MSAN 601: Linear Regression Analysis
Regression Case Study

Due: Friday, 10/6 at 9:00 AM

In this regression case study, you will be applying the ideas you learned in class to analyze the value of residential homes in Ames, Iowa. A major component of online companies like Redfin (https://www.redfin.com) and Zillow (https://www.zillow.com) is their ability to develop estimates for the cost of a house currently on the market. In this study, you are asked to think like an employee at one of these companies. In particular, you will use regression analysis to complete two major tasks:

- Explore the features of houses in in Ames, Iowa and build an explanatory regression model that describes which aspects of a home most strongly affects its value.

- Build a predictive model that most closely forecasts the value at which the house will be sold.

[**From kaggle.com**]: *Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.*

Directions: Work in groups of 4 (with students of your choice). As a group, you are tasked to write up a formal business report that accomplishes the two parts of this study in Sections 2 and 3. You are to turn in two separate files: (i) a knit .Rmd file with all the R code used to conduct any analysis done in the study but with no plots included, and (ii) a business report with plots, suggestions, and analyses that fully answer the questions posed in Sections 2 and 3.

1 Description of Data

The dataset you will be analyzing in this study is saved in the `housing.txt` file on the github site, and a description of each variable in the data set is provided in the `Data_Description.txt` file on the github site.

The data contains 1460 observations with 79 variables, including a house ID and the sale price of each house.

2 Part I: Explanatory Modeling

Task 1: Your first task is to determine what features of a house are *most* relevent in determining its expected sale price. Build an appropriate regression model to answer this question, and provide any exploratory plots and analyses that may be useful in your discussion. Discuss in detail your findings based on the fit of your regression model and model choices. Be sure to formally test *all* assumptions that are made for your final model. If assumptions are *not* met, then perform appropriate transformations of the data to ensure that a regression model is justified. Furthermore, be sure to consider outliers and points of influence that may be affecting your analysis.

Task 2: A customer, Morty, comes to you wondering if you can help him figure out the sell price of his house. The information on his house is available in the `Morty.txt` file on github. He has already been told from another firm (as shown in the data), that his house will sell for \$143,000. Based on the features of his house, provide Morty a maximum value of what you think he could possibly sell his house for as is. Justify your reasoning. Furthermore, based on the model that you built above, provide Morty the 3 most significant changes that he can make to his house to increase the value of his home before he puts it on the market and explain your reasoning.

3 Part II: Predictive Modeling

You are given a contract to develop the “best” regression model for predicting the sale price of a house for new homes on the market. Your pay for this job will be based on how close you are to the actual final sale price of a new home. Using the data in `housing.txt`, come up with the best predictive model you can. Consider OLS, Ridge regression, Lasso regression, as well as Elastic Net and consider how to go about variable selection when prediction is the primary goal. Write the final model and justify your model. In this study, you do *not* have to verify the validity of assumptions for your model as inference is not your key goal here, but you do need to justify any choices of variable selection and values of tuning parameters.