EXPLICIT CONTENT PREDICTOR IN SONGS

Introduction

We want to know if “popular songs” with explicit content in it can be statistically proven as different from clean songs based on their musical features, as well as the relations and correlations between this features.

To do this we will use a dataset of the top 100 Billboard songs from each week since 1958.

It shows 22 attributes related to the songs. From the author, song name or genre to other technical aspects like tempo, loudness or danceability.

It is curious how the language has some expressions or words marked as “not suitable for every audience”, but do these expressions affect the musical properties of a song itself? Our objective is to predict if a song uses explicit content based on its instrumental features. We, as music consumers (we all consume music daily), are motivated to find out if there are correlations between the variables that we have (song features), and in specific, the explicitness of the songs.

After having developed our models, we will want to check our results by applying the predictions in songs that we are interested in (taking into account that the study will be performed on popular songs, therefore the result should, in theory, be more accurate if the song is well known). We will use the spotify “Audio Features for Tracks” tool for developers, with which we will make tests for songs we like or that we are looking forward to analyze.

Dataset

Our data set contains 22 parameters but for the analysis we are interested in the last 15 ones corresponding to:

1. Explicitness *- factor*
2. Duration (in ms)
3. Popularity (between 0 and 100)
4. Danceability (between 0 and 1): an estimation based on other parameters about how danceable a song seems to be.
5. Energy (between 0 and 1): a value describing how energetic a song is.

6. Key: in which key is the song made (0 corresponds to C, and so on for the following keys in musical scale).
7. Loudness (in dB): the average loudness of the song, usually varies from -60 to 0 dB.
8. Mode *- factor* : Mode indicates the modality (major or minor). Major is represented by 1 and minor is 0.
9. Speechiness (from 0 to 1): a measure of spoken words or spoken-style singing found in the song (rap → 1, melodic → 0).
10. Acousticness (from 0 to 1): how acoustic (not electronic) a song is.
11. Instrumentalness (from 0 to 1): a measure of the amount of vocals a song has.
12. Liveness (from 0 to 1): probability that the song was recorded live.
13. Valence (from 0 to 1): measures the positivity of a sound, if it is happy, cheerful, euphoric etc.
14. Tempo (in Beats Per Minute)
15. Time signature: beats per bar of the song.

Missing data

The dataset that we've chosen isn't complete (like most datasets), there are many rows with missing parameters. Most of the rows with missing parameters had many of them, almost every time a column is missing, the rest of them are also missing. This is why we have chosen to delete the rows where the target variable (explicit) or "danceability" are missing.

Exploration process

In a first approach to understand the data we are using we will extract some information before the data treatment.

We are going to set some *a priori* hypothesis from simple histogram based analysis from the pre processed data to see the difference between the distribution of these variables on explicit vs non explicit.

By seeing the distribution of the initial data we can estimate some possible outcomes and detect singular distributions in the data. In the following plots the red bars correspond to explicit songs and the green to non-explicit.

When observing the danceability histogram (*see fig.1*) it can be seen that the explicit songs tend to a higher level of danceability centered around 0.75 while non-explicit songs is centered at 0.6.

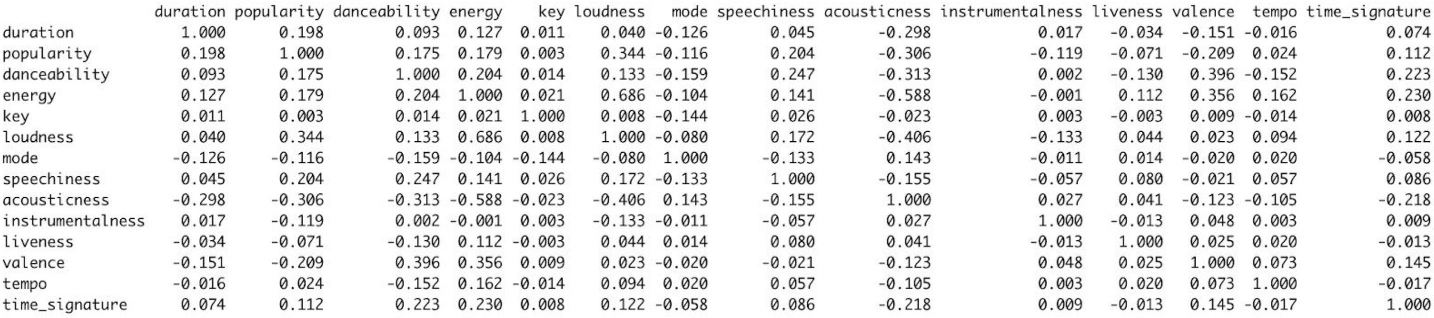
Looking at the popularity histogram (*see fig.2*) it can be seen that explicit songs tend to be more popular than non explicit ones. In the non-explicit case the popularity seems uniform at the beginning but then tends to zero after passing the value 60. In the other case the explicit songs shows kind of a gaussian curve centered at 65 with a high concentration of values between 50 and 80.

Looking at the data from the key values (*see fig.3*) we see that there is a general trend, specially in the explicit songs for the 0 key (C). The remaining data shows a more or less uniform distribution with a minimum on the 2nd key.

Looking at the loudness of the data (*see fig.4*) it can be seen that non-explicit songs are less loud in general but have a greater variance. Explicit songs are louder and use a smaller range of values.

Regarding to the speechiness of the songs (*see fig.5*) we see that in the non-explicit case, songs are mostly non-spoken at all. The explicit songs are widely distributed over the axis which can be thought as urban songs that usually use a stronger language and have a greater speechiness index.

We will also take a look at the correlation matrix of the data.



```
                 duration popularity danceability energy    key loudness   mode speechiness acousticness instrumentalness liveness valence  tempo time_signature
duration            1.000      0.198        0.093  0.127  0.011    0.040 -0.126       0.045       -0.298            0.017   -0.034  -0.151 -0.016          0.074
popularity          0.198      1.000        0.175  0.179  0.003    0.344 -0.116       0.204       -0.306           -0.119   -0.071  -0.209  0.024          0.112
danceability        0.093      0.175        1.000  0.204  0.014    0.133 -0.159       0.247       -0.313            0.002   -0.130   0.396 -0.152          0.223
energy              0.127      0.179        0.204  1.000  0.021    0.686 -0.104       0.141       -0.588           -0.001    0.112   0.356  0.162          0.230
key                 0.011      0.003        0.014  0.021  1.000    0.008 -0.144       0.026       -0.023            0.003   -0.003   0.009 -0.014          0.008
loudness            0.040      0.344        0.133  0.686  0.008    1.000 -0.080       0.172       -0.406           -0.133    0.044   0.023  0.094          0.122
mode               -0.126     -0.116       -0.159 -0.104 -0.144   -0.080  1.000      -0.133        0.143           -0.011    0.014  -0.020  0.020         -0.058
speechiness         0.045      0.204        0.247  0.141  0.026    0.172 -0.133       1.000       -0.155           -0.057    0.080  -0.021  0.057          0.086
acousticness       -0.298     -0.306       -0.313 -0.588 -0.023   -0.406  0.143      -0.155        1.000            0.027    0.041  -0.123 -0.105         -0.218
instrumentalness    0.017     -0.119        0.002 -0.001  0.003   -0.133 -0.011      -0.057        0.027            1.000   -0.013   0.048  0.003          0.009
liveness           -0.034     -0.071       -0.130  0.112 -0.003    0.044  0.014       0.080        0.041           -0.013    1.000   0.025  0.020         -0.013
valence            -0.151     -0.209        0.396  0.356  0.009    0.023 -0.020      -0.021       -0.123            0.048    0.025   1.000  0.073          0.145
tempo              -0.016      0.024       -0.152  0.162 -0.014    0.094  0.020       0.057       -0.105            0.003    0.020   0.073  1.000         -0.017
time_signature      0.074      0.112        0.223  0.230  0.008    0.122 -0.058       0.086       -0.218            0.009   -0.013   0.145 -0.017          1.000
```

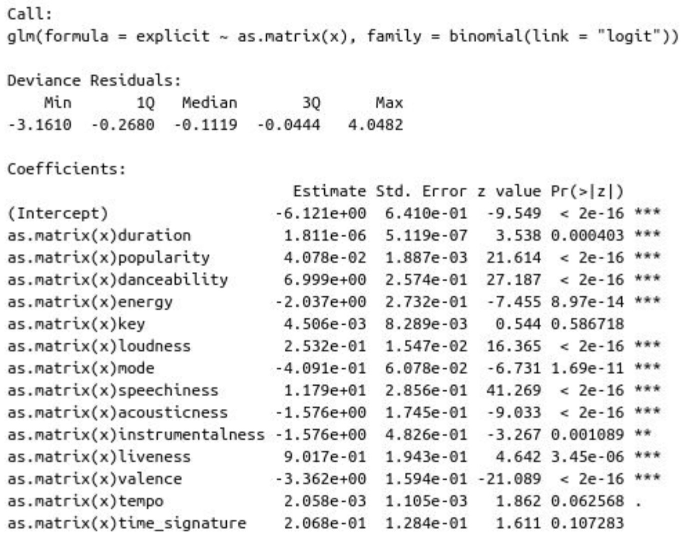
If we take a closer look we can see that the variables are in general low correlated but there are some special cases where the variables seem to have some kind of correlation. We can see this between the variables energy and loudness with a correlation value of 0.686 and energy and acousticness with -0.588. It is logical to think that, as shown, energetic songs are loud and tend to be more electronic and less acoustic.

Variable selection

In order to reduce the amount of variables we are using we will define a first GLM of the data. By doing this we can get a first raw approach of how the variables behave corresponding to their effect on the explicitness of the songs.

For our following models and tests we will be using an alpha value of 90%.

We have defined the following model:



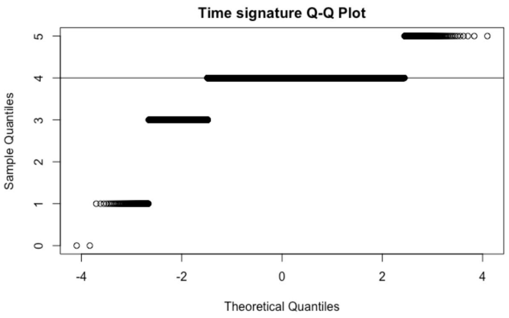
```
Call:
glm(formula = explicit ~ as.matrix(x), family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1610  -0.2680  -0.1119  -0.0444   4.0482

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.121e+00  6.410e-01  -9.549  < 2e-16 ***
as.matrix(x)duration         1.811e-06  5.119e-07   3.538 0.000403 ***
as.matrix(x)popularity       4.078e-02  1.887e-03  21.614  < 2e-16 ***
as.matrix(x)danceability     6.999e+00  2.574e-01  27.187  < 2e-16 ***
as.matrix(x)energy          -2.037e+00  2.732e-01  -7.455 8.97e-14 ***
as.matrix(x)key              4.506e-03  8.289e-03   0.544 0.586718
as.matrix(x)loudness         2.532e-01  1.547e-02  16.365  < 2e-16 ***
as.matrix(x)mode            -4.091e-01  6.078e-02  -6.731 1.69e-11 ***
as.matrix(x)speechiness      1.179e+01  2.856e-01  41.269  < 2e-16 ***
as.matrix(x)acousticness    -1.576e+00  1.745e-01  -9.033  < 2e-16 ***
as.matrix(x)instrumentalness -1.576e+00  4.826e-01  -3.267 0.001089 **
as.matrix(x)liveness         9.017e-01  1.943e-01   4.642 3.45e-06 ***
as.matrix(x)valence         -3.362e+00  1.594e-01 -21.089  < 2e-16 ***
as.matrix(x)tempo            2.058e-03  1.105e-03   1.862 0.062568 .
as.matrix(x)time_signature   2.068e-01  1.284e-01   1.611 0.107283
```

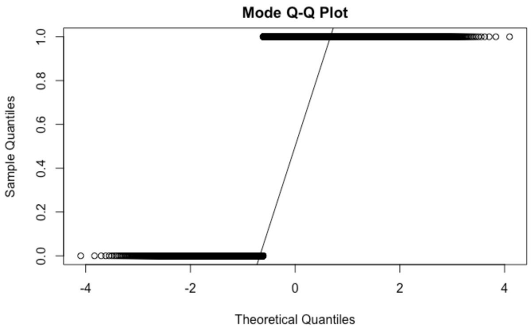
Looking at the summary of the model and focusing on its ANOVA test we can see that the variables key and time signature are not significant to the model. It is logical to think that the key in which the song is made does not imply the existence of explicitness and that the time structure of the song doesn’t affect the resulting language used. We can see that variables like danceability and popularity, among others, are strongly significant.

This is not our final model, yet it has given us important information.

Another important thing to see is the data distribution. We assume that the data follows a normal distribution but we need to check this for all variables.

With the *qqnorm* and *qqline* functions in R we can see how well the data adjusts compared to a normal distribution. We have tested this on all variables and we have seen two special cases.





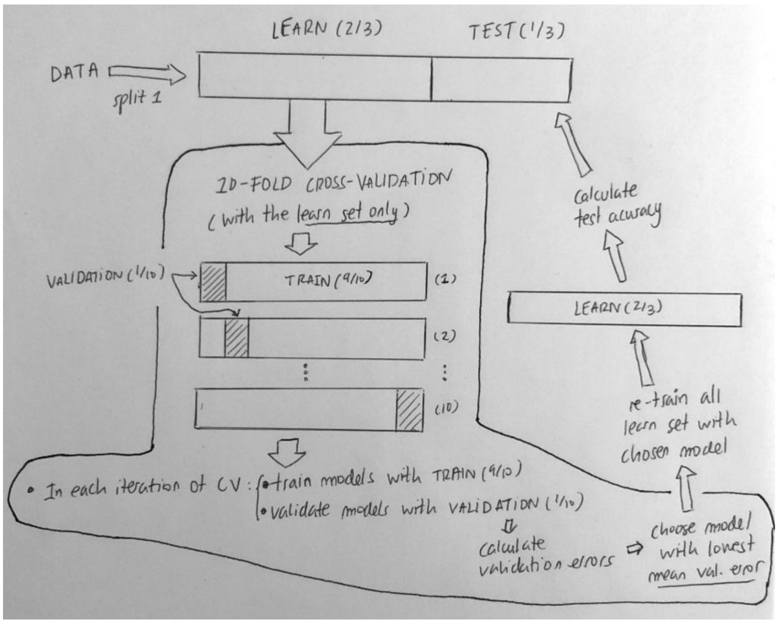
The variables time signature and mode differ from the normal distribution. In the mode case, our variable only takes 0 or 1 values with a preference on 1. It is clearly not normal but it is a factor that can be applied later on and can be useful, furthermore the ANOVA test proved its significance.

For the other variable, time signature, we can see how the values do not correctly adjust to the normal distribution. In addition to the fact that it was not significant this variable is not suitable for our project.

From now on we will not work with the variables key and time signature.

Training the dataset

We are going to consider a k-fold cross-validation to choose the model that best fits our data to then use it to classify the output to decide whether a song is explicit or not. To do so, we will first split the data in two sets, the learning one which has ⅔ of the songs and will be used to apply the 10-fold cross-validation to choose the best model, and the test set which has ⅓ of the songs and will be used to test the goodness of fit (test accuracy) of our final classifications with the final model chosen by the cross-validation which will have previously been trained with all the ⅔ of learn data.


LEARN (2/3)
TEST (1/3)
DATA
split 1
10-FOLD CROSS-VALIDATION
(with the learn set only)
VALIDATION (1/10)
TRAIN (9/10)
(1)
(2)
(10)
Calculate test accuracy
LEARN (2/3)
re-train all learn set with chosen model
• In each iteration of CV: • train models with TRAIN (9/10)
• validate models with VALIDATION (1/10)
Calculate validation errors
choose model with lowest mean val. error

These are the models that we will use in the cross-validation, from which we will choose the best one to train the whole learn set with it:

a) Model 1 → Classification using a two-group Dichotomizer, with the only assumption that the data is normal, which we have proven to be for most explanatory variables as we have seen with the *qqplots*. The dichotomizer is built with the following discriminant functions for each group k:

$$g_k(x) = ln\{P(w_k)p(x|w_k)\} = ln\,P(w_k) - ln\{(2\pi)^{d/2}|\Sigma_k|^{1/2}\} - 1/2(x-\mu_k)'\,\Sigma_k^{-1}(x-\mu_k)$$

and defining the dichotomizer as $g(x) = g_1(x) - g_2(x)$, where a positive result means the result of the classification is explicit and otherwise for the negative case.

This first model would be the equivalent to applying QDA classification (R function) but we have decided to implement it ourselves to understand better the process behind.

We learned a lot by doing this because we spent a whole afternoon looking for errors in the code and it was only a parenthesis.

b) Model 2 → Regression through a GLM using the binomial family (1 → explicit, 0 → non explicit) with link *logit* with witch we will use it's predictions to classify to both groups using the *round* function (we get numbers between zero and one, and classify to the closest group to the prediction).

c) Model 3 → We will consider using a third model of classification using k-nearest-neighbours (knn). In order to speedup our program we will fix our k-value before executing the model selection (10-fold cross validation). Using the full learning data and contrasting with the test set we have found the best value (least error) to be k=11.

Results obtained

After the 10-fold cross validation, all the models gave almost the same results in terms of mean validation errors (~ 7%) and F1 scores (~96%)

CV Mean Error	CV Mean F1 score
0.0705032	0.9606019
0.06899012	0.9619048
0.06772964	0.9625241

The variances of the errors

	Error Variance	F1 Variance
Model 1 (qda dichotomizer)	5.519058e-05	1.721999e-05
Model 2 (glm binomial)	7.213955e-05	2.253331e-05
Model 3 (11-nn)	3.571098e-05	1.159235e-05

The Central limit theorem tells us that the mean and the variance are asymptotically normal, so we are going to test if the mean errors of each model are significantly different.

To see this, we are going to test if the difference in the mean of errors in every iteration of the c.v. is equal to 0. We can do that with the R function t.test.

These are the results:

Hypothesis	p-value
Error_dichotomizer = Error_glm ?	0.6767
Error_dichotomizer = Error_11-NN ?	0.3703
Error_11-NN = Error_glm ?	0.7061

So we can conclude that there's no significant difference between the validation error of the three models as al p-values are > alpha = 0,1.

The Decision - Model Choice

In terms of error rates and variances we have seen how the best model is *knn* but by a small difference. Because the three models are so similar, if we wanted to minimize the complexity of the model, we would choose the second model (logistic regression).

Therefore, we will proceed to train our whole learning dataset (⅔ of the total) with the chosen model (depending on what we prioritize {complexity vs few less error}) and test its accuracy with the test set. Furthermore, once we have the model trained, we will be able to download track features from the Spotify Developer API (url in *References*) from which we can acquire the explanatory variables of our data and therefore use the already trained and tested model to classify that song.

References

The data is available at: Billboard Hot weekly charts - dataset by kcmillersean

Spotify audio features obtainment: Get Audio Features for a Track

Some of the concepts used in the project not presented in the course: F1 score

Most of the mathematicals models and procedures used in this project are based on the lectures of Machine Learning 1.

Annex

fig.1

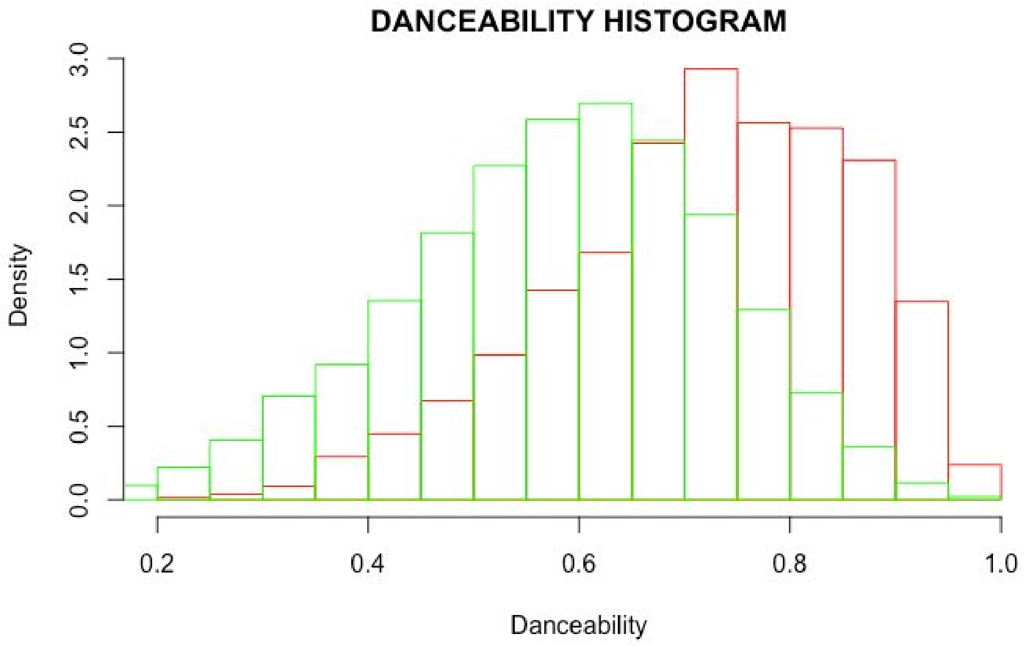


fig.2

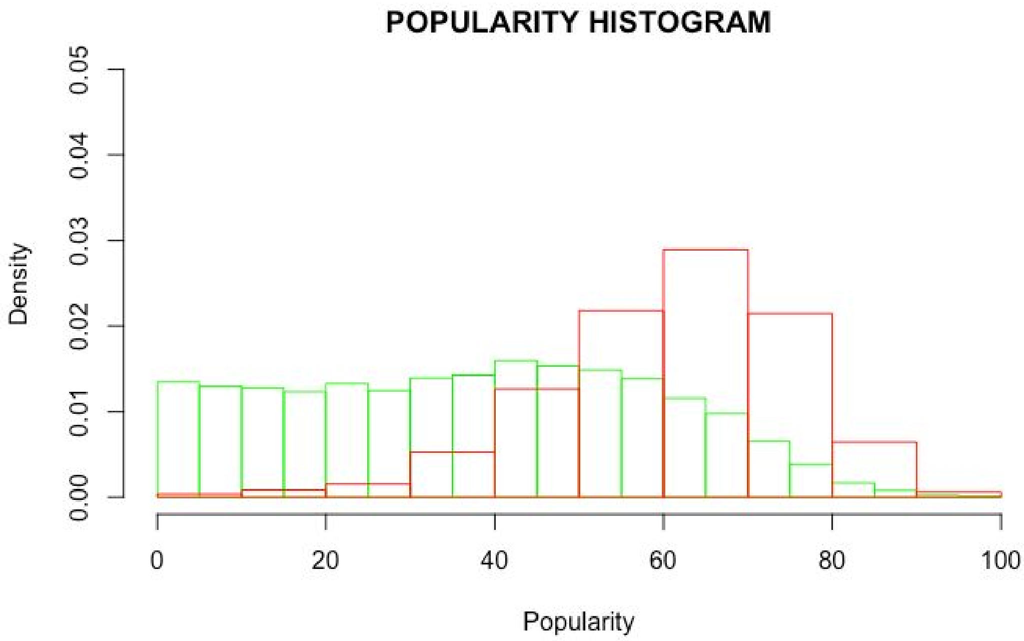


fig.3

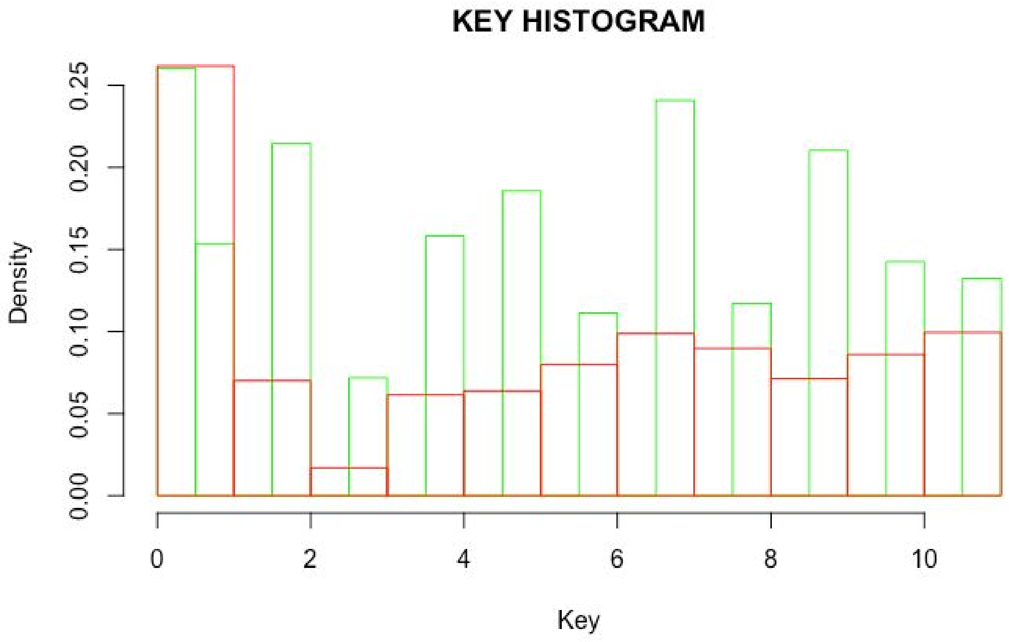

KEY HISTOGRAM
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
Key

fig.4

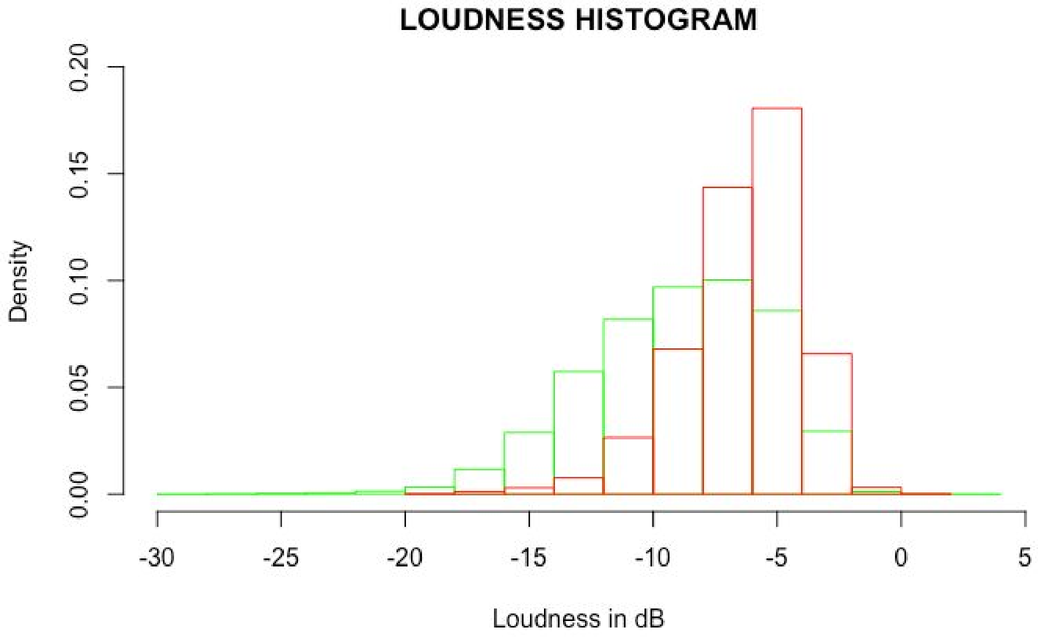
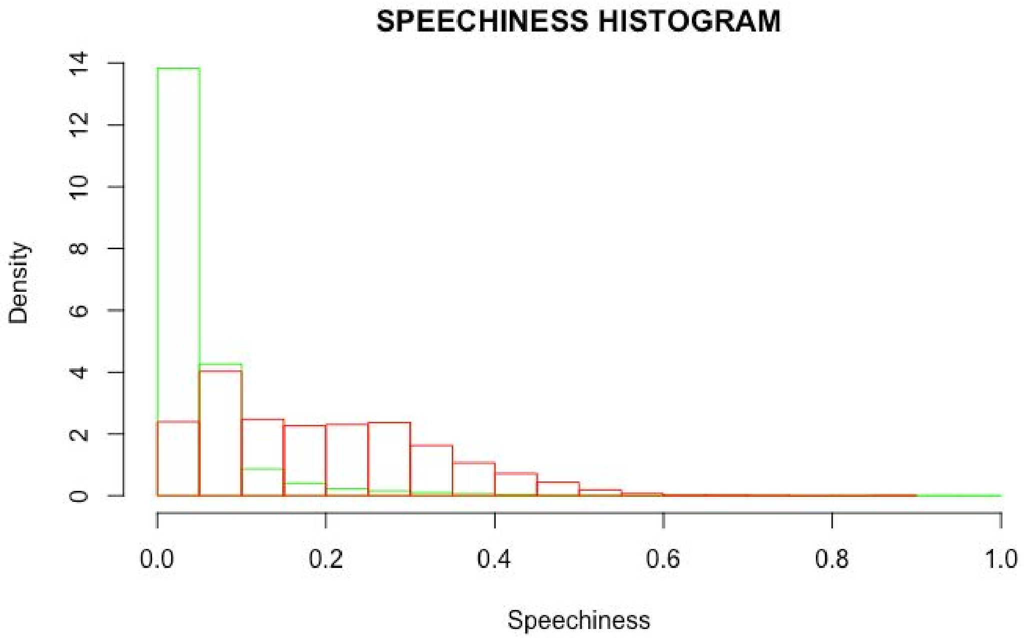

LOUDNESS HISTOGRAM
Density
0.00
0.05
0.10
0.15
0.20
-30
-25
-20
-15
-10
-5
0
5
Loudness in dB

fig.5


SPEECHINESS HISTOGRAM
Density
0
2
4
6
8
10
12
14
0.0
0.2
0.4
0.6
0.8
1.0
Speechiness