Statistical Methods II (MATH 375) Final Project

- **Data Set Submission Date: Wednesday, May 1 at 9:00 AM to jcwagaman@wcu.edu**

  Your data set must have prior approval. You must analyze a data set with a single response variable and at least three potential explanatory variables. Your response variable should be a quantitative variable or a binary response, include at least a complete sentence in the e-mail message that states your response variable is [some column in the data] and your potential explanatory variables are [at least three columns in the data, not including response variable]. Send an e-mail with an attachment of your data set by this deadline, and I will approve/disapprove by Friday 4/28 at 11:59 PM (hopefully by class time on Friday). Data options for final project are below; if I do not approve your data, then you will analyze the home sales data from Project 2. Data set submission is worth 3% of your final grade.

- **Slide and R Code Submission Date: Wednesday, May 8 at 9:00 AM to jcwagaman@wcu.edu**

  You must submit and use electronic slides in your presentation using Powerpoint or Beamer. Submit an electronic copy of your slides as one file (.pdf or .ppt or .pptx) and your R code as a separate file (.R or .txt) to my e-mail address by the deadline. An outline of the slides is below. Slides and R code are worth 12% of your final grade. **THERE IS NO PAPER – ONLY SLIDES AND R CODE**

- **Presentation Date: Wednesday, May 8 at 12:00-2:30 PM in Coulter 301**

  You will make a presentation to the class (order selected at random) during this time block in conjunction with your prepared and submitted slides. Your planned speaking time should be 5-7 minutes, not including questions. Presentation is worth 10% of your final grade.

Outline of talk and slides:

1. Introduction
   (a) Describe why you are interested in this data.
   (b) Describe the variables under consideration and explicitly state your response variable and your potential explanatory variables.
2. Model Selection with Training Data: First, you will select a training data set comprising 80% of the data set's observations selected at random using the last 3 numbers of your student ID number. Then do the following steps in some order using ONLY your training data:
   (a) Present scatterplots which show potential associations (or lack thereof) among variables.
   (b) Use a stepwise selection procedure or an all-subsets selection procedure to obtain a subset of explanatory variables.
   (c) Identify whether any observations are outliers and/or influential and potentially refit the model.
   (d) Evaluate the model assumptions based on residuals.
3. Model Confirmation with Validation Data: You will use a validation data set to confirm your model using the remaining 20% of the data that were not previously selected to fit the your final model from the previous step.
4. Conclusion: Summarize what you learned and future directions or improvements.