Tarefa 1: Base de Dados

12 de agosto de 2025

> Estude bastante o que mais lhe interessa, da maneira mais indisciplinada, irreverente e original possível.
>
> *Richard Feynmann*

1 Introdução

Nesta primeira tarefa, aprenderemos como escolher uma base de dados para estudo e como realizar as primeiras análises exploratórias de suas variáveis. Todo o trabalho será feito utilizando a linguagem R, que nos oferece um ambiente adequado para aplicar as ferramentas e métodos do aprendizado estatístico.

2 Banco de Dados

Escolha do Banco de Dados

O estudo e a prática do aprendizado estatístico exigem o uso de bases de dados. Para esta atividade, acesse os sites listados abaixo e selecione um conjunto de dados que desperte seu interesse. Verifique como realizar o download, salve-o no formato `.csv` e leia-o no R.

Evite conjuntos de dados muito grandes ou com excesso de variáveis. O ideal é escolher uma base suficientemente rica para permitir análises relevantes, mas simples o bastante para ser manipulada em um computador comum. Tente escolher uma base com diversidade de comportamento das variáveis.

Após baixar a base, revise os nomes das variáveis e, se necessário, ajuste-os para que fiquem mais claros e consistentes. Por fim, registre o motivo da escolha dessa base de dados, justificando seu interesse nela.

UCI Machine Learning Repository

Tidy Finance – Financial Data in R

Análise Macro – Baixando dados da B3 com R

Base dos Dados – Documentação

Hugging Face – GigaVerbo (Portuguese Corpus)

Kaggle – Conjuntos de Dados

Prefeitura de Santos – Dados Abertos

Banco Central do Brasil – Dados Abertos

IBGE – Centro de Estudos Estatísticos

Porto de Santos – Estatísticas Online

Banco Nacional de Dados Oceanográficos (BNDO)

Governo Federal – Portal de Dados Abertos

Fatec – Dados Abertos

Pacote

Para instalar um pacote no R, utiliza-se a função:

```
install.packages("readr")
```

Uma vez instalado, é necessário carregá-lo com o comando:

```
library("readr")
```

Assim, todas as funções do pacote ficam disponíveis para uso.

O pacote `readr` é voltado para a leitura e escrita de arquivos de texto contendo dados, como `.csv`, `.tsv`.

- `read_csv()`: Lê arquivos delimitados por vírgula.
- `read_csv2()`: Lê arquivos delimitados por ponto e vírgula (comum no Brasil e na Europa).
- `read_tsv()`: Lê arquivos delimitados por tabulação.
- `read_delim()`: Lê arquivos delimitados por qualquer caractere especificado. Neste caso use a opção `delim = "..."`.
- `read_table()`: Lê arquivos com colunas separadas por espaços em branco.
- `write_csv()`: Salva um *data frame* ou *tibble* em um arquivo CSV.
- `write_tsv()`: Salva um *data frame* ou *tibble* em formato TSV.

Caso sua base de dados esteja no formato `.xlsx`, utilize o pacote `readxl`. Após instalar o pacote, use a função `read_excel()` para realizar a leitura do arquivo.

Se desejar importar uma aba específica da planilha, utilize o argumento `sheet = "NomeDaAba"`.

Para salvar uma base no formato `.xlsx`[1], utilize a função `write_xlsx()` do pacote `writexl`.

Assim que sua base de dados estiver carregada, é fundamental examiná-la diretamente. Para isso, utilize a função `head(base_de_dados, 10)` para visualizar as 10 primeiras linhas da tabela. Caso deseje visualizar as 10 últimas linhas, utilize a função `tail(base_de_dados, 10)`.

3 Variáveis

Com a base de dados em mãos, é importante compreender detalhadamente as variáveis que ela contém. Para isso, comece consultando a documentação disponível da base escolhida. Procure responder às seguintes questões:

[1] Recomenda-se cautela ao salvar nesse formato, pois pode haver limitações; prefira formatos como `.csv` para maior compatibilidade.

- Qual foi o método utilizado para a coleta dos dados?
- A base já foi empregada em trabalhos acadêmicos ou pesquisas científicas?
- A base possui alguma certificação oficial ou validação reconhecida?
- Em sua opinião, a base selecionada é confiável? Justifique sua resposta.

Após responder a estas perguntas, vamos olhar mais atentamente as variáveis. Para verificar os nomes das variáveis, use a função `names(base_de_dados)`. Os nomes precisam ser coerentes com as suas práticas. Se for preciso alterar algum nome use o comando

```
names(base_de_dados$variavel) <- "novo-nome"
```

Uma sugestão: evite usar o underscore em nomes.

Precisamos agora fazer uma descrição qualitativa das variáveis da base de dados. Para isso, organize uma tabela com as seguintes colunas:

- **Nome da Variável**: Identificação da variável exatamente como consta na base de dados.
- **Tipo**: Classe da variável, como numérica, categórica, texto, data, lógica etc.
- **Unidade**: Unidade de medida da variável, se aplicável (por exemplo, metros, segundos, reais). Caso não se aplique, deixar em branco ou informar "não se aplica".
- **Descrição**: Breve explicação sobre o que a variável representa.
- **Observação**: Informações adicionais relevantes, como detalhes sobre a coleta dos dados, possíveis limitações, valores ausentes ou cuidados na interpretação.

Para auxiliar na descrição qualitativa das variáveis, você pode utilizar algumas funções básicas do R que fornecem informações sobre os dados:

- `str()`: Mostra a estrutura do objeto, exibindo o tipo de cada variável e uma prévia dos seus valores.
- `class()`: Retorna a classe do objeto ou variável (por exemplo, `numeric`, `factor`, `character`).
- `typeof()`: Indica o tipo interno do objeto no R (por exemplo, `double`, `integer`, `character`).
- `summary()`: Fornece um resumo estatístico básico para cada variável, como mínimos, máximos, medianas e quartis para variáveis numéricas, e frequências para variáveis categóricas.

O uso combinado dessas funções facilita a identificação do tipo, distribuição e características principais de cada variável, auxiliando na elaboração da tabela de descrição qualitativa.