

Análise dos Dados com R

Introdução

James R. Hunter, PhD

26 de setembro de 2023

Introdução

Onde tudo começa

O Que É Nossa Objetivo

- Aprender análise de dados **em prática**
 - Fazer análises de cabo ao rabo
- Utilizar a Linguagem **R**
- Atualizar nossa habilidade em bioestatística básica

Professor James Hunter

- Professor Afiliado, DIPA
- DSc., Laboratório de Retrovírologia, DIPA, UNIFESP
- Pós-Doc, Projeto de Cura de HIV
- Carreira anterior em consultoria de negócios e urbanismo
- Foco em Estatística e Métodos Quantitativos desde 1973
- Trabalho com R desde 2010

Contato com o Professor

- email: jameshunterbr@gmail.com
- Twitter: [jimhunterbr](#)
- Threads: [jameshunterbr](#)
- cel: 11-9-5327-5656
- Office Hours:
 - Tues (3^a feira) 14h - 16h
 - EP2, Rua Pedro de Toledo 669, 6º Andar Fundos

Filosofia da Matéria

- Única maneira para aprender uma linguagem de computação é escrever ela
- Mais código que você escreve, mais fácil será a próxima vez
- Solucionar problemas práticos com R

Don't Panic...



Perguntas

- Fazem muitas!
- Se você tiver uma dúvida, outros na turma terão também
- **Não existe perguntas burras**

Carl Sagan sobre Perguntas Burras

- Astrofísico que escreveu e era apresentador do programa de TV original **Cosmos**
- Livro : **The Demon-Haunted World: Science as a Candle in the Dark (O mundo assombrado pelos demônios)**

There are naive questions, tedious questions, ill-phrased questions, questions put after inadequate self-criticism. But every question is a cry to understand the world. There is no such thing as a dumb question.

Sempre Existe uma Segundo Ponto de Vista



Quanta Matemática Você Precisa Dominar?

- O que aprendeu no colegial suficiente
- Não precisa cálculo
- Somas (Σ), logaritmos e exponentes
- Equação de uma linha reta

$$y = b_0 + b_1 x$$

Informação e Conhecimento

"We are drowning in information, but we are starved for knowledge". -- John Naisbitt

Apesar esta frase seja atribuído a futurólogo John Naisbitt, esta citação tem muitos pais e mães. Usei aqui do livro de Danielle Navarro, **Learning statistics with R: A tutorial for psychology students and other beginners**, 2020, <http://compcogscisydney.org/learning-statistics-with-r>

Porque Nós Precisamos Análise dos Dados?

- Podemos ver as coisas que estudamos? NÃO
 - Vírus, bactérias, células, nucleotídeos, proteínas
- Máquinas que produzem os dados genômicas que estudamos são probabilísticos
 - Palavra "*calling bases*" - sugestão de erro
- Processo natural de replicação celular ou viral - propenso a erros
- Resposta humana às doenças, remédios, tratamentos
 - Nível alto de incerteza e variancia
 - Diferenças naturais entre pessoas

Estatística Ajuda a Encontrar Verdades Subjacentes

- Desenvolver conjunto das regras para processar informação que recebemos
 - Script/Programa
- Tirar conclusões que outros podem entender, concordar ou discordar
- Como alunos, precisam poder conduzir análises básicas
 - Modelos e métodos mais avançados ficam com especialistas

Habilidade Necessária para Todo Ciêntista

- Entender as estatísticas que você vê em papers e livros
- Separar o que é importante do que não é importante
- Separar a verdade de falsidade
- "Call Bullshit"* quando você está sendo enganado
- Resultado: precisamos maneiras probabilísticas para achar essas verdades subjacentes

*CT Bergstrom & JD West, *Calling Bullshit: the Art of Skepticism in a Data-driven World, New York: Random House, 2020.*

17 / 64

Trabalho para a Matéria

- 3 Lições de Casa
 - Trabalho individual
- Participação
 - Perguntas/comentários

R - Uma Ferramenta para Manipulação e Análise dos Dados

CRAN: The Comprehensive R Archive Network

- Uma ONG educacional quem é o dono do código mãe de R
- Fonte oficial para cópias do software base e pacotes averiguados por eles

R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files.

Historia de R

- Baseada na linguagem de programação estatística ("S")
 - S desenvolvida por Bell Labs em 1976
 - Ainda existe como um produto comercial
- R desenvolvida por Ross Ihaka e Robert Gentleman em 1995 em Nova Zelândia
- Comunidade ativa de desenvolvedores e usuários
- Mais que 19.800 pacotes adicionais disponíveis no repositório de CRAN
 - Muitos úteis para as análises biológicas
 - Bioconductor -- outro 2.000 pacotes
 - Muitos outros espalhados em vários repos (e.g. GitHub)

Virtudes de R para Análise de Dados

- Analisar via programas e scripts invés de clicar botões
 - Controlar a sequência e opções de operações em sua análise
- Programas sempre fazem a mesma coisa - produzem mesmo resultado
 - Sem surpresas porque você clicou em um botão que mudou sua análise
 - Só usam opções e parâmetros que você entende
- Criar um registro de como você chegou no resultado
- **De Graça** Sem custo, para sempre!
 - Não tem uma versão "estudantil" estupidamente cara
 - Nem precisa cópias piratas do software

A Crise de Reprodutibilidade

- Sendo capaz de reproduzir análises em tempos diferentes e em labs diferentes
- Maioria dos estudos científicos não podem ser reproduzidos
- *Nature's Checklist de Reprodutibilidade*
 - | Workflows based on point-and-click interfaces, such as Excel, are not reproducible.
Enshrine your computations and data manipulation in code.*
- R e Python trunfa Excel, Graphpad e seus amigos

*Perkel. Challenge to Scientists Nature 584, no. 7822 (2020).

R - Difícil de Aprender

- Se você nunca programou antes, todas as linguagens de computação parecem difíceis ao início
- R muito mais fácil que a maioria
- Passos Iniciais
- Criar vetores e conjuntos de dados ("*data frames*")
 - Executar funções estatísticas e matemáticas
- Vamos começar hoje escrever código
- R torna mais difícil quando você começa de escrever suas próprias funções
 - Quando não pode achar eles nos pacotes que tem

O Que Vocês Devem Fazer

- Investir tempo entre as aulas
- Instalar os softwares (R e RStudio) nos seus laptops
- Ler o material sugerido aqui
- Experimente um dos cursos de R Básico no internet
 - Ter um segundo olhar sobre o mesmo material

RStudio -- Comunicação Sofisticada com R

- Integrated Development Environment ("IDE") para R
- Disponível desde 2010
- Sede de *Tidyverse*
- Onde vocês vão fazer seu trabalho em R
- Também **De Graça**

R & Python

- Python - outra linguagem bastante popular
 - Baseada em conceitos similares aos do R
 - Outra linguagem de alto-nível interpretada
- Lançado em 1991
 - Guido van Rossum de Holanda
 - Nome vem do grupo comédico inglês, "Monty Python's Flying Circus"
 - Não a espécie de cobra
- Para estatística, mais fraco de R
 - Precisa funções de vários módulos para conseguir completar operações básicas de estatísticas

Recursos para a Matéria

Arquivos, Slides, etc.

- Arquivado na página do curso no Google Classroom

Leituras Chaves

- Textos de Estatística
 - Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics 4**
 - Navarro, D. **Learning statistics with R: A tutorial for psychology students and other beginners**
- Livros sobre R - Nível Básico
 - Wickham & Grolemund, **R for Data Science**
 - Ismay & Kim, **Statistical Inference via Data Science: A moderndive into R and the Tidyverse**
 - Irizzary, **Introduction to Data Science**
 - Frank E. Harrell, **R Workflow** (<http://hbiostat.org/rflow/>)

RStudio "Cheat Sheets"

Série de resumos de 1 e 2 páginas de um número de pacotes de funções em R

Base R Cheat Sheet

Getting Help

- ?mean**: Get help of a particular function.
- help.search('weighted mean')**: Search the help files for a word or phrase.
- help(package = 'dplyr')**: Find help for a package.
- More about an object**
- str(iris)**: Get a summary of an object's structure.
- class(iris)**: Find the class an object belongs to.

Using Packages

- install.packages('dplyr')**: Download and install a package from CRAN.
- library(dplyr)**: Load the package into the session, making all its functions available to use.

Vectors

Creating Vectors	
<code>c(2, 4, 6)</code>	<code>2 4 6</code>
<code>2:6</code>	<code>2 3 4 5 6</code>
<code>seq(2, 3, by=0.5)</code>	<code>2.0 2.5 3.0</code>
<code>rep(1:2, times=3)</code>	<code>1 2 1 2 1 2</code>
<code>rep(1:2, each=3)</code>	<code>1 1 1 2 2 2</code>

Programming

For Loop	
<code>for (variable in sequence){</code>	<code>Do something</code>

While Loop

While Loop	
<code>while (condition){</code>	<code>Do something</code>

Vector Functions

<code>sort(x)</code>	<code>rev(x)</code>
Return x sorted.	Return x reversed.
<code>table(x)</code>	<code>unique(x)</code>

Selecting Vector Elements

By Position	
<code>x[4]</code>	The fourth element.
<code>x[-4]</code>	All but the fourth.
<code>x[2:4]</code>	Elements two to four.

If Statements

If Statements	
<code>if (condition){</code>	<code>Do something</code>
<code>} else {</code>	<code>Do something different</code>

Functions

Functions	
<code>function_name <- function(var){</code>	<code>Do something</code>
<code>} else {</code>	<code>Do something different</code>
<code>return(new_variable)</code>	

Reading and Writing Data

Also see the `readr` package.

Cursos Online

- edX - Cursos de Harvard sobre R com Prof. R. Irizzary
 - <https://www.edx.org/learn/r-programming/harvard-university-data-science-r-basics>
- Coursera - Cursos de Johns Hopkins sobre R e outros sobre R em aplicações biomedicos
 - <https://www.coursera.org/specializations/jhu-data-science>
- Coursera - Duke University - sequence of R courses by Cetinkaya-Rundel
 - <https://www.coursera.org/specializations/statistics?>

All excellent

Sites sobre R

- R Bloggers (<https://www.r-bloggers.com/>)
- Tidyverse (<https://www.tidyverse.org/learn/>)
- Stack Overflow (<https://stackoverflow.com/questions/tagged/r>)
- Twitter (#rstats)

Sistemas de Ajuda de R e RStudio

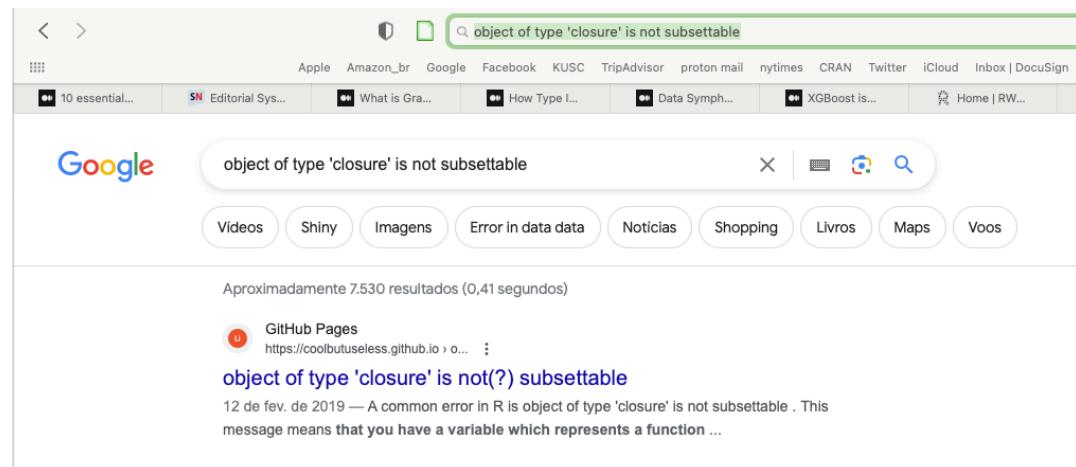
- Completo
- Toda função tem uma tela de ajuda
- Escrito por nerds para outros nerds
 - Explicações às vezes opacas
 - Especialmente mensagens de erro
- Último recurso: copiar a mensagem de erro e colar ele no Google
 - Alguém, em algum lugar, também não entendeu a mesma coisa que é problemática para você

Aplicando Google para um Erro

- O Erro

```
> mean[1:10]
Error in mean[1:10] : object of type 'closure' is not subsettable
```

- Último recurso: copiar a mensagem de erro e colar ele no Google
 - Alguém, em algum lugar, também não entendeu a mesma coisa que é problemática para você



Instalação dos Softwares

Instalar R

- Fica na página seguinte:
 - <https://cran.r-project.org/>



[CRAN
Mirrors](#)
[What's new?](#)
[Search](#)
[CRAN Team](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Task Views](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

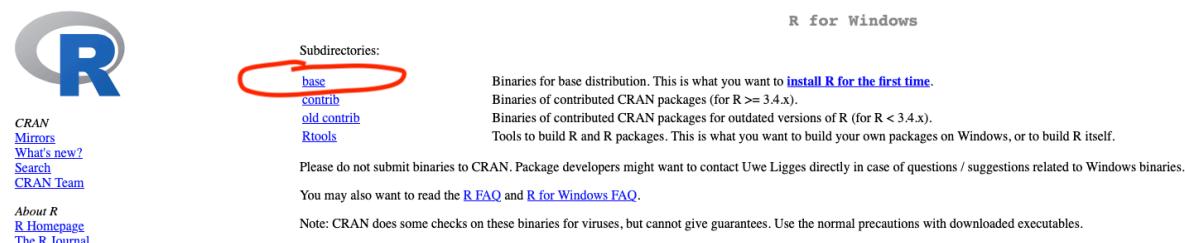
- The latest release (2023-06-16, Beagle Scouts) [R-4.3.1.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

Na Tela Inicial (Windows)

1. Clique no link "Download R for Windows"
2. Na próxima tela, clique no "base"
 - *Mac não faz este passo*



Proceder para Instalação

- Clicar on **Download R 4.3.1 for Windows**



- Programa vai aparecer no seu computador

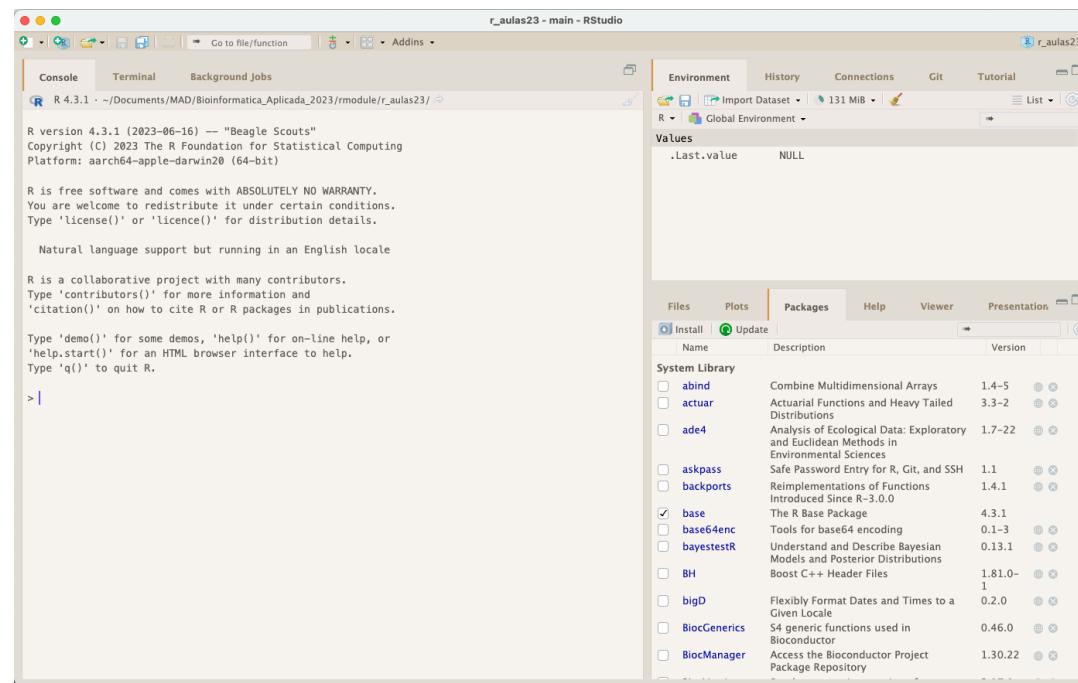
Instalação de RStudio

- Site: <https://posit.co/products/open-source/rstudio/>
- Role para baixo até um grande botão azul: "Download RStudio"
 - Vai informar o número de versão e tamanho do programa

Iniciar RStudio

- Na área de trabalho (ou através dos menus), *duplo click* no ícone de RStudio
 - **Não o ícone de R**
- RStudio abrirá
 - R automaticamente abrirá dentro do RStudio

Console Inicial de RStudio -- Ready to Rock!



Seu Programa Primeiro

Carregar os Pacotes Importantes

- Os pacotes mais importantes que potencializam R
- Usaremos a maioria durante estas 4 semanas
- Script simples (`pacotes_iniciais.r`)

```
packages <- c("tidyverse", "broom", "car", "caret", "corrr", "data.table",
            "descr", "devtools", "gapminder", "ggpubr", "ggsci",
            "glue", "gmodels", "gt", "gtsummary", "here", "Hmisc", "hms",
            "janitor", "jsonlite", "kableExtra", "knitr", "lattice", "lubridate",
            "mice", "nortest", "nycflights13",
            "outliers", "palmerpenguins", "pROC", "psych", "quarto",
            "Rcpp", "readxl", "ROCR", "shiny", "styler", "summarytools",
            "titanic", "usethis")  
  
install.packages(packages)
```

O Que Faz Este Script - Linha 1

- Linha 1: atribuição de conjunto de pacotes ao nome `packages`
 - Utiliza `<-` para fazer a atribuição
- Conjunto de pacotes é combinado num vetor de nomes de pacotes
 - Função `c()` cria um vetor de vários elementos
 - `c()` - *combinar* or *concatenar*
 - *vector* - matriz unidimensional
- Elementos de `packages` - "strings" de classe *character*
 - Entre aspas ("")
- Resultado da Linha 1

VSS: Operadores de Atribuição

- Principal: <-
- Pode usar (mas não é considerada uma boa prática) =
 - **Vai confundir com o sinal para igualdade lógica ==**
 - Vai acontecer! Todos nos fazemos

O Que Faz Este Script - Linha 2

- Instala os pacotes no vetor
- Procura no site de CRAN (espelho) no internet
- Faz os downloads e instala os pacotes
- Vários dos pacotes têm dependências
 - Instalará esses pacotes também
- Dependências: outros pacotes que um pacote precisa para executar as funções do pacote primeiro

Scripts vs. Console

- Escrever seus comandos num script de R Markdown invés do Console
 - Pode salvar seu trabalho
- Console é o lugar onde os comandos são executados
 - Mais fácil de salvar comandos em scripts que salvando a historia dos comandos do Console

Executar "pacotes_iniciais.r"

- Download o arquivo do repo da alua para sua pasta de R
- A aba `Files` no painel inferior direto do RStudio
 - Clicar em `pacotes_iniciais.r`
- Script abrirá no painel superior esquerdo
- Clicar no botão `Source` na barra de comandos
- Pode seguir o progresso no Console



```
## Program para carregar módulos iniciais para matéria MAD
## James R. Hunter, D.Sc.
## 2022-10-03

packages ← c("tidyverse", "broom", "car", "caret", "corrr", "data.table",
           "descr", "devtools", "gapminder", "ggpubr", "glue", "gmodels",
           "gt", "gtsummary", "here", "Hmisc", "hms", "janitor",
           "jsonlite", "kableExtra", "knitr", "lattice", "lubridate",
           "magrittr", "mice", "nortest", "outliers", "palmerpenguins",
           "pROC", "psych", "quarto", "Rcpp", "readxl", "ROCR", "shiny",
           "styler", "summarytools", "titanic", "usethis")

install.packages(packages)
```

R - Operações Básicas

R como um Calculadora

```
5 + 5
```

```
## [1] 10
```

```
36 * 2500000
```

```
## [1] 9e+07
```

```
5876/35.44320
```

```
## [1] 165.7864
```

```
2^25 # exponent
```

```
## [1] 33554432
```

```
25 * (12 + 27)
```

```
## [1] 975
```

Funções Matemáticas em R

Função	O Que Ela Faz
<code>abs(x)</code>	valor absoluto de x
<code>sqrt(x)</code>	raiz quadrado de x
<code>log(x)</code>	logaritmo natural (naperiano) de x
<code>exp(x)</code>	exponente natural de x
<code>log10(x)</code>	logaritmo base 10 de x
<code>round(x, n)</code>	arredondar x para n casas decimais

Funções em Operação

```
sqrt(9849)
```

```
## [1] 99.24213
```

```
log(377898)
```

```
## [1] 12.84238
```

```
exp(12.84238)
```

```
## [1] 377898.2
```

```
log10(377898)
```

```
## [1] 5.577375
```

```
round(exp(12.84238), 0)
```

```
## [1] 377898
```

Sobre `log()` e `exp()`

- No exemplo acima, exponente do 12.84238 é 377898.2, não 377898
- R relata 5 casas decimais na tela
 - Internamente, é 12.8423795969182 (13 casas decimais)
- Sabemos que $\log(x) = e^x$
- Não quebramos as leis da matemática.

```
x <- 377898
y <- log(x) # calcular o log de x e atribuir a y
y
```

```
## [1] 12.84238
```

```
exp(y)
```

Comentários

- Linha 2 do script tem um comentário à direta
- Comentários começam com um hashtag #
 - Tudo à direta do # não será interpretado (executado)
- Comentários nos lembra o que fizemos e porque
- **Hiper importantes**
- Usam eles MUITO

Ordem de Cálculo (**PEMDAS**)

Operation	Symbol	Example	PEMDAS
parentheses	()	$5 * (7 + 2) = 45$	P
exponents	\wedge	$5^2 = 25$	E
multiplication	*	$5 * 7 = 35$	M
division	/	$25/5 = 5$	D
addition	+	$5 + 7 = 12$	A
subtraction	-	$5 - 7 = -2$	S

- Se você retirar os parênteses de $5 * (7 + 2)$?
- $5 * 7 + 2 = 37$
- VSS: regras de matemática não mudam porque usamos um computador

Atribuição

- (nome de objeto) <- (definição do objeto)
- definição = valores que compõem o conteúdo do objeto

Atribuição - Estilos

- Esses servem

```
x <- 6  
x <- "Hi!"
```

- Esses funcionam mas não recomendo e não uso

```
x = 6  
6 -> x
```

- Esse produz um erro (não pode iniciar um comando com um número)

```
> 6 = x  
Error in 6 = x : invalid (do_set) left-hand side to assignment  
> |
```

Mensagens de Erro Estranhas?

- Consulte Dr. Google

A screenshot of a Google search results page. The search bar at the top contains the query "invalid (do_set) left-hand side to assignment". Below the search bar, there are several navigation links: Todas (selected), Vídeos, Notícias, Shopping, Maps, Mais, Configurações, and Ferramentas. A message indicates "Aproximadamente 1.340 resultados (0,40 segundos)". The first result is a link to a Stack Overflow post titled "invalid (do_set) left-hand side to assignment in R - Stack Overflow" with the URL <https://stackoverflow.com/.../invalid-do-set-left-hand-si...>. It includes a "Traduzir esta página" option. Below the link, it says "1 resposta" and provides a snippet of the answer: "22 de mai de 2017 - Problem is with your basics of R knowledge. You cannot name your parameter starting with number. And if you want to start it try this". There are several other search results listed below, each with a title, a snippet, and a timestamp.

Título da busca	Respostas	Data
"invalid (do_set) left-hand side to assignment ..."	1 resposta	5 de jun de 2017
pass a list by reference in R function	2 respostas	5 de jul de 2017
Making variables immutable in R	1 resposta	16 de ago de 2019
r dplyr transmute_string input error	2 respostas	17 de nov de 2016

Atribuição -- Nomes das Variáveis

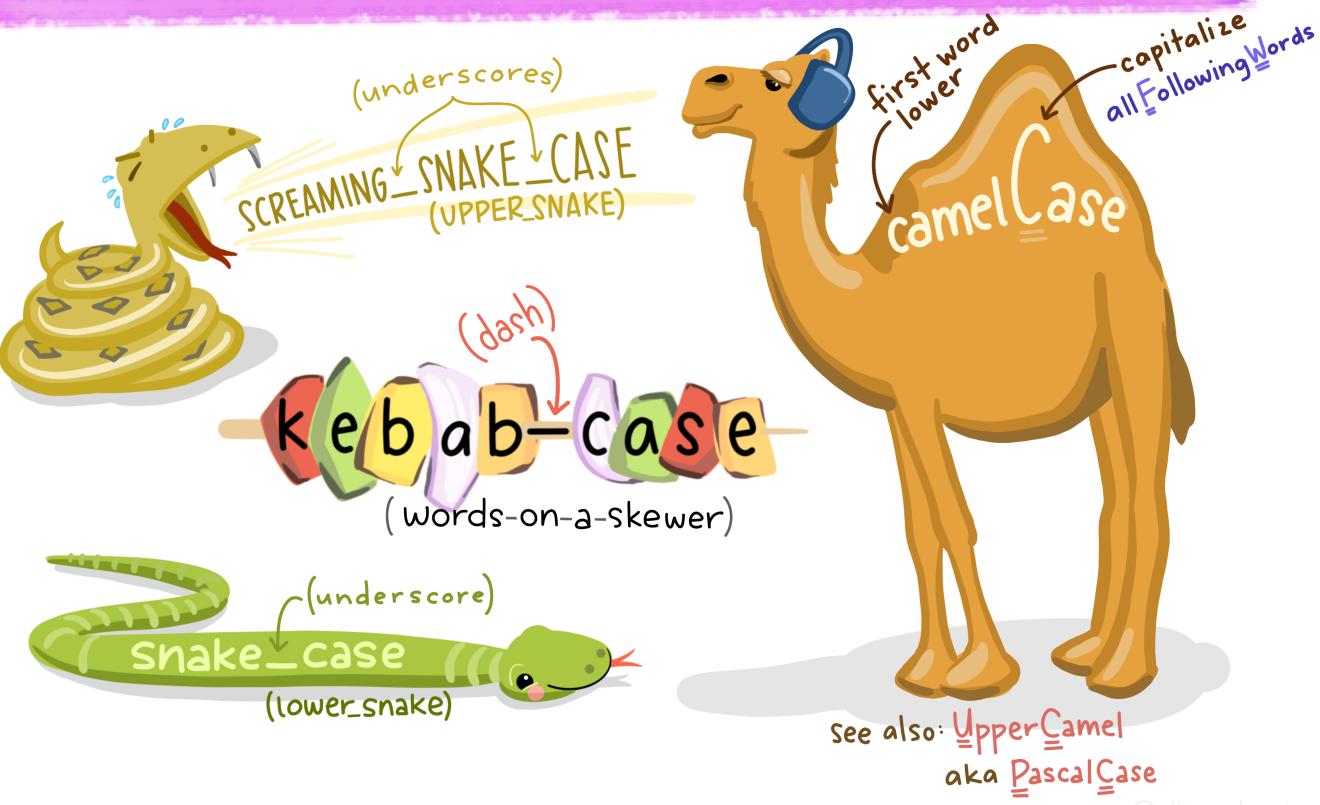
- Regras importantes de R
1. Deve usar só letras (maiúsculas ou minúsculas), números ou símbolos . or _.
 2. Deve iniciar o nome com uma **letra**

Nomes das Variáveis -- Corolários

- Não devem incluir espaços
 - "Snake case" supera essa restrição
 - Conectar palavras com sublinhar "_"
- Palavras reservadas de R não podem ser usados como nomes de variáveis
 - Exemplos: TRUE, FALSE, if, else, for, function
- Nomes de variáveis diferenciam maiúsculas de minúsculas
 - Variable e variable são 2 nomes diferentes
 - Mesmo para x e X

Casos em R

in that case...



Nomes das Variáveis - Ainda Mais

- Usar nomes claros e informativos
 - x, apesar de ser popular, é inútil como um nome

```
## 1a versao
peso <- 55 ## Pessoa pesa 55 kg.

## 2a versao
peso_kg <- 55 ## Mais claro

## 3a versao, pode converter às libras
peso_lb <- peso_kg * 2.2
peso_lb

## [1] 121
```

Nomes das Variáveis - Último

- Faça um diccionário dos dados
 - Tabela dos nomes das variáveis, qual tipo de dados, e o intervalo dos valores
- Tente de fazer os nomes mais curtos possíveis
- "*Camel case*" - alternativa a "*snake case*"