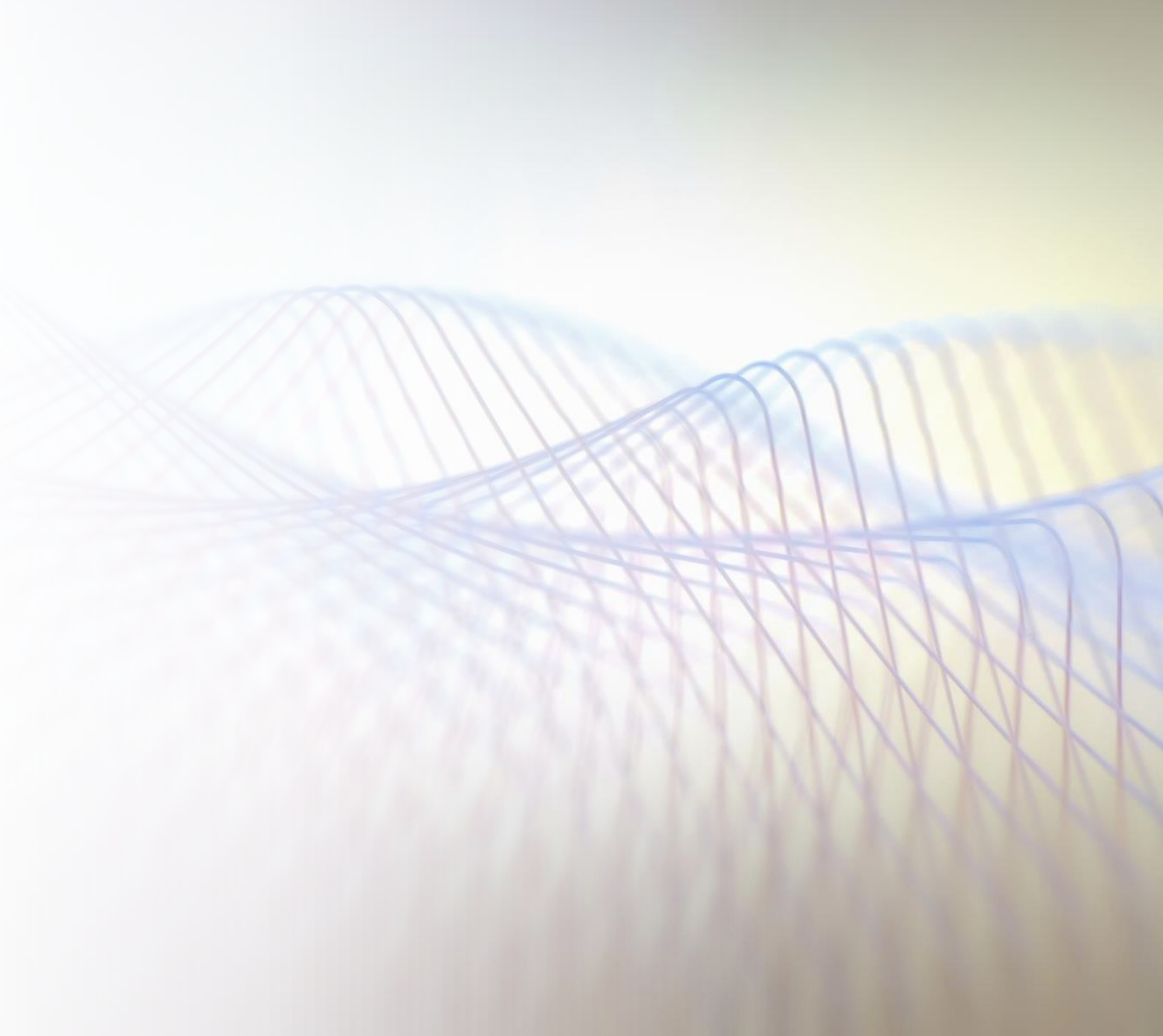
MUSIC DATA ANALYSIS



BY

HARSHAVARDHAN NILAKANTAN

The Problem

- Music industry is tough to break through.
- Too many different forms and genres of music.
- Highly preferential.

Solution:

- Would be useful to know what kind of music would be more lucrative to produce.
- Audio characteristics of more “popular” music.

The Client & How Data Science Can Help

The Clients:

- Music Producers
- Musicians
- Music Students

How Data Science Can Help:

- Breakdown the audio in terms of quantifiable characteristics
- Characterize “mainstream” music
- Produce recommendations on audio characteristics that might have a better chance at becoming lucrative

The Data & Preprocessing (Wrangling)

- Spotify data was obtained from Kaggle, with quantified audio features.
- Supplemental data required for analysis for downloaded and appended to the data using the SpotiPy library to connect to Spotify web API endpoints.
- Supplemental data was appended to existing data.
- Duplicates and invalid data were removed
- Removed unnecessary columns
- Key data metrics – Danceability, Instrumentalness, Popularity, Time Signatures, Durations, Acousticness, Loudness, Valence, Tempo, Genre

Exploratory Data Analysis (EDA)

Initial analysis found that:

- High Danceability, High Instrumentalness (Fig. 1), Durations and Non-standard time signatures (Fig. 2) had an effect on the Popularity measure.
- Tempo, Loudness, Acousticness did not directly have an effect on the Popularity metric.

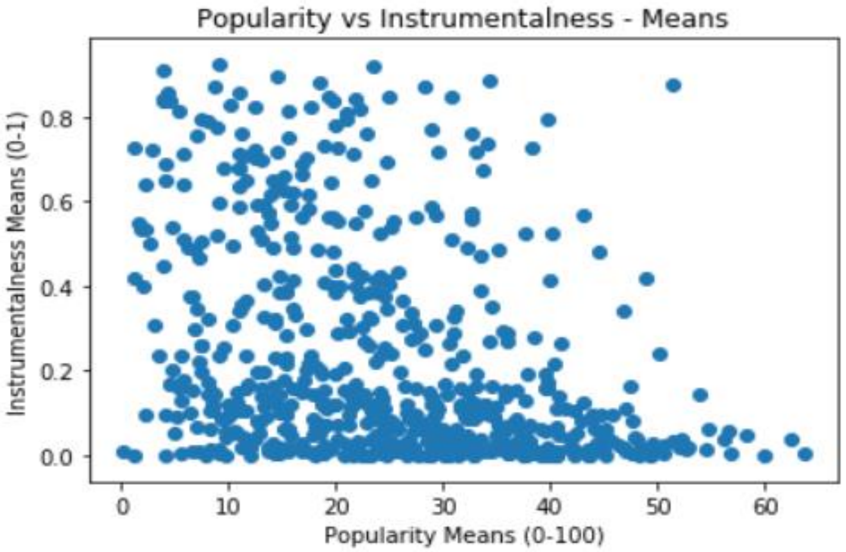
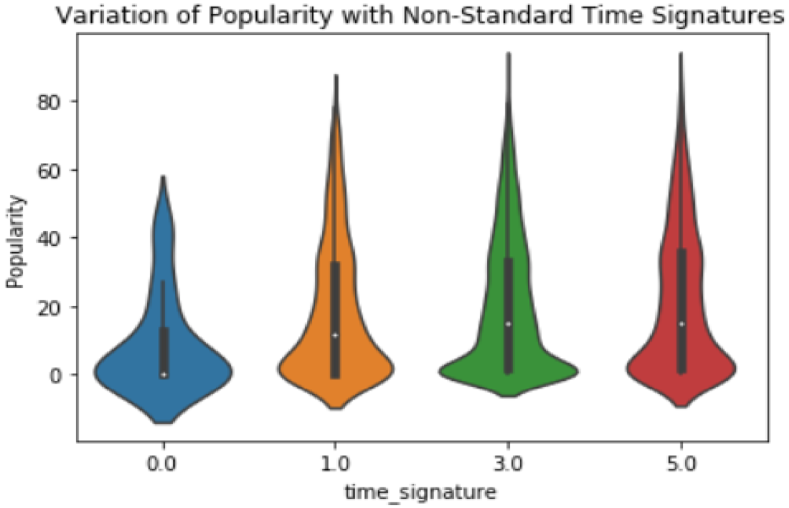


Fig. 1 (left): Means of Popularity vs Means of Instrumentalness

Fig. 2 (right): Non-standard time signatures vs Popularity



EDA Continued..

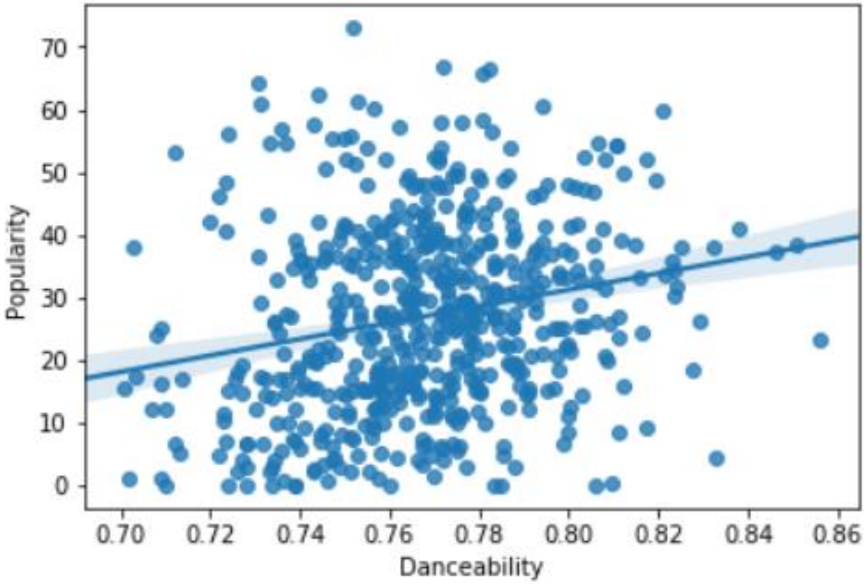
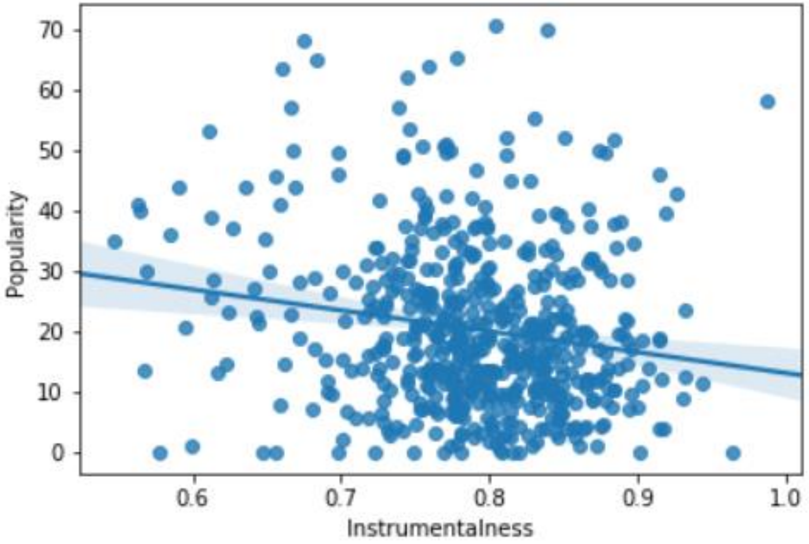


Fig. 3 (left): Danceability vs Popularity

Further analysis found:

- Danceability was positively correlated with popularity (Fig. 3)
- Instrumentalness was strongly anticorrelated with popularity (Fig. 4)
- The 3.0 and 5.0 non-std time signatures were similarly distributed, while 1.0 and 5.0 were differently distributed.

Fig. 4 (right): Instrumentalness vs Popularity



Machine Learning Model

- Chosen Model – Unsupervised Learning Model – K-Means clustering
- Unsupervised because this will allow us to examine how the data clusters naturally, and **find** any underlying structures.
- Found Optimal K with K=10 using "Elbow" Method (Fig. 5).
- Beyond 10, variability reduces overall, inflection point.
- Followed by PCA into 2 components, for visualization

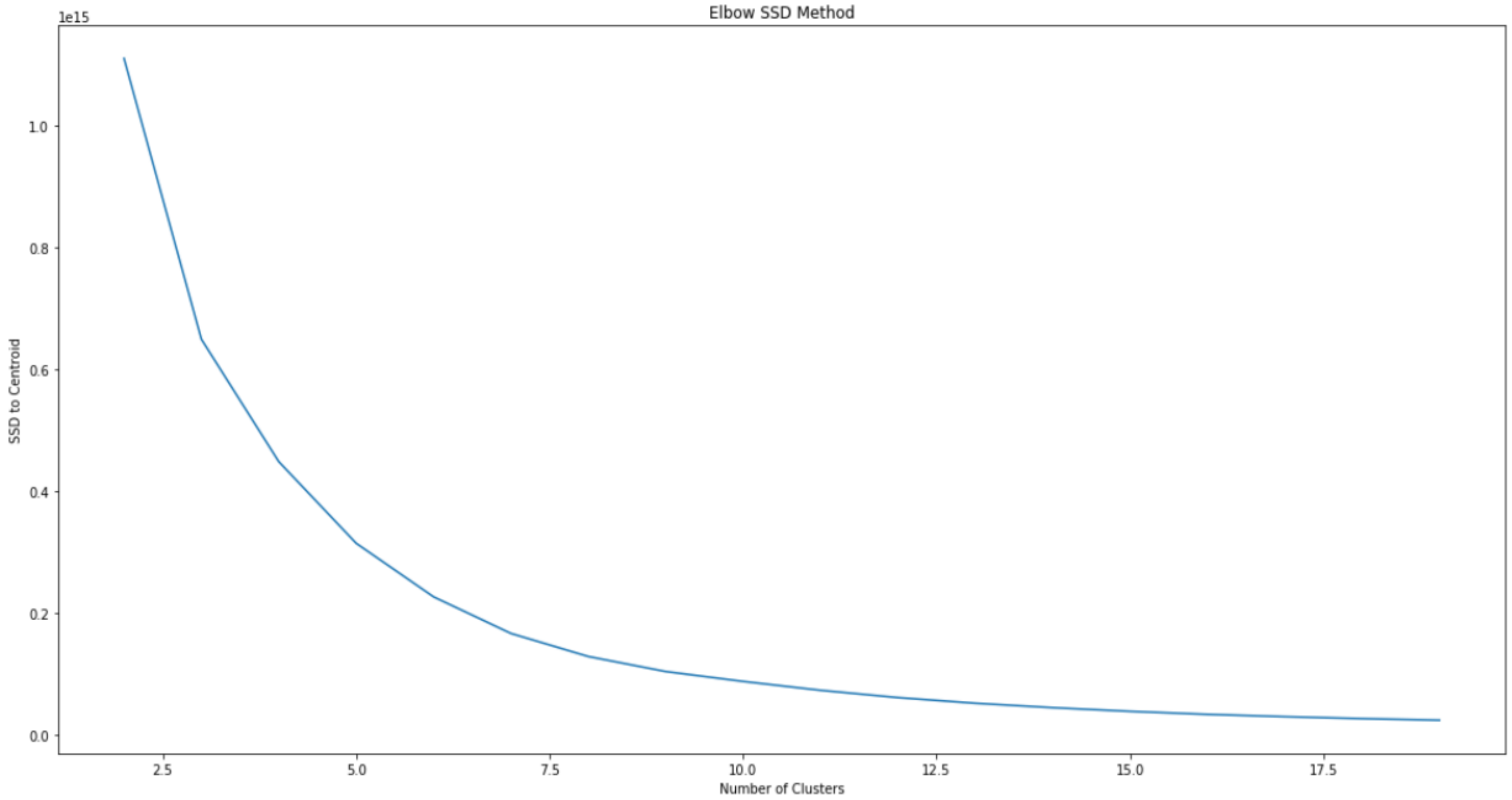


Fig. 5 (above): No. of clusters vs SSD to Centroid, to find the "elbow" or inflection point.

Machine Learning Model

- Clusters naturally separate into bands.
- Taller, thinner bands on left (group B), higher popularity.
- Shorter, wider bands on right (group A), lower popularity.
- Bands to left of blue line – clusters belonging to group B – higher popularity
- Bands to right of blue line – clusters belonging to group A – lower popularity

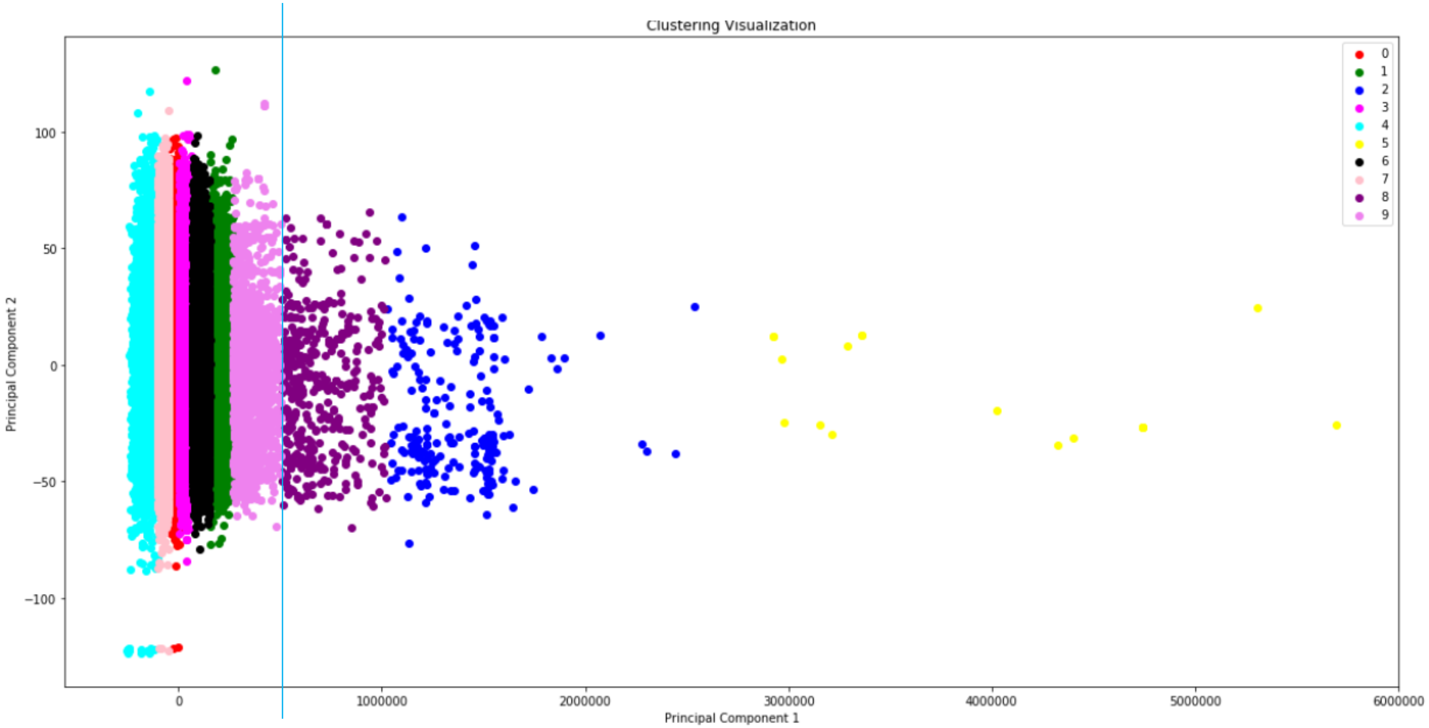


Fig. 6 (above): K=10; Clusters split into bands; Popularity increases from left to right, along x-axis.

Findings

- Some audio features showed cluster separation over smaller differences (subtle), some over bigger differences (stronger), both very important.
- Important audio features for clustering are:
  - Subtle: Danceability, More, Speechness, Liveness, Popularity
  - Stronger: Energy, Loudness, Acousticness, Instrumentalness, Tempo, Duration
- Cluster separation and, in turn, model performance would have been better with smaller K.
- Genres did not separate well into clusters. However, group A (lower pop) had only 276 genres, while group B had all 625 genres.

Recommendations

- Avoid entering into production/composition with the more fickle genres (present in low popularity super group) as they tend to collapse, and either lose popularity quickly or only retain popularity in smaller pockets. This would not be very lucrative.
- Even within the fickle genres Mine the specific characteristics of the higher popularity clusters and ensure that the audio features of your tunes fall within a range of those audio features, as various combinations of the same show higher popularity.
- Generally, as a trend seen in our data, keep your music more danceable, less instrumental.
- Stick to standard time signatures, as popularity has been shown to go up with those. This makes sense as the untrained ear may find it difficult to follow odd time signatures or something more complex.

Conclusions & Future Work

Conclusions:

- Popularity played in important role in further separation of clusters into groups A and B.
- Genre has limited bearing on clustering, when K is high.
- Further grouping (more clustering) showed better and more characterizable separations

Future Work:

- Reduce K – fewer clusters – better separation
- Remove some audio features that did not weigh into the clustering too much
- Genre classification was way too granular. Collapse subgenres into parent genres. Re-cluster.
- Other clustering algorithms