Some Guidelines for STAT 542 Projects

-It should involve the intake and analysis of a real data set.
-I want to leave you a lot of leeway for the type of data set it is, but it should be something you are interested in, so that you can raise interesting questions about the data and make sensible conclusions based on your analysis.
-In terms of where to find data: There are government and financial data available on public websites that can be found with a Google search. Searching for "historical weather data" brings up options for past weather data. We also have learned how to scrape data from html tables on websites like Wikipedia. Major sports like baseball, basketball, and football have websites with large amounts of current and historical data, such as the collections at
https://www.sports-reference.com/
and many other websites.
There are plenty of places with large, ready-to-use data files like
https://vincentarelbundock.github.io/Rdatasets/datasets.html
and
https://archive.ics.uci.edu/
However, I'd rather that you pick a data set not because it was easy to import into R, but rather because you had a genuine interest in it. If you don't have a particular interest in your data set and the information you get from it, this will show in your report and presentation.
-One thing to keep in mind is that if there is a lot of work needed to import and preprocess the data (or even to merge more than one data set), then that can be a large part of the project. On the other hand, if getting the data set ready to analyze is fairly trivial for the data set you choose, then more will be expected of you in terms of visualization, summarization, and data analysis.
-The book covers the analysis of text data in Chapter 19 and the analysis of geospatial data in Chapter 17-18. It is very possible to choose a data set of one of those types if you'd like. However, we won't be getting to those chapters until the end of the semester (if at all) so you would have to read ahead for ideas about how to import, manage, and summarize data sets of those types. This would be an impressive type of project to undertake, if you would like to do so!
-The project should involve importing the data and doing some type of analysis to reveal information about the data set. This sort of analysis should include exploratory tools like plots and graphs and numerical summaries and tables. This course doesn't focus on more advanced analytic tools like regression, classification, statistical inference, etc., so these

tools are not necessary, but if you have previous knowledge of these and wish to include them in your project, you may. But it’s not at all a requirement.

-Each group will turn in (electronically) a typed report, the main report being 3-4 pages in length. The main report should include

-Some background about the data set, including descriptions of what the individuals represent and what the variables are

-Some justification about why the data set is interesting or important to study

-A description of the source of the data set (i.e., how it was collected and where you found it)

-A description of any steps that you needed to take to import, preprocess, or transform the data to get it ready for analysis

-A brief description of your choice of graphical visualizations and data summaries you did (and why you chose to do those)

-Most importantly: Present your findings! What did you learn about the data from your analysis? Provide relevant graphs, tables, etc.

-If you include several interesting graphs, you don’t have to count the space for those graphs against the page limit. But think carefully about what graphs you present and don’t overwhelm the report with dozens of graphs: Make sure the ones you include tell us something interesting.

-In addition, each group should turn in a text file (you can use R Markdown or Quarto if you already know those, but this is NOT a requirement by any means) containing the R code used to carry out the project work. This text file should include DETAILED COMMENTS so that an intermediate-level R user could understand what is being done in the code.

-Finally, each group will make an oral presentation in class to present to their classmates (and me) what they did and learned about their data. You can prepare slides (PowerPoint or pdf or Google Slides) and use the classroom computer to display them. Each group member must speak during the group presentation. Aim for about 12 minutes for your presentation, followed by about 3 minutes for questions. We will do up to 3 presentations in a 50-minute class period. These will be done during the last week of classes, and I will set up a schedule for these in advance.

-Be aware that group members should communicate well with each other during the progress of the project and that each member should contribute substantively, both behind-the-scenes and for the oral presentation. I will be sending a FULLY CONFIDENTIAL form to every group member, asking each individual to assess whether all the group members made a substantial contribution, or whether one or more group members did not put in a good effort. The answers to this form will be *for my eyes only* and in some situations, could affect the project grade of students who did not put in a strong effort for their group.

-The project overall will count 8% of a student’s course grade for STAT 542. I’ll explain later the breakdown of the project grade, but it is likely to include components for: the written report; the oral presentation; brief comments you will make on your classmates’ presentations; and your initial proposal of your project data set.