

# The truth behind the lyrics

**Author: Jinxu Xiang**

This report is prepared with the following environmental settings.

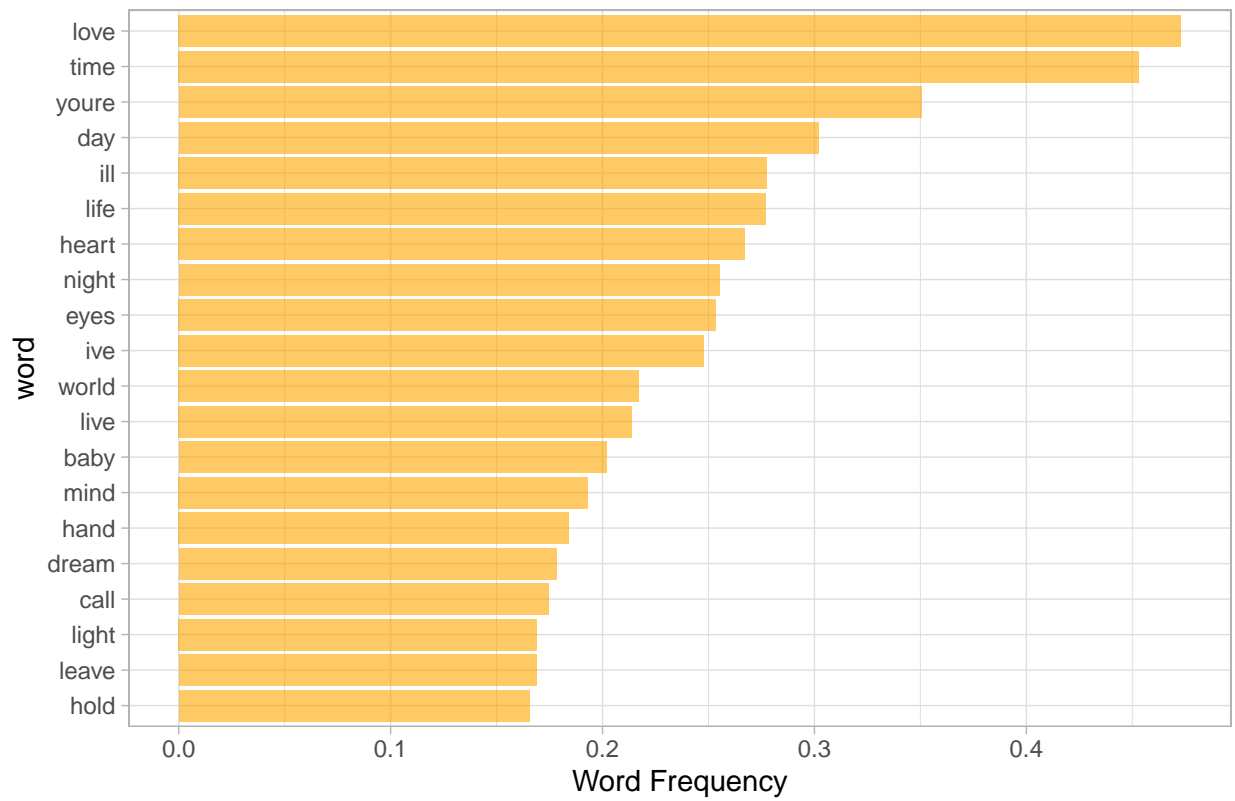
```
print(R.version)
```

```
##  
## platform      x86_64-w64-mingw32  
## arch          x86_64  
## os            mingw32  
## system        x86_64, mingw32  
## status  
## major         3  
## minor         6.2  
## year          2019  
## month         12  
## day           12  
## svn rev       77560  
## language      R  
## version.string R version 3.6.2 (2019-12-12)  
## nickname      Dark and Stormy Night
```

First, I processed the raw textual data 'lyrics.RData' saved in 'data' file by cleaning data, removing stopwords and creating a tidy version of texts which is saved in 'output' file.

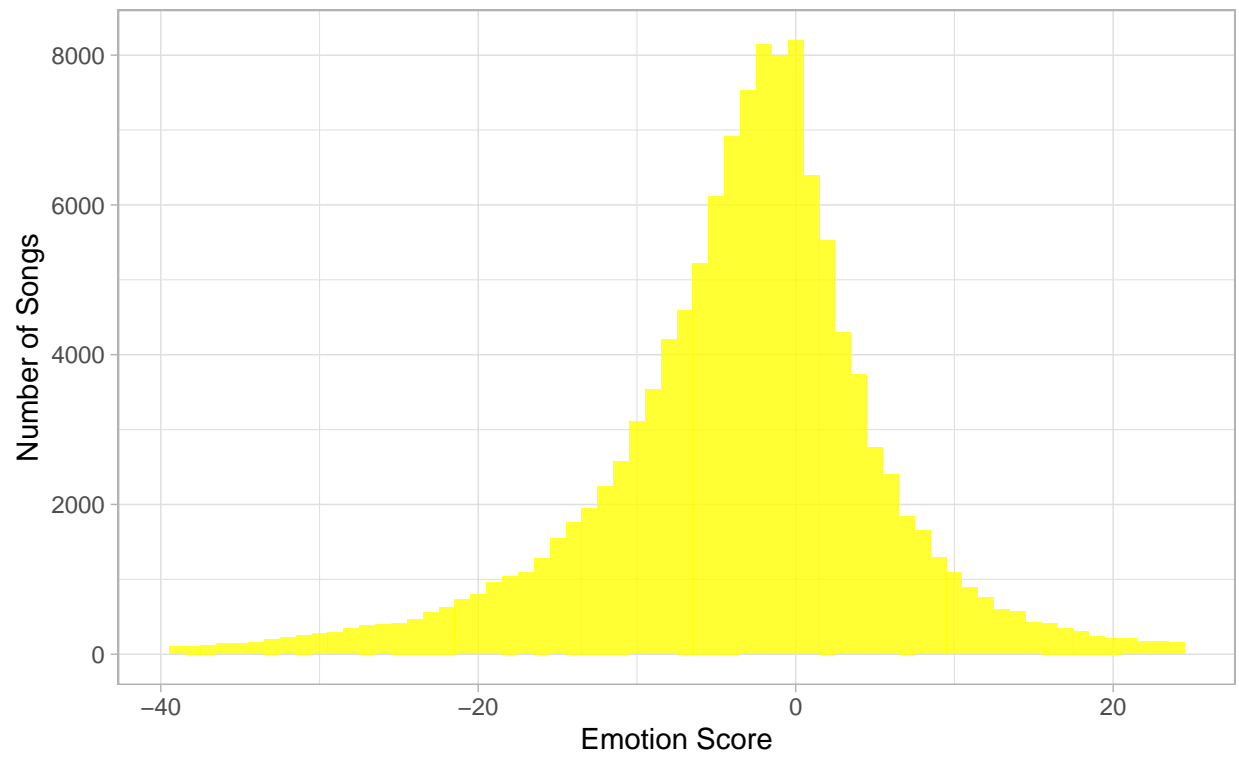
Then, I combined the processed text with artist information 'artists.csv' and saved the joint data in 'output' file. The 'Origin' column of joint data contains the name of city and country (or state in America). So I extracted the names of each region and saved it as 'Precessed\_country'.

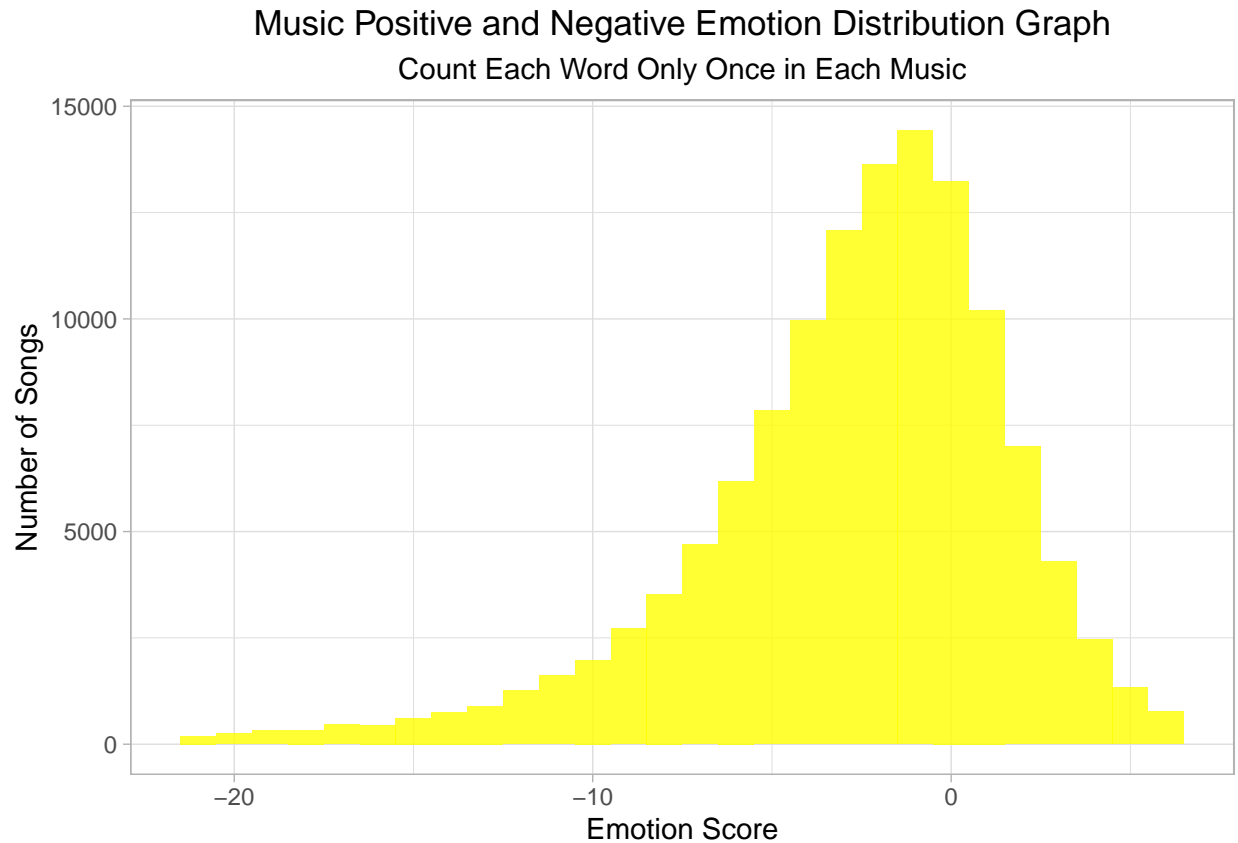
Word Frequency of All Lyrics





Music Positive and Negative Emotion Distribution Graph  
Count Each Word Repeatedly in Each Music

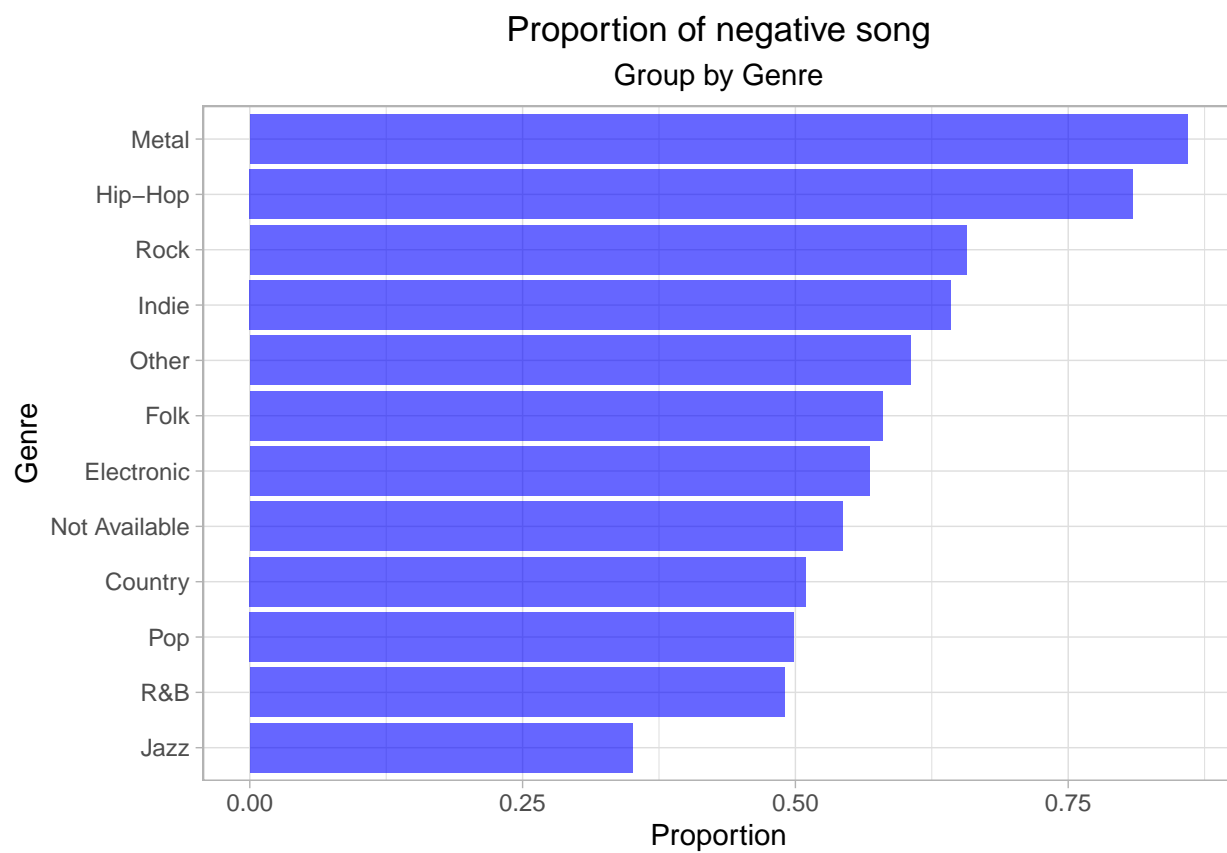


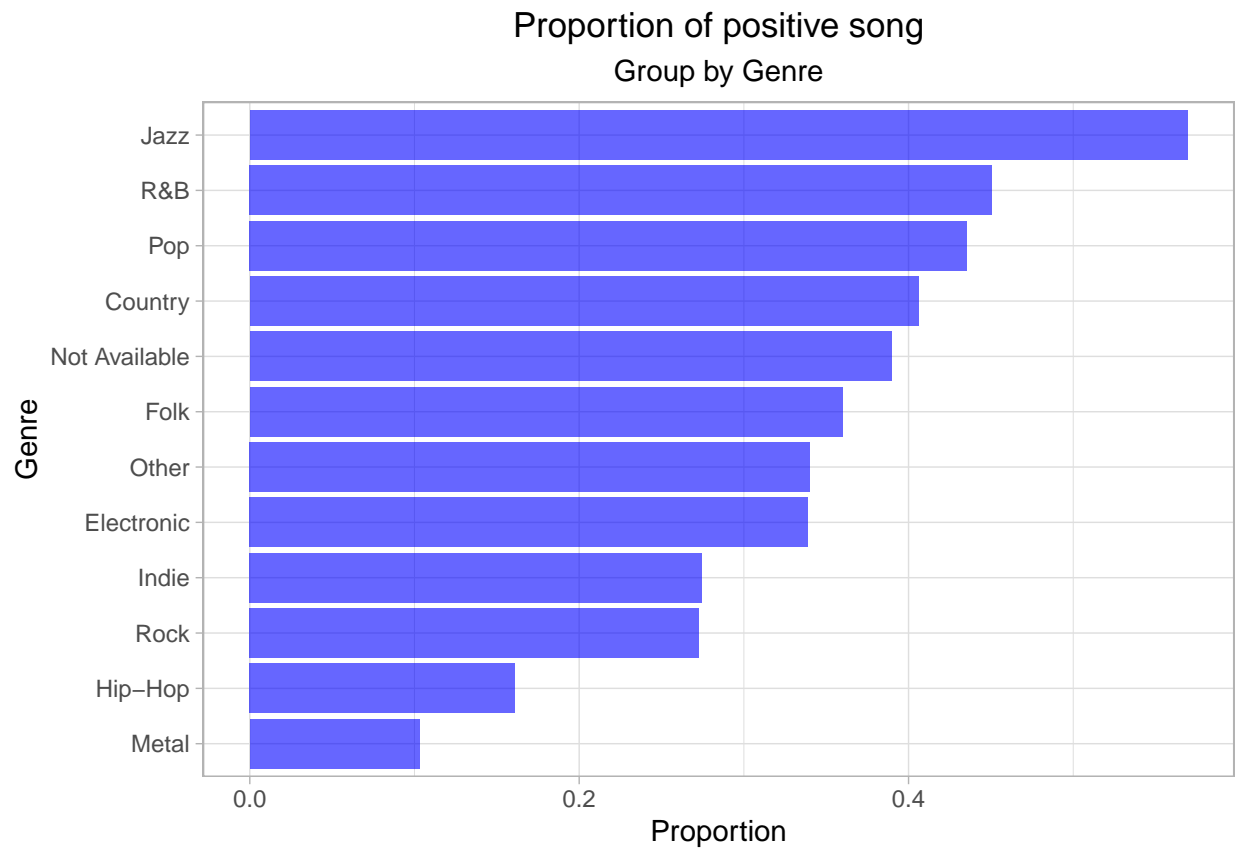


These two graphs above are the opposite of what I said before. Even though positive words occupy most of the wordcloud, about 70% music use more negative words. In addition, the emotion score which each word only count once looks like Poisson distribution. This property may be used in more places.

The statistical model of this data is too good, it is difficult to imagine that the 120,000 data can get a distribution that looks a lot like the real distribution. In addition, considering that there are many strange author names, song names and lyrics in this dataset, it will increase people's doubt about the reality of this dataset.

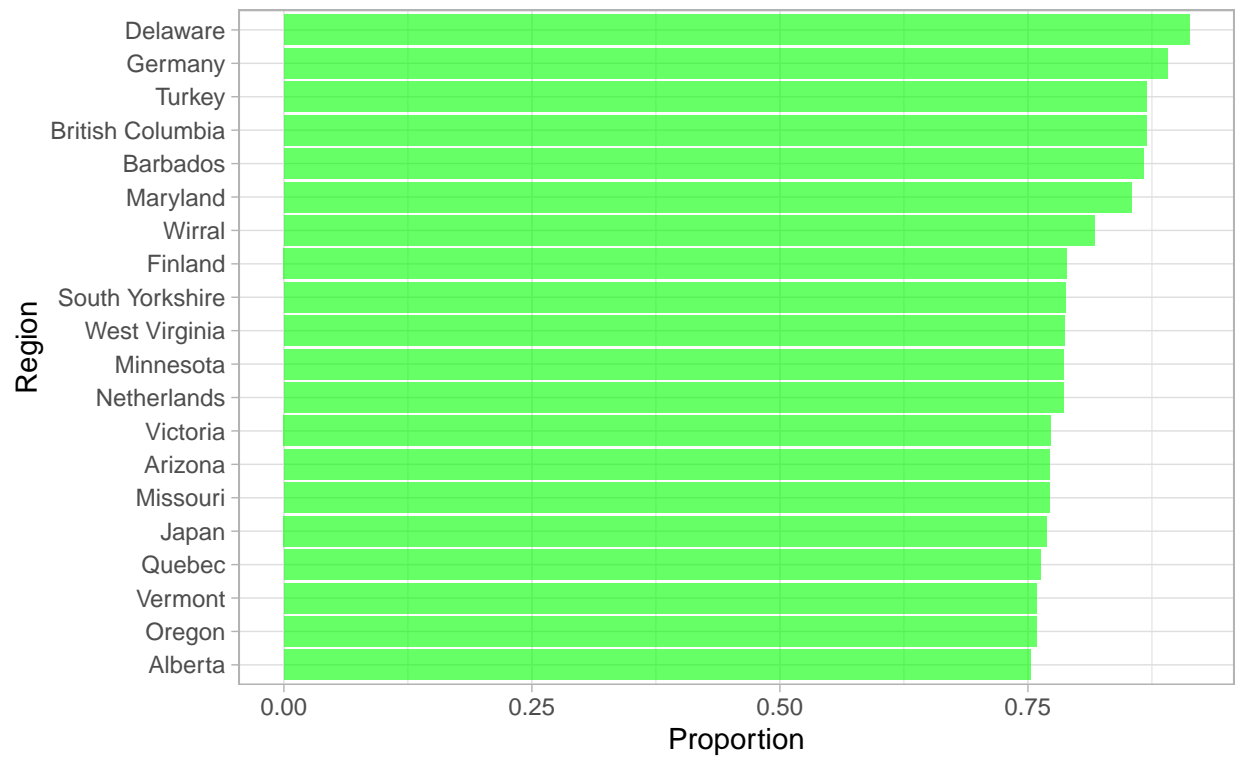
Since more negative words are shown below the appearance of positive words, so where, when and what kind of music use negative words more? Where, when and what kind of music use positive words more? The following graphs make statistics of different categories of lyrics.



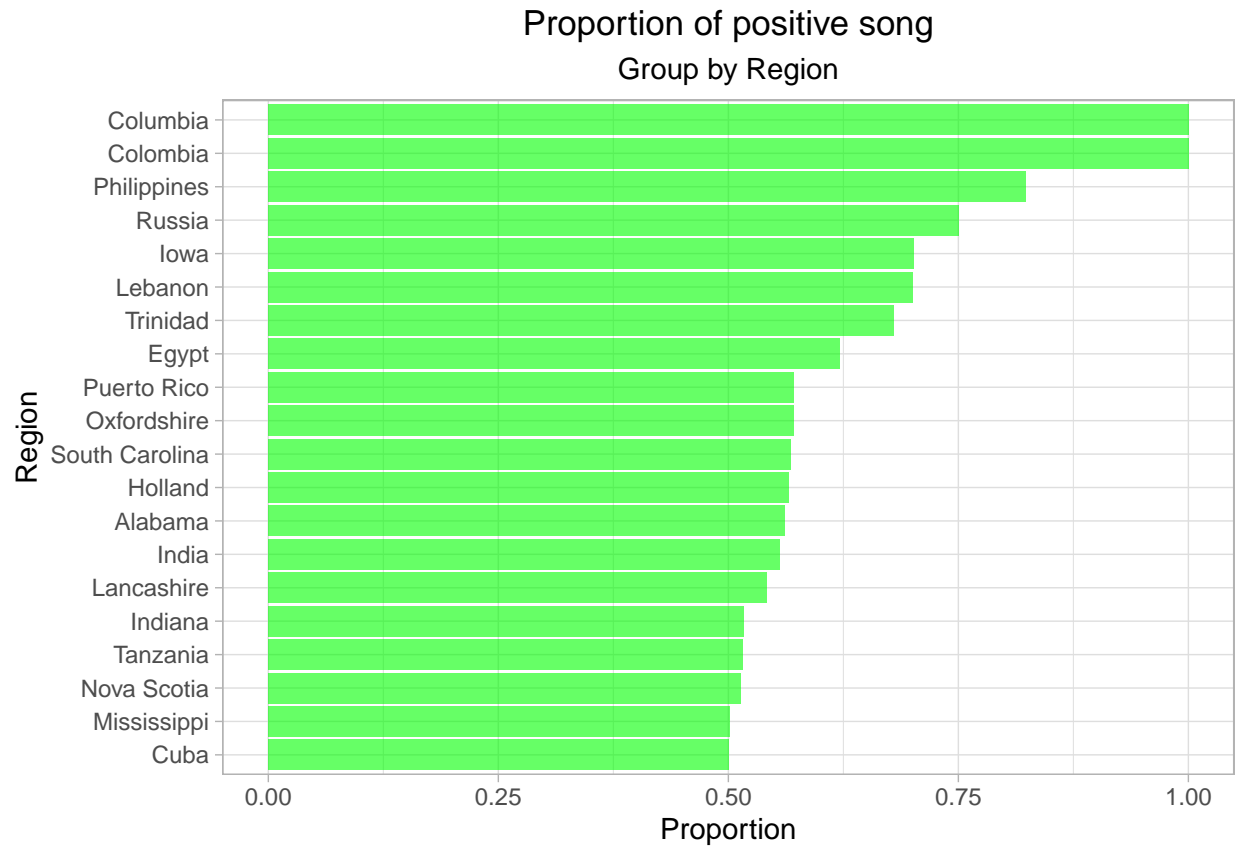


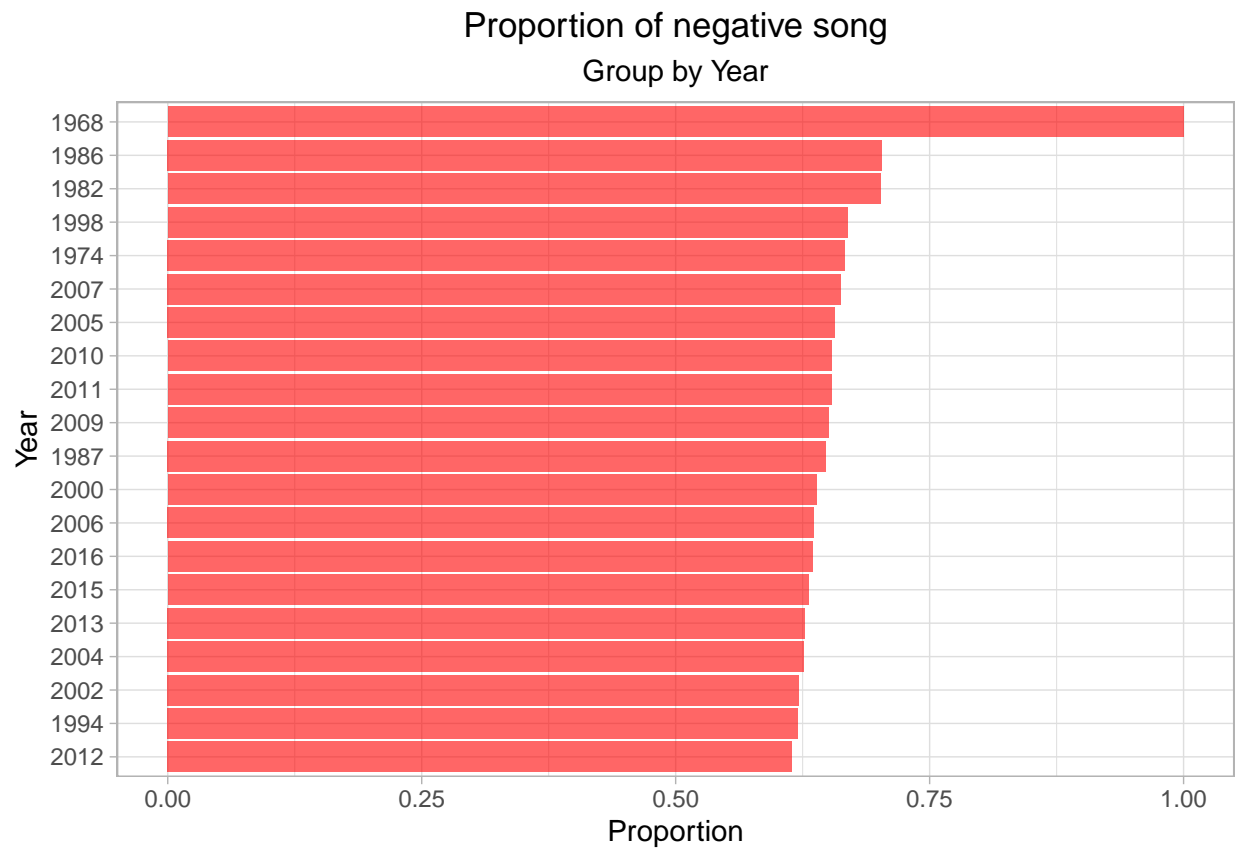
## Proportion of negative song

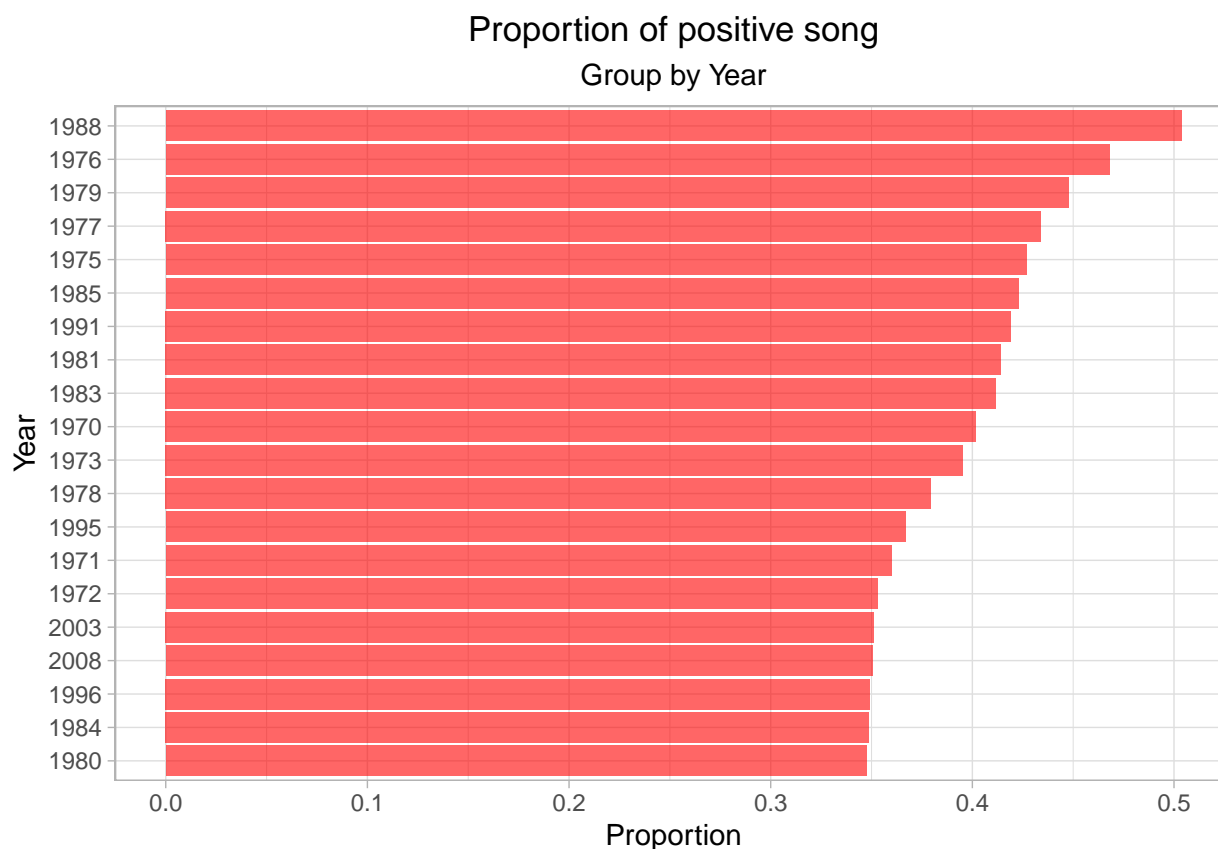
Group by Region











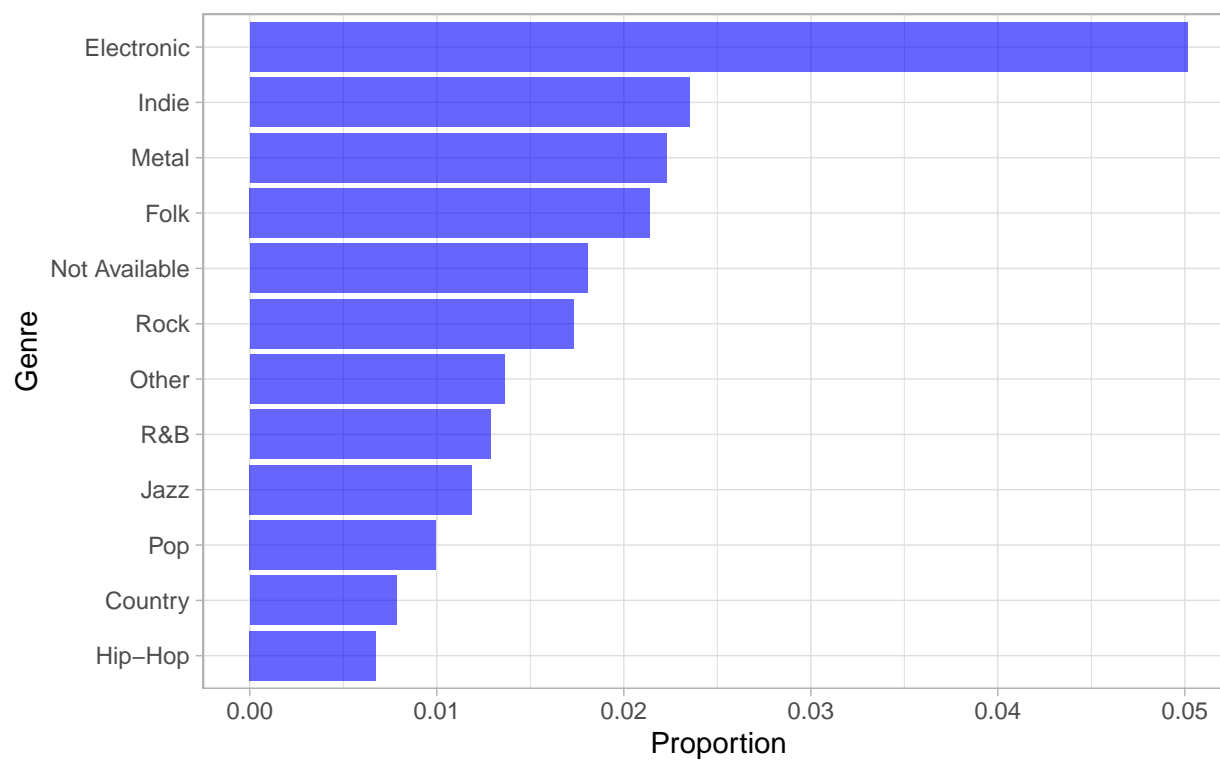
Metal and Hip-hop are the music genres with the most negative words and Jazz is with the least. Columbia, Phillippines and Russia are the regional using positive words while most of others use negative words more. Since there was only one song in 1968, the proportion reached 100%. In addition, the negative words were used the most in 1986 and 1982 and the least used in 1988.

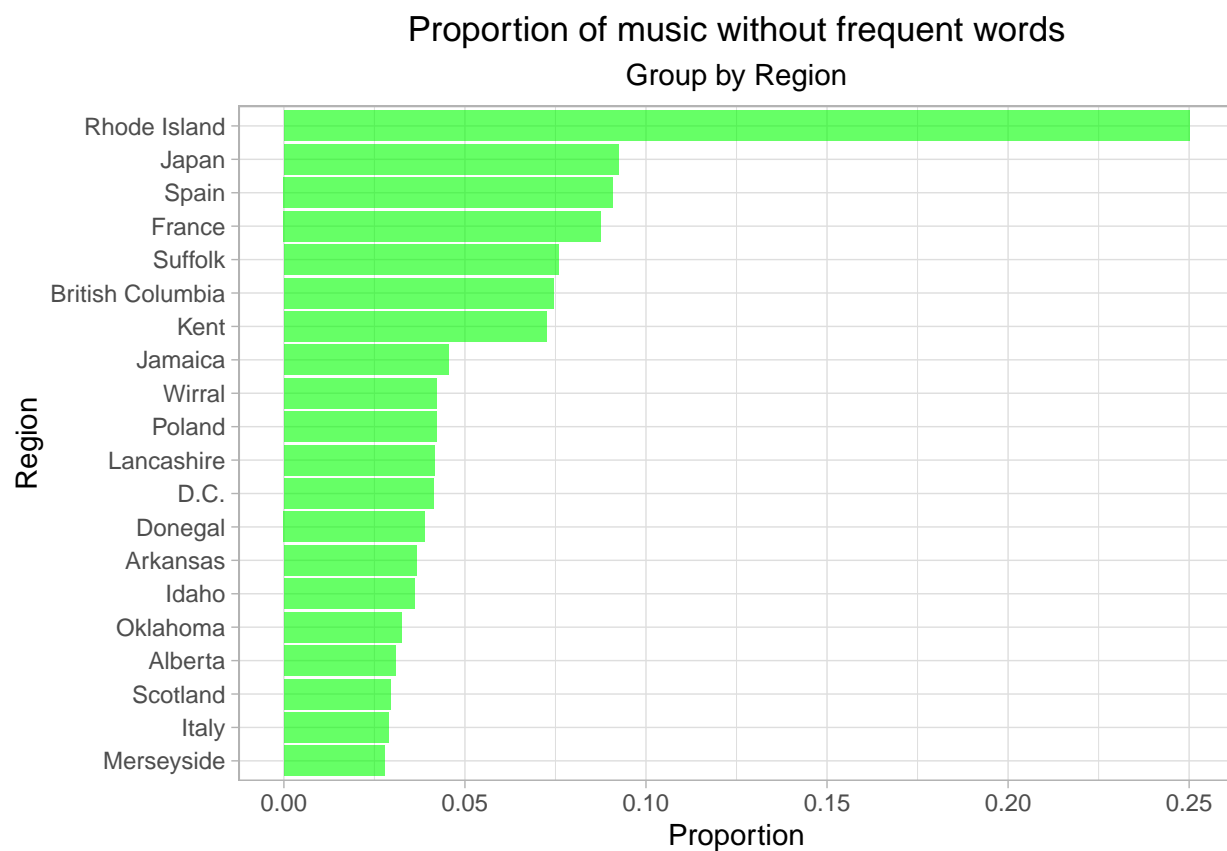
## Question 2 - What kind of music is out of the ordinary?

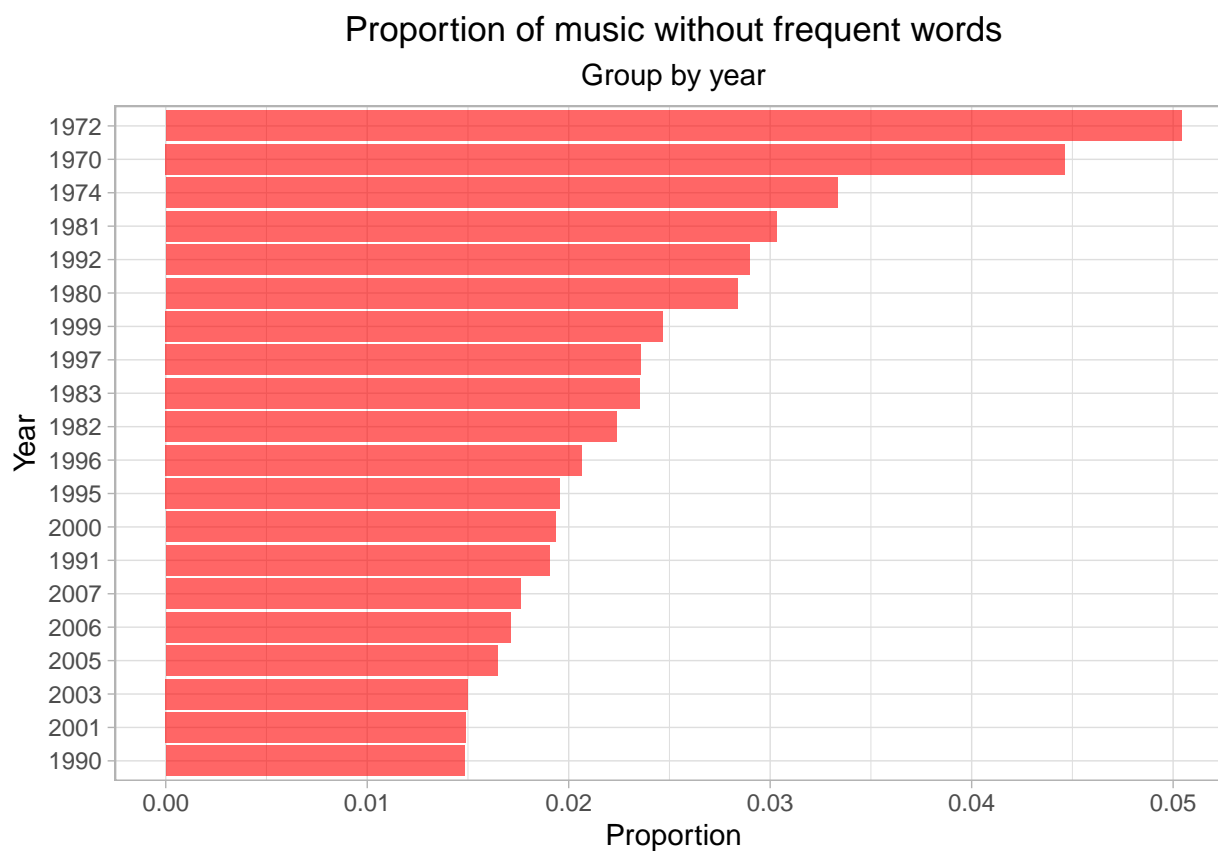
From the shinyapp and statistical chart above, we can see that most of music uses words like 'love' and 'time'. we will be curious about such a question, what kind of music doesn't contain the most frequent words. After selecting the music that doesn't contain the most frequent 30 words, we can get music statistics in different categories as follows.

## Proportion of music without frequent words

Group by Genre







These statistical graphs show that electronic music uses the least high-frequency words when classifying music genres. Rhode Island, Japan, Spain, France, Suffolk and British Columbia use the least frequent words when classifying by country and people used the least frequent words in 1972, 1970 and 1974. Then, what are the musical keywords for these genres, places and times?



