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

Hadley Wickham es estadístico y uno de los miembros más destacados de la comunidad de desarrolladores de lenguaje R, especialmente por su papel en la creación de bibliotecas de software que facilitan la vida de quienes trabajan con datos. En un artículo de 2014, sintetizó parte de la filosofía detrás de sus contribuciones en un artículo, ya clásico, titulado "Tidy Data" (data ordenada).
Wickham comienza el artículo señalando que hay poca investigación sobre el proceso de limpieza y preparación de datos, aunque este paso consume mucho tiempo cuando los analizamos. Su artículo se centra en un aspecto pequeño pero importante de este proceso: ordenar los datos (data tidying).
Pero, ¿qué significa ordenar los datos? Antes de explicar la idea de "datos organizados" o "datos ordenados", comencemos con algunas definiciones básicas.
La mayoría de los conjuntos de datos estadísticos son tabulares. Es decir, están representados por una tabla que contiene filas y columnas. Sin embargo, los mismos datos se pueden representar de diferentes maneras. Tomemos, por ejemplo, las dos tablas siguientes: los valores son los mismos, pero su estructura es diferente.

[Tabla 1 y Tabla 2 - página 3]

Por lo tanto, un conjunto de datos (data set)  es una colección de valores, ya sea numéricos o textuales. A su vez, cada valor pertenece a una variable y una observación. La variable es una medida del mismo atributo en diferentes unidades, por ejemplo, altura, temperatura, duración del tiempo, etc. Cada observación es una unidad de lo que se está midiendo, por ejemplo, personas o días.
La siguiente tabla reorganiza los datos de la primera tabla para hacer que los valores, las variables y las observaciones sean más evidentes. En este caso, tenemos tres variables:
1. persona, con tres valores diferentes (John, Mary y Jane)
2. tratamiento, con dos valores distintos (A y B)
3. resultado, con los valores numéricos asociados con cada tratamiento

[Tabla 3 - página 4]

Si bien es fácil identificar qué variables y observaciones son para una tabla específica, es difícil hacer una definición general. La propuesta de Wickham para "datos ordenados" (tidy data) es precisamente para fijar un patrón que relaciona la estructura de los datos y sus significados.
Este patrón es simple, solo cumple tres reglas fundamentales:
1. Cada variable es una columna;
2. Cada observación es una línea;
3. Cada tipo de unidad de observación es una tabla;
Por lo tanto, cualquier otro dato que no se ajuste a este formato podría considerarse desordenado. La Tabla 3 es la versión organizada de la Tabla 1.
A continuación, Wickham enumera 5 de los problemas más comunes en la organización de datos de acuerdo con estos principios.
1. Cuando los encabezados (la primera fila de la tabla) son valores, no el nombre de las variables;
2. Cuando varias variables se almacenan en la misma columna;

3. Cuando las variables están almacenadas en filas y columnas;
4. Cuando los diferentes tipos de unidades de observación se almacenan en la misma tabla;
5. Cuando una sola unidad de observación se almacena en diferentes tablas.
Veamos cada uno de estos problemas.
1. Cuando los encabezados (la primera fila de la tabla) son valores, y no el nombre de las variables;

[Tabla 4]

Si bien reconoce que dicha estructura de almacenamiento de datos puede ser extremadamente eficiente, dependiendo del propósito del análisis, Wickham muestra que esta forma de estructurar los datos no sigue los principios definidos anteriormente. El conjunto de datos anterior tiene tres variables (religión, ingresos y frecuencia), y, para ser ordenado, Wickham muestra que necesitaría "fundirse" (melt) o "apilarse" (stack). Es decir, debe convertir las columnas en filas.

[tabla 6]

En la tabla 6, puede ver cómo se vería el conjunto de datos en este nuevo formato.
Otro ejemplo es esta tabla con una clasificación de canciones. Observe las últimas tres columnas, que representan la semana (wk1, wk2, etc.) y la posición en el ranking.

[tabla 7]

En los datos anteriores, el encabezado de la tabla contiene un valor: el número de la semana en cuestión. Ya en la tabla a continuación, los mismos datos están organizados en el formato "tidy". Ahora el número de semanas se representa como valores de una variable llamada "semana".

[tabla 8]

2. Cuando varias variables se almacenan en la misma columna;
[tabla 9]

La tabla anterior es de la Organización Mundial de la Salud. Los registros muestran el número de casos confirmados de tuberculosis por país, año y grupo demográfico. Estos últimos están representados en las columnas y están divididos por género ("m" para hombre y "f" para mujer) y edad (0-14,15-25-25-34). Para ahorrar espacio, no todas las columnas se muestran arriba.
En la tabla a continuación, vemos estos mismos datos de manera organizada, donde las variables "género" y "edad" se dividen en dos variables reales, representadas por diferentes columnas.

[tabla 10b]

3. Cuando las variables están almacenadas en filas y columnas;
Según Wickham, esta sería la forma más complicada de "datos desordenados". La tabla a continuación muestra datos para una estación climática en México (MX17004) durante cinco meses en 2010. Existen

variables como columnas individuales (id, año, mes) pero también hay variables distribuidas en varias columnas (día, d1-31) y entre líneas (tmin, tmax - indicando la temperatura mínima y máxima).

[tabla 11]

Para organizar estos datos, lo primero que se requiere es "apilar" las variables demográficas. Eso nos llevaría a la tabla de la izquierda. Luego, la columna se divide en dos: una para el género y otra para la edad. Ahí, como se muestra en la tabla de la derecha, los datos están en formato "ordenado" o “tidy”.

[tabla 12a y tabla 12b]

4. Cuando los diferentes tipos de unidades de observación se almacenan en la misma tabla;
Los conjuntos de datos (data sets o bases de datos) a menudo incluyen valores recopilados en diferentes niveles con diferentes tipos de unidades de observación. Los datos a continuación muestran dos tipos de unidades de observación: la música y su posición en el ranking cada semana.

[tabla 8]

En formato "ordenado" o “tidy”, los datos se dividirán en dos tablas. En la primera, tenemos la información de cada canción -como el artista, el nombre de la pista y la duración-, así como un identificador único, representado en la primera columna. En la segunda tabla, en lugar de repetir toda la información de cada canción, solo se inserta el identificador, la fecha y la posición de la canción en el ranking.

[tabla 13]

5. Cuando una sola unidad de observación se almacena en diferentes tablas.
También es común encontrar datos sobre el mismo tipo de unidad de observación distribuidos en múltiples pestañas, páginas o archivos, que a menudo están separados por alguna variable. Es decir, casos en los que cada tabla representa un año, persona o ubicación. Si el formato es siempre el mismo, este es un problema fácil de resolver.
1.- Inserte cada archivo en una lista de tablas;
2.- Para cada una, agregue una nueva columna que registre el nombre del archivo original. Esto se debe a que a menudo, como se dijo, el nombre de archivo es una variable importante;
3.- Fusione todas las tablas en una.
En el artículo, Wickham muestra un código R que realiza esta operación. En las otras secciones del texto, presenta herramientas para la manipulación, visualización y modelado, así como también presenta un estudio de caso.