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

   This paper is about creating a clean dataset. Despite Data Scientists spending large amounts of time cleaning datasets to make them ready for analysis, there is little research on how to clean data effectively and efficiently. Based on the abstract, this paper goes into detail about the guidelines for creating tidy data.

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

   The “tidy data standard” is meant to create a standardized data cleaning method. Since data cleaning is understudied, but Data Scientists spend time and effort cleaning data, is it important to ensure the process is standardized. With the “tidy data standard,” it will be easier and more efficient to clean data.

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."

   The first sentence means that each dataset has its own unique properties that have to be cleaned. However, once cleaned all datasets have similar properties.

   The second sentence touches on the idea we discussed in class. Where data is often organized in a fashion that makes it easy to store but not to use. We talked about the example of storing data about a country and years for a specific variable. In this case, an observation would be country-year, however, in practice the data would be separated with country in the rows and year in the columns.

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

   Value: numeric, categorical, or strings. Every value belongs to a variable and an observation.

   Variable: a collection of values that measure the same attribute or property.

   Observation: a collection of variables that measure it.

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

   In “Tidy Data,” each variable is a column, each observation is a row, and each type of observational unit is in a table. If the data is not organized this way (ie not Tidy), it is messy.

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?

   5 most common problems with messy datasets:

   1. Column headers are values, not variables
   2. Multiple variables are stored in one column

3. Variables are stored in both rows and columns
4. Multiple observational units are stored in one table
5. A single observational unit is stored in multiple tables

Table 4 is messy because the columns are variables (income).

“Melting” a dataset is the process of converting column-value variables into rows.

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

   Table 11 is messy because it has the days as the columns, which is incorrect as they are values not variables. Table 12 is “molten” because it melts the days into a variable: date, however, the dataset still isn’t tidy because the element variable contains variable names not values.

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 “chicken and egg” problem is if tidy data is only as useful as the tools that we work with, then tidy tools will be inextricably linked to tidy data. This causes the issue that independently changing data structures or data tools will not improve workflow.

   Wickham hopes in the future that tidy data cleaning isn’t just training people to use the tools effectively, but instead creating a more robust collection of ideas and tools for data cleaning.