R!

Prof. Dr. Alexandre Chiavegatto Filho

Faculdade de Saúde Pública
Universidade de São Paulo

Introdução ao R para a Análise de Dados – EPI5713
Docentes Responsáveis: Alexandre Chiavegatto Filho, Ana Paula Sayuri Sato, Gleice Conceição.
Docentes Colaboradores: Francisco Chiaravalloti Neto, Dirce Zanetta.
Sala: Laboratório de Informática da FSP – 2º Andar.

03/06 (4af.)	**Introdução** Instalação do R, RStudio, estrutura dos dados, pacotes e abertura de bancos de dados de outras fontes. Prof. Alexandre
08/06 (2af.)	**Gráficos com o ggplot2** Gráficos de dispersão, pirâmides populacionais, boxplots e histogramas. Prof. Alexandre
10/06 (4af.)	**Bancos de dados** Criação, renomeação e exclusão de variáveis, alteração de valores, identificação de valores missing, alteração do tipo de variável, fusão de bancos de dados. Prof. Alexandre
15/06 (2af.)	**Análise descritiva** Descrição das variáveis, análise de frequência e tabelas bivariadas. Profa. Gleice

Introdução ao R para a Análise de Dados – EPI5713
Docentes Responsáveis: Alexandre Chiavegatto Filho, Ana Paula Sayuri Sato, Gleice Conceição.
Docentes Colaboradores: Francisco Chiaravalloti Neto, Dirce Zanetta.
Sala: Laboratório de Informática da FSP – 2º Andar.

17/06 (4af.)	**Análise bivariada** Chi-quadrado, teste-t e correlação. Profa. Gleice
22/06 (2af.)	**Regressão** Regressão linear e logística. Profa. Gleice
24/06 (4af.)	**Mapas** Shapefiles e visualização de dados em mapas. Prof. Francisco
29/06 (2af.)	**Entrega e apresentação dos trabalhos**

AVALIAÇÃO

Trabalho: Apresentação de resultados preliminares da sua pesquisa com os conceitos aprendidos na disciplina. Peso 7.
Exercícios durante as aulas: Peso 3.

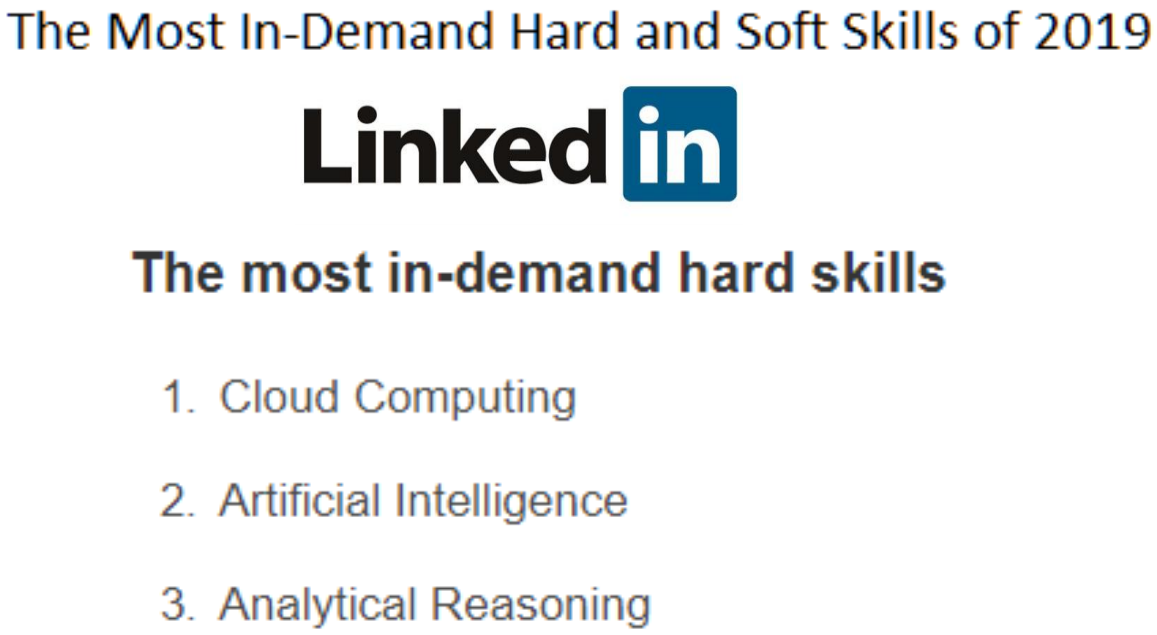
A ascensão do cientista de dados

- Aumento da disponibilidade de dados:
  - Empresas querem conhecer melhor seus clientes e aumentar seus lucros.
  - Cientistas querem desenvolver novos produtos e fazer novas descobertas.

- Antes: decisões baseadas em chutes.
- Hoje: cada vez mais baseadas em dados.
  - “Sábios” vs. Cientistas de dados.
    - Os “sábios” (normalmente a pessoa mais velha de um grupo) estão sendo rapidamente trocados pelo cientista de dados (o profissional capaz de analisar os dados e tomar uma decisão).

Tem também crescido o interesse em cientistas de dados pelo mercado

Segundo a consultoria McKinsey: EUA terão déficit de 140 a 190 mil cientistas de dados até 2018.



CARREIRA - VOCÊ S/A

Cientista de dados: a profissão do futuro continua em alta

Profissionais que atuam com métodos estatísticos e computacionais são cada vez mais disputados pelas empresas

Por **Abril Branded Content**
27 maio 2019, 09h00

2019:
- Salário médio no Brasil: 9 000 reais, podendo chegar a 20 000 reais.

EXAME.COM

Melhor profissão do ano nos EUA deve explodir no Brasil. Entenda

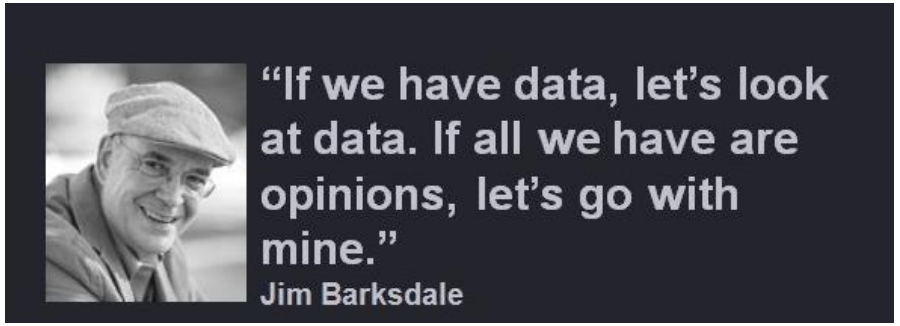
No mercado americano, a carreira de estatístico registra altos níveis de satisfação em quesitos como renda e perspectivas de ascensão. Veja como é no Brasil

Por **Claudia Gasparini**
17 maio 2017, 09h25 - Publicado em 16 maio 2017, 15h00

- Estatístico já é a segunda profissão com maior salário no Brasil.
- Empresas e governo: pressão por menos decisões baseadas em “achismos”.
- Imune à crise econômica (até ajudou, mais pressão por decisões mais eficientes).

Área acadêmica

- Ciência hoje é análise de dados.
  - Alternativa é opinião.

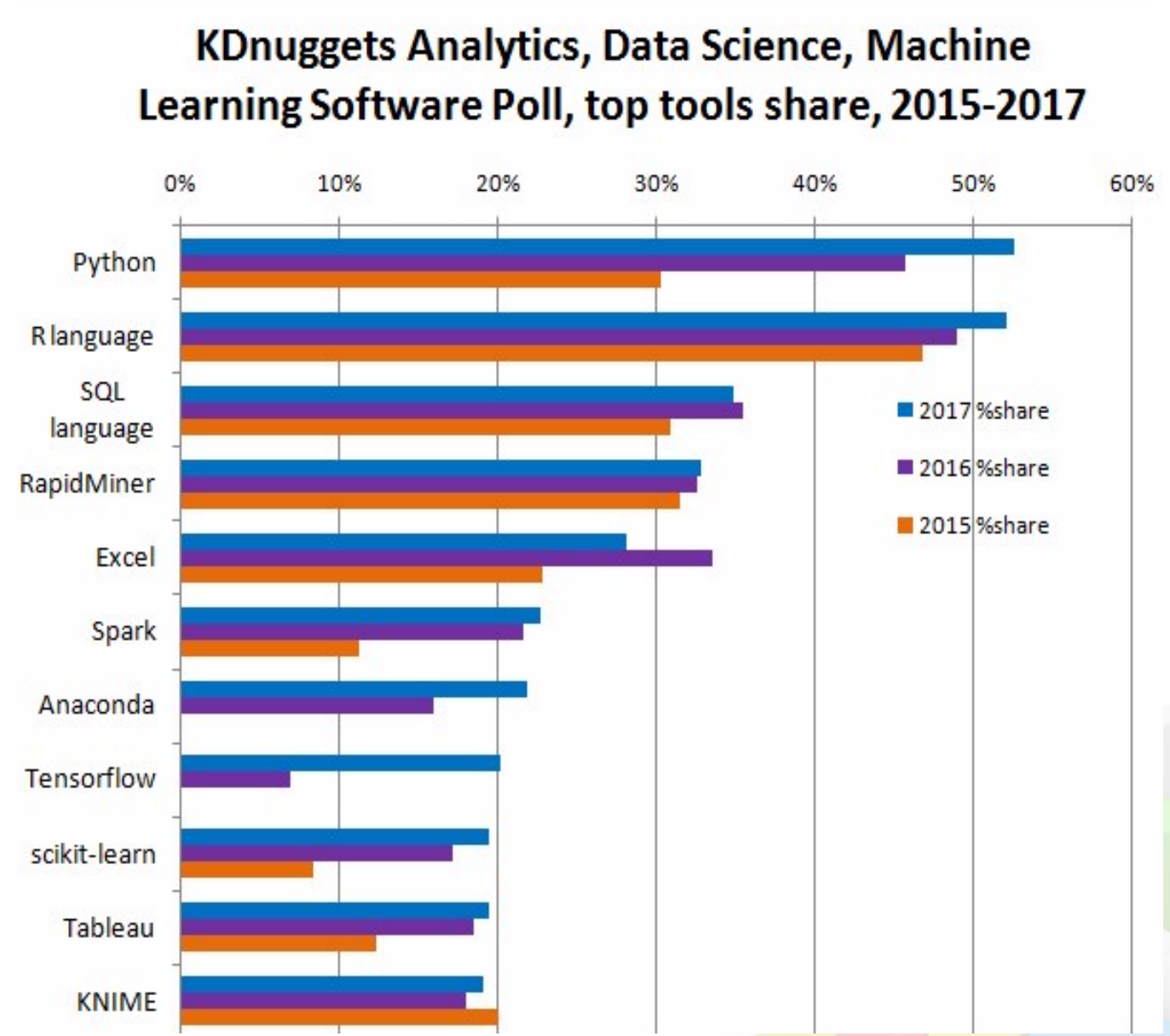


- Faça um teste: procure as especialidades dos <u>novos</u> professores das melhores faculdades do mundo
  - A grande parte dos jovens professores é especialista em análise de dados.

- Sonho:
  - Consenso sobre qual linguagem/software utilizar em ciência de dados!
    - Evitar ter de aprender várias linguagens de programação.

- Realidade:
  - Caminhando para consenso no uso do R e Python.
  - Gratuitos e com comunidade ativa de jovens programadores.
  - Estamos na torcida para que virem consenso.
    - Estamos fazendo a nossa parte!

- Kdnuggets annual poll:
- 2017: 2.900 cientistas de dados.
- “Quais softwares você usou para analisar dados nos últimos 12 meses?”
- R: 52,1%.
- Python: 52,6%.



- Linguagem de programação:
  - Muito fácil fazer alterações nas análises (1 linha de código).
  - Facilita análises colaborativas.
  - Garante reprodutibilidade dos resultados em novas amostras.
- Gratuito.
- Comunidade ativa de programadores.
- Foco na análise de dados.

- Linguagem de programação:
  - Muito fácil fazer alterações nas análises (1 linha de código).
  - Facilita análises colaborativas.
  - Garante reprodutibilidade dos resultados em novas amostras.
- Gratuito.
- Comunidade ativa de programadores.
- Linguagem de programação geral.

nature International weekly journal of science

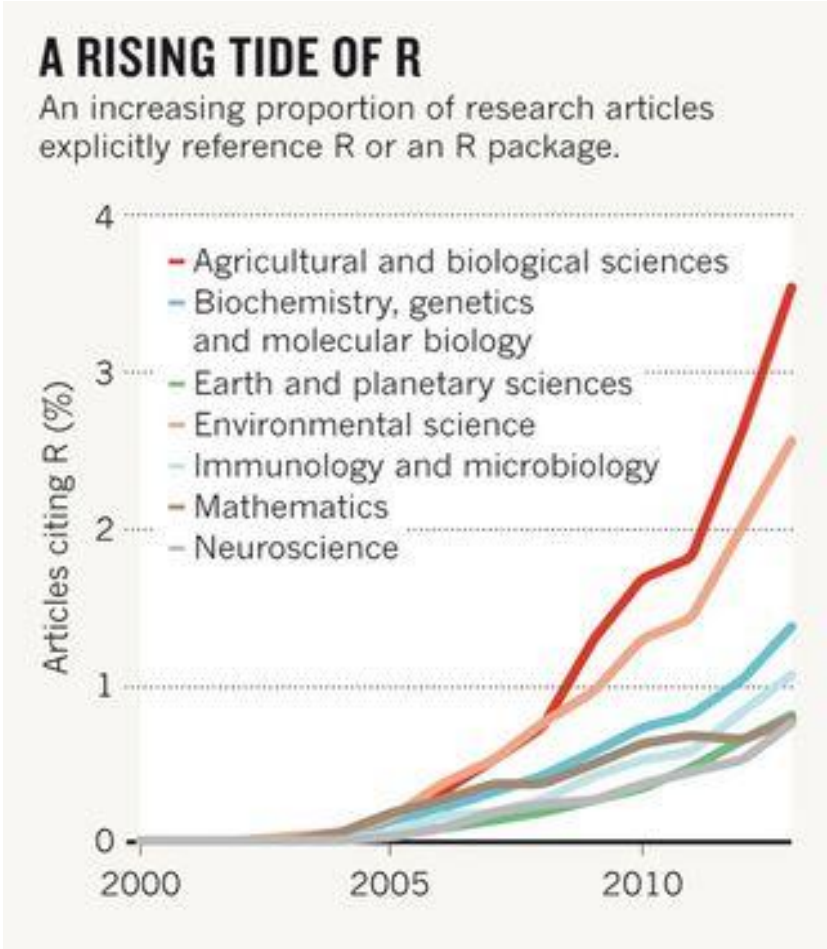
NATURE | TOOLBOX

Programming tools: Adventures with R

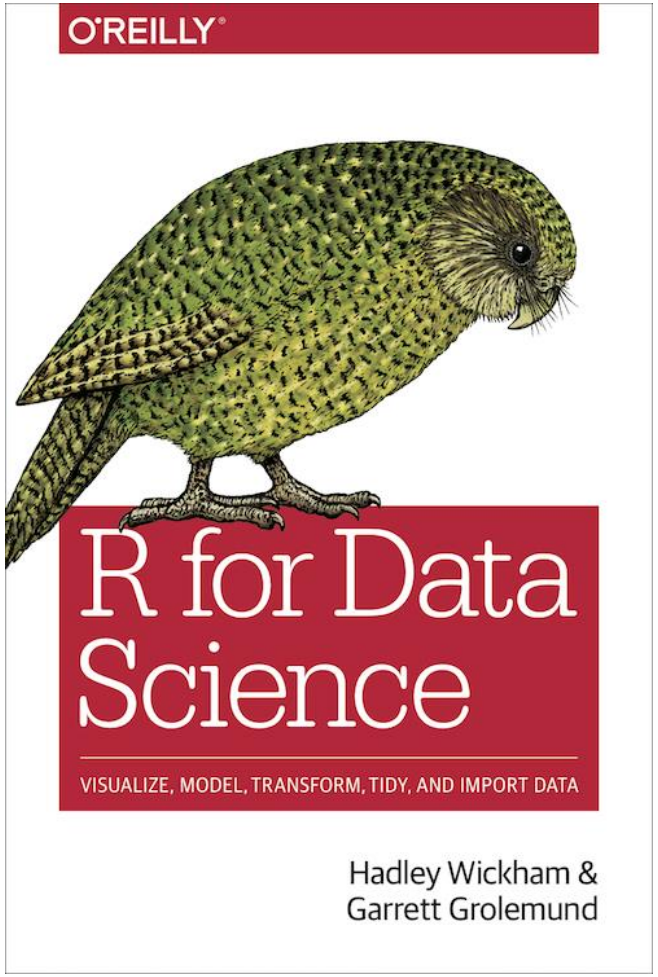
A guide to the popular, free statistics and visualization software that gives scientists control of their own data analysis.

Sylvia Tippmann

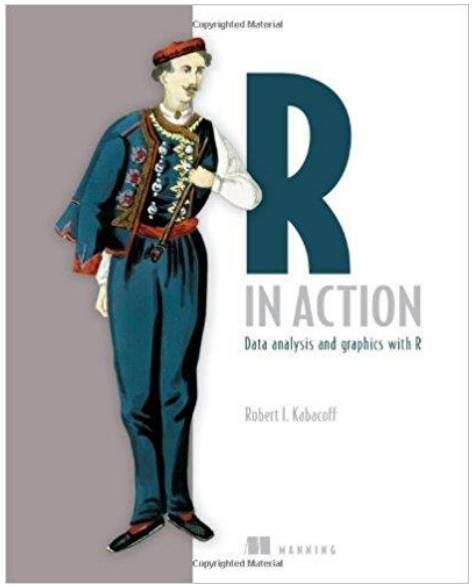
29 December 2014

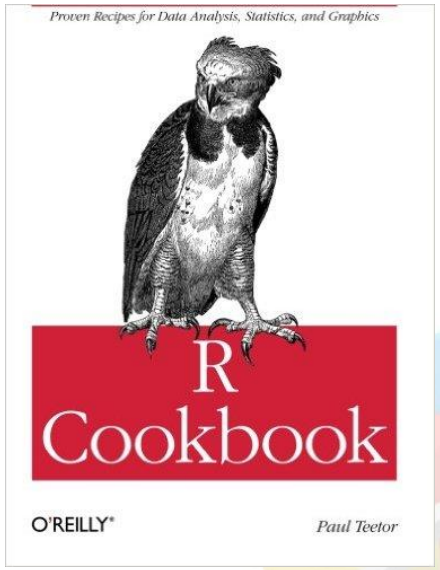


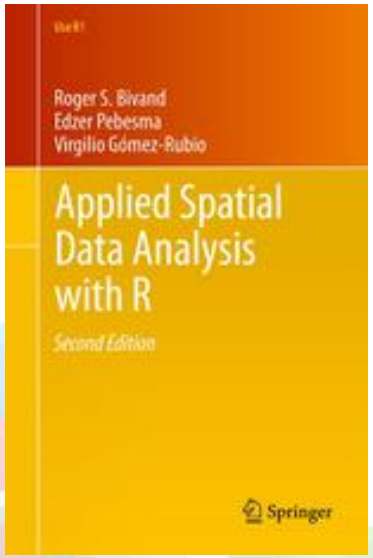
Livro texto



Livros complementares





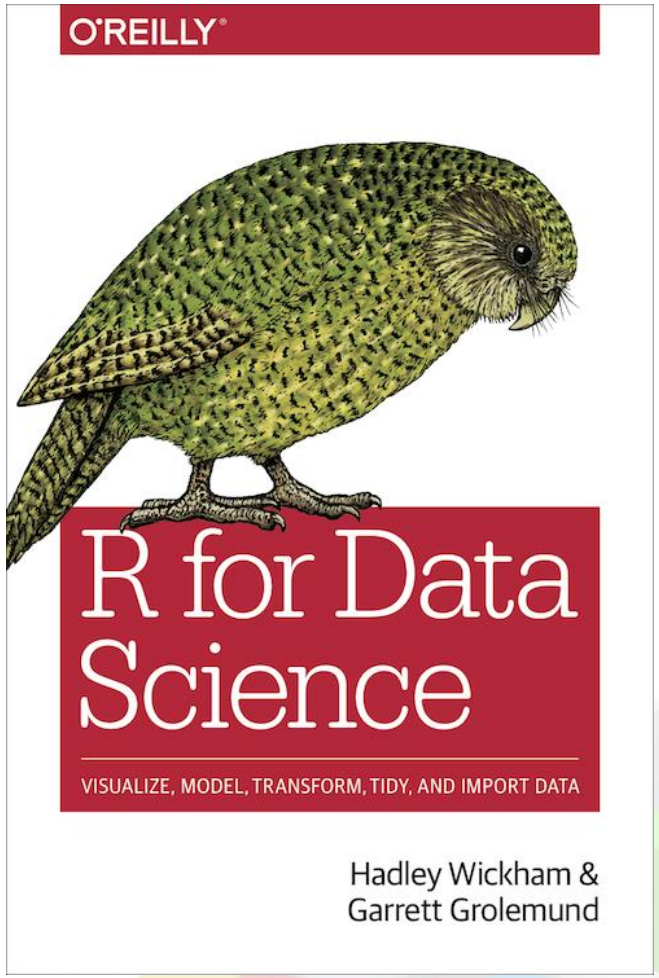


- Disponível completamente online (recomendamos a compra):

  http://r4ds.had.co.nz

- Utiliza o tidyverse (anteriormente conhecido como hadleyverse).

Hadley Wickham

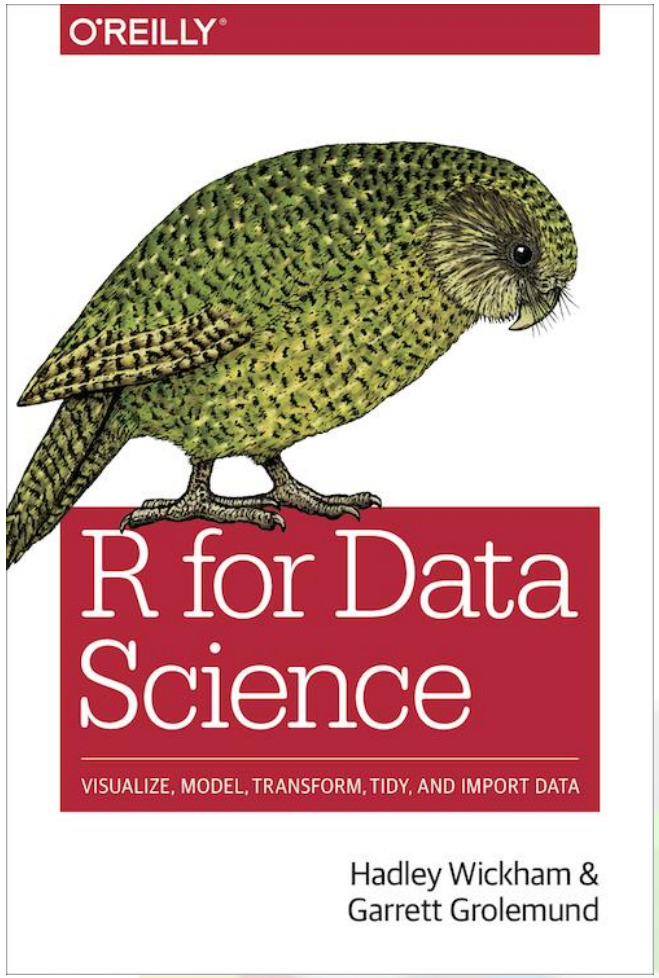


- Tidyverse: conjunto de pacotes que permitem a importação, limpeza e visualização de dados.

- Resolvem um problema importante do R: pacotes com estruturas de comandos diferentes.

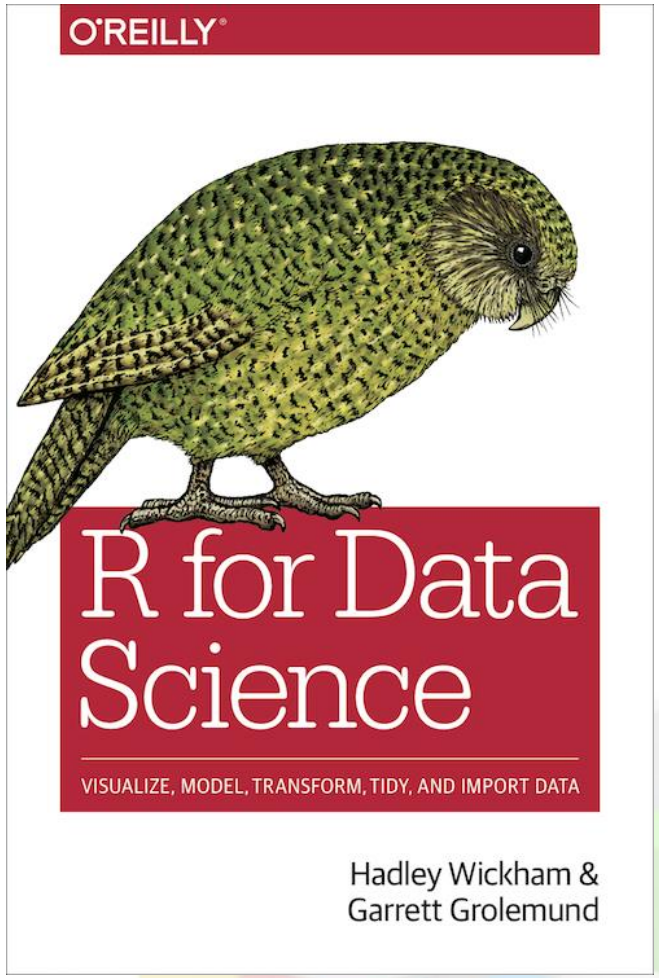
```
install.packages("tidyverse")
```

Instala automaticamente:
- readr: importar dados.
- tidyr: limpeza de dados.
- dplyr: manipulação dos dados.
- ggplot2: visualização de dados.
- Entre outros...

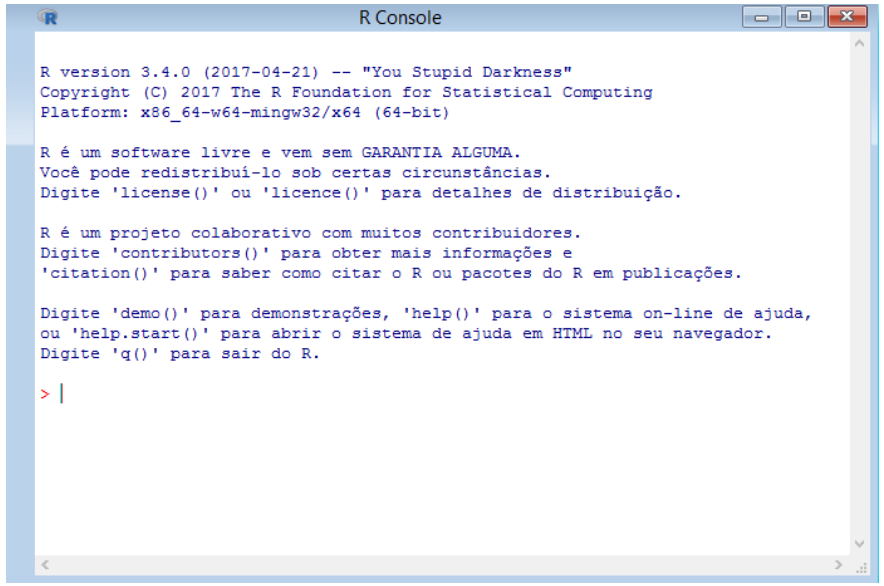


- Existia uma grande demanda.

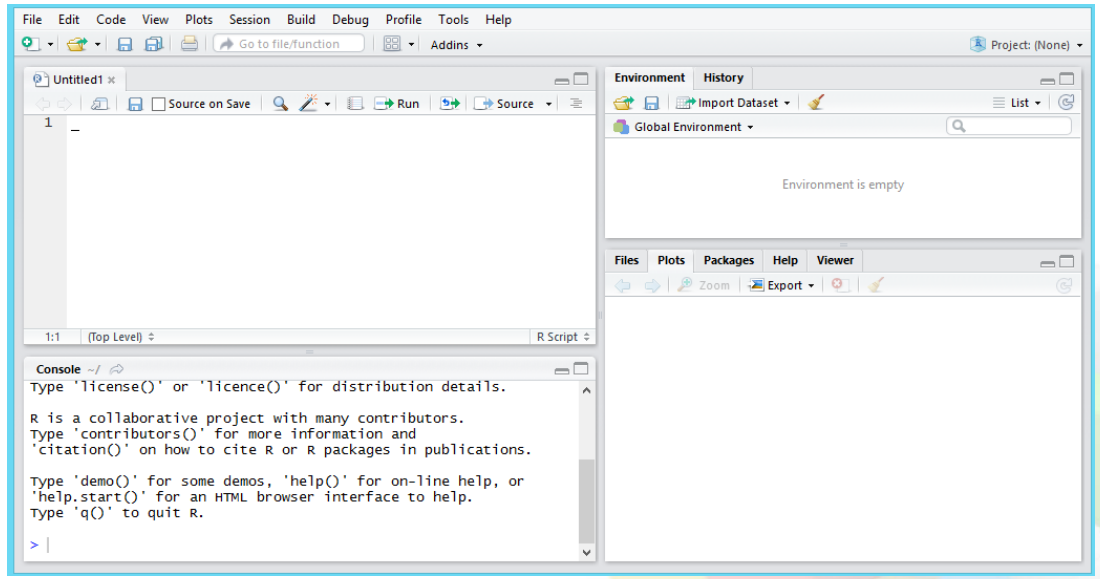
- Positiva e rápida absorção pela comunidade de R:
  - Lançado em janeiro 2017.
  - Bestseller na Amazon em ciência de dados.
  - Cursos da USP: verão do IME, veterinária, FSP...



- Baixar o R.
- Última versão: 4.0.0.
- https://cran.r-project.org/

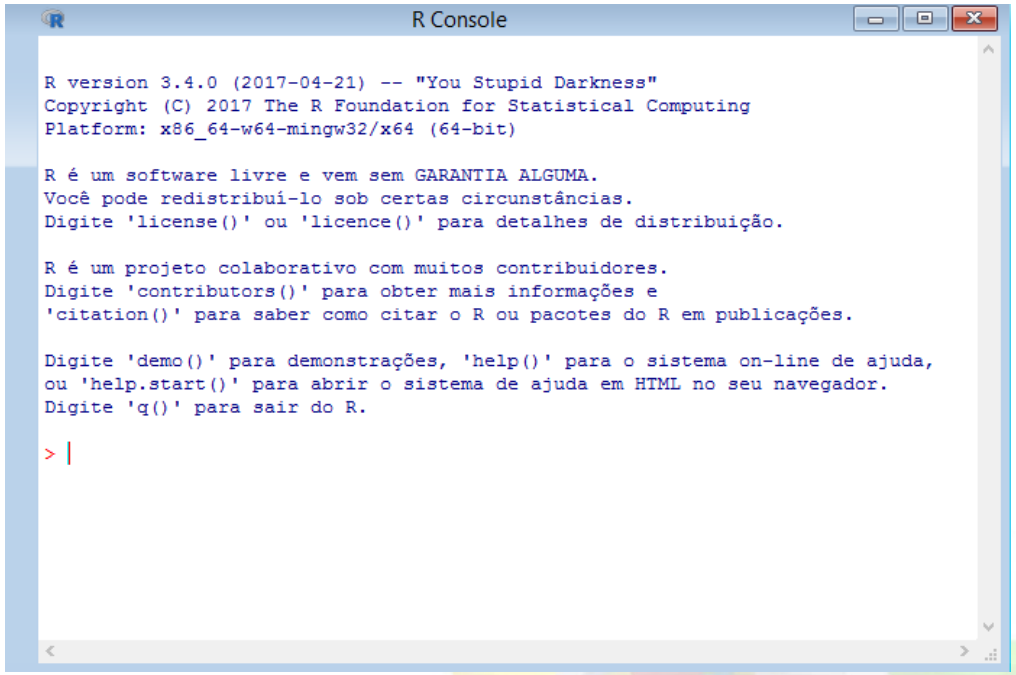


- Baixar o RStudio.
- Última versão: 1.3.959
https://www.rstudio.com/products/rstudio/download/



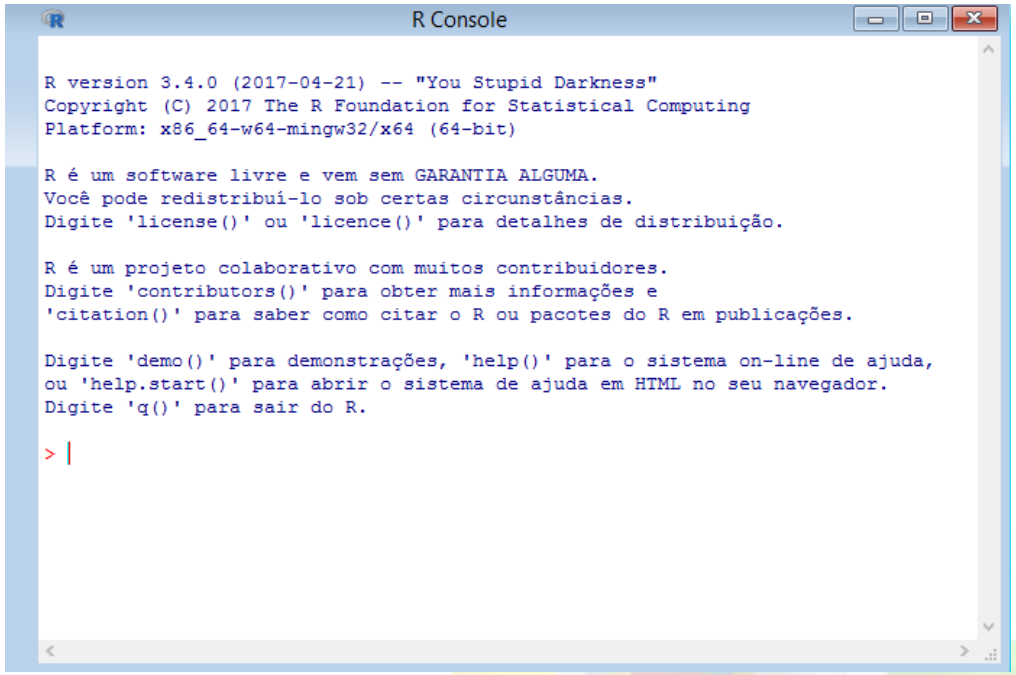
R

- Linguagem de programação.
- Software para rodar códigos.
  - Software composto por uma única janela (“Console”).
  - O software do R é pouco utilizado na prática.



R

- Vem apenas com um conjunto básico de pacotes.
  - Pacote: conjunto de funções, dados e códigos com um objetivo em comum (análise espacial, testes psicológicos...).
  - É necessário instalar os pacotes específicos de interesse.



R

- Para utilizar um pacote é necessário:

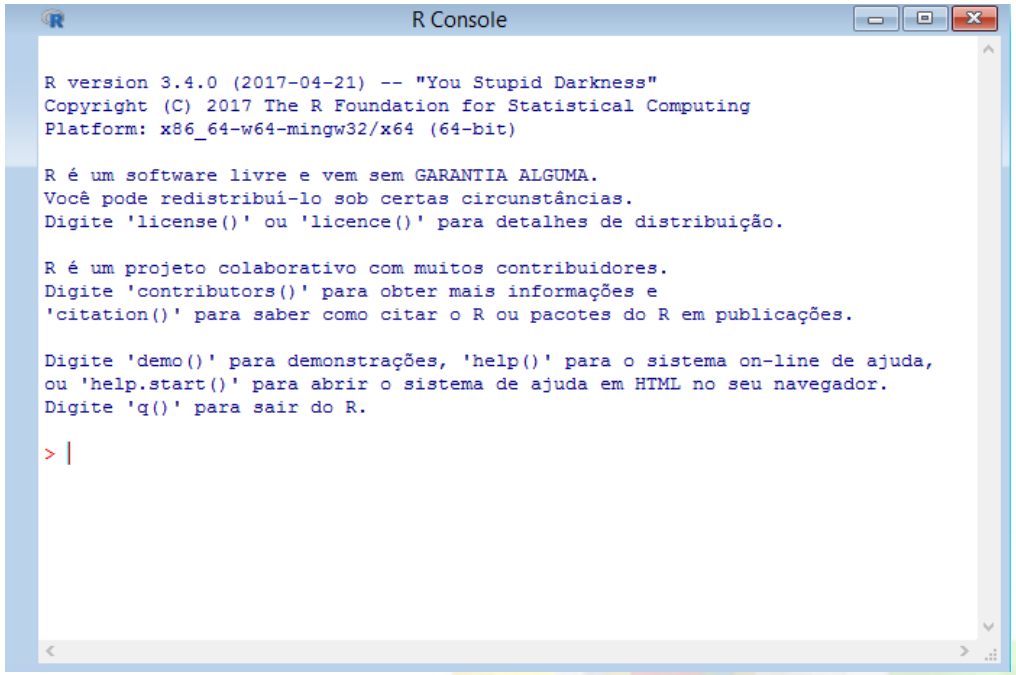
1 – Baixar o pacote (apenas uma vez).

install.packages (“tidyverse”)

2 – Chamar o pacote (ao início de toda sessão).

library (tidyverse)

- Ver todos baixados: library()
- Ver todos chamados: (.packages())

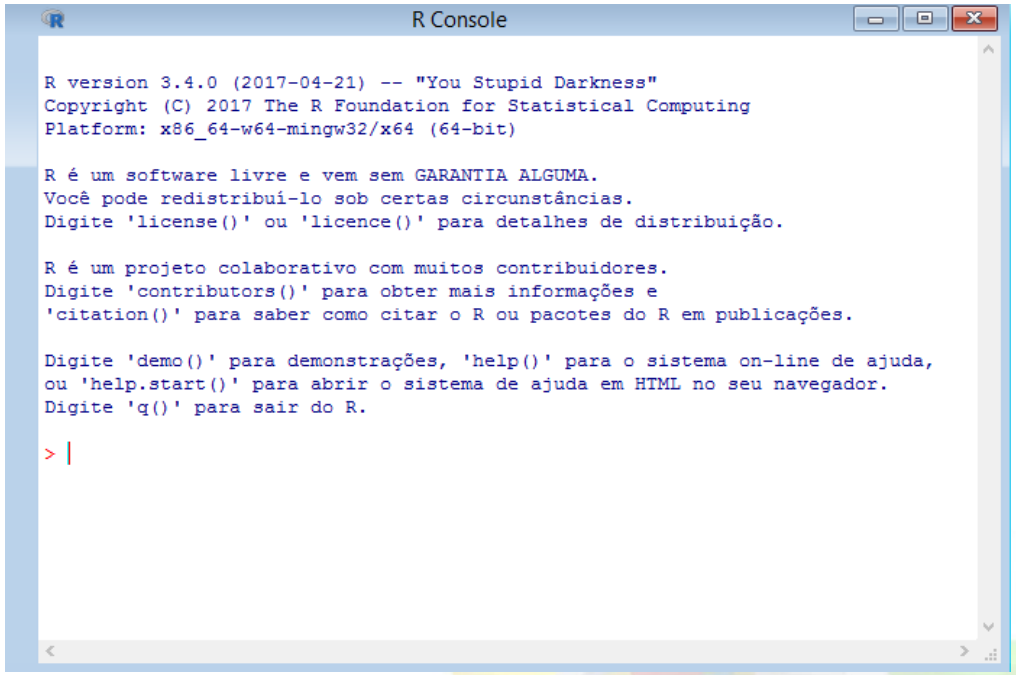


R

- A linguagem do R diferencia maiúscula de minúsculas.

- Missing no R: NA
  - Stata: .
  - MlWin: *

- Objetos criados em uma sessão são armazenados apenas temporariamente.



R

- Estrutura fundamental dos dados:
  - Vetores: conjunto de valores de um mesmo tipo:
    - Principais:
      - Logical: TRUE ou FALSE.
      - Numeric: valores discretos ou contínuos.
      - Character: categorias ou palavras.
      - Missing.

R

- Dataframe: conjunto de vetores com o mesmo número de observações.
  - Podem ter vetores de diferentes tipos.
  - Equivalente à planilha do Excel.

- Listas: não precisam ter o mesmo tamanho.

- Função mais simples do R: calculadora.
  - Digitar após “>”
  - Calcular IMC de pessoa com 89kg e 1,76m:
    89/1.72^2
  - Se “+” é porque faltou alguma coisa.

RStudio

- É um ambiente de desenvolvimento integrado (IDE, em inglês).
- Software que possibilita um ambiente didático para programar e visualizar resultados.
- Necessita ter o R baixado.