

# Introdução à Ciência de Dados e mundo *tidyverse*

Jadson Pessoa

Professor do DECON | Membro GAPE



# Plano de Trabalho

Introdução

Coleta e Tratamento

Visualização

Modelagem

Comunicação

O mundo *tidyverse*

# Introdução

- Ciência de Dados (*Data Science*);
- Conjunto de técnicas de coleta, tratamento, análise e apresentação de dados;
- Sistematização de dados em busca informações úteis.
- **melhorar apresentação inicial**

# Introdução

- Sistematização de projeto em DS:
  1. Coleta;
  2. Tratamento;
  3. Análise;
  4. Visualização e apresentação.

# Introdução

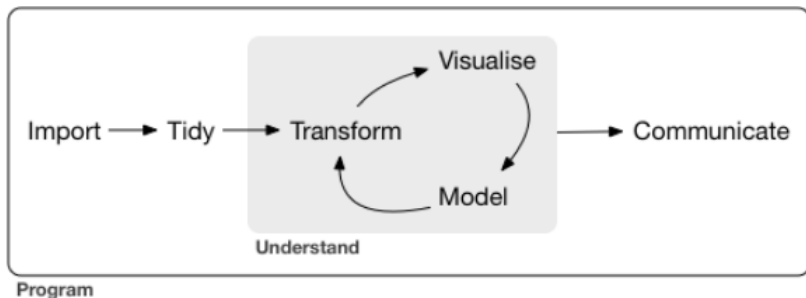
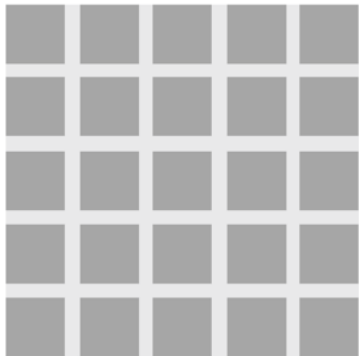


Figure 1: O processo de compreensão dos dados (Grolemund and Wickham [2017])

# Coleta e Tratamento

## Structured data



## Unstructured data



Figure 2: Dados Estruturados e Não estruturados

# Coleta e Tratamento

A **coleta** dos dados podem em diferentes formatos:

- Excel;
- XML;
- JSON;
- txt;
- HTML;
- MySQL;
- Formatos proprietários (Stata, Minitab, SPSS, SAS, etc).

# Coleta e Tratamento

O dados precisam ser **tratamentos**:

- Limpeza de dados;
- Tratamento de *missing values* ou NA;
- Construção de números índices;
- Deflacionar valores correntes;
- Obtenção de taxas de crescimento;
- Tratando tendências;
- Dessazonalização;
- Subconjuntos (*subsetting*);
- Classificação;
- Utilização de *lags*.



# Visualização

Uma vez que seus dados estejam arrumados, podemos passar para a parte de **exploração dos dados**. A exploração de dados é a arte de analisar seus dados, gerando hipóteses rapidamente, testando-os rapidamente, repetindo-os várias vezes. O objetivo da exploração de dados é gerar muitos leads promissores que você poderá explorar mais tarde com mais profundidade. Em geral, faz-se exploração de dados por meio da *visualização* dos mesmos.

Um bom processo de visualização de dados permite que possamos nos concentrar naquilo que realmente importa, deixando de lado relações não tão importantes.

# Modelagem

Uma vez que tenhamos conseguido propor uma *hipótese de trabalho* através da etapa de exploração/visualização de dados, o próximo passo é propor um **modelo** entre as variáveis do nosso conjunto de dados. O objetivo da modelagem é capturar a essência de um conjunto de dados.

# Comunicação

A última etapa do processo de *data science* é comunicar os resultados para clientes, gestores ou demais interessados. É uma fase absolutamente crítica do projeto. Isto porque, ao menos que você consiga se comunicar com a sua audiência, de nada valeu todo o trabalho realizado nas etapas anteriores.

# O mundo *tidyverse*

De modo a fazer cada uma dessas etapas dentro do R, nós vamos utilizar a família de pacotes *tidyverse*. Assim, antes de qualquer coisa, certifique-se que você tenha o tenha instalado na atual versão do R.

```
install.packages('tidyverse')  
require(tidyverse)
```

# A família *tidyverse*

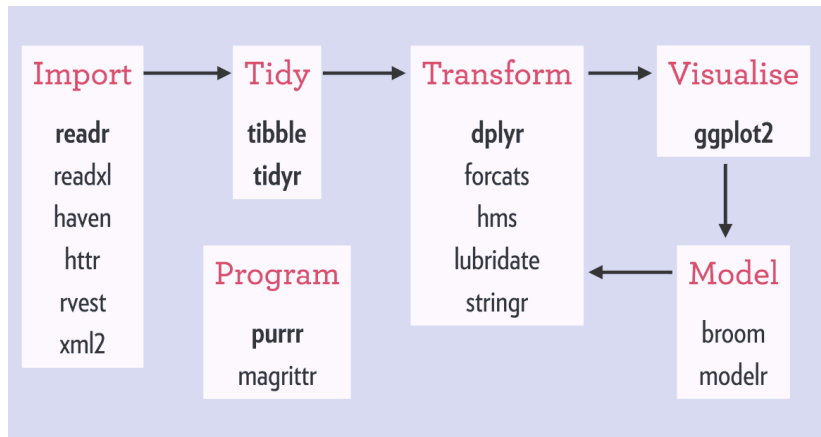


Figure 3: Os pacotes tidyverse (Grolemund and Wickham [2017])

# Referências I

G. Grolemund and H. Wickham. *R for Data Science*. O'Reilly Media, 2017.