Peyton Bailey

April 20th, 2025

D600 – Task 1

B. Describe the purpose of this data analysis by doing the following:

1. Propose one research question that is relevant to a real-world organizational situation captured in the provided dataset that you will answer using multiple linear regression in the initial model.

How do various factors impact the housing prices in this market? I will explore key factors such as square footage, number of bedrooms, number of bathrooms, crime rate, school rating, age of home, renovation quality, previous sale price, presence of garage, and whether the house is classified as luxurious.

2. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

The goal is to develop a predictive model to estimate housing prices accurately based on key property features. This will serve as a tool for both real estate professionals and prospective home buyers to make informed decisions.

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:

1. Identify the dependent and all independent variables that are required to answer the research question and justify your selection of variables.

The independent variables are square footage, number of bedrooms, number of bathrooms, crime rate, school rating, age of home, renovation quality, previous sale price, presence of garage, and luxury status. The dependent variable is the price of the house. I chose my independent variables

based on factors that I know can affect the price of a house, but I also wanted to include variables that could influence price both positively and negatively. For example, houses with a large square footage and a greater number of bedrooms and bathrooms usually have higher prices. However, a large house may not be as expensive if it's located in a neighborhood with a high crime rate or a lower school rating, since those factors can drive prices down. Age of home is another variable I included because I expected newly built homes to be more expensive, so I hypothesized a negative correlation between age and price. Luxury homes are more expensive than non-luxury homes, so I wanted to see by how much and how strong of a predictor that classification would be. The presence of a garage was included because homes with garages are often found in more affluent neighborhoods with higher-priced homes, though not having a garage doesn't always mean a home is lower quality or in a bad area. Renovation quality is important because a well-renovated home will generally be priced to reflect its condition, even in less affluent neighborhoods. Finally, I included previous sale price because real estate professionals often price homes above their last purchase price to ensure a profit. Overall, I selected these variables to capture a range of factors—both structural and neighborhood-related—that can influence housing prices in different ways.

2. **Describe the dependent variable and all independent variables from part C1 using descriptive statistics (counts, means, modes, ranges, min/max), including a screenshot of the descriptive statistics output for each of these variables.**

```
count    7000.000000
mean     1048.947459
std       426.010482
min       550.000000
25%       660.815000
50%       996.320000
75%      1342.292500
max      2874.700000
Name: SquareFootage, dtype: float64
```

```
count    7000.000000
mean        3.008571
std         1.021940
min         1.000000
25%         2.000000
50%         3.000000
75%         4.000000
max         7.000000
Name: NumBedrooms, dtype: float64
```

```
count    7000.000000
mean        2.131397
std         0.952561
min         1.000000
25%         1.290539
50%         1.997774
75%         2.763997
max         5.807239
Name: NumBathrooms, dtype: float64
```

```
count    7000.000000
mean       46.797046
std        31.779701
min         0.010000
25%        20.755000
50%        42.620000
75%        67.232500
max       178.680000
Name: AgeOfHome, dtype: float64
```

```
count    7000.000000
mean        6.942923
std         1.888148
min         0.220000
25%         5.650000
50%         7.010000
75%         8.360000
max        10.000000
Name: SchoolRating, dtype: float64
```

IsLuxury	Frequency	Ratios
1	3528	0.504
0	3472	0.496

```
count    7000.000000
mean       31.226194
std        18.025327
min         0.030000
25%        17.390000
50%        30.385000
75%        43.670000
max        99.730000
Name: CrimeRate, dtype: float64
```

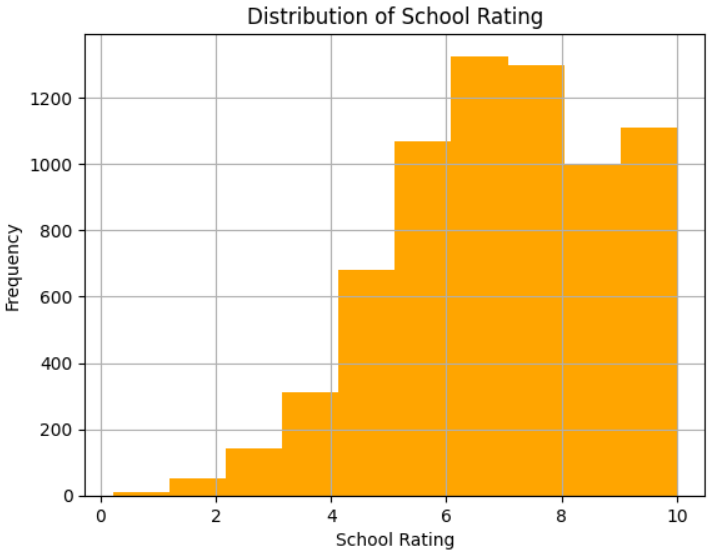
```
count       7,000.00
mean      307,281.97
std       150,173.43
min        85,000.00
25%       192,107.53
50%       279,322.95
75%       391,878.13
max     1,046,675.64
Name: Price, dtype: float64
```

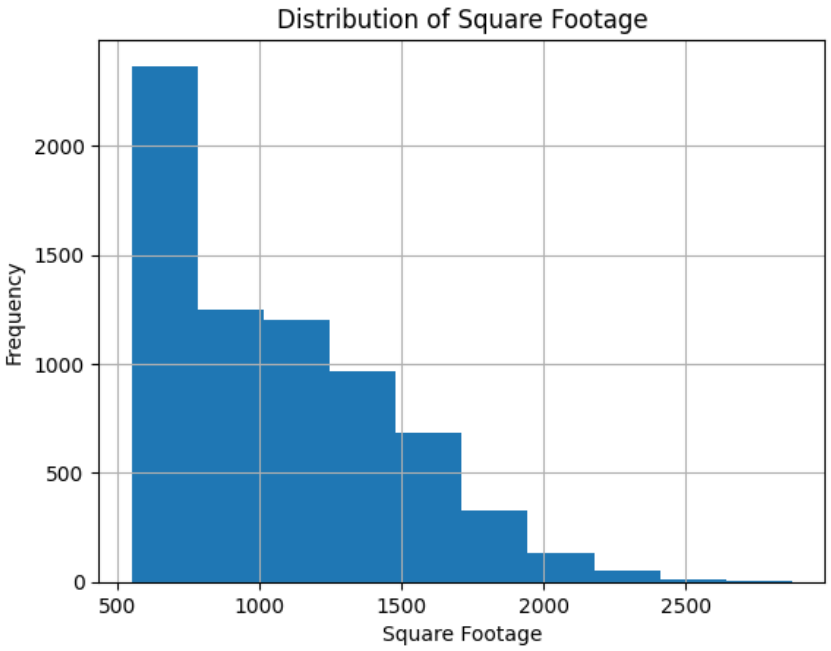
```
count       7,000.00
mean      284,509.35
std       185,734.02
min        -8,356.90
25%       142,013.98
50%       262,183.13
75%       396,121.17
max     1,296,606.69
Name: PreviousSalePrice, dtype: float64
```

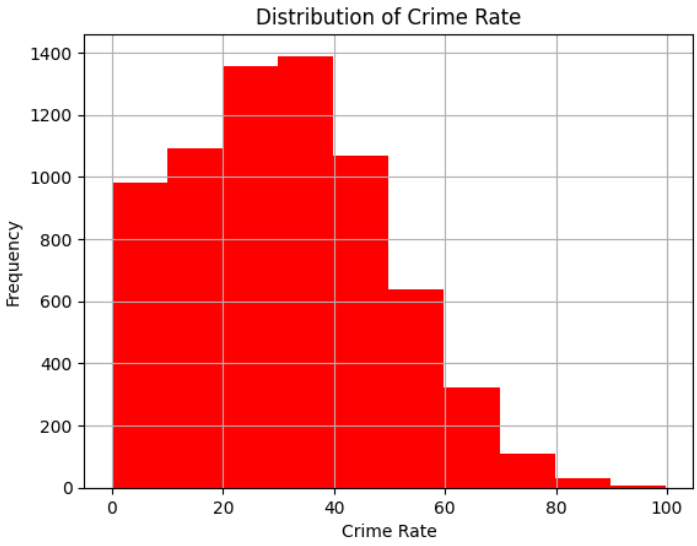
```
count    7000.000000
mean        5.003357
std         1.970428
min         0.010000
25%         3.660000
50%         5.020000
75%         6.350000
max        10.000000
Name: RenovationQuality, dtype: float64
```

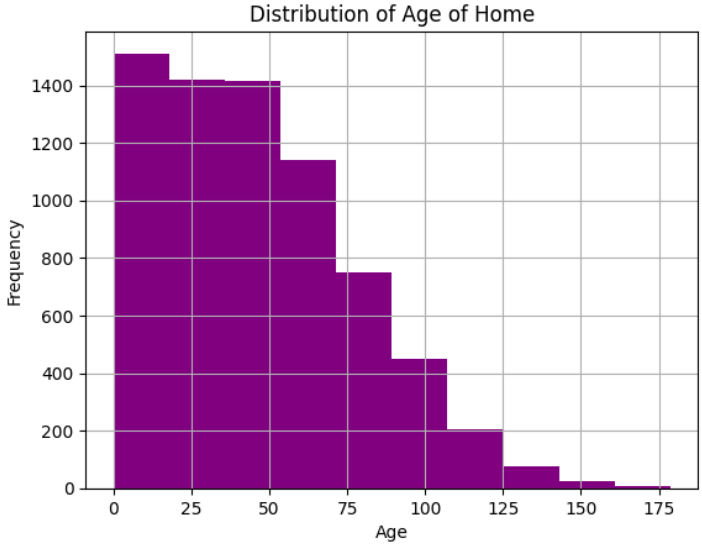
Garage	Frequency	Ratios
No	4488	0.641143
Yes	2512	0.358857

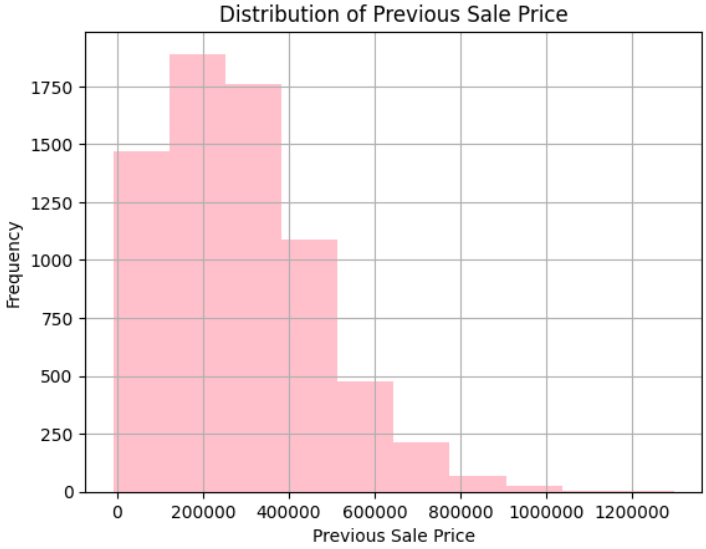
3. **Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables from part C1, including the dependent variable in the bivariate visualization.**

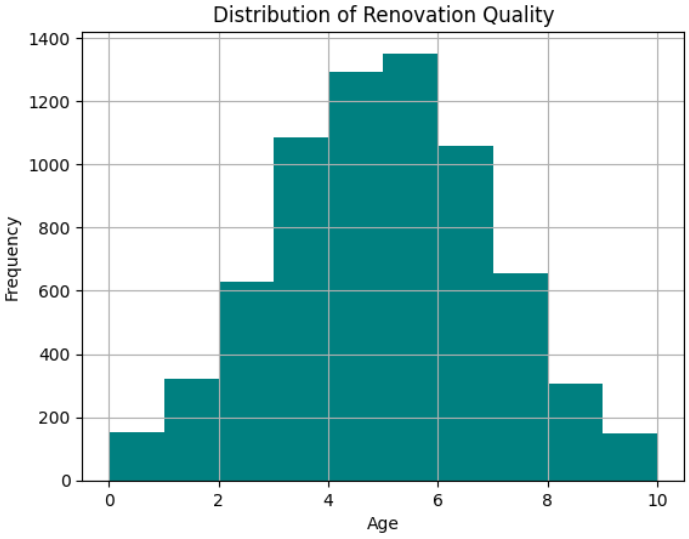


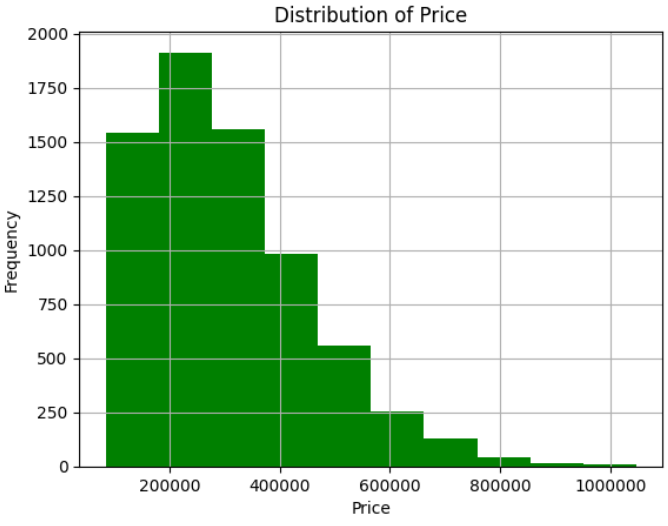


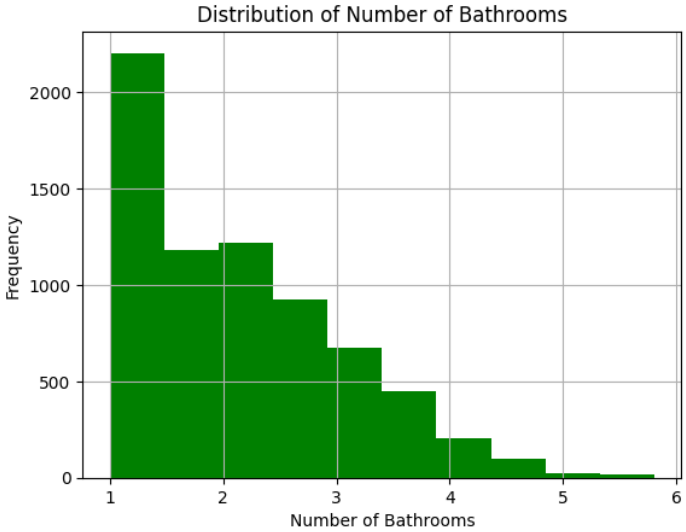

Distribution of Crime Rate
Frequency
Crime Rate
0
200
400
600
800
1000
1200
1400
0
20
40
60
80
100

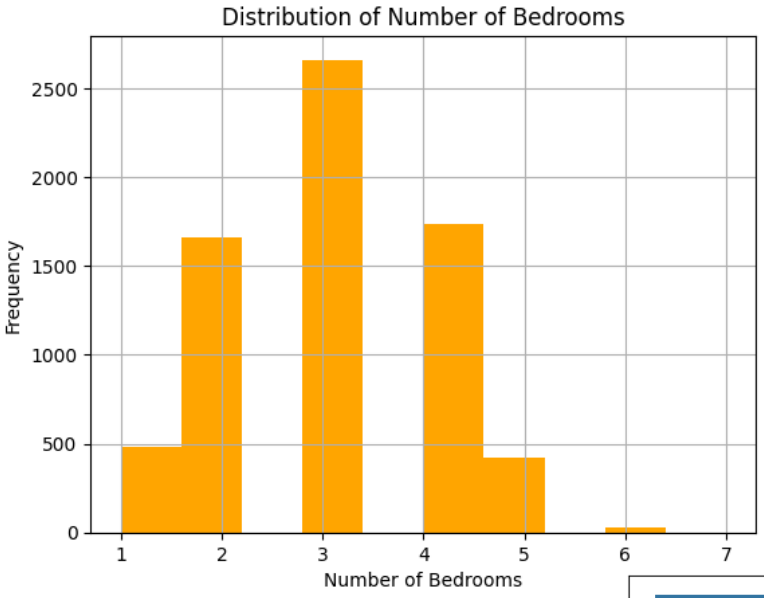

Distribution of Age of Home
Frequency
Age
0
200
400
600
800
1000
1200
1400
0
25
50
75
100
125
150
175

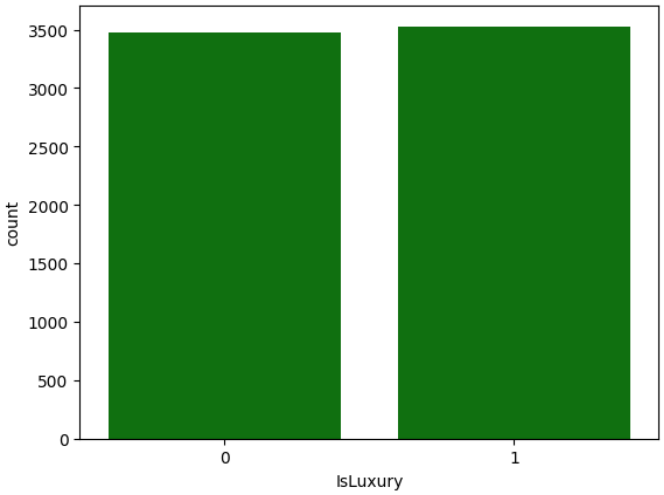

Distribution of Previous Sale Price
Frequency
Previous Sale Price
0
250
500
750
1000
1250
1500
1750
0
200000
400000
600000
800000
1000000
1200000

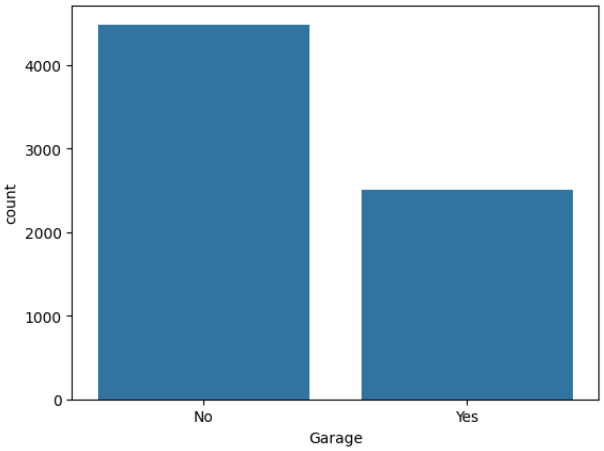

Distribution of Renovation Quality
Frequency
Age
0
200
400
600
800
1000
1200
1400
0
2
4
6
8
10


Distribution of Price
Frequency
Price
0
250
500
750
1000
1250
1500
1750
2000
200000
400000
600000
800000
1000000

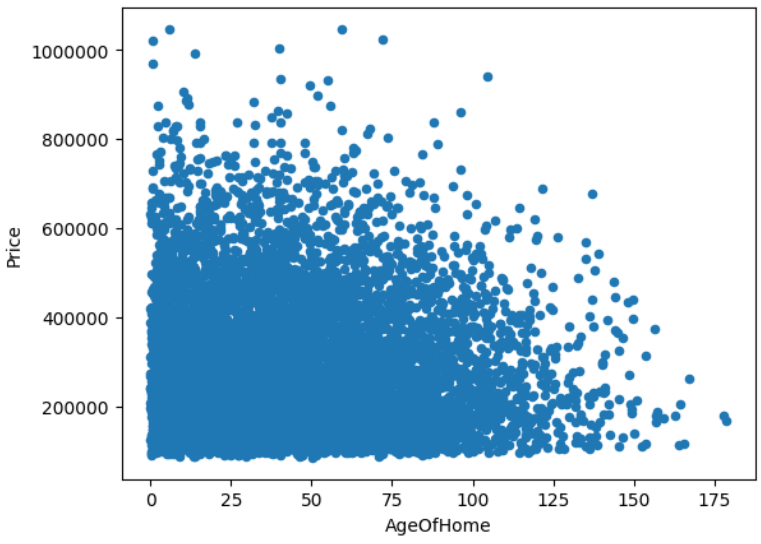

Distribution of Number of Bathrooms
Frequency
Number of Bathrooms
0
500
1000
1500
2000
1
2
3
4
5
6

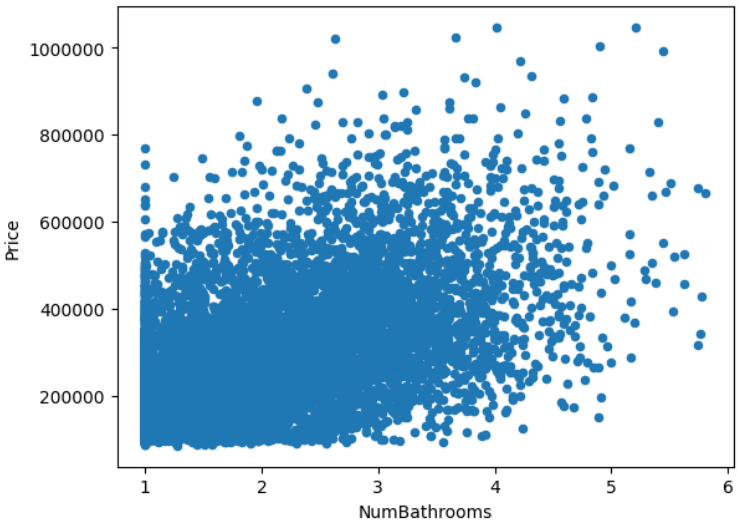


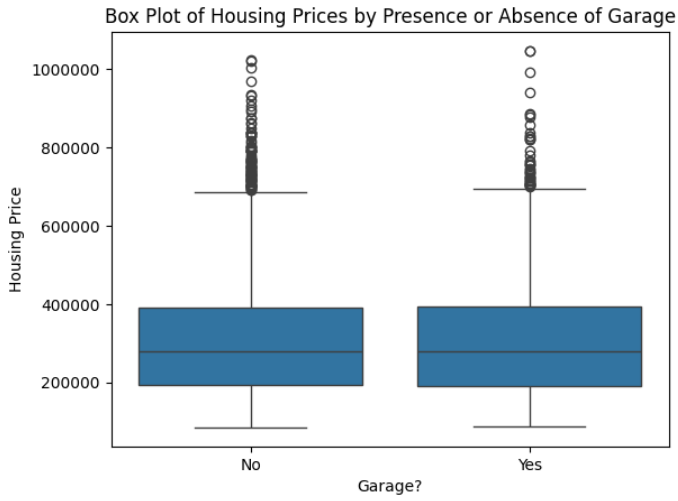


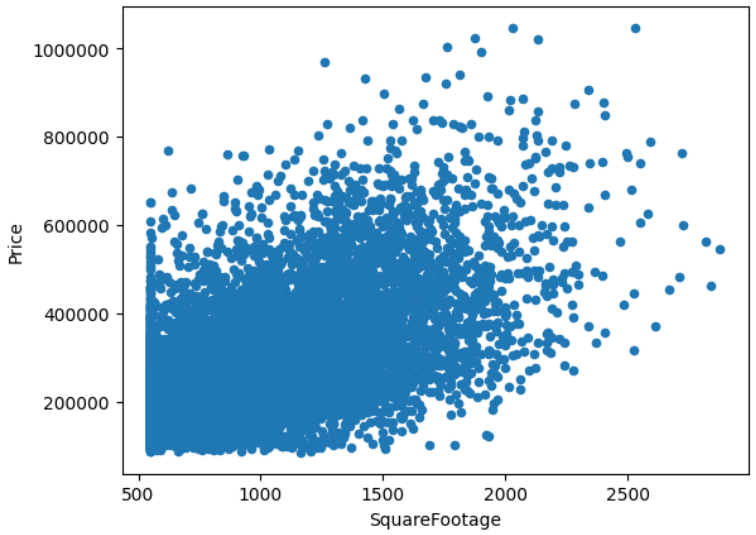


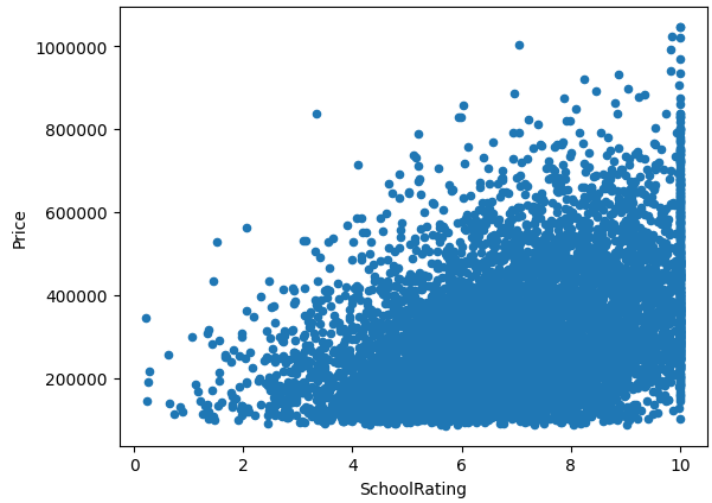
Bivariate Visualizations

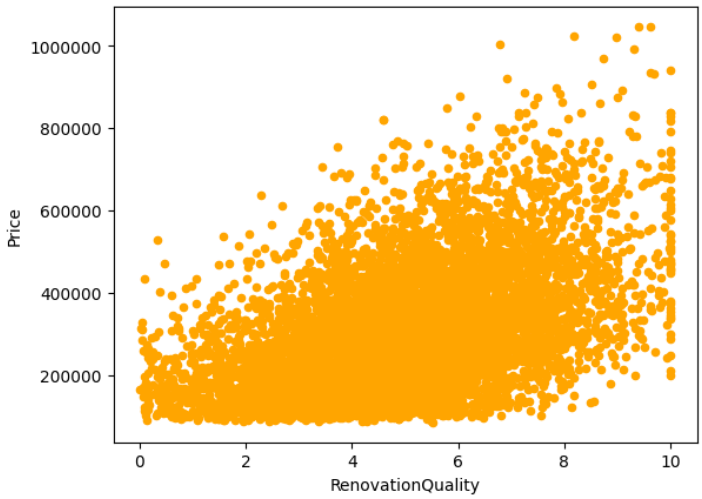


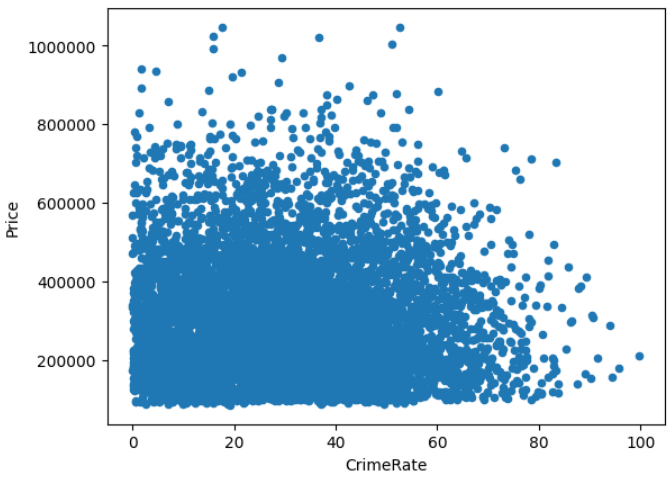


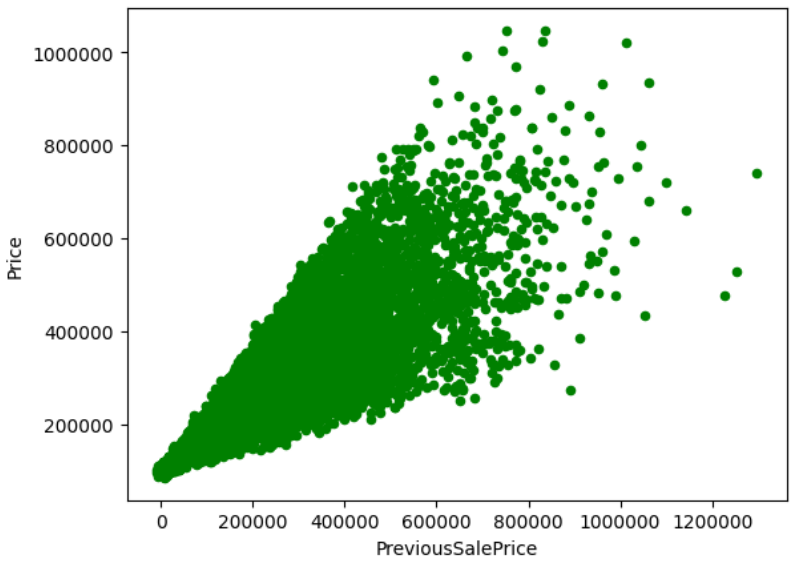

Box Plot of Housing Prices by Presence or Absence of Garage
Housing Price
1000000
800000
600000
400000
200000
No
Yes
Garage?

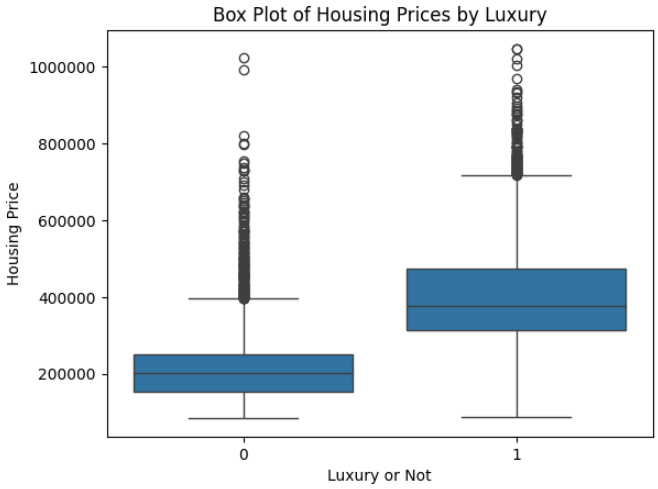

Price
1000000
800000
600000
400000
200000
500
1000
1500
2000
2500
SquareFootage

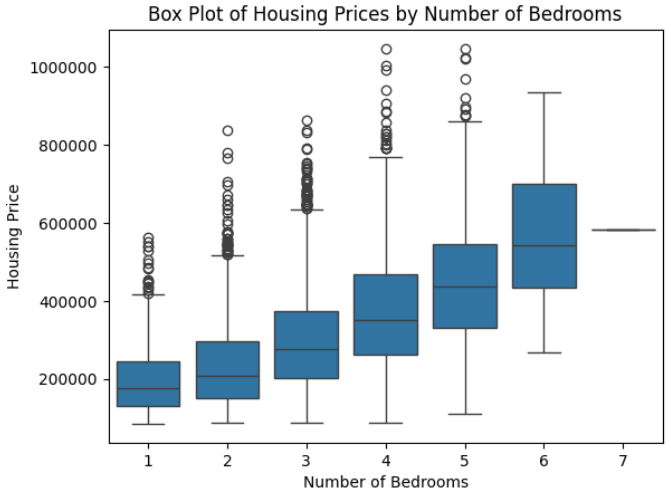

Price
1000000
800000
600000
400000
200000
0
2
4
6
8
10
SchoolRating


Price
1000000
800000
600000
400000
200000
0
2
4
6
8
10
RenovationQuality


Price
1000000
800000
600000
400000
200000
0
20
40
60
80
100
CrimeRate


Price
1000000
800000
600000
400000
200000
0
200000
400000
600000
800000
1000000
1200000
PreviousSalePrice





D. Perform the data analysis and report on the results by doing the following:

1. **Split the data into two datasets, with a larger percentage assigned to the training dataset and a smaller percentage assigned to the test dataset. Provide the files.**
2. **Use the training dataset to create and perform a regression model using regression as a statistical method. Optimize the regression model using a process of your selection, including but not limited to, forward stepwise selection, backward stepwise elimination, and recursive selection. Provide a screenshot of the summary of the optimized model or the following extracted model parameters:**
   - **adjusted R2**
   - **R2**
   - **F statistics**
   - **probability F statistics**
   - **coefficient estimates**
   - **p-value of each independent variable**

```
=== Final Model Summary ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.735
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     2588.
Date:                Wed, 16 Apr 2025   Prob (F-statistic):               0.00
Time:                        16:27:33   Log-Likelihood:                -70974.
No. Observations:                5600   AIC:                         1.420e+05
Df Residuals:                    5593   BIC:                         1.420e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.048e+04   5649.690      3.624      0.000    9401.341    3.16e+04
SquareFootage     50.9125      3.231     15.758      0.000      44.579      57.246
NumBathrooms   1.801e+04   1318.368     13.664      0.000    1.54e+04    2.06e+04
NumBedrooms    1.735e+04   1261.006     13.758      0.000    1.49e+04    1.98e+04
PreviousSalePrice     0.3931      0.010     38.896      0.000       0.373       0.413
AgeOfHome          -93.0360     32.664     -2.848      0.004    -157.070     -29.001
IsLuxury_1     6.977e+04   2508.285     27.815      0.000    6.49e+04    7.47e+04
==============================================================================
Omnibus:                      450.269   Durbin-Watson:                   2.017
...
Notes:
```

4. Give the mean squared error (MSE) of the optimized model used on the training set.

```
Mean Squared Error on Training Set: 5969309812.8282
```

E. Summarize your data analysis by doing the following:

1. List the packages or libraries you have chosen for Python or R and justify how each item on the list supports the analysis.

Pandas – This package allowed me to import and manipulate the dataset within a data frame. This includes isolating each variable, calculating its descriptive statistics, and creating summary tables for the categorical variables.

Numpy – This is a fundamental library for numerical and scientific computing which aided my analysis.

Matplotlib – This package supports the creation of visualizations. I used this package to create histograms that depicted the distribution of my quantitative variables and scatterplots to visualize the bivariate distribution of each independent quantitative variable and "Price," which was the dependent quantitative variable.

Seaborn – This package also supports visualizations. I used this to create box plots as well as bar graphs.

Statsmodels – This package directly supports regression analysis. This allowed me to add a constant to my independent variables, input the X and Y variables, and fit them into a model using Ordinary Least Squares. This model provided a summary that included the coefficients for each independent variable, the R-squared, adjusted R-squared, F-statistic, AIC, and BIC. Furthermore, this package also consists of the variance inflation factor (VIF), which I used to detect multicollinearity before fitting the model.

Sklearn – This package also supports regression analysis, emphasizing machine learning. Through this package, I could import a train-test split feature, allowing me to split my data into training and testing data, which allowed me to evaluate the accuracy of the prediction model. This package also allowed me to calculate the mean squared error between the predicted vs. actual values.

2. Discuss the method used to optimize the model and justification for the approach.

This model was first checked for multicollinearity by detecting the VIF values for each independent variable. Any VIF value over 10 was considered too high. The variable with the highest VIF value was removed. The VIF values were then recalculated, and the next variable was removed. This process was completed until all VIF values were under 10. Next, I used backward stepwise elimination to eliminate variables that were not significant to the prediction

of the dependent variable. This was also done one at a time using the p-value as a reference. A variable with a p-value over 0.05 was considered insignificant and thus removed from the model equation. I used this approach because I already knew what factors affect housing prices, so I knew which variables I wanted to include in my initial model. Backward stepwise elimination allowed me to do this, remove the factors that weren't significant, and retain the ones that were.

3. Discuss the verification of assumptions used to create the optimized model.

The assumptions of multiple linear regression are linearity, independence, normality, and homoscedasticity.

Linearity

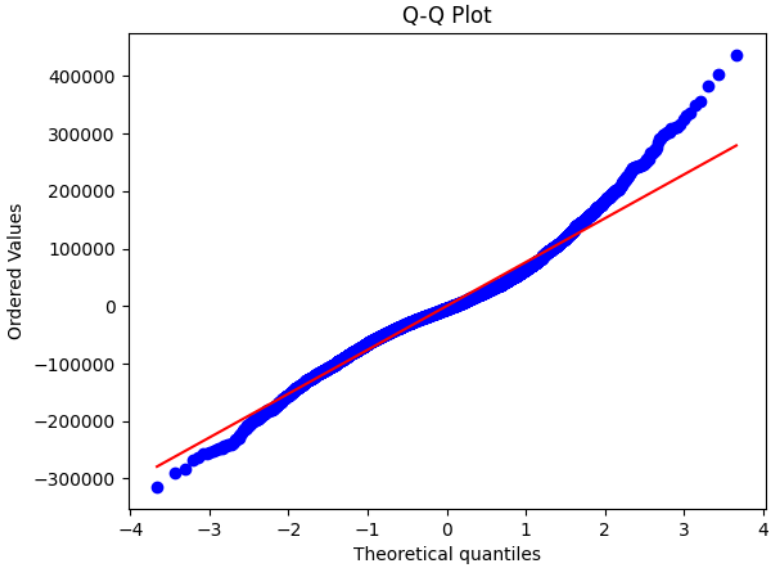
Linearity was verified with the bivariate relationships depicted on the scatterplots of each independent variable paired with the dependent variable. Each showed a fairly linear relationship, allowing me to confirm this assumption.

Independence

Independence was verified through the Durbin-Watson test. The Durban-Watson test is a statistical test used to detect autocorrelation in the residuals from regression analysis. This is done by examining whether the residuals from the regression model are independent or follow a pattern over time. A statistic of 2 indicates that no autocorrelation exists. I performed this test on my final model, and a statistic of 2.01715 was produced, meaning that this independence assumption was strongly satisfied.

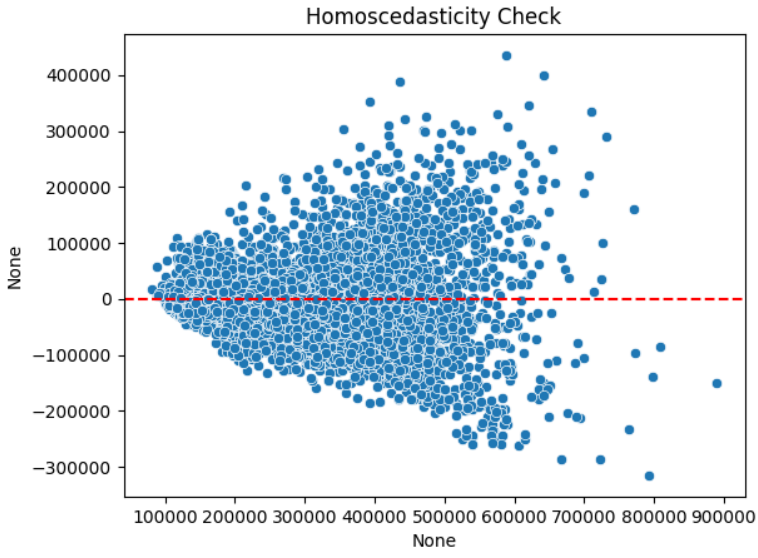
Normality

Normality was verified via a Q-Q plot. The plot compares the quantiles of my residuals in blue to the theoretical quantiles of a normal distribution in red. As shown, the plots mainly follow the red line in the center. However, they deviate on the tails, indicating some extreme outliers. This is expected in real-world data and is still acceptable for regression analysis.



Homoscedasticity

Homoscedasticity means that the scatter or spread of the residuals between observed and predicted values should be roughly the same across all values of the explanatory variables. This means the prediction errors should have consistent variability throughout the data range. This can be verified graphically. As shown in the image, this assumption was not satisfied. The megaphone shape of the plot of residuals indicated heteroscedasticity in the model. This was due to an increasing variance pattern. The lower values were relatively clustered around zero; however, the higher values showed a much more defined vertical spread.



4. **Provide the regression equation and discuss the coefficient estimates.**

=== Multiple Linear Regression Equation ===

Price = 20476.93 + 50.91*SquareFootage + 18013.65*NumBathrooms +
17348.64*NumBedrooms + 0.39*PreviousSalePrice + -93.04*AgeOfHome +
69768.87*IsLuxury_1

The optimization process of the model removed school rating, renovation quality, garage, and crime rate as variables. The square footage, number of bathrooms, number of bedrooms, previous sale price, age of home, and luxury remained in the equation. The luxury variable had the largest coefficient (69768.87). The only two possible values for this category are 0 for non-luxury and 1 for luxury, meaning any house classified as luxury automatically will have this price increase of nearly $70,000. The number of bathrooms and bedrooms variables also had relatively large coefficients (~17k-18k); however, the possible values for these variables are single digits, usually between 2 and 4. This still positively impacts the house price for properties with multiple bedrooms and bathrooms. Even though the coefficient for square footage is relatively small (50.91), this has an enormous impact on price since square footage in the dataset had a mean of 1,048. The only negative coefficient in the model was for the age of the home variable (-93.04). It intuitively makes sense that newer houses are more expensive than older ones.

5. **Discuss the model metrics by addressing each of the following:**
    - **the R^2 and adjusted R^2 of the training set**

The R^2 value represents the proportion of variance in the dependent variable explained by the model's independent variables (predictors). The adjusted R^2 is a modification of the R^2 that accounts for the number of predictors in the model and penalizes the addition of variables that do

not significantly improve it. The R^2 and adjusted R^2 are roughly 0.735, suggesting that the model's predictors are substantial and have not led to overfitting.

- **The comparison of the MSE for the training set to the MSE of the test set.**

The training MSE was 5,969,309,812.8282 and the test MSE was 6,011,563,127.2063. This is a difference of 42,253,314.38, which is only an increase of 0.71% from the training data. This suggests an excellent balance between both sets of data and no evidence of either overfitting or underfitting. For further confirmation, I ran an independent t-test on them and returned a p-value of 0.978. This provides strong evidence that no significant difference exists between the model's performance on the training and test datasets.

6. **Discuss the results and implications of your prediction analysis.**

In conclusion, the results of the prediction analysis indicate that the optimized regression model is accurate and reliable for estimating housing prices in this market. The model explains approximately 73.5% of the variance in housing prices, as shown by the R^2 and adjusted R^2 values. The minimal difference between the training MSE (5,969,309,812.83) and the test MSE (6,011,563,127.21)—an increase of just 0.71%—demonstrates that the model generalizes well to new data and is not overfitting. This is further supported by the independent t-test, which produced a p-value of 0.9078, confirming no statistically significant difference between the model's performance on the training and test sets.

Furthermore, the model's coefficients provide valuable insights for understanding which factors most strongly influence housing prices. The luxury classification has the most significant impact, with homes classified as luxury increasing in price by nearly $70,000. Square footage, number of bathrooms, and number of bedrooms also have substantial positive effects, while the age of

the home negatively impacts price. The model provides a practical and statistically sound tool for predicting housing prices based on key property features.

7. **Recommend a course of action for the real-world organizational situation from part B1 based on your results and implications discussed in part E6.**

Based on the result, I recommend that real estate professionals and organizations use this regression model as a decision-support tool when estimating property values and setting listing prices. The model can help identify which property features add the most value, allowing agents and sellers to prioritize renovations or highlight specific attributes in marketing materials. For example, investments in upgrades that move a home into the luxury category or increase usable square footage are likely to yield the highest returns. Additionally, the model can provide data-driven guidance to buyers, helping them understand which factors most influence price and allowing for more informed negotiations.