Statistics 3301: Course Project
Due Wednesday, December 5th by end of day (11:59pm)

The data analysis project is based on a real estate data set, described below. The time/effort required to complete the data analysis project is intended to be similar to that of about two challenging homework assignments.

Project Goals
The general goal of the data analysis project is to demonstrate competence with the regression modeling and data analysis methods we have discussed this semester. Specifically, the goals of the data analysis project are to

1. Show how the various variables are related to the sale prices of houses and to each other by performing an exploratory data analysis.
2. Take a large set of potential predictor variables and turn them into useful regressors.
3. Use the concepts discussed in class to build a model that can be used to predict the sale prices of houses that is complex enough to be useful but that is simple enough to be interpretable.
4. Use the data and your fitted model(s) to describe how the square footage of a house is related to sale price in the context of the other variables in the data set.
5. Convey the results of the analysis in writing in such a way that the technical details are clear while maintaining readability and interpretability.
6. Produce visualizations of the data and fitted model(s) that are professional and readable and that convey useful and correct information.

Group Project
You will work in groups of three to complete the data analysis project. All group members are expected to contribute to all aspects of the data analysis, including the writeup. You will need to indicate your group partnership on Carmen by **Friday, November 9, 2018** (details on how to do this will follow).

Data Set
The data set contains information on approximately 1600 individual residential properties in Ames, Iowa. There is one response variable of interest (sale price of each house) and 8 potential explanatory variables. The data set, called `ames_housing_data_PROJECT2018.csv`, is available on Carmen; the file `ames_housing_data_description_PROJECT2018.txt` contains a description of all 9 variables.

Logistics

1. The project is due on Wednesday, December 5 by end of day (midnight).

2. The completed project will consist of a write-up and `R` code (details for both are provided below). The write-up should be submitted in **pdf** format to Carmen by one of the group members. The `R` code should be submitted as a text file (or .R, .Rmd file) to Carmen as well.

Analysis Requirements
The project consists of the following data analysis requirements. Address each requirement as thoroughly as possible. Guidelines for formatting your answers are given in the *Write-Up Requirements* section below.

AR-1 [20 points] Perform an exploratory analysis where you investigate the pair-wise relationships between the variables `SalePrice`, `LotArea`, `YearBuilt` and `SqFt`. The sale price of houses is our variable of interest, so focus on relationships where one of the variables is sale price (although do report on any interesting relationships between the predictors). Use appropriate (and readable) graphs to help convey this information. We will be building linear models where sale price is the outcome of interest; **make sure you explore whether any transformations of the response and/or the predictors are useful/necessary for building linear models. Use any such transformed variables in the rest of the analysis.**

AR-2 [30 points] Using appropriately transformed versions of the variables `SqFt`, `LotArea` and `YearBuilt`, build a regression model to predict the appropriately transformed `SalePrice` of houses. You may explore interactions and/or other higher-order terms (quadratic terms, etc.) if they seem warranted, however be sure to avoid overfitting and avoid using superfluous higher-order terms.

After finding a reasonable model, report the fitted model and use the estimated parameters to describe how square footage is related to the response in the context of the model. Make appropriate plots to help visualize this relationship.

Use the residual diagnostic techniques we have discussed this semester to help guide your process. Make and report on appropriate residual plots to assess the fit of your final model. The model may indicate some lack of fit. If the lack of fit can be easily remedied using the data at hand, try to do so. Otherwise, just be sure to indicate what aspects of the model do not fit well.

There is no one correct answer to this component of the project, but simply fitting a single model with no discussion will not receive full credit. Experiment with a few different models using the predictors above and report on your process.

AR-3 [20 points] Starting with the final model fit in AR-2 above, update the model to include other regressors related to the variables `Bedrooms`, `Rooms`, `FullBath`, `HalfBath` and `Style`. You will need to make decisions about how to handle these predictors (continuous vs. factor, creating new variables, etc.).

Report the final model you decide on and the process you used to arrive at that model.

Describe how square footage is related to the response in the context of this model. Make appropriate plots to help visualize this relationship.

Use the residual diagnostic techniques we have discussed this semester to help guide your process. Make and report on appropriate residual plots to assess the fit of your final model. The model may indicate some lack of fit. If the lack of fit can be easily remedied using the data at hand, try to do so. Otherwise, just be sure to indicate what aspects of the model do not fit well.

AR-4 [10 points] Say a newspaper reporter asks you to estimate the value of a typical house in Ames, Iowa (a pretty vague question!). Use your analysis and the final model you fit in AR-3 above to formalize the question (i.e., make it specific enough to answer) and provide an answer, along with a description of uncertainty about your answer (i.e., a range of plausible values). Make sure you are very clear about how you are performing the calculations required to answer the question.

Up to [20 points] will be awarded for the quality and clarity of the write-up (see details below).

Total points: 100

Write-Up Requirements

1. Your write-up should have five sections: one section for each of the four analysis requirements (AR-1 through AR-4) above and a fifth Appendix section.

2. Sections 1-4 of your write-up should contain at most 6-7 pages and be written in complete sentence/paragraph form. Use these sections to explain what you have done to answer the questions. Only include the most important figures, equations, tables, etc., in these sections. All R output should appear in the Appendix, along with supporting figures, etc.

3. Section 5 should contain supporting figures, R output, etc., that does not appear in the main part of the write-up. Label the figures/plots and any tables/`R` output in your appendix so that you can reference them in the main text (e.g., Figure 1, etc.). Do not include extraneous information that is not discussed in the main part of the writeup. The material in this section should be formatted in way so that it is easily readable.

4. Make sure all figures look professional. Axes should be appropriately labeled. The plots should be easily readable and sized appropriately. (You do not have to use `ggplot2`; base `R` graphics are fine.)

`R` Code

You will need to turn in a file with your `R` code on Carmen. Make sure your `R` code works and that you have cleaned it up enough for it to be clear to me what you are doing. Comments (# in `R`) help. Delete extraneous code or code from analyses that didn't make it into the final write-up. I should be able to follow the `R` code without too much trouble.