▷ Dados Sujos

↳ São dados incompletos, incorretos ou irrelevantes para o problema que se está tentando resolver.

- Tipos de dados Sujos

1 - **Dados duplicados**

Qualquer registro que apareça mais de uma vez

2 - **Dados desatualizados**

Dados antigos que devem ser substituídos por informações mais novas e mais precisas.

3 - **Dados Incompletos**

Dados que faltam campos importantes

↳ Coleta inadequada

4 - **Dados Incorretos / Imprecisos**

Dados completos, mas imprecisos

5 - **Dados Inconsistentes**

Dados que usam formatos diferentes para representar a mesma coisa.

- Impacto dos dados Sujos

- Decisões imprecisas
- Perda de receita
- Insights imprecisos

➡ Erros comuns a serem evitados

- Não verificar erros de ortografia.
- Esquecer de documentar os erros.
- Não verificar se há valores com campos incorretos.
- Ignorar os valores ausentes.
- Analisar apenas o subconjunto dos dados.
- Acompanhamento dos objetivos comerciais.
- Não corrigir a origem do erro.
- Não analisar o sistema antes da limpeza de dados.
- Não fazer backups
- Não contabilizar a limpeza de dados nos prazos / Processos

Lista de Verificação de Limpeza

1- Determinar o Tamanho do conjunto de dados

↳ Conjuntos de dados grandes podem ter mais problemas de qualidade de dados e levar mais tempo para serem processados.

2- Determinar o número de categorias ou rótulos

↳ Para entender melhor a diversidade do conjunto de dados

3- Identificar dados ausentes

4- Identificar dados não formatados

5- Explorar os diferentes tipos de dados

↳ Compreender os tipos de dados em seu conjunto de dados ajuda a selecionar métodos de limpeza apropriados e aplicar técnicas de análise de dados relevantes.