Analyzing Spotify Trends and Predicting Song Popularity

By Danail Krzhalovski and Sandra Andovska

Music streaming services thank its growth to that they are able to react to new expectations of listeners, who want searchable music collections, automatic playlist suggestions, music recognition systems and more (Casey et al. 2008). Next to importance of recommendations, there is the importance of prediction of music popularity. In the music industry too, all parties have an interest in connecting consumers with content they will like and buy and it remains one of the biggest mysteries in the industry why some songs become popular while other songs fail to do so.

MOTIVATION

1. Data Collection
2. Data Exploration
3. Regression Models
4. Classification Models
5. Main Results and Discussion
6. Technical Appendix

OVERVIEW

1. DATA COLLECTION

Spotify is a digital music service that enables users to remotely source millions of different songs on various record labels. To recommend new music to users, and to be able to internally classify songs, Spotify assigns each song values from 13 different attributes/features. Spotify also assigns each song a popularity score, based on total number of clicks/listens. This dataset contains approximately 19000 songs from different Spotify playlists, which includes a popularity score and a set of metrics for each one. The dataset is available on Kaggle.

1. DATA COLLECTION

Using Excel, we merged the two .csv files into one, and carefully analyzed which columns to work with. Some were duplicates, such as the names of the song and artists present in both sheets, and some we found to be redundant – album name and playlist. We are working the following:

- **song_name**: name of the song observed
- **arist_name**: name of the artist.
- **song_duration**: duration of the song in miliseconds, changed to minutes
- **acousticness**: a confidence measure from 0.0 to 1.0 of whether the song is acoustic. 1.0 represents high confidence the song is acoustic.

1. DATA COLLECTION

- **liveness**: detects presence of an audience in recording
- **danceability**: describes how suitable a song is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity
- **energy**: a perceptual measure from 0.0 to 1.0 of intensity and activity
- **instrumentalness**: predicts whether a song contains no vocals
- **key**: represents the overall musical key the song is composed in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C#/Db, 2 = D, etc.

1. DATA COLLECTION

- **loudness**: overall loudness of a song in decibels (dB)
- **speechiness**: detects presence of spoken words in a song
- **tempo**: estimated tempo of a song in beats per minute (BPM)
- **time signature**: estimated overall time signature of a track (how many beats in each bar)
- **valence**: measure from 0.0 to 1.0 describing musical positiveness
- **popularity**: dependent variable in the range [0,100]

```
## 'data.frame':    18835 obs. of  16 variables:
##  $ song_name       : Factor w/ 13070 levels "'Til I Get It Right",..: 1607 5660 9804 1817 5093 1698
##  $ artist_name     : Factor w/ 7564 levels "$ain't","$uicideBoy$",..: 2575 3911 6858 5472 4817 2068
##  $ song_duration_ms: int  262333 216933 231733 216933 223826 235893 199893 213800 222586 203346 ...
##  $ acousticness    : num  0.00552 0.0103 0.00817 0.0264 0.000954 0.00895 0.000504 0.00148 0.00108 0.
##  $ danceability    : num  0.496 0.542 0.737 0.451 0.447 0.316 0.581 0.613 0.33 0.542 ...
##  $ energy          : num  0.682 0.853 0.463 0.97 0.766 0.945 0.887 0.953 0.936 0.905 ...
##  $ instrumentalness: num  2.94e-05 0.00 4.47e-01 3.55e-03 0.00 1.85e-06 1.11e-03 5.82e-04 0.00 1.04e
##  $ key             : int  8 3 0 0 10 4 4 2 1 9 ...
##  $ liveness        : num  0.0589 0.108 0.255 0.102 0.113 0.396 0.268 0.152 0.0926 0.136 ...
##  $ loudness        : num  -4.09 -6.41 -7.83 -4.94 -5.07 ...
##  $ audio_mode      : int  1 0 1 1 1 0 0 1 1 1 ...
##  $ speechiness     : num  0.0294 0.0498 0.0792 0.107 0.0313 0.124 0.0624 0.0855 0.0917 0.054 ...
##  $ tempo           : num  167 105 124 122 172 ...
##  $ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ audio_valence   : num  0.474 0.37 0.324 0.198 0.574 0.32 0.724 0.537 0.234 0.374 ...
##  $ song_popularity : int  73 66 76 74 56 80 81 76 80 81 ...
```

1. DATA COLLECTION

Data was already cleaned up and tidied, but some additional modifications had to be made. We created factors for the 'audio_mode' and 'key' variables to make the data easier to interpret. Anyone with at least a bit of musical knowledge would prefer and actually find it easier to understand the analysis if the person could see the keys (C, C#, etc.) and modes (major, minor) instead of going back to the description of features to check their numerical values. We too transform the song duration from miliseconds to minutes, and factorize the time signature since it does not provide us a true numerical property.

2. DATA EXPLORATION

After checking for missing values and duplicates, we are now working with 14926 observations. Data was already cleaned up and tidied, but some additional modifications had to be made.

We created factors for the 'audio_mode' and 'key' variables to make the data easier to interpret. Anyone with at least a bit of musical knowledge would prefer and actually find it easier to understand the analysis if the person could see the keys (C, C#, etc.) and modes (major, minor) instead of going back to the description of features to check their numerical values. We too transform the song duration from miliseconds to minutes, and factorize the time signature since it does not provide us a true numerical property.

```
'data.frame':   14926 obs. of  16 variables:
$ song_name       : Factor w/ 13070 levels "'Til I Get It Right",..: 1607 5660 9804 1817 5093 1698 6327 772 7595 9815 ...
$ artist_name     : Factor w/ 7564 levels "$ain't","$uicideBoy$",..: 2575 3911 6858 5472 4817 2068 5089 3169 6624 3602 ...
$ song_duration   : num  4.37 3.62 3.86 3.62 3.73 ...
$ acousticness    : num  0.00552 0.0103 0.00817 0.0264 0.000954 0.00895 0.000504 0.00148 0.00108 0.00172 ...
$ danceability    : num  0.496 0.542 0.737 0.451 0.447 0.316 0.581 0.613 0.33 0.542 ...
$ energy          : num  0.682 0.853 0.463 0.97 0.766 0.945 0.887 0.953 0.936 0.905 ...
$ instrumentalness: num  2.94e-05 0.00 4.47e-01 3.55e-03 0.00 1.85e-06 1.11e-03 5.82e-04 0.00 1.04e-02 ...
$ key             : Factor w/ 12 levels "C","C#","D","D#",..: 9 4 1 1 11 5 5 3 2 10 ...
$ liveness        : num  0.0589 0.108 0.255 0.102 0.113 0.396 0.268 0.152 0.0926 0.136 ...
$ loudness        : num  -4.09 -6.41 -7.83 -4.94 -5.07 ...
$ audio_mode      : Factor w/ 2 levels "major","minor": 1 2 1 1 1 2 2 1 1 1 ...
$ speechiness     : num  0.0294 0.0498 0.0792 0.107 0.0313 0.124 0.0624 0.0855 0.0917 0.054 ...
$ tempo           : num  167 105 124 122 172 ...
$ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
$ audio_valence   : num  0.474 0.37 0.324 0.198 0.574 0.32 0.724 0.537 0.234 0.374 ...
$ song_popularity : int  73 66 76 74 56 80 81 76 80 81 ...
```

2. DATA EXPLORATION

We begin our analysis by checking which artists are a part of Spotify playlists the most. By looking at the entire list, we can see that all artists are from different time periods, genres and cultures, so we can say that our data is not biased towards a specific type of music nor a time period.

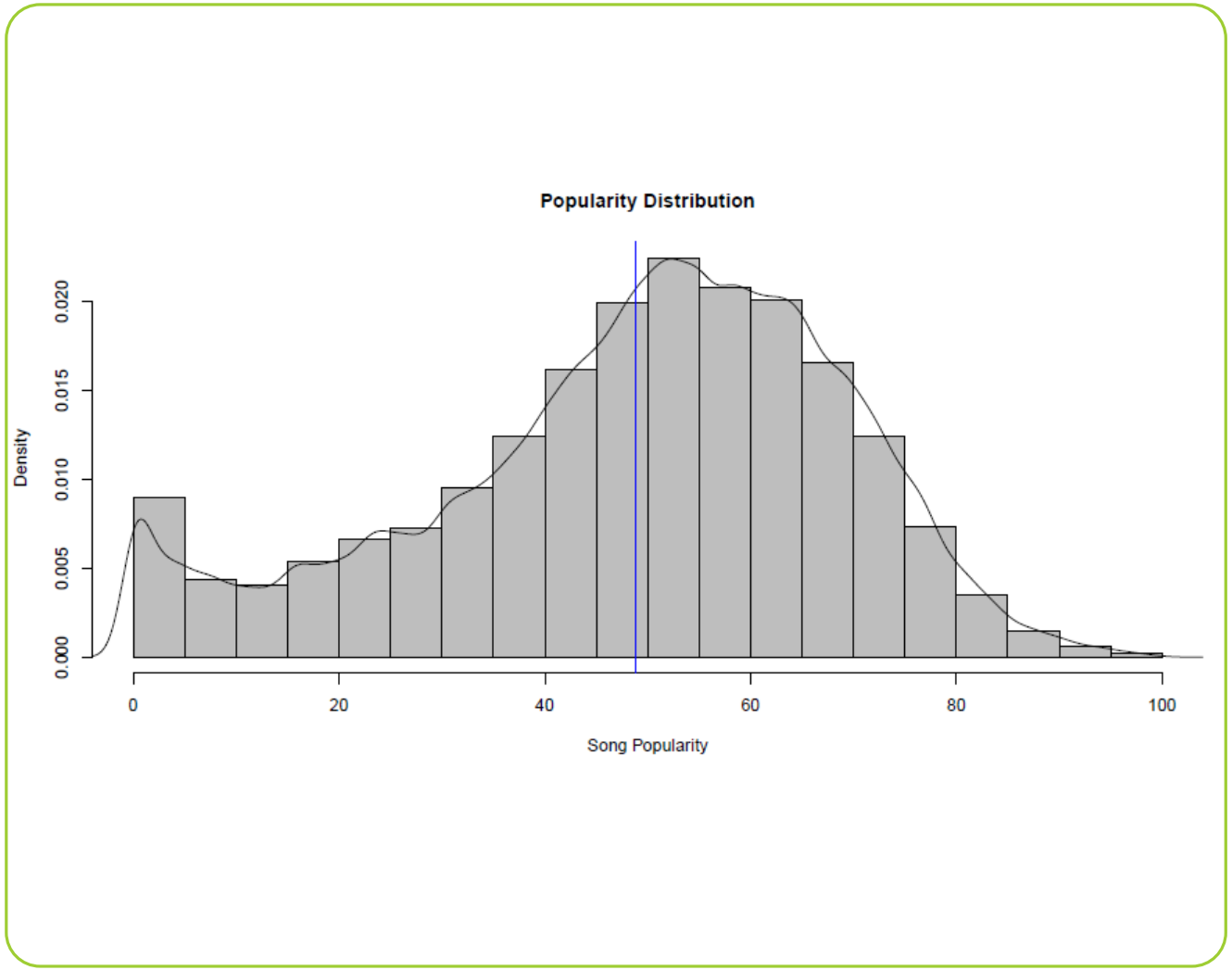
```
##                 Number of Songs
## Lady Gaga                    57
## Drake                        45
## Kanye West                   44
## Eminem                       25
## Khalid                       25
## Ed Sheeran                   23
## David Guetta                 22
## Gucci Mane                   22
## Kendrick Lamar               22
## Celia Cruz                   20
## Maroon 5                     20
## Future                       19
## Major Lazer                  19
## The Beatles                  19
## Imagine Dragons              18
## R3HAB                        18
## Rihanna                      18
## The Weeknd                   18
## Hank Williams                17
## Johnny Cash                  17
```

SUMMARIES OF THE DATA

```
  song_name         artist_name    song_duration     acousticness       danceability        energy
Fire   :    8   Lady Gaga :   57   Min.   : 0.200   Min.   :0.000001   Min.   :0.0000   Min.   :0.00107
Heaven :    8   Drake     :   45   1st Qu.: 3.066   1st Qu.:0.023600   1st Qu.:0.5240   1st Qu.:0.49600
Alright:    7   Kanye West:   44   Median : 3.531   Median :0.139000   Median :0.6360   Median :0.67200
Better :    7   Eminem    :   25   Mean   : 3.649   Mean   :0.270452   Mean   :0.6245   Mean   :0.63976
Breathe:    7   Khalid    :   25   3rd Qu.: 4.079   3rd Qu.:0.458000   3rd Qu.:0.7400   3rd Qu.:0.81800
Fall   :    7   Ed Sheeran:   23   Max.   :29.989   Max.   :0.996000   Max.   :0.9870   Max.   :0.99900
(Other):14882   (Other)   :14707
instrumentalness          key          liveness          loudness       audio_mode    speechiness
Min.   :0.0000000   C      :1734   Min.   :0.0109   Min.   :-38.768   major:9432   Min.   :0.00000
1st Qu.:0.0000000   G      :1654   1st Qu.:0.0930   1st Qu.: -9.389   minor:5494   1st Qu.:0.03720
Median :0.0000208   C#     :1594   Median :0.1220   Median : -6.750                Median :0.05410
Mean   :0.0920668   A      :1410   Mean   :0.1804   Mean   : -7.677                Mean   :0.09942
3rd Qu.:0.0051050   D      :1399   3rd Qu.:0.2240   3rd Qu.: -4.991                3rd Qu.:0.11300
Max.   :0.9970000   F      :1257   Max.   :0.9860   Max.   :  1.585                Max.   :0.94100
                    (Other):5878
    tempo          time_signature audio_valence    song_popularity
Min.   :  0.00   0:    3         Min.   :0.0000   Min.   :  0.00
1st Qu.: 98.12   1:   67         1st Qu.:0.3320   1st Qu.: 37.00
Median :120.02   3:  684         Median :0.5270   Median : 52.00
Mean   :121.11   4:13977         Mean   :0.5270   Mean   : 48.75
3rd Qu.:139.94   5:  195         3rd Qu.:0.7278   3rd Qu.: 63.75
Max.   :242.32                   Max.   :0.9840   Max.   :100.00
```

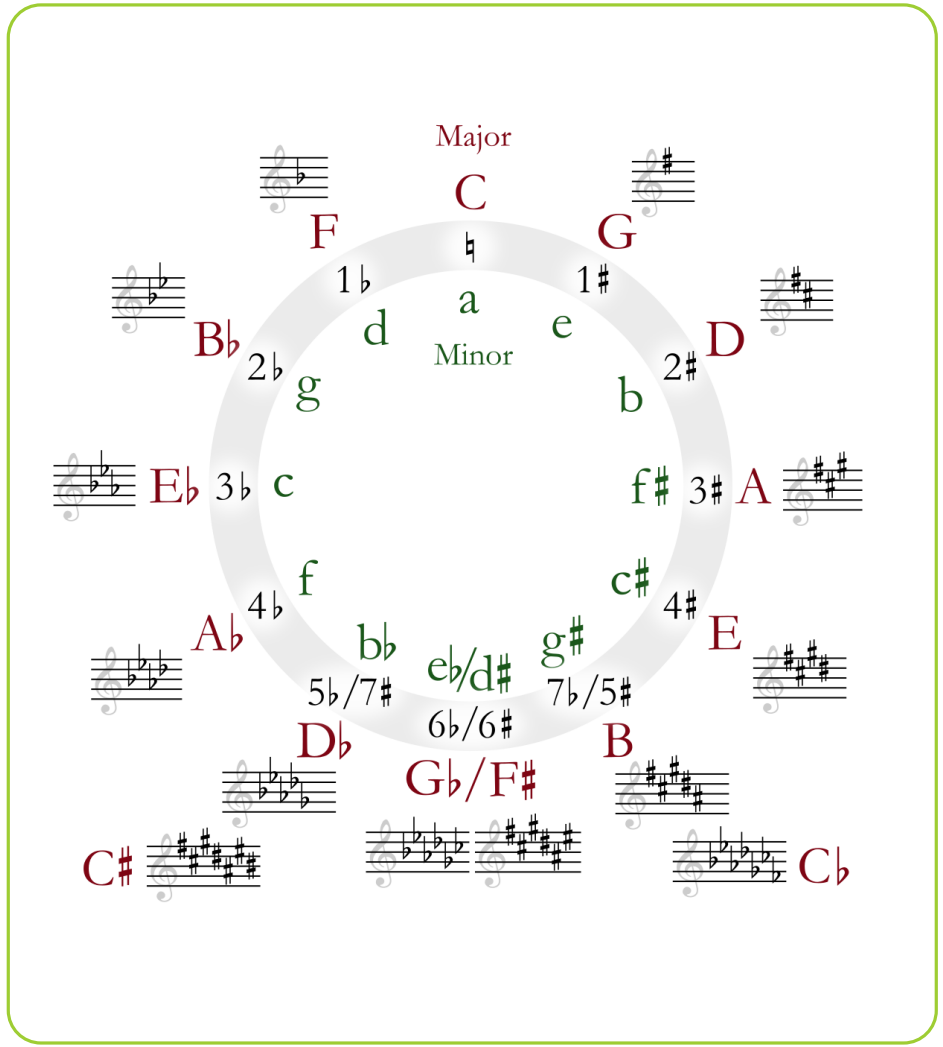
POPULARITY DISTRIBUTION

By taking a look at how the popularity scores are distributed, one thing is noticeable instantly. The data appears to be slightly negatively skewed, with majority of the songs having a popularity score more than 40. The mean popularity score at a value of 48.75 and median is 52.



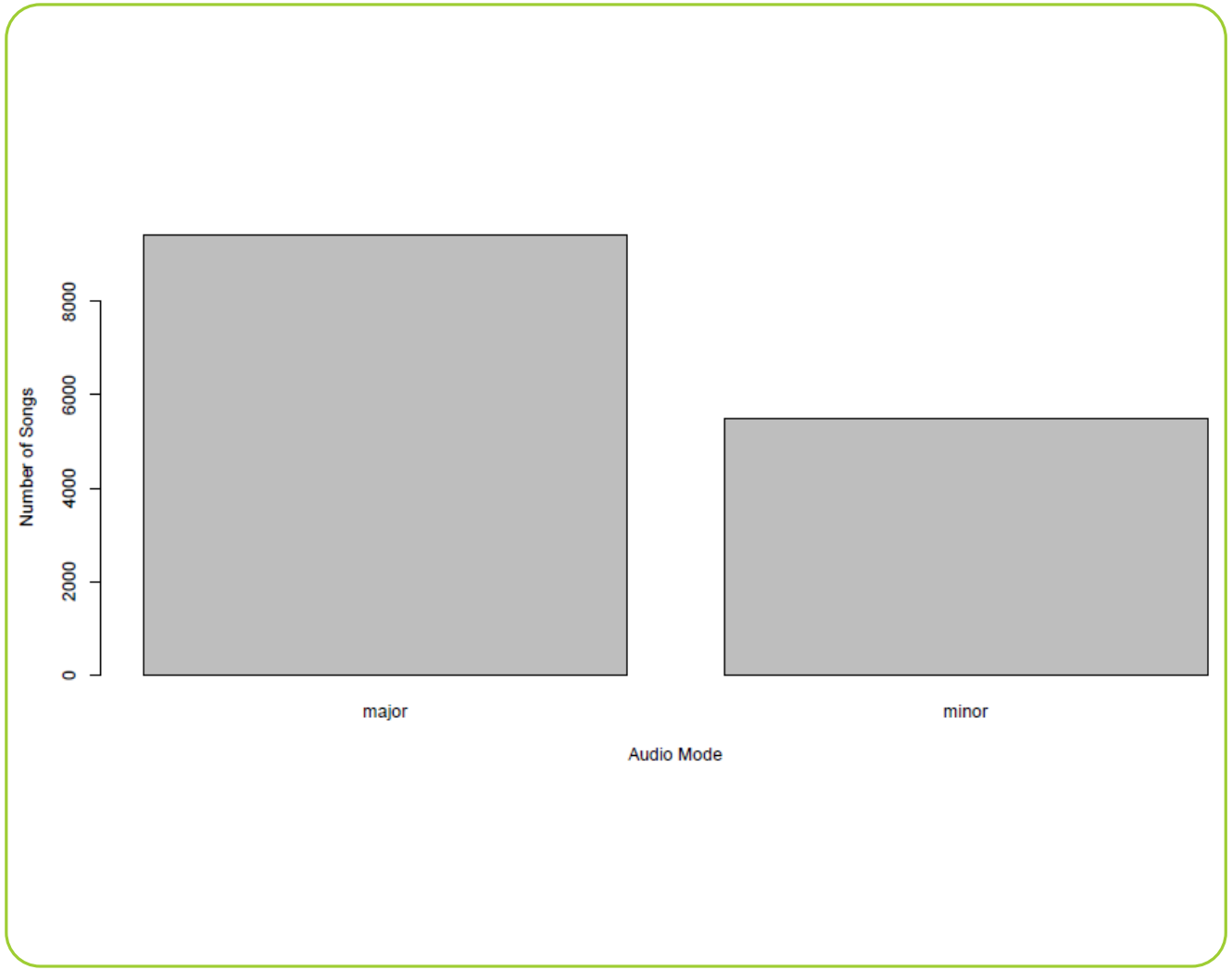
ASSUMPTION

We can assume that the audio mode, key, tempo and time signature of a song are composition elements, which we can make us of to describe and classify music.



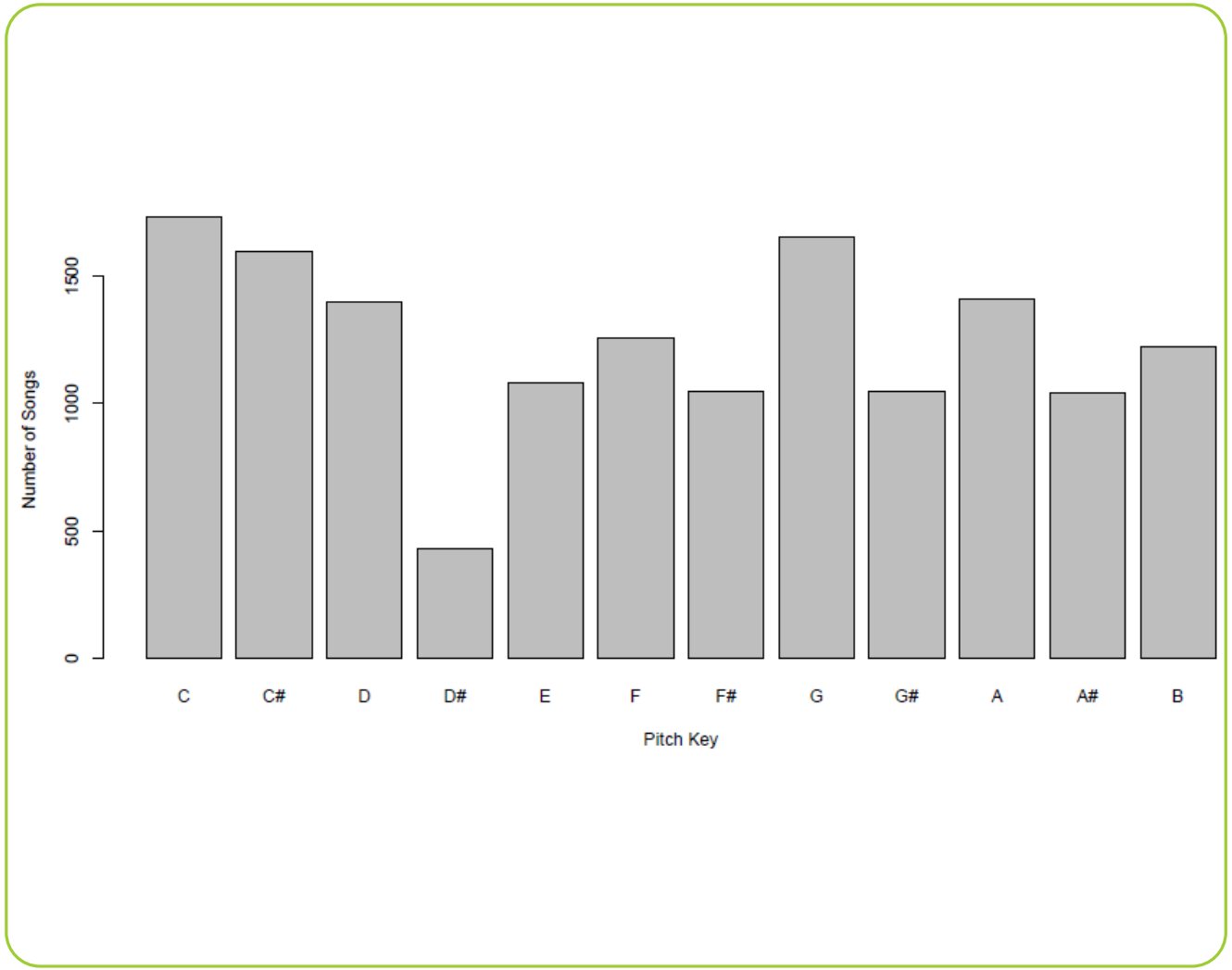
AUDIO MODE

The emotional center of music comes from one of two places: the major chord or the minor chord. If you're listening to music and you can sense happiness and you're at ease, you're probably listening to a song that uses mostly major chords to create that feeling.



PITCH KEYS

In addition to the audio mode, the association of musical keys with specific emotional or qualitative characteristic was fairly common prior to the 20th century. When Mozart wrote a piece in a Ab major, for example, they were well aware that this was the ‘key of the grave’.



PITCH KEY INTERPRETATION

The most common key in our dataset is C:

- **C major**: Completely pure. Its character is: innocence, simplicity, naivety, children's talk.
- **C minor**: Declaration of love and at the same time the lament of unhappy love. All languishing, longing, sighing of the love-sick soul lies in this key.

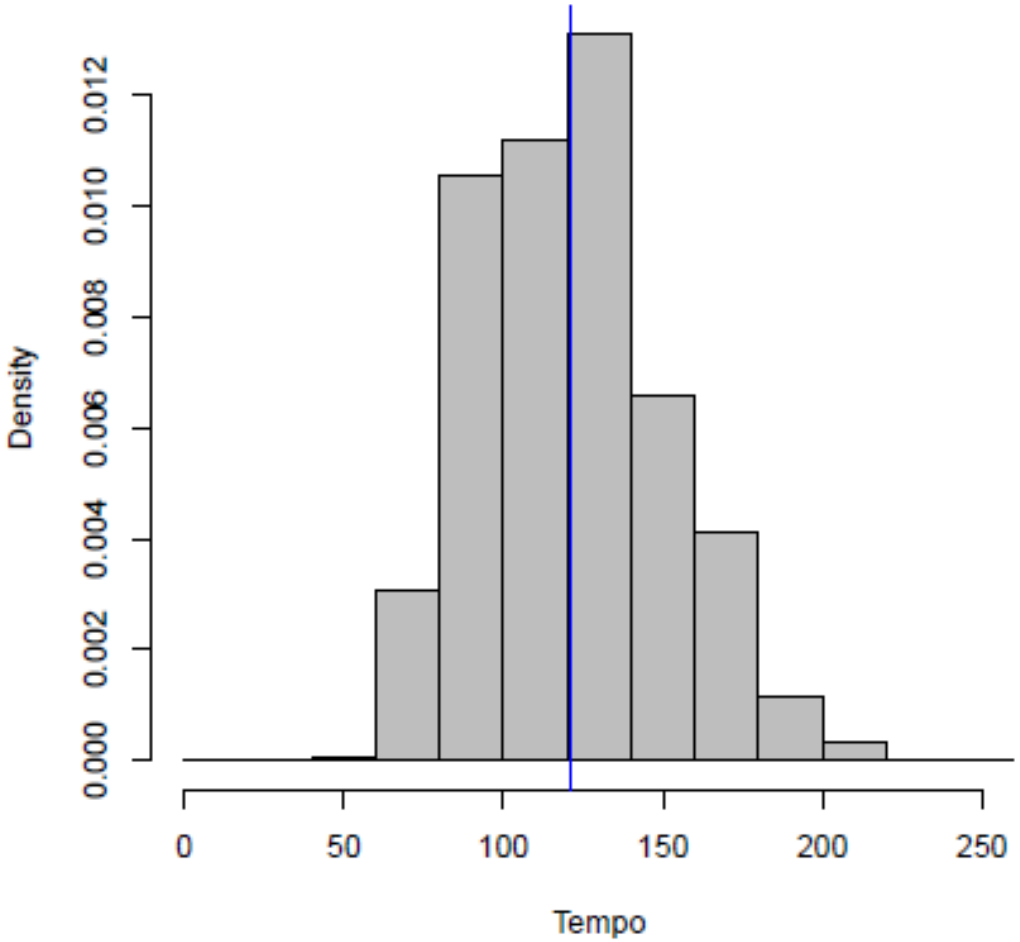
The least common key is D#:

- **D# major** (equivalent to E-flat): The key of love, of devotion, of intimate conversation with God.
- **D# minor**: Feelings of the anxiety of the soul's deepest distress, of brooding despair, of blackest depression, of the most gloomy condition of the soul. Every fear, every hesitation of the shuddering heart.

TEMPO

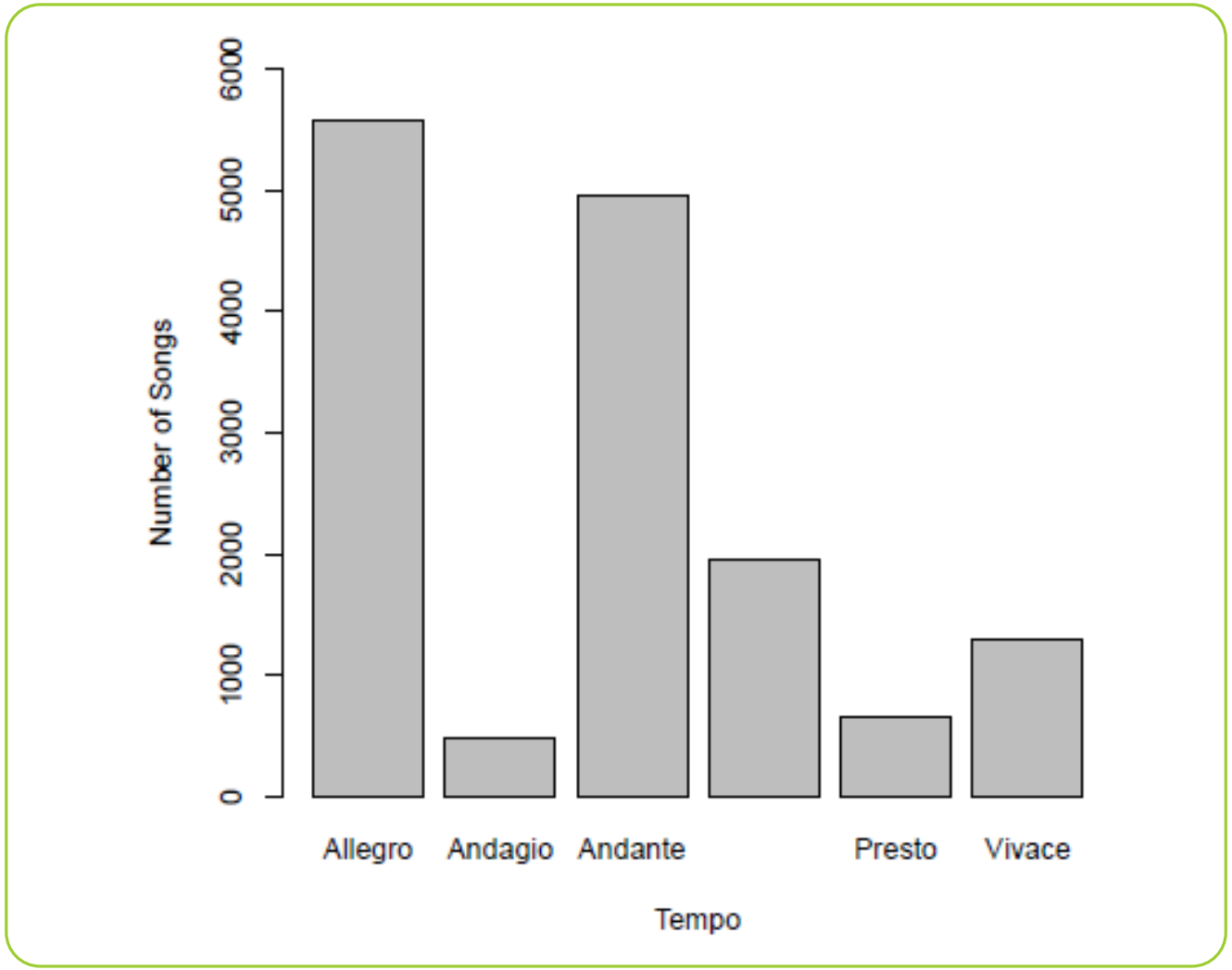
Popular songs tend to have a faster tempo with an average BPM of 121.105. This is commonly known as ‘allegro’, ranging from 120 – 156 BPM. Modern music tempos are in this range:

- techno (120-140 BPM)
- house (115-130 BPM)
- hip-hop (80-120)

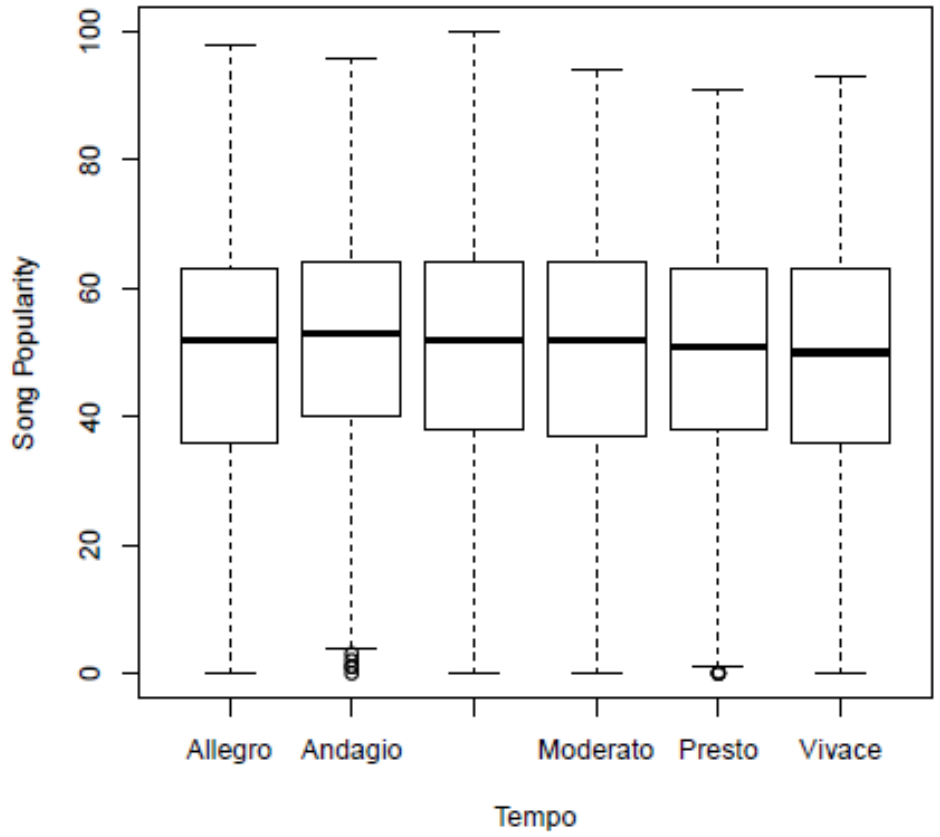


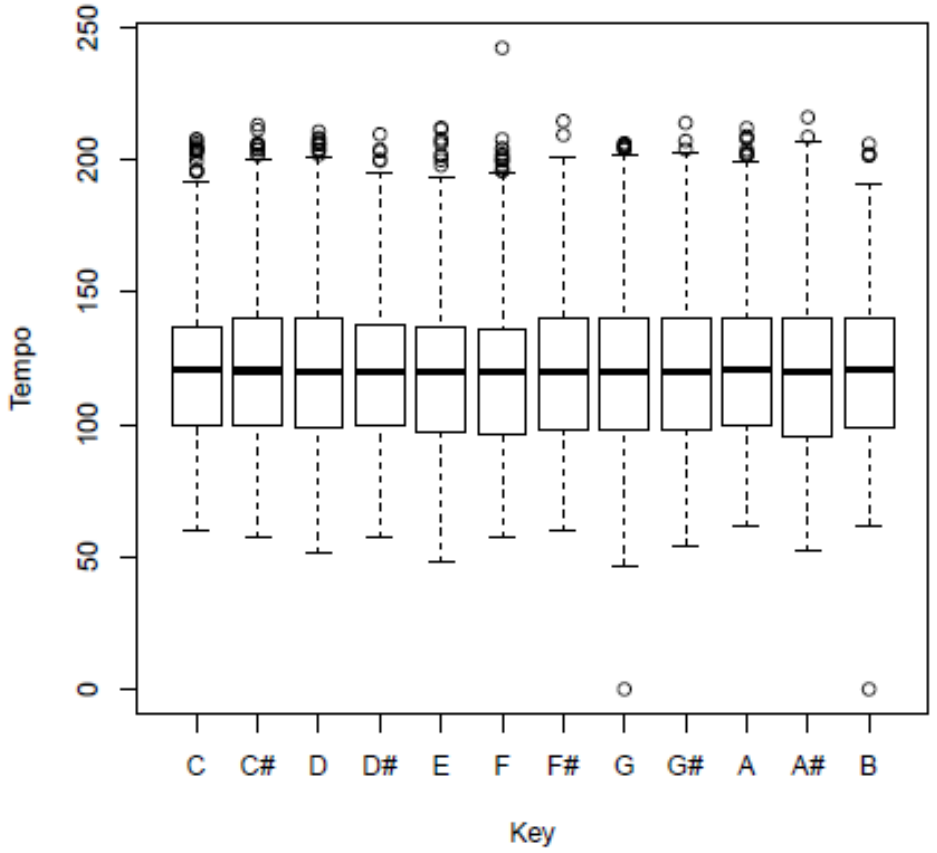
TEMPO CATEGORIES

- **Adagio** - slowly with great expression (66-76 bpm)
- **Andante** - at a walking pace (76-108 bpm)
- **Moderato** - at a moderate speed (108-120 bpm)
- **Allegro** - fast, quickly, and bright (120-156 bpm)
- **Vivace** - lively and fast (156-176 bpm)
- **Presto** - very, very fast (168-200 bpm)

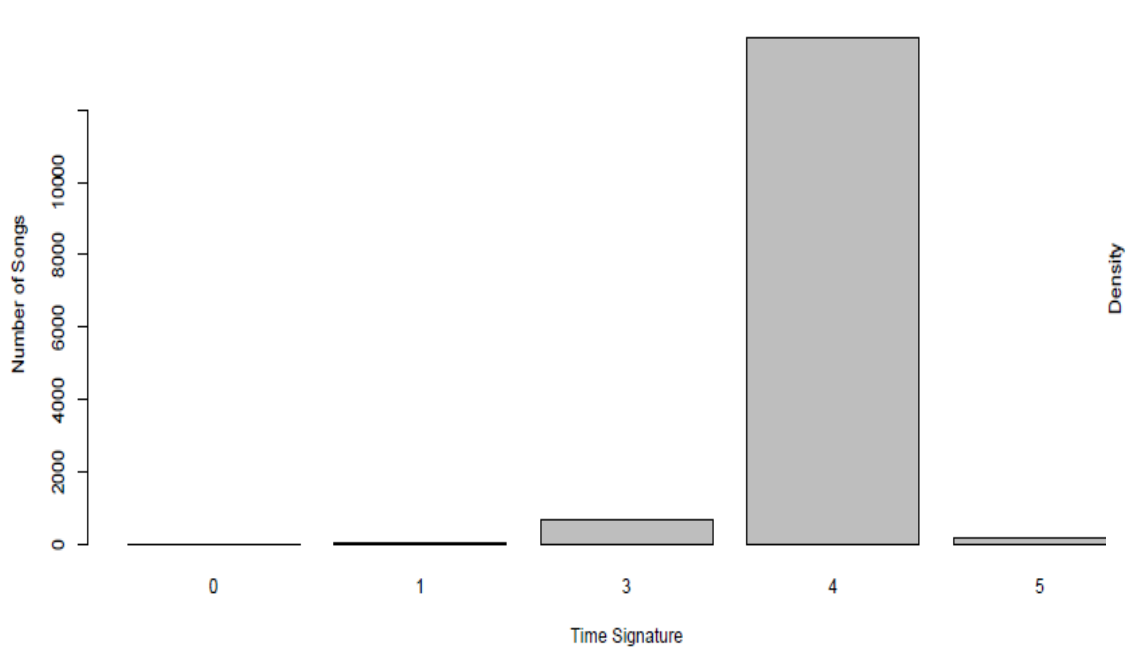


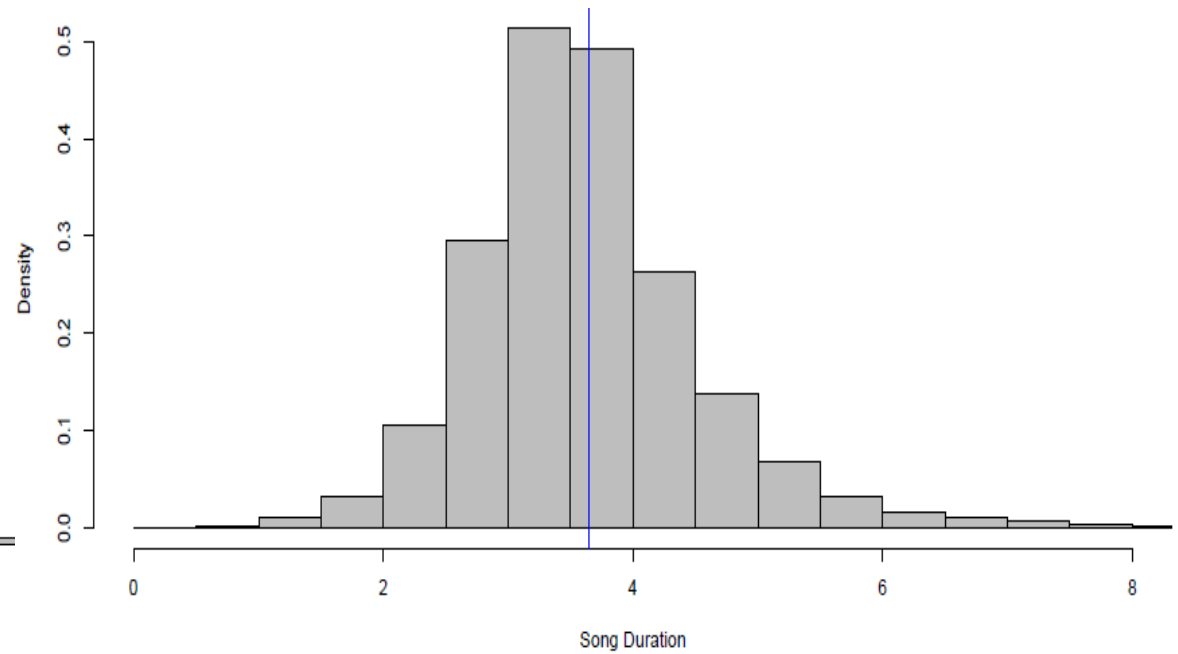
We also have a clear indication of the consistency of mean tempos across all songs and their popularities, with a symmetrical distribution. We could say that the data is approximately normally distributed in terms of the songs' tempo. We too can observe a consistency of mean tempos across all keys.



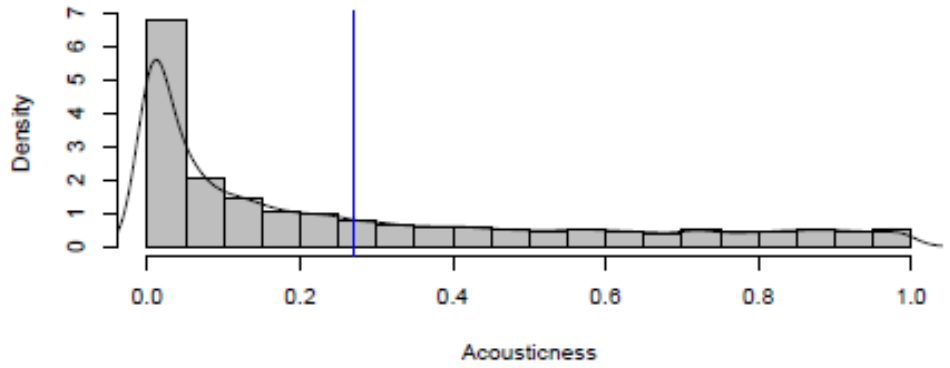


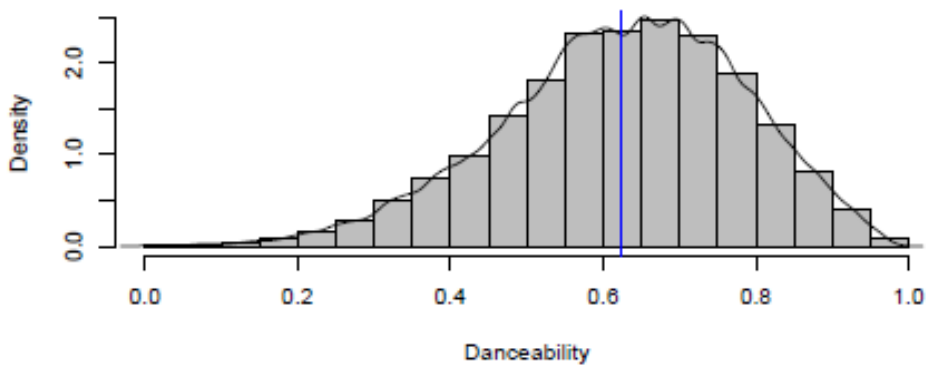
We can see that the most common time signature used in music is the 4/4 meter, known as 'common time', and that a song has an average duration of 3-4 minutes.

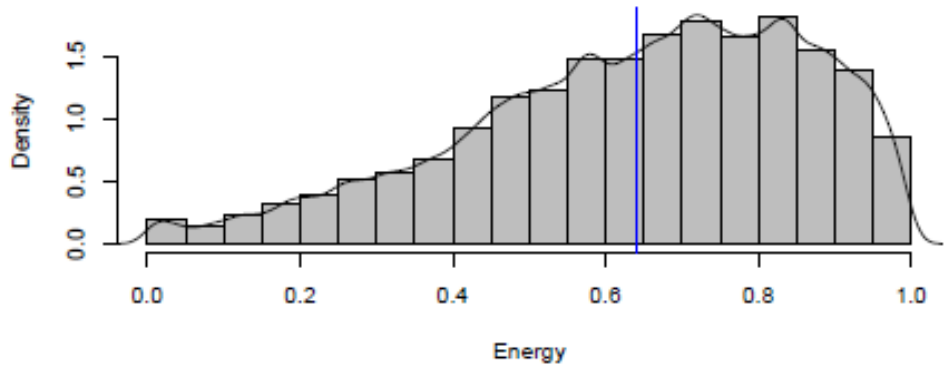


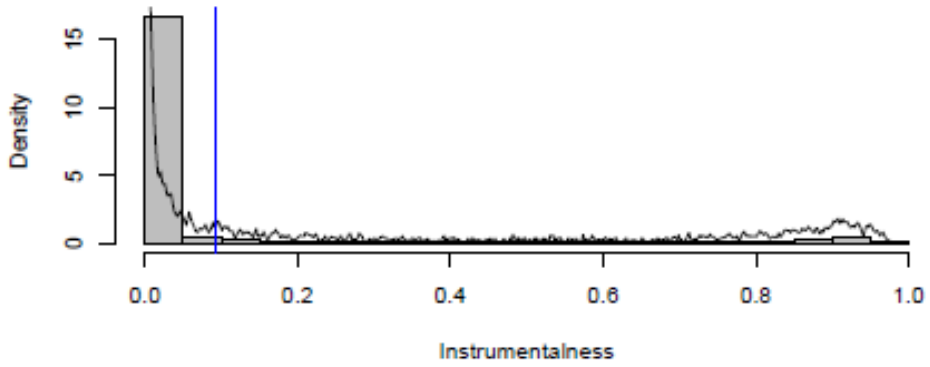


Distribution of acousticness, danceability, energy and instrumentalness of our data

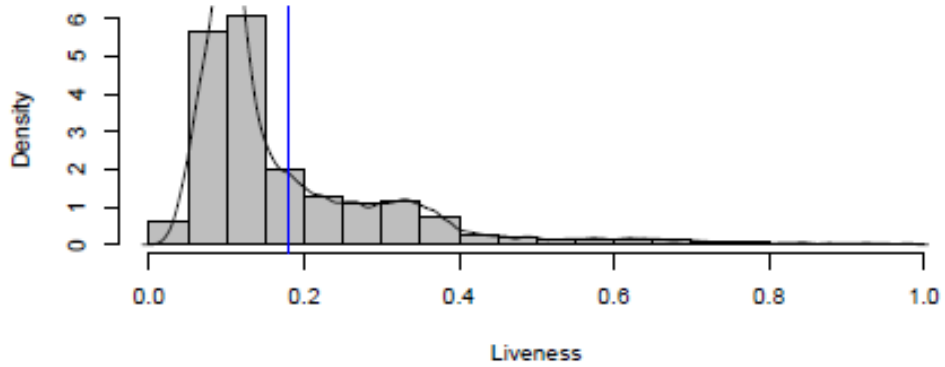


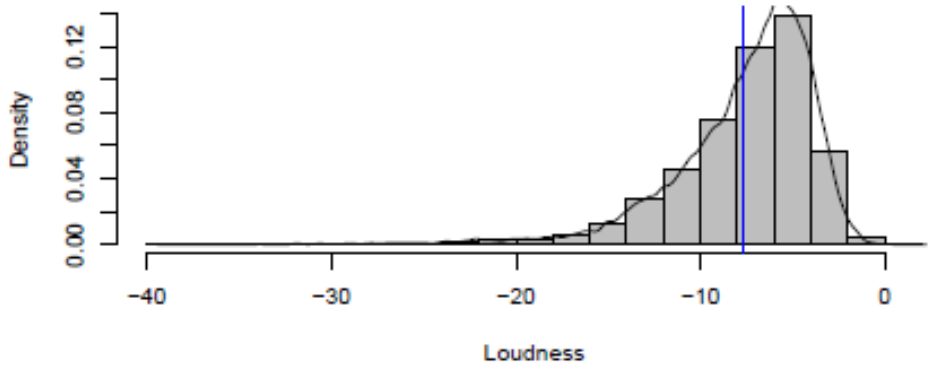


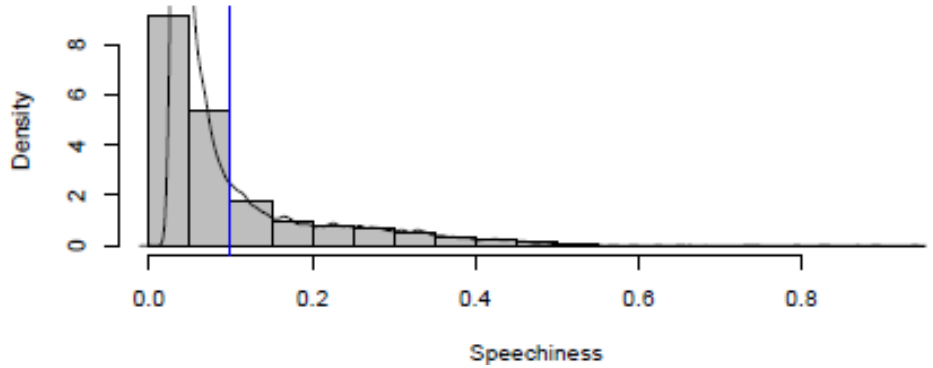


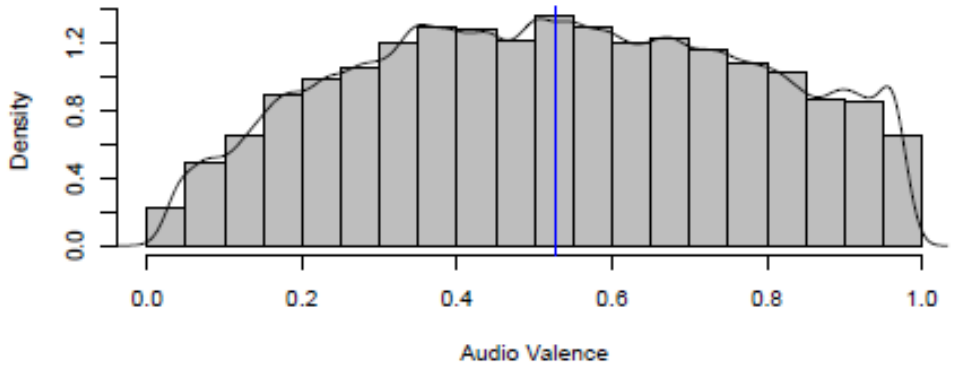


Distribution of liveness, loudness, speechiness and valence of our data

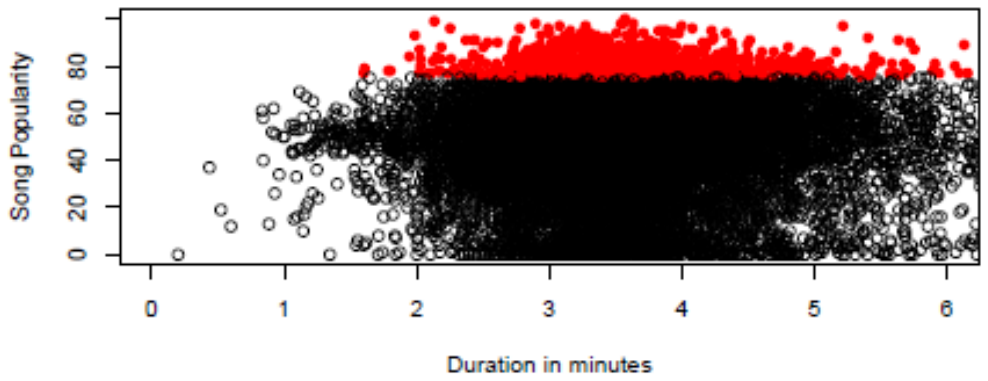


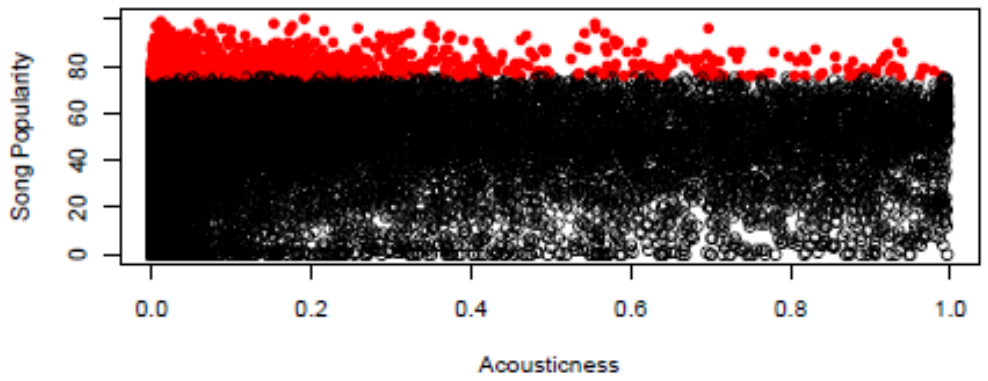


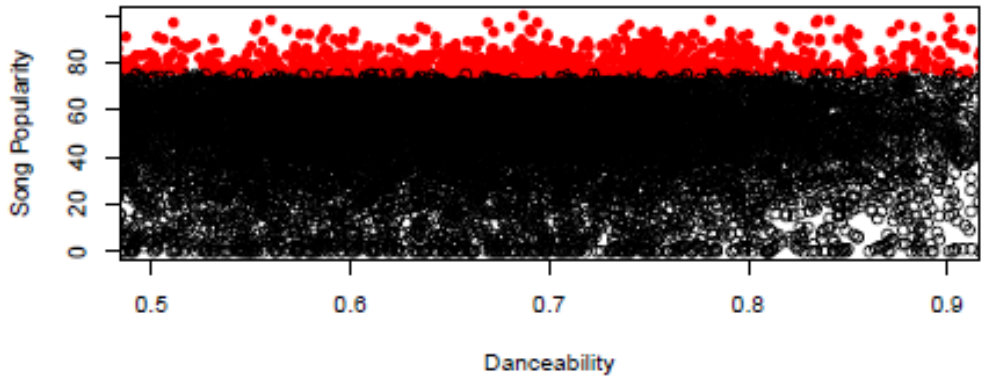


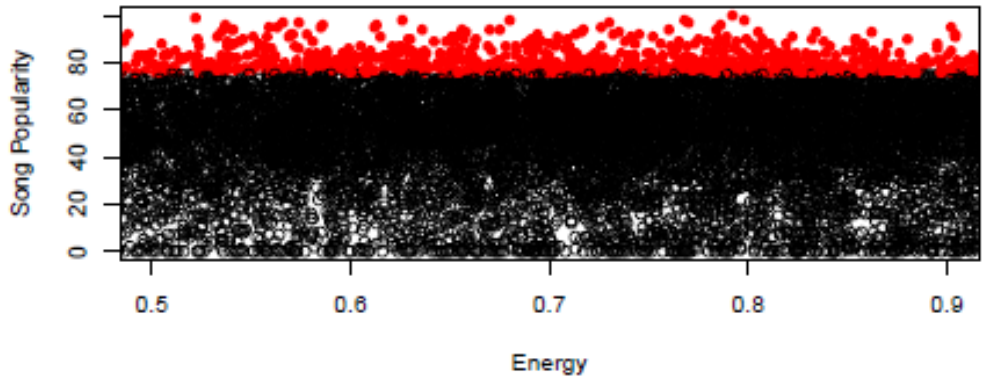


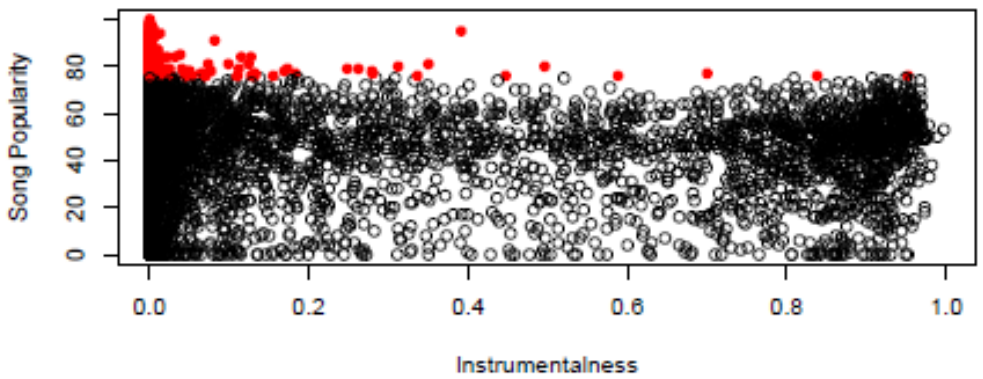
Because we deal with classification in the second part, we divide the song popularity into low and high. We choose a popularity score of 75, since this is the border of the top 25% of songs.

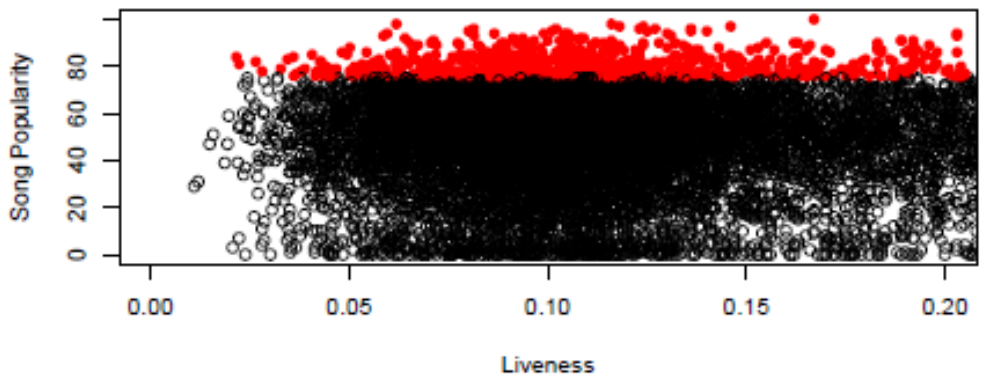


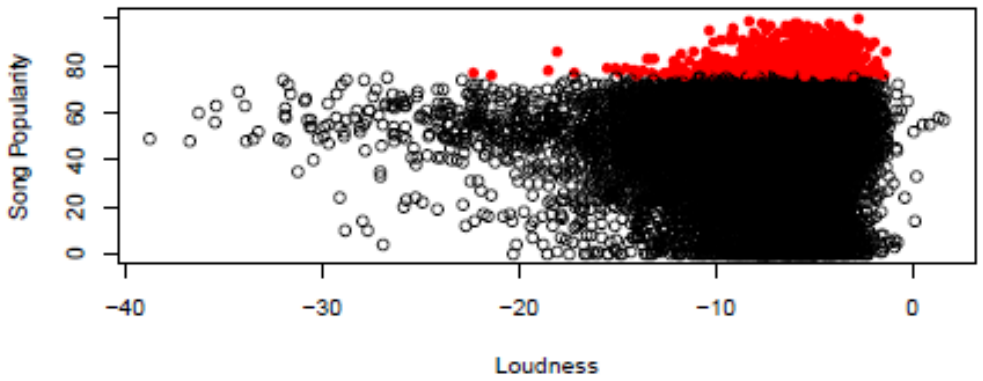


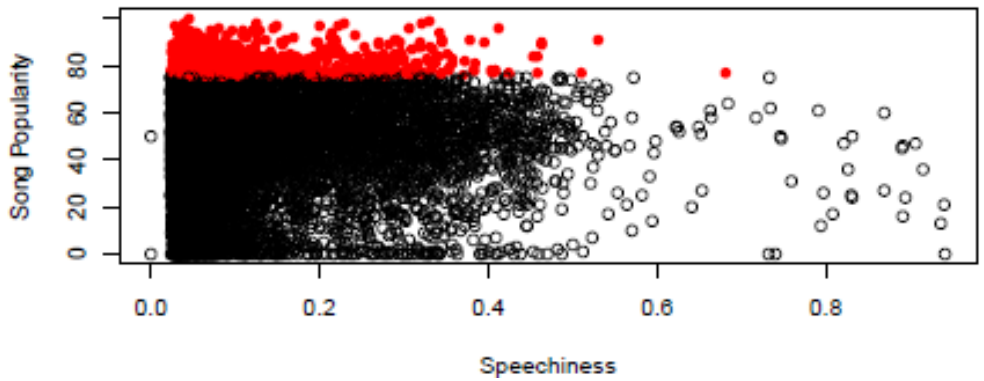


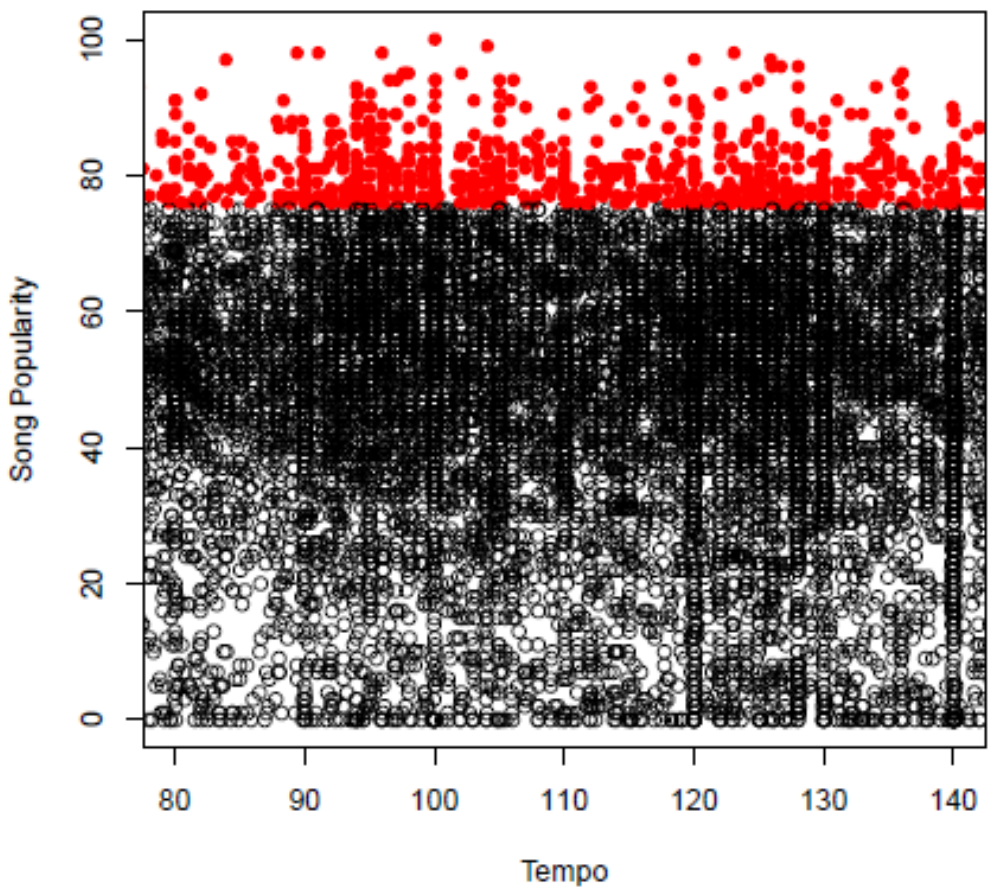


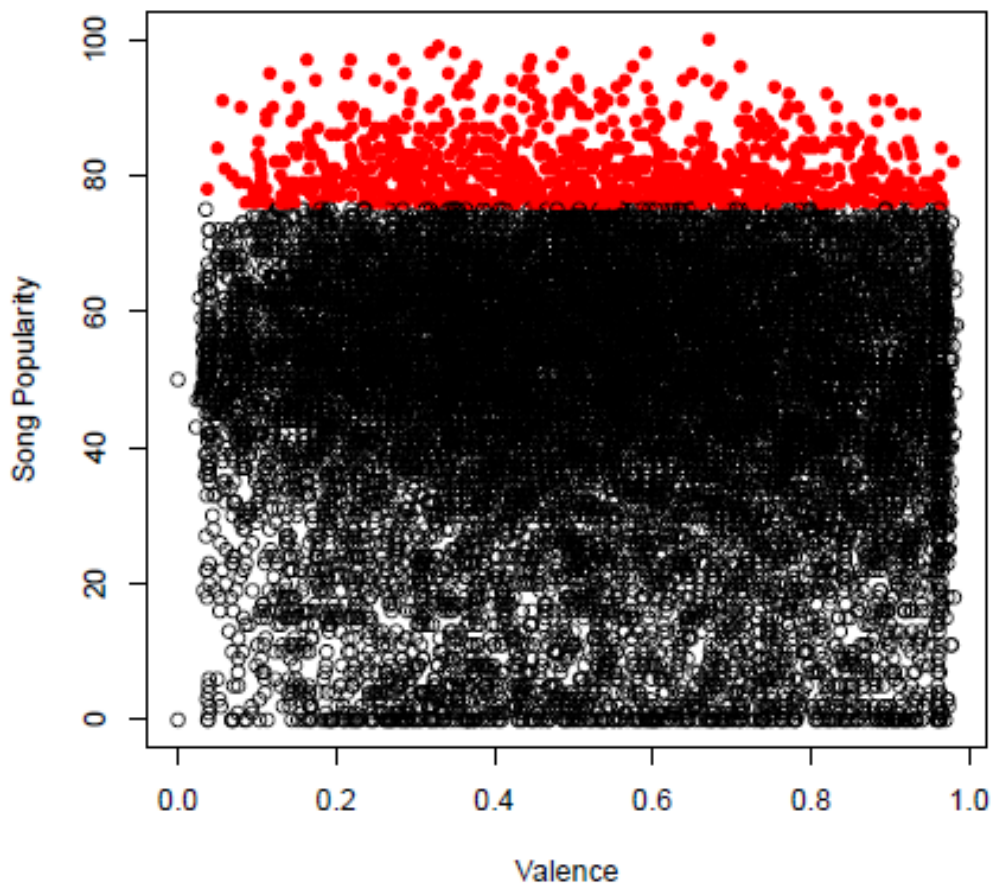

Song Popularity
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
Instrumentalness


Song Popularity
0
20
40
60
80
0.00
0.05
0.10
0.15
0.20
Liveness


Song Popularity
0
20
40
60
80
−40
−30
−20
−10
0
Loudness

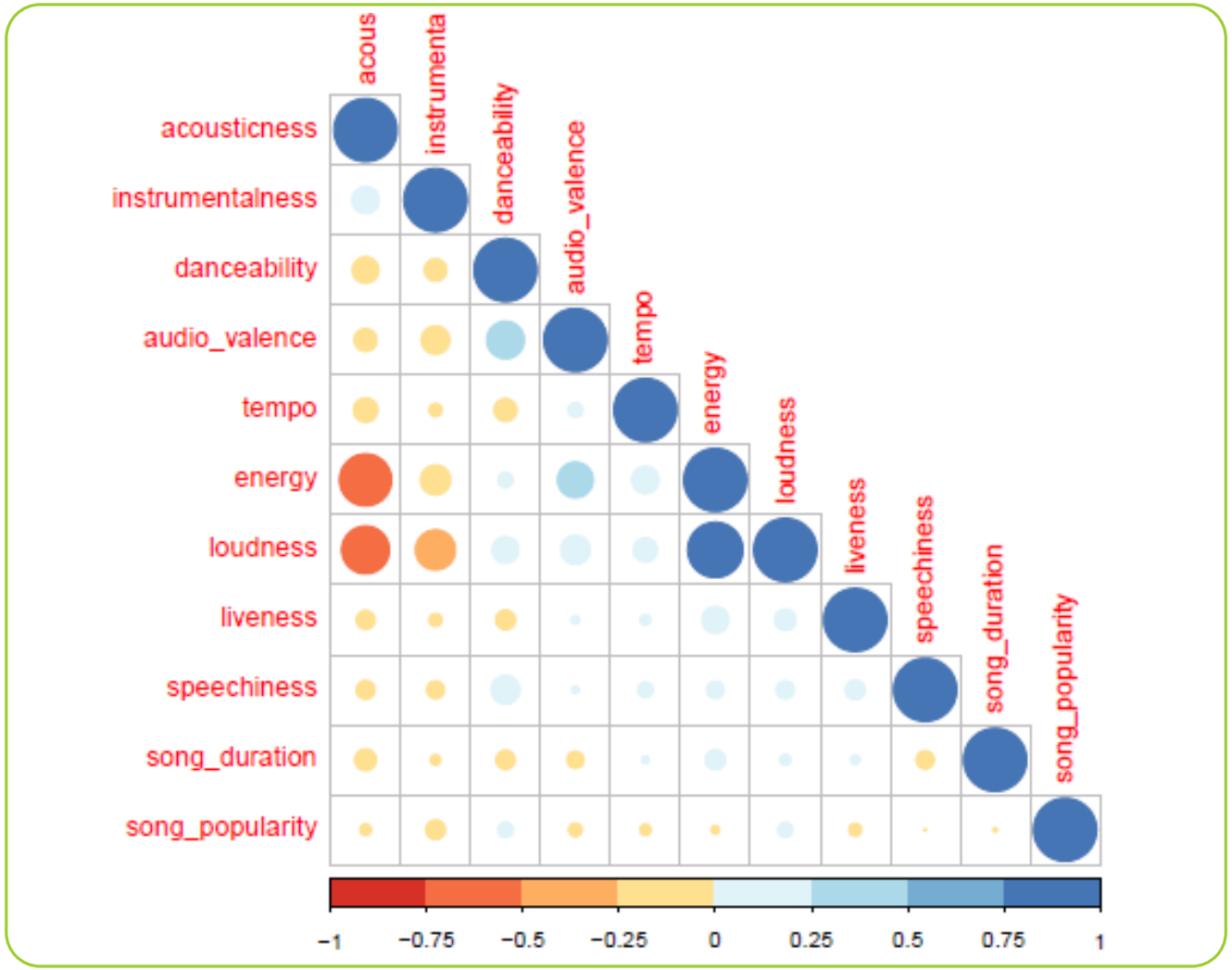

Song Popularity
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
Speechiness


Song Popularity
0
20
40
60
80
100
80
90
100
110
120
130
140
Tempo


Song Popularity
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Valence

COLLINEARITY

We further need to check the covariance matrix to see the direction of the linear relationship between variables, but since correlation measures both its direction and strength, we use a colour correlation plot just to see how our data behaves.



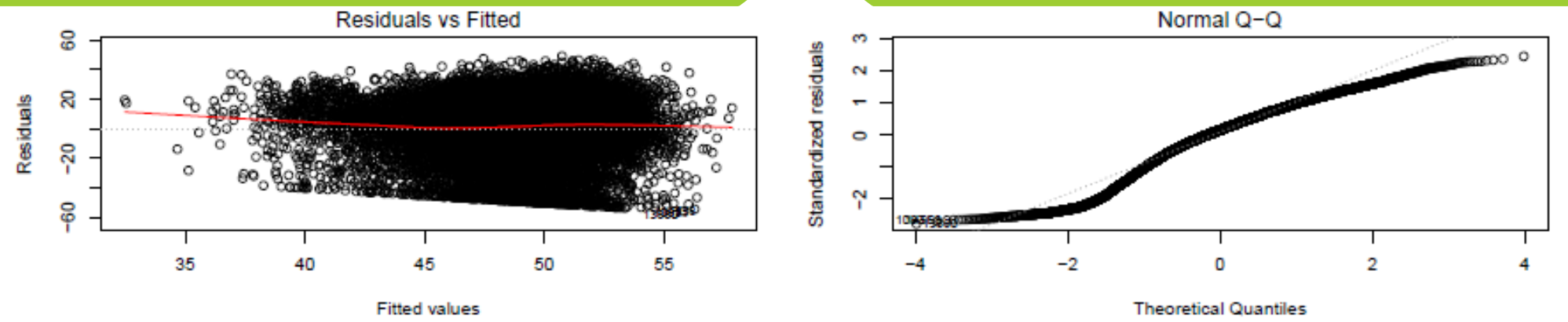
3. REGRESSION MODELS

Judging by the looks of our data, and the relationship among the predictors and the response variable, we do not expect a linear model to help us predict a song's popularity, but we do try out several:

- Linear Regression
- Ridge Regression
- LASSO
- Bayes Regression

To evaluate the results, we use Root Mean Squared Error. Mean Squared Error tells us how close a regression line is to a set of points, and we are simply taking its root.

3.1 LINEAR REGRESSION



95% CONFIDENCE INTERVALS OF ESTIMATES

The estimates of the regression coefficients are subject to sampling uncertainty. Therefore, we will never exactly estimate the true value of these parameters from sample data in an empirical application.

$$popularity = \beta_0 + \beta_1 duration + \beta_2 acousticness + \beta_3 danceability + \beta_4 instrumentalness + \beta_5 key + \beta_6 liveness + \beta_7 loudness + \beta_8 mode + \beta_9 speechiness + \beta_{10} tempo + \beta_{11} signature + \beta_{12} valence$$

```
##                       2.5 %       97.5 %
## (Intercept)       21.02335848 66.840219245
## song_duration     -0.55379036  0.081527415
## acousticness      -1.85634654  0.874759608
## danceability       5.91327365 10.653819518
## instrumentalness  -9.51570149 -6.534399439
## keyC#              0.09317867  2.860324122
## keyD              -1.93140269  0.912922988
## keyD#             -2.41969938  1.849930373
## keyE              -1.55046241  1.543478253
## keyF              -1.02261604  1.930135468
## keyF#              0.21804128  3.344740038
## keyG              -1.89174101  0.827743480
## keyG#             -1.69710143  1.407885949
## keyA              -1.94634586  0.905511025
## keyA#             -0.47494157  2.663235599
## keyB              -0.60585479  2.399644041
## liveness          -7.19640631 -2.629527459
## loudness           0.01782611  0.231851260
## audio_modeminor   -1.51143870 -0.095417543
## speechiness       -7.68089694 -1.072140490
## tempo             -0.02768363 -0.004563098
## time_signature1  -15.51391182 31.225121872
## time_signature3  -14.30007516 31.562886962
## time_signature4  -13.29221530 32.482206635
## time_signature5  -12.40553251 33.691644389
## audio_valence     -8.92484711 -6.042788933
```

3.1 LINEAR REGRESSION

Backward variable selection gave us the following model to work with:

$$popularity = \beta_0 + \beta_1 duration + \beta_2 key + \beta_3 mode + \beta_4 speechiness + \beta_5 tempo + \beta_6 loudness$$
$$+\beta_7 liveness + \beta_8 danceability + \beta_9 valence + \beta_{10} instrumentalness$$

The attributes that have the largest effect on song popularity using this model are:

- **Danceability** with 8.481296, meaning with each unit increase, song popularity increases by roughly 8.48;
- **Instrumentalness** with -8.017524, meaning with each unit increase, song popularity decreases by roughly 8.02;
- **Audio valence** with -7.460624, meaning with each unit increase, song popularity decreases by roughly 7.46.

3.1.1 REVIEW

Once we are done with training our model, we can not just assume that it is going to work well on data that it has not seen before. In other words, we can't be sure that the model will have the desired accuracy and variance in production environment. We validate our model using 5-fold Cross Validation to calculate RMSE:

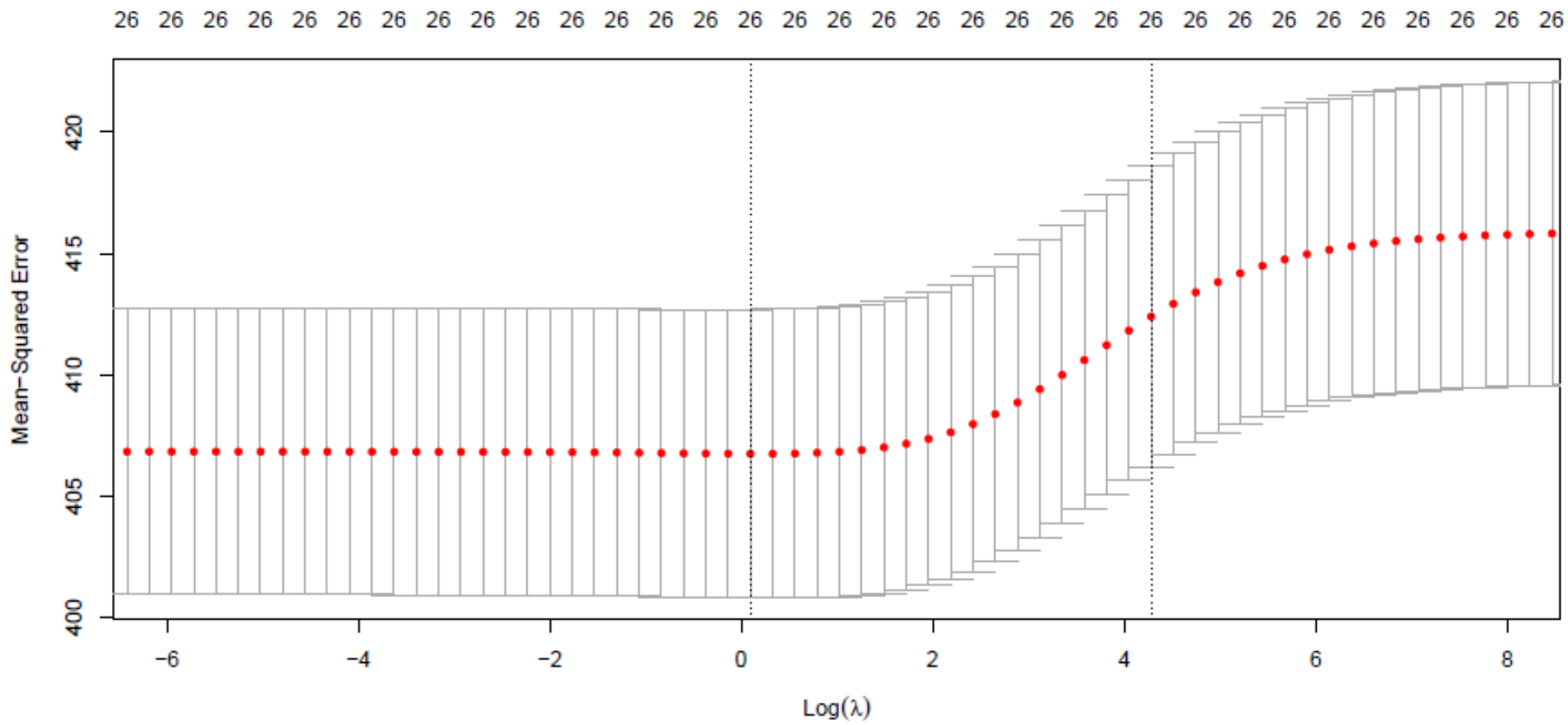
```
##                Root Mean Squared Error
## Full model                    20.18663
## Reduced Model                 20.17511
```

Because we couldn't get a lower error, we suspect that the reason for this is either removing the energy variable (although it seemed justified at the time), or that the variable selection is not good enough, so we further use Ridge regression to impose a penalty term on predictors to impose a bias on the estimators and LASSO for the latter.

3.2 RIDGE REGRESSION

The least square method shown previously finds the coefficients that best fit the data. We are trying to find coefficient estimates that would minimize the error in the previous linear model, by adding a shrinkage penalty called L2-norm, which is the sum of the squared coefficients $\lambda \sum_{j=1}^{p} \beta_j^2$, where λ is a constant that can fine-tune amount of the penalty. We use all predictors in hope to bias the estimates and yield a better result.

Since the optimal value for $\log(\lambda)$ obtained by 5-fold CV is very close to 0, this model is equivalent to the previous.

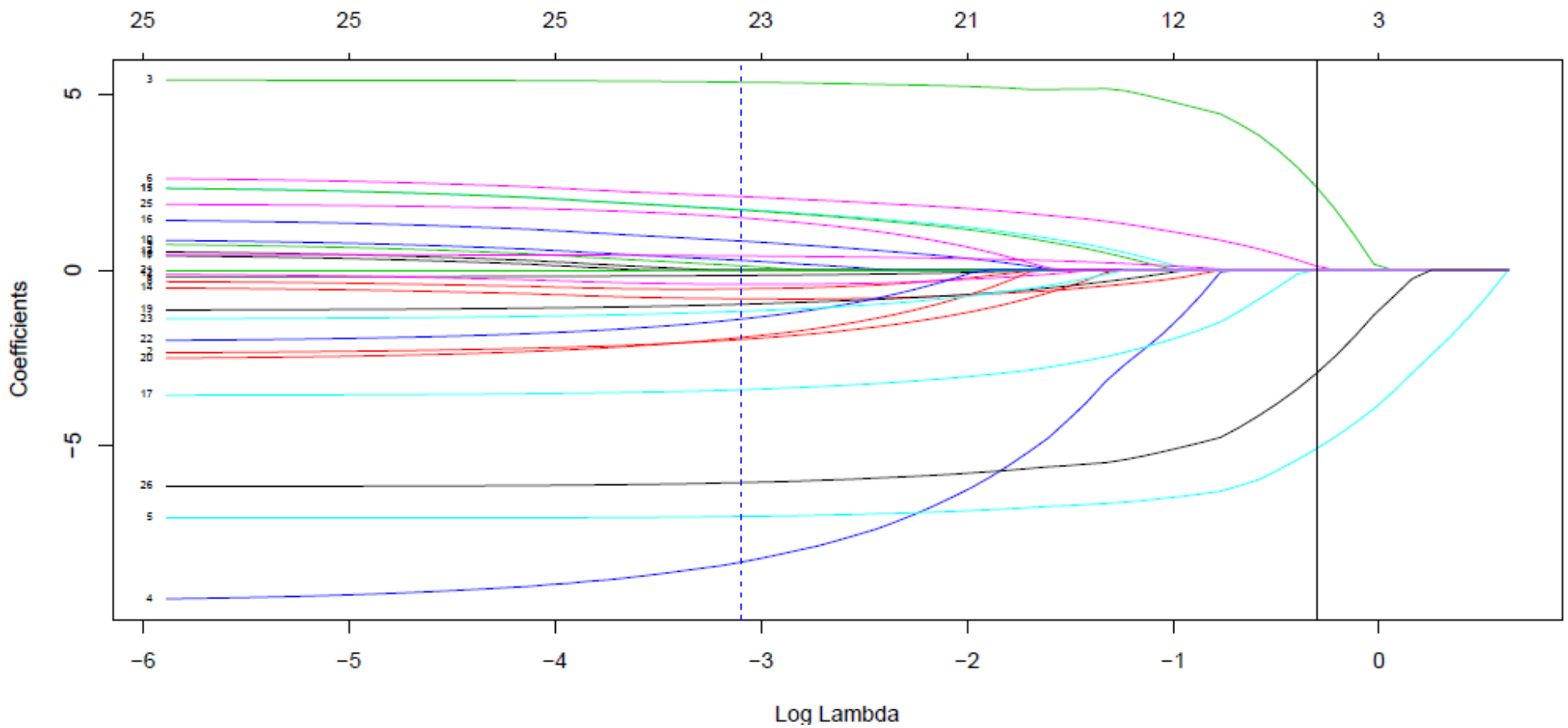


3.3 LASSO

We saw that ridge regression with a wise choice of lambda didn't really outperform least squares. Now, we're interested in whether the LASSO can yield either a more accurate and a more interpretable model. We are trying to find coefficient estimates that would minimize the error in the linear model, by adding a shrinkage penalty called L1-norm, which is the sum of the absolute coefficients $\lambda \sum_{j=1}^{p} |\beta_j|$, where λ is a constant as in Ridge Regression.

Using 5-fold CV the optimal $\log(\lambda)$ obtained is nearly -3, and the model obtained is

$$popularity = \beta_0 + \beta_1 dacenability + \beta_2 instrumentalness + \beta_3 key_C + \beta_4 valence$$



3.4 REVIEW

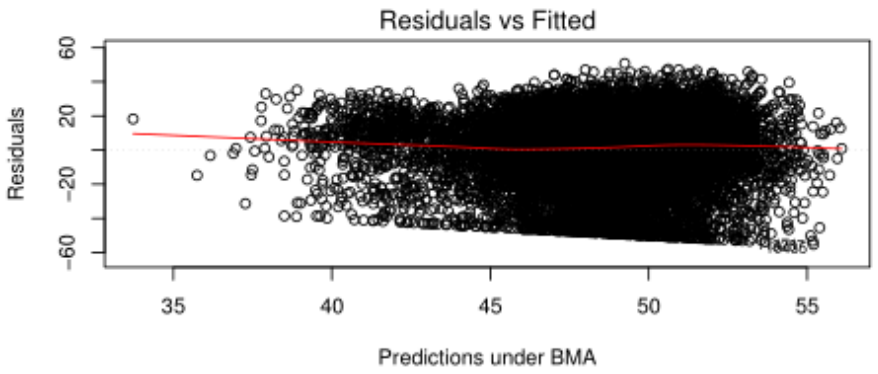
We see that what makes a song popular, according to LASSO, is people's perception of danceability and valence. As mentioned previously, songs that are more danceable, such as club songs which we sing to, end up being more popular than the ones containing no vocals. We see LASSO considers whether a song has an overall key of C as an important factor to our analysis. We see that 80% of the observations with an overall musical key C are in fact with an audio mode in major. This means that people tend to enjoy innocently happy, childlish-like songs, and LASSO studies this as important. The results from these models are the following:

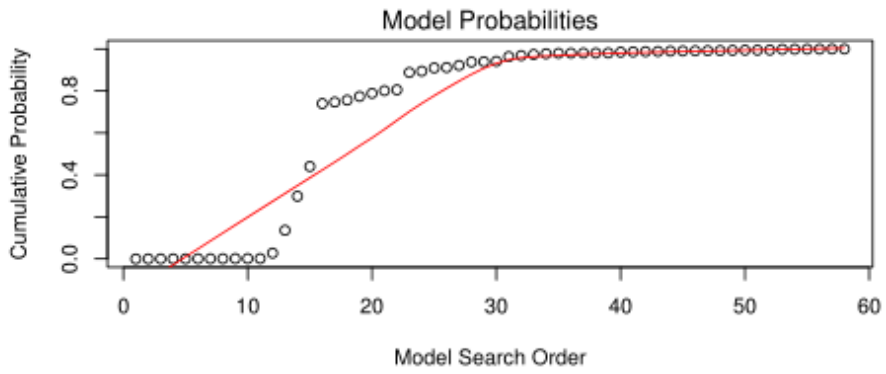
```
##                   Root Mean Squared Error
## Linear Model                     20.18596
## Ridge Regression                 20.24823
## LASSO                            20.20546
```

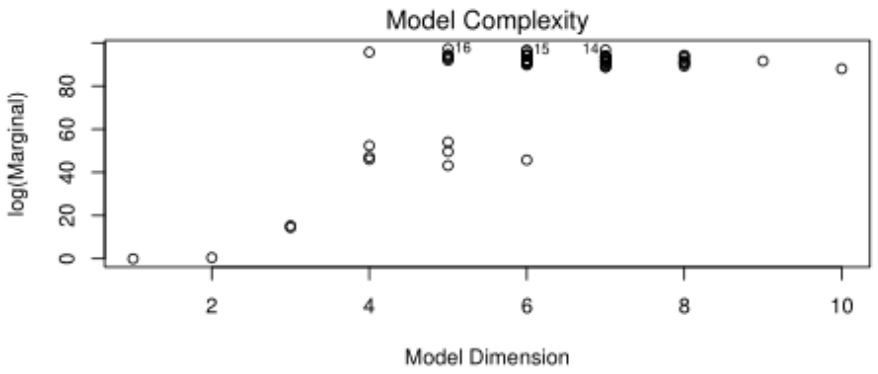
3.5 BAYESIAN APPROACH

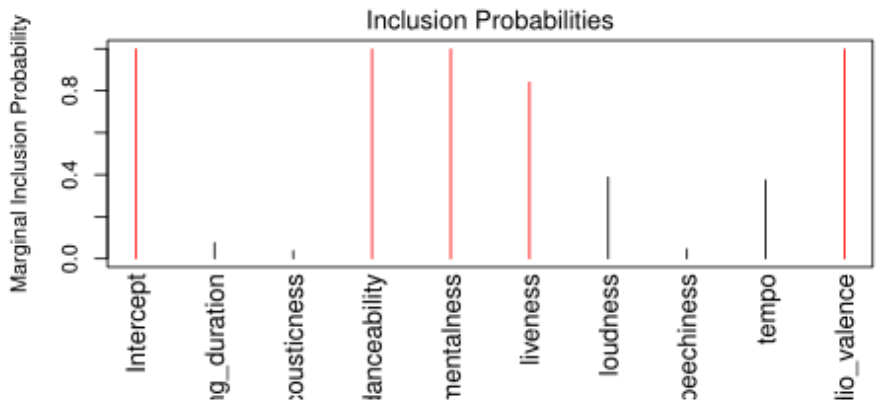
We show another approach which is more intuitive and close to how we think about probability in everyday life and yet is a very powerful tool: Bayesian statistics, which has its foundations on conditional probability and Bayes theorem. There is a key element when we want to build a model under Bayesian approach: the Bayes factor - the ratio of the likelihood probability of two competing hypotheses (usually null and alternative hypothesis) and it helps us to quantify the support of a model over another one.

We analyze the residuals and see how they differ from the regular multivariate linear regression observed previously.

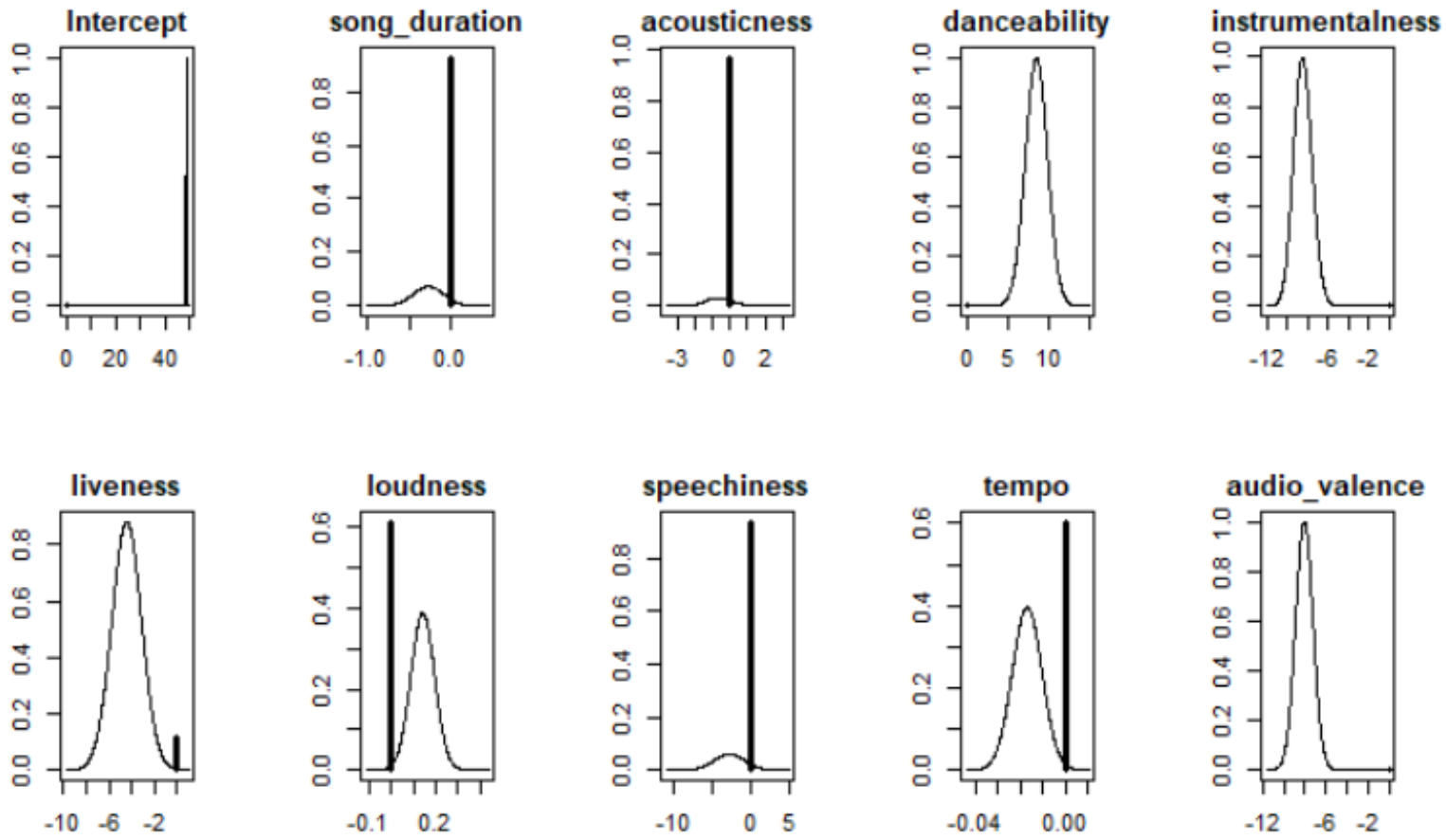








Obtaining the coefficient estimates and standard deviations to be able to examine the marginal distributions for the significant predictors:



3.6 RESULTS AND DISCUSSION

MODEL	RMSE
LINEAR	20.19
RIDGE	20.25
LASSO	20.21
BAYES	20.14

3.6 RESULTS AND DISCUSSION

What we can conclude is that with an increase in the value of danceability, a song gains in popularity. This has been confirmed by all models. As has the fact that if a song is more instrumental it gains less popularity. What perplexed us was the fact that valence had a negative effect on our target variable. This means that the more happy a song sounds, the less popularity score it will obtain. This might be due to the fact that music features gradual changes along these dimensions (song attributes) within a broader emotion category remaining the same throughout the musical segment. Sometimes, listeners do not even perceive any discrete emotion in the music. They merely perceive a certain level of arousal. What may be of greater impact, according to LASSO, is whether or not the songs has an overall key C#: giving us a strange, but pleasant feeling, sometimes uneasy but freeing. What is certain is that even regularizing our linear regression fit did not yield better results, not even when the statistical analysis is undertaken within the context of Bayesian inference. This gives us the motivation to try predicting popularity from another angle.

Songs in C#

```
##                                                song_name       artist_name
## 9                                          Mr. Brightside       The Killers
## 18                                            Best of You      Foo Fighters
## 28                                           American Idiot         Green Day
## 48                                              Miss Murder               AFI
## 56                                                  The Red          Chevelle
## 62  Taki Taki (with Selena Gomez, Ozuna & Cardi B)          DJ Snake
## 63                                                White Boi    Dillon Francis
## 73                      Banana Clip - Spanish Version             Miguel
## 95                            DÃganle - Tainy Remix      Leslie Grace
## 96                                         Sangria Wine Pharrell Williams
## 100                          Scooby Doo Pa Pa (Remix)            Dj Kass
## 102    1, 2, 3 (feat. Jason Derulo & De La Ghetto)       Sofia Reyes
## 103                               Pa Mi (with Ozuna)        Tory Lanez
## 116                                       Belong To U       Fancy Cars
## 128                                 Hey, Soul Sister             Train
## 134                                    We Can't Stop       Miley Cyrus
## 151                                           No One       Alicia Keys
## 155                 Single Ladies (Put a Ring on It)          BeyoncÃ©
## 157                                 Lose My Breath   Destiny's Child
## 160                                      Die Young             Kesha
```

4. CLASSIFICATION MODELS

We now use classification models to predict the probability a song is highly popular among users (i.e. has a popularity score above 75). We fit our models to predict the probability of a song being very popular: P(Y = high popularity | X), via:

- Logistic Regression
- Quadratic Discriminant Analysis
- K – Nearest Neighbours
- Bayes Classification

The evaluation metrics used are: accuracy, recall (True Positive Rate) and precision. Instead of cross validation, we see how the models generalize using a single validation (test) set approach.

4.1 CHOICE OF MODEL

Opposite to the linear regression model at the beginning, different predictors turned out significant when we fit the full logistic regression model. Not all had a great contribution to the response variable, so post appropriate hypothesis testing – the following model obtained through backward variable selection was chosen:

$$popularity = \beta_0 + \beta_1 duration + \beta_2 valence + \beta_3 acousticness + \beta_4 danceability + \beta_5 energy + \beta_6 instrumentalness + \beta_7 loudness$$

4.1 LOGISTIC REGRESSION

The estimates now give the change in the log odds of the outcome for a one unit increase in the predictor variable. Here we report the most significant:

- Valence, the log odds of a song being popular decreases by 0.9;
- Acousticness, the log odds of a song being popular decreases by 0.75;
- Danceability, the log odds of a song being popular increases by 1.87;
- Energy, the log odds of a song being popular decreases by 2.53;
- Instrumentalness, the log odds of a song being popular decreases by 3.82.

95% CONFIDENCE INTERVALS IN TERMS OF THEIR EFFECT ON THE LOG ODDS OF A SONG BEING POPULAR

```
##                       2.5 %      97.5 %
## (Intercept)      0.359482462  1.93826954
## song_duration    1.007119619  1.17305811
## audio_valence    0.284098701  0.58063451
## acousticness     0.319900597  0.69279170
## danceability     3.767588602 11.30269482
## energy           0.040027017  0.15653507
## instrumentalness 0.005713239  0.06429316
## loudness         1.208426705  1.31519304
```

RESULTS ON TRAINING AND TESTING SET

TRAINING PREDICTIONS

TP = 565	FN = 220
FP = 4398	TN = 6758

TESTING PREDICTIONS

TP = 144	FN = 52
FP = 1114	TN = 1675

RESULTS ON TRAINING AND TESTING SET

METRIC	TRAINING	TESTING
AUC	71.7%	72%
ACCURACY	61.33%	60.94%
TPR/RECALL	71.97%	73.47%
PRECISION	11.38%	11.45%

4.2 QDA

We now try to fit a non-linear boundary between classifiers. To do so, we perform a Quadratic Discriminant Analysis on our dataset, assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. It assumes that each class has its own covariance matrix.

```
Prior probabilities of groups:
       low       high
0.93426011 0.06573989

Group means:
     song_duration audio_valence acousticness danceability    energy instrumentalness  loudness
low       3.648959     0.5274530    0.2758200    0.6218511 0.6387022       0.09778559 -7.772139
high      3.683711     0.5036536    0.2003206    0.6707682 0.6615371       0.01084694 -6.161431
```

RESULTS ON TESTING SET

TP = 28	**FN = 168**
FP = 144	**TN = 2645**

METRIC	SCORE
ACCURACY	89.55%
TPR/RECALL	14.29%
PRECISION	16.28%

4.3 Naïve Bayes Classification

We additionally decide to fit a generative model - Naive Bayes, which is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

RESULTS ON TESTING SET

TP = 71	**FN = 125**
FP = 459	**TN = 2230**

METRIC	SCORE
ACCURACY	80.44%
TPR/RECALL	36.22%
PRECISION	13.4%

4.4 K-Nearest Neighbours

Since we suspect that songs that share similar features have approximately the same popularity score, we perform a K-Nearest Neighbour classification for the test set from training set. Unlike most algorithms, this algorithm is non-parametric which means that it does not make any assumptions about the dataset. This makes it more effective since it can handle realistic data.

We set K=1, meaning that our model will classify a song based on most similar features with another already classified observation in the training set.

RESULTS ON TESTING SET

TP = 24	**FN = 172**
FP = 172	**TN = 2617**

METRIC	SCORE
ACCURACY	88.48%
TPR/RECALL	12.24%
PRECISION	12.24%

4.5 RESULTS AND DISCUSSION

METRIC	LOGISTIC	QDA	BAYES	1-NN
ACCURACY	60.94%	89.55%	80.44%	88.48%
TPR/RECALL	73.47%	14.29%	36.22%	12.24%
PRECISION	11.45%	16.28%	13.4%	12.24%

4.5 RESULTS AND DISCUSSION

Looking at the results of the classifiers, we can say with certainty that QDA performed the best in term of accuracy. We get a similar accuracy using 1-NN, but we suspect that setting K=1 induces high variance and an overly flexible decision boundary. Since we have a sufficient number of training examples, and the variance of the classifier is not a major concern, QDA serves as a compromise between K-NN and the logistic regression approach. Though not as flexible as KNN, it still yielded better results, including the ones for precision and recall. But, by fitting the Naive Bayes Classification model, we obtained an accuracy rate of somewhere in between these, but a recall (TPR) which outperformed both QDA and 1-NN. This might be due to the naive assumptions, i.e. independence among features, but speed does come at a cost. Logistic regression gives a recall rate almost 5 times larger than QDA and twice as large in comparison to Bayes. This does comes as a surprise, since it supports only linear solutions, but we suspect that this is due to the classes being unbalanced.

5. NOTES

There appear to be other factors missing that could potentially help in determining whether a song is popular or not, as well as it popularity score in the regression case.

Factors that might contribute could be:

- Whether an artist has had previous hits;
- Collaborations between artists;
- Genre of the song;

and similar.

THANK YOU