Problem: While Real Estate is typically regarded as a sound investment, primarily because it is difficult for it to depreciate, and has strong potential for growth due to the potential rise in the value of the statistical area in which it resides; the United States is rife with examples of small bungalows built five decades ago in markets that have taken off selling for five to six figures, the market can often be unpredictable, but mainly in the direction of how much the property may appreciate. In turn, there is of course conventional wisdom on the matter, larger houses selling for more, the land's proximity may affect its value, however this cannot be taken as a perfect model. Size may lend itself to diminishing returns. More holistic matters such as noise, smell, and view may complicate the matter of proximity and define neighborhoods better than simple proximity to a primary business or desirable district. The good news is that this is not an unsolvable problem. Statistical modeling can aid in clarifying these problems and improving our understanding of the situation.

Case Study: Ames, Iowa: Ames is a strong candidate market to explore the possible nuances of the housing market. It is a mid-sized college town, with a population estimate of approximately 66,424 in 2021, as well as the presence of Iowa State University, Iowa's largest higher learning institution. As a result, there is a strong, consistent, and stable demand for housing, offering an excellent case study in residential real estate, as it is a stable market with fairly inelastic demand. Moreover, the market is fairly balanced between the buyer and the seller, making it a strong choice for speculation, where the optimizing individual can expect to be on both sides of that equation. With an average home price around $338,000, the moderate investment is also ripe for experimentation on that level as well. The market, with recent graduates working in startups, the presence of tenured professors, as well as a reliable supply of newly-minted upperclassmen or graduate students seeking affordable short-term housing. Both house flipping and rental properties are viable forms of housing speculation, and while they may seem disparate in concept, the underlying traits desirable in both are often quite similar, as on a fundamental level, both engage simply with the demand of people seeking a place to live.

The Model Itself: The initial dataset can be rather intimidating, with a massive 79 distinct variables associated with the Ames Housing Market. Thankfully, data can easily be cleaned with the right hyperparameters. To begin, this model ruled out those variables without a significantly high correlation with the dependent, or target variable, which was the eventual sale price of each individual house. To wit, the six variables demonstrated to have at least a 50% correlation, positive or negative were used.[1] Joining those six was a single qualitative or "dummy" variable, which framed the year remodeled as a binary, "0" if the last remodel was before the year 2000, and "1" if after. The other significant hyperparameter is the "training" vs. "testing" data set, using the standard split of 20% vs. 80%. In doing so, a random sample of the data can be taken, and can in many ways become more "representative" of the dataset than the dataset itself. With these, a more specific model of housing in Ames can be implemented.

[1] In some cases, a variable can have too high of a correlation with the target variable,thus telling us very little about the relationship between it and the target, but as the correlations here never rose above a moderately high 79%, this does not apply here.

Model Selection: For this analysis, multiple models were considered, but two in particular stand out as illustrations of the challenge model selection presents: Random Forest and Ordinary Least Squares. Random Forest, the first of the two models to be practically tested for this analysis, is one that creates a model tailored to the data it observes, optimizing it based on decision trees around the regression models it develops, and "voting" on the best model. Surprisingly, this model proved somewhat ineffective, as after learning on the training data, the model engaged with the test data at a success rate of only 30% of 1% of the observations. This is likely a consequence of overfitting, that is, the model was given information in the training data that proved less than representative of the data ultimately used in the test case. This is known as "overfitting." In turn, the second model, the simpler ordinary least squares, simply takes all relevant observations, and sums their relationship to the average of all observations of that variable to determine its impact on the target class. This proved strongly relevant, as the OLS model's own accuracy measure, the R-Squared, demonstrates a 67% accuracy rate.

Model Findings: Better news yet emerges in the regression itself, as all variables showed a significant impact on housing prices, and all except for one meeting statistical significance on at least an 85% confidence interval, and two that showed absolute certainty of a causative relationship between those two. Of these, the most notable are that of Garage Area and the house's overall quality rating. Also of note is the relationship between the recency of a remodel and its sale price. While a more recent remodel seems to lower the sale price in total, those remodels added after the year 2000 were associated with a significant rise in sale price. This implies that a greater recency is helpful to a re-sale, and that a remodel may eventually fall out of fashion. In conclusion, this implies an expansion of the garage, and any action which may improve the overall rating of the house are to . To determine profitable actions, consider the impact for each variable on the sale price, and keep costs below the resultant number.