1. Wickham is interested in taking a more abstract approach to data cleaning, a process that often doesn’t get the attention it deserves. While removing NA values is helpful, what should the data look like to say the cleaning is truly done? He introduces a clear set of guidelines: each row should be an observation, each column should represent a variable, and each type of observational unit should have its own table. He then explores what happens when you follow these rules.
2. Despite how much time and effort data cleaning takes in analysis, it’s still an underappreciated skill. The "tidy data" standard is designed to help make data cleaning easier by giving everyone a shared goal and a clear set of steps to get there. This way, everyone knows what clean data should look like and how to achieve it.
3. The main point here is that messy data can come with unique challenges, while clean data tends to have similar characteristics. You’ve probably seen this yourself—whoever created the data usually organized it for their own purposes, which often makes it hard to work with in R or any other tool. This situation can be compared to Russian literature, which is famous for its complexity. The second idea is that while the concept of a data frame (with rows as observations and columns as variables) seems intuitive, it’s actually a deliberate choice that can benefit the analyst if used thoughtfully. For example, if you have data for counties over several years, is an observation a county or a county-year? Treating it as a county-year might not be obvious to someone new to this kind of data, but it can prevent bad decisions when cleaning and organizing.
4. A dataset is essentially a collection of values, which can be either numbers or categories (like strings). Every value belongs to both a variable and an observation. A variable is a group of values that measure the same thing (such as height, color, or temperature), and an observation is a collection of values that describe a particular item or entity.
5. In tidy data, each variable is a column, each observation is a row, and each type of observational unit is its own table. If data doesn’t follow this structure, it’s considered messy.
6. Messy data can happen in a few ways. One common issue is when column headers are actually values rather than variable names. For example, if you have unemployment data by county and year, the years might be column headers, but those are values (like 2012), not variables like "Year." Another issue is when multiple variables are stored in one column, like dates that include the month, day, and year all in one. Time variables, in particular, often need to be converted into a standard format. Sometimes variables are scattered across both rows and columns, or different types of observational units, like firms and workers, are lumped together in the same table instead of separated into their own datasets. Finally, a single observational unit might be split across multiple tables, which makes the data harder to clean. For instance, in Table 4, the columns are actually values for a hidden variable, income. In Table 6, this gets fixed by creating a separate column for income, alongside other variables like religion and frequency. The process of "melting" a dataset is essentially taking column-value pairs and converting them into rows to make the data tidy.
7. Take Table 11, which lists days along the top as values. Table 12 fixes this by melting the days into a single "date" variable, but it’s still not quite tidy because the element column contains names like tmax and tmin, which are actually variable names, not values. In Table 12(b), the data is tidy because every entry is an actual value, not a variable name, making the structure more consistent.
8. Wickham’s goal is to promote a bigger picture view of data cleaning. If the tidy framework is just about making certain tools easier to use, it risks becoming nothing more than a marketing tool. What Wickham hopes for is a broader concept that trains people not just in using tools like ggplot2 but in thinking more deeply about the overall process of data cleaning and analysis.