Introdução à Ciência de Dados como Prática na Pesquisa Acadêmica

Seção 02 - Ciência de Dados e a Pesquisa Acadêmica

Jadson Pessoa

Professor do DECON | Membro GAPE

Plano de Trabalho

Objetivos

Método Científico como Teória e Prática

DS e a Pesquisa Teórica-Empírica

O mundo *tidyverse*

Objetivos

- Discutir o método científico e sua relação com a sistematização de um projeto de Data Science (DS), assim como, apresentar os modelos canônicos da pesquisa acadêmica, tendo como foco pesquisas que utilizam aboragens empíricas.

Método Científico como Teória e Prática

- Ciência como prática (técnica);
- Ciência como método;
- Ciência como epistemologia.

  "Ciência é sempre um enlace de uma malha teórica com dados empíricos, é sempre uma articulação do lógico com o real, do teórico com o empírico, do ideal com o real" Severino [2017], *p. 100.*

Método Científico

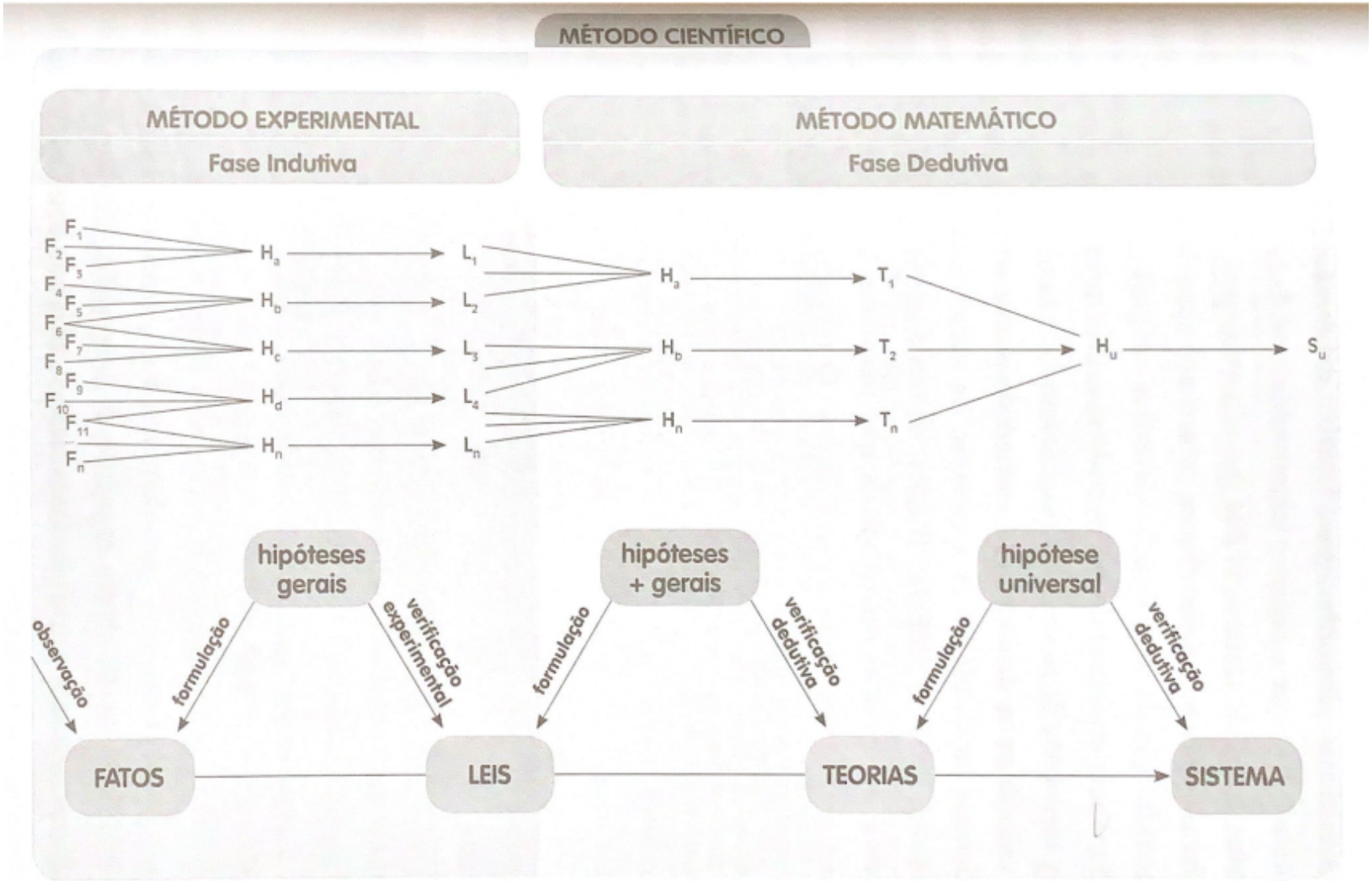


Figure 1: Estrutura Lógica do método científico Severino [2017], p.101.

Método Científico Revisitado

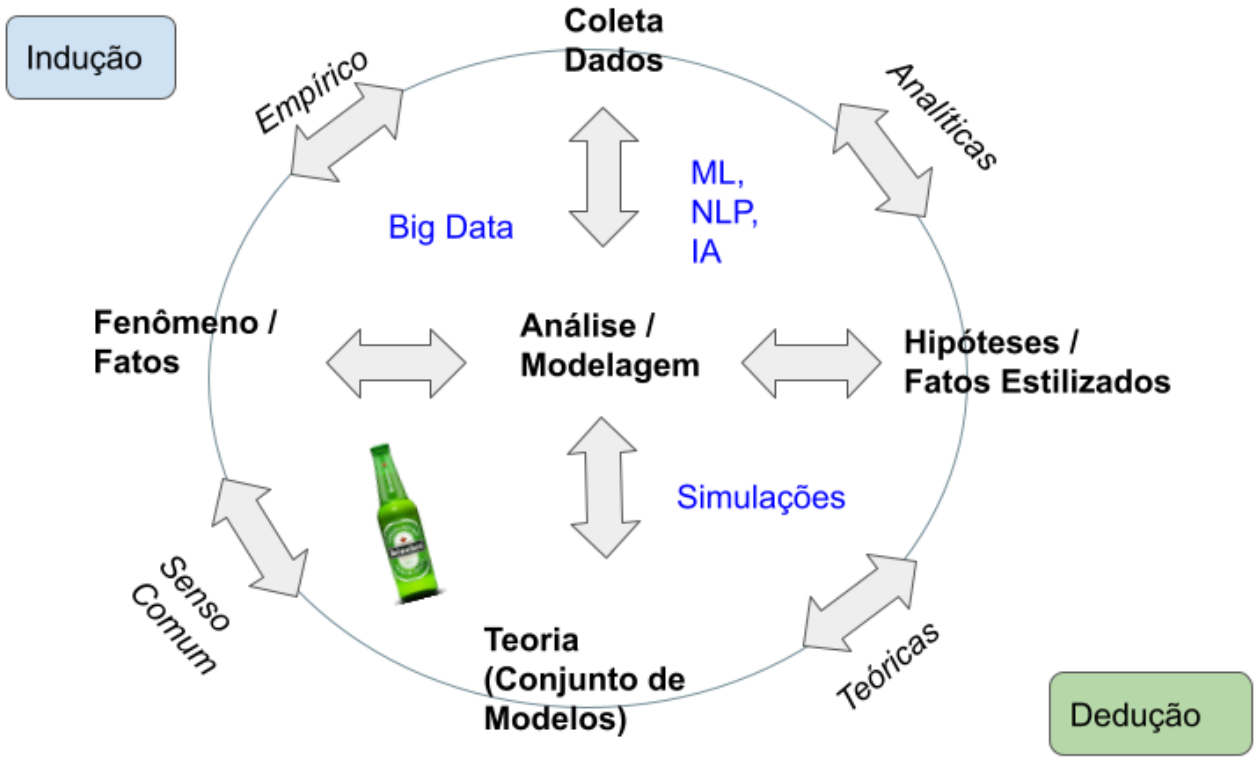


Figure 2: Computational Social Science Methods - Hilbert [2021]

Metodologias de Pesquisa Científica

Ciência:
Aplica Técnicas $\rightarrow$ Métodos $\rightarrow$ Fundamentos Epistemológicos

- Modalidades de Pesquisa Acadêmicas:
  - Pesquisa Teórica
  - Pesquisa Teórica-Empírica

Elementos de um Artigo Empírico

INTRODUÇÃO	Discussão Geral	Justificativa	Problema de Pesquisa
MÉTODO	Base de Dados	Desenho da Pesquisa	Execução
RESULTADOS E DISCUSSÃO	Resultados Obtidos	Discussão com a teoria ou com outros trabalhos	Resposta a questão
CONCLUSÃO	Achados (*Findings*)	Possíveis Contribuições	Novas questões

Figure 3: Elementos - Artigo Empírico

Exemplos: Artigos 1; Revista Economia Aplicada - USP; Journal of Applied Economics

DS e a Pesquisa Teórica-Empírica

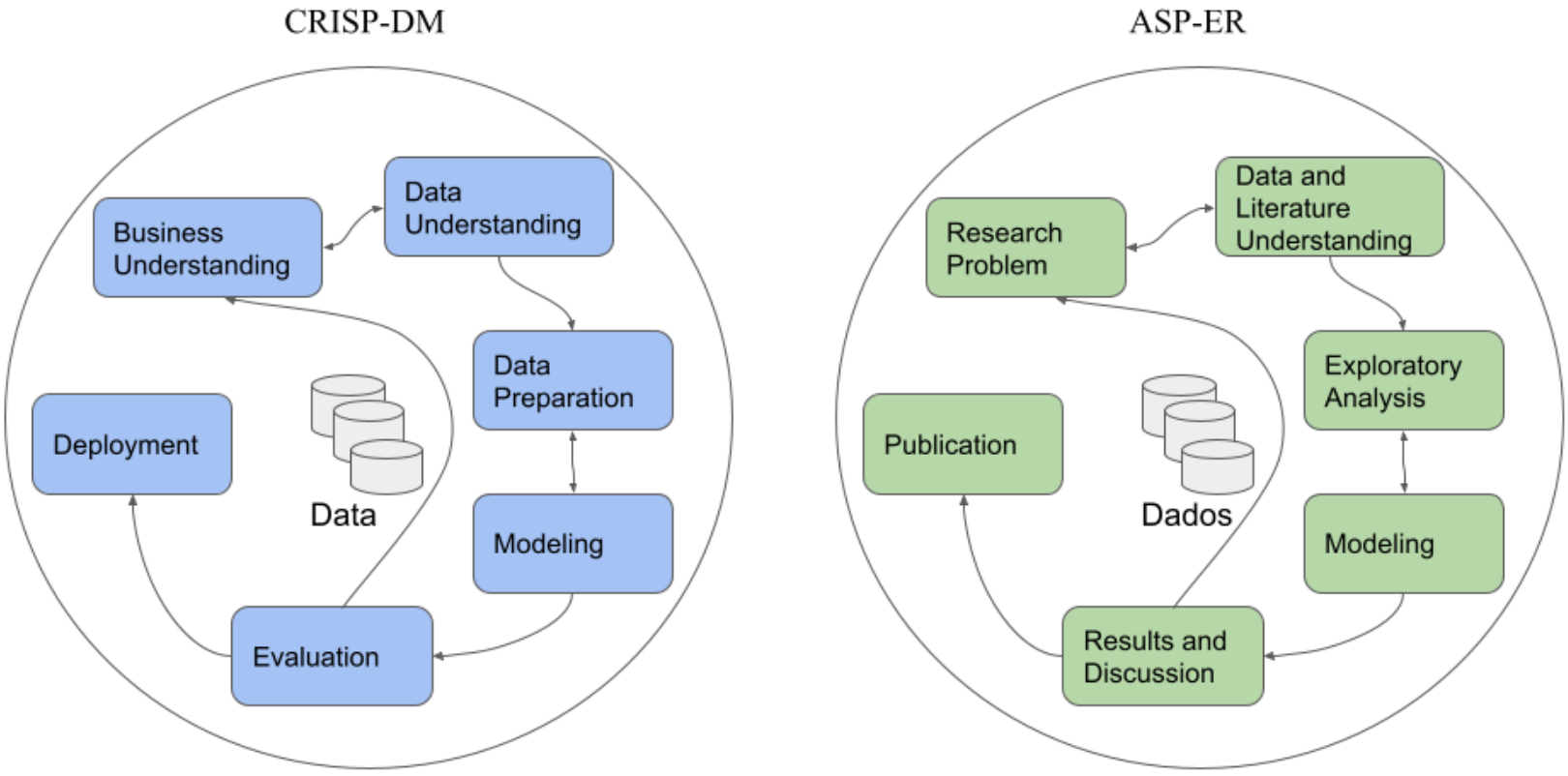


Figure 4: CRISP-DM IBM [2021] e ASP-ER

Coleta e Tratamento

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

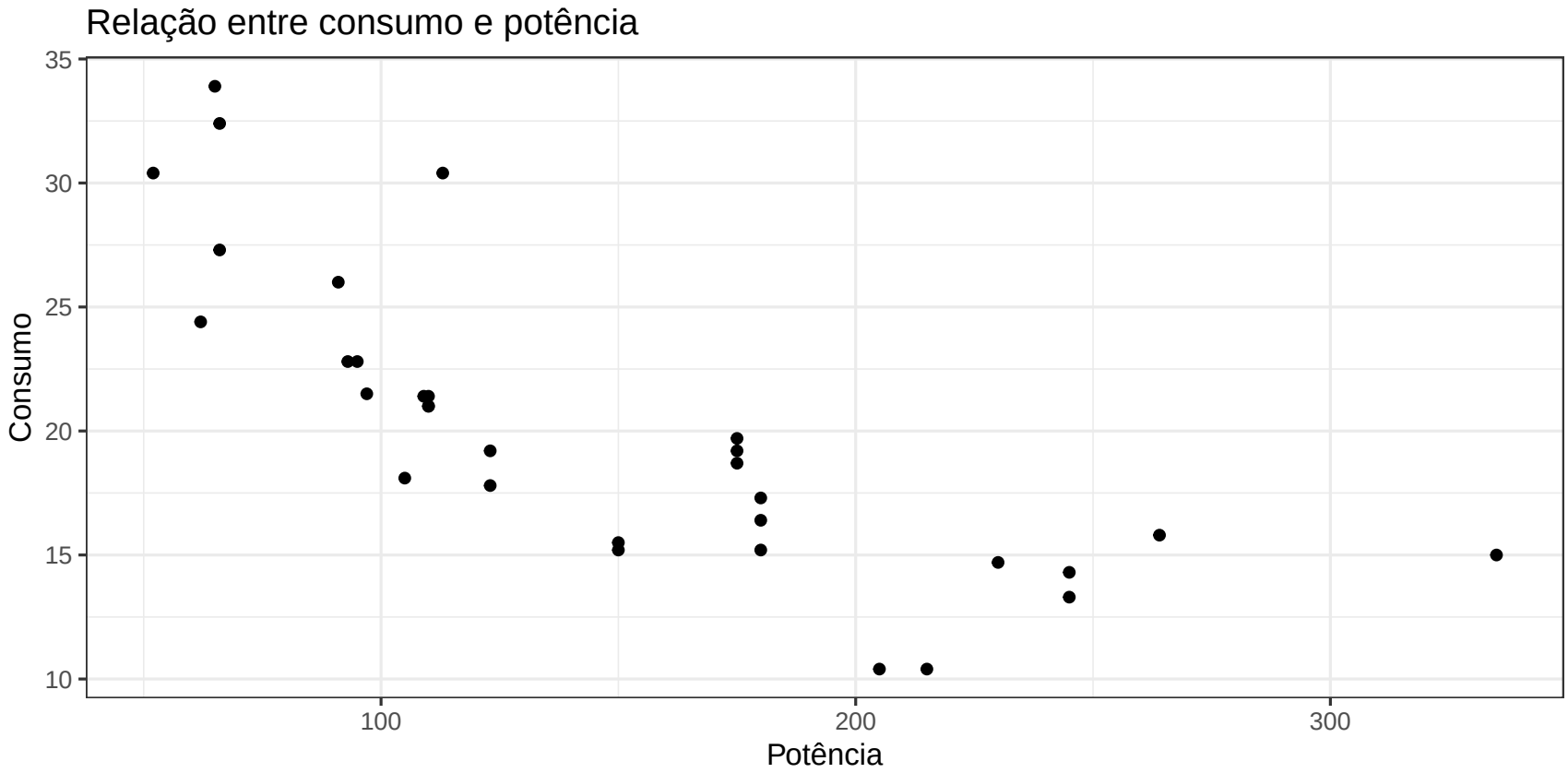
- Excel;
- XML;
- JSON;
- txt;
- HTML;
- MySQL;
- Formatos proprietários (Stata, Minitab, SPSS, SAS, etc).

Coleta e Tratamento

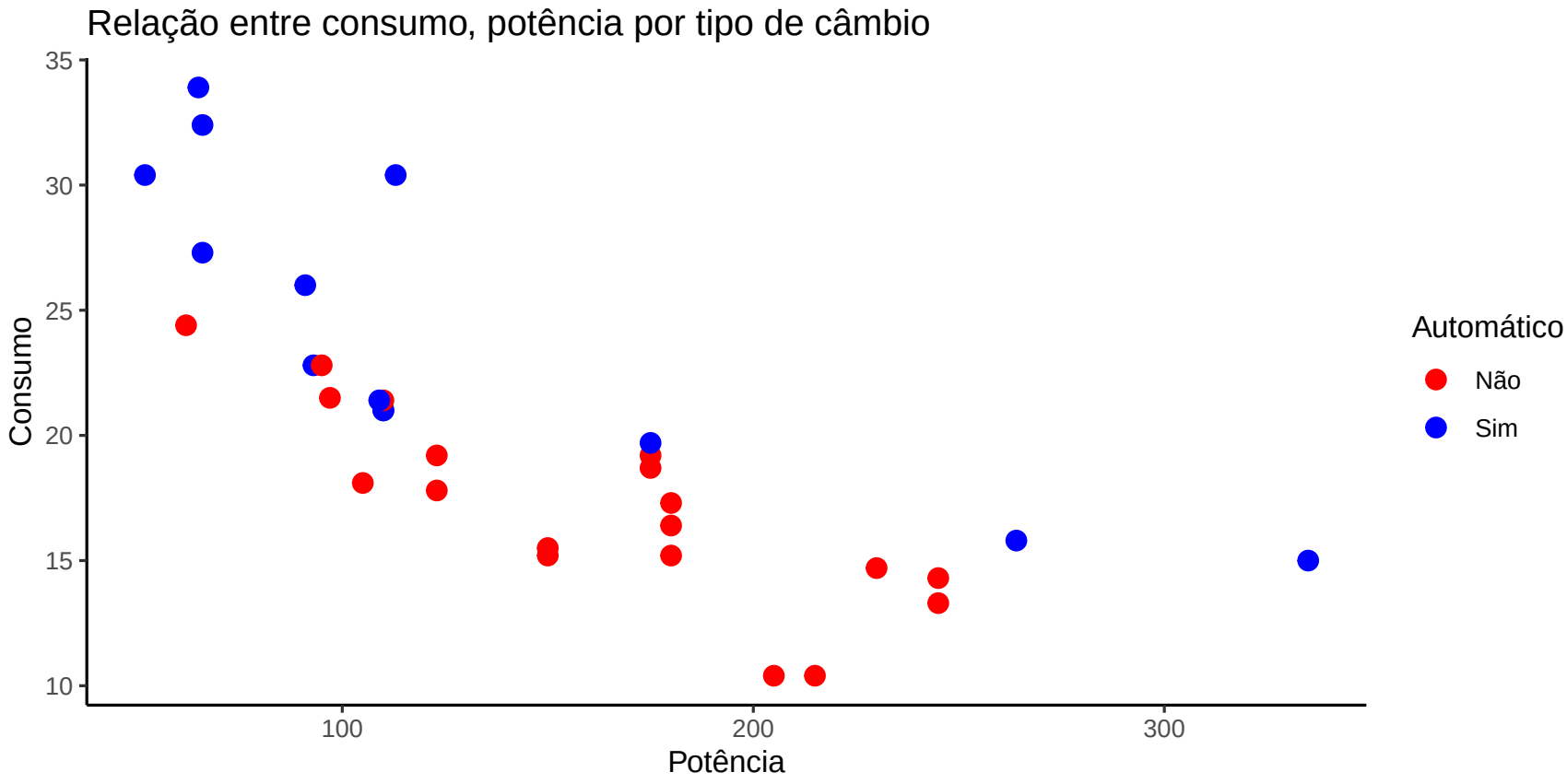
O dados precisam ser **tratamentodos**:

- 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



Visualização



Modelagem

MQO:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, i = 1, \ldots N.$$

Table 1: Regressão por MQO

	Dependent variable:		
	mpg		
	(1)	(2)	(3)
hp	-0.068^{***}	-0.059^{***}	-0.059^{***}
	(0.010)	(0.010)	(0.013)
Constant	30.099^{***}	26.625^{***}	31.843^{***}
	(1.634)	(1.616)	(1.993)
Observations	32	19	13
R^2	0.602	0.691	0.641
Adjusted R^2	0.589	0.673	0.608
Residual Std. Error	3.863 (df = 30)	2.192 (df = 17)	3.859 (df = 11)
F Statistic	45.460^{***} (df = 1; 30)	38.088^{***} (df = 1; 17)	19.647^{***} (df = 1; 11)
Note:			*p<0.1; **p<0.05; ***p<0.01

É possível escrever dentro do texto utilizando marcação do R, como o seguinte exemplo: 30.099, o intercepto para 26.6248479 e do 31.8425012.

Rmarkdown

- Literate Programs - LitPro;
- Exemplo

O mundo *tidyverse*

- Todas as etapas podem ser desenvolvidas em um único ambiente (RStudio);
- Bibliotecas, bibliotecas e seus próprios pacotes.

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

O mundo *tidyverse*

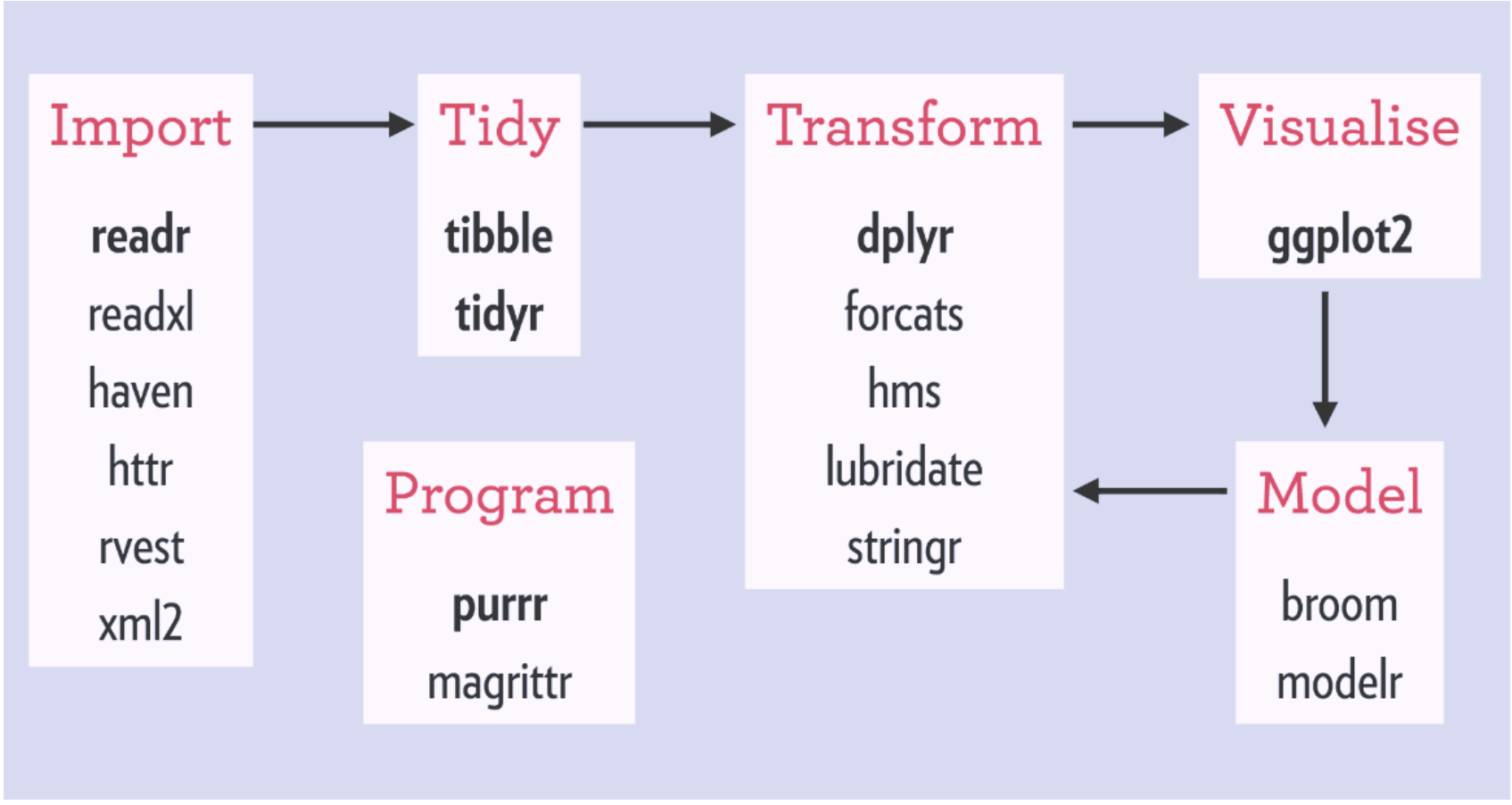


Figure 5: Os pacotes `tidyverse` (Wickham and Grolemund [2017])

Referências I

Martin Hilbert. Computational Social Science Methods, 2021. URL https://www.coursera.org/learn/computational-social-science-methods/home/welcome.

IBM. IBM SPSS Modeler CRISP-DM Guide, May 2021. URL https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/spss-modeler/SaaS?topic=SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html.

Antônio Joaquim Severino. *Metodologia do trabalho científico*. Cortez editora, 2017.

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