Influence of Proximity to Railway on House Prices

36-617: Applied Linear Models

Jason Andriopoulos, Yahan Yang

2024-10-24

Executive Summary

The project adopted a linear regression method to investigate the impact of proximity to rail trails on nearby residential property values in Northampton, Massachusetts in 2007. Using data including amenities, housing price, distance and other neighborhood factors, the research is aimed to detect whether distance determines house prices. The report uses linear model regression method to examine this problem.

Key findings from our regression analysis indicate that distance to the rail trail negatively influences home prices. For every unit increase in distance from the trail, house prices decrease by approximately 6.39%. The confidence interval is between 3.26% and 9.52%. Other features are also influencing house price. Home square footage (with a 38.75% price increase for every 1000 sq. ft.) and number of bedrooms positively impact house prices (the influence is significant only for houses with 3 bedrooms), while the number of garage spaces and certain zip codes have lesser or no significance.

In conclusion, the analysis reveals that rail trails appear to be a desirable feature for homebuyers, making them a significant consideration in real estate development strategies.

Introduction

Since the late 19th century, the landscape of American transportation has seen many transformations. In the late 1800s and early 1900s, the United States introduced a significant expansion of rail lines which was used for connecting cities for both passenger and cargo services. These rail networks would soon become obsolete with the emergence of automobiles and the Interstate Highway System. More recently, many towns are seeing the conversion of these abandoned rail lines into a repurposed rail trail. These new rail trails offer neighborhoods new ways to enjoy their atmospheres through walking and biking on these new, long, continuous, and gentle slopes. We suspect this transformation has sparked interest in homeowners and may increase the desirability and value of nearby homes.

This study aims to investigate the relationship between proximity to these new rail trails and residential property values in Northampton, Massachusetts. More specifically, we will answer Mr W. E. Coyote's question: Are rail trails appealing to those looking to buy a home, and if so, what influence does the distance to a rail trail have on a given house's value?

This research is important because it can inform insights in real estate development strategies, and the public policies surrounding the decisions behind what to do with abandoned rail trails. Understanding economic impacts of these rail trails on home values can provide valuable insights for Acme Homes, LLC when setting home prices.

Data

The data was collected from houses sold in Northampton, Massachusetts in 2007. A new rail trail was opened in Northampton in 1984, offering an opportunity to compare the values of homes near the trail and farther from the trail.

We start our statistical analysis with descriptive statistics to summarize the data. There are 104 observations and 20 variables in total. The dataset includes the following variables:

- housenum: The unique number for each house
- adj1998: The estimated house price from Zillow in 1998 (in thousands of 2014 USD)
- adj2007: The estimated house price from Zillow in 2007 (in thousands of 2014 USD)
- adj2011: The estimated house price from Zillow in 2011 (in thousands of 2014 USD)
- price1998: The estimated house price from Zillow in 1998 (in thousands of USD)
- price2007: The estimated house price from Zillow in 2007 (in thousands of USD)
- price2011: The estimated house price from Zillow in 2011 (in thousands of USD)
- distance: The distance (miles) to the nearest rail trail network
- acre: Number of acres of the given property
- bedrooms: Number of bedrooms in each property
- bikescore: The friendliness of biking in the area with a 0-100 scale.
- walkcore: The friendliness of walking in the area with a 0-100 scale.
- garage_spaces: Number of garage spaces in each property
- latitude: The latitude of the given property
- longitude: The longitude of the given property
- squarefeet: The square footage of a property's inside space
- streetname: The name of the street where the property is located
- streetno: The number of the house on a given street
- Zip: The zipcode of the given property (binary variable)

Before conducting a formal Explotary Data Analysis (EDA), we plot a Directed Acyclic Graph (DAG) to visualize the relationships between various property characteristics and the adjusted house prices in 2007, with a focus on the impact of the distance to the rail trail.

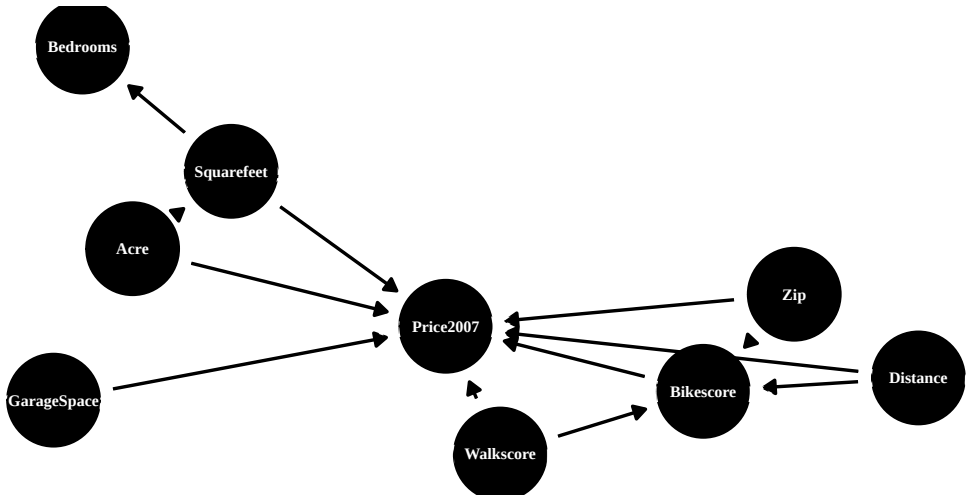

Figure 1: DAG of Variables of Interest

The DAG provides an intuitive way to start analyzing the data. To proceed from here, some key relationships are worth noting in regards to the distance to the rail trail influences walk score and bike score. This makes sense because those scores are essentially calculated by how close they are to amenities, such as the rail trail. We also have reason to believe that square footage is impacted by the property size, which in turn has an impact on the number of bedrooms and garage spaces. We will investigate walk scores and bike scores with distance to rail trails to see if we have reasonable evidence to remove them from our final model.

In our further analysis, we chose to exclude bike score and walk score from the regression model. These variables are highly correlated with the distance to the rail trail (see graph below), as they are calculated based on proximity to the rail trail itself. Including them could introduce multicollinearity, leading to unreliable coefficient estimates. By focusing on the distance to the rail trail, we aim to provide a clearer and more interpretative model that accurately captures the relationship between proximity to the rail trail and house values.

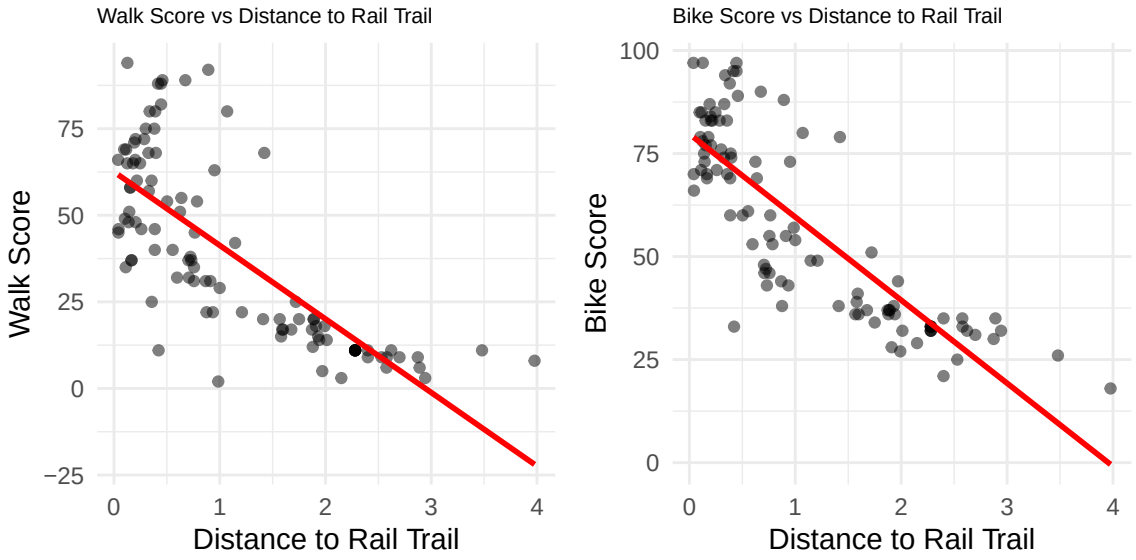

Figure 2: Walk Score and Bike Score vs. Distance to Rail Trail

Next, we examine the distributions of our response variables and covariates to ensure they are normally distributed. This step is crucial for validating the assumptions of our statistical models and ensuring the reliability of our analysis. More specific, the key assumption is for the observations to follow a normal distribution.

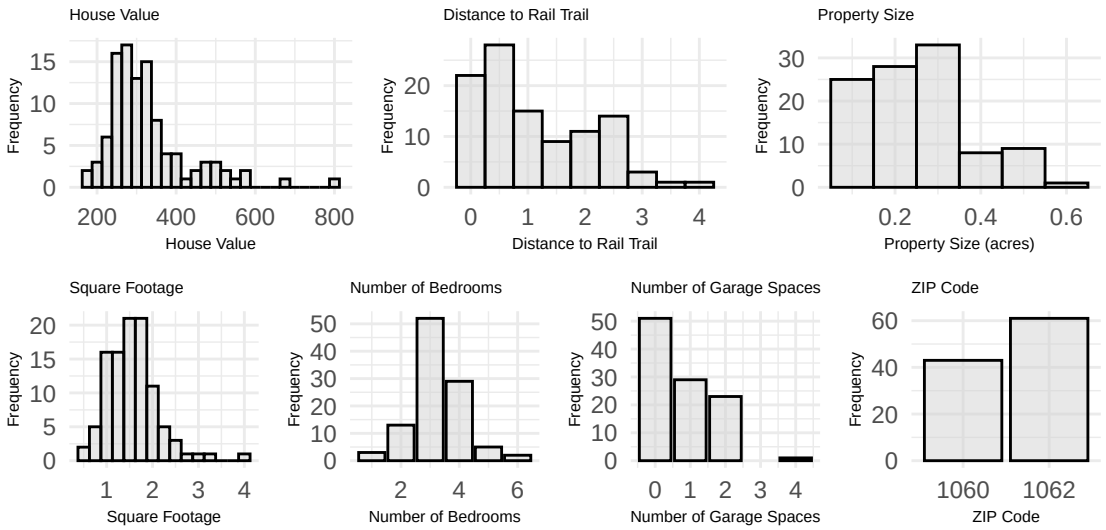

Figure 3: Histograms showing the distributions of key housing characteristics

We noticed from Figure 3 that all of our variables appear to be right-skewed. More concerningly, there are very few homes with 5 and 6 bedrooms, and only one home with 4 garages and none with 3 garages. This made us curious about the price distributions of these homes compared to homes with fewer bedrooms

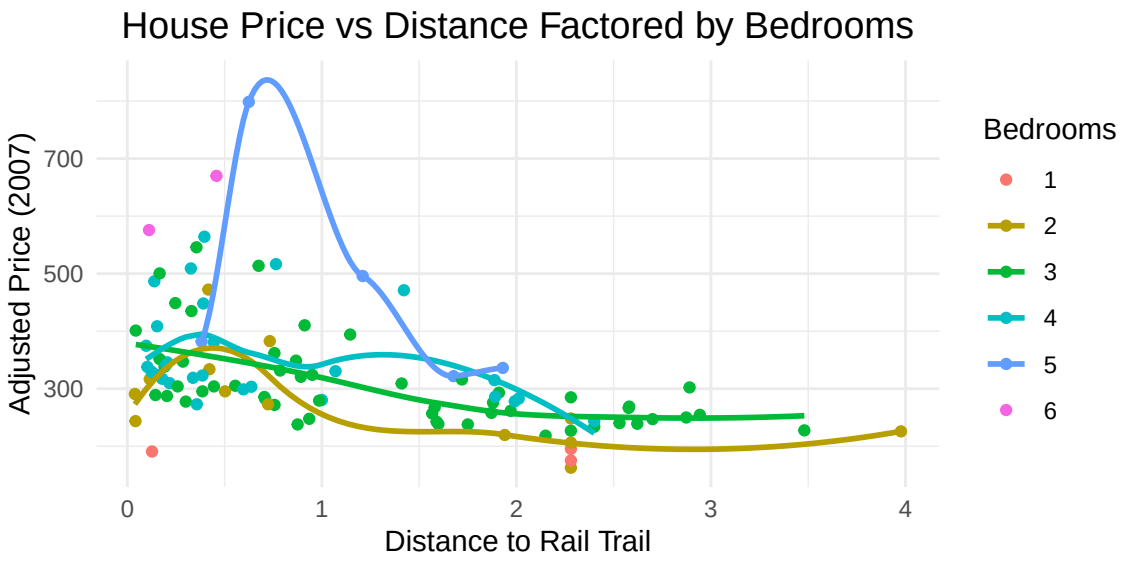

Figure 4: House Prices vs. Distance to Rail Trail by Number of Bedrooms

We can observe significant differences in price distribution for homes with 5 and 6 bedrooms compared with homes with 1 to 4 bedrooms. The curves for homes with 5 and 6 bedrooms, display much higher price variation compared to home with fewer bedrooms. This large fluctuation could indicate these homes are outliers, likely catering to a different segment of the housing market. Including these outliers would distort the results as they might not follow the same market trends as the majority of the homes.

Since the primary goal of the study is to assess whether rail trails affect house prices, focusing on the more typical segment of the housing market (1 to 4 bedrooms) will provide a clearer and more accurate analysis. The price distribution for 1 to 4 bedrooms is more stable and consistent, allowing for a better understanding of how proximity to rail trails influences pricing. In addition to this, there are only 7 homes with more than 4 bedrooms so we do not lose too much information about the population. The limitation to this is that our study will be constrained to analyzing homes with less than 5 bedrooms. We will re-examine these histograms with the outliers removed. After removing the data points with bedrooms being 5 and 6, we observed that the many distributions become less right-skewed.

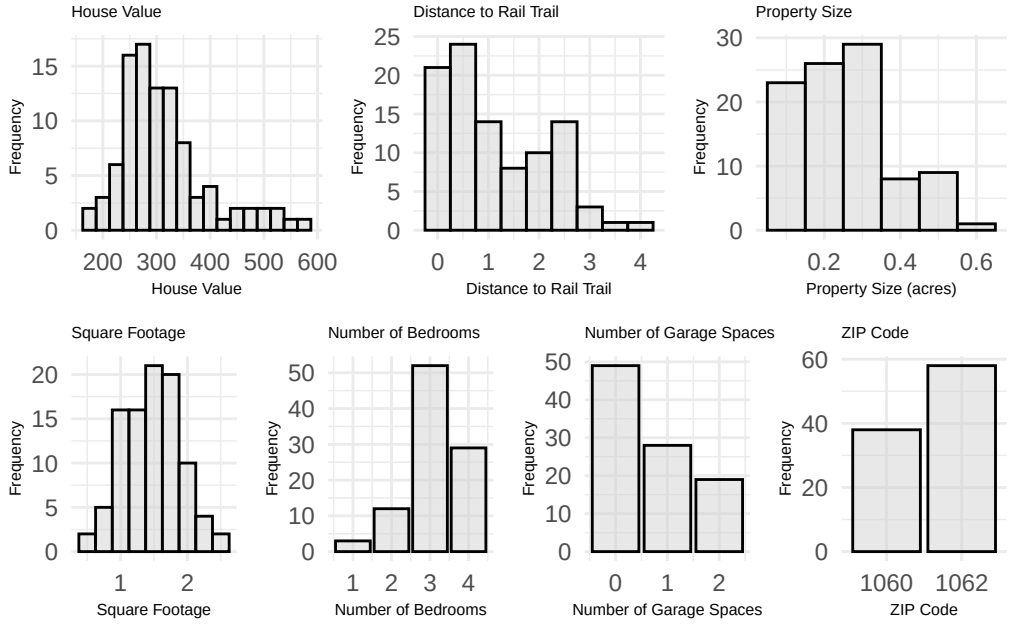

Figure 5: Histograms of Variables of Interest (Excluding Bedroom = 5&6)

While the distribution of house prices is less right-skewed, we believe it still requires a log transformation to achieve better normality and improve the accuracy of our analysis. When we applied a log transformation to house price, the resulting distribution appears normal. This transformation is crucial for our analysis, as the key assumption is normal distribution of the response variable. By ensuring that our data meets this assumption, we improve the accuracy and reliability of our analysis. Figure 5 shows us a distribution with better normality. Thus, we will use house price with log transformation for future analysis.

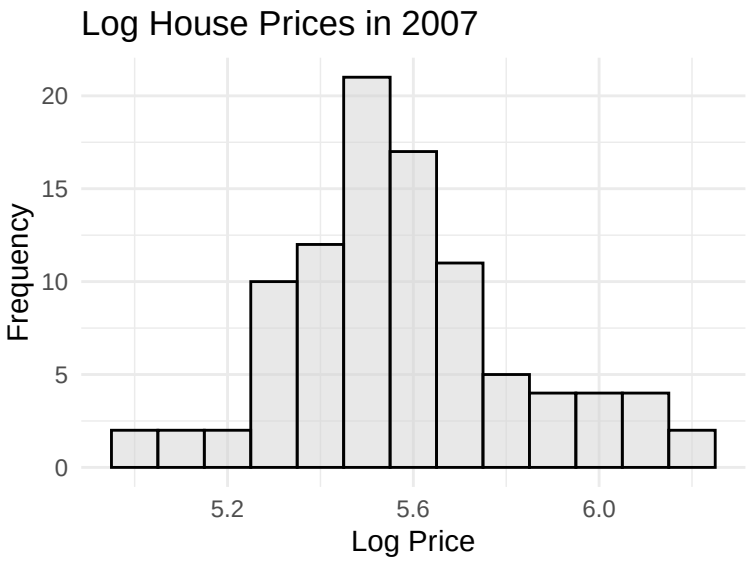

Figure 6: House Price Distribution with Log Transformation

We further our understanding of the relationship between various predictor variables and house values through looking at their respective scatter plots and box plots (Figure 5). We found a strong negative correlation between the distance to the rail trail and house values, indicating that homes closer to the rail trail tend to have higher values. Although there was no significant association between property size and house values, we included property size in our model due to its general importance in determining house value. We observed a strong positive correlation between square footage and house value, as well as a positive correlation between the number of bedrooms and house value. Similarly, the number of garage spaces positively correlated with house values, with notable variability in prices for homes with two garage spaces. Additionally, houses in zip code 1060 were generally more expensive but exhibited higher price variance compared to those in zip code 1062, possibly due to their proximity to the rail trail.

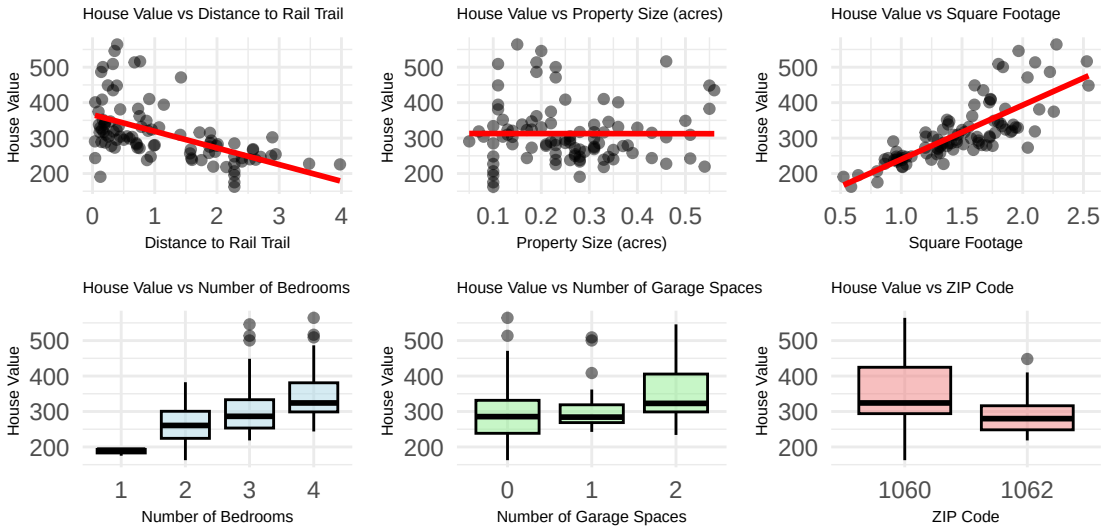

Figure 7: Relationships of Main Variables of Interest

Methods

After preparing the data, we will adopt the linear regression method with the log-transformed house price as the response variable and six covariates: distance to the rail trail, property size, square footage, number of bedrooms, number of garage spaces, and zip code. Here, variables distance, acre, square footage are treated as continuous variables, while number of garage spaces, number of bedrooms and zip code are treated as categorical variables. As we run the model, we will check the assumptions of the linear regression model, including linearity, normality, homoscedasticity, and independence of residuals, and perform diagnostic checks to validate the model.

The final model is specified as:

$$\log Price = \beta_0 + \beta_1 Distance + \beta_2 Acre + \beta_3 Sqft + \beta_4 B2 + \beta_5 B3 + \beta_6 B4 + \beta_7 G1 + \beta_8 G2 + \beta_9 Z + \epsilon$$

In our model, the notation of each variable is as follows:

- LogPrice: The logged form of price of each property
- Distance: The distance from each property to the nearest rail trail
- Acre: The size measurement of each property
- Sqft: The inner area measurement of each property
- B2: Dummy variable, which equals 1 if the property has two bedrooms, 0 otherwise
- B3: Dummy variable, which equals 1 if the property has three bedrooms, 0 otherwise
- B4: Dummy variable, which equals 1 if the property has four bedrooms, 0 otherwise
- G1: Dummy variable, which equals 1 if the property has one garage, 0 otherwise
- G2: Dummy variable, which equals 1 if the property has two garages, 0 otherwise
- Z: Dummy variable, which equals 1 if the zip code is 01062 and 0 if the zip code is 01060

Table 1, 2 and 3 below illustrate the value of each categorical given different attributes of each house.

Table 1: Dummy Variables for Number of Bedrooms

Bedrooms	B2	B3	B4
One Bedroom	0	0	0
Two Bedrooms	1	0	0
Three Bedrooms	0	1	0
Four Bedrooms	0	0	1

Table 2: Dummy Variables for Number of Garage Spaces

Garage_Spaces	G1	G2
No Garage	0	0
One Garage	1	0
Two Garages	0	1

Table 3: Dummy Variables for Zip Code

Zip_Code	Z
1060	0
1062	1

We used a partial F-test to determine whether the variable Distance significantly improves the model's ability to predict house prices, given the other predictors (Acre, SquareFeet, Bedrooms, Garage_Spaces, and Zip code). To do this we fit two models, the first is the full model which includes all the predictors—Distance, Acre, SquareFeet, Bedrooms, Garage_Spaces, and Zip. The second is the reduced model which excludes the predictor Distance.

H_0: The coefficient for Distance is equal to zero.

H_A: The coefficient for Distance is significantly different from zero.

We calculate the partial F-statistic based on the sum of squared residuals (SSR) from both models:

$$F = \frac{(SSR_{\text{reduced}} - SSR_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSR_{\text{full}}/(n - p_{\text{full}})}$$

- SSR_{reduced} is the sum of squared residuals from the reduced model.
- SSR_{full} is the sum of squared residuals from the full model.
- p_{full} and p_{reduced} are the number of parameters in the full and reduced models, respectively.
- n is the sample size.

Results

Table 4: Model Coefficients

Term	Estimate	Lower_Bound	Upper_Bound	Std_Error	T_Value	Pr_t	Significance
(Intercept)	5.0548	4.8858	5.2238	0.0862	58.632	< 2e-16	***
Distance	-0.0639	-0.0953	-0.0325	0.0160	-3.998	0.000135	***
Acre	0.1305	-0.1292	0.3902	0.1325	0.985	0.327370	
SquareFeet	0.3874	0.2945	0.4803	0.0474	8.175	2.34e-12	***
Bedrooms2	0.1564	-0.0147	0.3275	0.0873	1.791	0.076740	.
Bedrooms3	0.2124	0.0444	0.3804	0.0857	2.480	0.015096	*
Bedrooms4	0.1315	-0.0553	0.3183	0.0953	1.379	0.171419	
Garage_Spaces1	0.0143	-0.0506	0.0792	0.0331	0.431	0.667379	
Garage_Spaces2	0.0314	-0.0435	0.1063	0.0382	0.822	0.413281	
Zip1062	-0.0976	-0.1693	-0.0259	0.0366	-2.668	0.009124	**

Table 5: Model Summary Statistics

Metric	Value
Residual Standard Error	0.1258
Multiple R-Squared	0.7722
Adjusted R-Squared	0.7484
F-Statistic	32.39 on 9 and 86 DF
P-Value	< 2.2e-16

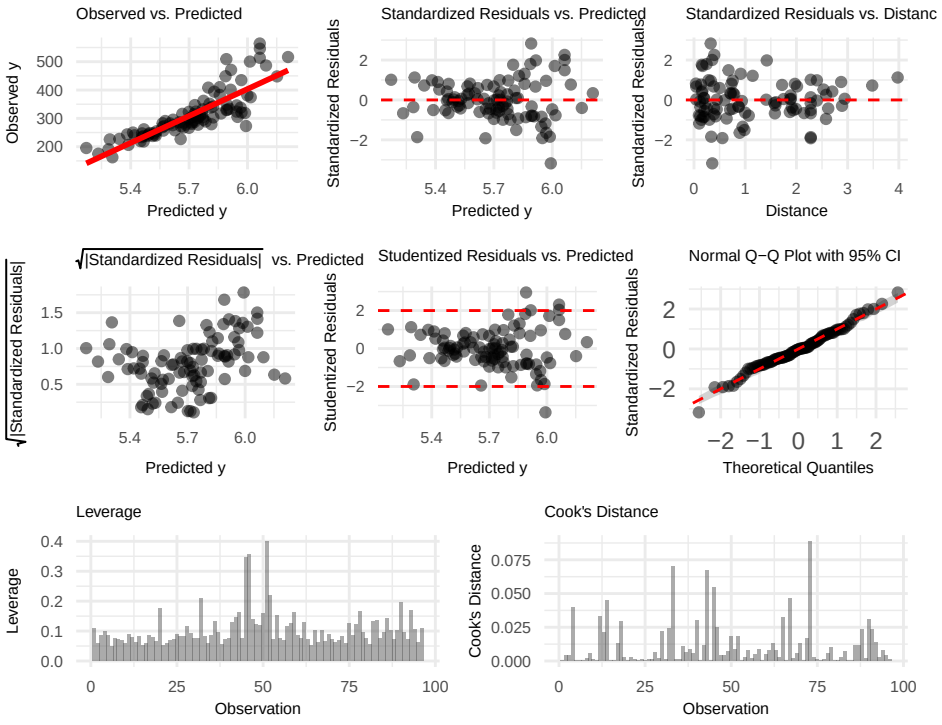

Figure 8: Diagnostic Plots of the Final Model

Discussion

The model had a multiple R-squared of 0.7722 which suggests that our model explains 77.22% of the variation in house values. Additionally, the residuals ranged from -0.38317 to 0.34117, with a median close to zero, which also indicates a good fit of the model.

The resulting p-value from the F-test shows that Distance is significant because the p-value associated with Distance is very small ($p = 0.000135$), much lower than the typical significance threshold of 0.05. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that Distance is a significant predictor of house prices. However, we did not formally check for multicollinearity, meaning there could be correlations between other variables that act as mediators, potentially affecting the accuracy of our estimates. While we attempted to address this by carefully considering and removing potential mediators in our DAG, it is possible that some remain. If so, the estimated coefficients may be unstable, making it difficult to fully isolate the effect of distance on house prices.

The diagnostic plots of the regression model suggest that the model performs well, and most assumptions are satisfied. The Observed vs Predicted plot shows the model captures the trend well, as there is mostly a linear trend between the two and no obvious curvature. The Standardized Residuals vs Predicted plot shows some slight fanning out of residuals as the predicted values increase. Mild heteroscedasticity in linear regression can lead to inefficient estimators, unreliable standard errors, and potentially invalidate our inferential tools (hypothesis tests and confidence intervals). It also compromises the model's predictive power, as it suggests the model doesn't fully capture the data patterns. We will be weary of this when using our model to interpret results. In the Q-Q plot, the residuals generally follow the trendline, supporting the assumption of normality. There are few points outside the 95% confidence intervals which is to be expected for such a large number of data points. The Scale-Location plot shows relatively even spread of residuals, although a subtle downward trend once again hints at possible heteroscedasticity. The Residuals vs Leverage plot reveals a few high leverage points, which may have a disproportionate effect on the model but the Cook's Distance Plot shows there are no significant influential points that we need to be concerned about.

Our regression analysis provides several key insights into the factors influencing house prices in Northampton, Massachusetts, and specifically addresses the impact of proximity to rail trails.

Interpretation of Results

Proximity to Rail Trail:

- The coefficient for distance to the rail trail is -0.06390, with a 95% confidence interval ranging from -0.09522 to -0.03258. This means that if we were to take 100 different samples and compute a 95% confidence interval for each sample, approximately 95 of those intervals would contain the true effect of distance on house prices. Since the entire confidence interval is negative, we can be confident that increasing the distance from the rail trail is associated with lower house prices. Specifically, the interval suggests that for each additional mile from the rail trail, house prices decrease by between 3.26% and 9.52%. This finding reinforces the hypothesis that homes closer to the rail trail are more attractive and thus command higher prices.

Square Footage:

- The coefficient for square footage is 0.38745, with a 95% confidence interval ranging from 0.29455 to 0.48035. The interval indicates a strong and statistically significant positive relationship between square footage and house prices. For every 1000 square feet of additional space, house prices increase between 29.46% and 48.04%. The fact that the confidence interval is well above zero reinforces the reliability of this positive relationship, consistent with market expectations that larger homes tend to have higher values.

Number of Bedrooms:

- The coefficient for homes with 3 bedrooms is 0.21241, with a 95% confidence interval ranging from 0.04455 to 0.38027. Since the confidence interval does not include zero, this indicates a statistically significant positive impact of having 3 bedrooms on house prices. Homes with 3 bedrooms are estimated to be valued between 4.46% and 38.03% higher than homes with fewer bedrooms, highlighting the importance of this number of bedrooms in the housing market.
- The coefficient for 2 bedrooms is 0.15637, with a confidence interval from -0.01472 to 0.32746, and the coefficient for 4 bedrooms is 0.13146, with a confidence interval from -0.05550 to 0.31843. Both intervals cross zero, meaning the effects of having 2 or 4 bedrooms are not statistically significant at the 95% confidence level. While the number of bedrooms is treated as a categorical variable, this suggests that 3-bedroom homes stand out in significance, while homes with 2 or 4 bedrooms do not significantly affect house prices compared to 1 bedroom homes.

Number of Garage Spaces:

- The coefficient for 1 garage space is 0.01426, with a 95% confidence interval ranging from -0.05123 to 0.07974. Since the interval includes zero, this indicates that the number of garage spaces does not have a statistically significant effect on house prices. The coefficient for 2 garage spaces is 0.03142, with a confidence interval from -0.04399 to 0.10684. Again, the interval crosses zero, showing that the number of garage spaces does not have a significant impact on house prices at the 95% confidence level.

Zip Code:

- The coefficient for zip code 01062 is -0.09759, with a 95% confidence interval ranging from -0.16952 to -0.02566. The interval does not include zero, confirming a statistically significant negative effect on house prices for homes in this zip code compared to those in zip code 01060. Homes in 01062 are priced between 2.57% and 16.95% lower than those in 01060, which could reflect the latter's closer proximity to the rail trail, making it a more desirable location.

The primary question of our study was whether rail trails are attractive for people buying homes, who might be willing to pay more for a house closer to a rail trail. Our analysis provides strong evidence that proximity to the rail trail is indeed a significant factor in determining house prices. Homes closer to the rail trail tend to have higher values, suggesting that rail trails are an attractive feature for home buyers. This finding is important for Acme Homes, LLC, as it indicates that developing homes near rail trails could be a profitable strategy.

However there are some limitations to our study. As an observational study, we can identify correlations between proximity to rail trails and house prices, but we cannot establish causality. Unobserved factors such as neighborhood quality or proximity to other amenities mau also influence house prices, which we were unable to control for. Additionally, our analysis focuses on homes with fewer than 5 bedrooms and up to 2 garage spaces, so the results may not generalize to larger properties. We also did not include interaction terms, which assumes the effect of distance from the rail trail is consistent across homes of different sizes and characteristics—an assumption that may not always hold.

Despite these limitations, our results highlight the importance of considering proximity to rail trails in real estate development and pricing strategies. By focusing on areas near rail trails, Acme Homes, LLC can potentially increase the value and attractiveness of their properties.