MATH 40853 – Project 2 Spring 2023
Dr. Nelis Potgieter Due: 4/28 (Fri) @ 6pm

The file *housing_data.csv* contains data collected from a sample of residential properties sold in Ames, Iowa between 2006 and 2010. A new start-up company *Zillot* wants to develop a statistical model for predicting the sales price of a house based on readily available predictor variables. For each home in the sample, in addition to the selling price (`SalePrice`), the following variables are observed:

- `LotArea`: the total area of the lot in square feet
- `OverallQual`: overall material and finish quality of the house, on a scale of 1 (Poor) to 10 (Excellent)
- `OverallCond`: overall condition of the house, on a scale of 1 (Poor) to 10 (Excellent)
- `YearBuilt`: the original construction date of the house
- `CentralAir`: whether the house has central air conditioning (coded as ”Y” or ”N”)
- `Flr1SF`: the total square footage of the first floor of the house
- `Flr2SF`: the total square footage of the second floor of the house
- `Fireplaces`: the number of fireplaces in the house
- `GarageArea`: the total square footage of the garage
- `YrSold`: the year the house was sold

Zillot asks you to develop a multivariable regression model for predicting the selling price of a home in Ames. They emphasizes model parsimony, meaning that they prefer models with fewer variables over models with many variables. They ask you to develop three candidate regression models. All models should have **seven or fewer** slope parameters. In constructing these models, you may want to consider the following during your model building process:

- You may want to evaluate whether log-transformations are appropriate for some of the variables, including `SalePrice`.
- Provide visual and numeric summaries of all the variables. Consider which variables are numeric, and which variables are categorical.
- Use stepwise selection for at least one of your models. Start with a model using only a single variable, and then use an appropriate metric to determine whether or not to add another variable to the model.
- At least one of your models should utilize one of the categorical variables `OverallQual` or `OverallCond`. Note however, that you should not include categories with fewer than 50 observations as predictors. Think of a way to handle this without losing any data.

- The data provides total square footage of the first and second floors. You should calculate two new variables based on these two numbers: a variable recording total square footage, and a variable recording whether a home is single- or two-story. At least one of your models should include these two variables, along with an interaction term (product of the numeric and binary variables).
- Develop at least one model that does not use square footage as an explanatory variable.

You are furthermore asked to summarize the process you follow in performing the data analysis in a short report. This report should also include your (statistical) evaluation of the estimated models, including appropriate summary information. For all of your models, report adjusted R^2 as well as one of AIC or BIC. Comment on which model is preferred using the two different metrics. Also perform a residual analysis for each of your models. In your report, carefully discuss your thought process as you build these models. Also make a recommendation to *Zillot*. This recommendation should also consider how *Zillot* could improve their model by discussing potential variables not included in the model, and how you anticipate these would affect price prediction.

Your report should not be less than three pages and should not exceed six pages of typed content with font size 12. This page limit includes any tables and figures you report. Thus, before including any figure or table, ask yourself if it is of benefit to the report/reader. Assume that your reader has little statistical knowledge. **Also know that you are free to ask questions of the senior statistician (Dr. Nelis Potgieter, who happens to teach at TCU), but beyond that the work should be your own. No collaborations will be allowed.**

Include the `R` code you use for the analysis as an appendix to your report. Please *annotate* your code. Points will be deducted from code that is not annotated and is difficult to read/follow.