Project 1 GSS Data: Political Preferences & Personal Background

Group 7: Anna Brown, Adaire Burnsed, Tamera Fang, Diana Nguyen, Elle Park, Carol Wu

School of Data Science, University of Virginia

DS 3001: Foundations of Machine Learning

Professor Johnson

March 2, 2024

Abstract

Is there a correlation between an individual's background and their political preferences? Utilizing data from the General Social Survey (GSS), we initially researched the potential correlation between age and political preferences. However, as our research progressed, we also found exploring the relationship between race and political ideologies compelling. Therefore, we centralized our analysis on the dataset's *polviews* (political view) and *age* variables. Our methodology began with importing these variables and conducting a preliminary review of the dataset's initial columns. Then, we assessed the dataset's dimensions and categorized political views, both qualitatively and as numerical variables. Further analysis was directed towards the *racerank1* variable, which categorizes voters' political views across different races and ethnicities. In our data cleaning process, we replaced missing values with placeholders and mapped the *racerank1* values to a new column with corresponding numeric representations. To effectively communicate our findings, we employed a variety of visualizations, including histograms, stacked bar charts, scatter plots, joint plots, bar plots, and kernel density plots. Our analysis revealed that a significant portion of the population identifies as moderate or centrist. We also noted that most voters surveyed in the dataset identified as White, with African Americans constituting the second largest demographic group. The kernel density plot, in particular, indicated that individuals between 40 and 60 years old tend to lean slightly conservative. By assigning numerical values to political views, we observed that, excluding the White demographic to focus on minorities, African Americans—the predominant minority group—most frequently aligned with the moderate or centrist position. Overall, our exploration into how both age and race, along with other ethnic factors, influence political views in the

United States via GSS data was an interesting topic to research.

Data

Our data frame includes the variables *racerank1*, *racerank1_numeric*, *age*, *polview*s, and *polviews_numeric*. The *age* variable corresponds to the respondent's age. There are 591 NAN values in the *age* variable. To avoid skewing the data analysis result due to the large number of missing values, we dropped all NAN values in *age* from our analysis.

The *racerank1* corresponds to the answer to the question, "If you had to choose, which of these races do you identify most with?" The *racerank1_numeric* variable was created by mapping the GSS codebook assigned value to each race category. We used these two variables to analyze the sixteen different races that the survey received data from. Further into the project, we merged the *racerank1* variable into six categories: White, Hispanic, American Indian or Alaska Native, Black or African-American, Asian and Pacific Islander, and some other races. The *racerank1_numeric* is the numeric variable where each race from *racerank1* was assigned a number from 1-16. There are a total of 20,639 NAN values in *racerank1*, which can be a result of choosing one of the following answers based on the GSS codebook: 'Don't Know,' 'No Answer,' 'Skipped on Web,' 'Not Applicable.' To avoid skewing the data analysis result due to the large number of missing label groups, we dropped all NAN values in *racerank1* and *racerank1_numeric* from our analysis.

The *polviews* variable was the political views that the participants aligned: slightly/extremely liberal, liberal, moderate/ middle of the road, slightly/extremely conservative, and conservative. The *polviews_numeric* variable provides a numeric scale between 1 and 7 that aligns with participants' political views. To treat all variables consistently, we dropped the 98 NAN values in the *age* variable and all NAN values in *the polviews* variable from our analysis.

Results

According to the histograms Figures 1A and 1B, there is a distribution skewed toward White respondents, with a notably smaller representation from Black or African American individuals and significantly fewer from other racial categories. As indicated by the mode of *racerank1*, most of the survey participants identify as 'White' (2,514 respondents). The second highest group of respondents identified as 'Black or African American.' The categories {'Asian Indian,' 'Chinese,' 'Filipino,' 'other Asian,' 'Korean,' 'Vietnamese,' 'Japanese,' 'other Pacific Islander,' 'Samoan,' 'Guamanian or Chamorro'} have low individual counts compared to other major race categories. To reduce complexity in data visualization, we merged these categories into one single category, 'Asian and Pacific Islanders,' for the percent stacked bar chart of political affiliations by race.

To take political views into account, the histogram in Figure 2 allows one to view the total counts per category. However, this data is not very clean, so the political affiliation categories are further organized into three major ones for simplicity of visualization, as seen in Figure 3: liberal, moderate, and conservative. Since the number of respondents from each racial group varies drastically, we scaled the counts of the three political affiliations for each race as percentages of the total respondents in each race group. The stacked bar chart in Figure 3 indicates that "moderate" is the predominant political affiliation across all racial groups. Furthermore, it shows that the percentage of respondents identifying as liberal outweighs those identifying as conservative across all racial groups. The data also highlights variations in the percentages of liberal and conservative respondents among different racial groups, with the most significant gap observed in the "Asian and Pacific Islander" group and the smallest in the

"White" group. The trend suggests that while moderates are the most prevalent political affiliation across all racial groups, there is a consistently higher proportion of liberals compared to conservatives, with variations in the magnitude of this difference across different racial demographics.

Figure 4A is a kernel density plot that visualizes the relationship between *age* and the six political affiliations in the *polviews* variable. According to the visualization, 'moderate, middle of the road' political affiliation is most frequent for all age groups overall (0.007). There are the most 'moderate, middle of the road' affiliated individuals at about 30-35 years old. The data also highlights how 'extremely conservative' or 'extremely liberal' have the lowest frequencies compared to other political affiliations. There seems to be a pattern where 'moderate, middle of the road' and every liberal category are more frequent with younger age groups (<50 years and under) than their older age groups (>50). And this pattern seems to be the opposite for all conservative political affiliations. Older age groups (>50) have more individuals identifying with all conservative political affiliations than their younger age groups (<50).

The histogram in Figure 4B depicts the age distribution of survey respondents, which could be a critical factor in understanding political preferences. Based on this figure, the dataset appears to have a normal distribution, with the largest number of respondents between ages 30 and 40 and the lowest number between ages 80 and 90.

Figure 5A presents a scatterplot of *age* and *polviews* across all races, highlighting the distribution of political affiliations within different age and racial groups. This plot demonstrates the uniformity of political orientation across ages within each racial group, which suggests that the age of respondents does not significantly affect the distribution of political views within

racial groups. This finding might indicate a cross-generational consensus in political leanings within these communities. In Figure 5B, the scatter plot distribution excludes White respondents to focus on minority groups, clarifying the political views of these demographics. The plot helps view political diversity within minority age cohorts, showing that political preferences are not monolithic within racial categories other than White.

Meanwhile, Figure 5C isolates the White respondents, providing a focused look at the largest demographic group in the dataset. This concentration allows for a detailed examination of political views within the White population across different ages. It is a critical comparison point to Figure 5B, highlighting potential differences or similarities in political affiliation trends between White and non-White groups.

Figure 5D's kernel density plot showcases the distribution of fx by race across all ages. This visualization helps us understand which political affiliations are most densely populated within racial groups and how they compare across these groups. The density peaks for moderate views across races, yet the spread and shape of the curves for liberal and conservative views can inform on the diversity and intensity of political leanings within each racial category.

The joint plots in Figure 6 are beneficial because they allow us to visualize political views across age and race. They show us the distribution of the variables, especially how the race variable covaries with both political views and age. In that way, these plots summarize our entire project! We made three of these plots: one with all the races included, one with just the white respondents, and one that excluded the white respondents. The details and patterns noticed are as follows: Figure 6A is a joint plot distribution of age and political views across all races. This figure demonstrates a tendency towards moderate political views across the racial spectrum,

regardless of age. Given the extremes, the distribution tails suggest that younger respondents lean slightly more liberal while older respondents lean somewhat more conservative. Figure 6B is crucial as it excludes White respondents, thereby focusing on minority groups to reveal the distribution of political preferences amongst these populations. The pattern observed from the previous figure is consistent, maintaining a central moderation in political views across age groups. Our last joint plot in Figure 6C addresses the political views of only White respondents. The importance of this figure lies in its ability to isolate the political trends within the most represented racial group in the survey, in which we also notice that central moderation is maintained.

Conclusion

Our analysis of the General Social Survey (GSS) data brings us to several important conclusions about the American electorate's political preferences, given their age and race. The uniformity of political orientations across ages within racial groups, as demonstrated in Figure 5A, and the general age-related progression toward more conservative views suggest a need for political strategies that address the distinct concerns and values across the age spectrum. However, the category of political views with the largest number of self-identified members was ultimately the moderate category, as clearly depicted by the joint plots in Figure 6A. This centrality within the political spectrum challenges the standard narrative of a highly polarized electorate, suggesting that the existing two-party system may underserve a substantial portion of the population. It would be interesting to examine the views and identities of the people who identify as moderate more deeply to see how satisfied they are with our current government and if they feel their representatives are serving them. Although this further research may require additional surveying, it would have more comprehensive implications than this project's scope. Additionally, it could provide information beneficial to politicians or political scientists and help people understand the current political climate in the United States.

Appendix

Figure 1A: Histogram showing the distribution of respondents by ethnicity, with a predominant peak for 'White' respondents, illustrating the racial composition of the survey participants.

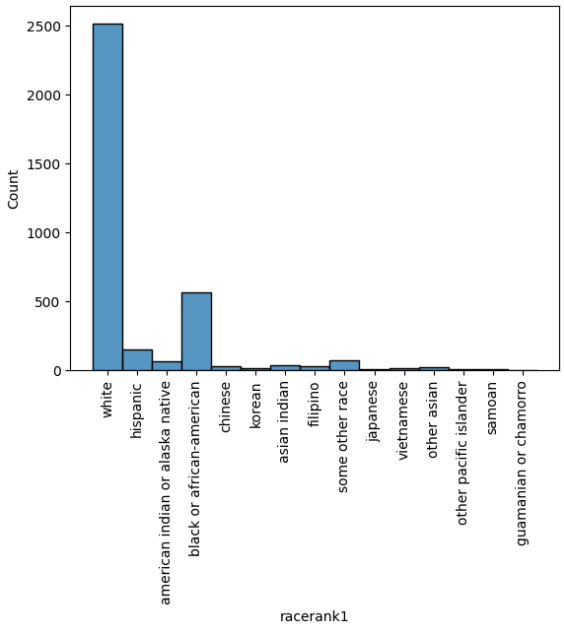

Figure 1B: Similar to Figure 1A, but this histogram is of each ethnicity assigned as a specific number as per the GSS codebook.

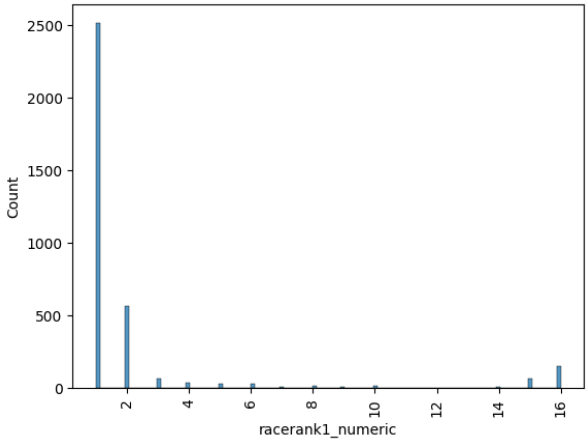

Figure 2: Histogram of the counts of each political view.

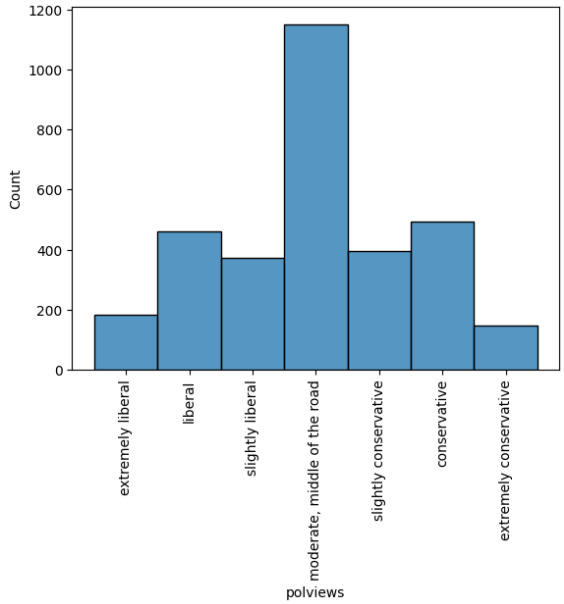

Figure 3: Stacked bar chart further exploring respondent demographics, in which the percentage of each political affiliation within different races helps correlate different political views with racial identity.

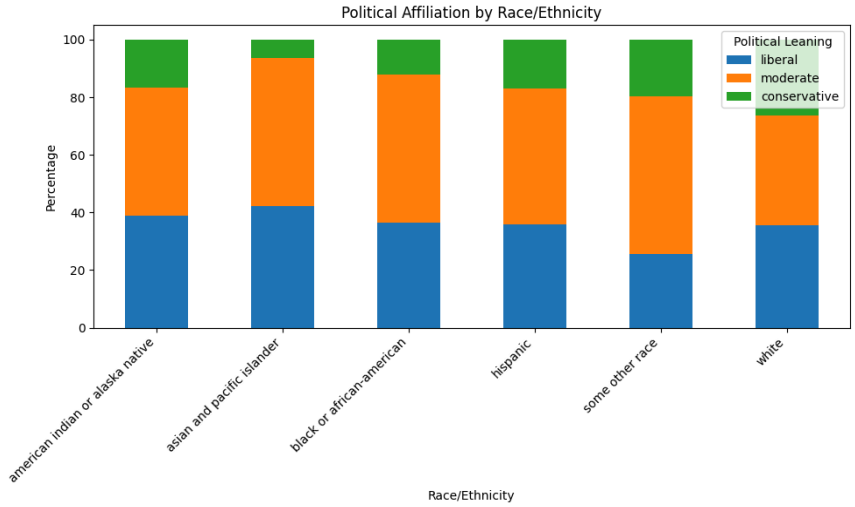

Figure 4A: Kernel density plot displaying the density of political orientations across age groups.

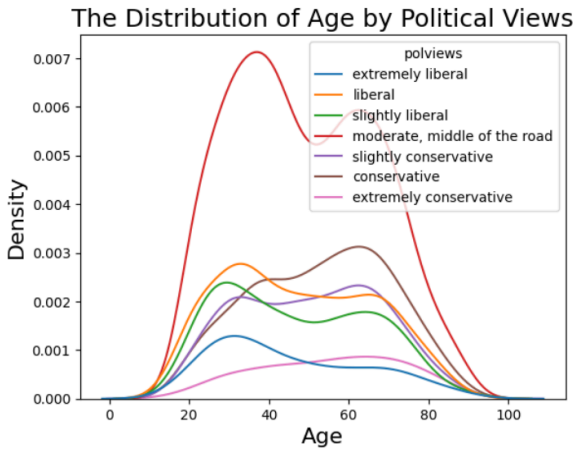

Figure 4B: Histogram showing distribution of respondents by age.

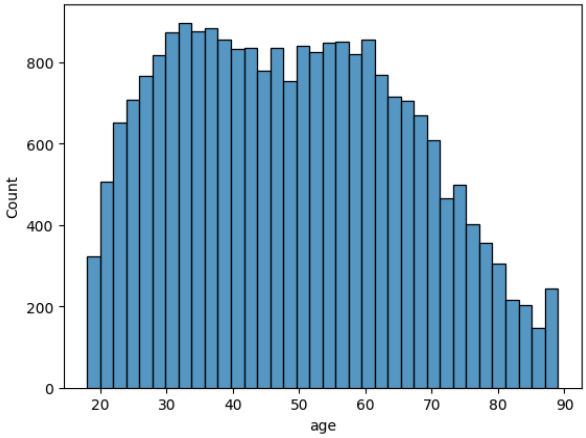

Figure 5A: Scatterplot distribution of age and political views by race across all races.

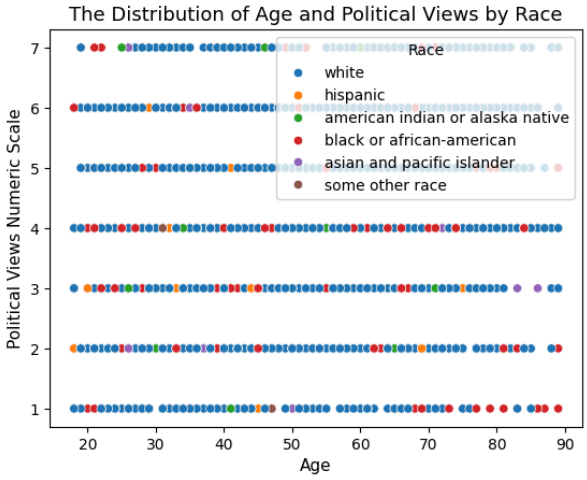

Figure 5B: Scatterplot distribution of age and political views by race across all races except for 'White'.

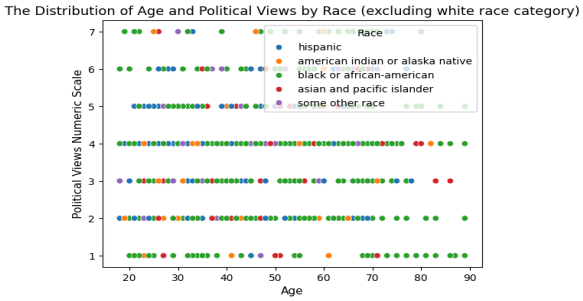

Figure 5C: Scatterplot distribution of age and political views only for the 'White' race category.

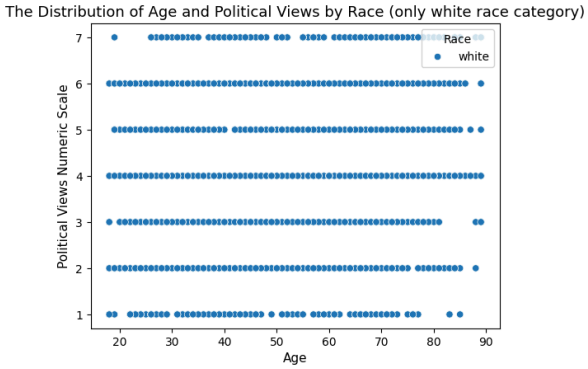

Figure 5D: Kernel density plot for the distribution of age and political views by race across all races.

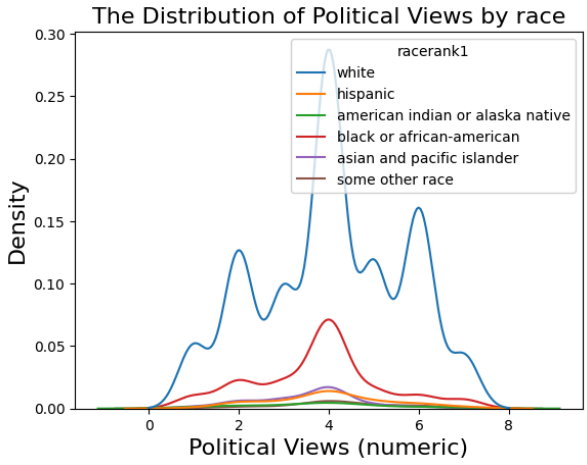

Figure 6A: Joint plot distribution of age and political views by race across all races.

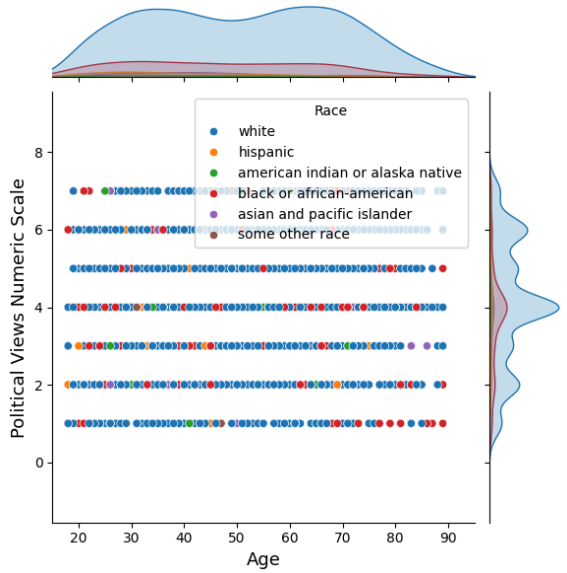

Figure 6B: Joint plot distribution of age and political views by race across all races except for 'White'.

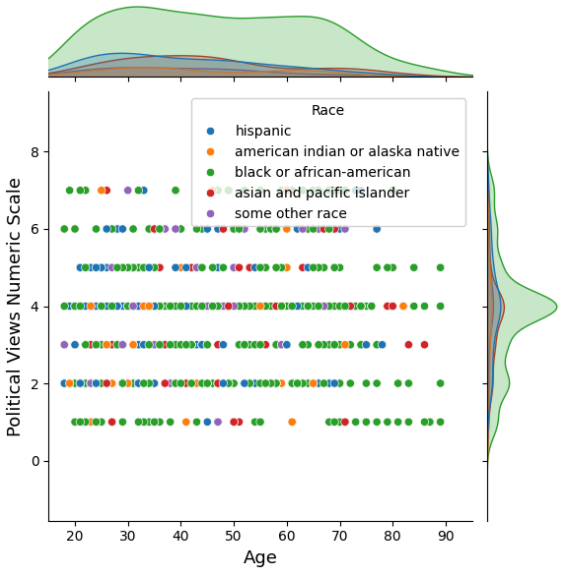

Figure 6C: Joint plot distribution of age and political views only for the 'White' race category.