Capítulo 1

Introdução

É difícil encontrar uma definição para o que é música, uma vez que ela possui um caráter de abstração e individualidade, mas de maneira geral, podemos dizer que música são sons organizados e padronizados que fazem parte de todas as culturas do mundo. Segundo o relatório da Federação Internacional da Industria Fonográfica (IFPI) divulgado no dia 4 de maio de 2020, a receita total do mercado mundial de música gravada cresceu 8,2%, subindo para US$ 20,2 bilhões referente ao ano de 2019. Além disso, pela primeira vez representou mais da metade (56,1%) da receita mundial de música gravada. A partir desses dados, percebe-se que o mercado musical é muito valioso e em constante evolução. Saber usá-lo e conhecer como essa grande engrenagem funciona é de grande interesse para aqueles que querem se aventurar nesse ramo. Será que é possível identificar um padrão nas músicas que mais repercutiram no mundo e replicá-los? Seria possível identificar um gênero musical mais popular em todos os países do mundo?

Capítulo 2

Banco de dados

Selecionamos um banco de dados, com 484 observações e 14 variáveis, referente as 484 músicas mais ouvidas no mundo pelo aplicativo Spotify entre o período de 2010 a 2019.

Das 14 variáveis, temos 4 qualitativas, e 10 covariáveis quantitativas.

As 4 variáveis qualitativas são:

- Nome da música
- Nome do artista
- Gênero (Gênero da musica)
- Ano de lançamento da música

As 10 covariáveis quantitativas são:

- Batidas por minuto (O ritmo da música)
- Energia (quanto mais alto o valor, mais enérgico é a música)
- Habilidade de dança (Quanto mais alto o valor, mais fácil é dançar essa música)
- Intensidade em dB (decibéis)
- Vivacidade (Quanto mais alto o valor, maior a probabilidade de a música ser uma gravação ao vivo.)
- Valência (Quanto mais alto o valor, mais positivo será o clima da música.)
- Comprimento (A duração da música em segundos)
- Acústica (Quanto mais alto o valor, mais acústica é a música)
- Discurso (Quanto mais alto o valor, mais palavra falada a música contém.)
- Popularidade (Quanto maior o valor, mais popular é a música)

	title	artist	top genre	year	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop
1	Hey, Soul Sister	Train	neo mellow	2010	97	89	67	-4	8	80	217	19	4	83
2	Love The Way You Lie	Eminem	detroit hip hop	2010	87	93	75	-5	52	64	263	24	23	82
3	TiK ToK	Kesha	dance pop	2010	120	84	76	-3	29	71	200	10	14	80
4	Bad Romance	Lady Gaga	dance pop	2010	119	92	70	-4	8	71	295	0	4	79
5	Just the Way You Are	Bruno Mars	pop	2010	109	84	64	-5	9	43	221	2	4	78
6	Baby	Justin Bieber	canadian pop	2010	65	86	73	-5	11	54	214	4	14	77
7	Dynamite	Taio Cruz	dance pop	2010	120	78	75	-4	4	82	203	0	9	77
8	Secrets	OneRepublic	dance pop	2010	148	76	52	-6	12	38	225	7	4	77
9	Empire State of Mind (Part II) Broken Down	Alicia Keys	hip pop	2010	93	37	48	-8	12	14	216	74	3	76
10	Only Girl (In The World)	Rihanna	barbadian pop	2010	126	72	79	-4	7	61	235	13	4	73
11	Club Can't Handle Me (feat. David Guetta)	Flo Rida	dance pop	2010	128	87	62	-4	6	47	235	3	3	73
12	Marry You	Bruno Mars	pop	2010	145	83	62	-5	10	48	230	33	4	73
13	Telephone	Lady Gaga	dance pop	2010	122	83	83	-6	11	71	221	1	4	73
14	Like A G6	Far East Movement	dance pop	2010	125	84	44	-8	12	78	217	1	45	72
15	OMG (feat. will.i.am)	Usher	atl hip hop	2010	130	75	78	-6	36	33	269	20	3	72
16	Eenie Meenie	Sean Kingston	dance pop	2010	121	61	72	-4	11	83	202	5	3	71
17	The Time (Dirty Bit)	The Black Eyed Peas	dance pop	2010	128	81	82	-8	60	44	308	7	7	70
18	Alejandro	Lady Gaga	dance pop	2010	99	80	63	-7	36	37	274	0	5	69
19	Your Love Is My Drug	Kesha	dance pop	2010	120	61	83	-4	9	76	187	1	10	69
20	Meet Me Halfway	The Black Eyed Peas	dance pop	2010	130	63	80	-7	32	40	284	0	7	68
21	Whataya Want from Me	Adam Lambert	australian pop	2010	186	68	44	-5	6	45	227	1	5	66
22	Take It Off	Kesha	dance pop	2010	125	68	73	-5	9	74	215	0	3	66
23	Misery	Maroon 5	pop	2010	103	81	70	-5	22	73	216	0	4	65
24	All The Right Moves	OneRepublic	dance pop	2010	146	95	53	-4	28	65	238	26	5	65
25	Animal	Neon Trees	indie pop	2010	148	83	48	-6	38	74	212	0	4	65
26	Naturally	Selena Gomez & The Scene	dance pop	2010	133	90	61	-5	5	88	203	2	5	64
27	I Like It	Enrique Iglesias	dance pop	2010	129	94	65	-3	6	73	231	2	9	63

Figura 2.0.1: Breve visualização dos dados

Capítulo 3

Objetivo

O objetivo deste trabalho, é realizar e apresentar uma análise multivariada, visando estudar e compreender a relação e o comportamento das covariveis, suas medidas descritivas e observações dos dados referente as 484 músicas mais ouvidas no mundo pelo aplicativo Spotify entre o período de 2010 a 2019.

Capítulo 4

Resultados

4.1 Análise Descritiva

Tabela 4.1: Dispersão dos dados de cada variável

Variável	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Amplitude
bpm	43.0	100.0	120.0	130.0	206.0	163
enrg	14.0	62.0	73.0	81.0	98.0	84
danc	23.0	58.0	66.0	74.0	97.0	74
dB	-12.0	-6.0	-5.0	-4.0	-2.0	10
viva	3.0	9.0	12.0	23.3	74.0	71
vale	4.0	36.0	52.0	70.3	98.0	94
comp	115.0	201.0	219.5	239.2	424.0	309
acus	0.0	2.0	6.0	16.3	97.0	97
disc	3.0	4.0	6.0	10.0	46.0	4
popu	0.0	66.0	72.0	79.0	95.0	95

4.2 Vetor de médias

Tabela 4.2: Resultados obtidos pelo Vetor de Médias.

Covariáveis	**Média**
bpm	119.7
enrg	70.5
danc	64.9
dB	-5.4
viva	17.5
vale	52.9
comp	223.3
acus	13.8
disc	8.9
popu	71.7

4.3 Matriz de Variância e Covariância (Cov)

$Cov =$

	bpm	enrg	danc	dB	viva	vale	comp	acus	disc	popu
bpm	652.65									
enrg	23.78	239.03								
danc	-66.00	11.88	170.20							
dB	0.51	15.19	1.58	2.55						
viva	11.27	34.12	-9.75	1.34	159.30					
vale	-28.34	125.37	131.90	11.34	-3.80	489.74				
comp	-12.22	-32.07	-62.72	-5.55	57.00	-160.62	1191.75			
acus	-26.65	-168.95	-45.87	-8.32	-35.67	-90.12	20.34	380.44		
disc	26.23	3.61	-6.11	-2.25	10.48	16.20	6.92	6.50	65.34	
popu	5.26	-44.73	26.13	-2.31	-20.24	-5.52	-71.01	31.40	4.69	142.33

Na matriz de variância e covariância, os elementos fora da diagonal contêm as covariâncias de cada par de variáveis, como as covariâncias são diferentes de zero, temos indícios de que as variáveis são correlacionadas. Já os elementos da diagonal contêm as variâncias de cada variável, medindo como os dados estão espalhados em torno da média. A matriz de variância e covariância reflete a dispersão dos dados, em outras palavras, a relação linear entre os vetores de covariaveis.

Além da matriz de variância e covariância, temos duas outras métricas, que também podem ser usadas pra refletir a dispersão dos vetores de dados, elas são a Variância Total e a Variância Generalizada:

- Variância Total : $tr(Cov) = 3493.33$, ou seja, a soma das variâncias das covariáveis é 3493.33.
- Variância Generalizada : $det(Cov) = 3.854161e{+}21$, ou seja existe uma variabilidade de 3.854161e+21 em nosso conjunto de variáveis.

4.3.1 Matriz de Correlação

$Cor =$

	bpm	enrg	danc	dB	viva	vale	comp	acus	disc	popu
bpm	1.00									
enrg	0.06	1.00								
danc	-0.20	0.06	1.00							
dB	0.01	0.62	0.08	1.00						
viva	0.03	0.17	-0.06	0.07	1.00					
vale	-0.05	0.37	0.46	0.32	-0.01	1.00				
comp	-0.01	-0.06	-0.14	-0.10	0.13	-0.21	1.00			
acus	-0.05	-0.56	-0.18	-0.27	-0.14	-0.21	0.03	1.00		
disc	0.13	0.03	-0.06	-0.17	0.10	0.09	0.02	0.04	1.00	
popu	0.02	-0.24	0.17	-0.12	-0.13	-0.02	-0.17	0.13	0.05	1.00

A correlação é uma associação estatística usualmente utilizada para medir o grau de relação entre um par de variáveis. Os valores de correlação vão de -1 até 1. Se a correlação for positiva, significa que o aumento de uma variável é acompanhado pelo aumento da outra, já com uma correlação negativa, temos o inverso, o aumento de uma variável é acompanhado pela diminuição da outra e vice-versa.

Na matriz de correlação, temos os elementos fora da diagonal principal como as correlações entre os pares de variáveis. Como podemos observar, não temos nenhum valor zero, o que significa que todos os pares de variáveis possuem uma correlação, podendo ser positiva ou negativa, conforme explicito na matriz.

A Figura 4.3.1 abaixo, representa visualmente a matriz de correlações entre as variáveis, as cores mais intensas indicam um maior grau de correlação.

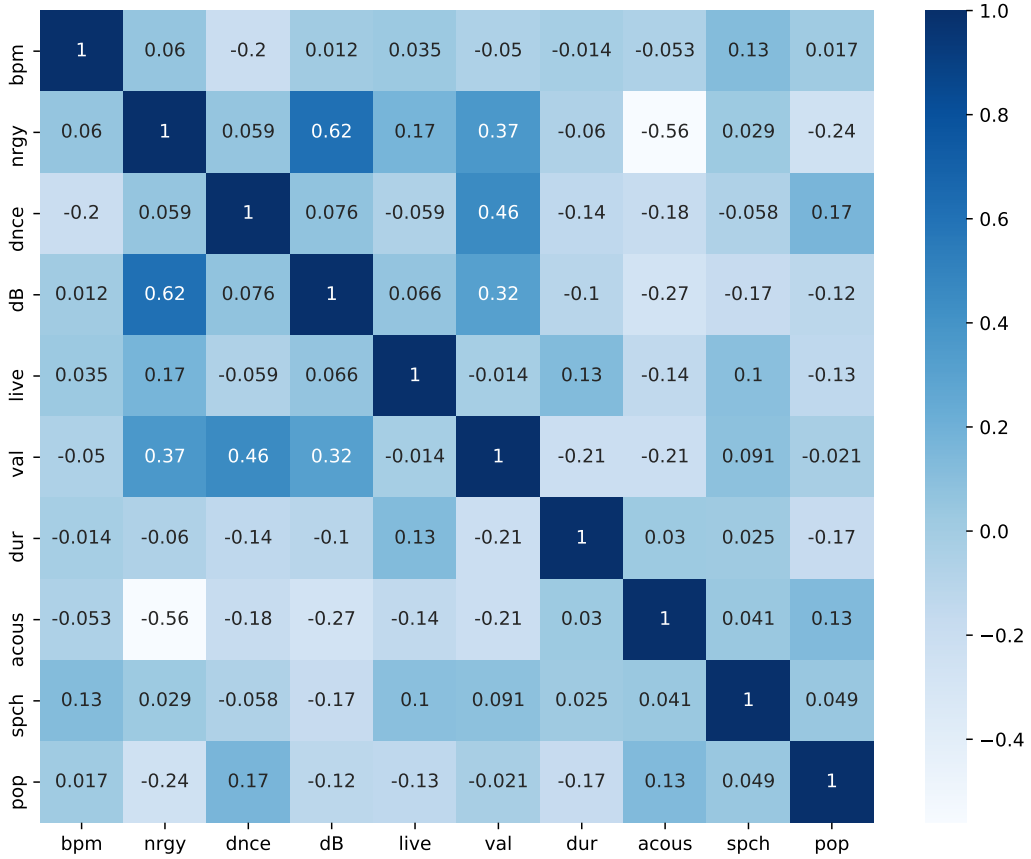

Figura 4.3.1: Mapa de calor contendo a matriz de correlação dos dados

Com o mapa de calor da matriz de correlação acima, é possível perceber que a maior correlação positiva se apresenta entre as variáveis Energia e Intensidade (em decibéis), ou seja, a medida que uma cresce a outra acompanha seu crescimento. Também nota-se

que a maior correlação negativa ocorre entre as variáveis Acústica e Energia, ou seja, a medida que uma aumenta a outra diminui. Vale ressaltar que, como a correlação entre diferentes covariáveis não é zero, temos que essas covariáveis não são independentes.

4.3.2 Visualização dos Dados

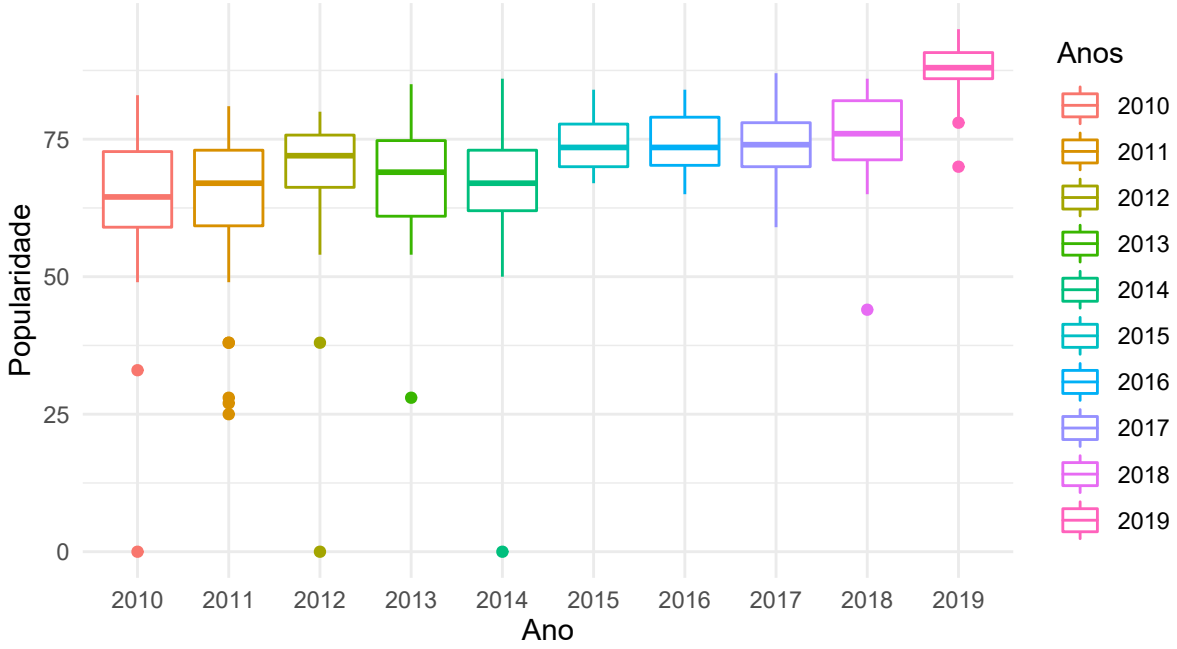

Figura 4.3.2: Dispersão dos dados da variável popularidade da música por ano.

Ao analisar o gráfico da Figura 4.3.2, podemos dizer que há evidências de que o nível de popularidade foi aumentando ao longo dos últimos 10 anos. Os dados do ano de 2019 estão mais concentrados, apresentando uma variabilidade menor em relação aos outros anos e alocados nos pontos mais altos do eixo y.

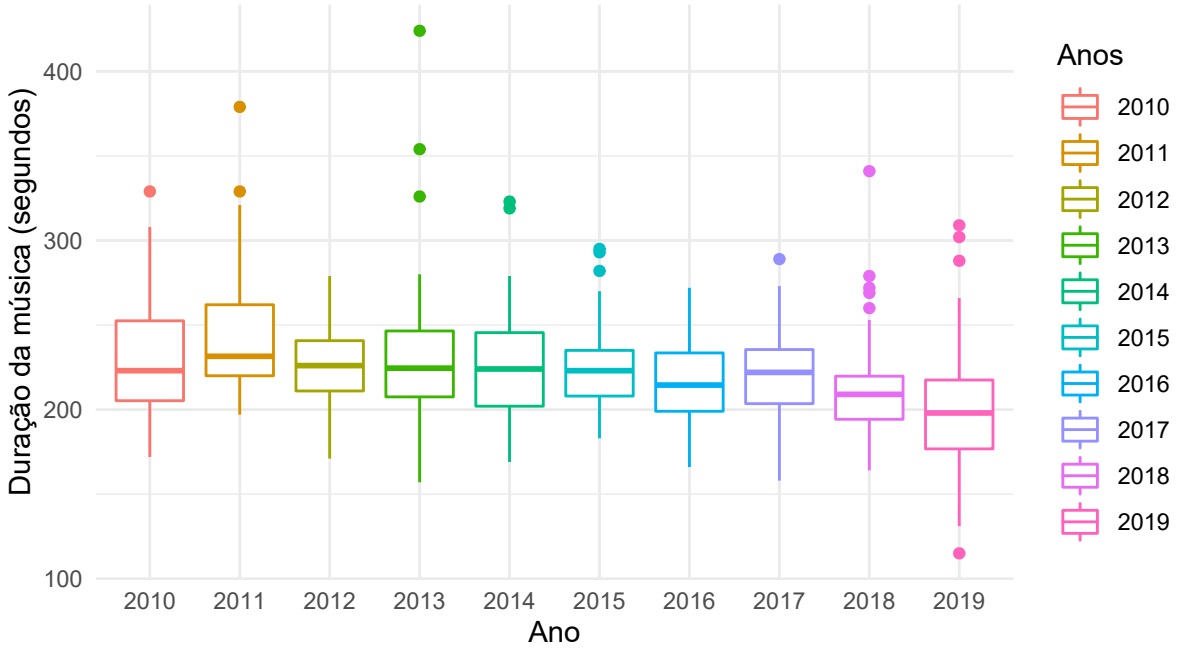

Figura 4.3.3: Dispersão dos dados da variável comprimento da música por ano.

Já a Figura 4.3.3, expressa a duração da música em comparação aos anos. Ao analisar os boxplots, percebe-se que não existem grandes evidências de que o tamanho das canções se alterou. No ano de 2019 observa-se minimamente que a distribuição dos dados estão mais abaixo de que as demais caixas do boxplot. O ano de 2013 tem a música com maior duração, ultrapassando 400 segundos enquanto que em 2019 temos a música de menor duração, sendo ela, próxima a 100 segundos.

4.4 Distância estatística.

A distância estatística é a distância entre dois pontos no espaço multivariado, mesmo para pontos correlacionados, uma vez que ela considera a sua covariância.

Para realizar a distância entre o conjunto de dados e o vetor de médias, consideramos apenas as variáveis definidas como quantitativas na Página 3 e a matriz de covariâncias.

Dessa forma,

Seja $P_i = (x_{1i}, x_{2i}, x_{3i}, \ldots, x_{10i})$ e $Q = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{10})$, com $i = 1, 2, 3 \ldots, 484$

Em que:

x_{ji} é a $i - ésima$ observação da variável j;
$\bar{x}_1$ é a média da $i - ésima$ variável;
S é a matriz de covariância da amostra.

A distância entre P e Q é dada por:

$$D(P_i, Q) = \sqrt{(P_i - Q)^t \cdot S^{-1} \cdot (P_i - Q)}$$

E parte dos resultados estão apresentados na Tabelaa 4.3 abaixo.

$D(P_i, Q)$	**Distância**
$D(P_1, Q)$	7.37
$D(P_2, Q)$	20.36
$D(P_3, Q)$	5.29
$D(P_4, Q)$	11.21
$D(P_i, Q)$	3.56
...	...
$D(P_{485}, Q)$	29.72
$D(P_{481}, Q)$	10.56
$D(P_{482}, Q)$	5.82
$D(P_{483}, Q)$	9.38
$D(P_{484}, Q)$	5.32

Tabela 4.3: Distância estatística entre os pontos P_i e Q.

A partir dos resultados obtidos, temos que a maior distância (46.82) entre os pontos P e Q ocorreu para o ponto P_{50}, correspondente a canção *Hello* do artista Martin Solveig, enquanto que a menor distância distância (1.50) se deu para o ponto P_{314}, correspondente a música *Close* do artista Nick Jonas.

4.5 Testes de normalidade multivariada

Nesta etapa vamos verificar se nosso banco de dados possui variáveis que apresentam distribuição normal, também iremos verificar se nossos dados são de uma amostra normal multivariada de dimensão 10, com vetor de medias μ e matriz de covariâncias Σ.

Inicialmente afim de checar a normalidade univarida, para cada variável separadamente das demais, fizemos o QQ-Plot de cada variável :

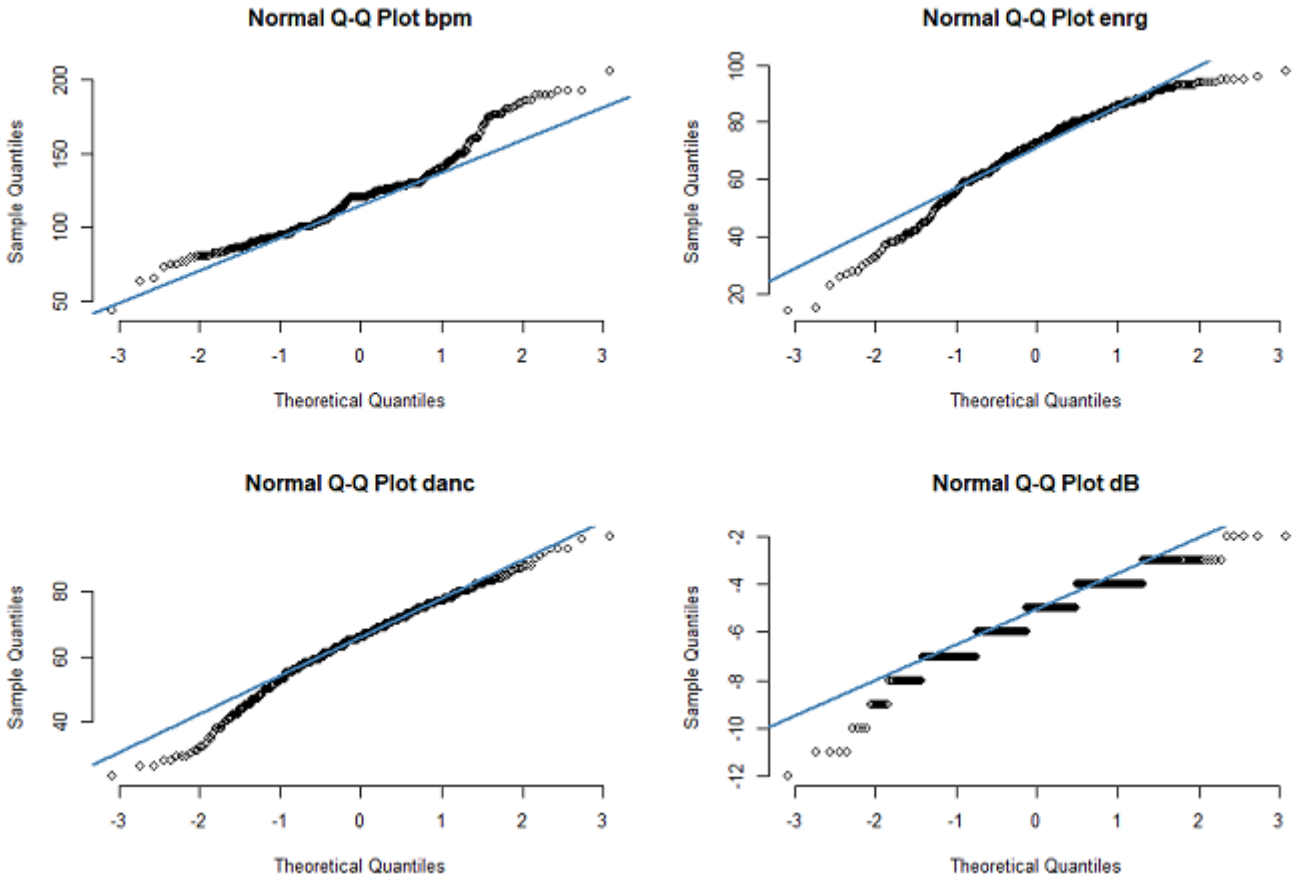

Figura 4.5.1: Gráficos QQ-Plot das covariaveis bpm,enrg,danc e dB.

Ao observarmos a Figura 4.5.1, notamos que, nos gráficos QQ-Plot bpm, QQ-Plot enrg, QQ-Plot danc e QQ-Plot dB, os pontos não parecem se distribuir muito bem ao longo da reta identidade, indicando que as covariáveis bpm, enrg, danc e dB não possuem distribuição normal.

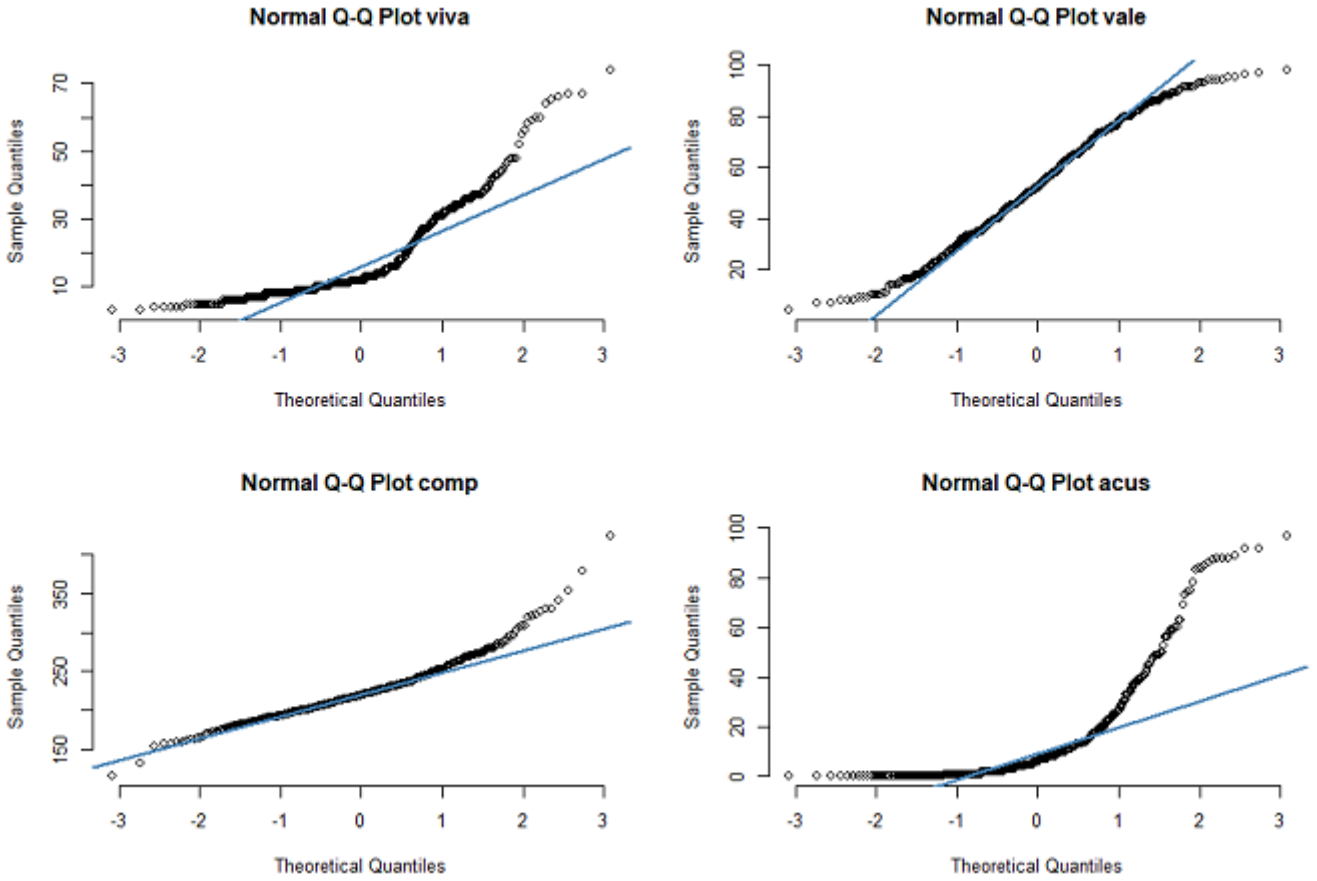

Figura 4.5.2: Gráficos QQ-Plot das covariaveis viva,vale, comp e acus.

Da mesma forma, ao observarmos a Figura 4.5.2, notamos que, nos gráficos QQ-Plot viva, QQ-Plot vale, QQ-Plot comp e QQ-Plot acus os pontos não parecem se distribuir muito bem ao longo da reta identidade, indicando que as covariáveis viva, vale, comp e acus não possuem distribuição normal.

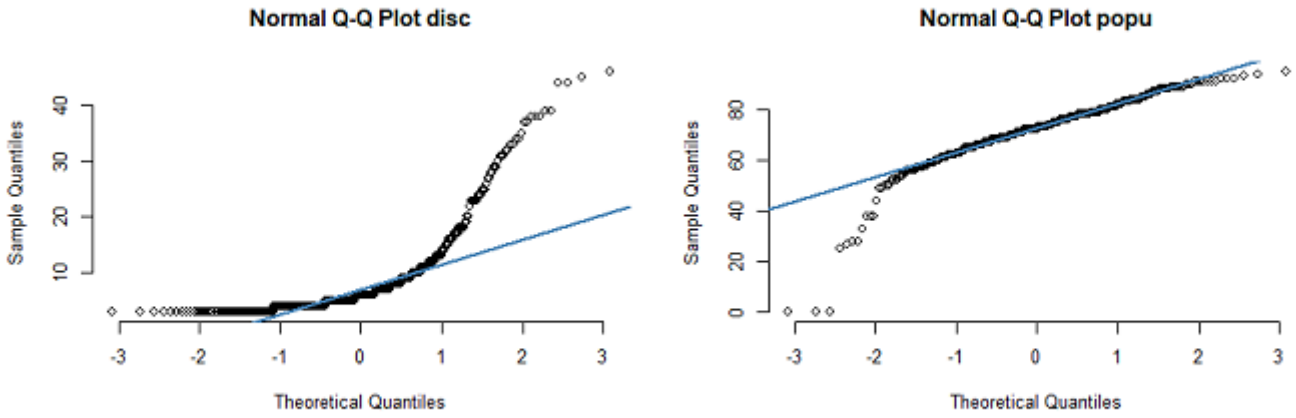

Figura 4.5.3: Gráficos QQ-Plot das covariaveis disc e popu.

Ao observarmos a Figura 4.5.3, notamos que, nos gráficos QQ-Plot disc os pontos não parecem se distribuir muito bem ao longo de uma reta, quanto ao QQ-Plot popu, temos que os pontos se concentram perto da reta, mas se afastam nas extremidades, indicando que as covariáveis disc e popu não possuem distribuição normal.

Assim, pela análise gráfica, temos indícios que nenhuma das covariáveis em estudo possui normalidade, para verificar se os dados de fato possuem ou não distribuição normal, a priori, para fins de estudos, foi utilizado o auxílio dos testes de normalidade vistos em aula, como o de Shapiro-Wilk, Anderson-Darling e Lilliefors (Kolmogorov-Smirnov), esses realizam o teste para cada variável separadamente das demais, e todos são baseados nas seguintes hipóteses:

- H_0 : Os dados provém de uma distribuição Normal.
- H_1 : Os dados não provém de uma distribuição Normal.

Tabela 4.4: Resultado dos Testes de Normalidade

	P - Valor		
Covariável	**Lilliefors (Kolmogorov-Smirnov)**	**Anderson-Darling**	**Shapiro-Wilk**
bpm	2,20E-16	1,912E-11	1,74E-07
enrg	4,59E-05	9,81E-11	3,71E-08
danc	1,27E-02	5,60E-04	9,15E-04
dB	2,20E-16	2,20E-16	2,83E-10
viva	<2,2E-16	<2,6E-16	<2,2E-16
vale	0,009542	0,0005703	1,84E-02
comp	5,01E-07	1,03E-09	1,30E-09
acus	<2,6E-16	<2,6E-16	<2,6E-16
disc	<2,6E-16	<2,6E-16	<2,6E-16
popu	6,05E-10	<2,6E-16	<2,6E-16

Através dos resultados obtidos acima, podemos notar que nenhuma covariável que compõe nosso banco de dados provém de uma distribuição Normal. Ao nível de significância de $\alpha = 0.05$ rejeitamos a hipótese nula para todas as variáveis presentes.

É importante que se tenha um estudo Multivariado para os testes de Normalidade, para isto foi aplicado os testes de Shapiro-Wilk e Anderson-Darling, ambos para Normalidade Multivariada.

Sob as seguintes Hipóteses:

- H_0 : Os dados provém de uma distribuição Normal.
- H_1 : Os dados não provém de uma distribuição Normal.

Tabela 4.5: Resultado do Teste de Shapiro-Wilk para Normalidade Multivariada

MVW	**P - Valor**
0.90892	2.2e-16

Tabela 4.6: Resultado do Teste de Anderson-Darling para Normalidade Multivariada

Teste	**Variável**	**Estatística**	**P - Valor**	**Normalidade**
Anderson-Darling	bpm	58.629	<0.001	Não
Anderson-Darling	nrgy	55.659	<0.001	Não
Anderson-Darling	dnce	27.690	<0.001	Não
Anderson-Darling	dB	98.369	<0.001	Não
Anderson-Darling	live	316.693	<0.001	Não
Anderson-Darling	val	15.403	6.00e-04	Não
Anderson-Darling	dur	51.392	<0.001	Não
Anderson-Darling	acous	489.521	<0.001	Não
Anderson-Darling	spch	503.433	<0.001	Não
Anderson-Darling	pop	74.984	<0.001	Não

Novamente, através dos P-Valores obtidos em ambos os testes, podemos afirmar que os dados não provém de uma distribuição Normal. Ao nível de significância de $\alpha = 0.05$ rejeitamos a hipótese nula para a Normalidade Multivariada.

Como a normalidade univariada foi rejeitada para todas as variáveis, esperava-se que a normalidade multivariada também fosse, assim como foi visto nos resultados dos testes Shapiro-Wilk e Anderson-Darling.

Uma alternativa a se pensar nessa situação, seria a utilização do Teorema Central do Limite, que nos diz que,

Teorema Central do Limite Seja $x_1, x_2, ..., x_n$ vetores aleatórios independentes de alguma distribuição com vetor de medias μ e matriz de var-cov Σ, então :

$$\sqrt{n}(\bar{x} - \mu) \overset{\text{a}}{\sim} N(0, \Sigma) \tag{4.1}$$

, para grandes amostras $n > p$.

Porém, não temos informações relevantes sobre as covariáveis serem ou não vetores aleatórios independentes, e como foi observado na seção 4.3.1, pela matriz de correlação dos dados e pelo Mapa de calor da matriz de correlação, não temos indícios de que as covariáveis sejam independentes. Sendo assim, como não podemos confirmar a suposição do Teorema Central do Limite de que os vetores aleatórios são independentes, não podemos, neste caso, utilizar o teorema.

Referências Bibliográficas

Report ifpi 2020. https://www.ifpi.org/resources/. acessado em 29/08/2021.