Script 3 — `3_Data_Understanding.R`

Objetivo do capítulo (CRISP-DM: Data Understanding)

- **estrutura** do dataset (dimensões, tipos, `str`)
- **qualidade** dos dados (duplicados, missing, coerência do target)
- **distribuições** (target e variáveis-chave)
- **binárias** (proporções e detecção de variáveis raras / quase constantes)
- **relações** (scatter target vs variável chave)
- **correlações** (entre numéricas e com o target)

Este script regista outputs e guias auxiliares, em formato .txt e .csv, para a reliação do treinamento de modelos

2) Função principal: `run_cap3(df, out_tables, out_figs)`

Inputs

- `df`: dataset a analisar (no pipeline é `df_raw`)
- `out_tables`: diretório para outputs tabulares/texto (CSV/TXT)
- `out_figs`: diretório para outputs gráficos (PNG)

Output

Lista com:

- `n_dup`: nº de linhas duplicadas
- `missing_summary`: tabela de missing por variável
- `bin_cols`: lista de colunas binárias 0/1 identificadas

---

3) Passos do capítulo (3.0 a 3.10)

3.0 Snapshot da estrutura do dataset

```
write_text_snapshot(df, "cap3_estrutura_dataset.txt")
```

Gera um TXT com:

- dimensões (`dim`)
- nomes das variáveis
- `str(df)`

Ficheiro gerado

- cap3_estrutura_dataset.txt

3.1 Duplicados

```
n_dup <- sum(duplicated(df))
```

Conta duplicados “brutos” (linhas exatamente iguais).

Ficheiro

- cap3_duplicados.txt

3.2 Missing values

```
missing_summary <- missing_summary_df(df)
```

Cria tabela com `n_missing` e `pct_missing` por variável.

Ficheiro

- cap3_missing_summary.csv

3.3 Coerência básica

```
msgs <- coherence_messages(df, target="score_review", ...,
numeric_var="logavaliacoes")
```

Gera mensagens (TXT) com:

- range do target
- nº de valores fora de [1,5]
- resumo de `logavaliacoes` (se existir)
- avisos se variáveis não existirem

Ficheiro

- cap3_coerencia_basica.txt

3.4 Distintos por coluna (cardinalidade)

```
uniq_df <- unique_counts_df(df)
```

Conta nº de valores distintos por variável.

Ficheiro

- `cap3_unique_counts.csv`

Como usar no relatório

- variáveis com **muito poucos valores** podem ser binárias/categóricas
- variáveis com **muitos valores únicos** podem ser IDs/strings livres (potencial ruído)
- ajuda a explicar feature engineering no Cap. 4

3.5 Summary geral

```
write_summary_txt(df, "cap3_summary_geral.txt")
```

Guarda `summary(df)` num TXT (min, median, mean, quartis, etc. para numéricas; contagens para fatores/strings).

Ficheiro

- `cap3_summary_geral.txt`

Uso

- rápido para entender escala, assimetria e valores extremos.

3.6 Distribuição do target + frequências

Se `score_review` existir:

- cria tabela de frequências do target
- guarda gráfico de barras

Ficheiros

- `cap3_score_review_frequencias.csv`
- `cap3_dist_score_review.png`

Se não existir:

- `cap3_target_ausente.txt` (com aviso)

Interpretação

- detecta "desbalanceamento" (ex.: muitos 4 e 5, poucos 1 e 2)
- justifica **estratificação** no split e folds (Cap. 4 e Cap. 5)

3.7 `logavaliacoes`: histograma e boxplot

Se a variável existir:

- histograma → forma da distribuição (assimetria)
- boxplot → outliers e amplitude

Ficheiros

- `cap3_hist_logavaliacoes.png`
- `cap3_box_logavaliacoes.png`

Se não existir:

- `cap3_logavaliacoes_ausente.txt`

Porque é relevante

- `logavaliacoes` tende a ter caudas; isto ajuda a justificar **winsorização**.

3.8 Binárias 0/1 (proporções + "NZV" simples)

Aqui há uma parte metodologicamente importante:

(a) Coerção segura para numérico

```
df_num <- coerce_all_to_numeric_safely(df)
```

Tenta transformar tudo em numérico para permitir detetar binárias mesmo se vieram como texto.

(b) Proporções e raridade

```
bin_info <- binary_props_df(df_num, thr_nzv = 0.02)
```

Gera:

- `bin_cols`: colunas binárias identificadas
- `props_df`: proporção de 1 em cada binária
- `nzv_df`: binárias "raras" (min(prop,1-prop) < 0.02)

Ficheiros

- `cap3_colunas_binarias_01.txt`
- `cap3_proporcoes_binarias.csv`
- `cap3_binarias_raras_NZV.csv`
- `cap3_proporcoes_binarias.png`

Se não houver binárias:

- cap3_binarias_01_vazias.txt

Interpretação / ligação ao Cap. 4

- binárias com proporção muito baixa (quase sempre 0) são candidatas a:
  - remover (pouca informação)
  - agrupar/engenharia
  - manter se forem "raras mas fortes"
- isto antecipa decisões sobre "near zero variance" (NZV) e estabilidade do modelo

---

3.9 Relação score_review vs logavaliacoes

Se ambas existirem:

```
plot_scatter(df, xvar="logavaliacoes", yvar="score_review")
```

Ficheiro

- cap3_scatter_score_vs_logavaliacoes.png

Como usar

- avalia visualmente se há tendência (linear ou não)
- dá uma pista se modelos não-lineares (árvores/RF/GBM) podem capturar padrões que modelos lineares não apanham tão bem

---

3.10 Correlações (numéricas válidas)

Passo em duas fases:

(a) filtra numéricas "válidas"

```
df_num_only <- valid_numeric_cols(df_num)
```

Remove:

- colunas não numéricas (após coerção)
- colunas constantes / sem variância
- colunas com dados insuficientes

(b) calcula correlações

```
cor_out <- correlation_outputs(df_num_only, target="score_review")
```

Se executar sem erro:

- grava `cap3_matriz_correlacao.csv`
- grava `cap3_correlacao_com_score_review.csv` (ranking por correlação com o target)

Se ocorrer algum erro:

- `cap3_correlacao_erro.txt` com mensagem

Se não houver correlação com target:

- `cap3_correlacao_target_indisponivel.txt`