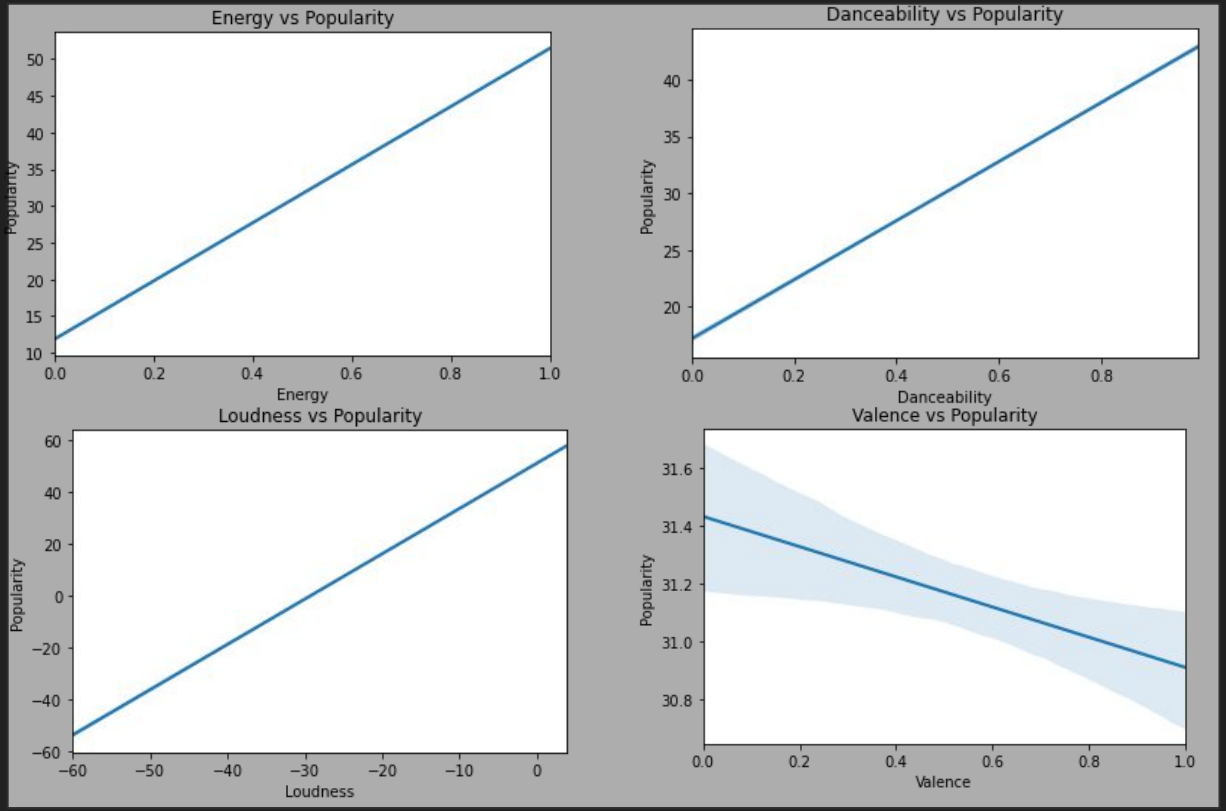
Relatively clean dataset, no missing values

Duplications - Artists sometimes re-release the same track on different albums or as a single

- These were removed using both track name and artist name

Columns such as SpotifyID were dropped

Exploratory Data Analysis

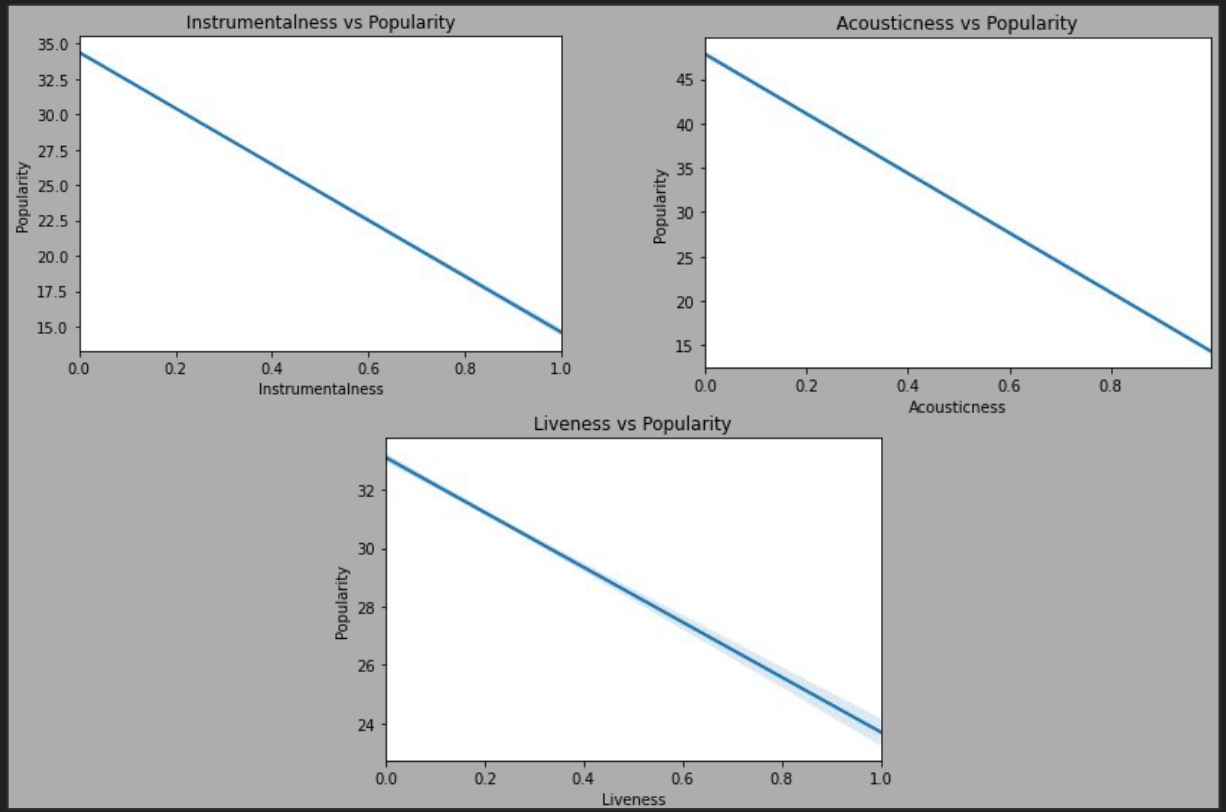


People seem to prefer songs they can dance to, songs that are loud and energetic.

These attributes have positive correlations with popularity.

Surprisingly songs with high valence (a metric Spotify uses to measure happiness in songs) has a negative correlation with popularity.

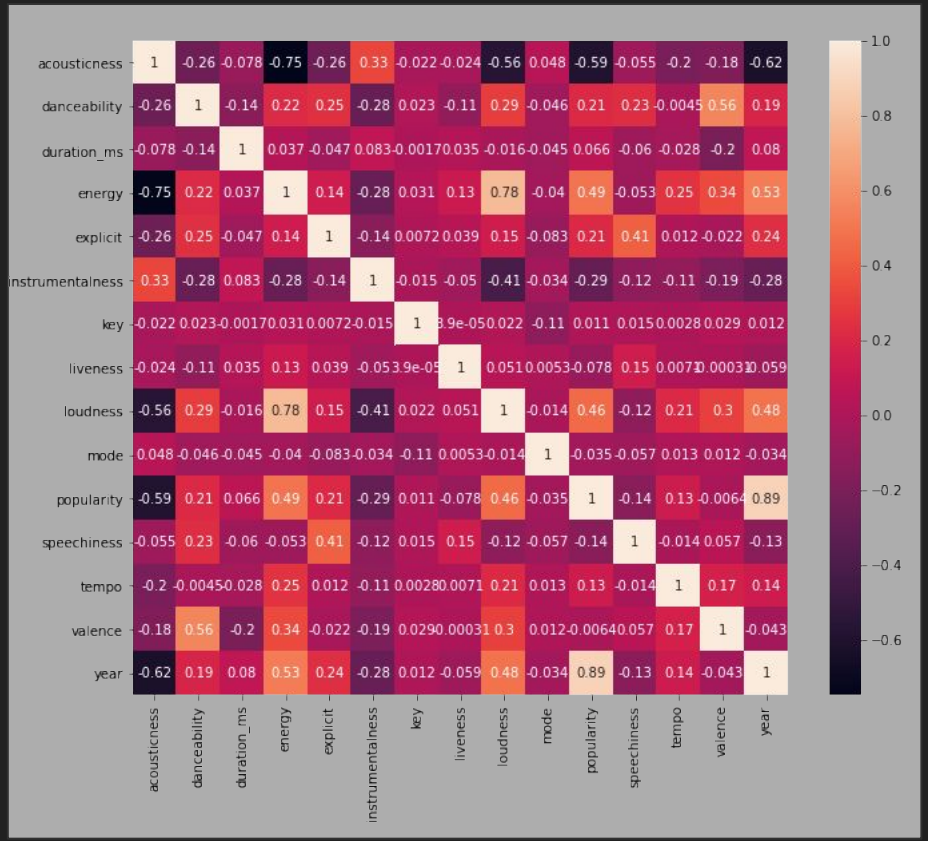
Exploratory Data Analysis



Overall, people do not prefer songs with high levels of instrumentalness or acousticeness.

Based on that knowledge it is not surprising that songs that are more likely to be "live" tracks are also not as popular.

Correlation Matrix



sns.heatmap() was used to view the correlations between all attributes

Results and Conclusions

The attributes can be used to predict popularity with an average 77.5% accuracy.

The model was fit using all the data except the targeted tracks.

Release date of the track did not have an effect on the average accuracy of the prediction.

Popularity, as defined by the spotify metric, is relatively predictable based on the attributes.

Linear Regression had an average 77% accuracy and was much faster at fitting and predicting.

Model Comparison

Compared Linear Regression, Decision Tree Regressor, and Random Forest Regression models and then optimized each using GridSearchCV

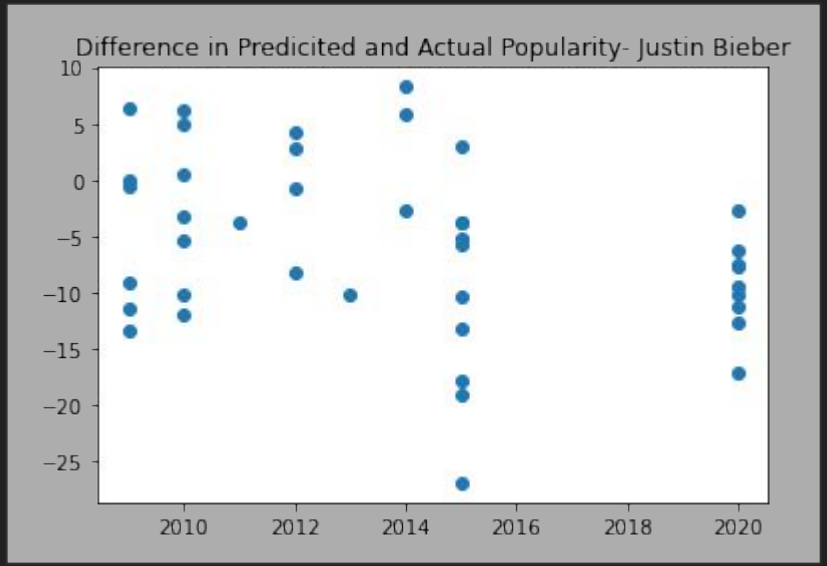
train_test_split() was used to split the data for training

Test_size = 0.3, random state = 42

Model	Best Parameters	CV Score
Linear Regression	StandardScaler: None	0.794
Decision Tree Regressor		0.650
Random Forest Regressor	N_estimators: 195 StandardScaler: Standard Scaler	0.833

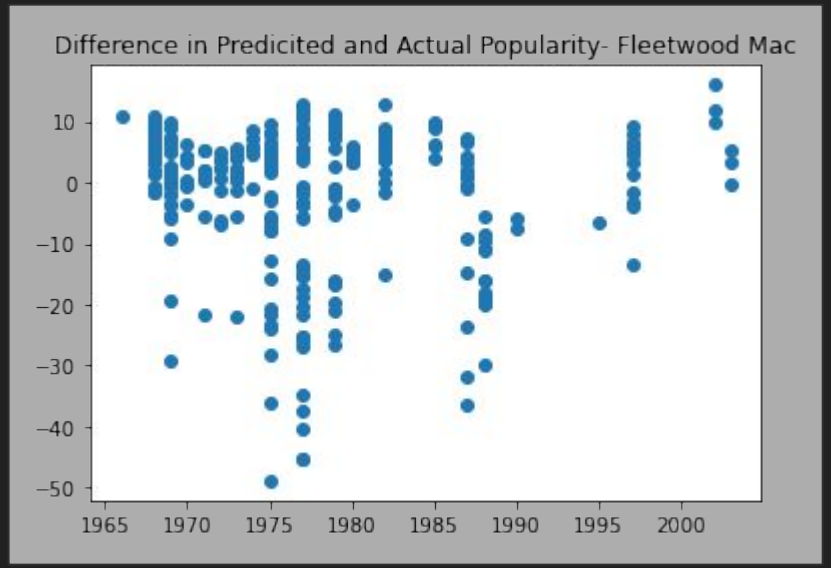
Comparison of Model Prediction Popularity vs Actual Popularity of Tracks by Justin Bieber

The model was used to predict popularity of tracks by Justin Bieber. The difference in predicted popularity vs reported popularity was plotted by release year.



Comparison of Model Prediction Popularity vs Actual Popularity of Tracks by Fleetwood Mac

The model was used to predict popularity of tracks by Fleetwood Mac. The difference in predicted popularity vs reported popularity was plotted by release year.



Conclusions

With 77.5% accuracy on population predictions we can begin to target new tracks with the model and push those tracks to new playlists for listener discovery playlists as well as target tracks for marketing uses.

Spotify concerts can also be better targeted, looking at artists that have higher predicted popularity based on their tracks.

Further Analysis Possible

Can we use different popularity metric (one that doesn’t favor recent tracks) to predict the next trends in music?

Podcasts have boomed in popularity:

How does the prediction work if we separate music tracks from podcasts?

Can we predict popularity of podcasts from current attributes? What additional/replacement attributes would be helpful?