Q4. 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?

The 2020 US Census gathered data on race through asking individuals to self identify. From census.gov: “The data on race are based on self-identification and the categories on the form generally reflect a social definition of race. The categories are not an attempt to define race biologically, anthropologically, or genetically. Respondents can mark more than one race on the form to indicate their racial mixture.” The question itself on the US Census lists around a dozen pre-defined races along with an “other” category for individuals to put down what otherwise was not listed.

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

The US Census is taken every 10 years to get an accurate picture on the citizens and non-citizens currently residing in each household in the country. This data helps influence politics and public life– through allowing the government to appropriately apportion representatives, districts, funding, planning, and emergency response systems. The data quality matters because accurate, representative data allows for fairer decision-making– letting resources be given based on actual needs, not assumptions or biases.

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

The 2020 Census made changes to the self-identification portions allowing individuals to be more accurately represented. For example, separate questions were asked to gauge if the individual was of Hispanic origin or of race, where previously this distinction was not made. Additionally, as this census could be done digitally, it had an overall more accessible platform. However, what I would criticize about the census is that it specifically gathers data on each household. Likely, the homeless and immigrant populations were extremely underrepresented. Additionally, the gender (and lack of sexuality) portions did not reflect the nuances present in today’s society. For future censuses, I would recommend attempting to reflect the current society’s range of labels and categories. Additionally, I believe that there should be a greater attempt to reach out to underrepresented groups. I think the best thing about the US Census that should be adopted more broadly is transparency. I want to know how, where, and why you collected your data.

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

As mentioned a bit above, you are told to mark off the sex you were at birth (male or female). Already, this is an incorrect question. Those who are both intersex or with chromosome patterns that do not match with the typical ‘XY’ or ‘XX’ pattern have been excluded from this section entirely, forced to pick the gender role they were most closely brought up with compared to their actual, genetic sex. The US census in 2020 did not collect any data on gender or sexuality.

5. 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?

Lots of protected characteristics will likely be either left blank by those who are “outside the norm” or, possibly, individuals may purposefully misclassify themselves to the “norm.” Likely this is due to the stigma and discrimination that comes with identifying outside of “societal standards.” So, with this in mind, it is likely that there will not be an accurate representation of minority groups in these fields- lots of blank responses or misidentifications. Cleaning this data will be difficult– do you fill in the “norm” or attempt to match them to one of the minority groups they may have chosen to not identify as? Do we leave these people out entirely? I can see a lot of people adopting the “just out the most popular answer in the NaNs” or “fill the NaNs to match the distribution of the non-NaNs” which could, if a larger portion of the minority groups chose to leave the questions blank, not accurately portray the actual population.

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

If I had to bet, I would say that there is likely no survey that has ever been done on a large enough population that accurately portrays the intricacies or gender, race, sex, and sexuality in such a way that no population is underrepresented. Because of this, using an algorithm to impute values for these characteristics will certainly be biased and massively incorrect– overrepresenting and underrepresenting those groups in which are already over and underrepresented. Additionally, it feels ethically immoral to assign someone a value in these protected categories if they’ve intentionally left these categories blank– especially so if that person left it blank because of some societal concern.