Search

Home

Library

STAT 4214 Final Project:

Predicting Successful Songs

Ashley Covitz, Ethan Appiah, Lidia Pinkevich, Nami Jain

Table of contents

Search

Home

Library

Our Music Playlists

01

Introduction to Problem & Data

Problem Context

Who Collected?

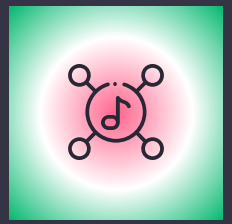
3:15

Spotify compiles the top 100 most-played songs of that year.

What is Collected?

3:20

Data on Top Songs of 2018 that includes additional audio features.

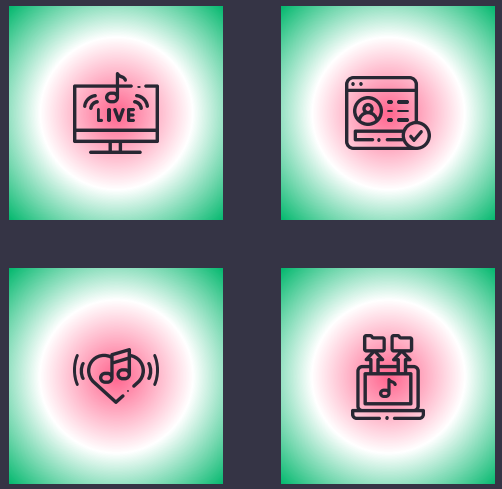


Why Important?

3:10

Producers, artists, music industry can predict and optimize.

Dataset Introduction



Top Songs of 2018

- 100 observations
- 13 explanatory variables
- Response variable of Rank

Using the audio feature included in the file (explanatory variables), we are hoping to predict where a song would end up in the charts.

Variables

#1 Danceability
0-1 scale, 1 most

#2 Key
Number representation

#3 Speechiness
0-1 scale, lyrical

#4 Instrumentalness
0-1 scale, 1 most

#5 Valence
0-1 scale, positiveness

#6 Energy
0-1 scale, 1 most

#7 Loudness
-10 to -3, Decibels

#8 Acousticness
0-1 scale, how acoustic

#9 Liveness
0-1 scale, performance

#10 Tempo
Speed, beats per min.

#11 Others
Mode, Duration, Time Signature

Search

Home

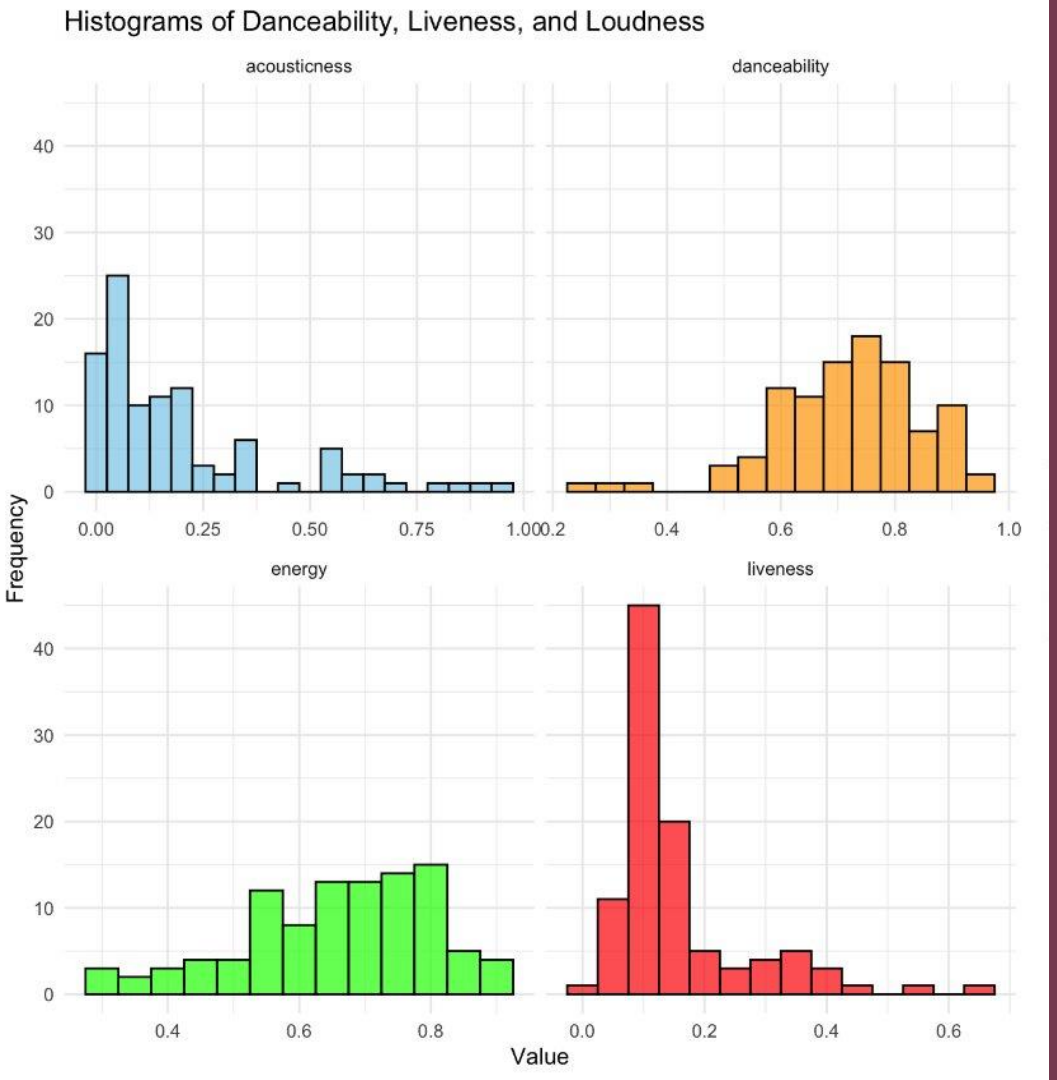
Library

Our Music Playlists

02

Exploratory Data Analysis

Graphic 1:
Initial Variables Investigation

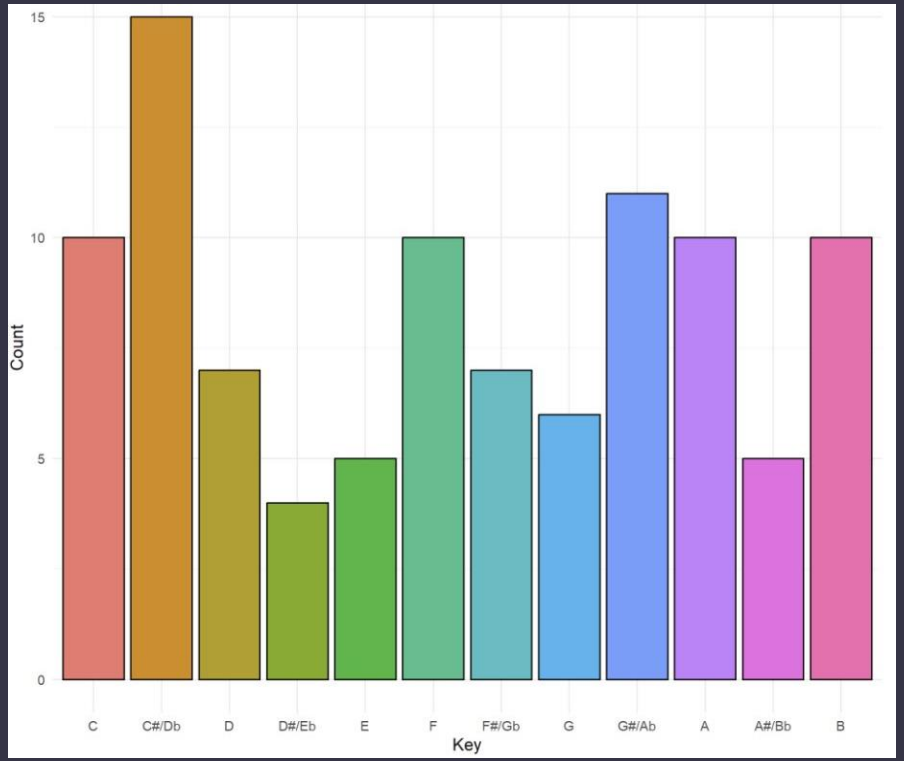


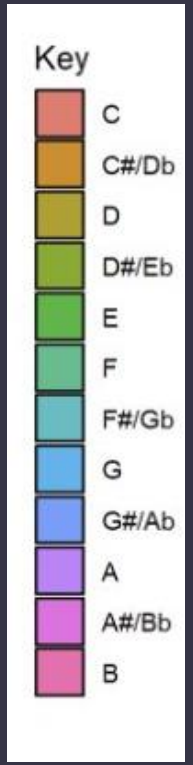
Search

Home

Library

Graphic 2: Investigation of Key



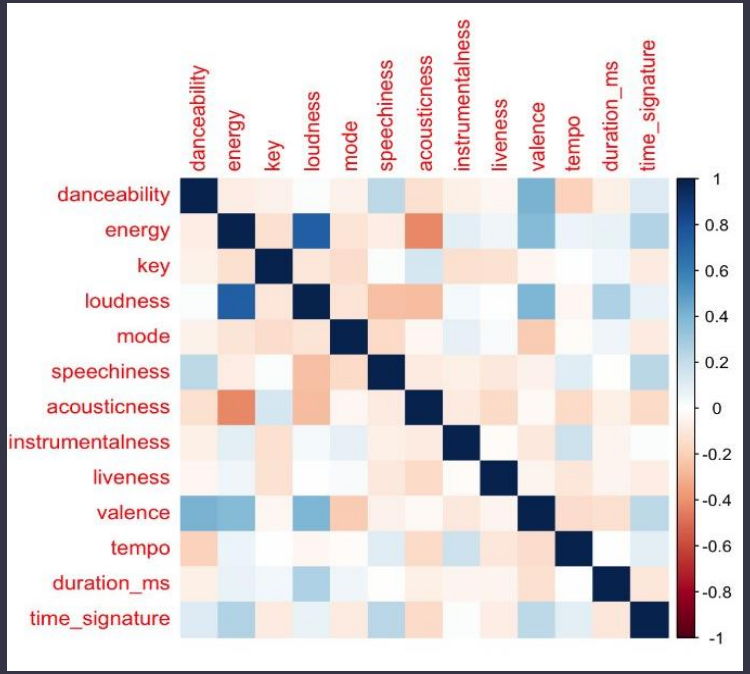


Search

Home

Library

Graphic 3: Multicollinearity Heatmap



Search

Home

Library

Our Music Playlists

03

Model Fitting Process

Linear Model Results

0.1584
R-Squared

Built via AIC

Loudness, Liveness, and Key significant at 10% level

0.0644
R-Squared

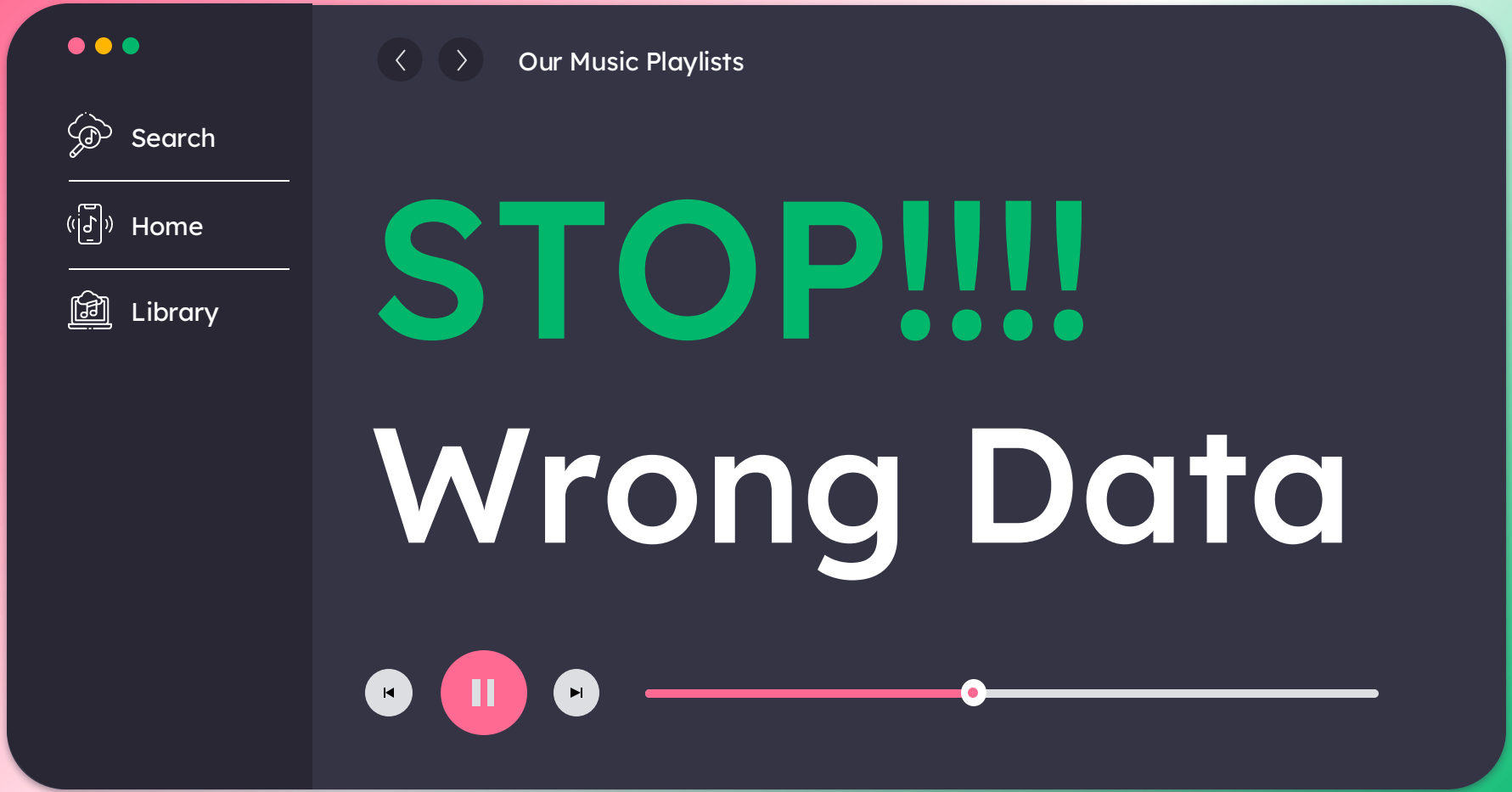
Built via BIC

Only Liveness included, significant variable

Influential Outlier Songs

Order by

#	Title	Artist	Released
92	Yes Indeed	Lil Baby	May 17
91	Promises (ft. Sam Smith)	Calvin Harris	August 17
44	Thunder	Imagine Dragons	April 27 (2017)
99	Dusk Till Dawn	ZAYN	September 7 (2017)
13	Nice For What	Drake	April 6
85	Perfect Duet (ft Beyonce)	Ed Sheeran	December 1 (2017)
98	No Brainer	DJ Khaled	July 27


Search
Home
Library
Our Music Playlists
STOP!!!!
Wrong Data

What is Ordinal Logistic Regression?

Ordered Categories

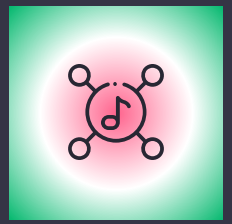
3:15

Order matters, but difference isn't exactly measurable; more options than binary

Estimation

3:20

Probability of being in an ordered category based on cumulative odds of being in a higher category



Split into "bins" of success

3:10

Compared top 10 to top 20 vs between songs 18 and 19

Logisitic Model Results

487.379
AIC

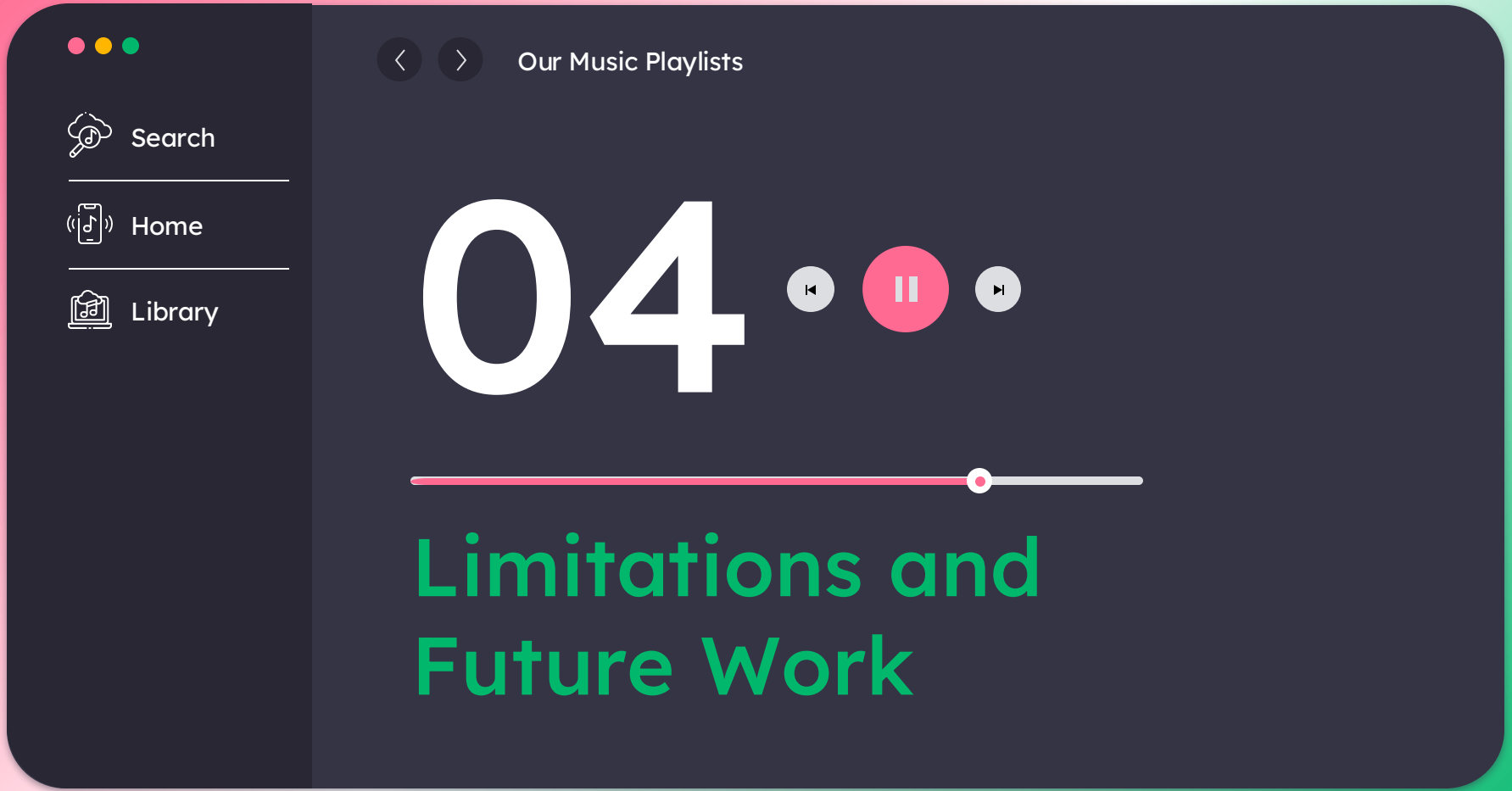
Full Model

No significant variables

478.521
AIC

Model via Step

Loudness, Speechiness, Valence
significant at 10% level


Search
Home
Library
Our Music Playlists
04
Limitations and
Future Work

Future Work

Assumptions

You listened 4h35min

- Proportional Odds Assumption
- Linearity of Log Odds
- Independence
- Multicollinearity

Outliers

You listened 3h15min

- Box plots or z-scores
- Cook's Distance
- Predicted probabilities

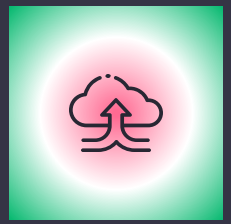
Limitations

Dataset

Only 100 data points, all from 2018

Sample Bias

Already established successful songs



Release Date

Success measured by accumulated streams

Artist

Dedicated fanbases skew streaming

Conclusion

Problem
What qualities determine a successful song?

Linear
Terrible results, recognize improper data

Logistic
Ordinal Logistic Regression
- Loudness, Speechiness, Valence

Future
Test assumptions and outliers to check validity of model

Limitations
Limited biased sample, external influences

Search

Home

Library

Our Music Playlists

Questions?

AIC-Built Linear Model

```
Call:
lm(formula = id ~ danceability + key + loudness + liveness +
    valence + tempo, data = spotify)

Residuals:
    Min      1Q  Median      3Q     Max
-55.051 -22.705  -0.738  21.766  60.395

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.85788   24.94195   1.798  0.07534 .
danceability -33.89882   23.85911  -1.421  0.15872
key           -1.41551    0.76279  -1.856  0.06666 .
loudness      -3.51768    1.74069  -2.021  0.04617 *
liveness     -67.56711   25.13911  -2.688  0.00852 **
valence       23.57732   16.44615   1.434  0.15504
tempo          0.13991    0.09874   1.417  0.15984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.46 on 93 degrees of freedom
Multiple R-squared:  0.1584,    Adjusted R-squared:  0.1041
F-statistic: 2.917 on 6 and 93 DF,  p-value: 0.01188
```

Used AIC starting with full model, used direction "both" completed in 10 steps
k (penalty factor) = 2

Rank ID = 44.9
- 33.9 * dance
- 1.4 * key
- 3.5 * loudness
- 67.6 * liveness
+ 23.6 * valence
+ 0.14 * tempo

Appendix 1

BIC-Built Linear Model

```
Call:
lm(formula = id ~ liveness, data = spotify)

Residuals:
    Min      1Q  Median      3Q     Max
-50.827 -24.099   0.632  22.013  51.488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    60.94       4.91  12.411   <2e-16 ***
liveness      -65.92      25.39  -2.597   0.0109 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.21 on 98 degrees of freedom
Multiple R-squared:  0.06437,   Adjusted R-squared:  0.05482
F-statistic: 6.742 on 1 and 98 DF,  p-value: 0.01086
```

Used BIC starting with full model, completed in 1 step

k (penalty factor) = 4.60517

Rank ID = 60.94 - 65.92 * liveness

Appendix 2

Full Logistic Model

- Split into 10 "popularity" bins
  - 1-10 in bin 1, …, 91-100 in bin 10

Log-odds(popularity) = 1.44 * energy + 0.03 * key + 0.18 * loudness – 0.223 * mode + 2.82 * speechiness + 0.82 * acousticness – 7.08 * instrumentalness – 0.70 * liveness – 1.98 * valence

No significant variables

AIC: **487.379**

Appendix 3

Logistic Model Built Step-Wise

- Split into 10 "popularity" bins
  - 1-10 in bin 1, …, 91-100 in bin 10

Log-odds(popularity) = 0.217 * loudness + 2.99 * speechiness – 1.655 * valence

All significant variables at 10% level

AIC: **478.521**

Appendix 4

Music streaming icon pack