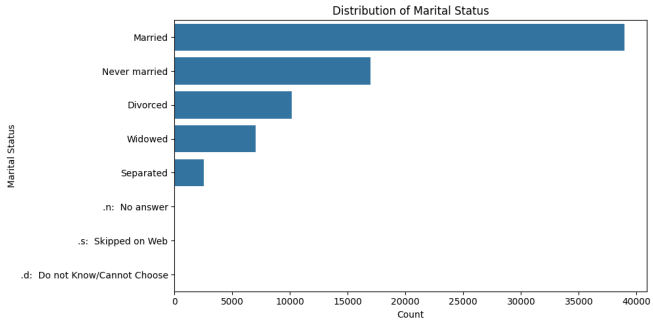
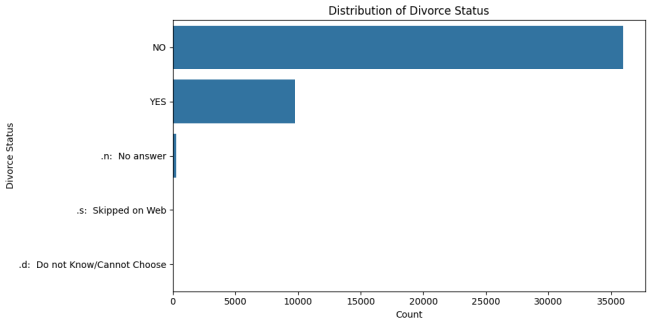
Write a short description of the data you chose, and why (1 page)

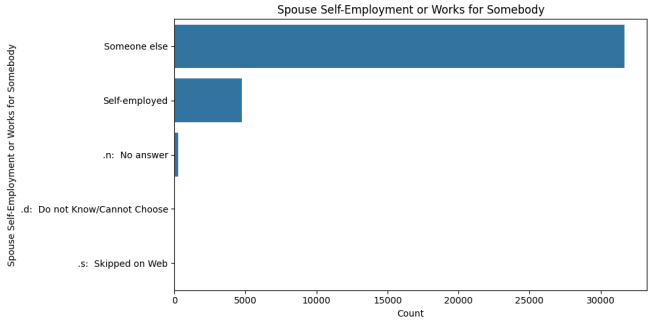
For the month of Valentine's Day, I chose to analyse variables that relate to marital history, marriage timing, family composition, and the socioeconomic aspects of significant others. The chosen variables from the General Social Survey (GSS) include divorce experience, current marital status, age at first marriage, number of children, marital happiness, spouse work arrangement (self-employed or employed by someone else), spouse occupational prestige, and the survey year. When analyzed together, these variables offer a multifaceted view of marriage outcomes and the social and economic characteristics associated with them.

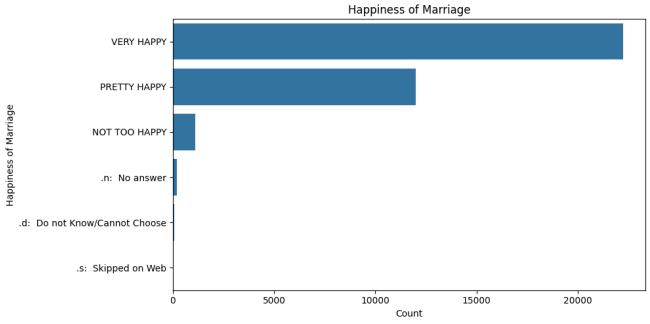
The variables relating to marriage offer different insights, subjective and objective, into relationship dynamics. The question "Ever divorced or separated" captures the subject's marriage history. "Marital status" identifies the subject's current relationship position such as remarried and stably single. "Happiness of marriage" offers qualitative insight of the satisfaction and contentment of relationships across various marital histories and demographic groups. Variables relating to timing and family structure such as age at first marriage and number of kids offers temporal and demographic information. For instance, earlier marriages potentially affect the educational and career paths of and different risks of divorce. The number of children reflects household characteristics and responsibilities that may potentially impact marital satisfaction for better or worse. Moreover, variables about the socioeconomic status of spouses help grant insight on relationships within different economic realities. A spouse's self-employment status can bring obstacles in work-life balance and income stability. Occupational prestige can offer insight on socioeconomic standing and job quality. The survey year enables and ballot form helps with the analysis of how these dynamics evolve across changing social and economic conditions.

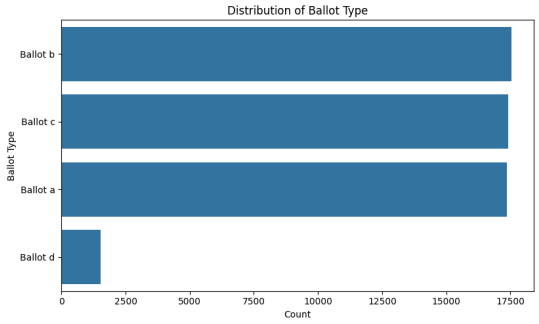
Produce some numeric summaries and visualizations. (1-3 pages)





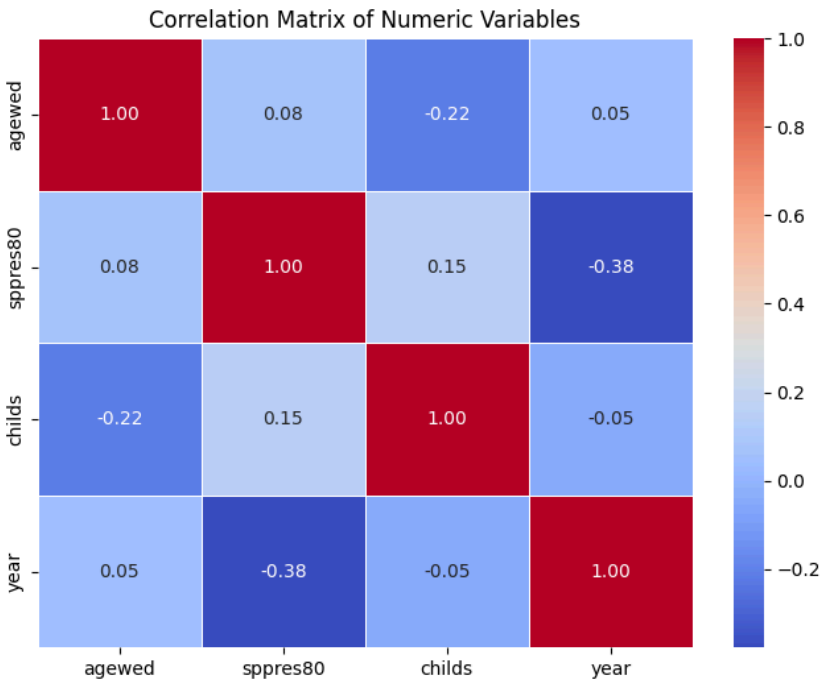


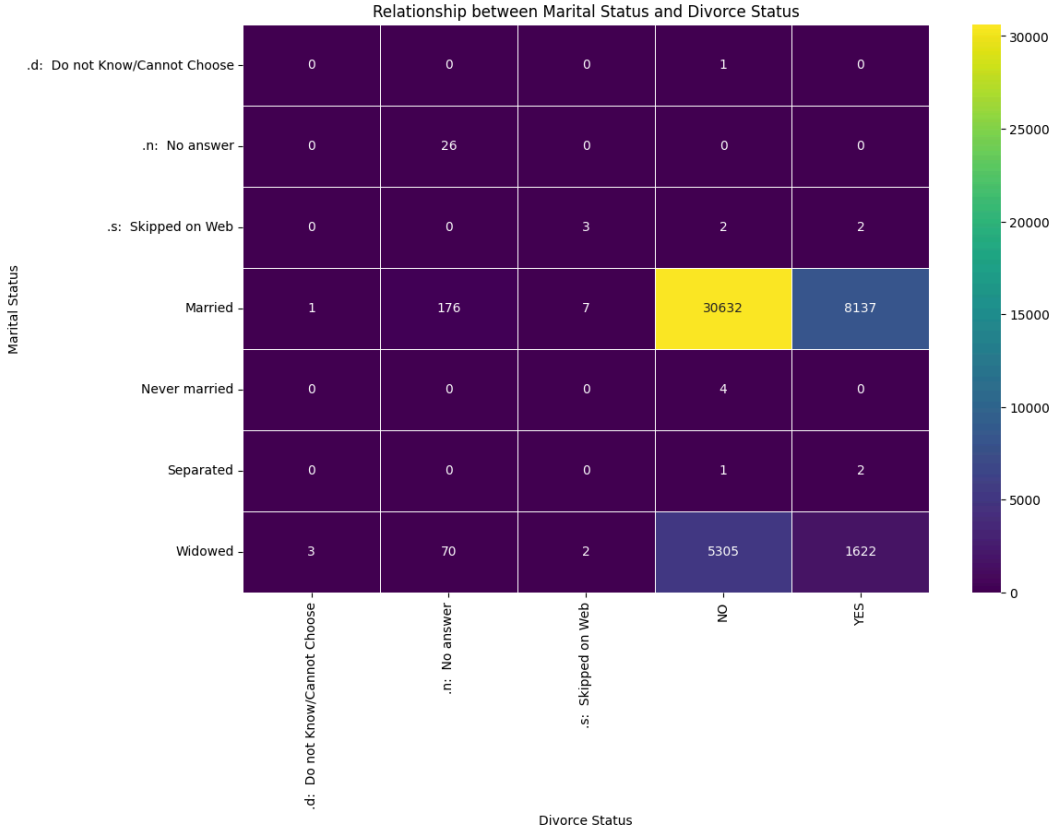


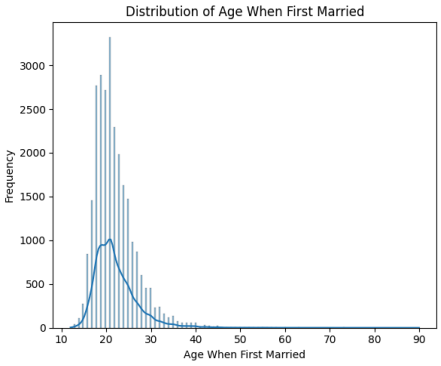


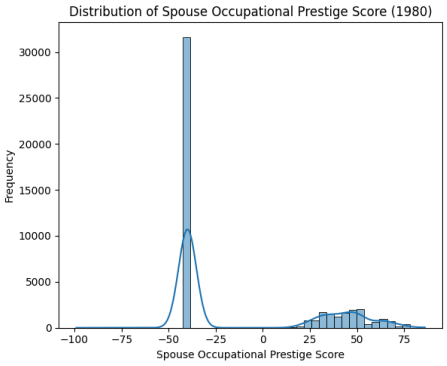
Descriptive statistics for 'agewed', 'sppres80', and 'childs' columns:

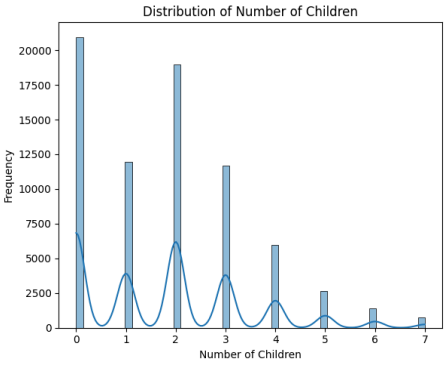
	agewed	sppres80	childs
count	26543.000000	46971.000000	74375.000000
mean	22.150058	-12.256988	1.824847
std	4.885114	40.650909	1.613581
min	12.000000	-99.000000	0.000000
25%	19.000000	-40.000000	0.000000
50%	21.000000	-40.000000	2.000000
75%	24.000000	34.000000	3.000000
max	90.000000	86.000000	7.000000

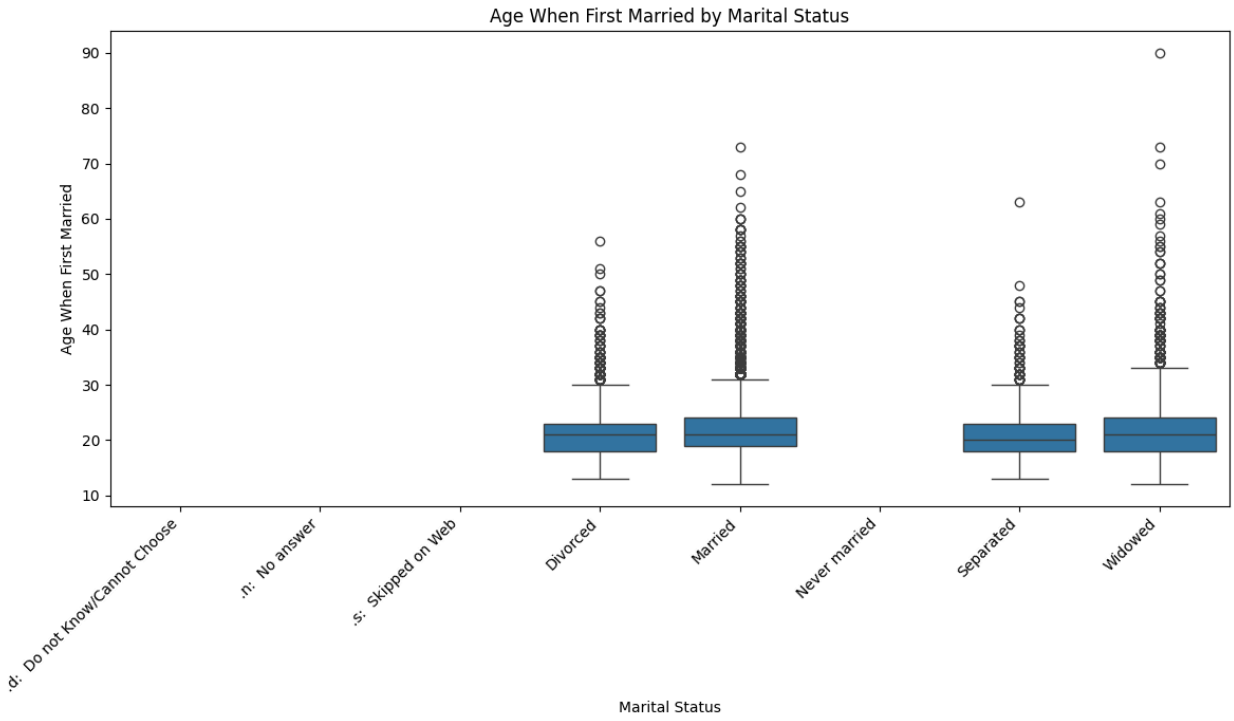

Correlation Matrix of Numeric Variables
agewed
sppres80
childs
year
1.00
0.08
-0.22
0.05
0.08
1.00
0.15
-0.38
-0.22
0.15
1.00
-0.05
0.05
-0.38
-0.05
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Relationship between Marital Status and Divorce Status
Marital Status
.d: Do not Know/Cannot Choose
.n: No answer
.s: Skipped on Web
Married
Never married
Separated
Widowed
0 0 0 1 0
0 26 0 0 0
0 0 3 2 2
1 176 7 30632 8137
0 0 0 4 0
0 0 0 1 2
3 70 2 5305 1622
.d: Do not Know/Cannot Choose
.n: No answer
.s: Skipped on Web
NO
YES
Divorce Status
30000
25000
20000
15000
10000
5000
0


Distribution of Age When First Married
Frequency
Age When First Married


Distribution of Spouse Occupational Prestige Score (1980)
Frequency
Spouse Occupational Prestige Score


Distribution of Number of Children
Frequency
Number of Children


Age When First Married by Marital Status
Age When First Married
.d: Do not Know/Cannot Choose
.n: No answer
.s: Skipped on Web
Divorced
Married
Never married
Separated
Widowed
Marital Status

Describe your findings in 1-2 pages.

Data cleaning was an important part of the process. Missing values were not consistently formatted throughout the dataset. The exploratory data analysis shows patterns in marriage timing, family structure, and spouse characteristics. It also shows that missing data is highly prevalent in the dataset. For instance, many variables are only relevant to certain respondents such as those regarding people with spouses and who have ever been married. Thus, there is a large number of NaN values that likely reflects inapplicability rather than random data problems.

The numeric variables show distinct patterns. “Age when first married” is right-skewed with a mean of approximately 24.8 years and standard deviation of 4.9. This suggests most people marry in their 20s, while a smaller group marries later. “Number of children” is also right-skewed with a mean of 1.9 and standard deviation of 1.6, showing that of the surveyed people, most of them have 0-2 kids and few have larger families. “Spouse's occupational prestige score (1980),” is only available and applicable for married respondents with complete, valid occupational data.

The categorical variables showed expected patterns. Most people that were surveyed are “Married,” followed by “Never married” and “Divorced.” The data also shows that most people have never been divorced, which makes sense since the dataset includes both never married and married respondents. Out of the respondents who are in relationships, most indicated that the spouse works for someone else while self-employment is less common. The “Spouse's occupational prestige score” variable shows that married respondents are generally happy. “Pretty happy” is most common, followed by “Very happy,” with “Not too happy” being rare.

The visualizations show some clear relationships between variables. “Age at wedding” has a negative correlation of -0.22 with “Number of children.” This suggests that those who

marry later tend to have fewer children. “Spouse occupational prestige” also has a negative correlation of -0.38 with survey year, possibly reflecting shifts in who was surveyed over time. Other relationships between variables are not as strong, such as a small positive correlation of 0.15 between “Spouse prestige” and “Number of children.”

For future work, analysis should focus on respondents eligible for each question. For example, one could analyze marital happiness and spouse prestige only among residents who are married rather than applying the same treatment to the whole dataset. This would help further explore how factors such as age at marriage, family size, economic position, and marital satisfaction relate to marital status over time.