Identify a research question similar to questions we've talked about in this course. Use the General Social Survey (GSS) dataset (provided below).

All analysis must be completed using the R programming language via RStudio, and your write up must be an R Markdown document. To help you get started we provide a template Rmd file below (see Rmd template in the Required files section below). Download this file, and fill in each section.

**IMPORTANT:** Analyses completed using software other than R, or not written up using R Markdown, will receive a 0 on the project regardless of their content.

## Required files

- Data - Save this file in the same directory as the Rmd template (provided below).

**NOTE:** If you are using Chrome as your browser you might need to change the .gz at the end of the extension to .Rdata in the file you downloaded.

gss.Rdata

- Codebook - Review this file to find out what each column in the data represents.

gss.html

- Rmd template - You must use this template to write up your project. Save the data and this file in the same directory.

stat_inf_project.Rmd

- Assessment rubric - You might want to review the assessment rubric while working on your project so that you have some idea of how your peers will evaluate your work.

stat_inf_project_rubric.html

## Instructions

Your project will consist of 4 parts:

1. **Data**: (3 points) Describe how the observations in the sample are collected, and the implications of this data collection method on the scope of inference (generalizability / causality). Note that you might will need to look into documentation on the GSS to answer this question. Seehttp://gss.norc.org/ as well as the "More information on the data" section below.
2. **Research question**: (3 points) Come up with a research question that you want to answer using these data. You

should phrase your research question in a way that matches up with the scope of inference your dataset allows for. You are welcomed to create new variables based on existing ones. Along with your research question include a brief discussion (1-2 sentences) as to why this question is of interest to you and/or your audience.

3. **EDA**: (10 points) Perform exploratory data analysis (EDA) that addresses the research question you outlined above. Your EDA should contain numerical summaries and visualizations. Each R output and plot should be accompanied by a brief interpretation.
4. **Inference:** (28 points) Perform inference that addresses the research question you outlined above. Each R output and plot should be accompanied by a brief interpretation.

In addition to these parts, there are also 6 points allocated to format, overall organization, and readability of your project. Total points add up to 50 points. See the assessment rubric (provided above) for more details on how your peers will evaluate your work.

You can begin working on the project immediately. Please save your work as you go along. When you're ready to submit your work for evaluation, remember to click the "Submit" button.

After you submit your proposal, please provide feedback to others on their projects. Please assess at least 3 projects. This peer assessment will not only provide you with experience with a data set and research questions but also prepare for your projects in the future courses in the Specialization.

## More information on inference

**INFERENCE:** Statistical inference via hypothesis testing and/or confidence interval.

- State hypotheses
- Check conditions
- State the method(s) to be used and why and how
- Perform inference
- Interpret results
- If applicable, state whether results from various methods agree

It is your responsibility to figure out the appropriate methodology. What techniques you use to conduct inference will depend on the type of data you're using, and your sample size. All of you should conduct at least a hypothesis test, and report the associated p-value and the conclusion. Those of you comparing two means, two medians, or two proportions should also calculate a confidence interval for the parameter of interest. Those of you working with categorical variables with more than two levels will need to use methods like ANOVA and chi-square testing for which there is no associated confidence interval, and that's ok. If your data fails some conditions and you can't use a theoretical method, then you should use an appropriate simulation based method.

- If you **can** use both theoretical and simulation based methods, then choose one and stick with it. You don't have to do both. However if you **can't**use both, then you need to decide which is appropriate.
- If you **can** do both a hypothesis test and a confidence interval, do both, and comment on agreement of the results

from the two methods. However if your variables do not lend themselves to a confidence interval, that's ok.

- It's essential to make sure the method you're using is appropriate for the dataset and the research question you're working with.

## More information on the data

Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups; to compare the United States to other societies in order to place American society in comparative perspective and develop cross-national models of human society; and to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting.

GSS questions cover a diverse range of issues including national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior.

Source: Duke University Data and Visualization Services

## Frequently Asked Questions

**Do I have to use R for my project?** Yes. While there are other statistical packages and/or programming languages that may be perfectly appropriate for your project, since one of the goals of this course is to learn R, all analysis **must** be completed in R and using the Rmd template provided above. Projects completed using other statistical packages and/or programming languages will receive a 0 on the project.

**Where can I find a list of R commands that might be useful for the project?** Refer to the previous labs and see the RStudio cheatsheets for dplyr, ggplot2, and RMarkdown.

**Who am I writing for?** Write as if you are explaining your results to whomever would be interested in your research question, whether this is another scholar in your field or peers sharing your interest in the topic. This audience may not have taken a statistics course. You must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that anyone can understand.

**Does my project have to be written in English?**Yes, your project must be written in English; this is the only way to ensure that the students who are assigned to review your project can understand it.

**What is a peer assessment?** Peer Assessment is when students in a course evaluate a fellow student's work. First, each student submits an assignment. Then, the students who have submitted an assignment are given other students' assignments to evaluate, according to provided criteria. Finally, each student receives a grade that is based on the other students' evaluations.

**Can I use a paper I've worked on for another course or purpose?** No. Please create a unique project for this course. Do not use your master's thesis, work you have published elsewhere or work you have submitted for another course. In the past, students who have submitted work they used elsewhere were reported as submitting plagiarized work.

**What if I think the project I am assessing has been plagiarized?**

1. Assess the project according to the Evaluation/Feedback directions.
2. Report plagiarism to Courserahttps://learner.coursera.help/hc/en-us/articles/209818863-Coursera-Honor-Code

**How do I avoid plagiarism?** "In an instructional setting, plagiarism occurs when a writer deliberately uses someone else’s language, ideas, or other original (not common-knowledge) material without acknowledging its source." - The Council of Writing Program Administrators Therefore, please give credit for all of the sources you have used. Copying and pasting from a site without giving the source is plagiarism, and will be reported. For more information, see this tutorial on avoiding plagiarism. In your own project, give credit to all sources you used, even if you have paraphrased them or if they are your own work but published elsewhere