

# **Aprendizagem Supervisionada Aula 2**

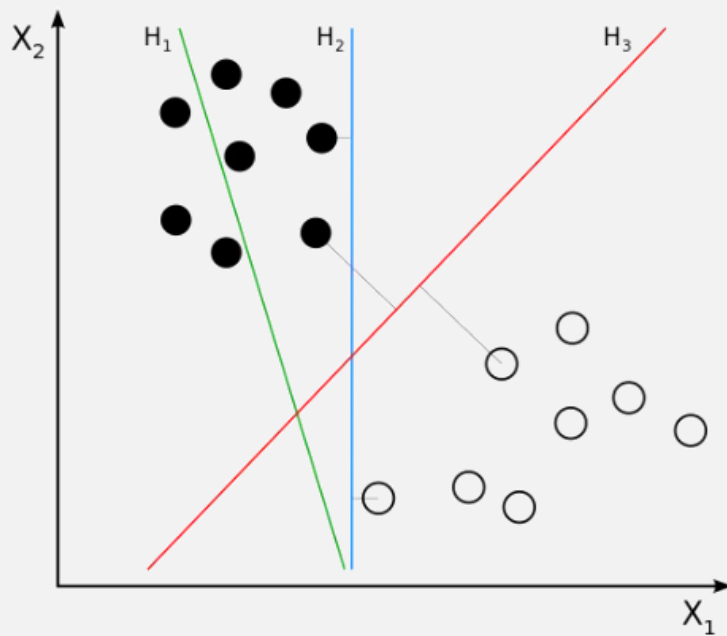
Análise Exploratória de  
Dados

Bernard da Silva  
Orientador: Rafael Parpinelli  
21/11/2022

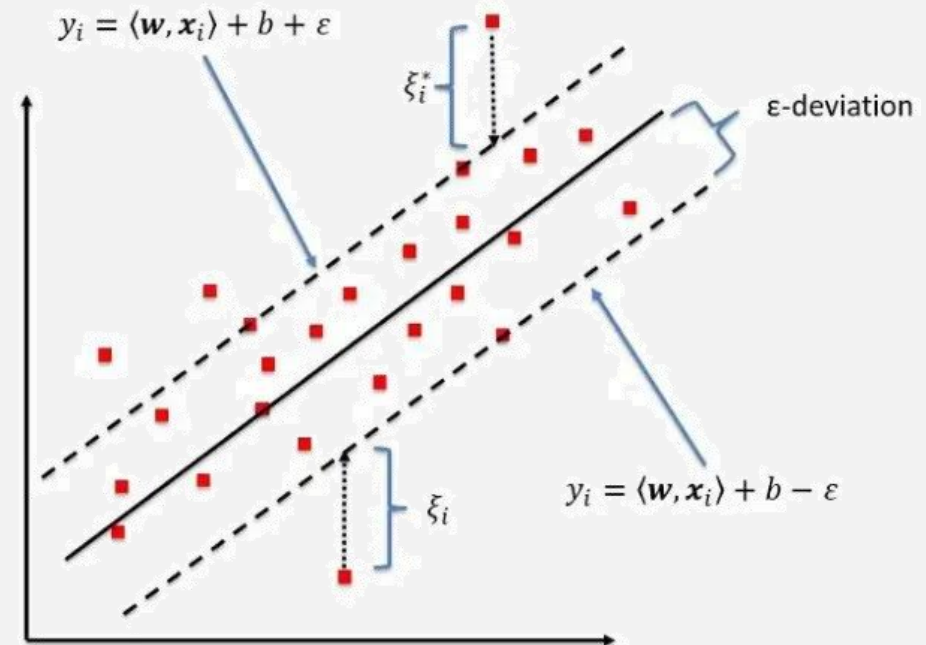
# Dúvidas aula passada

SVM – Support Vector Machine

Classificação

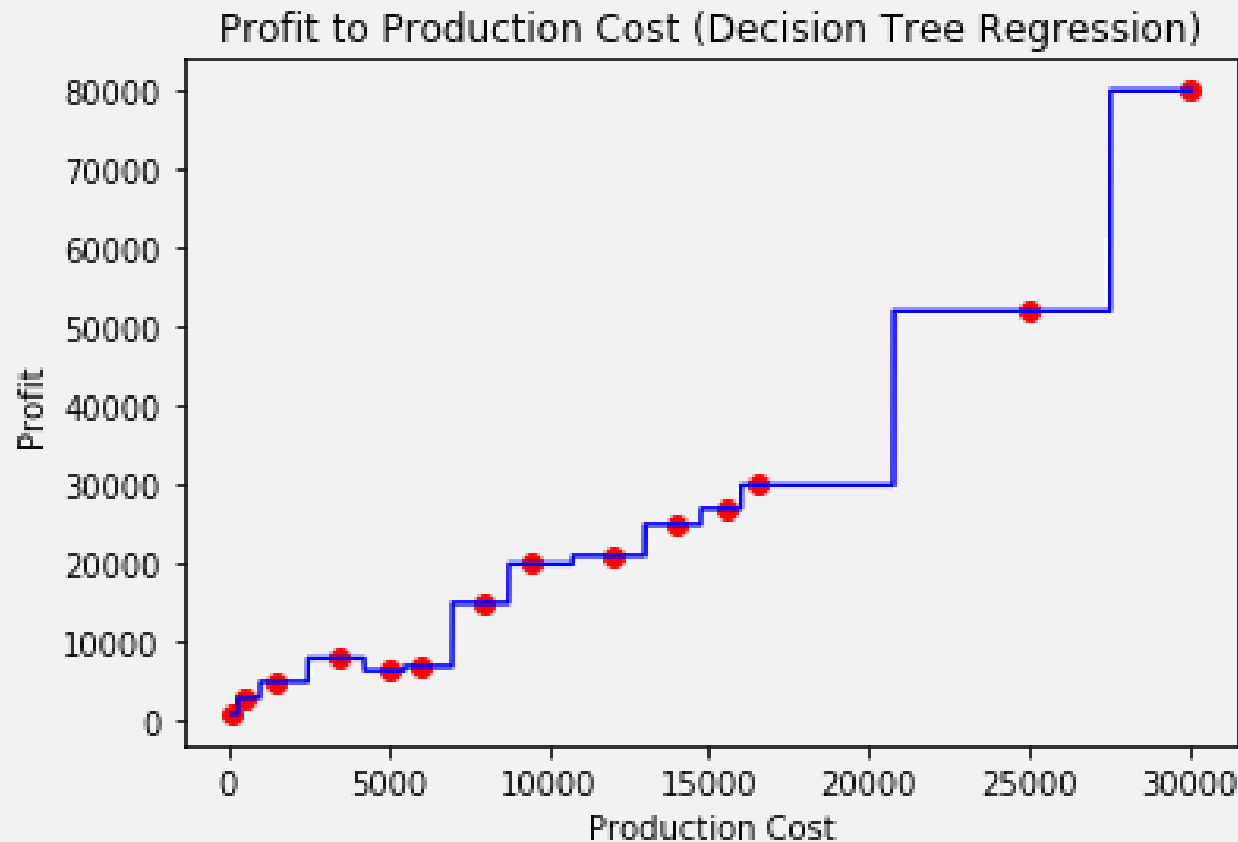


Regressão



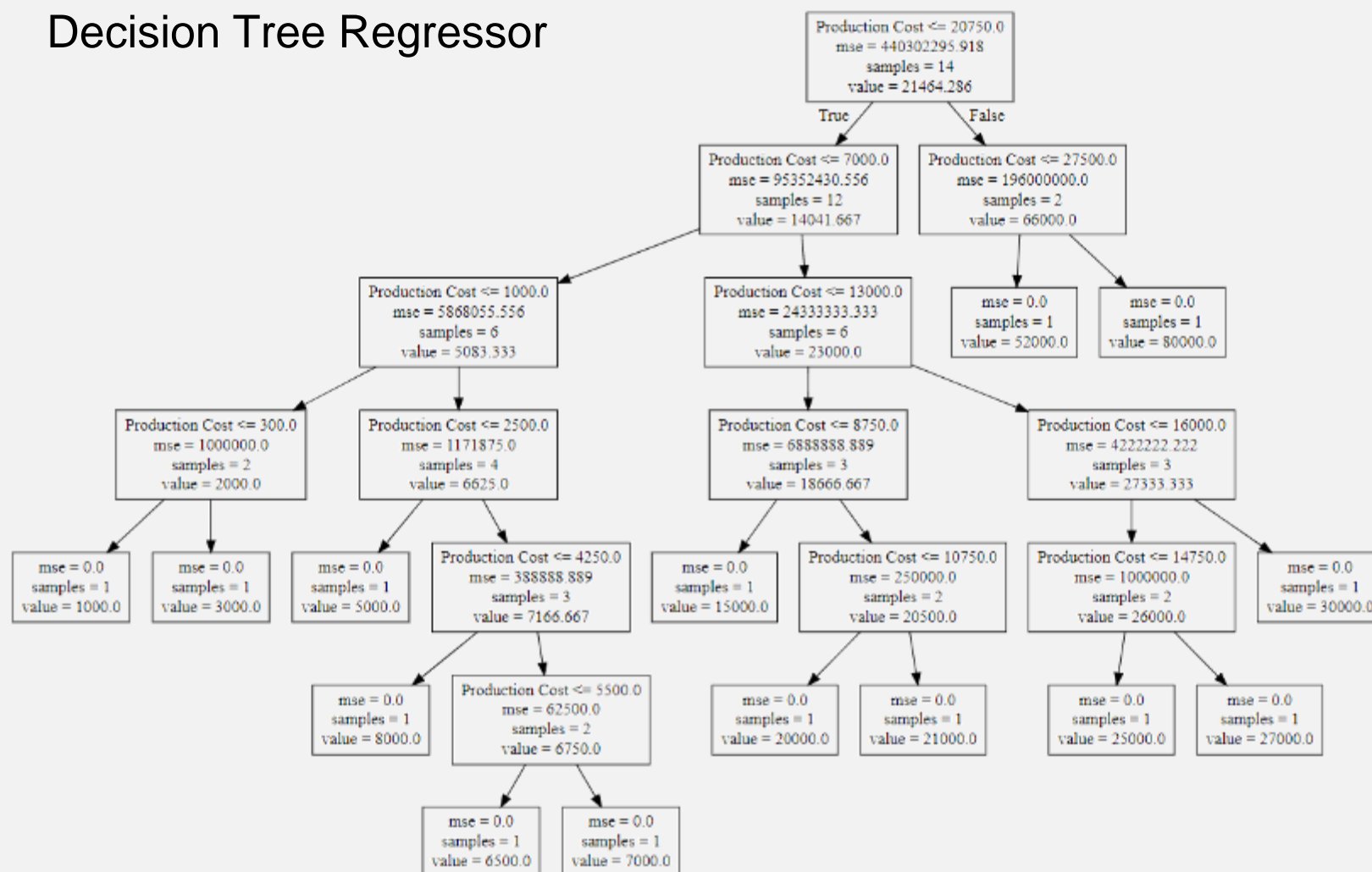
# Dúvidas aula passada

## Decision Tree Regressor



# Dúvidas aula passada

## Decision Tree Regressor

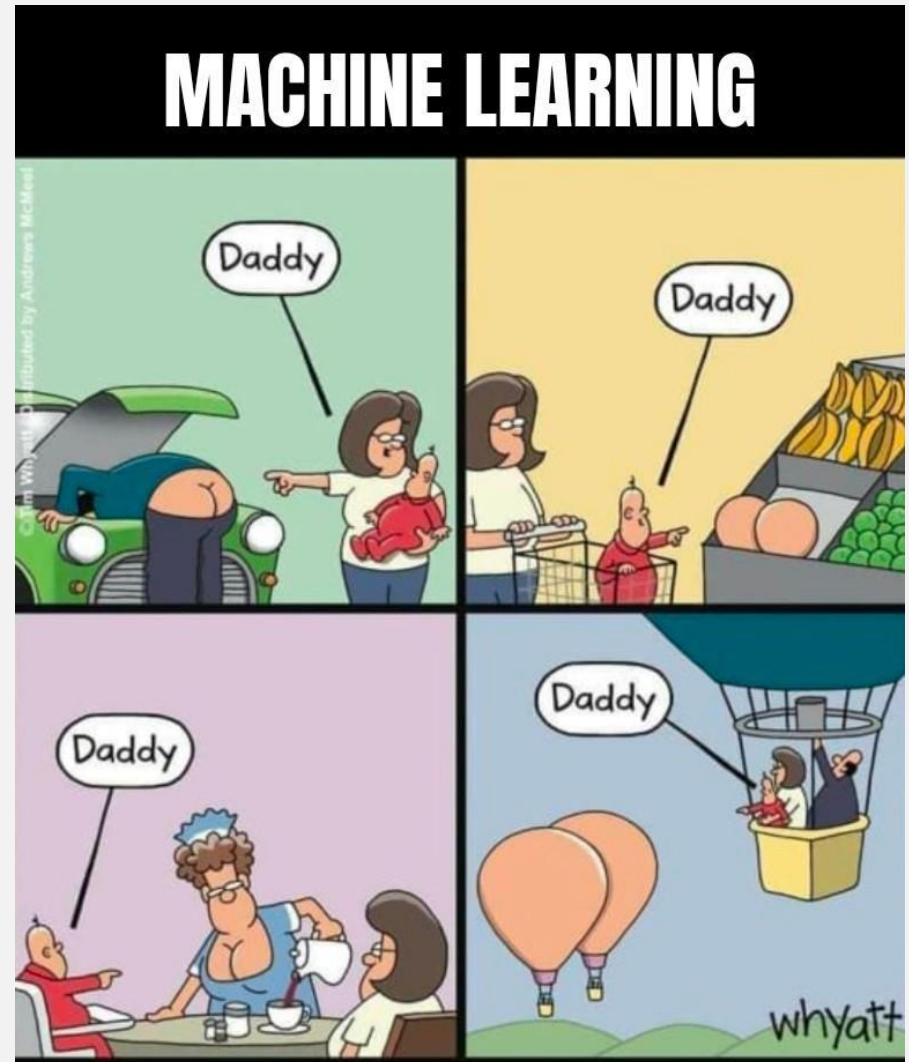


# Objetivo dessa aula

- Ter uma noção de como fazer uma Análise Exploratória de Dados e pré-processar dados.
- Entender seu PROBLEMA.

# A importância de dados em aprendizagem de máquina

- Um modelo de Machine Learning é tão bom quanto os dados usados na sua modelagem.
- Dados não balanceados podem gerar modelos não balanceados
- Dados com tendência podem gerar modelos tendenciosos
- Por que isso é perigoso?



# A importância dos dados

Em março de 2016, a Microsoft descobriu que usar as interações do Twitter como dados de treinamento para algoritmos de Machine Learning pode ter resultados desanimadores.



Em 16 horas, o chatbot postou mais de 95.000 tweets, e esses tweets rapidamente se tornaram abertamente racistas, misóginos e antissemitas. A Microsoft suspendeu rapidamente o serviço para ajustes e, por fim, retirou o plugue.



# A importância dos dados

Amazon desiste de ferramenta secreta de recrutamento que mostrou viés contra mulheres.



Isso porque os modelos de computador da Amazon foram treinados para examinar os candidatos observando padrões em currículos enviados à empresa durante um período de 10 anos. A maioria veio de homens, um reflexo do domínio masculino em toda a indústria de tecnologia.





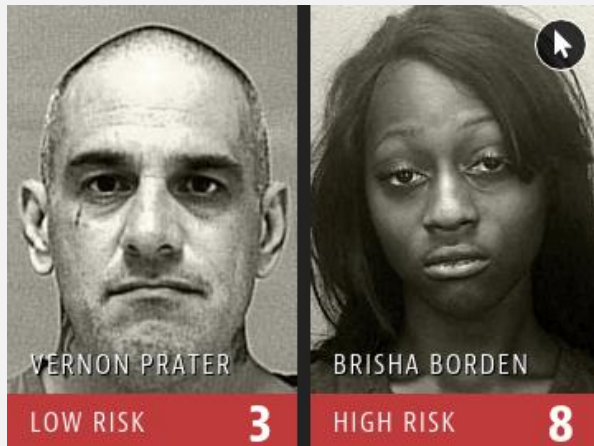
# A importância de dados em aprendizagem de máquina

- Às vezes pode parecer cômico...



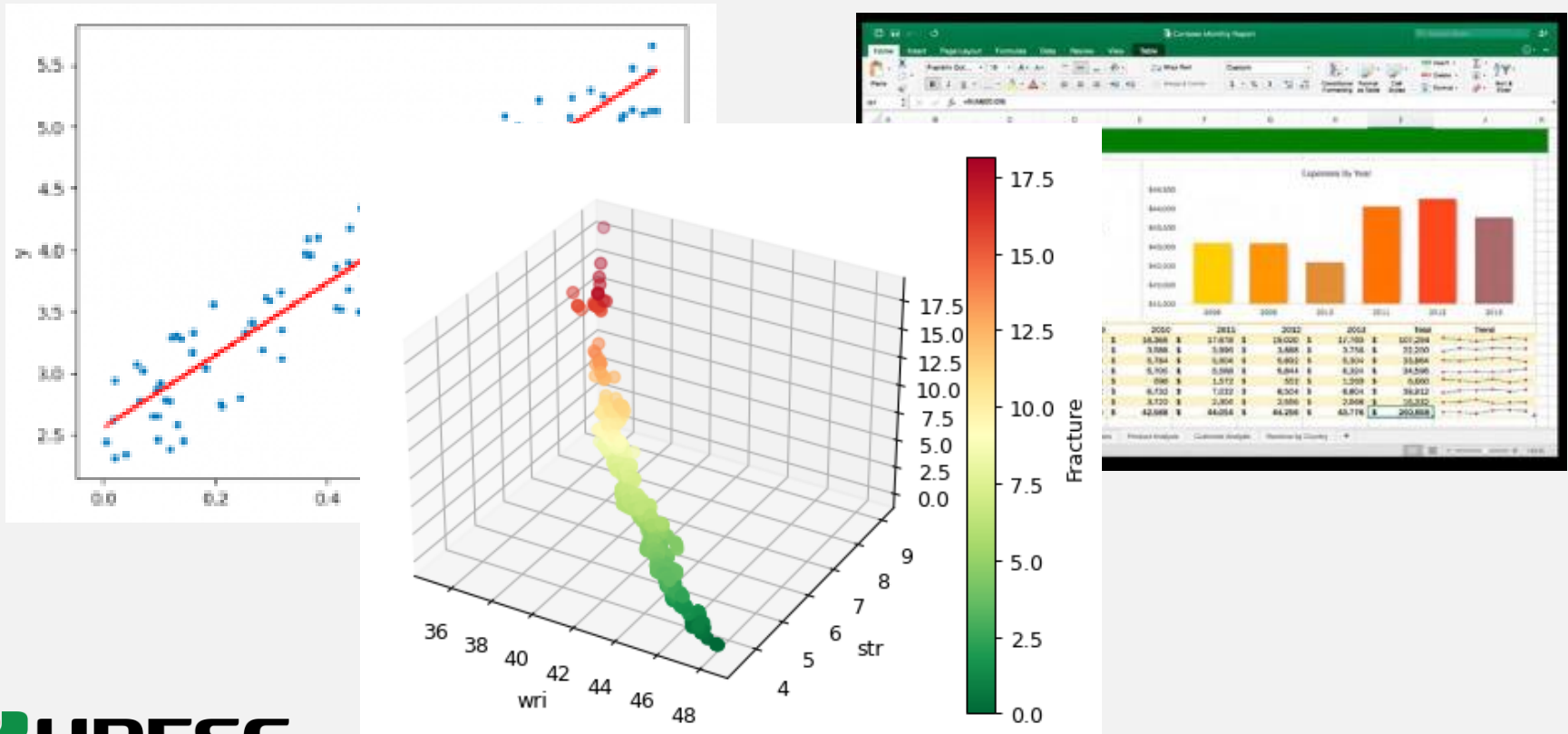
# A importância de dados em aprendizagem de máquina

- Um modelo de ML em 2016 que calculava o “risco” de um “criminoso” pode inferir que pessoas negras têm maior risco que pessoas brancas.



# Mas afinal... o que são dados?

- Possíveis definições:
  - Dado é o registro do atributo de um ente, objeto ou fenômeno.
  - Unidades de informação, coletados através de observação.



# Tipo de dados

- Dados numéricos vs Dados categóricos
  - O que são, e qual a diferença?



# Tipo de dados

- Numéricos:
  - Contínuos
    - Admitem qualquer valor numérico
      - [..., -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, ... , 1000000.0, ...]
  - Discretos
    - Admitem apenas valores inteiros
      - [..., -2, -1, 0, 1, 2, 3, ... ]
- Categóricos:
  - Nominais
    - Não existe ordenação entre as categorias
      - [Masculino, Feminino], [Sim, Não], [Alto, Médio, Baixo]
  - Ordinais
    - Existe ordenação entre as categorias
      - Mês/ano/dia de observação, Nível de escolaridade, Estágio de doenças, etc.

# Distribuição de variáveis

Análise exploratória de dados são divididas em duas etapas:

1. **Inspeção visual** dos dados, para entender como eles estão distribuídos;
2. Obtenção de **medidas estatísticas** que descrevam os dados.

# Inspeção Visual

As ferramentas mais comuns para inspeção visual dos dados:

- Diagrama ramo-e-folhas (stem-and-leaf);
- Gráfico de dispersão (scatter plot);
- Histograma;
- Diagrama de caixa (boxplot);
- Gráfico da distribuição empírica.

# Inspeção Visual

## Diagrama ramo-e-folhas

São úteis para mostrar a densidade relativa e a forma dos dados, dando ao leitor uma rápida visão geral da distribuição.

No Exemplo 1, os tempos de execução do programa B são os seguintes:

125 130 125 126 121 130 123 121 123 131

O diagrama ramo-e-folhas para esses tempos de execução ficaria assim:

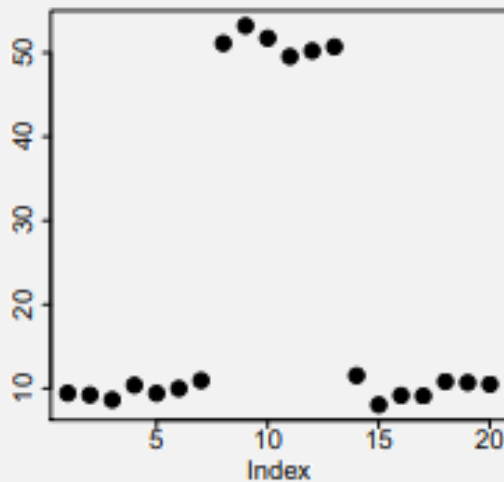
<i>ramo</i>	<i>folhas</i>
12	1133556
13	001



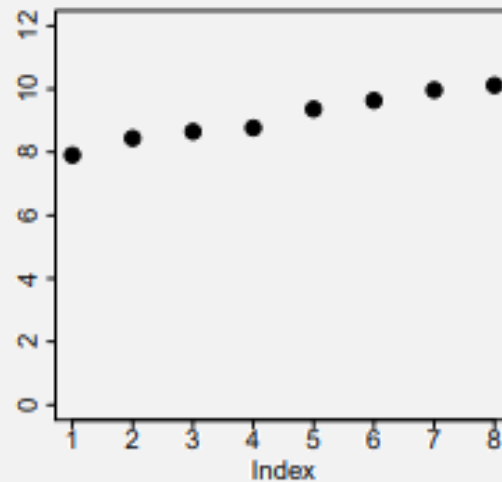
# Inspeção Visual

## Gráfico de dispersão (Scatter Plot)

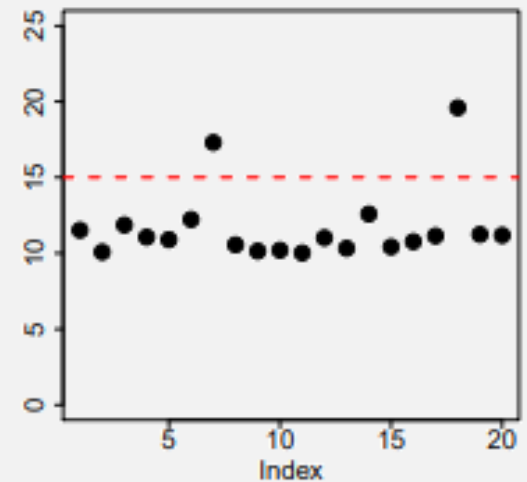
Gráficos de dispersão (scatter plots) são muito usados para analisar a relação entre variáveis. Eles também permitem identificar padrões (tendências, anomalias) nos dados e visualizar aderência a uma curva.



(a)



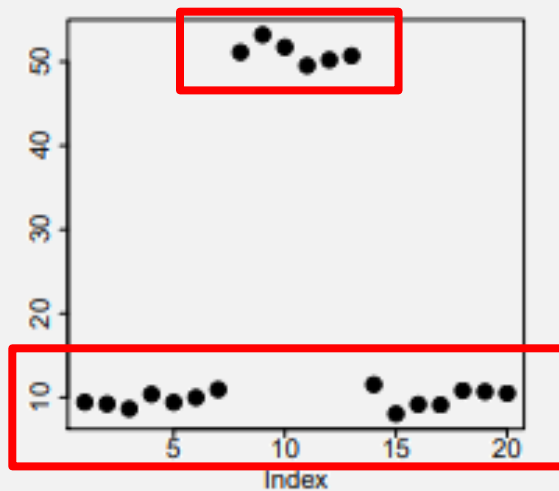
(b)



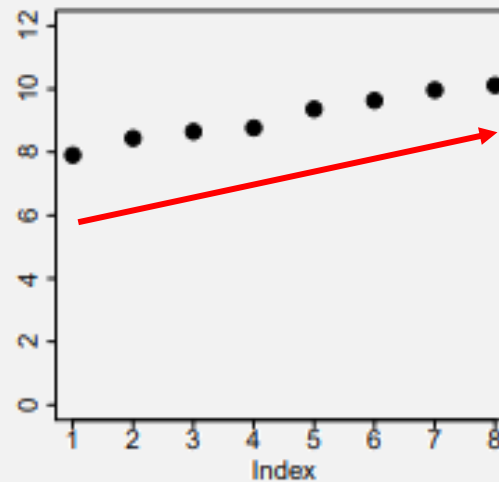
(c)

# Inspeção Visual

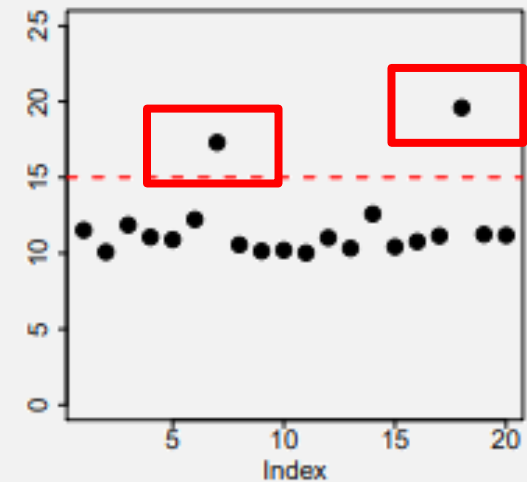
Como avaliar um gráfico de dispersão (Scatter Plot)?



(a)



(b)

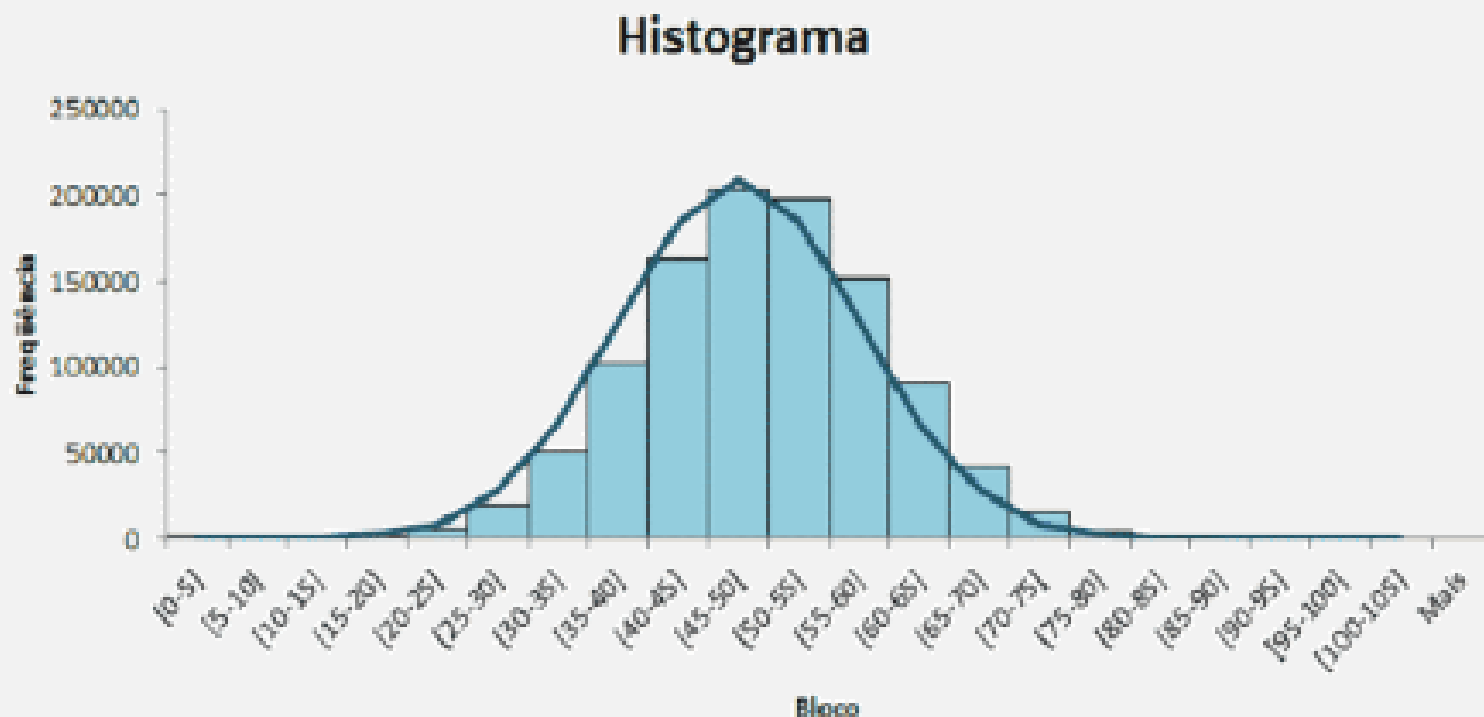


(c)

# Inspeção Visual

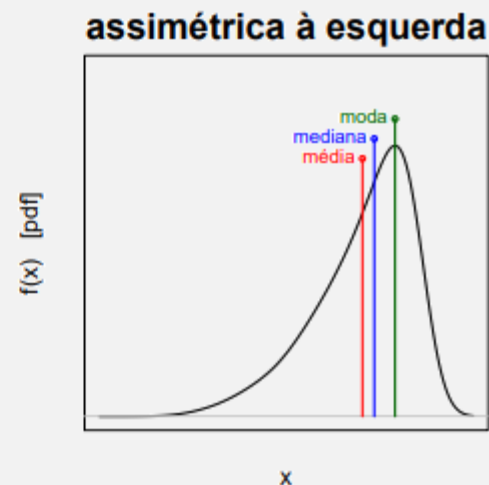
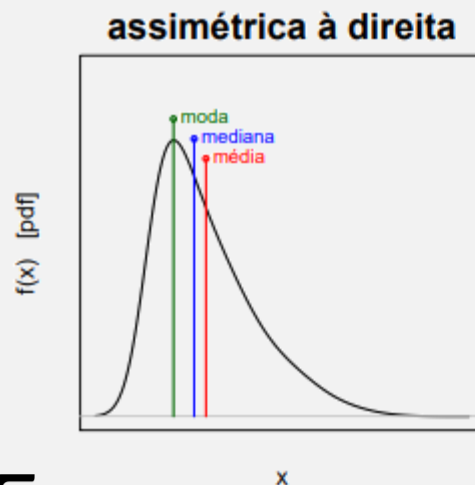
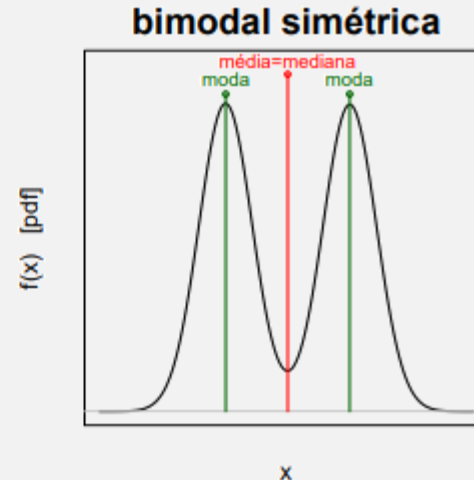
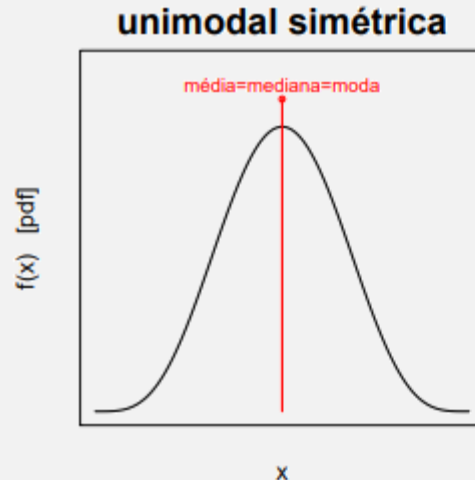
## Histograma

É um gráfico de colunas que mostra dados de frequência.



# Inspeção Visual

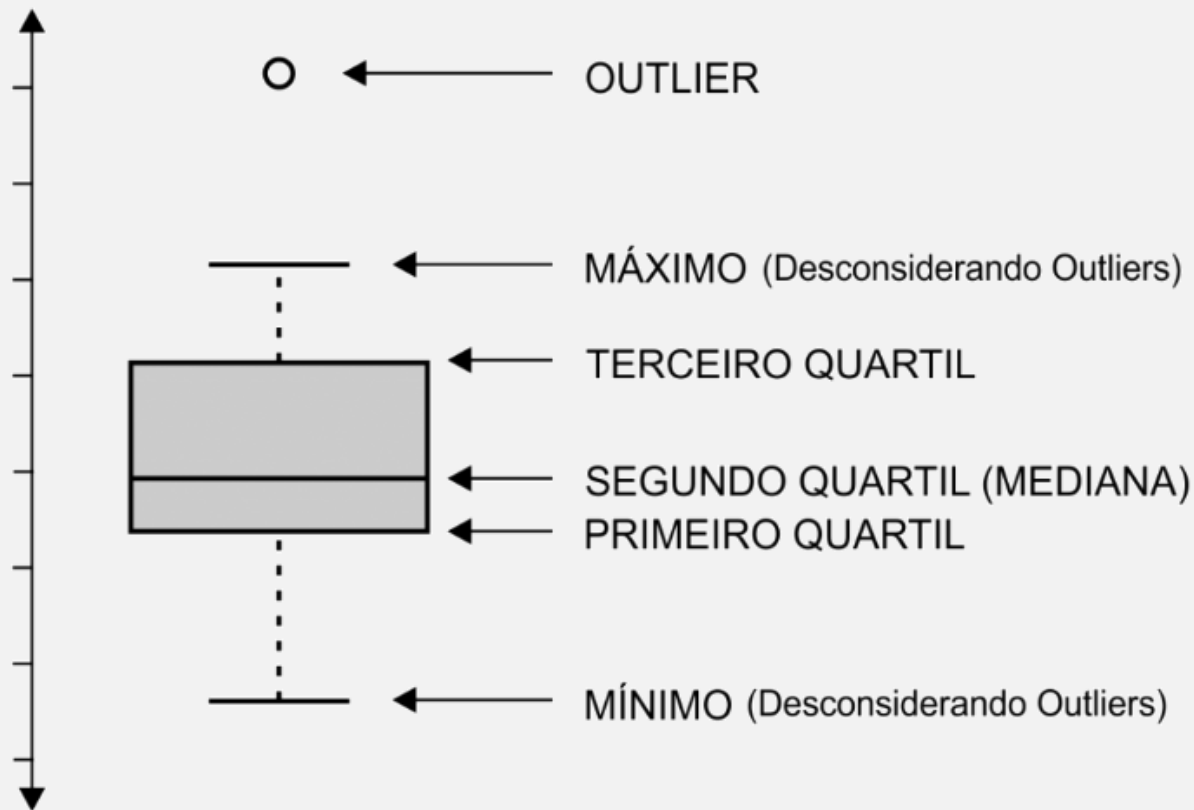
Como avaliar um Histograma?



# Inspeção Visual

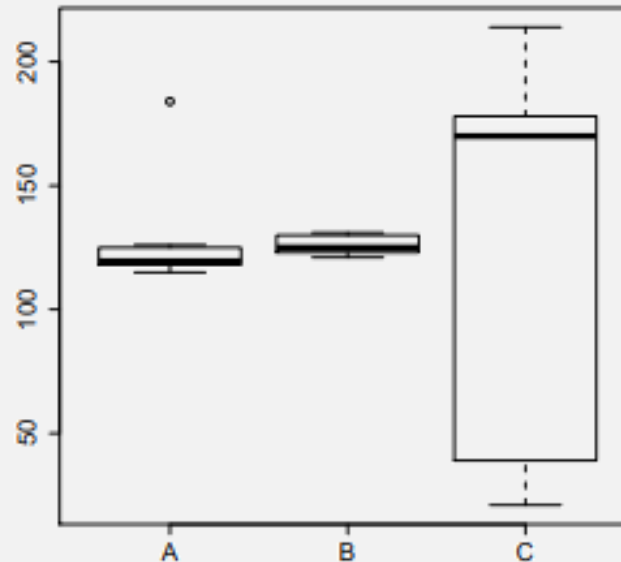
## Diagrama de Caixa (Boxplot)

É uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes (outliers) dos dados.



# Inspeção Visual

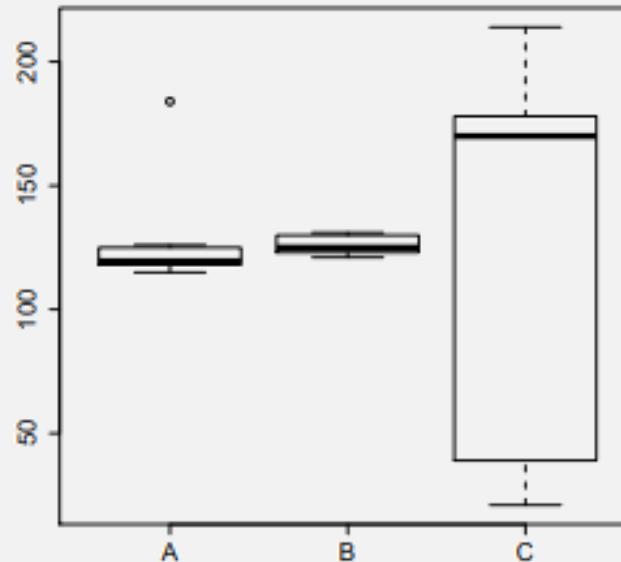
Como avaliar um Diagrama de Caixa (Boxplot)?



1 - Os tempos de execução para o programa A são concentrados entre 115 e 125 ms, aproximadamente, com um outlier com cerca de 185 ms.

# Inspeção Visual

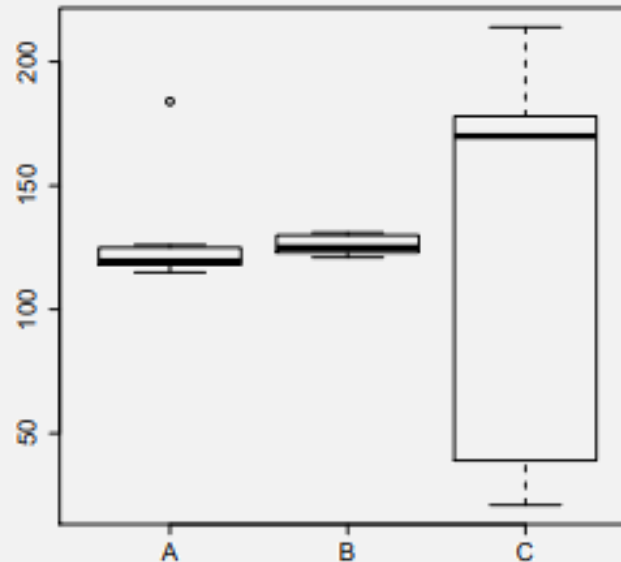
Como avaliar um Diagrama de Caixa (Boxplot)?



2 - Os tempos de execução para o programa B são concentrados entre 120 e 130 ms, aproximadamente, com mediana levemente superior à do programa A.

# Inspeção Visual

Como avaliar um Diagrama de Caixa (Boxplot)?



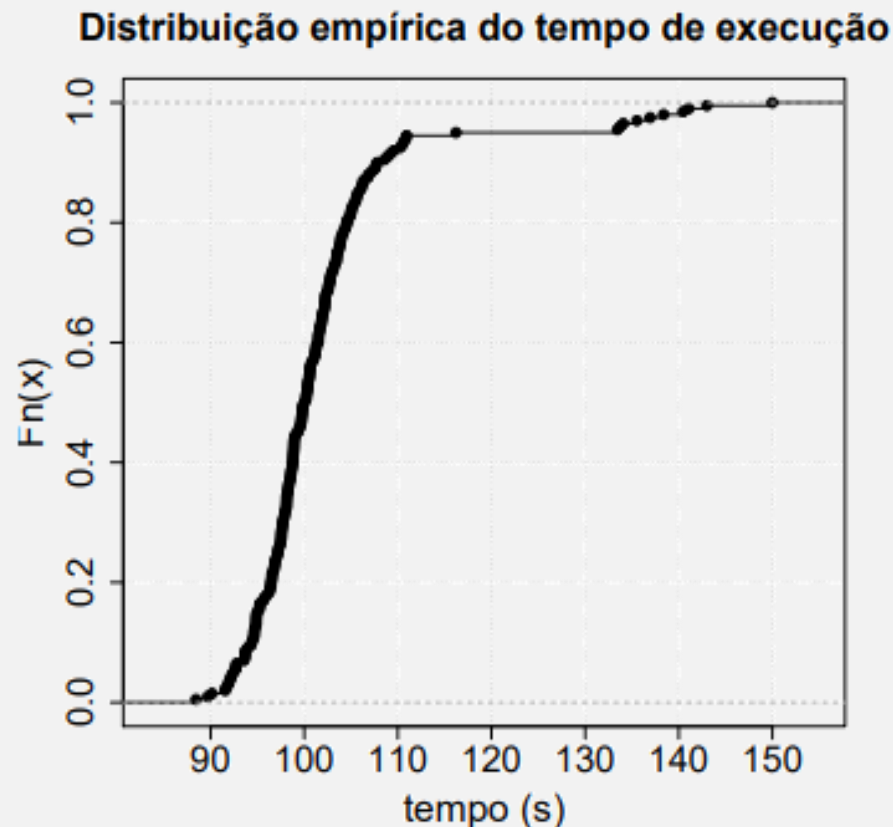
3 - Os tempos de execução para o programa C são bastante dispersos entre 20 e 210 ms, aproximadamente. A mediana e o 3º quartil ficam perto de 175 ms enquanto o 1º quartil fica perto de 40 ms, o que sugere uma distribuição com cauda à esquerda ou bimodal.



# Inspeção Visual

## Gráfico de Distribuição Empírica

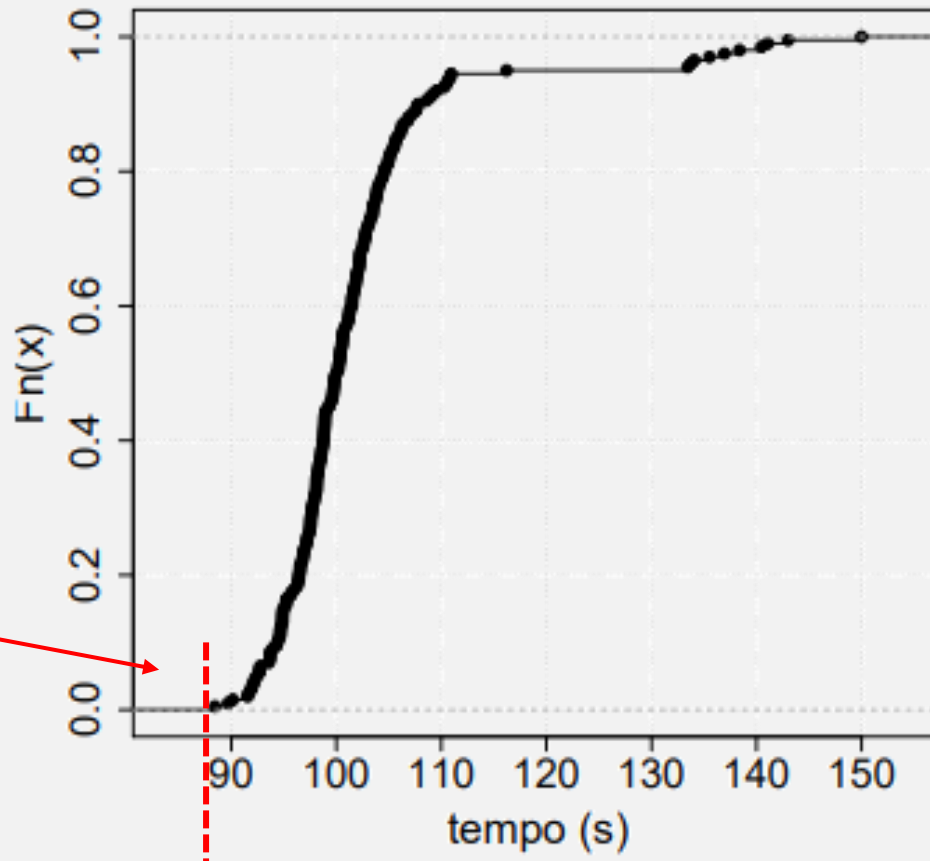
O gráfico da função de distribuição empírica permite visualizar a distribuição da variável sem precisar determinar o número de classes usadas no histograma.



# Inspeção Visual

Como avaliar o Gráfico de Distribuição Empírica?

Distribuição empírica do tempo de execução

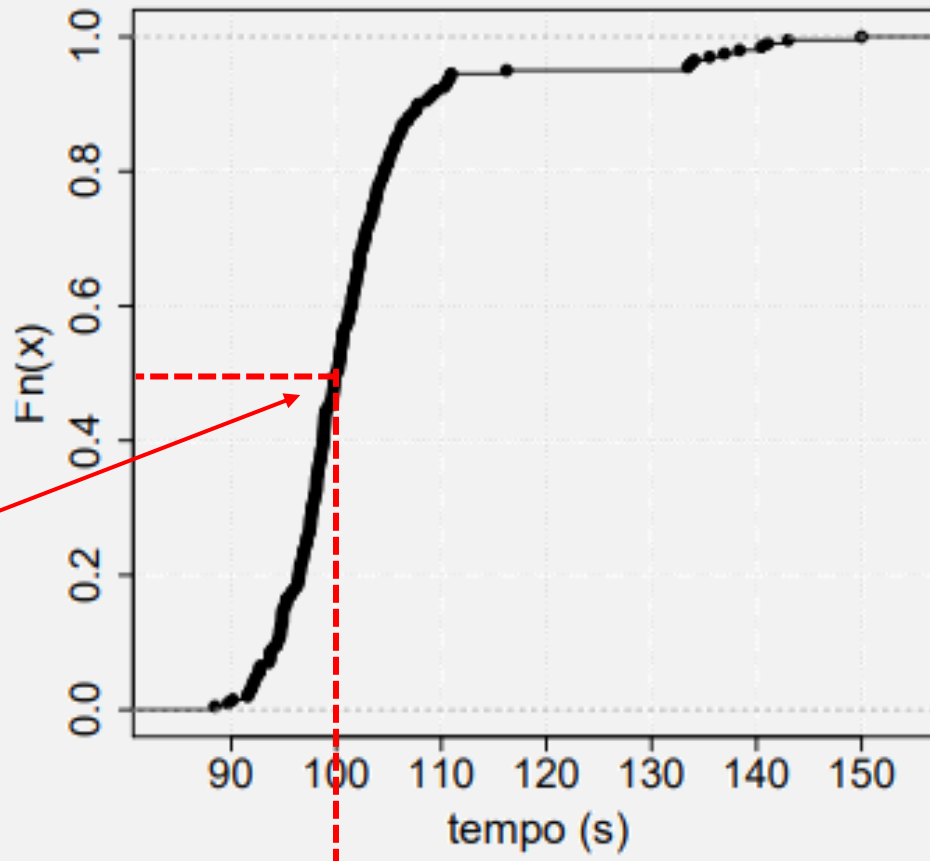


Tempo mínimo  
pouco abaixo  
de 90 s

# Inspeção Visual

Como avaliar o Gráfico de Distribuição Empírica?

Distribuição empírica do tempo de execução

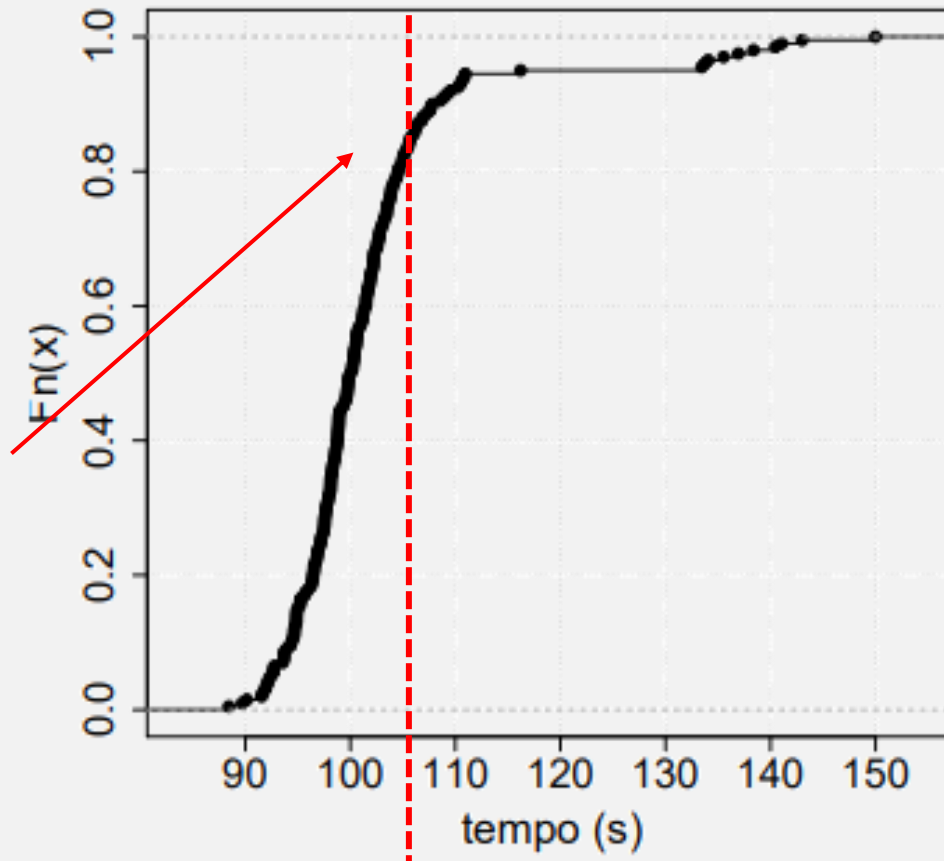


Mediana (50%)  
de  
aproximadamente  
100 s

# Inspeção Visual

Como avaliar o Gráfico de Distribuição Empírica?

Distribuição empírica do tempo de execução

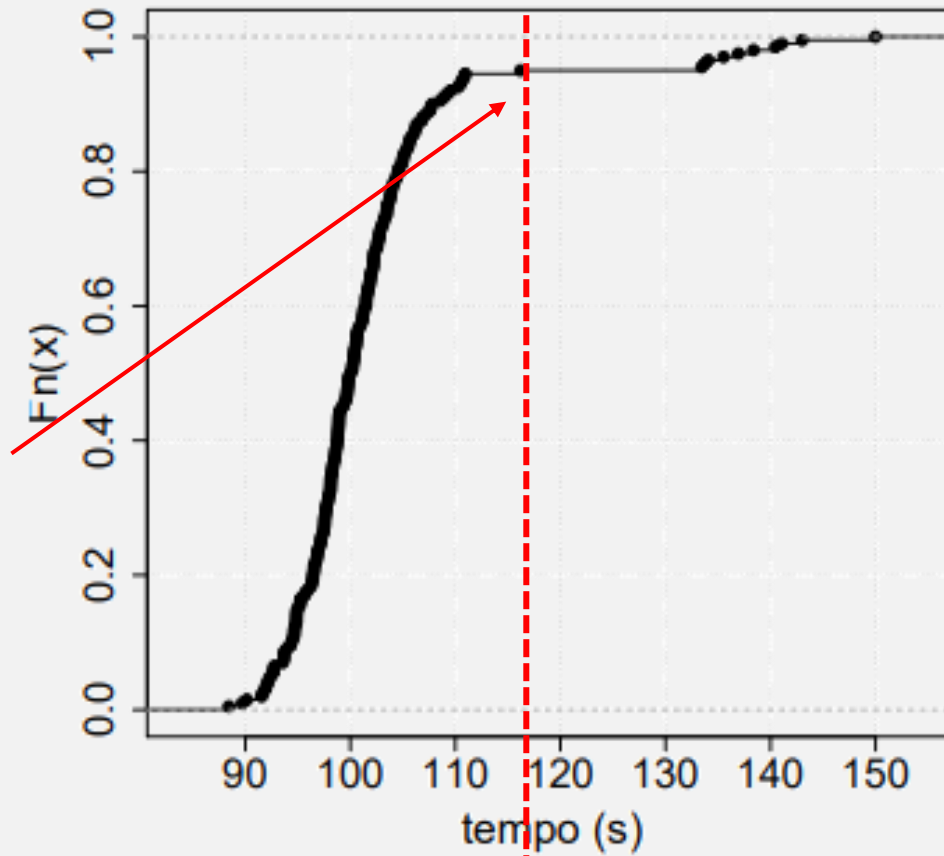


80% dos tempos inferiores a 105 s

# Inspeção Visual

Como avaliar o Gráfico de Distribuição Empírica?

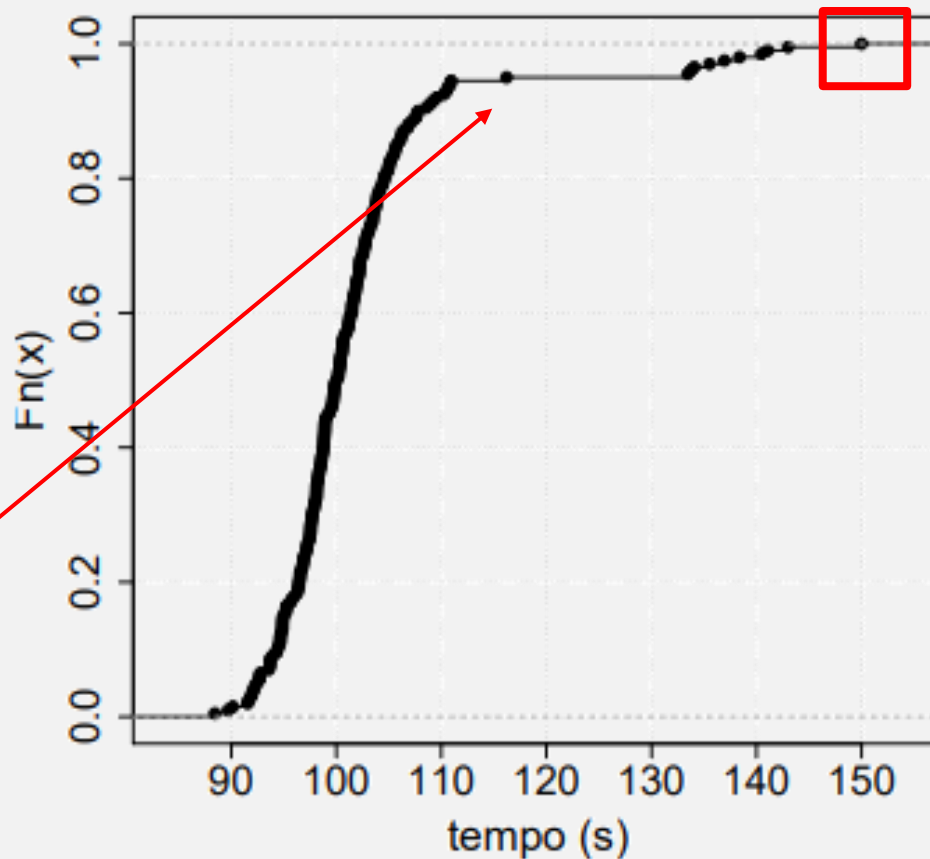
Distribuição empírica do tempo de execução



# Inspeção Visual

Como avaliar o Gráfico de Distribuição Empírica?

Distribuição empírica do tempo de execução



Tempo máximo  
de  
aproximadamente  
150 s

# Medidas Estatísticas

Essas medidas dividem-se em dois grupos, medidas de tendência central e medidas de dispersão.

As medidas de tendência central (ou de posição) resumem os dados em um único número, e incluem:

- Média (aritmética, truncada, harmônica);
- Mediana;
- Moda;
- Quantis (quartis, percentis).

# Medidas Estatísticas

Média Aritmética (dados uniformes e comportados):

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

Média Truncada (desconsiderar outliers):

$$A: \bar{x} = \frac{966}{8} = 120,8 \text{ ms}$$

Média Harmônica (vazão ou taxa de transferência):

$$MH = \frac{1}{\frac{1}{N} \left( \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N} \right)} = \left( \frac{1}{N} \sum_{i=1}^N \frac{1}{x_i} \right)^{-1}$$



# Medidas Estatísticas

## Mediana

A mediana é o valor intermediário em um conjunto de dados: metade dos valores estão abaixo da mediana e metade acima. A mediana é obtida a partir do conjunto ordenado. Se o número de elementos for ímpar, a mediana é o elemento central:

$$x = \{2, 4, 6, 8, 51\} \Rightarrow \text{med}(x) = 6$$

E se o quantidade de números for par?

$$x = \{2, 4, 6, 8, 51, 171\} \Rightarrow \text{med}(x) = (6 + 8)/2 = 7$$

A mediana é mais indicada que a média quando a distribuição é assimétrica.

# Medidas Estatísticas

## Moda

A moda é o valor que aparece com maior frequência no conjunto. Conjuntos de dados podem ser unimodais ou multimodais, ou mesmo amodais (não têm moda).

$$x = \{115, 117, 118, 118, 118, 121, 123, 125, 126, 184\}$$

A moda é mais indicada para dados categóricos, isto é, variáveis que podem assumir um dentre um conjunto finito de valores (numéricos ou não), que são as categorias.

# Medidas Estatísticas

## Quartis e Percentis

Quartis e percentis são outras medidas de posição que descrevem os dados de forma resumida.

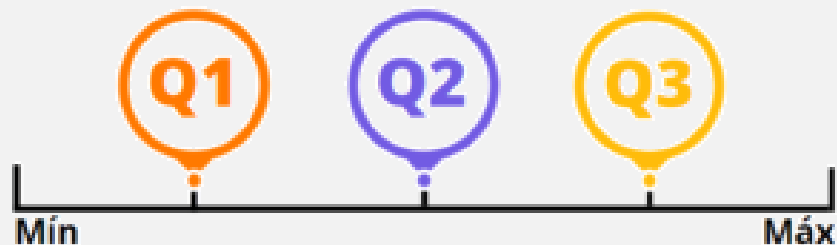
A: 118 125 121 118 184 115 123 118 126 117

$$P_{90} = 126$$

$$Q_1 = P_{25}$$

$$Q_2 = P_{50}$$

$$Q_3 = P_{75}$$



Os percentis e quartis são interessantes para termos uma ideia de como os dados estão distribuídos e tirarmos algumas conclusões sobre esses dados.

# Medidas de Dispersão

As medidas de dispersão, por sua vez, quantificam a variabilidade dos dados, e incluem:

- Amplitude;
- Variância;
- Desvio padrão;
- Coeficiente de variação.

# Medidas de Dispersão

Amplitude:

A amplitude de um conjunto é a diferença entre o maior e o menor valor. A amplitude interquartil (interquartil e range, IQR) é a diferença entre o 3º quartil e o 1º quartil:

$$IQR = Q_3 - Q_1$$

# Medidas de Dispersão

Variância Amostral:

É dada pelo somatório das diferenças quadráticas entre as observações  $x_i$  e a média, dividido por  $N - 1$ :

$$s^2 = \frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2$$

# Medidas de Dispersão

## Desvio Padrão Amostral:

A variância não é expressa na mesma unidade dos dados, o que dificulta a análise da variabilidade. Por isso, é mais comum trabalhar com o desvio padrão amostral ( $s$ ), que é a raiz quadrada (positiva) da variância amostral:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

A magnitude do desvio padrão deve ser interpretada em relação à média. Por exemplo, um desvio padrão de 1 pode ser considerado baixo quando a média for 100, mas alto quando a média for 2.

# Medidas de Dispersão

Coeficiente de Variação (CV):

Para facilitar essa interpretação é possível recorrer ao coeficiente de variação (CV), que é a razão entre o desvio padrão e a média:

$$CV = \frac{s}{\bar{x}}$$

O CV permite comparar a variabilidade de duas ou mais variáveis mesmo que as respectivas médias sejam muito diferentes, ou que essas variáveis sejam expressas em unidades distintas (o CV é adimensional). Por outro lado, quando a média é próxima de zero, o CV tende a infinito, e torna-se uma medida menos útil. O CV geralmente é expresso como uma porcentagem.



# Escolhendo Medidas

Uma questão que surge naturalmente é qual medida de tendência central e de dispersão utilizar, uma vez que existem várias opções.

- Dados categóricos  $\Rightarrow$  moda;
- Total de interesse  $\Rightarrow$  média;
- Distribuição assimétrica  $\Rightarrow$  mediana;
- Caso contrário, use a média.

# Estatística Univariada, Bivariada e Multivariada

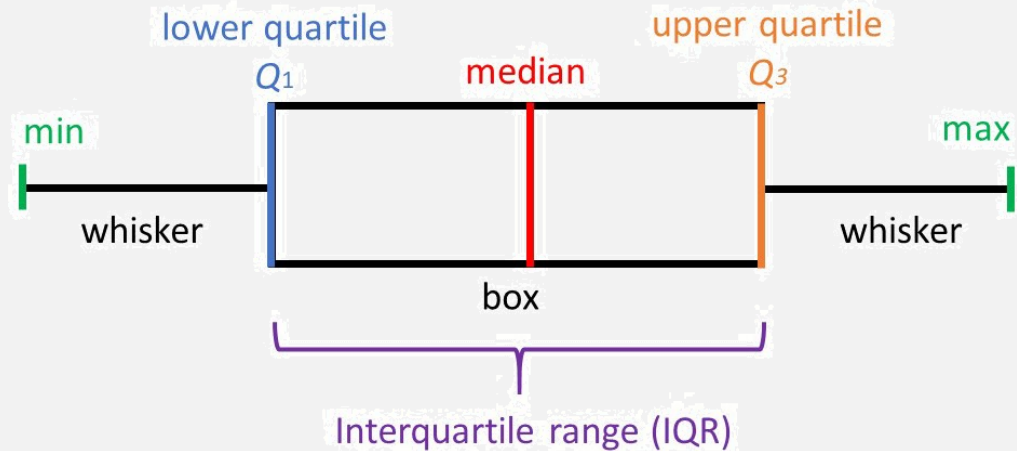
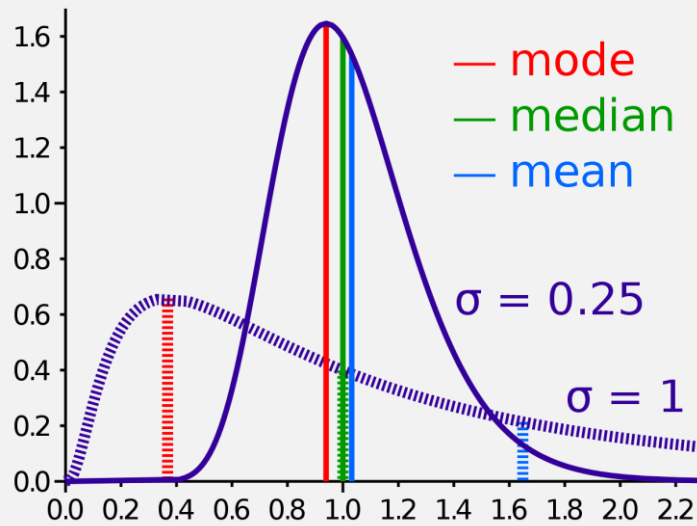
A **Estatística Univariada** inclui todos os métodos de estatística descritiva que permitem a análise de cada variável separadamente.

A **Estatística Bivariada** inclui métodos de análise de duas variáveis, podendo ser ou não estabelecida uma relação de causa/efeito entre elas

A **Estatística Multivariada** inclui os métodos de análise das relações de múltiplas variáveis dependentes e/ou múltiplas variáveis independentes, quer se estabeleçam ou não relações de causa/efeito entre estes dois grupos.

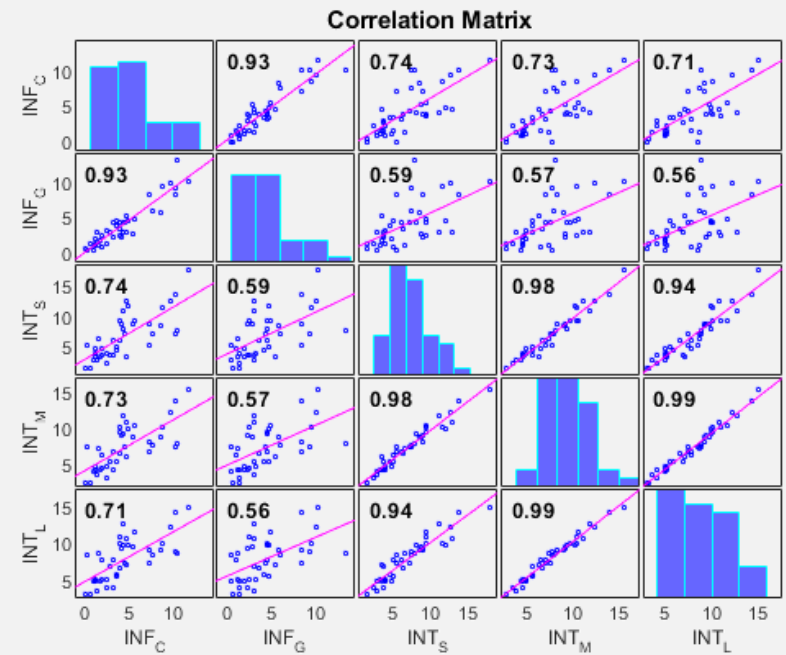
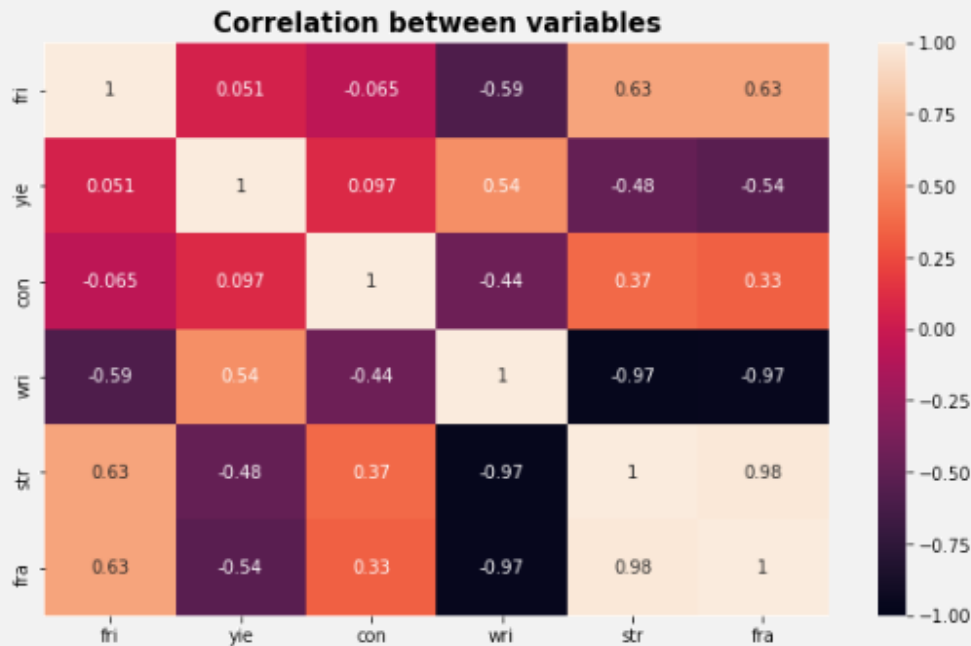
# Análise Univariada

- Para dados quantitativos
  - Média, Mediana, Moda, Mínimo, Máximo, Intervalo, Percentil, Amplitude Interquartil, Histograma, Box Plot.
- Para dados qualitativos
  - Moda, Frequência, Histograma



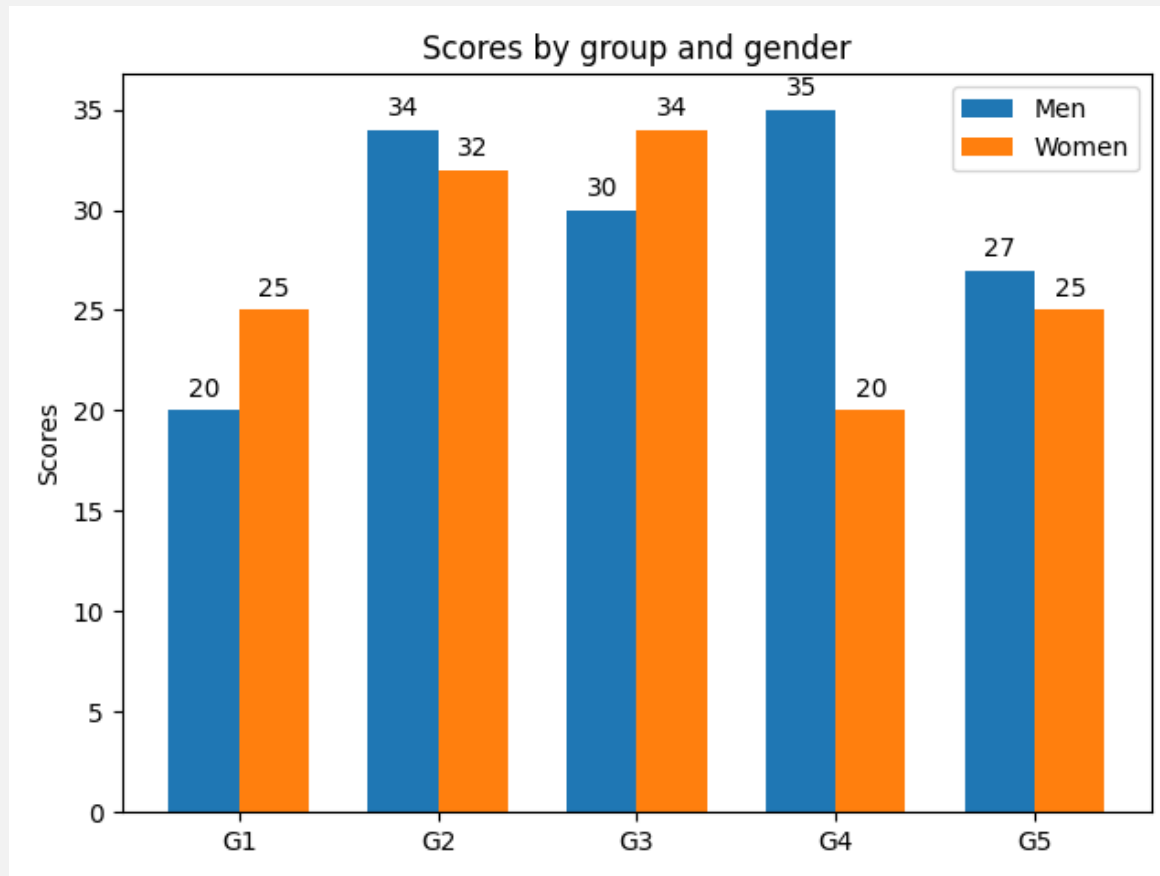
# Análise Bivariada

- Gráfico de correlação (Heatmap)



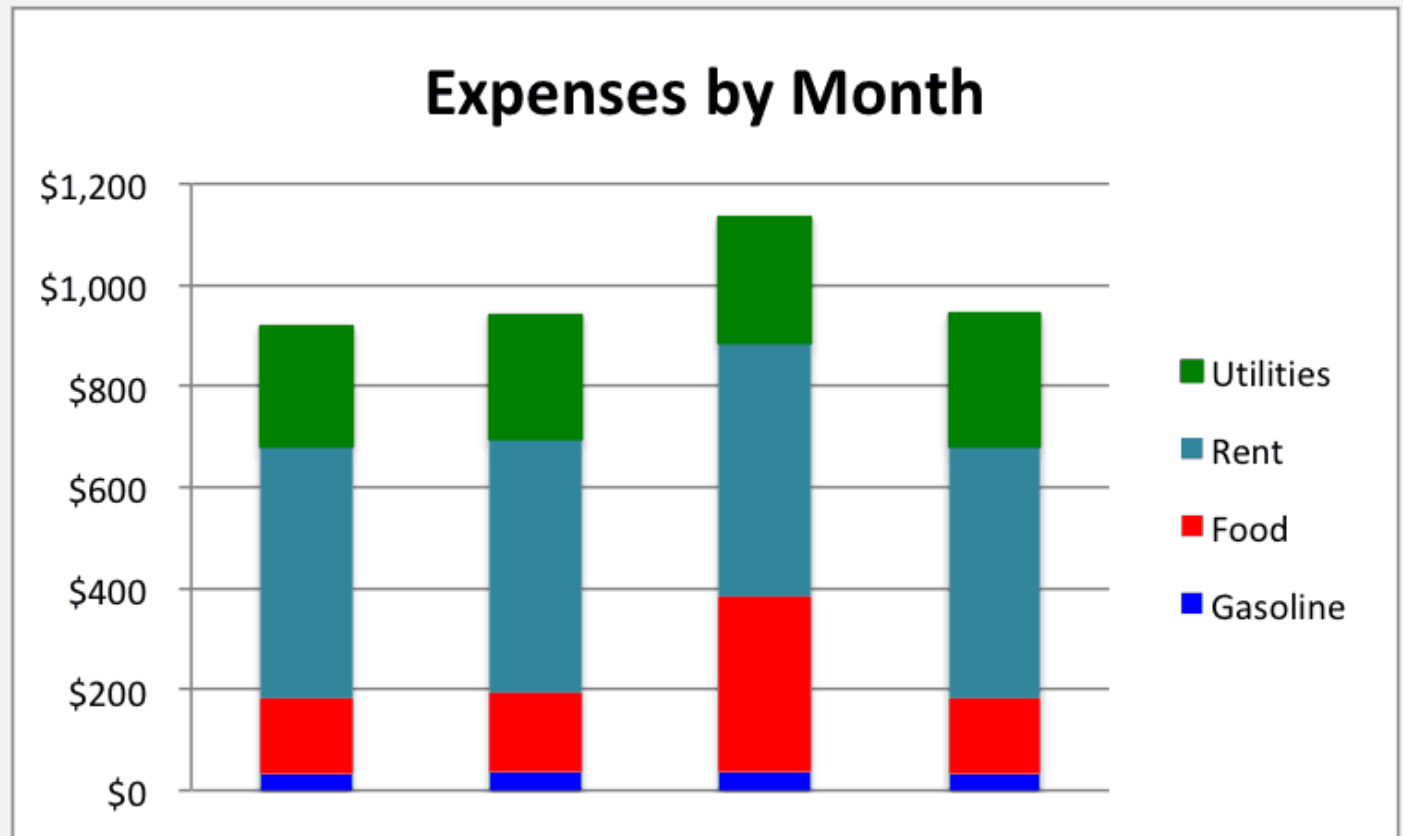
# Análise Bivariada

- Gráfico de barras e barras empilhadas



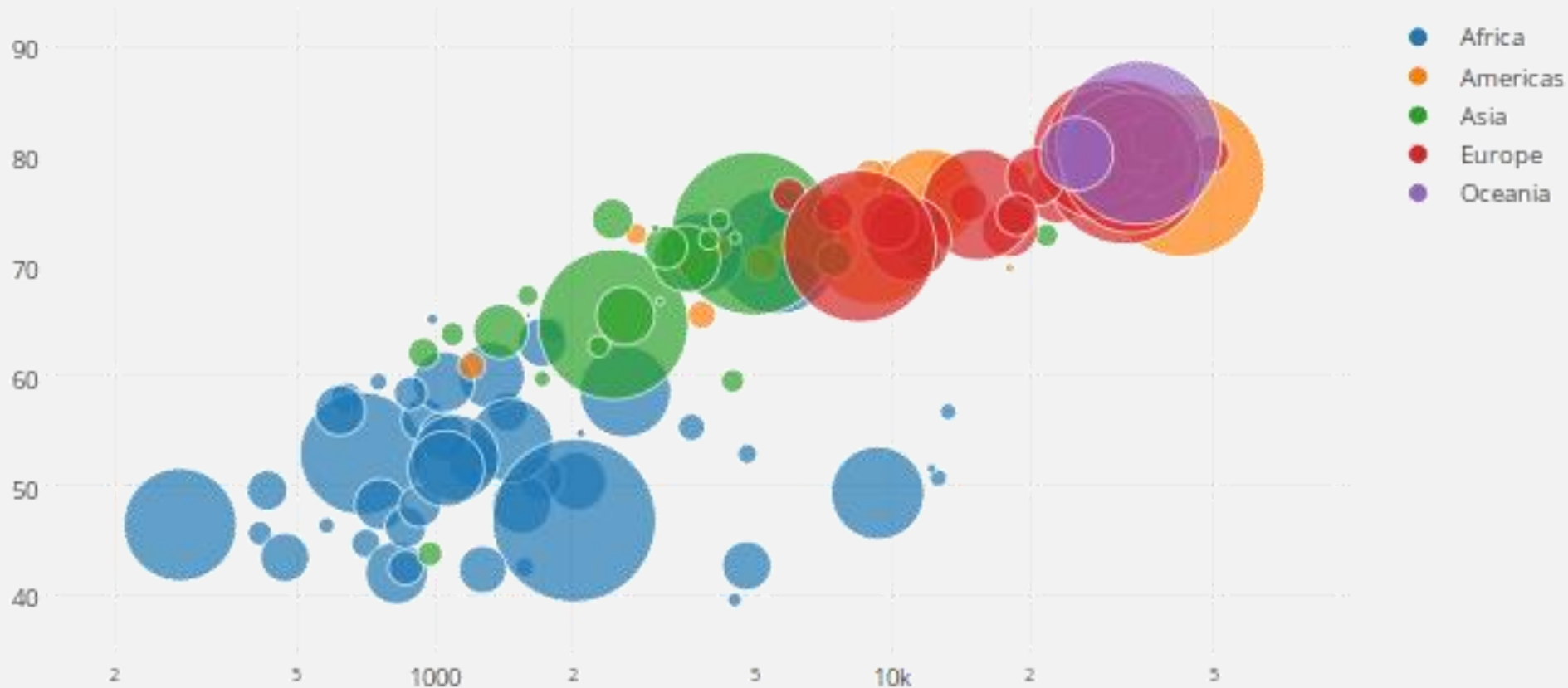
# Análise Bivariada

- Gráfico de barras e barras empilhadas
  - Atenção com o contexto!



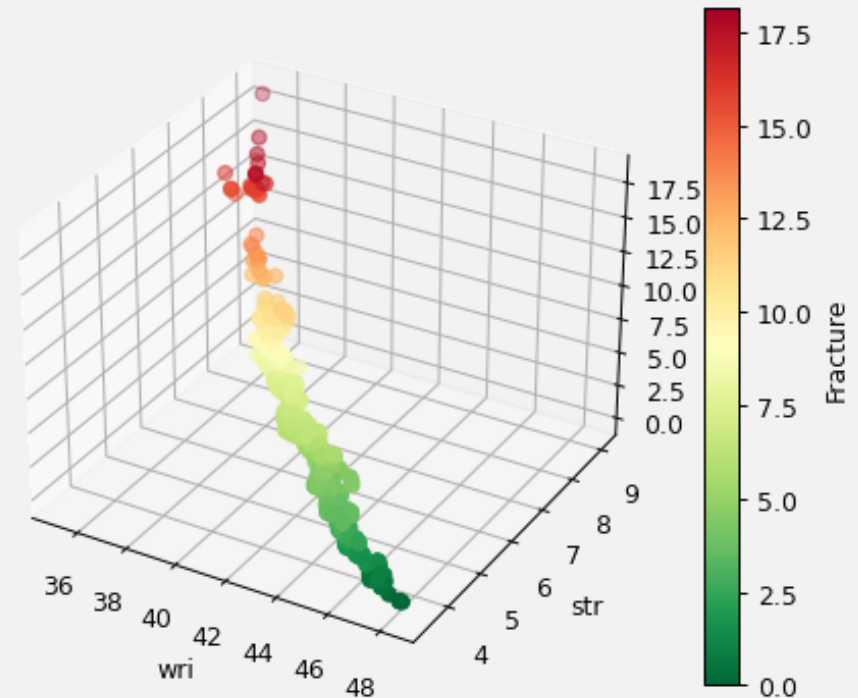
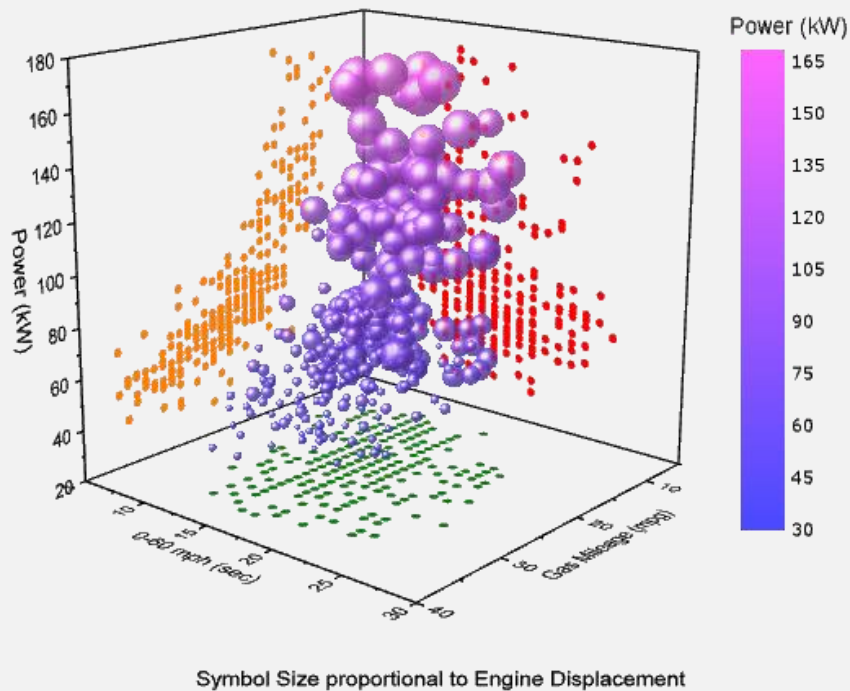
# Análise Multivariada

- Gráfico de Bolhas (Bubble plot)
  - É uma variação do Gráfico de Dispersão onde cores, símbolos e tamanho também representam variáveis.



# Análise Multivariada

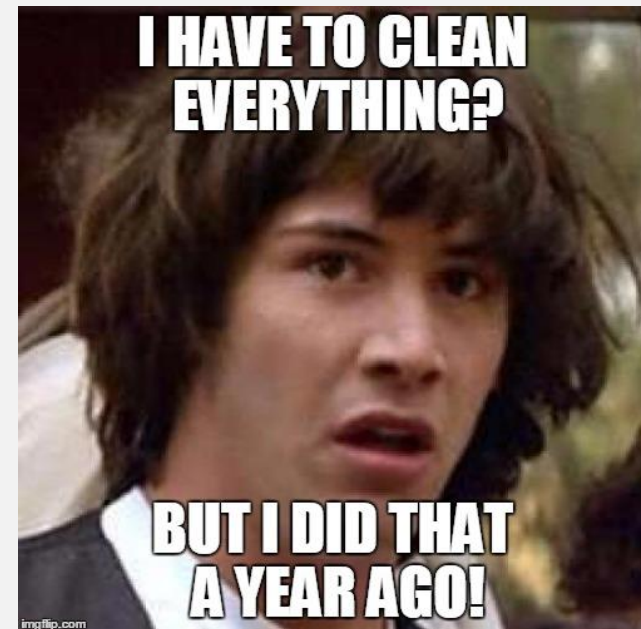
- Gráfico de Dispersão (Scatter plot 3D)








# ■ Limpeza de dados

- O que significa “limpar dados” ?
- Limpeza de dados (ou Data cleaning) consiste em detectar, remover OU corrigir dados corruptos, incorretos, ausentes ou irrelevantes.
- Limpeza de dados (Data cleaning) é muitas vezes confundido com Pré processamento de dados (Data preprocessing)
- Principais causas de sujeira:
  - Valores ausentes
  - Caracteres especiais
  - Inconsistências
  - Outliers



# Limpeza de dados

- Valores ausentes
  - Em Python normalmente são variáveis do tipo **None**.
  - Dependendo do módulo/biblioteca que estiver usando, pode admitir outros valores, como **NaN** ou **np.nan** (numpy)
  - Mas dependendo do caso pode ser valores “especiais”, exemplo:
    - “-” (string)
    - “empty” (string)

entree	pets	emergency_contact
		
shrimp		Pepper
beef		Jane
chicken	62	Janet
beef		Henry
		NA
veggie		n/a
chicken		None
shrimp	3	empty
shrimp		-
		""
veggie	1	
chicken		null

# Limpeza de dados

## O que fazer com valores ausentes?

- Deleta toda a linha ou faz uma imputação?
- Tipos de imputação
  - Hot-Deck
    - Substitui o dado ausente por algum outro dado do banco de dados
  - Cold-Deck
    - Substitui o dado ausente por um dado de OUTRO banco de dados
  - Imputação através da média
    - Calcula a média dos dados presentes e substitui nos dados ausentes
  - Imputação através da regressão
    - Cria uma regressão com base nos dados presentes para prever os dados ausentes

# Limpeza de dados

## Inconsistências

- Valores que são claramente incompatíveis com o resto dos dados ou que fogem do bom senso.
  - Ex: Altura de uma pessoa ser 10 metros
  - Ex:  $-10^{\circ}\text{C}$  em Fortaleza em pleno verão
  - Ex:  $0^{\circ}\text{C}$  em um forno que opera na faixa de  $500^{\circ}\text{C}$
  - Ex: \$1.000.000,00 na minha conta bancária

# Limpeza de dados

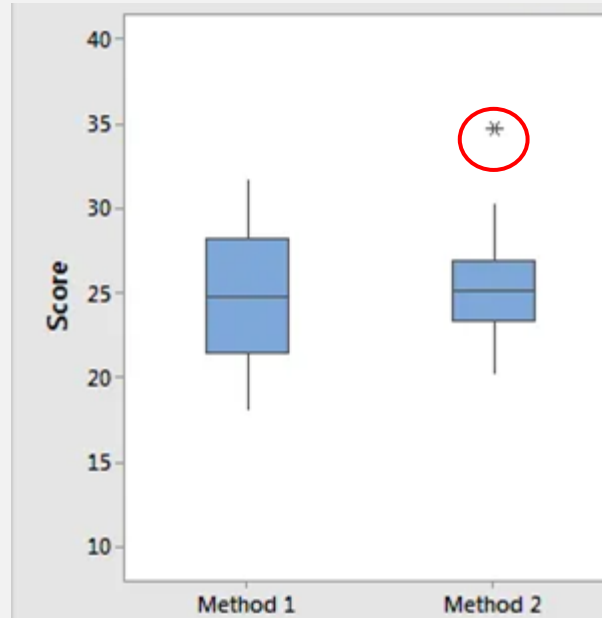
## Outliers

- Valores que são incompatíveis com o resto dos dados, mas que não necessariamente estão incorretos.
  - Ex: Altura de uma pessoa ser 2.3 metros
  - Ex: 15°C em Fortaleza em pleno verão
  - Ex: 300°C em um forno que opera na faixa de 500°C
  - Ex: A conta bancária do Bill Gates

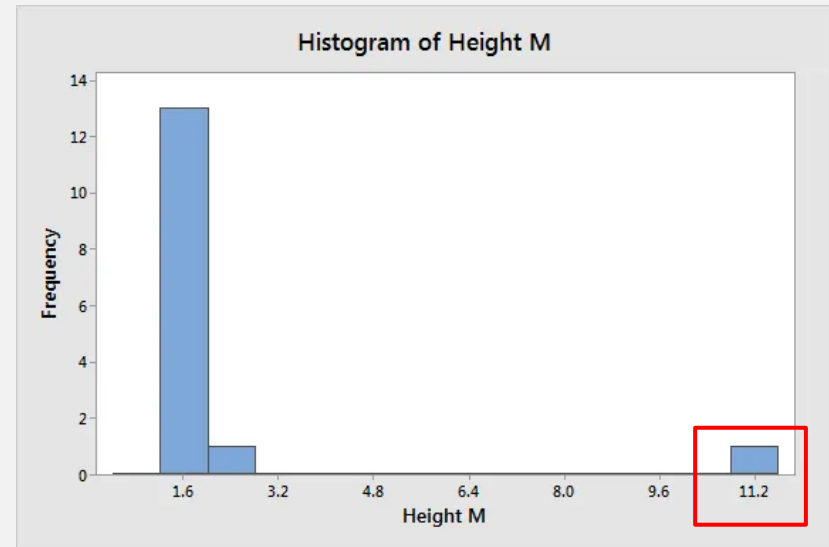
# Limpeza de dados

## Como detectar Outliers?

Height M
1.5895
1.6508
1.7131
1.7136
1.7212
1.7296
1.7343
1.7663
1.8018
1.8394
1.8869
1.9357
1.9482
2.1038
10.8135



Boxplot



Histogramas

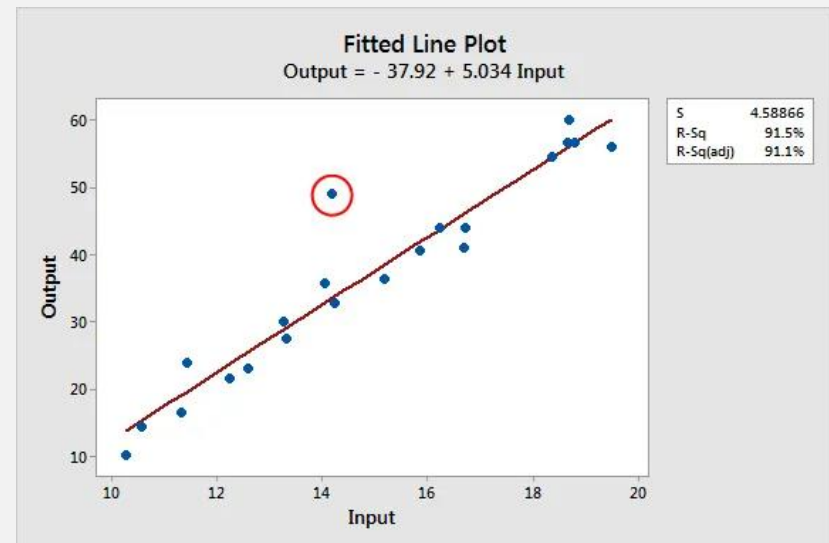


Gráfico de Dispersão

Ordenando dados

# Limpeza de dados

## Como detectar Outliers?

Se os dados seguirem uma distribuição normal, existe uma técnica que pode ser utilizada de forma QUANTITATIVA de forma robusta: Z-test ou T-test.



# Limpeza de dados

## Como detectar Outliers?

Consiste em aplicar uma fórmula para calcular o quão “dentro do esperado” é um valor em uma distribuição

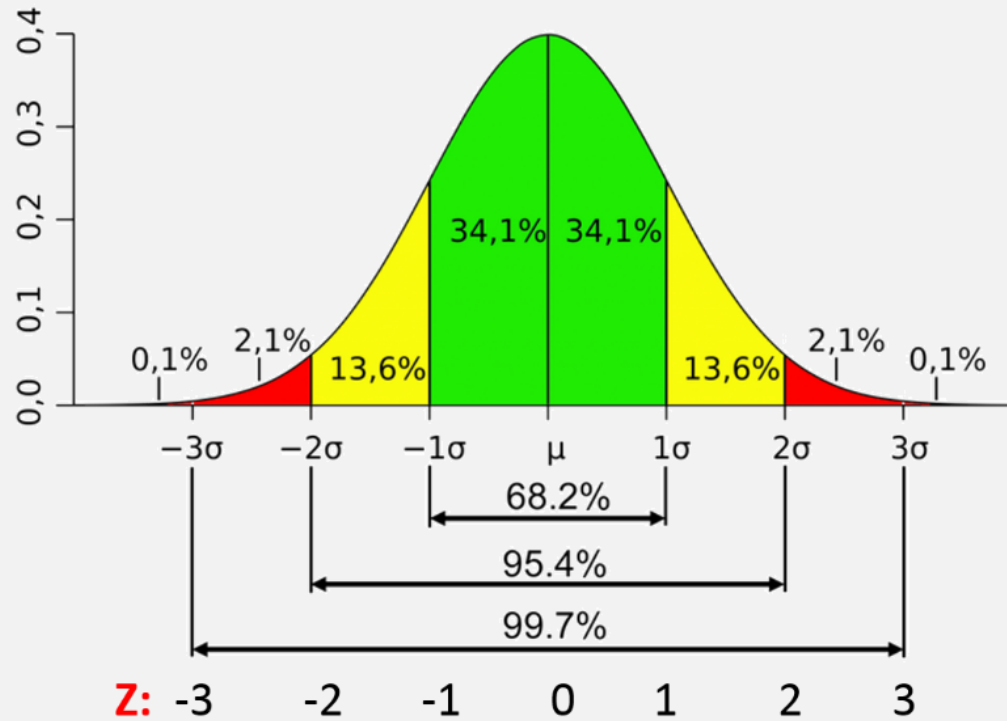
$$Z = \frac{X - \mu}{\sigma}$$

Onde:

X = valor

$\mu$  = média

$\sigma$  = desvio padrão



Height M	Z-score
1.5895	-0.34603
1.6508	-0.31975
1.7131	-0.29301
1.7136	-0.29283
1.7212	-0.28954
1.7296	-0.28595
1.7343	-0.28394
1.7663	-0.27020
1.8018	-0.25501
1.8394	-0.23888
1.8869	-0.21852
1.9357	-0.19757
1.9482	-0.19223
2.1038	-0.12551
10.8135	3.60910



# Trabalho Final (parte 1)

Análise Exploratória de Dados

Data de entrega do relatório final: 07/12

Dúvidas e envio do trabalho:

e-mail: [bernarddss62gmail.com](mailto:bernarddss62gmail.com)

# Referências

- Gunther, N. J. (2008). What the harmonic mean means.  
<http://perfdynamics.blogspot.com/2008/11/what-harmonic-mean-means.html>.
- Jain, R. (1991). The Art of Computer Systems Performance Analysis. Wiley.
- Jelihovschi, E. (2014). Análise Exploratória de Dados Usando o R. Editora da UESC. Disponível em  
[http://www.uesc.br/editora/livrosdigitais2/analiseexploratoria\\_r.pdf](http://www.uesc.br/editora/livrosdigitais2/analiseexploratoria_r.pdf).
- Kabacoff, R. I. (2015). R in Action, 2nd Ed. Manning.
- Lilja, D. J. (2000). Measuring Computer Performance: A Practitioner's Guide. Cambridge University Press.
- Montgomery, D. C. and Runger, G. C. (2012). Estatística Aplicada e Probabilidade para Engenheiros, 5a Ed. Livros Técnicos e Científicos.
- Patil, S. and Lilja, D. J. (2010). Using resampling techniques to compute confidence intervals for the harmonic mean of rate-based performance metrics. IEEE Computer Architecture Letters, 9(1):1–4.
- Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.
- Verzani, J. (2004). Using R for Introductory Statistics. Chapman and Hall/CRC