Brendan Kilduff

Q1.

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

The paper is about data tidying, which is a method used to make data cleaning as easy and efficient as possible. Tidy datasets have a specific structure making them easy to manipulate, visualize, and model.

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

The tidy data standard is intended to make initial data cleaning easier and facilitates the initial analysis of the data. It also helps simplify the creation of other data analysis tools, allowing tools to work well with each other.

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 talks about the inherent differences between tidy datasets. Even though their structure is standardized, their observations and variables will always have different meanings.

The second sentence describes the notion that it is easier to describe functional relationships between variables than between rows. In addition, this sentence implies that it is easier to make comparisons between groups of observations rather than groups of columns. For instance, the readings describe the differences between the variables height and weight versus height and width. Height and weight would most likely be clear cut variables while height and width could imply some sort of dimensional variable.

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

Values are defined as strings (qualitative) or numbers (quantitative) that belong to both a variable and an observation.

Variables are characterized as containing values that all measure the same attribute across units. Examples include: height, weight, temperature, income, etc.

Observations contain all values that all measure the same unity across attributes.

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

Tidy data is defined as a standardized way to structure a dataset with its meaning. In a tidy data set:

- Each variable forms a column
- Each observation forms a row

Brendan Kilduff

- Each type of observational unit forms a table.

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 datasets include:

1. Column headers are values, not variables
2. Multiple variables are stored in one column
3. Variables are stored in rows and columns
4. Multiple observational units are stored in the same table
5. A single observational unit is stored in multiple tables

The data in table 4 is messy because there are 3 variables that form tabular data where variables form both the rows and columns and column headers are values.

Melting a dataset refers to the process of turning columns into rows. It involves turning the initial list of columns into a single variable and adding two new variables

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

Table 11 is messy because there are variables stored in both rows and columns. For instance, the variables id, year, and month are individual columns while day, d1-d31 are spread across columns. Table 12 is tidy and molten because the data has been rearranged so that one row represents a single day and each column represents a variable.

8. Read Section 6. What is the "chicken-and-egg" problem with focusing on tidy data?

The issue revolves around the fact that tidy data is only as useful as the tools that work with it. This can cause a local maxima effect where changes in either data structures or data tools won’t improve workflow.

9. What does Wickham hope happens in the future with further work on the subject of data wrangling?

In the future, Wickham hopes to use methods from fields such as user-testing and ethnography to improve the cognitive side of data analysis. He also supports the development of new frameworks to improve other data cleaning tasks such as parsing dates and numbers.

Brendan Kilduff

Q5. Many important datasets contain a race variable, typically limited to a handful of values often including Black, White, Asian, Latino, and Indigenous. This question looks at data gathering efforts on this variable by the U.S. Federal government.

1. How did the most recent US Census gather data on race?

Data on race was collected through a self-identification process where citizens selected all of the following races that applied to them. These races include: White, Black or African American, American Indian, or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander)

2. Why do we gather these data? What role do these kinds of data play in politics and society?

Race data plays several roles in politics and society. These roles include:

- Redistricting: To prevent gerrymandering, race data is used to redraw electoral districts to ensure representation.
- Policy Formulation: Race data helps policymakers implement legislation aimed at addressing the needs of specific racial or ethnic groups. Public health intervention and policies also need race data to monitor the health disparities between different groups.
- Resource allocation: Race data plays an important role in understanding where additional resources for education, housing, and healthcare are needed.
- Equal employment: Race data allows for the monitoring and enforcement of equal employment opportunities under the Civil Rights Act of 1964.

3. Why does data quality matter?

If race data is incomplete or inaccurate, it can have serious effects on the economic and political state of different communities. Many policies aim to reduce the disparity between races and if this data is inaccurate, so will the policies that become passed. If some racial groups are underrepresented or overrepresented, unfair political representation, health disparities, resource misallocation, and inequitable policy formulations will follow.

4. Please provide a constructive criticism of how the Census was conducted: What was done well?

The 2020 census marked the first time that everyone had the opportunity to respond online. Invitations were sent out by mail, online, or by phone. One criticism is the inherent non-response bias that comes with a census, especially in 2020 with the pandemic. Young adults are less likely to respond to the census, causing nonresponse and undercoverage to be more prominent for these demographics. However, for being in a pandemic, the census was able to succeed in reaching out to people online, preventing some selection bias. In addition, the inclusive, self-identification question design succeeds in allowing citizens to have multiple responses.

Brendan Kilduff

5. What do you think was missing? How should future large scale surveys be adjusted to best reflect the diversity of the population? Could some of the Census' good practices be adopted more widely to gather richer and more useful data?

In the 2020 census, six states suffered undercounts of more than 5%. In addition, people of color were undercounted at disproportionate rates. Overall census response rates have remained stagnant. To best reflect diversity in the population, the census needs to be better by modernizing data collection. One idea could be removing statutory limits on data collection methods. This could help increase underrepresented demographics in the census. I believe that having an increasing online presence could continue to be helpful if they were adopted more widely.

6. How did the Census gather data on sex and gender? Please provide a similar constructive criticism of their practices.

The 2020 Census gathered data on sex with only two options, male or female. In addition, they had an optional question about gender where respondents could write-down whatever gender they identified as. One criticism is the lack of inclusivity in the sex question of the census that doesn’t account for individuals that don’t consider themselves either male or female. This can cause response bias and affect potential policy implementations.

7. When it comes to cleaning data, what concerns do you have about protected characteristics like sex, gender, sexual identity, or race? What challenges can you imagine arising when there are missing values? What good or bad practices might people adopt, and why?

One concern regarding protected characteristics and cleaning data is the privacy and transparency concerns. If these data cleaning practices are not mentioned explicitly to the public, bias and accuracy will be questioned. When there are missing values, that means that there is non-response bias present. The challenge is figuring out what demographic is not responding to particular questions to better avoid the bias from the data. One bad practice people might adopt is biased, arbitrary data cleaning. If there is bias in not only the data but also how it is cleaned, the accuracy of the census decreases greatly.

8. Suppose someone invented an algorithm to impute values for protected characteristics like race, gender, sex, or sexuality. What kinds of concerns would you have?

One concern I have is using these qualitative variables by nature in an almost quantitative way. There could be underlying biases in what numbers are assigned to different races, genders, etc.