Análise exploratória de dados

Sara Mortara, Andrea Sanchez-Tapia, Diogo S. S. Rocha

aula 5

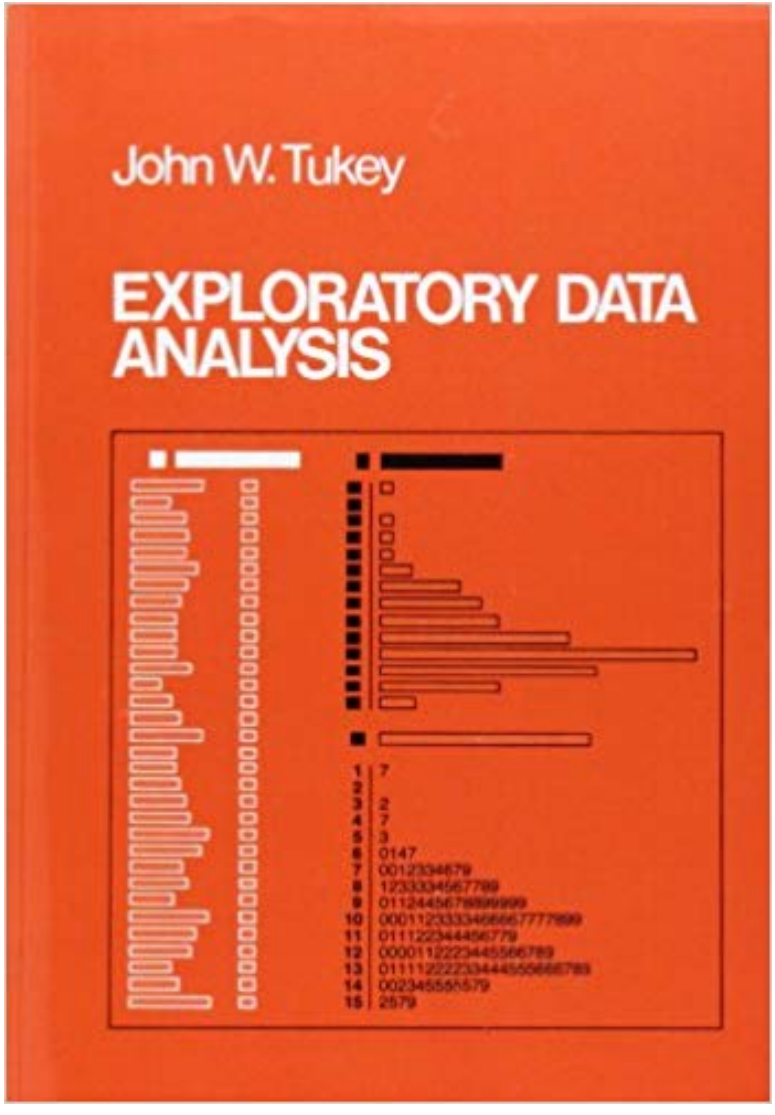
sobre a aula

1. análise exploratória de dados
2. estatísticas descritivas
3. gráficos
4. relações entre variáveis
5. extra: PCA & regressão linear

1. análise exploratória de dados (AED)

a vida sem análise exploratória de dados

Explanatory Data Analysis de John Tukey



conheça seus dados!

objetivos da AED

objetivos da AED

1. controlar a qualidade dos dados

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados
3. apoiar a escolha dos procedimentos estatísticos de testes de hipótese

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados
3. apoiar a escolha dos procedimentos estatísticos de testes de hipótese
4. avaliar se os dados atendem às premissas dos procedimentos estatísticos escolhidos

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados
3. apoiar a escolha dos procedimentos estatísticos de testes de hipótese
4. avaliar se os dados atendem às premissas dos procedimentos estatísticos escolhidos
5. indicar novos estudos e hipóteses

alerta!

alerta!

análise exploratória não é tortura de dados

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

alerta!

análise exploratória não é tortura de dados

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

assume-se que pesquisador(a) formulou *a priori* **hipóteses** plausíveis amparadas pela **teoria**

dicas

dicas

- pode levar entre 20 e 50% do tempo das análises

dicas

- pode levar entre 20 e 50% do tempo das análises
- deve ser iniciada ainda durante a coleta de dados

dicas

- pode levar entre 20 e 50% do tempo das análises
- deve ser iniciada ainda durante a coleta de dados
- utiliza-se largamente técnicas visuais

importânica do gráfico e quarteto de Anscombe

- criado pelo matemático Francis Ascombe
- 4 conjuntos de dados com as mesmas estatísticas descritivas, mas muito diferentes graficamente

os dados de Anscombe

```r
# claro que o conjunto já existe dentro do R
data("anscombe")
```

```r
# média dos dados
apply(anscombe, 2, mean)
```

```
##       x1       x2       x3       x4       y1       y2       y3       y4
## 9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
```

```r
# variância dos dados
apply(anscombe, 2, var)
```

```
##        x1        x2        x3        x4        y1        y2        y3
## 11.000000 11.000000 11.000000 11.000000  4.127269  4.127629  4.122620
##        y4
##  4.123249
```

vamos olhar para os dados

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

correlação entre x e y

```r
# correlação
cor(anscombe$x1, anscombe$y1)
```

```
## [1] 0.8164205
```

```r
cor(anscombe$x2, anscombe$y2)
```

```
## [1] 0.8162365
```

```r
cor(anscombe$x3, anscombe$y3)
```

```
## [1] 0.8162867
```

```r
cor(anscombe$x4, anscombe$y4)
```

```
## [1] 0.8165214
```

coeficientes da regressão linear de x e y

```r
# coeficientes da regressão
coef(lm(anscombe$y1 ~ anscombe$x1))
```

```
## (Intercept) anscombe$x1
##   3.0000909   0.5000909
```

```r
coef(lm(anscombe$y2 ~ anscombe$x2))
```

```
## (Intercept) anscombe$x2
##    3.000909    0.500000
```

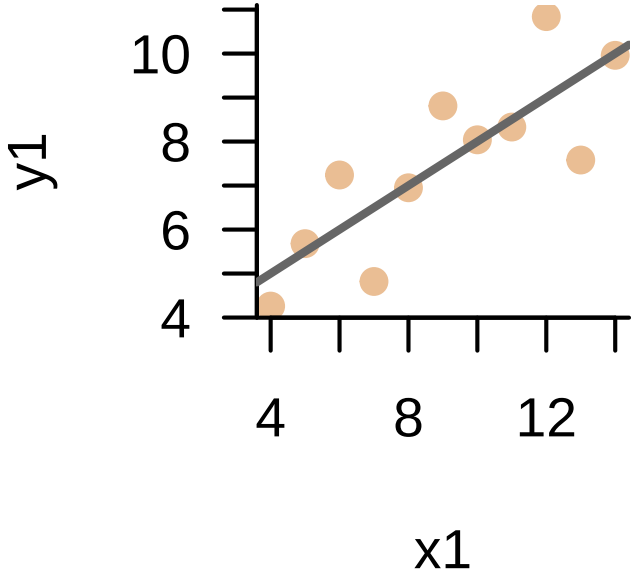
```r
coef(lm(anscombe$y3 ~ anscombe$x3))
```

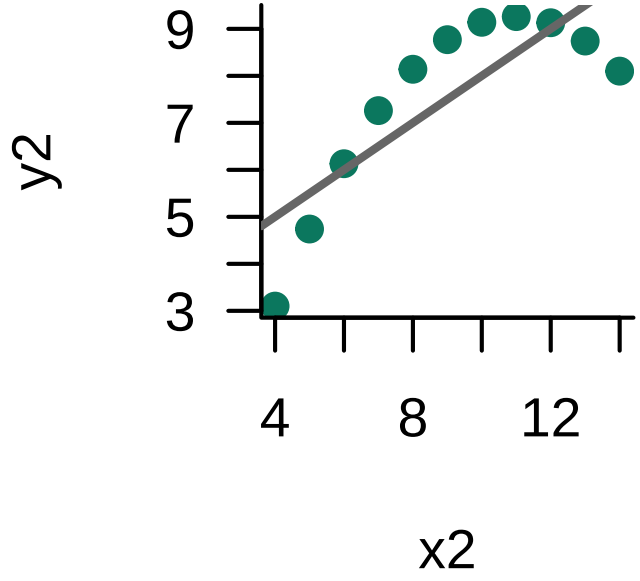
```
## (Intercept) anscombe$x3
##   3.0024545   0.4997273
```

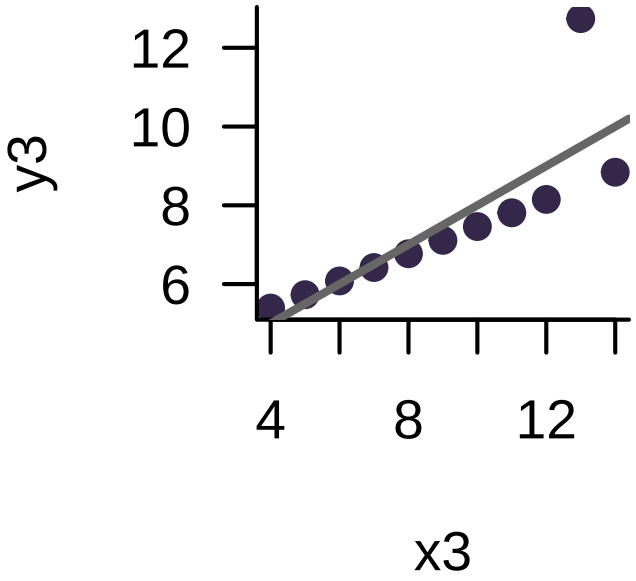
```r
coef(lm(anscombe$y4 ~ anscombe$x4))
```

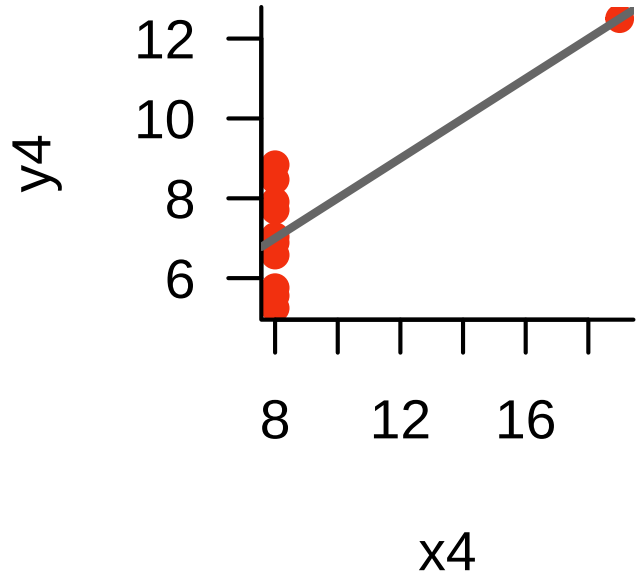
```
## (Intercept) anscombe$x4
##   3.0017273   0.4999091
```

agora sim vamos olhar para os dados do Anscombe









perguntas que nos devemos fazer

1. Onde os dados estão centrados? Como os dados estão distribuídos? Os dados são simétricos, assimétricos, bimodais?
2. Existem outliers?
3. As variáveis seguem uma distribuição normal?
4. Existem relações entre as variáveis? As relações entre variáveis são lineares?
5. As variáveis precisam ser transformadas?
6. O esforço amostral foi o mesmo para cada observação ou variável?

2. estatísticas descritivas

conferência de dados no R

```
# lendo os dados da idade da população que usa fraldas
#fraldas <- read.csv("../data/idade_fraldas.csv")
```

3. gráficos

4. relações entre variáveis

5. extra: PCA & regressão linear