# Project 2 Overview

The goal of this project is for you to apply skills you’ve learned about regression and statistical inference (confidence intervals and hypothesis tests) on a real-world issue of your choosing. Like Project 1, you will also need to explain how you’ve demonstrated your command of STS 2300 topics. There are two options for this project:

1. Use the data **project2\_data.csv** (posted in the data folder on our code repository), which includes information on dozens of variables from a random sample in 2022 of 30,000+ US residents. The data is from the National Health Interview Survey, which is conducted annually by the Centers for Disease Control and Prevention (CDC). I do **NOT** want you to use the full 30,000+ observations in the dataset. Instead, I want you to think of a question that interests you and use a subset of the data to investigate your question of interest question. A variable list with brief explanations is provided on Moodle and Github in the sts2300-project2-data-variables.docx document.
2. You may use another data set (or collect one) that can reasonably be considered a random sample from a population of interest. If you have an idea for this option, you must discuss it with me and get approval before moving forward.

For this project, you will work in randomly assigned groups of 3-4 students. You should read the project description and guidelines first, then meet with your group to assign roles and delegate responsibilities to group members.

There will be three main products associated with this project:

1. A report, poster, or presentation that presents your work and results to someone outside of this class who is interested in learning about your topic or question of interest. This person would likely not have strong statistics or data analytics skills, so the product should not include jargon.
2. An RMarkdown file (.Rmd) with associated knitted document (i.e., .html) that explains, alongside your code, what you did, why you did it, and how it demonstrates your proficiency in the data analytics topics mentioned earlier. The intended audience for this product is your professor (or an interviewer) who is assessing your ability to write R code for data management and analysis.
3. An individual assessment worksheet (to be posted later) that demonstrates your proficiency in all aspects of the project, not just the parts you worked on. Because you will be asked about all aspects of the project, you should make sure you understand what the others in your group are doing.

This project will provide the opportunity for you to practice many data analytics and statistical skills:

* Proposing a research question that can be addressed with data analytics.
* Collecting and wrangling relevant data for your question.
* Using R to analyze data.
* Applying regression models and/or performing statistical inference on data.
* Communicating your findings to audiences of differing technical backgrounds.

## Project Tasks

1. Meet with your group to discuss the project and how to divide tasks. I recommend creating a plan of when certain tasks/subtasks must be completed and by whom.
2. Submit a proposal for your project (see Section 1 below). This is **due by the start of class on Friday, April 11**.
3. Present your findings (see Section 2) to both a general audience (product 1) and a technical audience (product 2). This is **tentatively** **due by the start of class on Friday, May 2.**
4. Each student will complete an individual project assessment and peer evaluation **tentatively** **due by the start of class on Friday, May 9.** The details for this assessment will be provided in a separate document later.

# Section 1: Project Proposal (Choose Data and Question of Interest)

Decide whether you will use the project2\_data.csv file or your own data. Then choose a question of interest that will let you do at least two of the following: linear regression, confidence interval, or hypothesis test. The two analyses you choose don’t have to answer the exact same question, but they should be related to a central question of interest.

If you use the project2\_data.csv, I encourage you to focus on a subset of the data as your sample (e.g., 18 – 22-year-olds, people who smoke, people who work over 40 hours per week, etc.). At the end of this document (Section 3), I have included some *possible* topics of interest, but I encourage you to choose something that interests you (whether it is from that list or not).

Once you have chosen your data and a question of interest, complete the project topic form and submit it.

There are many, many different directions you can pursue for this project. **This is not a trivial task**, and it will require thought and discussion with your group. You will need to think about what story you want to tell, what variables you might want to include in your analysis, which observations you might want to include, etc. You do not need to have all the details decided for the proposal, but proper, advanced planning will make your project much smoother and easier to complete later.

**Submission**: A hardcopy of the Project 2 Topic Form (sts2300-project2-topic-form.docx on Github and Moodle) that answers the questions in the form.

**Audience**: The instructor.

# Section 2: Main Products (Article/Poster/Presentation and RMarkdown Report)

Once you have chosen your data set, you will then create two main products: (1) an article, poster, or presentation to someone who is interested in learning about your topic or research question, and (2) an RMarkdown report to your professor or interviewer who is assessing your ability to write R code for data management and analysis.

For this project, you will have to do at least two of the following three things:

1. Create a regression line related to your question of interest.
2. Calculate a confidence interval related to your question of interest.
3. Conduct a hypothesis test related to your question of interest.

## 2.1 Article/Poster/Presentation to General Audience

Your group’s first task is to report your results in way that someone outside of class (likely with little statistics and data analytics background) could understand your findings. There is flexibility in both the format (e.g., article, poster, presentation), and the style and tone of your choice. Regardless of your decision, you need to present your findings in a way that someone without statistics and data analytics knowledge can understand what you did and found. For your product, you must:

1. Introduce the topic, provide background information, and clearly state your question of interest related to it. This should include a citation of the dataset and citations of any other sources you use to provide background information. A typical (written) length for this section would be a full paragraph (~5 sentences) – though it can be longer, it shouldn’t go much past one page double-spaced.
2. Briefly discuss the data and methods you used for the analysis. The reader of this document won’t care about all the details (e.g., variable names), but you can explain the types of things you’re doing and why they are needed. You can choose to cite the data in this section if it works better for your product.
3. Present your results. This should include:
   1. [if appropriate, but this is not to be the main focus of your product] Descriptive statistics and graphs to help orient the reader to your data and better communicate your story.
   2. [if chosen as one of your two analyses] Regression
      1. Calculate and write out your regression line equation.
      2. Interpret the y-intercept and (at least one) slope.
      3. Calculate and interpret the R2 value.
      4. Make a reasonable prediction with your equation.
   3. [if chosen as one of your two analyses] Confidence interval
      1. Calculate an estimate for your parameter and discuss why a confidence interval is needed (as opposed to just using your estimate).
      2. Generate and graph a bootstrap distribution. Describe what it shows.
      3. Use the bootstrap distribution to calculate a confidence interval. You must choose a confidence level.
      4. Interpret your confidence interval in the context of the problem.
   4. [if chosen as one of your two analyses] Hypothesis test
      1. State your hypotheses. Define any parameters as needed. You must choose a significance level you will use.
      2. Summarize your data (with an estimate and a graph).
      3. Generate a null distribution with a graph that includes your p-value.
      4. Calculate the p-value and write a sentence about what it means in context.
      5. Provide a formal conclusion for your test.

1. Provide a discussion on the results and why someone else should be interested. State any overall takeaways as they relate to your question of interest. Discuss possible ramifications your results may have for someone interested in your topic. What new questions do you have? Are there limitations to how someone should use your report? What future students could be done to further explore your topic?
2. Provide your references in a professional formatting style. If you used the project2\_data.csv file, you should the following citation (you can update the style):

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Annie Chen, Stephanie Richards, and Michael Westberry. IPUMS Health Surveys: National Health Interview Survey, Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2023. <https://doi.org/10.18128/D070.V7.3>.

The product can take several forms (article, poster, presentation, etc.). Whichever format you choose, you must include the information listed above. There is no strict page or length requirement if your sufficiently address the points above. Below I give some general recommendations for articles and presentations, but please talk with me if you have questions or other ideas that you’re unsure about.

1. A written article.
   1. You can think of this being written for Elon’s The Pendulum or the Elon News Network in general, though other options (e.g., op-ed) are acceptable if cleared with the instructor first.
   2. The article should be well-written (i.e., no grammar/spelling errors) for a non-technical audience.
   3. The article should have a descriptive and interesting title (i.e., not “Project 1 Report”), and all group members’ names should be listed. If you prefer, you can include a cover page.
   4. The article should be double-spaced, 11 or 12-point font, 1-inch margins, and a standard font (e.g., Times New Roman, Calibri).
   5. The article should be well-organized, though no specific format is required if you include the information listed previously.
2. A recorded presentation.
   1. You can think of this as being posted as a YouTube video on Elon News Network, a podcast, or some other website, though other options are acceptable if cleared with the instructor first.
   2. The presentation should have well-made slides that are easy to read and understand. The slides/talk should be geared towards a non-technical audience.
   3. The presentation should have a descriptive and interesting title (i.e., not “Project 1 Report”), and all group members’ names should be listed.
   4. I recommend keeping your slides simpler (e.g., bullet points, graphs, pictures) and avoiding unnecessary special effects/transitions.
   5. Every group member must speak in the presentation for a roughly equal amount of time.

**Submission:** A hardcopy **and** an emailed copy of your product (poster will be small on a hardcopy – that’s ok; a presentation hardcopy will only be slides).

**Audience**: Someone outside the course who is interested in your topic or question of interest.

## 2.2 RMarkdown Report to Technical Audience

Your group’s second task is to explain to your professor or interviewer what you did, why you did it, and how it demonstrates your proficiency in the analyses chosen. For each topic, you will have flexibility in how you demonstrate your abilities, but you will need to make your case for why what you did demonstrates proficiency.

You must do all coding in an RMarkdown file. This file should, alongside the code, explain your code and decisions. This will likely mirror some of the content from your article/poster/presentation in Section 2.1, but from a different perspective.

Here is a summary of what someone would be looking for:

1. A very brief introduction to the topic that clearly states your question of interest. Give background information about your data set to make sure the reader has a general understanding of the type of variables and information is in your data, how big the data set is, and where it comes from. This should include a citation of the dataset.
2. Data management that is appropriate for your question of interest (e.g., subsetting rows or columns, renaming variables, creating new variables, joining data frames, converting between wide and long format).
3. Results and analysis appropriate for your question of interest:
   * [if appropriate, but this is not to be the main focus of your product] Descriptive statistics and graphs that orient the reader to your data.
   * [if chosen as one of your two analyses] Regression
     + Calculate and write out your regression line equation.
     + Interpret the y-intercept and (at least one) slope.
     + Calculate and interpret the R2 value.
     + Make a reasonable prediction with your equation.
   * [if chosen as one of your two analyses] Confidence interval
     + Calculate an estimate for your parameter and discuss why a confidence interval is needed (as opposed to just using your estimate).
     + Generate and graph a bootstrap distribution. Describe what it shows.
     + Use the bootstrap distribution to calculate a confidence interval. You must choose a confidence level.
     + Interpret your confidence interval in the context of the problem.
   * [if chosen as one of your two analyses] Hypothesis test
     + State your hypotheses. Define any parameters as needed. You must choose a significance level you will use.
     + Summarize your data (with an estimate and a graph).
     + Generate a null distribution with a graph that includes your p-value.
     + Calculate the p-value and write a sentence about what it means in context.
     + Provide a formal conclusion for your test.

1. A very brief discussion on overall takeaways as they relate to your question of interest.
2. At the end of the document, you should include a brief section explaining the contributions of each member in the group. This is standard in many fields and becoming common practice.

**Submission:** A hardcopy of the knitted .html file **and** an emailed copy of the corresponding .Rmd file.

**Audience**: The instructor or an interviewer assessing your ability to code and demonstrate data analytics knowledge and skills

# Section 3: Possible Topics of Interest

You are not required to choose one of these topics. These are only for groups who need help deciding on a topic.

* What proportion of people who have had depression take medication for it?
* Are people who work full time more likely to have had COVID-19?
* What proportion of college graduates are female?
* How fast do 10- to 18-year0olds grow taller? (looking at AGE vs. HEIGHT)
* How does health status differ among people with and without paid sick leave (or workplace health insurance)?
* Are sexual orientation rates different across different regions of the US?
* Are US citizens more or less likely to have graduated high school than non-US citizens?
* Are people who smoke also more likely to drink?