ARTIST & SONG POPULARITY

IST687 Final Project

By: Courtney Smith, Jeremy Wallner, John Fields, Juan Castro


THE MASKED SINGER
American Idol
Altogether Legendary
the Voice
SONGLAND
LA BANDA
UM GRUPO, UM SONHO
THE FOUR
BATTLE FOR STARDOM

OVERVIEW

- The purpose of this project is to analyze the Million Song Database to predict "Hot-Warm-Cold" artists and songs based on attributes such as familiarity, artist location, loudness, terms used, and other variables.

- The analysis was done using R software on a 10,000 track subset of the data and our model was able to predict ratings of "Hot-Warm-Cold“ for artists with ~80% accuracy and songs with ~50% accuracy.


DATA

DATA

- **Song Dataset**
  - As the original Million Song Dataset (MSD) is incredibly large (~280GB), we based this analysis on a subset of 10,000 songs (1.8GB) for ease of data parsing, manipulation and modeling. The dataset was downloaded from CORGIS.

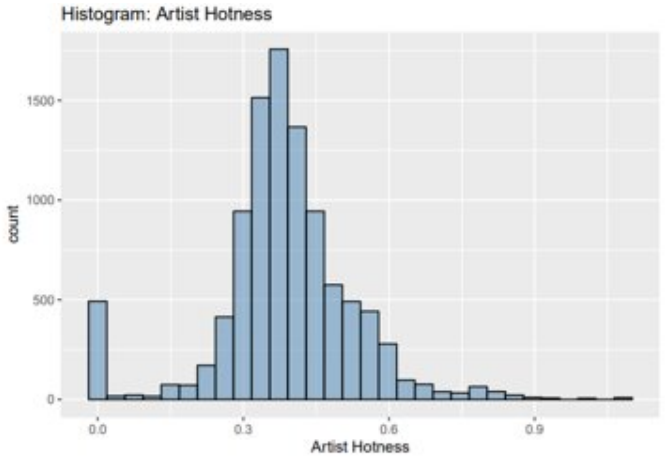
- **Artist Dataset**
  - The dataset contains 36 variables and 9,996 observations. The table below identifies the various fields and provides a description for each as defined by millionsongdataset.com.

```
##                          vars     n      mean       sd    median
## artist.hotttnesss           1 10000      0.39     0.14      0.38
## artist.id*                  2 10000   1907.24  1123.21   1882.00
## artist.name*                3 10000   2206.28  1270.75   2195.00
## artist_mbtags*              4 10000     50.04    79.26      1.00
## artist_mbtags_count         5 10000      0.52     0.88      0.00
## bars_confidence             6 10000      0.24     0.29      0.12
## bars_start                  7 10000      1.07     1.72      0.79
## beats_confidence            8 10000      0.61     0.32      0.69
## beats_start                 9 10000      0.43     0.81      0.33
## duration                   10 10000    240.62   246.08    223.06
## end_of_fade_in             11 10000      0.76     1.86      0.20
## familiarity                12  9996      0.57     0.16      0.56
## key                        13 10000      5.37     9.67      5.00
## key_confidence             14 10000      0.45     0.33      0.47
## latitude                   15 10000     37.16     9.54     37.16
## location*                  16 10000    596.95   238.92    705.00
## longitude                  17 10000    -63.93    30.89    -63.93
## loudness                   18 10000    -10.48     5.40     -9.38
## mode                       19 10000      0.69     0.46      1.00
## mode_confidence            20 10000      0.48     0.19      0.49
## release.id                 21 10000 371024.06 236777.83 333103.00
## release.name*              22 10000   3923.10  2258.13   3904.00
## similar*                   23 10000   1417.80   823.19   1402.00
## song.hotttnesss            24  5649      0.34     0.25      0.36
## song.id*                   25 10000   5000.50  2886.90   5000.50
## start_of_fade_out          26 10000    229.88   112.02    213.86
## tatums_confidence          27 10000      0.51     0.33      0.50
## tatums_start               28 10000      0.30     0.51      0.19
## tempo                      29 10000    122.90    35.20    120.16
## terms*                     30 10000    215.30   129.17    214.00
## terms_freq                 31 10000    224.89 22392.16      1.00
## time_signature             32 10000      3.56     1.27      4.00
## time_signature_confidence  33 10000      0.60     8.99      0.55
## title*                     34 10000   4865.28  2800.26   4861.50
## year                       35 10000    934.70   996.65      0.00
```

MILLION SONGS DATASET: PREDICTING SONG/TRACK POPULARITY

Goals:

- To Predict: whether an artist or song will be "popular" or not
- To Answer: What factors make an artist/song "popular"?
- Categorical labels: 3 or 5 levels of popularity




```
##Function to create descriptive statistics for artist hotness
descriptive_stats <- function(vector) {
library(moments)
  result <- c(
  Mean = mean(vector),
  Median = median(vector),
  Min = min(vector),
  Max = max(vector),
  SD = sd(vector),
  Quantile = quantile(vector, probs = c(0.25,.50, 0.75, 0.95)),
  Skewness = skewness(vector))
  print(result) }

descriptive_stats(music$artist.hotttnesss)

##Methodology for assigning artist hotness levels
#95% Quantile: 0.6011861 - Hot
#75% Quantile: 0.453858  - Warm
#50% Quantile: 0.3807423 - Tepid
#25% Quantile: 0.3252656 - Cool

##Code for assigning labels based on above quantiles
music$artist.hotness.label <- ifelse(
  music$artist.hotttnesss >= 0.6011861,
  "Hot", ifelse(
    music$artist.hotttnesss >= 0.453858 & music$artist.hotttnesss < 0.6011861,
    "Warm", ifelse(
      music$artist.hotttnesss >= 0.3807423 & music$artist.hotttnesss < 0.453858,
      "Tepid",  ifelse(
        music$artist.hotttnesss >= 0.3252656 & music$artist.hotttnesss < 0.3807423,
        "Cool", ifelse(
          music$artist.hotttnesss < 0.3252656,
          "Frigid", "Else")))))
```

METHODOLOGY

- Assign Categorical Artist & Song Popularity Levels:
  - 3 Levels
    - Hot (>.4590)
    - Warm (<.4590 and >.3357)
    - Cold (<.3357)
  - 5 Levels
    - Hot: >95% Percentile (~0.6012)
    - Warm: >75% Percentile (~0.4539)
    - Tepid: >50% Percentile (~0.3807)
    - Cool: >25% Percentile (~0.3253)
    - Frigid: <25% Percentile (~0.3253)

- Utilized two models to predict artist/song popularity:
  - Linear Regression
  - Random Forest

Results:

Model	Dataset	Accuracy (%)
Linear	Artist	64% (R^2)
RandomForest	Artist	80%
Linear	Song	6% (R^2)
RandomForest	Song	48%

Best Variables for Predicting Popularity:

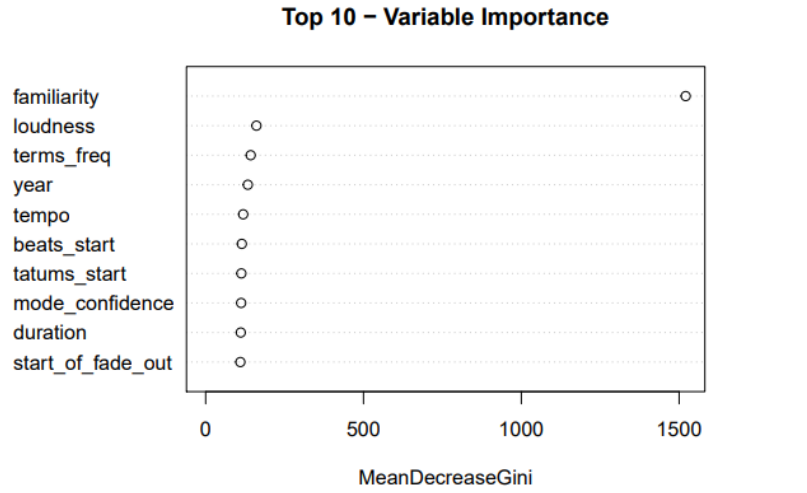
- Artists – Familiarity
- Songs – Loudness

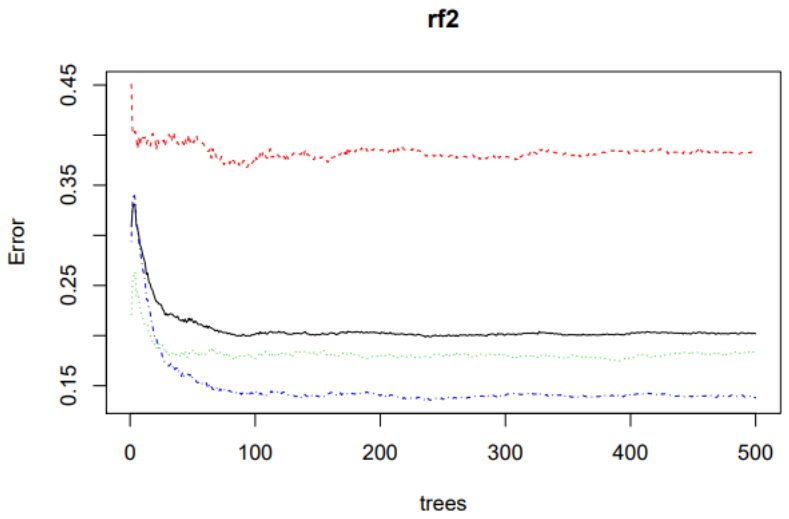
METHODS

Random Forest – Artist Popularity

```
Call:
 randomForest(x = music[, c(-1, -22, -23)], y = music[, 22])
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 20.18%
Confusion matrix:
     Cold  Hot Warm class.error
Cold  728    4  448   0.3830508
Hot     6 1289  284   0.1836605
Warm  200  198 2491   0.1377639
```



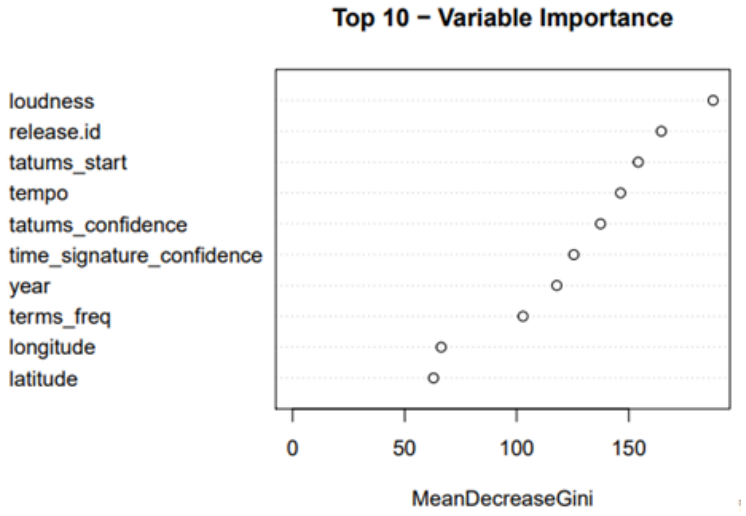


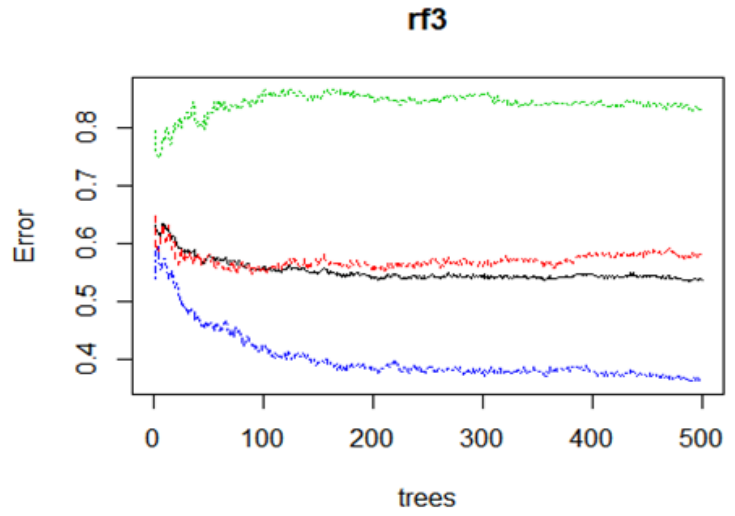
METHODS

Random Forest - Song Popularity

```
Call:
 randomForest(x = cmbomusic5[, -12:-13], y = cmbomusic5[, 13])
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 51.74%
Confusion matrix:
      Cold Hot Warm class.error
Cold   297  29  381   0.5799151
Hot     76 118  246   0.7318182
Warm   244  78  568   0.3617978
```





METHODS

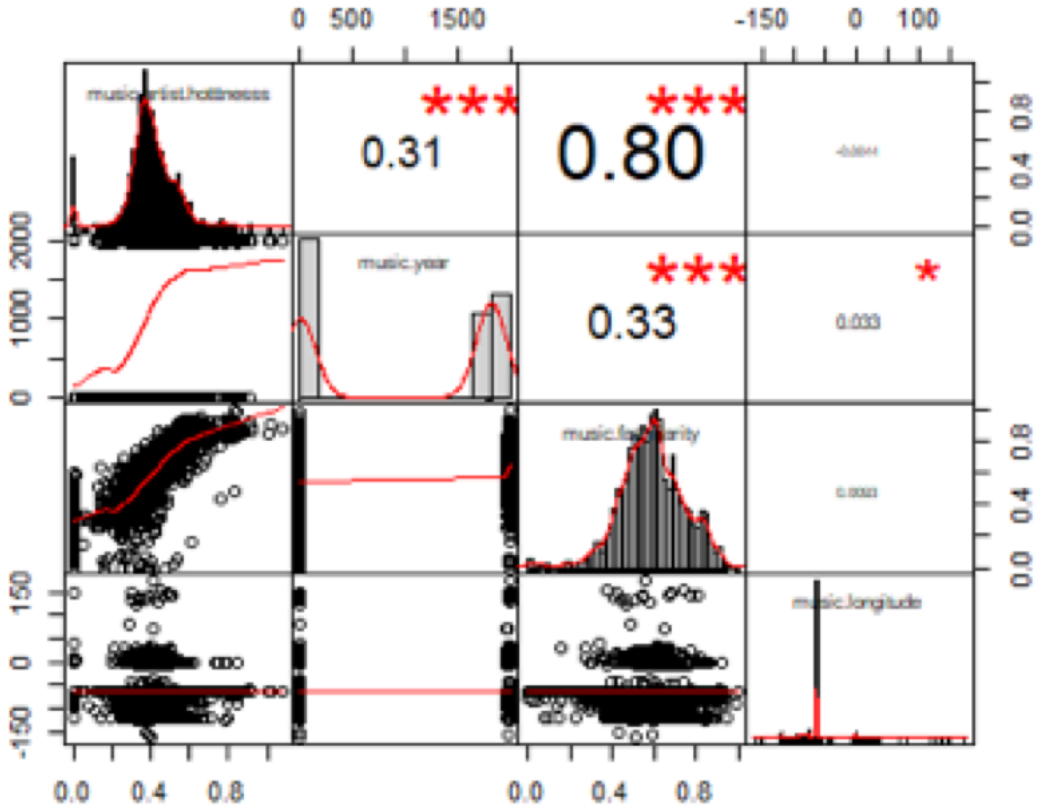
Regression and Correlation

```
Call:
lm(formula = music$artist.hotttnesss ~ music$year + music$bars_confidence +
    music$tempo + music$duration + music$start_of_fade_out +
    music$tatums_start + music$familiarity + music$latitude +
    music$tempo + music$longitude + music$beats_start + music$beats_confidence
+
    music$end_of_fade_in)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41865 -0.03239 -0.00136  0.03219  0.50014

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.500e-02  8.045e-03   1.865   0.0622 .
music$year               6.911e-06  1.081e-06   6.392 1.77e-10 ***
music$bars_confidence   -4.242e-04  3.754e-03  -0.113   0.9100
music$tempo             -3.122e-05  3.082e-05  -1.013   0.3111
music$duration           1.842e-05  1.881e-05   0.979   0.3276
music$start_of_fade_out -2.842e-05  2.121e-05  -1.340   0.1803
music$tatums_start      -5.004e-03  7.041e-03  -0.711   0.4773
music$familiarity        6.625e-01  7.156e-03  92.582   < 2e-16 ***
music$latitude          -1.039e-04  1.006e-04  -1.033   0.3015
music$longitude         -5.606e-05  3.190e-05  -1.758   0.0789 .
music$beats_start        5.494e-03  6.748e-03   0.814   0.4155
music$beats_confidence  -2.277e-03  3.227e-03  -0.706   0.4804
music$end_of_fade_in     9.355e-05  6.367e-04   0.147   0.8832
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07602 on 5635 degrees of freedom
Multiple R-squared:  0.6443,    Adjusted R-squared:  0.6436
F-statistic: 850.6 on 12 and 5635 DF,  p-value: < 2.2e-16
```



LESSONS LEARNED

LESSONS LEARNED

Data Munging really does take 70-80% of the time for a project!

N/A and Zero variables will cause more problems if not handled appropriately

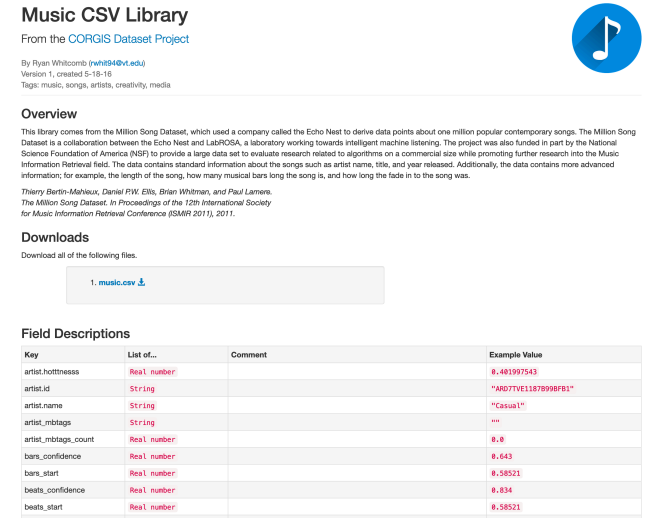
As the longitude increased, the hotness of the artist would be smaller

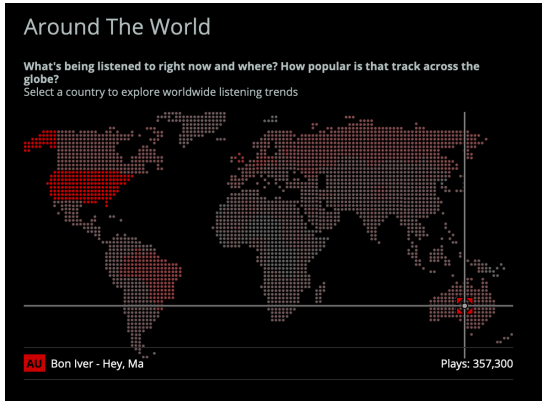
A data that has relative song popularity per year may be more dependable when using year to predict the popularity

Commenting your code is incredibly important!

REFERENCES

- Virginia Tech - https://think.cs.vt.edu/corgis/csv/music/music.html
- Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.
- https://www.last.fm/api
  - The Last.fm API allows anyone to build their own programs using Last.fm data





RECOMMENDED MUSIC

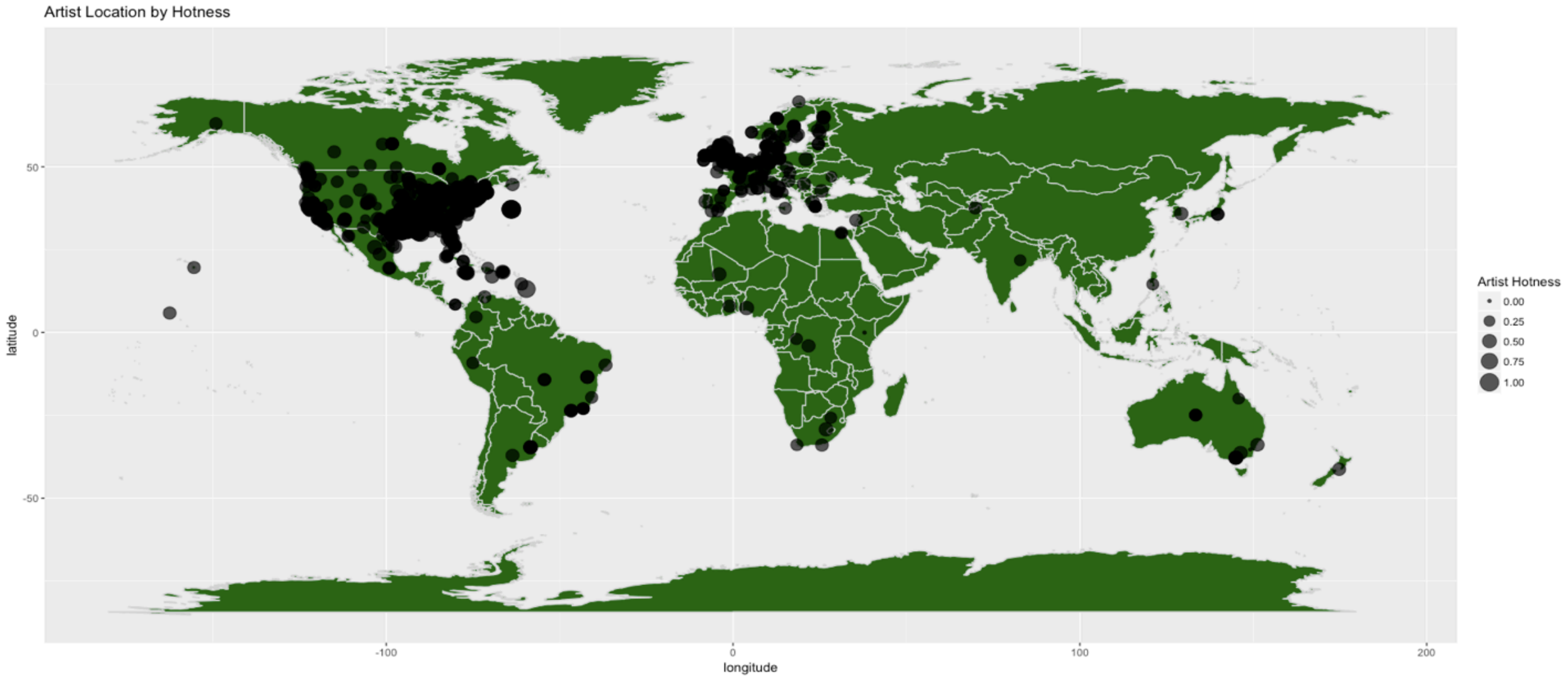
- John – Claire Ernst, John The Blind, Maggie Rogers, Tom Misch, Tyler Childers
- Juan – Sia, Bob Dylan, Shakira, Carlos Vives, Merengue(for dancing)
- Courtney – Old Dominion, Dierks Bentley, Chase Rice, Brothers Osborne
- Jeremy – Dierks Bentley, Ed Sheeran, Fall Out Boy, Eminem, Cary Brothers

Note: No algorithms were harmed in the development of this list. This is simply music we like. Enjoy!

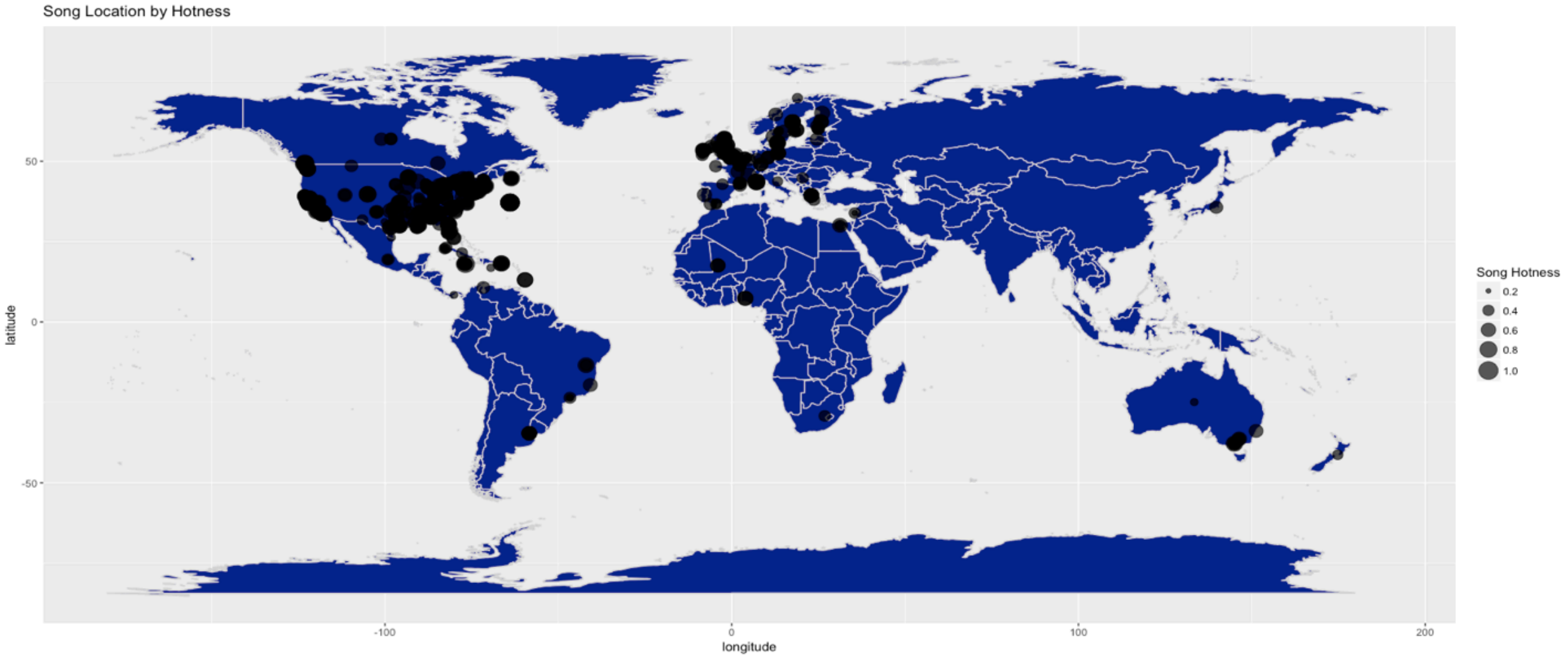

Music is forever; music should grow and mature with you,
following you right on up until you
~Paul Simon
Music is an outburst of the soul. ~ Frederick Delius

APPENDIX

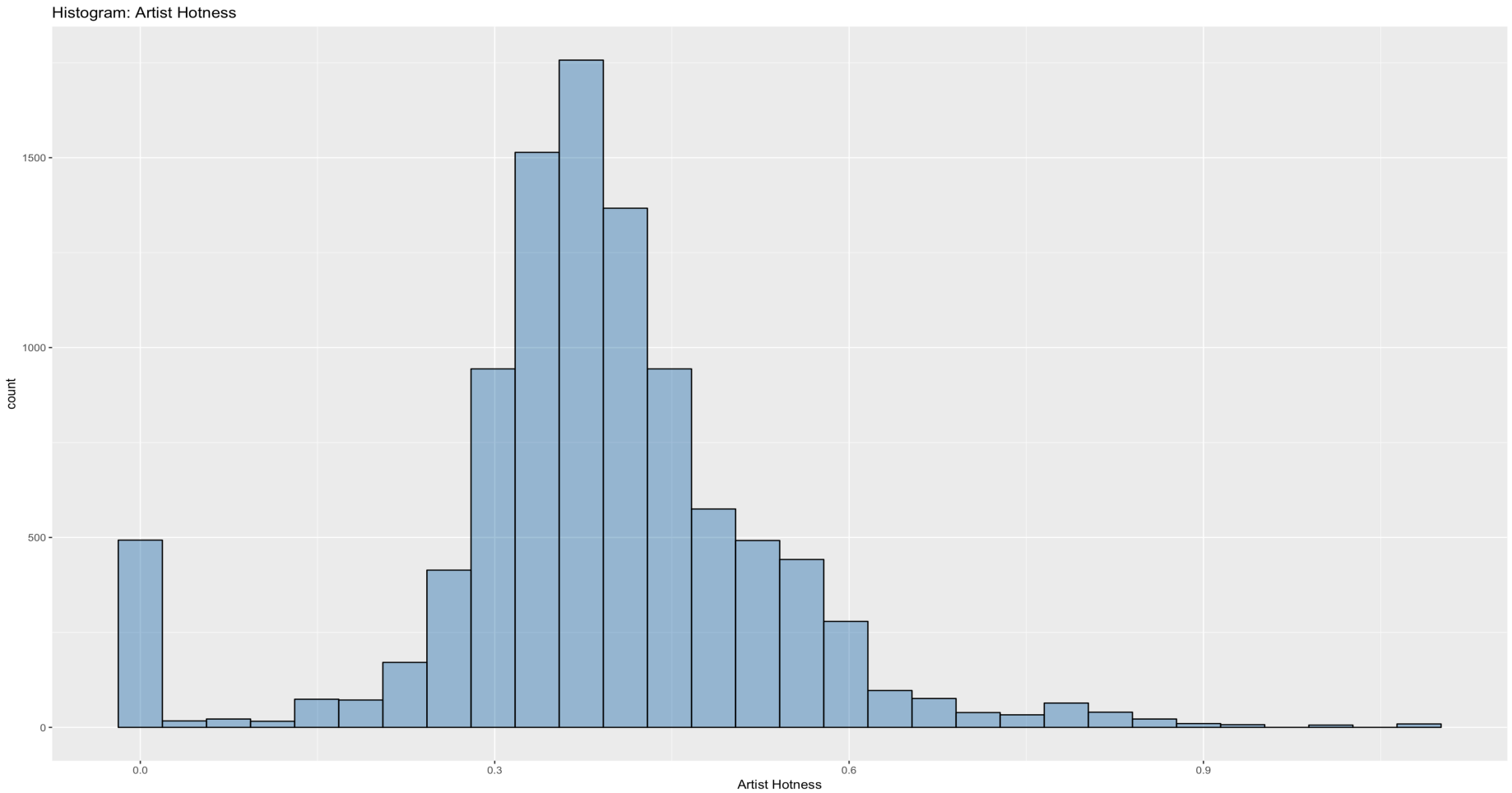
ARTIST LOCATION AND HOTNESS



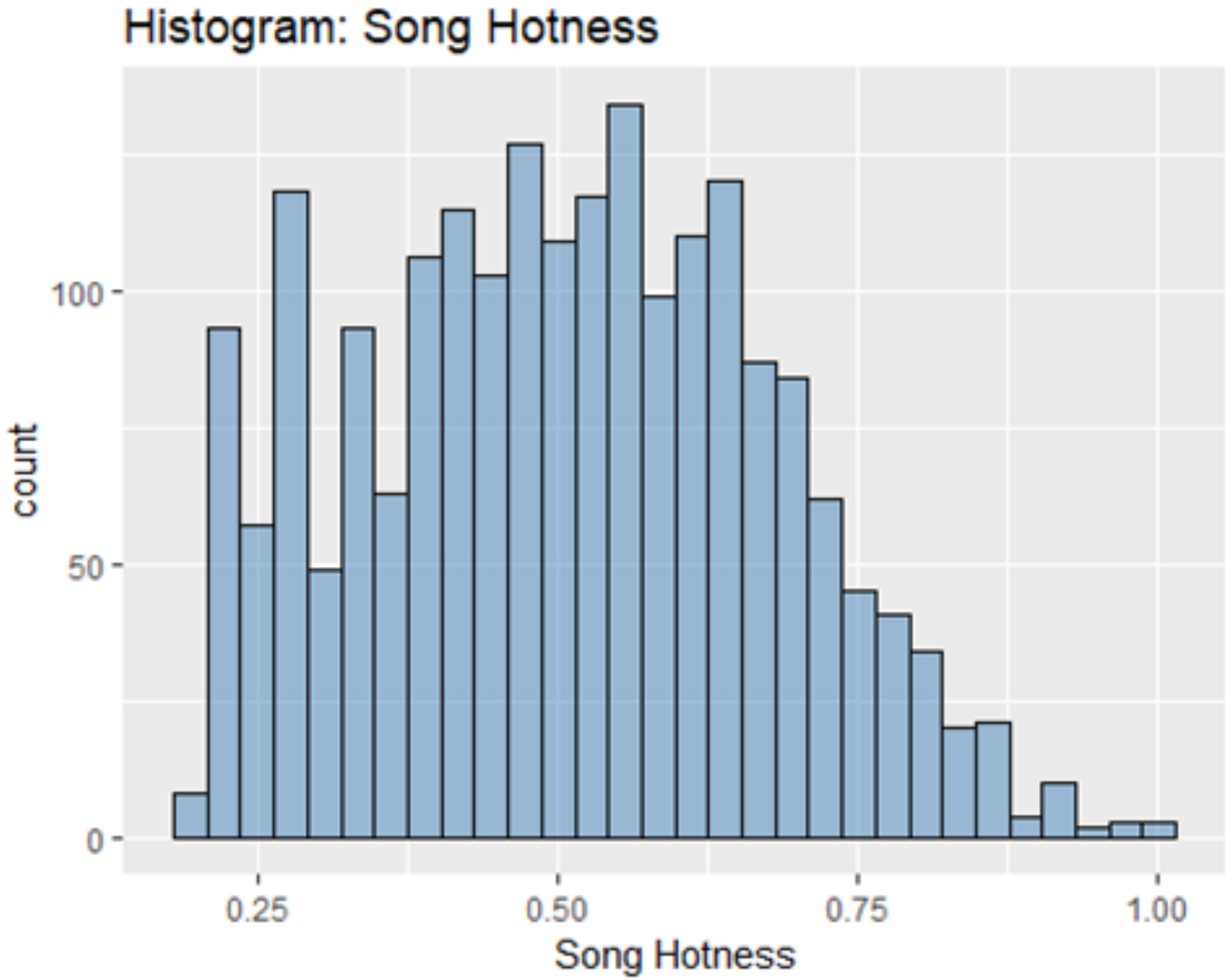
SONG LOCATION AND HOTNESS



ARTIST HOTNESS



SONG HOTNESS



ARTIST HOTNESS
HOT – WARM - COLD

Call:
randomForest(x = music[, -21:-22], y = music[, 21])
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 4

OOB estimate of error rate: 20.18%
Confusion matrix:

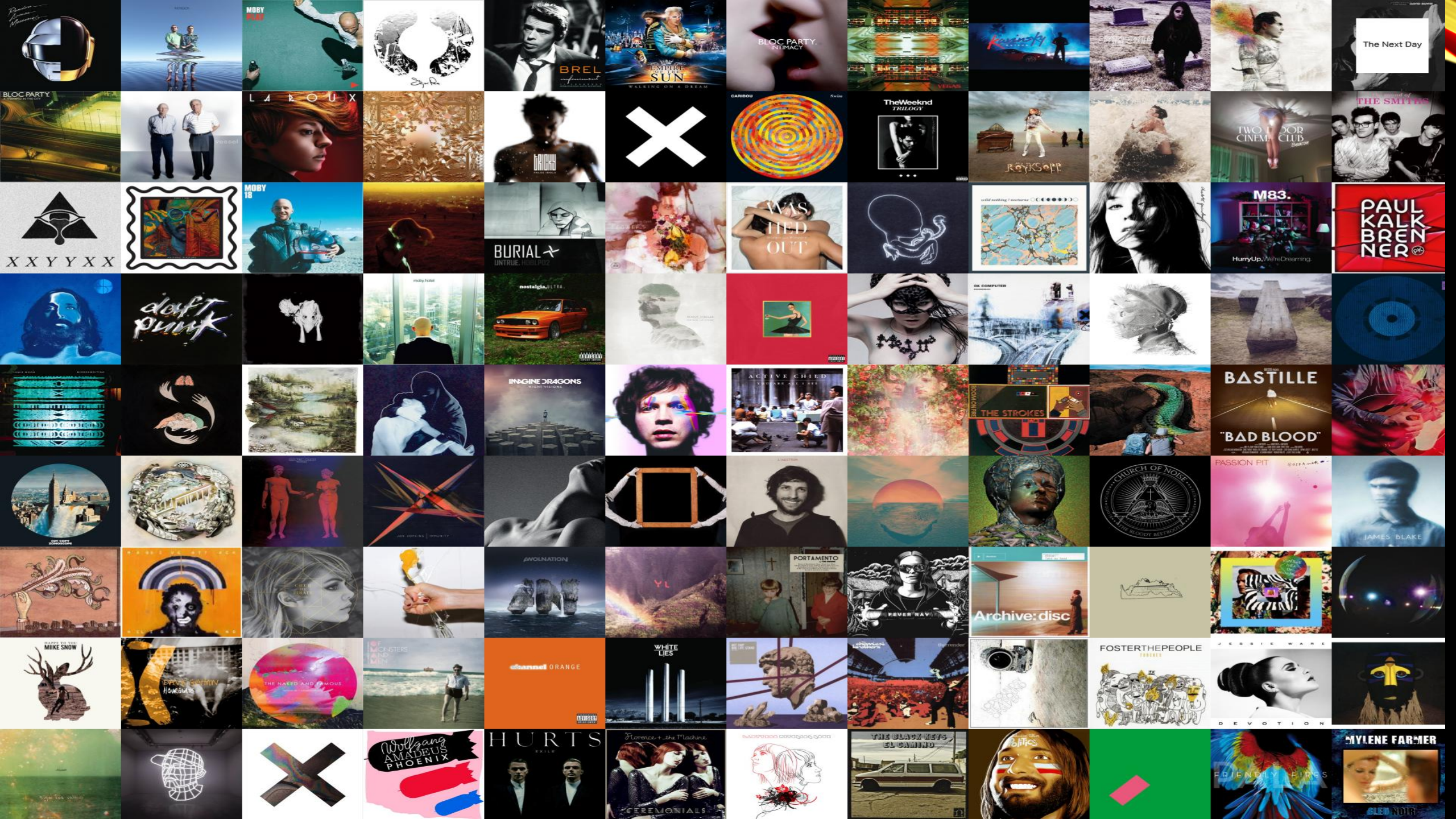
	COLD	HOT	WARM	class.error
COLD	728	4	448	0.3830508
HOT	6	1298	284	0.1836605
WARM	200	198	2491	0.1377639

SONG HOTNESS
HOT – WARM - COLD

Call:
randomForest(cmbomusic5[,-13:-14],cmbomusic5[,14])
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 3

OOB estimate of error rate: 46.49%
Confusion matrix:

	COLD	HOT	WARM	class.error
COLD	392	22	293	0.3830508
HOT	44	169	233	0.1836605
WARM	248	107	535	0.1377639


The Next Day
BLOC PARTY.
INTIMACY
BREL
MOBY
LA ROUX
THE SMITHS
TheWeeknd
TRILOGY
RÖYKSOPP
M83.
HurryUp,WereDreaming.
PAUL KALKBRENNER
XXYYXX
BURIAL
daft punk
IMAGINE DRAGONS
NIGHT VISIONS
ACTIVE CHILD
THE STROKES
BASTILLE
"BAD BLOOD"
PASSION PIT
JAMES BLAKE
CHURCH OF NOISE
AWOLNATION
PORTAMENTO
FEVER RAY
Archive:disc
MIIKE SNOW
THE NAKED AND FAMOUS
channel ORANGE
WHITE LIES
FOSTERTHEPEOPLE
DEVOTION
Wolfgang AMADEUS PHOENIX
HURTS
Florence + the Machine
CEREMONIALS
THE BLACK KEYS
EL CAMINO
FRIENDLY FIRES
MYLENE FARMER