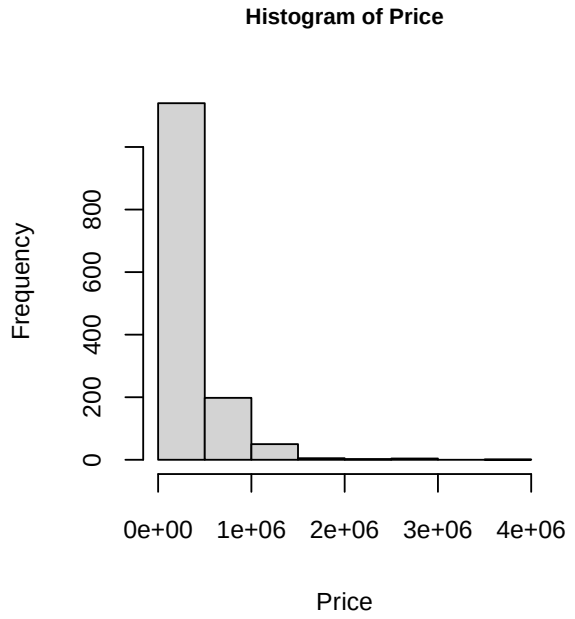
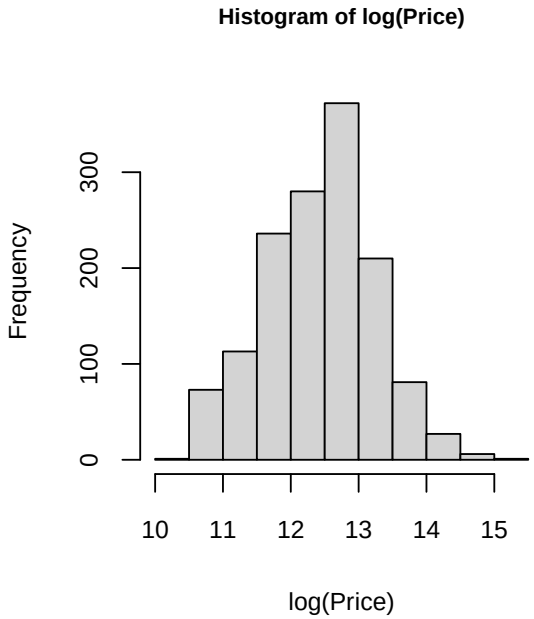
STAT 1361 Final Project Technical Report by Clare Cruz

The first step in developing the PA-VA Realty house pricing model is to explore the historical data set of houses. At the beginning of the exploration, the 17 variables are separated into different lists so they can be appropriately summarized and further analyzed. After sorting the variables, the data set has one response variable (price), one unique identifier, six categorical predictor variables, and quantitative variables. With these variables sorted, a summary is performed for the quantitative variables, and a table count is produced for the categorical variables. The quantitative summaries show evidence of outliers. With some properties having high prices, room counts, lot area, and square feet. This skewness is further highlighted by the extreme right-sided tails in the boxplot distributions. In the qualitative tables, there are also signs of outliers. Very few properties have concrete and log exterior finishes, only one property is a mobile home, and some zipcode areas have very few property counts. Additionally, the fireplace variable is the only variable that has blank values. While the remainder of the data did not stand out, it is noteworthy to mention that the data set is evenly distributed between VA and PA (687 and 713, respectively). Additionally, a correlation matrix shows moderately strong positive correlations between the response and total rooms, bathrooms, and square feet. There are also some moderate positive relationships between bedrooms, bathrooms, total rooms, and square feet. These correlations are unsurprising given that they all relate to the size of the property. Lastly, the variance inflation factor shows that state and zipcode are likely to have multicollinearity. But, this aligns within the data's context since they both account for location.





In the exploratory analysis, there were some data oddities mentioned that required action. Firstly, the fireplace variable had NA's for nearly half of the data points. Further analysis showed that every property in Virginia did not have data for fireplaces. The fireplaces in PA are recorded as 0 for no fireplace and integer values for every fireplace afterward, so it was impossible to fill in or infer the missing values. As a result, the fireplace column is excluded from the data set. Next, the summaries and boxplots showed several outliers

and skewed distributions. To further identify these observations, leverages, studentized residuals, and Cook's distances are calculated with the full linear model. No properties are found to have a high Cook's distribution (>0.5) and only two observations have a high leverage value. These two high-leverage observations are excluded from the data set. Moreover, several properties have a studentized residual value above an absolute value of 2 (51 properties). Since 51 data points are substantial, only the residuals above an absolute value of 4 are excluded from the data set. At the end of the outlier investigation, eight observations are removed from the data set. Also, there is only one property that is a mobile home. While this was not perceived as an issue initially, the leverage calculations conclude that this mobile home property is a high leverage value. Nonetheless, the test set does include a mobile home property, so the data set includes this observation. Lastly, the distribution of price is significantly skewed to the right. A log transformation is performed to mitigate this effect in the modeling process. A Shapiro-Wilk test showed that this transformation did not make the response variable normally distributed. But a histogram of the response shows that the log transformation significantly improves the distribution of the model (see graph above).

There are several models that are tested in this analysis. These models are grouped and summarized below:

Model Group	Model
Linear Models	Forward, Backward, Best Subset Selection
Shrinkage & Regularization	Lasso, Ridge Regression, KNN, PCR, PLS
Nonlinear Models	Polynomial, Splines, GAMs
Trees & Ensembles	Regression Tree, Random Forest (RF), Bagging RF, Boosted RF

The model's predictive accuracy is measured using mean squared error. Based on MSE, the Trees and Ensembles Group perform the best, with the random forest model performing the best overall. In this model, the most significant variables were five out of the 14 variables: the year built, square feet, bathrooms, lot area, and zip code. The importance of these variables is measured using the out-of-bag importance measurement with MSE. While every modeling method produced a different result, square feet, zipcode, lot area, bathrooms, and year built were the most common variables among all the tree-based methods.

In this data set, there were three significant challenges. The first obstacle revolves around the missing values in the fireplace variable. The details of the problem and the solution are highlighted earlier in the report. It was challenging because it was complex to figure out the pattern to the missing values and understand the effects of excluding the variable. The second major obstacle is with the zip code variable. Because zip code is a categorical variable, it was converted into a factor. However, turning the 45 zip codes into a factor variable produced many predictor variables since each of them becomes an indicator variable. This problem was critical for the tree function because it is limited to how many factors can be included in the model. The problem was also present for the PCR and PLS modeling functions. As a solution, the zip code variable was deleted from the data set for those three models. The last problem was figuring out how to build the polynomial and spline models. A plot matrix showed that there were some possible nonlinear relationships between the predictors and the response. But both modeling methods required only quantitative variables and have multiple possible tuning parameters. For simplicity purposes, only one predictor is evaluated at a time, the cubic spline is used, and only the first four polynomial degrees are analyzed. While both models can be expanded to multiple predictors, it is unnecessary since a GAM would accomplish the same thing. Overall, the data set was difficult to handle from a nonlinear perspective.

I do trust that the final model is the best model for this **particular data set**. The model has a lot of generalizability and has proven to be the most accurate prediction model. However, I would be extremely worried if my job depended on this model because of the data. I do not think that this data set encompasses all the data points that would contribute to the price of a house such as time and the stock market. I also do not think it would perform well on properties outside the scope of the historical data since it's a tree based model. In other words, if my job was in this perfect world where the housing market was less volatile and the properties predicted were very similar to the ones in the historical data set, I would feel more confident in the model's prediction ability. But in the real world, I do not think it's nearly robust enough to be a useful model.