Dados sujos (Dirty data) 🗳

Alguns tipos de dados sujos são:

- Dados duplicados (Duplicate data): qualquer dado que aparece mais de uma vez
- Dados desatualizados (Outdated data): dados que deveriam ser substituídos por uma versão mais recente
- Dados incompletos (Incomplete data): dados com informações importantes faltando
- Dados incorretos (Inaccurate data): dados completos porém incorretos
- Dados inconsistentes (Inconsistent data): qualquer dado que use diferentes formas de representar a mesma coisa

Limpando dados 🗭



Ao limpar os dados, certifique-se de:

- 1. Checar erros de gramática
- 2. Documentar como corrigir os erros que encontrar: com isso é possível agilizar o processo de limpeza na próxima vez
- 3. Checar dados em colunas erradas (misfielded values): pode acontecer de um dado correto ter sido colocado na coluna errada, por exemplo um nome de um país ter sido colocado na coluna de cidade e vice-versa.
- 4. Não ignorar valores nulos: caso contrário, essa prática pode resultar em análises incorretas ou incompletas
- 5. Limpar os dados como um todo: olhar apenas um pedaço dos dados, ou apenas alguns campos pode levar a informações repetidas e outros problema
- 6. Mantenha o foco no objetivo: a limpeza deve ser focada no objetivo final da análise
- 7. Corrija a raiz do problema: ao encontrar um problema, além de corrigi-lo, é importante corrigir o que causou ele, para não perder mais tempo no futuro com o mesmo erro
- 8. Analisar o sistema antes de começar a limpeza: é importante entender da onde aquele erro veio e como foi causado, para saber como corrigir corretamente.
- 9. Sempre fazer um Backup antes da limpeza
- 10. **Considerar o tempo de limpeza no prazo final:** o processo de limpeza toma tempo, então é sempre importante considerar ele para se manter dentro dos prazos.

Lista de problemas mais comuns:

Dados nulos

- Erros de gramática
- Erros de casa decimal em números
- Espaços em branco extras
- Dados duplicados
- Tipo de dado incorreto
- Nome de colunas incorreto ou sem sentido
- Dados formatados de forma inconsistente

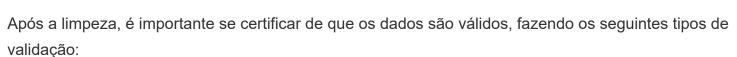
Guias de limpeza em Planilhas:

- · Limpando os dados no Excel
- Limpando do dados no Google Sheets

Automatizando a limpeza de dados <a> ©

- Automatizando análise de dados
- Automatizando Big-Data
- 10 ferramentas para automatização

Validando dados limpos



- 1. **Tipo do Dado | Data Type**: checar se o tipo de dado bate com o tipo do campo *(ex.: um valor de quantidade deve ser do tipo numérico)*
- 2. Extensão dos Dados | Data Range: checar se o valor está dentro dos limites permitidos pelo campo (ex.: um valor para mês deve estar entre 1 e 12)
- Restrição dos Dados | Data Constraints: verificar se o dado cumpre as restrições do campo (ex.: campo de CPF com 11 dígitos)
- 4. Consistência dos Dados | Data Consistency: verificar se o dado faz sentido no contexto (ex.: a data de entrega de um produto não pode ser antes da data de envio do produto)
- 5. **Estrutura do Dado | Data Structure**: o dado deve estar conforme a estrutura do campo (ex.: um dado para uma tabela de um site deve ter a estrutura de tabela)
- 6. Validação do Código | Code Validation: checar se o código já faz qualquer uma das validações acima na coleta de dados do usuário (ex.: limitar a entrada de notas de uma prova para valores entre 0 e 10)