What makes a song popular?

Christian Shadis

Dataset Attributes and Instances

- From kaggle.com*
- Data taken from Spotify
- One instance = one song
  - 50,000 instances
  - 5,000 per genre
- 12 Attributes
  - 9 Numeric
  - 3 Nominal

*https://www.kaggle.com/vicsuperman/prediction-of-music-genre?select=music_genre.csv

Attribute	Type (# categ.)	Values	Example
Popularity	Numeric	[0, 99]	65
Acousticness	Numeric	[0, 1)	.99
Danceability	Numeric	(0, 1)	.001
Energy	Numeric	(0, 1)	.5
Instrumentalness	Numeric	[0, 1)	0
Liveness	Numeric	(0, 1]	1
Loudness	Numeric	(-48, 4)	-3
Speechiness	Numeric	(0, 1)	.2
Valence	Numeric	[0, 1)	.99
Key	Nominal (12)	A, A#, B, etc.	F#
Mode	Nominal (2)	Major, Minor	Minor
Music Genre	Nominal (10)	Pop, Blues, etc.	Jazz

Preprocessing

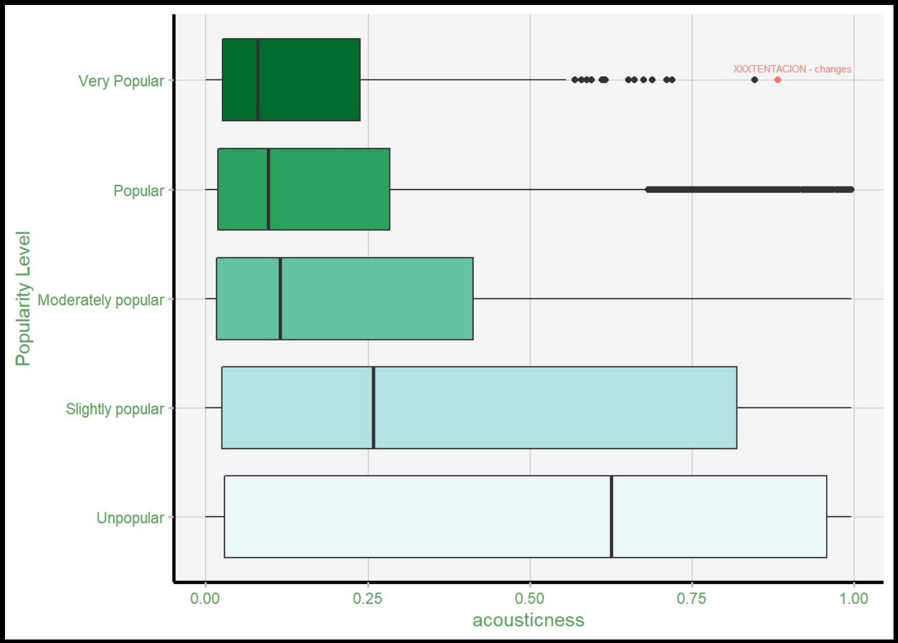
- Discretized popularity variable into 5 bins:
  - Unpopular, Slightly Popular, Moderately Popular, Popular, and Very Popular
    - Each bin is a range of 20 points
- During analysis, discard String attributes (artist name, track name, etc.)
- Removed problematic variables (missing values and unimportant for analysis
  - Song duration in ms
  - Tempo
  - Obtained date (date the song was scraped from Spotify)

Summary Statistics

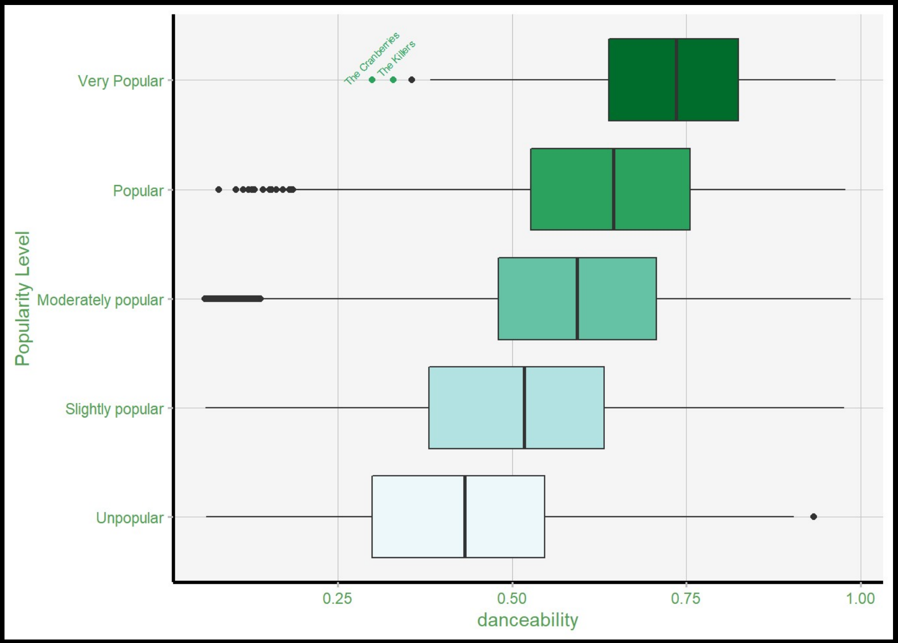
Variable	Minimum	Q1	Median	Q3	Maximum
Popularity	0	34	45	56	99
Acousticness	0	0.02	0.14	0.55	0.99
Danceability	0.06	0.44	0.57	0.69	0.99
Energy	7.92e-4	0.43	0.64	0.82	0.999
Instrumentalness	0	0	1.58e-4	0.16	0.99
Liveness	0.01	0.10	0.13	0.24	1
Loudness	-47.05	-10.86	-7.28	-5.17	3.74
Speechiness	0.02	0.04	0.05	0.10	0.94
Valence	0	0.26	0.45	0.648	0.99

Numeric Distributions

Acousticness

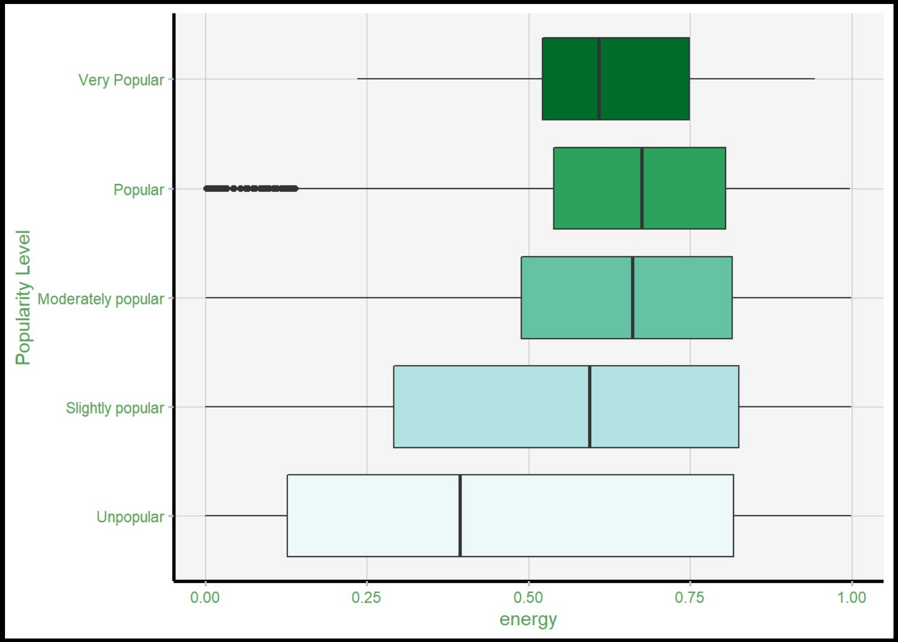


Danceability

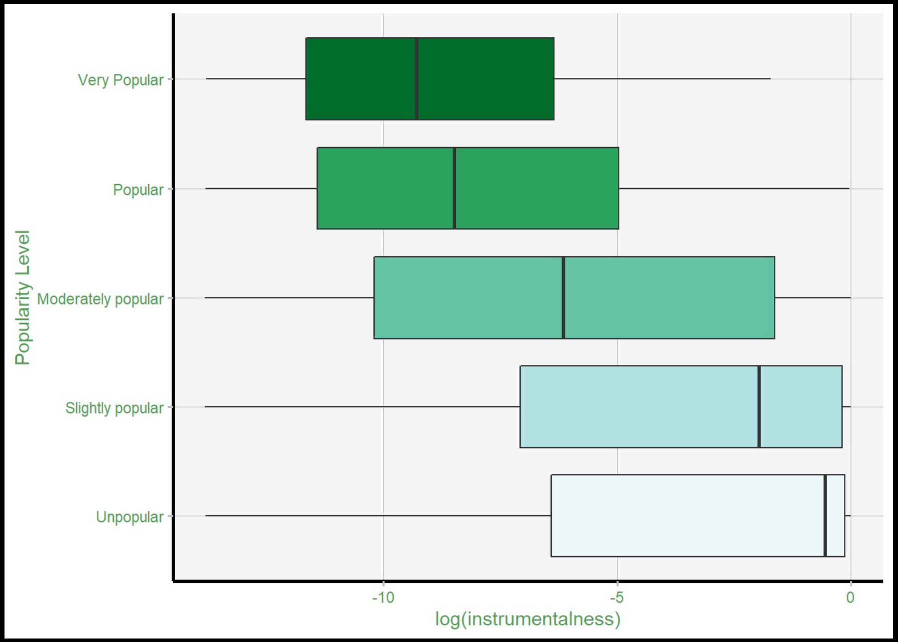


Numeric Distributions

Energy

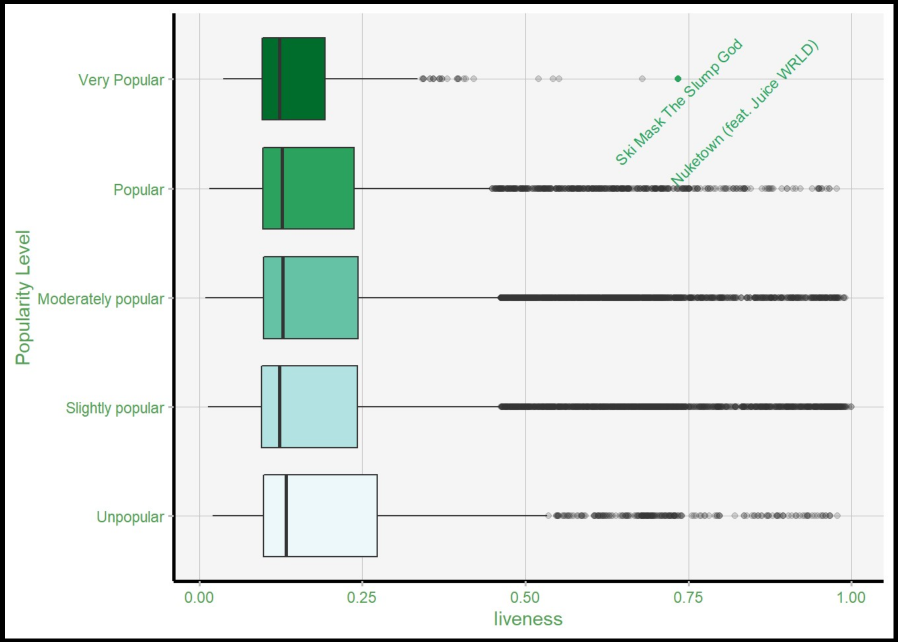


Instrumentalness (log-transformed)

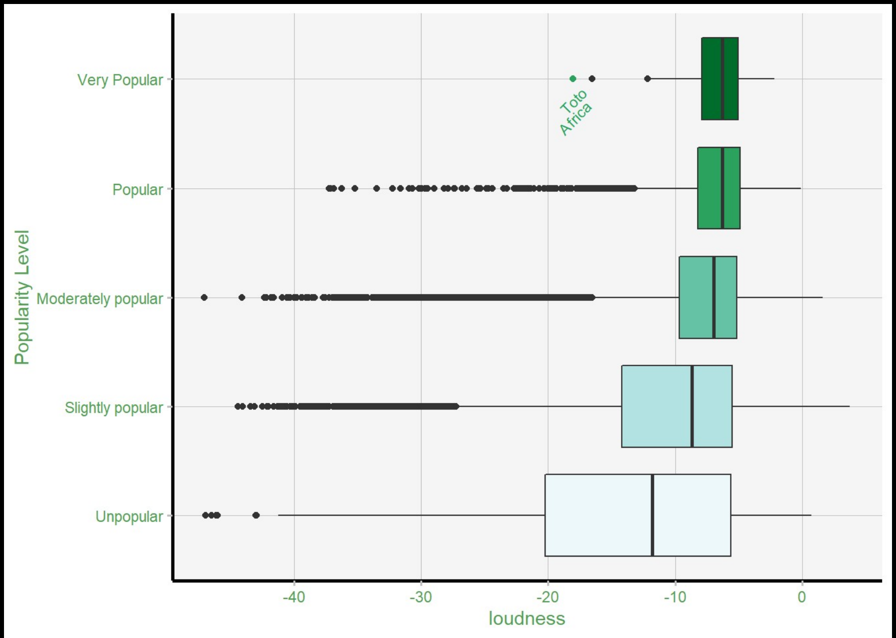


Numeric Distributions

Liveness

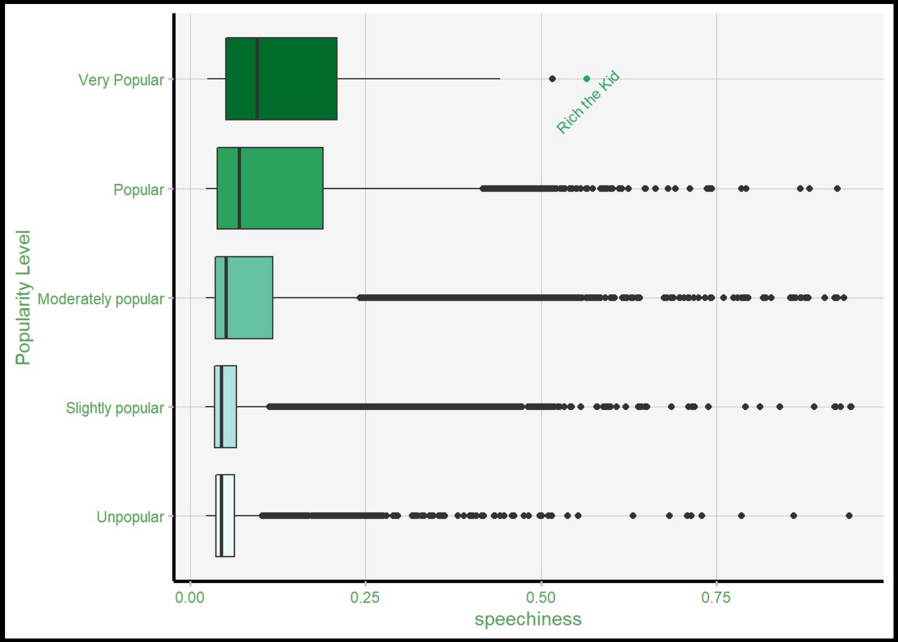


Loudness

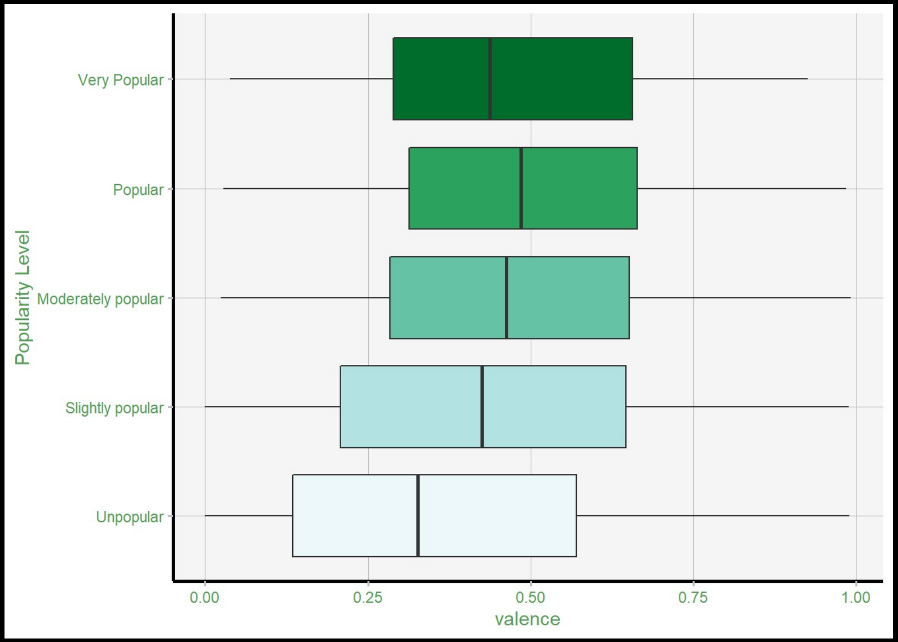


Numeric Distributions

Speechiness



Valence



Data Mining Questions

1. Can we predict a song's popularity using numeric attributes?
   - Predict Popularity using Multiple Linear Regression
   - Predict discretized popularity using Classification algorithm
2. Can we classify a song's genre based on numeric attributes?
   - Predict genre using classification algorithm

Predict Popularity Using Numeric Attributes: Multiple Linear Regression

- Weka:
  - Linear Regression: Relative absolute error 84.6%
  - Simple Linear Regression: Relative absolute error 90.9%
- R:
  - Best Subsets Regression (leaps) – Chose 3 models with highest R^2
    - Popularity ~ acousticness + danceability + energy + instrumentalness + liveness + loudness + speechiness + valence (R^2 = 0.233)
    - Popularity ~ acousticness + danceability + energy + instrumentalness + liveness + loudness + speechiness (R^2 = 0.231)
    - Popularity ~ acousticness + danceability + instrumentalness + liveness + speechiness + valence (R^2 = 0.229)
- Multiple Linear Regression **Not Effective**

Lack of Linear Correlation

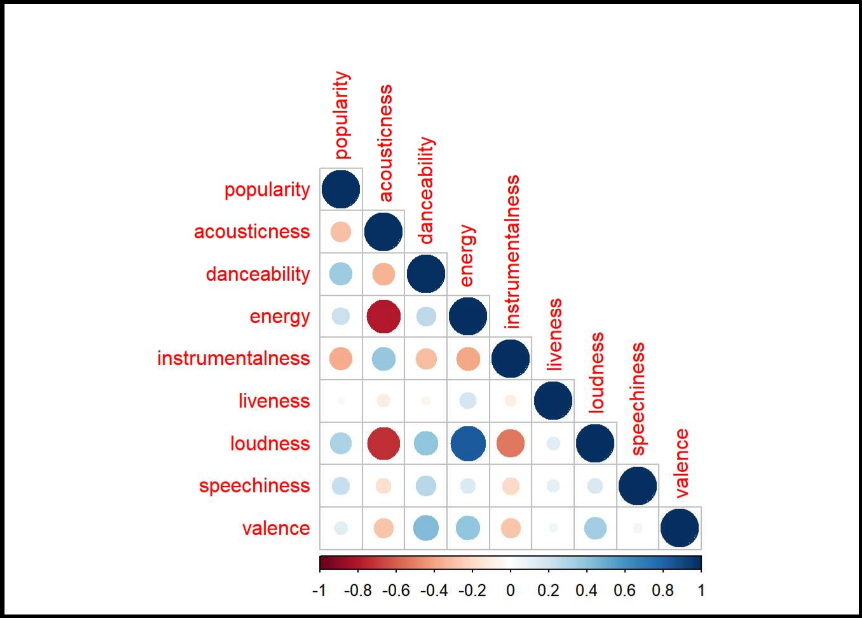
Results

- Strong Negative Correlation:
  - Energy / Acousticness
  - Loudness / Acousticness
- Moderately Strong Negative Correlation:
  - Loudness / Instrumentalness
  - Energy / Instrumentalness
- Moderately Strong Positive Correlation:
  - Valence / Danceability
- Strong Positive Correlation:
  - Loudness / Energy

No interesting or unexpected correlations

This likely caused poor performance of Multiple Linear Regression

Correlation Map



Predict Popularity (Discretized): Classification

- Algorithms:
  - Trees: J48, Random Tree, Random Forest, REPTree
  - Bayes: NaiveBayes, BayesNet
  - Rules: OneR, PART
- Variables considered:

Attribute	Type (# categ.)	Values
Popularity	**Nominal**	**Unpopular – Very Popular**
Acousticness	Numeric	[0, 1)
Danceability	Numeric	(0, 1)
Energy	Numeric	(0, 1)
Instrumentalness	Numeric	[0, 1)
Liveness	Numeric	(0, 1]
Loudness	Numeric	(-48, 4)
Speechiness	Numeric	(0, 1)
Valence	Numeric	[0, 1)

Popularity Classification Accuracy

Test Option / Algorithm	J48	RepTree	Random Tree	Random Forest	NaiveBayes	BayesNet	OneR	PART
Cross – Validation: 5	49.83%	51.18%	48.45%	57.75%	35.37%	45.66%	45.84%	51.68%
Cross – Validation: 10	50.15%	51.40%	49.04%	58.36%	35.42%	45.59%	46.13%	51.83%
Cross – Validation: 20	50.55%	51.63%	49.37%	58.63%	35.38%	45.66%	46.36%	51.92%
Percent Split: 66	49.11%	50.73%	47.13%	56.68%	34.94%	46.14%	45.49%	51.33%
Percent Split: 80	50.16%	51.97%	48.11%	58.11%	35.53%	46.22%	45.65%	52.07%
Percent Split: 90	50.76%	52.16%	48.52%	58.06%	35.44%	46.48%	46.28%	52.22%

Popularity Classification Results

- Best performance: Random Forest, 20 cross-validation folds
- **Classification of popularity unsuccessful**

Predict Music Genre: Classification

- Algorithms:
  - Trees: J48, Random Tree, REPTree
    - Random Forest excluded – computing limitations
  - Bayes: NaiveBayes, BayesNet
  - Rules: OneR, PART
- Variables considered:

Attribute	Type (# categ.)	Values
Genre	**Nominal**	**Pop, Blues, Etc**
Acousticness	Numeric	[0, 1)
Danceability	Numeric	(0, 1)
Energy	Numeric	(0, 1)
Instrumentalness	Numeric	[0, 1)
Liveness	Numeric	(0, 1]
Loudness	Numeric	(-48, 4)
Speechiness	Numeric	(0, 1)
Valence	Numeric	[0, 1)

Genre Classification Accuracy

Test Option / Algorithm	J48	RepTree	Random Tree	NaiveBayes	BayesNet	OneR	PART
Cross – Validation: 5	33.55%	39.14%	29.79%	33.57%	40.11%	21.08%	33.33%
Cross – Validation: 10	33.60%	39.36%	29.21%	33.54%	40.17%	21.00%	33.54%
Cross – Validation: 20	33.52%	39.72%	29.10%	33.53%	40.29%	21.06%	33.64%
Percent Split: 66	33.75%	39.41%	29.53%	33.60%	40.24%	20.86%	33.71%
Percent Split: 80	33.57%	38.29%	29.38%	33.22%	39.42%	21.12%	33.05%
Percent Split: 90	33.82%	39.14%	29.10%	32.98%	39.12%	21.02%	33.32%

Genre Classification Results

- Best performance: BayesNet, 20-fold cross-validation
- **Classification of genre unsuccessful**

- **Bonus Question: can we distinguish Rock vs. Jazz?**
  - **Refined dataset to Rock and Jazz only**
  - **PART algorithm (80/20 split) achieved 81.6% accuracy**

Conclusion

- There is no “formula” for a successful song
  - Though distributions suggest that popular songs would tend to have:
    - Many instruments (low acousticness)
    - Relatively high danceability
    - Moderate energy
    - Vocals (low instrumentalness)
    - Studio Recording quality (low liveness)
    - Little to no speech
- Music genres are not strictly divided by numerical attributes
- Two genres can be distinguished, but not all genres at once

References

Tools:

- R Programming Language
- Rstudio
- WEKA
- Past Experience

R packages:

- Tidyverse
- GridExtra
- Corrplot
- Leaps
- Ggrepel

R package documentation:

- https://www.tidyverse.org/
- https://cran.r-project.org/web/packages/gridExtra/index.html
- https://www.rdocumentation.org/packages/corrplot/versions/0.84/topics/corrplot
- https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/leaps
- https://cran.r-project.org/web/packages/ggrepel/ggrepel.pdf

Dataset:

- https://www.kaggle.com/vicsuperman/prediction-of-music-genre?select=music_genre.csv