STA 101 Spring 2018
Project I - Due Friday, May 5^{th} **by the end of lecture**

Read the following instructions carefully:

- **You may work by yourself, or in a team of two people total.**
- **You are not allowed to discuss the questions with anyone other than the instructor or TA.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (no comparing answers), posting the questions to homework help sites, etc.**
- **You are allowed to use or modify your previous code, or the instructors code that is posted online.**
- **Formatting will be a significant portion of your grade for this take home portion. There should be an appendix of code, and no code, or R results (results that are directly copied and pasted from R with no additional formatting) in the body of the report.**
- **Your report should be in full paragraph form.** You are allowed to have tables, and/or use R Markdown, but it should have clearly labeled sections. You can also use Word or Google docs (or latex) if you are more comfortable with those.

The Dataset

The dataset we will be working with is: `HospFull.csv` which describes characteristics of United States Hospitals. The source of this data is the text: *Applied Linear Statistical Models, fifth edition, Kutner, Nachtsheim, Neter, and Li.* It has the following columns:

Column 1: Length: The average length of stay of all patients in the hospital, in days

Column 2: Infect: The average estimated probability of acquiring infection in hospital (in percentages)

Column 3: Culture: Ratio of number of cultures performed to number of patients, times 100

Column 4: Bed: The average number of beds in hospital during study period

Column 5: MedSchool: `Y` if the hospital was associated with a medical school, `N` not

Column 6: Region: Geographical region, with categories `NE` (North East), `NC` (North Central), `S` (South), `W` West.

Your response variable (your `Y`) should be either `Length`, so that we are predicting the length of stay of a patient, **or** Infect, so that we are predicting the estimated probability of infection. You get to choose, **and you should only consider one or the other, not both**. You should have one response variable, and 5 explanatory variables.

3. The Report Format

The Goal: The goal is to fit the best linear model, where you may choose if your goal is prediction, or model correctness. You should write up a full, paragraph form report on your findings, which should include the following sections:

I: Introduction: A small introduction about the goal, what data you are using, and what model you are using.

II: Summary. This should include summary plots of describing the relationship between your explanatory and response variable, and any numerical summaries you find interesting.

III: Data Preparation. This section should include finding and removing any outliers in preparation for a model. How many outliers you found (if any) should be noted, and the rows with those outliers in them should be shown.

IV: Model fitting. This section should include the results of your model fitting, including which model selection criteria you used, what your final model was, any confidence intervals or hypothesis tests you will interpret in a later section, and the estimated regression line. If you have a categorical variable, consider writing down the separate models it suggests. For this section, it is your choice whether or not to include interaction terms. Or, you may first determine which single terms are important, then see if any interactions to do with those terms are also important.

V: Model Diagnostics: Perform diagnostics to see if the assumptions of linear regression hold. If you do not think they do, state this, but continue with the report.

VI: Interpretation: Interpret the coefficients and any confidence intervals or p-values that you calculated.

VII: Prediction: Use your model to predict the following (note, if you did not use all of the X variables you may ignore them):

- •(If you chose this question): The length of stay for a hospital with Infect= 4, Culture= 14, Bed=190 , MedSchool= No, Region= W
- (If you chose this question): The probability of infection hospital with Stay= 8, Culture= 14, Bed=190 , MedSchool= No, Region= W

VIII: Conclusion: One or two sentences on what variables you found were most important to your model, and how they affected your outcome.

4. Details

Your report should be the following format:

i. Typed.

ii. A title page including your name/s, the name of the class, and the name of your instructor (me).

iii. Treat each question as a small, stand alone report. Then staple them together at the end.

iv. Double-sided pages.

v. An appendix of your R code used to produce the results. Do not include in R code in the body of your report.

For example, your project should be put together in the following order (stapled):

```
Cover Page
Parts I-V
Code appendix
```

Feel free to make your cover page "unique" so that it is easy to find when I hand them back.

Notice: your project will be graded as a group effort (if you have two people). This means that you are responsible for your own work, and your partners work. I will not assign two different grades to one project.