



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Data Science com R - II

Primeira Parte

Pedro Sousa e João Lopes

Setembro 2024





INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Programa





INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Programa



- **Transformação dos dados**
- **Exploração dos dados**
- **Modelação**
- Relatórios e apresentações
- Comunicação





Programa



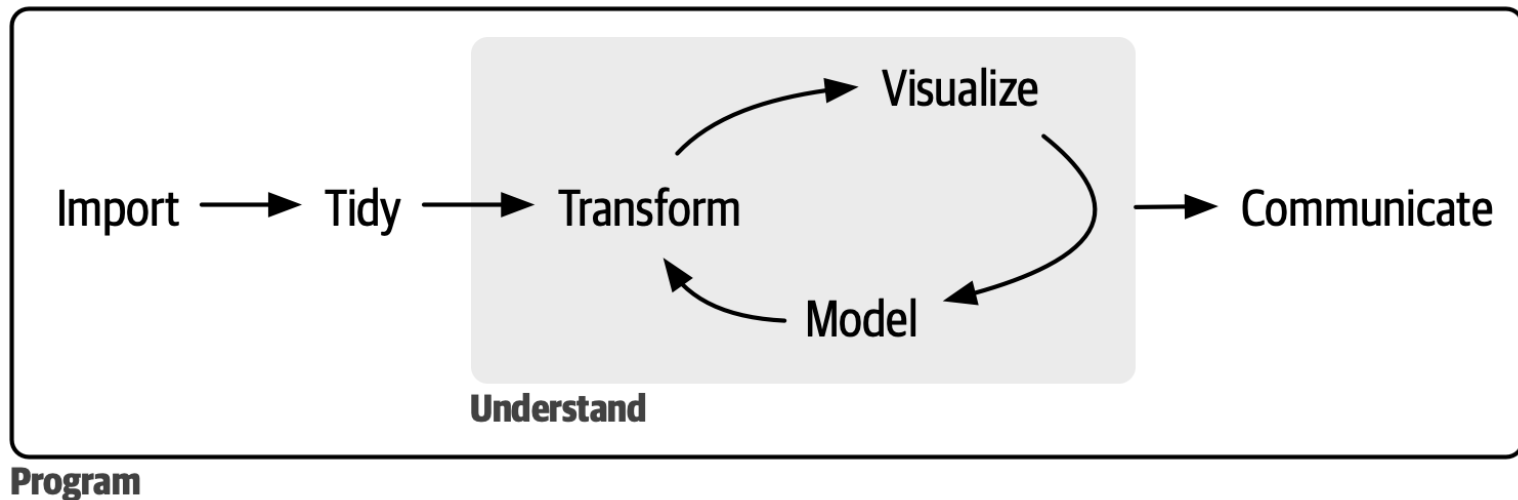
- **Transformação dos dados**
 - Vetores numéricos
 - Fatores
 - Vetores lógicos
- **Exploração dos dados**
 - Variação
 - Valores em falta
 - Covariação
 - Padrões e Modelos
- **Modelação**
 - Construir modelo simples
 - Construir modelo com workflow()
 - Construir vários modelos com workflow()



Recapitulação



» Esquema geral



H. Wickham M. Çetinkaya-Rundel & G. Grolemund (2023)

Dia 1 - Transformação





Dia 1 - Transformação



- **Vetores numéricos** → 09:30 - 11:30 e 11:45 - 12:30
 - Contagens (c/ exercícios)
 - Transformações numéricas (c/ exercícios)
 - Transformações genéricas (c/ exercícios)
 - Estatísticas descritivas (c/ exercícios)
- **Fatores** → 14:00 - 15:00
 - Operações básicas
 - Base de dados gss_cat (c/ exercícios)
 - Alterar ordem dos fatores (c/exercícios)
 - Alterar os fatores (c/ exercícios)
- **Vetores lógicos** → 15:15 - 17:00
 - Comparações (c/ exercícios)
 - Álgebra booleana (c/ exercícios)
 - Sumarização (c/ exercícios)
 - Transformações condicionais (c/ exercícios)



Recapitulação



» Pacote **dplyr**

- `filter()` Selecionar linhas (i.e. observações);
- `arrange()` Ordenar linhas (i.e. observações);
- `select()` Selecionar colunas (i.e. variáveis);
- `mutate()` Criar novas colunas (i.e. variáveis);
- `summarize()` Calcular estatísticas descritivas.
- `group_by()` Criar grupos de observações para manipulação.

<https://dplyr.tidyverse.org/reference/index.html>

<https://rstudio.github.io/cheatsheets/html/data-transformation.html>





» Pacote dplyr

```
library("tidyverse")

?diamonds

diamonds |>
  select(price, carat, cut) |>
  filter(carat < 3) |>
  mutate(lprice = log10(price)) |>
  group_by(cut) |>
  summarize(
    mean_price = mean(price)
    mean_carat = mean(carat)
  ) |>
  arrange(desc(mean_price))

#use data "diamonds"
#select "price", "carat" and "cut"
#filter for smaller diamonds
#create variable "lprice"
#group by "cut"
#calculate mean of "lprice"
#calculate mean of "carat"
#arrange by "mean_lprice"
```



Recapitulação



» Pacote dplyr

```
# A tibble: 5 × 3
  cut      mean_lprice mean_carat
<ord>      <dbl>      <dbl>
1 Fair      3.51      1.03
2 Premium   3.45      0.889
3 Good      3.41      0.847
4 Very Good 3.39      0.806
5 Ideal     3.32      0.702
```



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Dia 2 - Exploração





Dia 2 - Exploração



- **Variação** → 09:30 - 11:30
 - Valores típicos (c/ exercícios)
 - Valores invulgares
- **Valores em falta** → 11:45 - 12:30
 - Explícitos (c/ exercícios)
 - Implícitos (c/ exercícios)
 - Fatores e grupos vazios
- **Covariação** → 14:00 - 16:00
 - Uma categórica e uma numérica (c/ exercícios)
 - Duas categóricas (c/ exercícios)
 - Duas numéricas (c/ exercícios)
- **Padrões e modelos** → 16:15 - 17:00



Recapitulação



» Pacote **ggplot2**

- Data;
- Aesthetic mapping (**aes**);
- Geometric object (**geom**);
- Statistical transformation (**stat**);
- Scale;
- Themes.

<https://ggplot2.tidyverse.org/reference/index.html>

<https://rstudio.github.io/cheatsheets/html/data-visualization.html>



Recapitulação



» Pacote ggplot2

```
set.seed(1984)

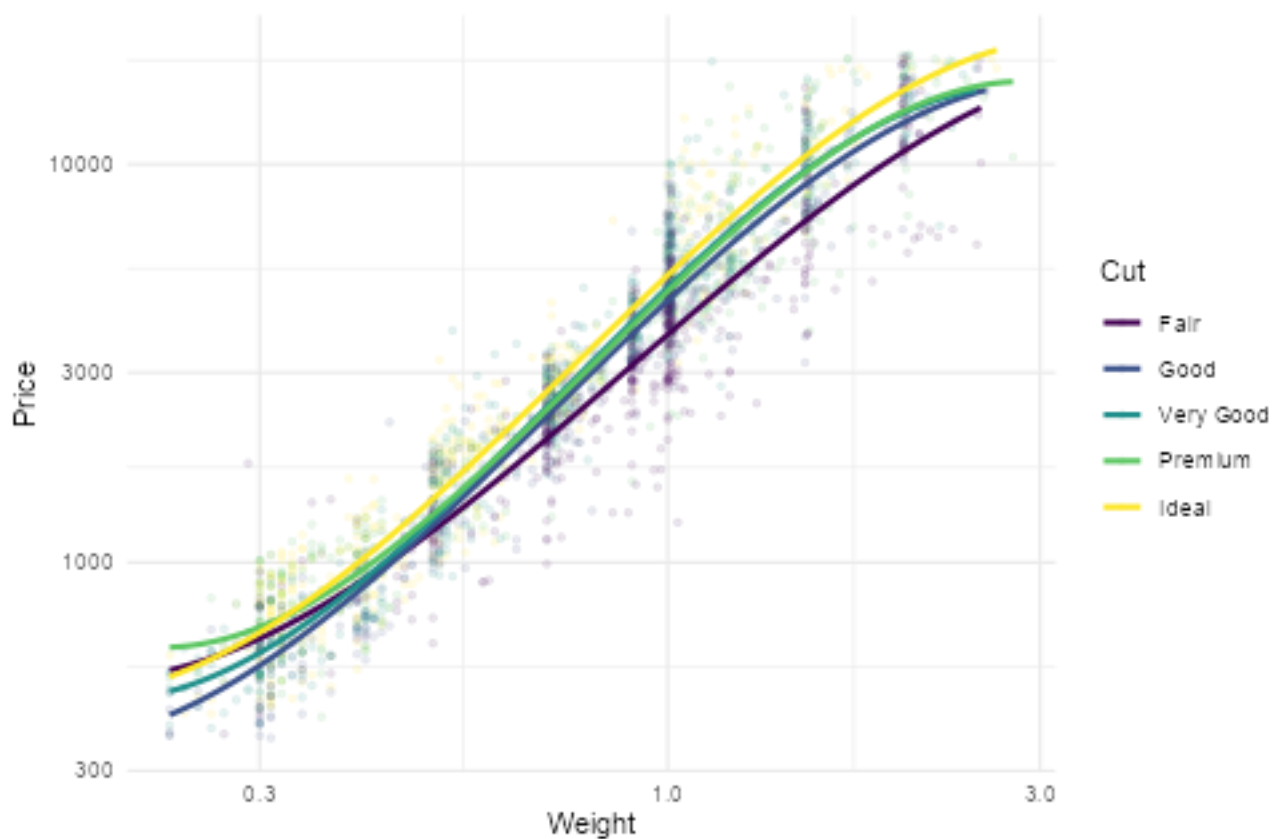
diamonds |>
  filter(carat < 3) |>
  slice_sample(n = 500, by = cut) |>
  ggplot(aes(x = carat, y = price, color = cut)) +
  geom_point(alpha = 0.1, size = 1) +
  stat_smooth(
    method = "lm",
    formula = "y ~ x + I(x^2) + I(x^3)",
    se = FALSE) +
  scale_x_continuous(trans = "log10") +
  scale_y_continuous(trans = "log10") +
  labs(x = "Weight", y = "Price", color = "Cut") +
  theme_minimal()
```

#use data "diamonds"
#filter for smaller diamonds
#sample for 500 obs per "cut"
#aesthetics mapping
#geometric object
#statistical transformation
#scale for x-axis
#scale for y-axis
#scale for labels
#change theme

Recapitulação



» Pacote ggplot2





INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Dia 3 - Modelação





Dia 3 - Modelação



- **Construir modelo simples** → 09:30 - 11:00
 - Explorar dados
 - Ajustar modelo
 - Usar modelo para previsão
- **Construir modelo com *workflow*** → 11:15 - 12:30
 - Explorar dados
 - Dividir dados
 - Criar *workflow*
 - Ajustar modelo
 - Avaliar modelo
- **Construir vários modelos com *workflow*** → 14:00 - 15:00 e 15:15 - 17:00
 - Explorar dados
 - Dividir dados
 - Criar *workflow* e ajustar modelo 1
 - Criar *workflow* e ajustar modelo 2
 - Avaliar o melhor modelo



Recapitulação



» Pacote `stats`: funções `lm()` e `summary()`

- Data;
- Formula (e.g. `formula()`);
- Fit model (eg. `lm()`, `glm()`, `aov()`, ...);
- Extract parameters (eg. `residuals()`, `predicted()`, `coef()`, ...);
- Testing assumptions (eg. `plot()`, ...);
- Evaluate model (eg. `summary()`, `AIC()`, `logLik()`, ...).

<https://www.datacamp.com/tutorial/linear-regression-R>

<https://rpubs.com/abigailpayne/743827>



Recapitulação



» Pacote `stats`: funções `lm()` e `summary()`

```
diamonds2 <- diamonds |>                                #use data "diamonds"
  filter(
    (x > 0) & (y > 0 & y < 20) & (z > 0 & z < 10), #filter out outliers
    carat < 3) |>                                         #filter in smaller diamonds
  slice_sample(n = 1000) |>                             #sample for 1000 observations
  select(price, carat, cut) |>                          #select "price", "carat" and "cut"
  mutate(
    lprice = log10(price),                               #create variable "lprice"
    lcarat = log10(carat),                               #create variable "lprice"
    fct_cut = factor(cut, ordered = FALSE))             #make variable "cut" into factor

mod1 <- formula("lprice ~ lcarat + fct_cut")             #specify model
res1 <- lm(mod1, data = diamonds2)                     #fit model to data

summary(res1)                                           #get summary
```

Recapitulação



» Pacote `stats`: funções `lm()` e `summary()`

Residuals:

Min	1Q	Median	3Q	Max
-0.33696	-0.06961	0.00255	0.07358	0.36898

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.56450	0.02020	176.449	< 2e-16	***
lcarat	1.70134	0.01384	122.938	< 2e-16	***
fct_cutGood	0.07818	0.02347	3.331	0.000896	***
fct_cutVery Good	0.09259	0.02164	4.279	2.06e-05	***
fct_cutPremium	0.08857	0.02144	4.131	3.91e-05	***
fct_cutIdeal	0.12437	0.02111	5.891	5.24e-09	***

Multiple R-squared: 0.9396, Adjusted R-squared: 0.9393

F-statistic: 3093 on 5 and 994 DF, p-value: < 2.2e-16

Recapitulação



» Pacote **performance**

MLR.1 O modelo é linear nos parâmetros;

MLR.2 Amostra aleatória (e.g. não há outliers, valores omissos aleatórios);

MLR.3 Não há multicolinearidade entre preditores;

MLR.4 Erro com valor esperado zero dado qualquer valor dos preditores;

MLR.5 Erro com variância constante dado qualquer valor dos preditores;

MLR.6 Erro é independente dos preditores e tem distribuição normal.

Wooldridge J, Introductory Econometrics: A Modern Approach, 7 ed. Thomson



» Pacote performance

```
library("performance")

check_model(res1, check = c(
  #MLR.1 The population model is linear in the parameters
  "linearity",
  #MLR.3 Random sample (e.g. no outliers, missing at random)
  "vif",
  #MLR.5 The error has constant variance given any values of the parameters
  "homogeneity",
  #MLR.6 The error is independent of the predictors and is normally distributed
  "qq"
))
```

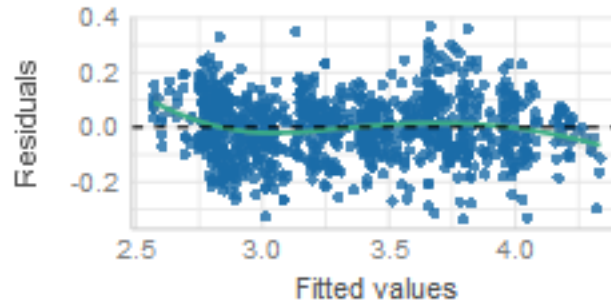
Recapitulação



» Pacote performance

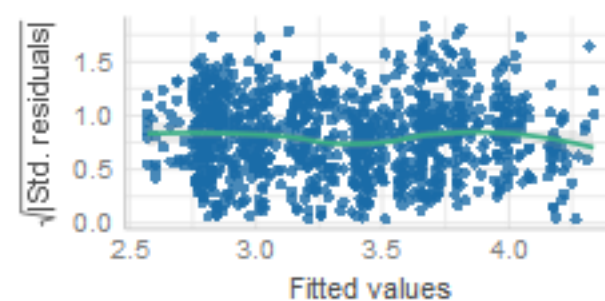
Linearity

Reference line should be flat and horizontal



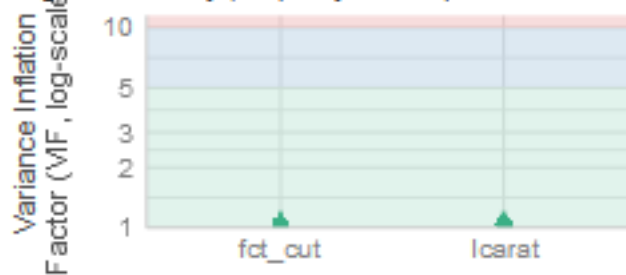
Homogeneity of Variance

Reference line should be flat and horizontal



Collinearity

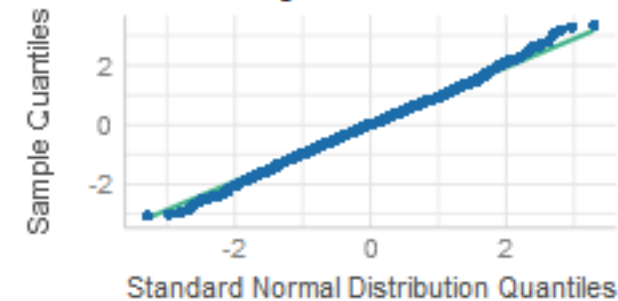
High collinearity (VIF) may inflate parameter uncer



Low (< 5)

Normality of Residuals

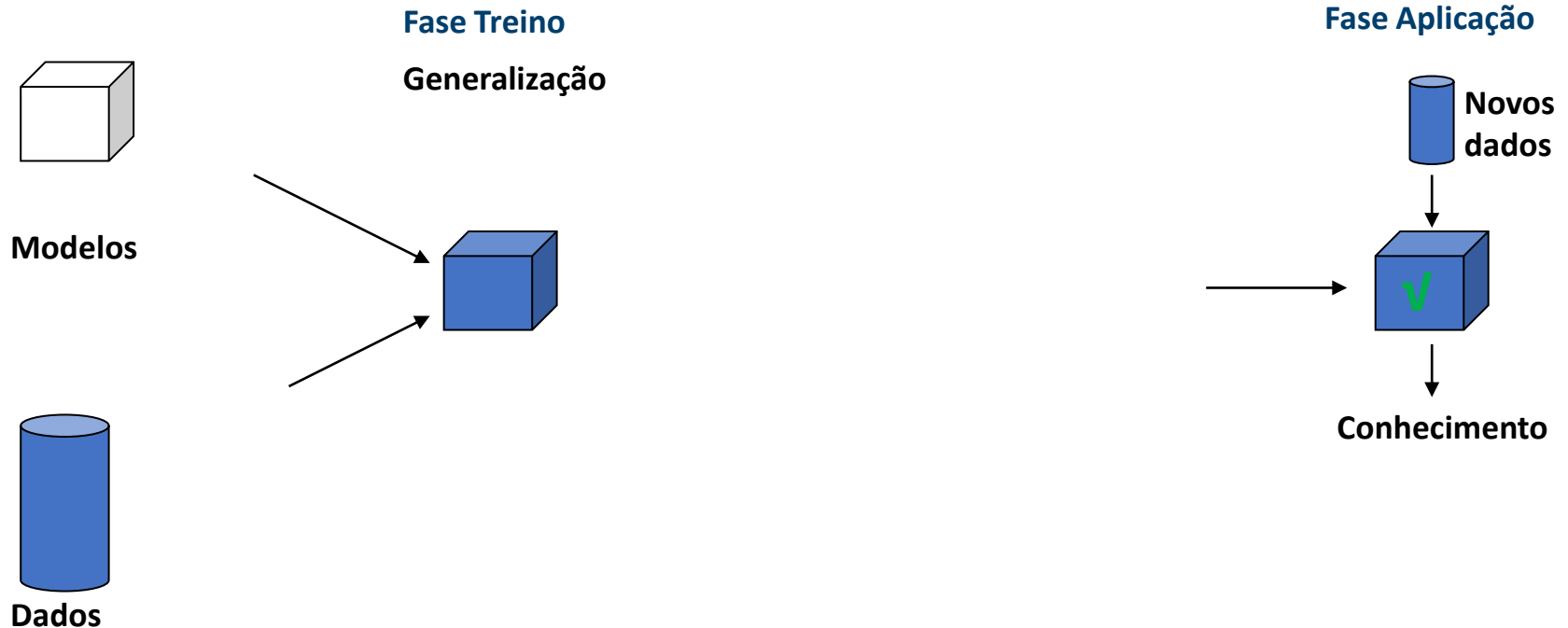
Dots should fall along the line



Avaliação de modelos



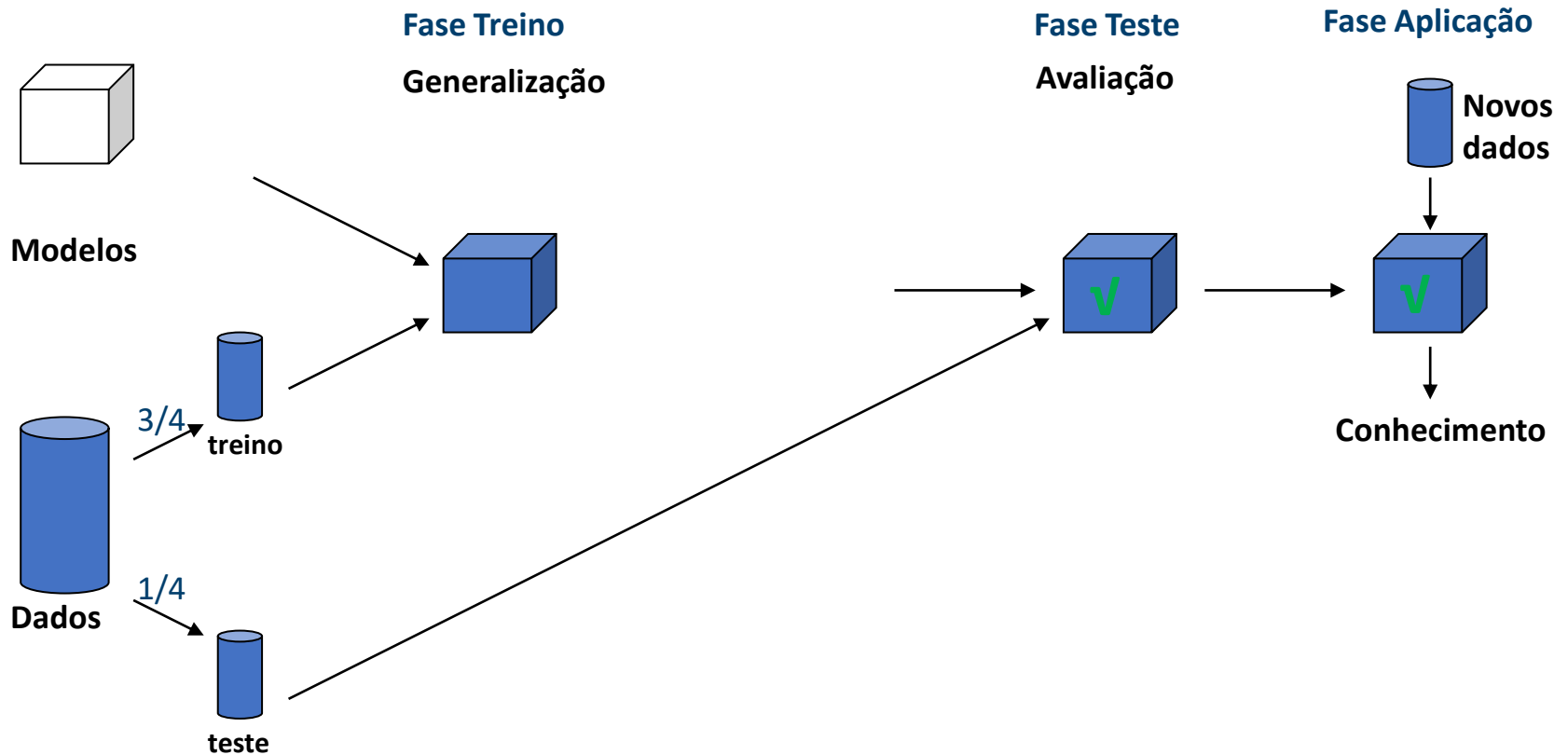
» Visão Geral



Avaliação de modelos



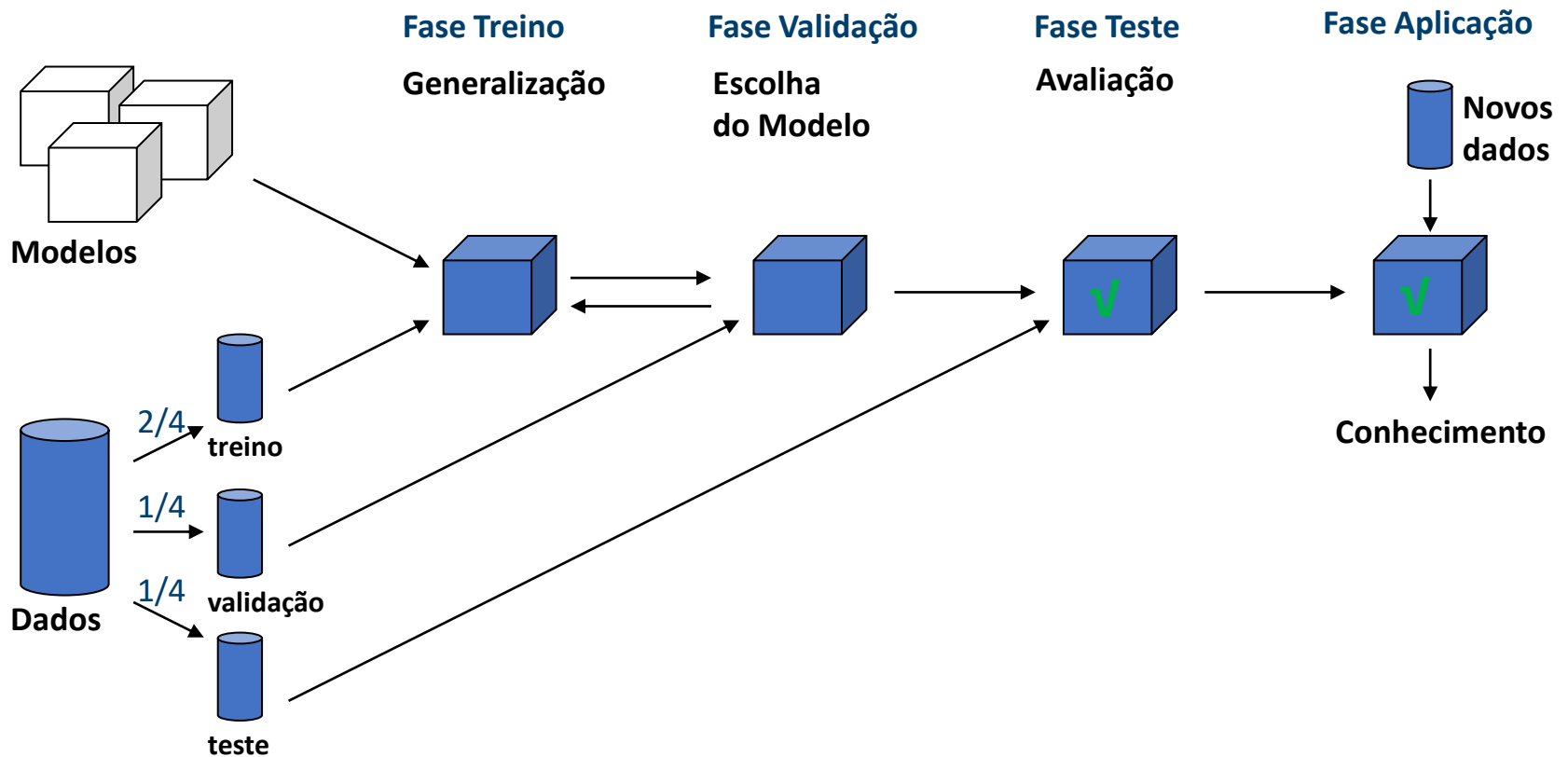
» Visão Geral



Avaliação de modelos



» Visão Geral



Avaliação de modelos



» Métricas: variável categórica (binária)

True Positive Rate (Sensitivity)

$$= TP/P$$

False Positive Rate (1 - Specificity)

$$= FP/N = 1 - TN/N$$

	Predicted		
Real	Positive	Negative	
Positive	True P.	False N.	P.
Negative	False P.	True N.	N.

» Métricas: variável contínua

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^n (\hat{y} - y)^2 / n}$$

Mean Absoute Error (MAE)

$$MAE = \sum_{i=1}^n |\hat{y} - y| / n$$



Avaliação de modelos



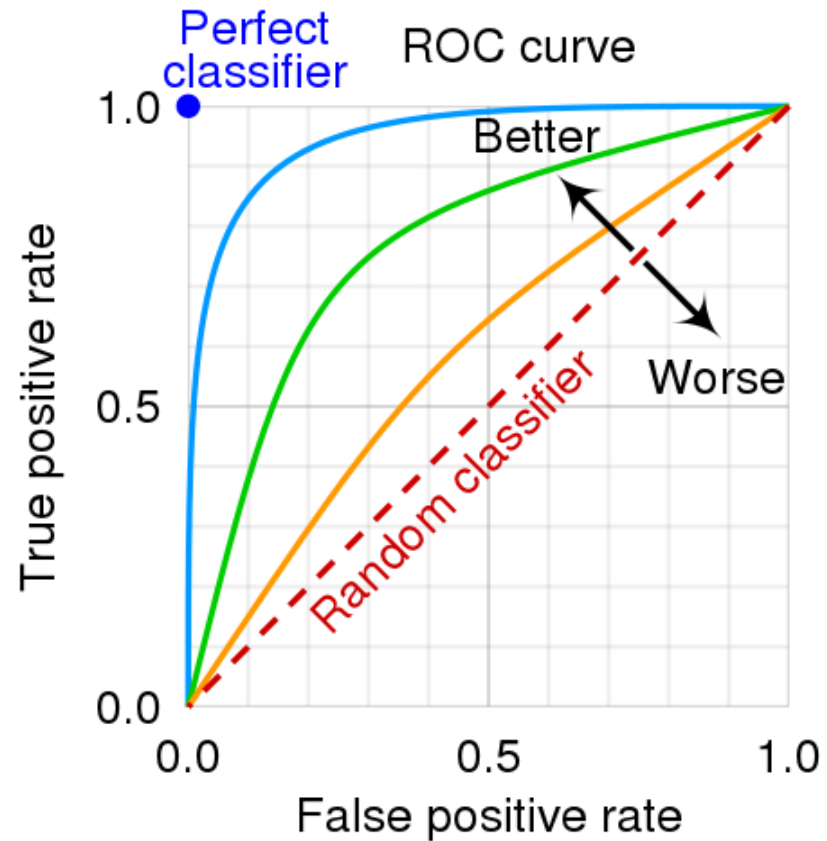
» Curva de ROC (Receiver Operating Characteristics)

True Positive Rate (Sensitivity)
= TP/P

False Positive Rate (1 - Specificity)
= $FP/N = 1 - TN/N$

Trade-off:

Sensitivity vs. Specificity



M. Thoma (2018)



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Informações finais



Informações finais



» Gestão de projectos

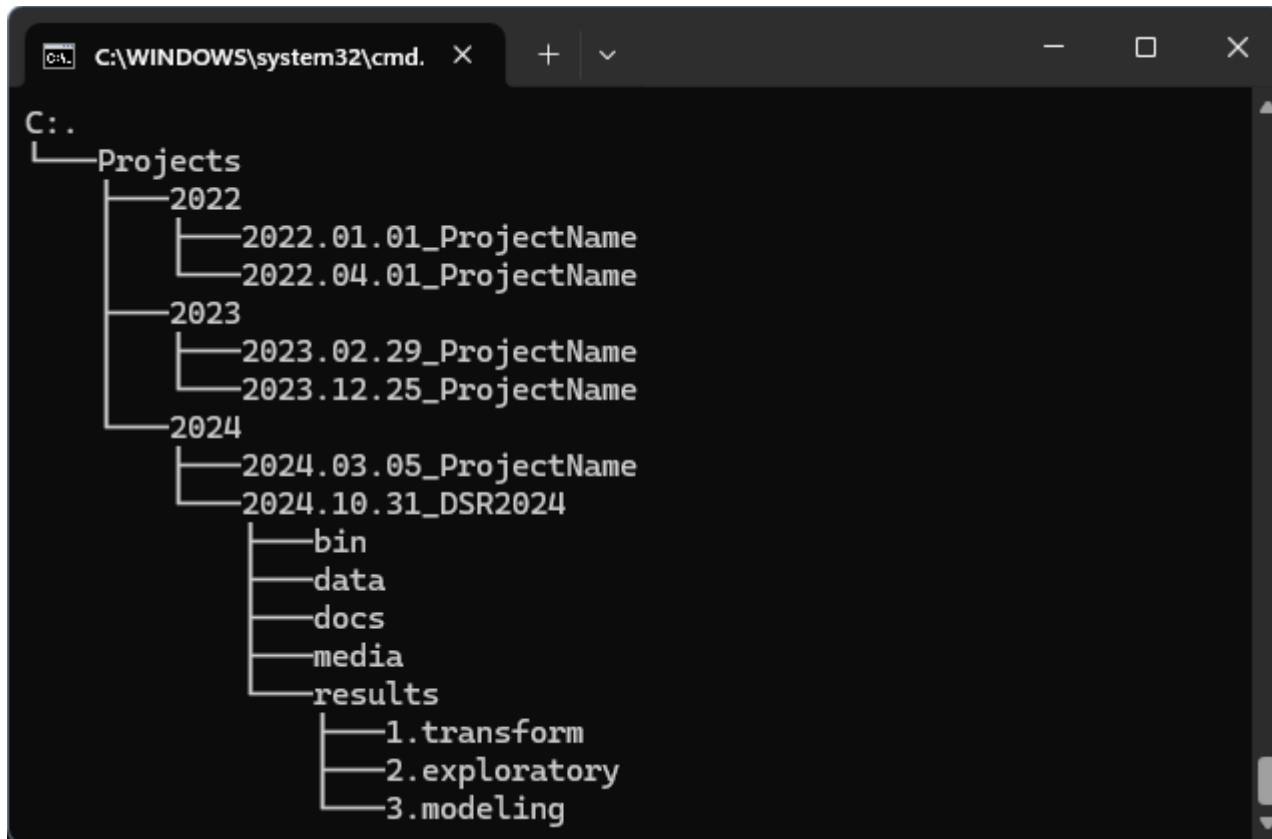
- Estrutura de pastas;
- Ficheiro README (e nomeação de ficheiros).



Informações finais



» Estrutura de pastas



```
C:\WINDOWS\system32\cmd. X + v - □ X  
C:.\n├── Projects\n│   ├── 2022\n│   │   ├── 2022.01.01_ProjectName\n│   │   └── 2022.04.01_ProjectName\n│   ├── 2023\n│   │   ├── 2023.02.29_ProjectName\n│   │   └── 2023.12.25_ProjectName\n│   └── 2024\n│       ├── 2024.03.05_ProjectName\n│       └── 2024.10.31_DSR2024\n│           ├── bin\n│           ├── data\n│           ├── docs\n│           ├── media\n│           ├── results\n│           │   ├── 1.transform\n│           │   ├── 2.exploratory\n│           │   └── 3.modeling
```

Informações finais



» Ficheiro README

```
#autor:      Joao Sollari Lopes
#local:      INE, Lisboa
#criado:     30.10.2023
#modificado: 06.05.2024

+bin
  |0.examples.r           #exemplos para recapitulacao
  |0.install_packages.r  #instalar pacotes necessarios
  |1.transform.r          #transformacao de dados
  |2.exploration.r        #exploração de dados
  |3.modelling.r          #modelacao de dados

+docs
  |DSR-II2024_program.pdf #programa
  |DSR-II2024_slides.pdf  #slides [versao final]
  |DSR-II2024_slides_20230807.pptx #slides [v2023-08-07]
  |DSR-II2024_slides_20240430.pptx #slides [v2024-04-30]

+results
  +1.tranform             #resultados de "1.transform.r"
  +2.exploration          #resultados de "2.exploration.r"
  +3.modelling            #resultados de "3.modelling.r"
README.txt               #Este ficheiro
```


Informações finais



» Comunidade R

- <https://www.r-project.org/>
- <https://www.tidyverse.org/>
- <https://www.tidymodels.org/>
- <https://education.rstudio.com/learn/>
- <https://www.r-project.org/help.html>
- <https://hour.ine.pt/>

Informações finais



» Bibliografia

» Wickham H & Grolemund G (2017) R for Data Science. O'Reilly Media Inc., Sebastopol.

URL: <https://r4ds.had.co.nz/>

» Wickham H, Çetinkaya-Rundel M & Grolemund G (2023) R for Data Science. O'Reilly Media Inc., Sebastopol. O'Reilly Media Inc., Sebastopol.

URL: <https://r4ds.hadley.nz/>

