Aula 9 – Limpeza e Preparação de Dados (Data Cleaning)

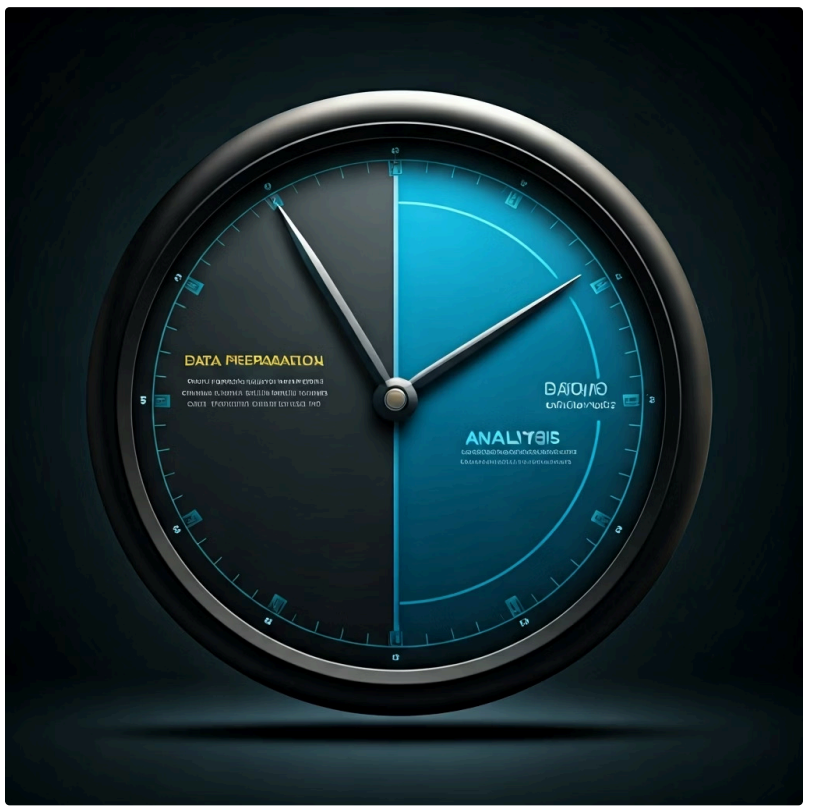
Seja bem-vindo(a) à nona aula do nosso curso de Análise de Dados para Negócios! Imagine que você está prestes a preparar uma refeição importante, mas os ingredientes estão espalhados, alguns estragados, outros com rótulos ilegíveis. Seria impossível cozinhar algo de qualidade, certo? No mundo dos dados, a situação é muito parecida. Antes de extrair qualquer insight valioso, precisamos garantir que nossos "ingredientes" – os dados – estejam limpos, organizados e prontos para uso.

Esta aula é um mergulho profundo na etapa que, muitas vezes, é subestimada, mas que consome a maior parte do tempo de um analista: a limpeza e preparação de dados, ou Data Cleaning. Você descobrirá por que essa fase é tão crucial e como ela impacta diretamente a confiabilidade das suas análises e decisões de negócio. Nosso objetivo é que, ao final, você seja capaz de identificar, diagnosticar e aplicar as principais técnicas para transformar dados brutos e caóticos em informações estruturadas e prontas para gerar valor.

Vamos explorar desde a famosa regra 80/20, que explica a dedicação necessária a essa etapa, até as estratégias para lidar com dados ausentes, duplicados e inconsistências de formato. Prepare-se para desvendar os segredos que garantem a qualidade dos dados, pavimentando o caminho para análises mais precisas e resultados mais confiáveis.

A Regra 80/20: Por Que a Preparação Consome a Maior Parte do Tempo

Você já ouviu falar da Regra de Pareto, ou a regra 80/20? Ela diz que, em muitos eventos, 80% dos efeitos vêm de 20% das causas. No universo da análise de dados, essa regra se manifesta de uma forma um tanto surpreendente e, para muitos iniciantes, frustrante: cerca de **80% do tempo de um analista é dedicado à limpeza e preparação dos dados**, e apenas 20% à análise propriamente dita e à geração de insights. Parece desproporcional, não é? Mas há uma razão muito lógica para isso.



Pense na construção de um prédio. A maior parte do trabalho inicial não é erguer as paredes ou pintar, mas sim preparar o terreno, fazer as fundações, garantir que a base seja sólida e nivelada. Sem essa etapa meticulosa, qualquer estrutura construída sobre ela estará fadada a ter problemas. Com os dados, é idêntico. Se a base estiver cheia de erros, inconsistências ou lacunas, qualquer modelo analítico ou relatório que você criar será, na melhor das hipóteses, impreciso, e na pior, completamente enganoso.

Essa dedicação intensa à preparação não é um luxo, mas uma necessidade. Dados vêm de diversas fontes – sistemas legados, planilhas preenchidas manualmente, APIs de terceiros – e raramente chegam em um formato perfeito. Eles podem ter erros de digitação, formatos diferentes para a mesma informação, valores ausentes ou registros duplicados. Ignorar essas questões é como tentar construir um castelo de cartas em um terreno instável: o resultado será frágil e insustentável.

Entendendo os Dados Ausentes (Missing Values)

O Problema

Imagine montar um quebra-cabeça sem várias peças. Como ter a imagem completa?

O Impacto

Distorção de análises, conclusões erradas e falhas em algoritmos de machine learning.

A Solução

Identificar, entender a natureza e aplicar estratégias adequadas de tratamento.

Um dos desafios mais comuns e traiçoeiros na limpeza de dados são os famosos "missing values", ou dados ausentes. Imagine que você está montando um quebra-cabeça, mas percebe que várias peças simplesmente não estão na caixa. Como você pode ter a imagem completa e clara se há lacunas significativas? No contexto dos dados, valores ausentes são exatamente isso: informações que deveriam estar lá, mas por algum motivo, não estão.

Essas lacunas podem surgir por diversas razões: um usuário que não preencheu um campo em um formulário, um sensor que falhou em registrar uma leitura, um erro na extração de dados, ou até mesmo dados que não eram aplicáveis a um determinado registro. Independentemente da causa, a presença de dados ausentes pode distorcer suas análises, levar a conclusões erradas e até mesmo impedir que certos algoritmos de machine learning funcionem corretamente.

> É crucial não apenas identificar onde estão esses buracos, mas também entender a natureza deles. Existem diferentes tipos de dados ausentes, e a forma como você decide tratá-los dependerá dessa compreensão.

Por exemplo, se os dados estão ausentes completamente ao acaso (MCAR - Missing Completely At Random), o tratamento pode ser mais simples. Mas se a ausência está relacionada a outros dados (MAR - Missing At Random) ou ao próprio valor que está faltando (MNAR - Missing Not At Random), a abordagem precisa ser mais sofisticada para evitar vieses.

Estratégias para Lidar com Dados Ausentes

Agora que compreendemos a natureza dos dados ausentes, a pergunta que surge é: o que fazemos com eles? Não existe uma solução única e perfeita, mas sim um conjunto de estratégias que devem ser aplicadas com discernimento, considerando o contexto do seu conjunto de dados e os objetivos da sua análise. A escolha errada pode ser tão prejudicial quanto ignorar o problema.

1

Remoção

Quando usar: Poucos dados ausentes, ausência MCAR

- Remover linhas com valores ausentes
- Remover colunas com muitos valores ausentes
- Simples, mas pode causar perda de informação

2

Imputação Simples

Quando usar: Dados numéricos/categóricos, ausência MCAR

- Preencher com média, mediana ou moda
- Usar valor constante (zero, "desconhecido")
- Preserva tamanho do dataset

3

Imputação Avançada

Quando usar: Ausência MAR/MNAR, alta importância

- Regressão para prever valores
- Algoritmos de machine learning
- Mais precisa, preserva relações

Uma das abordagens mais diretas é a **remoção**. Você pode optar por remover as linhas (registros) que contêm valores ausentes. Isso é viável quando a quantidade de dados ausentes é pequena em relação ao total do dataset e quando a remoção não causará uma perda significativa de informações ou um viés na amostra. Por outro lado, se uma coluna inteira tem muitos valores ausentes, talvez seja melhor remover a coluna, se ela não for essencial para a análise. No entanto, a remoção excessiva pode levar à perda de informações valiosas e reduzir o poder estatístico da sua análise.

A outra estratégia principal é a **imputação**, que consiste em preencher os valores ausentes com estimativas. Isso pode ser feito de várias maneiras: preencher com a média, mediana ou moda da coluna (para dados numéricos e categóricos, respectivamente), usar um valor constante (como zero ou "desconhecido"), ou aplicar métodos mais avançados como regressão ou algoritmos de machine learning para prever os valores ausentes com base em outras variáveis. A imputação é poderosa porque preserva o tamanho do seu dataset, mas exige cuidado para não introduzir vieses ou distorcer a distribuição original dos dados.

Exemplo prático: Imagine uma pesquisa de satisfação onde alguns clientes não informaram a idade. Se a idade é importante para segmentação, remover esses registros pode distorcer a análise demográfica. Imputar a idade com a mediana do grupo pode ser uma solução mais adequada, mantendo a integridade do conjunto de dados.

Estratégia	Vantagens	Desvantagens	Cenário de Uso
Remoção de Linhas	Simples, evita vieses de imputação	Perda de dados, pode gerar vieses se ausência não for aleatória	Poucos dados ausentes, ausência MCAR
Remoção de Colunas	Simplifica o dataset	Perda de informação de uma variável inteira	Coluna com muitos dados ausentes e pouca relevância
Imputação (Média/Mediana/Moda)	Preserva o tamanho do dataset, fácil de implementar	Pode reduzir a variância, distorcer distribuições	Dados numéricos/categóricos com ausência MCAR
Imputação (Avançada)	Mais precisa, preserva relações entre variáveis	Mais complexa, exige mais recursos computacionais	Ausência MAR ou MNAR, alta importância da variável

Identificação e Remoção de Registros Duplicados

O Problema dos Duplicados

Além dos dados ausentes, outro vilão silencioso que pode comprometer seriamente suas análises são os registros duplicados. Pense em uma lista de contatos onde a mesma pessoa aparece duas ou três vezes com pequenas variações. Se você tentar enviar um e-mail para essa lista, a pessoa receberá a mensagem múltiplas vezes, o que é ineficiente e irritante. No contexto dos dados de negócio, duplicatas podem levar a contagens erradas, relatórios inflacionados e decisões baseadas em informações superestimadas.

Causas Comuns:

- Erros de entrada de dados (digitação dupla)
- Fusão de bancos de dados com sobreposição
- Problemas em processos de extração
- Falhas na lógica de negócio (IDs idênticos)

Impacto em Clientes

Cliente duplicado = ofertas múltiplas para a mesma pessoa, métricas de retenção inflacionadas

Impacto em Vendas

Transações duplicadas = superestimação da receita, relatórios financeiros incorretos

Impacto em Análises

Contagens erradas = decisões baseadas em dados distorcidos, perda de credibilidade

Registros duplicados surgem por uma variedade de motivos: erros de entrada de dados (um operador digitou a mesma informação duas vezes), fusão de bancos de dados (onde registros de diferentes sistemas se sobrepõem), problemas em processos de extração ou replicação de dados, ou até mesmo falhas na lógica de negócio que permitem a criação de IDs idênticos. O problema é que, muitas vezes, essas duplicatas não são idênticas em todos os campos, o que as torna mais difíceis de detectar.

A presença de duplicatas pode ter consequências sérias. Em um banco de dados de clientes, um cliente duplicado pode significar que você está enviando ofertas para a mesma pessoa várias vezes, ou pior, que suas métricas de retenção e aquisição estão inflacionadas. Em um sistema de vendas, transações duplicadas podem levar a uma superestimação da receita. Identificar e remover esses "clones" é um passo fundamental para garantir a integridade e a confiabilidade do seu conjunto de dados.

Técnicas de Remoção de Duplicados

Compreender a existência de duplicatas é o primeiro passo; o próximo é saber como eliminá-las de forma eficaz. A remoção de registros duplicados não é apenas uma questão de apertar um botão, mas sim de aplicar uma lógica que respeite a integridade dos seus dados e os objetivos da sua análise. Existem diferentes abordagens, dependendo do que você considera um "duplicado".

Duplicatas Exatas

Linhas idênticas em *todos* os campos. Boa primeira varredura, mas insuficiente.

Baseada em Chaves

Define colunas únicas (ex: Nome + CPF). Mais sofisticada e eficaz.

Decisão de Manutenção

Qual registro manter? Primeiro, último ou mais completo?

A forma mais simples é a **remoção de duplicatas exatas**, onde o sistema busca por linhas que são idênticas em *todos* os campos. Essa é uma boa primeira varredura, mas muitas vezes insuficiente, pois pequenas variações (como um espaço extra ou uma letra maiúscula/minúscula) podem fazer com que registros que são conceitualmente duplicados não sejam identificados.

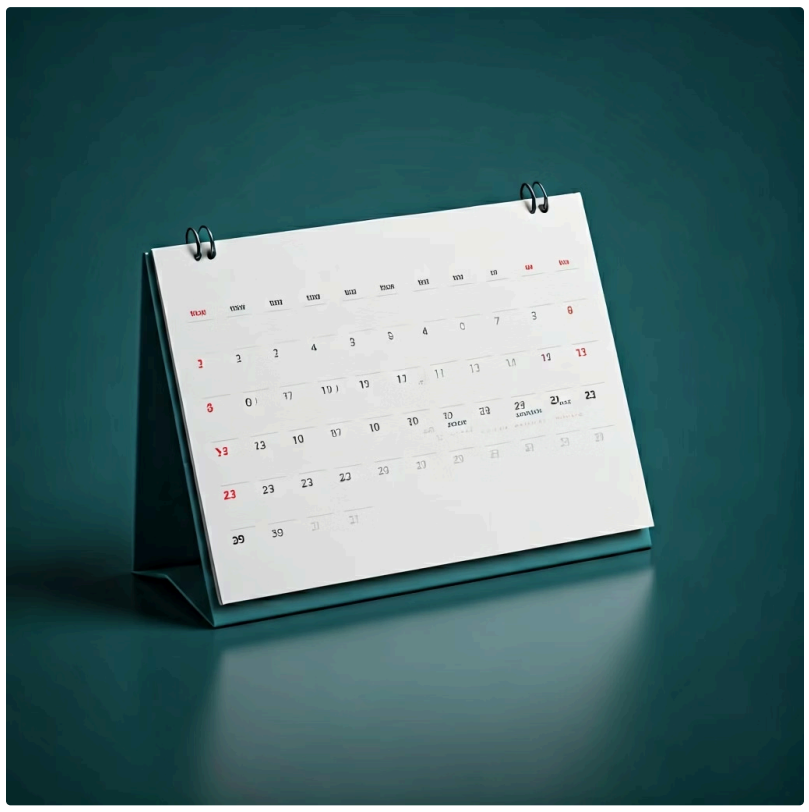
Uma abordagem mais sofisticada é a **remoção baseada em chaves ou um subconjunto de colunas**. Aqui, você define quais colunas, em conjunto, devem ser únicas para identificar um registro. Por exemplo, em uma tabela de clientes, a combinação de "Nome", "Sobrenome" e "CPF" pode ser considerada uma chave única. Se houver duas linhas com a mesma combinação desses três campos, uma delas é uma duplicata. É importante decidir qual registro manter (o primeiro, o último, o mais completo) caso haja diferenças em outros campos. Ferramentas como Excel e SQL oferecem funcionalidades robustas para essa tarefa. No Excel, a função "Remover Duplicatas" permite selecionar as colunas a serem consideradas. No SQL, a cláusula DISTINCT ou a combinação de GROUP BY com funções de agregação e HAVING são poderosas.

> **Exemplo prático:** Em uma lista de pedidos, se você tem ID_Pedido, Data, ID_Cliente e Valor, e percebe que o ID_Pedido deveria ser único, pode usar essa coluna como chave para identificar e remover duplicatas. Se houver dois registros com o mesmo ID_Pedido, mas com Valor diferente, você precisará investigar qual é o correto antes de simplesmente remover.

Ferramentas Práticas:

- **Excel:** Função "Remover Duplicatas" com seleção de colunas
- **SQL:** DISTINCT, GROUP BY + HAVING, DELETE com subconsultas
- **Power BI:** Power Query Editor com "Remover Linhas Duplicadas"

Padronização de Formatos – Datas



O Desafio das Datas

A padronização de formatos é uma etapa crucial na limpeza de dados, e as datas são, sem dúvida, um dos campos que mais causam dor de cabeça. Imagine que você está analisando as vendas mensais de uma empresa, mas em seu conjunto de dados, as datas aparecem de diversas formas:

- "01/01/2023"
- "Jan-01-23"
- "2023-01-01"
- "Primeiro de Janeiro de 2023"

Como você pode agrupar essas vendas por mês ou ano se o sistema não consegue reconhecer todas essas variações como datas válidas?

> **O problema com a inconsistência de datas é que ela impede que os softwares de análise as interpretem corretamente como valores temporais.** Em vez de tratá-las como datas que podem ser ordenadas, filtradas ou agregadas, o programa as vê como simples strings de texto.

01

Identificar Variações

Detecte todos os formatos de data presentes no dataset

02

Escolher Padrão

Defina um formato único (ex: YYYY-MM-DD)

03

Converter Formatos

Use funções específicas para transformação

04

Validar Resultado

Confirme que todas as datas estão no padrão escolhido

Isso significa que um filtro para "vendas de 2023" pode falhar, ou que a ordenação cronológica se torna impossível. É como ter várias pessoas falando sobre o mesmo evento, mas cada uma em uma língua diferente; a comunicação se torna inviável.

Para resolver isso, é fundamental estabelecer um formato padrão e converter todas as entradas para esse padrão. Por exemplo, decidir que todas as datas serão no formato YYYY-MM-DD (Ano-Mês-Dia) é uma prática comum e robusta. Ferramentas como Excel e SQL oferecem funções específicas para conversão de formatos de data. No Excel, você pode usar a função TEXTO ou as opções de "Formatar Células". No SQL, funções como CONVERT ou FORMAT são essenciais. A padronização garante que todas as datas sejam reconhecidas e tratadas uniformemente, permitindo análises temporais precisas e eficientes.

Padronização de Formatos – Textos e Números

A padronização não se limita apenas às datas; ela é igualmente vital para campos de texto e numéricos. A inconsistência nesses tipos de dados pode ser tão prejudicial quanto nas datas, levando a erros de contagem, falhas em filtros e agrupamentos, e uma visão distorcida da realidade. É como tentar organizar uma biblioteca onde alguns livros estão com o título em maiúsculas, outros em minúsculas, alguns com erros de digitação e outros com espaços extras: a busca por um título específico se torna um pesadelo.

Dados de Texto

Variações de Caixa

"Produto A", "produto a", "PRODUTO A" tratados como diferentes

Solução: Padronizar para uma única caixa

Espaços Extras

" Produto A " ou "Produto A" criam distinções artificiais

Solução: Remover espaços no início, fim e duplicados

Erros de Digitação

"São Paulo" vs. "Sao Paulo" vs. "S. Paulo"

Solução: Correspondência de strings ou correção manual

Sinônimos

"Celular" vs. "Telefone Móvel"

Solução: Criar dicionário de termos para unificar

Dados Numéricos

Separadores Decimais

Vírgula (1,50) vs. Ponto (1.50)

Solução: Padronizar para um único formato

Símbolos de Moeda

"R$ 1.000,00", "1000.00 USD", "1000 unidades"

Solução: Remover símbolos ou tratar separadamente

Valores como Texto

Números importados como texto ("'123'")

Solução: Converter para tipo numérico

Ferramentas Práticas:

Excel: MAIÚSCULA, MINÚSCULA, ARRUMAR, SUBSTITUIR, "Localizar e Substituir"

SQL: UPPER, LOWER, TRIM, REPLACE, CAST, CONVERT

Power BI: Transformações no Power Query Editor

A padronização de textos e números garante que suas análises sejam consistentes e que os dados sejam interpretados corretamente por qualquer ferramenta. No Excel, funções como MAIÚSCULA, MINÚSCULA, ARRUMAR, SUBSTITUIR e "Localizar e Substituir" são poderosas. No SQL, UPPER, LOWER, TRIM, REPLACE e CAST ou CONVERT são frequentemente utilizados.

Validação e Correção de Inconsistências

Após lidar com dados ausentes, duplicados e padronização de formatos, ainda há uma etapa crítica: a validação e correção de inconsistências. Esta fase é como a inspeção final de um produto antes de ser entregue ao cliente. Não basta que as peças estejam lá e no formato certo; elas precisam fazer sentido lógico e estar dentro dos parâmetros esperados. Dados inconsistentes são aqueles que, embora presentes e formatados corretamente, violam regras de negócio, limites lógicos ou padrões de integridade.

Exemplos de Inconsistências

- Quantidade vendida negativa
- Data de entrega anterior à data do pedido
- Idade com valor "200"
- Campo de gênero com valor não permitido

Tipos de Validação

- Verificação de limites (idade 0-120)
- Consistência entre campos
- Valores permitidos (categorias)
- Detecção de outliers

Métodos de Correção

- Correção manual para casos críticos
- Scripts automatizados com regras
- Investigação da fonte de dados
- Ajustes baseados em lógica de negócio

Pense em um registro de vendas onde a quantidade vendida é negativa, ou a data de entrega é anterior à data do pedido. Ou ainda, um campo de idade com o valor "200" ou um campo de gênero com "X" em um sistema que só aceita "M" ou "F". Esses são exemplos de inconsistências que, se não corrigidas, podem levar a análises completamente erradas. Uma média de idade distorcida, um cálculo de estoque negativo ou uma previsão de vendas baseada em dados ilógicos são resultados diretos da falta de validação.

> **A validação envolve a aplicação de regras de negócio e lógicas para verificar a integridade dos dados.** A correção pode ser manual (para casos pontuais e críticos) ou automatizada, utilizando scripts e regras para ajustar os dados. Em alguns casos, a inconsistência pode indicar um erro na fonte de dados, exigindo uma investigação mais profunda.

Ferramentas e Abordagens para Data Cleaning

A teoria da limpeza de dados é fundamental, mas a aplicação prática exige o domínio de ferramentas adequadas. Felizmente, o mercado oferece uma gama de opções, desde as mais acessíveis e amplamente utilizadas até as mais robustas e especializadas. Para o nosso público, com foco em Excel, SQL e uma introdução ao Power BI, as possibilidades são vastas e poderosas.

Excel

Arsenal para o dia a dia:

- Remover Duplicatas
- Texto para Colunas
- Localizar e Substituir
- Funções: ARRUMAR, MAIÚSCULA, MINÚSCULA
- Validação de Dados
- Filtros e Classificação

SQL

Poder para grandes volumes:

- DISTINCT para registros únicos
- GROUP BY e HAVING
- UPDATE e DELETE
- TRIM, UPPER, LOWER, REPLACE
- CAST e CONVERT
- CASE WHEN para lógicas complexas

Power BI

Transformação visual e replicável:

- Power Query Editor
- Remover colunas e renomear
- Alterar tipos de dados
- Preencher valores ausentes
- Remover duplicatas
- Dividir colunas

No Excel, você tem um arsenal de funcionalidades para o dia a dia: Remover Duplicatas, Texto para Colunas, Localizar e Substituir, funções de texto (ARRUMAR, MAIÚSCULA, MINÚSCULA, ESPAÇOS), Validação de Dados, e Filtros e Classificação para identificar rapidamente anomalias e padrões.

Com SQL, a capacidade de manipular grandes volumes de dados é incomparável: DISTINCT para selecionar apenas registros únicos, GROUP BY e HAVING para identificar e agrupar duplicatas ou inconsistências, UPDATE e DELETE para corrigir ou remover registros, funções de string (TRIM, UPPER, LOWER, REPLACE), CAST e CONVERT para mudar tipos de dados e formatos, e CASE WHEN para aplicar lógicas condicionais complexas.

No Power BI, a etapa de "Transformar Dados" (Power Query Editor) é um ambiente poderoso para limpeza e preparação. Ele permite aplicar uma série de transformações de forma visual e registrar cada passo, criando um fluxo de trabalho replicável. Você pode remover colunas, renomear, alterar tipos de dados, preencher valores ausentes, remover duplicatas, dividir colunas e muito mais, tudo com uma interface intuitiva.

O Impacto da Data Literacy na Limpeza de Dados

Data Literacy

A diferença entre executar e compreender

A limpeza de dados não é meramente uma tarefa técnica; ela é profundamente influenciada pela sua "Data Literacy", ou alfabetização em dados. Imagine que você está limpando um quarto. Se você não sabe o que é lixo e o que é um objeto de valor, ou onde cada coisa deve ser guardada, sua "limpeza" pode acabar jogando fora algo importante ou guardando itens no lugar errado. Da mesma forma, sem uma boa compreensão dos dados, a limpeza pode ser ineficaz ou até prejudicial.

> **A Data Literacy é a capacidade de ler, trabalhar, analisar e comunicar com dados.** Quando aplicada à limpeza, ela significa que você não apenas sabe *como* usar as ferramentas (Excel, SQL, Power BI), mas também *por que* está realizando cada etapa.

Compreender o Contexto

De onde vieram os dados? O que cada coluna representa?

Conhecer Regras de Negócio

Quais são as políticas e padrões da organização?

Tomar Decisões Informadas

Escolher estratégias adequadas para cada situação

Definir Objetivos

Qual o propósito final da análise?

Você entende o contexto de onde os dados vieram, o que cada coluna representa, quais são as regras de negócio associadas e qual o objetivo final da análise. Essa compreensão profunda permite que você tome decisões informadas sobre: qual tipo de dado ausente você está lidando e qual a melhor estratégia de imputação ou remoção; o que constitui um registro duplicado em um contexto específico de negócio; qual o formato padrão ideal para datas, textos e números, considerando as necessidades da análise e dos usuários; quais inconsistências são críticas e precisam de correção imediata, e quais são apenas "ruído" aceitável.

> **Uma pessoa com alta Data Literacy não apenas executa as tarefas de limpeza, mas questiona a qualidade dos dados na fonte, sugere melhorias nos processos de coleta e entende o impacto de cada decisão de limpeza nos resultados finais.** Isso transforma a limpeza de dados de uma tarefa mecânica em um processo estratégico, garantindo que os insights gerados sejam realmente confiáveis e relevantes para o negócio.

Desafios Comuns e Melhores Práticas

A jornada pela limpeza de dados raramente é um caminho reto e fácil. Existem desafios inerentes que todo analista enfrentará, mas também há melhores práticas que podem transformar essa tarefa árdua em um processo mais eficiente e menos propenso a erros. Reconhecer esses obstáculos e adotar as estratégias corretas é fundamental para o sucesso.

Desafios Comuns

Volume e Variedade

Terabytes de dados de múltiplas fontes, cada uma com particularidades

Dados Legados

Sistemas antigos com formatos inconsistentes e documentação precária

Erros Humanos

Digitação manual como fonte constante de erros e inconsistências

Falta de Padrões

Ausência de políticas claras de entrada e armazenamento

Ambiguidade

Campos com múltiplos significados ou interpretações diferentes

Manutenção Contínua

Dados continuam chegando e se degradando ao longo do tempo

Melhores Práticas

1 **Entenda o Negócio**

Compreenda o que os dados representam e o objetivo da análise

2 **Documente o Processo**

Registre cada etapa, decisões e justificativas

3 **Comece Pequeno**

Teste técnicas em amostras antes de aplicar em todo o dataset

4 **Automatize**

Use scripts para tarefas repetitivas, reduzindo erros

5 **Valide Regularmente**

Monitore a qualidade dos dados ao longo do tempo

6 **Comunique-se**

Reporte problemas às fontes para melhorias na coleta

7 **Use Ferramentas Adequadas**

Escolha as ferramentas certas para cada situação

8 **Itere e Refine**

Revisitar etapas conforme descobre novas inconsistências

Consolidação e Próximos Passos

Chegamos ao fim de uma das aulas mais fundamentais para qualquer aspirante a analista de dados. Percorremos a importância crítica da limpeza e preparação de dados, desvendando a regra 80/20 que governa essa etapa. Aprendemos a identificar e tratar dados ausentes, a caçar e remover registros duplicados, e a padronizar formatos de datas, textos e números para garantir a consistência. Exploramos a validação de inconsistências e as ferramentas práticas que nos auxiliam nessa jornada, como Excel, SQL e Power BI.

Em prática:

Lembre-se que **a qualidade da sua análise é diretamente proporcional à qualidade dos seus dados**. Dedique tempo à limpeza, encare-a como um investimento, não como um fardo. Documente seus passos, automatize o que puder e sempre questione a integridade dos seus dados. Uma base sólida é o segredo para insights confiáveis e decisões de negócio assertivas.

Autoavaliação

1. Qual das seguintes afirmações melhor descreve a "Regra 80/20" no contexto da análise de dados?
   - a) 80% dos dados são irrelevantes e 20% são cruciais.
   - b) 80% do tempo é gasto na análise e 20% na preparação.
   - c) 80% do tempo é gasto na preparação de dados e 20% na análise.
   - d) 80% dos problemas de dados são causados por 20% das fontes.
2. Ao lidar com dados ausentes, qual das seguintes estratégias é mais adequada quando a perda de dados é pequena e a ausência é completamente aleatória (MCAR)?
   - a) Imputar com valores de regressão.
   - b) Remover as linhas com valores ausentes.
   - c) Preencher com um valor constante como "0".
   - d) Ignorar os valores ausentes e prosseguir com a análise.
3. Em uma tabela de clientes, você identifica que a combinação de "Nome", "Sobrenome" e "CPF" deveria ser única, mas há registros com a mesma combinação. Qual técnica de limpeza de dados você aplicaria?
   - a) Padronização de formatos de data.
   - b) Validação de limites numéricos.
   - c) Remoção de registros duplicados baseada em chaves.
   - d) Imputação de dados ausentes.
4. Qual das seguintes ferramentas é mais indicada para realizar transformações visuais e registrar o fluxo de trabalho de limpeza de dados de forma replicável, especialmente para volumes maiores?
   - a) Microsoft Word.
   - b) Bloco de Notas.
   - c) Power BI (Power Query Editor).
   - d) Calculadora do Windows.

Gabarito

Questão 1

Resposta: c)

Questão 2

Resposta: b)

Questão 3

Resposta: c)

Questão 4

Resposta: c)

Questão Discursiva

> Explique a importância da Data Literacy (Alfabetização em Dados) para o processo de limpeza de dados, detalhando como ela influencia as decisões tomadas durante essa etapa e o impacto nos resultados finais da análise.

Continue sua jornada

Próxima Aula: Definição de KPIs e Métricas de Negócio

Na nossa próxima aula, a **Aula 10**, daremos um passo fundamental para transformar dados limpos em inteligência de negócio, explorando a "Definição de KPIs e Métricas de Negócio". Você aprenderá a escolher os indicadores certos para medir o desempenho e o sucesso das suas estratégias.

Recursos Adicionais

- **Artigo sobre Data Cleaning (Kaggle):** Para aprofundar nas técnicas e exemplos práticos.
- **Documentação Power Query (Microsoft):** Para explorar as funcionalidades de transformação de dados no Power BI.
- **Livro "Storytelling with Data" (Cole Nussbaumer Knaflic):** Embora não seja sobre limpeza, reforça a importância de dados confiáveis para uma boa narrativa.

NOTA IMPORTANTE: As informações regulatórias/legais/técnicas desta aula estão atualizadas até 2025. Consulte sempre fontes oficiais para verificar alterações.