Preliminaries:
There are two tasks for the HIL predoctoral fellowships. First is a mandatory exercise involving General Social Survey data using Stata or R. Second is an optional, **bonus** OCR exercise using Python.

Please return a zip file consisting of:
- a Stata do-file for Task 1.
- a Stata logfile for Task 1.
- a cleaned, final version of the dataset used for Task 1 (.dta file).
- a PDF document with outputs of task 1 (highlighted in grey above)
  - name the PDF file: [lastname]_[firstname])
- a Jupyter Notebook script for bonus Task 2 (downloaded from Google Colab).
- three text files for bonus Task 2.

Task 1
Data Preparation

1. Download data from https://gssdataexplorer.norc.org/gss_data
   a. Specifically use the GSS 1972-2022 Cross Sectional Cumulative Data (release 3a April. 2024)
   b. Open a log file
   c. Make all variables lowercase
2. Create a new indicator variable *vote_pres* that is 1 if the respondent voted in presidential elections in 1972, 1976, 1980, 1984, 1988, 1992, 1996
   a. Replace with 0 if the respondent did **not** vote
   b. Only use the reply from two years after the election cycle ended to avoid recall bias (e.g., use the response in 1982 to the question on voting in presidential election in 1980)
   c. Label all values of the variable.
3. Only keep observations whose responses correspond to whether they voted in 1972, 1976, 1980, 1984, 1988, 1992, 1996.
4. Create a new indicator variable *repub_v_dem* that is 1 if the respondent voted for the republican candidate in the presidential elections.
   a. Replace with 0 if the person voted democrat.
   b. Only use the reply from two years after the election cycle ended to avoid recall bias (e.g., use the response in 1982 to the question on which presidential candidate they voted in 1980)
   c. Label all values of the variable.
5. Create an indicator variable named *male* equal to 1 if respondent identifies as male and 0 if female.
   a. Label all values of the variable.
6. Create a categorial variable name *religion*, using GSS variable *relig* for the following categories: Protestant (1) , Catholic (2), No Religion (3), and Other (4).

    a. Label all values of the variable.
7. Create a categorical variable named *age_cat* for the following age categories (and label values)
    a. 18 to 29
    b. 30 to 49
    c. 50 to 64
    d. 65+
8. Create an indicator variable named *less_highschool* using GSS variable educ.
    a. Code as 1 if less than 12 years of school and 0 if more than or equal to 12 years of school.
    b. Be careful not to include missing values in your variable.

Data Analysis

1. Create a summary table with the statistics of total number of observations, mean, min, max and standard deviation for the variables you created above.
    a. Describe the data structure and comment on the summary statistics.
2. Regression whether someone voted on the following controls adding in the order specified one at a time until all controls are included: religion, age, male, less than high school, and year voted.
    a. Make sure to use provided weights, *wtssall*, and use probability weighting (add *pweight* as an option in your regressions)
    b. All categorical variables should be included nonparametrically (e.g., an indicator for each religious category, year etc.)
    c. Organize output in a table for submission.
    d. Interpret the coefficient on *Catholic* in the regression with full controls.
3. Repeat step 2 immediately above for the outcome of voted Republican vs. Democrat. Organize output in a table for submission. Compare the coefficient on *Catholic* to step 2d) in the above and include in your write-up.
4. In 1979, the US was experiencing high inflation due to an oil crisis and other international issues. This may have moved those with less education towards the Democratic party.
    a. To test this hypothesis, estimate a difference-in-differences regression with Republican vs. Democrat as the outcome and an indicator for post 1979 and less than High school education, be sure to include the main effects of each and the controls listed in 2. And continue using the provided weights, *wtssall*, with *pweight* option.
        i. Create a table of regression coefficients.
    b. Create an event study version of 4a) by interacting each year indicator with less than high school education and using the election year 1976 as the reference year.
        i. Plot the coefficients and standard errors.

Task 2 (bonus and optional)

OCR the gun advertisements from a magazine page scan using Python. Please see Jupyter Notebook on Google Colab for instructions and template. Make a copy of the Jupyter Notebook. When complete download and save the .ipynb file to your submission folder.