Introdução ao Processamento de Dados

Fundamentos de Coleta de Dados

Professor Douglas Castro

O que é Processamento de Dados?

Conjunto de atividades de coleta, organização, transformação, análise e armazenamento de dados.

Exemplos práticos: relatórios de vendas, análise de perfil de clientes, monitoramento de redes sociais.

Geração de informações úteis para a tomada de decisão.

Tipos de Dados

Dados Estruturados:
- Formato tabular (Excel, CSV, bancos SQL)

Dados Semi-Estruturados:
- JSON, XML, logs de aplicativos

Dados Não Estruturados:
- Textos, imagens, vídeos, áudios

Fases do Processamento de Dados

1. Coleta de Dados
2. Limpeza de Dados
3. Transformação de Dados
4. Armazenamento
5. Análise e Visualização
6. Modelagem (etapas avançadas)
7. Compartilhamento/Distribuição

Notebook 1

Limpeza e Análise Exploratória de Dados

Limpeza de dados (ou *data cleaning*) é o processo de **identificar, corrigir ou remover erros e inconsistências** em um conjunto de dados antes da análise.

Dados brutos geralmente contêm erros, valores faltantes, duplicações e formatos inconsistentes.
Se esses problemas não forem tratados, podem levar a resultados incorretos, modelos de baixa qualidade ou até tomadas de decisão erradas.

Principais problemas encontrados:

Dados ausentes (valores nulos)

Duplicatas

Inconsistências de formato (ex: datas em formatos diferentes)

Outliers (valores muito fora da média)

Erros de digitação ou codificação

Dados Faltantes (Missing Values)

São células/valores que deveriam ter um dado, mas estão **em branco** ou com **indicadores de ausência**, como NaN, null, None, ou até strings como "N/A".

Causas comuns:

- Problemas na coleta
- Dados não informados
- Erros de integração entre sistemas

O que fazer com dados faltantes?

- Excluir linhas/colunas
- Preencher com médias, medianas ou valores padrão
- Preencher com valor mais frequente (moda)
- Deixar como estão (casos raros)

Dados Inconsistentes

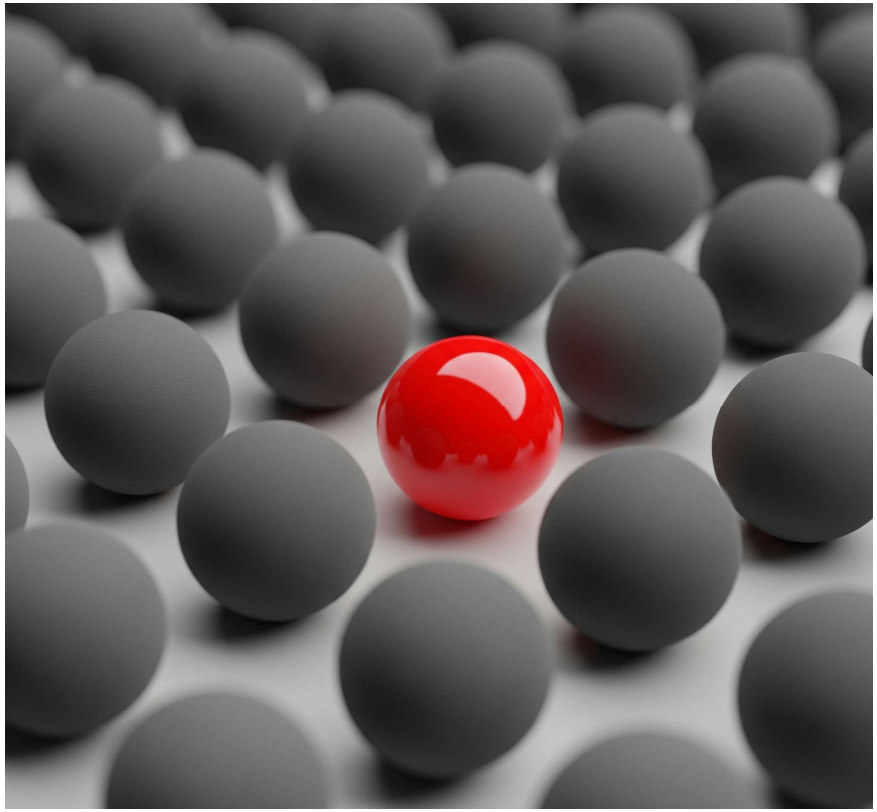
- Formatos de data diferentes: "01/06/2025", "2025-06-01", "Junho 1, 2025"
- Valores numéricos como texto: "1000" (string) e 1000 (inteiro)
- Categorias com grafia errada: "SP", "sp", "São Paulo"

O que fazer?

- Padronizar formatos
- Corrigir erros de digitação
- Transformar tipos de dados (ex: string para número)

Outliers (Valores Fora do Padrão)

São valores **extremos ou fora do padrão esperado**, que podem indicar **erros de registro**, **anomalias** ou **casos especiais**.

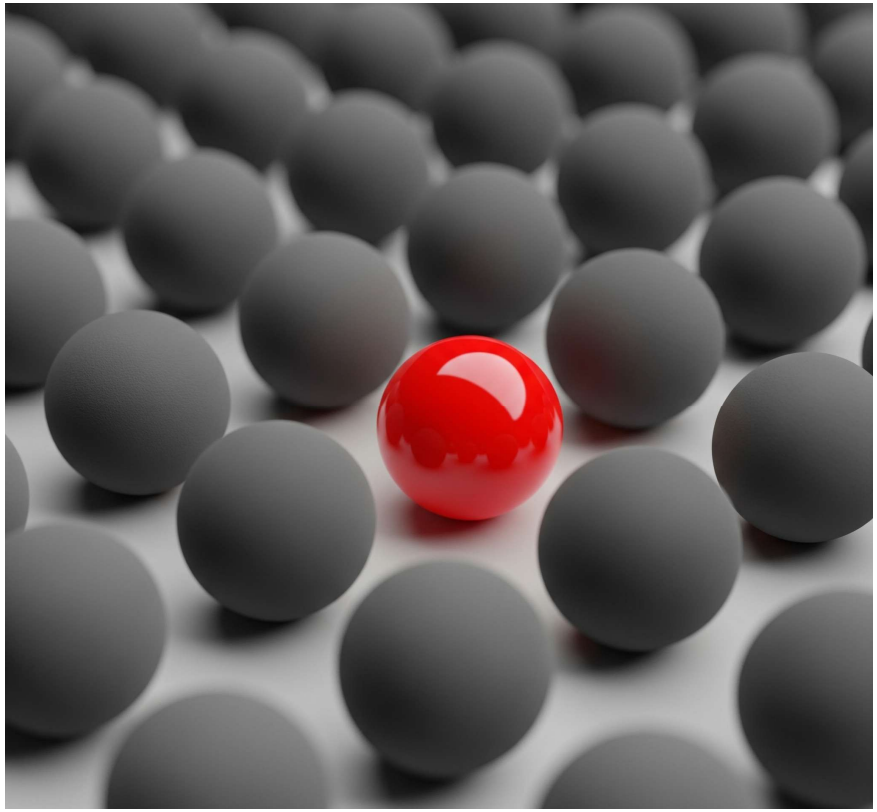


- Uma pessoa com 300 anos de idade num cadastro
- Um salário de R$ 1.000.000,00 em um grupo onde a média é R$ 3.000,00

Outliers (Valores Fora do Padrão)

Como identificar?

- Estatísticas descritivas (média, desvio padrão)
- Boxplots
- Regras de intervalo (ex: valores fora de 1.5xIQR)



O que fazer?

- Analisar caso a caso
- Corrigir se for erro
- Excluir se for dado incorreto
- Manter se for um caso real e relevante

Notebook 1