# Description and Why I Chose These Variables

The General Social Survey is given to Americans bi-anually. These surveys are a mix of questions regarding demographic, economic, and social status. Moreover, these questions also ask more personal details- like political views and income. GSS promises to its respondents that the information provided is fully anonymous and doesn't ask for identifiable information ( name, employee ID, address. I thought this dataset was very interesting, but one thing to be cautious of is that these results could likely be subjected to response bias- leading to skewed results and inaccurate reference. However, I think the goal of the lab is to explore the different variables and make an effort to make connections to different results from survey questions and to gain valuable insights about the lifestyle of Americans.

These are the variables I extracted from the dataset

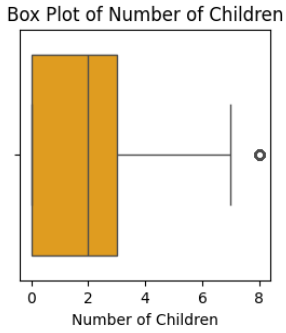
1. Year: This is the year that respondents participated in the survey
2. Educ: Highest Year of School completed
3. Income: Total Family income
4. Age: Age of respondent
5. Wrkstat: status of respondents' employment
6. Polviews: Respondent's political ideology
7. Hrs2: Number of Hours the respondent worked per week
8. Commute: Travel time to Work
9. Industry: Where the respondent's work
10. Childs: Number of children the respondent has
11. Paeduc- Highest year of school completed by their father
12. Maeduc- Highest year of school completed by their mother

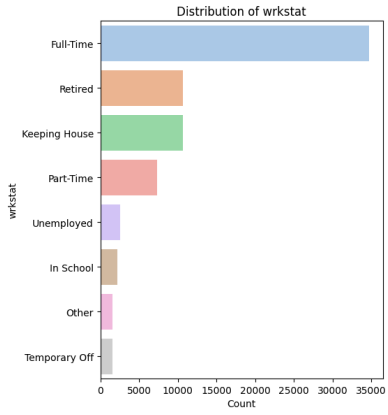
Description: I've explained what each variable represents/what they measure.Now the question is: why did I choose these particular variables? I was interested in exploring a few different relationships/correlations to see possible trends/patterns- then possibly making data driven ideas/ comments. In my mind, I've grouped these into a few measurements/relationships. The main one being if the Total family income has a correlation with commuting time ( longer time with low income could indicate the respondent has unreliable means of transportation).

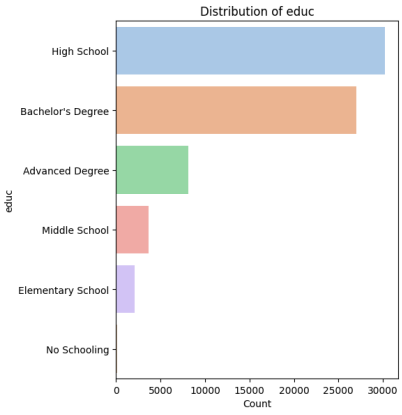
Another relationship I'd like to see is the relationship between income and educ/ paeduc/maeduc- I want to see if the education status of parents has influence over the completion status of the respondent's . Moreover, I was interested to see the relationship between income and polviews ( if income has a strong/ weak influence/**correlation** with their political affiliations. Additionally, I included variables like wrkstat (employment status), hrs2 (hours worked per week), and industry to explore potential links between job type, work hours, and income levels. I think what drew me to being interested in learning correlations/ insights with income and school completion is because ( being a first generation limited income student in college), I was curious to know if income distributions through the years are correlated with school completion. Please note, some of the values discussed will not be used depending on if the percentage of missing value is over 30% ( by which we will need to eliminate the variable from the study).

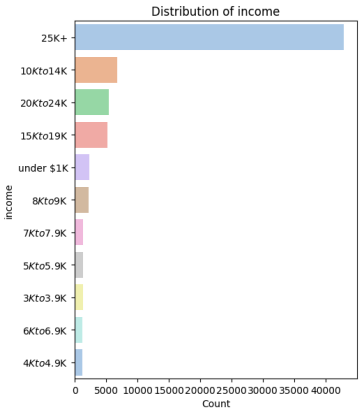
NUMERIC SUMMARIES: ON GOOGLE IPYNB file

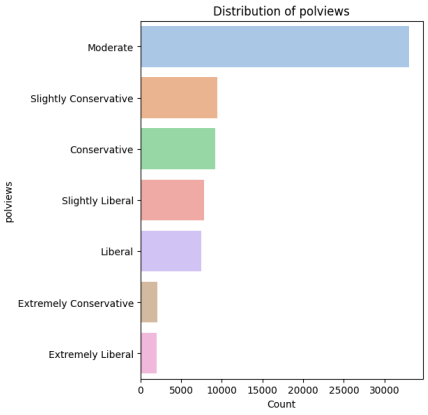
VISULIZATIONS!

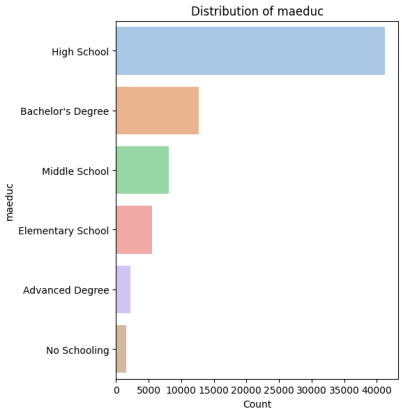


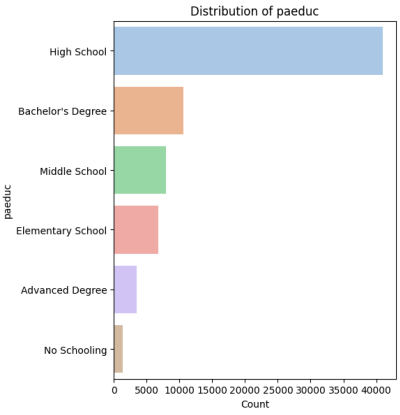



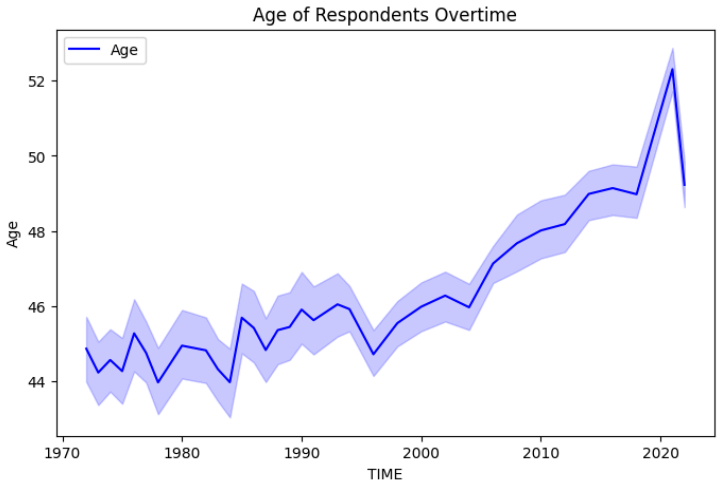



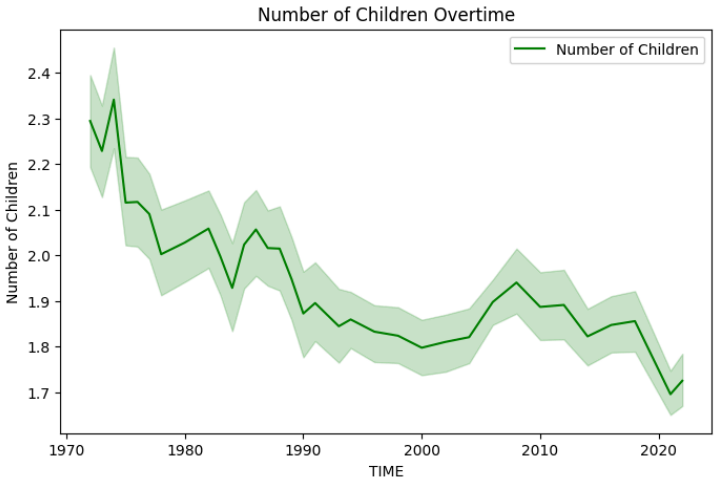



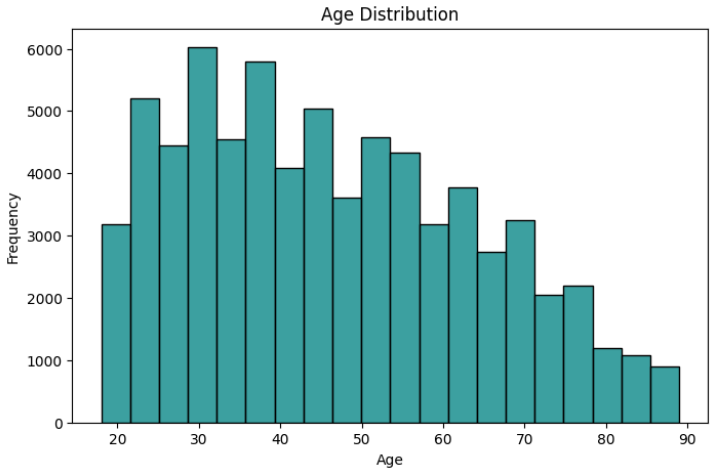



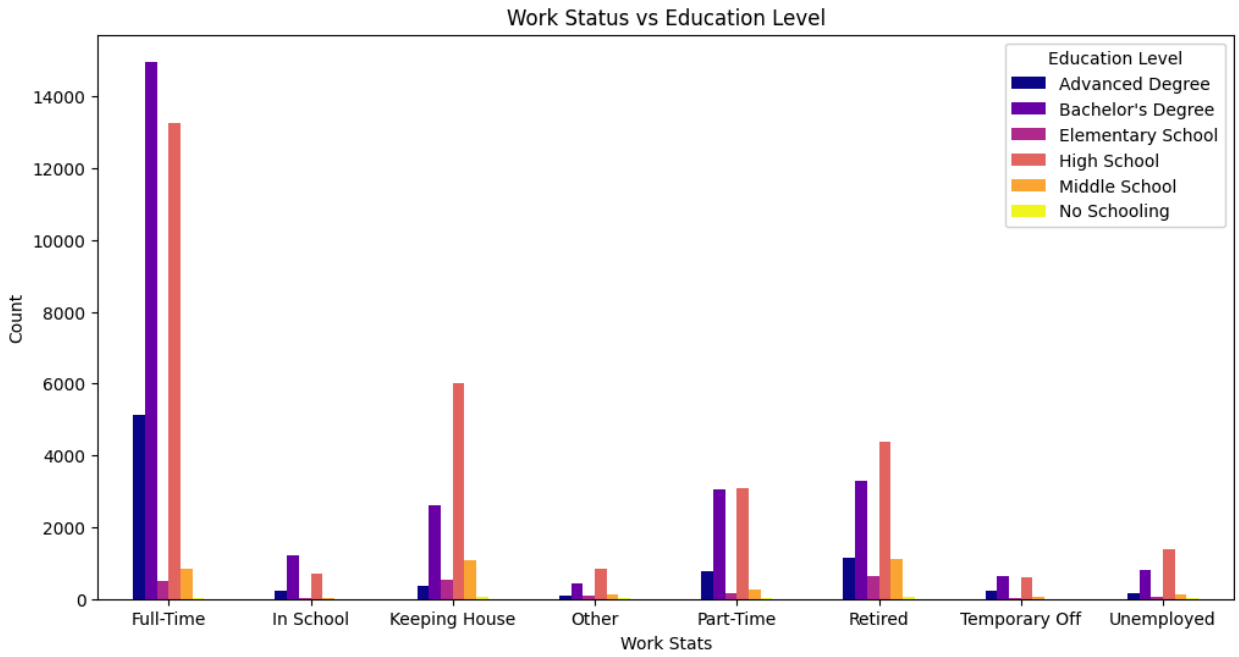




Age of Respondents Overtime
Age
52
50
48
46
44
Age
1970
1980
1990
2000
2010
2020
TIME


Number of Children Overtime
Number of Children
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
Number of Children
1970
1980
1990
2000
2010
2020
TIME


Age Distribution
6000
5000
4000
3000
2000
1000
0
Frequency
20
30
40
50
60
70
80
90
Age




INSIGHTS/ OBSERVATIONS

Let's start by describing what we saw from the box plot of the number of children distribution of respondents. It looks like the Max value was 8 and the minimum was 0 children. This type of plot is good at telling outliers. It looks like the plot is positively skewed (the tail being longer on the right side). This is interesting because through the years 1972-2022 ( measured from the graph labeled "Number of Children Overtime" - it looks like there was a relative decreasing trend of respondents having more than 3 children. From this, we could assume that the number decreased over time because of population dynamics (the majority of respondents are of single age- not married yet), and it may be due to economic pressures ( it's expensive to raise children in the modern economy).

Next, I graphed histogram/count plots for each categorical variable because I wanted to see the proportion/count makeup for each category. Something I found interesting is that it looks like the highest education completed by the respondent is undergrad and high school. However, the respondents' parents (the majority of them completed high school)- had different proportions/ratios of undergrad. Again, we could assume that older respondents completing the survey probably had the most counts of parents not completing undergrad ( limited accessibility/ different family dynamics during the 20th century). For example, I'm a first-generation college student ( meaning both my parents didn't go to college). My parents couldn't fulfill requirements/ balance family needs (In India) so they couldn't go to college.

Another observation I found interesting is that the highest total family income was 25K+ ( which makes sense because the modern economy is different from financial dynamics in the 1970s and we could see a dramatic increase in median household income). Moreover, the histogram that plotted the age distribution - indicates that most respondents were between the ages 18-50 ( meaning older people didn't contribute too much).

Finally, I graphed a stacked bar plot which looks at different work stats and how it's related to education level. Not surprisingly, it looks like a lot of respondents who work full-time have completed advanced degrees and bachelor's degrees. However, I'm curious to see the different types of full-time jobs for each education level- the industry data column had too many missing values to impute so I took it out from the study. A final observation I'd like to make is that the no schooling category is not observable with any of the bar plots- I'm interested to see the types of conditions/jobs that respondents with no schooling are doing. All in all, this was very insightful data to work with- and I

can understand the need for individuals to complete national surveys like these because working with missing values makes us neglect important information that could be used to help make data-driven decisions/ suggestions.