Missing data are defined as not available values, and that would be meaningful if observed.

missing sequence, incomplete feature, files missing, information incomplete, data entry error etc.

**missing completely at random**,

c. I think the data is missing values because most datasets don’t possess all the meaningful available values. There is also a large window for what counts as “missing data,” as it could be anything from missing a file or variable to a simple data entry error. By the design of big data, the data that’s missing in a dataset can be described as missing completely at random, so we don’t know exactly what variable or feature is missing or has an error, but we assume there to be some flaws in the data. In this dataset, I think there could be flaws in the census collection because often, census’s yield two types of errors: inclusion of people who should not have been included and omission of people who should have been included. False reporting could have led to inclusion of people in demographic categories that they didn’t truly belong in. Even if people didn’t deliberately false report, the demographic categories are too broad to encapsulate the true nationalities of each person, as a large amount of people come from mixed backgrounds. The dataset strived to account for this with a “Two or more races” category, but it would be a more descriptive dataset to know their mixed races. It’s very possible that some people in each state were missed by the census because they were out of town, or their home was too remote to be located. Maybe these people were excluded from the “No Response” category because they were completely undetected. For example, I could imagine it would be difficult to complete a census for Alaska because the population is very dispersed. Therefore, I am under the impression that the data is incomplete.

d. You can tell a broader story on diversity in America with the dataset. We can compare the states’ diversity with the data in census\_cleaned. We can also compare the diversity of students taking the ap test by state from ap\_census. We strived to make these comparisons with our function pct\_calc, which created a dictionary of proportions of each demographic category in each state out of the total population of that state. We could make further conclusions from these dictionaries such as which state has the highest or lowest proportion of each demographic category to reveal the diversity in that state. We could also set a numeric benchmark to define diversity and create an if statement with the benchmark as the criteria to yield a list of the states that have sufficient diversity and which don’t. These are just a few ways to use pct\_calc to draw conclusions about the diversity levels across states in America.

We could also make conclusions about diversity in education by evaluating the dictionary returned from pct\_dif, which gives the absolute value of the difference in proportion of student test takers in a state for a demographic group and the proportion of people in that demographic living in that state. The higher the difference, the less that demographic is represented in taking the AP test. The results of this dictionary could reveal which demographic groups per state are most involved in STEM. We observe that most states have the lowest difference in white students, so they are the most represented on the AP test. It could also reveal which demographic groups are excelling at school because taking AP classes is an indication of academic excellence. We could create research questions to explore the results of this dictionary outside the dataset to explore *why* certain groups make up a high percent of the state population and make up a low percent of AP test takers. max\_min helped evaluate the states with the most and least differences in test takers and demographic population. It helps tell the story of diversity in America by showing which states have the most and least proportion of AP test takers per state. To reveal which states have the most and least diversity, we could count how many times each state appears in the max key and min key. This would help reveal state’s culture and how diversity is represented throughout the nation. Overall, this dataset helps reveal diversity levels across the nation for the AP test. We are limited in making conclusions because we don’t know *why* certain states have more/less diversity and it would be more revealing on the climate of the nation to have more factors to draw conclusions on. For example, it might be helpful to know the average AP test grades of each demographic to assess whether overall academic performance is an indicator of high AP test-taking for a demographic, as predicted.