Introduction to Data Science – Final Project

Spotify Data Analysis

Julian Schmocker

INTRODUCTION

Outlook

Research Questions

What were the questions asked?

Data Sources

How was the data acquired?

Dataset A

Differences between chart songs and random songs

Dataset B

Popularity on Spotify

RESEARCH QUESTIONS

Research Questions

1. What are the main differences between songs that were Number 1 in the UK charts and random songs?

2. What is the best method to classify the two groups?

3. Which variables have the largest influence on the popularity (on Spotify) of a song?

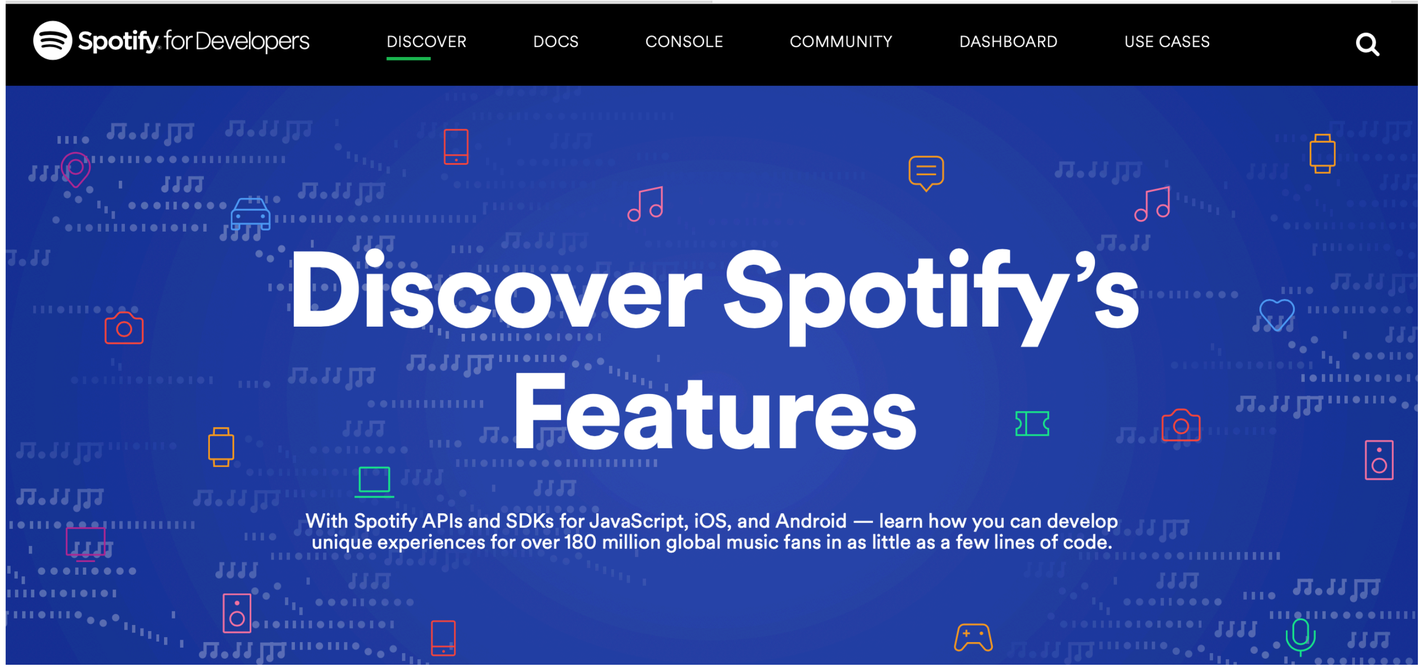
4. What is the best method to predict a song’s popularity?

What properties does a song need to have to be *popular*?

DATA SOURCE

Spotify API

The data used in this project was acquired from the Spotify API. I worked with the library *Spotipy* to access the API.



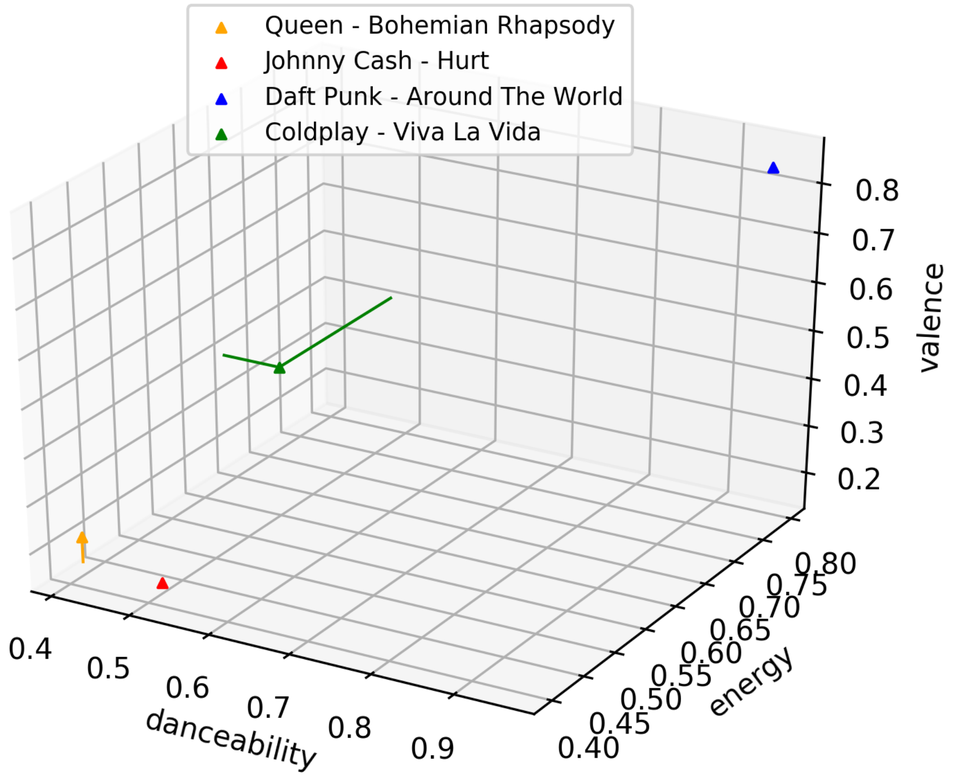
DATA SOURCE

Audio Features

Variable	Value type	Description
duration	int	Duration of the track in milliseconds
key	int	Estimated overall key of the track. C = 0, C# = 1, D = 2
mode	int	Modality of the track. Major = 1, Minor = 0
acousticness	float	A confidence measure from 0.0 to 1.0 of weather the track is acoustic.
danceability	float	Describes how suitable a track is for dancing (0.0 to 1.0).
energy	float	Measures the intensity and activity (0.0 to 1.0).
instrumentalness	float	Predicts whether a track contains no vocals (0.0 to 1.0).
liveness	float	Detects the presence of an audience in the recording (0.0 to 1.0).
loudness	float	Overall loudness of a track in decibels.
speechiness	float	Detects the presence of spoken words (0.0 to 1.0).
valence	float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (happy, cheerful, euphoric).
tempo	float	The overall estimated tempo of a track in beats per minute.

DATA SOURCE

Example

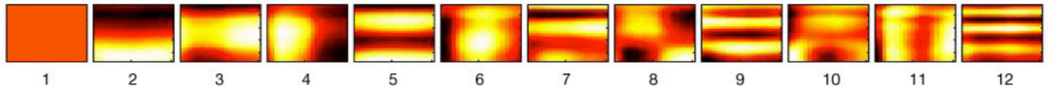


track	valence	energy	danceability	mode	key
Queen – Bohemian Rhapsody	0.224	0.404	0.414	0	0
Johnny Cash – Hurt	0.171	0.401	0.518	0	9
Daft Punk – Around The World	0.841	0.795	0.956	1 (0)	7 (4)
Coldplay – Viva La Vida	0.416	0.619	0.485	0	5

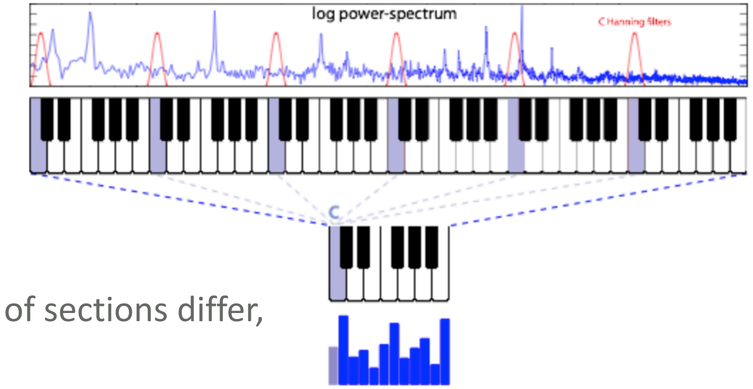
DATA SOURCE

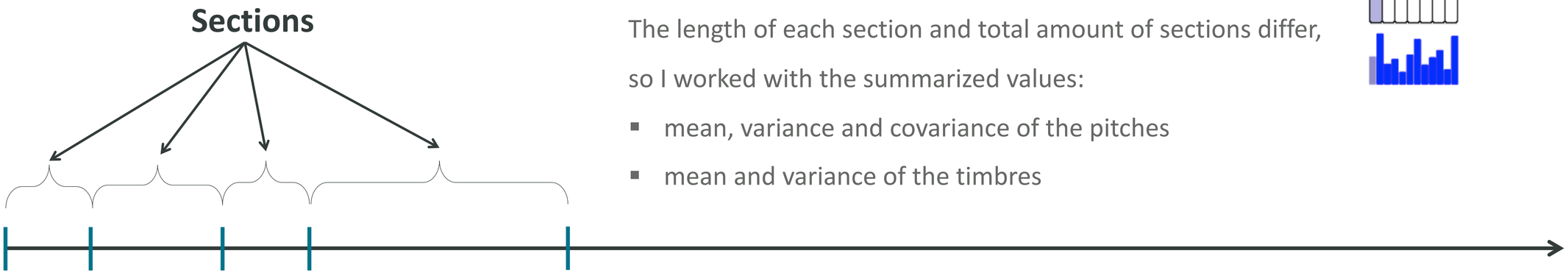
Audio Analysis - Sections

Variable	Value type	Description
pitch	float vector	Pitch content is given by a vector, corresponding to the 12 pitch classes, indicating the relative dominance in every pitch in the chromatic scale.
timbre	float vector	Timbre is the quality of a musical note or sound that distinguishes different types of musical instruments, or voices. It is a complex notion also referred to as sound color, texture, or tone quality.



12 basis functions for the timbre vector: x = time, y = frequency, z = amplitude





The length of each section and total amount of sections differ,

so I worked with the summarized values:

- mean, variance and covariance of the pitches
- mean and variance of the timbres

DATA SOURCE

Datasets

Dataset A

Variables:	All Audio Features, summarized Audio Analysis, year of release and chart_random
Focus on:	chart_random (indicates to which group a song belongs)

Chart songs from playlist “Every Official UK Number 1 Ever”

Dataset B

Variables:	All Audio Features, summarized Audio Analysis, popularity and year of release
Focus on:	popularity (depends on the total number of plays the track has had on Spotify and how recent those plays are)

DATASET A

1’300 Number 1 songs
1’300 random songs

DATASET B

5’000 random songs
5’000 recommended songs

INTRODUCTION

Outlook

Research Questions

What were the question asked?

Data Sources

How was the data acquired?

Dataset A

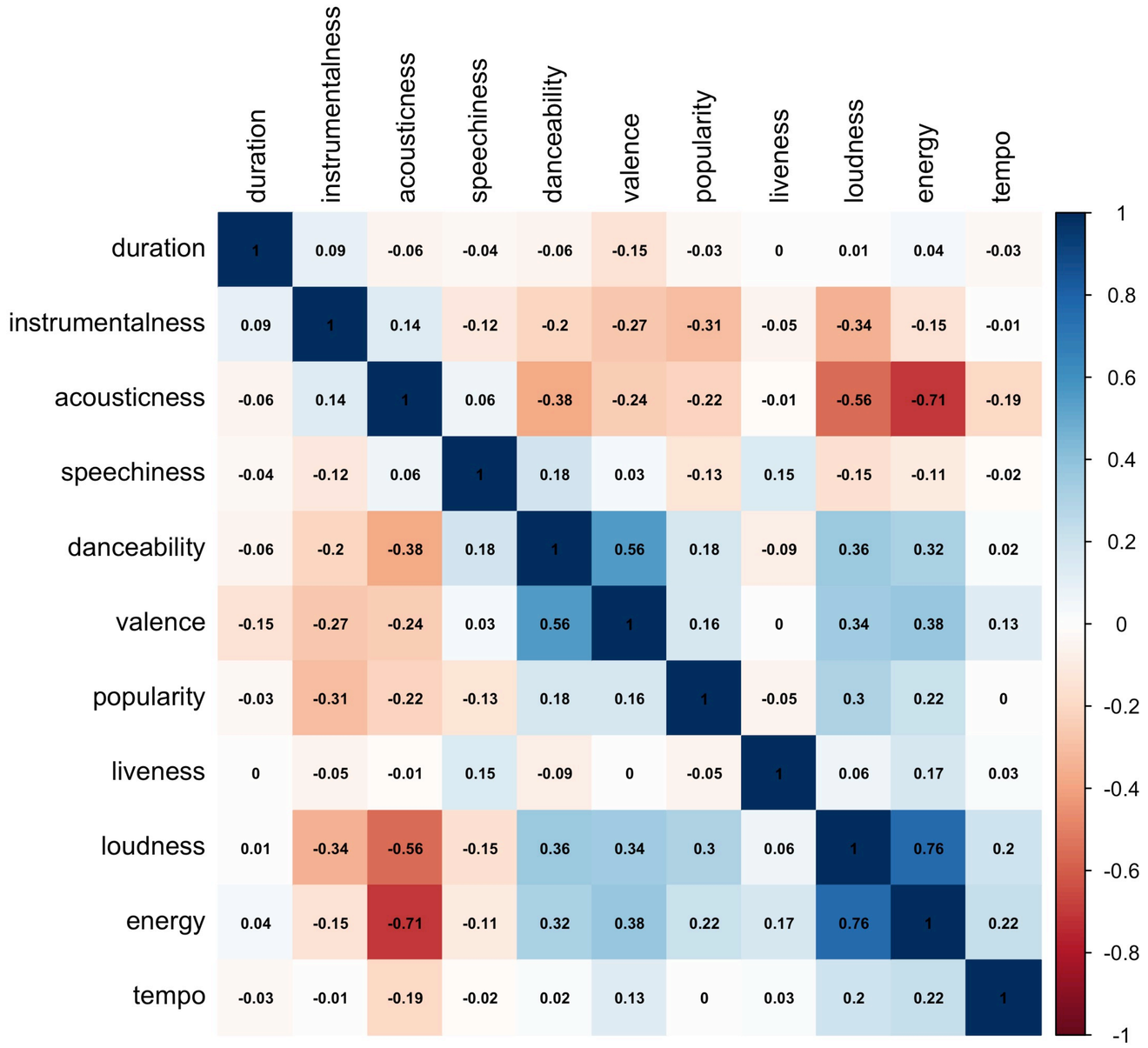
Differences between chart songs and random songs

Dataset B

Popularity on Spotify

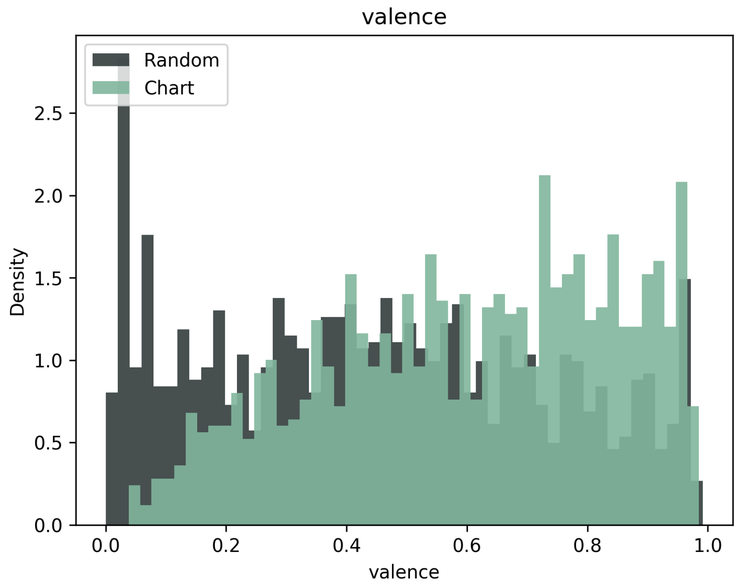
EXPLANATORY DATA ANALYSIS

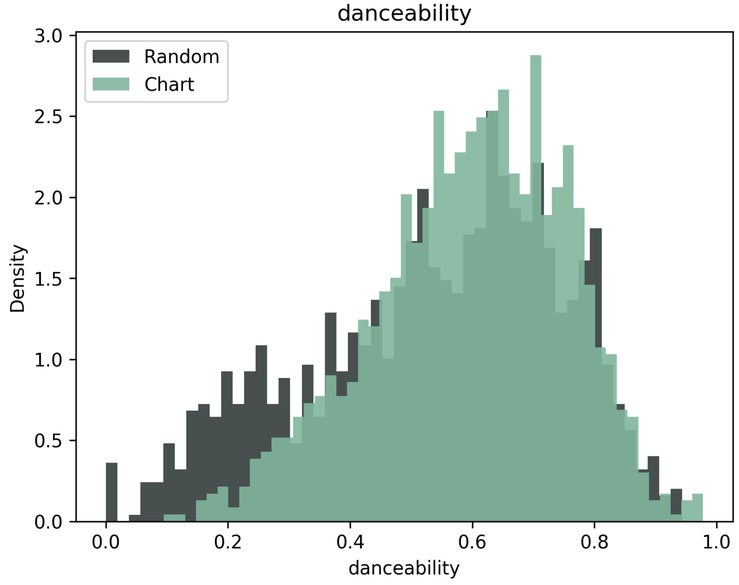
Correlations

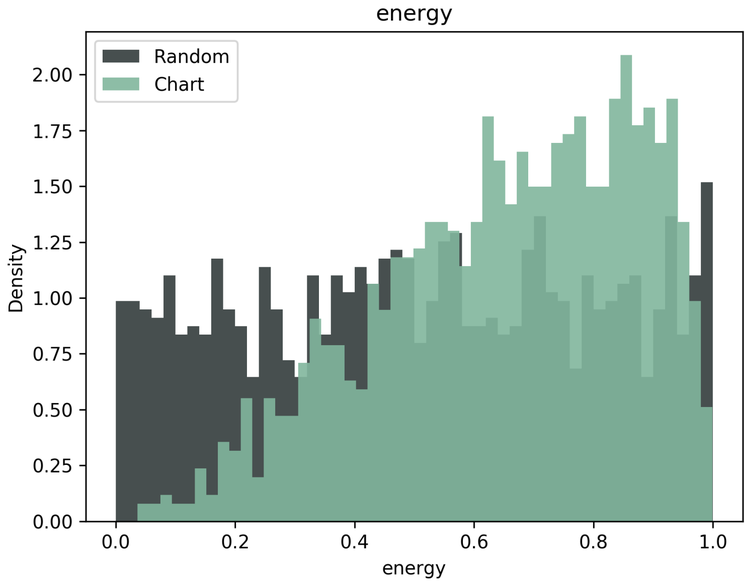


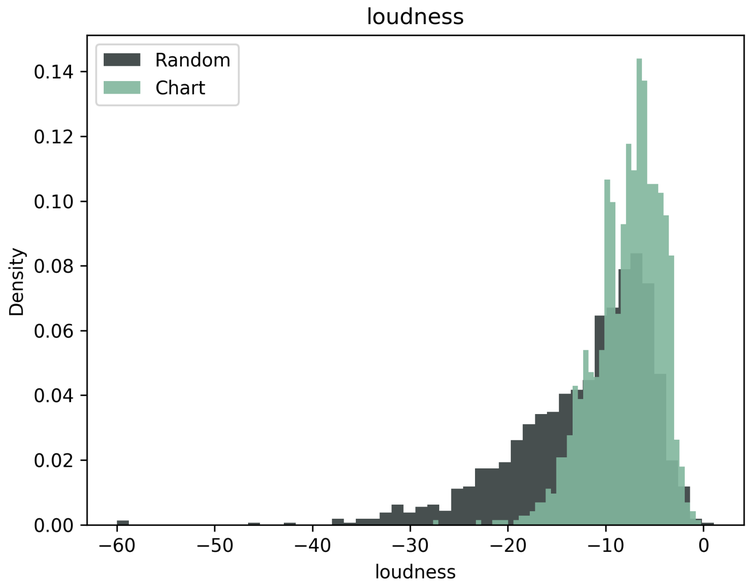
EXPLANATORY DATA ANALYSIS

Histograms



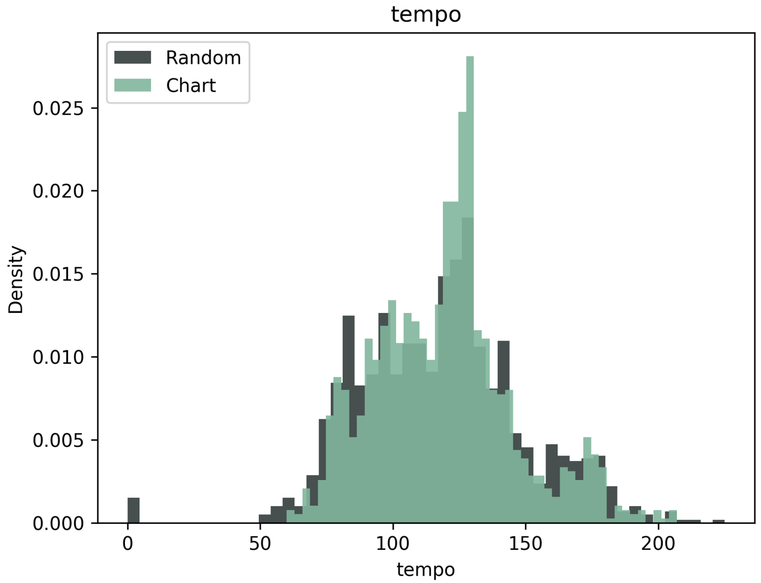


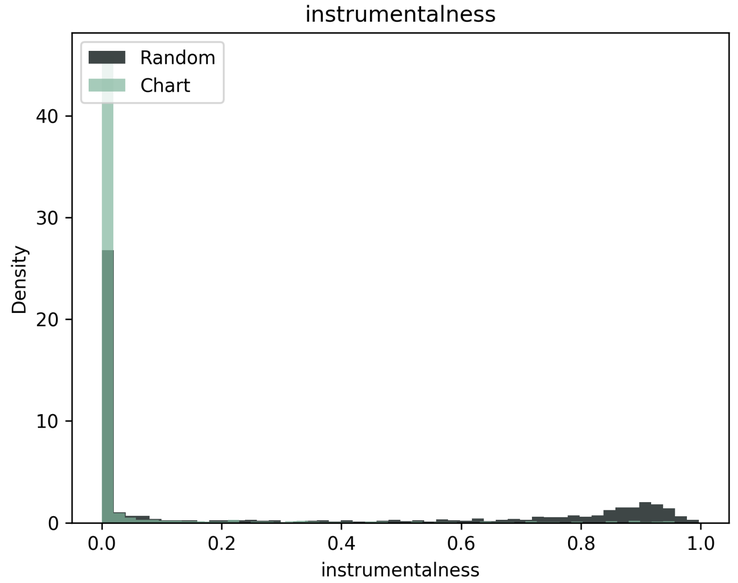


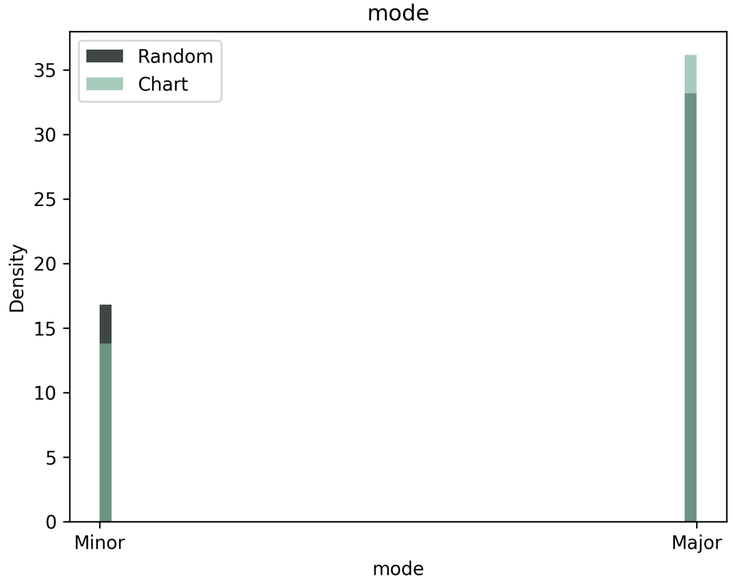


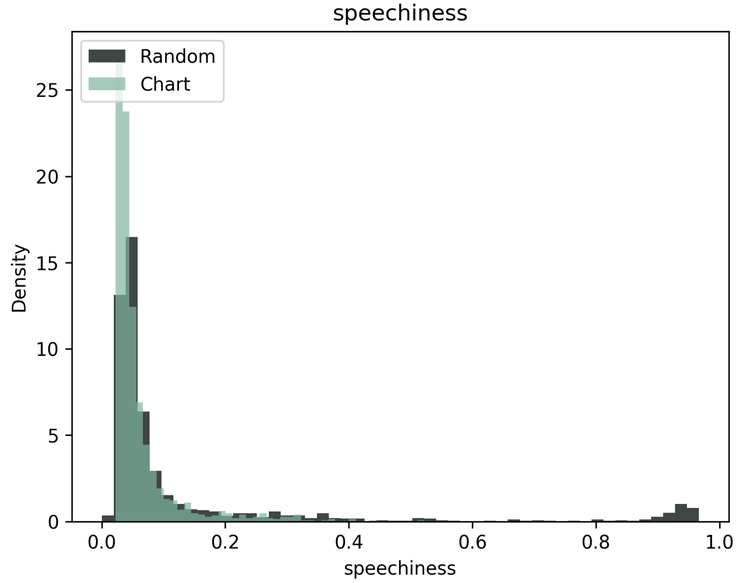
EXPLANATORY DATA ANALYSIS

Histograms









CLASSIFICATION

Logistic Regression

Audio Features

```
Call:
glm(formula = chart_random ~ acousticness + instrumentalness +
    loudness + speechiness + valence, family = binomial(link = "logit"),
    data = dataA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1187  -0.8319   0.4779   0.8437   2.9274

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.29103    0.16890   7.644 2.11e-14 ***
acousticness     -0.59557    0.17363  -3.430 0.000603 ***
instrumentalness -3.32072    0.21955 -15.125  < 2e-16 ***
loudness          0.09232    0.01295   7.132 9.93e-13 ***
speechiness      -6.67797    0.64554 -10.345  < 2e-16 ***
valence           1.43006    0.18788   7.612 2.71e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3652.9  on 2634  degrees of freedom
Residual deviance: 2663.9  on 2629  degrees of freedom
AIC: 2675.9

Number of Fisher Scoring iterations: 6
```

The other variables are not significant (tested with Chisq-test, see example below)

```
Analysis of Deviance Table

Model 1: chart_random ~ acousticness + instrumentalness + loudness + speechiness +
    valence
Model 2: chart_random ~ acousticness + instrumentalness + loudness + danceability +
    speechiness + valence
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2629     2663.9
2      2628     2663.8  1 0.030245   0.8619
```

CLASSIFICATION

Logistic Regression

$$\hat{odds} = e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\ldots+\hat{\beta}_p X_p} = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 X_1} \cdot \ldots \cdot e^{\hat{\beta}_p X_p}$$

```
exp(log_reg$coefficients[c(6,4,2,3,5)])
      valence          loudness      acousticness instrumentalness       speechiness
4.178944609       1.096716484       0.551247143       0.036126698       0.001258328
```

+1

The estimated odds are multiplied by the above factor if only the corresponding variable is increased by 1.

Chart Songs

Random Songs

not significant
danceability
energy
liveness
key
mode
tempo
duration

CLASSIFICATION

Best Classifier – Audio Features

The goal was also to find the classifier, which can best classify the the two groups. For this I used 10-fold-cross-validation.

To compare the models I looked at the fraction of incorrect classifications (averaged over all folds): $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$

Included variables: Audio Features (16 variables)

Model	Percentage of incorrect classification
K-nearest-neighbors (Python, sklearn), k = 6	23.60 %
Logistic Regression (R, glm)	20.95 %
Boosting (R, adabag), mfinal = 11	19.74 %
Bagging (R, adabag), mfinal = 11	19.66 %
Neural Network (Python, keras), 3 layers {(32, activation: relu), (16, activation = relu), (2, activation = softmax)}	19.43 %
Support Vector Machine (Python, sklearn)	**19.13 %**

CLASSIFICATION

Best Classifier – Audio Analysis

Included variables: Audio Analysis (summarized values of the pitches and timbres)

Model	Percentage of incorrect classification
Logistic Regression (R)	27.43 %
Boosting (R, adabag), mfinal = 11	27.09 %
Support Vector Machine (Python, sklearn)	**24.33 %**

CLASSIFICATION

Best Classifier – Full Dataset

Included variables: Full dataset (127 variables)

Model	Percentage of incorrect classification
Logistic Regression (R)	20.08 %
Boosting (R, adabag), mfinal = 11	18.92 %
Support Vector Machine (Python, sklearn)	**18.73 %**

81.27 %

This is the percentage of songs that were classified correctly with the Support Vector Machine on the full dataset.

Outlook

Research Questions

What were the question asked?

Dataset A

Find main differences between chart and random songs.

Data Sources

How was the data acquired?

Dataset B

Popularity on Spotify

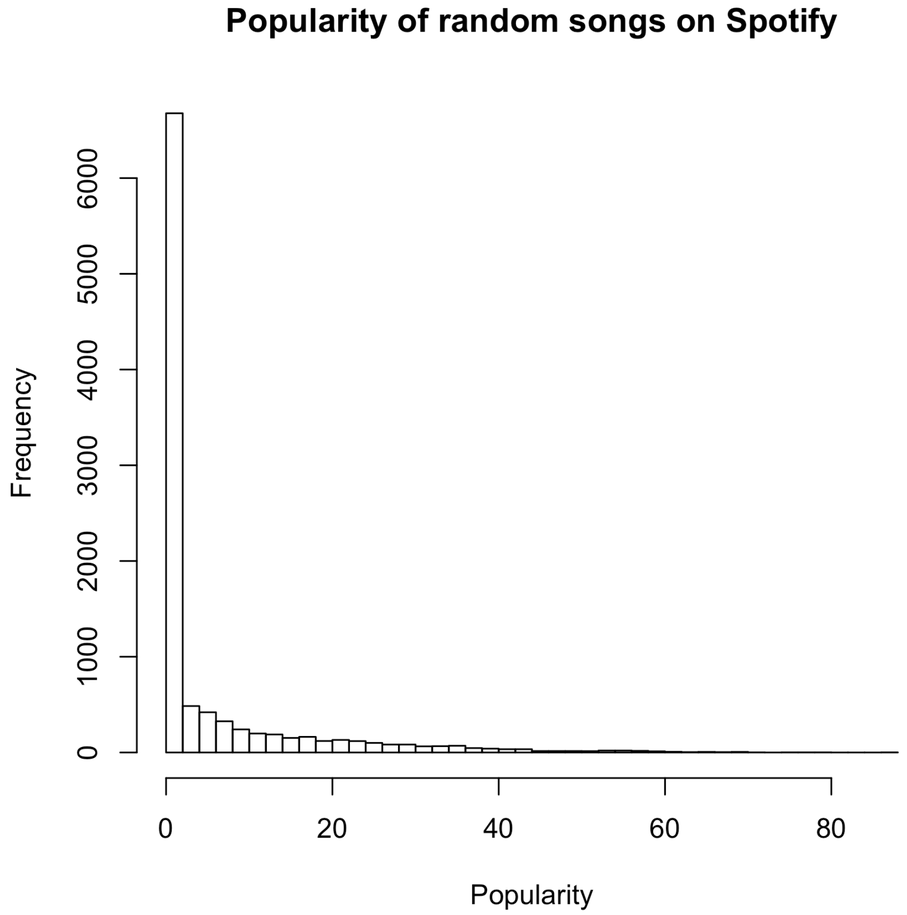
EXPLANATORY DATA ANALYSIS

Random Songs

It was not possible to pick completely random songs from the Spotify catalogue. So I first generated a random word (from a dictionary). Then I searched for tracks which have this word in their name. One of these tracks got chosen by a random offset.

Summary of Artist

```
                     artist
Johann Sebastian Bach  :  32
Various Artists        :  28
George Frideric Handel :  23
Giuseppe Verdi         :  21
Wolfgang Amadeus Mozart:  21
Claude Debussy         :  18
(Other)                :9857
```



Problems:

- Distribution of the variable *popularity*
  (~70% of the tracks have a *popularity* of 0)
- Randomness
  Most of the artists are classical composers. The problem is that rarely used words are picked with the same probability as often used word.

EXPLANATORY DATA ANALYSIS

Recommended Songs

I tried also an other approach to get tracks for my analysis. I got recommendation for different genres.

Here the distribution of the popularity is not problematic anymore.

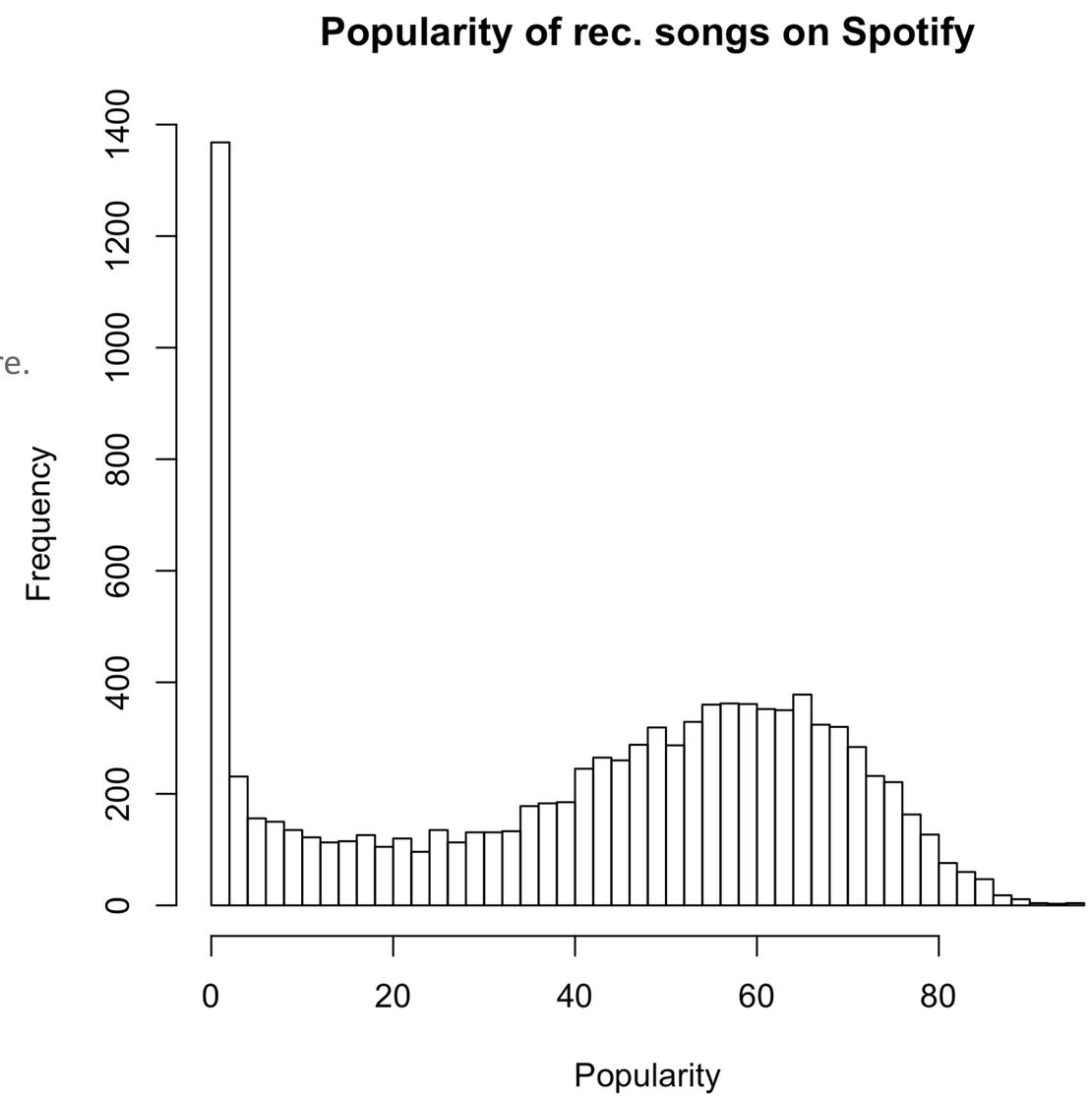
But they are of course not randomly chosen.

So I combined the random and the recommended song for my analysis.

Dataset B

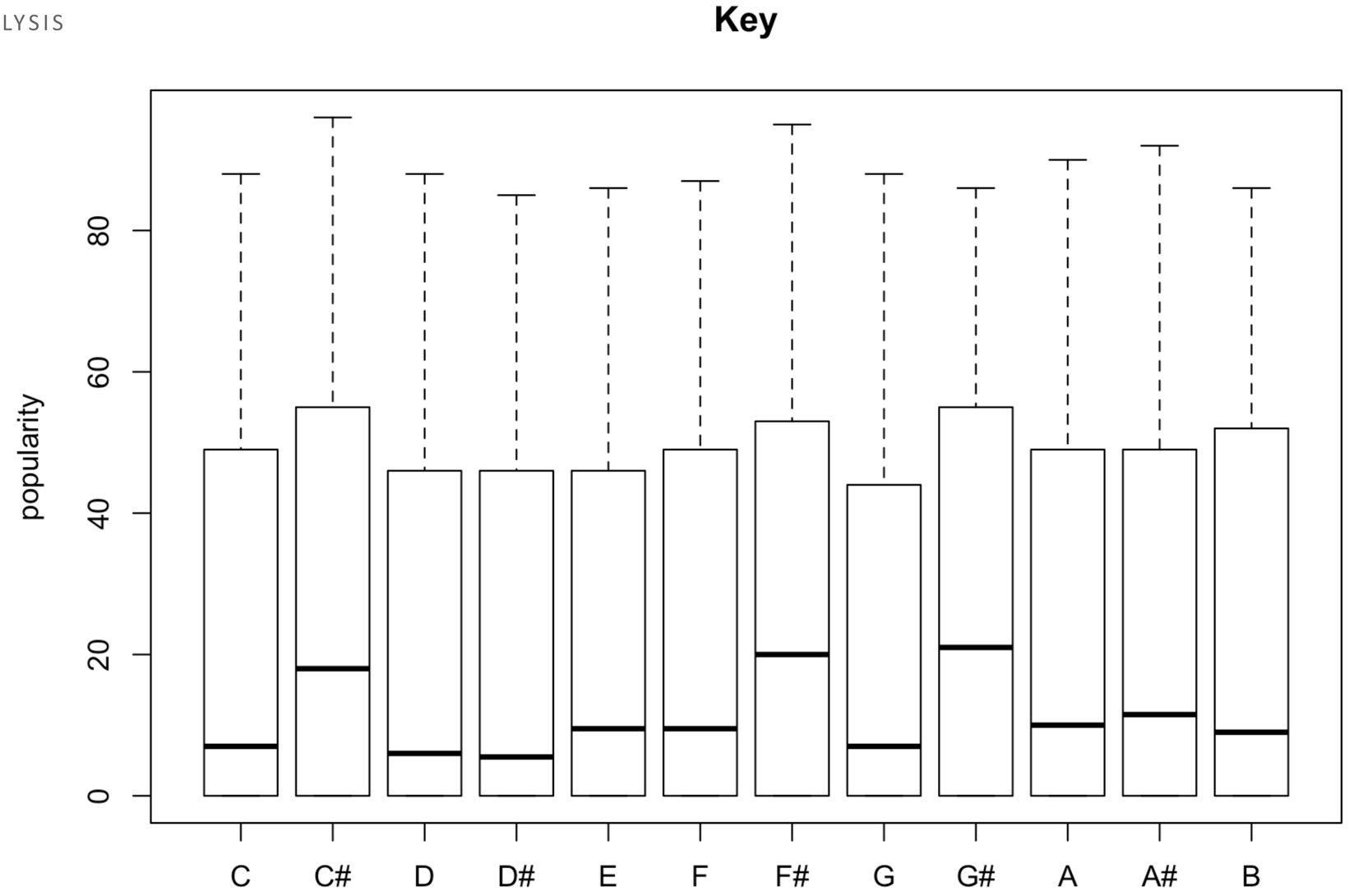
5’000 random songs

5’000 recommended songs



EXPLANATORY DATA ANALYSIS

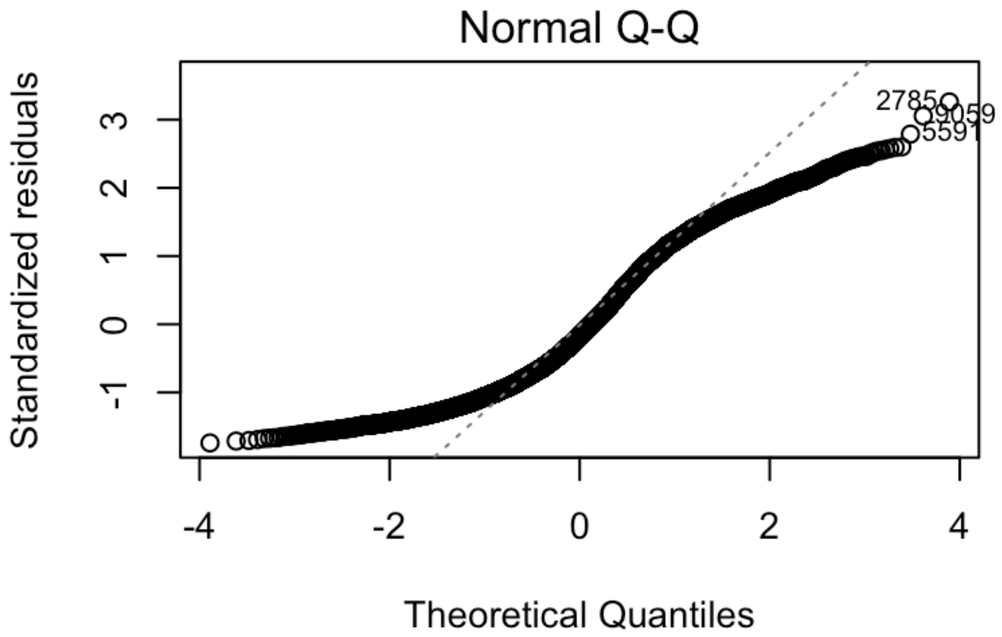
Boxplot



PREDICTION

Linear Regression

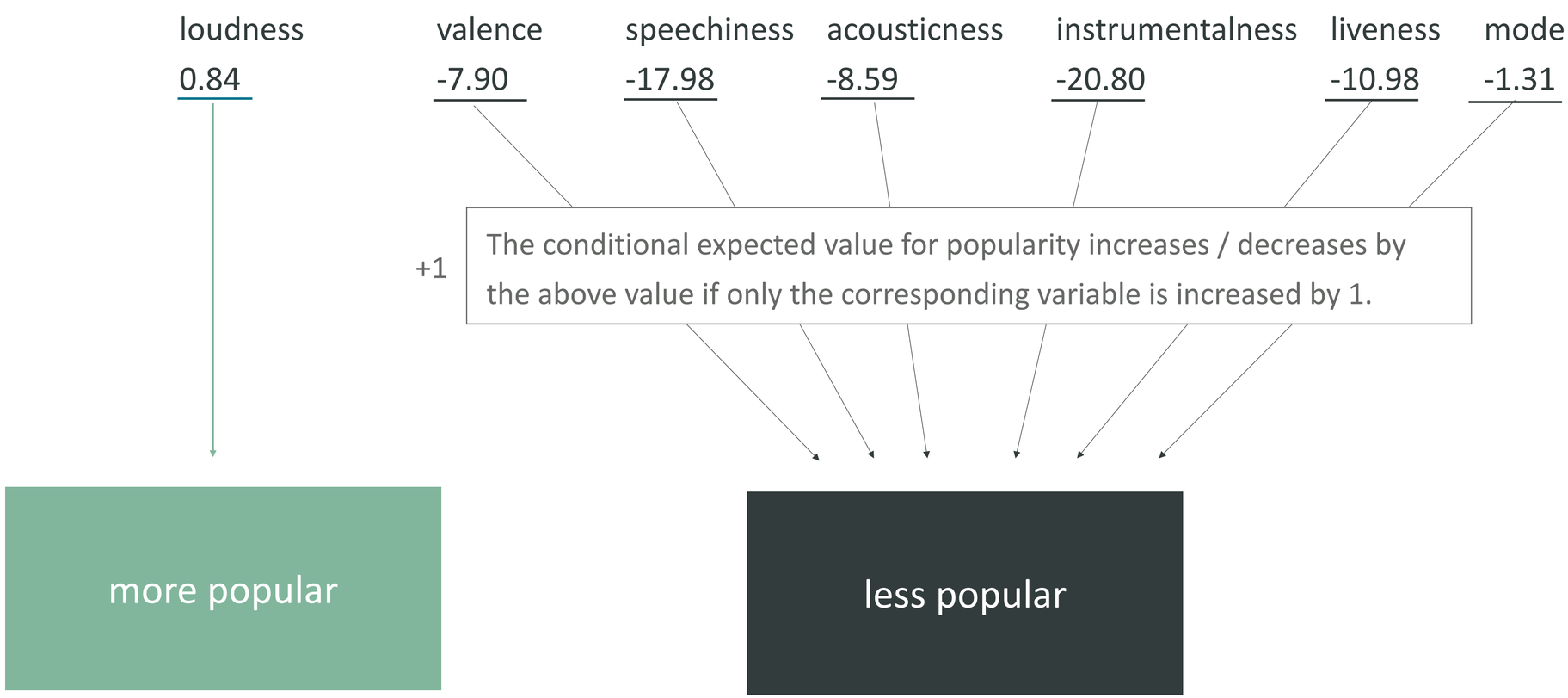
Audio Features




```
Call:
lm(formula = popularity ~ valence + loudness + acousticness +
    instrumentalness + speechiness + liveness + mode + tempo +
    duration + key, data = data_B)

Residuals:
   Min     1Q Median     3Q    Max
-42.42 -21.13  -4.14  20.49  79.46

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.006e+01  1.505e+00  33.275  < 2e-16 ***
valence          -7.899e+00  1.006e+00  -7.853 4.46e-15 ***
loudness          8.433e-01  5.755e-02  14.654  < 2e-16 ***
acousticness     -8.592e+00  8.951e-01  -9.599  < 2e-16 ***
instrumentalness -2.080e+01  8.007e-01 -25.977  < 2e-16 ***
speechiness      -1.798e+01  1.622e+00 -11.084  < 2e-16 ***
liveness         -1.098e+01  1.460e+00  -7.519 5.99e-14 ***
mode             -1.308e+00  5.187e-01  -2.522   0.0117 *
tempo            -1.198e-02  8.449e-03  -1.418   0.1562
duration          2.264e-07  1.631e-06   0.139   0.8896
key              -6.974e-02  6.924e-02  -1.007   0.3139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.39 on 9989 degrees of freedom
Multiple R-squared:  0.1859,    Adjusted R-squared:  0.1851
F-statistic:   228 on 10 and 9989 DF,  p-value: < 2.2e-16
```

PREDICTION

Linear Regression



not significant
key
tempo
duration

not included (to prevent multicollinearity)
energy
danceability

PREDICTION

Best Model for Prediction

The goal was also to find the best model to predict a song's popularity. For this I used 10-fold-cross-validation.

To compare the models I use the following errors (averaged over all folds):

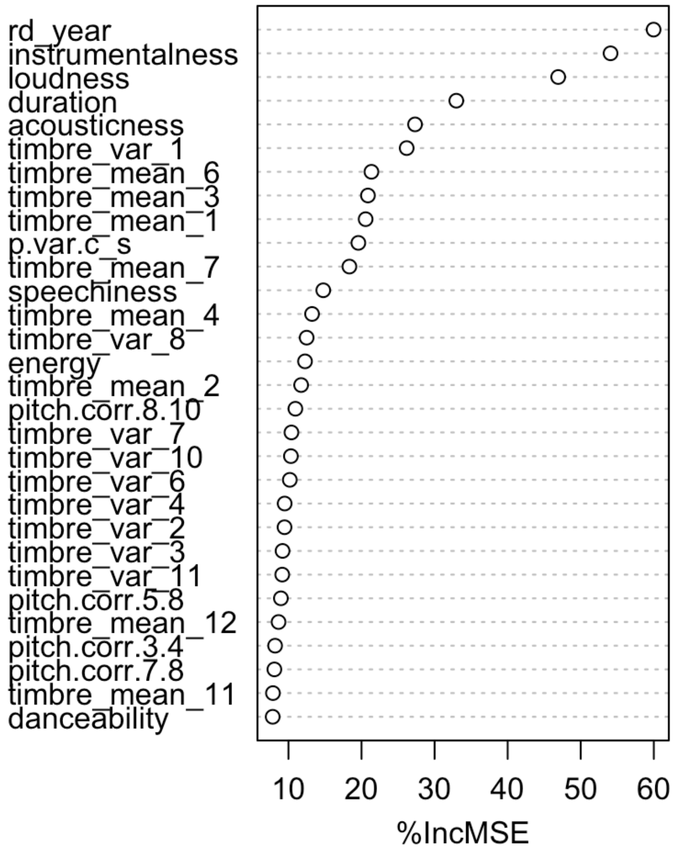
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

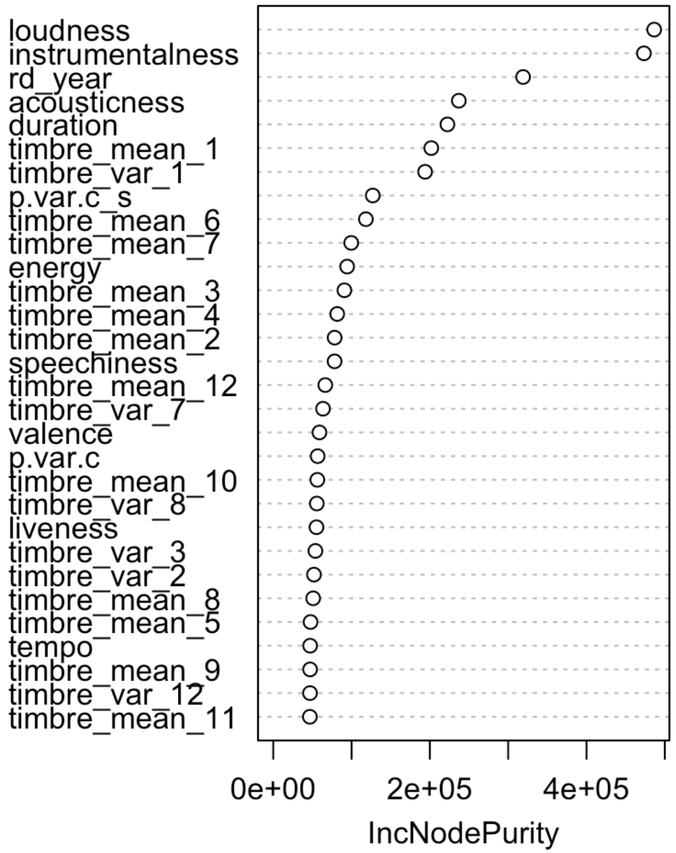
Model	Mean Absolute Error	Mean Squared Error
Linear Regression (R) with only intercept	24.29	730.06
Linear Regression (R) with Audio Features	21.03	648.77
Lasso (R, glmnet) (60.9% of coeff = 0)	19.96	541.32
Linear Regression (R), full dataset	19.71	538.25
Random Forest (R, randomForest), ntree = 100	18.76	492.05

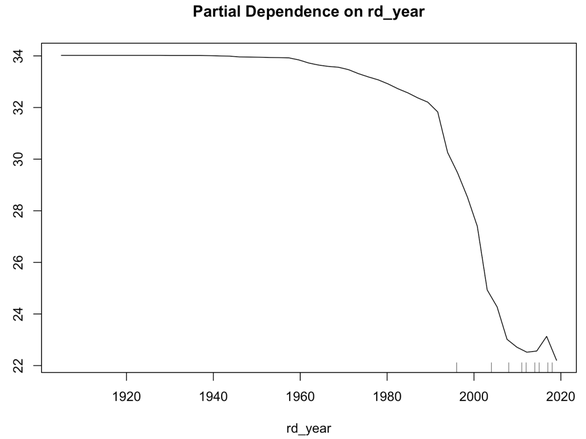
PREDICTION

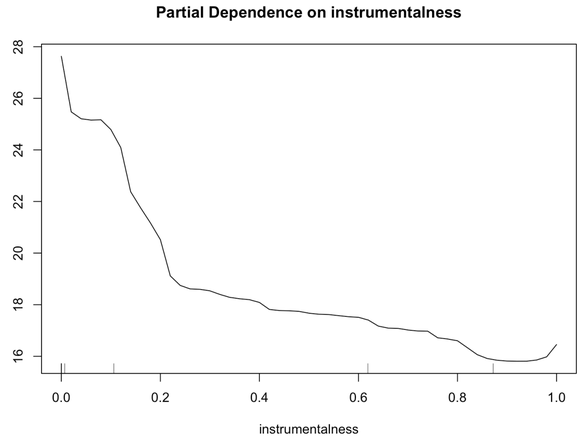
Random Forest

Variable importance









Interpretation:

“Mean predicted popularity” vs year of release and instrumentalness when averaging over all observed constellation of additional covariates .

Conclusion

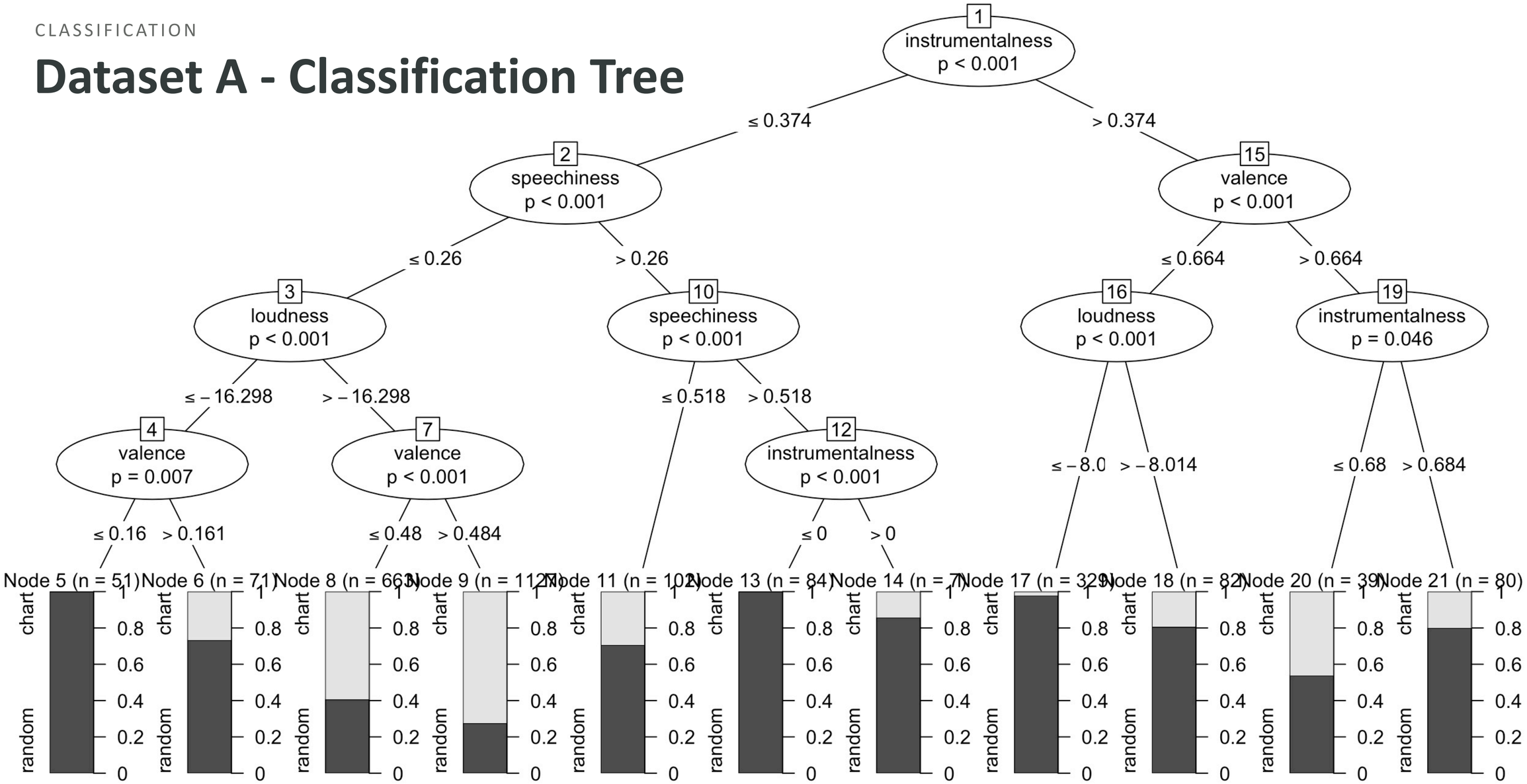
The Audio Features are helpful to classify the songs. With the current data it is difficult to predict a song’s popularity on Spotify.

The popularity depends also on various other things, like:

- rhythm
- lyrics
- song promotion
- artist

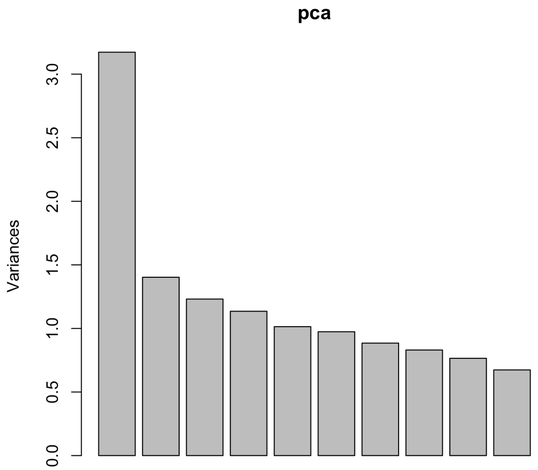
CLASSIFICATION

Dataset A - Classification Tree

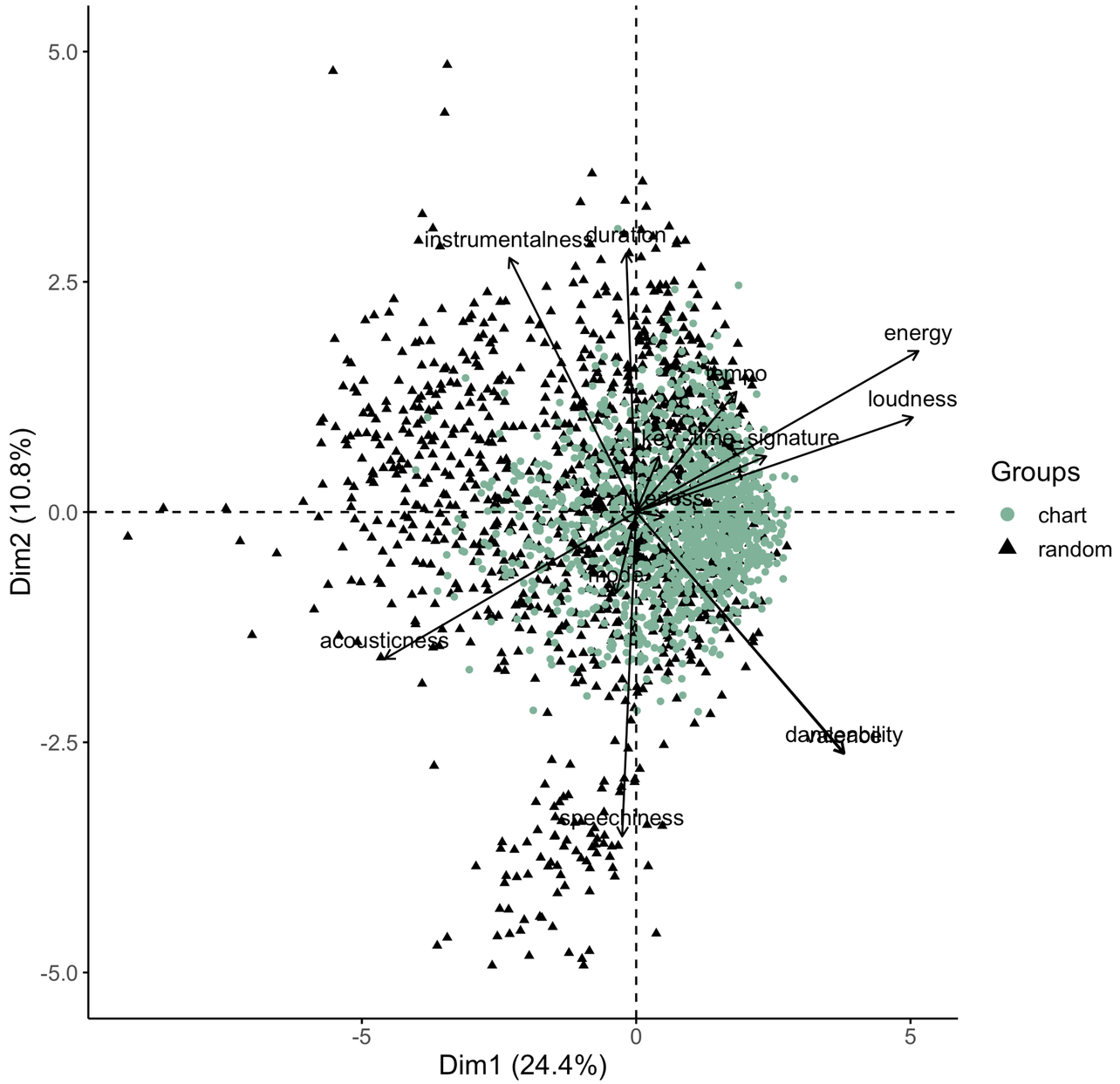


EXPLANATORY DATA ANALYSIS

Dataset A - PCA



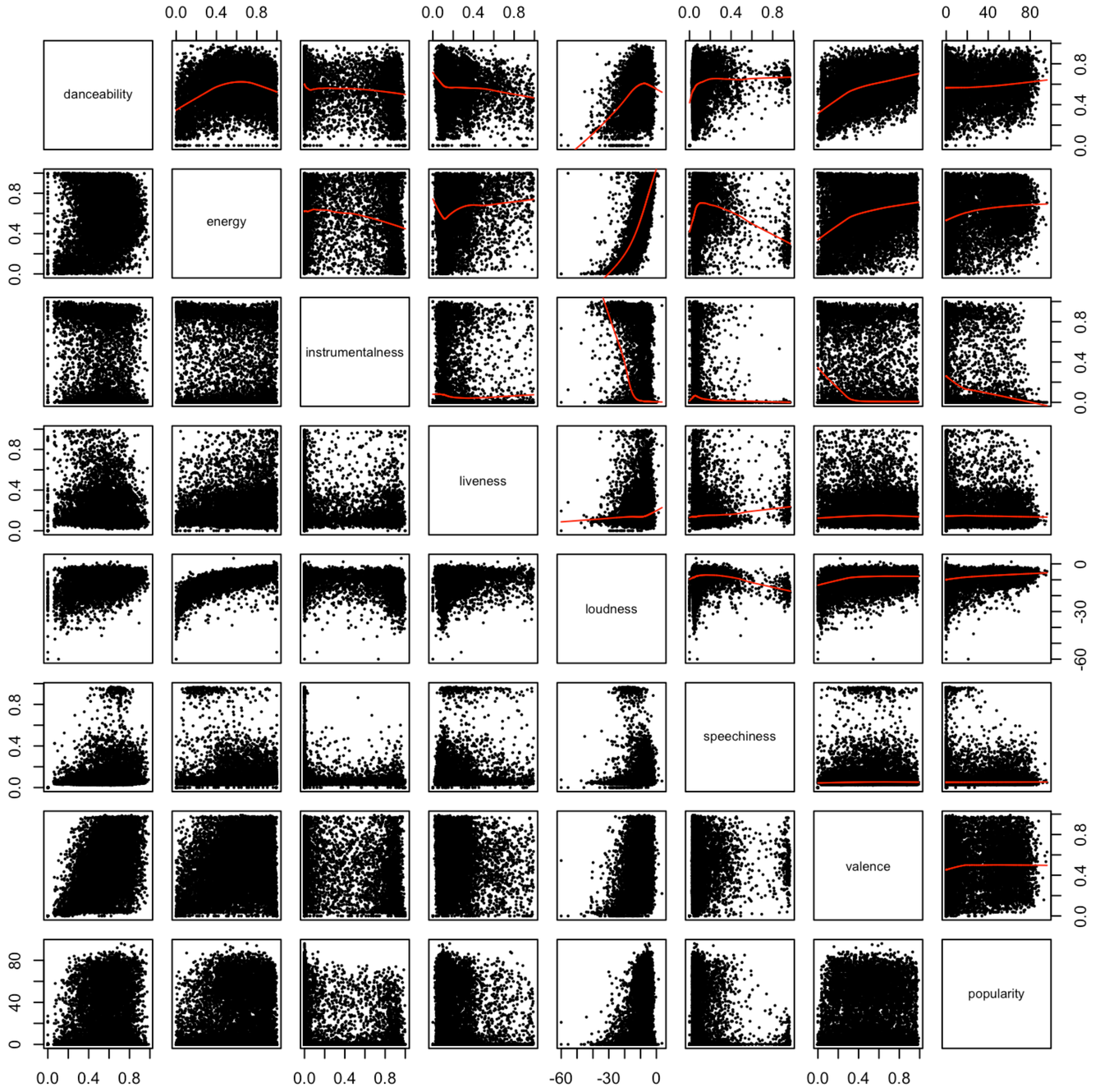
	PC2	PC3	PC4
acousticness	-0.42437918	-0.22205968	0.04618085	-0.004603094
danceability	0.35153893	-0.36449297	-0.09418094	-0.321739244
energy	0.47647789	0.24358175	-0.02589057	0.148690775
instrumentalness	-0.21403281	0.38449320	-0.12571093	-0.204198559
key	0.03843227	0.08406132	-0.53885204	-0.166217497
liveness	0.04678942	-0.00703968	-0.22824569	0.783617684
loudness	0.46711039	0.14322062	0.08835151	0.094001452
mode	-0.03412401	-0.12268657	0.62827444	0.214380194
speechiness	-0.02366533	-0.49061435	-0.43807986	0.234910303
tempo	0.16925177	0.18196122	0.04939822	0.167556658
time_signature	0.21943157	0.08435684	0.03271892	-0.148745118
valence	0.35045685	-0.36581213	0.05977753	-0.173653380
duration	-0.01679835	0.39145928	-0.16891807	-0.025256886
Cum. sum of Var:	24.4%	35.2%	44.7%	53.4%



EXPLANATORY DATA ANALYSIS

Dataset B - Pairplot

upper.panel = panel.smooth



PREDICTION

LASSO

LASSO = Least Absolute Shrinkage and Selection Operator

Uses instead of the ordinary least squares

$$\hat{\beta}_{\mathrm{LS}} = \mathrm{argmin}_{\beta} \| Y - X\beta \|_2^2,$$

the l1-norm regularization:

$$\hat{\beta}(\lambda) = \mathrm{argmin}_{\beta} (n^{-1} \| Y - X\beta \|^2 + \lambda \underbrace{\|\beta\|_1}_{\sum_{j=1}^{p} |\beta_j|})$$