Coleta de Dados e Qualidade dos Dados

Planejar a Coleta de Dados e Identificar Problemas de Qualidade

Márcio Nicolau

2025-10-29

Table of contents

List of Figures

Introdução e Objetivos

Nas aulas anteriores, exploramos as distribuições de probabilidade, amostragem, estimação e a lógica dos testes de hipóteses. Todos esses conceitos, por mais sofisticados que sejam, dependem fundamentalmente de uma base crucial: **dados de boa qualidade**. De nada adianta aplicar as técnicas mais avançadas de análise se os dados de entrada são falhos, incompletos ou viesados.

A **coleta de dados** é a primeira e uma das mais importantes etapas em qualquer projeto de análise. Um planejamento cuidadoso nesta fase pode prevenir inúmeros problemas futuros. No entanto, mesmo com o melhor planejamento, os dados raramente vêm perfeitos. A **qualidade dos dados** refere-se à sua adequação para o uso pretendido, e a identificação e tratamento de problemas de qualidade são habilidades essenciais para qualquer cientista de dados.

Nesta aula, abordaremos o processo de planejamento da coleta de dados, os métodos comuns de aquisição e, em seguida, focaremos nas dimensões da qualidade dos dados e nos problemas mais frequentes que podem surgir, ilustrando como identificá-los usando Python e R.

Objetivos de Aprendizagem

Ao final desta aula, você será capaz de:

- Compreender a importância do planejamento na coleta de dados.
- Identificar diferentes métodos para coletar dados.
- Definir as principais dimensões da qualidade dos dados.
- Reconhecer problemas comuns de qualidade nos dados (ausentes, outliers, inconsistências, duplicatas).
- Utilizar Python e R para realizar verificações básicas de qualidade dos dados.
- Entender a relação entre a qualidade dos dados e a validade da análise estatística.

O Processo de Coleta de Dados

A coleta de dados é o processo sistemático de reunir e medir informações de uma variedade de fontes para obter um quadro completo e preciso de uma área de interesse. Uma coleta bem-sucedida garante que os dados sejam relevantes, precisos e confiáveis.

Etapas Essenciais no Planejamento da Coleta

1. **Definição Clara do Objetivo e da Pergunta de Pesquisa:**
   - O que você quer descobrir? Quais perguntas você quer responder? Isso direciona todo o processo.
2. **Identificação da População de Interesse:**
   - Quem ou o que você quer estudar? (Pessoas, empresas, transações, etc.).
3. **Escolha do Método de Coleta:**
   - Como você vai obter os dados? (Pesquisas, observação, sensores, dados existentes, etc.).
4. **Seleção da Amostra (se aplicável):**
   - Se um censo não for viável, como você selecionará um subconjunto representativo da população? (Revisar Amostragem e Tipos de Amostras).
5. **Desenvolvimento dos Instrumentos de Coleta:**
   - Formulários, questionários, roteiros de entrevista, protocolos de observação.
6. **Definição do Formato e Armazenamento dos Dados:**
   - Como os dados serão estruturados e onde serão armazenados? (Planilhas, bancos de dados, arquivos de texto).
7. **Plano de Qualidade e Limpeza:**
   - Como você garantirá a qualidade dos dados durante e após a coleta?
8. **Considerações Éticas e de Privacidade:**
   - Como garantir a confidencialidade e o consentimento dos participantes?

Diagrama do Fluxo de Planejamento da Coleta de Dados

Métodos de Coleta de Dados

A escolha do método depende da natureza do estudo, dos recursos disponíveis e do tipo de dados que se deseja coletar.

1. **Pesquisas (Questionários e Entrevistas):**
    - **Questionários:** Conjunto padronizado de perguntas, autoaplicável ou assistido, para coletar dados de um grande número de pessoas.
        - *Vantagens:* Eficiente, padronizado.
        - *Desvantagens:* Rigidez, baixa taxa de resposta para autoaplicáveis.
    - **Entrevistas:** Interação direta com os participantes, permitindo aprofundamento e flexibilidade.
        - *Vantagens:* Riqueza de detalhes, adaptabilidade.
        - *Desvantagens:* Demorado, caro, viés do entrevistador.
2. **Observação:**
    - O pesquisador observa comportamentos, eventos ou fenômenos em seu ambiente natural, sem intervenção.
        - *Vantagens:* Coleta dados em tempo real, capta nuances.
        - *Desvantagens:* Pode ser influenciado pela presença do observador, demorado.
3. **Dados Secundários (Bases Existentes, Web Scraping):**
    - Utilização de dados que já foram coletados por outras fontes para outros propósitos.
        - *Vantagens:* Rápido, baixo custo, acesso a grandes volumes de dados.
        - *Desvantagens:* Qualidade e relevância podem ser limitadas, dados podem estar desatualizados ou não serem adequados ao objetivo.
        - **Web Scraping:** Extração de dados de websites.
            - *Cuidado:* Respeitar termos de serviço e legalidade.
4. **Experimentos:**
    - Manipulação de uma ou mais variáveis (independentes) para observar o efeito em outras variáveis (dependentes) em um ambiente controlado.
        - *Vantagens:* Permite estabelecer relações de causa e efeito.
        - *Desvantagens:* Pode ser artificial, questões éticas, caro.
5. **Sensores e IoT (Internet das Coisas):**
    - Dispositivos que coletam dados automaticamente do ambiente físico (temperatura, umidade, localização, atividade).
        - *Vantagens:* Coleta contínua e em tempo real, grande volume.
        - *Desvantagens:* Custo de infraestrutura, desafios de armazenamento e processamento.

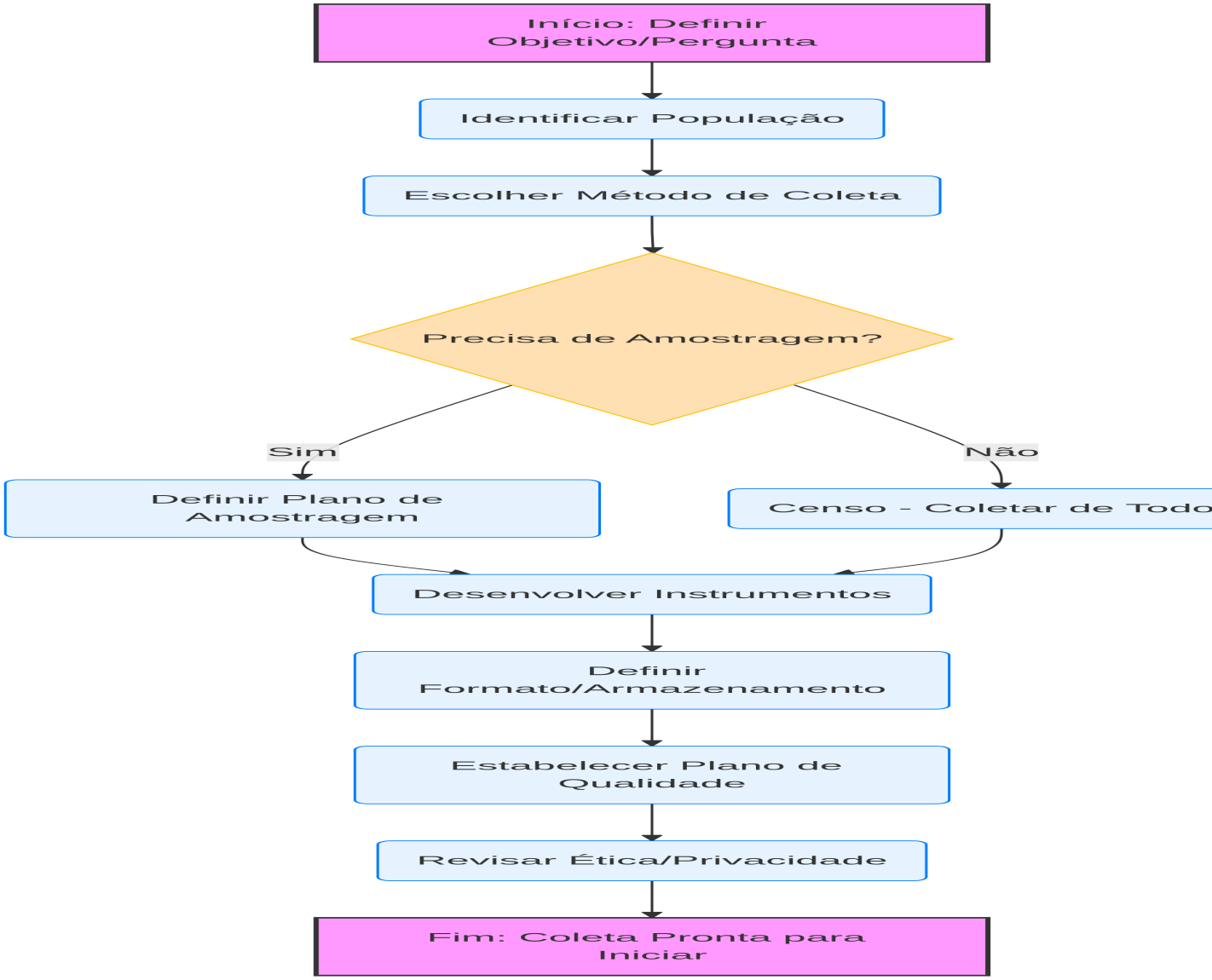

Figure 1: Fluxo de Planejamento da Coleta de Dados

Qualidade dos Dados

A **qualidade dos dados** refere-se ao grau em que os dados são adequados para um propósito específico. Dados de baixa qualidade podem levar a análises incorretas, insights enganosos e decisões ruins.

Dimensões Chave da Qualidade dos Dados

(Bussab; Morettin, 2017, p. 11–12)

1. **Acurácia (Exatidão):** Os dados refletem a realidade de forma verdadeira e correta?
   - *Exemplo:* Um registro de salário realmente corresponde ao salário pago ao funcionário.
2. **Completude:** Todos os valores esperados estão presentes? Não há valores ausentes onde deveriam existir.
   - *Exemplo:* Nenhuma entrada para "idade" está vazia em um formulário obrigatório.
3. **Consistência:** Os dados são coerentes entre si e com outras fontes? Não há contradições.
   - *Exemplo:* O CEP de um endereço corresponde à cidade e estado informados. A data de nascimento não é posterior à data atual.
4. **Pontualidade (Atualidade):** Os dados são recentes o suficiente para o propósito da análise?
   - *Exemplo:* Dados de estoque refletem a quantidade atual de produtos disponíveis.
5. **Validade:** Os dados estão em conformidade com as regras de negócio e os formatos definidos?
   - *Exemplo:* Um número de telefone tem 9 dígitos, um campo "gênero" só aceita 'M' ou 'F'.
6. **Unicidade:** Não há registros duplicados para a mesma entidade.
   - *Exemplo:* Cada cliente é representado por um único registro no banco de dados.

Problemas Comuns na Qualidade dos Dados e Como Identificá-los

Mesmo com um bom planejamento, problemas de qualidade são inevitáveis. A capacidade de identificá-los é o primeiro passo para a limpeza.

Valores Ausentes (Missing Values)

Dados que não foram registrados ou estão indisponíveis. Podem ser representados por `NaN`, `NA`, `None`, 0, espaços em branco, ou valores específicos como `-99`.

- **Impacto:** Podem causar erros em cálculos, viesar análises ou reduzir o poder estatístico.
- **Identificação:** Verificação de `null` ou valores sentinela.

Exemplo: Identificando Valores Ausentes

Python

```
\ImportTok{import}\NormalTok{ pandas }\ImportTok{as}\NormalTok{ pd}
\ImportTok{import}\NormalTok{ numpy }\ImportTok{as}\NormalTok{ np}

\NormalTok{data }\OperatorTok{=}\NormalTok{
 ↪  \{}\StringTok{\textquotesingle{}ID\textquotesingle{}}\NormalTok{:,}
        \StringTok{\textquotesingle{}Nome\textquotesingle{}}\NormalTok{:
         ↪  []\StringTok{\textquotesingle{}Alice\textquotesingle{}}\NormalTok{,
         ↪  }\StringTok{\textquotesingle{}Bob\textquotesingle{}}\NormalTok{,
         ↪  }\StringTok{\textquotesingle{}Charlie\textquotesingle{}}\NormalTok{,
         ↪  }\StringTok{\textquotesingle{}David\textquotesingle{}}\NormalTok{,
         ↪  }\StringTok{\textquotesingle{}Eve\textquotesingle{}}\NormalTok{],}
        \StringTok{\textquotesingle{}Idade\textquotesingle{}}\NormalTok{:
         ↪  []\DecValTok{25}\NormalTok{, np.nan, }\DecValTok{30}\NormalTok{,
         ↪  }\DecValTok{28}\NormalTok{, np.nan],}
        \StringTok{\textquotesingle{}Renda\textquotesingle{}}\NormalTok{:
         ↪  []\DecValTok{50000}\NormalTok{, }\DecValTok{60000}\NormalTok{, np.nan,
         ↪  }\DecValTok{70000}\NormalTok{, }\DecValTok{55000}\NormalTok{]\}}
\NormalTok{df }\OperatorTok{=}\NormalTok{ pd.DataFrame(data)}

\BuiltInTok{print}\NormalTok{(}\StringTok{"DataFrame original com valores
 ↪  ausentes:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df)}

\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Verifi
 ↪  cação de valores ausentes (True para ausente):"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df.isnull())}

\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Contag
 ↪  em de valores ausentes por coluna:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df.isnull().}\BuiltInTok{sum}\NormalTok{())}

\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Propor
 ↪  ção de valores ausentes por coluna:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df.isnull().}\BuiltInTok{sum}\NormalTok{() }\OperatorTok{/}
 ↪  \BuiltInTok{len}\NormalTok{(df) }\OperatorTok{*} \DecValTok{100}\NormalTok{)}
```

R

```
\NormalTok{df }\OtherTok{\textless{}-} \FunctionTok{data.frame}\NormalTok{(}
```

```
  \AttributeTok{ID =} \FunctionTok{c}\NormalTok{(}\DecValTok{1}\NormalTok{,
  ↪  }\DecValTok{2}\NormalTok{, }\DecValTok{3}\NormalTok{, }\DecValTok{4}\NormalTok{,
  ↪  }\DecValTok{5}\NormalTok{),}
  \AttributeTok{Nome =} \FunctionTok{c}\NormalTok{(}\StringTok{\textquotesingle{}Alice\t
  ↪  extquotesingle{}}\NormalTok{,
  ↪  }\StringTok{\textquotesingle{}Bob\textquotesingle{}}\NormalTok{,
  ↪  }\StringTok{\textquotesingle{}Charlie\textquotesingle{}}\NormalTok{,
  ↪  }\StringTok{\textquotesingle{}David\textquotesingle{}}\NormalTok{,
  ↪  }\StringTok{\textquotesingle{}Eve\textquotesingle{}}\NormalTok{),}
  \AttributeTok{Idade =} \FunctionTok{c}\NormalTok{(}\DecValTok{25}\NormalTok{,
  ↪  }\ConstantTok{NA}\NormalTok{, }\DecValTok{30}\NormalTok{,
  ↪  }\DecValTok{28}\NormalTok{, }\ConstantTok{NA}\NormalTok{),}
  \AttributeTok{Renda =} \FunctionTok{c}\NormalTok{(}\DecValTok{50000}\NormalTok{,
  ↪  }\DecValTok{60000}\NormalTok{, }\ConstantTok{NA}\NormalTok{,
  ↪  }\DecValTok{70000}\NormalTok{, }\DecValTok{55000}\NormalTok{)}
\NormalTok{)}

\FunctionTok{cat}\NormalTok{(}\StringTok{"DataFrame original com valores
 ↪  ausentes:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(df)}

\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Verificação de valores ausentes (TRUE para
 ↪  ausente):}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(}\FunctionTok{is.na}\NormalTok{(df))}

\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Contagem de valores ausentes por
 ↪  coluna:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(}\FunctionTok{colSums}\NormalTok{(}\FunctionTok{is.na}\No
 ↪  rmalTok{(df)))}

\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Proporção de valores ausentes por
 ↪  coluna:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(}\FunctionTok{colMeans}\NormalTok{(}\FunctionTok{is.na}\N
 ↪  ormalTok{(df)) }\SpecialCharTok{*} \DecValTok{100}\NormalTok{)}
```

Outliers (Valores Atípicos)

Observações que se afastam significativamente dos demais dados. Podem ser erros de registro ou eventos genuinamente raros.

- **Impacto:** Distorcem médias, desvio padrão, variância e podem influenciar negativamente modelos

estatísticos.

- **Identificação:** Box plots, histogramas, análise de desvio padrão (valores a mais de 2 ou 3 desvios padrão da média), métodos estatísticos (Z-score, IQR).

Exemplo: Identificando Outliers (Visualmente com Box Plot)

Python

```
\ImportTok{import}\NormalTok{ pandas }\ImportTok{as}\NormalTok{ pd}
\ImportTok{import}\NormalTok{ matplotlib.pyplot }\ImportTok{as}\NormalTok{ plt}
\ImportTok{import}\NormalTok{ seaborn }\ImportTok{as}\NormalTok{ sns}
\ImportTok{import}\NormalTok{ numpy }\ImportTok{as}\NormalTok{ np}

\CommentTok{\# Criando um dataset com outliers}
\NormalTok{data\_outliers }\OperatorTok{=}\NormalTok{
 ↪  pd.DataFrame(\{}\StringTok{\textquotesingle{}Valor\textquotesingle{}}\NormalTok{:\})}

\NormalTok{plt.figure(figsize}\OperatorTok{=}\NormalTok{(}\DecValTok{8}\NormalTok{,
 ↪  }\DecValTok{5}\NormalTok{))}
\NormalTok{sns.boxplot(y}\OperatorTok{=}\NormalTok{data\_outliers[}\StringTok{\textquote
 ↪  single{}Valor\textquotesingle{}}\NormalTok{])}
\NormalTok{plt.title(}\StringTok{\textquotesingle{}Identificação de Outliers com Box
 ↪  Plot\textquotesingle{}}\NormalTok{)}
\NormalTok{plt.ylabel(}\StringTok{\textquotesingle{}Valor\textquotesingle{}}\NormalTok{)}
\NormalTok{plt.grid(axis}\OperatorTok{=}\StringTok{\textquotesingle{}y\textquotesingle{}
 ↪  }\NormalTok{,
 ↪  linestyle}\OperatorTok{=}\StringTok{\textquotesingle{}{-}{-}\textquotesingle{}}\Norm
 ↪  alTok{, alpha}\OperatorTok{=}\FloatTok{0.7}\NormalTok{)}
\NormalTok{plt.show()}

\CommentTok{\# Identificação numérica (ex: usando Z{-}score)}
\NormalTok{mean\_val }\OperatorTok{=}\NormalTok{ data\_outliers[}\StringTok{\textquotesi
 ↪  ngle{}Valor\textquotesingle{}}\NormalTok{].mean()}
\NormalTok{std\_val }\OperatorTok{=}\NormalTok{ data\_outliers[}\StringTok{\textquotesin
 ↪  gle{}Valor\textquotesingle{}}\NormalTok{].std()}
\NormalTok{data\_outliers[}\StringTok{\textquotesingle{}Z\_Score\textquotesingle{}}\Norm
 ↪  alTok{] }\OperatorTok{=}\NormalTok{
 ↪  (data\_outliers[}\StringTok{\textquotesingle{}Valor\textquotesingle{}}\NormalTok{]
 ↪  }\OperatorTok{{-}}\NormalTok{ mean\_val) }\OperatorTok{/}\NormalTok{ std\_val}
\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{DataFr
 ↪  ame com Z{-}scores:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(data\_outliers)}
\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Outlie
 ↪  rs (Z{-}score \textgreater{} 2 ou \textless{} {-}2):"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(data\_outliers[(data\_outliers[}\StringTok{\textquotesingl
 ↪  e{}Z\_Score\textquotesingle{}}\NormalTok{].}\BuiltInTok{abs}\NormalTok{()
 ↪  }\OperatorTok{\textgreater{}} \DecValTok{2}\NormalTok{)])}
```

R

```
\FunctionTok{library}\NormalTok{(ggplot2)}

\CommentTok{\# Criando um dataset com outliers}
\NormalTok{df\_outliers }\OtherTok{\textless{}{-}}
 ↪  \FunctionTok{data.frame}\NormalTok{(}\AttributeTok{Valor =}
 ↪  \FunctionTok{c}\NormalTok{(}\DecValTok{10}\NormalTok{, }\DecValTok{12}\NormalTok{,
 ↪  }\DecValTok{11}\NormalTok{, }\DecValTok{13}\NormalTok{, }\DecValTok{10}\NormalTok{,
 ↪  }\DecValTok{12}\NormalTok{, }\DecValTok{100}\NormalTok{, }\DecValTok{9}\NormalTok{,
 ↪  }\DecValTok{11}\NormalTok{, }\DecValTok{13}\NormalTok{))}

\FunctionTok{ggplot}\NormalTok{(df\_outliers,
 ↪  }\FunctionTok{aes}\NormalTok{(}\AttributeTok{y =}\NormalTok{ Valor))
 ↪  }\SpecialCharTok{+}
  \FunctionTok{geom\_boxplot}\NormalTok{(}\AttributeTok{fill =}
   ↪  \StringTok{"lightblue"}\NormalTok{) }\SpecialCharTok{+}
  \FunctionTok{labs}\NormalTok{(}\AttributeTok{title =}
   ↪  \StringTok{\textquotesingle{}Identificação de Outliers com Box
   ↪  Plot\textquotesingle{}}\NormalTok{, }\AttributeTok{y =}
   ↪  \StringTok{\textquotesingle{}Valor\textquotesingle{}}\NormalTok{)
   ↪  }\SpecialCharTok{+}
  \FunctionTok{theme\_minimal}\NormalTok{()}

\CommentTok{\# Identificação numérica (ex: usando Z{-}score)}
\NormalTok{mean\_val }\OtherTok{\textless{}{-}}
 ↪  \FunctionTok{mean}\NormalTok{(df\_outliers}\SpecialCharTok{$}\NormalTok{Valor)}
\NormalTok{std\_val }\OtherTok{\textless{}{-}}
 ↪  \FunctionTok{sd}\NormalTok{(df\_outliers}\SpecialCharTok{$}\NormalTok{Valor)}
\NormalTok{df\_outliers}\SpecialCharTok{$}\NormalTok{Z\_Score
 ↪  }\OtherTok{\textless{}{-}}\NormalTok{
 ↪  (df\_outliers}\SpecialCharTok{$}\NormalTok{Valor }\SpecialCharTok{{-}}\NormalTok{
 ↪  mean\_val) }\SpecialCharTok{/}\NormalTok{ std\_val}
\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  DataFrame com
 ↪  Z{-}scores:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(df\_outliers)}
\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Outliers (Z{-}score \textgreater{} 2 ou \textless{}
 ↪  {-}2):}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(df\_outliers[}\FunctionTok{abs}\NormalTok{(df\_outliers}\
 ↪  SpecialCharTok{$}\NormalTok{Z\_Score) }\SpecialCharTok{\textgreater{}}
 ↪  \DecValTok{2}\NormalTok{, ])}
```

Inconsistências e Formatos Incorretos

Erros de digitação, uso de diferentes formatos para a mesma informação, unidades de medida misturadas, etc.

- **Impacto:** Dificultam a análise, agregação e comparação de dados.
- **Identificação:** Análise de valores únicos (`value_counts()` em Pandas, `table()` em R), expressões regulares, limites de validação.

Exemplo: Identificando Inconsistências

Python

```
\ImportTok{import}\NormalTok{ pandas }\ImportTok{as}\NormalTok{ pd}

\NormalTok{data\_inconsistente }\OperatorTok{=}\NormalTok{
 ↪  \{}\StringTok{\textquotesingle{}País\textquotesingle{}}\NormalTok{:
 ↪  []\StringTok{\textquotesingle{}Brasil\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}Brazil\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}ARGENTINA\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}Chile\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}brasil \textquotesingle{}}\NormalTok{],}
                      \StringTok{\textquotesingle{}Capital\textquotesingle{}}\NormalTok{:
                       ↪  []\StringTok{\textquotesingle{}Brasília\textquotesingle{}}\Nor
                       ↪  malTok{,
                       ↪  }\StringTok{\textquotesingle{}Brasilia\textquotesingle{}}\Norm
                       ↪  alTok{, }\StringTok{\textquotesingle{}Buenos
                       ↪  Aires\textquotesingle{}}\NormalTok{, }\StringTok{\textquotesin
                       ↪  gle{}Santiago\textquotesingle{}}\NormalTok{,
                       ↪  }\StringTok{\textquotesingle{}Brasilia\textquotesingle{}}\Norm
                       ↪  alTok{]\}}
\NormalTok{df\_inc }\OperatorTok{=}\NormalTok{ pd.DataFrame(data\_inconsistente)}

\BuiltInTok{print}\NormalTok{(}\StringTok{"DataFrame com inconsistências:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df\_inc)}

\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Valores
 ↪  únicos na coluna \textquotesingle{}País\textquotesingle{} antes da
 ↪  limpeza:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df\_inc[}\StringTok{\textquotesingle{}País\textquotesingle
 ↪  {}}\NormalTok{].value\_counts())}
```

```
\CommentTok{\# Exemplo de limpeza simples}
\NormalTok{df\_inc[}\StringTok{\textquotesingle{}País\_Limpo\textquotesingle{}}\NormalTo
 ↪  k[] }\OperatorTok{=}\NormalTok{
 ↪  df\_inc[}\StringTok{\textquotesingle{}País\textquotesingle{}}\NormalTok{].}\BuiltInT
 ↪  ok{str}\NormalTok{.strip().}\BuiltInTok{str}\NormalTok{.capitalize()}
\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Valores
 ↪  únicos na coluna \textquotesingle{}País\_Limpo\textquotesingle{} após limpeza
 ↪  simples:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df\_inc[}\StringTok{\textquotesingle{}País\_Limpo\textquot
 ↪  esingle{}}\NormalTok{].value\_counts())}
```

R

```
\NormalTok{df\_inc }\OtherTok{\textless{}{-}} \FunctionTok{data.frame}\NormalTok{(}
\NormalTok{  País }\OtherTok{=} \FunctionTok{c}\NormalTok{(}\StringTok{\textquotesingle{
 ↪  }Brasil\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}Brazil\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}ARGENTINA\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}Chile\textquotesingle{}}\NormalTok{,
 ↪  }\StringTok{\textquotesingle{}brasil \textquotesingle{}}\NormalTok{),}
  \AttributeTok{Capital =} \FunctionTok{c}\NormalTok{(}\StringTok{\textquotesingle{}Bras
   ↪  ília\textquotesingle{}}\NormalTok{,
   ↪  }\StringTok{\textquotesingle{}Brasilia\textquotesingle{}}\NormalTok{,
   ↪  }\StringTok{\textquotesingle{}Buenos Aires\textquotesingle{}}\NormalTok{,
   ↪  }\StringTok{\textquotesingle{}Santiago\textquotesingle{}}\NormalTok{,
   ↪  }\StringTok{\textquotesingle{}Brasilia\textquotesingle{}}\NormalTok{)}
\NormalTok{)}

\FunctionTok{cat}\NormalTok{(}\StringTok{"DataFrame com
 ↪  inconsistências:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(df\_inc)}

\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Valores únicos na coluna \textquotesingle{}País\textquotesingle{} antes da
 ↪  limpeza:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(}\FunctionTok{table}\NormalTok{(df\_inc}\SpecialCharTok{$
 ↪  }\NormalTok{País))}

\CommentTok{\# Exemplo de limpeza simples}
\NormalTok{df\_inc}\SpecialCharTok{$}\NormalTok{País\_Limpo }\OtherTok{\textless{}{-}}
 ↪  \FunctionTok{tolower}\NormalTok{(}\FunctionTok{trimws}\NormalTok{(df\_inc}\SpecialCh
 ↪  arTok{$}\NormalTok{País))}
\NormalTok{df\_inc}\SpecialCharTok{$}\NormalTok{País\_Limpo
 ↪  }\OtherTok{\textless{}{-}}\NormalTok{ Hmisc}\SpecialCharTok{::}\FunctionTok{capitali
 ↪  ze}\NormalTok{(df\_inc}\SpecialCharTok{$}\NormalTok{País\_Limpo) }\CommentTok{\#
 ↪  Necessita do pacote Hmisc}
```

```
\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Valores únicos na coluna \textquotesingle{}País\_Limpo\textquotesingle{} após limpeza
 ↪  simples:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(}\FunctionTok{table}\NormalTok{(df\_inc}\SpecialCharTok{$
 ↪  }\NormalTok{País\_Limpo))}
```

Duplicatas

Registros idênticos ou muito semelhantes que representam a mesma entidade.

- **Impacto:** Inflam o tamanho do dataset, viesam contagens e podem distorcer análises estatísticas.
- **Identificação:** `duplicated()` em Pandas/R, identificação por um subconjunto de colunas.

Exemplo: Identificando Duplicatas

Python

```
\ImportTok{import}\NormalTok{ pandas }\ImportTok{as}\NormalTok{ pd}

\NormalTok{data\_duplicada }\OperatorTok{=}\NormalTok{
 ↪  \{}\StringTok{\textquotesingle{}ID\textquotesingle{}}\NormalTok{:,}
                  \StringTok{\textquotesingle{}Nome\textquotesingle{}}\NormalTok{:
                   ↪  []\StringTok{\textquotesingle{}Alice\textquotesingle{}}\NormalTok{,
                   ↪  }\StringTok{\textquotesingle{}Bob\textquotesingle{}}\NormalTok{, }
                   ↪  \StringTok{\textquotesingle{}Charlie\textquotesingle{}}\NormalTok{,
                   ↪  }\StringTok{\textquotesingle{}Alice\textquotesingle{}}\NormalTok{,
                   ↪  }\StringTok{\textquotesingle{}David\textquotesingle{}}\NormalTok{]
                   ↪  ,}
                  \StringTok{\textquotesingle{}Idade\textquotesingle{}}\NormalTok{:\}}
\NormalTok{df\_dup }\OperatorTok{=}\NormalTok{ pd.DataFrame(data\_duplicada)}

\BuiltInTok{print}\NormalTok{(}\StringTok{"DataFrame com duplicatas:"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df\_dup)}

\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Verifi
 ↪  cação de linhas duplicadas (True se a linha é uma duplicata
 ↪  posterior):"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df\_dup.duplicated())}

\BuiltInTok{print}\NormalTok{(}\StringTok{"}\CharTok{\textbackslash{}n}\StringTok{Regist
 ↪  ros duplicados (incluindo o primeiro):"}\NormalTok{)}
\BuiltInTok{print}\NormalTok{(df\_dup[df\_dup.duplicated(keep}\OperatorTok{=}\VariableTo
 ↪  k{False}\NormalTok{)]) }\CommentTok{\# keep=False mostra todas as ocorrências de
 ↪  duplicatas}
```

R

```
\NormalTok{df\_dup }\OtherTok{\textless{}{-}} \FunctionTok{data.frame}\NormalTok{(}
  \AttributeTok{ID =} \FunctionTok{c}\NormalTok{(}\DecValTok{1}\NormalTok{,
    ↪  }\DecValTok{2}\NormalTok{, }\DecValTok{3}\NormalTok{, }\DecValTok{1}\NormalTok{,
    ↪  }\DecValTok{4}\NormalTok{),}
  \AttributeTok{Nome =} \FunctionTok{c}\NormalTok{(}\StringTok{\textquotesingle{}Alice\t
    ↪  extquotesingle{}}\NormalTok{,
    ↪  }\StringTok{\textquotesingle{}Bob\textquotesingle{}}\NormalTok{,
    ↪  }\StringTok{\textquotesingle{}Charlie\textquotesingle{}}\NormalTok{,
    ↪  }\StringTok{\textquotesingle{}Alice\textquotesingle{}}\NormalTok{,
    ↪  }\StringTok{\textquotesingle{}David\textquotesingle{}}\NormalTok{),}
  \AttributeTok{Idade =} \FunctionTok{c}\NormalTok{(}\DecValTok{25}\NormalTok{,
    ↪  }\DecValTok{30}\NormalTok{, }\DecValTok{35}\NormalTok{, }\DecValTok{25}\NormalTok{,
    ↪  }\DecValTok{28}\NormalTok{)}
\NormalTok{)}

\FunctionTok{cat}\NormalTok{(}\StringTok{"DataFrame com
 ↪  duplicatas:}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(df\_dup)}

\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Verificação de linhas duplicadas (TRUE se a linha é uma duplicata
 ↪  posterior):}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(}\FunctionTok{duplicated}\NormalTok{(df\_dup))}

\FunctionTok{cat}\NormalTok{(}\StringTok{"}\SpecialCharTok{\textbackslash{}n}\StringTok{
 ↪  Registros duplicados (incluindo o
 ↪  primeiro):}\SpecialCharTok{\textbackslash{}n}\StringTok{"}\NormalTok{)}
\FunctionTok{print}\NormalTok{(df\_dup[}\FunctionTok{duplicated}\NormalTok{(df\_dup)
 ↪  }\SpecialCharTok{|} \FunctionTok{duplicated}\NormalTok{(df\_dup,
 ↪  }\AttributeTok{fromLast =} \ConstantTok{TRUE}\NormalTok{), ])}
```

Dados Desatualizados

Informações que não são mais relevantes ou precisas devido à passagem do tempo.

- **Impacto:** Leva a decisões baseadas em informações incorretas sobre o estado atual.
- **Identificação:** Campos de data/hora, comparação com fontes externas, regras de negócio para validade temporal.

Vieses na Coleta

Introdução de erro sistemático na amostra ou no processo de coleta, fazendo com que a amostra não seja representativa da população.

- **Impacto:** Distorce completamente a capacidade de generalizar os resultados.
- **Identificação:** Requer um entendimento profundo do processo de amostragem (revisitar a aula de Amostragem) e do contexto do negócio.

Relação com Outros Conceitos

A qualidade dos dados é a base para a confiabilidade de todas as análises subsequentes.

- **Estatística Descritiva e AED:** Medidas de tendência central e dispersão, assim como visualizações, podem ser severamente distorcidas por dados de baixa qualidade (outliers, valores ausentes).
- **Amostragem:** Vieses na coleta comprometem a representatividade da amostra, invalidando qualquer inferência.
- **Estimação e Testes de Hipóteses:** Intervalos de confiança e p-valores dependem da suposição de dados válidos e aleatoriamente coletados. Dados ruins levam a estimativas imprecisas e conclusões de testes errôneas.
- **Modelagem (ML/IA):** Modelos aprendem com os dados. Se os dados são "lixo", o modelo também será "lixo" (GIGO - Garbage In, Garbage Out).

Verificação de Aprendizagem

Responda às perguntas e realize a tarefa prática para verificar sua compreensão sobre coleta e qualidade dos dados.

1. **Conceitos de Qualidade:**

   a) Você está analisando um dataset de vendas e percebe que algumas datas de transação estão no futuro (ex: 2026). Qual dimensão da qualidade dos dados está sendo comprometida?
   b) Ao coletar dados de clientes, você observa que o campo "email" está vazio para 40% dos registros. Qual dimensão da qualidade dos dados é mais afetada aqui?
   c) Explique brevemente por que a unicidade dos dados é importante.

2. **Métodos de Coleta:**

   Qual método de coleta de dados seria mais apropriado para cada cenário e por quê?

   a) Estimar o tempo médio que os carros levam para passar por um semáforo durante o horário de pico.

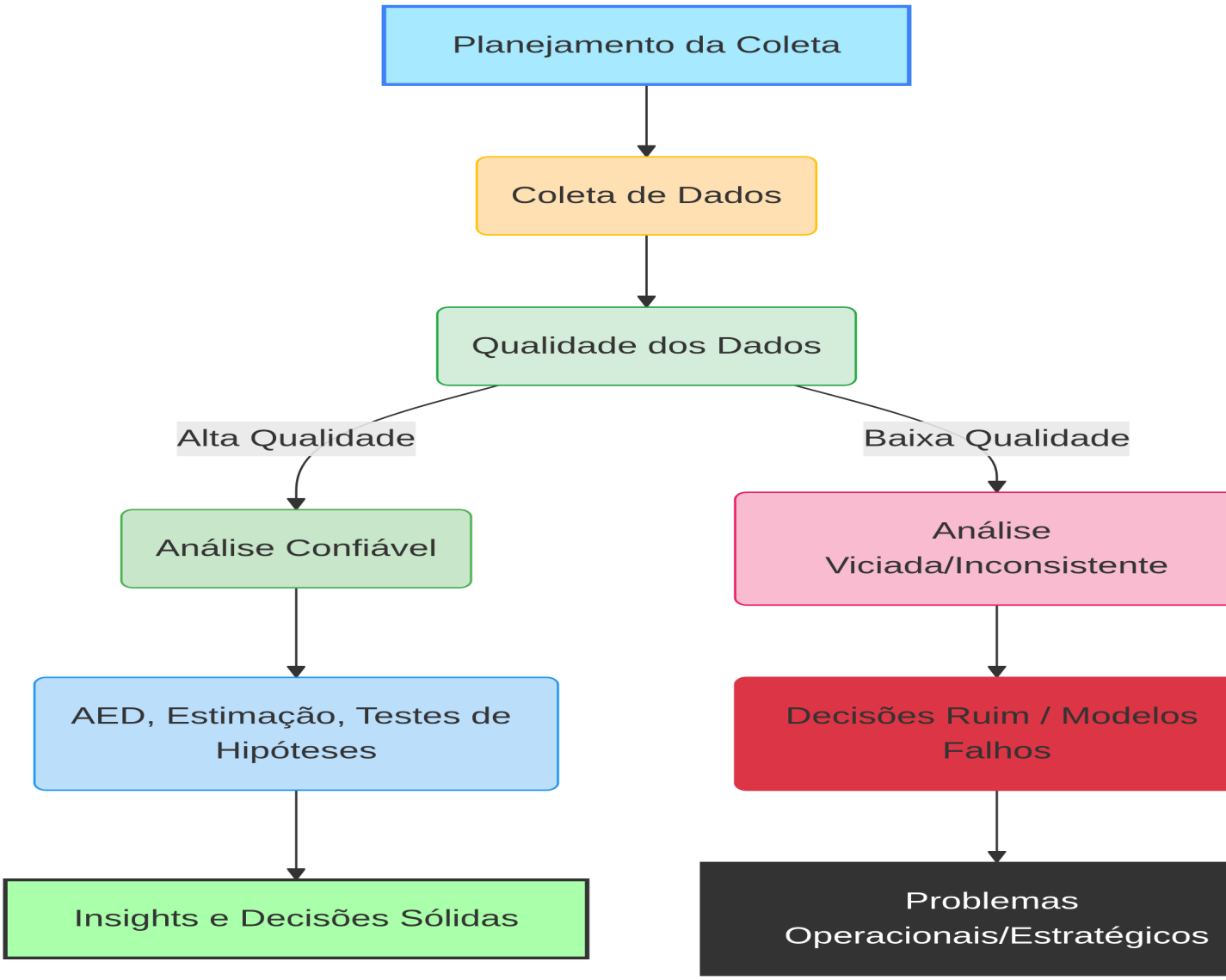

Figure 2: Relação da Qualidade dos Dados com Outros Conceitos Estatísticos

b) Entender as percepções profundas e experiências de pacientes com uma nova terapia.
c) Obter dados históricos de preços de ações de uma empresa nos últimos 10 anos.

3. **Problema Prático (Identificação de Problemas de Qualidade):**

   Considere o seguinte conjunto de dados (simulado) de informações de clientes. Sua tarefa é carregá-lo em Python ou R e identificar os problemas de qualidade.

```
ID,Nome,Idade,Renda,Cidade,Status_Cliente
1,Ana Silva,30,5000,São Paulo,Ativo
2,Bruno Costa,NA,6000,Rio de Janeiro,Inativo
3,Carla Dias,25,4500,São Paulo,Ativo
4,Ana Silva,30,5000,São Paulo,Ativo
5,Daniel Alves,40,7000,BH,Ativo
6,Eduarda F.,22,NaN,Curitiba,Ativo
7,Fábio G.,999,8000,São Paulo,Ativo
8,Helena I.,35,5500,são paulo,Ativo
9,IgOR J.,28,6200,RIO DE JANEIRO,Ativo
10,Julia K.,50,4800,São Paulo,Ativo
11,Luís L.,-5,7500,Porto Alegre,Ativo
12,Mariana N.,45,120000,Recife,Ativo
```

   a) Carregue os dados em um DataFrame/data.frame.
   b) Identifique e conte os **valores ausentes**.
   c) Identifique e conte os **registros duplicados** (considerando todas as colunas).
   d) Identifique possíveis **outliers** ou valores inválidos nas colunas `Idade` e `Renda`.
   e) Identifique **inconsistências de capitalização/formato** nas colunas `Cidade` e `Nome`.

Referências Bibliográficas

BUSSAB, Luiz O. de M.; MORETTIN, Pedro A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.