Relatório – Projeto Final da Matéria COE241 (2019.2)

Thiago Guimarães Rebello Mendonça de Alcântara
DRE: 118053123
Código: https://github.com/guim4dev/FinalProjectCOE241

Objetivo e Introdução

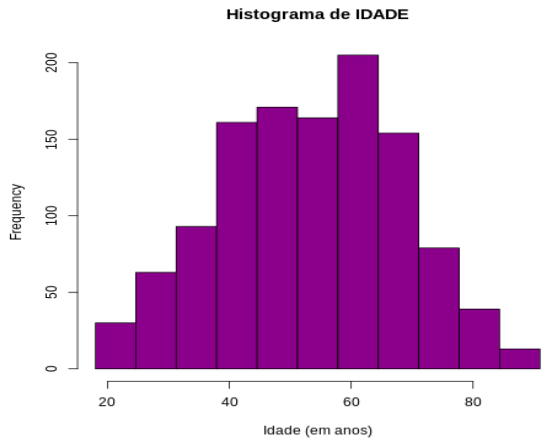
O objetivo deste trabalho é estudar um conjunto de dados aplicando a teoria aprendida em classe. Utilizamos dados reais obtidos a partir de uma extensa base de dados do Prof. Claudio Gil, coletada durante muitos anos e usada em suas pesquisas. Os dados mostram uma medida da condição aeróbica do paciente (o VO2 max) (por quilo de peso do indivíduo) e ainda as variáveis idade, peso e a carga máxima atingida durante um teste ao qual o paciente foi submetido. Todos os pacientes são masculinos. O VO2 max é a taxa máxima de consumo de oxigênio medida durante um teste de esforço, e reflete a capacidade aeróbica do paciente, expressa em volume de oxigênio por massa corporal por minuto (ml/(Kg.min).
Para realização das análises, utilizei a linguagem R e o Rstudio. Os gráficos gerados estão, também, no repositório no github.
Referências da linguagem R: https://www.rdocumentation.org/

Análises

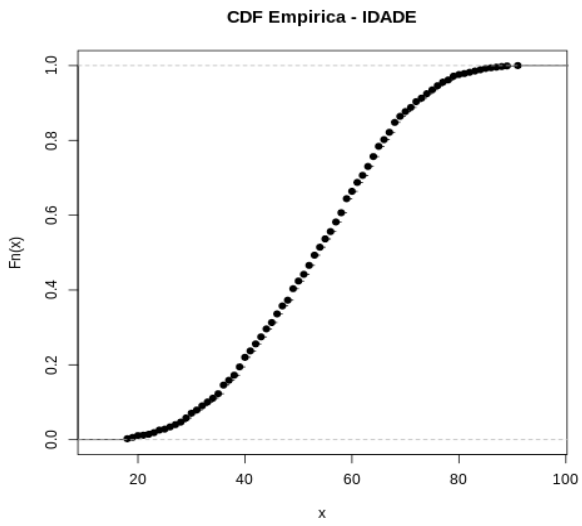
1) Histogramas, pdf e cdf das variáveis

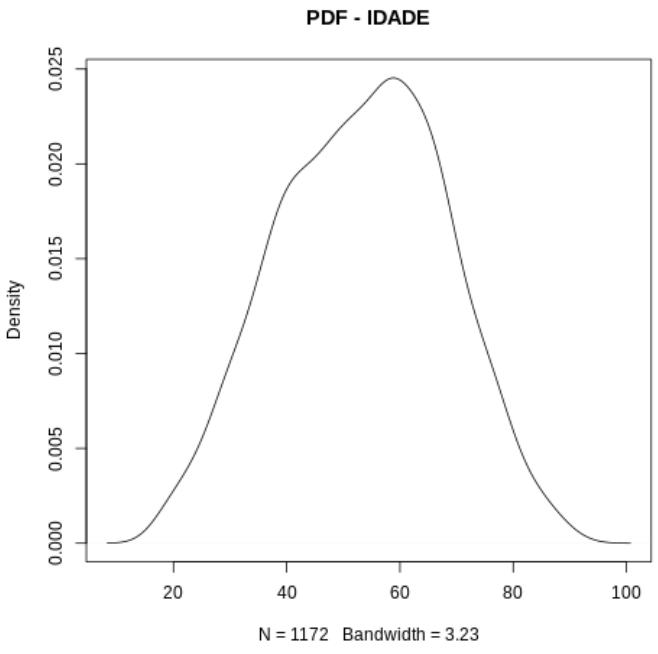
O bin dos histogramas foi estimado a partir da Regra de Sturge (bin = 1 + 3.3*log10(n), sendo n = número de medidas). Tendo em vista que todos os dados possuem o mesmo número de medidas, o bin é o mesmo para os 4 histogramas. O valor encontrado foi bin = 11.

- *IDADE*:

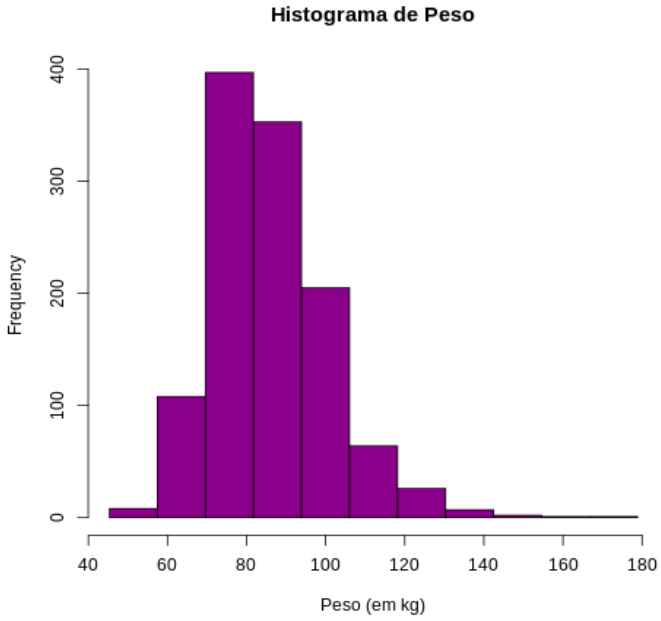


→ A partir do histograma e da PDF, fica notável que: possuímos uma variância grande para a variável idade e não há uma grande concentração da variável em uma área específica dos gráficos.

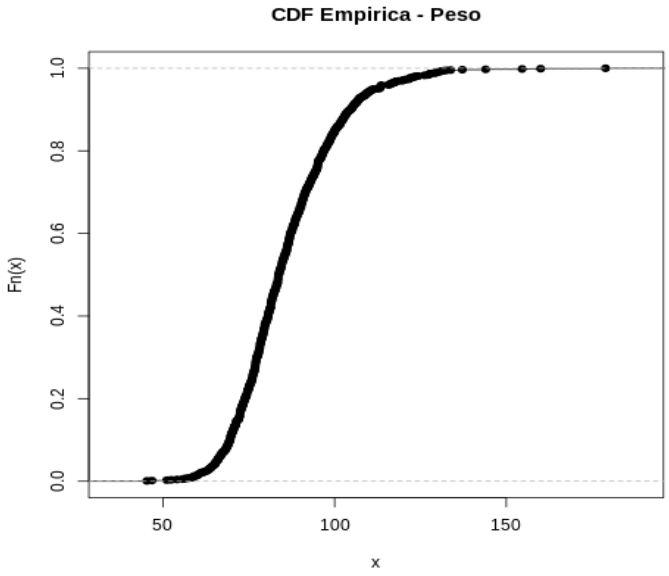


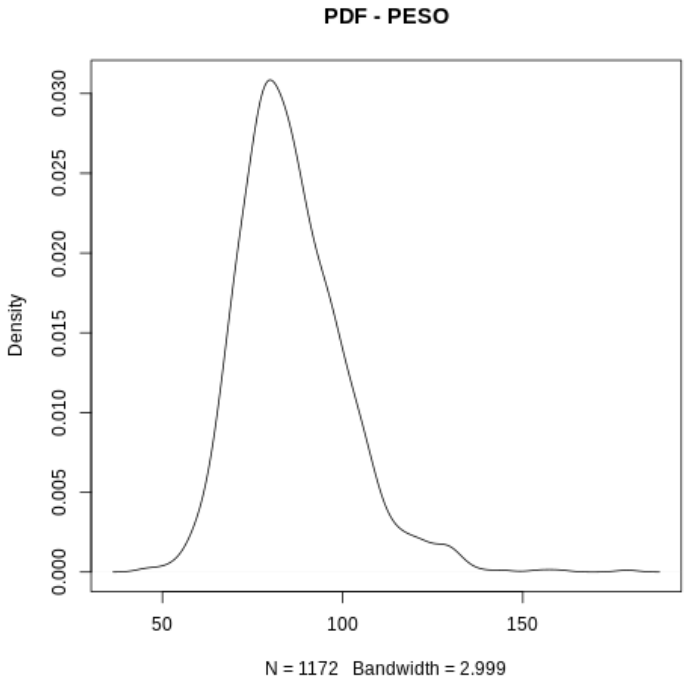


➢ *PESO:*



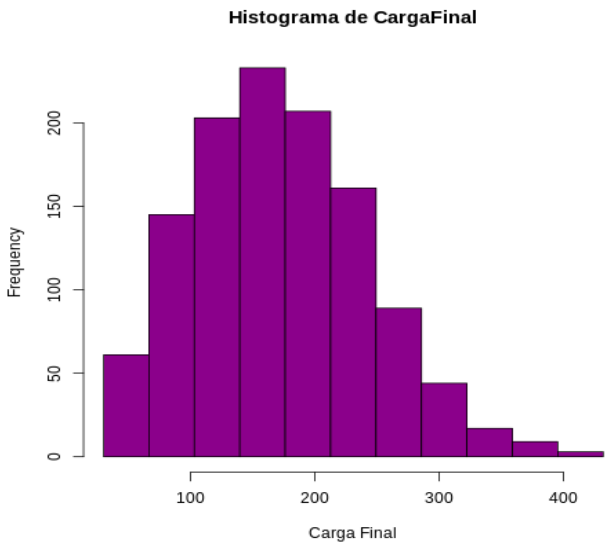
→ Diferentemente da variável idade, percebe-se uma grande concentração de dados. Neste caso, entre 70 e 95 kgs, o que faz sentido, tendo em vista que os dados foram coletados em cima de uma amostra de homens adultos, majoritariamente entre 40 e 70 anos de idade

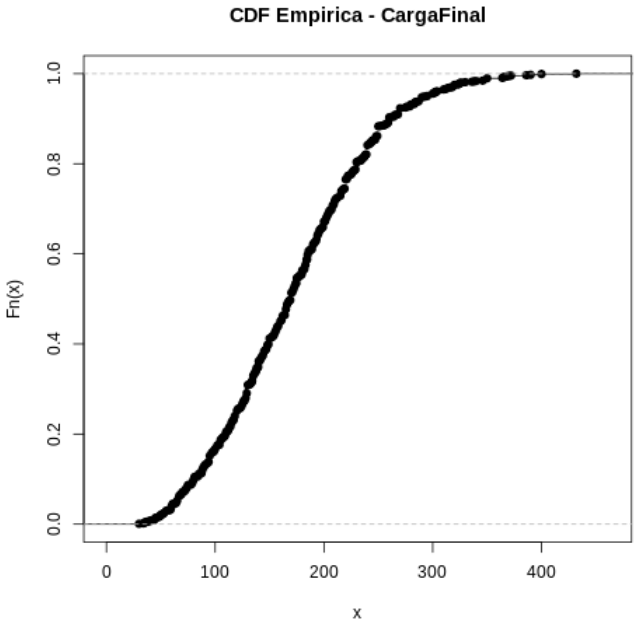


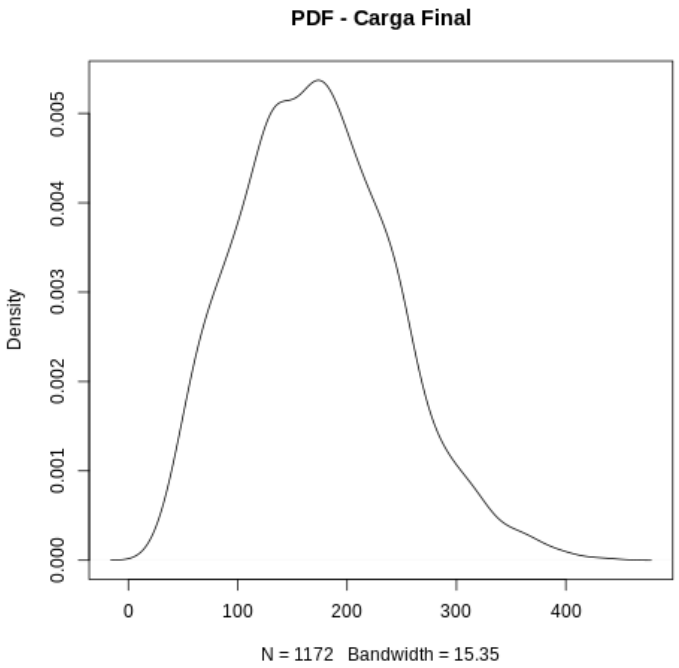


➢ *Carga Final:*

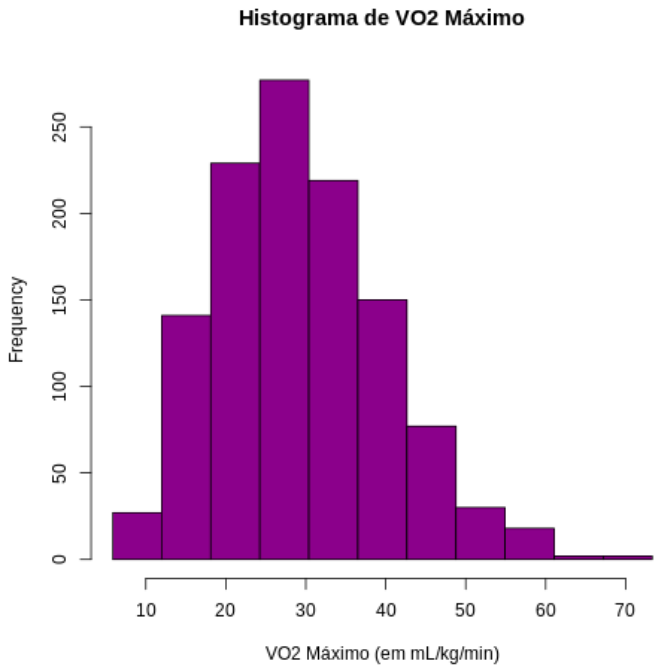
→ Na variável Carga Final, repara-se uma concentração maior na faixa dos 50 aos 250 watts e alguns indivíduos acima, provavelmente pessoas com melhor condicionamento físico.



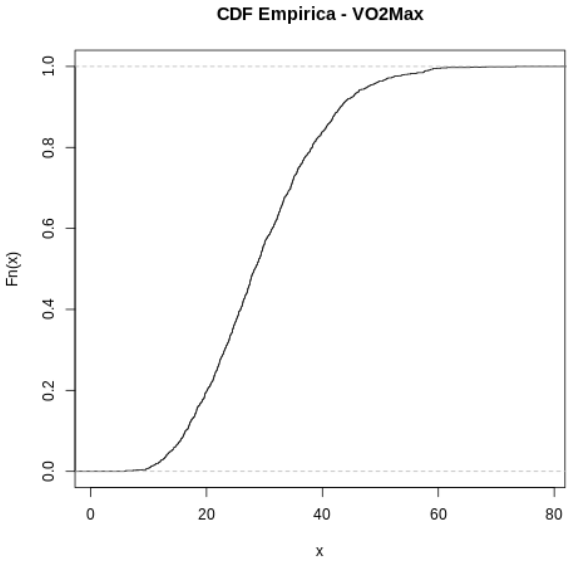


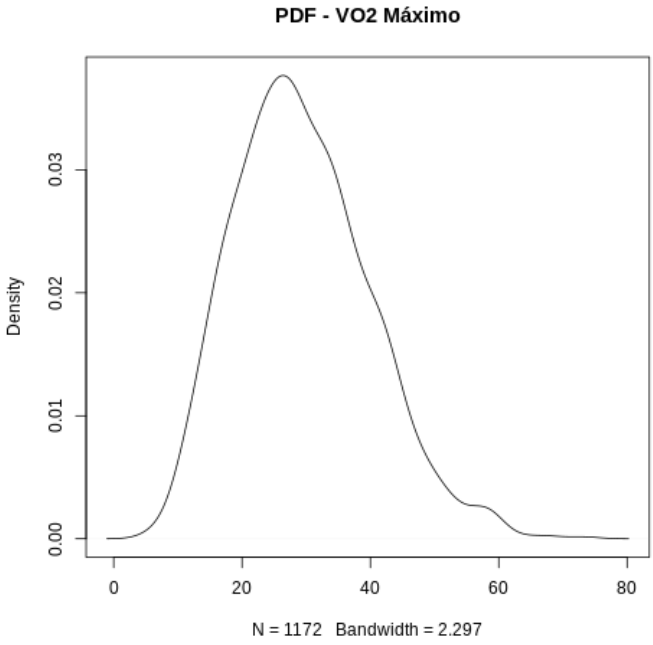


➢ *VO2 Máximo:*



→ O histograma de VO2 Máximo é muito semelhante ao histograma de Carga Final, o que parece sugerir uma conexão, correlação entre as duas variáveis. Percebe-se uma maior concentração na faixa de 15 a 45 ml/kg/min.





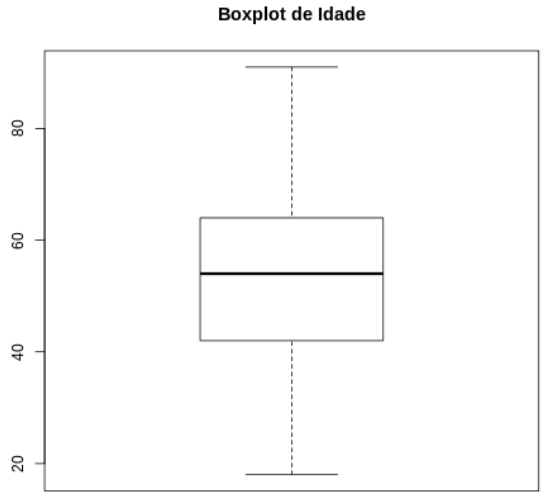
2) BoxPlot, Média e Variância das Variáveis

- *Idade:*

media_idade	53.2909556313993
variancia_idade	217.453274235434

→ Pelo Boxplot, fica claro que a variável idade não possui outliers, ou seja, não possui valores atípicos nas medidas.

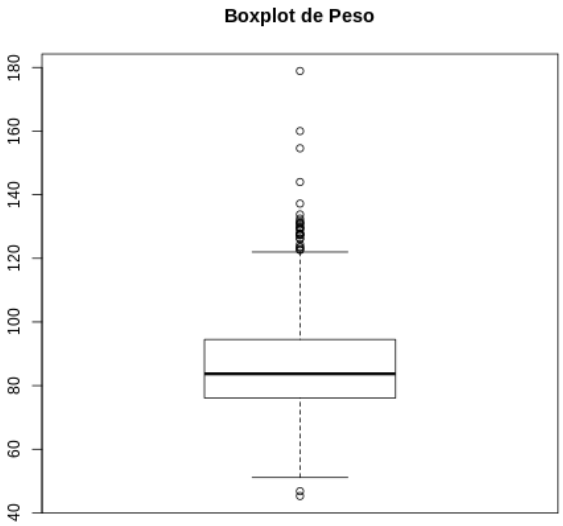
Também percebe-se em uma análise conjunta com a média e a variância que as amostras estão concentradas na média e bem distribuídas entre si.



- *Peso:*

media_Peso	85.9257764505119
variancia_Peso	219.013756954253

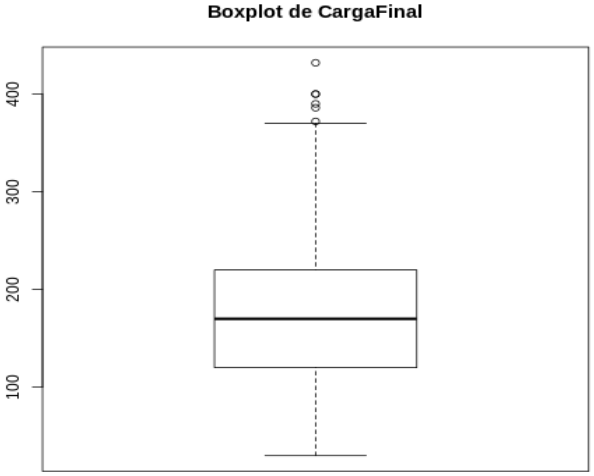
→ Pelo Boxplot, percebe-se inicialmente a presença de outliers tanto abaixo do lower quartile – 1.5*IQR como acima do upper quartile + 1.5*IQR. Ademais, na parte acima possuímos muitas medidas e, inclusive, medidas extremamente acima do upper quartile + 1.5*IQR. Tais medidas devem influenciar nos resultados finais obtidos. A média do peso está consideravelmente acima da média do peso nacional dos homens e a variância está num valor contundente com a amostragem.



- *Carga Final:*

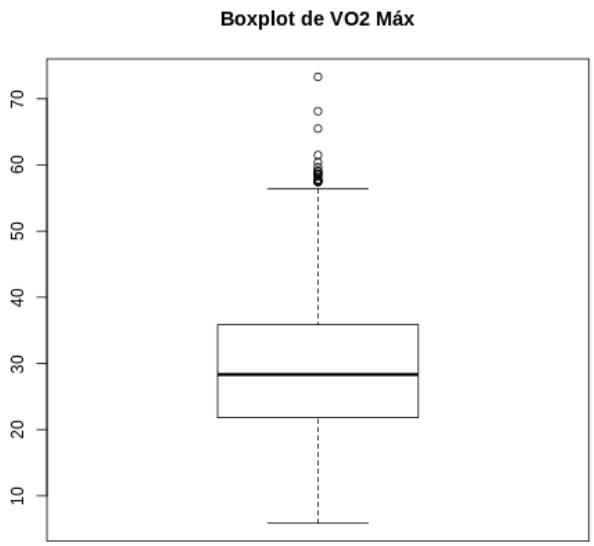
media_CargaFinal	172.271501706485
variancia_CargaFinal	4913.04598476259

→ A média e a variâncias indicam uma distribuição contudente dos dados. Já pelo Boxplot, percebemos a presença de outliers acima do upper quartile + 1.5*IQR, representando aqueles indivíduos com melhor porte físico, como mencionado anteriormente.



➢ *VO2 Máximo:*

media_VO2Max	29.3947279231532
variancia_VO2Max	110.192255325032



→ A média indica um valor menor do que o esperado para os homens (35), o que demonstra que essa amostragem possuiu muitos indivíduos com condições fisiológicas abaixo do comum. A variância indica uma maior concentração dos dados do redor da média, se comparada a variância das outras variáveis aleatórias analisadas. Já a BoxPlot revela a presença dos outliers acima do upper quartile + 1.5*IQR, representando, novamente, os indivíduos com melhor condicionamente físico.

3) Comparando com distribuições da literatura com parâmetros estimados via MLE

Utilizamos o MLE (Maximum Likelihood Estimator) para estimar os parâmetros das distribuições: Exponencial, Gaussiana, Lognormal e Weibull para cada variável aleatória observada. Depois, comparamos essas funções em um único gráfico.
Definição de Likelihood: $$\mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

Utilizando o MLE com log (achamos o log da função, visando transformar multiplicação de termos em adição de termos depois derivamos e igualamos a zero para encontrar o máximo dá função) temos:

Log Likelihood: $$\ell(\theta\,;\mathbf{y}) = \ln L_n(\theta\,;\mathbf{y}).$$

derivando para cada parâmetro: $$\frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell}{\partial \theta_k} = 0,$$

Utilizando estas definições, chegamos as seguintes fórmulas para os parâmetros das ditribuições:

→ Exponencial: $\lambda = 1/\varphi$, sendo φ = média das amostras

→ Gaussiana:

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu)^2$$

→

→ LogNormal:

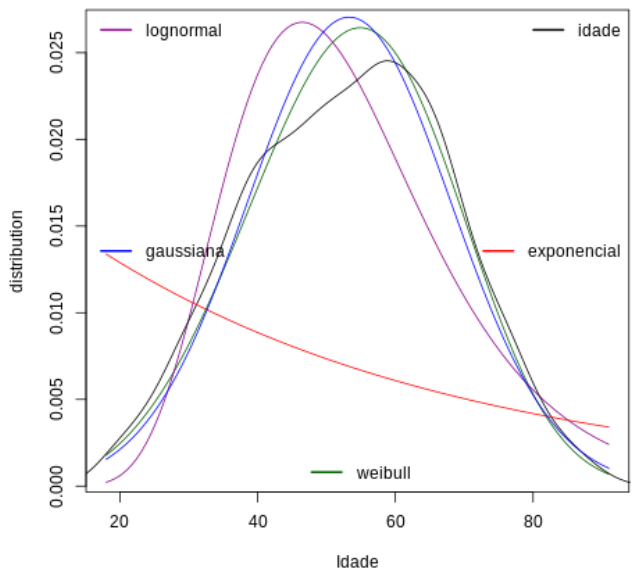
$$\widehat{\sigma}^2 = \frac{\sum_k (\ln x_k - \widehat{\mu})^2}{n} \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} \ln(X_i)}{n}$$

→ Weibull:
No caso da Weibull, utilizei a biblioteca EnvStats (função eweibull com method = ‘mle’) para estimar seus parâmetros, dado que estes não são dados por uma fórmula trivial como as demais distribuições apresentadas anteriormente. Porém, a fórmula encontrada via MLE é:

$$\widehat{\lambda}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

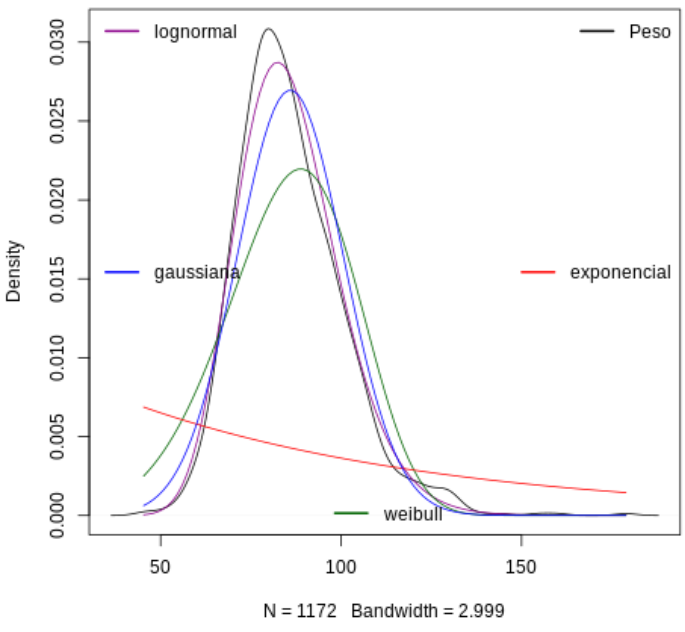
Os parâmetros podem ser verificados um a um, para cada variável, em suas respectivas variáveis globais no arquivo script.r no repositório do github, a partir da linha 118 até a 339.

➢ *Idade:*



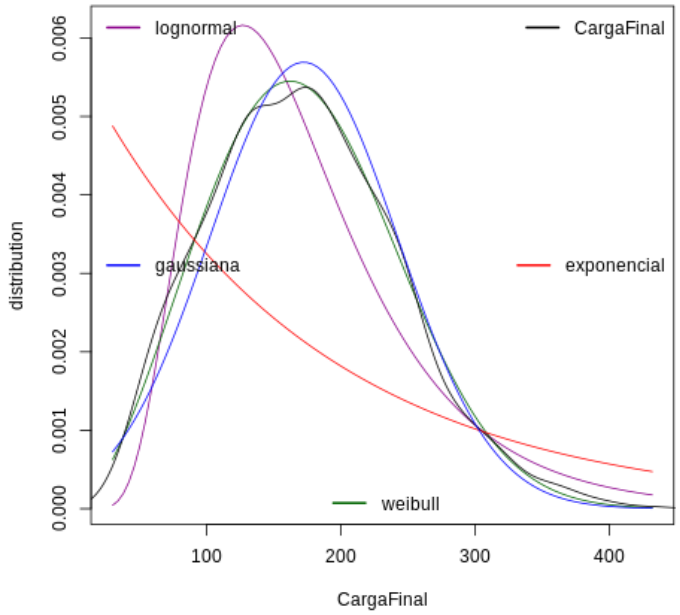
→ No Caso da variável Idade, percebemos que as distribuições que melhor podem lhe representar são as distribuições weibull ou gaussiana.

➢ *Peso:*



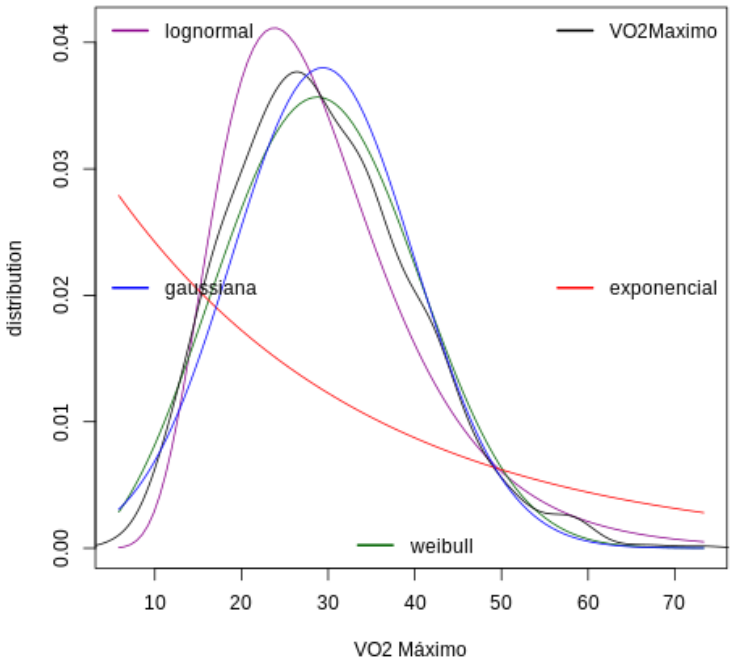
→ No caso da variável Peso, percebemos que a distribuição que melhor pode lhe representar é a distribuição lognormal, seguida pela distribuição gaussiana.

➢ *Carga Final:*



→ No caso da variável Carga Final, a distribuição Weibull parece ser a distribuição que a melhor representa, seguida pela Gaussiana.

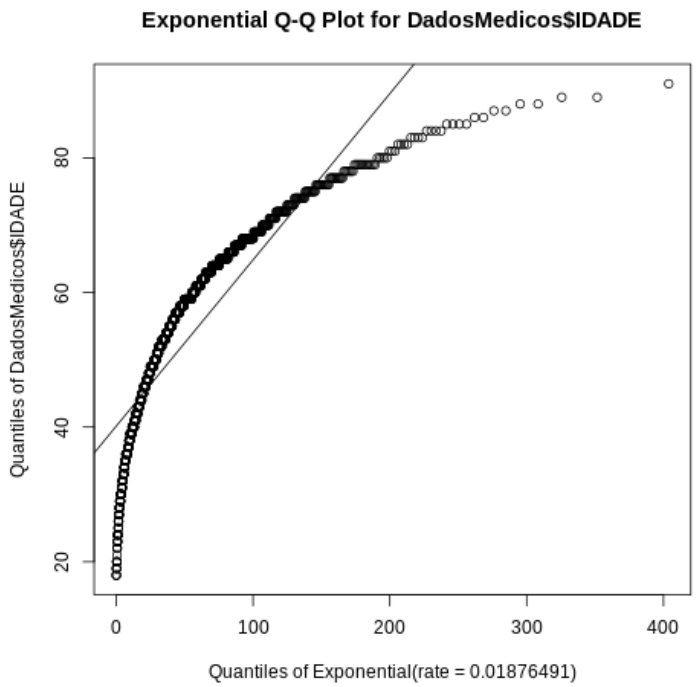
➢ *VO2 Máximo:*

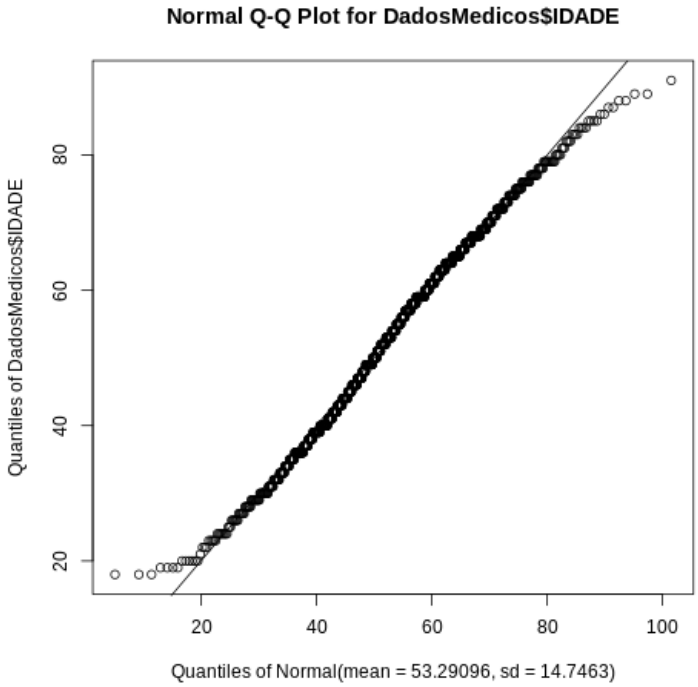


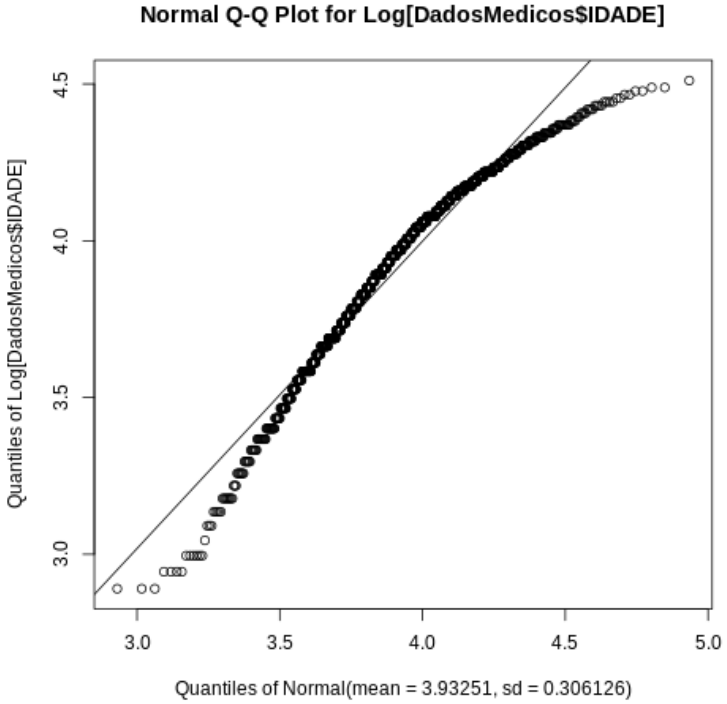
→ No caso da variável VO2 Máximo, tanto a distribuição gaussiana como a weibull parecem representar bem a variável.

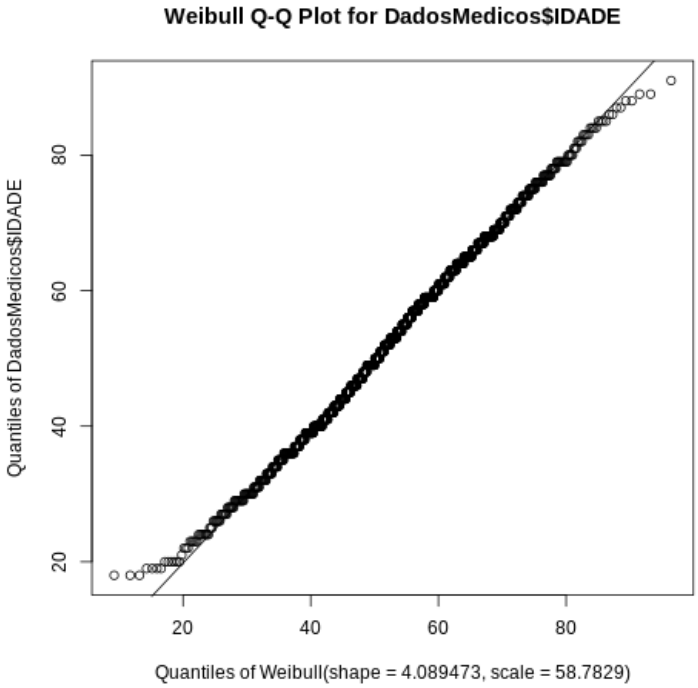
4) Comparando via QQPlot

- *Idade:*



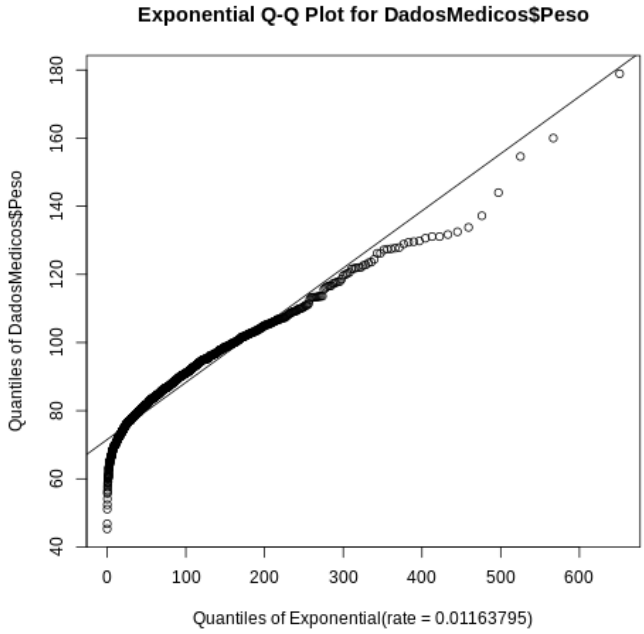


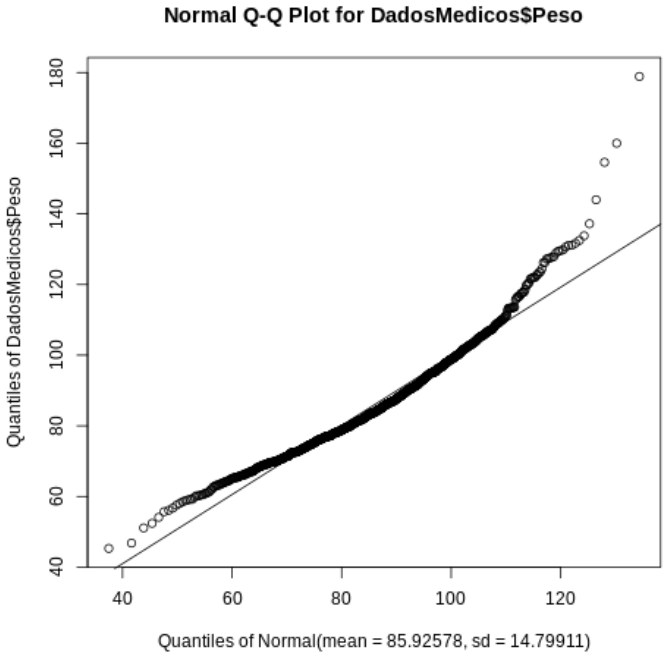


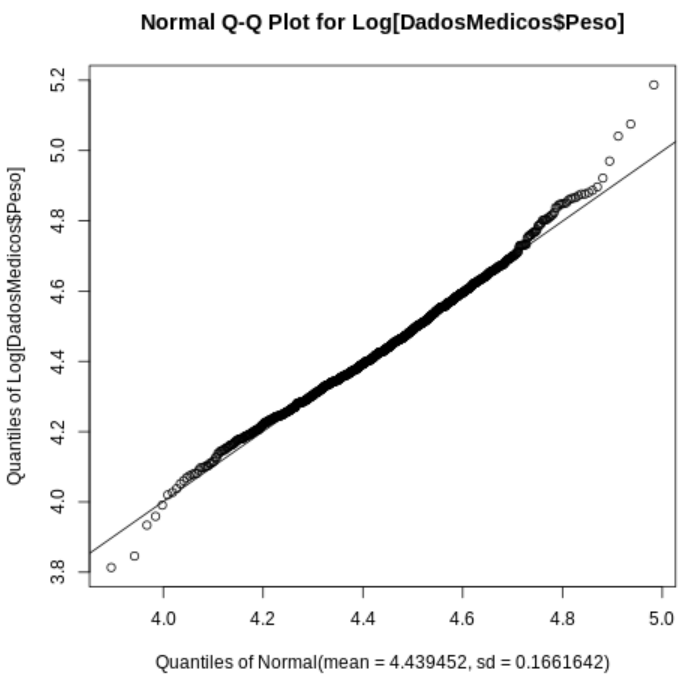


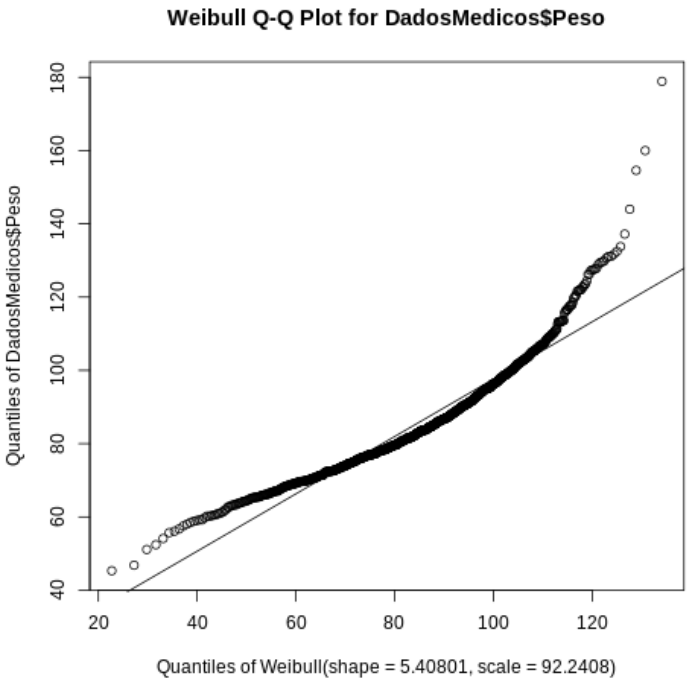
Os gráficos QQPlot confirmam nossas suspeitas anteriores a respeito da variável idade. Porém parece que a Weibull é levemente melhor, pois há menor diferença entre as curva das amostras e a reta de comparação.

- *Peso:*



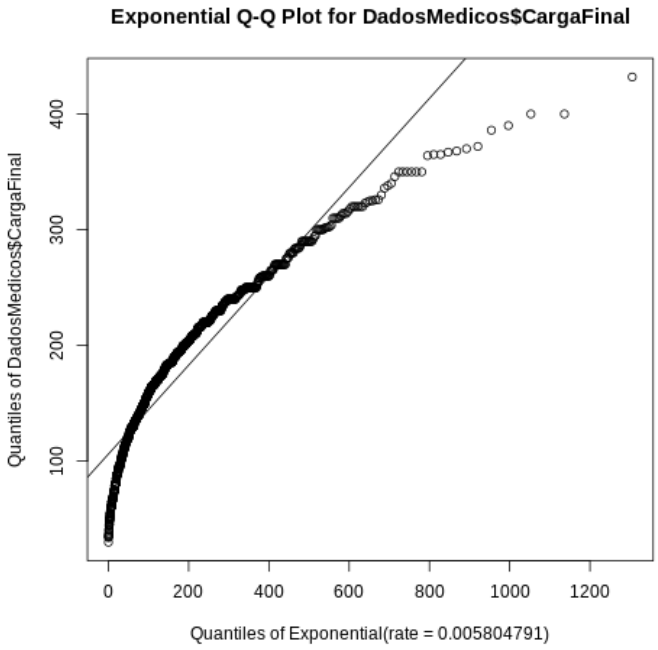


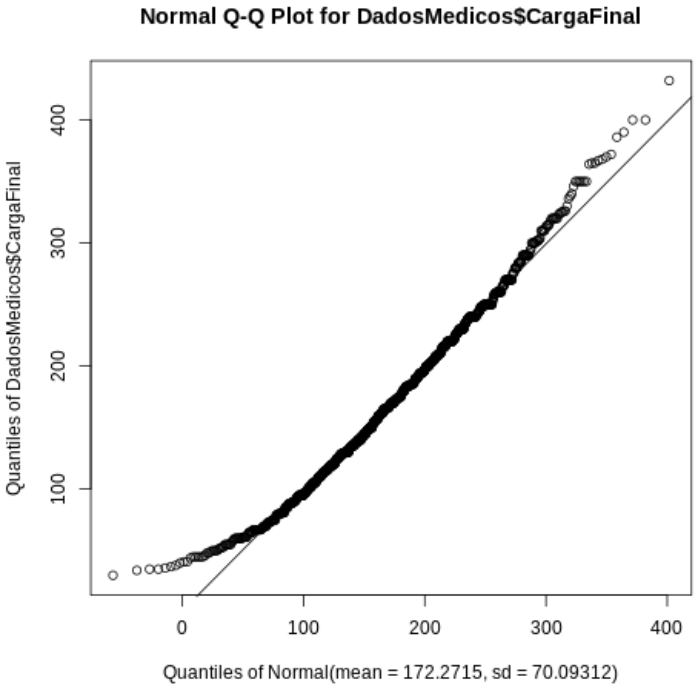


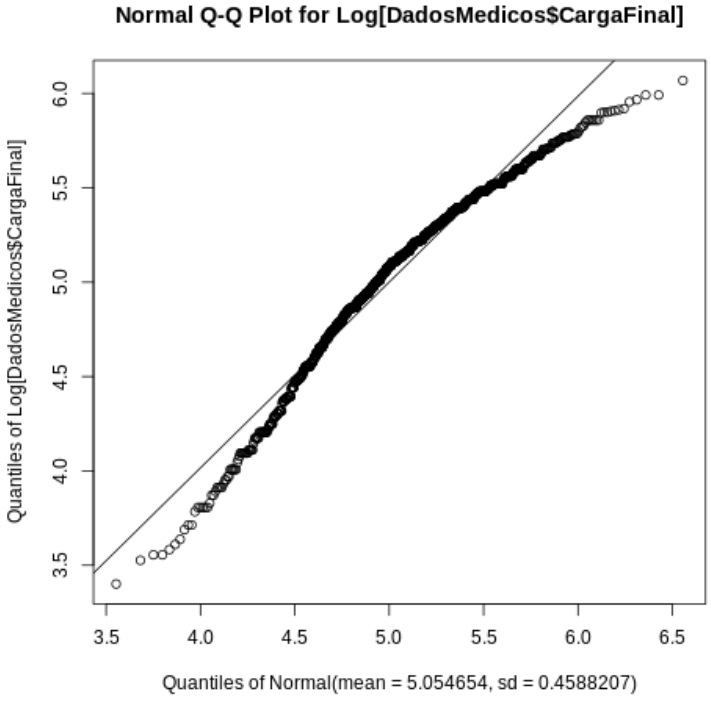


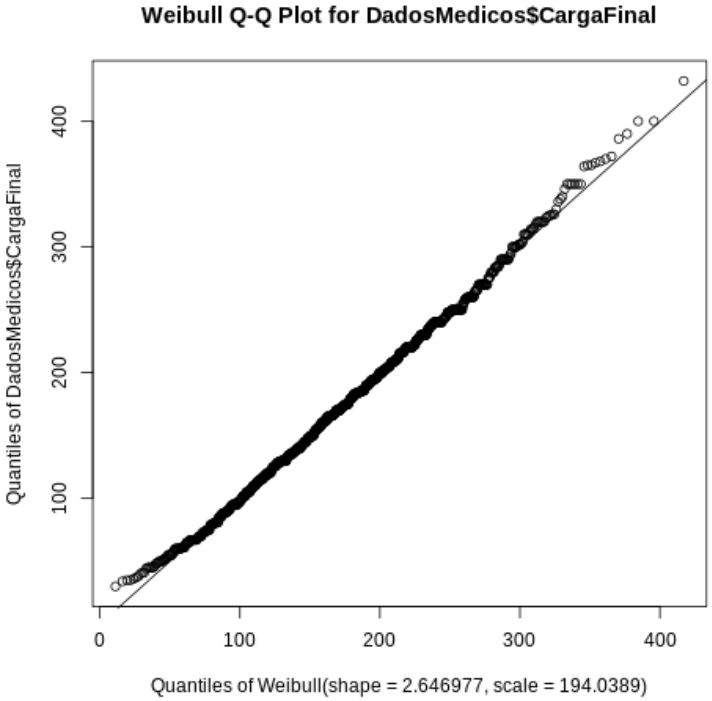
Os gráficos QQPlot, mais uma vez, confirmam nossas suspeitas anteriores . O QQPlot da Lognormal (Normal QQPlot for Log(DadosMedicos$Peso)) é o QQPlot que possui a melhor semelhança entre a reta e os dados dentre os 4 QQPlot criados. Há uma maior discrepância nos valores extremos somente(localização dos outliers).

- *Carga Final:*



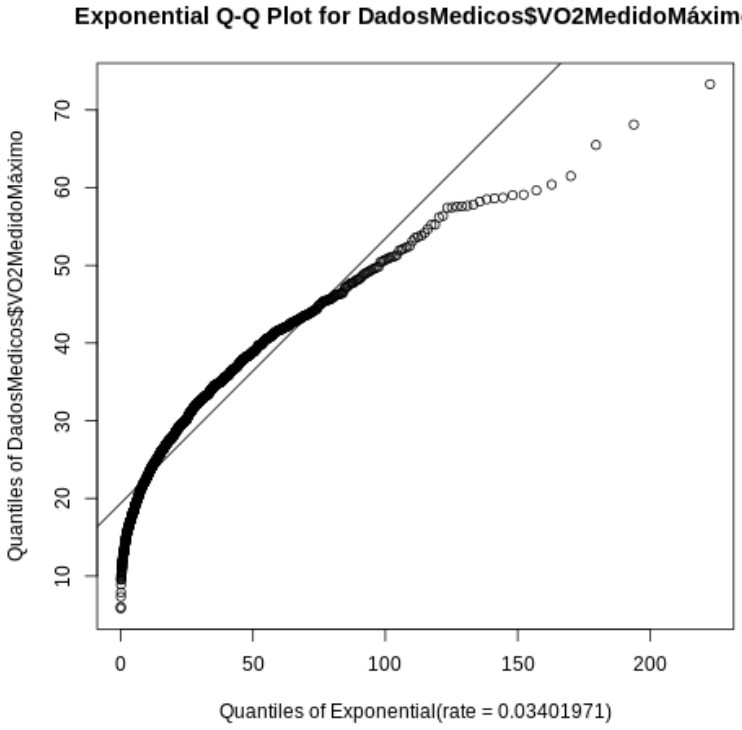


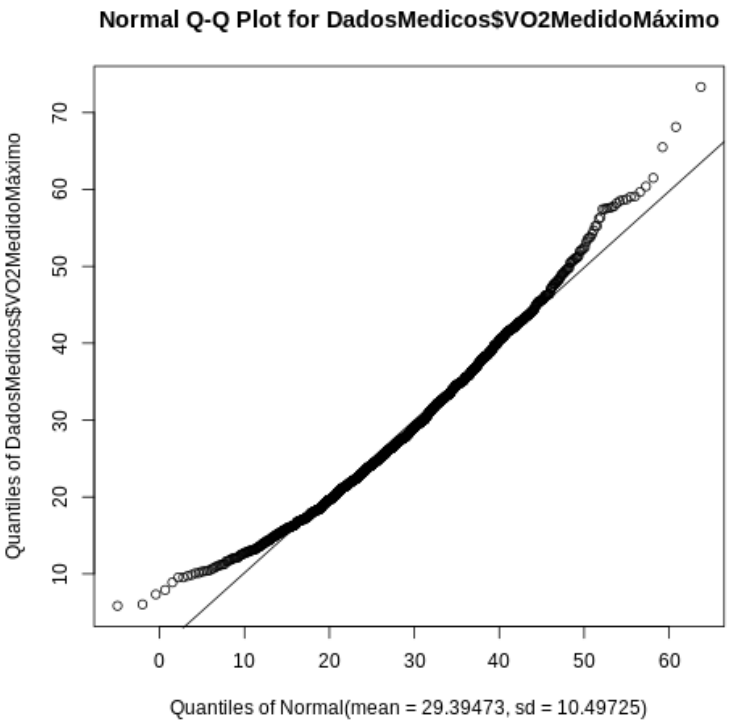


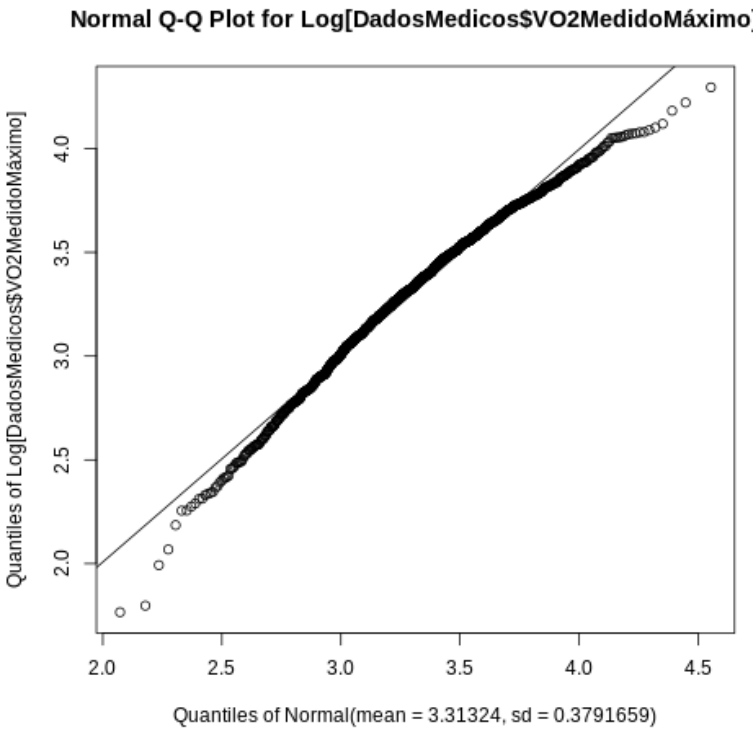


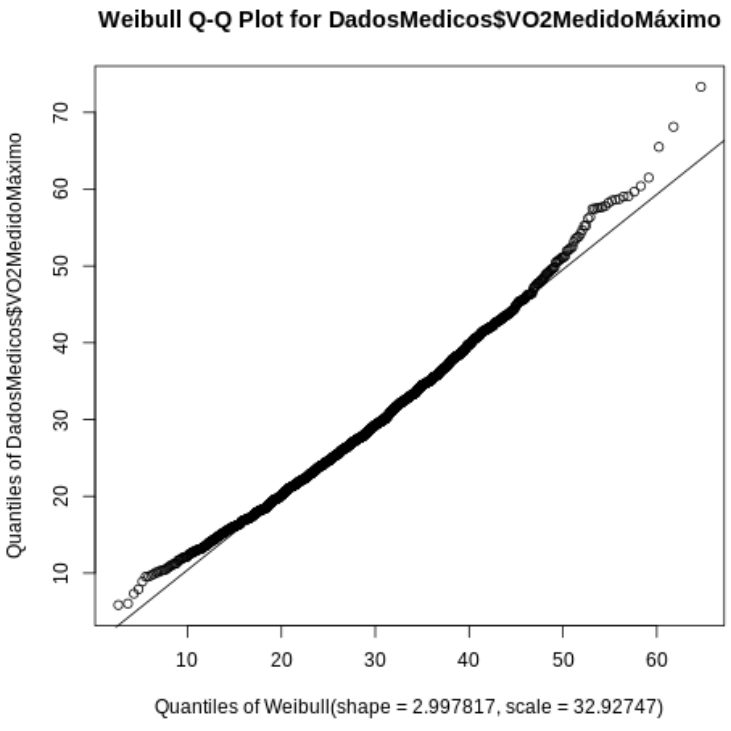
Novamente, nossas suspeitas anteriores foram confirmadas pelos QQPlots. A Weibull representa muito bem a variável aleatória Carga Final, o que se mostra evidente na grande semelhança entre a reta e os dados. A distribuição Gaussiana perde principalmente nos extremos das medidas (outliers).

- *VO2 Máximo:*









No caso da variável VO2 Máximo, a Weibull parece ter uma singela vantagem sobre a Gaussiana, tendo em vista a diferença na discrepância nos extremos inferiores dos dois gráficos QQPlot, implicando que a Weibull parece ser a melhor representação da variável neste caso.

5) Teste de Hipótese – Komogorov-Smirnov

Utilizei a função ks.test(), presente no R base para a realização do teste de Komogorv-Smirnov, que possui como estatísica:

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$

Realizamos um teste de hipótese bilateral com **α = 0.05**. A função ks.test() retorna a estatística Dn e o p-valor em questão. Os logs dos testes podem ser encontrados no arquivo hypoteshis_testing.log no repositório github.

- *Idade:*

"Teste de hipótese - IDADE X Exponencial"

Results of Hypothesis Test

Test Statistic: *D = 0.3727556*

P-value: *0*

Rejeitada, pois p-value < 0,05 , logo, a variável idade não é bem representada por distribuição exponencial.

"Teste de hipótese - IDADE X Gaussiana"

Results of Hypothesis Test

Test Statistic: *D = 0.04402252*

P-value: *0.02129072*

Rejeitada, pois p-value < 0,05 , logo, a variável idade não é bem representada por distribuição gaussiana. (diferentemente do que era esperado)

"Teste de hipótese - IDADE X Lognormal"

Results of Hypothesis Test

Test Statistic: *D = 0.08467285*

P-value: *1.005974e-07*

Rejeitada, pois p-value < 0,05 , logo, a variável idade não é bem representada por distribuição lognormal.

“Teste de hipótese - IDADE X Weibull"

Results of Hypothesis Test

Test Statistic: *D = 0.0330373*

P-value: *0.1547867*

Aceita, pois p-value > 0,05 , logo, a variável idade é bem representada por distribuição weibull. Dentro do esperado.

- *Peso:*

"Teste de hipótese - Peso X Exponencial"

Results of Hypothesis Test

Test Statistic: *D = 0.495441*

P-value: *0*

Rejeitada, pois p-value < 0,05 , logo, a variável peso não é bem representada por distribuição exponencial.

“Teste de hipótese - Peso X Gaussiana"

Results of Hypothesis Test

Test Statistic: *D = 0.06663052*

P-value: *6.047165e-05*

Rejeitada, pois p-value < 0,05 , logo, a variável peso não é bem representada por distribuição gaussiana.

“Teste de hipótese - Peso X Lognormal"

Results of Hypothesis Test

Test Statistic: *D = 0.03231203*

P-value: *0.1729427*

Aceita, pois p-value > 0,05 , logo, a variável peso é bem representada por distribuição lognormal.

"Teste de hipótese - Peso X Weibull"

Results of Hypothesis Test

Test Statistic: *D = 0.1032176*

P-value: *2.854561e-11*

Rejeitada, pois p-value < 0,05 , logo, a variável peso não é bem representada por distribuição weibull.

- *Carga Final:*

"Teste de hipótese - CargaFinal X Exponencial"

Results of Hypothesis Test

Test Statistic: *D = 0.2865163*

P-value: *0*

Rejeitada, pois p-value < 0,05 , logo, a variável Carga Final não é bem representada por distribuição exponencial.

"Teste de hipótese - CargaFinal X Gaussiana"

Results of Hypothesis Test

Test Statistic: *D = 0.03916339*

P-value: *0.05491148*

Aceita, pois p-value > 0,05 , logo, a variável Carga Final é bem representada por distribuição gaussiana. Dentro do esperado, mas surpreendente.

“Teste de hipótese - CargaFinal X Lognormal"

Results of Hypothesis Test

Test Statistic: *D = 0.08034078*

P-value: *5.373948e-07*

Rejeitada, pois p-value < 0,05 , logo, a variável Carga Final não é bem representada por distribuição lognormal.

"Teste de hipótese - CargaFinal X Weibull"

Results of Hypothesis Test

Test Statistic: *D = 0.02457064*

P-value: *0.4788417*

Rejeitada, pois p-value < 0,05 , logo, a variável Carga Final não é bem representada por distribuição weibull. Fora do esperado.

Obs: Os resultados do KSTest para a variável Carga Final são os mais interessantes, tendo em vista que eles contradizem o que pensamos anteriormente na análise dos QQPlots. Segundo este teste, a melhor representação deve ser via Gaussiana, não Weibull.

- *VO2 Máximo*:

"Teste de hipótese - VO2 Máximo X Exponencial"

Results of Hypothesis Test

Test Statistic: D = 0.3348897

P-value: 0

Rejeitada, pois p-value < 0,05 , logo, a variável VO2 Máx não é bem representada por distribuição exponencial.

"Teste de hipótese - VO2 Máximo X Gaussiana"

Results of Hypothesis Test

Test Statistic: D = 0.04450684

P-value: 0.01925495

Rejeitada, pois p-value < 0,05 , logo, a variável VO2 Máx não é bem representada por distribuição gaussiana.

"Teste de hipótese - VO2 Máximo X Lognormal"

Results of Hypothesis Test

Test Statistic: D = 0.04056923

P-value: 0.04222412
Rejeitada, pois p-value < 0,05 , logo, a variável VO2 Máx não é bem representada por distribuição lognormal.

"Teste de hipótese - VO2 Máximo X Weibull"

Results of Hypothesis Test

Test Statistic: D = 0.03674681

P-value: 0.08440764

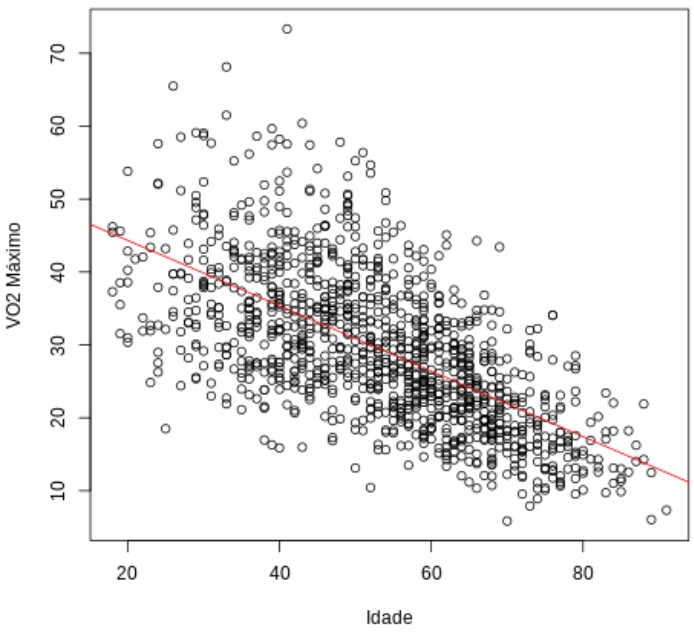
Aceita, pois p-value > 0,05 , logo, a variável VO2 Máx é bem representada por distribuição Weibull, como esperado via análises anteriores.

6) Modelo de Regressão e ScatterPlots

Os coeficientes de correlação foram obtidos via função cor() do base R, porém a formúla a ser utilizada para a sua obtenção é a seguinte: Onde x e y são as amostras e Xbarrada e Ybarrado suas médias. Os coeficientes lineares foram obtidos via regressão linear com a criação de um modelo linear para cada caso no R.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

- *Idade e VO2 Máx:*

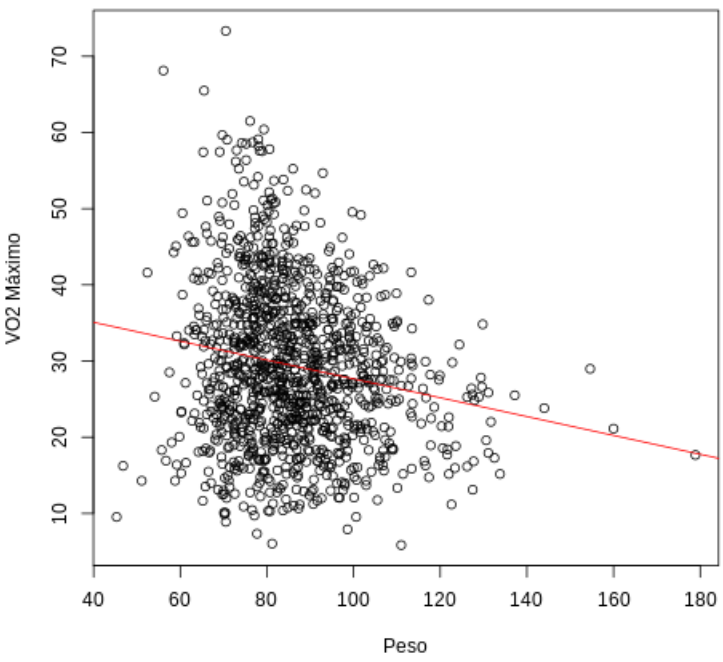


correlacao_idade_VO2Max -0.630072019250342

Coeficientes lineares: *a* = -0.448521
b = 53.296839

→ O Scatterplot demonstra a tendência de piora do condicionamento físico com o aumento da idade, o que parece algo normal num primeiro momento, tendo em vista que pessoas mais jovens tendem a ser mais ativas que pessoas mais velhas. Porém, o coeficiente de correlação não se aproxima tanto de -1, o que implica que seja apenas uma tendência, não uma correlação forte e absoluta. Neste caso, não faz um pouco de sentido a criação de um modelo de regressão, mas não tão em definitivo.

➢ *Peso e VO2 Máx:*



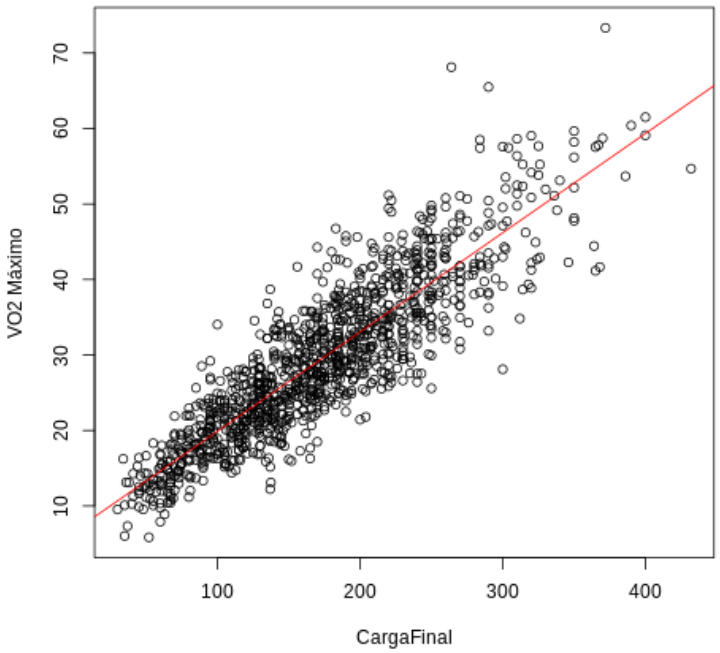
correlacao_Peso_VO2Max -0.174400618296308

Coeficientes lineares:

a = -0.1237052
b = 40.0241909

→ O ScatterPlot não fala muita coisa neste caso, tendo em vista que os dados de peso estão muito concentrados. Isto se traduz no índice de correlação, que demonstra uma correlação que se aproxima de 0 entre as variáveis, demonstrando que não há uma relação específica entre elas. Isto faz sentido, tendo em vista que alguém com muito peso pode tanto ser alguém muito atlético como alguém que possui um condicionamento físico ruim, tendendo a estar com grande acúmulo de gordura. Fora a altura do indivíduo, que não é levada em conta e tem grande impacto no peso. Neste caso, não sentido sentido algum na criação de um modelo de regressão.

➢ *Carga Final e VO2 Máx:*



correlacao_CargaFinal_VO2Max 0.878325609405962

Coeficientes Lineares:

a = 0.1315393
b = 6.7342478

→ O ScatterPlot demonstra uma grande correlação entre as duas variáveis. O que faz total sentido, tendo em vista que indivíduos que aguentam uma maior carga final tendem a possuir melhor condicionamente físico e, consequentementes, melhores taxas de VO2 Máximo. Isto revela-se de forma muito clara no coeficiente de correlação, que é muito próximo de 1. Neste caso, é muito interessante a criação de um modelo de regressão, tendo em vista a correlação aparente entre as variáveis.

7) Inferência Bayesiana entre Carga Final e VO2 Máximo

Escolhemos a Carga Final pois é a variável com maior correlação com a variável VO2 Máximo.
Os cálculos desta parte do trabalho foram realizados via criação de matrizes na linguagem R. O log dessas matrizes pode ser encontrado no arquivo bayesian_inference.log no repositório github.
Foram criadas 2 tabelas, uma com VO2Máx < 35 e outra com VO2Máx ⩾ 35. Também dividimos os dados em 5 intervalos de hipóteses, para facilitação dos cálculos.

Tabela Bayesiana para VO2Máx < 35:

Hypothesis – Carga Final	Prior	Likelihood	Bayes Num	Posterior
H(30.0 - 110.4)	0.20392491	1.00000000	0.203924915	0.282505910
H(110.4 - 190.8)	0.41723549	0.91820041	0.383105802	0.530732861
H(190.8 - 271.2)	0.30204778	0.43502825	0.131399317	0.182033097
H(271.2 - 351.6)	0.06655290	0.05128205	0.003412969	0.004728132
H(351.6 432.0)	0.01023891	0.00000000	0.00000000	0.00000000

Total da Prior = 1, Total da Posterior = 1, Total dos Bayes Num = 0.721843
Probabilidade VO2 Máximo < 35.0 = 0.721843 (Soma dos Bayes Numerator)

A partir desta tabela, observamos que a maioria(em torno de 53%) dos pacientes com VO2Máx < 35.0 estão na faixa de 110.4 até 190.8 da Carga Final.

Tabela Bayesiana para VO2 Máx ⩾ 35:

Hypothesis – Carga Final	Prior	Likelihood	Bayes Num	Posterior
H(30.0 - 110.4)	0.20392491	0.00000000	0.00000000	0.00000000
H(110.4 - 190.8)	0.41723549	0.08179959	0.03412969	0.12269939
H(190.8 - 271.2)	0.30204778	0.56497175	0.17064846	0.61349693
H(271.2 - 351.6)	0.06655290	0.94871795	0.06313993	0.22699387
H(351.6 432.0)	0.01023891	1.00000000	0.01023891	0.03680982

Total da Prior = 1, Total da Posterior = 1, Total dos Bayes Num = 0.278157

Probabilidade VO2 Máximo ⩾ 35.0 = 0.278157 (Soma dos Bayes Numerator)
A partir desta tabela, observamos que a maioria(em torno de 61%) dos pacientes com VO2Máx ⩾ 35.0 estão na faixa de 190.8 até 271.2 da Carga Final.

Tabela de Bayes Predict (Prever VO2Máx ⩾ 35.0 após ter VO2Máx < 35.0)

Hypothesis – Carga Final	Prior 1	Likelihood 1	Bayes Num 1	Posterior 1	Likelihood 2	Predictions
H(30.0 - 110.4)	0.20392491	1.00000000	0.203924915	0.282505910	0.00000000	0.00000000
H(110.4 - 190.8)	0.41723549	0.91820041	0.383105802	0.530732861	0.08179959	0.043413731
H(190.8 - 271.2)	0.30204778	0.43502825	0.131399317	0.182033097	0.56497175	0.102843558
H(271.2 - 351.6)	0.06655290	0.05128205	0.003412969	0.004728132	0.94871795	0.004485664
H(351.6 432.0)	0.01023891	0.00000000	0.000000000	0.000000000	1.00000000	0.00000000

Probabilidade de melhorar, ou seja, P[VO2Máx ⩾ 35 | VO2Máx < 35] = 0.150743 (Soma das Predictions para cada hipótese)

Portanto, percebemos que após ter tido um valor abaixo de 35.0, é muito improvável que o indivíduo melhore sua marca, tendo apenas aproximadamente 15% de chance de melhora.

Conclusão

O trabalho trouxe uma experiência prática sobre os conceitos que foram ensinados ao longo do curso, principalmente em sua parte final. Os conhecimentos aprendidos foram muito reforçados em minha cabeça durante o projeto.
Realizamos uma análise básica das variáveis inicialmente e depois partimos para uma análise mais comparativa em relação a literatura, o que nos permitiu definir a melhor distribuição para cada amostragem obtida, o que em um caso entrou em conflito com o resultado do teste de hipótese, demonstrando que análises de gráficos via olho não são 100% corretas, assim como as de somente teste de hipótese também não. O conjunto em si que deve ser analisado.
Me pareceu muito relevante esta análise no âmbito médico, esta análise deve ter muito valor para um mapeamento de grupos com maiores chances de desenvolver certos tipos de doenças relacionadas com as variáveis estudadas no projeto.