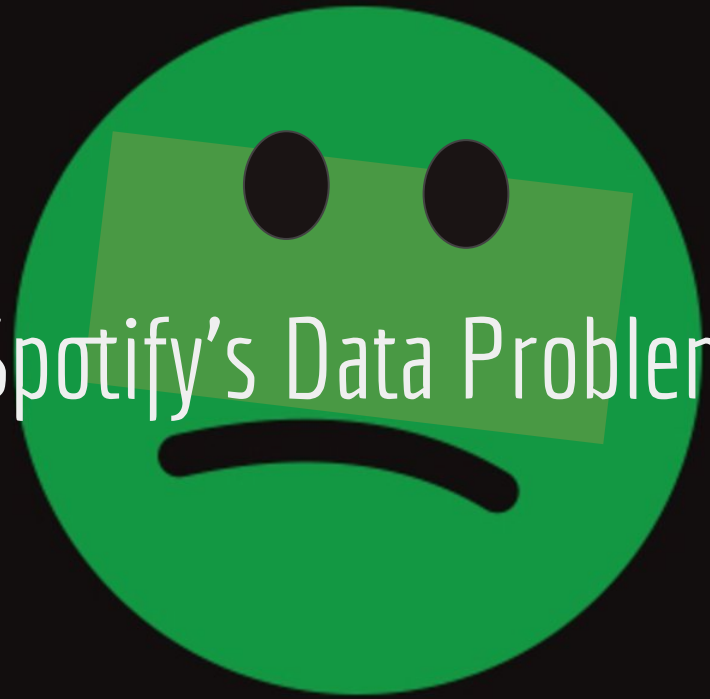


# Spotify's Data Problem

By
Quinn Dizon
Allan Gayahan

# Our Data

## Primary:

- Kaggle Spotify genre & feature dataset
- https://www.kaggle.com/grasslover/spotify-music-genre-list

## Secondary:

- **Spotify API** (for time series data)

## Music Data Usefulness:

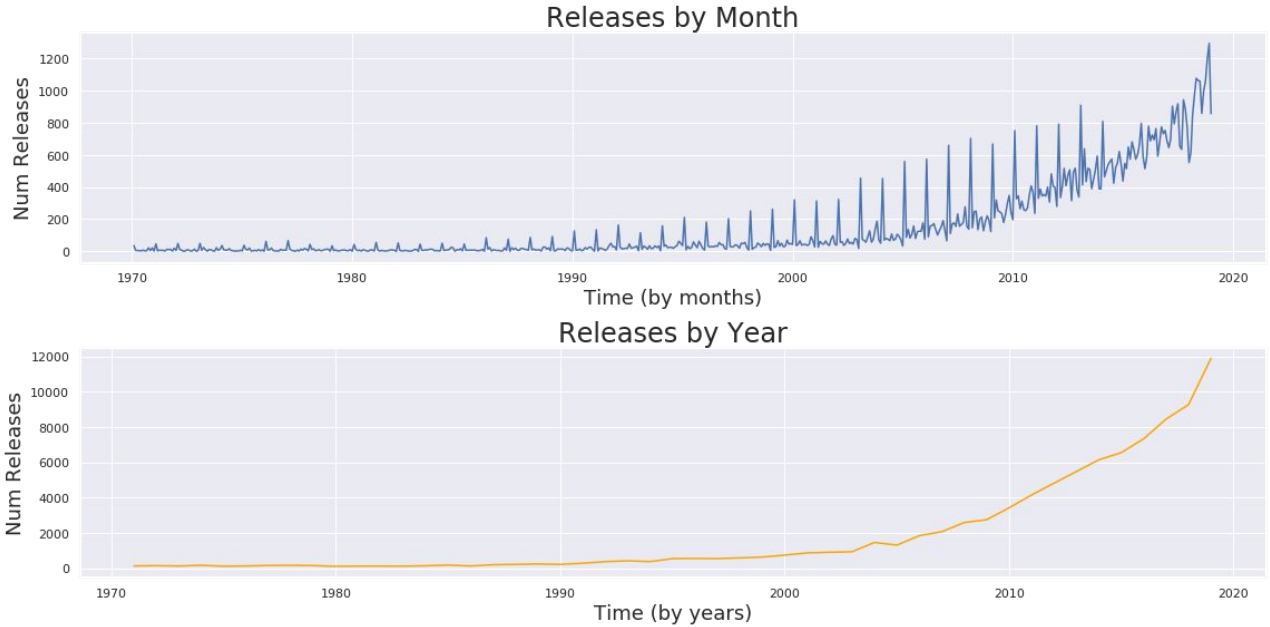
- **Recommending** songs/style based on user preference
- Predicting the next hit song/genre through **trend analysis**
- **AI** generated music
- And More!!!

# Album Releases Over Time

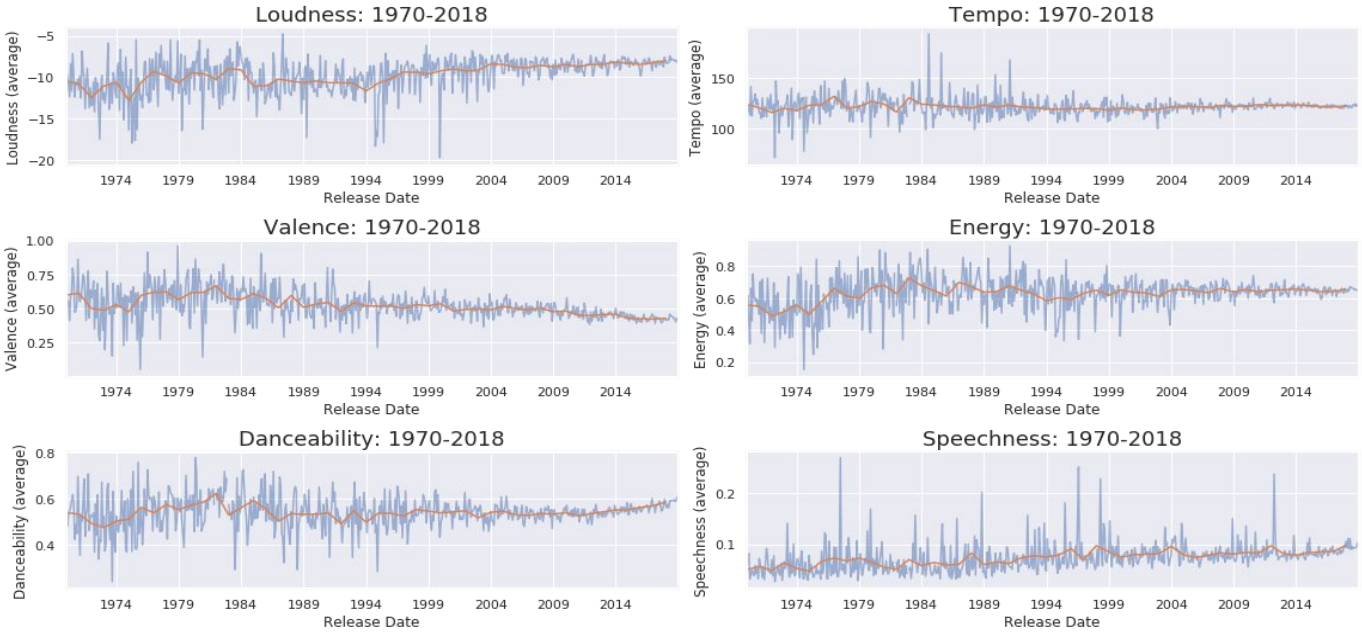
Clear **upwards trend** after ca. 2000

Placeholder values of "*xxxx-01-01*" **skew monthly release data**

Difficult to assess seasonality



# Time Series Analysis: Feature Convergence




**Variance reduces** for nearly all features over time

As more tracks are released, the **more similar their features become**

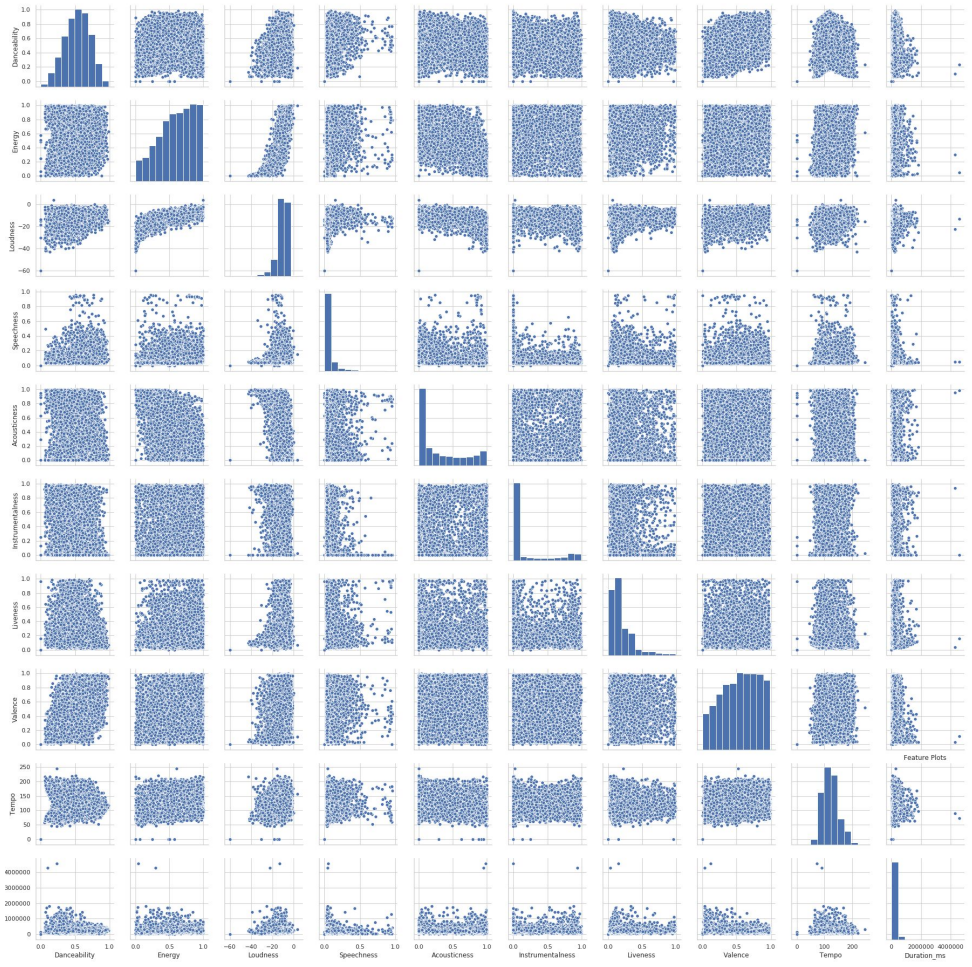
# Linear Regression

Dataset was very clean - no nulls

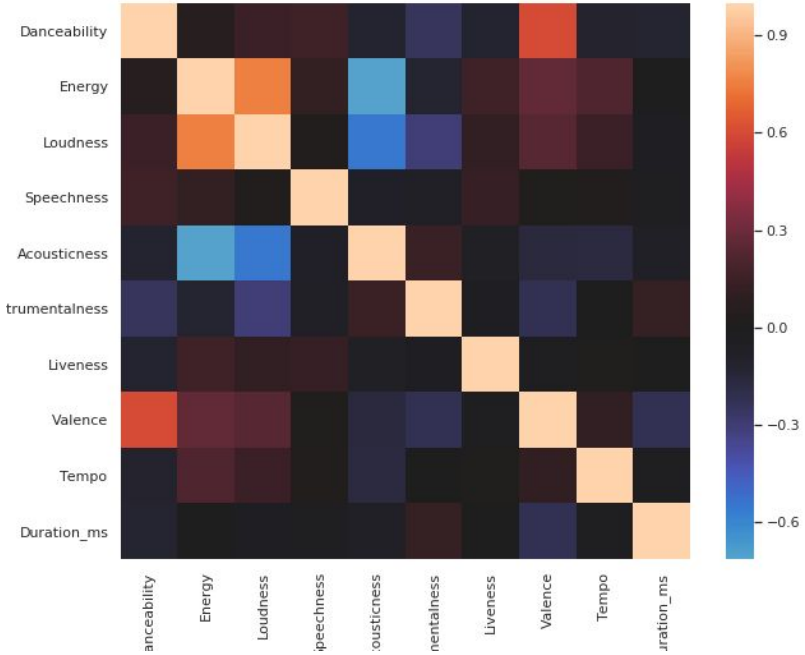
Combination of numerical and categorical features

100k+ entries

Almost all on similar scales  - [0,1]



# Linear Regression




All inclusive model (with dummies)

No Transformation

Target: Energy (of the song)

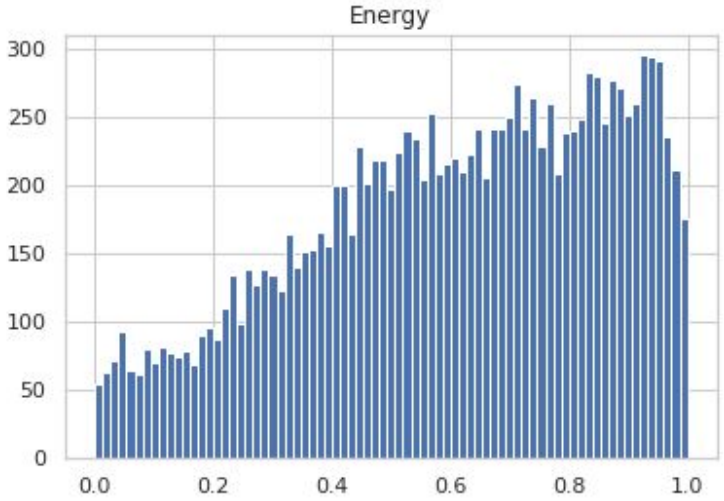
Predictors: All features except Energy

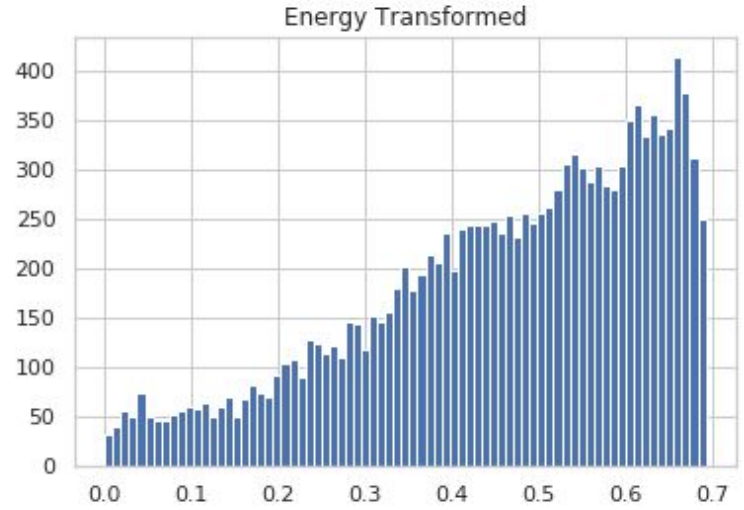
Adjusted R-Squared: 0.774

→ The model accounts for 77% of the data

# Linear Regression: Model Evaluation/Improvement

Np.log1p ⇒ log(1 + Energy)



# Linear Regression: Model Evaluation/Improvement

Sklearn - PowerTransform()

Loudness

Tempo

Liveness

Valence

Duration

**All Inclusive model:**

Adj R-squared: 0.801

**Train Model:**

Adj R-squared: 0.801

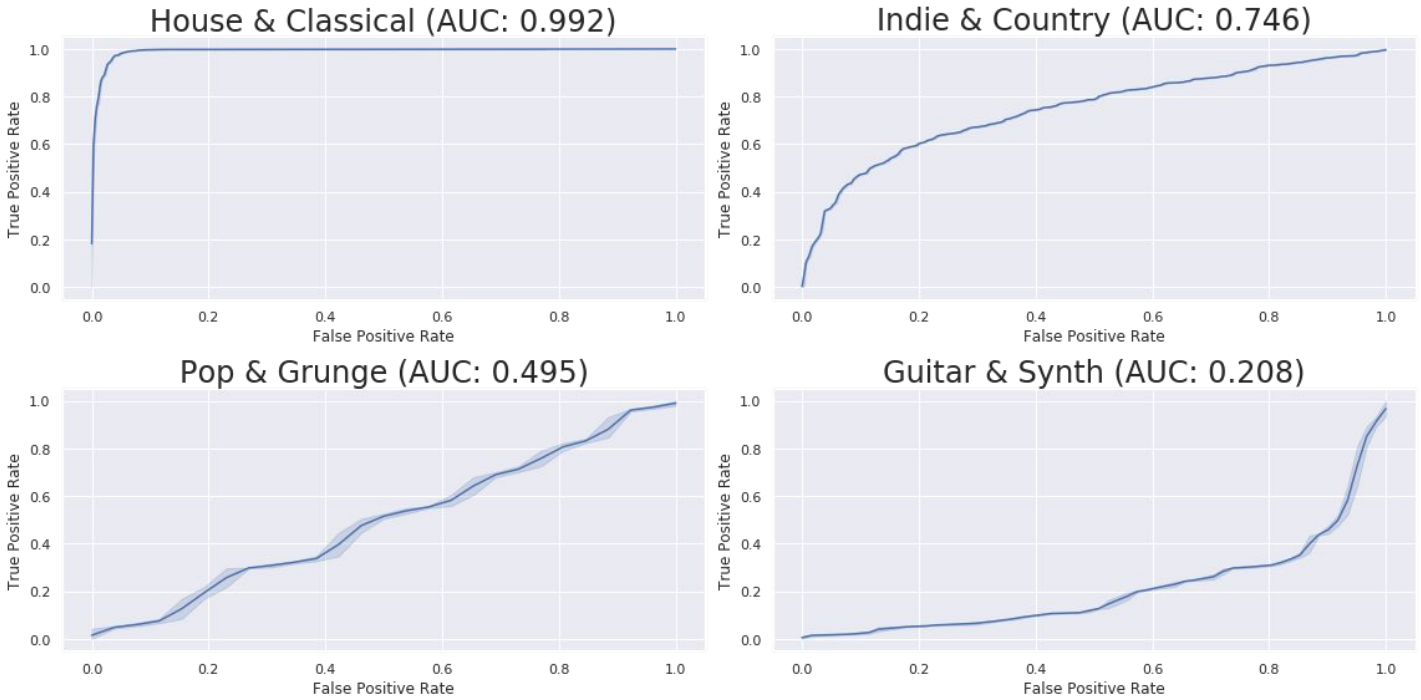
**Test Model:**

Adj R-squared: 0.799

# Logistic Regression: Genre/Style Prediction

Predictive power changes based on genres compared

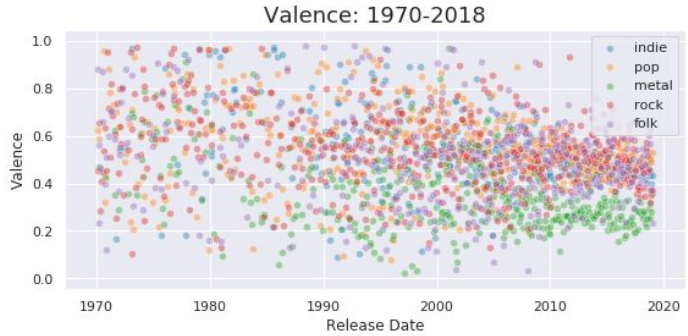
**Multinomial Logistic Regression** - mean accuracy of 0.44

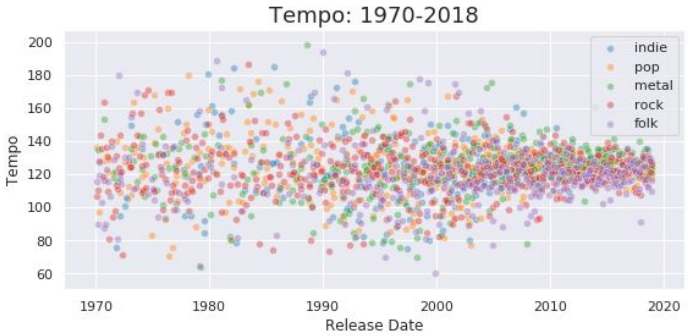


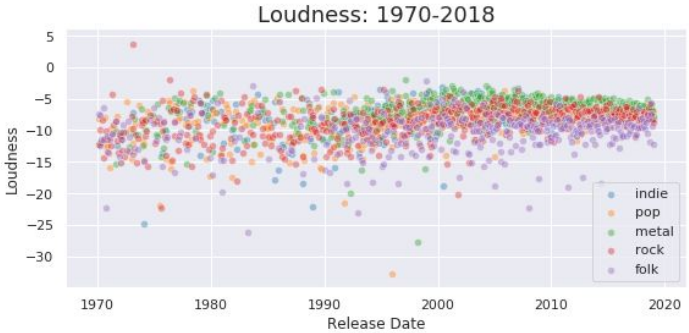
# Features of Most Released Genres Over Time

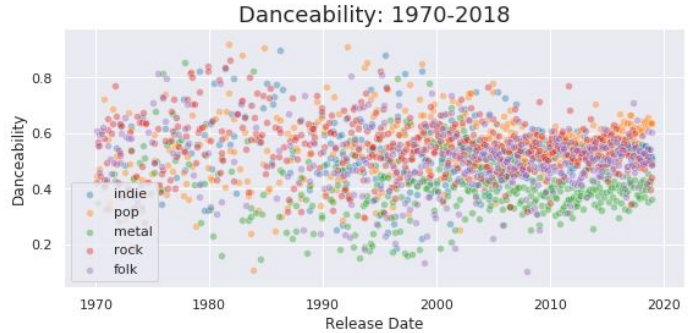
**Features** of styles appear to become **more similar over time**

Supports previous observations of **feature convergence**



# Conclusions

Business Insights

- Spotify should investigate other metrics by which to quantify musical features
- If they don’t, convergence of present feature set will likely lose predictive power over time

Other possible analytic techniques:

- Perhaps deep learning could pull more from these features
- Unsupervised methods (K-means clustering) could help better define genre categories