# Onde a Ciência de Dados Começa?

**"Onde a ciência de dados começa?"**

Antes de qualquer análise ou tomada de decisão baseada em dados, precisamos ter **dados de alta qualidade** . A qualidade dos dados é o alicerce para todo o processo de ciência de dados.

> **Lembre-se:** Dados de baixa qualidade levam a análises imprecisas e decisões equivocadas, independentemente da sofisticação dos algoritmos utilizados.


## Conceitos Fundamentais

### Dados Primários

Coletados diretamente para um objetivo específico

Ex: Pesquisas, entrevistas, experimentos

### Dados Secundários

Já existentes, coletados por terceiros

Ex: Bases governamentais, relatórios, redes sociais

# Fontes Primárias vs. Secundárias

## Fontes Primárias

Dados coletados diretamente para um objetivo específico.

**Exemplos:** Pesquisas, experimentos, entrevistas, observações diretas.

**Vantagens**

- Controle total sobre a coleta
- Dados específicos para o objetivo
- Maior confiabilidade

**Desvantagens**

- Processo caro e demorado
- Requer expertise em coleta
- Amostra geralmente menor

## Fontes Secundárias

Dados já existentes, coletados por terceiros para outros propósitos.

**Exemplos:** Bases governamentais (IBGE, DataSUS), relatórios, redes sociais.

**Vantagens**

- Acesso rápido e econômico
- Grande volume de dados
- Séries históricas disponíveis

**Desvantagens**

- Qualidade e relevância limitadas
- Falta de controle sobre a coleta
- Possível desatualização

# Fontes Internas vs. Externas

## Fontes Internas

Dados gerados dentro da própria organização durante suas operações regulares.

**Características:** Maior controle, acesso facilitado, específicos para o negócio.

**Exemplos de Fontes Internas**

- Registros de vendas e transações
- Dados de RH (funcionários, salários)
- Registros de produção e estoque
- Dados de CRM e atendimento ao cliente
- Logs de sistemas e aplicações
- Pesquisas internas com funcionários

*Vantagem principal: Dados específicos para o contexto da organização, com acesso direto e controle sobre a qualidade.*

## Fontes Externas

Dados provenientes de fora da organização, coletados e disponibilizados por terceiros.

**Características:** Maior volume, perspectiva externa, complementam dados internos.

**Exemplos de Fontes Externas**

- Dados governamentais (dados.gov.br, IBGE)
- Dados de saúde pública (DataSUS)
- Dados econômicos (Banco Central)
- Redes sociais e mídia online
- Pesquisas de mercado de terceiros
- Dados de parceiros comerciais

*Vantagem principal: Fornecem contexto mais amplo e permitem comparações com o mercado e tendências externas.*

# Atividade Prática 1: Mapeamento e Seleção de Fontes

**Objetivo:** Desenvolver a capacidade de identificar, avaliar e selecionar fontes de dados adequadas para um problema específico.

## Instruções

### 1 Escolha do Tema

Cada escolherá um dos seguintes temas:

| | |
|---|---|
| **Hábitos de Consumo** | **Mobilidade Urbana** |
| **Saúde Pública** | **Educação Digital** |

### 2 Brainstorm de Fontes

Cada um deve listar pelo menos:

- 2 fontes de dados primárias que poderiam usar para o tema
- 2 fontes de dados secundárias que poderiam usar para o tema

*Exemplo para "Saúde Pública": Uma fonte primária seria uma pesquisa com a comunidade local e uma secundária seria o DataSUS.*

### 3 Justificativa

Para cada fonte listada, o grupo deve justificar:

- Relevância para o tema (por que esta fonte é útil?)
- Confiabilidade da fonte (como avaliar sua qualidade?)
- Possíveis limitações ou vieses da fonte
- Fontes selecionadas e justificativas
- Critérios utilizados para avaliar a qualidade das fontes
- Desafios encontrados durante o processo de seleção

# Atividade Prática 2: Simulação de Coleta Experimental

**Objetivo:** Realizar uma coleta de dados primários através de um experimento simples utilizando o Google Forms.

## Passo a Passo

### 1 Criação do Experimento

Cada grupo deve criar um experimento simples que possa ser realizado em sala de aula.

**Exemplo:**

"Qual o tempo médio que os alunos levam para responder a uma pergunta de múltipla escolha no Google Forms?"

### 2 Criação do Formulário

Utilize o Google Forms para criar um formulário que colete os dados necessários:

- Acesse forms.google.com
- Crie um novo formulário
- Adicione perguntas relevantes para o experimento
- Configure as opções de coleta de respostas

### 3 Aplicação e Coleta

Os grupos aplicam o formulário entre si e coletam os dados:

- Compartilhe o link do formulário com os outros grupos
- Cada aluno deve responder aos formulários dos outros grupos
- Acompanhe as respostas em tempo real na aba "Respostas" do Google Forms

**Dicas para um bom formulário:**

- Mantenha o formulário curto e objetivo
- Use perguntas claras e diretas
- Evite perguntas tendenciosas
- Teste o formulário antes de aplicar
- Inclua instruções claras no início

# Atividade Prática 3: Limpeza e Tratamento dos Dados

**Objetivo:** Trabalhar com os dados coletados na atividade anterior, identificando e corrigindo erros comuns.

**Fixação**

Após a limpeza, discuta com a turma:

- Quais erros foram mais comuns?
- Como esses erros poderiam afetar análises futuras?
- Quais estratégias poderiam prevenir esses erros na coleta?

## Passo a Passo (Google Sheets/Excel)

### 1 Transferência dos Dados

Exporte os dados do Google Forms para uma planilha

Verifique se todos os campos foram transferidos corretamente

### 2 Identificação de Erros

**Valores Nulos**

Campos deixados em branco pelos respondentes

*Dica: Use filtros para identificar células vazias*

**Duplicatas**

Respostas enviadas mais de uma vez

*Dica: Use "Remover duplicatas" no menu Dados*

**Inconsistências**

Formatos diferentes para o mesmo tipo de resposta

*Ex: "sim", "s", "SIM" para a mesma pergunta*

**Valores Atípicos**

Respostas que fogem muito do padrão esperado

*Ex: Tempo de resposta de 300 segundos quando a média é 30*

### 3 Limpeza dos Dados

Corrija os erros identificados manualmente ou usando fórmulas

Padronize formatos (datas, números, textos)

Documente todas as alterações feitas

# Comandos Python/Pandas para Limpeza de Dados

O Python, com a biblioteca Pandas, oferece ferramentas poderosas para automatizar a limpeza e o tratamento de dados. Abaixo estão alguns comandos essenciais:

## 1. Identificação de Valores Nulos

```
# Verificar valores nulos em todo o DataFrame
df.isnull().sum()
```

```
nome 0
idade 2
cidade 1
dtype: int64
```

## 2. Remoção de Linhas com Valores Nulos

```
# Remover linhas que contêm valores nulos
df_limpo = df.dropna()

# Remover linhas com valores nulos apenas em colunas específicas
df_limpo = df.dropna(subset=['idade', 'cidade'])
```

## 3. Preenchimento de Valores Nulos

```
# Preencher valores nulos com um valor específico
df['idade'] = df['idade'].fillna(0)

# Preencher com a média da coluna
df['idade'] = df['idade'].fillna(df['idade'].mean())
```

## 4. Remoção de Duplicatas

```
# Identificar duplicatas
df.duplicated().sum()

# Remover linhas duplicadas
df_sem_duplicatas = df.drop_duplicates()
```

## 5. Tratamento de Inconsistências

```
# Padronizar texto (converter para minúsculas)
df['cidade'] = df['cidade'].str.lower()

# Remover espaços extras
df['nome'] = df['nome'].str.strip()
```

# Atividade Prática 4: Discussão de Casos Éticos e LGPD

**Objetivo:** Desenvolver o pensamento crítico sobre questões éticas relacionadas à coleta e uso de dados, considerando a Lei Geral de Proteção de Dados (LGPD).

## Caso A: Marketing Digital e Geolocalização

Uma empresa de marketing digital utiliza dados de localização de clientes para enviar publicidade direcionada sem consentimento explícito. Os usuários recebem notificações de ofertas ao passar próximo a lojas parceiras, sem terem sido informados sobre esse tipo de rastreamento.

## Caso B: Vazamento em Sistema de Saúde

Um sistema de saúde utiliza dados anônimos de pacientes para pesquisa médica. No entanto, uma falha de segurança permite que terceiros consigam cruzar informações e identificar pacientes específicos, expondo condições médicas sensíveis.

## Perguntas para Discussão

**É permitido coletar esse dado?**

Analise se há base legal para a coleta segundo a LGPD.

**Qual a base legal (LGPD) para essa coleta?**

Identifique qual das bases legais previstas na LGPD poderia justificar (ou não) a coleta.

**Como garantir a privacidade e o anonimato?**

Proponha medidas técnicas e organizacionais para proteger os dados.

# LGPD e Proteção de Dados

**Objetivo principal:** Proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural.

## Bases Legais para Coleta de Dados

A LGPD estabelece que o tratamento de dados pessoais só pode ocorrer nas seguintes hipóteses:

**Consentimento**
Manifestação livre, informada e inequívoca do titular

**Obrigação Legal**
Cumprimento de obrigação legal ou regulatória

**Execução de Contrato**
Necessário para execução de contrato com o titular

**Interesse Legítimo**
Atender aos interesses legítimos do controlador

**Proteção da Vida**
Proteção da vida ou da incolumidade física

**Tutela da Saúde**
Tutela da saúde, em procedimento realizado por profissionais

## Princípios da LGPD

- Finalidade
- Adequação
- Necessidade
- Livre acesso
- Qualidade dos dados
- Transparência
- Segurança
- Prevenção
- Não discriminação
- Responsabilização


**Implicações para Coleta de Dados**

- Obter consentimento explícito quando necessário
- Informar claramente a finalidade da coleta
- Coletar apenas dados necessários (minimização)
- Garantir a segurança dos dados coletados
- Permitir acesso, correção e exclusão pelos titulares
- Documentar todo o processo de tratamento

# Avaliação da Qualidade dos Dados

A qualidade dos dados é fundamental para garantir análises confiáveis e decisões acertadas. Dados de baixa qualidade podem levar a conclusões equivocadas, independentemente da sofisticação das técnicas de análise utilizadas.

> *"Garbage in, garbage out"* - Princípio fundamental da ciência de dados que ressalta a importância da qualidade dos dados de entrada.

## Critérios de Qualidade dos Dados

### Precisão

Os dados representam corretamente a realidade que pretendem descrever? Estão livres de erros?

### Completude

Os dados contêm todas as informações necessárias? Existem valores ausentes ou campos incompletos?

### Consistência

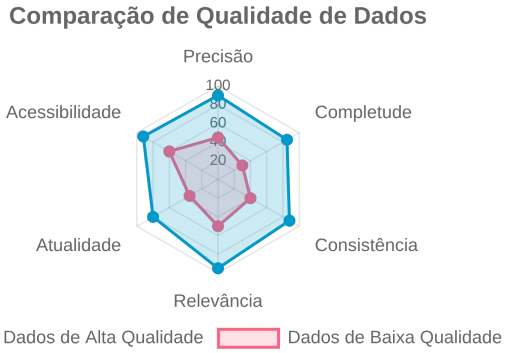
Os dados são coerentes entre si? Existem contradições ou informações conflitantes?

### Relevância

Os dados são úteis para o problema que se pretende resolver? Atendem às necessidades da análise?

## Impacto da Qualidade dos Dados

- Decisões de negócio mais acertadas
- Redução de custos operacionais
- Maior confiança nos resultados analíticos
- Melhor experiência do cliente
- Conformidade com regulamentações



# Encerramento e Próximos Passos

## Principais Aprendizados

A coleta e limpeza de dados são etapas fundamentais para análises confiáveis

A escolha das fontes de dados deve ser baseada no problema a ser resolvido

Dados primários oferecem controle, mas dados secundários são mais acessíveis

A qualidade dos dados impacta diretamente a qualidade das análises

A ética e a LGPD devem ser consideradas em todo o processo de coleta

**A coleta e a limpeza de dados não são apenas etapas técnicas, mas processos que exigem responsabilidade, ética e atenção aos detalhes.**