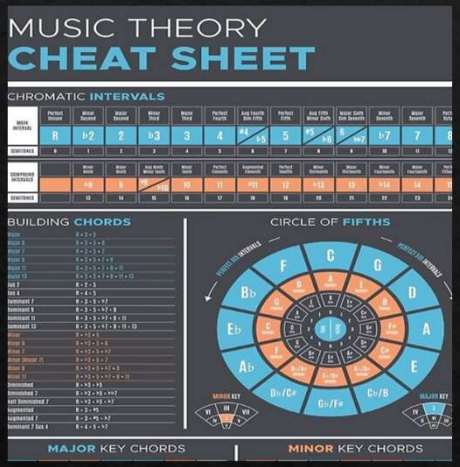
Clustering Songs by Attributes (to find additional song value)

***Supervised and Unsupervised Learning Algorithms for Genre Classification and Song Popularity**

The Project – Subject Matter Expertise

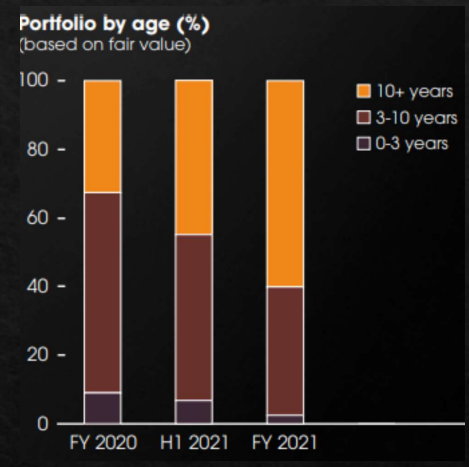


Basic Music Theory

Understanding songs attributes

Investment Funds

Recent emergence of songs as asset class – Hipgnosis Songs Fund and Roundhill Music Fund



Music Copyright

Music Royalties – maximizing value of Songs for investors

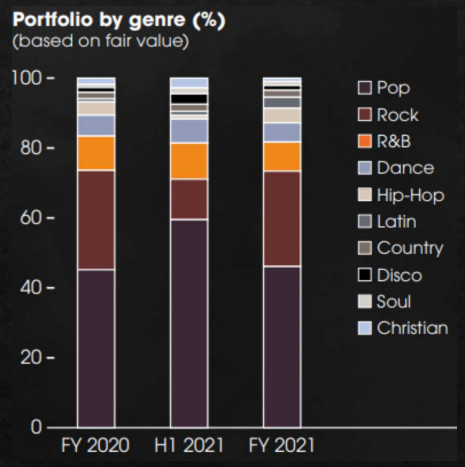
Songs as Assets – Investing in Songs

Portfolio Management

Emergene of Song Funds

Songs as Investments

Balanced Song Portfolio

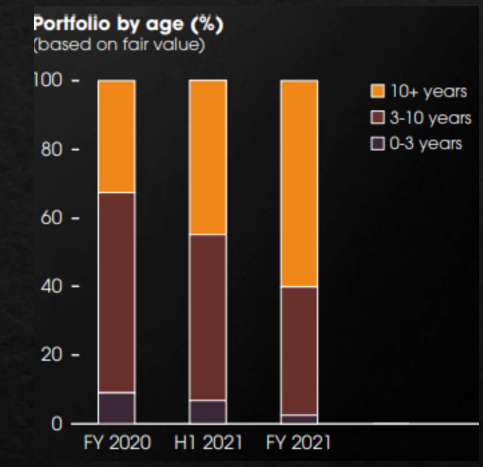


Genre

Investors understand ‘Genre’

Genre Mix

Genre Subjectivity



Vintage

Investors understand ‘Vintage’

Vintage Mix

Vintage Earnings Profile

Business Question & Value Add

1. Can we weight a song's value by it's features i.e. by song genres, individual attributes or combinations of attributes?

2. Can we derive a song's TRUE genre - i.e. inspect song audio attributes > find clusters of similar song types?

We will implement 2 learning algorithms as follows:

- a supervised learning model – feed model the 'human' genres
- an unsupervised learning model – hide genres > find natural clusters

If we find natural songs clusters:

- **how do they relate to song value? Do certain features hold more value, i.e are more investable?**
- can we define them in a way that is understandable to the average human music listener?
- cross-reference them against the 'human' genre classes to determine how natural the 'human' genres are

Modeling

SUPERVISED LEARNING

The models we will look at are:

- KNN
- SVM (multiclass)
- Random Forests

For the supervised learning algorithm we will perform hyperparamter optimisation and model selection to choose a best fit model for multiclass classification.

UNSUPERVISED LEARNING

3 clustering methods:

- K-Means
- DBSCAN
- Hierarchical Clustering

For the unsupervised learning algorithm we will fit and compare 3 clustering methods.

Data Collection and Description

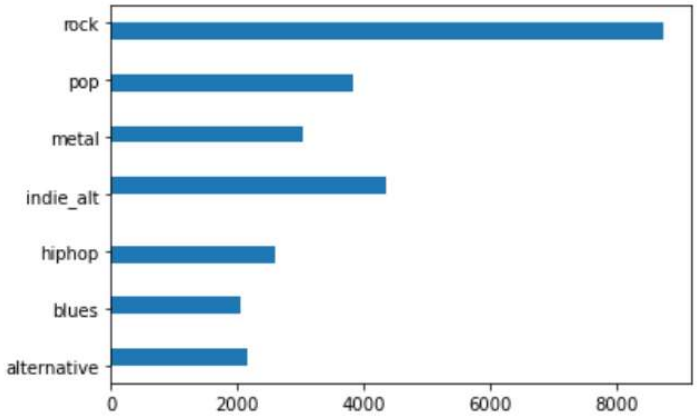
Two datasets were combined….

```
array(["['britpop', 'madchester', 'new wave', 'new wave pop', 'permanent wave', 'pop rock', 'rock']",
       "['modern alternative rock', 'modern rock', 'rock']",
       "['alternative dance', 'indie rock', 'modern alternative rock', 'modern rock', 'new rave', 'oxford indie', 'rock']",
       ...,
       "['album rock', 'classic rock', 'country rock', 'folk rock', 'hard rock', 'mellow gold', 'new wave pop', 'pop rock',
'rock', 'soft rock']",
       "['neo-psychedelic', 'psychedelic soul']",
       "['anime', 'j-metal', 'visual kei']"], dtype=object)
```

1,144 messy genre tags

1. Dataset of songs in Spotify

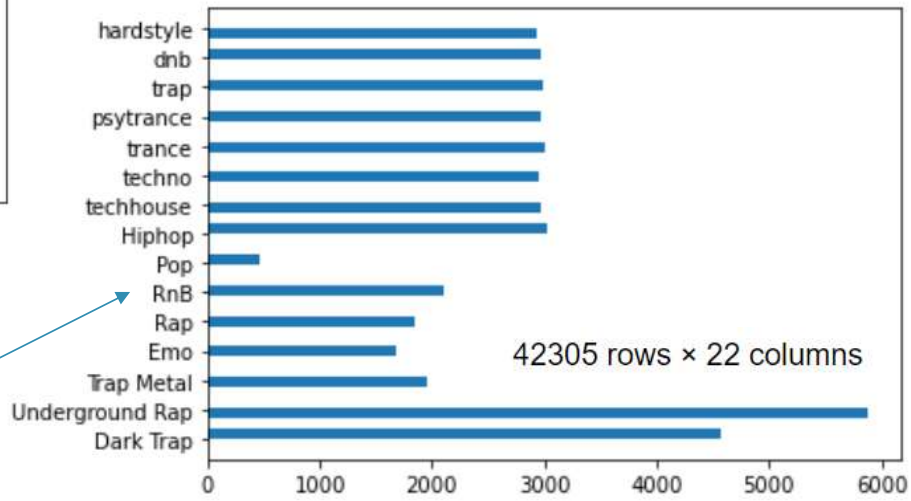
- alternative_music_data.cs
- blues_music_data.csv
- hiphop_music_data.csv
- indie_alt_music_data.csv
- metal_music_data.csv
- pop_music_data.csv
- rock_music_data.csv



kaggle

2. Dataset of songs in Spotify

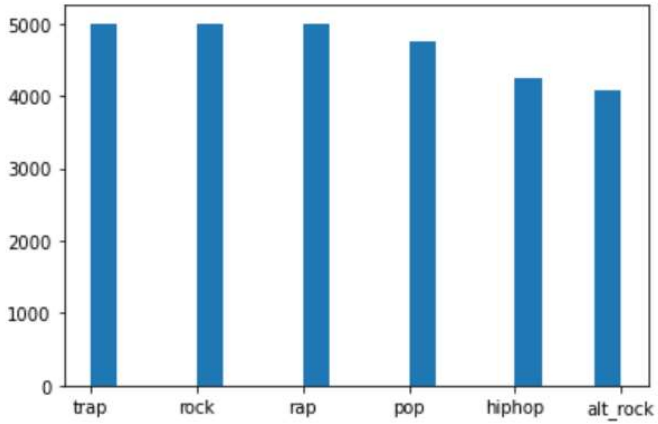
- genres_v2.csv



Data Collection and Description

…and cleaned to match up

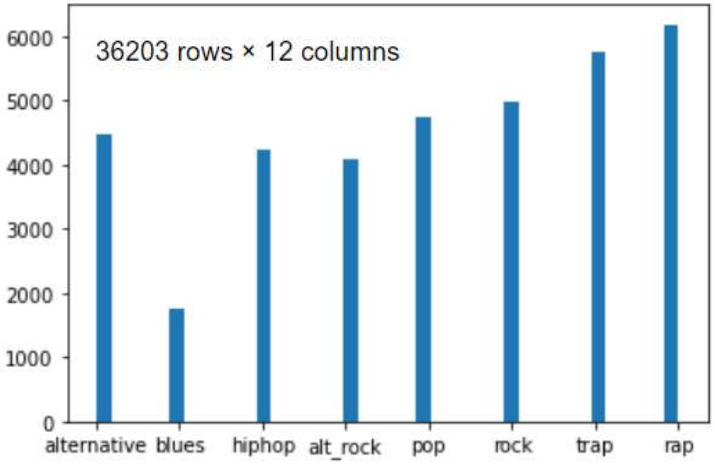
SUPERVISED LEARNING DATASET



28056 rows × 12 columns

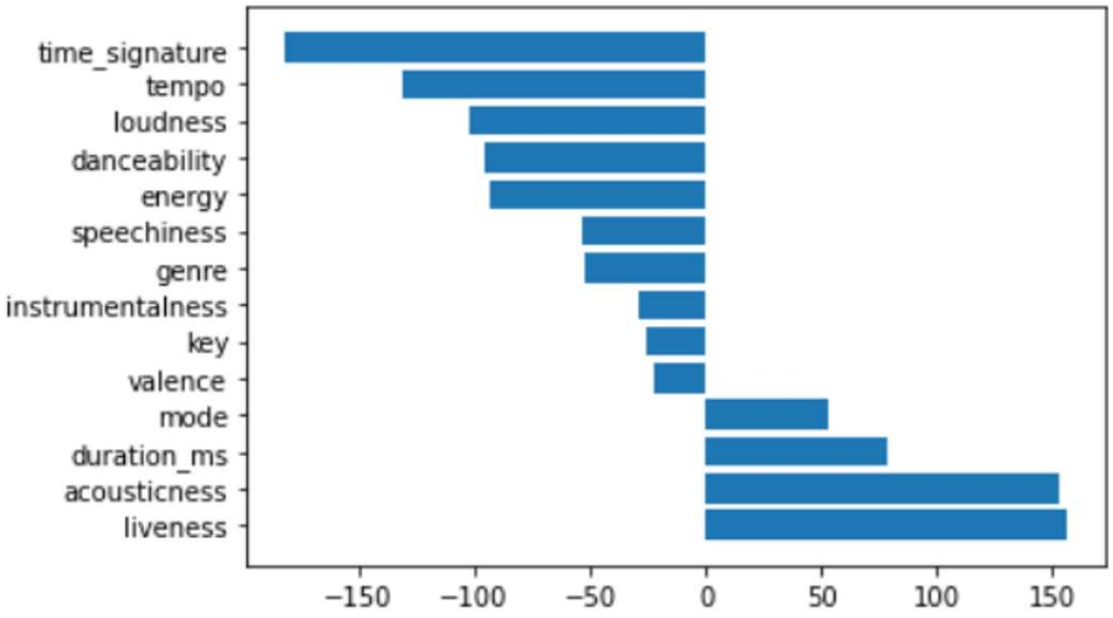
Column Name
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
tempo
time_signature

UNSUPERVISED LEARNING DATASET

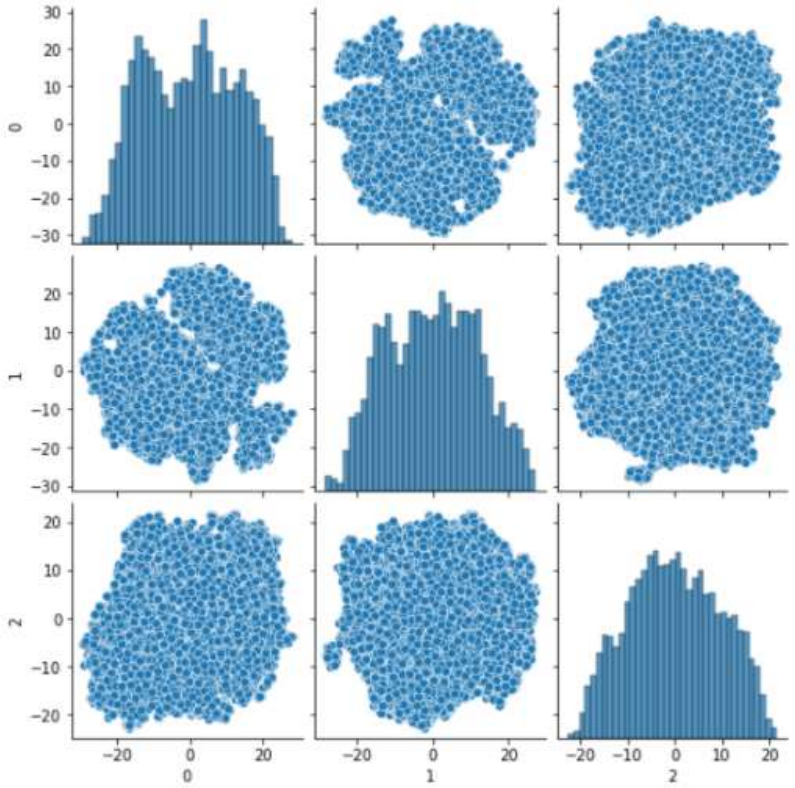


Data Visualization – map genres in space using PCA(n_7) and t_SNE

Feature weight within PCA Components:



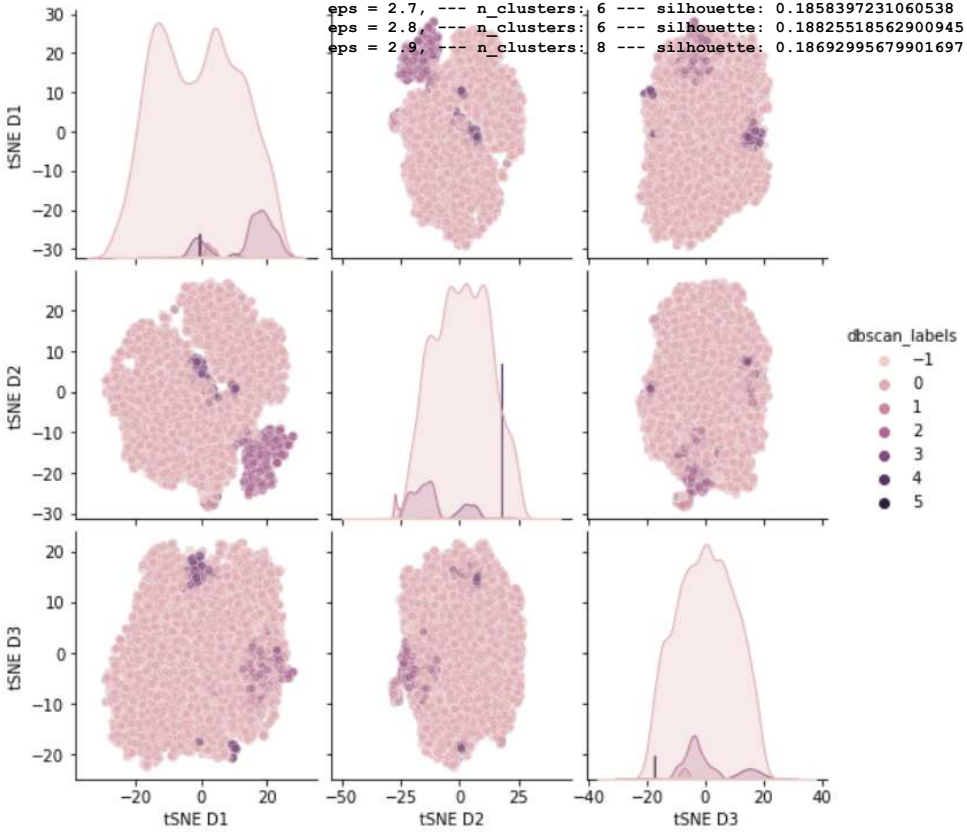
t-SNE 3 Components:



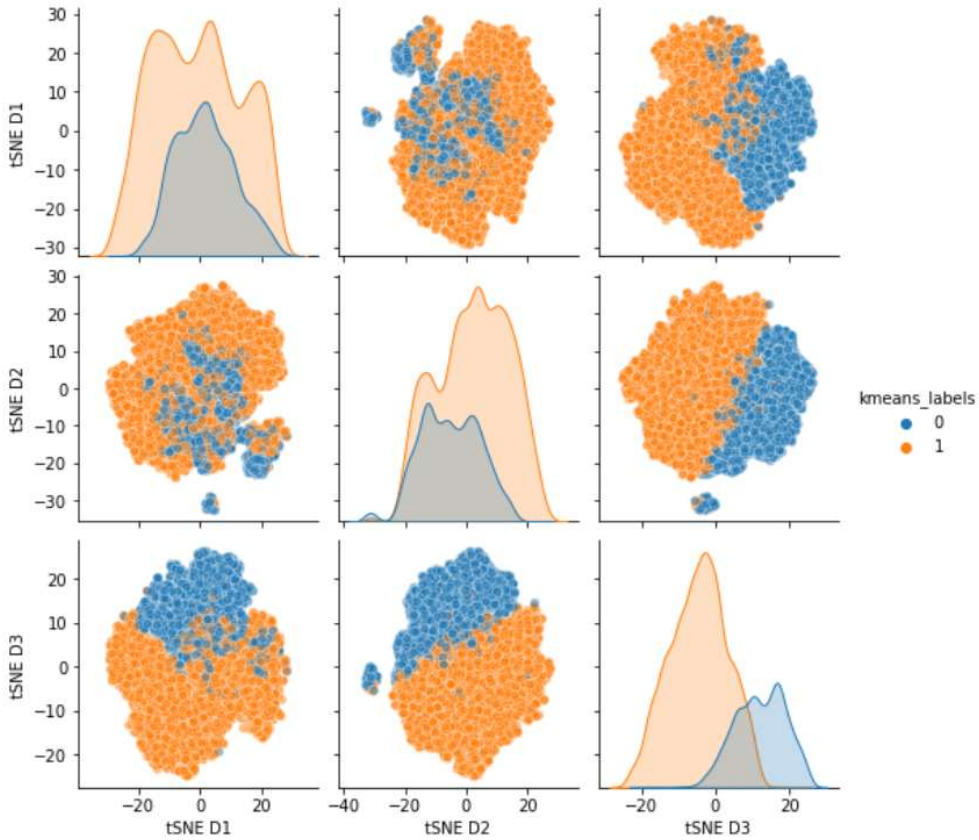
Clustering – First Pass Findings

DBSCAN:

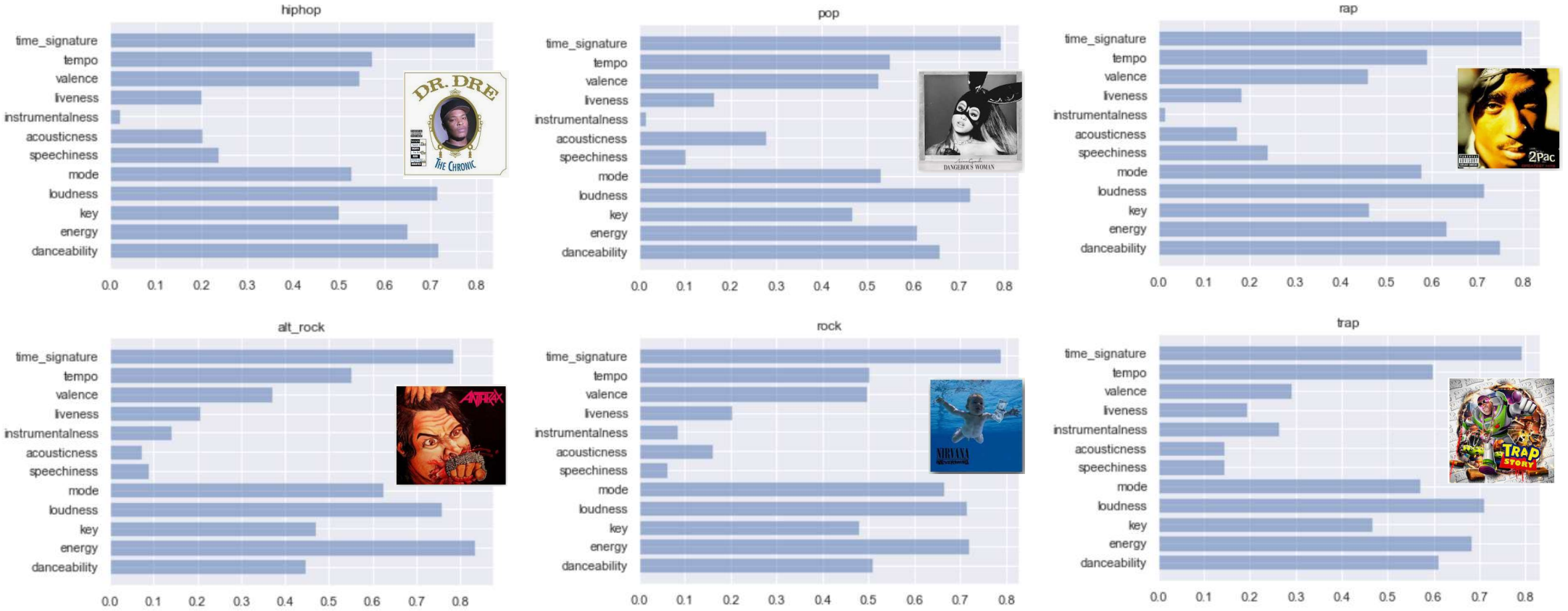
```
eps = 2.0, --- n_clusters: 20 --- silhouette: 0.058890698902765846
eps = 2.1, --- n_clusters: 13 --- silhouette: 0.08095891305190783
eps = 2.2, --- n_clusters: 10 --- silhouette: 0.11531371749430948
eps = 2.3, --- n_clusters: 6 --- silhouette: 0.1280624970232836
eps = 2.4, --- n_clusters: 5 --- silhouette: 0.15236696044842565
eps = 2.5, --- n_clusters: 4 --- silhouette: 0.17971373551556463
eps = 2.6, --- n_clusters: 5 --- silhouette: 0.1805791712301547
eps = 2.7, --- n_clusters: 6 --- silhouette: 0.1858397231060538
eps = 2.8, --- n_clusters: 6 --- silhouette: 0.18825518562900945
eps = 2.9, --- n_clusters: 8 --- silhouette: 0.18692995679901697
```



K-Means:



Genre Feature Distributions - Scaled



TOO SIMILAR– DIVERSIFY GENRES!
WHAT FEATURES CAN WE ADD?

(….also, figure out FacetGrid)

Project Limitations and Future Developments

Project Limitations

- Limited genre / class examples
- Limited access to non-musical features of songs
  - income from 'Synchronisation' i.e. placement of songs in ads, tv, film
  - earnings profiles
  - consumption statistics by territory
  - etc

Future Developments

- **More diverse genres such as EDM, Country, Latin, Disco, Soul**
- Song vintage
- Song lyrics
- There is huge opportunity to develop this model further given access to basic data held privately by copyright owners
- Predict future consumption/earnings, access to half yearly or quarterly earnings is not publicly available

Finally… Next Two Weeks

- **More data from more diverse genres such as EDM, Country, Latin, Disco, Soul**
- **Combine HipHop and Rap, Rock and Metal**
- **Separate RnB, find some Soul to combine it with**
- **Remove ‘Alternative’ genre**
- **Run supervised learning models to teach algorithm how to predict genre from song input**
- **Define and compare clusters (if any) from unsupervised model**
- **Obtain song total consumption figures from Spotify and see if can attribute value to song features**