Summary: A one paragraph description of the question, methods, and results (about 350 words)

Project Concept: What is the research question? Is it well-defined? Is the data appropriate to address the question? References to data documentation or codebooks are essential in this part of the project. What is the research strategy, and is it appropriate to the research question?

Marital and relationship plays a large role in society - influencing peoples’ daily routines, the values they hold, and even their perceived status. However, these variables have been subject to change over the years. Based on data from 2015, it appears that both marriage and divorce are trending downward after experiencing large fluctuations due to both extreme social change (WWII, rising feminist movement) and gradual change in opinion. Since the GSS provides a robust set of data for a number of years spanning from 1972 to 2022, our group thought it would be appropriate to closely examine the question of how marital status influences their productivity and routines - more specifically the hours they are working each week - at particular points in time. On top of this, to gain a more comprehensive view, variables such as the type of work and the highest level of education earned were examined in regards to relational status as well.

As for our methods, we created separate dataframes, grouping survey participants by their eligibility and marital status. We then were able to compare these different participant groups by creating data visualizations with boxplots, scatterplots, histograms, and kernel density plots. These visualizations show that the participants’ workweeks have largely similar times of commitment, with all people regardless of marital status working an average of around 40 hours per week. It is seen that respondents are more likely to work more than 40 hours per week, as opposed to fewer hours; however, there are no visible trends relating average worked hours with marital status. The range variation provides more insight, with participants having been divorced and are now single having the largest range, while those who divorced and then remarried have the lowest range variation. One interesting relationship is captured in the linear regression plot, which shows that the hours worked in the previous week for divorced respondents is positively correlated with the same variable for their divorced spouse, while there is no such correlation with married respondents and their current spouse. Overall, more research could provide valuable insight into the relationship between one’s relationships and the work that they do.

<https://randalolson.com/2015/06/15/144-years-of-marriage-and-divorce-in-1-chart/>

Data: One to two pages discussing the data and key variables in the analysis, and any challenges in reading, cleaning, and preparing them for analysis

Exploratory Data Analysis and Visualization: Are basic quantitative features of the data addressed and explained?

Out of the variables provided in the GSS, these were ones that we decided were relevant for our research question:

* **year**: the year in which the participant responded to the survey
* **wkstat:** the status of labor that the participant holds
* **hrs1:** total number of hours worked last week
* **hrs2:** estimated average number of hours worked per week
* **evwork:** whether or not the participant has ever worked for as long as a year
* **wkrslf:** whether the participant works under an employer, or is self-employed
* **marital:** marital status of the participant
* **divorce:** whether or not the participant has ever been divorced before
* **spwrksta**: the status of labor that the participant’s spouse holds
* **sphrs1:** total number of hours the participant’s spouse worked last week
* **sphrs2:** estimated average number of hours the participant’s spouse works per week
* **spevwork:** whether or not the participant’s spouse has ever worked for as long as a year
* **spwrkslf:** whether the participant’s spouse works under an employer, or is self-employed
* **spocc10:** occupation of the participant’s spouse
* **spind10:** industry of the occupation of the participant’s spouse
* **degree:** highest level of education attained
* **spdeg**: highest level of education attained by the participant’s spouse
* **ballot**: ballot administered to the participant for the survey

Certain variables were selected because they are essential to our research question of how marital status and hours worked per week are related. Other variables were chosen because they provide supplementary data that could possibly be useful in explaining the relationship between the two. We gathered a comprehensive set of variables, cleaned them accordingly, and then narrowed the set down to key variables that we thought would be best to work with. In our view, it was better to have a fully cleaned dataset with a few unused supplementary variables than to have a dataset that is not fully cleaned with variables that might have to be cleaned further down the line if we ever decided to use them.

The process of cleaning and sorting the variables was standard, just making the data more presentable and suitable for visualization. For each variable, there were values that did not provide any substantial information, such as “Inapplicable” or “Do Not Know/Cannot Choose” in **hrs1** and **hrs2**. For these values, we decided to replace them with null values. Also, for variable pairs that were comparing data between the respondent and their spouse, there were certain values that were conveying exactly the same information, but the values themselves were not named properly. So, we decided to replace those values and match them with each other - for example, changing **spdeg**’s“LT HIGH SCHOOL” values to “Less than high school” to match up with the **degree** variable. Lastly, for the **hrs2** variable, the “Full Time”, “Part Time”, and “Unemployed” values seemed out of place, since they were not classified as numeric. So, we assumed that “Full Time” was 40 hours, “Part Time” was 25 hours, and “Unemployed” was 0 hours, changing them accordingly so that they would match with other values within **hrs2**.