

Introdução à Ciência de Dados

Aula Inaugural

Daniel Brito dos Santos

Antes de tudo

Vamos instalar o sistema R

- ☐ Baixe e instale o **Rstudio**
- ☐ Ele vai te orientar como instalar o **R**
- ☐ Já explico tudo

Sobre a matéria

Objetivos da matéria

- ▶ Apresentar a **ciência de dados**, seus principais **conceitos** e **ferramentas**.
- ▶ Oferecer uma **trilha pedagógica** bem definida e “opinionada”.
- ▶ Guiar os aventureiros nessa trilha, com muito exercício e direcionamento.
- ▶ Construir conhecimento útil pra cada participante.

Métodos de avaliação

▶ **Listas de exercícios**

- ▶ Toda aula vai ter dever de casa
- ▶ Responder no classroom

▶ **Aulas de revisão**

- ▶ Cada exercício será explicado por um aluno sorteado

▶ **Um projeto de dados**

- ▶ Trabalho final
- ▶ Vai ser tranquilo, confia

- ▶ Realmente tentar responder   mais importante que acertar
 - ▶ “N o consegui responder por n o ter entendido x. O que eu entendi dessa pergunta foi tal”   uma resposta v lida
- ▶ Entregar no prazo

- ▶ 6 semanas. Dessa segunda 09/01 até 17/02
- ▶ Vamos nos encontrar três vezes por semana, sendo a última aula da semana de revisão.
- ▶ Seg, qua, sex. 14h - 16h.

Conteúdo

Por que usar R?

- ▶ R não é apenas uma linguagem mas todo um ecossistema completamente dedicado desde o início para **Ciência de Dados**
- ▶ O sistema **S**, predecessor do R é tido como um dos principais responsáveis pela revolução na forma que o ser humano lida com dados:
the S system, which has forever altered the way people analyze, visualize, and manipulate data [...] VER COMO REFERENCIAR

- ▶ R ser muito flexível permitiu o surgimento de **mini-linguagens** pra lidar com partes específicas de processos na DS.
 - ▶ Essas mini-linguagens ajudam a **pensar nos problemas** como um cientista de dados, tornando mais fluída a interação do cérebro com o computador.
 - ▶ O maior exemplo dessas mini-linguagens é o dialeto **tidyverse**.

- ▶ coleção de pacotes R projetados para manipulação e visualização de dados.
- ▶ Ele é baseado no framework de dados “tidy”, que estrutura os dados de maneira fácil de trabalhar e analisar.
- ▶ Os pacotes do tidyverse são projetados para trabalharem juntos de forma flúida e integrada.
- ▶ Muito usado por cientistas de dados devido a sua simplicidade e eficácia

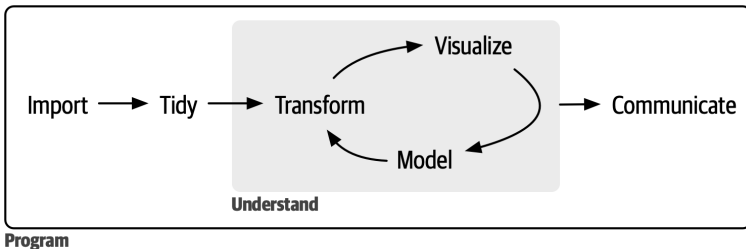
- ▶ Criador do **tidyverse**
- ▶ Estatístico, cientista da computação
- ▶ Vida dedicada a construir e ensinar métodos e ferramentas melhores, mais eficientes e mais humanos.
- ▶ Seu trabalho é altamente influente na maneira como fazemos DS atualmente.

- ▶ Vamos seguir religiosamente o livro **R for Data Science**.
 - ▶ Porque ele é pedagógico, atual, claro, lindo e cheiroso.
 - ▶ Escrito por **Hadley Wickham**
 - ▶ O livro é dividido em partes -> capítulos -> seções
 - ▶ Vamos nos focar na primeira parte: **whole game**
 - ▶ Resolver os exercícios propostos
 - ▶ E finalizar com um pequeno projetinho individual

Nosso modelo de ciência de dados

Ciência de dados

- ▶ Disciplina que permite transformar dados brutos em compreensão, insight e conhecimento
- ▶ Vamos utilizar o seguinte processo para isso:



- ▶ O primeiro passo é importar os dados para o nosso programa.
- ▶ Afinal, sem dados não tem ciência de dados!

- ▶ Depois de importar é uma boa ideia deixar “tidy”
 - ▶ Tidy significar organizar os dados de modo que sua estrutura combine com sua semântica.
 - ▶ Em resumo: **cada coluna é uma variável** e **cada linha uma observação**.
 - ▶ Esse princípio permite que a gente tenha uma **estrutura consistente** e possa se focar em **responder perguntas** sobre os dados ao invés de quebrar a cabeça tentando encaixar eles em funções diferentes.

- ▶ Depois de arrumar os dados vamos **transformá-los**
 - ▶ Selecionar uma cidade, selecionar determinado ano
 - ▶ Criar novas variáveis a partir das existentes
(*velocidade* \leftarrow *distncia/tempo*)
 - ▶ Calcular resumos estatísticos (médias e contagens)

- ▶ Com os dados arrumados e transformados temos dois motores para gerar conhecimento:
 - ▶ **Visualização e Modelagem.**
 - ▶ Como cada um tem vantagens e desvantagens complementares análises na vida real vão iterar muitas vezes entre elas.

- ▶ Atividade fundamentalmente humana.
- ▶ Uma boa visualização:
 - ▶ vai mostrar coisas **surpreendentes** ou
 - ▶ **levantar perguntas** sobre os dados.
 - ▶ Também pode indicar se você está fazendo a **pergunta errada**
 - ▶ ou que você precisa coletar mais dados.
- ▶ Pode te **surpreender** mas não escala bem.

- ▶ Quando tiver suas perguntas precisas o suficiente, você pode usar um modelo para responde-las.
- ▶ Fundamentalmente matemáticos ou computacionais, portanto escalam muito bem.
 - ▶ Mesmo quando não escalam bem é mais barato comprar mais computadores do que mais cérebro.
- ▶ Mas todo modelo faz suposições e, por sua própria natureza, um modelo não pode questionar suas próprias suposições. Isso significa que um modelo não pode realmente te surpreender.

- ▶ Parte absolutamente crítica de qualquer projeto de análise de dados.
- ▶ Não importa o quão bom são seus modelos e visualizações se você não pode comunicá-los.

Sobre o livro

- ▶ As ferramentas descritas no livro são utilizadas em **todos os projetos de dados**, mas não são suficientes para a maioria deles.
- ▶ Com elas conseguimos em média resolver **80%** de qualquer de qualquer projeto de dados, outras são necessárias para os 20% restantes.
- ▶ A ideia é construir uma **base sólida** que te permita ir atrás deles conforme precisar.

- ▶ Faria sentido ensinar os processos de dados na ordem em que eles acontecem
- ▶ Mas os autores perceberam que importar e arrumar os dados na maioria das vezes é **chato e protocolar**,
 - ▶ nas outras é **frustante e esquisito**.
- ▶ Péssimo lugar pra começar

- ▶ Vamos começar com **visualização** e **transformação** de dados já arrumadinhos
- ▶ Assim, quando passarmos pela **importação** e **arrumação** você já vai saber que vale a pena!

▶ **Visualização - Capítulo 2**

- ▶ Vamos fazer gráficos elegantes e informativos. Aprender a estrutura básica do ggplot2, e técnicas para transformar dados em plots.

▶ **Transformação - Capítulo 4**

- ▶ Vamos aprender verbos chave para selecionar variáveis importantes, filtrar observações fundamentais, criar novas variáveis e computar resumos.

▶ **Tidying - Capítulo 6**

- ▶ Vamos aprender sobre tidy data, uma forma consistente de armazenar os dados para tornar a transformação, visualização e modelagem mais fácil. Veremos os princípios e como deixar os dados “arrumadinhos”.

▶ **Importação - Capítulo 8**

- ▶ Vamos apresentar o básico da leitura de um csv pra dentro do R. Não é tão simples quanto parece.

- ▶ Além desses capítulos temos outros cinco (3,5,7, e 9) com enfoque em apresentar boas práticas do Workflow em R. Essas práticas de escrita e organização de código vai nos direcionar na trilha do sucesso a longo prazo, e nos manter organizados se e quando enfrentarmos projetos reais.