Introdução

author: Cristofer Weber date: 2018-04-09

About Me

```
library(lubridate)
library(dplyr)
neogrid_period <- ymd("2009-06-08") %--% ymd("2017-10-06")
sicredi_period <- ymd("2017-10-07") %--% (today() + duration(10, "years"))

works_at <- case_when( today() %within% neogrid_period ~ "na Neogrid",
                       today() %within% sicredi_period ~ "no Sicredi Digital" )

works_as <- paste("Cientista de Dados", works_at)
```

Cientista de Dados no Sicredi Digital

Mais sobre mim: https://www.linkedin.com/in/cristoferweber/

Contato: cristofer.weber@outlook.com

Sobre as aulas

- Programa:
  - Entregue em separado
- Exposição do conteúdo:
  - Misto entre apresentações, notebooks, textos para leitura e sugestão de conteúdos complementares (textos, vídeos)
- Avaliação:
  - Resposta aos exercícios nos notebooks (exceto da primeira aula)
  - Discussões sobre os textos (textos desde a primeira aula)
  - Projeto(s) de dados (uso de datasets públicos: CEASA/RS, TED talks, Airbnb, Youtube, etc)

Conteúdo

Análise de dados
Estatística Descritiva
Análise exploratória com gráficos
Álgebra Linear para análise de dados
Distribuições de probabilidade
Estatística Inferencial
Análise de regressão
Frequent Item Set Mining

R
Conceitos, introdução ao R (tipos)

R
Data Munging e *Tidyverse*
Data Science Workflow
Fontes de dados (arquivos, web)
Programação funcional, imutabilidade
Organização de código e dados
Reproducibilidade
Execução de *scripts*

Para combinarmos:

- Direcionamento da disciplina pelo conteúdo de R
  - com aplicações de Análise de Dados
- Direcionamento da disciplina pelo conteúdo de Análise de Dados
  - introdução gradual dos conteúdos de R
- Tópicos ou complementação de conteúdo
  - Se tudo correr bem com o conteúdo listado podemos adicionar um tópico de interesse

Disciplina: Análise de Dados com R

Análise de Dados

- Computação, Inferência, Previsão
- Organização dos dados

Com R

- Uso da linguagem R na Análise de Dados

Do site da UniRitter

Objetivo da Especialização em Big Data & Data Science

- Preparar profissionais para o mercado
  - Processo de análise de dados
  - Princípios de engenharia de software

Diferenciais

- Aulas ministradas por profissionais que atuam na indústria/mercado na área de Data Science e Big Data
  - Aproveitem para discussões sobre práticas!

Análise de Dados

Análise como ação: Examinar (minuciosamente), Criticar, Estudar.

"O processo de transformação de dados brutos em informações utilizáveis"

Do ponto de vista empresarial/comercial, Análise de Dados consiste no emprego de diferentes ferramentas computacionais para examinar dados. Conceitos como Business Intelligence e Data Mining são bastante vistos como sinônimos de Análise de Dados.

Pode-se ter a impressão de que a análise de dados *"favorece"* o emprego da estatística. Todo o ferramental estatístico é de grande valor para a Análise de Dados, tal como o ferramental computacional e matemático (geral), e o conhecimento do domínio de aplicação.

A linguagem R

Mais que uma linguagem, um ambiente para computação estatística e gráfica. Seus pontos fortes são as funcionalidades estatísticas, processamento gráfico, estruturas de dados, manipulação de objetos e computação numérica.

O que distingue uma linguagem de um ambiente? No caso do R, ele nasceu como um software interativo que integra (mostrar no RStudio):

- a linguagem de programação voltada à manipulação de dados e computação,
- o interpretador de comandos que executa as operações desta linguagem,
- o gerenciador de ambientes de escopo de dados que organiza os dados em memória,
- e as facilidades para exibição de gráficos em tela.

A linguagem e o ambiente R atendem diferentes propósitos:

- Pode ser utilizado de forma iterativa, similar ao emprego de ferramentas como SAS e Matlab
- É empregado também na criação de *scripts* para automação de etapas ou da totalidade de *pipelines* analíticos
- Possui bibliotecas que permitem a criação de aplicações e serviços Web
- Bastante utilizado na criação de relatórios e apresentações.

(vamos explorar um pouco disto hoje)

O ecossistema R

Ao longo dos anos formou-se um ecossistema neste ambiente, que atualmente consiste em 11.500 bibliotecas (Outubro de 2017) de contribuição espontânea da comunidade R.

Esta comunidade é formada por acadêmicos, grupos de pesquisa, empresas e profissionais que colaboram para que a linguagem e o ambiente possam ser aplicados em mais domínios e cenários.

Origem

R tem sua origem na linguagem S, criada na década de 70 por John Chambers no Bell Labs. Desde sua criação, a linguagem S passou por várias atualizações a ponto de ser nomeada "New S" a partir de 1988. Estas modificações, em conjunto com demais mudanças introduzidas até 1991, contemplam muitos dos recursos hoje vistos como fundamentais para a linguagem R.

- Vídeo: Forty years of S

R é tida como uma implementação moderna da linguagem S, assim como a linguagem S-PLUS, que é a 'herdeira' da linhagem do S.

Tanto S quanto R foram inicialmente desenvolvidos em linguagem C para o ambiente UNIX, com as operações computacionais mais pesadas sendo escritas em Fortran. R possui uma forte integração com a linguagem C++, muito usada no desenvolvimento de bibliotecas.

R é bastante empregado como um *front-end* para bibliotecas e serviços codificados em outras linguagens de melhor desempenho.

Peculiaridades

As características mais peculiares da linguagem são seu paradigma híbrido, que combina recursos de programação funcional com programação orientada à objetos, e seu modelo de dados ser vetorizado.

Apesar de não óbvio, R possui muitas semelhanças com a linguagem Scheme (um dialeto de Lisp). Seu modelo de orientação a objetos e de ambientes de escopo de dados são bastante similares.

R também possui similaridades com o Matlab, principalmente na vetorização de operações e na indexação de matrizes. O ambiente do Matlab também compartilha algumas similaridades com o R, tais como a exibição de resultados em linha de comando e a exibição de gráficos.

Leituras para discussão na próxima aula

- Por que R?
  - How R helps Airbnb make the most of its data
- Perfil e desafios de cientistas de dados (estudos na Microsoft)
  - The Emerging Role of Data Scientists on Software Development Teams
  - Data Scientists in Software Teams: State of the Art and Challenges
- Visão sistêmica em projetos de dados
  - You say data, I say system
- Trata das diferentes formas que o código afeta projetos de pesquisa
  - The roles of Code in Computational science

Material será disponibilizado no Blackboard!