1. Read the abstract. What is this paper about?

This paper discusses methods for simplifying the data cleaning process and making it as effective as possible through tidy datasets. Tidy datasets are ones that are far easier to manipulate and model compared to messier data sets. Overall tidy datasets save time from cleaning messy datasets.

2. Read the introduction. What is the "tidy data standard" intended to accomplish?

The tidy data standard is a method of data cleaning that serves to provide a universal way of describing and organizing data. It is intended to save time and make the process easier in order to facilitate data analysis.

3. Read the intro to section 2. What does this sentence mean: "Like families, tidy datasets are all alike but every messy dataset is messy in its own way." What does this sentence mean: "For a given dataset, it's usually easy to figure out what are observations and what are variables, but it is surprisingly difficult to precisely define variables and observations in general."

Since tidy data sets are standardized, they should be universal and share overlapping elements, in effect being consistently easy to model. With a messy data set, anything is fair game. If you have a survey then the same idea could be represented in several different ways(man, guy, dude, masc, M, AMAB all represent men for example). Whereas other data sets could just have an unreasonable amount of missing data. When looking at a dataset the observations in a dataframe are clearly defined as they're under the column for the variable. But as I discussed earlier it is very easy to define the same concept in many ways, and while I may understand what I mean, others may not (for example height/weight vs height/depth). It is less clear what the intent of the variable is.

4. Read Section 2.2. How does Wickham define values, variables, and observations?

Values make up a dataset, they are organized as either variables or observations and act as numbers or strings. Variables contain all values that measure the same underlying units of measurement (height/weight).An observation contains all values measured by the same unit, such as qualities of a person or a day.

5. How is "Tidy Data" defined in section 2.3?

Tidy data is a way of mapping data to a structure that is consistent across data sets. Tidy data is made up of three key criteria. Each variable forms a column, each observation forms a row, each type of observational unit forms a table. All of these serve to create a consistent format for data, any deviation from this is considered messy data. Good ordering can help to contribute to a more tidy form of a dataset.

6. Read the intro to Section 3 and Section 3.1. What are the 5 most common problems with messy datasets? Why are the data in Table 4 messy? What is "melting" a dataset?

The most common issues are that the column headers are variables and not values. Multiple variables are stored in one column, variables are stored in both rows and columns,multiple types of observational units are stored in the same table, and a single observational unit is stored across

several tables The main issue with the table is that there is a variable stored in both rows and columns. Melting a data set involves moving columns into rows where they are better fit as observations rather than variables.

7. Why, specifically, is table 11 messy but table 12 tidy and "molten"?

11 is messy because the element column is not a variable, it stores the name of variables. 12 is tidy because missing data is removed. Table 12 is not tidy in the book, they say it is almost tidy, but a key difference is each row now contains meteorological info for each day, whereas table 11 has days assigned to columns.

8. Read Section 6. What is the "chicken-and-egg" problem with focusing on tidy data? What does Wickham hope happens in the future with further work on the subject of data wrangling?

The focus on tidy data is linked to the tools that are used to create it. If the tools to create tidy data are no longer seen as tidy, then neither is the data considered tidy. As a result changing the tools or structures alone will not speed along the workflow. As a result the author hopes that better tools and storage strategies will develop to allow for more tidy data. Additionally hoping that elements of statistics and human centered design enter the field.