

Aula 01: Dados Tabulares

ME315 - Manipulação de Bancos de Dados

Benilton Carvalho, Guilherme Ludwig, Tatiana Benaglia

Com modificações por Carlos Trucíos

Arquivos Tabulares

- Arquivos tabulares têm forma retangular;
- Exemplo clássico de arquivo tabular: planilha Excel;
- Podem ser apresentados ao usuário/analista em diferentes versões;
- Sugestão para realização de análises:
 - Importação dos dados feita cautelosamente;
 - Assim, minimiza-se trabalhos posteriores na formatação dos dados;
- Colunas costumam representar variáveis e linhas, observações;

Formato Tidy

- Anteriormente, i.e. no SAS, conhecido como formato longo;
- É o melhor formato para análises estatísticas;
- Pode não ser o formato mais compacto, mas é o mais versátil;
- Métodos comumente implementados para ciência de dados costumam utilizar como entrada dados no formato tidy;

Formato Tidy

- Cada linha é uma única observação;
- Cada coluna é o nome de uma variável;
- Cada célula é um valor;

Produto	Dia	Valor
Gasolina	Segunda	4.19
Gasolina	Terça	4.19
Gasolina	Quarta	4.09
Etanol	Segunda	3.39
Etanol	Terça	3.39
Etanol	Quarta	3.09

Formato não-tidy

- Nomes de colunas possuem o valor de uma variável;

Produto	Segunda	Terça	Quarta
Gasolina	4.19	4.19	4.09
Etanol	3.39	3.39	3.09

Formato não-tidy

- Valores em uma coluna correspondem a duas variáveis;
- Uma célula pode corresponder a mais de um valor;

Produto-dia	Valor
Gasolina-Segunda	4.19
Gasolina-Terça	4.19
Gasolina-Quarta	4.09
Etanol-Segunda	3.39
Etanol-Terça	3.39
Etanol-Quarta	3.09

Formato não-tidy

Observação:

As funções `pivot_longer()` e `pivot_wider()` do pacote `tidyr` são bastante úteis para passar de um formato a outro.

Arquivos CSV

- Arquivo no formato texto;
- Cabeçalho opcional;
- Separador é vírgula;
- Separador decimal deve ser diferente de vírgula (por exemplo, ponto)
- Será problemático em países que utilizam a vírgula como separador decimal;

```
Produto,Dia,Valor  
Gasolina,Segunda,4.19  
Gasolina,Terça,4.19  
Gasolina,Quarta,4.09  
Etanol,Segunda,3.39  
Etanol,Terça,3.39  
Etanol,Quarta,3.09
```


Arquivos CSV2

- Arquivo no formato texto;
- Cabeçalho opcional;
- Separador é ponto-e-vírgula;
- Separador decimal deve ser diferente de ponto-e-vírgula (por exemplo, vírgula)

```
Produto;Dia;Valor  
Gasolina;Segunda;4,19  
Gasolina;Terça;4,19  
Gasolina;Quarta;4,09  
Etanol;Segunda;3,39  
Etanol;Terça;3,39  
Etanol;Quarta;3,09
```

Arquivos TSV

- Arquivo no formato texto;
- Cabeçalho opcional;
- Separador é o símbolo de tabulação (visível como espaço em branco)

Produto	Dia	Valor
Gasolina	Segunda	4.19
Gasolina	Terça	4.19
Gasolina	Quarta	4.09
Etanol	Segunda	3.39
Etanol	Terça	3.39
Etanol	Quarta	3.09

Arquivos Delimitados

- Arquivo no formato texto;
- Cabeçalho opcional;
- Separador é definido pelo criador do arquivo:
 - Se vírgula, então é um arquivo CSV;
 - Se ponto-e-vírgula, então é um arquivo CSV2;
 - Se tabulação, então é um arquivo TAB;

Arquivos de Largura Fixa

- Representação relativamente compacta de dados;
- A largura de cada campo é pré-especificada;
- Muito rápidos de serem importados, visto que a posição de cada campo é sempre fixa;
- Usuário precisa entender o posicionamento de cada campo;
- Essencialmente, é preciso conhecer as posições de início e fim de cada campo;

Arquivos XLS/XLSX

- Arquivos binários ou XML;
- Também conhecidos como "arquivos Excel";
- Versões antigas do Excel, restringem arquivos a terem, no máximo, 65.535 linhas;
- Versões recentes do Excel, restringem arquivos a terem, no máximo, 1 milhão de linhas;

Importação de Dados: R

- O R oferece múltiplas opções para importação de dados tabulares;
- Comandos do pacote básico:
 - CSV: `read.csv`
 - CSV2: `read.csv2`
 - TSV: `read.delim`
 - Delimitados: `read.table`
 - Largura fixa: `read.fwf`
- Formas aprimoradas para importação estão implementadas no pacote `readr`:
 - CSV: `read_csv`
 - CSV2: `read_csv2`
 - TSV: `read_tsv`
 - Delimitados: `read_delim`
 - Largura fixa: `read_fwf`

Observações - Importação de Dados

- Arquivos delimitados são os mais genéricos;
- São a base arquivos de conteúdo retangular;
- Casos especiais de `read_delim`:
 - `read_csv`
 - `read_csv2`
 - `read_tsv`

Dicas para Importação de Arquivos

- "Espie" o conteúdo do arquivo:
 - as primeiras linhas já podem ser suficientes;
- O que separa uma coluna da sua vizinha?
- Qual é o separador decimal utilizado?
- Qual é o separador de milhar utilizado?
- O arquivo possui cabeçalho?
- Que *string* define o que é um valor faltante?
- Existem linhas de comentário dentro do arquivo?
- Existem linhas no início do arquivo que devem ser puladas no momento da importação?
- Quantas linhas devem ser importadas?
- Quais são os tipos de cada coluna a ser lida?

Sugestões para Criação de Arquivos Tabulares

- Dados volumosos podem ser problemáticos em arquivos XLS/XLSX (limite de linhas);
- Quando possível, prefira arquivos de formatos mais simples (arquivos texto vs. XLS/XLSX ou outros binários);
- Não é preciso descompactar um arquivo texto antes de importá-lo;
- Ao escrever código, utilize nomes explícitos de argumentos (`read_delim('arq.txt', del=',')` vs. `read_delim('arq.txt', delim=',')`);
- Evite nomear colunas com expressões:
 - Iniciadas por números;
 - Que contenham espaços em branco;
 - Que contenham caracteres especiais (como letras acentuadas);

Importação de Dados em Python

- O Python possui um módulo chamado Python Data Analysis Library;
- Módulo amplamente conhecido como pandas;
- pandas = Panel Data
- Capaz de importar:
 - CSV;
 - TSV;
 - Arquivos delimitados em geral;
 - Até mesmo, SQL;
- Resultado da importação é um objeto DataFrame;
- Um DataFrame no Python comporta-se basicamente como um `data.frame` no R.

Atrasos em Vôos nos EUA

Origem: <https://www.kaggle.com/usdot/flight-delays>

- `airlines.csv` tem 359 b.
- `airports.csv` tem 23.3 kb.
- `flights.csv.zip` tem 185.8 Mb (zipped).

Importando Dados via R

```
library(readr)
in1 = read_csv('../dados/flights.csv.zip')
```

```
## Rows: 5819079 Columns: 31
## — Column specification _____
## Delimiter: ","
## chr (11): AIRLINE, TAIL_NUMBER, ORIGIN_AIRPORT, DESTINATION_AIRPORT, SCHEDUL...
## dbl (20): YEAR, MONTH, DAY, DAY_OF_WEEK, FLIGHT_NUMBER, DEPARTURE_DELAY, TAX...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Importando Dados via R

```
head(in1)[, 1:7]
```

##	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER
## 1	2015	1	1	4	AS	98	N407AS
## 2	2015	1	1	4	AA	2336	N3KUAA
## 3	2015	1	1	4	US	840	N171US
## 4	2015	1	1	4	AA	258	N3HYAA
## 5	2015	1	1	4	AS	135	N527AS
## 6	2015	1	1	4	DL	806	N3730B

Importando Dados via Python

```
library(reticulate)
```

```
## Warning: package 'reticulate' was built under R version 4.4.3
```

```
py_discover_config()
```

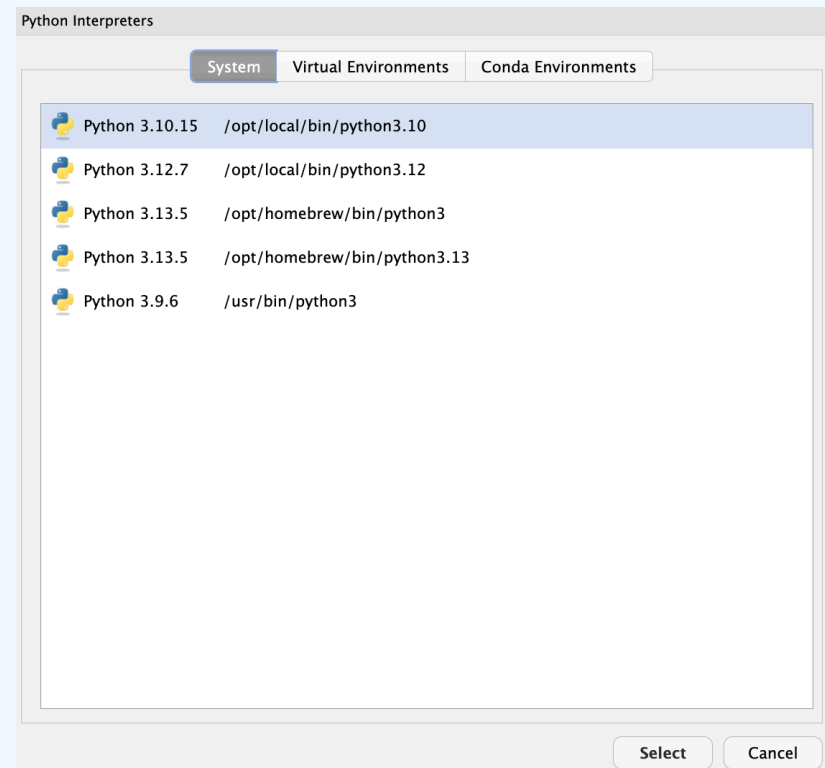
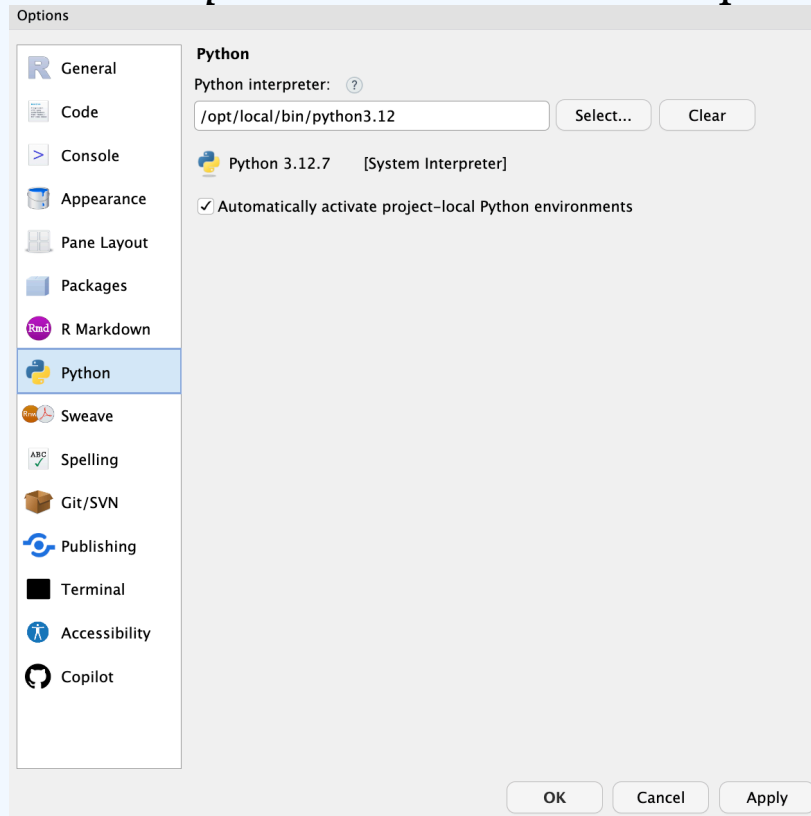
```
## python:           /Users/ctrucios/.virtualenvs/env_me315/bin/python
## libpython:        /opt/local/Library/Frameworks/Python.framework/Versions/3.12/lib/python
## pythonhome:       /Users/ctrucios/.virtualenvs/env_me315:/Users/ctrucios/.virtualenvs/env
## version:          3.12.7 (main, Oct  5 2024, 00:14:26) [Clang 15.0.0 (clang-1500.1.0.2.5)]
## numpy:            /Users/ctrucios/.virtualenvs/env_me315/lib/python3.12/site-packages/num
## numpy_version:    2.3.4
##
## NOTE: Python version was forced by VIRTUAL_ENV
```

```
#use_python("/opt/local/bin/python3.12")
```

Importando Dados via Python

Observação:

Se `py_discover_config()` retornar `NULL`, o problema pode ser que não selecionou o *Python interpreter*. Para solucionar o problema: Tools > Global Options >



Importando Dados via Python

```
virtualenv_create("env_me315", python = "/opt/local/bin/python3.12")  
use_virtualenv("env_me315", required = TRUE)  
py_install("pandas", envname = "env_me315")  
# virtualenv_remove("env_me315", confirm = FALSE)
```

```
use_virtualenv("env_me315", required = TRUE)
```


Confirmando Configuração Detectada de Python

```
py_config()
```

```
## python:          /Users/ctrucios/.virtualenvs/env_me315/bin/python
## libpython:       /opt/local/Library/Frameworks/Python.framework/Versions/3.12/lib/python
## pythonhome:      /Users/ctrucios/.virtualenvs/env_me315:/Users/ctrucios/.virtualenvs/env
## version:         3.12.7 (main, Oct  5 2024, 00:14:26) [Clang 15.0.0 (clang-1500.1.0.2.5)]
## numpy:           /Users/ctrucios/.virtualenvs/env_me315/lib/python3.12/site-packages/num
## numpy_version:   2.3.4
##
## NOTE: Python version was forced by use_python() function
```

Importando Dados via Python

```
import pandas as pd
in2 = pd.read_csv("../dados/flights.csv.zip", compression="zip",
                  header=0, sep=";", low_memory=False)
```

Importando Dados via Python

```
in2.head().iloc[:, list(range(1, 7))]
```

##	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER
## 0	1	1	4	AS	98	N407AS
## 1	1	1	4	AA	2336	N3KUAA
## 2	1	1	4	US	840	N171US
## 3	1	1	4	AA	258	N3HYAA
## 4	1	1	4	AS	135	N527AS

Questão para pensar

- Dados tabulares podem ser bastante grandes;
- Como trabalhar com um arquivo que possui 20GB de dados em um computador com 4GB de RAM?