MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA:
Análise de outliers e dados ausentes

Professor:

Luis E. Zárate

Preparação do conjunto de dados

- Nesta etapa é feito uma análises de *Outliers* e de dados aussentes.

- *Outliers* são dados com padrões muito diferentes aos demais que fogem ao padrão dos dados. Estes dados precisam ser identificados e analisados, pois possivelmente trata-se de erros no banco de dados.

- Deve ser feita também uma análise de dados ausentes com o intuito de verificar o impacto que esta ocorrência terá na descoberta de conhecimento.

- Cabe ao especialista KDD definir uma estratégia para tratar *outliers* e os dados ausentes.

Análise de Outliers

Quando as instâncias ou valores de um atributo são significativamente diferentes ou inconsistentes, são chamados de outliers.

O que pode produzir outliers:

- Erros de medição;

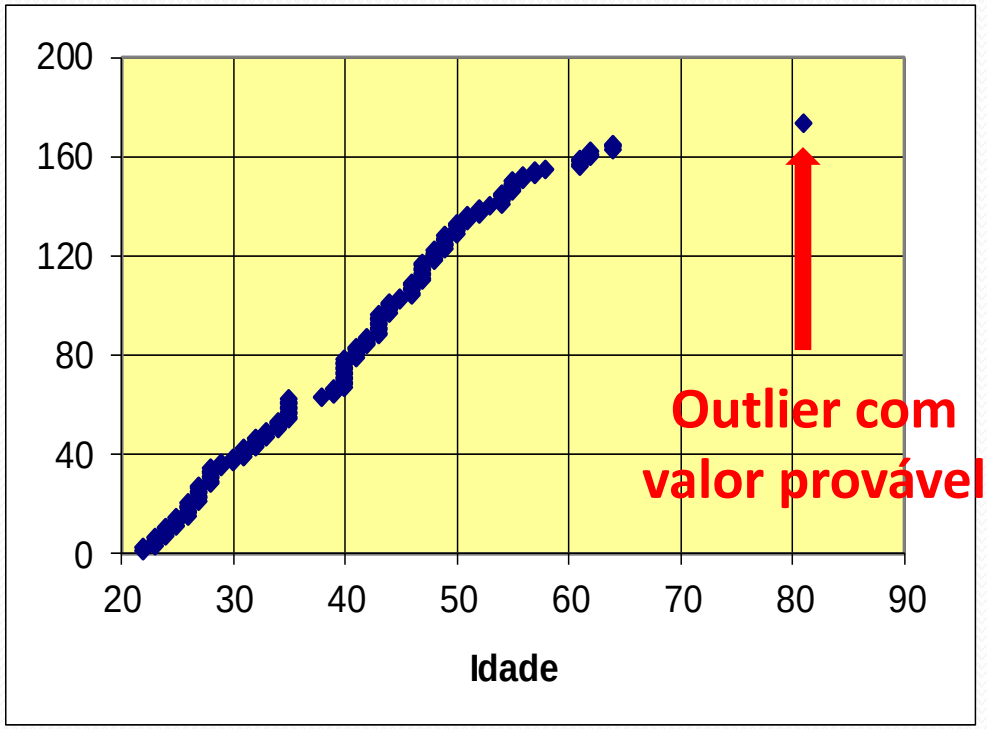
- Valores default assumidos durante o preenchimento de uma base de dados (para o campo salário o default pode ser 0,00)

- Podem corresponder a valores reais mas pertencentes a uma base de dados desbalanceada.

Análise de Outliers

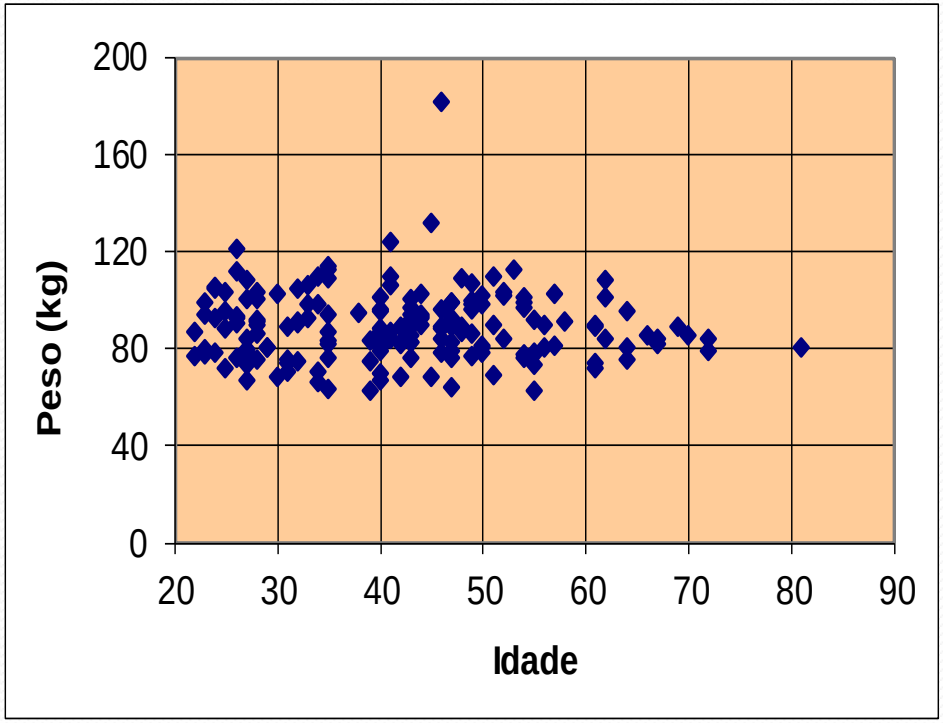
A detecção de outliers não é um processo trivial. Uma técnica utilizada é a inspeção visual, válida até 3 dimensões.

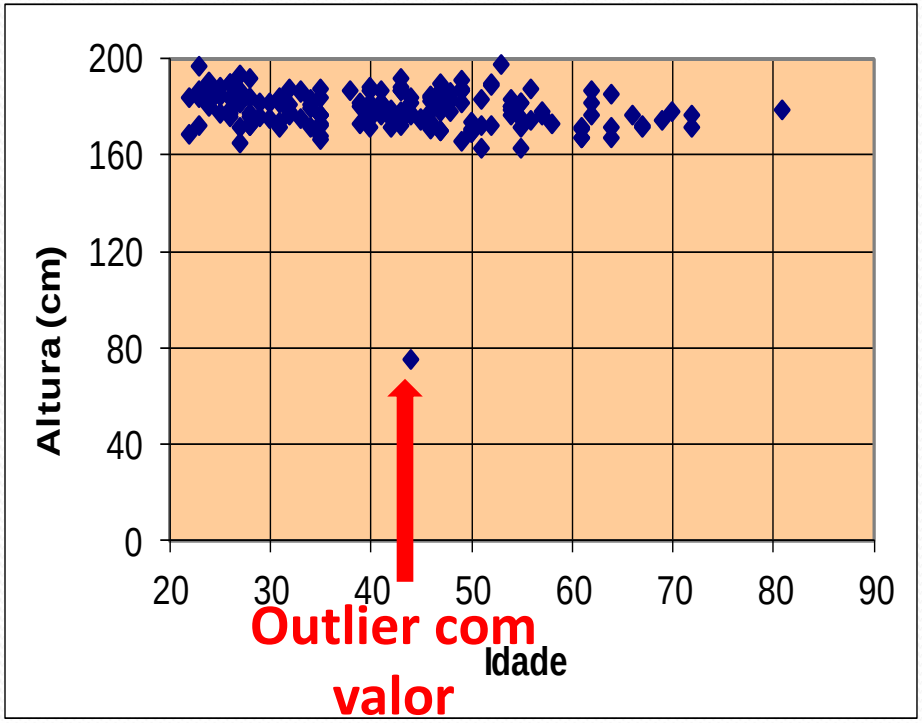
Por exemplo: considerando 173 registros de pessoas com idade mínima de 22 e idade máxima de 81 anos.



Análise de Outliers

Por exemplo: considerando 173 registros de pessoas com peso mínimo de 50, peso máximo de 185 kg. e altura mínima de 78 e máxima de 198 cm.





Análise de Outliers

1) Método estatístico para remover outliers:

Limiar = Média $\pm$ 3 * Desvio padrão

Para o atributo <idade>: 42,26 $\pm$18,75

Para o atributo <peso>: 88,97 $\pm$21,74

Para o atributo <altura>: 178,09 $\pm$15,60

Dados pouco dispersos

Intervalo	Probabilidade	
	Interna	Externa
$\mu \pm 1\sigma$	68,2%	31,74%
$\mu \pm 2\sigma$	95,46%	4,54%
$\mu \pm 3\sigma$	99,73%	0,27%

Limiar = Média $\pm$ 2 * Desvio padrão

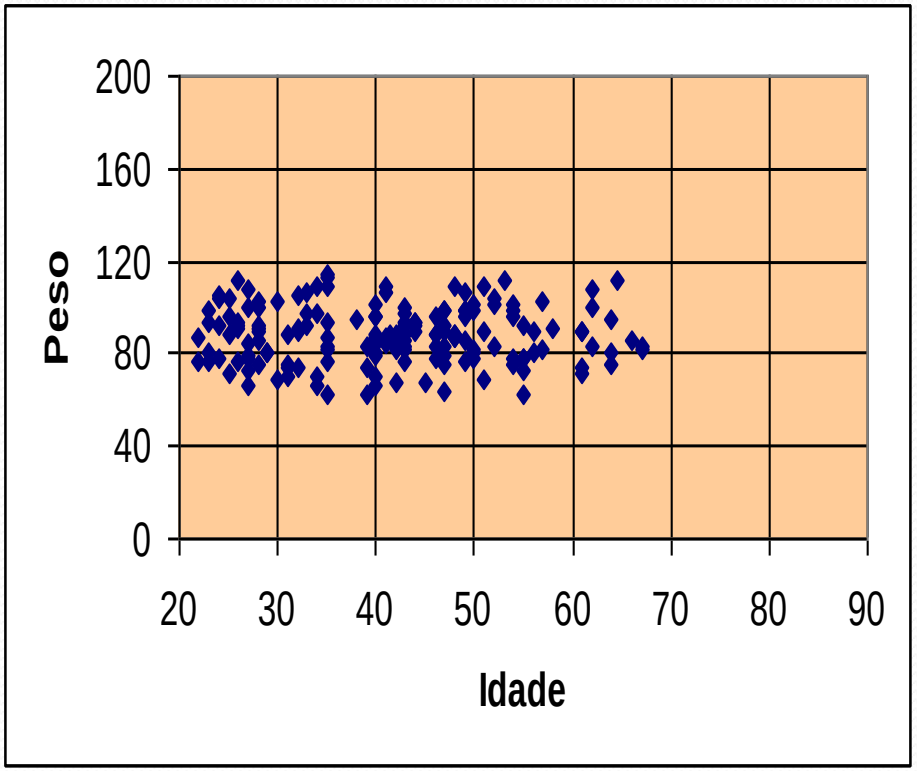
Para o atributo <idade>: 42,26 $\pm$12,51

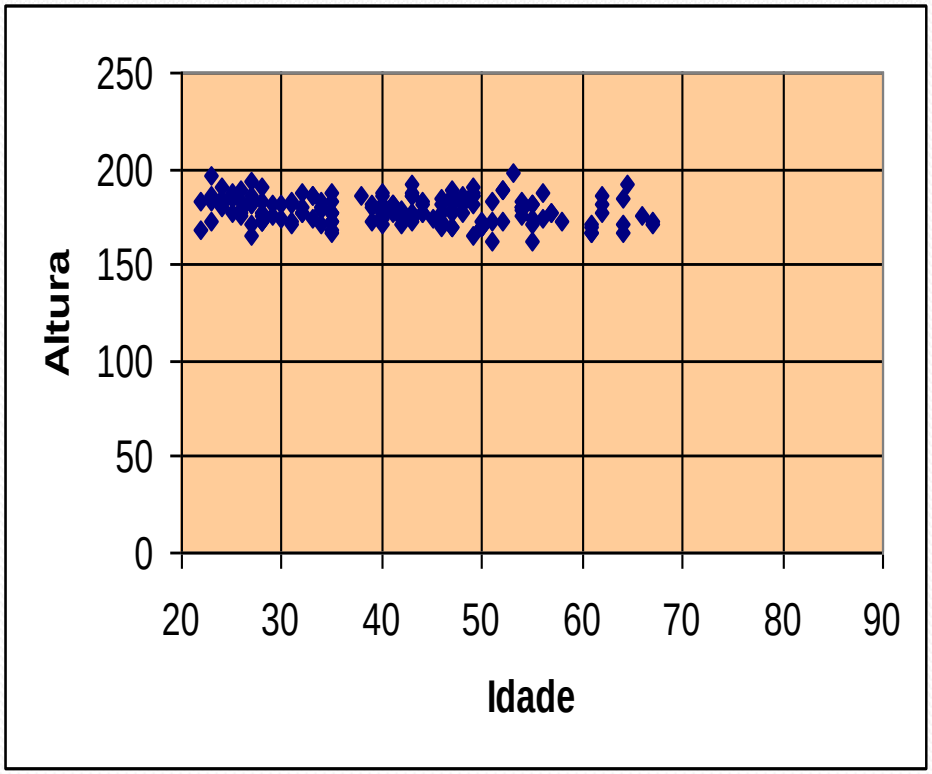
Para o atributo <peso>: 88,97 $\pm$14,57

Para o atributo <altura>: 178,09 $\pm$10,40

Dados muito dispersos

Análise de Outliers





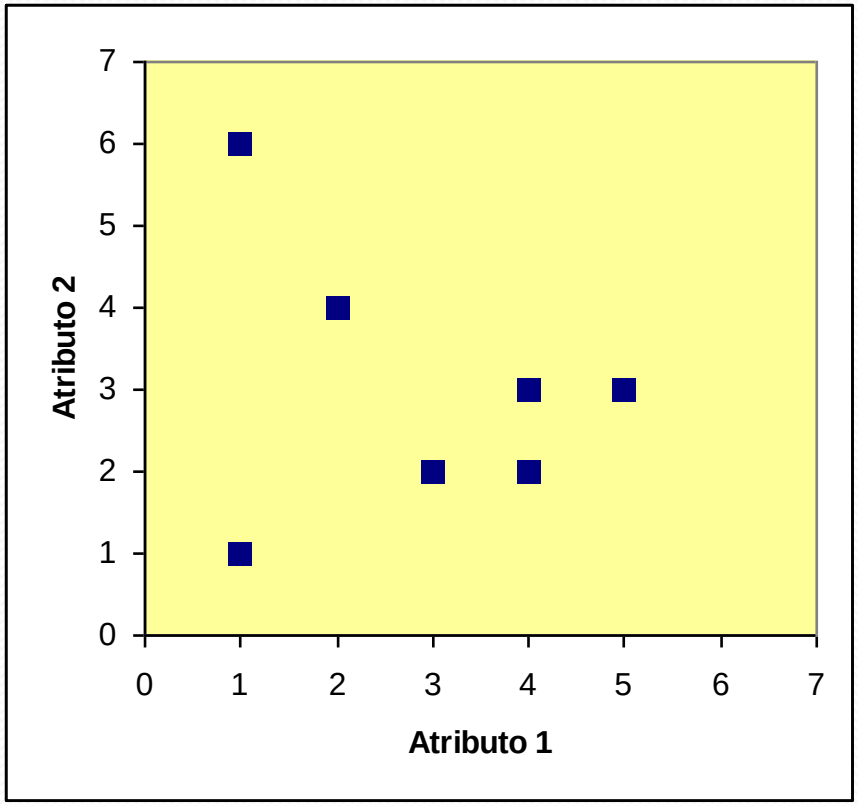
Análise de Outliers

2) Método da distância global na análise mutivariável:

Seja o conjunto "P" de amostras de duas dimensões:

P = {p1,p2,...,p7}

Atrib 1	Atrib 2
2	4
3	2
1	1
4	3
1	6
5	3
4	2



Análise de Outliers

A matriz das distâncias Euclideanas entre as amostras é mostrada a seguir:

	p1	p2	p3	p4	p5	p6	p7
p1	0	2,236	3,162	2,236	2,236	3,162	2,828
p2		0	2,236	1,414	4,472	2,236	1,000
p3			0	3,605	5,000	4,472	3,162
p4				0	4,242	1,000	1,000
p5					0	5,000	5,000
p6						0	1,414
p7							0

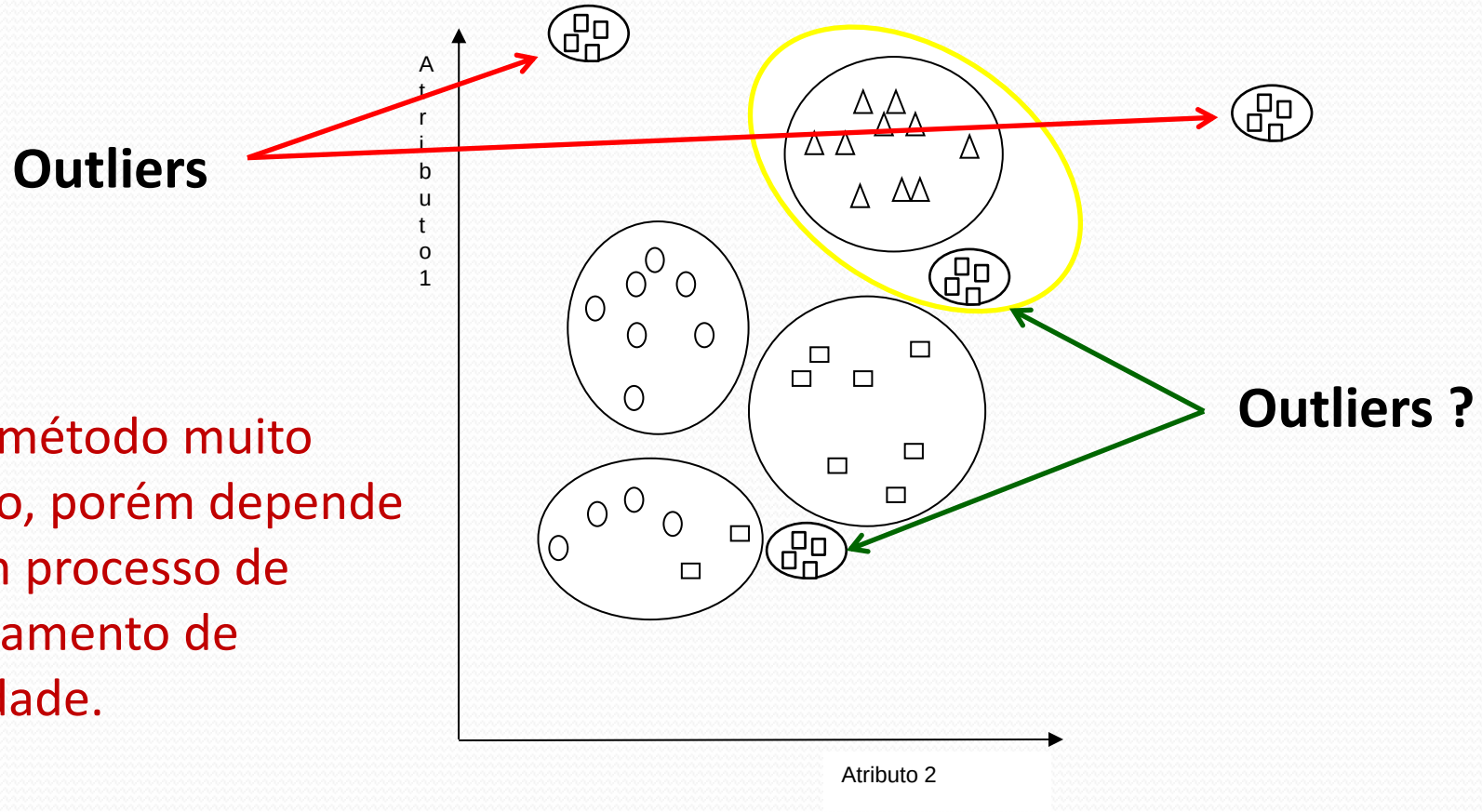
Amostras com distância > 3

Amostra	f
p1	2
p2	1
p3	**5**
p4	2
p5	**5**
p6	3
p7	2

p3 e **p5** podem ser considerados **outliers**

Análise de Outliers

3) Método da clusterização (análise mutivariável):



É um método muito efetivo, porém depende de um processo de agrupamento de qualidade.

Análise de dados ausentes

- **Variáveis com Valores Ausentes ou Vazios**
  - **Valor Ausente:** é aquele valor (possivelmente medido) que não foi inserido no conjunto de dados, mas seu valor atual existe no mundo.
  - **Valor Vazio:** é aquele valor que nenhum valor do mundo real pode ser suposto. Se estes existem podem distorcer os resultados.

Tratamento de valores Ausentes

- **Três pontos são observados quando se lida com dados ausentes:**

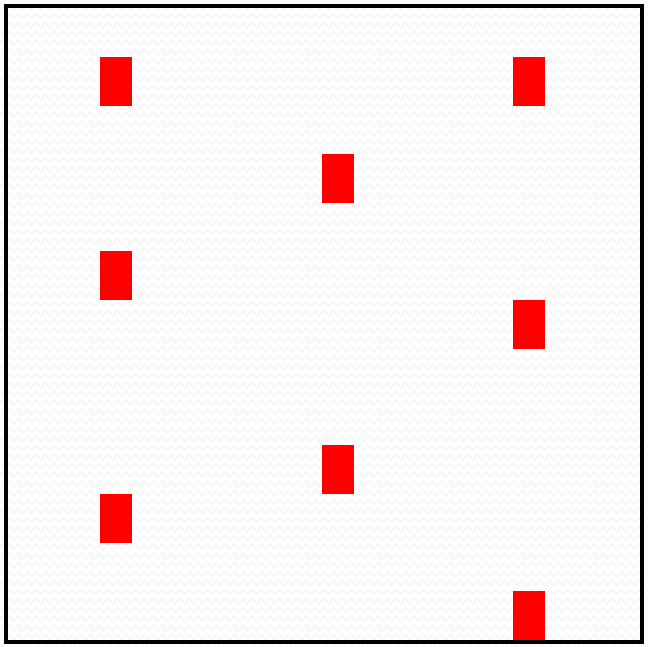
a) **A decisão pela eliminação ou não do atributo ou do registro, que contêm valores ausentes;**

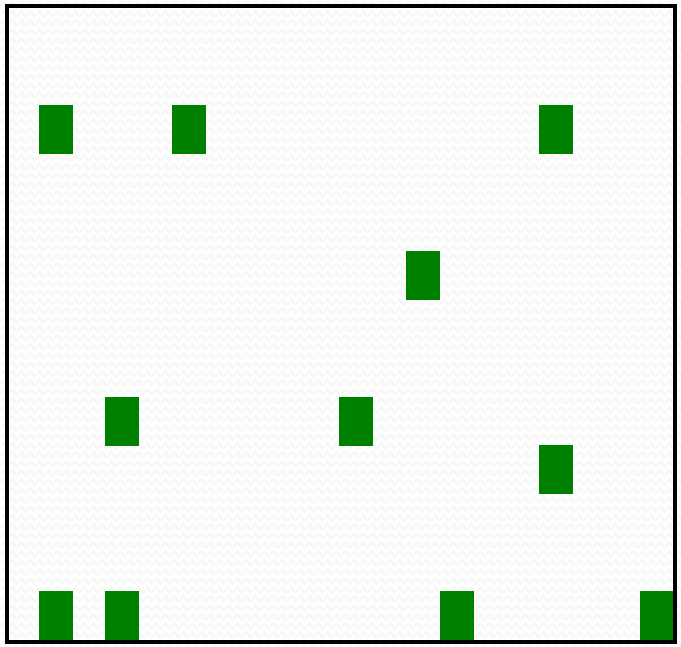
b) **A recuperação dos valores ausentes; e**

c) **Quais técnicas de Data Mining lidam melhor com valores ausentes e em que graus.**

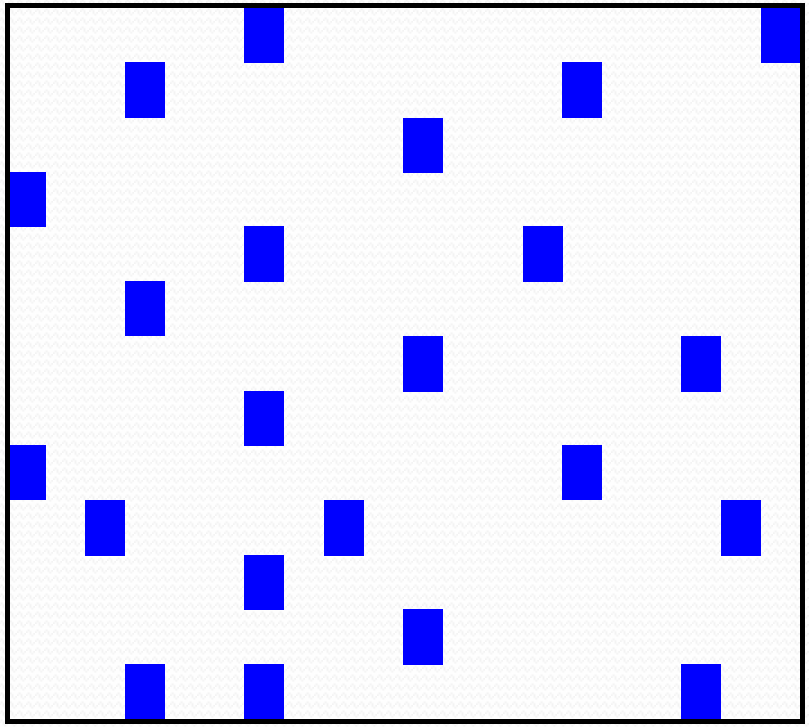
Tipos de dados ausentes

A presença de valores ausentes em uma base de dados é um fato comum podendo estar distribuído em diversos atributos, numa mesma instância (registro) ou de forma aleatória.

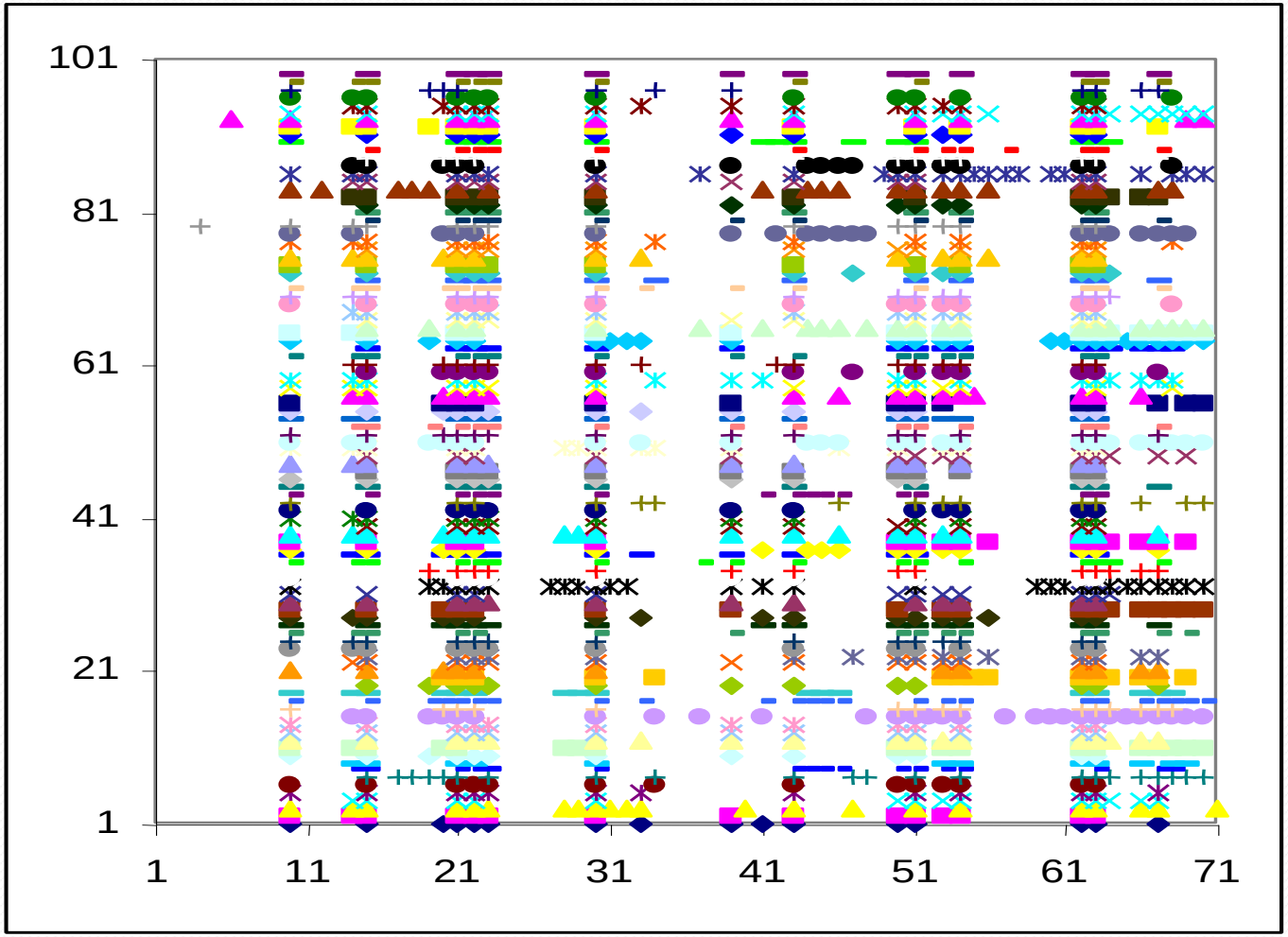




Valores Ausentes - aleatórios



Massivos dados ausentes



Mecanismos de ausência

- Uma importante contribuição, proposta em (RUBIN, 1976) foi a classificação dos mecanismos de ausência
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR)
- Esta classificação é baseada nas condições nas quais os valores ausentes foram produzidos

- MCAR: os valores ausentes estão distribuídos aleatoriamente, ou seja, a probabilidade de encontrar um valor ausente é a mesma para qualquer valor do atributo
- MAR: a probabilidade de encontrar um valor ausente depende de outro valor de outro atributo
- NMAR: a probabilidade de encontrar um valor ausente depende do próprio valor do atributo que possui dado ausente.

Mecanismos de ausência

- De uma forma geral, não se pode afirmar que o mecanismo de ausência é MCAR, MAR ou NMAR.
- Não há técnicas para garantir qual o mecanismo de ausência de um determinado conjunto de dados.
- Os mecanismos de ausência se confundem em determinadas situações

MCAR - aleatório

- Mesmo que o percentual de ausência seja massivo (95%), a amostra mantém suas características originais

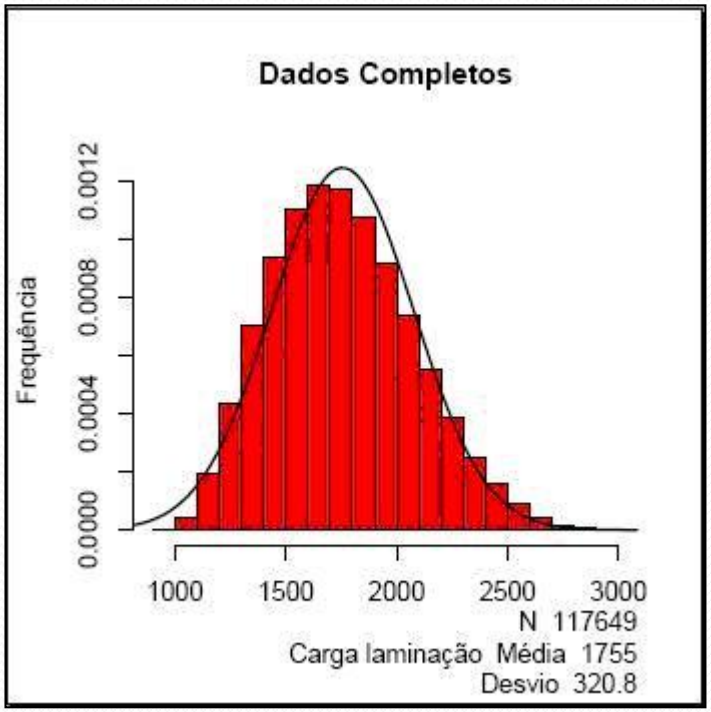


Gráfico 1 – Dados Completos

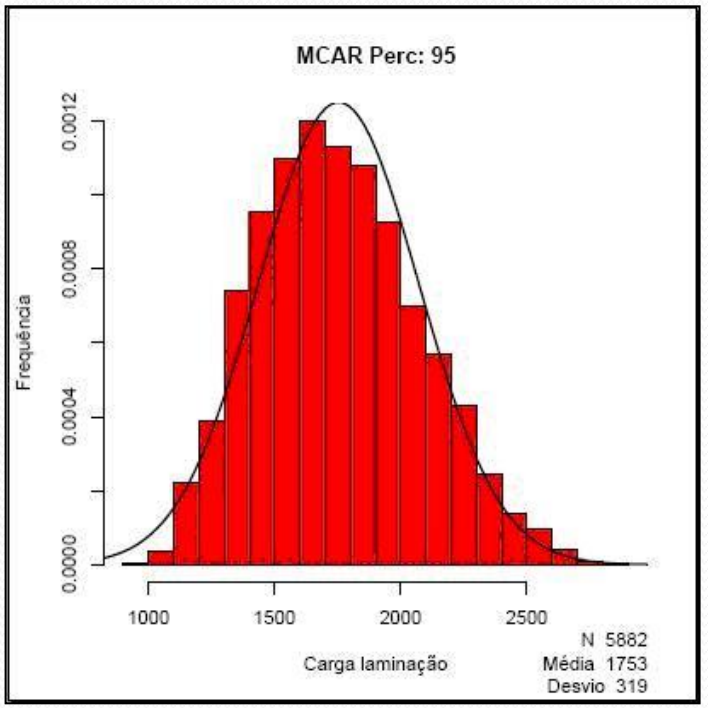


Gráfico 2 – MCAR 95% ausência

NMAR – depende de mesmo atributo

- Os dados ausentes foram gerados através de condições impostas a própria variável. Nota-se grande variaçãos nas características da amostra

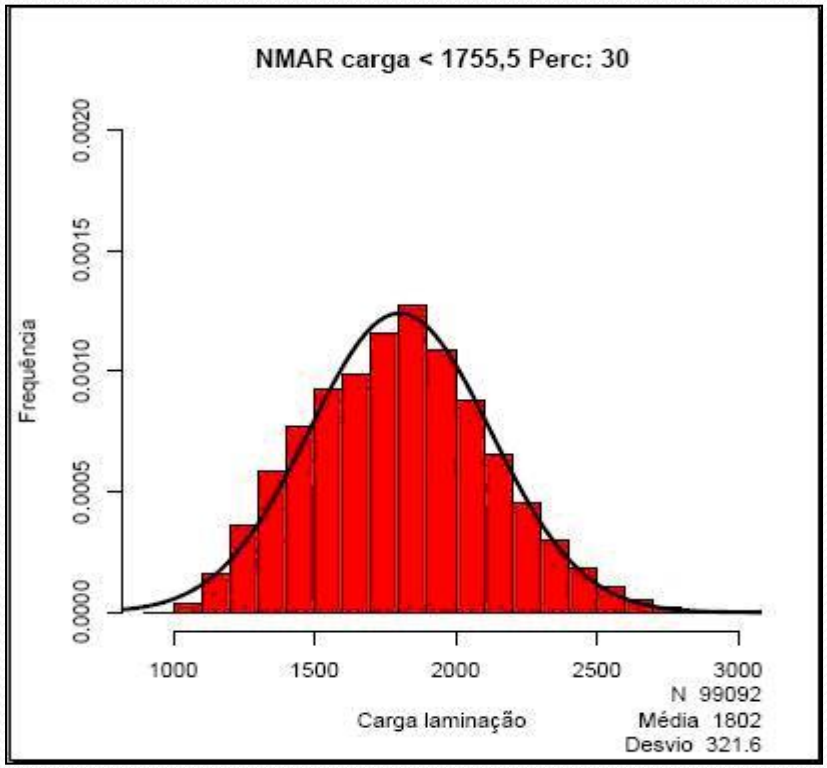


Gráfico 5 –NMAR condição carga < 1755,5
30% ausência

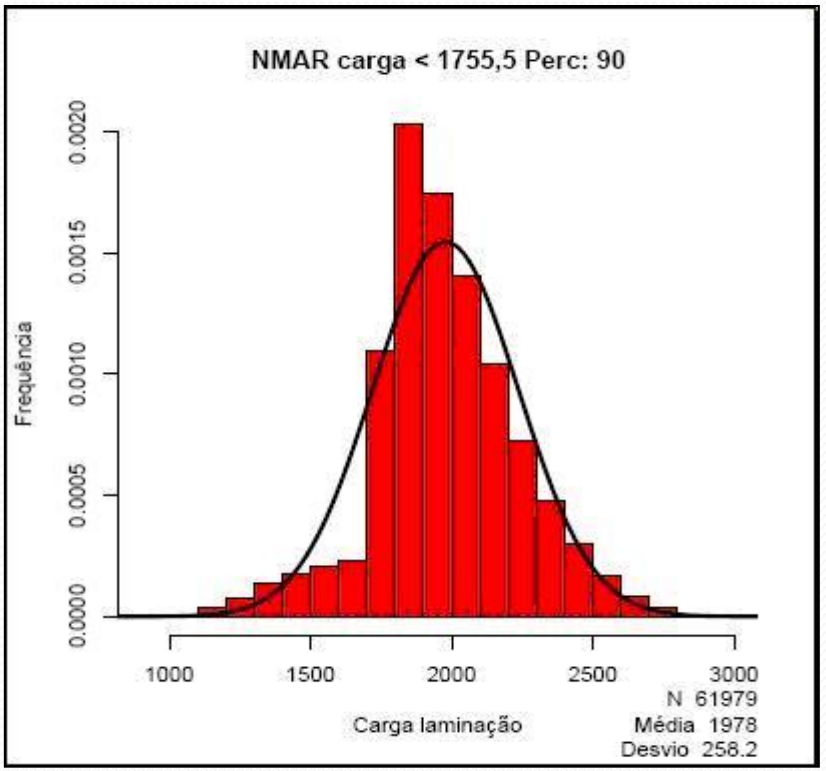


Gráfico 6 –NMAR condição carga < 1755,5
90% ausência

Valores Ausentes - comentários

1) A identificação do mecanismo de ausência não é uma tarefa trivial. Geralmente, o dado contém pouca informação que auxilia na identificação do mecanismo de ausência.

2) Eliminação de valores ausentes com o mecanismo NMAR (não ignorável) causa severas distorções nos resultados.

3) Na literatura muitos métodos de tratamento de dados ausentes são aplicáveis ao mecanismo MCAR ou MAR.

Valores Ausentes - comentários

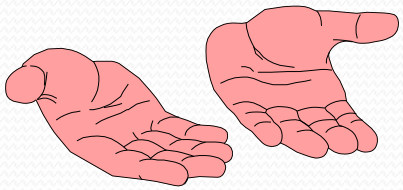
4) Um procedimento comum, para lidar com dados ausentes, consiste em eliminar o(s) atributo(s) ou a(s) instância(s) da base de dados, que apresentam esses valores, impondo desta forma restrições ao conhecimento extraído.

5) Outros procedimentos sugerem a substituição de valores ausentes por valores padrões ou valores médios em todas as ocorrências.

Valores Ausentes - comentários

6) A eliminação de instâncias e/ou atributos pode acarretar também a perda de informações importantes relativos aos valores que estão presentes.

7) A substituição por valor padrão, mesmo o mais criterioso, pode introduzir na base informações distorcidas, que não estão contidas no evento e nas circunstâncias que a gerou.



Valores Ausentes - comentários

8) A recuperação de dados ausentes torna-se, então, um ponto de extrema importância na descoberta de conhecimento em base de dados, requerendo predições cuidadosas dos valores, utilizando técnicas mais avançadas e elaboradas, além do conhecimento tácito de um especialista no domínio do problema.

9) Técnicas de Data Mining como: classificador por vizinho mais próximo nearest neighbor, classificadores bayesianos e diversas técnicas estatísticas, não conseguem lidar com valores ausentes, tornando seu uso inviável para determinadas bases de dados.

Valores Ausentes - comentários

10) Técnicas da Data Mining, como árvore de decisão podem lidar com bases de dados contendo pequeno número de valores ausentes.

11) Técnicas de aprendizado de máquinas como Redes Neurais Artificiais (RNA) conseguem aprender relações entre variáveis a partir das instâncias que lhe são mostradas e daí recuperar dados ausentes.

Obrigado

Professor:

Luis E. Zárate