James Kulp

Wrangling Assignment Questions 1 and 3

Question 1

- Part 1) The abstract states that the ensuing article is going to deal with data cleaning, specifically tidy data, which is a form of data that makes manipulation and visualization much easier. There is a framework for which you can convert messy data to this simpler version using a smaller set of tools to hopefully circumvent the usually prolonged process of data cleaning.

- Part 2) The tidy standard is meant to standardize what is a very time consuming but understudied portion of the data science process, the idea is making the process more streamlined and easier

- Part 3) The statement relating families to datasets is meant to describe the way in which both datasets and families have some underlying structure that traverses all types but each individual family or dataset also has unique qualities. The second statement encourages the reader to look more closely at the data frame itself and consciously choose what constitutes an observation and a variable. Columns in data imported from elsewhere are often named with titles that are too convoluted or not explanatory enough

- Part 4) Values are categorical strings or numeric inputs. When these values are compiled they constitute a dataset. One variable is a collection of values measuring the same thing while an observation is a collection of the values from variables pertaining to one entry.

- Part 5) Each variable is a column and each observation is a row and each type of observational unit is a table, if it’s not in this form then the data is not yet in a tidy format.

- Part 6) The five most common problems with messy sets are improper column titles, namely not variables, time variables being normalized to an arbitrary zero value, one column has multiple variables contained in it, the same table houses multiple types of observation units, or a single observation being stored in multiple tables. Melting a dataset is converting column value variables into rows. In table 4 columns are values of income in reality but since it's already a variable you need a new column.

- Part 7) Table 11 is messy because it has sayings along the top which are supposed to be values not variable columns. In table 12 they take these dates and melt them into the date variable.

- Part 8) Wickham is hoping to use the tidy framework to create a standard across the industry and make a universal system for the entire data science system. He wants the tidy system not to just help with making a single dataset more compatible with visualization techniques but to revolutionize the difficult and time consuming task of cleaning data

Question 3

- Part 1) Race is to check all that apply and they also include more specific origin countries as options. They don't collect gender and sexual orientation information in the census.

- Part 2) We gather these data to have a robust sense of the population demographics of the country. It's important to help give background information on the population that tracks trends and might inform changing political choices for example. The quality of census data is crucial to make sure we have accurate data regarding our countries population so that policy makers or curious users can draw accurate conclusions about it and not mislead the potential audience for their work. The data might be used to make cases to help advocate for inclusion of various races or to whom to help with government resources, how realistic social security benefits are to continue, etc.

- Part 3) The census was conducted in such a way that older versions do not have such a wide variety of choices which makes it difficult to glean data over multiple years and observe trends more effectively, however it could be argued that allowing more detailed information about country of origin adds to the detail of the census. The more important issue is the framing of questions, there's an entire course offered at UVA about how to conduct surveys and to preserve the integrity of the data so you don't allow respondents to answer in a manner what is subjective and not congruent with the rest of the data collected to avoid having unusable observations.

- Part 4) Sex is a binary choice Male/Female. The gender questions are an example of guarding against the above by not having options such as prefer not to say or other might result in non usable values but in doing so the trade off is less detailed information given the growing number of the population no longer identifying as a binary option.

- Part 5) While the census is trying to avoid null responses with the sex variable, they don't specify at birth or how they are currently identifying. This might lead to inaccurate interpretations of the data from the audience depending on what they perceive the response to refer to. When it comes to race, checking all that apply might lead to respondents identifying under a race to which they have only a slight percentage of

genetically speaking or the opposite and neglect parts of their race that they do not identify as, thereby losing the nuisance the options were designed to capture.

- Part 6) I'd be concerned about people cleaning data and aggregating certain country of origin designations to a larger continent: ie European instead of Irish descent which is more specific. In terms of the sex variable I'm not so sure cleaning is an issue so much as respondents inputting information without a universal understanding of what the question was looking for.