**Stt 3851: Project #1**

**Project Goal:** Working with a group of no less than two people total and no more than three people total (unless otherwise cleared with me), you will work with a data set suitable for a regression analysis, perform some regression analyses on the data, check underlying assumptions, and report your final results.

**Details:** The project will be accomplished through a sequence of steps. The final product will be an R Markdown document. We will introduce the use of the GitHub version control software as part of this project to facilitate revisions.

**First Project:**

Step 1: Reproducible Research Technology Setup

Reproducible research refers to a process by which a research project is organized and managed in a way so that another person can reproduce the results. It involves the organization of data input files, the storage and evolution of code files that are written to conduct data analyses, and project output, including figures, results, and written summaries of those results.

We will use the free software Github in our course, which will allow you to save your output in a repository (called a repo for short), make changes on your computer to the files in the repository, and then have those changes uploaded to the master project on GitHub. This will take some practice (at first) but there are strong benefits to working this way.

To setup your GitHub account, do the following:

* Sign up for a free account on [GitHub](http://www.github.com) (github.com)
* When you register for a free individual GitHub account, request a [student discount](https://education.github.com/) to obtain a few private repositories as well as unlimited public repositories. Please use something similar to **first\_name\_last\_name** as your username when you register with GitHub. For example, my username on GitHub is *RossGosky*. If you have a popular name such as John Smith, you may need to provide some other distinquishing characteristic in your username.
* Once you have a GitHub account, send an email to [goskyrm@appstate.edu](mailto:goskyrm@appstate.edu) with a Subject line of **STT3851 - GitHub Username**, and tell me in the body of your email your first name, last name, and your GitHub username. I will then manually add you as a team member to the repository in the organization for our course. This repository will be where you store all of your work for this project. If I can, I will also change your repository to a private repository.

Getting things set up so that you are able to use your GitHub repository for this project will take a little bit of upfront work. What your group should do is:

* Select a team leader for your group. Your group work will be stored in this person’s repository for this project. Let me know the team leader for your group.
* Once your team leader is determined and I’ve created a repository for them, I’ll work with the team leader to ensure that your project work can be pushed out to the GitHub repository, and that I am able to view your work in the repository.
* I would like to get this setup this week for each team leader if at all possible.

Step 2: Data Analysis and Writeup

For this project you will create a multiple regression model for a set of housing data from Oregon. The data will be sent to the team leader The variables in the data are:

* id: case number
* price: price of the house, in thousands of dollars
* size: size of the house, in thousands of square feet
* lot: a lot size variable; it measures from 1 to 11, with larger numbers indicating larger lot sizes
* bath: number of bathrooms
* bedrooms: number of bedrooms
* yearbuilt: the year the house was built
* agestandardized: calculated as (yearbuilt – 1970)/10
* garagesize: number of cars that can fit in the garage (0, 1, 2, or 3)
* status: categorical 🡪 sld = sold; pen = pending sale; act = active listing
* elem: elementary school district

The response variable for your model is the price of the house. Any and all of the explanatory variables, including transformations, nonlinear terms, and interactions between those variables can be included as predictors.

I suggest the following sections for your model building:

* Data Summary: Examine the statistics and values for each variable. Are any missing? Do any values need clarification or modification? If so, why and what did you do?
* Exploratory Data Analysis: Examine some of the variables relationships with price to help you determine which variables might be useful in an initial model. Explain your conclusions from this initial screening.
* Initial Modeling: Using your conclusions from the exploratory data analysis, build a regression model and report your findings.
* Model Modification: Consider modifying your model based upon your initial results and/or diagnostic plots from the model. Explain any modifications that you are making. Consider variance inflation factors for each predictor in your model and comment on that in your model selection process.
* Conclusions: Present your final model and diagnostic plots in support of that final model. In that presentation of the final model, comment on the R-squared value and its interpretation, give 95% confidence intervals for each of the β coefficients in your model, and illustrate your model’s use with a 95% confidence interval for the mean response and a 95% prediction interval for individual response for a hypothetical house of your choosing.

Writing Recommendations:

* All sections should be numbered. All section headings should be labeled appropriately.
* All equations should be written inline using R Markdown. Ask me if you have questions about how to do this.
* Explanations of output are required. You do not have to explain every number in a section of output, but output without **any** appropriate explanation will result in a lower grade.