

Aula 2 - Gerenciamento de Dados

Thiago S. F. Silva

Dado vs informação?

- Usabilidade
- Reproducibilidade
- Transparência
- Praticidade

Quem é o seu principal colaborador?

- Você mesmo daqui a 3 meses.

Dados estruturados vs. não-estruturados

- **Dados Estruturados:** planilhas, bancos de dados - dados organizados de maneira tabular.
- **Dados Não-Estruturados:** fotos, gravações, transcrições de entrevistas, paginas da web, etc.

Metadado: O dado sobre o dado. Informação adicional / documentação explicativa que permite o uso correto e reproduzível dos dados.

- Especialmente importante para dados não estruturados.
- Ex: Data, hora, local, coletor, instrumento, condições ambientais, unidades, nomes de variáveis, nomenclatura dos arquivos, organização das pastas. etc.

O gerenciamento de dados começa no campo/laboratório

- Protocolo de amostragem
- Caderno de laboratório
- Planilhas de campo

1. Faça um **estudo piloto**

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:
 - Ao final de cada dia/sessão, **fotografe** seu caderno / planilha de campo.

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:
 - Ao final de cada dia/sessão, **fotografe** seu caderno / planilha de campo.
 - Se houver acesso à internet, já sincronize com a nuvem.

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:
 - Ao final de cada dia/sessão, **fotografe** seu caderno / planilha de campo.
 - Se houver acesso à internet, já sincronize com a nuvem.
 - Dados digitais:

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:
 - Ao final de cada dia/sessão, **fotografe** seu caderno / planilha de campo.
 - Se houver acesso à internet, já sincronize com a nuvem.
 - Dados digitais:
 - Baixe com a maior frequência possível.

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:
 - Ao final de cada dia/sessão, **fotografe** seu caderno / planilha de campo.
 - Se houver acesso à internet, já sincronize com a nuvem.
 - Dados digitais:
 - Baixe com a maior frequência possível.
 - Mantenha pelo menos duas cópias, em dois dispositivos

Boas práticas

1. Faça um **estudo piloto**
2. Prepare suas planilhas / caderno **antes** de começar.
3. Anote observações sobre **tudo** o que desviar do protocolo.
Não, você não vai lembrar depois.
4. **Backup, backup, backup:**
 - Dados analógicos:
 - Ao final de cada dia/sessão, **fotografe** seu caderno / planilha de campo.
 - Se houver acesso à internet, já sincronize com a nuvem.
 - Dados digitais:
 - Baixe com a maior frequência possível.
 - Mantenha pelo menos duas cópias, em dois dispositivos

5. Tente digitar os dados o mais rápido possível - enquanto os detalhes do trabalho ainda estão frescos na sua cabeça.
6. Caderno /planilha vs. entrar dados direto no PC/tablet/telefone?

Qual software você usa pra digitar seus dados?

- Excel: cuidado com a formatação dos dados (exemplo)

Qual software você usa pra digitar seus dados?

- Excel: cuidado com a formatação dos dados (exemplo)
- Datas e horas: formato ISO8601

Qual software você usa pra digitar seus dados?

- Excel: cuidado com a formatação dos dados (exemplo)
- Datas e horas: formato ISO8601
- Separadores decimais e de milhar: vírgula ou ponto?

Qual software você usa pra digitar seus dados?

- Excel: cuidado com a formatação dos dados (exemplo)
- Datas e horas: formato ISO8601
- Separadores decimais e de milhar: vírgula ou ponto?
- Indicadores de dado faltante (*missing data*)

Exemplo 1 - Identifique e corrija os problemas

Prática: Configurando o RStudio para trabalhar

Uma parte **essencial** da análise de dados é a estruturação de um *projeto*:

- Conjunto de pastas que organiza o trabalho
- Ajuda na auto-documentação do projeto
- Diferentes análises - diferentes projetos

Organização mínima de um projeto:

```
meu_projeto
|_____ src
|_____ data
|           |_____ raw
|           |_____ processed
|_____ outputs
|_____ docs
```

Organização mínima de um projeto:

```
meu_projeto
|_____ scripts
|_____ dados
|          |_____ brutos
|          |_____ processados
|_____ resultados
|_____ notas
```

Boas práticas de nomenclatura de pastas/arquivos

- sem letras maiúsculas
- sem acentos
- sem espaços
- usar somente - e _ como caracteres extra

Estilos de nomenclatura de pastas/arquivos/variáveis

Snake case: meu_lindo_arquivo.txt

Kebab case: meu-lindo-arquivo.txt

Camel case: meuLindoArquivo.txt

Pascal case: MeuLindoArquivo.txt

Escolha sua favorita - mas seja **consistente!**

Exemplo: criando um projeto no RStudio

- Cada variável é uma coluna
- Cada observação é uma linha
- Cada tabela é um contexto

<https://www.jstatsoft.org/article/view/v059i10>

Qual o melhor formato para armazenamento de dados estruturados?

- Excel: xls, xlsx

Qual o melhor formato para armazenamento de dados estruturados?

- Excel: xls, xlsx
- CSV (Comma Separated Values)

Qual o melhor formato para armazenamento de dados estruturados?

- Excel: xls, xlsx
- CSV (Comma Separated Values)
- Parquet

Qual o melhor formato para armazenamento de dados estruturados?

- Excel: xls, xlsx
- CSV (Comma Separated Values)
- Parquet
- Estratégia Mista

E dados não-estruturados?

- Priorize formatos abertos

E dados não-estruturados?

- Priorize formatos abertos
- Use texto simples para os metadados

E dados não-estruturados?

- Priorize formatos abertos
- Use texto simples para os metadados
- Dados de texto não estruturados: JSON

Exemplo prático: lendo diferentes tipos de arquivos para o R.

Passo 1: preparando o arquivo `plot_florestal.xlsx`

Pacote: readxl (parte do tidyverse)

```
library(readxl)
```

```
dados <- 0
```