**Billboard Hot 100 Hit Prediction**

 Predicting Billboard's Year-End Hot 100 Songs using audio features from Spotify data.

**Overview**

Each year, Billboard publishes its Year-End Hot 100 songs list, which denotes the top 100 songs of that year. The objective of this project was to see whether or not a machine learning classifier could predict whether a song would become a hit *(known as*[*Hit Song Science*](https://en.wikipedia.org/wiki/Hit_Song_Science)*)* given its intrinsic audio features.

The goal of this project is to see if a song's audio characteristics can determine a song's popularity. Data and analytics aside, music listeners around the world probably have seen music trends change over time. Although each listener has custom interests in music, it is pretty clear when we listen to a hit song or soon to be hit song (consider Old Town Road). And over time, we see the characteristics of hit songs change. So, rather than using our intuition or "gut-feeling" to predict hit songs, the purpose of the project is to see if we can use intrinsic music data to identify hits.

Hit Song Science can help music producers and artists know their audience better and produce songs that their fans would love to hear.

Additionally, audio engineers can work with musicians to tweak intrinsic music qualities to make a song more popular catchy and likable.

Also, it can highlight unknown artists whose music is characteristic of top songs on the Billboard Hot 100.

**Data and Features**

A sample of 15000 Spotify and MSD songs was given , which included songs from various Spotify and MSD albums

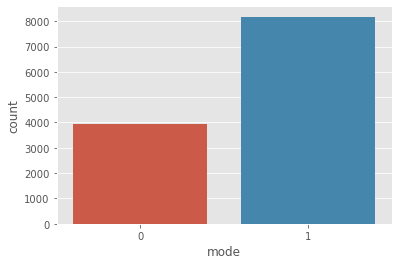
Using Spotify's Audio Features , the following [features](https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/) were collected for each song:

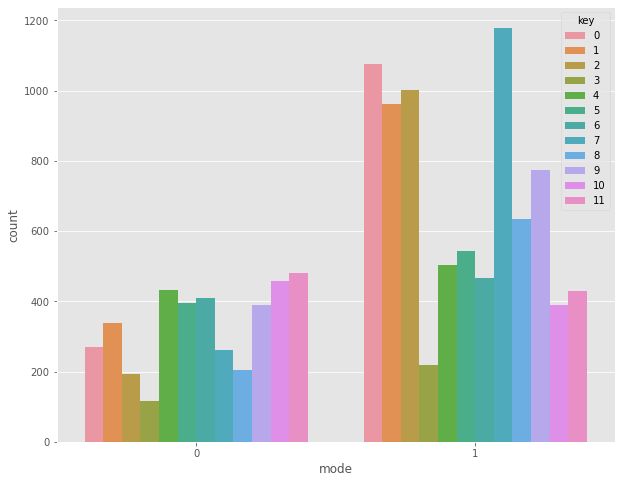
* **Mood**: Danceability, Valence, Energy, Tempo
* **Properties**: Loudness, Speechiness, Instrumentalness
* **Context**: Liveness, Acousticness

Spotify assigns each song a value between 0 and 1 for these features.

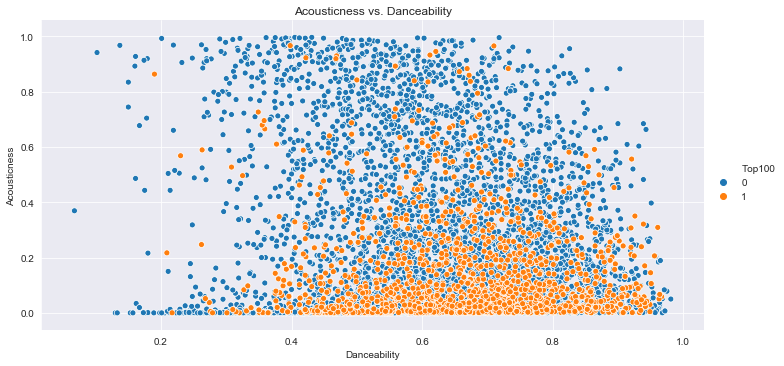
After cleaning the data, a dataset of approx. 12000 songs was created.

Distribution of Billboard songs





**Exploratory Data Analysis**

**Feature Comparisons** [](https://github.com/siddgood/billboard-hit-prediction/blob/master/images/acoustic-vs-dance.png) [](https://github.com/siddgood/billboard-hit-prediction/blob/master/images/acoustic-vs-loud.png)

The above graphs show the separability in the data when compared across two unique Spotify features; this suggests that data may separate across an n-dimensional feature space. Given this, the problem can alternatively be posed as an unsupervised learning problem where clustering methods can classify the data.

**Models and Results**

Given the unbalanced nature of the dataset, any model chosen would automatically yield high accuracy. So, in addition to aiming for high accuracy, another objective of modeling is to ensure a high AUC (so that TPR is maximized and FPR is minimized).

Here's a list of the models I tested:

1. Logistic Regression
2. Gaussian Discriminant Analysis (GDA)

**Model Summaries:**

| **Model** | **Accuracy** |
| --- | --- |
| Logistic Regression | 0.67 |
| GDA | 0.53 |

**Conclusion**

The best model after testing seems to logistic regression.

So we use logistic model for deployment.