Wranging HW Assignment - Question #1

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

This paper is about developing a set of tools and a structure that makes it easy to tidy data sets given the amount of effort needed to do this as a data scientist. The paper says that tidy datasets are easy to manipulate, and it lays out a mechanism to use a select few tools in order to accomplish this.

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

The “tidy data standard” is intended to give the user a process and tools on how to turn messy data sets into “tidy” ones that can be used for further analysis. Having a standard will allow users to align on this step by step process to make data cleaning more efficient, and have tools readily available to complete this.

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 a dataset that is “tidy” under the “tidy dataset standard” will be organized and cleaned in a way that is easily analyzed by the user. While all datasets are different, someone who is familiar with this tidy data standard would have an easier time being able to understand what the variables are, the observations, etc. Every messy dataset is messy in its own way because there are different ways in which the dataset veers off from the tidy data standard, and different things have to be done to get it to align with the standard. Additionally, every messy dataset is unable to be easily analyzed in a proper way.

The second sentence means that you can usually look at a dataset and determine in that specific context what are the variables that encompass certain observations. However, outside of that specific context, to give a general definition of what variables and observations are is quite difficult. The context of a dataset and what it is describing can usually lead to the user to determine what the specific variables and observations are.

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

Wickham defines values as numbers (if quantitative) or strings (if qualitative) that belong to a variable and an observation. He defined a variable as “contain[ing] all values that measure the same underlying attribute across units” and an observation as “contain[ing] all values measured on the same unit.” In the example, “person,” “treatment,” and “result” are the variables and the observations are the crosses between the people and treatments.

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

Tidy data is defined as a fixed structure of how data is arranged with every other structure being “messy” data. According to the text, in this structure each variable forms a column, each observation

forms a row, and each type of observed unit forms the table. The order of these elements doesn't matter, but a good order is easier to use.

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 5 most common problems with messy data sets are: 1) the column heads aren't variable names (they are values), 2) there are multiple variables in 1 column, 3) variables are stored in rows and columns, 4) multiple types of observed units are in the same table, and 5) a single observable unit is stored in multiple tables.

The data in Table 4 is messy because the column headers are values, not variables. While the dataset only has 3 variables (religion, income, and frequency), the column headers are religion, and then various values of ranges of income. To make it tidy, the user will "melt" the dataset which basically stacks the data and turns columns into rows.

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

Table 11 is messy because there is a column for the year, month, and each possible day in the month. Table 12 is "molten" because all of those columns have been combined into 4 variables: id, date, tmax, and tmin (measured variables). Additionally, the missing values were dropped to conserve space. The final table represents a "tidy" structure and can be used for further analysis.

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 with focusing on tidy data is that you need to improve both data structure and data tools for real improvement (both are linked to one another), just like a chicken is linked to the egg. Wickham referred to reaching and needing to break out of a "local maxima." In the future, Wickham hopes that better data storage strategies and tools are developing, and hopes to use methods from human factors and user-centered design to improve these tools. Additionally, he wants to look into other formulations of tidy data. The example he used was finding a formulation that looks at multidimensional arrays. And finally, he would like to develop other frameworks for tasks related to data cleaning like identifying missing values in a dataset.