­

**D600 – Statistical Data Mining**

**Performance Assessment #1 – Linear Regression Analysis**

Bernard Connelly

Master of Science, Data Analytics, Western Governors University

Dr. Keiona Middleton

January 24, 2025

**D600 – Statistical Data Mining: Linear Regression Analysis**

# Purpose of Analysis

## B1. Purpose of Research Question

The dataset presented provides a variety of metrics to review about specific houses in a housing market. Working under the assumption that this dataset comes from a realtor company or another interested party whose motivation is profit, a research question related to affecting the price of homes is most relevant. For this project, the research question is “How do the variables square\_footage, num\_bedrooms, crime\_rate, fireplace, house\_color, and garage impact the price of a home?”

## B2. Goal of Analysis

To maximize the value of the dependent variable, price, a set of variables that represent a direct relationship to improving this value should be selected. This analysis aims to identify variables that have an apparent effect on increasing the price and profitability of homes. In doing so, suggestions can be made based on the company on which specific metrics are most important when acquiring new houses in the future to turn a maximum profit.

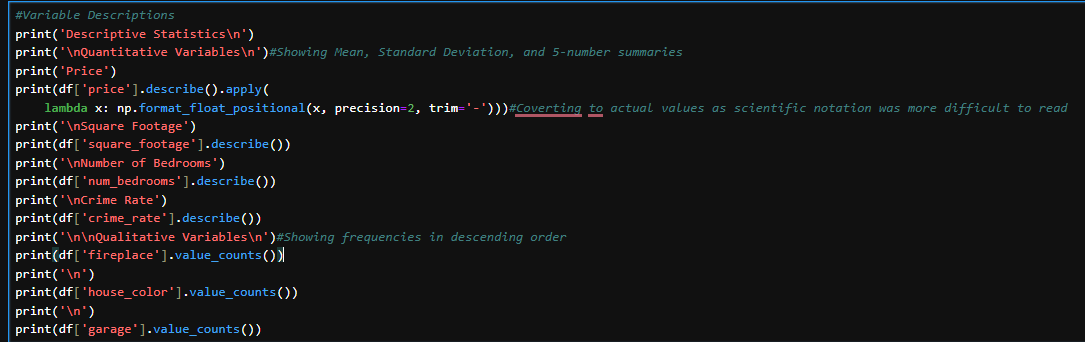
# Summarization of Data Preparation

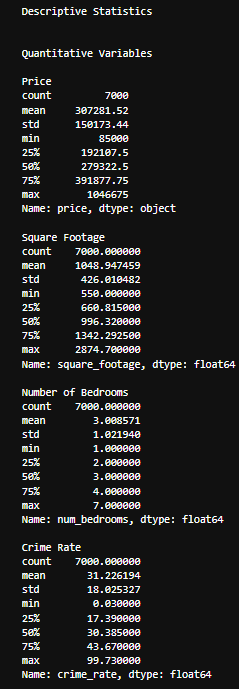
## C1. Identification of Variables

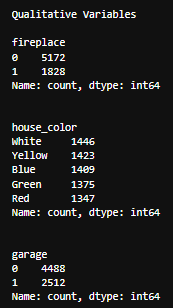
The dependent variable in this scenario is ‘price,’ which is the monetary value at which the home was most recently sold. All other variables are independent variables, which can be used to explain the changes in price. It was recommended in the lectures that we only select six total variables and utilize them for the project. As there was no data dictionary, the variables could only be selected based on their titles, which I used to try and achieve a variety of options. For this project, I selected square\_footage, num\_bedrooms, crime\_rate, fireplace, house\_color, and garage as my independent variables. These were selected based upon a few assumptions – some were quantitative continuous (square\_footage, crime\_rate), others were quantitative discrete (num\_bathrooms), some represented Boolean responses (garage and fireplace), and some were qualitative (house\_color).

## C2. Statistical Description of Variables

Below is a screenshot of the code and output descriptive statistics of each of the variables utilized for this analysis

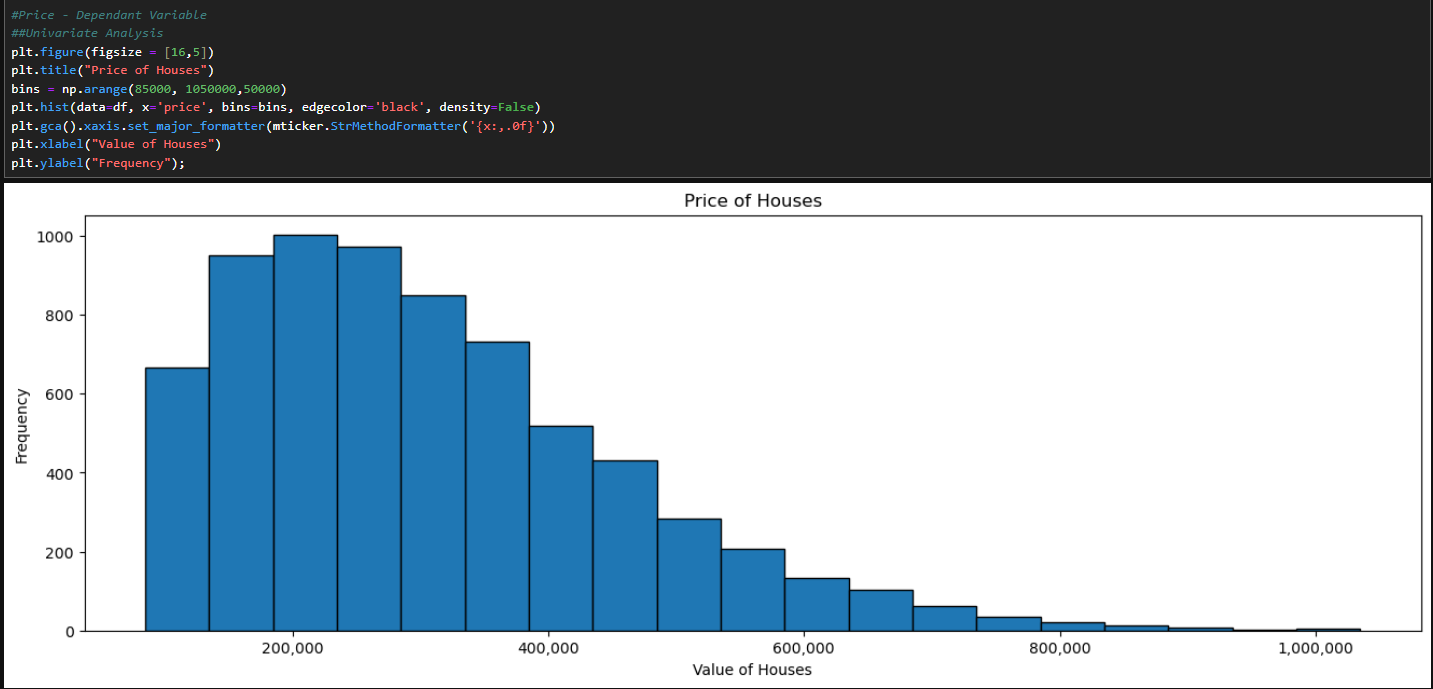






The above descriptive statistics outline the counts, means, standard deviations, min/max and quartiles for the quantitative data visible above, along with raw counts and distributions of the categorical data. Notably, the price column was formatted to be more readable using the lambda function (ChatGPT, OpenAI, 2025). This permits the values to be more readable overall to the user.

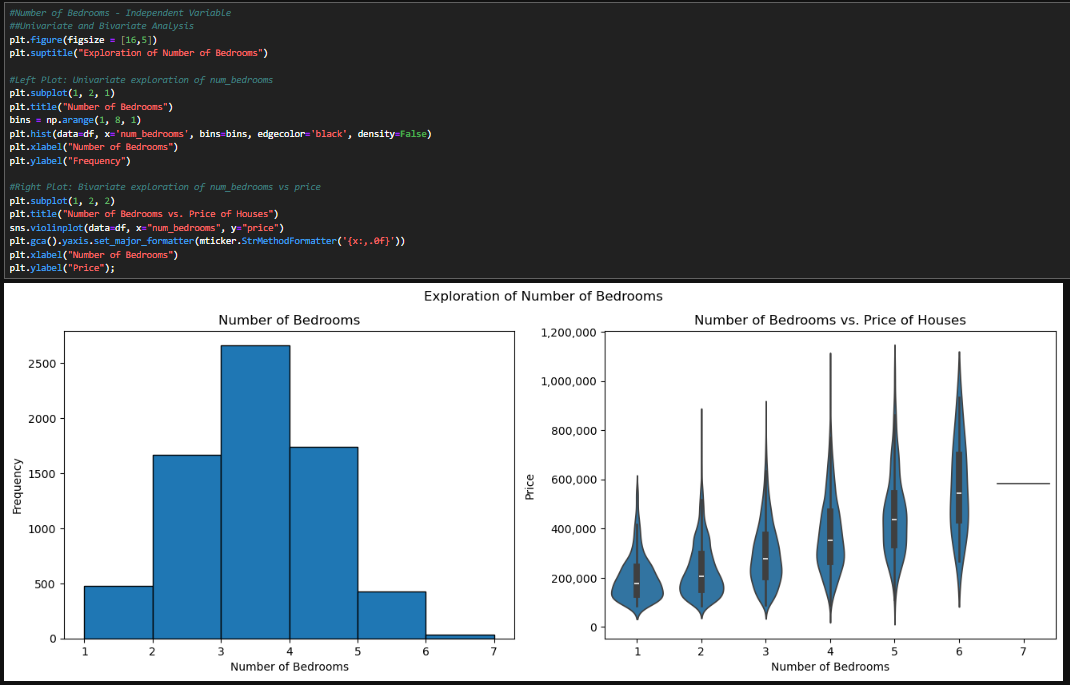
## C3. Statistical Visualizations

Below are the univariate and bivariate visualizations for the variables selected in the analysis. Matplotlib’s tick formatters were utilized to properly format the y-axes for the charts below by following guidance from Tutorialspoint (n.d.).

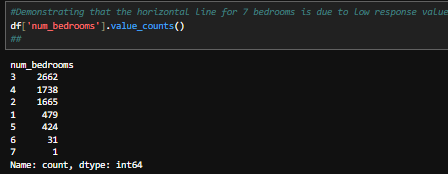
The price variable demonstrates a right-skewed distribution. This can be sub-optimal for a multiple regression analysis, so the residuals will be checked after the regression model is completed to ensure viability.

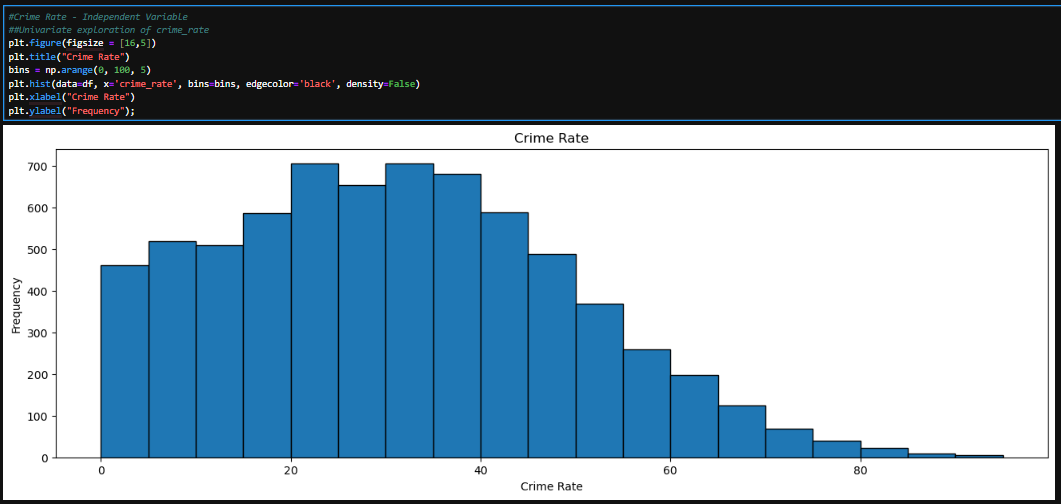


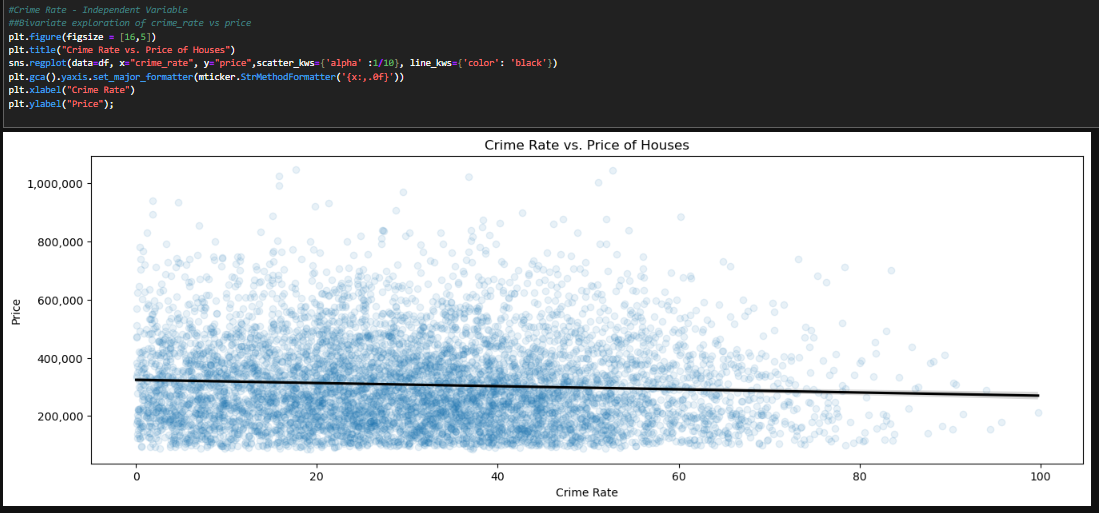
Square footage also appears to be an abnormal distribution, but the expectations of a housing dataset can explain this away. The square\_footage variable will not scale to 0, as a minimum square footage value must be maintained for a viable living area in a house. As a result, there is a cluster of values visible around 500 in both the univariate histogram and bivariate scatterplot. This is expected and will not affect the multiple regression analysis.



The num\_bedrooms variable demonstrated a normal distribution in the histogram and violin plot. Notably, since seven bedrooms represented an unusual violin plot as a horizontal line only, I confirmed this was due to a low response rate, which was confirmed as only one house returned a value of 7 for num\_bedrooms, noted in the code below.

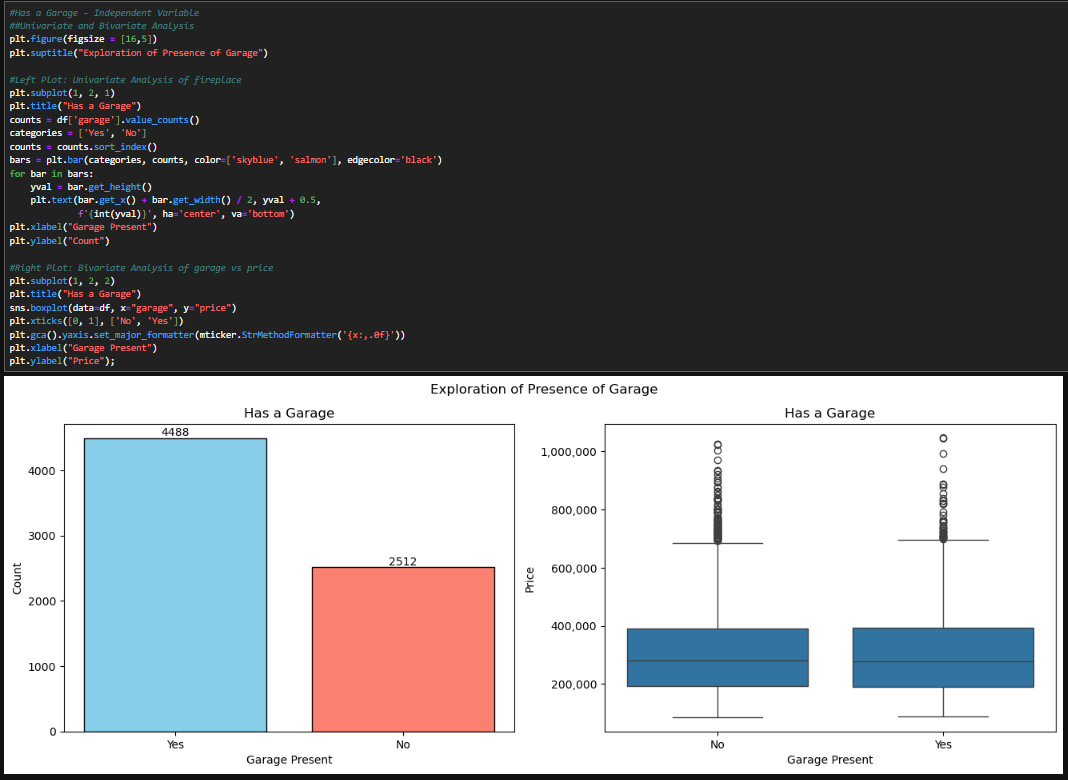




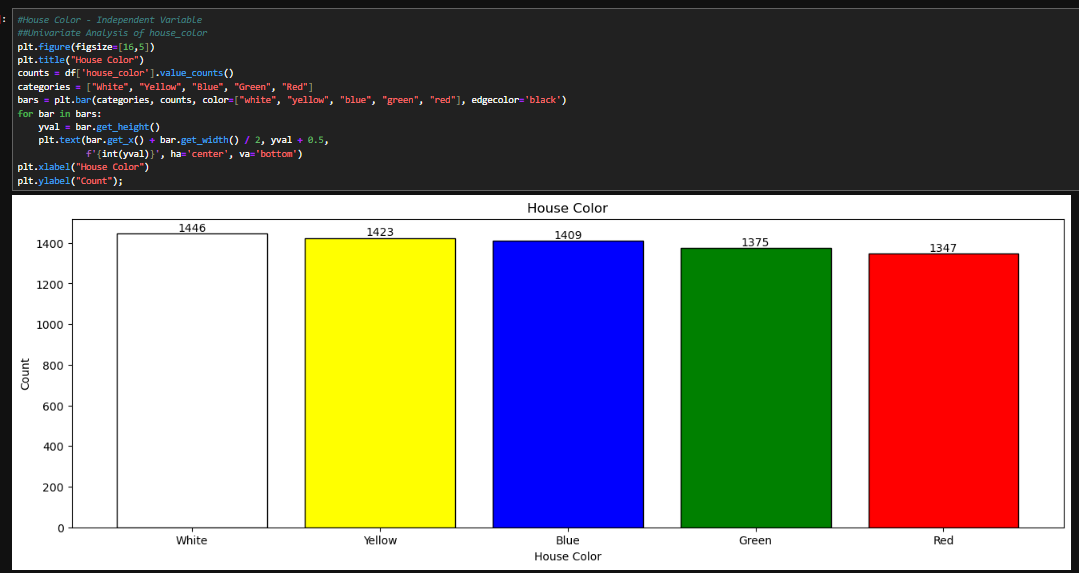


The crime\_rate variable demonstrated a close-to-normal distribution in both visualizations. However, clustering values on the lower end slightly skew the data. This will be reviewed as necessary as a part of the statistical analysis for this variable.





There is nothing notable about the distribution of has\_fireplace or has\_garage categorical variables. The boxplots represent an average spread of values with an acceptable amount of outliers.



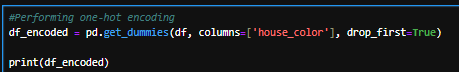


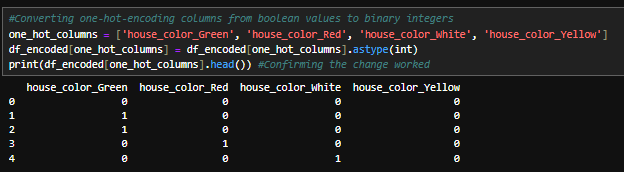
The uniform distribution of house color may have a negative effect on the multiple regression model, as this distribution of data may have lower predictive power, along with minimal contribution to optimizing the R² value.

# Linear Regression Modelling

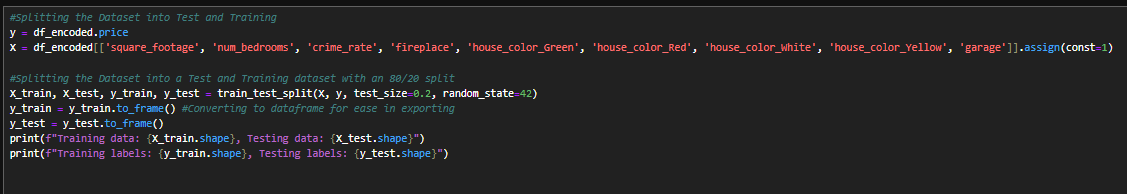
## D1. Creation of Training & Test Data Sets

Before splitting the datasets, one-hot encoding was performed on the categorical variable “house\_color,” and it was converted from boolean values to binary integers. Hence, it is usable in multiple linear regression.

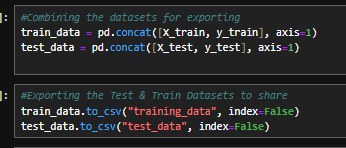




The next step was to split the data into two separate models, training and test, then exporting them to share as a part of the assessment. An 80/20 training to test split was utilized using the code below.

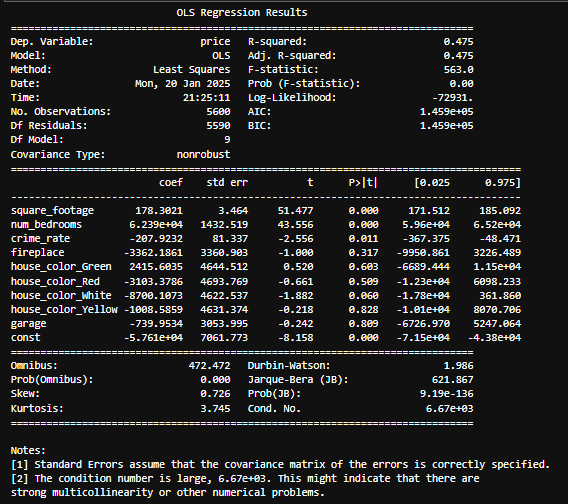


The following was used to export the data:

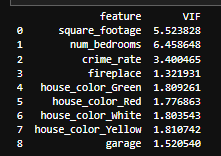


## D2. Optimization of the Model

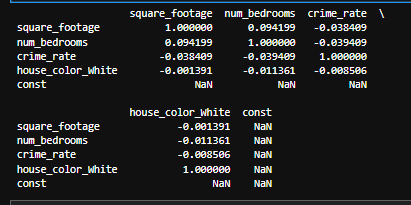
To optimize the model, backward stepwise elimination was used to remove the statistically insignificant variables to the research question. This method allows values to be removed one at a time and re-tests the model after removal. The initial model returned the details below as a result.



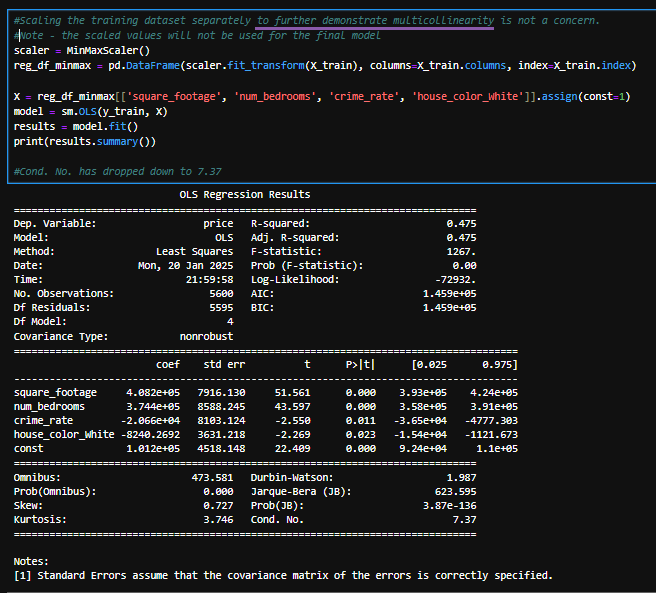
As one note indicated, there was a potential problem with multicollinearity; the Variance Inflation Factor was run on the dataset to identify any potential issues. The below results were returned.



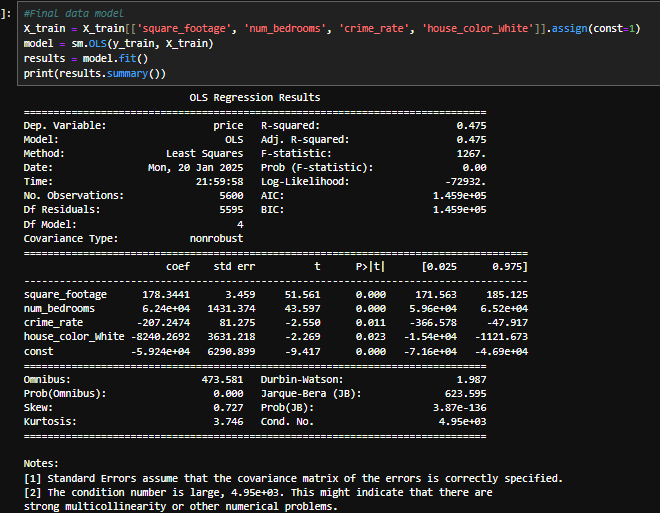
Both square\_footage and num\_bedrooms have VIF over 5, which presents a mild concern for multicollinearity. Neither has a value so high that it presents a significant concern, but to be sure, I wanted to rule this out. To do so, I reviewed the relationships between these variables using a correlation matrix and ultimately determined that the correlations were weak, so I kept all variables in the dataset. The results of the correlation matrix are below.



To further ensure multicollinearity was not a concern, the dataset was scaled to provide a better view of the continuous variables and assess the concerns more directly. After doing so, the cond No. dropped down to 7.37, and the note indicating multicollinearity was a concern. The scaled dataset was not utilized for the remainder of the analysis as it was not advised as a part of the instructions.



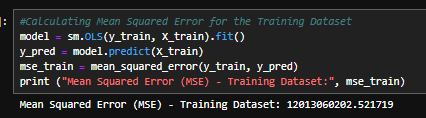
After ruling out multicollinearity, I initiated backward stepwise elimination. Using the p-value as the primary metric for whether a variable was statistically significant, I determined whether a variable should stay or be removed. I re-ran the model after each variable was removed in the following order: garage (.809 p-value), house\_color\_Yellow (.829 p-value), house\_color\_Red (.522 p-value), house\_color\_Green (.319 p-value), and fireplace (.326 p-value). This left the results below when the model was run, with no p-values outside an acceptable range.



The final data model included square\_footage, num\_bedrooms, crime\_rate, and house\_color\_White. Notably, the above output contains an R2 and adjusted R2 at .475, indicating that the variables in the model predict around 48% of the dependent variable of the price change. Also included are the F statistic, probability F Statistic, and coefficient estimates.

## D3. Mean Squared Error (MSE) Analysis

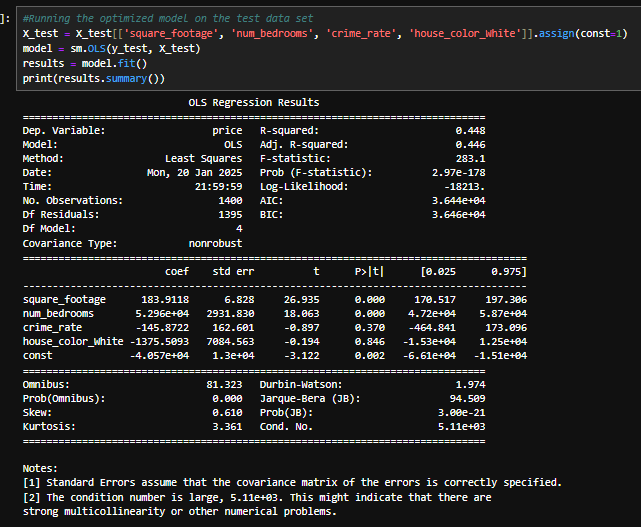
Below is the code and results of the Mean Squared Error after optimization



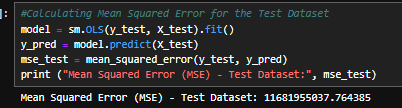
Taking the square root of the MSE above, the result is $109,604.11 – indicating the model’s predictions will differ from the actual house prices by around that value.

## D4. Prediction

Using the model developed using the training dataset, the test dataset was analyzed using only the variables deemed relevant to the analysis. The optimized model produced the following results in its output.



Additionally, the MSE was calculated for the test dataset and produced results in line with what was expected compared to the training dataset.



Comparing the MSE and RMSE of the two datasets demonstrated that the model was consistent across both datasets and generalized well to the new data. There is no indication of overfitting or underfitting, and the data was correctly split, resulting in little to no data leakage.

# Summary and Results

## E1. Libraries Utilized

## pandas & numpy – For general usefulness with data frames and Python coding

## matplotlib.pyplot - Used for data visualizations

## seaborn - Used for data visualizations

## statsmodels.api - Used to create Ordinary Least Squares (OLS) Regression Model

## matplotlib.ticker - Used to scale the axes in data plots properly

## sklearn.model\_selection: train\_test\_split - Used to split the datasets

## statsmodels.stats.outliers\_influence: variance\_inflation\_factor - Used to check for multicollinearity in the data models

## sklearn.preprocessing: MinMaxScaler - Used for a diagnostic test of VIF, scaling the data frame to solve for Multicollinearity

## sklearn.metrics: mean\_squared\_error - For calculating MSE

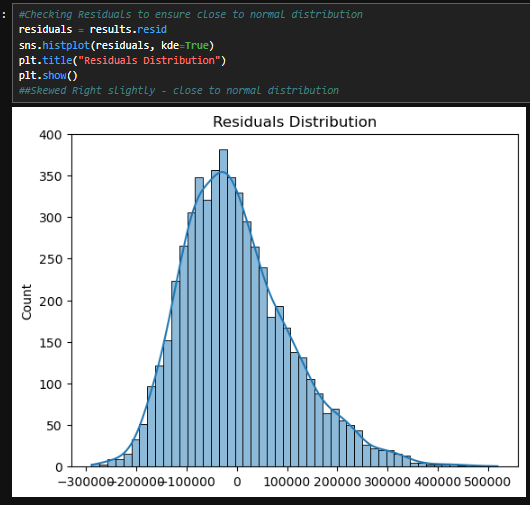
## E2. Method of Optimization

Backward Stepwise Elimination was utilized to optimize the model. Instructions were to begin with six independent variables and one dependent variable in the analysis, so Forward Stepwise Selection was not optimal as it would add additional variables to the data model. Backward Stepwise Elimination allowed me to eliminate variables one at a time when they were statistically insignificant to ensure only the most relevant variables remained. This was validated by checking for the mean squared error on both the training and testing datasets post-optimization to confirm that the error was reduced in the model.

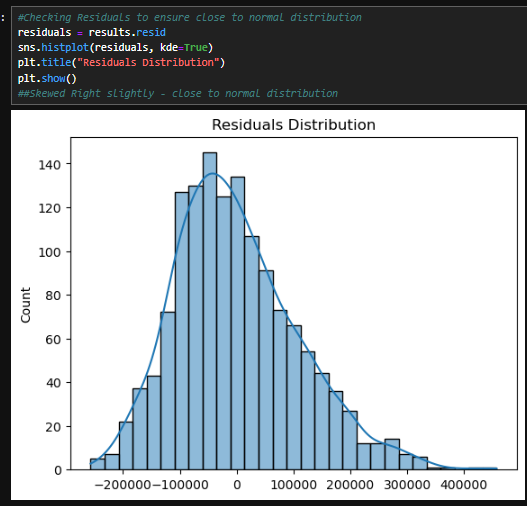
## E3. Verification of Assumptions

The major assumptions of backward stepwise elimination include linear relationships between predictor variables and their target, low multicollinearity, and a normal distribution of residuals. Linear relationships were confirmed using the visualizations listed above in section C3 – the scatterplots represented straight trend lines, and the correlation matrix presented demonstrated linear associations. For multicollinearity, VIF was calculated for the predictor variables, and the highest value was 6.4, which presented a moderate concern for multicollinearity. To rule this out, the dataset was temporarily scaled, and the model was re-run, demonstrating a significantly reduced condition number. To assess the normal distribution of residuals, histograms of the residuals were plotted after model optimization for both the training and test datasets; both demonstrated slightly skewed residuals but still a normal distribution.

Training Residuals:

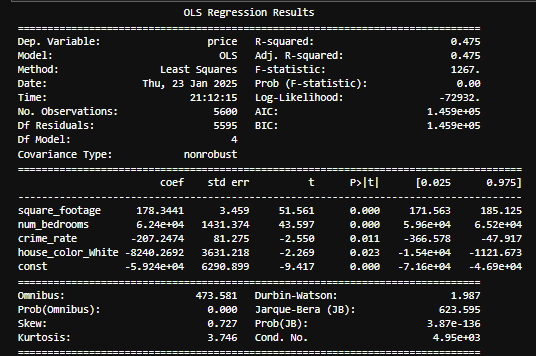


Test Residuals:



## E4. Regression Equation and Coefficient Estimates

Using the training data model, the final regression output is as follows:



From the above, the regression equation is:

price^​=−59,240+178.34(square footage)+62,400(num bedrooms)−207.25(crime rate)−8,240.27(house color White)

To explain, the general estimates are as follows:

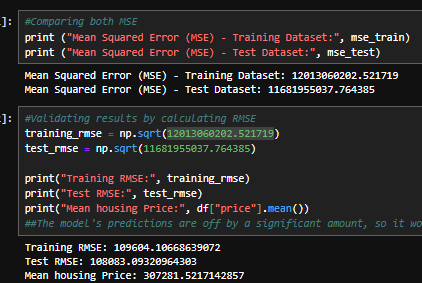
* Assuming all other variables remain constant for the below:
  + Every additional square foot increases the price of the house by $178.34
  + Every additional bedroom increases the price of the house by $62,400
  + A one-unit increase in crime rate decreases the price of the house by $207.25
  + The house being white decreases the price of the house by $8,240.27 compared to another color.

The intercept is represented at -59,240, which is not meaningful.

## E5. Model Metrics

Per the above output, the R2 and adjusted R2 are the same at 0.475. This means the model explains 47.5% of the variation of the pricing in houses. More broadly speaking, the variables selected have a definitive impact on the price of a house, but more factors were not included, which make up more than half the explanatory power. The p-values for all variables are less than 0.05, meaning they are likely to impact housing prices statistically. As stated above, the output indicates there would be multicollinearity concerns, but those have been ruled out separately.

When assessing the MSE for both datasets, it is most helpful to look at them alongside one another, and the mean price of houses was also included as a relevant comparison.



The two RMSE values for both datasets were within $1600 of one another, indicating the model was not overfit or underfit and generalized to data effectively. The relative size of the RMSE, however, was nearly 33% of the mean price of the house. This means the model has predictive power, but the prediction error is significantly higher than what would be considered an excellently fitted model.

## E6. Results and Implications

The results of the analysis indicate that square footage and the number of bathrooms increase the value of a home. Additionally, higher crime rates or white houses will reduce the value of the homes. All of these variables were statistically significant in terms of the price of the houses. The overall predictive power of the model was off by an average of around $108,000, which makes sense when considering the R2 value indicated that the model only accounted for around 47.5% of the variance. The model was not incorrect, but it was incomplete.

## E7. Recommendations

Based upon the analysis, the business should ensure they invest in houses that are in lower crime areas and whose houses are not white in color. Additionally, focusing on the number of rooms and square footage counts will yield direct increases in profit. The best way to optimize their earnings would be to run another regression model encapsulating additional variables. Using the model prepared and implementing forward stepwise selection to add these variables would increase the efficacy of the prediction and provide further relevant variables to calculate the overall changes in price.

**References**

OpenAI. (2025, January 16). Explanation of lambda functions in Python. ChatGPT. Retrieved from <https://chat.openai.com>

Tutorialspoint. (n.d.). Matplotlib tick formatters. Tutorialspoint. <https://www.tutorialspoint.com/matplotlib/matplotlib_tick_formatters.htm>