Ferramente de Apoio à Economia

Aula 2 - Recolher e Preparar dados

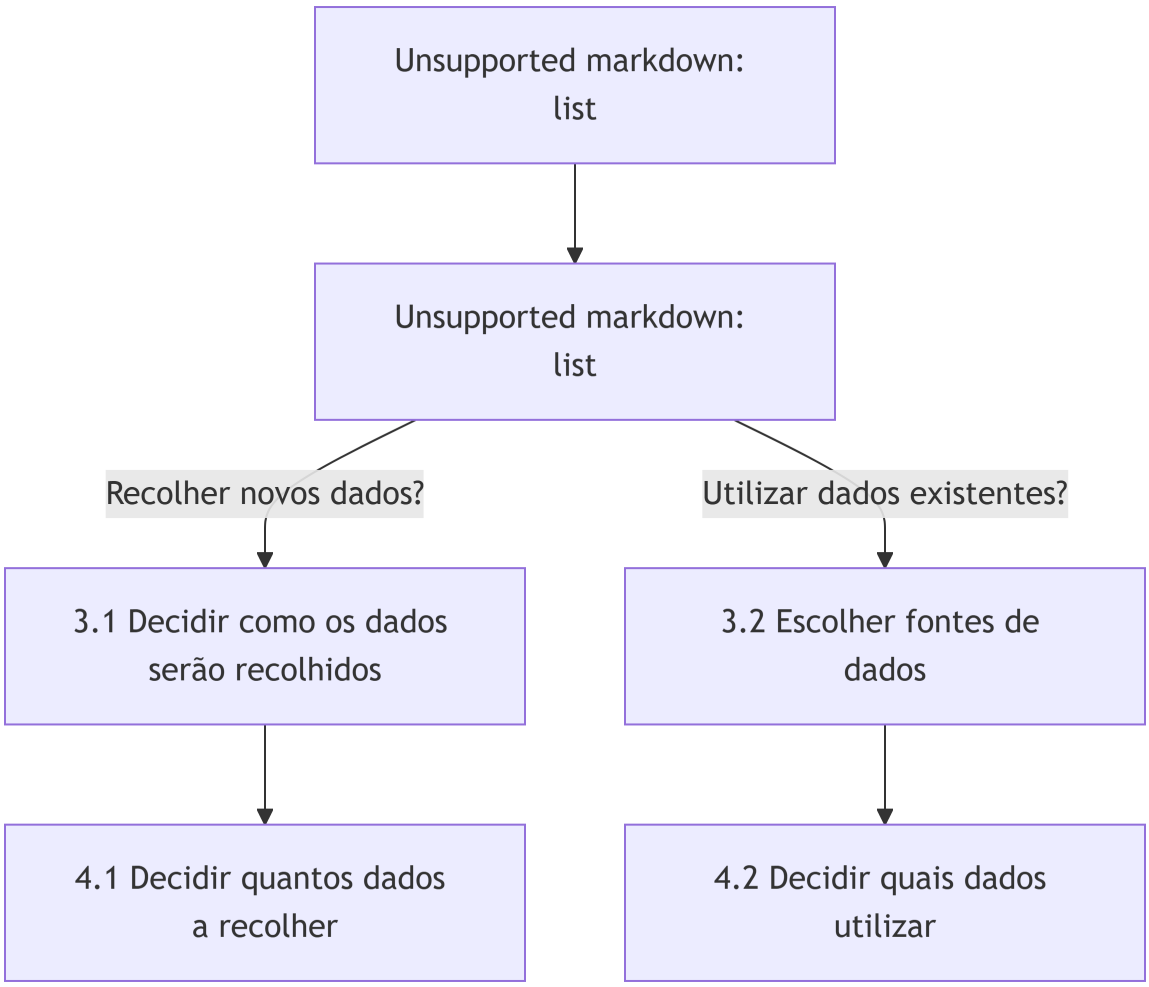
Tiago Afonso

2025-02-25

Table of contents

Considerações sobre a recolhas de dados

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



1. 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/estruturadps** 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.

2. 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, intradiariamente, 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, *inqueritos 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.

3. 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 enviesamnto 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 determindado 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 científicostida*, 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 utilziados em economia:

WDI- World Development Indicators

IMF- International Monetary Fund

OECD- Organisation for Economic Co-operation and Development

Eurostat- Statistical Office of the European Union

BP- Banco de Portugal

INE- Instituto Nacional de Estatística

PORDATA- Base de Dados Portugal

Outras bases de dados

Kaggle- Kaggle Datasets

UCI Machine Learning Repository- UCI Machine Learning Repository

Google Dataset Search- Google Dataset Search

Estrutura de dados

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

1. 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.

2. 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.

3. Estrutura de dados Económicos

Na área da economia, os dados podem ser organizados de diferentes formas, dependendo do tipo de análise que pretendemos realizar. Por exemplo, os dados macroeconómicos são normalmente organizados em séries temporais, onde cada linha representa uma observação ao longo do tempo.

Nos dados económicos existem 3 dimensões:

- **Entidades** $(i = 1, ...n)$ - Indivíduos, empresas, países, etc.
- **Variáveis** (x_1, x_2, x_3, x_j) - PIB, inflação, desemprego, etc.
- **Periodos de tempo** $(t = 1, ...T)$ - Anos, trimestres, meses, etc.

Dados seccionais

Várias entidades ($i = 1, ...n$), varias variáveis (x_1, x_2, x_3, x_j) e um periodo de tempo ($t = 1$)

Os *dados seccionais* são dados recolhidos numa determinada altura e referem-se a uma amostra de indivíduos, empresas, países, etc. Por exemplo, um inquérito de satisfação aos clientes de um supermercado num determinado dia/mês. Exemplo:

Cliente	Idade	Sexo	Profissão	Rendimento
1	25	M	Estudante	1000
2	35	F	Empresária	2000
3	45	M	Médico	3000

Séries Temporais

Uma entidade ($i = 1$), várias variáveis (x_1, x_2, x_3, x_j) e vários periodos de tempo ($t = 1, ...T$)

As *séries temporais* são dados recolhidos ao longo do tempo e referem-se a uma unidade (indivíduo, empresa, país, etc.). Por exemplo, o histórico de vendas de uma empresa ao longo de vários anos.

Exemplo:

Ano	PIB	POP
2022	1000	10
2023	1100	11
2024	1200	12

Dados em painel

Várias entidades ($i = 1, ...n$), várias variáveis (x_1, x_2, x_3, x_j) e vários periodos de tempo ($t = 1, ...T$)

Os *dados em painel* combinam as características dos dados seccionais e longitudinais, ou seja, são dados recolhidos ao longo do tempo e referem-se a uma amostra de indivíduos, empresas, países, etc. Por exemplo, o histórico de vendas de várias empresas ao longo de vários anos.

Exemplo:

País	Ano	PIB	POP
Portugal	2022	1000	10

País	Ano	PIB	POP
Portugal	2023	1100	11
Portugal	2024	1200	12
Espanha	2022	2000	20
Espanha	2023	2100	21
Espanha	2024	2200	22