STS 2300 – Project 1: Data Investigation Project

Project 1 Overview

The goal of this project is for you to apply skills you've learned in Notes 01 – 05 to a real-world data set of your choice and to be able to explain how you've demonstrated your command of STS 2300 topics. You must demonstrate your proficiency with appropriate data management strategies, data summaries (i.e., descriptive statistics), and data visualizations to explore a question of interest related to your data.

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 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. Imagine presenting your work at Elon's SURF Day. 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.
- Generating descriptive statistics and graphs to explore the data.
- Communicating your findings to audiences of differing technical backgrounds.

AI Policy Reminder: Based on the syllabus policy, AI tools (e.g., ChatGPT) **may <u>not</u> use AI tools to generate R code**. For text, "you **may not copy** AI-generated text verbatim as responses on…projects." Any code or text that is discovered to be AI-generated and not explicitly cited will (a) be treated as a violation of the Elon Honor Code and (b) result in the lowest marking for the associated portion of the project rubric.

STS 2300 – Project 1: Data Investigation Project

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 **tentatively due by the start of class on Friday, February 20**.

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, February 20.**

4. Each student will complete an individual project assessment and peer evaluation **tentatively due by the start of class on Friday, April 3.** The details for this assessment will be provided in a separate document later.

Section 1: Project Proposal (Choose a Data Set)

To start, you should choose a dataset that you plan to use. One option is to pick one from the Tidy Tuesday website (Note: Scroll down to see the list of datasets. You can click on the tabs for other years to see more data). You are also welcome to use datasets from other sources, but you should let the instructor know if you plan to do this.

In choosing a dataset, you should pick something that is complex enough to allow you to demonstrate your data management and data visualization skills sufficiently. Ideally you will have both quantitative and categorical variables for summary and graphing. You will be required to make a table of summary statistics and two different types of visualizations; keep this in mind when thinking about your data.

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: An .Rmd file (and knitted .html file) that demonstrates you can read your data into R and do one or two small things with the data (e.g., explore the structure, make a simple graph, summarize a variable, etc.). A template .Rmd file is available on the course Github page and on Moodle.

Audience: The instructor.

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

Once you have chosen your dataset, you will then create two main products: (1) a 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.

2.1 Poster or 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. You can imagine you are giving a presentation at Elon's SURF Day. You have two options for this (poster or presentation), and there is flexibility in the style and tone of your choice. However, for both, 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 both options, you must:

1. Introduce the topic, provide background information, and clearly state your question of interest related to it. This should include citations of any you use to provide background information. A typical (written) length for this section would be a full paragraph (around 5 full sentences).

2. Briefly discuss the data and methods you used for the analysis. The audience 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 should cite the data and R in this section.

3. Present your results. This should include:
    a. Descriptive statistics. Create well-formatted and visually appealing numeric summaries and tables appropriate for your question of interest (e.g., describing the center and spread of numeric variables for different groups, showing the count or proportion of observations within different categories). You must include **at least one** table containing descriptive statistics. Though you will need to use something like the kableExtra package to make a table in your RMarkdown document, you can choose to reformat the table in your poster or presentation using other software to make the table look better.
    b. Data visualizations. Create **at least two** different kinds of visually appealing and informative graphs using the ggplot2 package that are related to your question of interest. Your graphs should have appropriate labels, themes, scales, etc.

4. 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 studies could be done to further explore your topic?

5. Provide your references in a professional formatting style.

The two options for your product are:

1. Poster
    a. The poster should be well-written (i.e., no grammar/spelling errors) for a non-technical audience.
    b. The poster should have a descriptive and interesting title (i.e., not "Project 1 Report"), and all group members' names should be listed.
    c. I will provide poster templates, but the dimensions should be 48 inches by 36 inches (the standard size for SURF posters). You are welcome to look at other templates online or use ones you have seen before. To make your poster easier to read, I suggest avoiding being too text-heavy (e.g., someone should not have to spend more than around five minutes reading your poster).
    d. The poster should be well-organized, though no specific format is required if you include the information listed previously.

2. Recorded presentation
    a. 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.
    b. The presentation should have a descriptive and interesting title (i.e., not "Project 1 Presentation"), and all group members' names should be listed.
    c. I recommend keeping your slides simpler and avoiding unnecessary special effects/transitions. I also recommend keeping the total slide number under 15 (a good pace is around 1 slide per minute).
    d. Every group member must speak in the presentation for a roughly equal amount of time. The total presentation time should be **between 7 and 12 minutes.**
    e. The presentation should be well-organized, though no specific format is required if you include the information listed previously.

<u>Submission</u>: A hardcopy of the poster (on a regularly sized piece of paper) or slides **and** an emailed copy of your poster or presentation.

<u>Audience</u>: 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 data analytics topics from Notes 01 - 05. Imagine you are being interviewed for a job, and the interviewers want to know that you can use R/RStudio and RMarkdown to do data wrangling, summary statistics, and visualizations. They give you the dataset and question (which happens to be the one you chose for Section 2.1); they want you to demonstrate with evidence your R proficiency.

STS 2300 – Project 1: Data Investigation Project

You must do all coding in an RMarkdown file. This file should explain your code and decisions. This will likely mirror some content from your product in Section 2.1, but from a different perspective. 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. Typically, this means more than a couple lines of code using only 1 or 2 things learned about that topic.

Here is a summary of what you should include:

1. Alongside an informative title (which could be the same as Section 2.1), list the authors and a (very) short statement of the goal or purpose of the document.

2. Data wrangling 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). Alongside your code, you should include in RMarkdown explanations of what you did and how it demonstrates proficiency in data wrangling.

3. Results and analysis appropriate for your question of interest:
    a. Descriptive statistics. Create well-formatted summaries and tables appropriate for your question (e.g., describing the center and spread of numeric variables for different groups, showing the count or proportion of observations within different categories). You must include **at least one** table containing descriptive statistics using the kableExtra or janitor packages (or something similar). Alongside your code, you should include in RMarkdown explanations of what you did and how it demonstrates proficiency in creating numerical summaries.
    b. Data visualizations. Create **at least two** different kinds of visually appealing and informative graphs using the ggplot2 package that are related to your question. Your graphs should have appropriate labels, themes, etc. Since the goal is to demonstrate proficiency, make sure to go beyond the basics when creating your graphs. Alongside your code, you should include in RMarkdown explanations of what you did and how it demonstrates proficiency in creating visualizations.

4. At the end of the document, you should include a brief section explaining the contributions of each member in the group, including the contributions to the poster or presentation from Section 2.1. This is standard in many fields and becoming common practice.

5. Your script (.Rmd) should reproduce your results from both Sections 2.1 and 2.2. I will check for this by running your files. I recommend having someone else run your final script to ensure it runs **without modification**.

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

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