



Introdução à programação em R para jornalismo

1ª parte - Não deixe o R te intimidar

Link para esta apresentação
bit.ly/codabelem_23_introR




Você se enquadra em alguma situação abaixo?

- Sou de humanas, não de exatas kkkk
- Eu era ruim em matemática
- Já tentei aprender R no passado, mas esqueci tudo
- Sei só um básico de Excel
- Sou apenas um jornalista.

98% das reportagens com dados podem ser apuradas com Excel e Google Sheets. Para todas as outras, utilize o R.

—



Quando devo usar o R em uma apuração?

- Quando algo fica muito complexo e você pode perder o fio da meada facilmente, correndo grandes riscos de cometer erros.
- Quando a escala da sua apuração supera a capacidade das planilhas de lidar com seus dados.
- Quando você quer documentar uma metodologia para que outras pessoas possam seguir seus passos. O R é reprodutível.
- Quando você tem uma hipótese complexa e precisa descobrir o caminho para apurá-la enquanto apura.

Vamos decifrar juntos

Vocês conseguem a partir do texto dessa reportagem finalista do Pulitzer, descrever pra mim como isso foi apurado???

The Washington Post
Democracy Dies in Darkness

THE AMAZON, UNDONE

THE GOD OF SÃO FÉLIX

He's been called a deforester and killer. Now he's called mayor.

are, in fact, the representatives of those who deforest.

“It has reached such an absurd point that once, during an active investigation in Rorainópolis, in the south of Roraima, the mayor came to the police station with the people we were investigating and asked me to dismiss the case,” he said.

The Post analysis identified 1,189 officeholders over the past 20 years in the Amazon who have been cited for federal environmental infractions. Many won more than one election, and more than 3 in 4 were accused of deforestation or a deforestation-related offense. The examination, which analyzed all environmental infractions in federal databases and anyone who had held municipal, state or gubernatorial office, found that at least one-third of the politicians were cited for environmental abuse while they held office.

Because of data limitations, the findings are almost certainly an undercount.

were owners of companies that had been.

The amount of money donated to Amazon political campaigns by people and companies that have been cited for environmental infractions is far more than previously known. More than 1,590 people and 717 companies cited for environmental wrongdoing made at least 5,546 contributions over the past two decades, amounting to nearly \$37 million. (In 2015, Brazil prohibited campaign donations by companies.)

The names of the politicians and the cities they have governed provide a road map through the most deforested swaths of the Amazon. Many are within what's known as the "arc of deforestation," a section of more-intense deforestation along the forest's southern sweep. Many cities in this arc — Novo Progresso, Feliz Natal, Cotriguaçu — have repeatedly elected officials accused of environmental wrongdoing.



Vamos começar a primeira parte da nossa apuração

Entrar no site do ICMBIO e pegar uma base de dados

Abra o arquivo R chamado PERGUNTAS.

<https://www.icmbio.gov.br/cpb/index.php/conteudo-do-menu-superior/31-dados-abertos>

Ué, cadê os dados?



Vamos tentar de novo

Buscar novamente no google

dados abertos icmbio

<https://www.google.com/search?q=dados+abertos+icmbio&oq=dados+abertos+icmbio&aqs=chrome.0.0i512j69i57j69i64j69i60l3.4057j0j4&sourceid=chrome&ie=UTF-8>



O que o R é..... e o que ele não é

- Uma calculadora diferente
- Uma ferramenta estatística
- Um estagiário super poderoso, que faz tarefas repetitivas sem errar, mas fala uma língua diferente da sua
- Uma ferramenta de **apuração de hipóteses**

- vai substituir suas fontes
- fazer a matéria para você
- fazer a matemática para você
- dizer se a sua apuração está certa ou não



O que é possível fazer com R?

- Análise de dados - Estatística, modelagem, etc.
- Visualização de dados
- Apresentações
- Relatórios dinâmicos
- Escrever livros, blogs, tutoriais estruturados
- Mineração de dados
- **Apurações para reportagens**
- Mucho más...



Porque o R pode ser útil a jornalistas?

Para quê se esforçar tanto para aprender uma linguagem de programação?

- Fazer apurações cuja escala e/ou complexidade inviabilizam de realizar no Excel ou Google Sheets.
- Descobrir o que você não notou antes.
- Confrontar fontes técnicas.
- Trazer novas perguntas e abordagens.
- Testar hipóteses que **não** estão nos sumários das assessorias de imprensa.
- Confrontar erros dos órgãos públicos.



Por quê usar linguagens de programação para analisar dados?

As vantagens de analisar dados usando linguagens de programação

- O código é uma linguagem, então podemos documentar a nossa análise (e sermos transparentes, fiscalizados e cobrados por isso)
- O código é texto, então podemos copiar e colar
- As principais linguagens de programação para ciência de dados são de código aberto (nesse caso, grátis)



O que é? (analogia com culinária)

Rstudio - é como uma cozinha

Scripts - como uma receita de bolo

Projetos - a vasilha em que o bolo será preparado

Pacotes - são como catálogos de temperos e procedimentos para você usar.

Funções - são os procedimentos que você faz para avançar na sua receita de bolo. (verbos)

Dataset, dataframe ou base de dados - é o ingrediente que você vai transformar no seu bolo.



Boas práticas

Os nomes das suas variáveis devem fazer sentido: Ao nomear suas variáveis, dê nomes que tenham significado para seres humanos. Pense que o código que você escreve hoje deve ser claro para você daqui 1 ano e também deve ser claro para algum(a) colega seu(ua).

Comente bem o seu código: É possível fazer comentários usando o símbolo '#'. É sempre bom explicar o que uma variável armazena, o que uma função faz, porque alguns parâmetros são passados para uma determinada função, qual é o objetivo de um trecho de código, etc.

Evite linhas de código muito longas: Usar linhas de código mais curtas ajuda na leitura do código.

Escreva um código organizado: Por exemplo, adote um padrão no uso de minúsculas e maiúsculas, uma lógica única na organização de pastas e arquivos, pode ser adotada uma breve descrição (como comentário) indicando o que um determinado script faz.



R como calculadora

Tente fazer isso no seu console

- O R permite realizar muitas operações aritméticas em seu console!

```
2 + 5      # adição
## [1] 7
9 - 4      # subtração
## [1] 5
5 * 2      # multiplicação
## [1] 10
7 / 5      # divisão
## [1] 1.4
```



R como calculadora

Tente fazer isso no seu console

```
9 %% 4 # resto da divisão de 9 por 4
## [1] 1
7 %/% 4 # parte inteira da divisão de 7 por 4
## [1] 1
8 ^ 2 # potenciação
## [1] 64
sqrt(1024) # radiciação
## [1] 32
```

A ordem matemática das operações também vale no R.



Desafio

traduzindo seu pensamento para a máquina

Como você resolveria esse problema usando o R?

O Rei Théoden tem 7 dias para chegar a Minas Tirith e salvar a cidade das forças da Sombra. Sua guarda possui 600 homens e uma mulher escudeira e sabe que, se esperar, os cavaleiros leais virão se juntar ao exército. Entretanto, movimentar as tropas leva 3 dias, se chegar atrasado, não haverá nada para ser salvo.

58 cavaleiros chegam por hora ao acampamento. Théoden espera conseguir 13 mil lanças totais para a batalha. Qual o tamanho real do exército de Théoden?

$$600+1+(58\times 24\times (7-3))$$



Desafio real iniciante

Como você responderia a essa pergunta do editor?

Qual a unidade de conservação que mais sofreu desmatamento de corte raso em 2021???



Operadores relacionais

- Igual a: ==

Exemplos:

```
TRUE == TRUE
## [1] TRUE
TRUE == FALSE
## [1] FALSE
```

- Diferente de: !=

Exemplos:

```
TRUE != TRUE
## [1] FALSE
TRUE != FALSE
## [1] TRUE
```




Operadores relacionais

- Maior que: >
- Maior ou igual: >=
- Menor que: <
- Menor ou igual: <=

Exemplos:

```
3 < 5
## [1] TRUE
10 >= 10
## [1] TRUE
10 > 10
## [1] FALSE
```



NA not available

Uma característica importante do R que pode dificultar a comparação são os valores ausentes ou **NA**s (não disponíveis).

NA representa um valor desconhecido.



Classes básicas ou atômicas

São os tipos básicos de dados que podem ser representados na linguagem R. Em uma tabela, cada coluna é uma variável e pode ser de uma classe. Um dataset pode carregar variáveis de diferentes classes.

- **Integer:** números inteiros
- **Numeric:** números racionais
- **Complex:** números complexos (raramente usados para Análise de Dados)
- **Logical:** TRUE, FALSE ou NA
- **Factor:** variáveis categóricas
- **Character:** texto puro e está sempre entre aspas.



O que é uma variável

- Ao se desenvolver um projeto, você irá trabalhar com diversos tipos de arquivos, além de informações que serão repetidas ao longo do script.
- Para reutilizar essas informações ao longo do script utilizamos o que chamamos de **variável**
- Uma variável é um espaço de memória que retém e representa um valor ou expressão
- Em uma tabela bidimensional, cada coluna é uma variável



Tipos de objeto

- **Vetor:** armazena elementos de mesma classe.
- **Matriz:** vetores de duas dimensões que armazenam elementos de mesma classe.
- **Lista:** tipo especial de vetor que aceita elementos de classes diferentes.
- **Data.frame:** são tabelas de dados com linhas e colunas, como uma tabela do Excel. Como são listas, essas colunas podem ser de classes diferentes.



Funções úteis para utilizar com dataframes

- `tail()`: mostra as últimas 6 linhas.
- `names()`: mostra os nomes das colunas.
- `View()`: mostra o dataframe
- `glimpse()`
- `describe ()`

Conheça a flecha



para atribuir coisas a um objeto

Conheça o pipe

>%>

encadeia o resultado de uma função com a próxima função. Equivale ao “isso então aquilo” da nossa linguagem falada



Operador pipe >%>

Imagine uma receita que tenha as instruções: junte os ingredientes, misture e leve ao forno. Na forma usual do R, essas instruções provavelmente seriam assim:

```
forno(misture(junte(ingredientes)))
```

Dessa forma temos que pensar “de dentro para fora”. O primeiro comando que lemos é forno, sendo que essa é a última operação que será realizada. Com o operador pipe seria algo assim:

```
ingredientes %>% junte %>% misture %>% forno
```

É mais intuitivo!



Introdução à programação em R para jornalismo

2ª parte - Um exemplo realista



Vamos colocar a mão na massa com dados ambientais?

Pacotes que vamos utilizar

Dplyr

data.table

Tidyr

Principais funções:

`select()`

`filter()`

`arrange()`

`group_by()`

Quando não entender o que faz uma função

digita o nome dela no google, junto com “r documentation”

Isso vai geralmente lhe dar acesso a uma explicação detalhada sobre como a função funciona e seus argumentos.



Expectativa vs realidade

**É preciso ter
paciência,
principalmente
quando estamos
aprendendo.**

As bases de dados são limpinhas durante cursos e exemplos, mas quando enfrentamos uma situação sozinhos elas podem parecer bem selvagens.

É preciso ter também uma intimidade anterior com a base de dados para navegar por ela com mais tranquilidade e confiança. Para ajudar nisso há os dicionários de dados.



Desafio final

Como você responderia isso usando uma planilha? Vamos construir um script, mexer com um dataset e aplicar funções nele.

Qual o bioma brasileiro que mais sofreu desmatamento em 2021? E em toda a série histórica?

Link para esta apresentação
bit.ly/codabelem_23_introR



Para aprender mais

- <https://github.com/R-Ladies-Sao-Paulo>
- <https://datalabe.org/dados-sem-cao/>
- <https://curso-r.com/material/>
- <https://journalismcourses.org/course/introitor/> (inglês)
- <https://learn.r-journalism.com/en/> (inglês)



Obrigada a todos

Cecília do Lago

jornalista de dados no Estadão

@ceciliadolago

ceciliadolago@gmail.com