R para iniciantes
Aula VI Boas Práticas

Carlos Henrique Tonhatti

Universidade Estadual de Campinas

Dúvidas da última Aula?

Sumário

Sumário

Boas práticas para lidar com projetos de dados

- Comece cada programa com a descrição sobre o que ele faz
- Carregue todos os pacotes necessários logo no inicio
- Considere que você está no diretório de trabalho.
- Comente as sessões de seu código
- Coloque as definições de funções no inicio do arquivo, ou em um arquivo separado
- Use nomes e estilo de forma consistente
- Quebre o código em pedaços pequenos
- Não altere os dados brutos
- Sempre inicie um ambiente limpo ao invés de salvar o workspace (!)
- Mantenha o registro das sessões
- Tenha alguém para rever o código
- Use controle de versão

Princípios ao lidar com análise de dados

- Transparência: Organização lógica das unidades.
- Manutenção: Padronização e comentários objetivos.
- Modularidade: Separe em unidades lógicas.
- Portabilidade: use caminhos relativos, minimize dependências, deixe a dependências claras.
- Reprodutibilidade: Facilidade em repetir os resultados
- Eficiência para manter e modificar

Anatomia da pasta de trabalho

Use nomes de diretórios/ pastas de forma padrão para guardar os arquivos

Dados brutos Pasta com os dados mais simples, menos manipulado

Dados processados Dados limpos, transformados pronto para analise

Gráficos Guardar os gráficos

doc Documentação sobre o projeto, sobre os dados, artigos e rascunhos etc

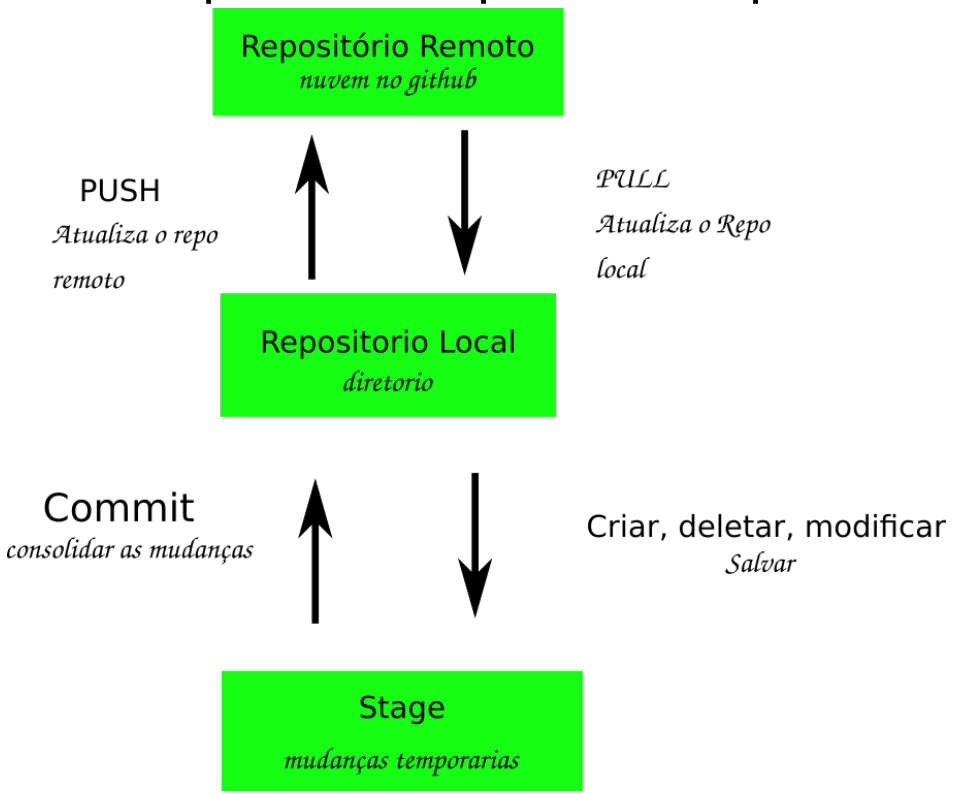
Relatórios Relatórios gerados sobre os dados

Além disso, vale a pena ter um arquivo Leiame.txt como um pequeno resumo do projeto. E um arquivo ParaFazer.txt.

Sumário

Controle de Versão

Muito mais que um simples backup em nuvem



Mais sobre o GitHub

- Você pode escolher e colocar o repo como privado. Há licença educacional para isso.
- Instalar no windows `https://dicasdeprogramacao.com.br/como-instalar-o-git-no-windows/`
- Um pequeno tutorial em PT-BR `https://rogerdudler.github.io/git-guide/index.pt_BR.html`
- O livro `https://git-scm.com/book/pt-br/v2`

Sumário

Rmarkdown

Uma linguagem de marcação de texto que gera relatórios dinâmicos. Facilita o trabalho de comunicação de resultados e aumenta a transparência da análise.

Sumário

Assuntos que não foram cobertos pelo curso

Assuntos que não foram cobertos pelo curso e podem ser necessários dependendo do projeto.

- Estatística (geral);
- Matemática (geral);
- Iteração e recursão (otimização, solução de problemas);
- Manipulação de palavras “strings” e expressão regular (trabalhando com texto);
- Interação com outras linguagens (geral);
- Banco de dados (trabalhando com muitos dados);
- Computação paralela (otimização).

Pacotes

- dplyr, tidyr, readr, ggplot2 tidyverse
- stringr, stringi, “regex”
- popgenreport, ape, ade4, pegas hierfstat
- vegan, MASS, caret
- doParallel
- roxigen2

Sumário

Trabalho final

O trabalho final da disciplina tem o objetivo de desenvolver programas que realizem análises mais completas que as vistas durante o curso. Cada trabalho dá ao aluno a chance de aprender algo novo.
A seguir algumas propostas de Projeto, se tiver alguma outra ideia pode falar

Proposta 1

Reconstrução filogenética

Reconstrução filogenética usando sequências do GenBank

Fonte: Analysis of Phylogenetics and Evolution with R. Emmanuel Paradis. páginas:46–50, 121–125 “Caso Sylvia”

Proposta 2

Modelagem de crescimento

Crescimento independente de densidade

Fonte : A primer of ecology with R.

M. Henry H. Stevens. Cap. 1. Problema 1.1

Proposta 3

Seleção de modelos

Seleção de modelos por verossimilhança

Fonte: `http://cmq.esalq.usp.br/BIE5781/doku.php?id=07-selecao:07-selecao`

Proposta 4

Análise Exploratória de dados

Requer adaptação do roteiro

- Compreender como são os dados. Estatísticas de sumário. Como estão distribuídos, há padrões?
- dados `https://archive.ics.uci.edu/ml/datasets/student+performance`
- Se quiser tentar uma PCA `https://www.r-bloggers.com/computing-and-visualizing-pca-in-r/`

Proposta 5

Dados bibliográficos

Bibliometria

- Usar o pacote bibliometrix
  `https://cran.r-project.org/web/packages/bibliometrix/vignettes/bibliometrix-vignette.html`
- Usar uma palavra-chave da tua areá
- Criar gráficos: quem produz mais, mais citações, co-citação
- Gerar rede de co-citação.

Entrega

- 21 dias para entregar por email
- Em pdf (com Rmarkdown) com todos os arquivos necessários
- Repositório do Github ou pasta zipada
- Roteiro
  - Introdução (1 paragrafo)
  - Requerimentos (pacotes)
  - Desenvolvimento
  - Respostas encontradas (1 paragrafo)
  - Dificuldades encontradas
  - Bibliografia (sem formato definido)

  Pode acrescentar itens se quiser.

O que se espera

1. Dentro do prazo.
2. Uso de Rmarkdown/PDF.
3. Dentro do roteiro
4. No PDF, todas as linhas de código de R aparecerem sem cortes
5. Estilo de escrita do código, identação, clareza.
6. Comentários no código, se criar alguma função comentar o que faz, qual entrada e qual a saída
7. Github ou zipado com todos os arquivos necessários.
8. Reprodutibilidade, o arquivo .rmd gera o PDF apresentado?
9. Texto do documento.

Nota: Cada item vale um ponto. Todas as atividades do Swirl somam um ponto a média.