**Insper – Instituto de Ensino e Pesquisa**

**ESTATÍSTICA 1 – ATIVIDADE PRÁTICA SUPERVISIONADA 2**

**Professor: Tatiana Terabayashi Melhado**

**Monitor: Marina Bicudo de Almeida Muradian**

Texto

Descrição gerada automaticamente com confiança baixa

**Andreas Azambuja Barbisan - andreasab@al.insper.edu.br**

**Amanda Caroline Ribeiro da Costa - amandacrc@al.insper.edu.br**

**Lucas Vilela Vila – lucasvv2@al.insper.edu.br**

**APS 2**

21/05/2023

**São Paulo**

**2023**

Maio/2023

Andreas Azambuja Barbisan, Amanda Caroline Ribeiro da Costa e Lucas Vilela Vila

**Atividade Prática Supervisionada 2**

O mercado de streaming musical

Trabalho apresentado nos cursos de

Administração e Economia no Instituto de Ensino e Pesquisa - Insper, na disciplina de Estatística, ministrada pela professora Tatiana Terabayashi Melhado.

SÃO PAULO

2023

**Sumário**

[**1.** **Introdução APS2** 4](#_Toc135317092)

[**2.** **Análise** 5](#_Toc135317093)

# **Introdução APS2**

Na APS2, nos foi instruído, através do arquivo e da base de dados disponibilizados no BlackBoard, investigar “se há diferenças no tipo de música consumido em plataformas concorrentes”, especificamente no Spotify e Youtube.

Para verificar esse funcionamento, o arquivo nos dá uma breve explicação sobre o mercado de música e uma pesquisa específica (aps2.xlsx), que nos permitirá averiguar a existência de diferenças entre os estilos musicais consumidos.

Dentro desse banco de dados, foi possível encontrar diversas variáveis para análise, entre elas: “Artist” (nome do artista da música), “Track” (nome da música), “Album” (álbum que a música em questão pertence), “Genre” (gênero da música em questão), “Energy” (medida de 0 a 1 que representa a intensidade de uma música), “Loudness” (volume geral de uma música em decibéis), “Speechiness” (detecta a presença de palavras faladas em uma faixa), “Acousticness” (medida de confiança de 0 a 1 que demonstra o quão acústica é a música), “Liveness” (detecta a presença de uma audiência em uma gravação, quanto maior mais provável que a faixa tenha sido tocada ao vivo), “Duration\_ms” (duração da faixa em milissegundos), “Stream” (quantidade de streams da música no Spotify), “Likes” (quantidade de likes da música no Youtube) e “Views” (quantidade de visualizações do vídeo no Youtube).

Para fazer a investigação desses fatores, fomos instruídos a usar o sistema de programação “RStudio”, linguagem usada principalmente para análise de dados e estatística. Esse software fornece uma ampla gama de funcionalidades para esse trabalho, como por exemplo a visualização de dados em tabelas e criação de gráficos que podem ajudar no estudo.

# **Análise**

Para começar a análise, foi necessário separar os dados obtidos em categorias (a), que podem ser: variáveis qualitativas nominais – dados que são representados por nomes ou categorias, sem ordem ou qualquer tipo de hierarquia; variáveis qualitativas ordinais – dados representados por nomes ou categorias com uma ordem de hierarquia; variáveis quantitativas discretas – dados numéricos que representam contagens ou números inteiros; variáveis quantitativas contínuas – dados numéricos que podem assumir qualquer valor em um intervalo contínuo.

Tendo em vista essas possíveis categorias de dados, a primeira coisa a se fazer como solicitado no arquivo do BlackBoard foi separar os dados que foram fornecidos, o que foi feito da seguinte maneira: “Artist”, “Track”, “Album”, “Genre” foram classificadas como qualitativa nominal, “Energy”, “Loudness”, “Speechiness”, “Acousticness”, “Liveness” foram classificadas como qualitativa contínua, “Duration\_ms”, “Stream”, “Likes” e “Views” foram classificadas como qualitativa discreta.

Para observar essas variáveis no R Studio, baixamos o banco de dados e o nomeamos de “APS”. Assim, pudemos fazer nosso primeiro estudo (b) dessas informações, que partiu da seguinte pergunta: A duração da música pode ser um fator relevante para o sucesso de uma música?

Para responder essa pergunta, usamos, como foi instruído, o esquema de 5 pontos. Para isso, utilizamos o “boxplot”, um gráfico que mostra a posição central, dispersão, simetria dos dados de uma amostra, cumprimento de caudas e dados discrepantes, para resumir as informações de um conjunto de dados através do código:

**boxplot(APS$Duration\_ms/1000,**

**col = "lightblue",**

**ylab = "Tempo (segundos)",**

**xlab = "Duração da música",**

**main = "Esquema de 5 pontos (Boxplot) da duração das músicas",**

**ylim = c(90, 350))**

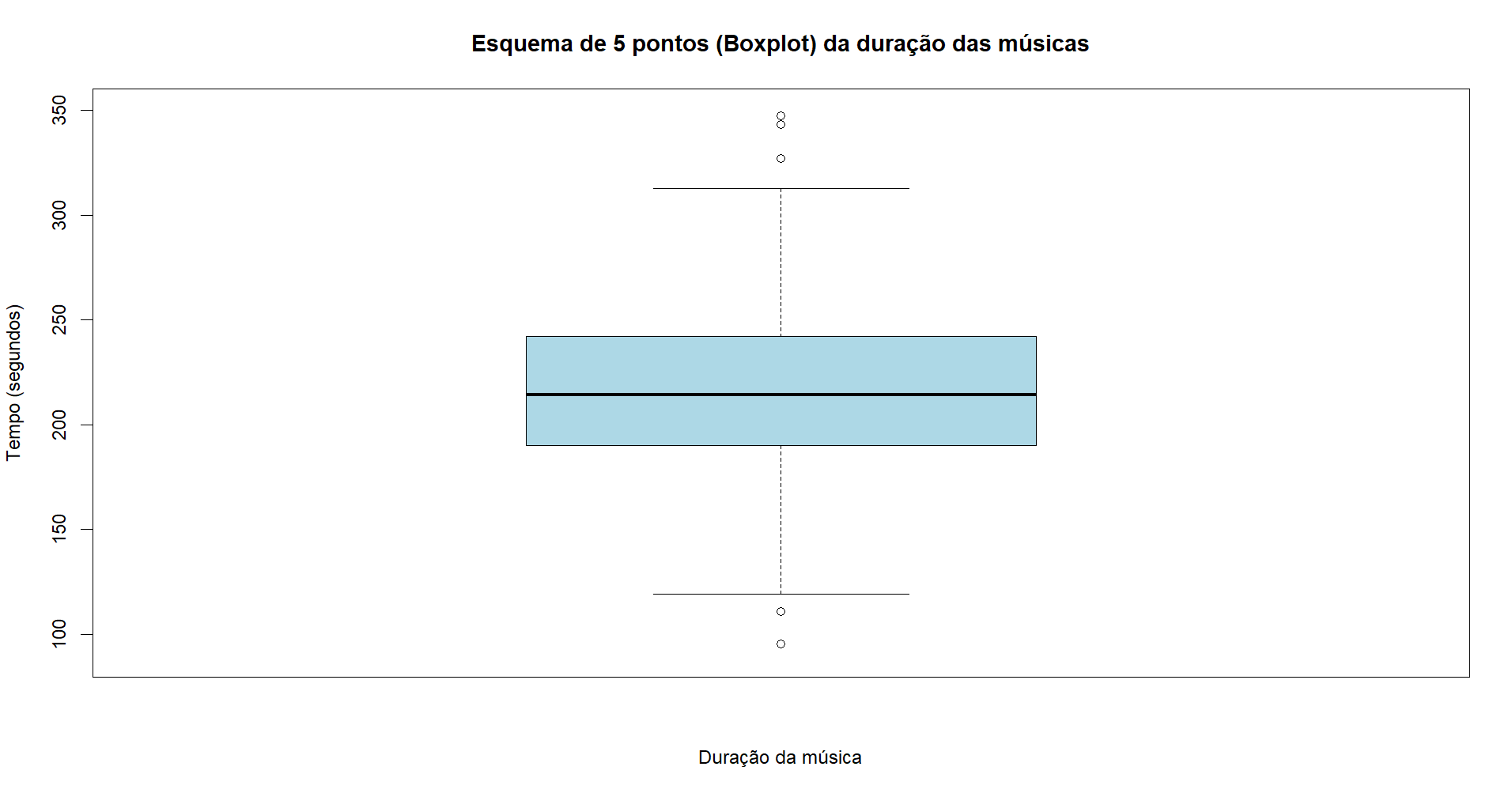
**quantileduration = quantile(APS$Duration\_ms)/1000**

**tabeladuration = round(cbind(quantileduration), 1)**

**colnames(tabeladuration) = c("Duração das músicas (segundos)")**

**View(tabeladuration)**

O qual gerou os seguintes resultados:

Imagem 1 - Boxplot da duração das músicas da base de dados. Confecção própria.

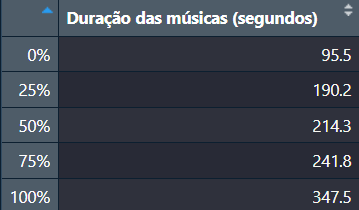


Imagem 2 - Quartis de duração das músicas. Confecção própria.

Sobre a distribuição dessa variável, podemos afirmar que existem 5 músicas com dados discrepantes e tempo maior em relação a todas as outras, 3 que possuem um tempo muito maior e 2 com um tempo consideravelmente menor.

Em relação às outras músicas, é possível perceber que o centro da distribuição, indicado pela linha preta no centro do retângulo, é próximo do centro absoluto da figura, deixando-a próxima de ter uma simetria perfeita, visto que o terceiro quartil é extremamente semelhante ao primeiro quartil: o Q3 com intervalo de 26.8 segundos e o Q1 com intervalo de 24.1. Porém, como estes não são iguais, já que o Q3 é maior que o Q1, ou como chamamos dentro de sala, que o “b” é maior que o "a”, podemos observar uma leve assimetria positiva. É avaliado também um intervalo interquartílico de 51.6 com uma mediana de 214.3, o que pode ser considerado de uma variabilidade expressiva. Podemos observar também que as 25% maiores observações possuem uma amplitude maior, de 105.7 contra 94.7 das 25% menores observações.

Já no ponto de análise (c), é solicitado para verificarmos através de boxplots “se há diferenças na duração das músicas para cada um dos gêneros musicais presentes na base de dados”.

Fizemos isso através do código:

**par(mfrow = c(1,1))**

**boxplot(APS$Duration\_ms[APS$Genre == "RnB"]/1000,**

**APS$Duration\_ms[APS$Genre == "Dark Trap"]/1000,**

**APS$Duration\_ms[APS$Genre == "Rap"]/1000,**

**col = c("darkblue", "grey", "purple"),**

**xlab = "Gêneros Músicais", ylab = "Tempo (s)",**

**main = "Duração das músicas por gênero musical",**

**ylim = c(90, 450),**

**cex.main = 1.3,**

**cex.lab = 1.3)**

**legend("topright",**

**legend = c("RnB", "Dark Trap", "Rap"),**

**fill = c("darkblue", "grey", "purple"),**

**cex = 0.70)**

O que gerou o seguinte resultado:

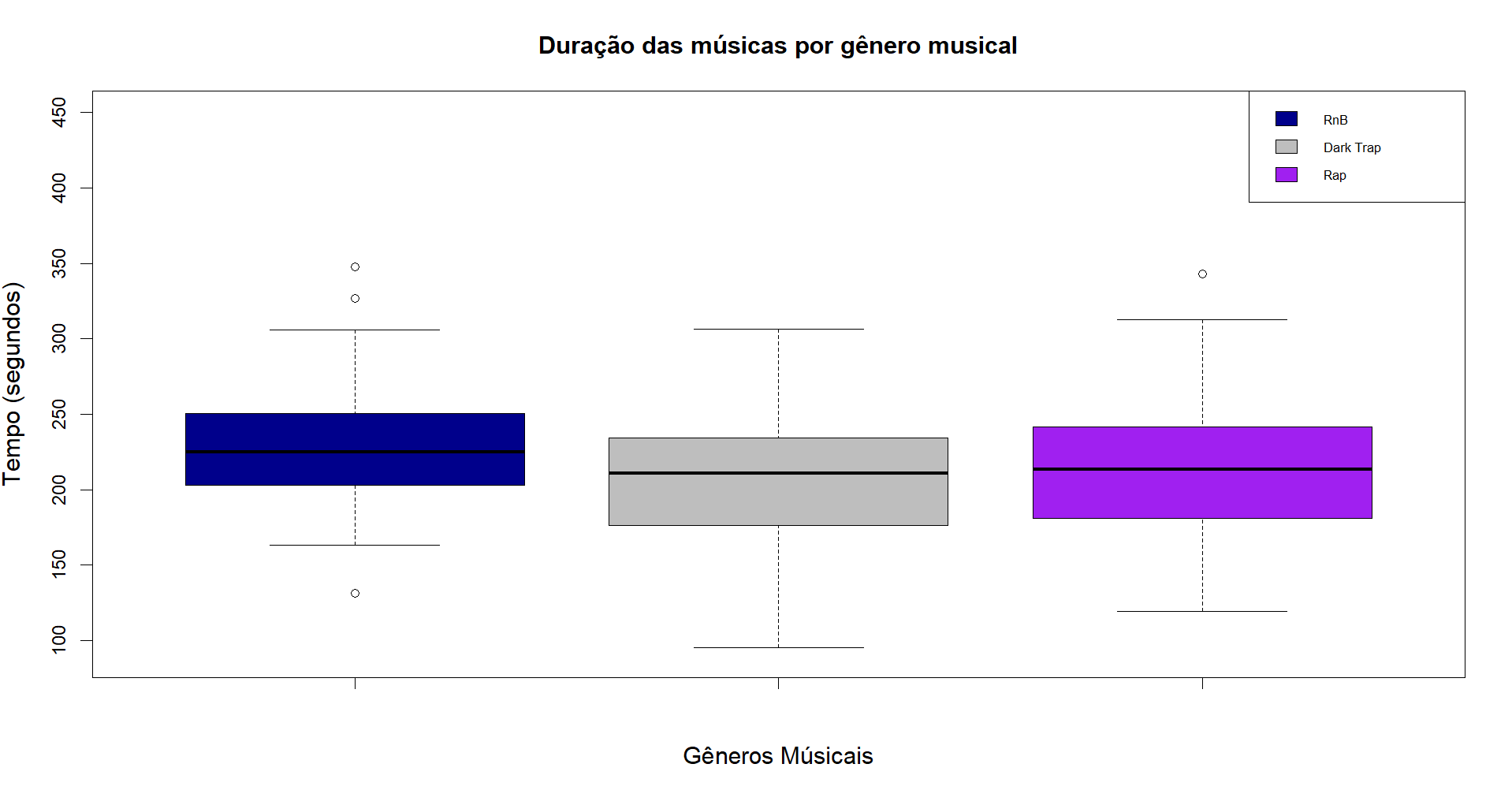


Imagem 3 - Boxplot da duração das músicas de cada gênero musical da base de dados. Confecção própria.

É possível ver nesse gráfico que as músicas possuem sim uma desigualdade de duração em relação ao gênero que estão inseridas, além de ser possível ver que a diferença de tempo que os diversos gêneros possuem entre a música com máximo e mínimo tempo é consideravelmente divergente. Apesar disso, cabe ressaltar que o tempo máximo das músicas é próximo, principalmente entre os gêneros “RnB” e “Dark Trap”, o que pode sugerir que no geral as músicas que têm mais sucesso têm um “limite de tempo” em que devem atuar.

Entre os pontos de análise (d) e (f) são separadas as músicas populares na plataforma Spotify. Dentro disso, no primeiro item é solicitada a análise de qual característica tida pelas músicas da plataforma parece ter o maior grau de associação com o sucesso da música (“variável Stream”).

Para isso, o grupo realizou a análise de correlação entre cada uma das características das músicas com o seu índice de sucesso através da variável “cor”, o que resultou nos códigos e tabela a seguir:

**# Item D**

**corenergy = cor(APS$Energy, APS$Stream)**

**corloudness= cor(APS$Loudness, APS$Stream)**

**corspeechiness = cor(APS$Speechiness, APS$Stream)**

**coracousticness = cor(APS$Acousticness, APS$Stream)**

**corliveness = cor(APS$Liveness, APS$Stream)**

**tabelacor = round(rbind(corenergy, corloudness, corspeechiness, coracousticness, corliveness), 2)**

**colnames(tabelacor) = c("Índice absoluto de correlação")**

**rownames(tabelacor) = c("Energy", "Loudness", "Speechiness", "Acousticness", "Liveness")**

**View(tabelacor)**

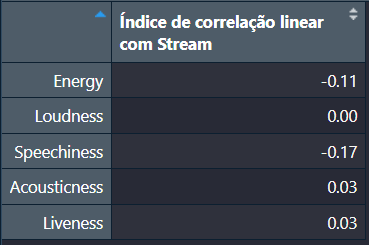


Imagem 4 - Tabela de correlação entre característica da música e índice de sucesso. Confecção própria.

Com a tabela feita, é possível afirmar que o índice absoluto com maior correlação entre o sucesso da música e as suas características é o aspecto “Speechiness”, com uma associação linear negativa entre as variáveis de -0,17. Também, é importante abordar que essa correlação é um índice de associação adequado, mas não necessariamente indica uma relação de causalidade, entretanto podem ter outros efeitos que fazem com que a música tenha sucesso, como por exemplo a data de lançamento ou sentimento da população em geral. O item (e), por sua vez, pede o mesmo que o item anterior, porém pede em relação a base de dados “Genre” e se volta em relação a pergunta: “Qual característica da música que parece ter maior grau de associação com seu sucesso. Há diferenças entre os gêneros?”. Para analisar o item, o grupo criou o código:

**# Item E**

**# RnB:**

**corrnbenergy = cor(APS$Energy[APS$Genre == "RnB"], APS$Stream[APS$Genre == "RnB"])**

**corrnbloudness = cor(APS$Loudness[APS$Genre == "RnB"], APS$Stream[APS$Genre == "RnB"])**

**corrnbspeechiness = cor(APS$Speechiness[APS$Genre == "RnB"], APS$Stream[APS$Genre == "RnB"])**

**corrnbacousticness = cor(APS$Acousticness[APS$Genre == "RnB"], APS$Stream[APS$Genre == "RnB"])**

**corrnbliveness = cor(APS$Liveness[APS$Genre == "RnB"], APS$Stream[APS$Genre == "RnB"])**

**tabelacorrnb = round(rbind(corrnbenergy, corrnbloudness, corrnbspeechiness, corrnbacousticness, corrnbliveness), 2)**

**colnames(tabelacorrnb) = c("RnB")**

**rownames(tabelacorrnb) = c("Energy", "Loudness", "Speechiness", "Acousticness", "Liveness")**

**# Dark Trap:**

**cordkenergy = cor(APS$Energy[APS$Genre == "Dark Trap"], APS$Stream[APS$Genre == "Dark Trap"])**

**cordkloudness = cor(APS$Loudness[APS$Genre == "Dark Trap"], APS$Stream[APS$Genre == "Dark Trap"])**

**cordkspeechiness = cor(APS$Speechiness[APS$Genre == "Dark Trap"], APS$Stream[APS$Genre == "Dark Trap"])**

**cordkacousticness = cor(APS$Acousticness[APS$Genre == "Dark Trap"], APS$Stream[APS$Genre == "Dark Trap"])**

**cordkliveness = cor(APS$Liveness[APS$Genre == "Dark Trap"], APS$Stream[APS$Genre == "Dark Trap"])**

**tabelacordk = round(rbind(cordkenergy, cordkloudness, cordkspeechiness, cordkacousticness, cordkliveness), 2)**

**colnames(tabelacordk) = c("Dark Trap")**

**rownames(tabelacordk) = c("Energy", "Loudness", "Speechiness", "Acousticness", "Liveness")**

**# Item E - Continuação**

**# Rap:**

**corrapenergy = cor(APS$Energy[APS$Genre == "Rap"], APS$Stream[APS$Genre == "Rap"])**

**corraploudness = cor(APS$Loudness[APS$Genre == "Rap"], APS$Stream[APS$Genre == "Rap"])**

**corrapspeechiness = cor(APS$Speechiness[APS$Genre == "Rap"], APS$Stream[APS$Genre == "Rap"])**

**corrapacousticness = cor(APS$Acousticness[APS$Genre == "Rap"], APS$Stream[APS$Genre == "Rap"])**

**corrapliveness = cor(APS$Liveness[APS$Genre == "Rap"], APS$Stream[APS$Genre == "Rap"])**

**tabelacorrap = round(rbind(corrapenergy, corraploudness, corrapspeechiness, corrapacousticness, corrapliveness), 2)**

**colnames(tabelacorrap) = c("Rap")**

**rownames(tabelacorrap) = c("Energy", "Loudness", "Speechiness", "Acousticness", "Liveness")**

**# União das três tabelas**

**tabelacorrelacoes = cbind(tabelacorrnb, tabelacordk, tabelacorrap)**

**View(tabelacorrelacoes)**

O que gerou o resultado:

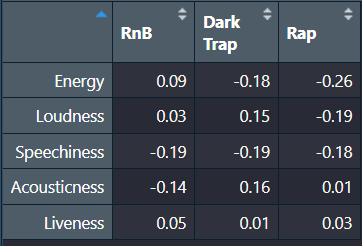


Imagem 5 - Tabela de Correlação entre característica e gênero musical. Confecção própria.

Existem grandes diferenças entre a correlações de cada característica musical e o seu gênero, como, por exemplo, o aspecto “Energy”, que para o gênero “Rap” tem a maior correlação (negativa) de toda tabela, e não só isso, mas para o gênero “RnB” passa a ter uma correlação consideravelmente menor e positiva. Diferenças similares acontecem em “Loudness” e “Acouticness”, porém, nas características “Speechiness” e “Liveness” as diferenças são menores, sendo que na primeira todos tem correlações negativas similares e no segundo todos tem correlações positivas pequenas.

Em relação ao ponto de análise (f), foi solicitado que estimássemos uma reta relacionando o sucesso de uma música com a característica selecionada como a mais importante para cada gênero musical. Para isso, fizemos o código a seguir:

**# Item F:**

**par(mfrow = c(1, 3))**

**#RnB**

**streamrnb = APS$Stream[APS$Genre == "RnB"]/1000000**

**speechinessrnb = APS$Speechiness[APS$Genre == "RnB"]**

**plot(streamrnb ~ speechinessrnb,main = "RnB",xlab = "Speechiness",ylab = "Sucesso da música (Streams) - em milhões de streams",cex = 1,pch = 18,col = "darkblue")**

**abline(lm(streamrnb ~ speechinessrnb))**

**coefrnb = round(coef(lm(streamrnb ~ speechinessrnb)), 2)**

**intercepto1 = coefrnb[1]**

**coef\_angular1 = coefrnb[2]**

**legend("topright", legend = c(paste("Intercepto:", round(intercepto1, 2)), paste("Coef. Angular:", round(coef\_angular1, 2))), bty = "i", cex = 0.8)**

**#Dark Trap**

**streamdk = APS$Stream[APS$Genre == "Dark Trap"]/1000000**

**speechinessdk = APS$Speechiness[APS$Genre == "Dark Trap"]**

**plot(streamdk ~ speechinessdk,main = "Dark Trap",xlab = "Speechiness",ylab = "Sucesso da música (Streams) - em milhões de streams",cex = 1,pch = 18,col = "purple")**

**abline(lm(streamdk ~ speechinessdk))**

**coefdk = round(coef(lm(streamdk ~ speechinessdk)), 2)**

**intercepto2 = coefdk[1]**

**coef\_angular2 = coefdk[2]**

**#Item F continuação:**

**legend("topright", legend = c(paste("Intercepto:", round(intercepto2, 2)), paste("Coef. Angular:", round(coef\_angular2, 2))), bty = "i", cex = 0.8)**

**#Rap**

**streamrap = APS$Stream[APS$Genre == "Rap"]/1000000**

**energyrap = APS$Energy[APS$Genre == "Rap"]**

**plot(streamrap ~ energyrap,main = "Rap",xlab = "Energia da música",ylab = "Sucesso da música (Streams) - em milhões de streams",cex = 1,pch = 18,col = "lightblue")**

**abline(lm(streamrap ~ energyrap))**

**coefrap = round(coef(lm(streamrap ~ energyrap)), 2)**

**intercepto3 = coefrap[1]**

**coef\_angular3 = coefrap[2]**

**legend("topright", legend = c(paste("Intercepto:", round(intercepto3, 2)), paste("Coef. Angular:", round(coef\_angular3, 2))), bty ="i",cex=0.8)**

O que gerou o gráfico a seguir:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Imagem 6: Gráfico de dispersão e reta estimada para relação entre sucesso e característica mais importante. Confecção própria.

Como é possível ver no gráfico, para as músicas do gênero “RnB” e “Dark Trap”, a característica mais importante a ser analisada é “Speechiness”, e, para o “Rap”, a mais importante é “Energy”. Para todas elas, observa-se que a reta resultante tem um grau de associação negativo. Porém, nota-se que a angulação é diferente, com uma tendo um impacto maior que a outra para cada 0.1 que passam no “eixo x”.

Partindo para o ponto de análise (g), foi solicitado um estudo do número de curtidas de um vídeo a fim de verificar "qual o gênero musical mais popular nesta plataforma". Com esta finalidade, o grupo escreveu o seguinte código:

**par(mfrow = c(1par(mfrow = c(1, 1))**

**boxplot(APS$Likes[APS$Genre=="RnB"]/1000000,**

**APS$Likes[APS$Genre=="Dark Trap"]/1000000,**

**APS$Likes[APS$Genre =="Rap"]/1000000,**

**main = "Número de likes por gênero musical (em milhões de likes)",**

**col= c("darkblue", "purple", "lightblue"),**

**ylim = c(0, 35))**

**legend("topright", legend = c("RnB", "Dark Trap", "Rap"), fill = c("darkblue", "purple", "lightblue"))**

Assim, obtivemos o seguinte resultado:

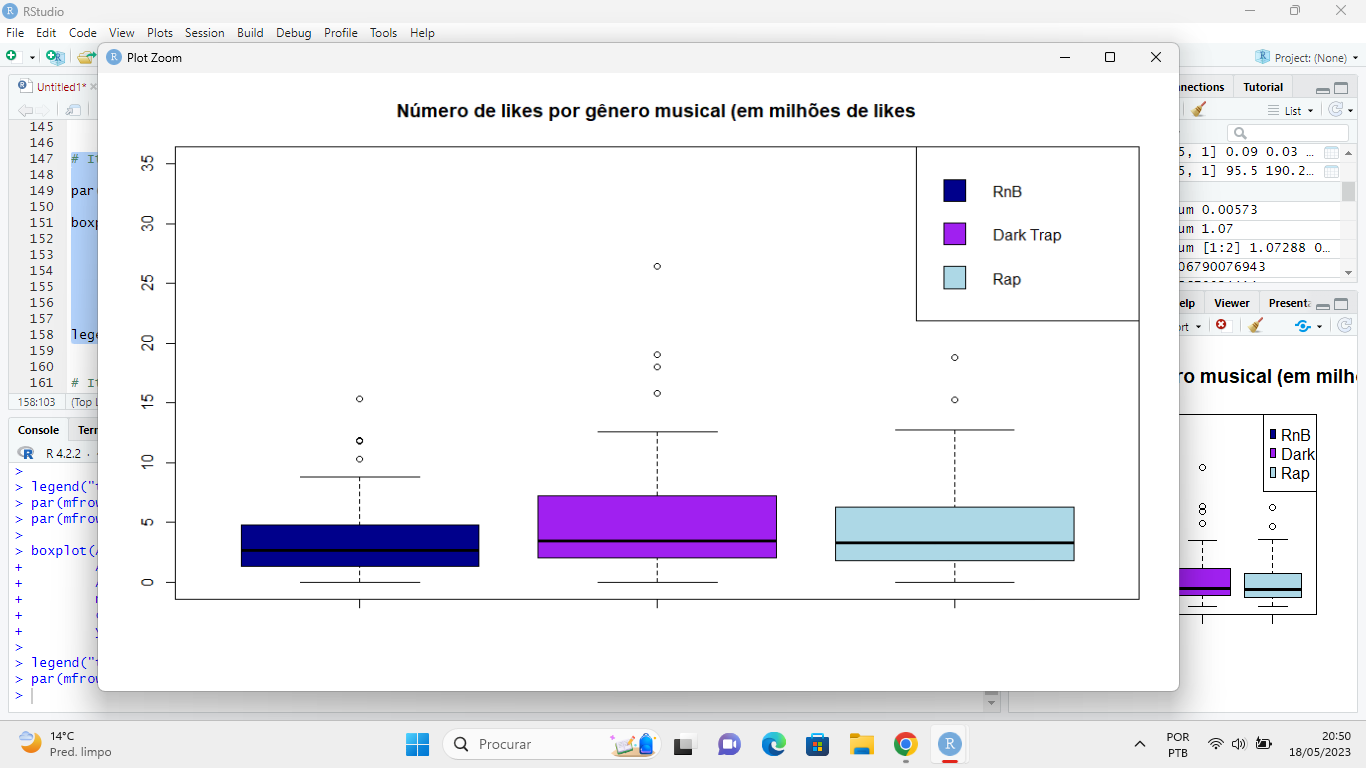


Imagem 7 - Boxplot do número de curtidas, em milhões, por gênero musical. Confecção própria.

A partir do gráfico, cabe julgar que, com base na posição do terceiro quartil dos 3 estilos musicais, há um favoritismo pelo gênero "Dark Trap", que, inclusive, possui o maior número atípico de curtidas. Ao mesmo tempo, este se assemelha ao “Rap” se compararmos seus valores máximo, mínimo e até mesmo a mediana, mas difere no alongamento do terceiro quartil. Também, vale estudar o "RnB", que possui um terceiro quartil bem mais estreito e um valor máximo bem menor do que dos outros dois gêneros, assim, sendo o estilo com menos curtidas na plataforma Youtube.

Por sua vez, o item (h) sugere uma nova hipótese para análise: “será que as mesmas músicas fazem sucesso em plataformas diferentes?”. Especificamente, abordaremos se há uma associação entre o desempenho de uma música na plataforma Spotify e o número de curtidas na plataforma Youtube. Produzimos um código a fim de responder esta pergunta:

**stream\_mi = APS$Stream/1000000**

**likes\_mi = APS$Likes/1000000**

**plot(likes\_mi ~ stream\_mi,**

**ylab = "Likes no YouTube (em milhões de likes)",**

**xlab = "Streams no Spotify (em milhões de streams)",**

**cex = 0.7,**

**cex.lab = 0.7,**

**pch = 18,**

**col = "darkblue",**

**main = "Streams no Spotify x Likes no YouTube")**

**abline(lm(likes\_mi ~ stream\_mi))**

**coefstream = round(coef(lm(likes\_mi ~ stream\_mi)), 2)**

**intercepto4 = coefstream[1]**

**coef\_angular4 = coefstream[2]**

**legend("topright", legend = c(paste("Intercepto:", round(intercepto4, 2)), paste("Coef. Angular:", round(coef\_angular4, 2))), bty = "i", cex = 0.7)**

**#Equação da reta:**

**coefreta = coef(lm(stream\_mi ~ likes\_mi))**

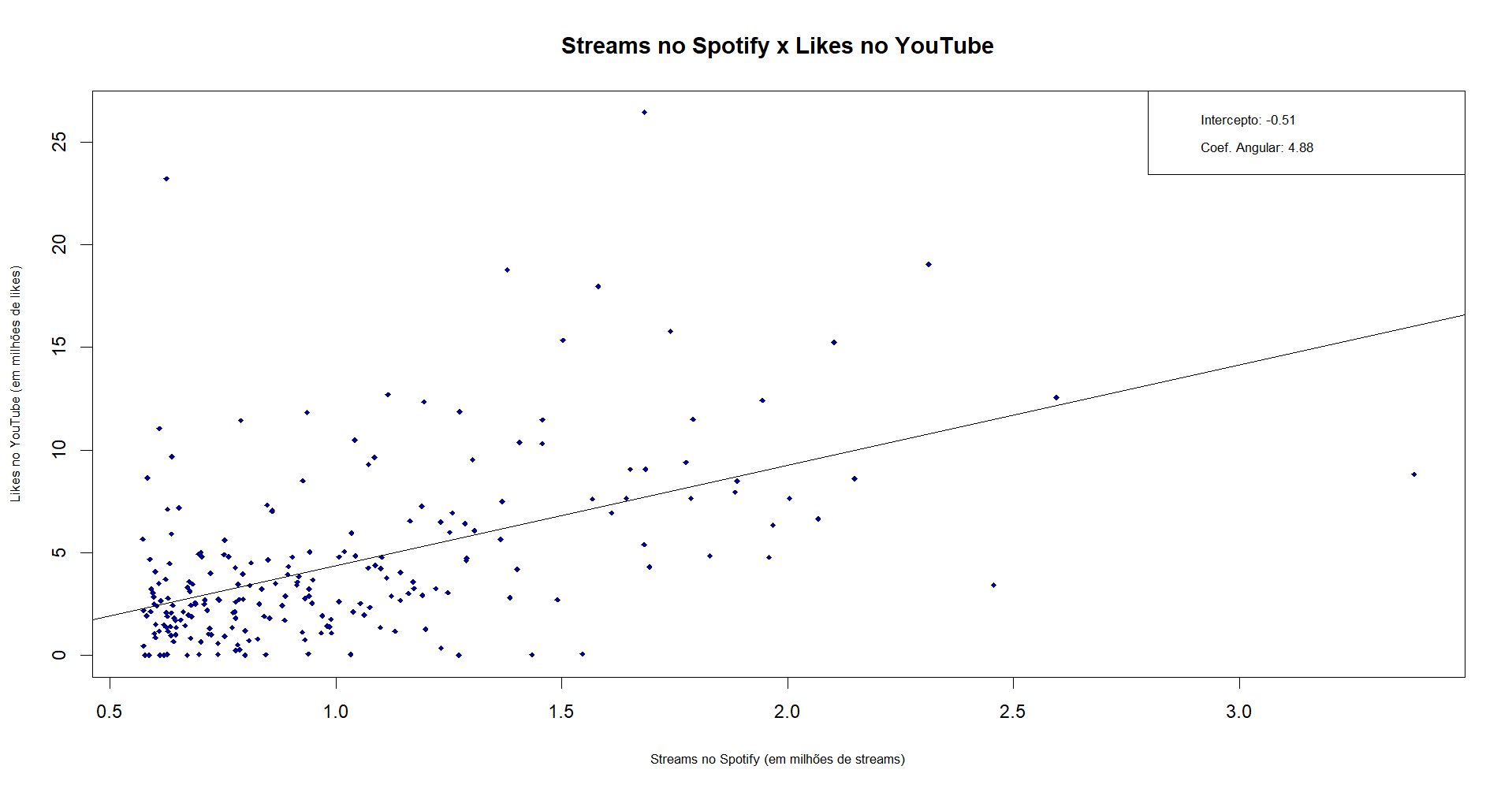
**a = coefreta[2]**

**b = coefreta[1]**

**#Correlação**

**correlH = cor(stream\_mi, likes\_mi)**

Com esta elaboração, obtivemos o seguinte resultado:

Imagem 8 - Gráfico de correlação entre streams no Spotify e curtidas no Youtube. Confecção própria.

Analisando o gráfico, podemos concluir que há uma correlação positiva entre streams no Spotify e curtidas no Youtube. E através do código elaborado pelo grupo, constatamos que esta correlação é de 0.5114. Ou seja, conclui-se que a correlação desta associação é moderada. Além do mais, pode-se afirmar, através do coeficiente angular de 4.88, que a cada aumento de um milhão no número de streams no Spotify, espera-se um acréscimo médio de 4.88 no número de curtidas (em milhão) no Youtube.

Abordando o item (i), foi-se pedido ao grupo que reavaliasse o item anterior, porém, agora, “utilizando o número de visualizações do vídeo como proxy para o sucesso deste no Youtube”. Com este objetivo, desenvolvemos o seguinte código:

**likes\_mi = APS$Likes/1000000**

**views\_mi = APS$Views/1000000**

**plot(likes\_mi ~ views\_mi,**

**ylab = "Likes no YouTube (em milhões de likes)",**

**xlab = "Views no Youtube (em milhões de views)",**

**cex = 0.7,**

**cex.lab = 0.7,**

**pch = 18,**

**col = "darkblue",**

**main = "Views no Yotube x Likes no YouTube")**

**abline(lm(likes\_mi ~ views\_mi))**

**coefstream2 = round(coef(lm(likes\_mi ~ views\_mi)), 2)**

**intercepto5 = coefstream2 [1]**

**coef\_angular5 = coefstream2 [2]**

**legend("topright", legend = c(paste("Intercepto:", round(intercepto5, 2)), paste("Coef. Angular:", round(coef\_angular5, 2))), bty = "i", cex = 0.7)**

**#Equação da reta:**

**coefreta = coef(lm(likes\_mi ~ views\_mi))**

**a = coefreta[2]**

**b = coefreta[1]**

**#Correlação**

**correlI = cor(likes\_mi, views\_mi)**

A partir disto, geramos o gráfico a seguir:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Imagem 9 - Gráfico de correlação entre visualizações no Youtube e streams no Spotify. Confecção própria.

Observa-se que há uma correlação positiva entre as visualizações e as curtidas no Youtube. Segundo os cálculos feitos através do código desenvolvido, podemos afirmar que esta correlação é de 0.4188, assim, considera-se uma correlação moderada. Com base no coeficiente angular da linha de tendência, espera-se um acréscimo médio de 626.2024 no número de streams (em milhão) no Spotify a cada aumento de um milhão no número de visualizações na plataforma.

Por fim, o ponto de análise (j) nos direciona a comparar as investigações dos dois últimos itens e constatar “qual medida de sucesso no Youtube parece ter maior grau de associação com o sucesso da música medido no Spotify”. De primeira, podemos afirmar que, por o número de curtidas ter uma correlação mais forte comparada ao do número de visualizações, a associação com maior grau é a likes-streams. Além de que o coeficiente angular da associação views-streams é muito grande, o que mostra o distanciamento entre os números de ambos.