PSTAT 126 Final Project

Akshara Kollu, Praveen Arunshankar, Kangyuan Li, Lloyd Lei, Bianca Ramos-Medina

2025-02-28

1 Abstract

In this analysis, we developed a MLR model to predict housing prices based on various housing attributes. Our findings revealed that key factors like location, square footage, number of bedrooms, and amenities (such as gardens and pools) significantly influence house prices. The model explained over 95% of the variance in prices, indicating a strong fit. However, the analysis is limited by assumptions of linearity and the exclusion of potential interaction effects. Future work could explore non-linear models, additional predictors, and more complex interactions to improve the model's predictive accuracy.

2 Introduction

The Housing Prices Regression Dataset from Kaggle contains several attributes of real estate that are useful for determining the price of a property. We aim to build a multivariable linear regression model that can predict the price of a property based on these attributes.

Table 1: Housing Prices Regression Dataset - Variable Breakdown

Name	Label	Type	Categories	Valid_Obs
Identifier	ID	numeric		500
Square Feet	Square_Feet	numeric		500
Number of Bedrooms	Num_Bedrooms	numeric		500
Number of Bathrooms	Num_Bathrooms	numeric		500
Number of Floors	Num_Floors	numeric		500
Year Built	Year_Built	numeric		500
Has Garden	Has_Garden	categorical	yes/no (0/1)	500
Has Pool	Has_Pool	categorical	yes/no (0/1)	500
Garage Size	Garage_Size	numeric		500
Location Score	Location_score	numeric		500
Distance to Center	Distance_to_center	numeric		500
Price	Price	numeric		500

Research Questions:
1. How does the distance to the center of the city impact property prices?
2. Is there a correlation between the year a property was built and its price?
3. Does having a garden or pool affect the price of a property?

Hypotheses:
H1: As the distance from the city center increases, the price of a property decreases.
H2: Newer properties (built in recent years) tend to have higher prices compared to older properties.

H3: Properties with a garden or pool have higher prices compared to properties without these features.

Connecting Research Questions to the Regression Model
Our research questions focus on understanding how various property attributes influence housing prices. To address these questions, we will employ a multiple linear regression model, which allows us to analyze the impact of multiple factors on property prices simultaneously. Each research question directly translates into predictor variables in our regression model. Beyond these core variables, we will incorporate additional predictors such as Square_Feet, Number of Bedrooms/Bathrooms, Garage_Size, and Location_score to account for other key factors affecting housing prices. This ensures our model provides a more accurate and comprehensive understanding of property valuation. By running the MLR model, we can quantify the influence of each factor on price, test our hypotheses, and derive insights that align with our research objectives.

3 Data Processing

Before building our MLR model, we conducted several preprocessing steps to ensure the dataset was clean and suitable for analysis. This included checking for missing values and duplicates, as well as checking for and handling any outliers.

```
Missing Values Count for Each Column:

                ID        Square_Feet       Num_Bedrooms      Num_Bathrooms
                 0                  0                  0                  0
         Num_Floors         Year_Built        Has_Garden           Has_Pool
                 0                  0                  0                  0
        Garage_Size     Location_Score Distance_to_Center              Price
                 0                  0                  0                  0

No duplicate rows found.


No outliers for ID

No outliers for Square_Feet

No outliers for Num_Bedrooms

No outliers for Num_Bathrooms

No outliers for Num_Floors

No outliers for Year_Built

No outliers for Has_Garden

No outliers for Has_Pool

No outliers for Garage_Size

No outliers for Location_Score

No outliers for Distance_to_Center

Outliers for Price :
```

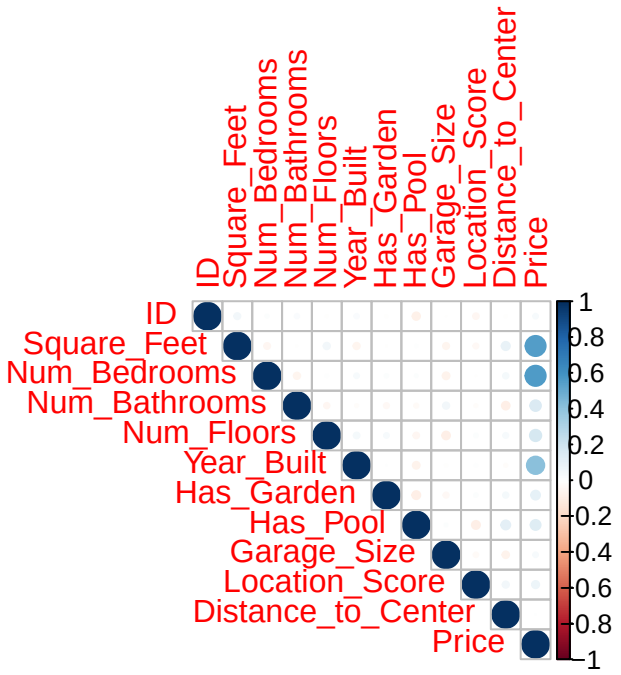
```
      Price
151 960678.3
```

There are no missing values in the dataset, so no imputation or removal of data is necessary. There are also no duplicated values. When checking for outliers in all variables, we only found one in Price. To prevent the model from being skewed unfairly by this outlier, we removed it, reducing the dataset size from 500 to 499 observations. Since our variables were already in appropriate numerical formats, no additional scaling was required. We ensured that categorical variables remained in a binary format, making them suitable for regression modeling.

4 Exploratory Data Analysis

```
                   vars   n      mean        sd    median   trimmed       mad
ID                    1 499    250.70    144.56    251.00    250.75    185.32
Square_Feet           2 499    174.44     74.61    178.02    174.41     96.06
Num_Bedrooms          3 499      2.95      1.44      3.00      2.94      1.48
Num_Bathrooms         4 499      1.98      0.82      2.00      1.97      1.48
Num_Floors            5 499      1.96      0.80      2.00      1.95      1.48
Year_Built            6 499   1957.48     35.42   1959.00   1957.10     45.96
Has_Garden            7 499      0.54      0.50      1.00      0.54      0.00
Has_Pool              8 499      0.49      0.50      0.00      0.49      0.00
Garage_Size           9 499     30.17     11.59     30.00     30.35     14.83
Location_Score       10 499      5.16      2.85      5.20      5.20      3.68
Distance_to_Center   11 499     10.46      5.59     10.86     10.57      6.82
Price                12 499 581451.18 121213.03 574652.92 578908.27 118450.92
                         min       max     range  skew kurtosis      se
ID                      1.00    500.00    499.00  0.00    -1.21    6.47
Square_Feet            51.27    298.24    246.98 -0.02    -1.26    3.34
Num_Bedrooms            1.00      5.00      4.00  0.06    -1.32    0.06
Num_Bathrooms           1.00      3.00      2.00  0.04    -1.52    0.04
Num_Floors              1.00      3.00      2.00  0.07    -1.45    0.04
Year_Built           1900.00   2022.00    122.00  0.06    -1.25    1.59
Has_Garden              0.00      1.00      1.00 -0.14    -1.98    0.02
Has_Pool                0.00      1.00      1.00  0.04    -2.00    0.02
Garage_Size            10.00     49.00     39.00 -0.07    -1.23    0.52
Location_Score          0.00     10.00      9.99 -0.08    -1.15    0.13
Distance_to_Center      0.06     19.93     19.87 -0.14    -1.09    0.25
Price              276892.47 909199.56 632307.09  0.17    -0.26 5426.24
```

```
corrplot 0.95 loaded
```

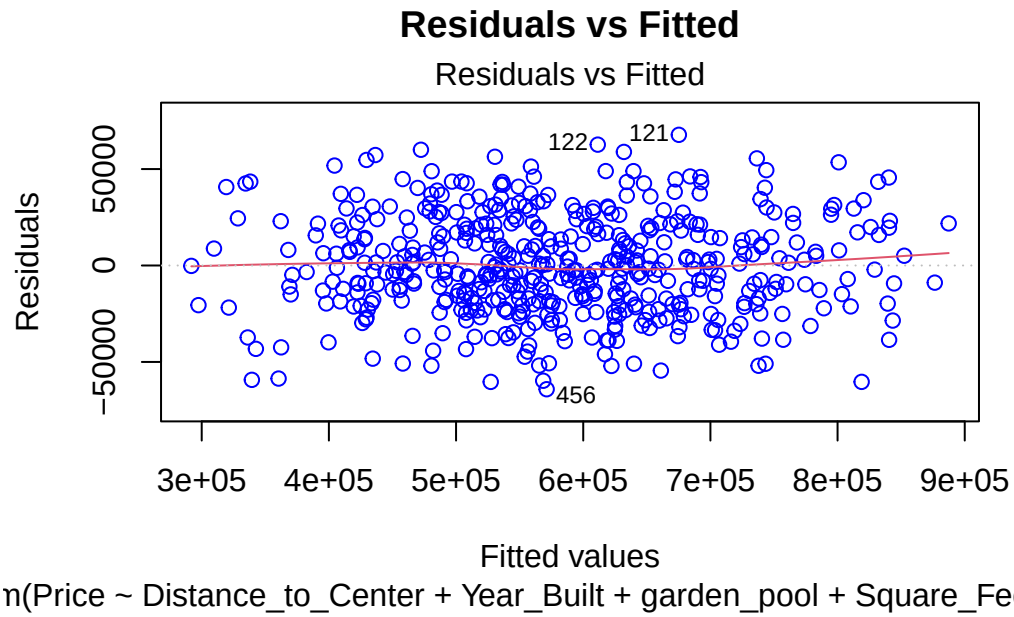


Price is the only variable highly correlated with the other factors, which is fine and will not cause multicollinearity because Price is the dependent variable, not a predictor. To confirm that MC is not an issue, we can also check the VIF.

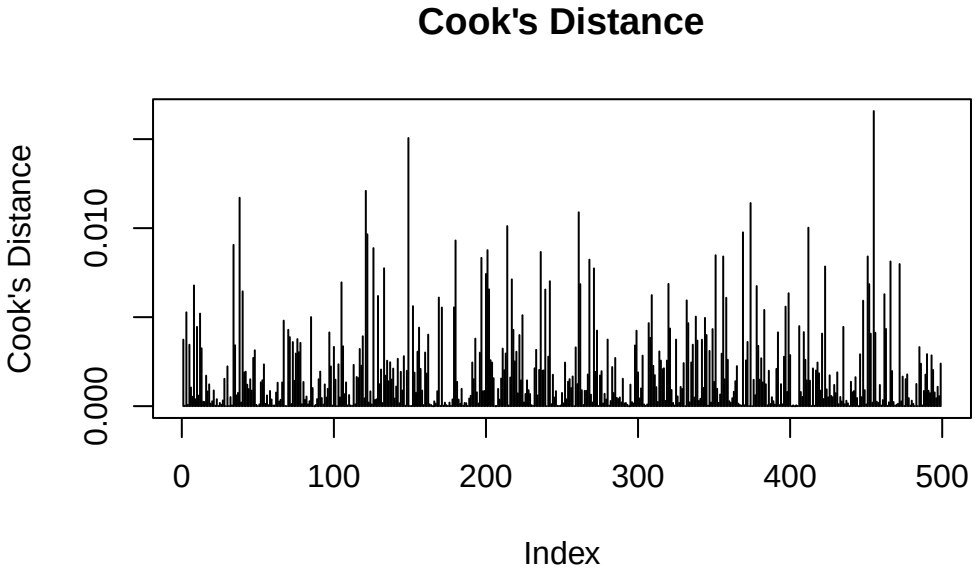
```
Variance Inflation Factor (VIF) for Independent Variables:

Distance_to_Center         Year_Built        garden_pool        Square_Feet
          1.034050           1.015149           1.023174           1.023132
      Num_Bedrooms      Num_Bathrooms         Num_Floors        Garage_Size
          1.012935           1.014048           1.013604           1.019591
    Location_Score
          1.009905
```

Because all VIF values are below 5, we have no multicollinearity issues. This means that the independent variables are not highly correlated with each other, allowing each predictor to contribute uniquely to explaining the variation in house prices

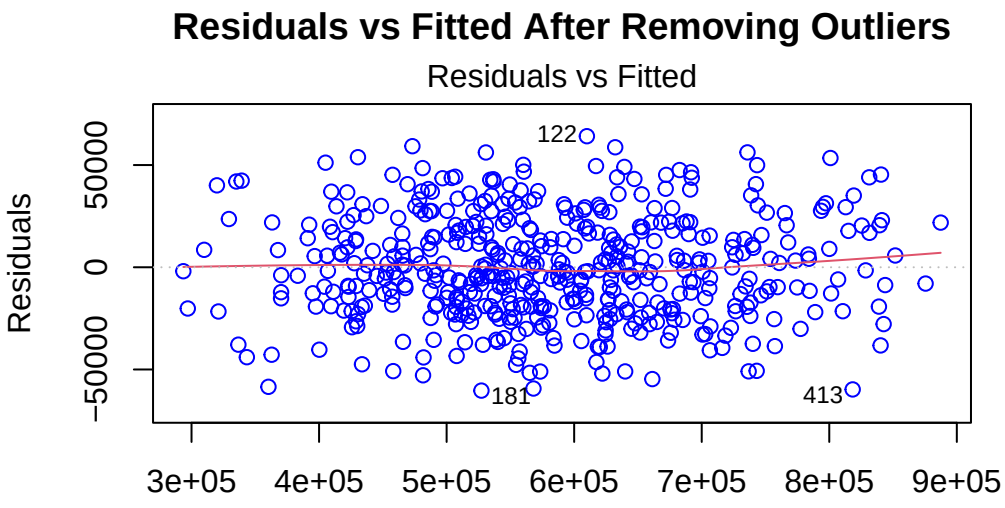
5 Visualizations



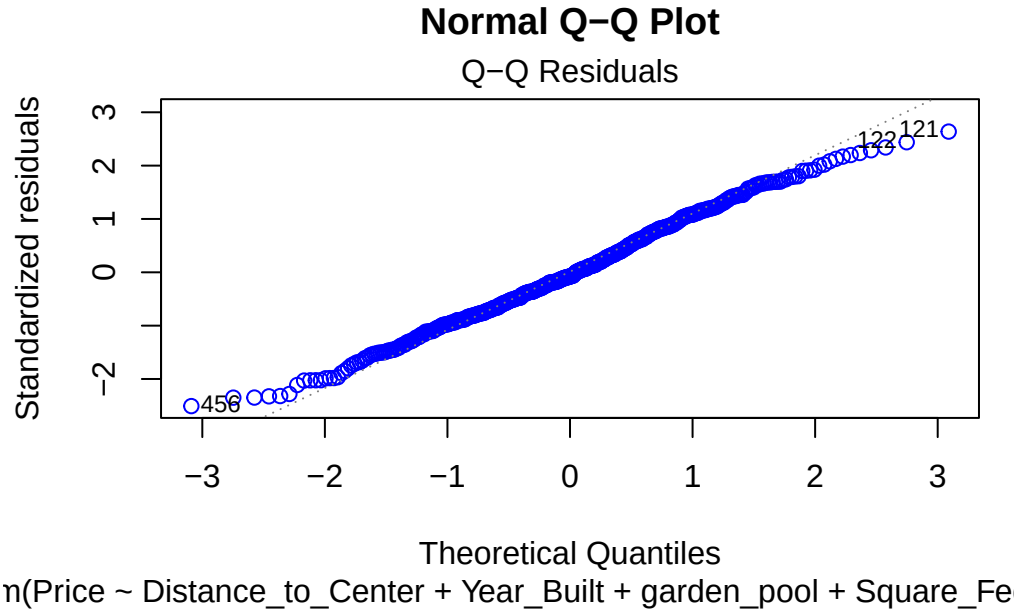
The residuals appear randomly scattered around zero, with no clear pattern or curvature, suggesting that the assumption of linearity is largely satisfied. The spread of residuals is relatively consistent across different fitted values, but there are some points where the variance appears slightly larger for higher fitted values. There is no clear funnel-shaped pattern, meaning heteroscedasticity does not seem to be a major issue. Some data points are labeled (122, 121, and 456), which suggests they might be influential observations. These outliers could be affecting the regression model disproportionately. We investigate these observations using Cook's distance to determine if they should be removed or handled differently.



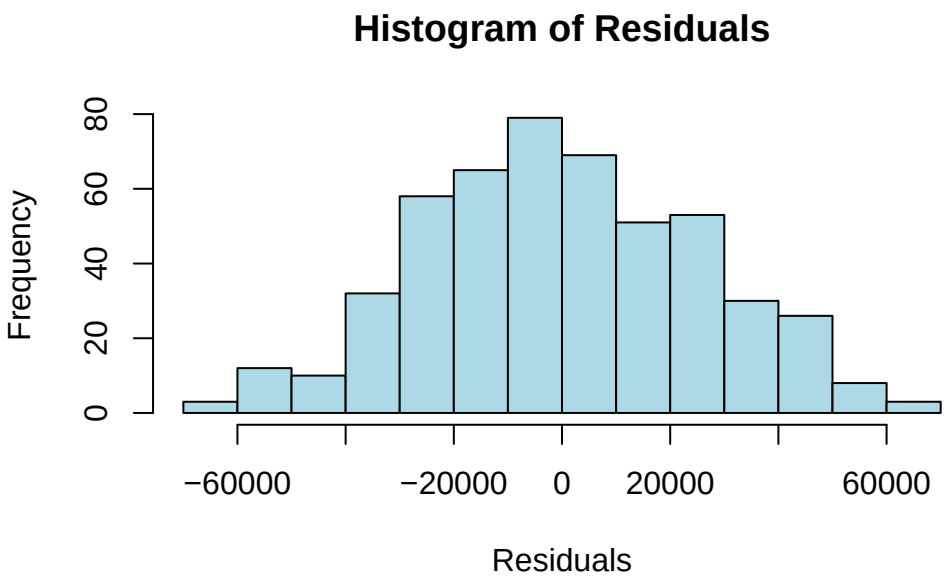
From this plot of Cook's Distance, we can see that although a few have higher influence than others, none produce a Cook's Distance greater than 0.02, indicating that there is no need to remove or handle the outliers differently. However, just as a check, let's see how our model changes if we remove the 4 most influential points by Cook's Distance.



Seeing that the spread of the residuals vs fitted plot barely changes in comparison to our original model, we can conclude that these points do not have a disproportionate influence on our model, and thus should not be removed or handled differently.



The points lie along the diagonal line, which is a good indication that the residuals follow an approximate normal distribution. However, we can further verify this with a histogram.



This plot also indicates that the residuals are approximately normally distributed, which gives us even stronger evidence that the normality of residuals assumption in our linear regression has not been violated.

6 Explaining MLR Model

```
Call:
lm(formula = Price ~ Distance_to_Center + Year_Built + garden_pool +
    Square_Feet + Num_Bedrooms + Num_Bathrooms + Num_Floors +
    Garage_Size + Location_Score, data = df_clean)

Residuals:
   Min     1Q Median     3Q    Max
-64146 -18468  -2088  19169  67753

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.905e+06  6.533e+04 -44.463  < 2e-16 ***
Distance_to_Center -1.738e+03  2.111e+02  -8.235 1.65e-15 ***
Year_Built          1.525e+03  3.302e+01  46.179  < 2e-16 ***
garden_poolYes      4.993e+04  2.858e+03  17.471  < 2e-16 ***
Square_Feet         1.019e+03  1.574e+01  64.716  < 2e-16 ***
Num_Bedrooms        5.060e+04  8.116e+02  62.347  < 2e-16 ***
Num_Bathrooms       2.825e+04  1.424e+03  19.843  < 2e-16 ***
Num_Floors          2.048e+04  1.457e+03  14.053  < 2e-16 ***
Garage_Size         1.132e+03  1.011e+02  11.200  < 2e-16 ***
Location_Score      4.503e+03  4.092e+02  11.005  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25900 on 489 degrees of freedom
Multiple R-squared:  0.9552,    Adjusted R-squared:  0.9543
F-statistic:  1157 on 9 and 489 DF,  p-value: < 2.2e-16
```

We utilized a MLR model because it allows us to analyze the relationship between Price and multiple independent variables that impact housing prices while still explaining the variability in the data. The factors we chose—Square Footage, Number of Bedrooms/Bathrooms, Number of Floors, Garage Size, Location Score, Distance to Center, Year Built, and Garden/Pool Presence—are all important variables that contribute to explaining the variation in house prices. All of these factors were found to be statistically significant, indicating that they have a meaningful impact on predicting the price. By verifying key assumptions such as linearity, homoscedasticity, and the absence of multicollinearity, we ensured that the model is valid and interpretable.

7 Interpreting Results

The R^2 value of 0.9552 indicates that approximately 95.52% of the variance in house prices is explained by the selected predictors. The Adjusted R^2 of 0.9543 shows a minimal drop when adjusting for the number of predictors, suggesting that the model is well-fitted and that the inclusion of additional variables does not overly complicate the model. The F-statistic of 1157, with a p-value $<$ 2.2e-16, indicates that the model is highly significant overall and that the predictors collectively explain a substantial portion of the variation in house prices.

Every predictor in the model was deemed statistically significant by p-value, but some have much greater influence than others. By inspecting the magnitude of the t-values, we find that square footage, number of bedrooms, and the year the home was built explained the greatest amount of variance in house prices. If we run the MLR again without these 3 variables, the R^2 drops from 0.95 all the way down to 0.08, demonstrating that those 3 predictors explain ~87% of the variance in home prices just by themselves.

```
Call:
lm(formula = Price ~ Distance_to_Center + garden_pool + Num_Bathrooms +
    Num_Floors + Garage_Size + Location_Score, data = df_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-278945  -83387   -5664   75976  315136

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       423639.0    29521.2  14.350  < 2e-16 ***
Distance_to_Center   -191.4      949.1  -0.202 0.840274
garden_poolYes      37997.9    12857.4   2.955 0.003273 **
Num_Bathrooms       23859.5     6424.6   3.714 0.000228 ***
Num_Floors          27533.5     6571.4   4.190 3.31e-05 ***
Garage_Size           433.0      455.5   0.951 0.342196
Location_Score       3051.3     1846.8   1.652 0.099124 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 117100 on 492 degrees of freedom
Multiple R-squared:  0.0786,    Adjusted R-squared:  0.06737
F-statistic: 6.995 on 6 and 492 DF,  p-value: 3.734e-07
```

This shows that just by knowing a home's square footage, the year it was built, and how many bedrooms it has, you can predict the home price with a very respectable degree of accuracy.

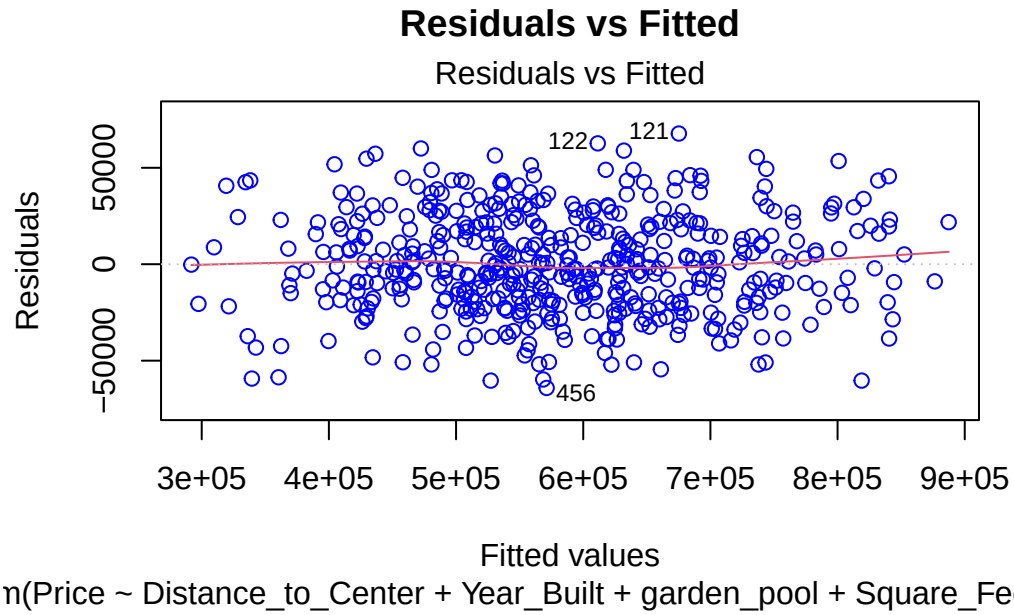
8 Checking Assumptions

The first assumption to verify is the normality of residuals. From our aforementioned Normal Q-Q Plot and residual histogram, we have a pretty good indication that the residuals are normal. However, just to verify, let's make sure using the Shapiro-Wilks Normality test under the null hypothesis that the data is normal.

```
    Shapiro-Wilk normality test

data:  residuals(model5)
W = 0.99456, p-value = 0.07361
```

Since the p-value is greater than $\alpha = 0.05$, we do not reject the null, and do not have evidence that the residuals are non-normal. In other words, there is no reason to believe the normality of residuals assumption was violated.

Next, we need to verify the assumption of linearity. For this, we use our residuals vs. fitted values plot from earlier and ensure the residuals are randomly scattered around 0.



The residuals do appear to be randomly distributed around 0, so we conclude the assumption of linearity is not violated.

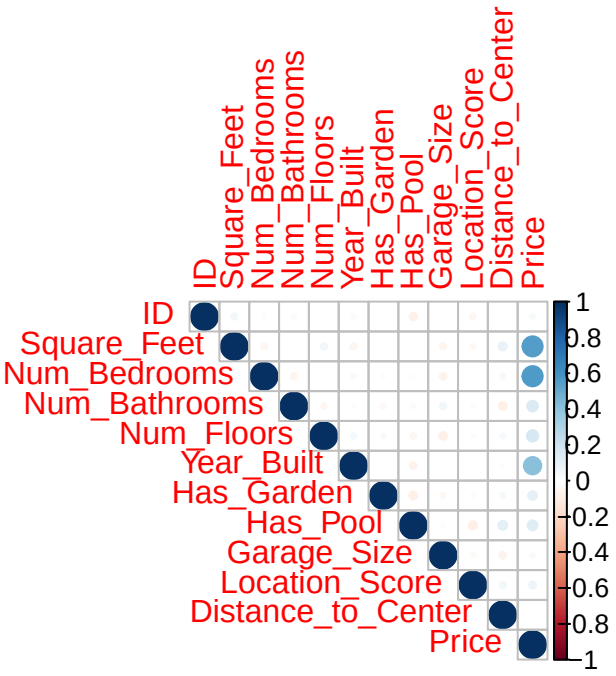
We can also use this plot to check the assumption of homoscedasticity, or constant variance between residuals. Since there is no funnel shape or any sort of pattern in the plot, there is no heteroscedasticity present, and thus we have no reason to believe this assumption is violated.

The next assumption to verify is that all the residuals are independent of each other. For this, we use the Durbin-Watson Test under the null hypothesis that there is no autocorrelation.

```
    Durbin-Watson test

data:  model5
DW = 1.8916, p-value = 0.1132
alternative hypothesis: true autocorrelation is greater than 0
```

Since the p-value is greater than $\alpha = 0.05$, we fail to reject the null and conclude that we do not have evidence that there is autocorrelation in our model. This confirms our assumption that the residuals are independent.

The final assumption we need to verify is that there is no multicollinearity between the independent variables. In our exploratory data analysis, we used this correlation matrix to investigate this:



No two predictor variables have a correlation coefficient over 0.8, which indicates no multicollinearity. To double check this, we can refer to the Variance Inflation Factor (VIF) like we did earlier.

```
Variance Inflation Factor (VIF) for Independent Variables:

Distance_to_Center         Year_Built        garden_pool        Square_Feet
          1.034050           1.015149           1.023174           1.023132
      Num_Bedrooms      Num_Bathrooms         Num_Floors        Garage_Size
          1.012935           1.014048           1.013604           1.019591
    Location_Score
          1.009905
```

As we found before, none of the variables have a VIF even close to 5, which verifies the assumption that there is no multicollinearity between independent variables in our model.

9 Key Insights to Predictors

Based on the multiple linear regression (MLR) model, we identified several key factors that significantly influence housing prices:

- Square Feet: The most influential factor-each additional square foot increases house prices substantially. Larger homes are generally valued higher.
- Num Bedrooms & Num Bathrooms: Both variables positively affect house prices. However, the effect of the number of bedrooms is less pronounced when house size is accounted for.
- Location Score: Highly positively correlated with price—houses in better-rated locations tend to be more expensive.
- Distance to City Center: Negative impact—houses farther from the city center tend to have lower prices.
- Garage Size: Larger garages increase house prices, although the effect is smaller than factors like location or size.
- Has Pool & Has Garden: Homes with pools and gardens have significantly higher values, indicating these features are desirable in the real estate market.

- Year Built: Newer homes generally have higher prices, but the effect is less significant compared to other factors.

These findings suggest that house size, location, and amenities (such as pools and gardens) are key determinants of housing prices.

One potential refinement would be to explore interaction effects between certain variables. For example, the impact of square footage may depend on the number of bedrooms and bathrooms. Similarly, the effect of location score might differ depending on distance to the city center. Future models incorporating interaction terms could provide deeper insights into these relationships.

10 Limitations

Our current model, while effective in predicting house prices, has several limitations that need to be addressed for more strategic real estate investment decisions. One of the key issues is the lack of critical external factors such as crime rates, school district quality, local economic conditions, and neighborhood demographics. These elements significantly influence property values, and their exclusion limits the model's predictive accuracy. Enhancing the model by incorporating these variables could align it more closely with real-world investment considerations.

Additionally, the assumption of a strictly linear relationship between predictors and prices oversimplifies the complexity of real estate markets. House pricing is highly nonlinear, with factors like diminishing returns on square footage and intricate location dependencies playing crucial roles. To capture these complexities, advanced machine learning models such as Support Vector Machines (SVM) for property classification and K-Means clustering for neighborhood segmentation could improve prediction accuracy and strategic insights.

Further refinement involves time-series analysis—since real estate prices fluctuate over time, incorporating historical pricing data and macroeconomic trends (such as interest rates or inflation) could further enhance predictive power. Moreover, applying feature selection techniques, such as LASSO regression, could help identify the most critical predictors, reducing model complexity while maintaining high accuracy.

Another significant limitation is the model's static nature, which fails to account for market trends and historical price fluctuations. Real estate markets are influenced by economic cycles, mortgage rates, and inflation, necessitating the integration of time-series data to enhance predictive capability. By incorporating historical data and employing regression-based forecasting or machine learning techniques like LSTM, we can perform price backtesting and project future market trends more accurately. Furthermore, the dataset's geographic constraints limit its generalizability, as pricing factors vary across regions due to zoning laws, rent control policies, and local economic conditions. Implementing regional segmentation analysis and training location-specific models would enhance the model's applicability across different markets. Future improvements should focus on expanding predictors to include key economic and social indicators, utilizing advanced machine learning algorithms to capture complex relationships, and integrating dynamic forecasting techniques to improve decision-making for real estate investments.

11 Recommendations

For buyers and investors, prioritizing location is crucial, as properties with higher location scores tend to appreciate more in value. Investing in larger homes generally yields higher returns, and properties with unique features such as gardens, pools, and larger garages command higher valuations. Proximity to city centers is another key factor, as homes closer to downtown areas tend to hold better long-term value. While newer homes are often priced higher, well-maintained older properties can also offer good investment potential. For real estate developers, focusing on high-scoring locations and emphasizing amenities such as gardens, pools, and ample garage space can justify premium pricing. Additionally, balancing house size with usability is important, as larger homes sell for more, but optimizing interior space efficiently remains crucial.

Market analysts should refine the model by exploring nonlinear relationships and interaction effects to improve accuracy. Monitoring macroeconomic indicators, including interest rates and inflation, can help in making informed pricing strategies. Furthermore, customizing the model for specific geographic markets allows for more precise and applicable insights, ensuring that investment decisions align with regional market conditions.

City planners and policymakers can use these findings to guide zoning laws and urban development strategies. Encouraging the construction of affordable, high-location-score housing near city centers could help balance market demand. Similarly, incentivizing home improvements, such as energy-efficient upgrades, pools, or gardens, might enhance property values in developing neighborhoods.

12 Conclusions

Our analysis successfully addressed the initial research questions. We found that property prices generally decrease as the distance from the city center increases, which supports Hypothesis H1. This highlights the importance of location, as homes closer to urban centers tend to be more valuable due to better accessibility and higher demand. Hypothesis H2 was also confirmed by the positive relationship between Year Built and Price. Meaning newer properties tend to be priced higher, likely due to modern construction standards and updated amenities. Lastly, Hypothesis H3 was strongly supported, showing that homes with features like a garden or a pool are significantly more expensive, because of their appeal to buyers.

These findings have practical implications for homebuyers, sellers, and real estate professionals. Buyers may prioritize properties closer to the city center, with more space and modern features, while sellers could focus on enhancing property amenities (like gardens or pools) to increase value. Real estate professionals could use these insights to better guide clients in setting prices or selecting homes based on their preferences.

While our model captures important factors that influence housing prices, future research could refine it further by integrating additional socioeconomic factors such as neighborhood safety, school district quality, and local economic conditions, which are known to significantly influence real estate markets. Moreover, implementing advanced spatial econometric models could provide deeper insights into geographic price clustering and help improve property valuation accuracy. Expanding the analysis through non-linear modeling techniques, such as Random Forest regression or neural networks, would allow for better detection of complex relationships between variables that linear models may not catch.

Another interesting direction for future studies would be using clustering techniques to identify distinct housing market segments. Methods like k-means or hierarchical clustering could help group properties based on shared characteristics, which offer deeper insights into different submarkets. Additionally, applying spatial regression could improve the accuracy of price predictions by considering regional economic factors.

By addressing these limitations and incorporating advanced methodologies, future research could refine property valuation models and provide more robust insights for homebuyers, sellers, investors, and policymakers. Understanding the complex relationships between housing features, location, and market trends is crucial for making data-driven real estate decisions that align with economic realities and buyer preferences.