BrazilSpeaks

Gabriel Saruhashi

3/17/2019

Intro

On April 1, 1964, the military organized a coup d'état that overthrew the government of president João Goulart. That day marked the beginning of the Military Dictatorship that lasted for twenty-one years. Under the pretext of eliminating the growing Communist threat, the regime suppressed freedom of speech and imposed rigorous censorship over the all forms of media. In the late 60s, with the popularization of television and radio stations, music began to have a lot of influence over society and, for this reason, it was heavily monitored by the regime's censors. On the one hand, there was a group of musicians that simply conformed to the oppressive rules of the regime. Inspired by the soft rock melodies by the Beatles, they avoided political themes and made fortunes composing songs about love and trivial, middle-class concerns. Yet, on the other hand, a group of musicians stood out in the fight against oppression. Through their music, they conveyed a message of criticism against the regime. Their "protest music" denounced blatant social injustices, mobilized political passions, praised the individual and collective heroes who fought the oppressors. In this project, I will be analyzing this dataset I built

DATA

Data Scraping & Collection

To collect the data, I followed the work done by Carocha (2006). I built a Python script that called the Spotify API and ran a data enrichment pielin

I compiled two Spotify playlists, one for each class of music. Through the Spotify API, I obtained key features of each song, such as speechness, danceability and energy, that are measured in a scale of 0.0 to 1.0. However, Spotify does not directly provide the lyrics for each of the songs. To circumvent this limitation, I built a parallel pipeline that, given a song name ands its author, scrapes song lyrics from Genius and Vagalume, two well-known music platform that provide lyrics and song annotations. The procedure yielded a corpus of 280 songs equally divided in the two categories: 140 censored and 140 uncensored songs.

Overview of the data

Upon loading the data, we observe the following structure:

1. song_sp_uri (chr): a unique identifies song in the spotify platform

- song_name (chr): the name of the song
- song_isrc (chr): the International Standard Recording Code for the song
- song_popularity (int) : provided by the Spotify API, "The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are"
- song_lyrics (chr): the lyrics of the song scraped from Genius and Vagalume
- class (chr): the class of the song (either protest or Young Guard) according to the definition presented in the intro
- danceability (num): provided by the Spotify API, "a value of 0.0 is least danceable and 1.0 is most danceable."

- energy (num): provided by the Spotify API, “energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity”
- key (int): provided by the Spotify API, “the estimated overall key of the track”
- loudness (int): provided by the Spotify API, “the overall loudness of a track in decibels (dB)”
- mode
- speechiness (num): provided by the Spotify API, “float Speechiness detects the presence of spoken words in a track”
- acousticness (num): provided by the Spotify API, “A confidence measure from 0.0 to 1.0 of whether the track is acoustic”
- instrumentalness (num): provided by the Spotify API, “predicts whether a track contains no vocals”
- liveness (num): provided by the Spotify API, “detects the presence of an audience in the recording. . Higher liveness values represent an increased probability that the track was performed live”
- valence (num): provided by the Spotify API, “a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track”
- tempo (num): provided by the Spotify API, “the overall estimated tempo of a track in beats per minute”
- id (chr): the Spotify ID for the artist
- duration_ms: provided by the Spotify API, “the duration of the track in milliseconds”
- time_signature (int)
- artist_genres (chr): provided by the Spotify API, “a list of the genres the artist is associated with
- artist_name (chr): the name of the artist
- artist_photo: (chr): url to the photo of the artist
- artist_popularity (int): provided by the Spotify API, “the value will be between 0 and 100, with 100 being the most popular.”
- artist_sp_followers (int): 542214 542214 542214 299597 829961 532021 542214 16490 2440436 299597

```
## [1] "Number of dimensions in our dataset (row, col):"
```

```
## [1] 200  26
```

```
unique(music$artist_name[music$class=='Protest'])
```

```
##  [1] "Chico Buarque"        "Belchior"             "Raul Seixas"
##  [4] "Elis Regina"          "Taiguara"             "Legião Urbana"
##  [7] "Zé Keti"              "Caetano Veloso"       "Os Mutantes"
## [10] "Maria Bethânia"       "Jards Macalé"         "Edu Lobo"
## [13] "Paulo Cesar Pinheiro" "Paulinho Da Viola"    "Geraldo Vandre"
## [16] "João Bosco"           "Sérgio Sampaio"       "Milton Nascimento"
## [19] "Gonzaguinha"          "Ivan Lins"            "Joao Do Vale"
## [22] "Gilberto Gil"         "Paulo Diniz"          "Vinícius de Moraes"
## [25] "Tim Maia"             "Djavan"               "Jair Rodrigues"
## [28] "Secos & Molhados"     "Gal Costa"            "Nara Leão"
## [31] "Novos Baianos"
```

```
jg_artists = unique(music$artist_name[music$class=='Jovem Guarda'])

capture.output(jg_artists, file = "data/artists_jg.txt")
```

Create corpus for text mining

```
library(tm)
```

```
## Loading required package: NLP
```

```
print("Creating corpus by collapsing together both protest music and Young Guard music")
```

```
## [1] "Creating corpus by collapsing together both protest music and Young Guard music"
```

```r
jg <- paste(music$song_lyrics[music$class=="Jovem Guarda"], collapse = '')
protest <- paste(music$song_lyrics[music$class=="Protest"], collapse = '')
docs <- Corpus(VectorSource(c(jg, protest)))
str(docs)
```

```
## List of 2
##  $ 1:List of 2
##   ..$ content: chr "terrivel bom parar desse jeito provocar voce nao sabe onde venho terrivel vou di
##   ..$ meta   :List of 7
##   .. ..$ author       : chr(0)
##   .. ..$ datetimestamp: POSIXlt[1:1], format: "2019-05-08 02:46:17"
##   .. ..$ description  : chr(0)
##   .. ..$ heading      : chr(0)
##   .. ..$ id           : chr "1"
##   .. ..$ language     : chr "en"
##   .. ..$ origin       : chr(0)
##   .. ..- attr(*, "class")= chr "TextDocumentMeta"
##   ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
##  $ 2:List of 2
##   ..$ content: chr "pai , afasta mim calice pai , afasta mim calice pai , afasta mim calice vinho ti
##   ..$ meta   :List of 7
##   .. ..$ author       : chr(0)
##   .. ..$ datetimestamp: POSIXlt[1:1], format: "2019-05-08 02:46:17"
##   .. ..$ description  : chr(0)
##   .. ..$ heading      : chr(0)
##   .. ..$ id           : chr "2"
##   .. ..$ language     : chr "en"
##   .. ..$ origin       : chr(0)
##   .. ..- attr(*, "class")= chr "TextDocumentMeta"
##   ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
##  - attr(*, "class")= chr [1:2] "SimpleCorpus" "Corpus"
```

Data Cleaning

Although the song features supplied by the Spotify API were already normalized, I had to perform some preprocessing of the lyric. First, I removed stopwords (e.g 'me', 'I', etc.) from the dataset given that they are so common in the language that their informational value is near zero. Second I cleaned up the Genius lyrics by removing the annotations, punctuations and number.

```r
# Load
library("tm")
library("SnowballC")
library("wordcloud")
```

```
## Loading required package: RColorBrewer
```

```r
library("RColorBrewer")

# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove Portuguese common stopwords
docs <- tm_map(docs, removeWords, stopwords("portuguese"))
# Remove punctuations
```

```
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("mim", "pra", "vai"))
```

Descriptive Plots & Summary Information

Lyric Analysis

Inspired by the analysis conducted by (http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-si
First, I plotted a word cloud with the most frequent words for each class of music. Uncensored music had much more positive lyrics, with words such as love, romance and joy standing out, whereas protest music had more descriptive words such as violence, blood, etc.

```
# Document matrix is a table containing the frequency of the words. Column names are words and row name
dtm_jg <- TermDocumentMatrix(docs[1])
m <- as.matrix(dtm_jg)
v <- sort(rowSums(m),decreasing=TRUE)
d_jg <- data.frame(word = names(v),freq=v)
head(d_jg, 10)
```

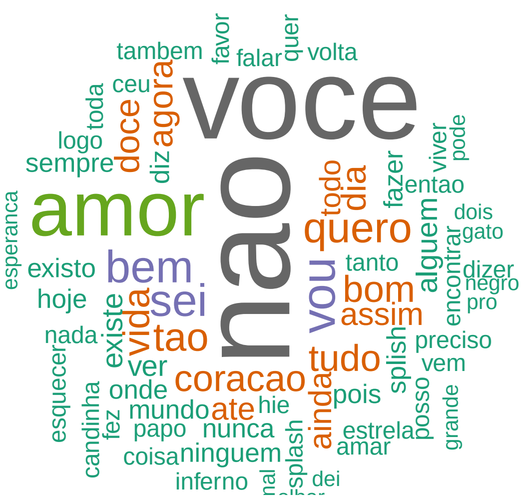
```
##            word freq
## nao         nao  268
## voce       voce  238
## amor       amor  146
## vou         vou   78
## bem         bem   73
## sei         sei   68
## quero     quero   63
## tao         tao   55
## coracao coracao   54
## tudo       tudo   54
```

```
dtm_protest <- TermDocumentMatrix(docs[2])
m <- as.matrix(dtm_protest)
v <- sort(rowSums(m),decreasing=TRUE)
d_protest <- data.frame(word = names(v),freq=v)
head(d_protest, 10)
```

```
##        word freq
## nao     nao  416
## voce   voce  179
## tudo   tudo   94
## gente gente   74
## dia     dia   71
## sera   sera   64
## todos todos   64
## amor   amor   60
## faz     faz   59
## quero quero   58
```

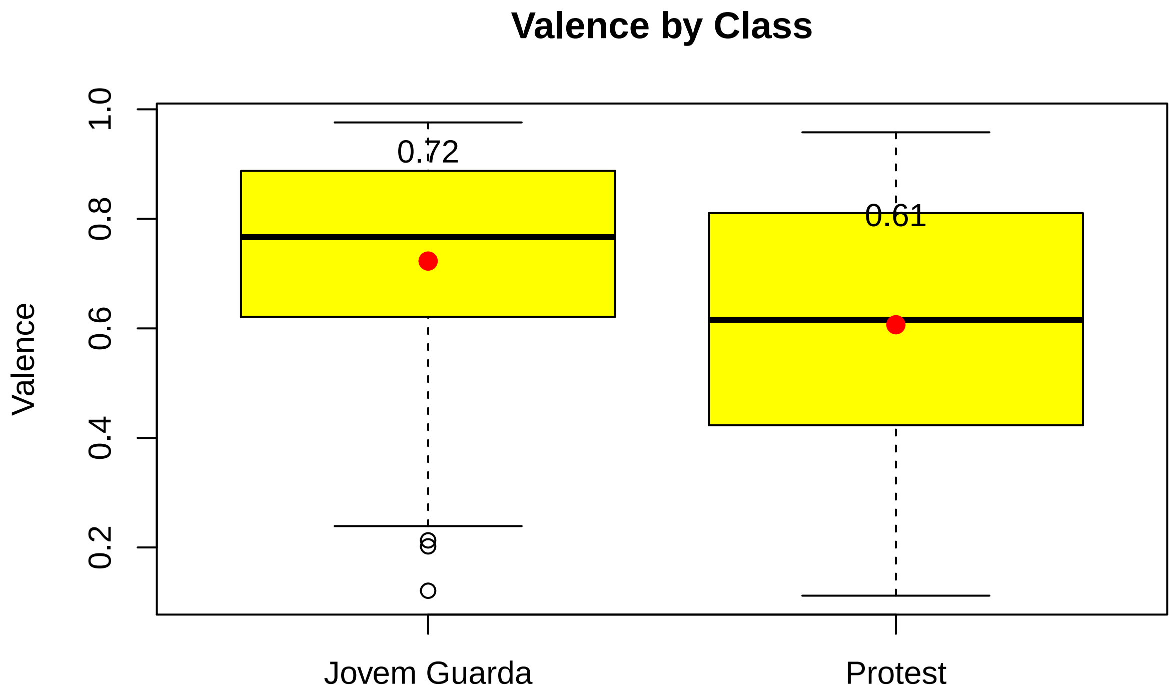
```r
set.seed(1234)
wordcloud(words = d_protest$word, freq = d_protest$freq, min.freq = 15,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

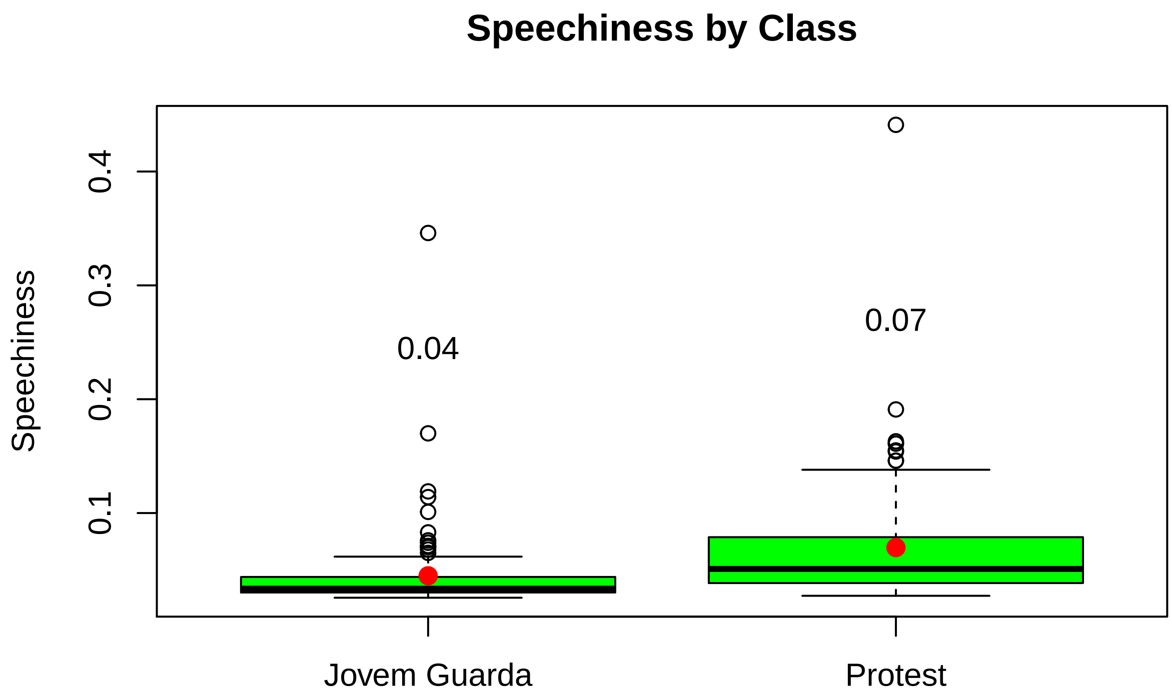

```r
wordcloud(words = d_jg$word, freq = d_jg$freq, min.freq = 15,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

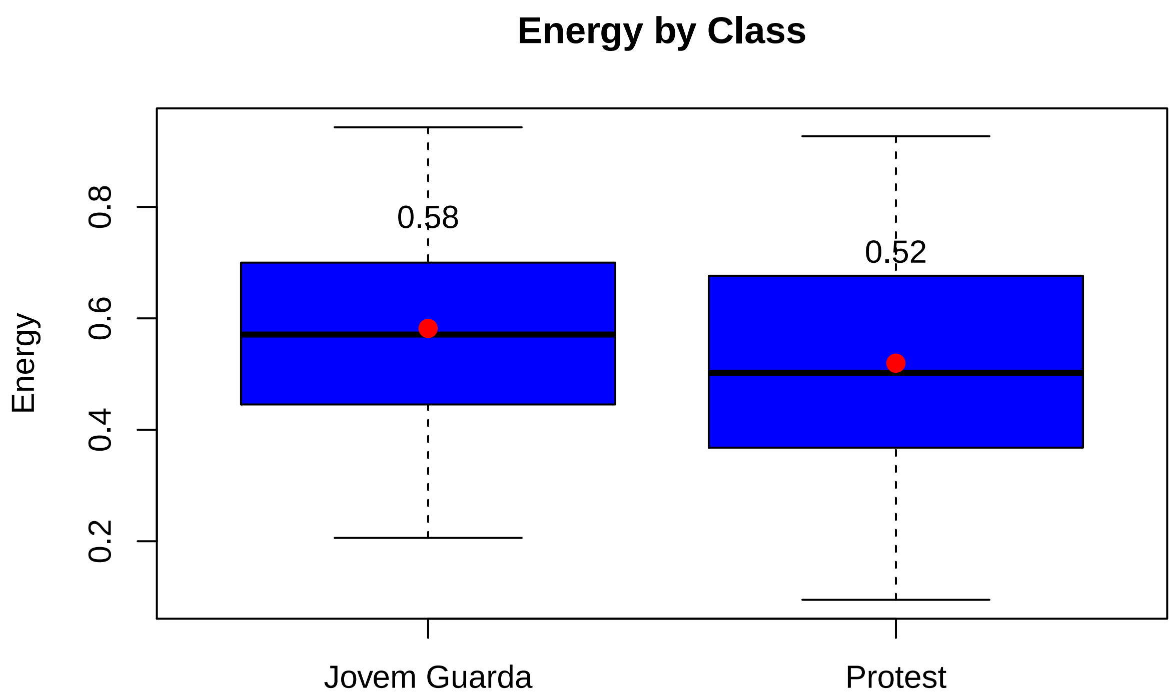


Then I performed ANOVA across four main song features, namely speechiness, energy, danceability and valence. As I imagined, protest music had higher speechiness given that the protest musicians prioritized the content of the message over form or harmonic features, whereas uncensored music had higher valence, danceability and energy. These characteristics were also in line with the insights gained from the historical study given that the Young Guard were known for their sappy songs that were popular in parties and bars (Table 1). All p-values were significant ($p < 0.05$).

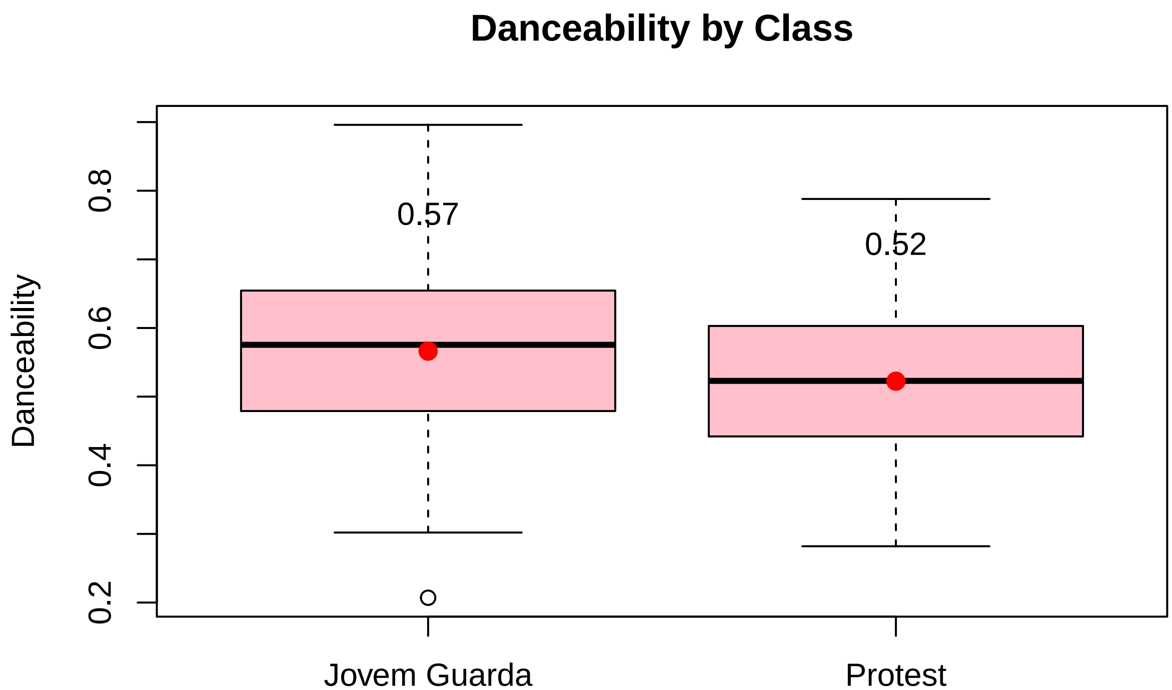
```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate
```


Valence by Class
Valence
1.0
0.8
0.6
0.4
0.2
0.72
0.61
Jovem Guarda
Protest


Speechiness by Class
Speechiness
0.4
0.3
0.2
0.1
0.04
0.07
Jovem Guarda
Protest



Energy by Class
Energy
0.2
0.4
0.6
0.8
0.58
0.52
Jovem Guarda
Protest



Danceability by Class
Danceability
0.2
0.4
0.6
0.8
0.57
0.52
Jovem Guarda
Protest

From the boxplots above, it seems that there is visual evidence for a significant differences between the two classes of music (Jovem Guarda and Protest). Let's conduct some t-tests to evaluate if these differences are significant.

```
## [1] 0.05272099 0.17973901
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] -0.03775548 -0.01186652
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 0.01013045 0.11459155
## attr(,"conf.level")
## [1] 0.95
```

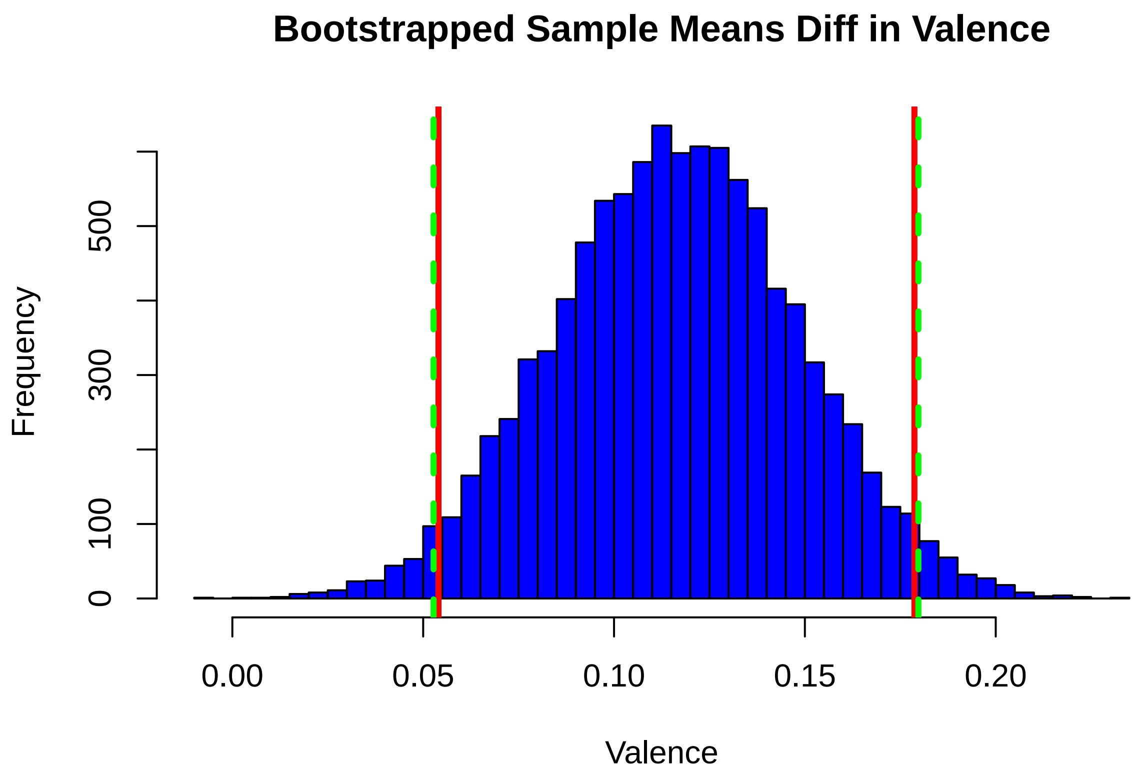
```
## [1] 0.009640185 0.077999815
## attr(,"conf.level")
## [1] 0.95
```

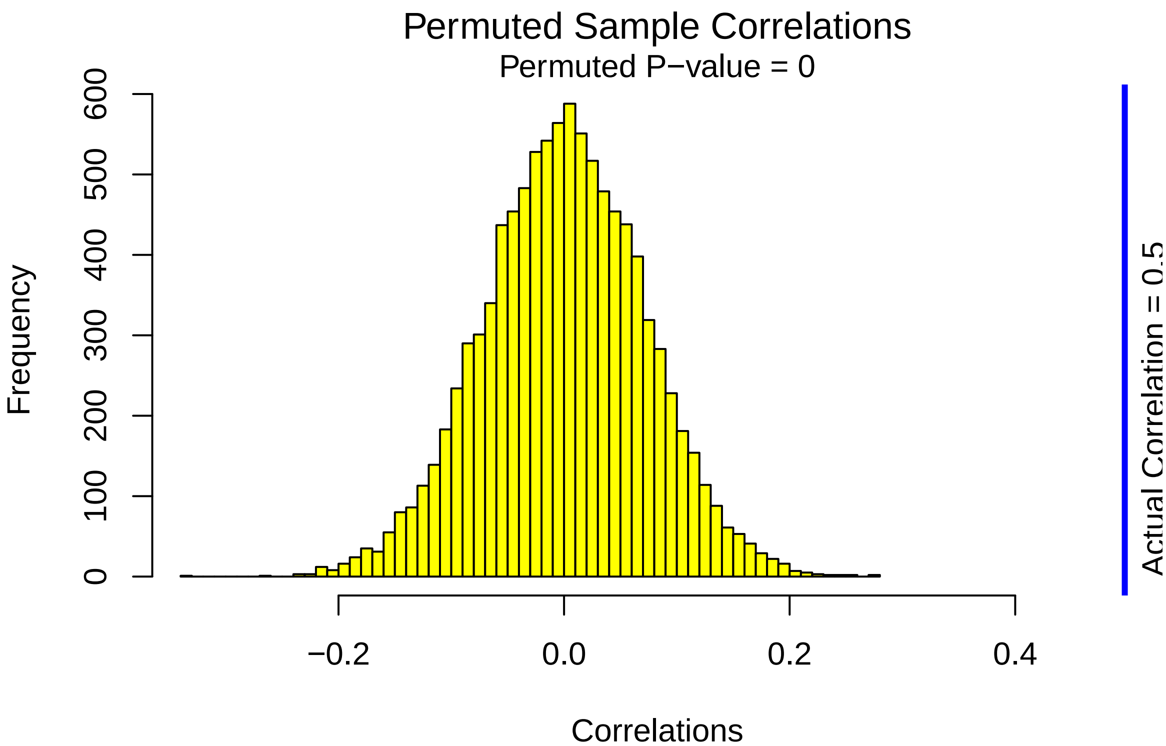
Permutation Test

Given that our t-tests p-values are significant, let's conduct a bootstrap test on Valence to look for the confidence intervals for the means difference in valence between the two classes.

```
N <- 10000
diffValence <- rep(NA, N)
set.seed(1)     #This is so we get same results every time

for (i in 1:N) {
  sA <- sample(music$valence[music$class == "Protest"],
              sum(music$class == "Protest"), replace = TRUE)
  sB <- sample(music$valence[music$class == "Jovem Guarda"],
              sum(music$class == "Jovem Guarda"), replace = TRUE)
```

```
  diffValence[i] <-  mean(sB) - mean(sA)
}

boot_ci <- quantile(diffValence, c(0.025, 0.975))

#Make histogram of bootstrap sample means
hist(diffValence, col = "blue", main = "Bootstrapped Sample Means Diff in Valence", xlab = "Valence", b

#Add lines to histogram for CI's
abline(v=boot_ci,lwd=3, col="red")
abline(v=test1,lwd=3, col="green", lty = 2)
legend(48,600, c("Original CI","Boot CI"), lwd=3, col = c("green","red"), lty = c(2,1))
```



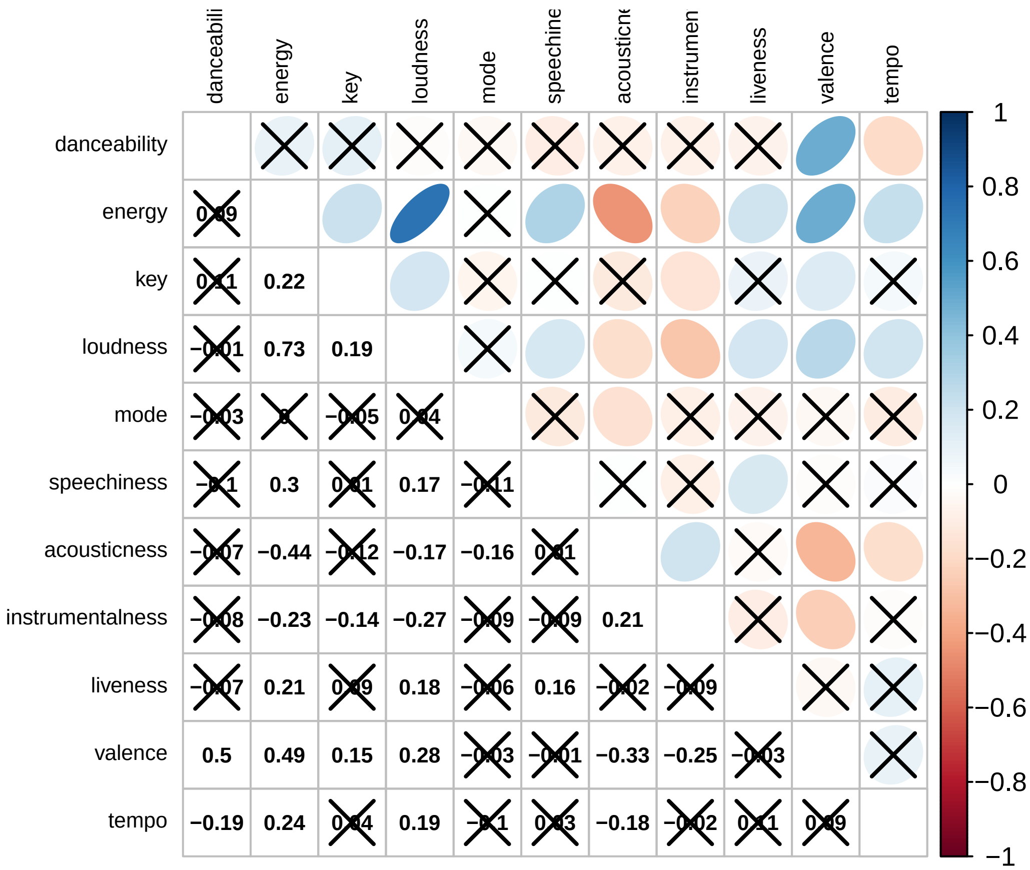


The null hypothesis is that there is no difference in the median of improvements between male and female runners. The alternative hypothesis is that there is a difference in the median of improvements between male and female runners. We cannot reject the null hypothesis in this case given that the difference is not statistically significant ($0.08 > 0.05$). In the context of this test, this p-value is the probability of finding a test statistic (i.e difference in mean) for the group comparison at least as high as the one observed, provided that there is no actual difference (i.e., null hypothesis is true). ## Basic tests with the different classes

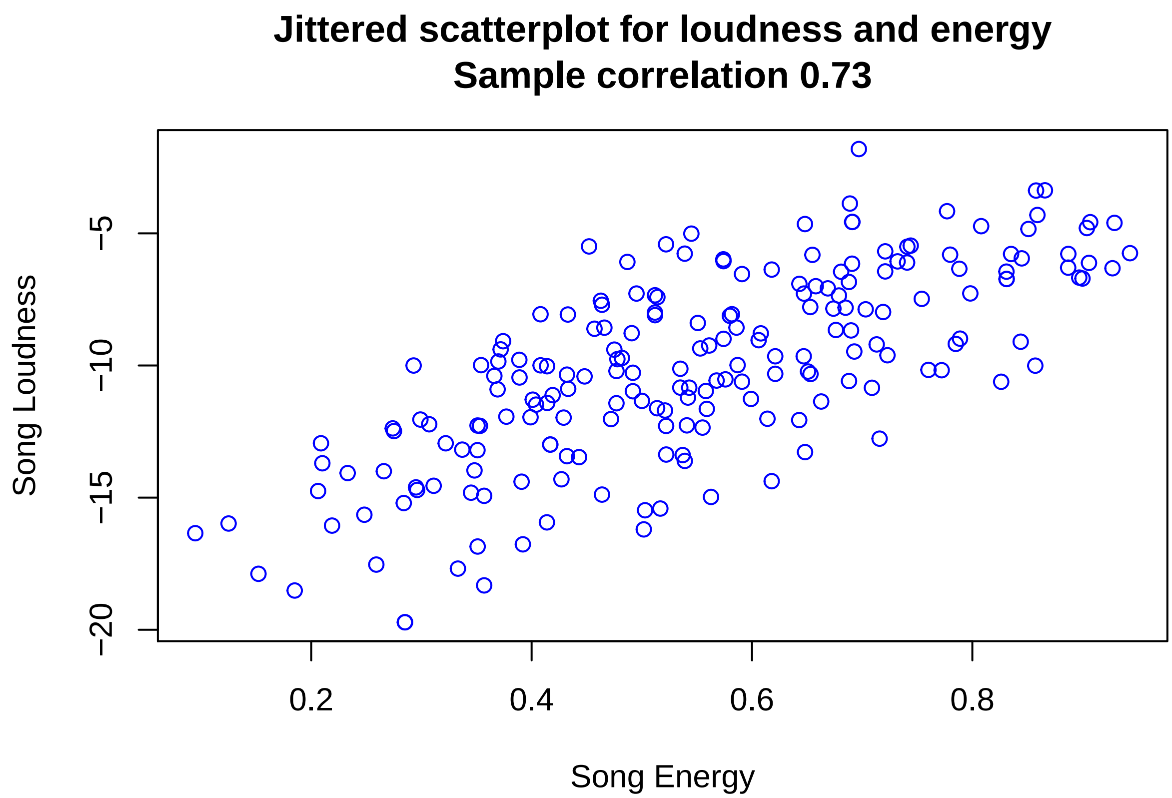
```
## corrplot 0.84 loaded
```

```
## [1] "The two column names of the two variables with the highest correlation:"
```

```
## [1] "loudness" "energy"
```



Now let's examine more closely the correlation between the two variables with highest correlation.



By

adding a small amount of random normally distributed noise, we can see observations and their densities more clearly, and now it looks like there is a strong correlation between the two song features (as demonstrated by the slighlty linear concentration in density).

Stepwise Regression

We are now going to proceed with performing stepwise regression. In particular, we're going to fit a model that looks at possible predictors of the class of the song. To do this, I'm making a new dataset called `music2` which contains the relevant columns (notice I'm putting the response variable FIRST). Be sure to remove the option `eval = F`.

```
#avoid multicolinaeartiy issues
music2 <- music[,c(5, 8:18)]

names(music2)
```

```
##  [1] "song_popularity"  "danceability"     "energy"
##  [4] "key"              "loudness"         "mode"
##  [7] "speechiness"      "acousticness"     "instrumentalness"
## [10] "liveness"         "valence"          "tempo"
```

```
dim(music2)
```

```
## [1] 200  12
```

```
total_vars <- dim(music2)[2]
```

Perform best subsets regression using the `regsubsets` function in the `leaps` package. Save the results in an object called `mod2`. Get the summary of `mod2` and save the results in an object called `mod2sum`. Display `mod2sum$which` to get a sense of which variables are included at each step of best subsets.

```
library('leaps')
mod2 <- regsubsets(song_popularity ~ ., data=music2, nvmax=total_vars)
mod2sum <- summary(mod2)
```

Now, let's examine the best model according to highest r-squared, etc.

```
(modnum = which.max(mod2sum$rsq))
```

```
## [1] 11
```

```
#Which variables are in model 12
names(music2)[mod2sum$which[modnum,]][-1]
```

```
##  [1] "danceability"     "energy"           "key"
##  [4] "loudness"         "mode"             "speechiness"
##  [7] "acousticness"     "instrumentalness" "liveness"
## [10] "valence"          "tempo"
```

```
#Fit this model and show results
musictemp <- music2[,mod2sum$which[modnum,]]

#summary(lm(song_popularity ~ .,data=musictemp))
```

```
(modnum <- which.max(mod2sum$adjr2))
```

```
## [1] 5
```

```r
#Which variables are in model 12
names(music2)[mod2sum$which[modnum,]][-1]
```

```
## [1] "mode"             "speechiness"      "acousticness"
## [4] "instrumentalness" "valence"
```

```r
#Fit this model and show results
musictemp <- music2[,mod2sum$which[modnum,]]
#summary(lm(song_popularity ~ .,data=musictemp))
```

BIC

```r
(modnum = which.min(mod2sum$bic))
```

```
## [1] 1
```

```r
#Which variables are in model 12
names(music2)[mod2sum$which[modnum,]][-1]
```

```
## [1] "acousticness"
```

```r
#Fit this model and show results
musictemp <- music2[,mod2sum$which[modnum,]]

#summary(lm(song_popularity ~ .,data=musictemp))
```

CP

```r
(modCP <- min(c(1:length(mod2sum$cp))[mod2sum$cp < c(1:length(mod2sum$cp))+1]))
```

```
## [1] 3
```

```r
#Which variables are in model 2
names(music2)[mod2sum$which[modCP,]][-1]
```

```
## [1] "mode"         "acousticness" "valence"
```

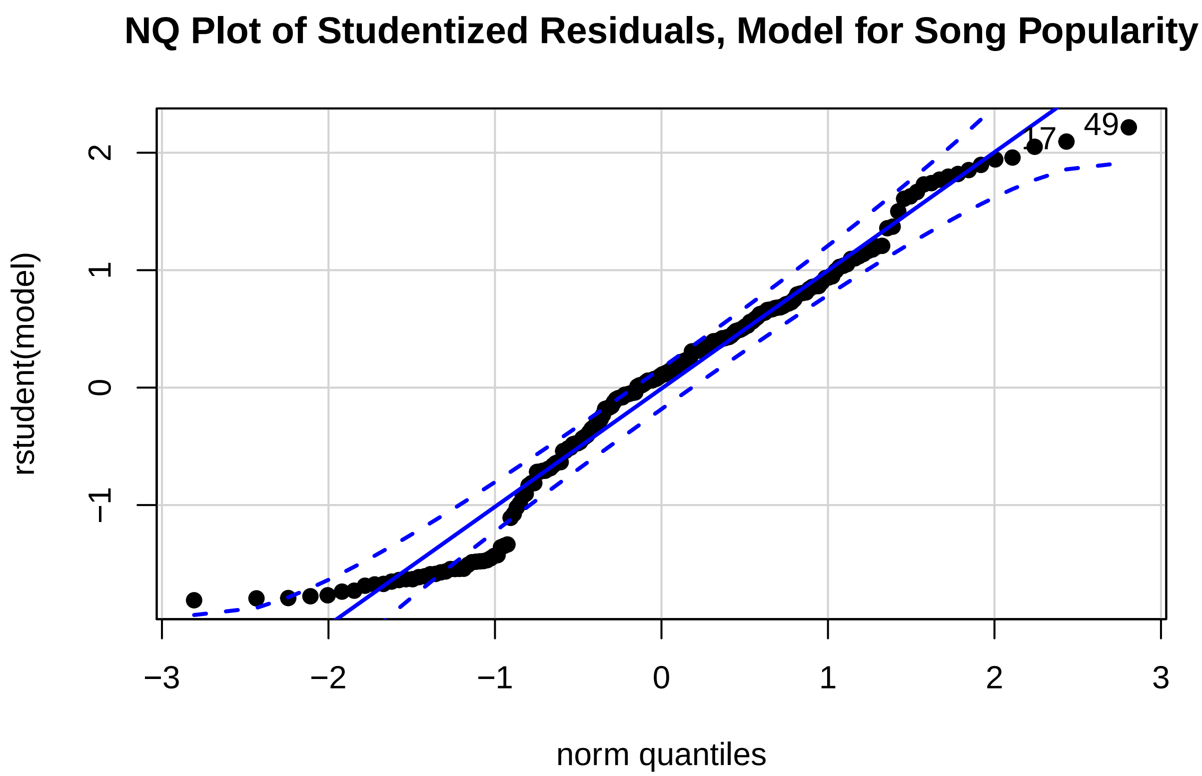
```r
#Fit this model and show results
musictemp <- music2[,mod2sum$which[modCP,]]
#summary(lm(song_popularity ~ .,data=musictemp))
```

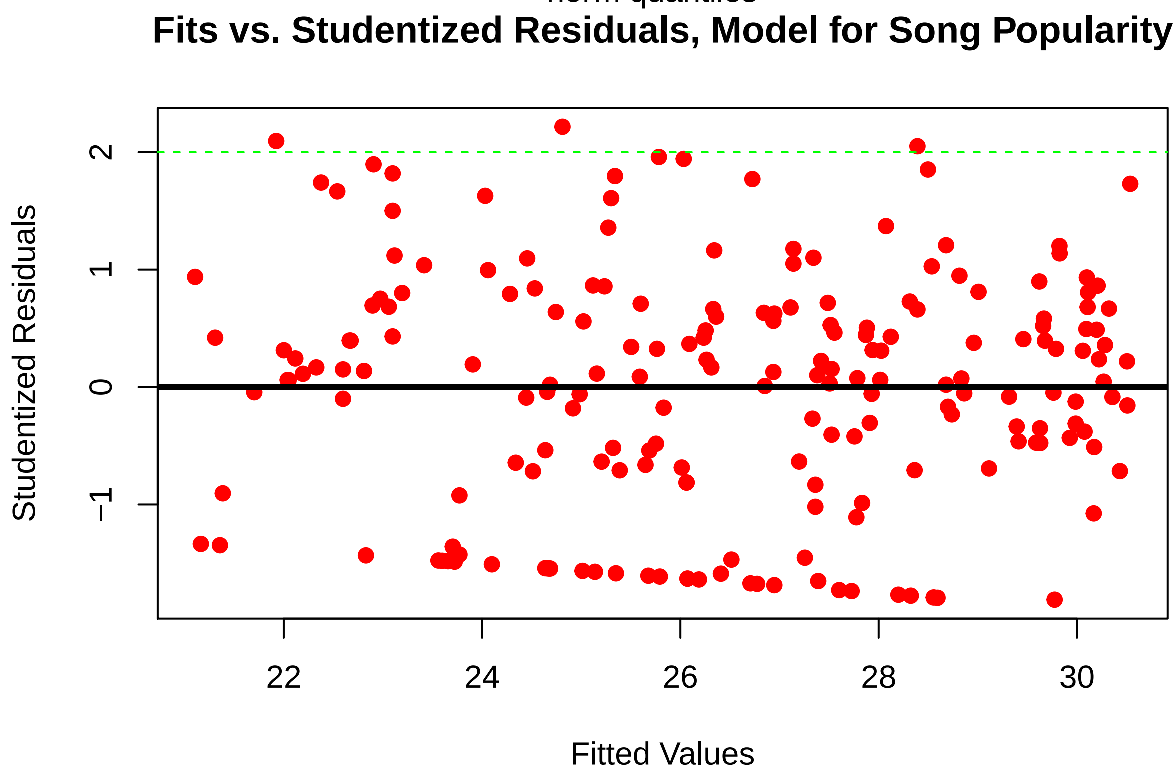
Now, let's evaluate the final model we have:

```r
musicfinal <- music2[,mod2sum$which[1,]]
modfin <- lm(song_popularity ~ .,data=musicfinal)

#get new function for pairs plotn AND get myResPlots function
source("http://www.reuningscherer.net/s&ds230/Rfuncs/regJDRS.txt")
```

```
## Loading required package: carData
```

```r
myResPlots2(modfin,"Model for Song Popularity")
```





The model assumptions do not seem be met given that the erros in the normal quantile plot are NOT normally distributed and the fits vs residuals plot show no evidence of heteroscedasticity (the distribution of positive and negative residuals appears symmetric across all the possible fitted values). in other words, the assumptions are met because there is constant variance across fitted values, few outliers, no clear trend.