Chinmayee Yerraguntla
January 27, 2026
EDA Extra Credit Lab

Description of Data

When choosing variables, I wanted to choose a niche or category of variables to make it easier to look through such a large survey. I wanted to hone in on the respondent's family situation as well as a few educational attributes. The table below shows each variable I chose to study, what it measures, and my rationale for choosing it.

Variable	What it Measures	Rationale
year	GSS year for this respondent	Baseline/default variable that may be useful in studying trends over time
age	Age of respondent	May provide context to some of the other answers, especially about the number of children, the highest education level, income, and the number of hours worked per week
race	Race of respondent	Another baseline variable that may be useful later on in studying the effect of race on family-building or education levels
childs	Number of children that the respondent has	An important variable when considering the respondent's family: the number of children a person has can greatly impact other parts of their life, like income or hours worked per week.
sibs	Number of brothers and sisters	I thought this would be an interesting variable to examine to see if people who have a lot of siblings choose to have more or fewer children.
educ	Highest year of school completed	This is another important variable when looking at education. The highest year of school completed can affect income, which can then impact how a respondent chooses to build their family.
padeg	Father's (or oldest same-sex parent's) highest degree	I want to see how someone's father's highest degree may impact their decision to pursue higher education. In data terms, I want to see the effect of this on educ.
madeg	Mother's (or youngest same-sex parent's) highest degree	Same rationale as above, but I think it would also be interesting to see the impacts of having two educated parents versus just one.
spdeg	Spouse's highest degree	I think it would be interesting to examine the

		respondent's spouse's highest degree of income. Do people who are highly educated seek partners who are also highly educated (and vice versa)? How does that change over time?
rincome	Respondent's income	I think this is an extremely important variable because it can tell us about why respondents chose to build a family a certain way to achieve a certain level of education. For example, someone with a higher income may choose to have more children because they can afford it. Likewise, they may have chosen to achieve a higher level of education to earn a high income.
hrs2	Number of hours usually worked a week	I think this would be an interesting variable to correlate with the number of children the respondent has. Do people with more children work more or less than people with fewer children?

As outlined in the table above, I made certain decisions based on the different relationships I sought to explore, but I wanted to explore family and education as my overall theme across all the variables.

Numeric Summaries and Visualizations

Table 1: Descriptive Statistics for Quantitative Variables

	year	age	childs	sibs	educ
count	72390.000000	72390.000000	72390.000000	72390.000000	72390.000000
mean	1997.715541	46.528830	1.916839	3.820887	13.030874
std	15.109995	17.508642	1.756343	3.110475	3.177196
min	1972.000000	18.000000	0.000000	0.000000	0.000000
25%	1985.000000	32.000000	0.000000	2.000000	12.000000
50%	1998.000000	44.000000	2.000000	3.000000	12.000000
75%	2010.000000	60.000000	3.000000	5.000000	16.000000
max	2022.000000	89.000000	8.000000	68.000000	20.000000

	hrs2
count	72390.000000
mean	39.984639
std	1.887419
min	0.000000
25%	40.000000
50%	40.000000
75%	40.000000
max	89.000000

Figure 1: Histogram showing the distribution of race

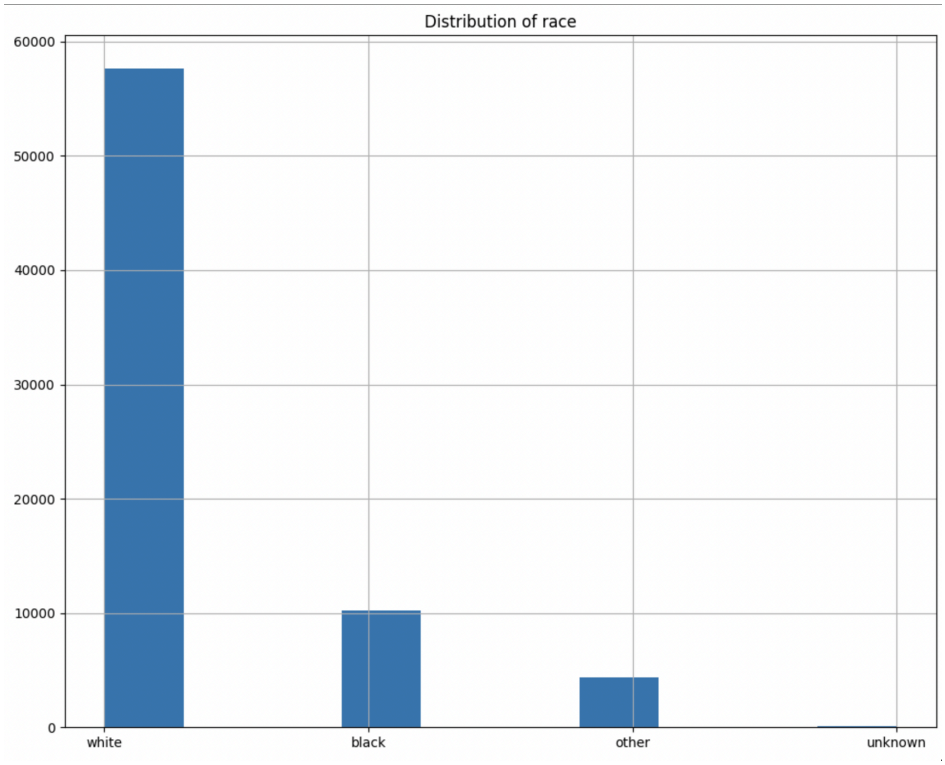


Figure 2: Histogram showing the distribution of fathers' highest degree achieved

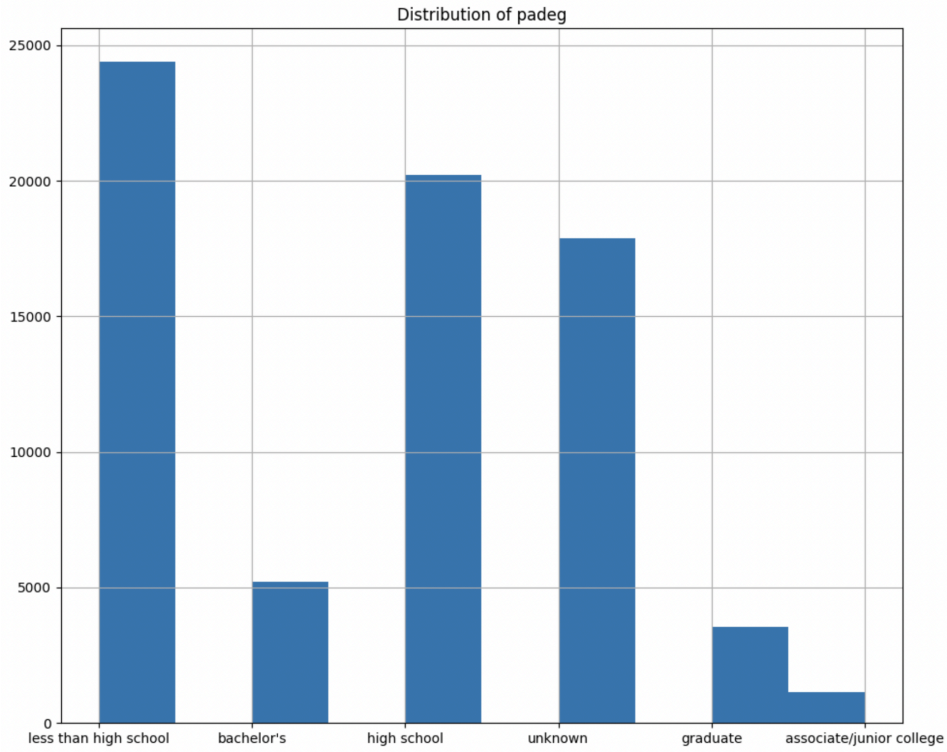


Figure 3: Histogram showing the distribution of mothers' highest degree achieved

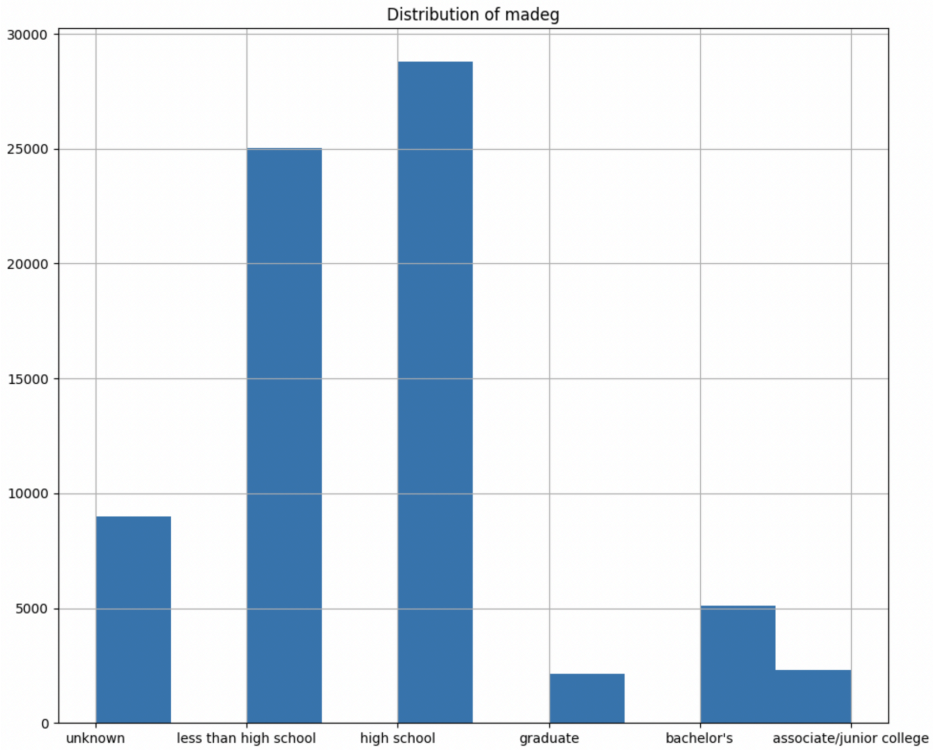


Figure 4: Histogram showing the distribution of spouses' highest degree achieved

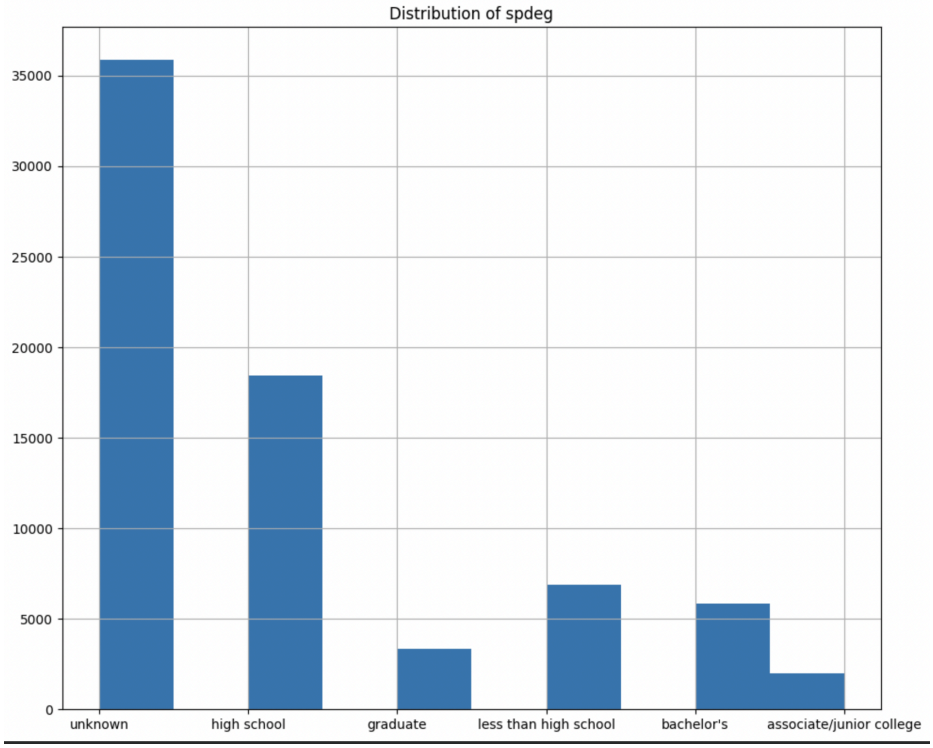


Figure 5: Stacked Bar Plot Showing Relationship between Father's Highest Degree and Mother's Highest Degree

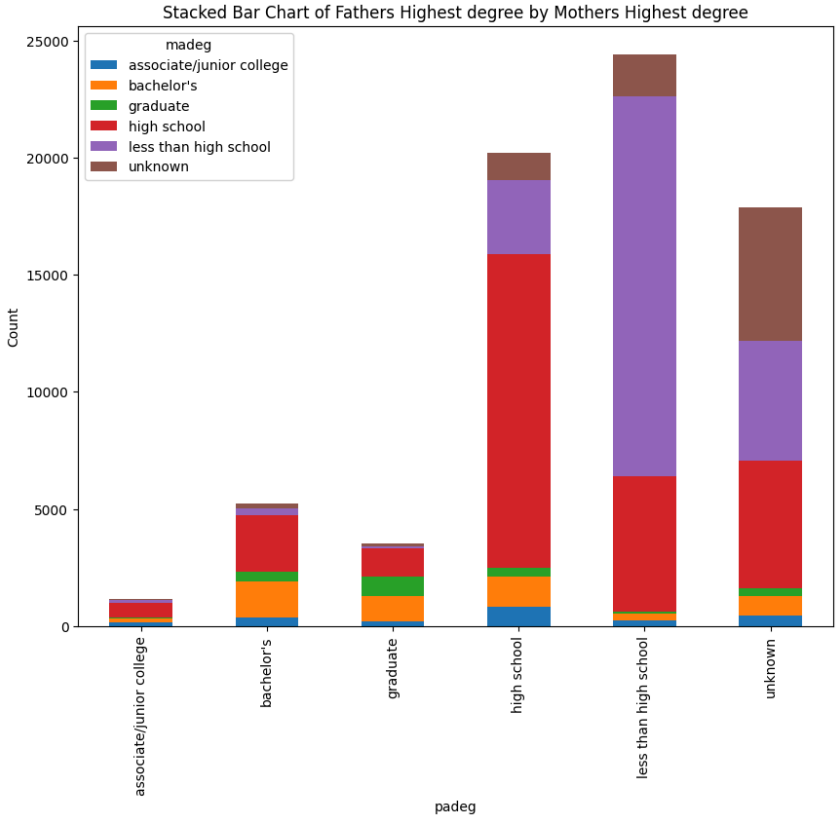
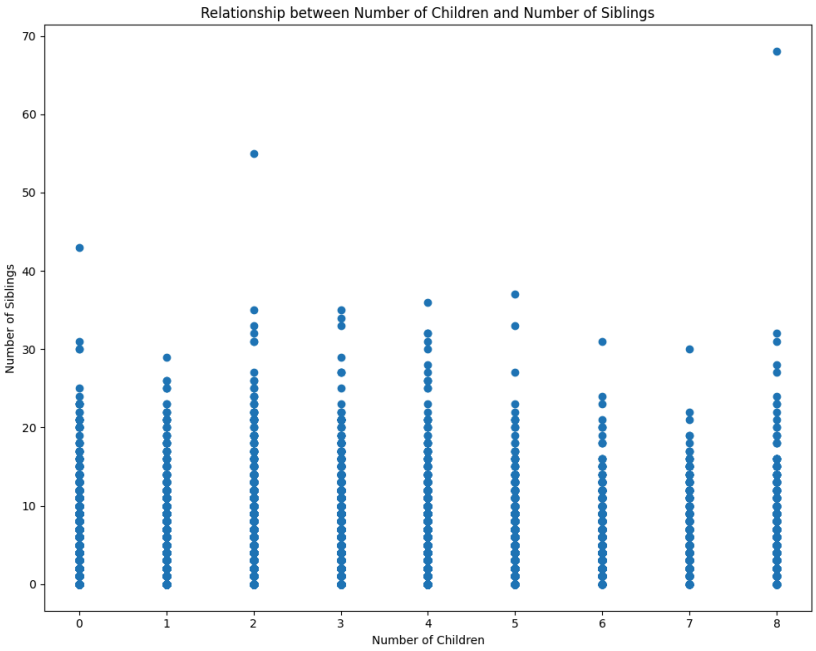


Figure 6: Scatterplot Showing the Relationship between Number of Children and Number of Siblings



Findings

Throughout my analysis, I found several notable things that illuminated information, not just about the data, but also about the survey in general. The first thing to note was that the 'rincome' variable has no data at all. Given that this was one of the most important variables that I thought would be the cornerstone of my analysis, I had to restructure my focus to explore different relationships, which I'll get into later. I also thought it was interesting that the GSS website included the indicator in the first place when it had no data at all.

Another thing I found interesting is that not all my education variables were of the same "type". I was able to classify all my variables as either quantitative or categorical. While 'madeg', 'padeg', and 'spdeg' are all categorical variables (Figures 1-4 show their distributions), the 'educ' variable is quantitative, with no easy coding mechanism to switch it to become a categorical variable. Since the education variables are measured in different ways, with 'educ' being the respondent's highest education level and arguably the most important education indicator, there wasn't a lot I could do with that variable. So, the 'rincome' and 'educ' variables were essentially unusable for my purposes.

To clean my data, I separated the variables by whether they were quantitative or categorical. For the quantitative variables, I got rid of any non-digit characters and converted the values from strings to numbers. I changed any missing values to the median value of the variable. For categorical variables, I stripped whitespace and changed the missing values to "Unknown" so that it would be easier to visualize the data.

Table 1 shows the numerical summaries for all the quantitative variables in the dataframe, and Figures 1-4 show the distributions of the categorical variables. After graphing the basic distributions/statistics for each of the variables, I wanted to graph the relationships between certain variables. The first one I chose to graph was the Father's highest degree and the Mother's

highest degree. This emulates the relationship I was originally trying to examine, between the respondent's highest level of education and their spouse's highest level of education. Figure 5, a stacked bar plot, displays this relationship. By far the largest category of respondents is where both the father and mother have less than a high school degree, but I think a large part of this is because of the generation the respondents' parents were born in. In the past, attaining higher education was even more strongly correlated with wealth than in present times. Furthermore, most of the madeg responses belong in the 'high school' or 'less than high school' category, which may be another reflection of the times.

The other relationship I wanted to graph was the one between the number of children and the number of siblings a respondent has. This was a relationship I wanted to examine from the very beginning, but the resulting relationship was a little bit confusing. Figure 6 shows the resulting plot. I think it's interesting how the scatterplot is in lines, and I also think it's interesting how the number of siblings variable has a max of 68 siblings, which seems a little counterintuitive. This may be indicative of inaccurate or false survey responses.

Overall, the process from choosing my variables until writing this findings section has been filled with insight that I hope to use in the future. Although my results and visualizations differed significantly from what I expected, they still provided valuable insights that could be useful for future research.