Checking boxes: Exploring how race and ethnicity are measured

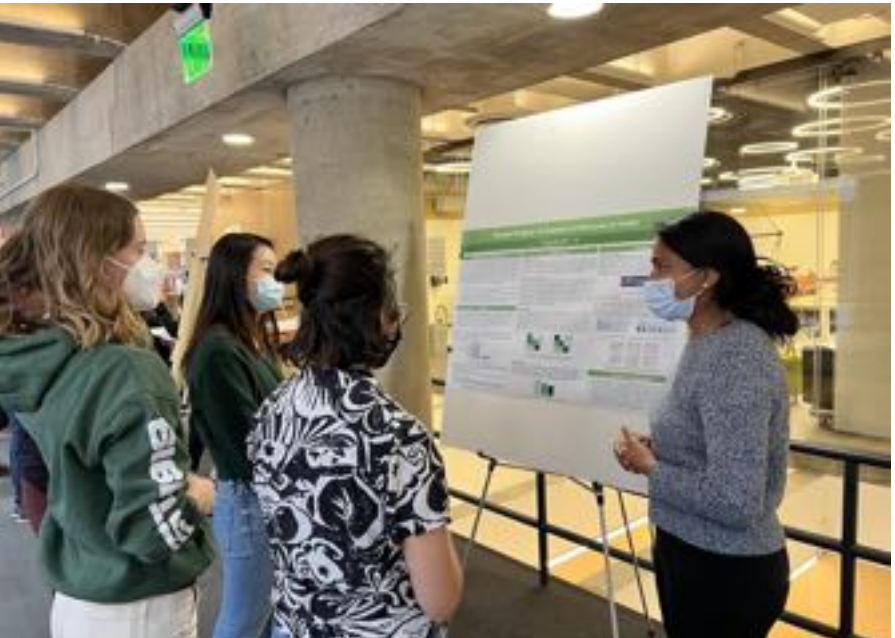
Cassandra Pattanayak
Wellesley College
eCOTS 2022

#ecots22-2-diversity-inclusion-social-justice-in-data-science-and-statistics

Wellesley is a private liberal arts college for women

51% identify as students of color

19% first generation college students



Second-level applied statistics and data science course

Brings together students who took any intro stat course

Students' majors include data science, computer science, economics, math, neuroscience, political science, psychology, sociology, plus statistics minors

Exposes students to parametric and non-parametric tests and robustness to assumptions; linear regression and regression trees; missing data; data cleaning; data ethics; visualization

Effort to tie course to relevant and current ideas

Fall 2021:

Spread over an auditorium twice a week (mostly lecture)

Outside in a tent once a week (mostly small-group activities)

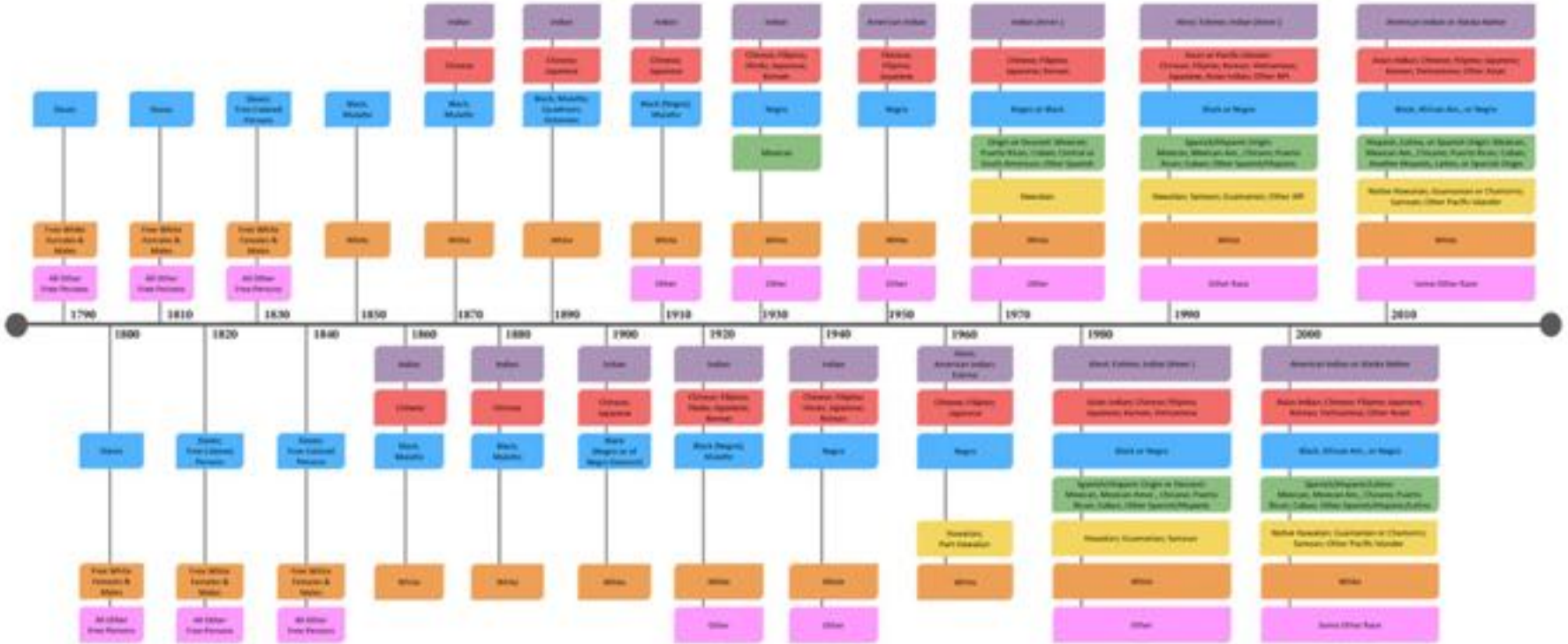
Check all that apply

- ❑ Asian
- ❑ Black
- ❑ Hispanic
- ❑ Native American
- ❑ White
- ❑ Other: _____

?

Measuring Race and Ethnicity Across the Decades: 1790–2010

Mapped to 1997 U.S. Office of Management and Budget Classification Standards

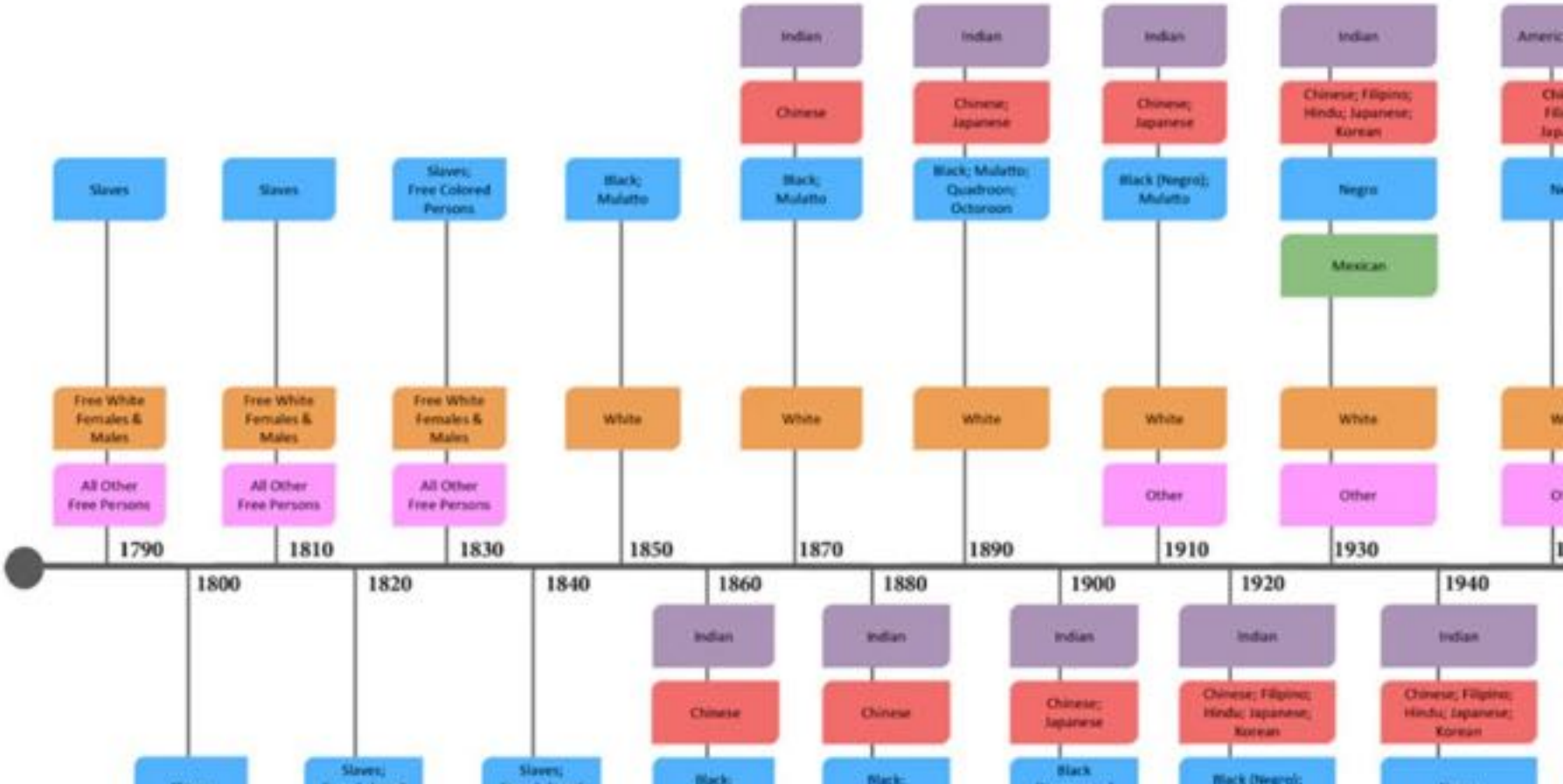


Gibson, Campbell, and Kay Jung. 2002. "Historical Census Statistics on Population By Race, 1790 to 1990, and By Hispanic Origin, 1790 to 1990, For The United States, Regions, Divisions, and States."
Humes, Karen, and Howard Hogan. 2009. "Measurement of Race and Ethnicity in a Changing, Multicultural America."
Humes, Karen R., Nicholas A. Jones, and Roberto R. Ramirez. 2011. "Overview of Race and Hispanic Origin: 2010."
Office of Management and Budget. 1978. "Statistical directive no. 15: Race and ethnic standards for federal agencies and administrative reporting."
Office of Management and Budget. 1997. "Revisions to the standards for the classification of federal data on race and ethnicity."
U.S. Census Bureau History Questionnaires. (2014, March 31).

Interactive image source: https://www.census.gov/data-tools/demo/race/MREAD_1790_2010.html

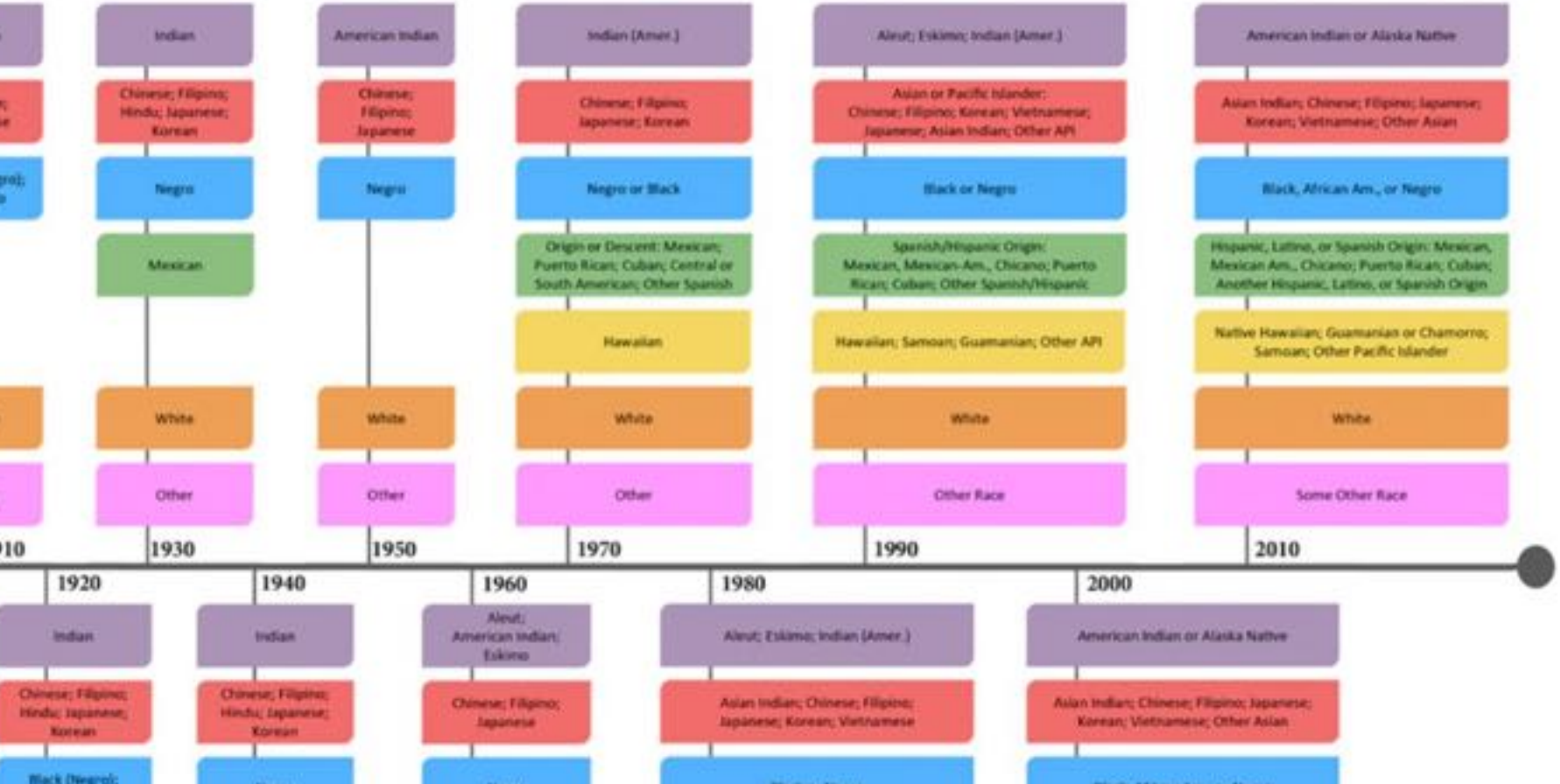
Measuring Race and Ethnicity Across t

Mapped to 1997 U.S. Office of Management and Bu



nicity Across the Decades: 1790–2010

Management and Budget Classification Standards



How the 2020 census asked about Hispanic origin and race

→ **NOTE: Please answer BOTH Question 6 about Hispanic origin and Question 7 about race. For this census, Hispanic origins are not races.**

6. Are you of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *Print, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

7. What is your race?

Mark ☒ one or more boxes ***AND*** *print origins.*

- ☐ White – *Print, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc.*
- ☐ Black or African Am. – *Print, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.*
- ☐ American Indian or Alaska Native – *Print name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.*

- ☐ Chinese
- ☐ Filipino
- ☐ Asian Indian
- ☐ Vietnamese
- ☐ Korean
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Samoan
- ☐ Chamorro
- ☐ Other Asian – *Print, for example, Pakistani, Cambodian, Hmong, etc.*
- ☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*

- ☐ Some other race – *Print race or origin.*

Source: Census Bureau paper questionnaire.
"Black and Hispanic Americans See Their Origins as Central to Who They Are, Less So for White Adults"

PEW RESEARCH CENTER

Exam Question that was later used for small-group discussion:

The U.S. Census allows people to identify themselves by checking multiple race categories. Separately, the Census asks about Hispanic origin. The format of these questions is shown.

Suppose that you are asked to create a simple visualization for a magazine that compares the racial makeup of MA to the racial makeup of NY, using Census data.

Clearly and concisely explain one way to define and visualize the race categories.

Compare to others' suggestions. What are the pros and cons?

Most responses did not initially include a way to deal with people who checked multiple races. The students did propose ideas for how to deal with the fact that there are two questions.

Summary of an ethnicity variable from a data set that was familiar to students from a homework

<u>Questions for exam/discussion/implementation:</u>

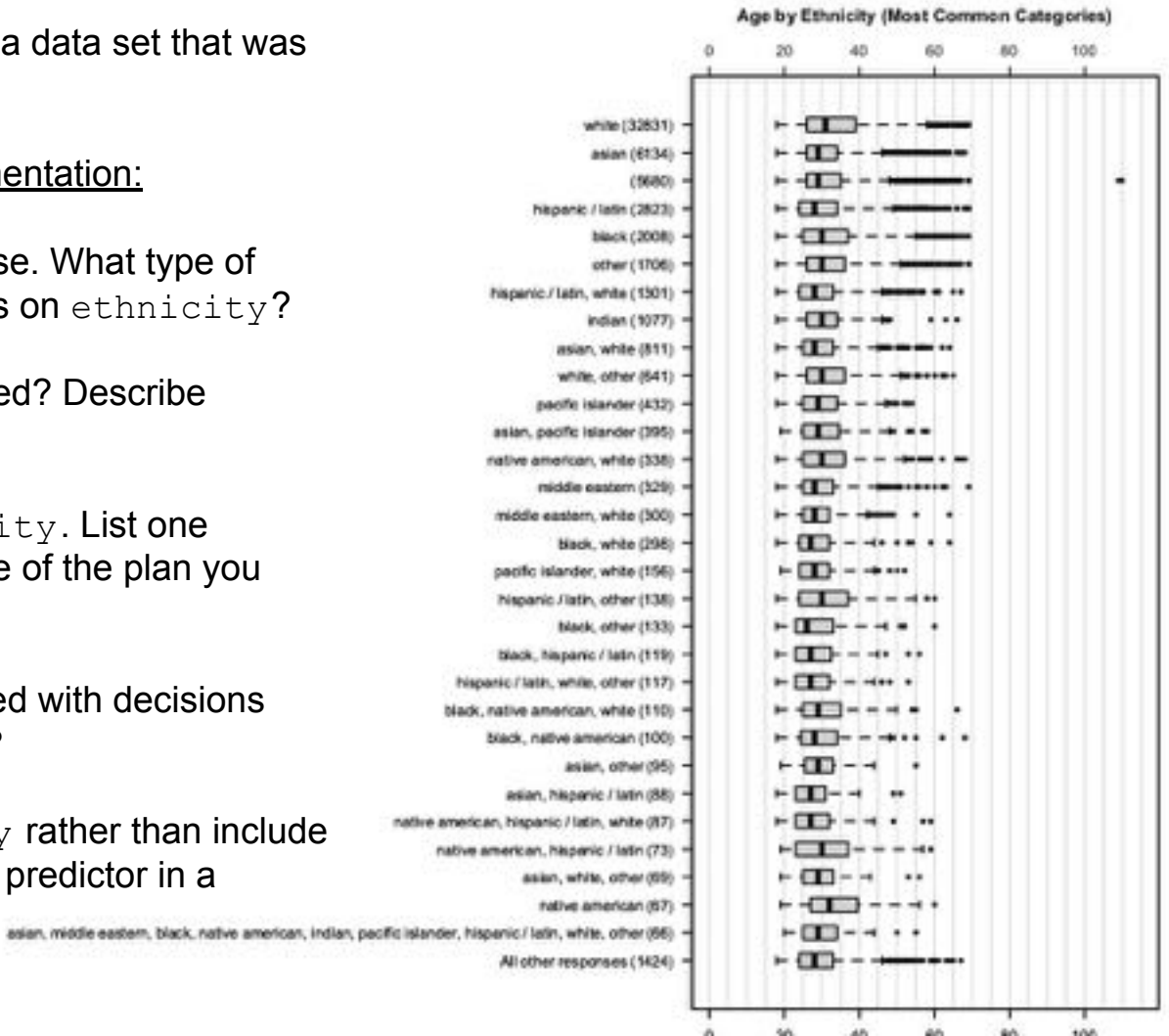
1. Not everyone provided a response. What type of missingness do you think there is on `ethnicity`?

2. How is missingness being handled? Describe another way to handle it.

3. Suggest a way to clean `ethnicity`. List one advantage and one disadvantage of the plan you suggest.

4. What ethical issues are connected with decisions about how to clean this variable?

5. Why would we clean `ethnicity` rather than include the raw categorical variable as a predictor in a model?



Funding opportunity that opened in fall 2021

National Science Foundation

National Center for Science and Engineering Statistics

Improving Surveys of the Science and Engineering Enterprise

Broad Agency Announcement

- Asking about Sexual Orientation and Gender Identity in the Spanish-Speaking Community
- Developing a Single Race Question Appropriate for the US Population
- STEM Workplace App
- Predicting the “Minimal” Survey Respondent

Excerpt from call for proposals:

“The most desirable outcome would be to combine the two questions into one race question with more inclusive categories. However, there are considerations as to how people prefer to classify themselves and how this differs by factors like age, level of education, and recency of immigration (or family’s immigration) to the United States. Additional sociological factors may influence self-identification, such as family influences and composition of the communities in which the respondent lives, works, socializes, etc. Context of the data collection is also a concern—do respondents identify differently depending on who is asking (such as their employer) or sponsoring the survey (e.g., the federal government), and the topic of the survey questions?”

Last day of class activity

Prompt:

1. Come up with at least one suggestion for improving the race/ethnicity question. What are the pros and cons?
2. With this example in mind, what stands out to you about the data ethics articles you read in this course? What will you remember?
3. What decisions did you make as you worked on your final project that were subjective? For hypothesis tests or models? Cleaning? Visualization?

Multiple groups landed on this question:

Why is it even necessary to categorize people by race/ethnicity?

(Is this a statistical question? Maybe! If you don't/can't measure an idea or a problem, does it exist?)

1. What is your preferred way to clean race/ethnicity data that is formatted as in the Census questions? What concerns do you think your students might raise about your preferred approach?

2. Which would be more challenging for your students: designing a race/ethnicity survey question or cleaning and summarizing race/ethnicity data? Why?

3. Which of the statistics/data science topics mentioned are most relevant for the students in a class you teach? We mentioned missing data, data ethics, visualization, data cleaning, model fitting…

4. Can you think of examples from your own teaching that have inspired discussions along these lines, about race/ethnicity or about simplifying complex data?

Topics

Data cleaning

Data ethics

Visualization

Survey question design

Missing data

Overfitting/interpretability when a variable has many categories

Inherent subjectivity of data collection/analysis

…

Examples of survey questions that motivate these topics

Race/ethnicity

Gender identity

Sexual orientation

Occupation

Education

…