Estatística e Modelos Probabilísticos - COE241

Trabalho Final

Professora: Rosa Maria Meri Leão
João Pedro Costa de Lacerda
DRE:116076670
Código: github.com/jpedrodelacerda/coe241-probest

Resumo

O Projeto final do curso tem como objetivo analisar os dados fornecidos pelo professor Cláudio Gil Soares de Araujo. O dataset em questão possui as seguintes informações: idade, peso, carga máxima em um determinado teste e a capacidade de VO2 Máxima de cada um dos 1172 indivíduos.

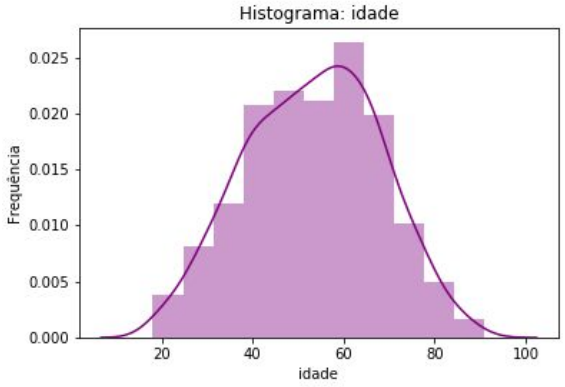
O código e os gráficos do relatório estão disponíveis em https://github.com/jpedrodelacerda/coe241-probest. As ferramentas utilizadas foram, primariamente, *Jupyter Notebook* com o *kernel Python 3* e bibliotecas como *pandas, numpy, scipy, seaborn e matplotlib.*

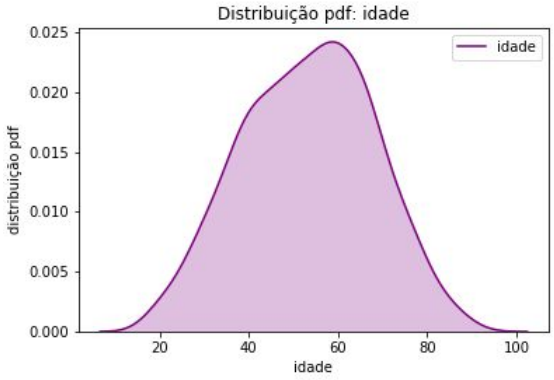
Análise inicial

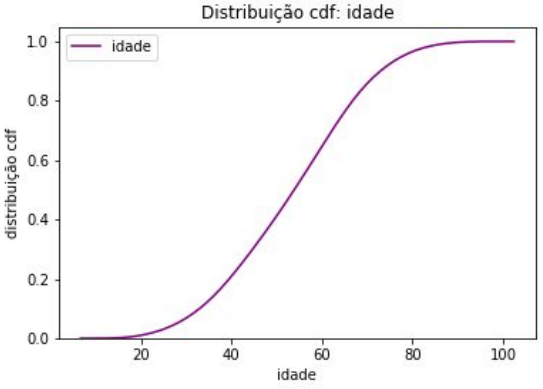
Num primeiro momento, foi necessário tratar os dados no contidos no *dataset* (*dadosMedicos.csv*). Então, foi possível utilizar a biblioteca *python-pandas* para fazer a leitura do *dataset*. Em seguida, foram extraídas do *dataset* uma série de estatísticas básicas, como os histogramas, médias, desvios padrões, distribuição pdf e cdf empíricas…

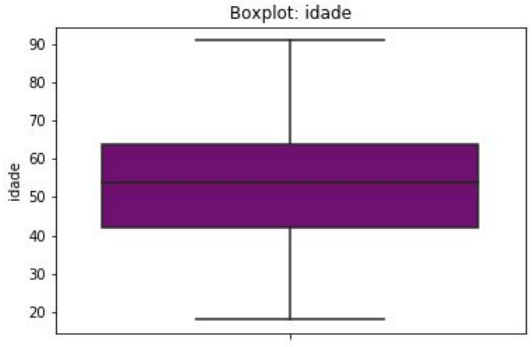
Para a montagem do histograma, o número de bins foi calculado utilizando a fórmula $m = 1 + 3.3\ log10(n)$, onde *m* representa o número de intervalos para o histograma e *n* o número de amostras no conjunto de dados. Como os valores de *n* se aproximam em todas as colunas do conjunto, o valor de *m* assumido foi 11.

- Idade







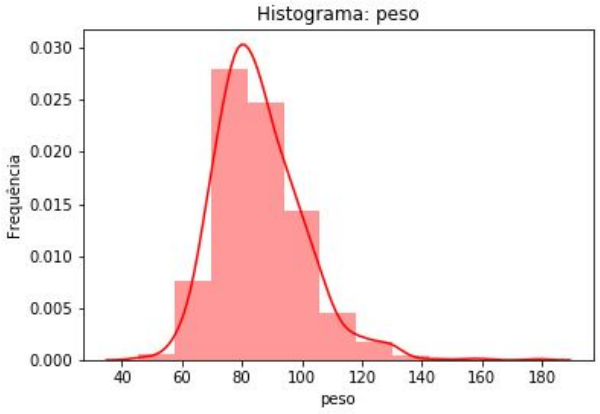


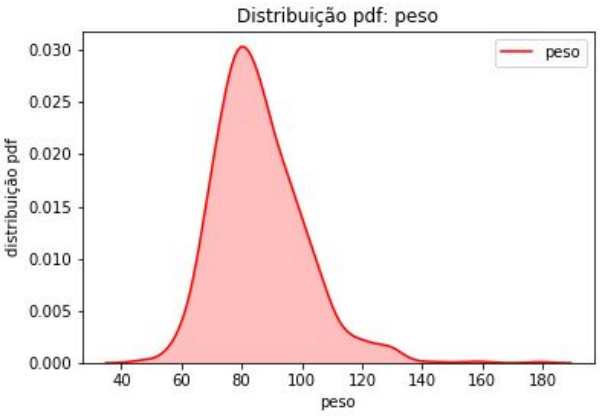
Média:	53.29095563139932
Desvio Padrão:	14.746296966880656
Variância:	217.45327423543367

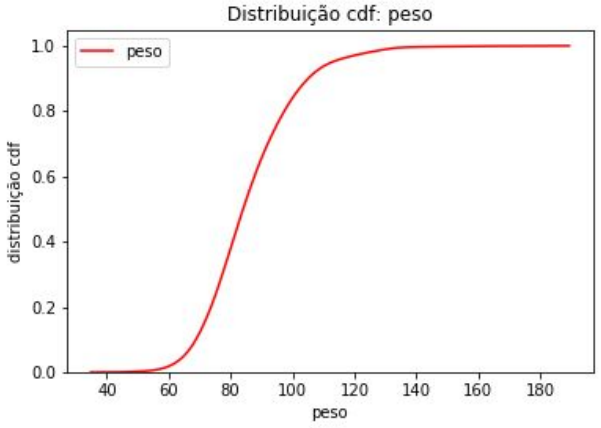
Com esses dados é possível perceber que as amostras possuem uma predominância na faixa dos 40 e 70 anos, porém, são bem distribuídas.

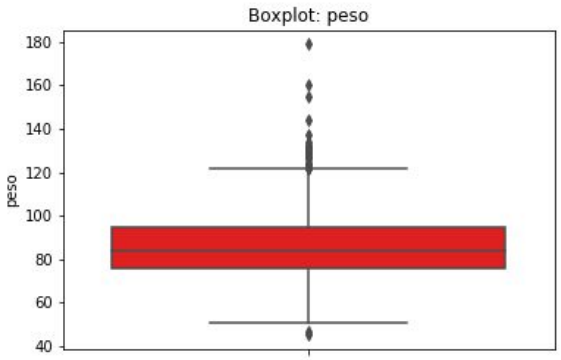
Podemos notar que não há presença de *outliers* pelo boxplot.

- Peso









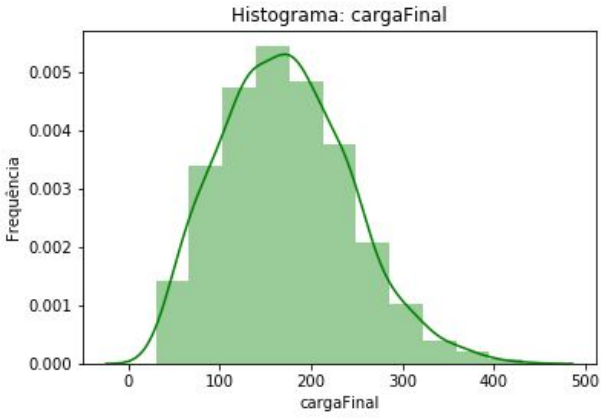
Média:	85.92577645051195
Desvio Padrão:	14.799113384059629
Variância:	219.0137569542528

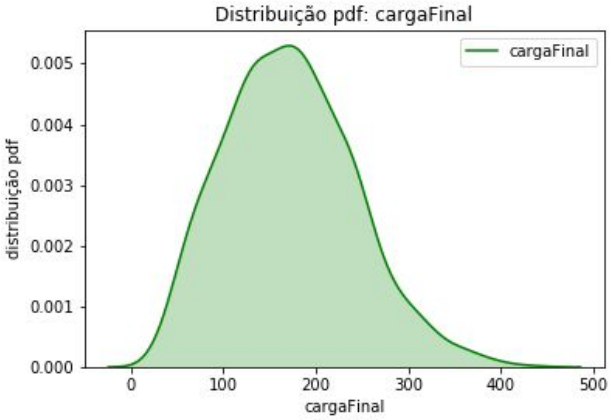
Análisando o histograma conseguimos ver que há uma concentração das amostras na faixa entre 70 e 100 kg.

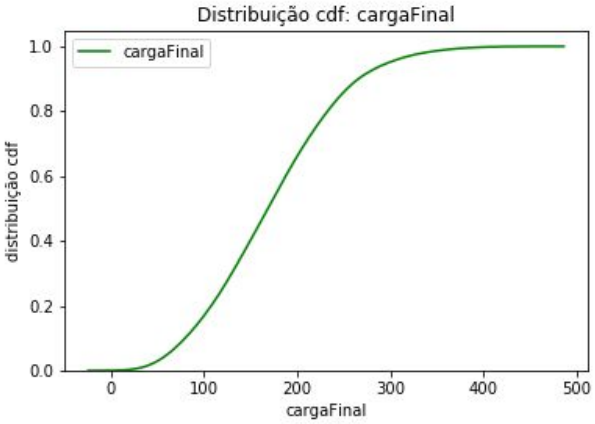
Pelo boxplot e pelo histograma também conseguimos perceber um número considerável de *outliers*, superiores em quantidade bem maior que os inferiores e isso pode afetar a análise dos dados e comparações com as distribuições que serão vistas logo mais.

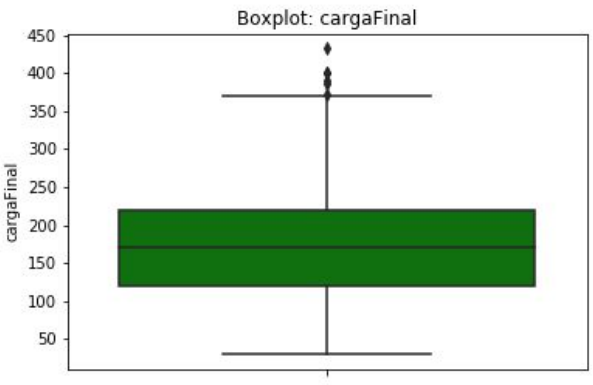
É necessário notar que essa concentração na faixa entre 70 e 100kg é compatível com o que vimos na análise da variável idade, visto que esta é a faixa comum para homens adultos.

- Carga Final









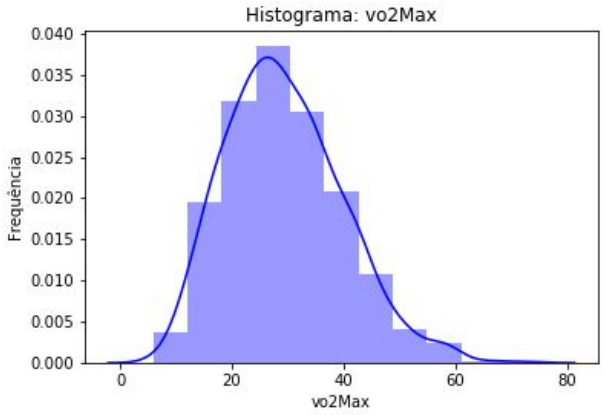
Média:	172.27150170648466
Desvio Padrão:	70.093123662472
Variância:	4913.0459847625925

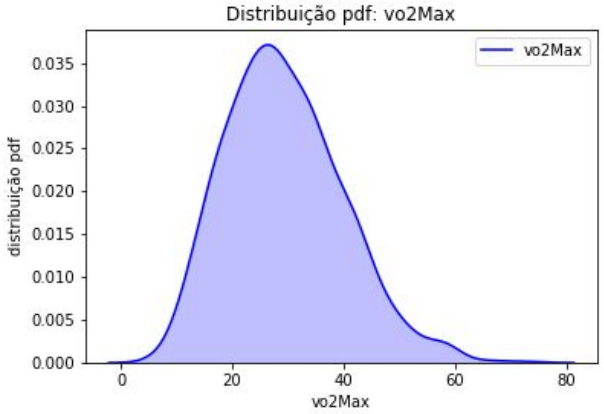
Analisando o histograma, pdf e boxplot podemos ver que há *outliers* superiores, porém é um grupo pequeno em relação aos *outliers* do peso, então não devem ter tanto impacto na análise.

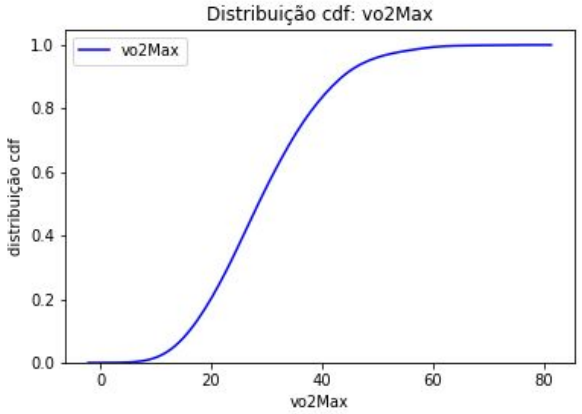
Se percebe também uma faixa entre 100 e 250 onde as amostras se concentram. Podemos perceber que a distribuição da carga final e do VO2 Máximo são semelhantes, o que pode ser um indicativo de dependência das variáveis.

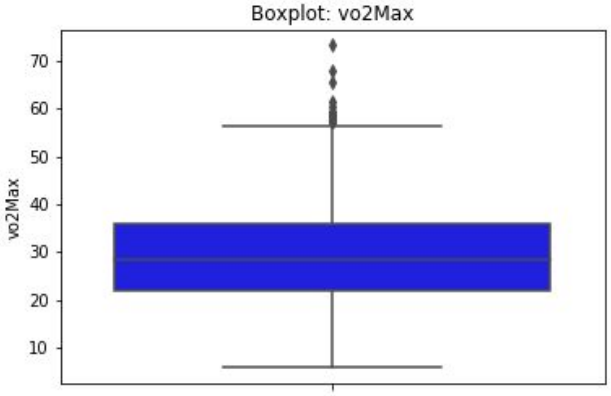
De modo geral, as amostras estão bem distribuídas.

- VO2 Máximo









Média:	29.39472792315316
Desvio Padrão:	10.49724989342601
Variância:	110.1922553250324

É possível notar também que os pacientes se concentram na faixa entre 15 e 40 mL/(kg * min).

Como citado anteriormente, a distribuição das amostras é bem parecida com a da carga final. Isso também se mostra na presença dos *outliers* superiores.

Parametrizando as distribuições

Essa etapa do trabalho consiste em comparar as distribuições empíricas com as distribuições da literatura. Para tal fim, foi utilizado o método da máxima verossimilhança (*MLE - Maximum likelihood estimation*) para fazer a estimação dos parâmetros de cada distribuição.

As distribuições tratadas aqui são: Exponencial, Gaussiana, Lognormal e Weibull.

- Exponencial
  - pdf:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

  - *likelihood*

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^n \exp(-\lambda n \bar{x}),$$

    where:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - então, λ é estimado por

$$\widehat{\lambda} = \frac{1}{\bar{x}} = \frac{n}{\sum_i x_i}$$

- Gaussiana
  - pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  - *likelihood*

$$\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

  - assim, μ e σ são estimados por:

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- Lognormal
  - pdf

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right], & \text{se } x > 0 \\ 0 \text{ caso contrário} \end{cases}$$

  - *likelihood*

$$\ell(\mu, \sigma \mid x_1, x_2, \ldots, x_n) = -\sum_i \ln x_i + \ell_N(\mu, \sigma \mid \ln x_1, \ln x_2, \ldots, \ln x_n).$$

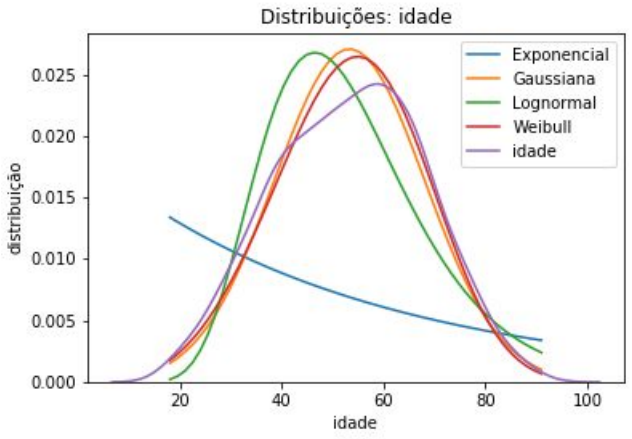
  - assim, μ e σ são estimados por:

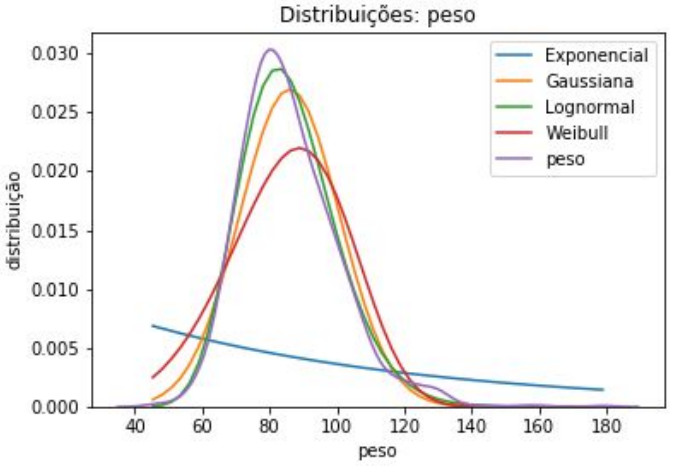
$$\widehat{\mu} = \frac{\sum_k \ln x_k}{n}, \qquad \widehat{\sigma}^2 = \frac{\sum_k (\ln x_k - \widehat{\mu})^2}{n}.$$

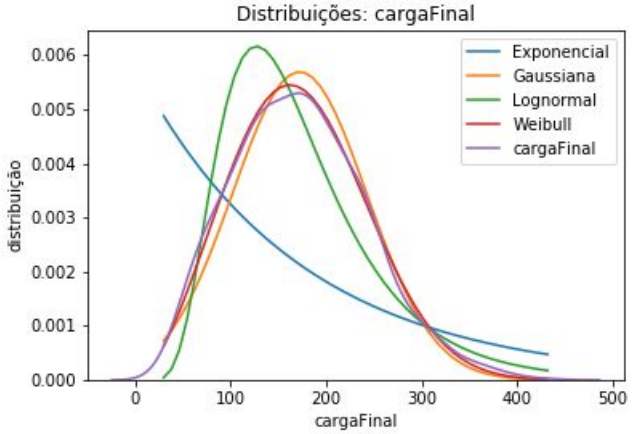
- Weibull

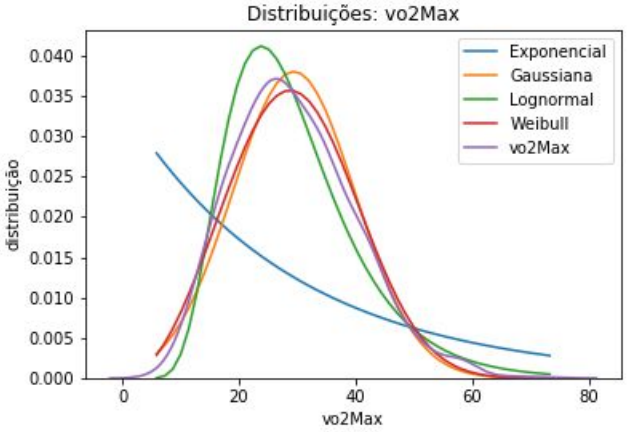
Foi utilizado a biblioteca *scipy.stats* com o método *weibull_min.fit()*.

Comparando as distribuições









Analisando os gráficos, percebemos que a distribuição exponencial não consegue representar nenhuma das variáveis.

Por outro lado, a Gaussiana, Lognormal e Weibull têm forma semelhantes em todas as variáveis:

- Idade:

  Percebemos que a Lognormal tem a pior aderência quando comparada com Weibull e Gaussiana. Destas duas restantes, é necessário uma análise mais profunda para determinar qual se adequa melhor

λ (Exponencial)	0.018764910258257682
μ (Gaussiana)	53.29095563139932
σ^2 (Gaussiana)	217.45327423543367
μ (Lognormal)	3.932509819486875
σ^2 (Lognormal)	0.0936331955793438

Constante (Weibull)	4.089481828645864
Loc (Weibull)	0
Scale (Weibull)	58.78289005707875

- Peso:

  A distribuição que melhor se assemelha é a Lognormal.

λ (Exponencial)	0.011637951279683105
μ (Gaussiana)	85.92577645051195
σ^2 (Gaussiana)	219.0137569542528
μ (Lognormal)	4.439451920143028
σ^2 (Lognormal)	0.027586997105752877
Constante (Weibull)	5.408013188534343
Loc (Weibull)	0
Scale (Weibull)	92.24080850317551

- Carga Final:

  As distribuições empíricas se assemelham bastante com exceção da Lognormal, assim como para a idade. Porém, a que melhor se assemelha com distribuição das amostras é a Weibull.

λ (Exponencial)	0.005804790636258545
μ (Gaussiana)	172.27150170648466
σ^2 (Gaussiana)	4913.0459847625925
μ (Lognormal)	5.0546544058509895
σ^2 (Lognormal)	0.2103368574854832
Constante (Weibull)	2.6469810001574725
Loc (Weibull)	0
Scale (Weibull)	194.0388415799269

- VO2 Máximo:

  Assim como para a idade, não é possível definir por enquanto qual é

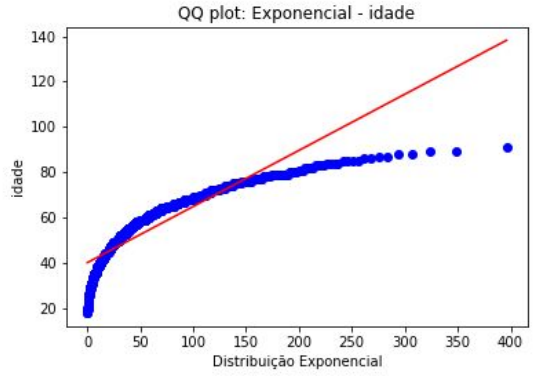
a melhor entre Gaussiana e Weibull, sendo necessária uma análise mais profunda.

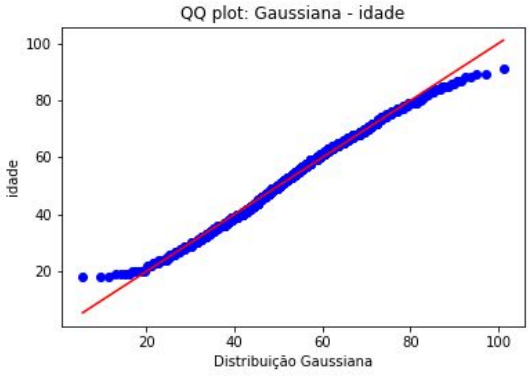
λ (Exponencial)	0.03401970593551017
μ (Gaussiana)	29.39472792315316
σ^2 (Gaussiana)	110.1922553250324
μ (Lognormal)	3.3132400746591215
σ^2 (Lognormal)	0.14364411960908474
Constante (Weibull)	2.9978221690896216
Loc (Weibull)	0
Scale (Weibull)	32.9274599599628

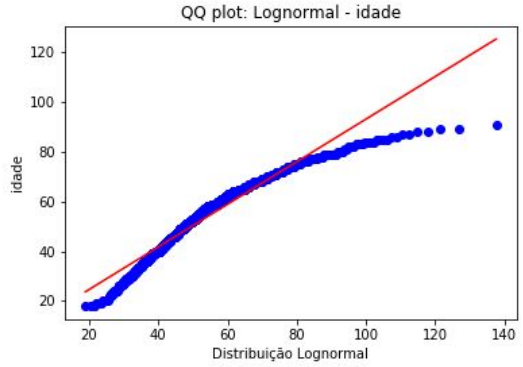
Gráficos QQplot

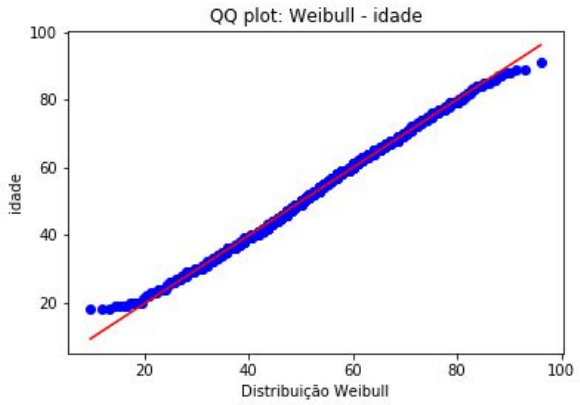
Os gráficos qqplot são uma ferramenta para determinar se dois conjuntos de dados possuem uma distribuição comum. A disposição dos pontos representa a proximidade entre os valores, quanto mais próximos à reta de 45º em relação aos eixos, maior a semelhança.

- Idade



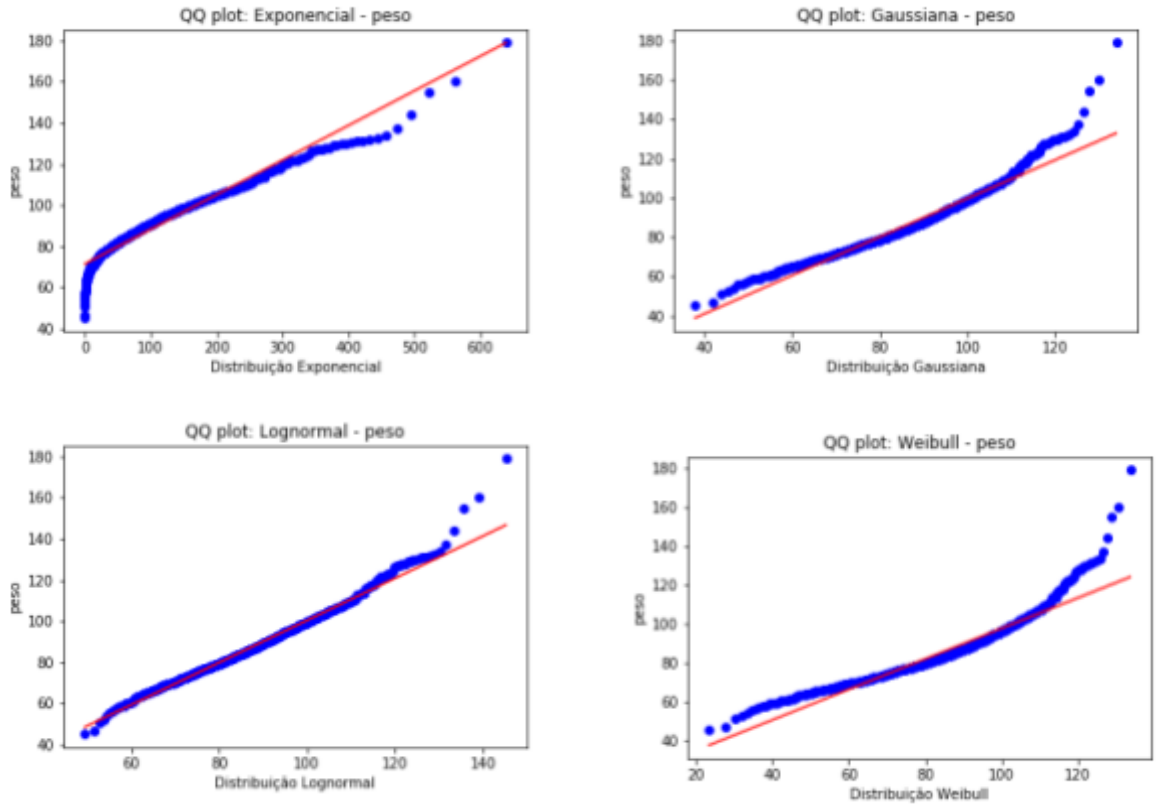






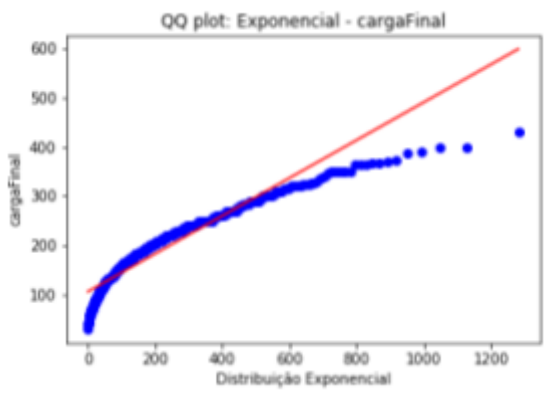
Mesmo com o QQplot, as distribuições Weibull e gaussiana se assemelham muito à empírica. Porém, é possível perceber uma semelhança maior com a Weibull na parte superior do gráfico.

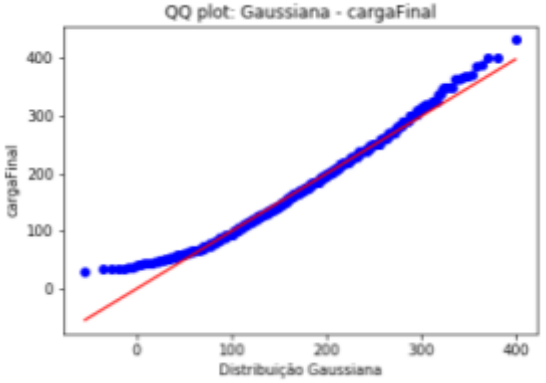
- Peso

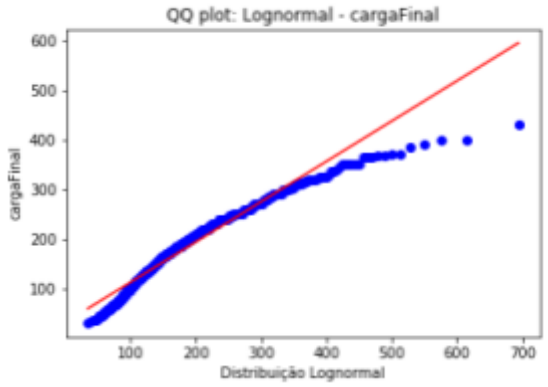


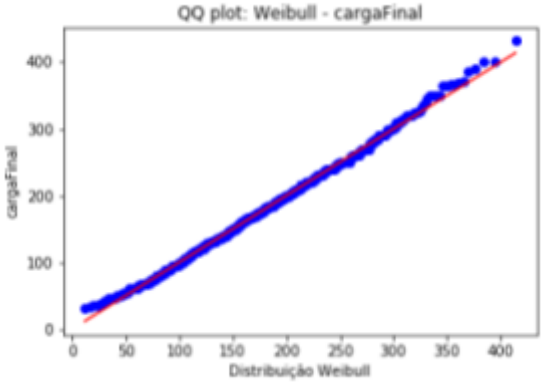
Como visto anteriormente, o número de *outliers* superirores acabou impactando consideravelmente na análise das distribuições. Porém, se desconsiderarmos esses valores, temos que a lognormal é a distribuição que mais se assemelha, como foi sugerido anteriormente.

- Carga Final



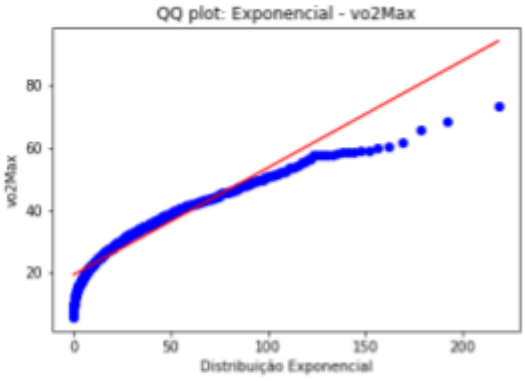


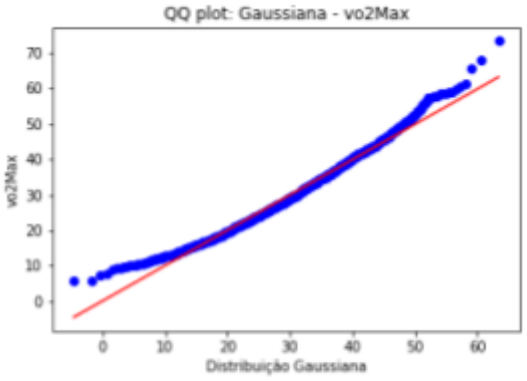


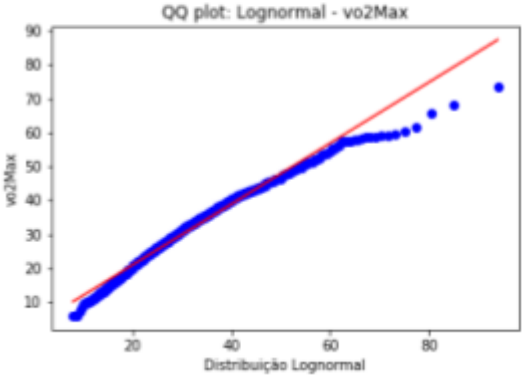


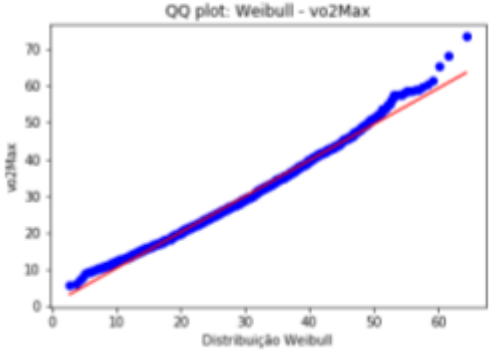
Para a carga máxima, claramente a distribuição que melhor se assemelha à distribuição empírica é a Weibull, onde se vê quase todos os pontos no eixo.

- VO2 Máximo









Assim como visto na análise do peso, os *outliers* impactaram as distribuições, porém conseguimos ter uma resposta mais assertiva em relação a distribuição Gaussiana e Weibull, onde a Weibull é a que melhor se assemelha à distribuição empírica se desconsiderarmos os *outliers* superiores, que começam na faixa dos 60.

Teste de hipótese

Nesta etapa do trabalho, foi utilizado o teste de Kolmogorov-Smirnov (KS) para validar ou descartar a compatibilidade das distribuições empíricas e da literatura.

- Idade

```
Exponencial:
        D       = 0.372755615059967
        p_value = 1.1092520093640775e-146

Gaussiana
        D       = 0.04408368872194113
```

Aqui com α = 0.05, rejeitar a hipótese de que a gaussiana também sirva não é possível. Porém, a Weibull continua apresentando os melhores resultados.

- Peso

Para o peso, com α = 0.05, a única distribuição não rejeitada, tanto para KS quanto p-valor é a lognormal. O que reforça o que vimos anteriormente.

```
Exponencial:
        D       = 0.4954410013455397
        p_value = 3.358574509201242e-266

Gaussiana
        D       = 0.06661818817785059
        p_value = 5.75842350736874e-05

Lognormal
        D       = 0.032285259002662436
        p_value = 0.17003957723540433

Weibull
        D       = 0.1032173331741221
        p_value = 2.5226220819374233e-11
```

- Carga Final

```
Exponencial:
        D       = 0.28651634266099946
        p_value = 1.1723966939662301e-85

Gaussiana
        D       = 0.039233911356943985
        p_value = 0.05277656069132966

Lognormal
        D       = 0.08035970386976421
        p_value = 4.962162909726044e-07

Weibull
        D       = 0.0245702256062538
```

Para os testes, ambas Gaussiana e Weibull são válidas. Entretanto, a Weibull mais uma vez possui os melhores resultados. Portanto, com o resultado dos testes e a análise dos QQplots, podemos dizer que a melhor distribuição para modelar a carga final é a Weibull.

- VO2 Máximo

```
Teste de Hipótese: vo2Max

Exponencial:
        D       = 0.3348896789424037
        p_value = 1.0598592744011172e-117

Gaussiana
        D       = 0.044531849851028094
        p_value = 0.018572422090605088

Lognormal
```

Para os testes, apenas a exponencial não é uma modelagem válida do conjunto. Porém, mais uma vez, a variável que melhor

representa o conjunto de dados é a Weibull. Que dá mais força a suspeita entre uma relação entre carga final e VO2 máximo, já que são melhor representadas pela mesma variável.

Análise de dependência entre as variáveis

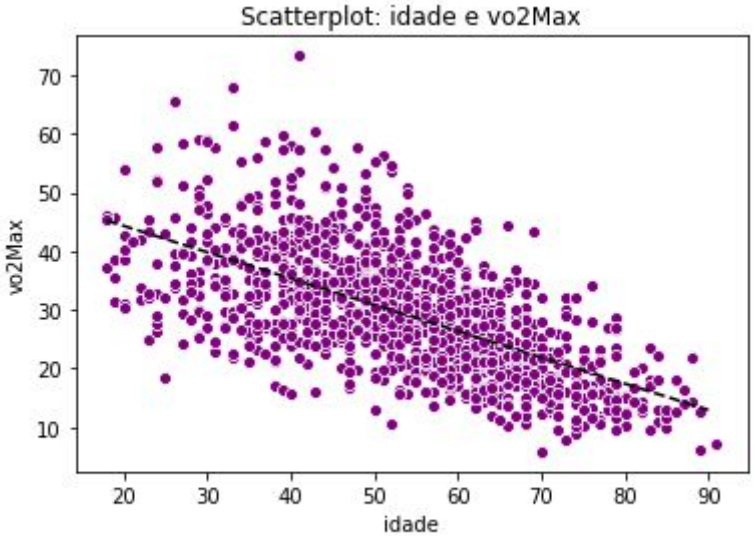
Esta parte do projeto tem por objetivo tentar descobrir possíveis relações entre as variáveis e posteriormente, tentar fazer predições.

O coeficiente de correlação entre variáveis é cálculado pela seguinte fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

, onde n é o número de amostras, r é o coeficiente de correlação e, {x,y} são as variáveis.

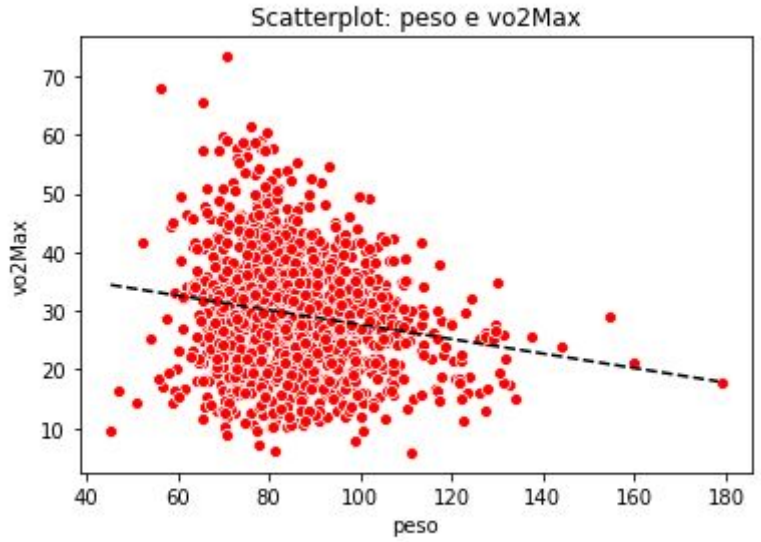
- Idade e VO2 Máximo




```
Regressão Linear:  [-0.44852097 53.2968391 ]
```

Pelo scatterplot, não podemos perceber uma relação linear explícita entre as duas variáveis.

- Peso e VO2 Máximo

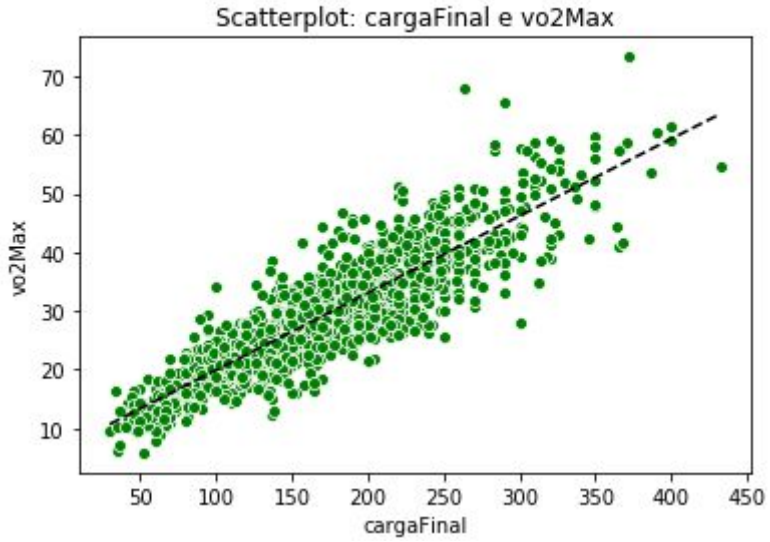



```
Regressão Linear:  [-0.12370517 40.02419091]
```

Para as variáveis peso e VO2 máximo, também não conseguimos falar de uma relação explícita entre elas.

- Carga Final e VO2 Máximo

```
Coeficiente de correlação: cargaFinal - vo2Max:
```



Aqui podemos perceber que há uma relação mais clara entre a Carga Final do paciente e a sua capacidade de VO2. É um resultado importante que permite uma exploração sobre a predição da capacidade de VO2 do paciente dada sua Carga final no teste.

Inferência Bayesiana

Esta parte do projeto não foi realizada por questões de tempo.

Conclusão

Apesar de encontrar uma relação importante entre Carga Final e o VO2 máximo, infelizmente não foi possível tratar da parte de predição de resultados. Apesar disso, foi um projeto em que os conceitos de modelagem estatística se fizeram presente de forma interessante.

Referências

https://www.land.ufrj.br/~/classes/est-prob-2019/
https://devdocs.io
https://seaborn.pydata.org/index.html