Trabalho AED - Top50 Spotify 2019

Luciano Martins, Sarah Nadaud, Déborah Gomes

20/02/2021

Resumo

Uma boa escolha de dados é uma etapa fundamental em qualquer pesquisa, pois inconsistência nela levarão a problemas nas fases seguintes, logo antes de escolher a base de dados com que iríamos trabalhar, verificamos se estava tudo certo com os dados. O tema entretenimento foi pré-definido. Já sabendo disso, decidimos usar a base de dados: Top 50 Spotify Songs – 2019, ou seja, as 50 músicas mais ouvidas no mundo pelo spotify. Este conjunto de dados possui várias variáveis sobre as músicas, o que nos ajudará a aplicar os conhecimentos adquiridos durante todo o curso. O que nos motivou a decidir por esses dados foi que a maior parte da população hoje em dia utiliza o spotify e ouve música o tempo inteiro, sem pensar na quantidade de pessoas que utilizam esse serviço de maneira simultânea, escuta as mesmas músicas, quais são os artistas/gêneros mais ouvidos, entre outros. Com a ajuda do R, iremos construir gráficos que nos ajudarão a melhor visualizar cada variável escolhida e tirar as estatísticas que serão explicadas com mais detalhes no decorrer do trabalho.

Descrição dos Dados

Estamos analisando uma base de dados referente a top 50 músicas do Spotify no período de 2018. A base está disponível no seguinte link: https://www.kaggle.com/leonardopena/top50spotify2019. A mesma consta com um total de 50 linhas e 13 variáveis.

Organizando os tipos de variáveis a seguir:

Variavel	Descrição	Tipo
ID	Número único referente a cada música no top50	Qualitativo Nominal
Track Name	Nome da Música	Categórica Nominal
Artist Name	Nome do Artista	Categórica Nominal
Genre	Gênero musical	Categórica Nominal
Beats per Minute	Quantidade de batidas por minuto da musica	Numerica Contínua
Energy	A energia de uma música, quanto mais alto o valor, mais animada ela é	Numerica Contínua
Dancebility	Quanto maior o valor, mais fácil é dançar com essa música.	Numerica Contínua
Loudness (dB)	Quanto maior o valor, mais alta é a música	Numerica Contínua
Livenss	Quanto maior o valor, mais chance dessa música ter sido gravada ao vivo	Numerica Contínua
Valence	Quanto maior o valor, mais um energia positiva a música passa.	Numerica Contínua

Variavel	Descrição	Tipo
Length	A duração da música.	Numerica Contínua
Acousticness	Quanto maior o valor mais acústico a música é	Numerica Contínua
Speechiness	Quanto maior o valor mais partes cantadas a música tem.	Numerica Contínua
Popularity	Quanto maior o valor mais popular a música é	Numerica Contínua

Após definirmos os tipos das variáveis podemos prosseguir fazendo a análise univariada correta para cada uma delas. No caso de variáveis númericas apresentaremos o boxplot e as estatísticas básicas de média, mediana, variância, moda, desvio padrão, quantidade de outlier e quartis. No caso das variáveis categóricas vamos trazer a quantidade de categorias únicas e suas frequências.

Comentário sobre as Análises

Quando formos analisar variáveis numéricas, vamos plotar seu histograma e seu boxplot. Caso seja interessante, vamos calcular também o coeficiente de curtose e assimetria. No caso de variaveis categoricas, vamos plotar os graficos de frequencia por categoria.

Id e Track Name

Como se trata de uma variável única para cada entrada não há o que ser analisado.

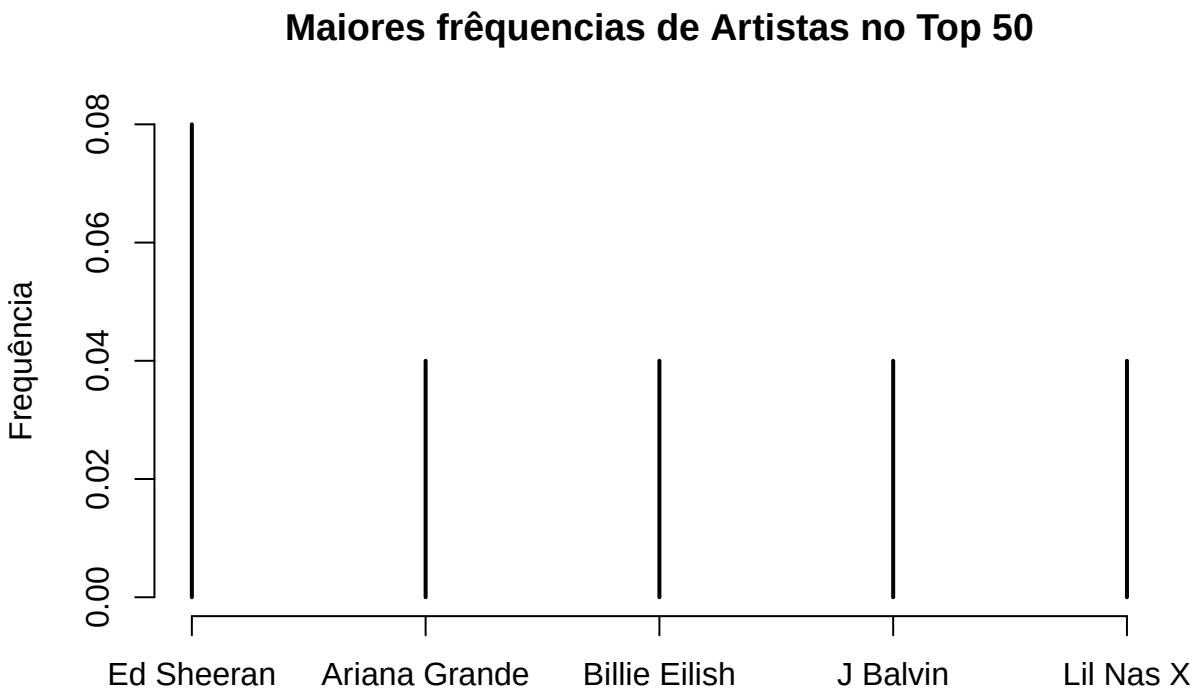
Artist Name

No top 50 temos ao todo 38 artistas, isso quer dizer que há artistas que aparecem mais de uma vez nesse top 50. Sendo assim procuramos analisar se existiam artistas com maior frequência no top 50.

```
length(unique(top50$Artist.Name))
```

```
## [1] 38
```

```
plot((sort(prop.table(table(top50$Artist.Name)), decreasing=TRUE)[1:5] ),type="h",
     ylab= 'Frequência',
     main= 'Maiores frêquencias de Artistas no Top 50')
```



O artista "Ed Sheeran" apresenta a maior frequeência entre eles, porém essa frequência é de apenas 0,08 , em conclusão percebemos que em geral não existem muitas músicas dos mesmos artistas no top50, sendo assim, não há uma hegemonia musical.

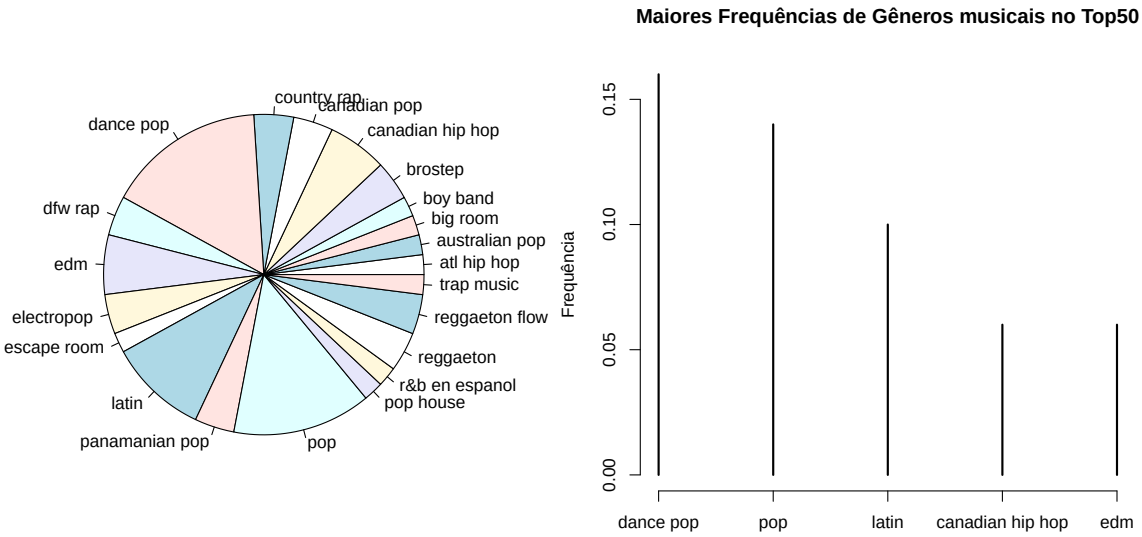
Gênero Musical

No top 50 temos ao todo 21 gêneros musicais, são relativamente poucos gêneros, sendo assim deve haver alguma relevância em alguns gêneros específicos que possam auxiliar para estar no top50. Com isso, procuramos analisar quais os gêneros musicais com maior frequência no top 50.

```
length(unique(top50$Genre))
```

```
## [1] 21
```

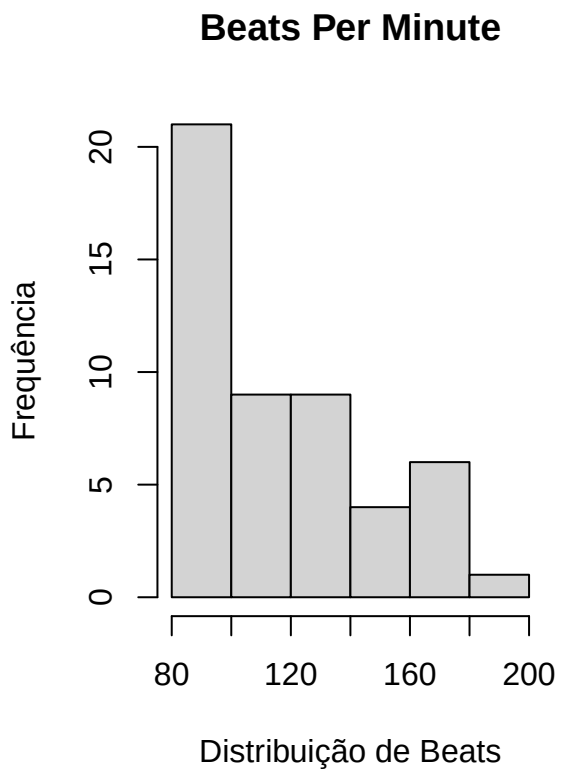
```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
pie(prop.table(table(top50$Genre)))
plot((sort(prop.table(table(top50$Genre)), decreasing=TRUE)[1:5] ),type="h", ylab = 'Frequência',
   main= 'Maiores Frequências de Gêneros musicais no Top50')
```

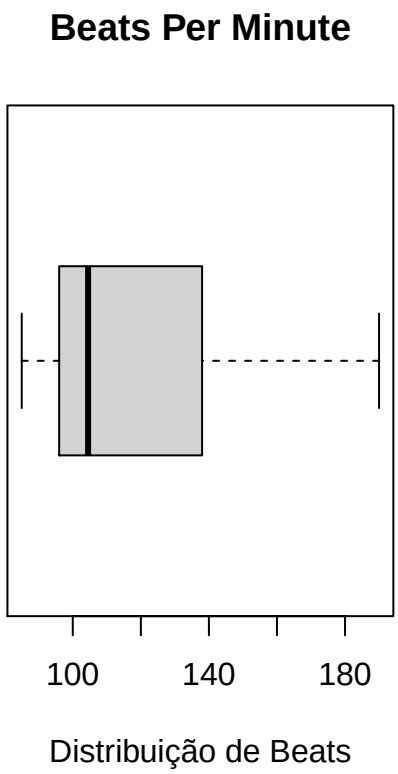


Os gêneros com maior frequência são os relacionados com o dance pop e o próprio pop, são gêneros de maior relevância midiática por serem introduzidos através da indústria americana, em contrapartida há o gênero latin que mostra uma crescente relevância do cenário latino na música. Outro ponto relevante é que os 5 gêneros são muito populares por carregarem músicas para dançar, assim esperamos que as músicas desses tais gêneros possam mostrar um valor de 'Dancebility' alto.

Beats per Minute

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Beats.Per.Minute, main = 'Beats Per Minute', xlab = 'Distribuição de Beats', ylab= 'Frequênc
boxplot(top50$Beats.Per.Minute, horizontal = TRUE ,
        main = 'Beats Per Minute',
        xlab= 'Distribuição de Beats')
```






```
(summary(top50$Beats.Per.Minute))
```

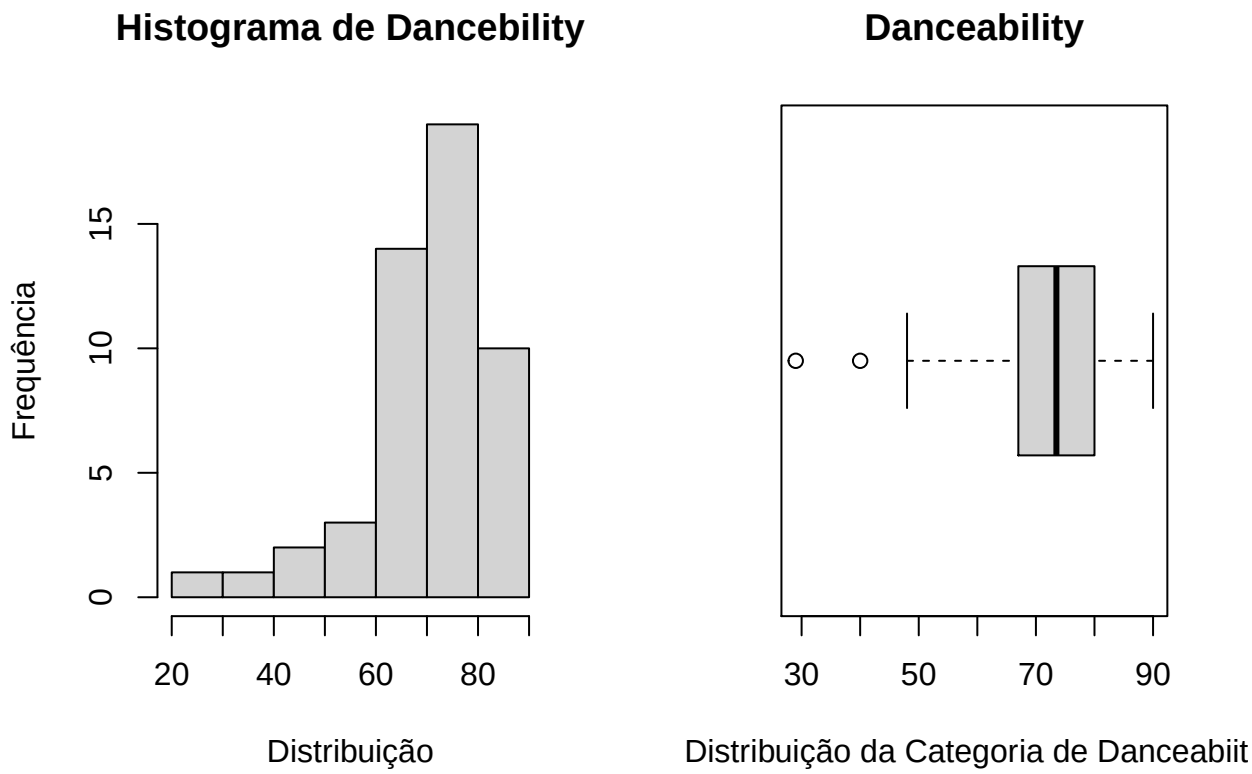
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    85.0    96.0   104.5   120.1   137.5   190.0
```

Tendo em vista que quanto maior o BPM maior é a velocidade da música, temos que uma faixa ideal para se ter uma música no top50 é entorno de 90 até 140, onde se localiza a maioria das músicas, mesmo assim, há uma preferencia por BPM menores.

Dancebility

Com essa variável buscamos analisar se as pessoas têm preferência por músicas mais animadas, ou mais lentas.

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Danceability, main = 'Histograma de Dancebility',xlab = 'Distribuição', ylab= 'Frequência')
boxplot(top50$Danceability, horizontal = TRUE ,
        main = 'Danceability',
        xlab= 'Distribuição da Categoria de Danceabiity')
```




```
(summary(top50$Danceability))
```

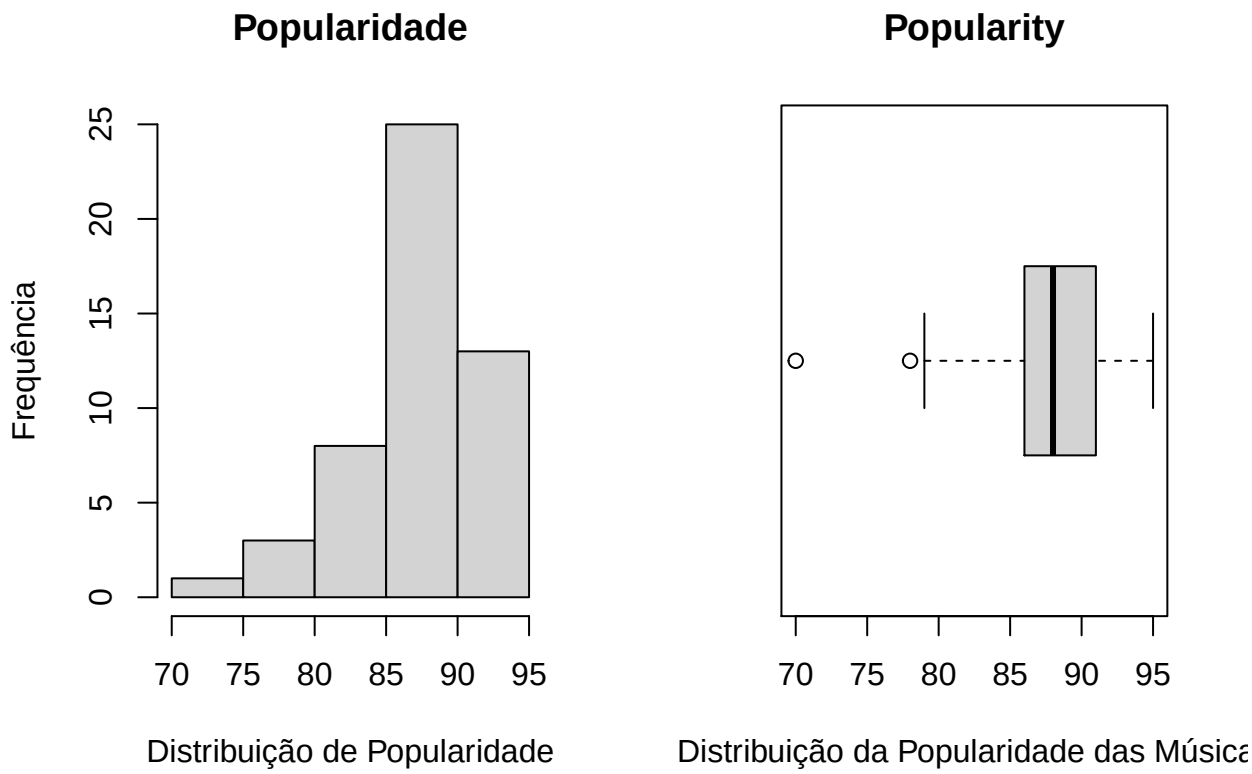
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   67.00   73.50   71.38   79.75   90.00
```

Observa-se que a maioria das músicas tem um valor de "Dancebility" alto, isso pode estar associado com a ideia de quanto melhor a música é para dançar, mais vezes vou escuta-la. Além disso, em todo top 50 há 2 outliers com os valores 40 e menor que 30, mostrando que há realmente uma preferencia por esses estilos de musicas e que as mais "lentas" não são tão suscetíveis a serem top 50.

Popularity

A popularity do spotify é baseada no número total de reproduções que a faixa teve e quão recentes são essas reproduções. Essa popularidade pode aumentar e diminuir com facilidade, visto que ela é baseada em um curto período de tempo. Essa é uma das variáveis mais importantes para que a música esteja no Top 50.

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Popularity, main = 'Popularidade', xlab = 'Distribuição de Popularidade', ylab= 'Frequência')
boxplot(top50$Popularity, horizontal = TRUE ,
        main = 'Popularity',
        xlab= 'Distribuição da Popularidade das Músicas')
```



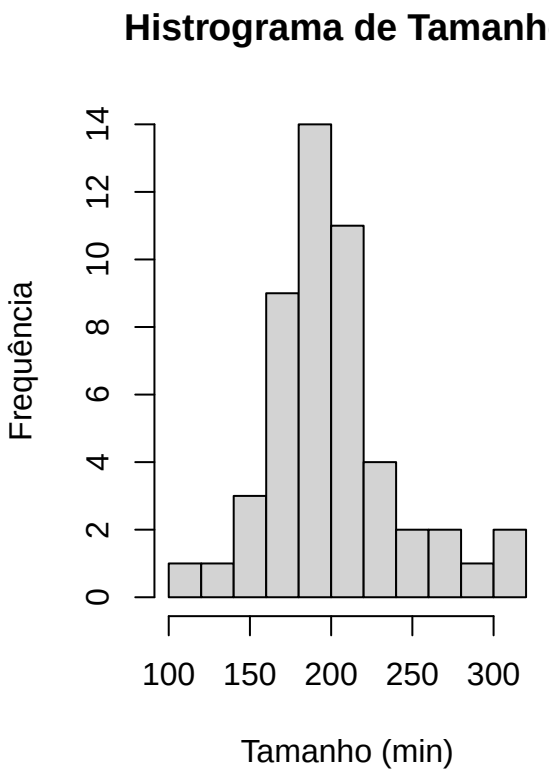

```
(summary(top50$Popularity))
```

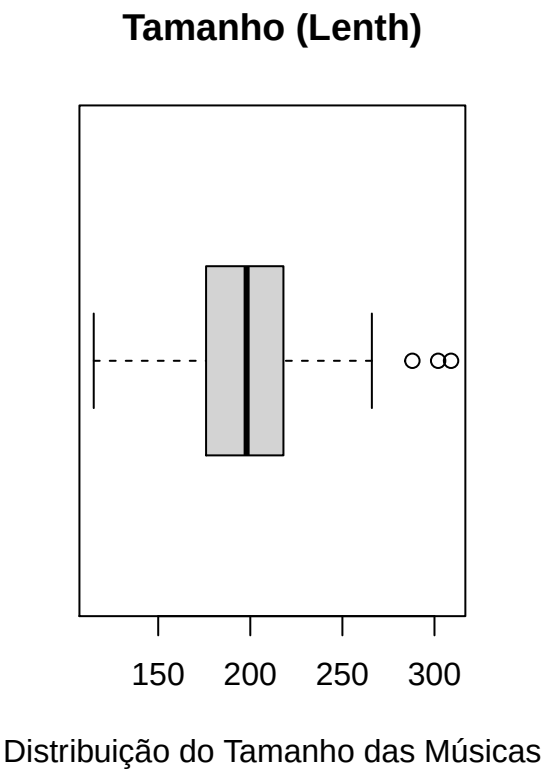
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.00   86.00   88.00   87.50   90.75   95.00
```

Apesar da popularidade ser um dos principais pilares para músicas estarem no top50 vemos que há duas músicas onde tem valores fora da normalidade, sendo uma delas com o valor de 70. Apesar disso boa parte das músicas respeita tal ideia e apresenta uma popularidade alta.

Lenth

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Length., main = 'Histrograma de Tamanho', xlab = 'Tamanho (min)', ylab= 'Frequência')
boxplot(top50$Length., horizontal = TRUE ,
        main = 'Tamanho (Lenth)',
        xlab= 'Distribuição do Tamanho das Músicas')
```






```
(summary(top50$Length.))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   115.0   176.8   198.0   201.0   217.5   309.0
```

```
var(top50$Length.)
```

```
## [1] 1532.243
```

```
kurtosis(top50$Length.)
```

```
## [1] 3.928966
```

```
skewness(top50$Length.)
```

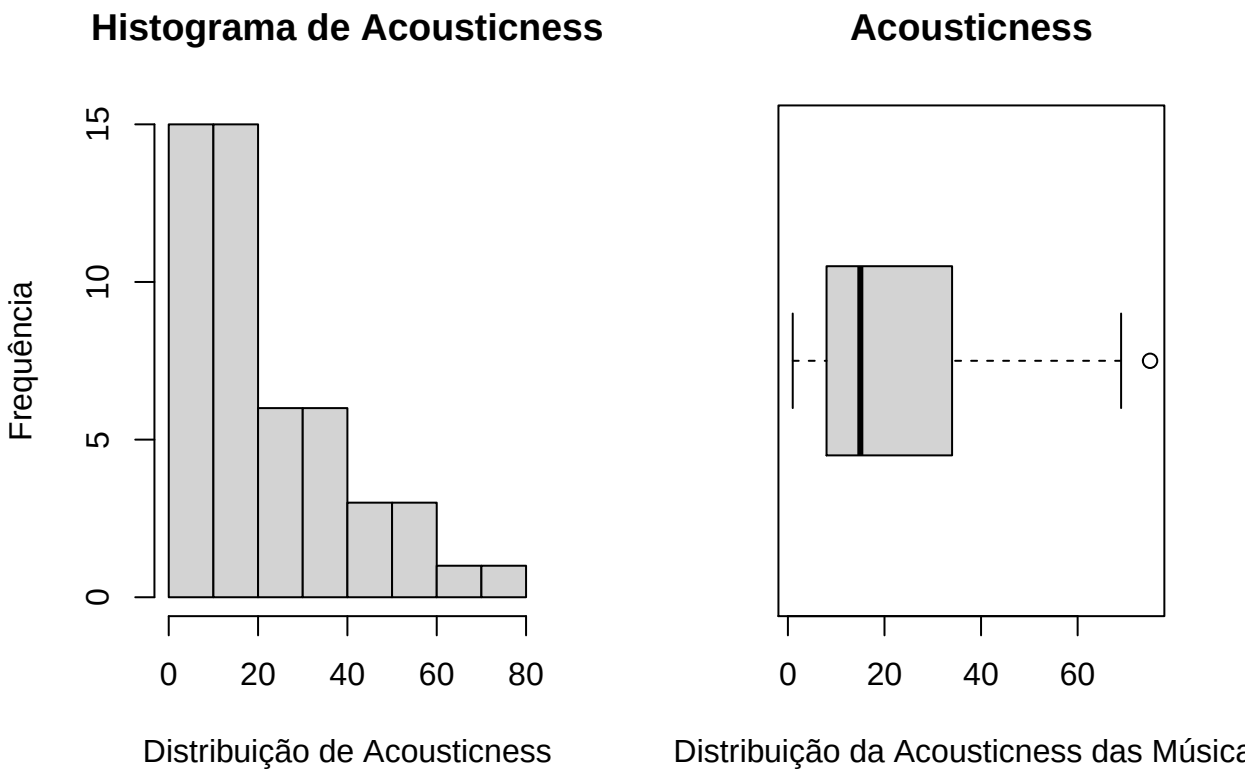
```
## [1] 0.7259074
```

Podemos notar que as músicas que geralmente se encontram no Top 50 tem entre 3 a 5 minutos de duração.

Acousticness..

—— ADICIONAR TEXTO ————

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Acousticness.., main = 'Histograma de Acousticness', xlab = 'Distribuição de Acousticness', y
boxplot(top50$Acousticness.., horizontal = TRUE ,
        main = 'Acousticness',
        xlab= 'Distribuição da Acousticness das Músicas')
```



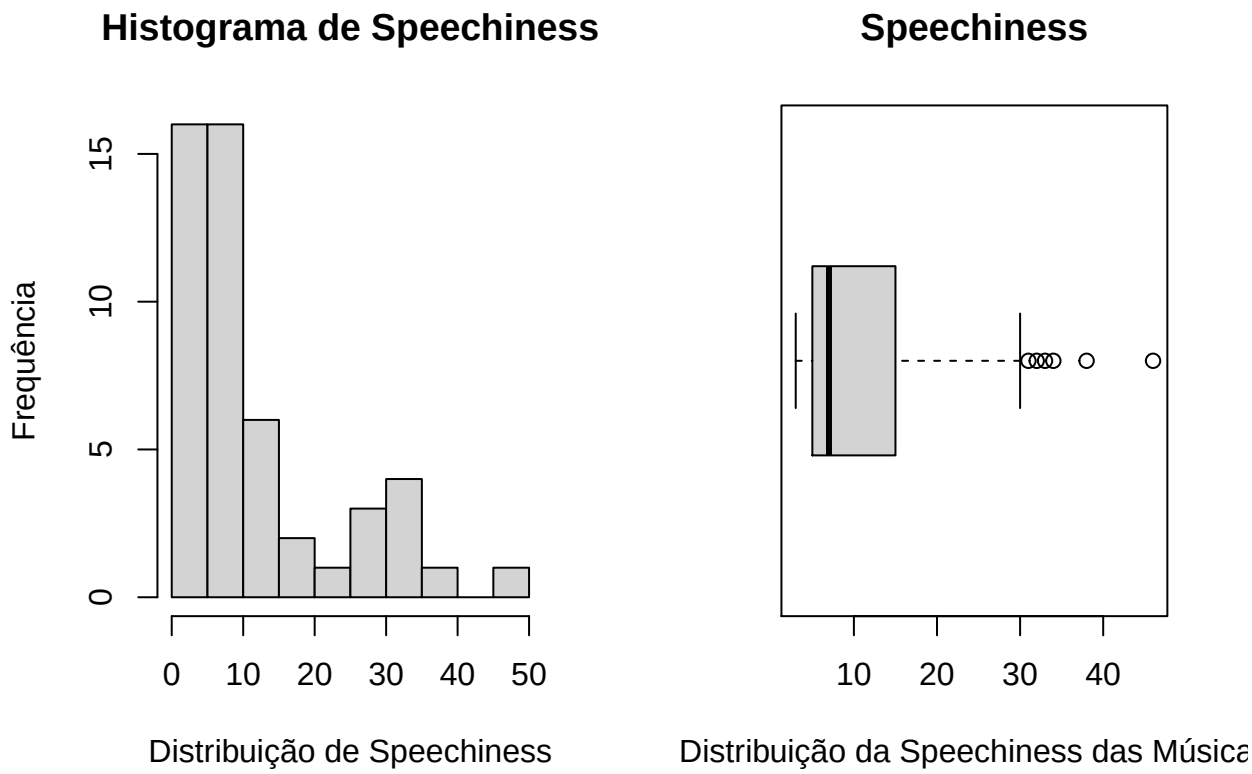

```
(summary(top50$Acousticness..))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.25   15.00   22.16   33.75   75.00
```

Observa-se que os valores não são tão altos, demonstrando que as músicas gravadas não foram gravadas de maneira acústica.

Speechiness.

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Speechiness., main = 'Histograma de Speechiness', xlab = 'Distribuição de Speechiness', ylab=
boxplot(top50$Speechiness., horizontal = TRUE ,
        main = 'Speechiness',
        xlab= 'Distribuição da Speechiness das Músicas')
```



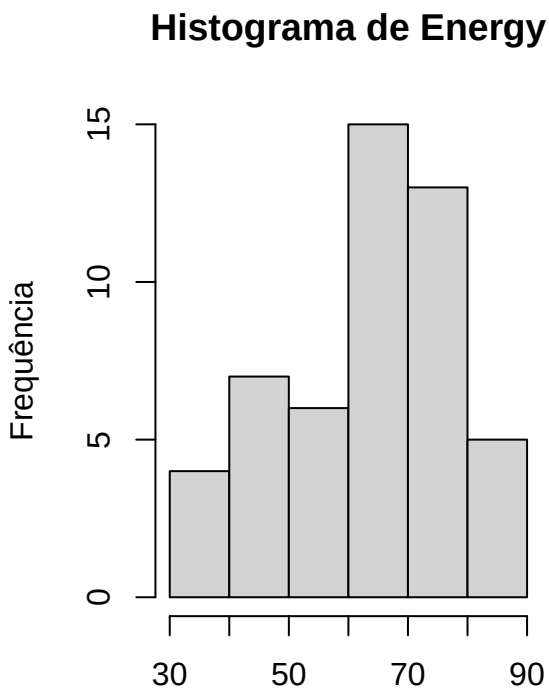

```
(summary(top50$Speechiness.))
```

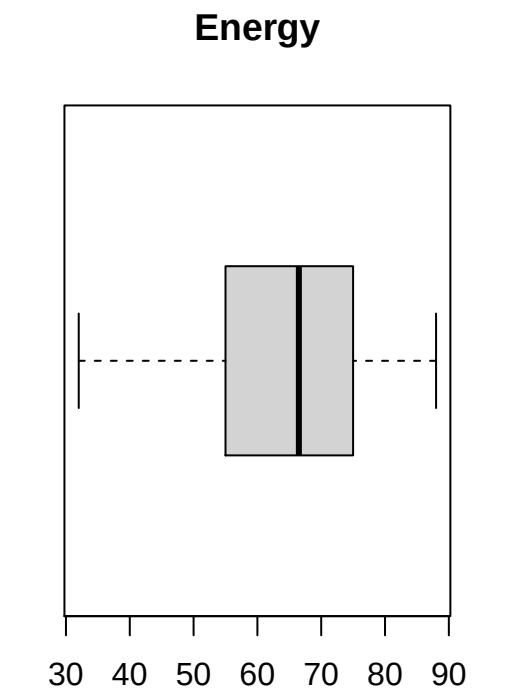
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    5.00    7.00   12.48   15.00   46.00
```

Podemos perceber que não há muitas músicas com uma grande quantidade de palavras faladas. Já que o valor não é tão grande.Essa informação corabora com a analise da variavel de dancabilidade e genero musical, onde prevalencem musicas pop e dançantes.

Energy

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Energy , main = 'Histograma de Energy', xlab = 'Distribuição de Energy', ylab= 'Frequência')
boxplot(top50$Energy, horizontal = TRUE ,
        main = 'Energy',
        xlab= 'Distribuição de Energy das Músicas')
```





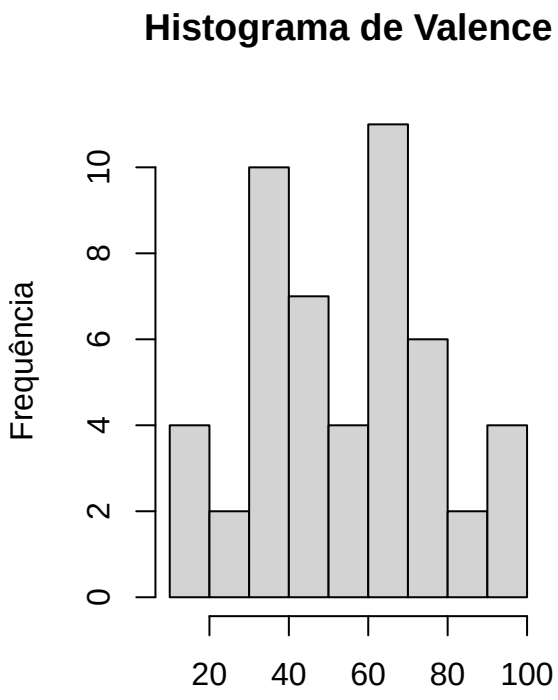

```
(summary(top50$Energy))
```

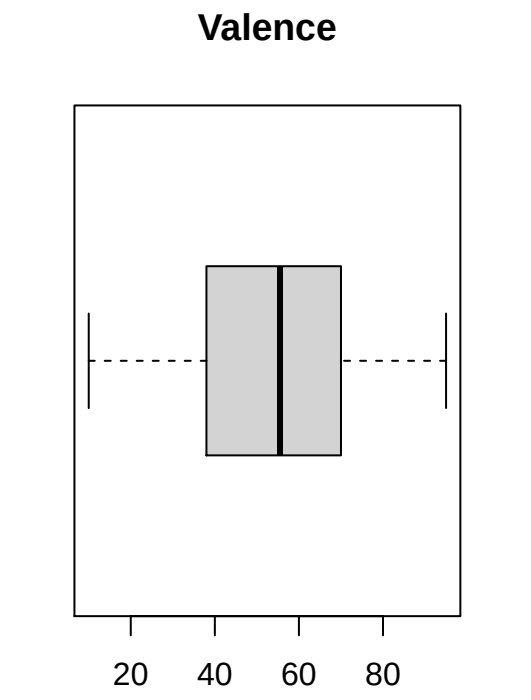
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.00   55.25   66.50   64.06   74.75   88.00
```

Podemos observar que há preferência por músicas mais animadas, visto que os valores de referência são altos.

Valence.

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Valence., main = 'Histograma de Valence', xlab = 'Distribuição de Valence', ylab= 'Frequênci
boxplot(top50$Valence., horizontal = TRUE ,
        main = 'Valence',
        xlab= 'Distribuição da Valence das Músicas')
```





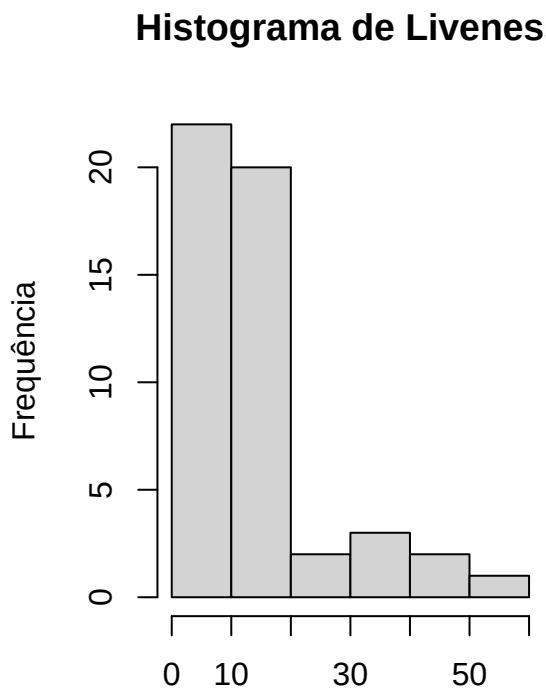

```
(summary(top50$Valence.))
```

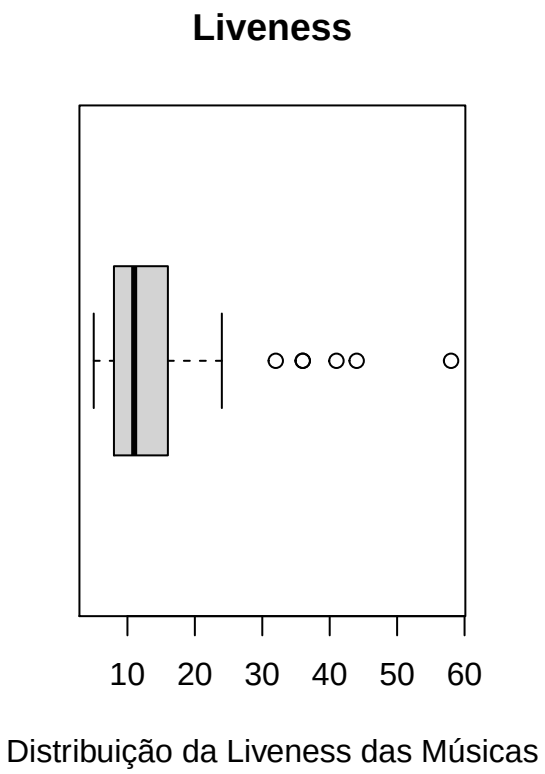
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   38.25   55.50   54.60   69.50   95.00
```

Podemos notar que há tanto músicas que passam uma energia positiva quanto músicas que passam uma energia negativa, visto que os valores ficam entre 40 e 70. Logo, o fato da musica ser ou nao "positiva" não influencia entrar no top50.

Liveness

```
par(mfrow=c(1,2))    # set the plotting area into a 1*2 array
hist(top50$Liveness, main = 'Histograma de Liveness', xlab = 'Distribuição de Liveness', ylab= 'Frequên
boxplot(top50$Liveness, horizontal = TRUE ,
        main = 'Liveness',
        xlab= 'Distribuição da Liveness das Músicas')
```






```
(summary(top50$Liveness))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    8.00   11.00   14.66   15.75   58.00
```

Podemos notar que não há tantas músicas gravadas com a presença de um público ao vivo. E que a grande maioria das musicas foi gravada em estudio.

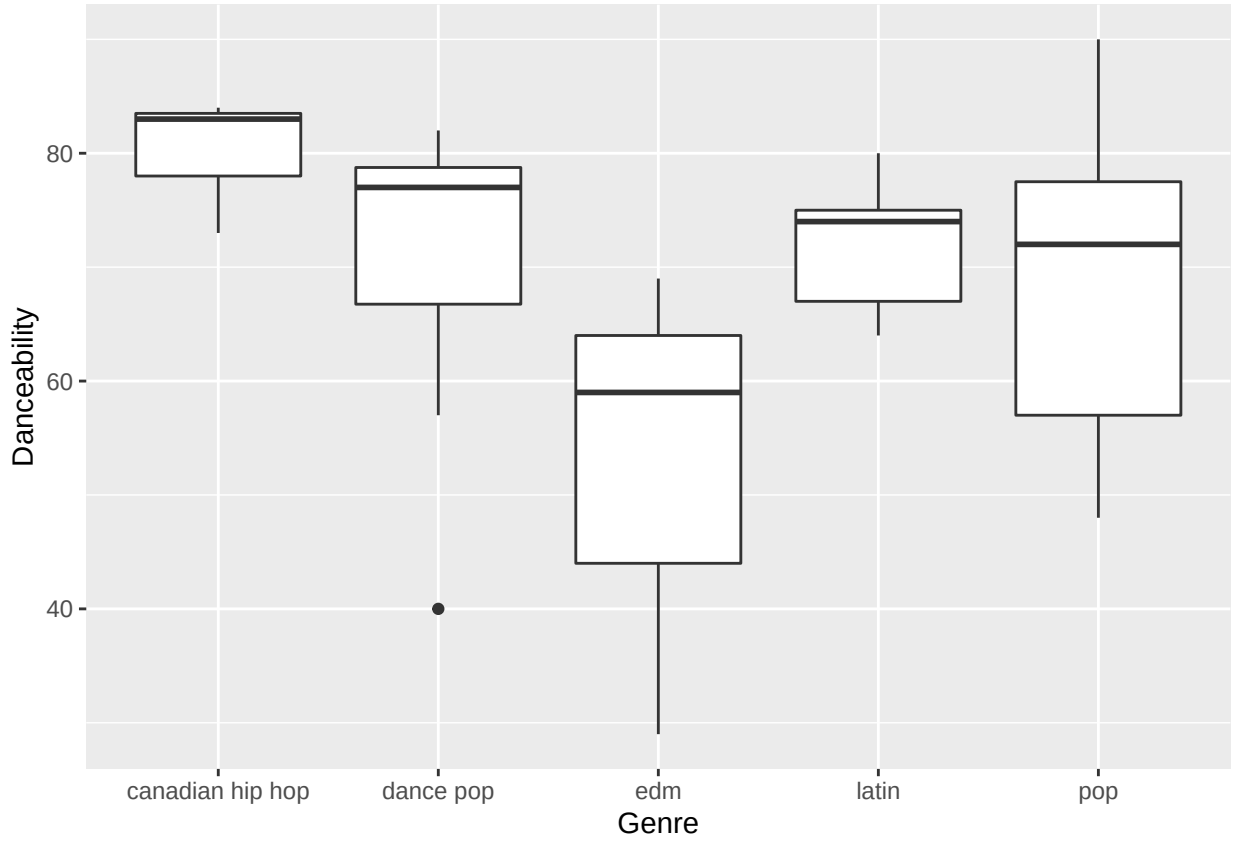
Análise Bidimensional

Seguiremos o trabalho com algumas analises bidimensionais entre nossas variáveis

Gênero x Dancebility

Gostariamos de verificar a premissa de que o genero influencia na "dançabilidade" da musica. COmo estamos tratando de uma naalise entre uma variável numérica e uma variável categorica, vamos plotar os boxplot por categoria, calcular as variancias por categoria e, por fim, calcular o grau de associatividade entre elas.

```
top5_gender = list('dance pop' ,'pop','latin','canadian hip hop', 'edm')
df_top5_gender = top50[top50$Genre %in% top5_gender,]
bi = ggplot( main='Análise Gênero x Dancebility',df_top5_gender,aes(x= Genre , y=Danceability))
bi+geom_boxplot()
```



Apenas com a análise do boxplot já podemos perceber uma grande mudança na distribuição da "dançabilidade" ao mudarmos de genero musical. É esperado então que o coeficiente de associatividade seja elevado.

```
# Calculo do coeficiente de associatividade
var_bar = 0
tot_lenght = 0
for(genero in top5_gender){
  df = top50[top50$Genre == genero,]
  var_bar = var_bar + var(df$Danceability)*length(df)
  tot_lenght = tot_lenght + length(df)
}
(var_bar/tot_lenght)/var(df_top5_gender$Danceability)
```

```
## [1] 0.9113492
```

Nosso coeficiente de associatividade é extremamente alto! Comprovamos que a "dançabilidade" de fato é muito influenciada pelo genero.

Danceability x Energy

Para analisar 2 variaveis numéricas, vamos calcular a correlacao entre elas, a covariancia e a reta de ajuste linear entre as variaveis.

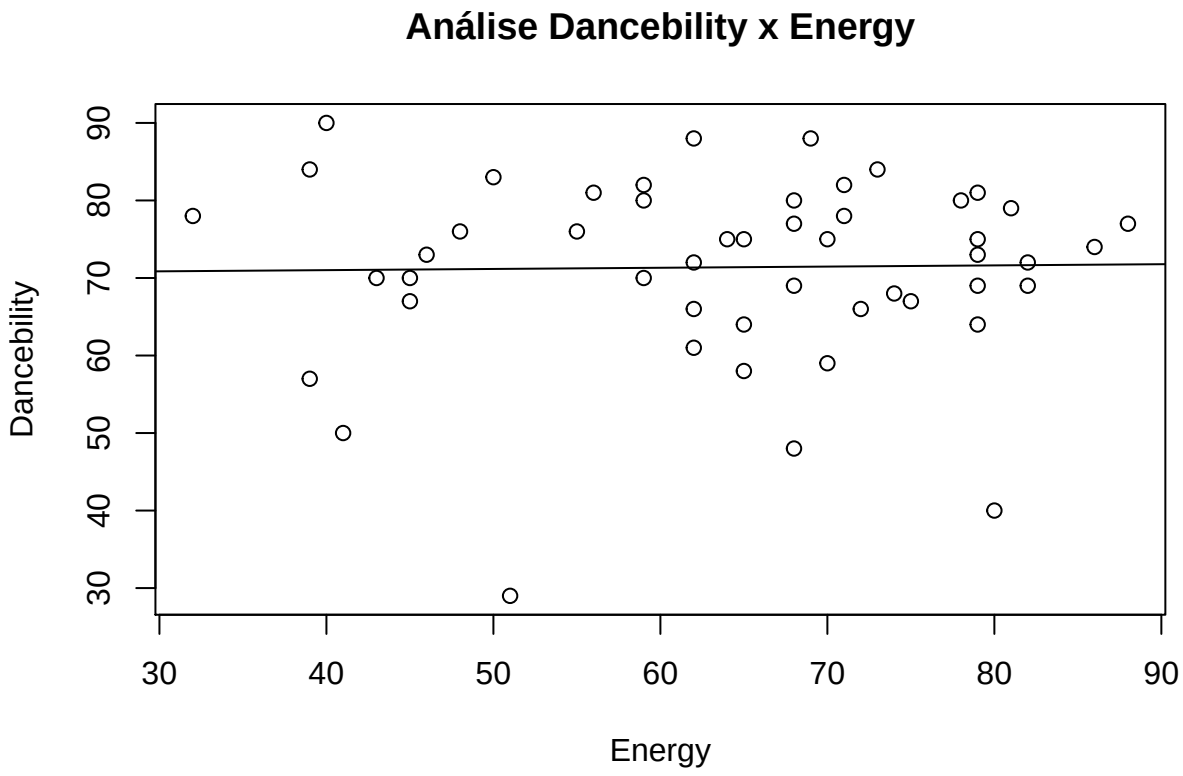
```
cor(top50$Danceability, top50$Energy)
```

```
## [1] 0.01825358
```

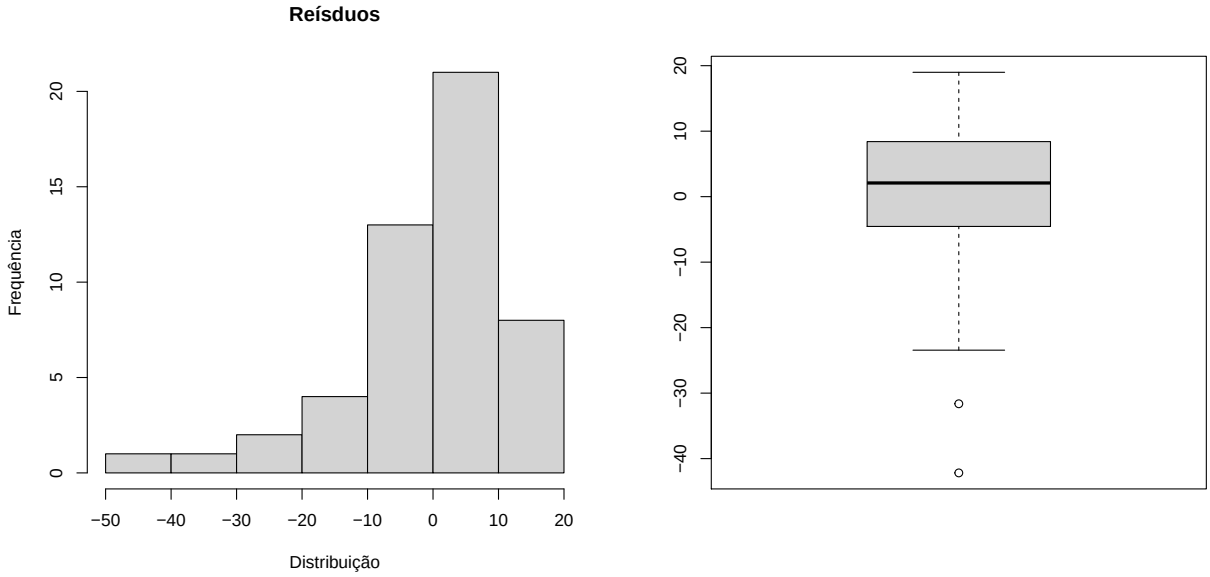
```
cov(top50$Danceability, top50$Energy)
```

```
## [1] 3.099184
```

```
fit = lm(top50$Danceability ~top50$Energy)
plot(x = top50$Energy , y = top50$Danceability, main='Análise Dancebility x Energy' , xlab='Energy' , yl
abline(fit)
```




```
par(mfrow=c(1,2))
hist(fit$residuals,ylab = 'Frequência', main = 'Reísduos' , xlab = 'Distribuição')
boxplot(fit$residuals)
```




```
print(kurtosis(fit$residuals))
```

```
## [1] 5.351196
```

```
print(skewness(fit$residuals))
```

```
## [1] -1.322673
```

Os residuos da regressão linear não seguem uma distribuição normal,pois existe uma assimetria a direita da media, logo, a premissa de dependencia linear entre as variaveis está errada. Calculamos também os coeficientes de curtose e assimetria, e ambos reforçaram que a distribuição dos residuos não é normal.

Genero vs Artista

No caso da análise entre duas variáveis categóricas, vamos calcular o coeficiente de Crames-V e as frequencias de cada subcategoria. Lembrando que , no nosso caso, essa naalise talvez não sej amuito interessanto pois em geral cada artista tem no máximo 3 musicas.

```
top5_artistas = list('Ed Sheeran','Ariana Grande','Billie Eilish','J Balvin', 'Lil Nas X')
df_top10_artistas = top50[top50$Artist.Name %in% top5_artistas,]
CrossTable(df_top10_artistas$Genre,df_top10_artistas$Artist.Name)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12
##
```

```
##
##                          | df_top10_artistas$Artist.Name
## df_top10_artistas$Genre  | Ariana Grande | Billie Eilish |    Ed Sheeran |      J Balvin |     Lil Nas
## -------------------------|---------------|---------------|---------------|---------------|------------
##              country rap |             0 |             0 |             0 |             0 |
##                          |         0.333 |         0.333 |         0.667 |         0.333 |         8.3
##                          |         0.000 |         0.000 |         0.000 |         0.000 |         1.0
##                          |         0.000 |         0.000 |         0.000 |         0.000 |         1.0
##                          |         0.000 |         0.000 |         0.000 |         0.000 |         0.1
## -------------------------|---------------|---------------|---------------|---------------|------------
##                dance pop |             2 |             0 |             0 |             0 |
##                          |         8.333 |         0.333 |         0.667 |         0.333 |         0.3
##                          |         1.000 |         0.000 |         0.000 |         0.000 |         0.0
##                          |         1.000 |         0.000 |         0.000 |         0.000 |         0.0
##                          |         0.167 |         0.000 |         0.000 |         0.000 |         0.0
## -------------------------|---------------|---------------|---------------|---------------|------------
##                electropop |             0 |             2 |             0 |             0 |
##                          |         0.333 |         8.333 |         0.667 |         0.333 |         0.3
##                          |         0.000 |         1.000 |         0.000 |         0.000 |         0.0
##                          |         0.000 |         1.000 |         0.000 |         0.000 |         0.0
##                          |         0.000 |         0.167 |         0.000 |         0.000 |         0.0
## -------------------------|---------------|---------------|---------------|---------------|------------
##                     latin |             0 |             0 |             0 |             2 |
##                          |         0.333 |         0.333 |         0.667 |         8.333 |         0.3
##                          |         0.000 |         0.000 |         0.000 |         1.000 |         0.0
##                          |         0.000 |         0.000 |         0.000 |         1.000 |         0.0
##                          |         0.000 |         0.000 |         0.000 |         0.167 |         0.0
## -------------------------|---------------|---------------|---------------|---------------|------------
##                       pop |             0 |             0 |             4 |             0 |
##                          |         0.667 |         0.667 |         5.333 |         0.667 |         0.0
##                          |         0.000 |         0.000 |         1.000 |         0.000 |         0.0
##                          |         0.000 |         0.000 |         1.000 |         0.000 |         0.0
##                          |         0.000 |         0.000 |         0.333 |         0.000 |         0.0
## -------------------------|---------------|---------------|---------------|---------------|------------
##              Column Total |             2 |             2 |             4 |             2 |
##                          |         0.167 |         0.167 |         0.333 |         0.167 |         0.1
## -------------------------|---------------|---------------|---------------|---------------|------------
##
##
```

A partir da tabela acima observa-se que cada artista tem seu gênero musical específico onde destaca-se no top50, sendo assim, seria menos provável que o mesmo artista aparece no top 50 com dois tipos de gêneros diferentes. Em exemplo, o cantor 'Lil Nax' aparece com o gênero 'Country Pop' , sendo seu único gênero musical prevalecente na lista.

Conclusão

Assim observa-se que primeiramente não há um domínio de um cantor ou de um gênero específicos no top 50. Contudo temos o dance pop. o pop e o latin como os principais gêneros musicais neste ranking. Além disso, observa-se que não há uma correlação entre as variáveis Dancebility x Energy, porém tais variáveis são importantes argumentos para saber se uma música tem maior probabilidade de estar no top ou não. Por fim, observa-se que os artistas focam em um gênero específico para estar no top 50, a probabilidade de um artista estar com dois gêneros diferentes é baixa.