**IST 718 Final Project**

**Team:** Bryan Crigger, Jason Rhodeman, Sydney Rickard, Grace Roehrig

**Topic:** Housing Data

**Summary and Recommendation:**

This project will focus on analyzing housing data to determine what factors influence home price. Various models including regression predictions, decision trees, and neural networks were utilized to predict home values based on these factors. In short, the recommendation is for Zillow to utilize the neural network model to predict home price based on its features. For homebuyers, we recommend keeping in mind that the square footage tends to be the biggest contributing factor to home value.

**Introduction:**

Purchasing a home is one the biggest financial decisions many people will make during their lives. Consequently, most people utilize various resources throughout their home buying process to make an educated decision. Our project aims to act as Zillow to help provide homebuyers with additional resources and valuable insight into trends in the housing market. Our project seeks to answer a key question central to the homebuying process: What characteristics of a house tend to have the biggest impact on house price? Providing an answer to this question will allow buyers to determine what they value most in a home and determine the most efficient way for them to spend their money. By providing customers with additional tools and insights during their home buying process, our real estate website will be able to stand out amongst competitors.

**Specification:**

The overarching problem our project is attempting to solve is that many prospective homebuyers lack the resources to truly make an educated decision on what may very well be the largest financial decision they make. Popular websites often show statistics on the home and price history of the property, but they do little to educate the buyer on what characteristics of the home are influencing the price or what features make a home a sound investment. Not only will this analysis allow Zillow to stand out as the premier house-hunting website, but it will also empower buyers to feel more confident about their decision. We have decided to narrow the scope to only include single family homes to make the project more manageable in our limited time frame.

To solve this problem, our team has identified various questions to investigate and analyses to perform. Our investigation will focus on key home features such as price, square footage, bedrooms, bathrooms, stories, presence of a basement, air conditioning, and others.

We will determine which features of a home add the most value, which will also allow us to make recommendations to homeowners as to what they can expect as their return on investment should they choose to add features before selling their home. This could be useful for people who are looking to move and sell their current home, or people looking to flip homes. Our team believes this project will give Zillow a competitive advantage by providing buyers with a superior toolset in their process.

As an initial hypothesis, we predict that bedrooms and bathrooms will be key in determining home value. We hypothesize that more bedrooms and bathrooms will result in a higher home price. We also think that homes located in preferred areas will be more expensive.

**The Data:**

Our team used the housing dataset from Kaggle for our analysis. The link can be found below:

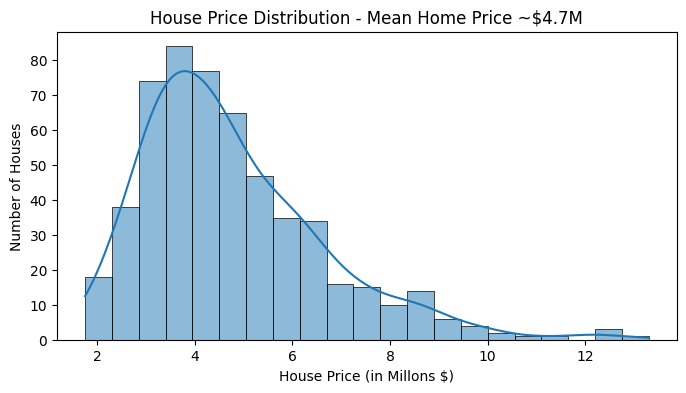
<https://www.kaggle.com/datasets/yasserh/housing-prices-dataset>

The dataset contains information about home prices and features for 545 different homes. We will have access to information including price, area, bedrooms, bathrooms, stories, guestrooms, basement presence, air-conditioning presence, and more. It is also important to note that the price range for the given dataset is between 1.75 million and 13.3 million which will account for attributes like higher mean squared error and averages.

**Cleaning the Data:**

The dataset our team chose to work with was relatively clean to begin with in that it was formatted and organized logically and consistently. However, as with most projects, we had to do some cleaning and formatting to have it ultimately in a form that we could perform our analyses on. We had one team member clean the data so the rest of the team could work with a consistent dataset.

The first step we took in cleaning the data was to create a column for the price in millions, which was calculated by dividing the initial price by one million. This allowed for better displays and graphics to be created based on price distribution. The distribution of the initial housing price data can be seen below:

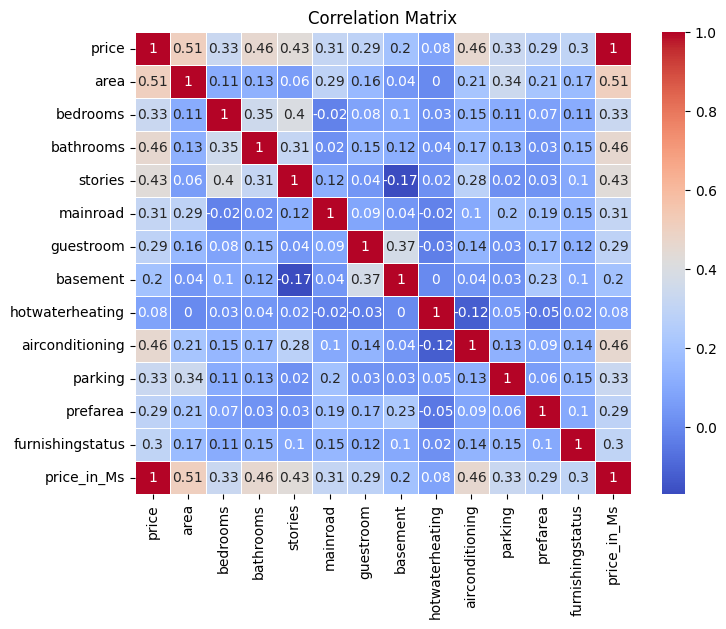


Next, it was opted to drop any houses that were outlier data points. This was accomplished by removing data points that fell significantly above or below the interquartile range. This step will help us create a more accurate model by not training the model with data that is an outlier in the dataset. There were only 15 houses removed as outliers, which brought our number of homes remaining in the dataset to 530.

The final step in cleaning the data was to convert any of the categorical variables to binary or integer values. The columns that were converted to binary included “main road”, “guest room”, “basement”, “hot water heating”, “air conditioning”, and “prefarea”. The variable “furnishing status” we converted to integers since there were three categorical attributes. This step will allow any models to include the presence or absence of these features in its calculations.

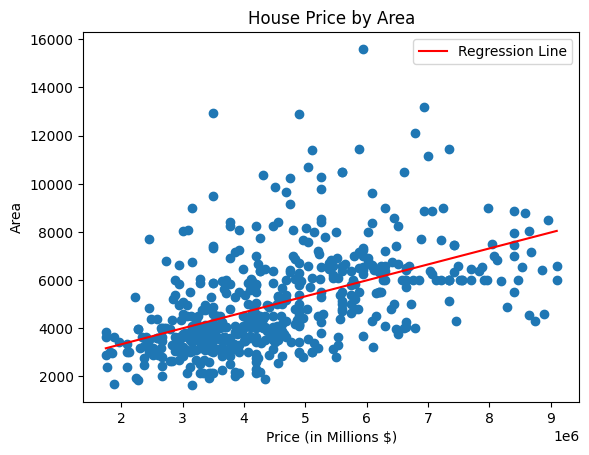
**Data Exploration and Observations:**

The next step in our project was to thoroughly explore the data to better understand any trends that may be present. The data exploration process was performed on the cleaned dataset that had the outliers removed. We started this process by creating a correlation matrix. An image of the correlation matrix can be seen below:



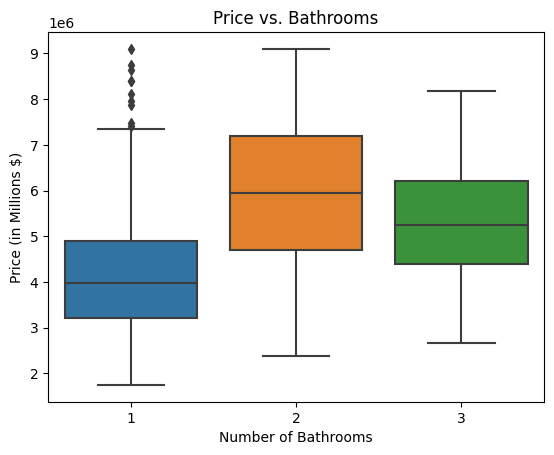
As shown in the matrix, price is most strongly correlated with area, bathrooms, and the presence of air conditioning. It is least correlated with hot water heating and the presence of a basement. The lack of correlation to the hot water heating of course may be due to the fact that most homes do indeed have hot water, limiting its ability to influence price.

Next, the team created a scatter plot to visualize the relationship between area and price. It is shown below:



Unsurprisingly, larger homes tend to be more expensive, though not exclusively. Based on the data points and regression line we can observe a general trend, but clearly many more factors in addition to square footage help to determine home price.

Another exploration completed was showing the range of home prices broken out by the number of bathrooms. The boxplots of number of bathrooms is shown below:



From the box plots we can see that homes with 1 bathroom tend to be cheaper than those that have 2 or 3 bathrooms, with a handful of exceptions shown by the outliers. Additionally, it seems like homes with 2 bathrooms are on average more expensive than those with 3 bathrooms in this dataset. We do not know why this is but we can speculate that maybe homes with 2 bathrooms tend to be nicer or perhaps the bathrooms are larger so less are necessary.

**Modeling: Regression:**

Now it is finally time to start creating models. The first set of models will be based on regressions. The data must be split into training and testing datasets. The training set will train the model to use the variables to calculate home price and the testing set will test the model created by the training set to see how it would perform on new data. The model will be based on the three variables most closely correlated with price. The tricky part of this will be accounting for multicollinearity, or variables that are correlated with one another.

The first model created included all the independent variables. It produced an R-squared of 0.655. This is not a convincing R-squared value. In addition to this, the large condition number indicates that there is likely strong multicollinearity.

The second model created included all the variables except bedrooms, but produced a similar R-squared and ran into the same issue with multicollinearity.

To better understand the collinearity issue, the variance inflation factor (VIF) was calculated for each of the independent variables. VIF is used to detect the severity of multicollinearity. The higher the VIF value, the more likely the variable is highly correlated with other variables. The VIF’s suggested that area, bathrooms, stories, and access to a main road were correlated. Apart from main road access this makes sense as a larger home tends to have more bathrooms and often more stories. We continued to remove variables based on VIF to test different variations of the model, however there was no significant improvement.

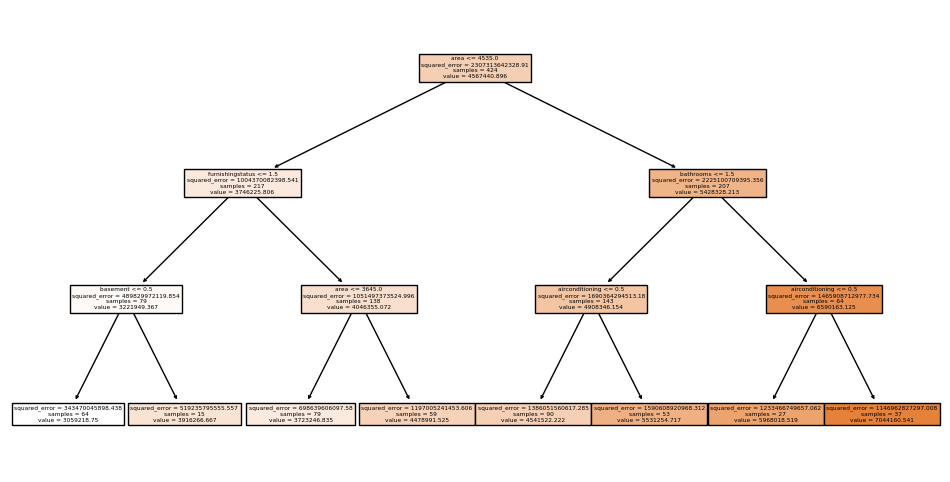
Looking at the R-squared and VIF’s, we can determine that none of the regression models are particularly accurate.

**Modeling: Random Forest:**

Another modeling technique chosen was random forests which is a powerful and often popular modeling technique for regression tasks such as this one. Considering our smaller data set, random forest appeared to be a good option in avoiding overfitting. The results of the random forest model were that area and bathrooms had the highest feature importance, and a mean squared error of 1,266,873,572,918.27 which is a mediocre accuracy.

**Modeling: Decision Trees:**

The next modeling technique implemented was creating decision trees. Again, the data was split into testing and training data and pruned to find the smallest possible mean squared error which arose with a max depth of 3 nodes. The decision tree is shown below:

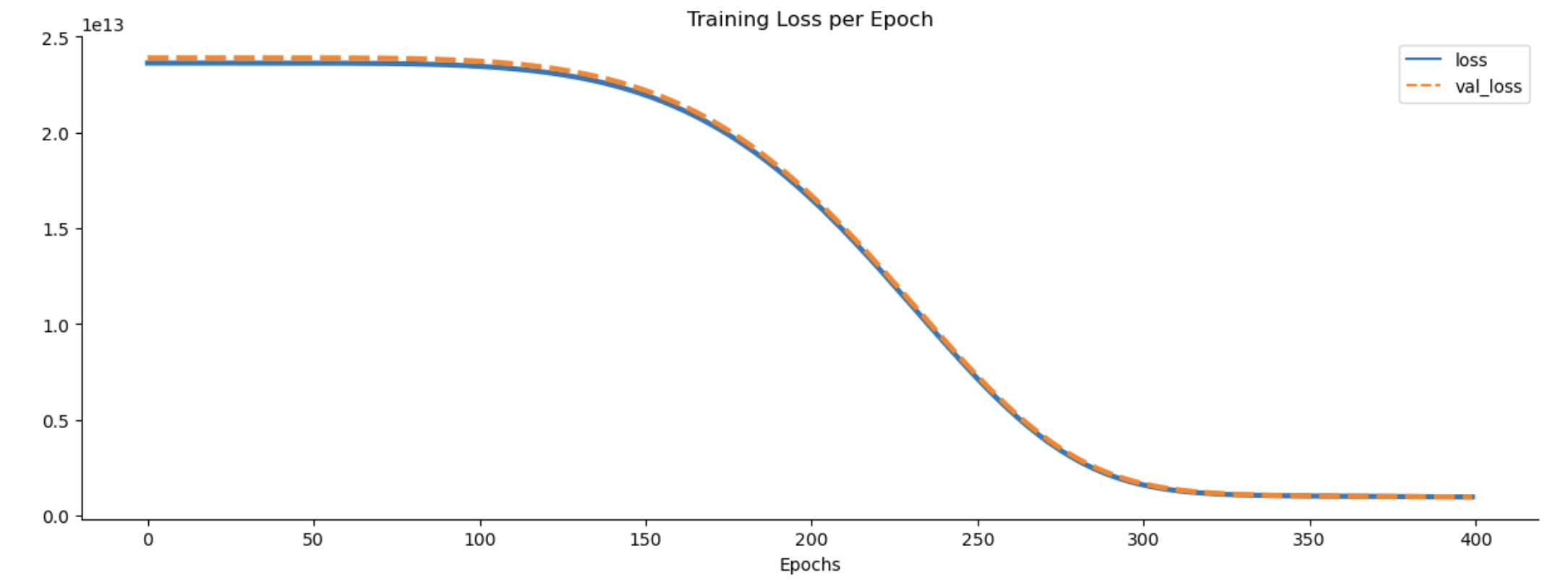


The decision trees did not produce an accurate model with Mean Squared Error of 2861643268348.62.

This model was chosen because a decision tree format would be helpful as a homebuyer who could easily see what criteria they meet and could then estimate what their house was worth or what they could sell it for. Additionally, decision trees are helpful for regression tasks and was a suitable option for this task of using regression on housing prices.

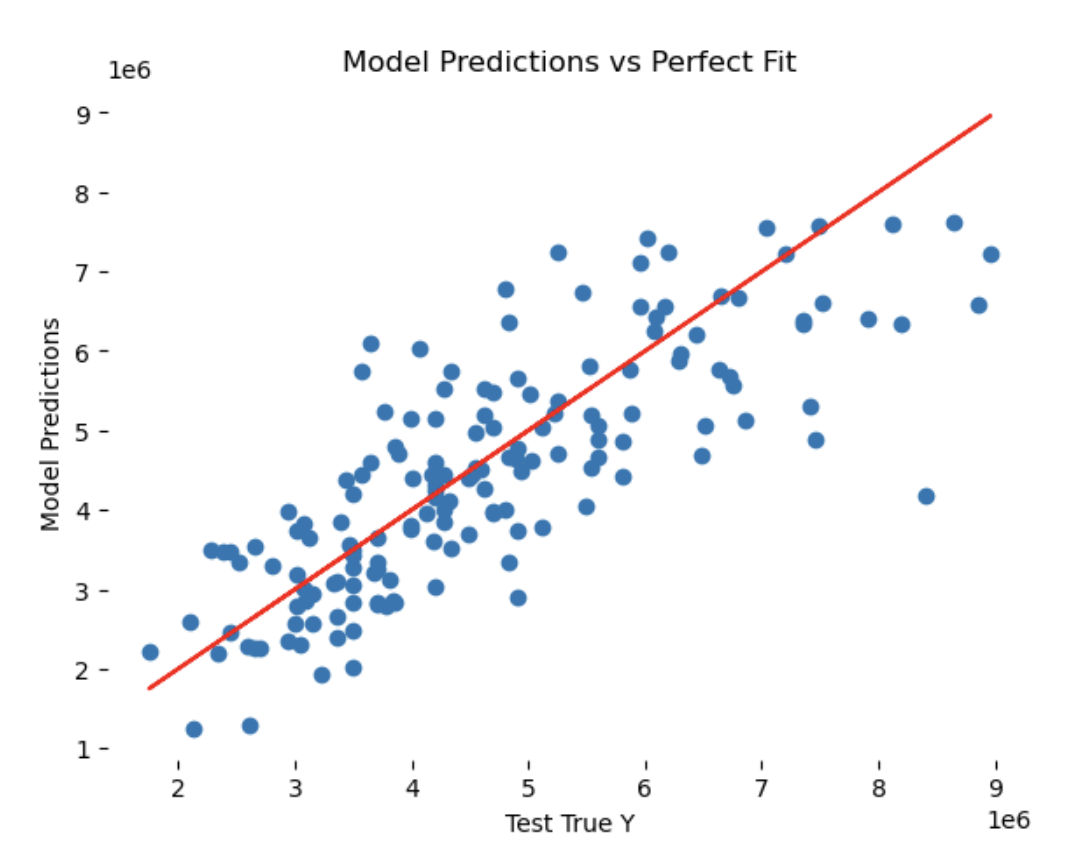
**Modeling: Using Keras:**

The next modeling technique tested was creating a neural network model using Keras. Again, in this method the first step was to split the data into testing and training sets. From there, the data was scaled to fit between 0 and 1. The model was then fit to the training data using 400 epochs. We created a plot to show training loss per epoch. As both lines go down at the same time, it suggests that we do not have an issue with overfitting. The plot is shown below:



The next step is to use the test set to generate predictions. From the predictions generated we will be able to obtain mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and a variance regression score. Our MAE was 763,165. This means our predictions were off by an average of $763,165. When you consider the mean home value in the test data set is $4.6 million, it means the prediction is off by about 16.5%, which is not great. The variance regression score indicates how much variance is able to be explained by the model with 1.0 being perfect. Our variance regression score was 0.60, which is again not great but could be worse.

Next, we wanted to look at the model predictions vs the perfect fit to hopefully gain more insight. A plot of this is shown below:



From the plot we can see that the points that stray farthest from the model’s perfect fit line tend to have values over $6 million. We could try altering the dataset to only include values up to $6 million to see if this improves our results.

**Conclusions and Recommendations:**

As this project demonstrates, there are many factors that determine the value of a home and there are no straightforward answers. Due to the nature of the data in home buying, much collinearity is present. This makes sense because many features are often correlated with one another (think a larger home tends to have more bedrooms and bathrooms).

However, we were still able to build various models to predict home value based on certain features. This will help guide home buyers to determine which features add value to their home as well as determining their priorities within their budget. Of all the modeling techniques we employed in this project, we identified neural network models as being the most effective to accomplish this goal. This model provided the highest accuracy and consistency in predicting home value based on features. It had the highest accuracy with approximately 16% error.

Our team recommends that Zillow move forward with this model as a baseline to continue developing its tool for prospective homebuyers. This will allow them to stand out amongst the competition by having an in depth tool to predict home value and demonstrate to buyers what features add the most value to a home. As for buyers and homeowners, we recommend keeping in mind that square footage will usually be the largest driver of value in their home. This means that their budget will greatly affect the size of home they can afford. Also, if they are looking for a long term investment opportunity, they can buy a home with a sizable square footage and add upgrades later as the area of the home will continue to drive its value.

As a final note, it should be observed that despite our best efforts no model truly captured the goal with complete accuracy. Our group completed additional research during this assignment and were unable to find any models from other sources that had exceptionally accurate predictions for the housing market. One future improvement that would likely increase the accuracy of the model would be to include more houses to the dataset. The original dataset included only 545 homes. Although we were able to get some insights from the data, increasing the number of homes to train the model on would likely provide more accurate results for all of our models. Though they can be improved by including more specific data, predicting home prices will always be difficult due to constantly changing market conditions and consumer preferences.