Midterm project:

Using what you have learned about R, prepare a dataset for analysis. The data for this assignment are from OSHA and were originally collected by NICAR. You can get the dataset exactly as it came to me on a CD from NICAR by cloning it from my GitHub:

https://github.com/havilandw/MA615-midterm-spring-17

Assume that you are preparing the data for analysis that will be used to report on the most dangerous places to work in Massachusetts. You don't need to do the analysis. You are just preparing the data.

You will need to read, clean, and organize the data using the principles and examples we have discussed in class. You should refer to Wickham's Tidy Data article. Your work must be repeatable. The steps you take and the code you use must be documented so that an independent third party could reproduce your steps and obtain identical results. Code you write must be well-commented.

Be sure to consider missing data. Will missing data be a problem in the planned analysis?

Explore the data, visualizing aspects of the data as necessary for you to explain what you did, why, and what you observed as you did your work.

You may ask me questions as you work. If you get help from anyone else, you must give them credit. This is not a project with a single right answer. If you talk with your classmates, give them credit in your report – and remember, there is not a single right answer. "Me too" work gets "me too" grades.

What you submit on blackboard should be just a link to the GitHub repository that contains your work. I will clone your repo, review your presentation, run your code, and read your code. Please make sure that I can do all three. Do not make assumptions about libraries I have.