Data Mining
Seleção e Processamento

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2023

www.inf.ufsm.br/~joaquim

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

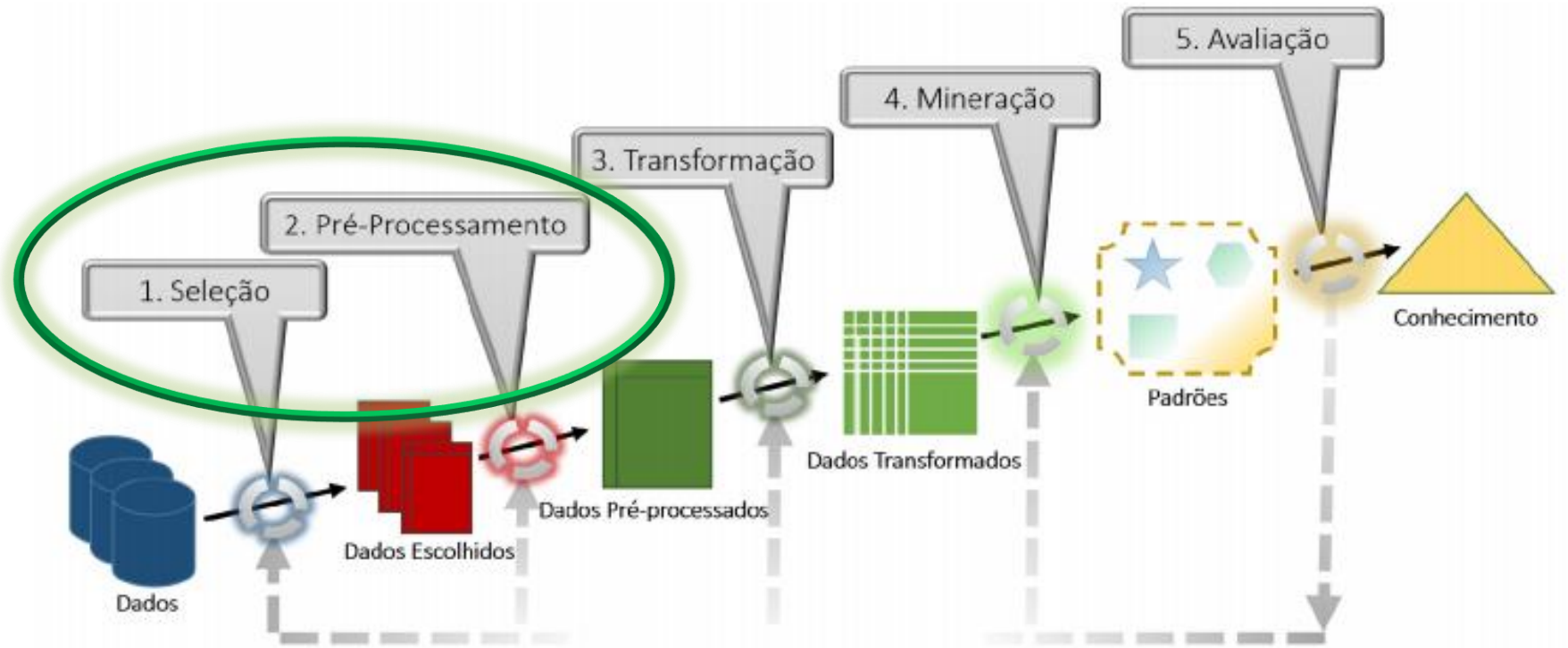
Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

*A maior parte deste material foi retirado do livro: "Joaquim V. C. Assunção. **Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R**. 192 páginas. Novatec. 2021. ISBN-10 : 6586057507."

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Seleção e processamento



* Figura mostrando o processo de KDD, proposto por Fayyad, 1996.

Seleção

- A seleção inicial dos dados é requerida onde grandes volumes de dados são armazenados.
- Se estes dados representam diferentes contextos, então uma prévia seleção se faz necessária.
- Isto geralmente é feito por meio de consultas SQL ou métodos de extração de dados em bancos noSQL.
- A seleção ajuda na próxima etapa, principalmente na qualidade dos dados.

Seleção

Na etapa de Seleção de dados está um grande conjunto de técnicas que não iremos explorar. Dentre elas:

I. SQL
   - Comandos DDL e DML,
   - Stored procedures,
   - triggers, etc.

II. Tecnologias de extração NoSQL

III. Programas especializados

IV. Sensores

V. Equipamentos IoT

VI. Etc.

Dados

“Os dados são valores de variáveis qualitativas ou quantitativas, pertencentes a um conjunto de itens”

→Conjunto de itens, também conhecidos como população; é o conjunto de objetos nos quais você está interessado.

→A população de dados pode ser dividida em dois tipos quanto ao refinamento, dados brutos e dados processados.

Dados brutos

- A fonte original dos dados;
- Frequentemente difícil de usar para análises de dados;
- A análise de dados, geralmente, inclui processamento.

Dados processados

- Dados que estão prontos para análise;
- O processamento pode incluir mesclagem, subconjunto, transformação etc.
- Pode haver padrões para processamento;
- Todas as etapas devem ser registradas.

Diretório de trabalho

- Um componente básico do trabalho com dados é conhecer seu diretório de trabalho.
- Em R, os dois principais comandos são `getwd()` e `setwd()`.

Diretório de trabalho

Esteja ciente dos caminhos relativos versus absolutos!

- Relativo - `setwd(".\data")`, `setwd("..\")`
- Absoluto - `setwd(" C:\Users\joaquim\DMining\")`
- Linux ou Mac use `/`

Verificando diretórios

`file.exists("nomeDiretorio")` verifica a existência de um diretório

`dir.create("nomeDiretorio")` irá criar um diretório

Baixando dados da internet

`download.file()` pode ser usado para fazer o download de um arquivo da internet

Parâmetros importantes: `url`, `destfile`, `method`

Útil para baixar arquivos delimitados por tabulação, csv e outros

Baixando dados da internet

```
arqUrl <- "http://www-
usr.inf.ufsm.br/~joaquim/UFSM/DM/ds/usr_data.c
sv"

download.file(arqUrl,destfile="./test.csv",
method="curl")
```

Considerações sobre seleção de dados

- Linguagens como R ou Python possuem diversidade de métodos nativos para selecionar dados e manter o código fácil de ser reproduzido.
- Porém, em diversos casos, podemos usar ferramentas externas para obter os dados de uma determinada fonte.

Considerações sobre seleção de dados

Podemos usar pacotes externos para coletar dados de uma determinada fonte.

Para compatibilidade entre ferramentas de mineração, podemos usar o pacote *foreign** que possui métodos de leitura e escrita para os vários formatos.

https://cran.r-project.org/web/packages/foreign/foreign.pdf

Considerações sobre seleção de dados

Podemos usar bibliotecas externas para coletar dados de uma determinada fonte.

Para compatibilidade entre ferramentas de mineração, podemos usar a biblioteca *foreign** que possui métodos de leitura e escrita para os vários formatos.

- **read.arff (Weka)**
- read.dta (Stata)
- read.mtp (Minitab)
- read.octave (Octave)
- read.spss (SPSS)
- read.xport (SAS)

https://cran.r-project.org/web/packages/foreign/foreign.pdf

Considerações sobre seleção de dados

R também possui pacotes de conexão e escrita SQL.

Alguns *providers:*

- PostgreQL,
- MySQL,
- Microsoft Access and
- SQLite
- MongoDb

https://cran.r-project.org/web/packages/foreign/foreign.pdf

Considerações sobre seleção de dados

R também possui bibliotecas para conexão com diversos formatos de dados, por exemplo:

- jpeg - http://cran.r-project.org/web/packages/jpeg/index.html
- readbitmap - http://cran.r-project.org/web/packages/readbitmap/index.html
- png - http://cran.r-project.org/web/packages/png/index.html
- EBImage (Bioconductor) - http://www.bioconductor.org/packages/2.13/bioc/html/EBImage.html
- raster - http://cran.r-project.org/web/packages/raster/index.html
- tuneR - http://cran.r-project.org/web/packages/tuneR/

Pré-processamento

- O pré-processamento pode ser dividido em Limpeza, Integração e Seleção.

Limpeza dos dados

Limpeza dos dados

- A limpeza de dados interfere na qualidade dos dados. Dados de qualidade devem possuir as seguintes características.
- Precisão (os dados são registrados corretamente)
- Integralidade (todos os dados relevantes são registrados)
- Exclusividade (sem registro de dados duplicados)
- Oportunidade (os dados não são antigos)
- Consistência (os dados são coerentes)

Limpeza dos dados

- A limpeza de dados tenta preencher os valores ausentes, suavizar o ruído ao identificar valores discrepantes e corrigir inconsistências nos dados.
- A limpeza de dados é geralmente um processo iterativo de duas etapas que consiste em detecção de discrepância e transformação de dados.

Limpeza dos dados

- O processo de limpeza dos dados, na maioria dos casos, é dividido em duas partes. São elas:

1. Verificar o conjunto de dados de origem para encontrar a discrepância.
2. Retirar ou alterar os dados discrepantes de modo com que estes dados possam ser usados com a qualidade esperada

Limpeza dos dados – Valores faltantes

- Considere um vetor $\boldsymbol{x}$ denotando o conjunto de salários de um grupo de funcionários de um mesmo setor; $\boldsymbol{x} = \{\mathbf{1200}, \mathbf{1400}, \mathbf{1200}, \mathbf{1500}, \mathbf{2100}, \mathbf{1200}, \mathbf{1400}, \boldsymbol{NA}, \mathbf{1500}\}$. Onde, *NA* é um valor faltante.
- Como você faria para preencher o dado deste vetor?
  - Pesquisar pelo funcionário e perguntar o salário do mesmo.
  - Usar 0 como valor.
  - Usar o valor mínimo ou máximo do set.
  - Usar a média para preencher o valor.
  - Usar um algoritmo que use as demais variáveis para prever o valor.

Limpeza dos dados – Valores faltantes

- Pesquisar pelo funcionário e perguntar o salário do mesmo.

  Valor exato, mas problemas para coleta como tempo, recursos, agendas, constrangimento, etc.

- Usar 0 como valor.

  Está errado neste contexto.

- Usar o valor mínimo ou máximo do set.

  Possivelmente errará com grande margem $\Delta\,(min, max)$

- Usar a média para preencher o valor.

  Fácil execução e garante baixo erro potencial.

Limpeza dos dados – Valores faltantes

- Usar um algoritmo que use as demais variáveis para prever o valor.

  Um algoritmo simples pode ser útil para inferir o valor. Se o cargo não for suficiente e a categoria não estiver disponível, talvez dados tais como tempo de contrato, alocação na empresa podem ajudar.

Limpeza dos dados – Valores faltantes

- Usar um algoritmo que use as demais variáveis para prever o valor.

  Um algoritmo simples pode ser útil para inferir o valor. Se o cargo não for suficiente e a categoria não estiver disponível, talvez dados tais como tempo de contrato, alocação na empresa podem ajudar.

  Se mesmo estes não estiverem disponíveis, podemos obter a correlacão com os demais funcionários.

  Outras técnicas mais avançadas podem ser usadas: Regressão, Classificação, etc., mas o esforço é bem maior do que uma simples média.

Limpeza dos dados – Valores faltantes

- Usar um algoritmo que use as demais variáveis para prever o valor.

  Um algoritmo simples pode ser útil para inferir o valor. Se o cargo não for suficiente e a categoria não estiver disponível, talvez dados tais como tempo de contrato, alocação na empresa podem ajudar.

  Se mesmo estes não estiverem disponíveis, podemos obter a correlacão com os demais funcionários.

- Outras técnicas mais avançadas podem ser usadas: Regressão, Classificação, etc., mas o esforço é bem maior do que uma simples média.

Technical help

- Em R, poderíamos usar: `mean(x, na.rm = TRUE)`. Onde, `na.rm` garante que não a média irá ignorar o *NA*
- Também podemos usar `complete.cases`() para verificar quais dados estão faltando.
- Usando a negação de `complete.cases`() como index podemos assimilar a média automaticamente.
  `x[!complete.cases(x)] <- mean(x, na.rm = TRUE)`

Technical help

- Em R, poderíamos usar as funções: `min(x) e max(x)` para obter o mínimo e o máximo de um vetor.
- A função `summary` retorna algumas estatísticas interessantes. Podemos obter algo específico usando o índice do valor.

```
summary(x)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 1200    1200    1400    1438    1500    2100       1
```

```
> summary(x)[3]
Median
  1400
```

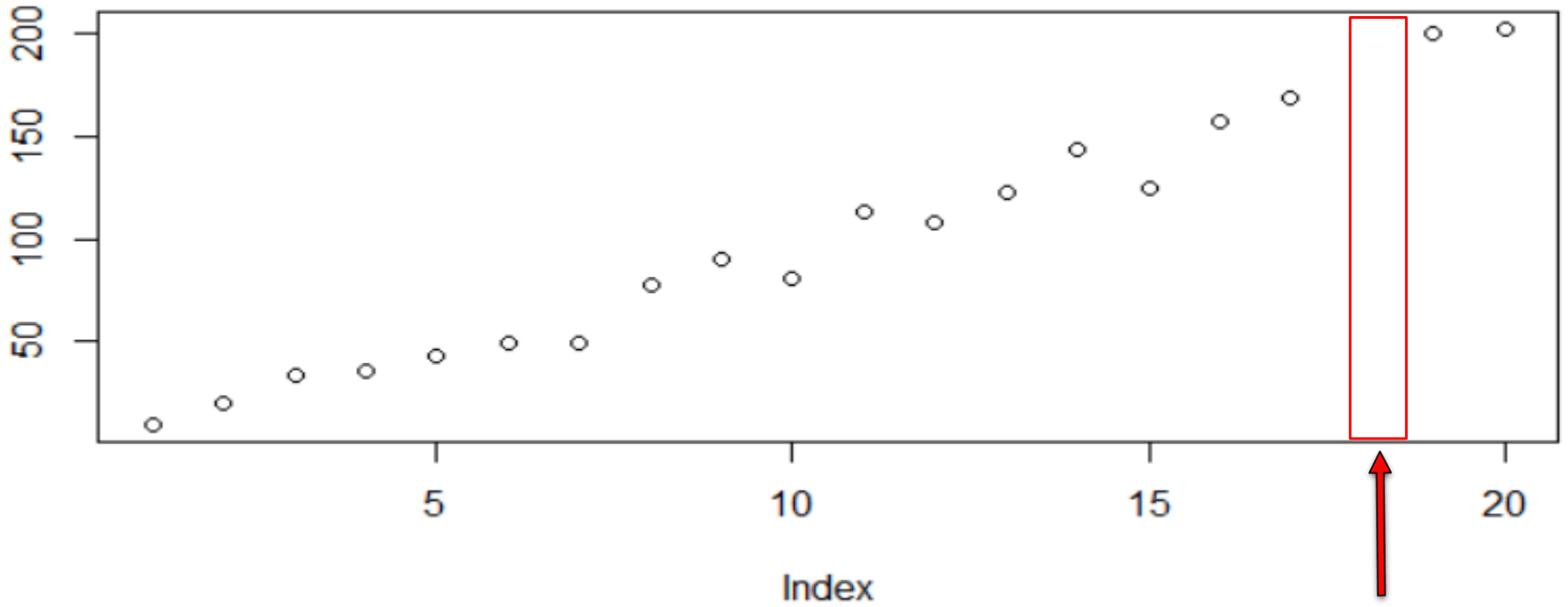
Limpeza dos dados – Valores faltantes

- Há casos em que uma boa opção é usar o valor imediatamente anterior ou posterior do conjunto.
- Exemplo: considere o conjunto $\boldsymbol{x}$ com os seguintes valores:

```
8.9 19.2 33.8 34.9 43.0 49.5 49.1 77.8 90.4 81.1 112.7 108.1
122.2 143.3 125.2 157.3 169.2 NA 200.5 203.0
```

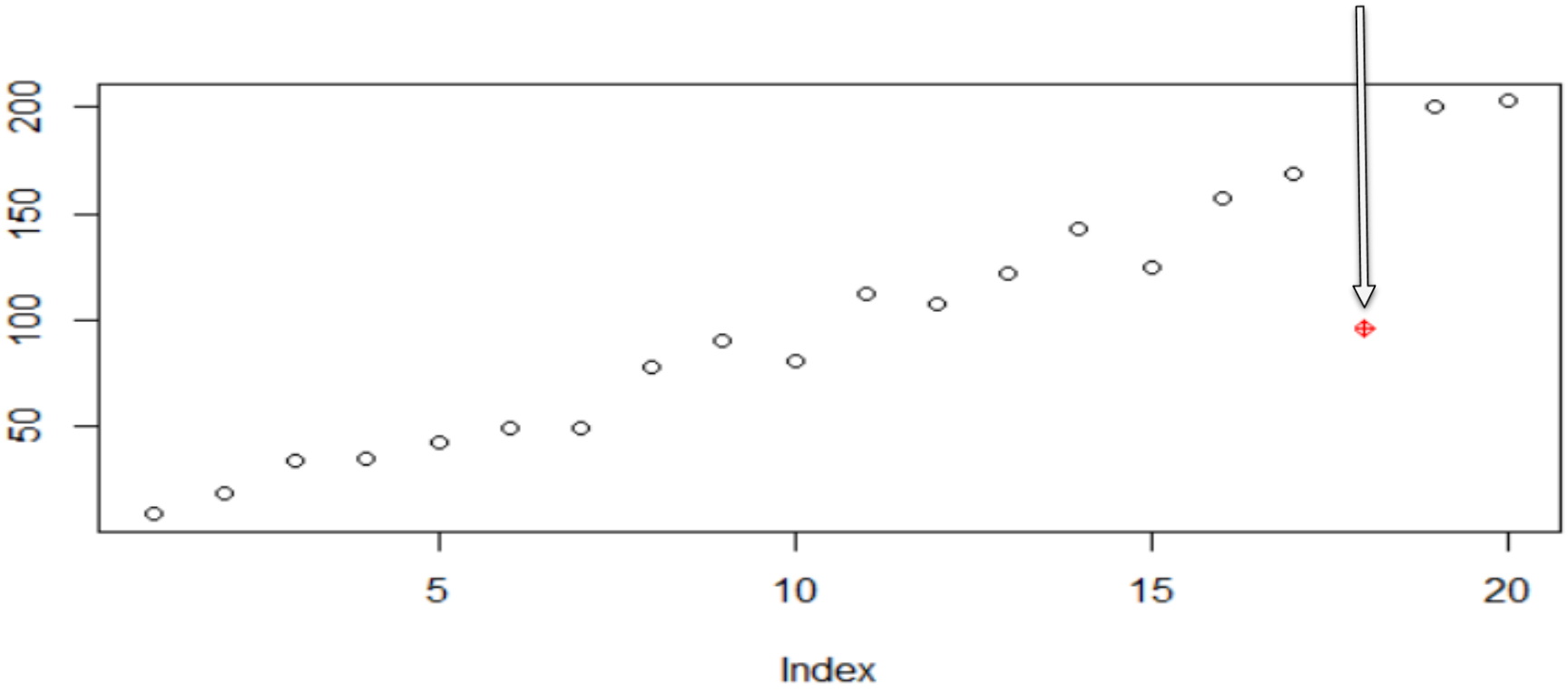
Limpeza dos dados – Valores faltantes

→ Qual seria um bom valor para $\boldsymbol{x}[\mathbf{18}]$?



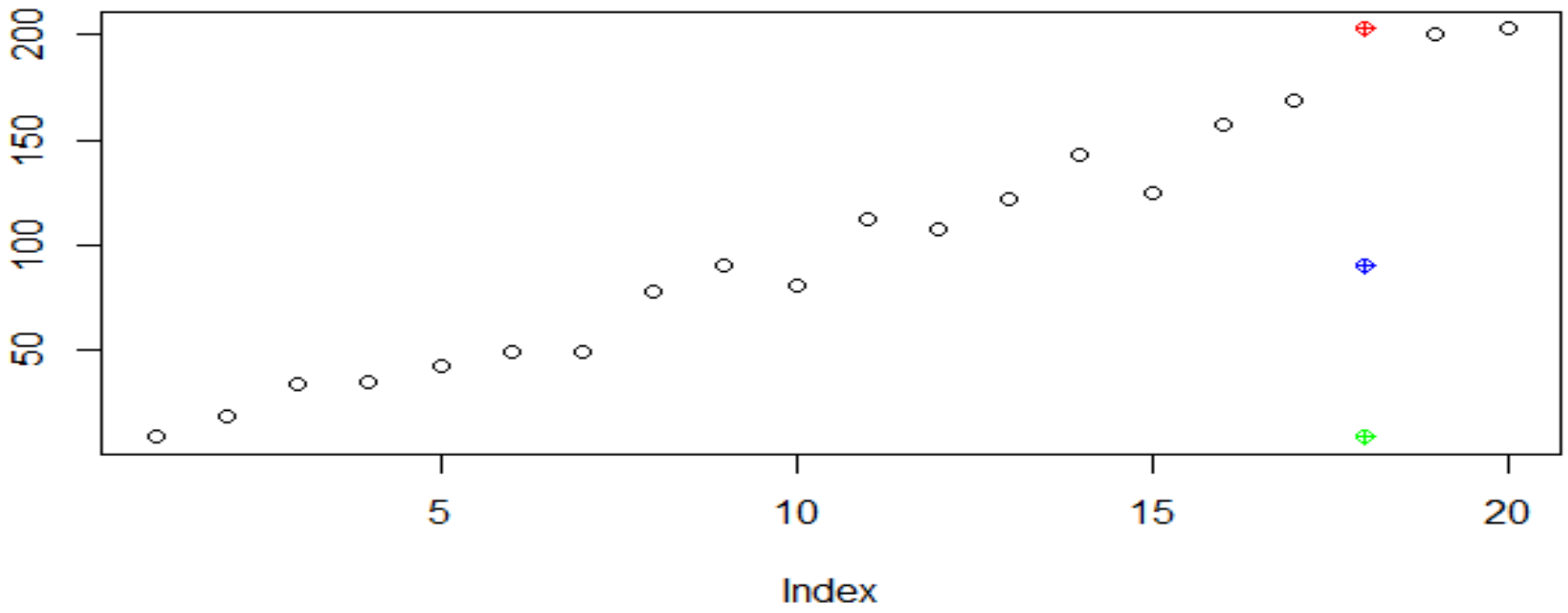
Limpeza dos dados – Valores faltantes

→ Vamos usar a média $\boldsymbol{x}[\mathbf{18}]$?



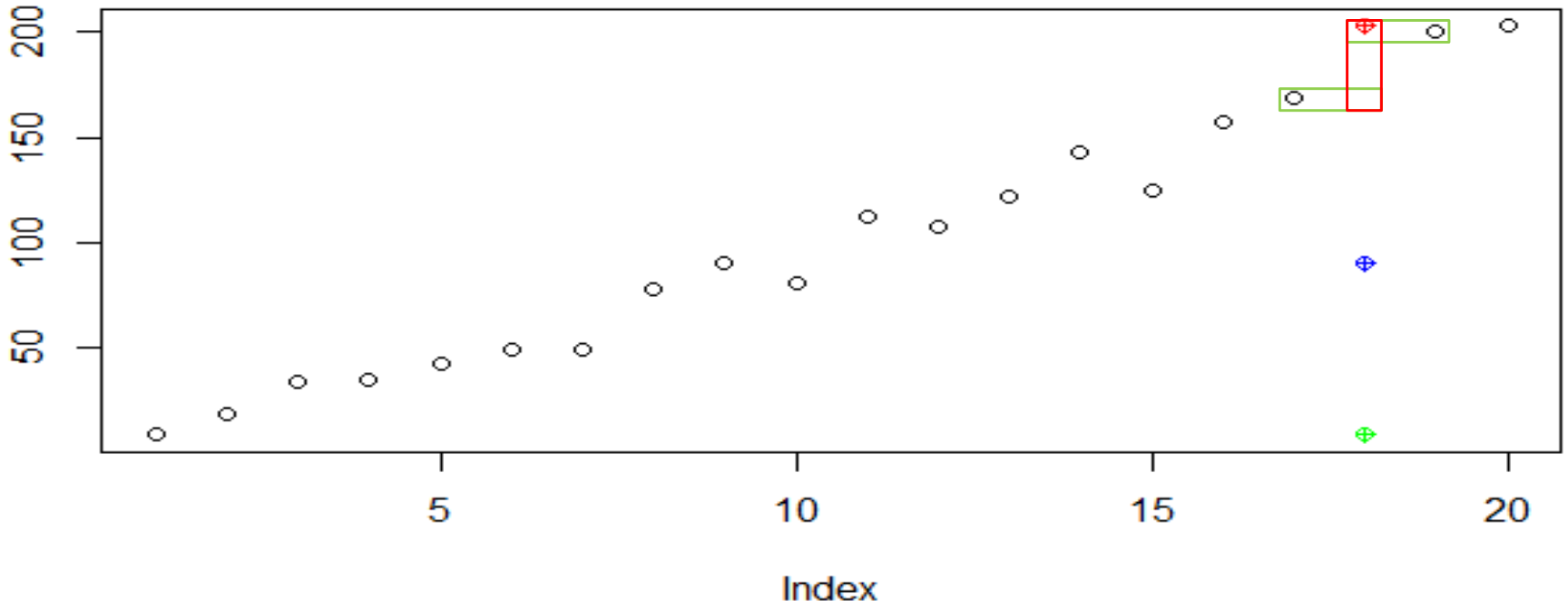
Limpeza dos dados – Valores faltantes

→ Talvez o máximo, mínimo, ou a mediana?



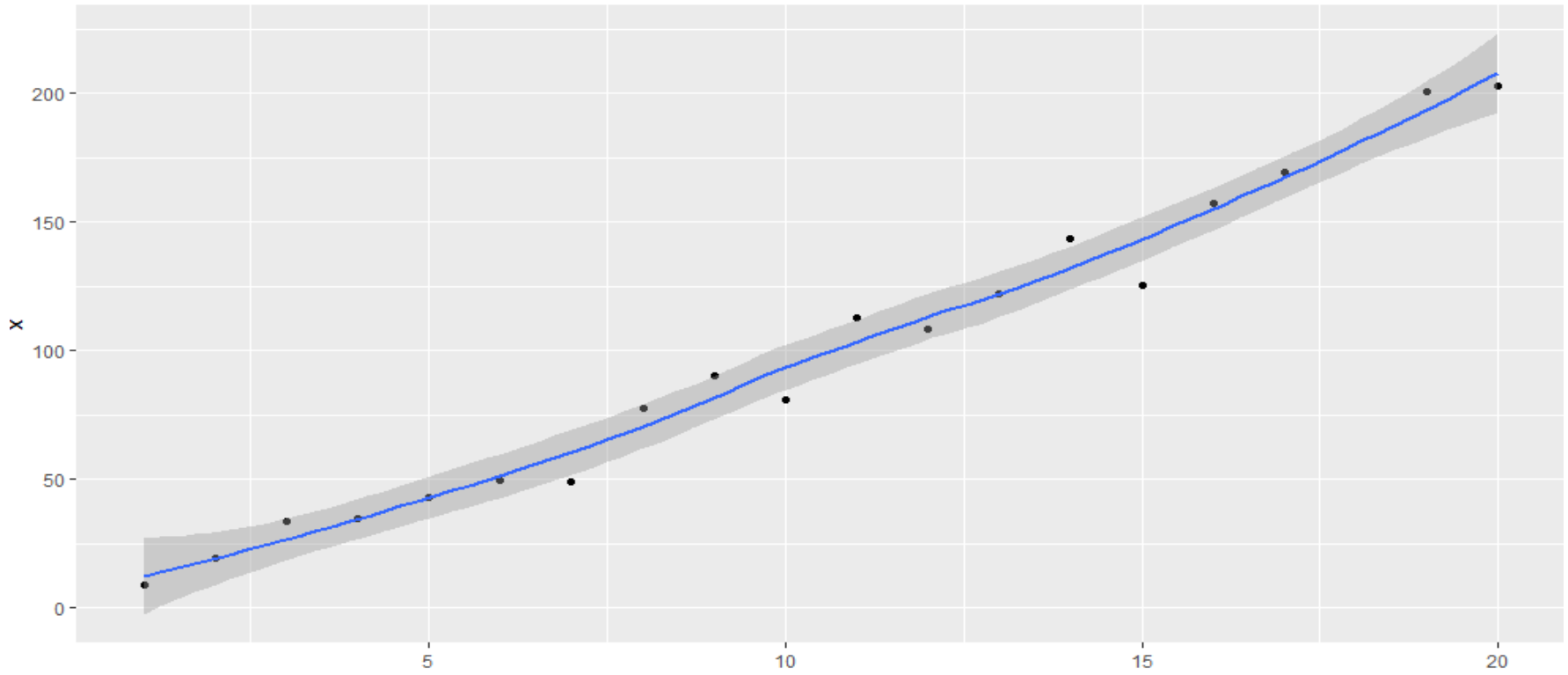
Limpeza dos dados – Valores faltantes

→ Ou podemos simplesmente usar o valor anterior ou posterior como referência.



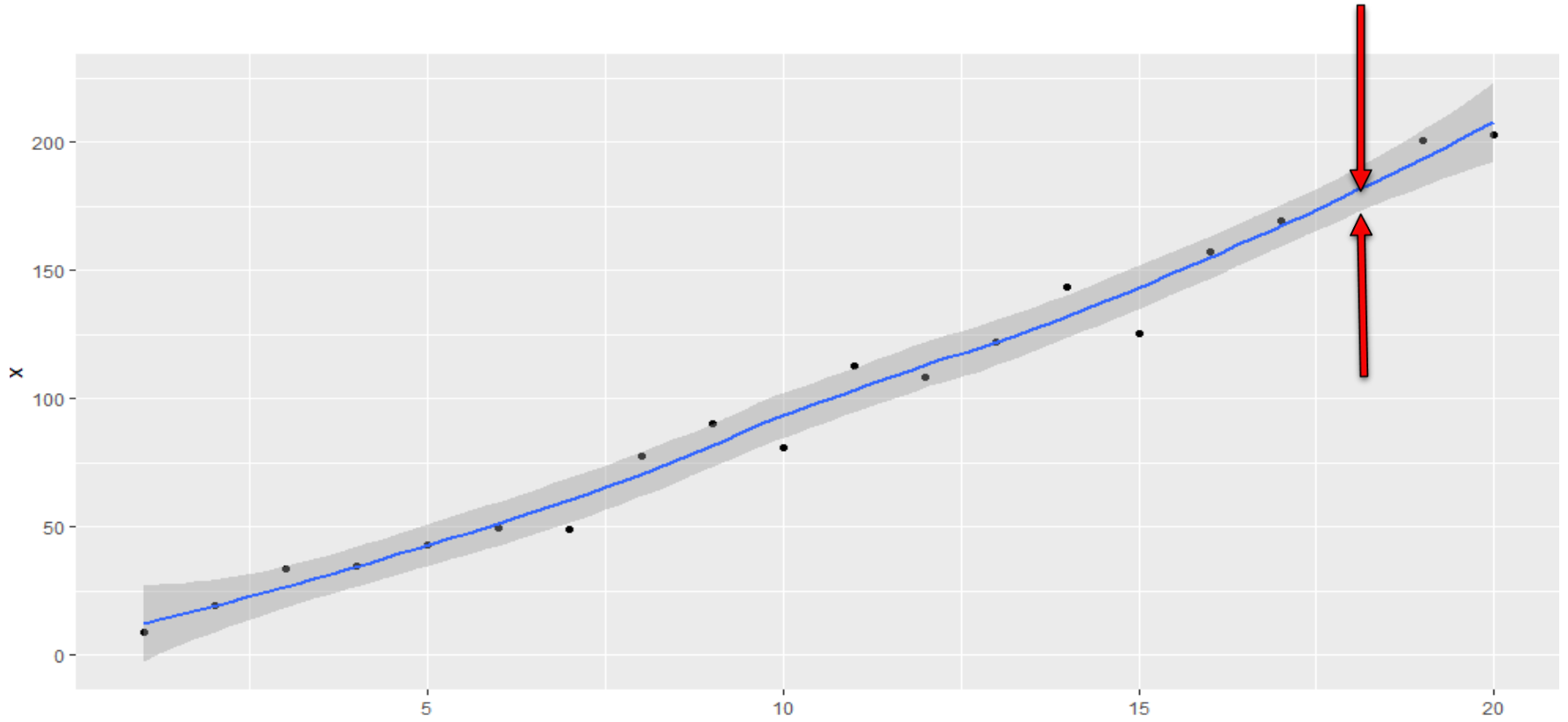
Limpeza dos dados – Valores faltantes

→ Uma solução mais acurada seria dada por um modelo de regressão.



Limpeza dos dados – Valores faltantes

→ No gráfico abaixo podemos extrair o valor ideal para $\boldsymbol{x}[\mathbf{18}]$.



Technical help

Em R:

- Assimilar valores para um item *i* do vetor: `x[i] <- x[j]`
- Ajuste de um modelo linear para dados x e y de um conjunto: `myfit <- lm(y ~ x)`
- *Plot* de um modelo linear sobre os dados de origem: `qplot(x,y, data = myfit, geom = c("point","smooth"))`
- Importante! `x` e `y` devem ter a mesma dimensão.

Hands on!

1) Dado os conjuntos

```
a={3,5,6,7,9,14, 16, 16, NA, 27, 34, 50, 61}

b={10, 9, 10, 11, 8, 11, 10, 12, NA, 11, 10, 8, 10}

c={0, 2, 3, 0, 4, 2, 0, 6, 2, 0, 4, 5, NA, 3, 4, NA, 5}
```

1) Use algumas das técnicas anteriores para estimar os valores faltantes para `a[9], b[9], c[13], e c[16]`

Limpeza dos dados – Ruídos

- Valores fora da escala normal.
- Tipo de dado não correspondente ao campo.
- Dados inválidos (ruídos de sensores).

Limpeza dos dados – Ruídos

Algumas soluções comuns são:

- Verificar e excluir valores fora da escala. Após, deve-se usar um meio apropriado de completar o *gap* caso seja útil para a análise.
- Alteração de tipos de dados, conversão para o tipo requerido, em alguns casos, reestruturação do conjunto.
- Detecção e exclusão dos ruídos, ou ampliação do sinal de modo a evidenciar o que não é ruído.

Hands on!

Abra o arquivo ***LD00c.csv***. Faça uma varredura rápida para detectar possíveis ruídos. Preencha os dados faltantes conforme julgar mais adequado. Na sequência obtenha:

1) A média, o total, o máximo, o mínimo para cada variável.
   I. ... Qual critério você usou para preencher os dados?
2) O dia que teve maior insolação.
3) O dia que teve a temperatura mais alta.
4) Um histograma (*hist*) da temperatura mínima.
5) Algo interessante foi encontrado com as temperaturas?

*http://www-usr.inf.ufsm.br/~joaquim/UFSM/DM/ds/

Hands on! //homework

Altere seu script e resolva os problemas anteriores usandoo dataset *LD00**b**.csv*.

Use um modelo linear para preencher os valores anômalos encontrados na temperatura.

*http://www-usr.inf.ufsm.br/~joaquim/UFSM/DM/ds/