Attitudes Towards Health Related Inquiries Across Demographics

DS 3001: Foundations of Machine Learning

University of Virginia

Eric Nguyen, Elaine Zhang, Hieu Vu, Cheryl Bai, Ashley Luk,

Adam Chow, Bryant Chow

Fall 2023

SUMMARY

In this project, we are exploring the question: *How do attitudes towards health-related internet usage vary among different demographic groups (age, location, education, race, etc.)?* After reading in the dataset from the GSS dataset, we first narrowed down the dataset to only include 2022 data. After this step, we were able to rename confusing variable names for clarity and drop all of the unnecessary columns that were filled with NaN values. Lastly, we decided it would make graphing easier to convert the “age” column to numeric and rename all of the remaining miscellaneous NaN values to unknown.

In our EDA process, we first identified the response variables that we wanted to test versus the exploratory categories in order to create sets of kernel density plots and count plots. Out of our six demographic variables, Age was our only continuous variable, leaving us with the ability to make the age variable a histogram with a kernel density plot overlaid. For the other five variables, we were able to create a for loop to produce a graph for each explanatory variable, given the different response variable. Because we had so many unknown values, we decided to take out these values to pinpoint our focus on the crux of the data. With the large number of unknowns, some of the patterns were either obscured or nonexistent, particularly in the income graphs. To the eye of the viewer, graphs without these large unknown values provide clarity when interpreting the data. For instance, it is much more apparent now that more frequent health website usage among higher income groups. While some demographic patterns emerged, especially around income, the prevalence of unknowns limited sharp conclusions.

Moving to our visualization, Group 2 primarily wanted to look at this to see if there were any correlations that someone who trusts the internet's health information would be more likely to use it. Furthermore, analyzed it by different demographics to get more insight into how different race or sex plays a role in answers. A heatmap of belief in reliability vs. visit frequency showed the most common response was visiting health sites several times a year while agreeing it's difficult to distinguish reliability. However, when breaking the counts down into percentages, separating by sex did not have a noticeable impact on this pattern. Respondents also largely agreed the internet helped explain doctor's advice and gauge symptom severity, though most found it difficult to distinguish reliable from unreliable information. While some demographic patterns emerged around income and frequency, reliability concerns were widespread. A heatmap of visit frequency and reliability by sex showed a similar pattern for both genders. Contrarily, looking at the attitudes towards health-related internet usage among different demographic groups revealed intriguing patterns. For Whites, their responses aligned closely with overall trends, showing consistent beliefs, especially in the "Several times a year" and "Agrees" categories, mirroring the general population. In contrast, Blacks and individuals categorized as "Other" displayed a distinct pattern. Their responses were scattered more evenly instead of being

concentrated between the "Several times a year" and "Agrees," indicating less defined beliefs about internet health information reliability. When examining degree levels, the graphs displayed that almost all of the degree categories exhibited uniform views, whereas individuals with less than a high school degree stood out significantly. Merely 1% in this group trusted online health information, and over 35% remained neutral, emphasizing the need for tailored health communication strategies addressing these disparities in trust and beliefs.

DATA

The data for our research question comes from the 2022 General Social Survey (GSS), which contains demographic information and attitudes on various topics for U.S. respondents. The initial dataset included variables from multiple years, but as a group, we decided to narrow the focus of the analysis only on the 2022 responses for consistency and to reflect the most current opinions on health website usage.

In this dataset, the key variables that we examined included demographic factors like age, sex, race, income, education level, and health status. Variables related to frequency of visiting health/lifestyle websites, medical sites, vaccination info, anxiety/stress sites were analyzed as healthy_lifestyle_visit_frequency, health_medical_visit_frequency, vaccination_visit_frequency, and anxiety_visit_frequency. Additionally, attitudinal variables that gauged agreement with statements about the internet's role in health decisions proved important: internet_affect_rating, internet_helped_explain_doctor, help_diagnose_serious_symptoms, help_check_doctor_advice, and is_reliable.

In order to read this data into Colab, there were several steps we had to take beyond the typical pd.read_csv statement. First, a curated list of specific variable names, denoted as `var_list`, was created. This list served as a filter, ensuring only the relevant data was retained, a step that required a deep understanding of the research objectives. Subsequently, we then designated an output file path, referred to as `output_file`. This location became the digital repository where the amalgamated data would be stored, representing a critical decision for our group, as it influenced data accessibility and future analyses.

The process that we chose to read in the data also demanded a keen understanding of file handling modes, distinguishing between writing the initial chunk and appending subsequent chunks. Defining precise write and append modes was a nuanced task for our group. The iterative loop through the 37 raw data chunks hosted on specific GitHub URLs was a formidable challenge. Each chunk was meticulously loaded into a dataframe using the `pd.read_csv()` function, necessitating careful consideration of file formats and ensuring data integrity during the import. Once loaded, each data frame was scrupulously subsetted to retain only the columns specified in `var_list`, as any other extraneous variables we determined would not add that much value to our future analysis.

Upon the completion of the looping process, the concatenated dataframe emerged from the specified output file path. This final step was both rewarding and challenging, as it represented the culmination of detailed planning and a fundamental understanding of the dataset's intricacies.

Now, preparing the data for the analysis of our research question involved extensive renaming and cleaning to make variables more interpretable. All variables with completely missing values were dropped - religion, weight, mobile_access, web_health_visit_frequency, used_health_web, had_anxiety, sought_information - as were respondents missing age information. Remaining missing NaN values were relabeled as "unknown" for clarity and graphing coherency. Furthermore, variables were renamed for more clarity. For example, rincome, compuse, and hlthwblif were altered to be more descriptive (respondent_income, computer_use, healthy_lifestyle_visit_frequency).

The 2022 data was ultimately selected over the 2000 data because it provided opinions from the same set of respondents in one time period rather than spanning multiple years. It also contained more relevant variables related to perceptions of health websites, like internet_helped_explain_doctor and is_reliable, that rely on recent technology advances. Though some cleaning was required, the 2022 data provided the most cohesive and applicable information to answer the research question on how different groups view online health resources. The main challenges came from importing and reading in the data, and renaming and aligning variables across multiple years to focus solely on 2022.

RESULTS

The main variable explored was the is_reliable variable. Of the given variables, we believed that a person's trust is the strongest indicator of their overall attitude towards health related inquiries. The initial distributions of is_reliable by demographic can be seen in the figures of appendix A.

Looking at the heatmap of responses for reliability of health data and frequency of visiting seen in figure 1, it is shown that the majority of the overall population is trusting of health data regardless of how often they visit these websites. Within each frequency category, 'agree' was noticeably the most common response for is_reliable. The responses 'disagree', 'neither agree nor disagree', and 'strongly agree' all had similar response rates; however, 'strongly disagree' was by far the least common response and is easily distinguished as being significantly different in number. Therefore, the majority of the responses indicated a trusting sentiment of health data.

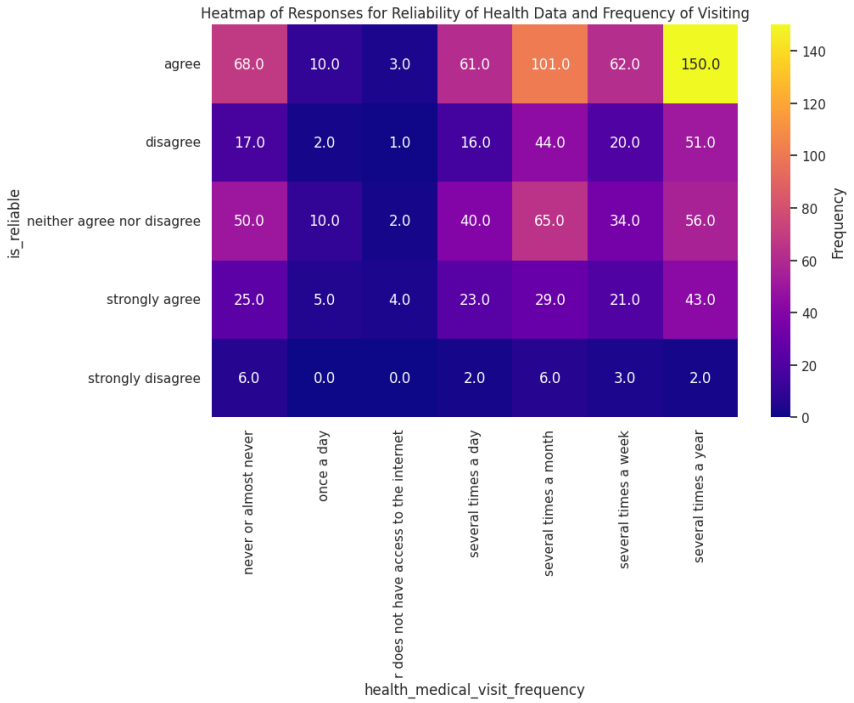

Figure 1: Responses for reliability of health data by frequency of visiting

When the responses are then split by sex, shown in figure 2, the difference in demographic shows no noticeable effect on trust in health data. In the heatmaps for reliability of health information and frequency by sex, it's seen that both heatmaps have extremely similar distributions and could be considered almost identical. Both heatmaps also nearly mirror the heatmap for the overall population; therefore, no effect on response due to sex is observed.

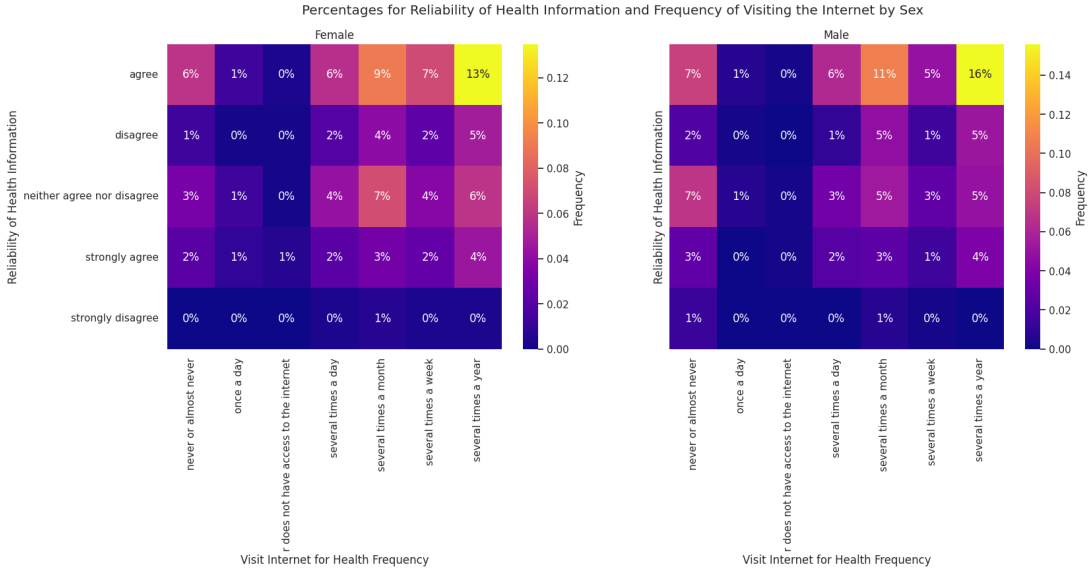

Figure 2: Responses for reliability of health data by frequency split by sex

Next, when the responses are split by race shown in figure 3, the difference in demographic does seem to have a small effect on the strength of opinion about health data reliability. As seen on all three heatmaps of health data reliability and frequency of visits by race, the overall sentiment about health data is still positive; however, minority respondents are more likely to hold a neutral opinion or not have one at all. It should also be noted that minority respondents showed a greater spread in response for the frequency that they visit health related websites. For white respondents, they were slightly more likely to respond that they visit 'several times a year' or 'several times a month'.

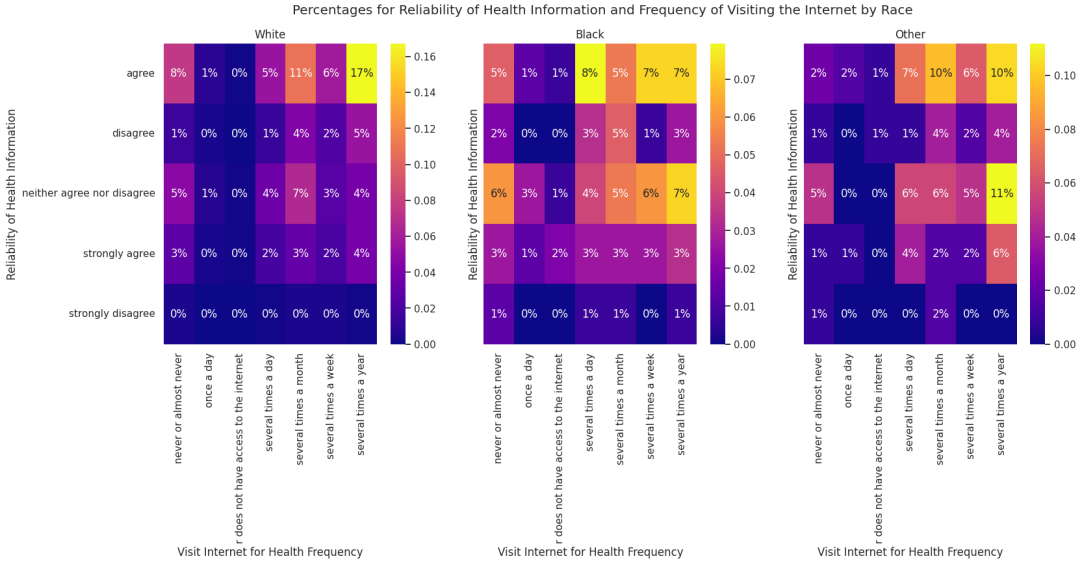

Figure 3: Responses for reliability of health data by frequency split by race

Overall, when the responses are split by education level shown in figure 4, there isn't much effect on trust of health data. Again, trust was the sentiment of the majority of the responses regardless of education level. Something of note would be that respondents with graduate level education who visit health related websites 'several times a year' are recognizably more likely to respond 'disagree' than other demographics. Also, it can somewhat be seen that the lower the education level, the more likely the respondent 'never or almost never' visits health related websites.

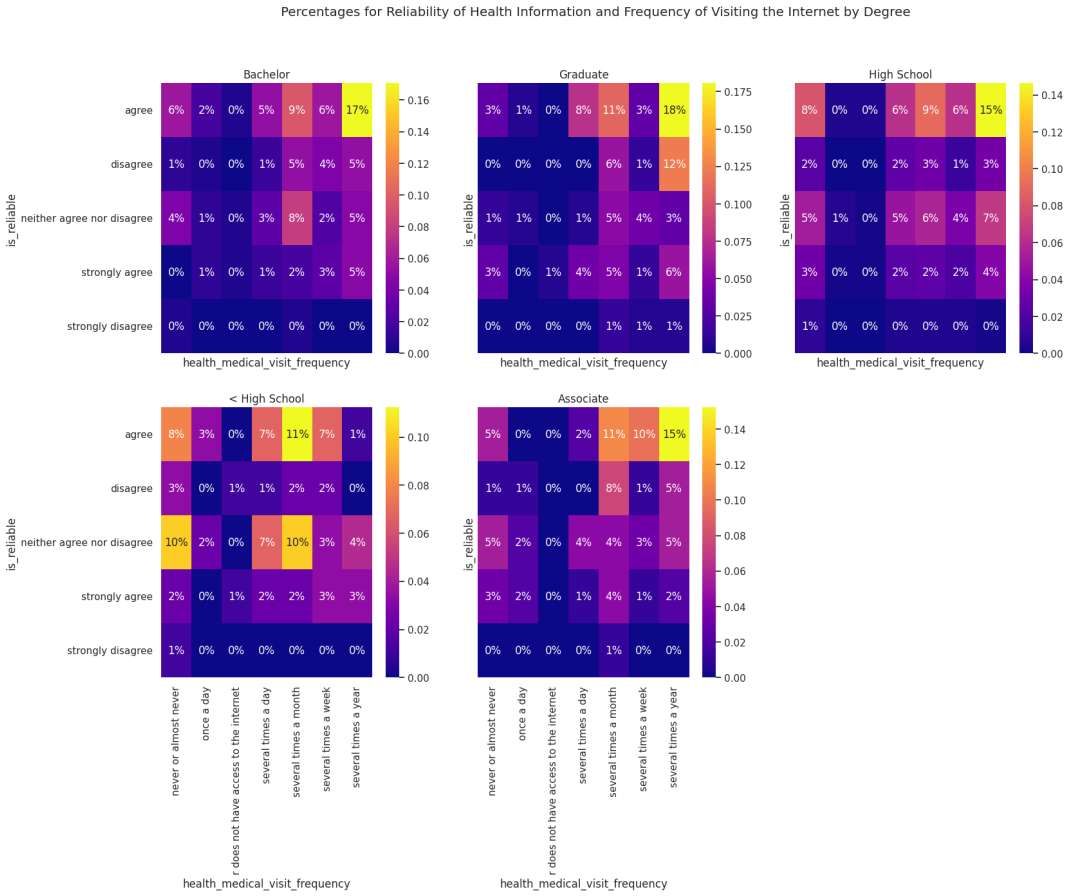

Figure 4: Responses for reliability of health data by frequency split by education level

To summarize thus far, demographics play nearly no role in determining a respondents opinion on the trustworthiness of health data. Sex, race, and education level have shown no major changes in response other than race playing a minor role in determining the strength of one's opinion and some variation in frequency of visits to health related websites.

We further explored the subject by investigating the relationship between frequency of visits to healthy lifestyle websites and whether or not a respondent feels they were affected positively or negatively. The mosaic plot shown in figure 5 shows a direct relationship between frequency and opinion of affect. The greater the frequency of visits of the respondents, the more often the respondents replied that they have been affected positively and vice versa.

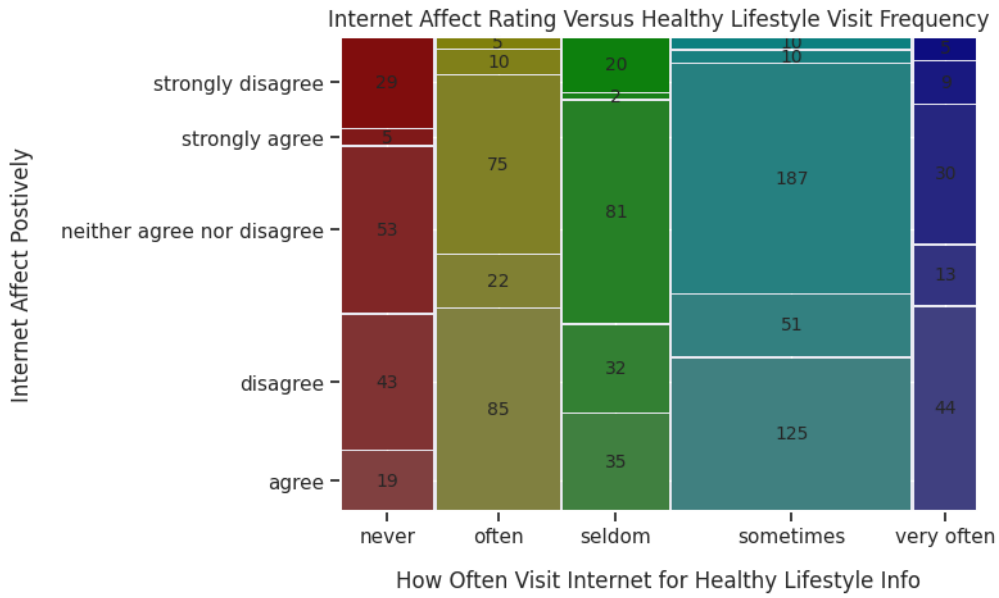

Figure 5: Responses to internet's effect by frequency of visit to healthy lifestyle websites

CONCLUSION

The overall scope of our project set out to explore variations in attitudes towards using the internet for health information across different demographic groups using the 2022 GSS dataset. The analysis focused specifically on frequency of visiting certain health-related websites as well as agreement with statements about the role of online resources in health decisions. Because both of these variables were surveyed as categorical variables, we selected heat maps and contingency tables as our main methods of visualization. After extensive data cleaning and preparation, visualizations and comparisons across groups revealed some noteworthy patterns to help answer our overall question, but also limitations.

Overall, the results suggest that whether or not someone agrees that it is not easy to distinguish reliable information on the internet may correlate with frequency of health website usage, but concerns about reliability spanned groups. Across ages, education levels, and sexes, this was the dominant perspective as it was consistently the lone yellow cell in the heat maps with few exceptions. This indicates that while usage habits may differ, most people share common judgments about distinguishing good and bad information online. More nuanced demographic differences did emerge around race and education when analyzing reliability in detail. Whites largely echoed average trends, while Blacks and "Other" categories showed less defined opinions. Those with lower education levels also displayed substantially more indifference to the statement: It is not easy to distinguish between reliable and unreliable health information on the internet.

As for our mosaic graph, we wanted to see if there is correlation with visiting the internet for healthy lifestyle purposes and using the internet for health has had a positive impact on a

person's health. We would want to check for a correlation between frequency of visiting health websites and believing the internet generally has positively impacted health behaviors to see if real-world usage aligns with perceived benefits. A total of 472 participants surveyed some combination of visiting the internet “often” or “sometimes” for healthy lifestyle information and “agree” or “neither agree nor disagree” on if the internet positively affects their health behavior. These four cells in our mosaic plot were clearly the most dense. The one exception to this trend was the seldom and neither group with a count of 81, meaning that a large majority of the people who rarely visit the internet for health information often do not have a particular feeling on whether the internet affects them positively or not.

Some would criticize our choice in using percentages in our heat maps. Directly comparing percentages of different populations can be dangerous because percentages can be misleading if given a small population size. Although this is true, in our EDA the split of the total data by demographic can be seen by comparing the heights of the bar graphs seen in figure A2 of appendix A. Therefore, the relative population sizes between demographics can be seen and a reader can determine if a demographic has too small of a population to yield trustworthy percentages.

One point of emphasis for a future change is that the self-reported nature of the survey data inherently limits its objectivity compared to logging actual internet activity. Although it would be extremely hard to gather opinions on certain questions or statements through completely “unbiased” means. However, If we were to survey people again, we as a group agree that it would be beneficial to take out the indifferent option of “neither agree nor disagree” or “sometimes” in some of these questions. Therefore, it would force participants to choose one side versus the other instead of having the option of answering a survey without any real effort.

Along with altering the answer choices, qualitative research like interviews or focus groups may help clarify the nuances in how different groups form perceptions of online health resources. Using these focus groups and interviews, we would be able to evaluate the causes and ethics of the reason why certain groups tend to think what they think. Factors like interface design, formatting, website credibility, and past experience likely influence these specific attitudes. Additionally, tracking real behavior through browsing data and search patterns could complement the self-reported frequencies relied on here.

As for improvements in our graphs, we would try to find a way to include both the raw counts and the percentage in each cell. Having both of these numbers together gives relative context to each cell rather than just having one or the other. Another similar critique that we would impose in a future draft would be to aggregate and display the total count of each category on the X and Y axes of our mosaic graph and heatmaps. This would allow us to clearly distinguish how dominant a category might be over another.

Overall, our project demonstrates a broadly shared sense that health websites can provide useful guidance, but with questionable reliability. Our heat maps in particular point towards income, race, and education as potential focuses for improving access and understanding. In our mosaic plot, we saw trends that we tend to expect: those who use the internet for their health behavior more frequently are more likely to have a more positive connotation towards the internet's effect on them and vice versa. But fully unpacking the relationship between demographics and health information literacy will require more detailed behavioral data and qualitative insights into decision-making.

Appendix A

Initial response distributions of is_reliable

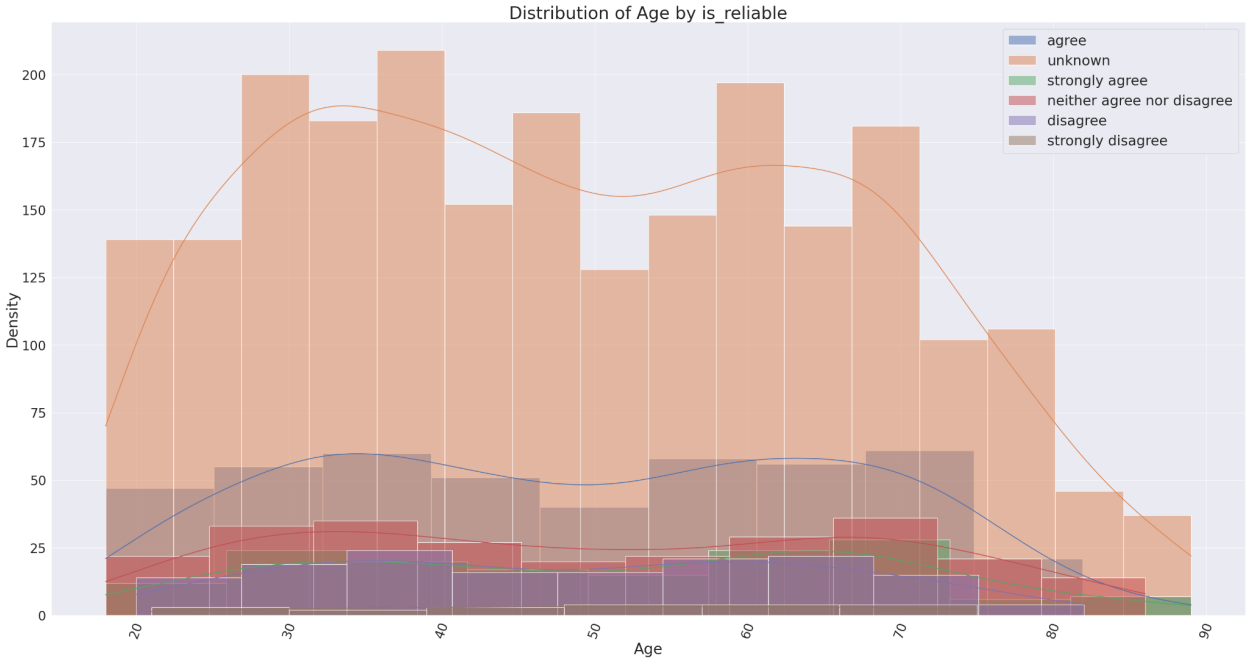

Figure A1: is_reliable responses by age

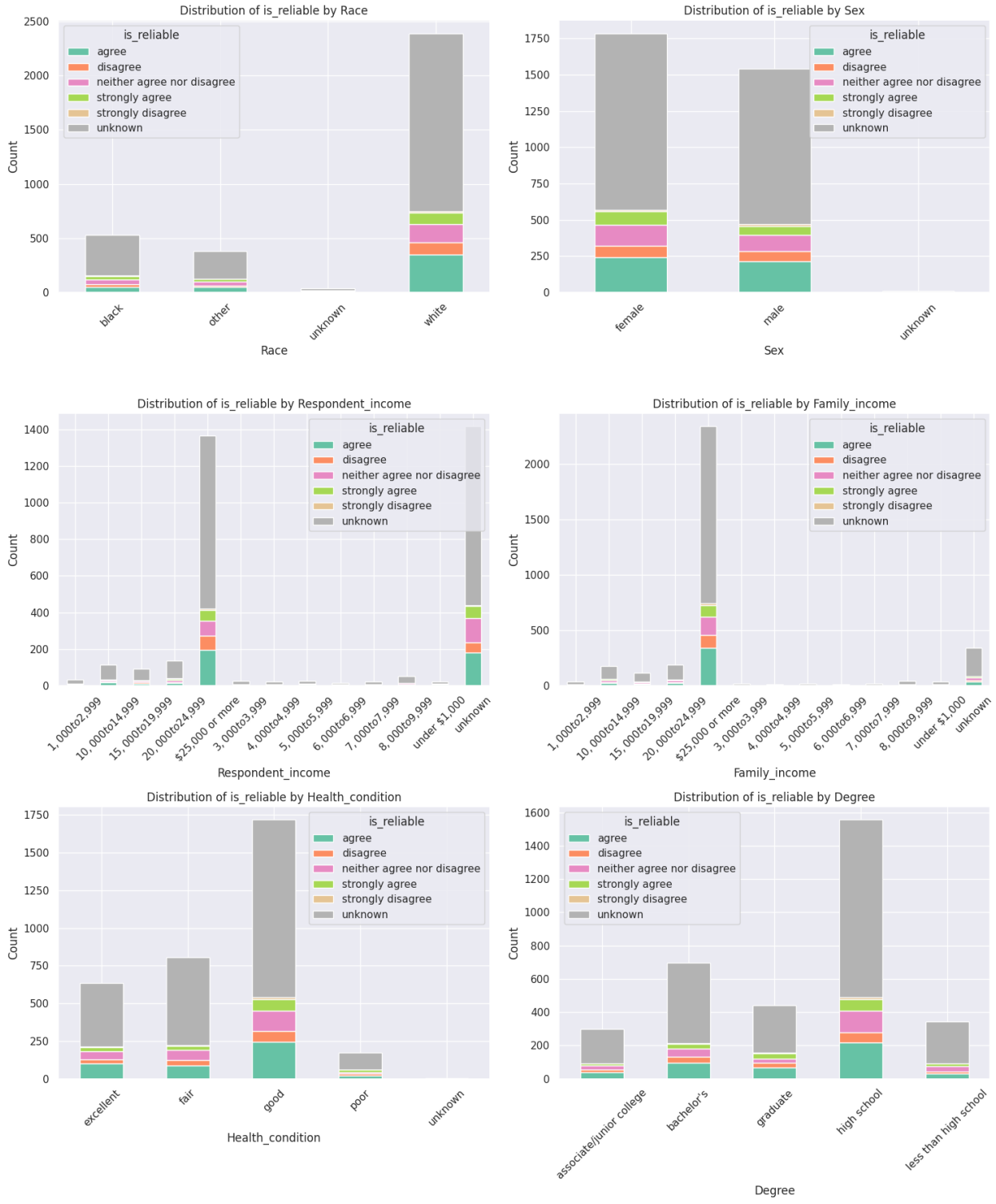

Figure A2: is_reliable distributions by race, sex, income, family income, health condition, and degree earned