Módulo 1

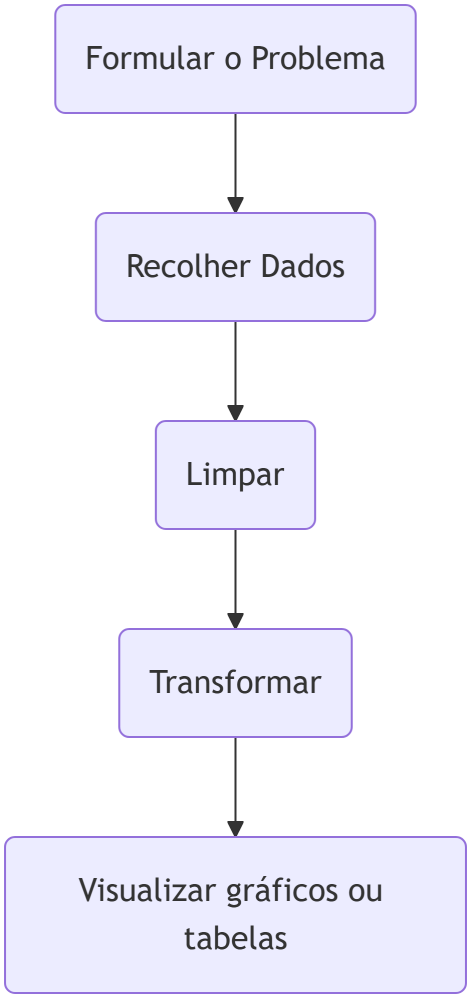
Introdução à Tomada de Decisão com Visualização de Dados

Tiago Afonso

2025-06-11

Módulo 1

Método de Análise de Dados



Dados na tomada de decisão

Qual a importância dos dados?

- **Tomada de Decião**: Os dados são a base da tomada de decisão informada.
- **Compreensão**: Os dados permitem compreender melhor a economia, identificar padrões e tendências, e prever o comportamento económico futuro.
- **Avaliação de Políticas**: Os dados são essenciais para avaliar o impacto das políticas económicas e sociais e para identificar áreas de melhoria.

- **Modelação e Previsão**: Os dados são utilizados para criar modelos económicos e prever o comportamento dos mercados, dos consumidores e da economia em geral.
- **Investigação**: Os dados são fundamentais para a investigação em diversas áreas, permitindo a análise de fenómenos, a validação de hipóteses e a descoberta de novos conhecimentos.

Extensões de ficheiros de dados

Os dados podem ser armazenados em diferentes formatos de ficheiros, cada um com as suas características e aplicações específicas. Aqui estão alguns dos formatos de ficheiros de dados mais comuns e as suas extensões associadas:

`.txt` - texto

`.docx` - microsoft word

`.xlsx` - microsoft excel

`.csv` - comma separated values

`.pbix` - power bi

`.r` - script rar

`.db` - database (sql)

`.json`- javascript object notation

`.xml` - extensible markup language

`.html` - hypertext markup language

`.pdf` - portable document format

`.pptx` - microsoft powerpoint

`.jpg` - imagem

`.png` - imagem

`.mp4` - video

`.mp3` - audio

`.zip` - comprimido "zippado"

`.rar` - comprimido "rar"

`.7z` - compactado "7z"

Nomes de Ficheiros

Ao trabalhar com ficheiros, é importante atribuir nomes de ficheiros simples e bem estruturados para facilitar a organização e a gestão. Aqui estão algumas dicas para atribuir nomes de ficheiros:

1. **Descritivo**: Utilizar nomes de ficheiros que descrevem claramente o conteúdo do ficheiro. Por exemplo, em vez de "dados1.xlsx", utilizar "Vendas_1T_2023.xlsx".
2. **Formatos Consistentes**: Escolha um formato padrão para os nomes de ficheiros e use-o consistentemente. Por exemplo, é possível utilizar "AAAA-MM-DD" para datas (2023-02-19).
3. **Evit arEspaços**: Em vez de espaços, utilizar sublinhados (_) ou hífens (-) para separar palavras. Ajuda a evitar problemas de compatibilidade em algumas linguagens de programação.
4. **Datas**: Para ficheiros que são atualizados regularmente, incluir a data no nome do ficheiro. Por exemplo, "Relatório_Financeiro_2023-02-19.xlsx".
5. **Versões**: Quando se trabalha em várias versões de um ficheiro, utulizar números de versão no nome do ficheiro. Por exemplo, "Projeto_versao1.0.docx". Não utilizar "final".
6. **Comprimento**: Manter os nomes dos ficheiros curtos, mas informativos. Evite nomes excessivamente longos que podem ser difíceis de ler e gerir.
7. **Caracteres Especiais**: Não utilizar caracteres especiais como `/ \ : * ? " < > |`, pois podem causar problemas em diferentes sistemas operativos e linguagens de programação.

Exemplos de Nomes adequados para ´Ficheiros‘

- **"Análise_Mercado_2023-02-19.xlsx"**
- **"Resumo_4T_2023.docx"**
- **"Estudo_Demográfico_versao2.0.pdf"**

Como Fazer Perguntas Orientadas a Dados

Especificidade: Em vez de perguntar "Como é que a economia está a mudar?", deve-se perguntar "Qual foi a taxa de crescimento do PIB de Portugal nos últimos cinco anos?".

Contextualização: Por exemplo, "Qual é a relação entre o nível de educação e o rendimento médio anual em Portugal?" é uma pergunta que oferece um contexto específico.

Quantificação: Por exemplo, "Quantos novos postos de trabalho foram criados na indústria tecnológica em 2024?" é mais claro do que "Houve muitos novos postos de trabalho na indústria tecnológica?".

Escolher de Métricas: Por exemplo, "Como variou a taxa de inflação anual nos últimos dez anos em comparação com a média europeia?" é uma pergunta com métricas específicas.

Exemplos de Perguntas/Hipóteses

Pergunta: "Qual foi a taxa de crescimento das exportações portuguesas no setor automóvel entre 2015 e 2020?"

Hipótese: "O aumento do turismo em Portugal leva a um crescimento significativo do setor da restauração."

Diagnóstico do Excel

Instalar Excel (apenas para computador pessoal)

1. **Site:** portal.office.com
2. **login:** conta UBI
3. Barra lateral esquera ->Apps
4. canto superior direito: `instalar o Office`
5. Utilizar o ficheiro `C1_xlsx` no moodle-UBI.

Funções Básicas do Excel

- **Soma**: `=SOMA(A1:A10)`
- **Média**: `=MÉDIA(A1:A10)`
- **Máximo**: `=MÁXIMO(A1:A10)`
- **Mínimo**: `=MÍNIMO(A1:A10)`

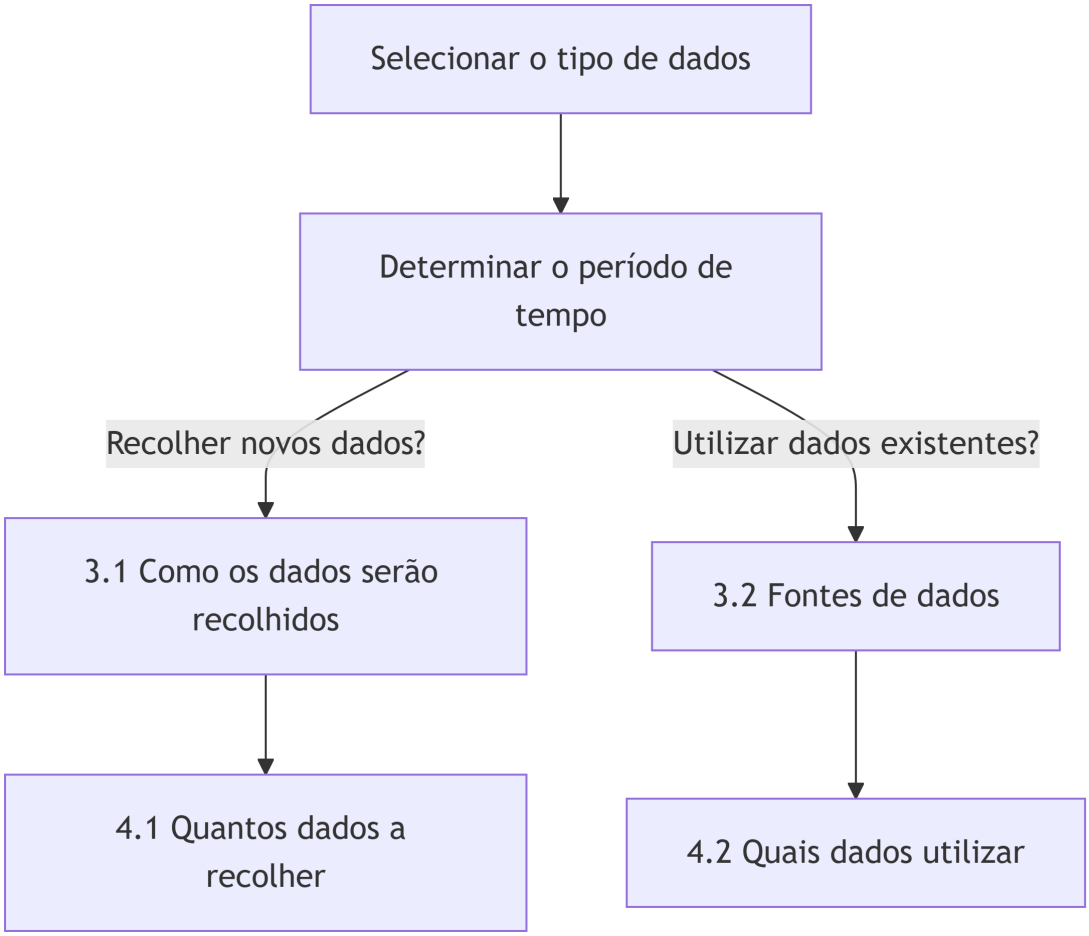
Alguns atalhos do teclado

- Soma automática: `Atl` + `=`
- Selecionar de células: `Shift` + `setas`
- Copiar: `Ctrl` + `C` | Colar: `Ctrl` + `V` | Cortar: `Ctrl` + `X`
- Anular: `Ctrl` + `Z` | Refazer: `Ctrl` + `Y`
- Selecionar conjunto de dados: `Ctrl` + `Shift` + `setas`
- Selecionar linha: `Shift` + `barra de espaço`

- Selecionar coluna: `Ctrl` + `barra de espaço`
- Eliminar coluna/linha: `Ctrl` + `-`
- Inserir coluna/linha: `Ctrl` + `+`
- Eliminar conteúdo selecionado: `Delete`
- Selecionar tudo: `Ctrl` + `T`
- Novo ficheiro: `Ctrl` + `N`
- Abrir ficheiro: `Ctrl` + `O`

Considerações sobre a recolhas de dados

O seguinte diagrama ilustra o processo de recolha de dados e as decisões a tomar em cada etapa.



Selecionar o tipo de dados correto

Tipos de dados

- **Dados Qualitativos** Não podem ser medidos, ou facilmente convertidos em números. Normalmente são listados como *Nomes*, *Descrições* ou *Categorias*.
  - **Ordenados** São dados que podem ser ordenados ou classificados numa escala. Por exemplo, a classificação de um filme ou a posição de um atleta numa corrida.
  - **Nominais** São dados que podem ser categorizados sem uma ordem. Por exemplo, a cor de um carro.
- **Dados Quantitativos** Podem ser medidos e expressos em números. Representam *quantidades*, *medidas* ou *intervalos*.

  - **Discretos** São dados que podem ser contados e que têm um número finito de valores possíveis. Por exemplo, o número de alunos numa sala de aula ou o número de carros numa rua. Não existe 1,5 carros.
  - **Contínuos** São dados que podem ser medidos e que têm um número infinito de valores possíveis. Por exemplo, a altura de uma pessoa ou a velocidade de um carro. Existe 1,5111111 metros de altura.

Formato dos dados

- **Dados Tabulares/estruturados** São dados organizados em linhas e colunas. Cada linha representa um registo e cada coluna representa um atributo.
- **Dados Não-estruturados** São dados que não estão organizados numa estrutura específica. Por exemplo, texto, imagens, vídeos, áudio, etc.

Determinar o período de tempo

Para determinar o período de tempo, é importante considerar:

- **Frequência de atualização** - Com que frequência os dados são atualizados? Anualmente, mensalmente, diariamente, intra diariamente, etc.
- **Granularidade dos dados** - Qual é a unidade de tempo dos dados? Segundos, minutos, horas, dias, semanas, meses, anos, etc.
- **Horizonte temporal** - Qual é o horizonte temporal dos dados? 2000-2020, 2010-2020, 2020-2025, etc.

Tendo em a periodicidade dos dados, podemos dividir os dados em duas categorias:

- **Dados Estáticos** - Dados que não mudam ao longo do tempo. Por exemplo, *inquéritos de satisfação*, *listas de clientes*, *listas de produtos*, etc.
- **Dados dinâmicos/Séries Temporais** - Dados que mudam ao longo do tempo. Por exemplo, *vendas diárias*, *temperatura diária*, *preço das ações*, etc.

Recolher dados

Considerando a recolha de dados, podemos dividir o processo em dois tipos:

- **Dados primários** - Dados recolhidos diretamente pelo investigador para um propósito específico. Por exemplo, *inquéritos*, *entrevistas*, etc.
  - Vantagens:
    * Controlo total sobre a recolha de dados.
    * Dados específicos para o propósito do estudo.
  - Desvantagens:
    * Custo e tempo associados à recolha de dados.
    * Possibilidade de enviesamento dos dados.

O enviesamento dos dados ou da amostra é um problema comum na recolha de dados primários. Pode ocorrer quando a amostra não é representativa da população ou quando os dados são recolhidos de forma tendenciosa (para obter um determinado resultado). Por exemplo: *amostra de conveniência*, *amostra de voluntários*, *amostra de amigos*, etc.

- **Dados secundários** - Dados que já foram recolhidos por outra pessoa ou organização para um propósito diferente. Por exemplo, *bases de dados públicas*, *relatórios de mercado*, *estudos científicos*, etc.
  - Vantagens:
    * Custo e tempo reduzidos na recolha de dados.
    * Dados de fontes credíveis e confiáveis.
    * Possibilidade de comparação com outros estudos.
    * garantia de metodologia adequada na recolha de dados.
  - Desvantagens:
    * Dados podem não ser específicos para o propósito do estudo.
    * Dados podem estar desatualizados ou incompletos.
    * Dados podem não estar disponíveis para o período de tempo desejado.

Quando recorremos a dados secundários, é importante avaliar a qualidade dos dados e a credibilidade da fonte. Por exemplo, verificar a metodologia de recolha de dados, a representatividade da amostra a fiabilidade dos dados.

Exemplos de fontes de dados secundários

Kaggle- Kaggle Datasets

UCI Machine Learning Repository- UCI Machine Learning Repository

Google Dataset Search- Google Dataset Search

Estrutura de dados

Os dados podem estar organizados de diferentes formas, dependendo do tipo de análise que pretendemos realizar. As duas formas mais comuns de organizar os dados são:

Dados em formato largo (*wide data*)

Os dados em formato largo são organizados de forma a que cada linha represente uma observação e cada coluna represente uma variável. Este formato é mais comum em bases de dados tabulares e é mais fácil de ler e interpretar.

Por exemplo, considere a seguinte tabela com dados de vendas de produtos:

Data	Produto A	Produto B	Produto C
2025-01-01	100	200	150
2025-01-02	120	180	160
2025-01-03	130	190	170

Neste formato, cada linha representa uma data e cada coluna representa um produto. Este formato é útil para análises que envolvem comparações entre produtos ou ao longo do tempo.

Dados em formato longo (*long data*)

Os dados em formato longo são organizados de forma a que cada linha represente uma observação única. Este formato é mais comum em análises estatísticas e é mais eficiente para armazenar grandes volumes de dados.

Por exemplo, considere a seguinte tabela com os mesmos dados de vendas de produtos, mas em formato longo:

Data	Produto	Vendas
2025-01-01	A	100
2025-01-01	B	200
2025-01-01	C	150
2025-01-02	A	120
2025-01-02	B	180
2025-01-02	C	160
2025-01-03	A	130
2025-01-03	B	190
2025-01-03	C	170

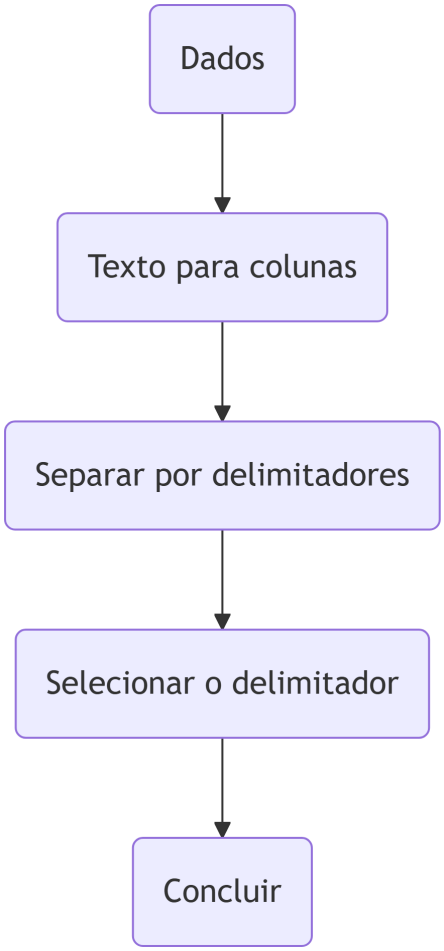
Neste formato, cada linha representa uma venda de um produto numa determinada data. Este formato é útil para análises estatísticas que envolvem a comparação de diferentes produtos ou datas.

Limpar dados

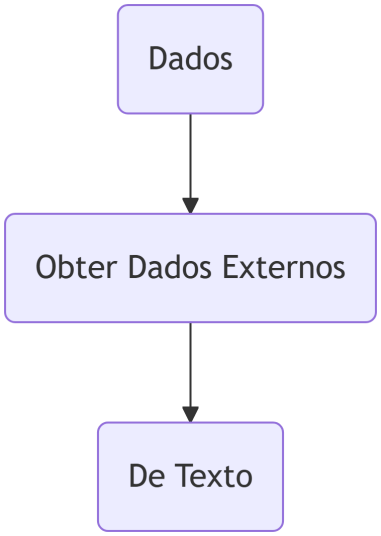
Limpar dados é uma tarefa fundamental para qualquer análise de dados. A limpeza de dados consiste em identificar e corrigir erros nos dados, como valores ausentes (*NA's*), valores duplicados, valores inconsistentes, entre outros.

Limpar dados no excel

No excel é relativamente simples limpar dados, pois os dados são apresentados de forma tabular e é fácil identificar os erros/padrões. No entanto, se os dados forem obtidos através de um ficheiro `.csv` ou `.txt`, é necessário importar os dados para o excel. Os dados podem aparecer todos na primeira coluna e necesse caso é necessário separar os dados em colunas através do menu:



Ou se os dados forem importados diretamente através de:



Os dados já estão separados por colunas.

Copiar -> colar valores

Esta é uma das formas mais simples de limpar dados no excel. Basta copiar os dados e colar como valores. Para isso, basta selecionar os dados, clicar com o botão direito do rato e selecionar a opção `Colar Valores` numa folha em branco. Desta forma, os dados são colados como valores e não como fórmulas e/ou formatação.

Localizar valores ausentes

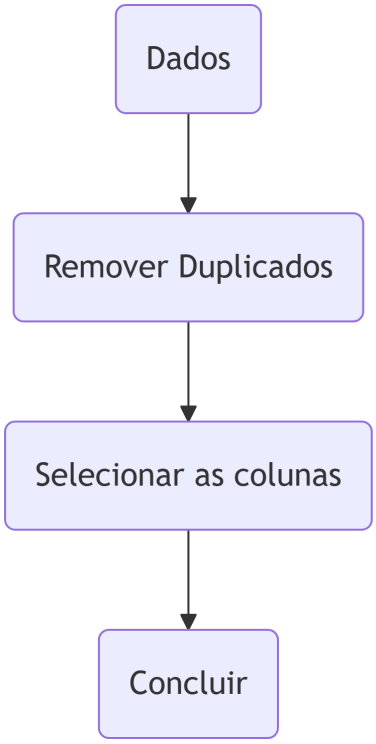
Um valor ausentes podem aparecer de diversas formas. Pode ser um espaço em branco, uma célula embranco, um `#N/A`, um `#DIV/0!`, entre outros. Para localizar valores ausentes, basta selecionar a coluna e clicar em `Localizar e Selecionar` -> `Localizar`. No campo `Localizar`, escrever o valor ausente e clicar em `Localizar Tudo`. Desta forma, é possível identificar todos os valores ausentes.

Outra forma de identificar valores ausentes é através da formatação condicional. Para isso, basta selecionar a coluna, clicar em `Formatação Condicional` -> `Nova Regra` -> `Usar uma fórmula para determinar quais células devem ser formatadas`. No campo `Formatar valores onde esta fórmula é verdadeira`, escrever a fórmula `=É.ERRO(A1)` e clicar em `Formatar`. Desta forma, os valores ausentes são destacados. Também é possível utilizar a fórmula `=É.CÉL.VAZIA(A1)` para identificar valores ausentes.

Outra forma muito eficiente é utilizar a fórmula `=CONTAR` ou algumas variações desta fórmula. Por exemplo, a fórmula `=CONTAR.SE(A1:A10; "NA")` conta o número de vezes que o valor `valor` aparece no intervalo `A1:A10`.

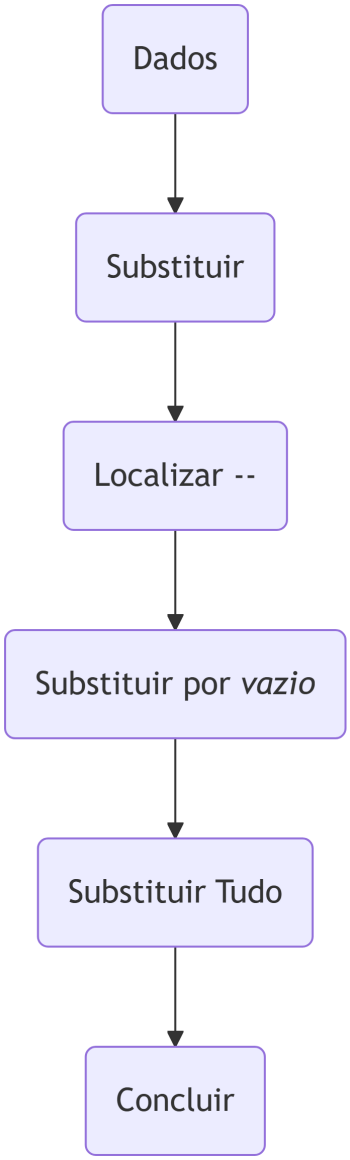
Remover duplicados

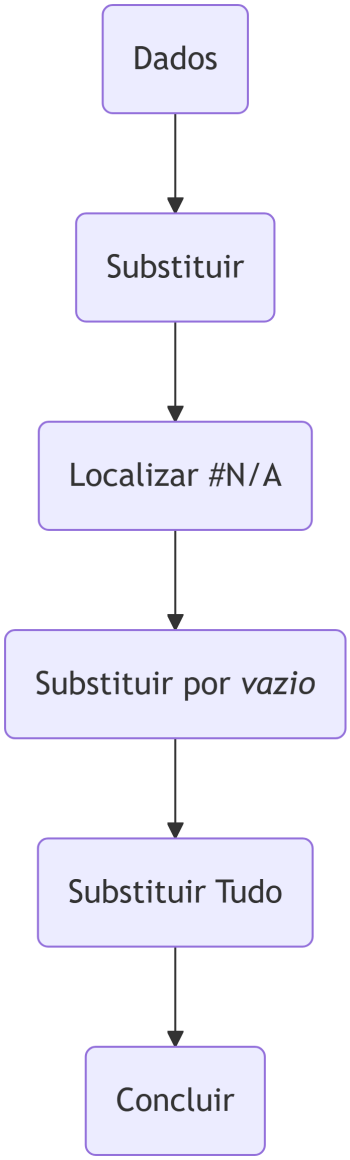
Como remover duplicados no excel:



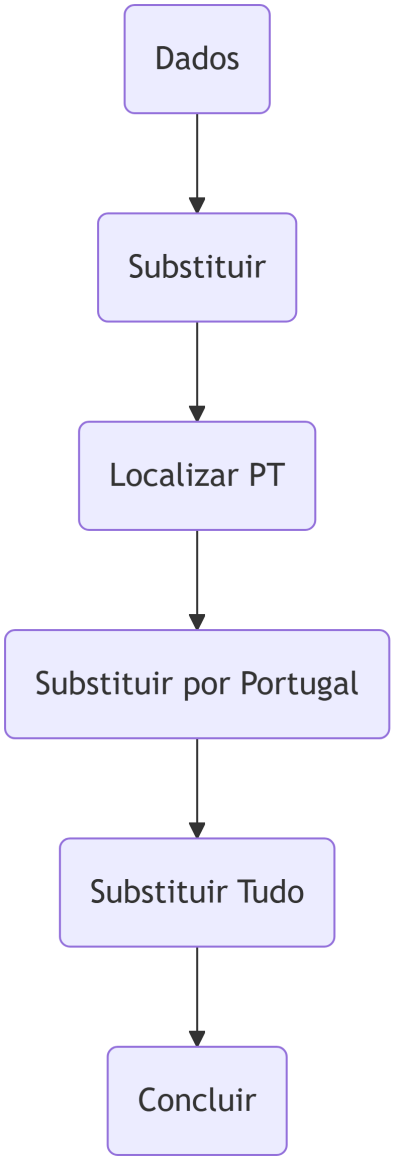
Limpar dados com `ctrl + L`

Com o atalho `Ctrl` + `L`, é possível abrir a caixa de diálogo `Localizar e Substituir`. Nesta caixa de diálogo, é possível substituir valores ausentes por outros valores, por exemplo:


Dados
Substituir
Localizar --
Substituir por *vazio*
Substituir Tudo
Concluir



Também é possível substituir valores inconsistentes por valores consistentes. Por exemplo, se tivermos “Portugal” e “PT” na mesma coluna, podemos substituir “PT” por “Portugal”.


Dados
Substituir
Localizar PT
Substituir por Portugal
Substituir Tudo
Concluir