Hunters Green is a gated residential community in Tampa about 10 miles northeast of USF. We have actual data on all homes sold in this community during the period 2015-2020. Download the data set “HuntersGreenHomeSales.xlsx" posted on Canvas. Your job is to create statistical models to predict: (1) agent days on market (adom), which is essentially the number of days it took to close the sale from date of listing, and (2) sale price of home (pricesold) based on relevant attributes in this data set.

You have to do a bit of research to understand what features are relevant and what are not relevant for the prediction. For example, the two primary kinds of roofs are shingle and tile; tile roofs tend to be a bit more expensive than shingle-roofs. You may have to write some code (in R) to extract and transform relevant features. You don't want to overcomplicate your model with too many factor variables. At the same time, you don't want to leave out important variables from your model either. There may be some highly correlated variables, which may introduce bias in your model if you include them all. You will have to create two sets of models for each of the dependent variable (DV) of interest. For each DV, please proceed as follows:

1. Create a table of relevant predictors, hypothesized direction of effect (+/-), and rationale for each hypothesized effect. (2 points)

2. Run three reasonable models for each DV. You may have to do some feature engineering before running the model. Present each model and summarize their output in a compact manner using stargazer. (2 points for feature engineering + 1 point for modeling)

3. Select the best model from each set and examine whether it meets the assumptions of the regression model. Which of the five regression assumptions are met for the final models? (2 points)

4. Using your best models, select the top three predictors of adom and pricesold, and explain their marginal effects on the dependent variables. Remember that we are interested in economic significance, not statistical significance. (2 points)

5. Aesthetics: Whether your analysis is presented in a nice, compact, summarized manner, without overburdening the reader with unnecessary details or analysis. (1 point)

Include your answers to the above questions in a Word document (or PDF file) and submit using the assignment link. You do NOT have to submit any R code file. Please be succinct, compact, and to the point. Please follow good statistical practices, and avoid common amateurish mistakes such as:

- Using kitchen-sink models (using every possible variable as a predictor).

- Dropping important variables from the model because they are statistically non-significant.
- Trying to artificially inflate adjusted R-square values.
- Running many random, arbitrary models to pick models that you think are "good".
- "Over-engineering" the data (e.g., converting numeric data into categorical factors).
- "Under-engineering" the data (e.g., using the provided data as-is without spending any effort to reduce the feature space or combine data into fewer meaningful features.
- Ignoring alternative model specifications, such as interaction effects or quadratic terms.

The points allocated to each question is indicative of the importance of that question. The midterms and final exams will carry similar point allocations. The PAGE LIMIT for this assignment is 5 single-spaced pages, and you will be penalized one point for each page you go over this limit. You will also lose points if you do not answer all questions asked within this page limit. This is an assignment not just on running regression models in R, but also on how well you understand what you are doing and why and how well you can explain your work in a succinct and comprehensive manner.

Every home needs one full bathroom, but having more than one will increase the homes desirability.