http://vita.had.co.nz/papers/tidy-data.pdf

Hadley Wickham é estatístico e um dos mais proeminentes membros da comunidade de desenvolvedores da linguagem R, em especial por seu papel na criação de bibliotecas de softwares que tornam mais fácil a vida de quem trabalha com dados. Em um artigo de 2014, ele sintetizou um pouco da filosofia por trás de suas contribuições em um artigo - já clássico - intitulado "Tidy Data".

Wickham inicia o artigo apontando que há poucas pesquisas sobre o processo de limpeza e preparação de dados, apesar desta etapa consumir boa parte do tempo na hora de analisá-los. Seu artigo foca em um pequeno, porém importante, aspecto deste processo: a organização dos dados (data tidying).

Mas o que significa isso? Antes de explicar a ideia de "dados organizados" ou "tidy data", vamos começar por algumas definições básicas.

A maioria dos conjuntos de dados estatísticos são tabulares. Ou seja, são representados por uma tabela, contendo linhas e colunas. No entanto, os mesmos dados podem ser representados de diferentes maneiras. Veja, por exemplo, as duas tabelas abaixo: os valores são os mesmos, mas a estrutura delas é diferente.

[Tabela 1 e Tabela 2 - pag.3]

Assim, um conjunto de dados (dataset) é uma coleção de valores, sejam eles numéricos ou textuais. Por sua vez, cada valor pertence a uma variável e a uma observação. Variável é uma medição de um mesmo atributo em diferentes unidades, por exemplo, altura, temperatura, duração temporal, etc. Já cada observação é uma unidade daquilo que está sendo medido, por exemplo, pessoas ou dias.

A tabela a seguir reorganiza os dados da primeira tabela para tornar os valores, variáveis e observações mais evidentes. No caso, temos três variáveis:

1. pessoa (person), com três valores diferentes (John, Mary e Jane)
2. tratamento (treatment), com dois valores distintos (A e B)
3. resultado (result), com os valores numéricos associados a cada tratamento

[Tabela 3 - pag. 4]

Apesar de ser fácil identificar o que são variáveis e observações para uma tabela específica, é difícil fazer uma definição geral. A proposta de Wickham dos "dados organizados" (tidy data) é justamente fixar um padrão que fornece um elo entre a estrutura dos dados e seus significados.

Este padrão é simples, basta atender a três regras fundamentais:

1. Cada variável é uma coluna;
2. Cada observação é uma linha;
3. Cada tipo de unidade observacional é uma tabela;

Assim, qualquer outro dado que não obedeça a este formato poderiam ser considerados bagunçados (messy data). A Tabela 3 é a versão organizada da Tabela 1.

Na sequência, Wickham lista 5 dos problemas mais comuns na hora de organizar os dados de acordo com estes princípios.

1. Os cabeçalhos (a primeira linha da tabela) são valores, não o nome das variáveis;
2. Diversas variáveis estão armazenadas na mesma coluna;
3. Variáveis armazenadas tanto em linhas, como em colunas;
4. Diferentes tipos de unidades observacionais estão armazenadas na mesma tabela;
5. Uma única unidade observacional está armazenada em diferentes tabelas.

Vejamos cada um destes problemas.

1. Os cabeçalhos (a primeira linha da tabela) são valores, não o nome das variáveis;
[Tabela 4]

Apesar de reconhecer que tal estrutura de armazenamento de dados pode ser extremamente eficiente, a depender do propósito da análise, Wickham mostra que esta forma de estruturar os dados não segue os princípios definidos anteriormente. O conjunto de dados acima possui três variáveis (religião, renda e frequência) e, para ser arrumado, Wickham mostra que ele precisaria ser "fundido" (melt) ou "empilhado" (stack). Ou seja, é preciso transformar as colunas em linhas.

[tabela 6]

Na tabela 6, é possível ver como o conjunto de dados se pareceria neste novo formato.

Outro exemplo é esta tabela com um ranking de músicas. Repare nas últimas três colunas, que representam a semana (wk1, wk2, etc) e a posição no ranking.

[tabela 7]

Nos dados acima, o cabeçalho da tabela contém um valor: o número da semana em questão. Já na tabela abaixo os mesmos dados estão arrumados no formato “tidy”. Agora, o número das semanas é representado como valores de uma variável chamada “week”.

[tabela 8]

2. Diversas variáveis estão armazenadas na mesma coluna;
[tabela 9]

A tabela acima é da Organização Mundial de Saúde. Os registros trazem a contagem de casos confirmados de tuberculose por país, ano e grupo demográfico. Estes últimos estão representados nas colunas e são divididos por sexo ("m" para masculino e "f" para feminino) e idade (0-14,15-25-25-34). Para poupar espaço, nem todas colunas são apresentadas acima.

Na tabela abaixo, vemos estes mesmos dados de forma organizada, onde as variáveis "sexo" e "idade" são decompostas em duas variáveis reais, representadas por colunas diferentes.

[tabela 10b]

3. Variáveis armazenadas tanto em linhas, como em colunas;
Segundo Wickham, esta seria a forma mais complicada, entre os “dados bagunçados”. A tabela abaixo mostra dados de um estação climática no México (MX17004) ao longo de cinco meses em 2010. Há variáveis como colunas individuais (id, ano, mês), mas também há variáveis espalhadas em várias colunas (dia, d1-31) e entre linhas (tmin, tmax - indicando a temperatura mínima e máxima).

[tabela 11]

Para organizar estes dados, a primeira coisa necessária é "empilhar" as variáveis sobre grupos demográficos. Com isso, chegaríamos à tabela da esquerda. Depois, a coluna é dividida em duas: uma para representar o sexo e outra para a idade. Com isso, como mostra a tabela da direita, os dados ficam no formato "tidy".

[tabela 12a e tabela 12b]

4. Diferentes tipos de unidades observacionais estão armazenadas na mesma tabela;
Conjuntos de dados (datasets) muitas vezes envolvem valores coletados em diferentes níveis, com diferentes tipos de unidades observacionais. Os dados abaixo mostram dois tipos de unidades observacionais: a música e sua posição no ranking de cada semana.

[tabela 8]

No formato "tidy", os dados seriam divididos em duas tabelas. Na primeira, temos as informações de cada música, como artista, nome da faixa e duração, além de um identificador único, representado na primeira coluna. Na segunda tabela, ao invés de repetir todas informações de cada música, é inserida apenas o identificador, a data e a posição da música no ranking.

[tabela 13]

5. Uma única unidade observacional está armazenada em diferentes tabelas.
Também é comum encontrar dados sobre um mesmo tipo de unidade observacional espalhados em múltiplas abas, páginas ou arquivos, que muitas vezes estão separadas por alguma variável. Ou seja, casos

onde cada tabela representa um ano, pessoa ou localização. Se o formato for sempre o mesmo, este é um problema fácil de resolver.

1. Insira cada arquivo em uma lista de tabelas;
2. Para cada uma, adicione uma nova coluna que registra o nome original do arquivo. Isto porque muitas vezes, como dito, o nome do arquivo é uma variável importante.
3. Mescle todas tabelas em uma só;

No artigo, Wickham mostra um código em R que realiza esta operação. Nas outras seções do texto,gm apresenta ferramentas para manipulação, visualização e modelagem, além de apresentar um estudo de caso.