MTH 209 Final Project

Greg Pologruto
University of Dayton
2025-12-02

Abstract

Housing Market Analysis

This study investigates the key factors that influence housing prices using a dataset of 545 residential properties containing structural, amenity, and location features. The goal is to identify which characteristics are most strongly associated with market value and to evaluate how well a multiple linear regression model can predict home prices. The analysis explores relationships between square footage and price, assesses whether the number of bathrooms and bedrooms meaningfully contributes to value, and examines the impact of location-based variables. Statistical methods including correlation analysis, visualization, and multiple linear regression are used to assess significance and predictive accuracy. Overall, this project aims to provide a clear understanding of how different home features contribute to price variation and to illustrate the usefulness of regression modeling in real estate price estimation.

Research Questions

Main Question

- How well can a multiple linear regression model predict home prices?

Subquestions

- Is there a linear relationship between square footage and house price?
- Do houses with more bathrooms have higher average market prices than houses with fewer bathrooms?
- Does the number of bedrooms still have a significant effect on price after accounting for square footage?
- Does furnishing status affect market prices?
- What factors are most strongly associated with housing price in this data?

Background/Significance

Understanding the factors that influence the housing prices is a main topic in real estate, urban planning, and more. The dataset used in this study provides detailed information on 545 homes. This information includes structural details like what amenities, the square footage, the number of bedrooms, and others. It also includes location based features like main-road access and area.

This data will allow for a comprehensive exploration of how different features contribute to the market price of a house. Exploring like square footage and number of bedrooms reflect on the physical size of a home while amenities such as air conditioning or parking represent the quality of life enhancements that may drive price. This data provides great variables for understanding housing price behavior by applying statistical methods.

Methods

Process

For this study, I want to study the relationship between price and the rest of the variables in the data. First, I will conduct an exploratory data analysis on the dataset. I will look through the distribution of price and furnishingstatus. I will also explore the relationship between area and price and look into the distribution of price by number of bathrooms.

Then, I will create a multiple linear regression using every variable to predict price. I will analyze this model before creating a second model using selected variables. I will compare the performance of these two models.

Cleaning Data

The data itself had no missing values. The only thing I needed to do was convert the character variables to factors.

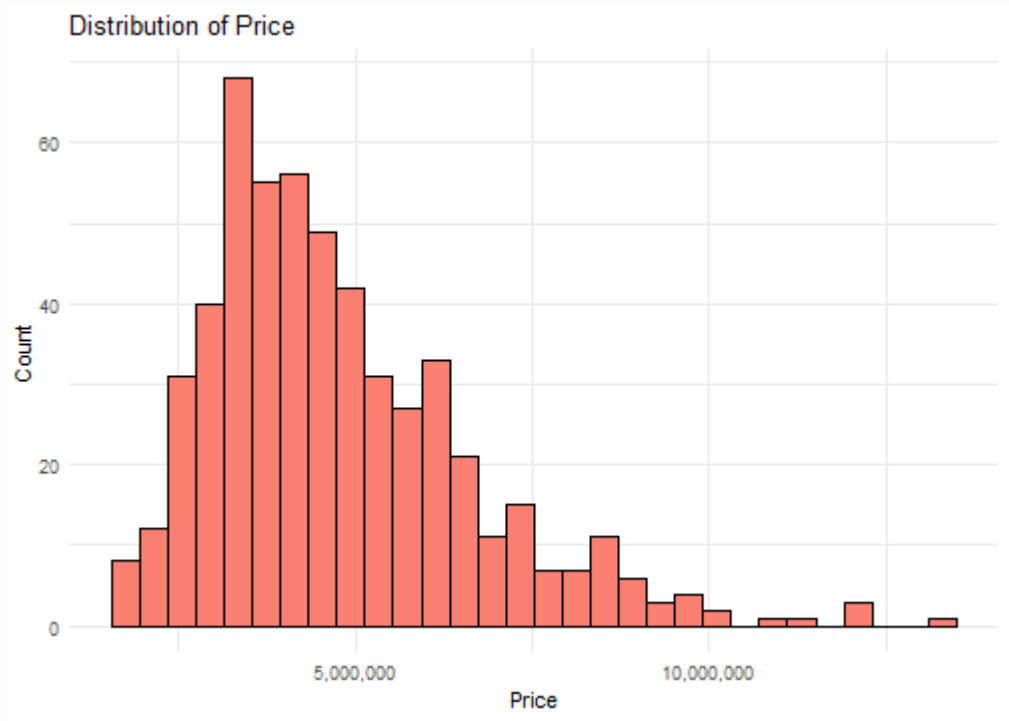
EDA

Dataset has 13 variables with 545 observations.

```
                                    ## Rows: 545
                                   ## Columns: 13
## $ price            <dbl> 13300000, 12250000, 12250000, 12215000, 11410000, 108…
## $ area             <dbl> 7420, 8960, 9960, 7500, 7420, 7500, 8580, 16200, 8100…
## $ bedrooms         <dbl> 4, 4, 3, 4, 4, 3, 4, 5, 4, 3, 3, 4, 4, 4, 3, 4, 4, 3,…
## $ bathrooms        <dbl> 2, 4, 2, 2, 1, 3, 3, 3, 1, 2, 1, 3, 2, 2, 2, 1, 2, 2,…
## $ stories          <dbl> 3, 4, 2, 2, 2, 1, 4, 2, 2, 4, 2, 2, 2, 2, 2, 2, 2, 4,…
## $ mainroad         <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes…
## $ guestroom        <fct> no, no, no, no, yes, no, no, no, yes, yes, no, yes, n…
## $ basement         <fct> no, no, yes, yes, yes, yes, no, no, yes, no, yes, yes…
## $ hotwaterheating  <fct> no, no, no, no, no, no, no, no, no, no, no, yes, no, …
## $ airconditioning  <fct> yes, yes, no, yes, yes, yes, yes, no, yes, yes, yes, …
## $ parking          <dbl> 2, 3, 2, 3, 2, 2, 2, 0, 2, 1, 2, 2, 1, 2, 0, 2, 1, 2,…
## $ prefarea         <fct> yes, no, yes, yes, no, yes, yes, no, yes, yes, yes, n…
## $ furnishingstatus <fct> furnished, furnished, semi-furnished, furnished, furn…
```

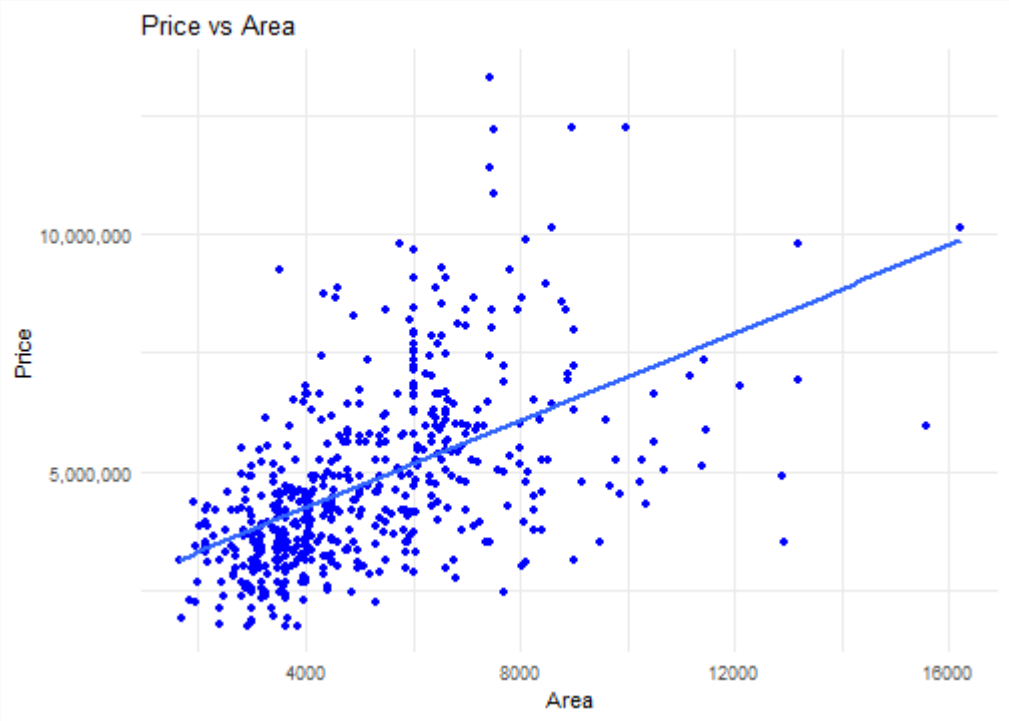
EDA Cont. (Price)

Distribution of price. Please note that price is not in USD, I am unsure of what exactly the currency is since it does not say in the description. The data is skewed right. It is unimodal and has a peak around 3 million. This histogram does show evidence of some outliers above the mean.



EDA Cont. (Price vs Area)

This scatterplot shows evidence of a linear relationship between area and price. There seems to be some points that have high residuals, they might be influential. There are also some points that are possibly leverage points.

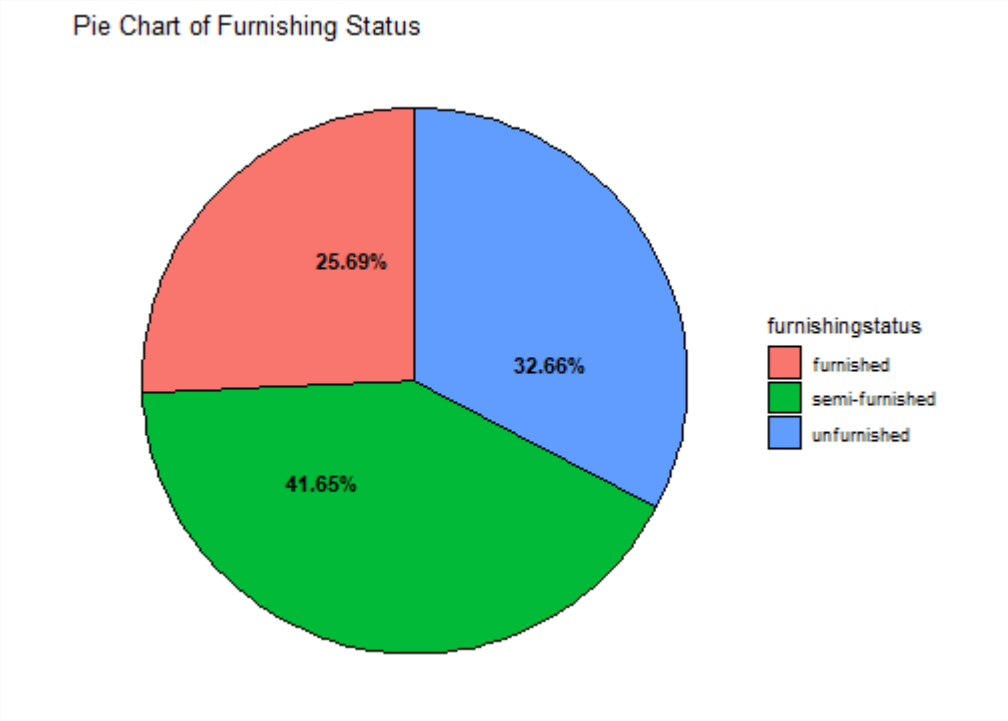


EDA Cont. (Bathrooms)

This boxplot clearly shows a relationship between the number of bathrooms and the market price. The median value increases each time the number of bathrooms increase. This clearly shows that the number of bathrooms are correlated to price.

EDA Cont. (Furnishing)

Semi-furnished homes make up 41.65% of the dataset. That is the most common category. It is also noteworthy that more houses are unfurnished than furnished.



Multiple Linear Regression

Before using a multiple linear regression model, these assumptions must be checked.

Assumptions

1. Linear Assumption: The relationship between the response and the regressors is linear.
2. Zero mean Assumption: The error term ϵ has zero mean.
3. Equal Variance Assumption: The error term ε has constant variance σ^2.
4. Independent Assumption: The errors are uncorrelated.
5. Normality Assumption: The errors are normally distributed.

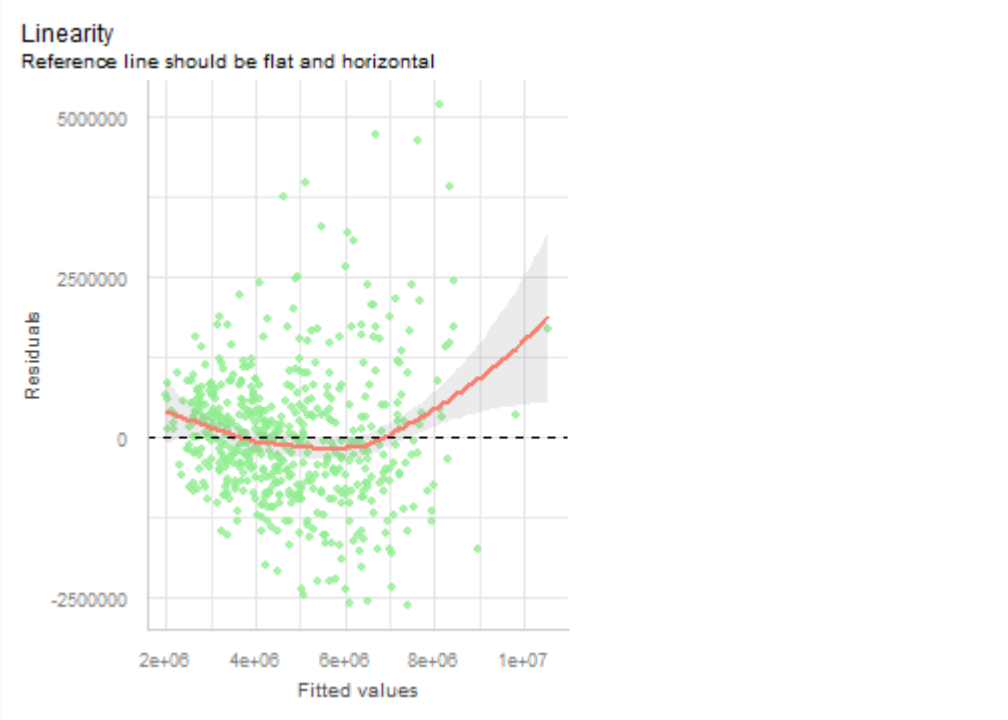
Importance

Assumptions must be check to ensure that conclusions based on the model reflect real patterns and that the model can be used to reliably make predictions on new or unseen data.

Linear and Zero Mean Assumption

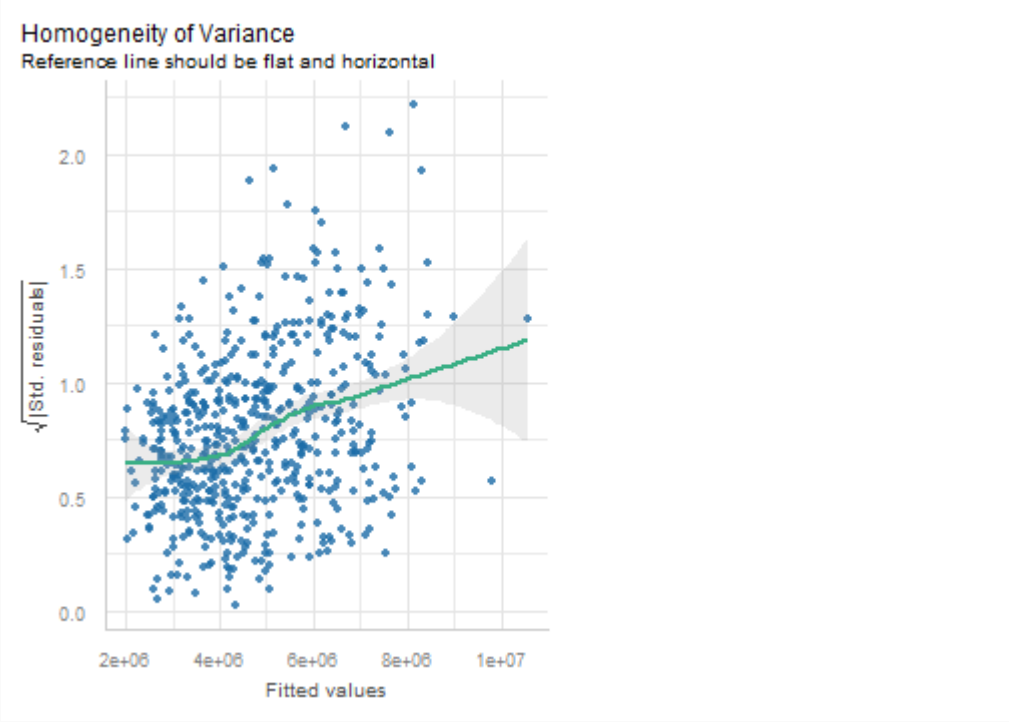
The linear assumption is checked using a Residuals vs Fitted plot. In this plot there is a lack of distinct patterns, that indicates a linear relationship.

The zero mean assumption is checked using the same plot. There should be roughly an equivalent number of points above and below the red line. In this case there is, so we can assume the error term has zero mean.



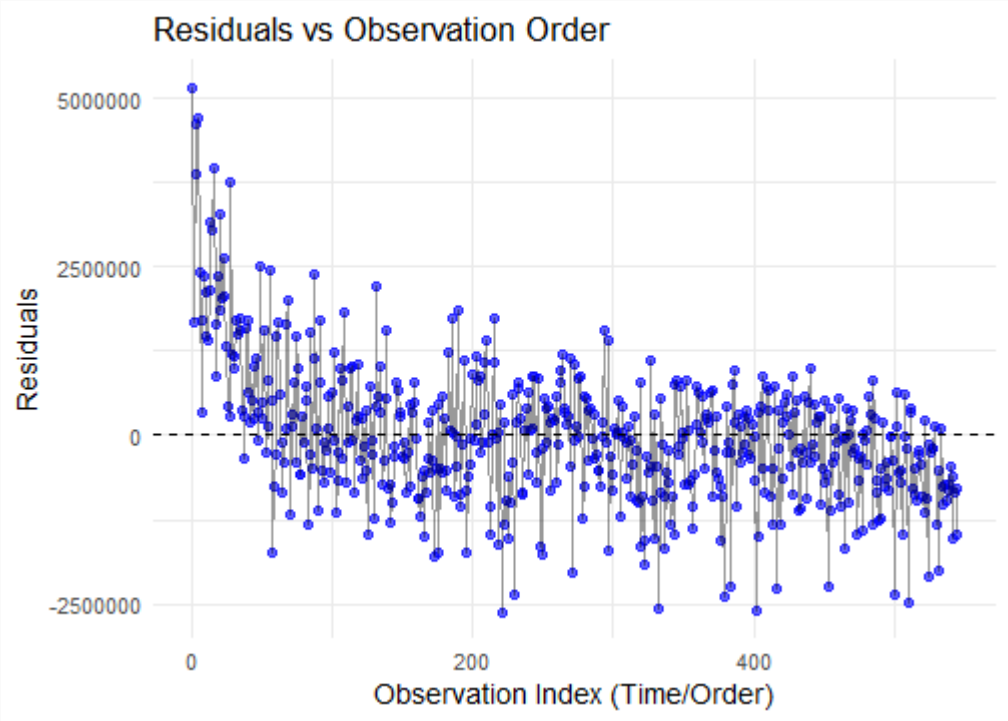
Equal Variance Assumption

The equal variance assumption is checked using the homogeneity of variance plot. The reference line is not entirely flat and horizontal however, this is real-world data, it will not look perfect on these plots. In this case, it is good enough to assume that the error term has a constant variance.



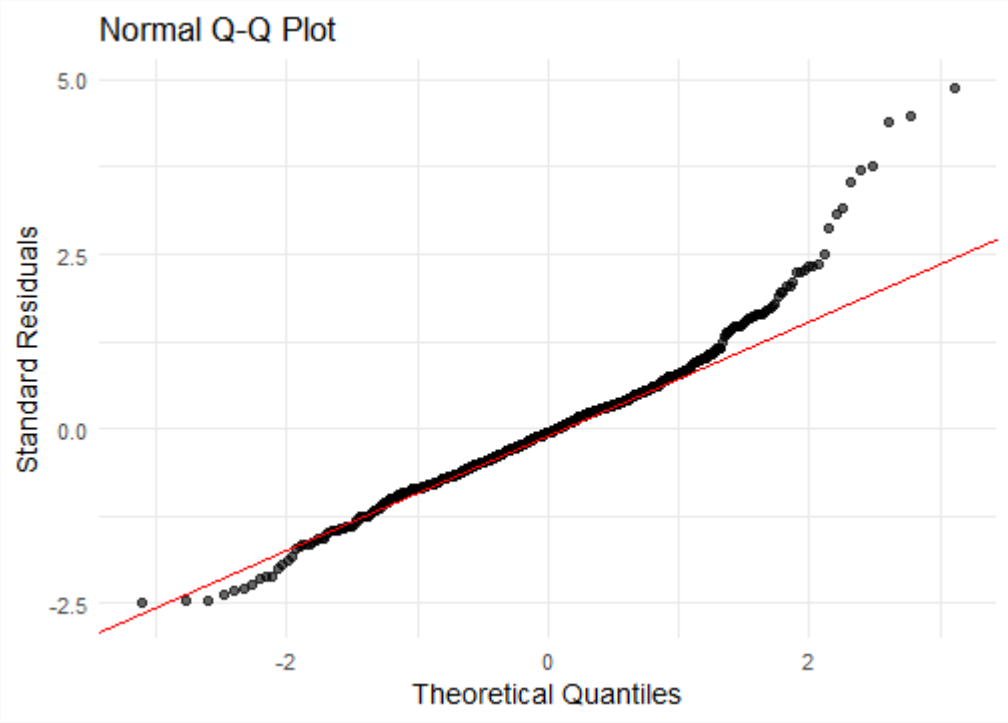
Independent Assumption

Since the data comes in order, check for evidence of patterns that suggest the observations may not be independent. In this plot, there do not seem to be any distinct patterns. We can assume that the observations are independent.



Normality Assumption

The normality assumption can be checked using the normal quantile-quantile plot. The plotted residuals should approximately follow the 45 degree red line. In this case, they do follow the line. Normality can safely be assumed.



Multiple Linear Regression Cont.

```
fit_1 ← lm(price ~ ., data=house)
```

Since all assumptions can safely be made, we can now evaluate the model. The model made from all the other variables has an adjusted R^2 value of 0.674. That means that the model accounts for 67.4% of the variation in price.

```
## [1] "Adjusted R-Squared:"
```

```
## [1] 0.6740117
```

Multiple Linear Regression Cont.

Get Significant Variables

From the first model, we should see which variables are significant and use them to create a new model.

```
## [1] "Significant Variables (p < 0.01)"
```

```
##  [1] "(Intercept)"                "area"
##  [3] "bathrooms"                  "stories"
##  [5] "mainroadno"                 "basementno"
##  [7] "hotwaterheatingno"          "airconditioningno"
##  [9] "parking"                    "prefareano"
## [11] "furnishingstatusunfurnished"
```

Select Variables

From this test we will use area, bathrooms, stories, mainroad, hotwaterheating, airconditioning and parking to create a new model and compare the two.

Multiple Linear Regression Cont.

New Model

```
fit ← lm(price ~ area + bathrooms + stories + mainroad + hotwaterheating +
airconditioning + parking, data=house)
```

The new model has an adjusted R-square value of 0.618, meaning the model accounts for 61.8% of the variation in prices. The coefficient for area is 280.2, that means for every square foot area increases, price increases by 280.2 units.

```
## [1] "Adjusted R-Squared:"
```

```
## [1] 0.6182895
```

```
## [1] "Coefficient for area"
```

```
##     area
## 280.2008
```

Compare Models

To compare the two models we can look at multiple performance indicators and select the model that is better.

Adjusted R-squared

The initial model's adjusted R^2 value is 0.674. The second model's value was 0.618. The first model does better in this metric.

AIC

For AIC, smaller values indicate better models. In this case, the first model has a smaller AIC than the second.

```
##       df      AIC
## fit_1 15 16693.01
## fit_2  9 16773.13
```

Compare Models Cont.

BIC

Smaller BIC values also indicate better models. The BIC value for the first model is smaller than the second.

```
##       df      BIC
## fit_1 15 16757.52
## fit_2  9 16811.84
```

Select Model

The original model using all the other variables to predict price is much better than the second model I created. It has a higher adjusted R^2 and lower AIC and BIC values. It provides a more accurate prediction of price.

Discussion of the Study

Main Question

The main topic in this study was to see if a multiple linear regression model can predict market prices in houses. The model we created had an adjusted R^2 value of 0.674. This number is not incredibly high however just because the adjusted R^2 is low that the model is bad. When using real data, the relationship is not always linear.

Limitations

The major limitation in this study was the dataset. It is relatively small and does not have a lot of other variables. More information on location, amenities and other details about the house might have helped improve the model.

Future Improvements

If I were to continue studying this data, I would look to improve my model in many ways. One strategy I could try is performing transformations on some variables. An example is instead of predicting price, predict log(price). I could also look into using nonlinear models or trying out different variables in a multiple linear regression model.

References

Sources

- Housing Dataset
- xaringan package
- MTH 209 Class Notes
- MTH 369 Class Notes

AI Usage

AI was used in this project to help create color palettes for the graphs, format slides, and debug code.

About Me

Bio

My name is Greg Pologruto. I am a Junior Computer Science student at the University of Dayton with minors in Mathematics and Data Analytics. I have experience working with an agile team and have skills in python, R, and PowerBI. My goal is to acquire a data science internship this summer.

Contact Me

- pologrutog1@udayton.edu
- LinkedIn
- GitHub