Análise Exploratória de Dados — Capítulo 1

Aprendizagem Estatística CC

Soraia Pereira

17 de setembro de 2025

EDA: objetivo

Análise Exploratória de Dados (EDA) visa caracterizar a distribuição e a estrutura dos dados (qualidade, heterogeneidade, dependências e forma), através de resumos numéricos e visualizações gráficas, para formular hipóteses e orientar as etapas seguintes da análise/modelação.

Ciclo (Grolemund & Wickham, 2017):

1. Colocar questões sobre os dados.
2. Explorar com visualização e medidas resumo.
3. Refinar questões com base no que foi descoberto e repetir.

Nota: o processo é iterativo e guiado pelas questões; não é um procedimento único.

Tipos de variáveis

Quantitativas

- **Discretas**: contagens (ex.: # transações).
- **Contínuas**: medidas em $\mathbb{R}$ (ex.: tempo, altura).

Qualitativas

- **Nominais**: categorias sem ordem (ex.: região, sexo).
- **Ordinais**: categorias com ordem (ex.: nível de educação).

Implicações: As medidas resumo e os gráficos adequados dependem do tipo de variáveis.

Dados de exemplo: `gapminder`

`library(gapminder)` (se pretender usar em R)

Visualização das 6 primeiras linhas:

```
> head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333       779.
2 Afghanistan Asia       1957    30.3  9240934       821.
3 Afghanistan Asia       1962    32.0 10267083       853.
4 Afghanistan Asia       1967    34.0 11537966       836.
5 Afghanistan Asia       1972    36.1 13079460       740.
6 Afghanistan Asia       1977    38.4 14880372       786.
```

Boa prática: documente unidade das variáveis, período temporal e origem dos dados.

Medidas resumo

Localização (tendência central)

Considere a amostra $(x_1, ..., x_n)$ de dimensão n.

- **Média**: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (ponto de equilíbrio; sensível a outliers)
- **Mediana**: medida de localização que divide a distribuição "ao meio".

$$\widetilde{x} = Q_{1/2} = \begin{cases} \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se n é par} \\ x_{(\frac{n+1}{2})}, & \text{se n é ímpar} \end{cases}$$

onde $x_{(i)}$ é o elemento da amostra ordenada que ocupa a posição i.

- **Moda**: valor mais frequente (pouco estável em amostras pequenas).

Localização

- **Quantil de probabilidade** p: valor Q_p tal que $p \times 100\%$ dos valores da amostra são inferiores a Q_p e $(1-p) \times 100\%$ são superiores a Q_p.

$$Q_p = \begin{cases} \frac{x_{(np)} + x_{(np+1)}}{2}, & \text{se np é inteiro} \\ x_{([np]+1)}, & \text{se np não é inteiro} \end{cases}$$

  Os quantis mais utilizados são os quartis $(Q_{1/4}, Q_{1/2}, Q_{3/4})$.

Média vs Mediana (assimetria)

- Distribuição **simétrica**: $\overline{x} \approx \widetilde{x}$.
- **Assimetria positiva**: $\overline{x} > \widetilde{x}$.
- **Assimetria negativa**: $\overline{x} < \widetilde{x}$.

Dispersão

- **Variância**: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sum x_i^2 - n\,\overline{x}^2}{n-1}$
- **Desvio-padrão**: $s = \sqrt{s^2}$
- **Coeficiente de variação**: $CV = s/\overline{x}$
- **Amplitude**: $R = x_{(n)} - x_{(1)}$
- **Intervalo inter-quartis**: $\mathrm{IQR} = Q_3 - Q_1$

Gráficos básicos

Gráfico de Barras

- Variáveis **nominais/ordinais**; alturas $\propto$ frequências (absolutas/relativas).
- Ordenar barras pode clarificar.

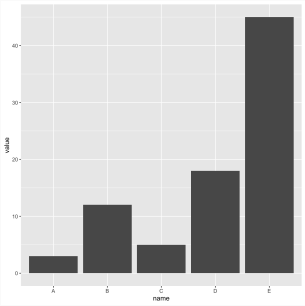

Figura 1: Exemplo de gráfico de barras

Histograma

- Variáveis **contínuas**; barras adjacentes indicam continuidade.
- Largura de classes (*binwidth*) influencia a leitura.
- Regras úteis: *Freedman–Diaconis* $h = 2\,\mathrm{IQR}\,n^{-1/3}$, *Scott*, *Sturges*.

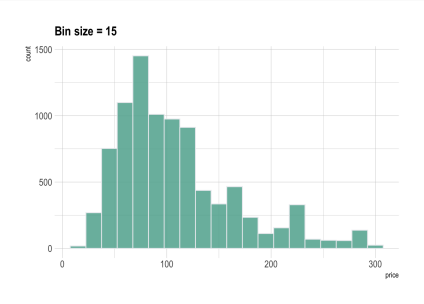

Figura 2: Exemplo de histograma

Boxplot

- Caixa: $[Q_1, Q_3]$; linha central: $\widetilde{x}$; bigodes até limites.
- $\mathrm{IQR} = Q_3 - Q_1$
- **Outliers**: pontos fora de $[Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}]$.

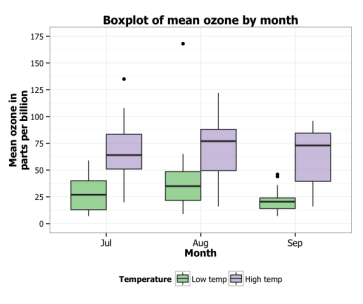

Figura 3: Exemplo de boxplot

Qualidade dos dados

Valores em falta e outliers

Valores em falta (NA):

- Quantificar por variável
- Estratégias: análise dos dados completos, imputação simples (mediana), métodos múltiplos (quando relevante).

Outliers:

- Diagnóstico via boxplot, análise contextual.
- Tratar *com justificação*: corrigir erros, manter e usar medidas robustas, ou análises com/sem.

Associação entre variáveis

Dispersão e associação

- **Gráfico de dispersão** é a primeira ferramenta para análise da associação entre duas variáveis quantitativas.
- **Homocedasticidade**: dispersão semelhante ao longo de X.
- **Cuidado**: padrões curvilíneos, clusters e pontos influentes.

Gráfico de Dispersão (scatterplot)

- Usado para **duas variáveis quantitativas** — pares (x_i, y_i).
- **Revela** forma da relação: linear/curvilínea, **direção** (+/−) e **força** (nuvem mais/menos concentrada).

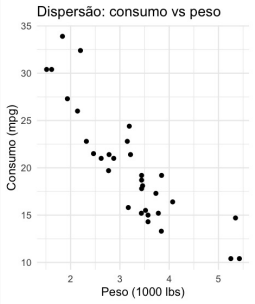

Figura 4: Exemplo de gráfico de dispersão.

Coeficiente de correlação amostral de Pearson (r)

- **Mede** a **força** e a **direção** da **associação linear** entre duas variáveis quantitativas (X, Y).
- **Intervalo**: $r \in [-1, 1]$.
  - $r \approx +1$: associação linear positiva muito forte.
  - $r \approx -1$: associação linear negativa muito forte.
  - $r \approx 0$: ausência de *associação linear* (pode existir relação não linear).

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{s_X \, s_Y}.$$

Coeficiente de correlação amostral de Pearson (r)

- **Propriedades**: adimensional; invariante a transformações lineares de escala
- **Atenção**: muito **sensível a outliers**; só capta **linearidade**; $r \approx 0$ não implica independência; **não implica causalidade**.
- **Boa prática**: inspecione o *scatterplot* antes de interpretar r.

Coeficiente de correlação amostral de Spearman (r_s)

- **Mede** a **força** e a **direção** da **associação monótona** (crescente/decrescente) entre duas variáveis *quantitativas ou ordinais*.
- **Baseia-se nas posições (ranks)** das observações.
- **Intervalo**: $r_s \in [-1, 1]$.
  - $r_s \approx +1$: monotonia crescente quase perfeita.
  - $r_s \approx -1$: monotonia decrescente quase perfeita.
  - $r_s \approx 0$: ausência de *associação monótona*.

$$\boxed{r_s \;=\; \mathrm{cor}_{\mathrm{Pearson}}\big(\mathrm{rank}(X),\, \mathrm{rank}(Y)\big)}$$

(sem empates) $\quad r_s \;=\; 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad d_i = rank(x_i) - rank(y_i).$

Coeficiente de correlação amostral de Spearman (r_s)

- **Empates**: usa *posições médias*
- **Propriedades**: mais **resistente a outliers** do que Pearson; invariante a transformações **monótonas estritamente crescentes**; adequado para dados **ordinais** e relações não lineares mas monótonas.
- **Boa prática**: ver *scatterplot* (ou dos *ranks*) antes de interpretar r_s.

Boas práticas e checklist

Transformações e escalas

- Transformações (log, raiz) podem reduzir assimetria/heterocedasticidade.
- Reporte resultados na **escala original** quando necessário para interpretação.
- Eixos e unidades **claros**.

Checklist de EDA

- **Contexto**: dicionário dos dados, unidades, período, fonte.
- **Qualidade**: NA, duplicados, intervalos plausíveis.
- **Univariada**: centro ($\widetilde{x}/\overline{x}$), dispersão ($\mathrm{IQR}/s$), forma e outliers.
- **Bivariada**: gráficos adequados; correlações; estratificar por grupos relevantes.
- **Documentação**: escolhas (binwidth, método de quantis), critérios de outliers e decisões tomadas.

Referências

- Wickham, H. & Grolemund, G. (2017). *R for Data Science*.
- Devore, J., Berk, K., Carlton, M. (2021). *Modern Mathematical Statistics with Applications*.