R: Uma Breve Introdução

DataZoomLAB Team and Daniel AC Barbosa

28/03/2021

Introdução

Essas notas introdutórias visam familiarizar os membros do DataZoom com a linguagem de programação R, commeçando com exemplos simples até ir avançando ao fluxo de desenvolvimento (workflow) atual dos produtos desenvolvidos pelo LAB. As notas são baseadas em duas notas de aulas (McDermott 2021; Oliveira and Cavalcante 2021), asim como em livros escritos para cientistas de dados em R (Hadley Wickham 2017; Wickham 2019). Wickham (2015) é uma referência fundamental ao worflow do LAB, porém será discutida em notas posteriores.

Instalação

Siga as instruções aqui link

Introdução ao R

Referências:

- McDermott (2021) Lecture 1 - Introduction
- McDermott (2021) Lecture 4 - R language basics

Manuseio de Dados

Vamos adotar como paradigma o Tidyverse. Ele é um conjunto de pacotes para o R voltados a Data Science, tendo como livro introdutório o R for Data Science (Hadley Wickham 2017).

É importante mencionar que, mesmo com o Tidyverse tendo ganho amplo escopo dentro da comunidade (inclusive entre economistas adeptos de R, veja McDermott (2021)), ele é *um* dos estilos de programação possíveis na linguagem R, não o único. Para ciência de dados, uma outra possibilidade é o Data.Table.

Não obstante, nas palavras de Oliveira and Cavalcante (2021): "*tidyverse* é o caminho. O R base não deixa de ser nosso lar, mas já tem mais de 20 anos e algumas atualizações são mais que bem vindas. De fato, as soluções fora do tidyverse são quase sempre mais complexas, com sintaxe menos limpa e por isso, pouco amigáveis com iniciantes"

Referências:

- McDermott (2021) Lecture 5 - Data Wrangling Tidyverse

Referêncis Adicionais:

- R for Data Science - Explore Hadley Wickham (2017)
- R for Data Science - Wrangle Hadley Wickham (2017)
- Tidyverse Styling Guide

Escrevendo Funções

[EM CONSTRUÇÃO]

Extraindo Dados da Web (*Webscrapping*)

[EM CONSTRUÇÃO]

Workflow

Uma vez que o básico da linguagem R e do Tidyverse foram vistos, é hora de entender o fluxo de trabalho desenvolvido para a criação e manutanção dos produtos desenvolvidos pelo DataZoom.

Version Control

Primeiro, vamos entender o conceito de controle de versão (*version control*), amplamente usado na comunidade de ciência de dados para assegurar a estabilidade e integridade dos códigos desenvolvidos, asssim como facilitar que diferentes membros trabalhem paralelamente em um mesmo projeto.

Para tal, vamos olhar para o Git e para o GitHub. Veja que os links já diferem entre si, e, mais importante, os nomes já são informativos! Git-Hub, significa que a plataforma (Hub) hospeda desenvolvimento de códigos e auxilia no controle de versionamento (Git) de uma forma mais natural para nós humanos. O DataZoom fica hospedado no GitHub como organização (link), enquanto os produtos ficam armazenados na forma de repositórios (DataZoom Amazônia.

OBS: Os outros repositórios que não o DataZoom Amazônia, estão em **STATA**, uma outra linguagem de programação, a qual não será abordada nesse documento. Entretanto, os princípios de uso do Git/GitHub são os mesmos entre as diferentes linguagens e o aprendizado aqui pode ser extendido, resguardada as devidas diferenças de sintaxe entre as linguagens.

Referências:

- McDermott (2021) Lecture 2 - Version control with Git(Hub)

Essa é uma ótima introdução ao Git(Hub). Porém, como o Grant é usuário de sistemas operacionais Linux, ele dá muita ênfase ao uso do prompt de comando (ou terminal) – mais sobre isso abaixo. A minha experiência prévia é que pessoas iniciando no mundo da programação tendem a achar frustrante começar com o prompt e preferem algo mais *user-friendly*. Por isso, recomendo fortemente a instalação do GitHub Desktop, uma interface para o sistema de controle de versão muito amigável.

- GitHub Desktop Getting Started

Referências Adicionais:

Caso você queira se aventurar e aprender a usar o prompt de comando, a referência abaixo é muito útil. No entanto, se esse nome pareceu estranho, melhor pular essa parte. Usuários de Windows geralmente não utilizam o prompt no seu fluxo de trabalho usual, mas caso você trabalhe em um sistema operacional Linux ou MACOSX e já esteja habituado, essa referência pode ser útil.

- McDermott (2021) Lecture 3 - Learning to love the shell

Pacotes no R

[EM CONSTRUÇÃO]

Documentando suas funções

Escrevendo boas vignettes

Programação Defensiva (ou como se prevenir do seu próprio erro)

Criar/Contribuir para um Repositório

Wrap Up dos anteriors + Pull Requests [EM CONSTRUÇÃO]

Bonus!

[EM CONSTRUÇÃO]

Análise de Dados Georeferenciados

Inferência Estatística

Visualização de Dados

Esse item terá um guia específico com incorporação de discussão de Shiny, WordPress/GitHub Pages e GGPlot. Aqui será uma breve discussão de um Ggplot básico para construção de boas vignettes.

Estilo de Programação (Paradigmas)

Programação Funcional

Eu quero mais! Próximos Passos =)

Bibliografia

Hadley Wickham, Garrett Grolemund. 2017. *R for Data Science*. O'Reilly UK Ltd. https://r4ds.had.co.nz/.

McDermott, Grant. 2021. "Data Science for Economists (Lecture Notes)." Lecture Notes, UOregon. https://github.com/uo-ec607/lectures.

Oliveira, João, and Pedro Cavalcante. 2021. "R in Rio Lecture Notes."

Wickham, Hadley. 2015. *R Packages*. O'Reilly UK Ltd. https://r-pkgs.org/.

———. 2019. *Advanced r, Second Edition*. Taylor & Francis Inc. https://adv-r.hadley.nz/.