# Evaluating the Effect of Proximity to Rail Trails on House Prices

## Executive Summary

This report examines the relationship between proximity to rail trails and house prices in Northampton, Massachusetts, based on a dataset of homes sold in 2007. The primary objective was to determine whether distance from a rail trail affects home values, providing insights for stakeholders in the real estate market. We considered several factors, including how far each home is from the rail trail, the number of bedrooms, the size of the home, and the number of garage spaces. Our analysis suggests that homes located closer to the rail trail tend to have higher prices; however, the financial impact is relatively modest, with an estimated decrease of about $390 for each additional thousand feet from the trail. These findings imply that while proximity to a rail trail may be a factor in home pricing, it is likely overshadowed by other significant variables, such as square footage. We also identified limitations, notably that our data only accounts for houses in Northampton sold in 2007, highlighting the need for a more diverse dataset to strengthen our conclusions.

## Introduction

In the late 19th and early 20th centuries, rail lines were a vital part of transportation in the U.S., connecting cities for both passenger and cargo services. However, as cars became more prevalent and highways expanded, rail use declined, leading to the closure and abandonment of many lines. Starting in the 1980s, some of these unused railways were transformed into rail trails—accessible paths for walking and biking. These trails are popular for their gentle slopes and continuous routes, and there is growing interest in how proximity to a rail trail may influence the value of nearby homes. We are contracted by Acme Homes LLC to evaluate the most profitable locations for building new homes. Specifically, we seek to determine if rail trails are attractive to homebuyers and how living near a rail trail impacts house values. Our analysis evaluates the hypothesis that proximity to rail trails leads to increased house prices. While our findings indicate that homes closer to rail trails may command higher prices, the actual impact is modest, suggesting that other factors play a more significant role in determining home values.

## Exploratory Data Analysis

### Data Summary

To answer these questions, we are using data collected from homes sold in Northampton, Massachusetts in 2007. A new rail trail was opened in Northampton in 1984, which offers an opportunity to compare the home values based on proximity to the trail. Our data features 104 homes in Northampton, Massachusetts spanning across two zipcodes. Our data set did not contain any missing values. Below is a table delineating the relevant variables we explored in our data, their abbreviation, and their meaning.

| Variable | Description |
|---|---|
| price2014 | Zillow's estimated value for the home in 2014, in thousands of dollars |
| distance | Distance (1000 feet) to the nearest entry to the rail trail network |
| acre | Number of acres of property |
| bedrooms | How many bedrooms the home has |
| bikescore | Bike friendliness of the area, estimated by WalkScore.com. 0-100 scale |
| walkscore | Walkability of the area, estimated by WalkScore.com. 0-100 scale |
| garage_spaces | Number of garage parking spaces (0-4) |
| latitude | House's latitude |
| longitude | House's longitude |
| squarefeet | Square footage of the home's interior finished space (in thousands of square feet) |
| streetname | Name of the street the house is on |
| zip | ZIP code of the house (leading 0 omitted). 1060 is Northampton, MA; 1062 is Florence, MA. |

#### Response Variable

For this analysis, we selected `Price2014`, the estimated value of these houses in 2014, as our response variable. Our developer manager, Mr. Cayote, aims to understand how this information influences his current choices for development locations. Although we could have chosen a variable reflecting the sale prices from 2007, we felt that data would be too outdated. It's also important to note that `Price2014` represents only an estimate of the houses' values in 2014.

## Exploratory Data Analysis

### Univariate Distributions

We begin with an exploratory data analysis of the relevant variables. Figure 1 displays the distribution of our response variable, the 2014 house prices. The distribution is unimodal, centered around $\$200,000$ to $\$250,000$, with a spread of approximately $\$400,000$. There is one outlier at around $\$800,000$, which will be excluded in subsequent analyses, as explained in the `Methods` section. Additionally, the distribution of our response variable is roughly symmetric, though the right tail (houses around $\$500,000$) shows a slightly higher frequency than the left tail (houses around $\$100,000$).

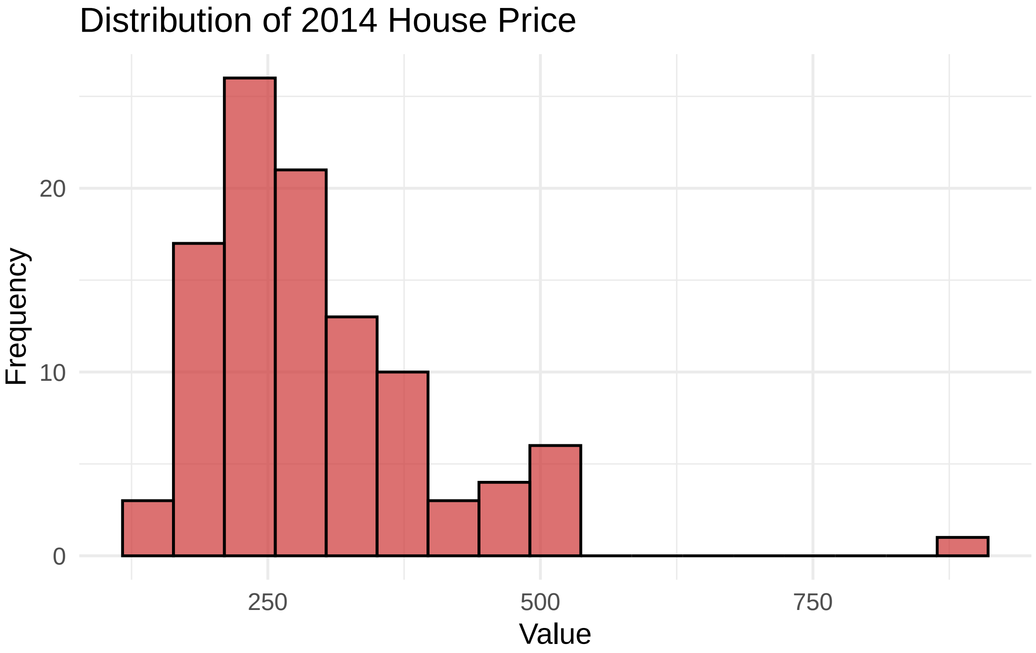

Figure 1: Unimodal Distribution of Response Variable (2014 Price)

Next, we examine the distributions of our potential continuous predictor variables: distance from the rail trail, lot acres, and square footage. Our primary variable of interest, distance, displays a bimodal distribution, with one group of houses located within 1,000 feet of the Northampton rail trail and another around 2,000 feet away. The histogram shows a higher concentration of houses within 1,000 feet of the rail trail. The lot acre distribution is approximately uniform between 0.1 and 0.4 acres, with fewer houses outside this range, giving the acres variable a spread of around 0.6. Lastly, the distribution of square footage is right-skewed, centered between 1,500 and 2,000 square feet, with a range extending up to approximately 4,000 square feet.

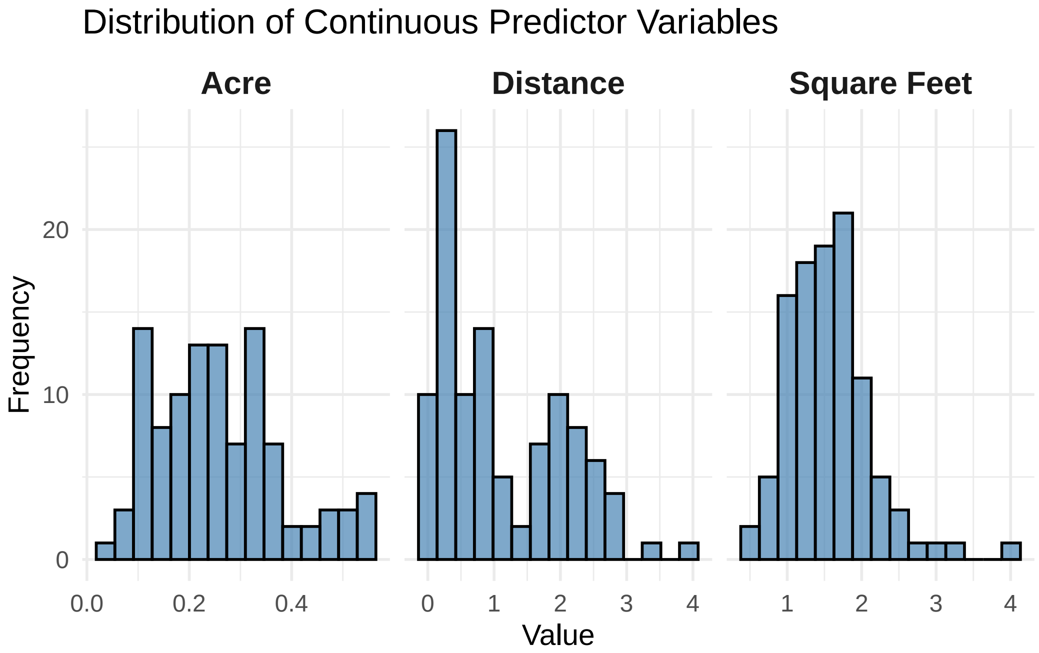

Figure 2: Distribution of Continuous Predictor Variables, Highlighting Central Tendency and Variation

Finally, we examine the univariate counts of our potential discrete predictor variables: number of bedrooms, number of garage spaces, and zipcode. We can see the number of bedrooms ranges between 1 and 6 with a center on 3. Due to there only being a very small amount of houses with one or six bedrooms, we remove these from our data all together and only focus on houses with 2-5 bedrooms. This decision is expanded on in the `Methods` section. The number of garage spaces appears right skewed with a mode around 0. There appears to be a single leverage point of 4 garage spaces. Lastly, when we examine the counts of the zipcode variable we see that one zipcode, 1062, appears to have more representation in this data set than the other zipcode, 1061.

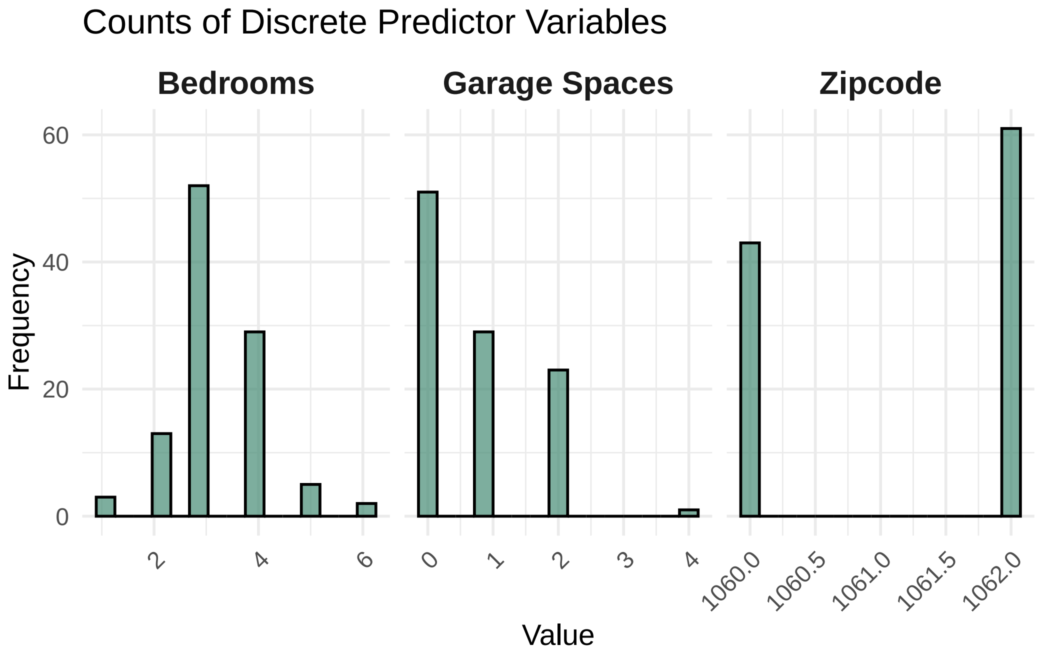

Figure 3: Distribution and Frequency of Discrete Predictor Variables

Please note that we did not examine the univariate distributions of latitude and longitude. Since those variables together describe a very specific location, it does not make sense to analyze their distributions separately. We instead, plot them together and examine their relationship with rail trail distance in the next section on multivariate distributions.

### Multivariate Distributions

Next, we evaluate how combinations of variables interact. Figure 4 presents a bivariate scatterplot of house price versus distance to the rail trail. This scatterplot suggests a roughly linear negative relationship with non-constant variance, indicating that houses farthest from the rail trail are generally priced the lowest. However, houses closer to the rail trail exhibit considerably more variation in price. This scatterplot suggests, as intuition would support, that distance to the rail trail alone does not fully explain the variation in house prices.

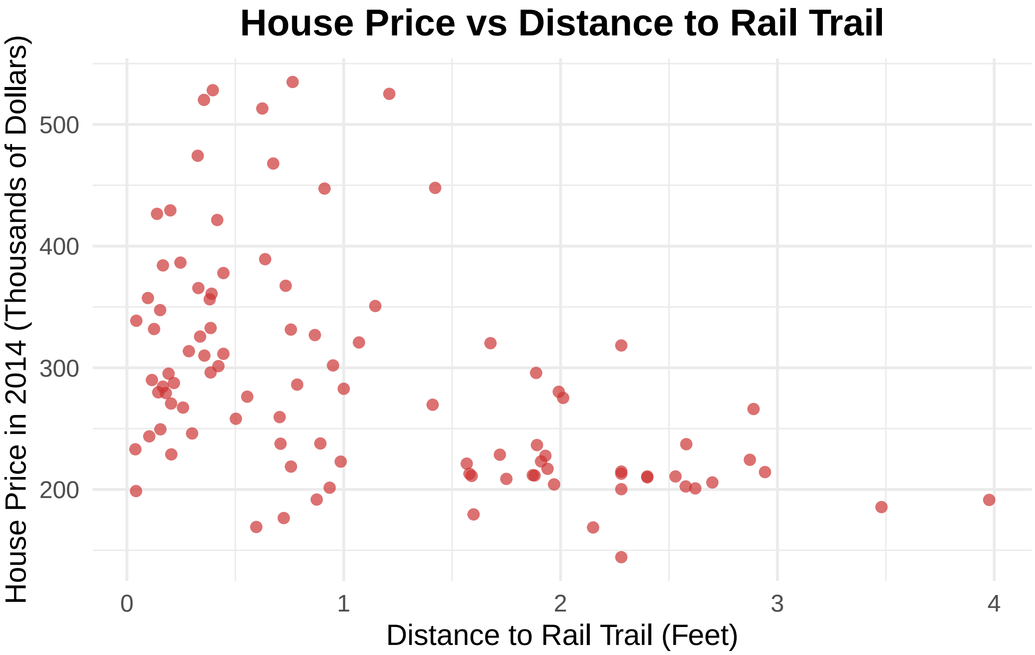

Figure 4: Weak Negative Correlation Observed Between House Price and Distance to Rail Trail

Now, we examine the relationship between house price, distance to the rail trail, and other predictor variables. First, we look at scatterplots of house price versus continuous predictor variables, with colors indicating proximity to the rail trail. For this, we created a binary variable to represent proximity by "binning" the original distance variable, setting a threshold of 1.4, which separates the two modes in the data. This binned variable is used only for exploratory analysis and is not included in the final model.

In the scatterplot of square footage, we see a strong positive linear relationship between house price and square footage. Notably, houses closer to the rail trail tend to have more square footage, which we will revisit in the discussion. A potential leverage point appears at 4,000 square feet. In the scatterplot of lot acres, no clear relationship with house price or distance to the rail trail emerges, though houses closer to the trail tend to be more expensive.

Finally, we examine a plot of longitude vs. latitude, colored by distance to the rail trail, mapping the area and the trail's location. A high density of houses is visible around a rail trail that runs diagonally across the map, with sporadic clusters of houses farther away, most of which are southwest of the trail.

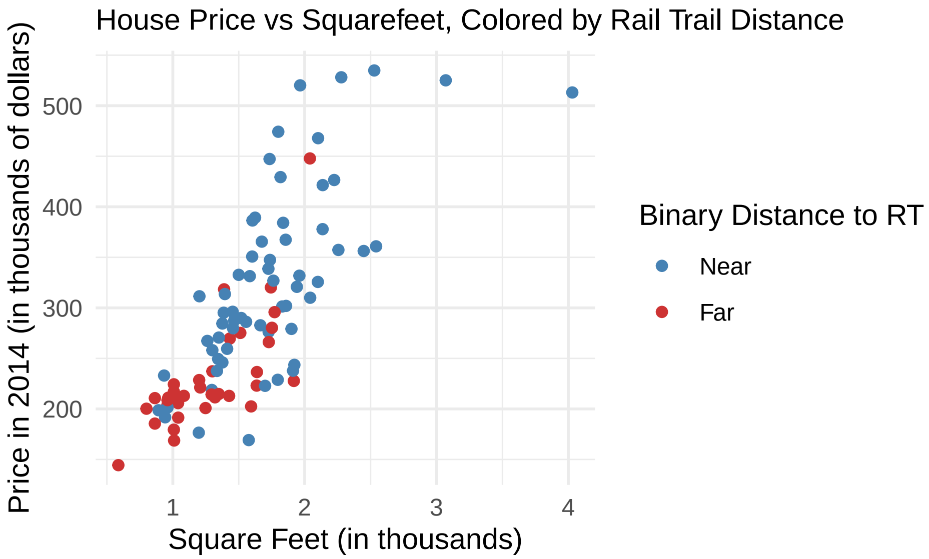

Figure 5: Strong Positive Correlation Observed Between House Price and Square Feet, Color-Coded by Rail Trail Distance

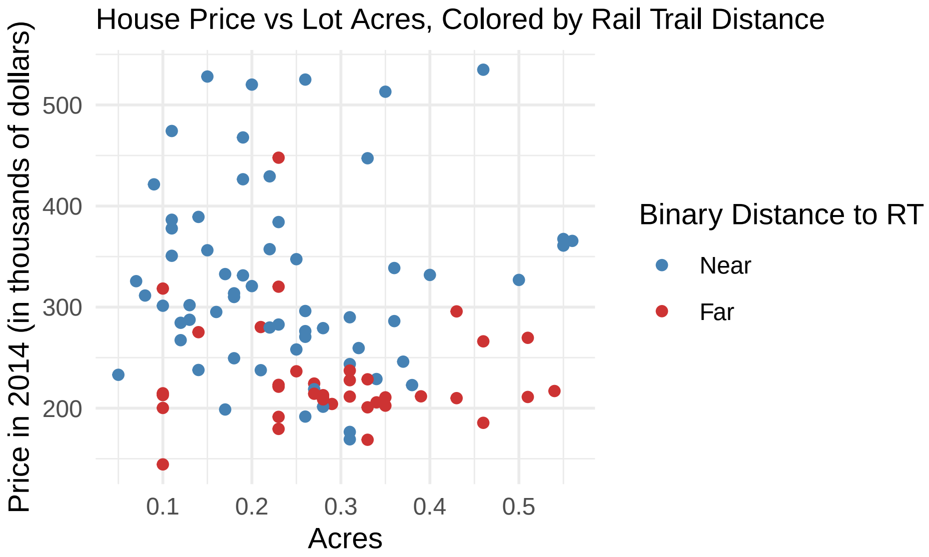

Figure 6: Weak Correlation Observed Between Price and Lot Acress, Color-Coded by Rail Trail Distance

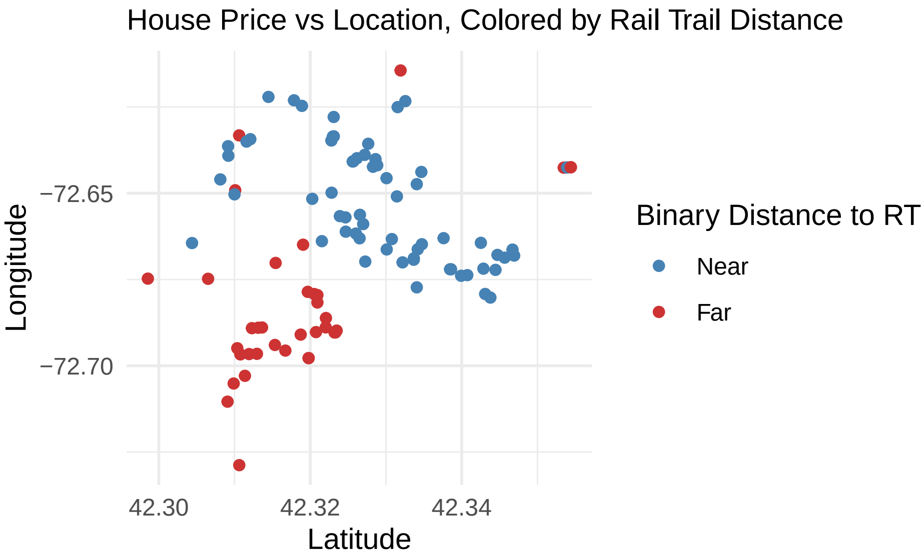

Figure 7: Map of Houses in the Data, Color-Coded by Rail Trail Distance (Highlighting Distribution Patterns and Spatial Relationships)

Finally, we explore the relationship between house price, distance to the rail trail, and our discrete predictors using boxplots. In the boxplot of house price by number of bedrooms, we observe that houses nearer to the rail trail are, on average, more expensive across all bedroom counts. As expected, house price generally increases with the number of bedrooms. For houses farther from the rail trail, the increase in price appears linear with bedroom count, while for houses nearer to the trail, prices seem to rise exponentially with more bedrooms. However, this apparent "exponential" trend may be unreliable due to only five observations with five bedrooms.

Similarly, in the boxplot of house price by garage spaces, prices increase with both proximity to the rail trail and garage space count. Here, we've binned garage spaces into two categories: fewer than 2 versus 2 or more garage spaces. Finally, examining the zipcode boxplot, we see that both zipcodes include houses near and far from the rail trail, though zipcode 1062 has a smaller proportion of houses farther away. In both zipcodes, houses near the rail trail tend to have higher average prices, though there remains a wide range of prices for houses near the trail in both areas. Overall, these findings suggest a consistent trend of higher house prices for properties closer to the rail trail across both zipcodes.

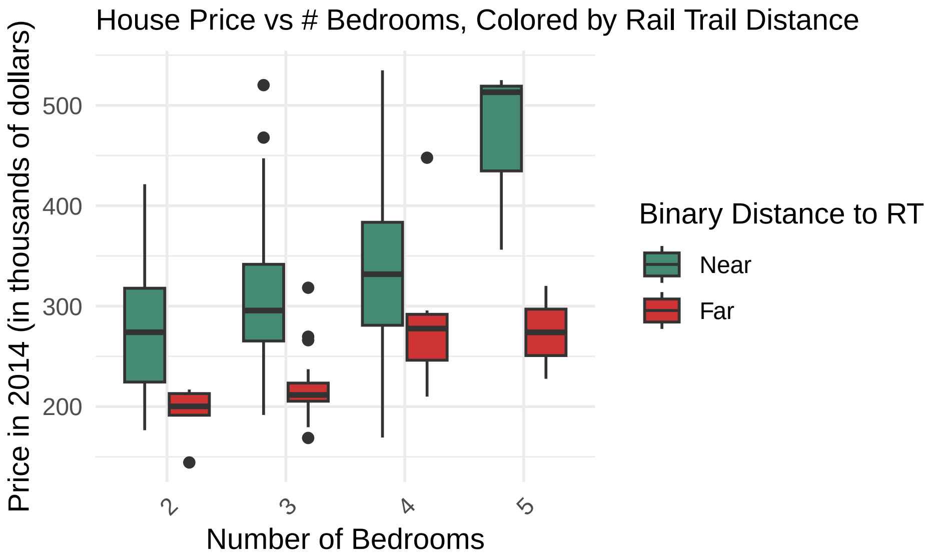

Figure 8: Boxplots Displaying Price Distributions Across Bedroom Counts, Color-Coded by Proximity to Rail Trail

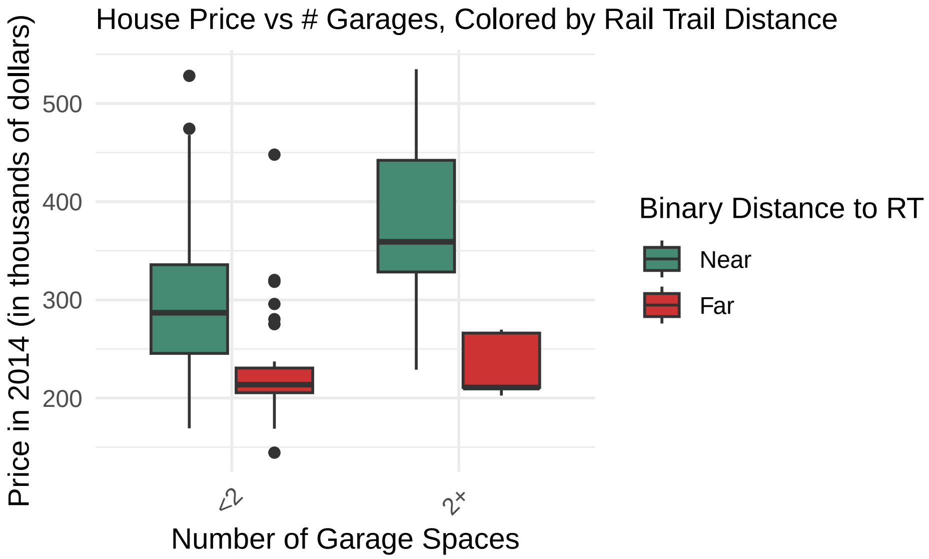

Figure 9: Boxplots Displaying Price Distributions by Garage Count, Color-Coded by Proximity to Rail Trail

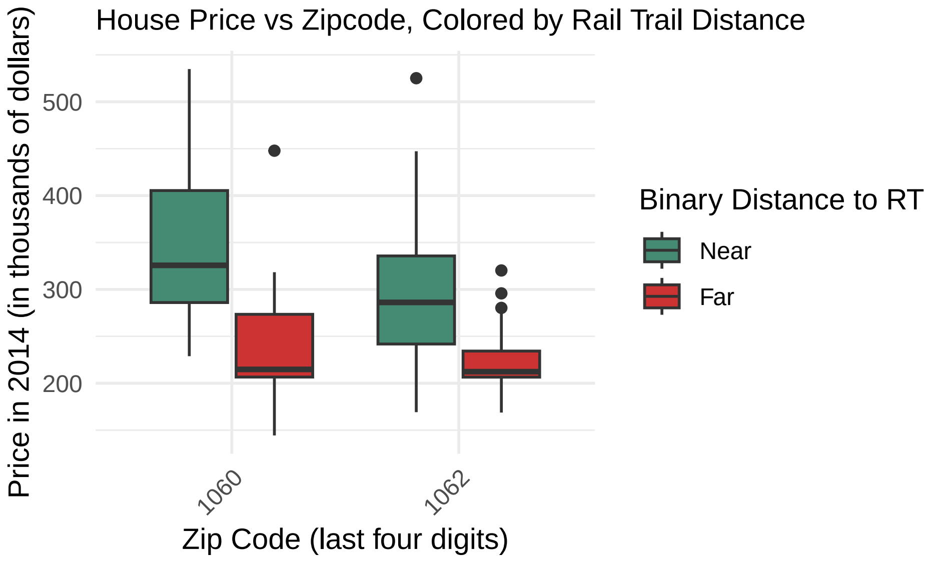

Figure 10: Boxplots Displaying Price Distributions by Zip Code, Color-Coded by Proximity to Rail Trail

## Data Transformations

The first data transformation we performed was simplifying the garage spaces variable from a four-level factor to a two-level factor. We grouped houses with fewer than two garage spaces into one category and houses with two or more garage spaces into another. Additionally, we removed five data points corresponding to houses with one or six bedrooms, as these were few and would have added unnecessary complexity to our model. We also removed a significant outlier in house price, valued at over $\$800,000$—substantially higher than the next highest point. Lastly, we identified, but did not remove (at this stage), one leverage point for square footage (house number 85) around 4,000 square feet.

## Limitations

It's important to recognize some limitations in our data. First, this dataset was last updated in 2014, making it outdated by at least a decade. Since then, various socioeconomic factors have influenced the housing market, which may not be reflected here. Additionally, this data focuses exclusively on one rail trail in a specific region of Massachusetts, meaning our analysis applies only to this area and is not generalizable to the broader U.S. housing market. Finally, several relevant factors—such as school district quality, proximity to grocery stores, and crime rates—are absent from this data. While some of these effects may be indirectly captured by location variables like latitude, longitude, and zipcode, direct measures of these influences would have been preferable.

# Methods

## Variable Selection

We now turn to our approach for selecting variables for the model. Since our aim is to estimate the effect of rail trail proximity on house price, we are addressing a causation problem. To isolate the causal impact of rail trail proximity, we must account for as many potential confounding variables as possible. Initially, we decided to exclude the walkscore and bikescore variables, as both are highly correlated with distance to the rail trail, suggesting that proximity to the rail trail strongly influences these scores.

Our dataset includes three location encodings: latitude/longitude coordinates, street name, and zipcode. While latitude and longitude provide the most granular geographic detail, they lack a straightforward linear relationship with other variables and must be considered together, complicating the modeling procedure. The street name variable is categorical with over 70 categories, each represented by only 1-5 data points. Given our dataset of roughly 100 observations, including this variable would result in a substantial loss of degrees of freedom, potentially leading to an uninformative model. Thus, we chose to represent location using zipcode.

To address confounders, we include acres, square feet, number of bedrooms, and number of garage spaces in the model. We treat number of bedrooms and number of garage spaces as categorical variables, with number of bedrooms divided into three factors and number of garage spaces into two.

## Modeling

### Model Choice

Based on these decisions, the final model we use is:

$$\text{Price2014} \sim \beta_0 + \beta_1 \cdot \text{distance} + \beta_2 \cdot \text{squarefeet} + \beta_3 \cdot \text{acre} + \beta_4 \cdot \mathbb{1}_{\text{bedrooms=3}} + \beta_5 \cdot \mathbb{1}_{\text{bedrooms=4}} + \beta_6 \cdot \mathbb{1}_{\text{bedrooms=5}} + \beta_7 \cdot \mathbb{1}_{\text{garage binary=1}} + \beta_8 \cdot \mathbb{1}_{\text{zip=1062}} + \varepsilon$$

**Removing Potential Outliers**

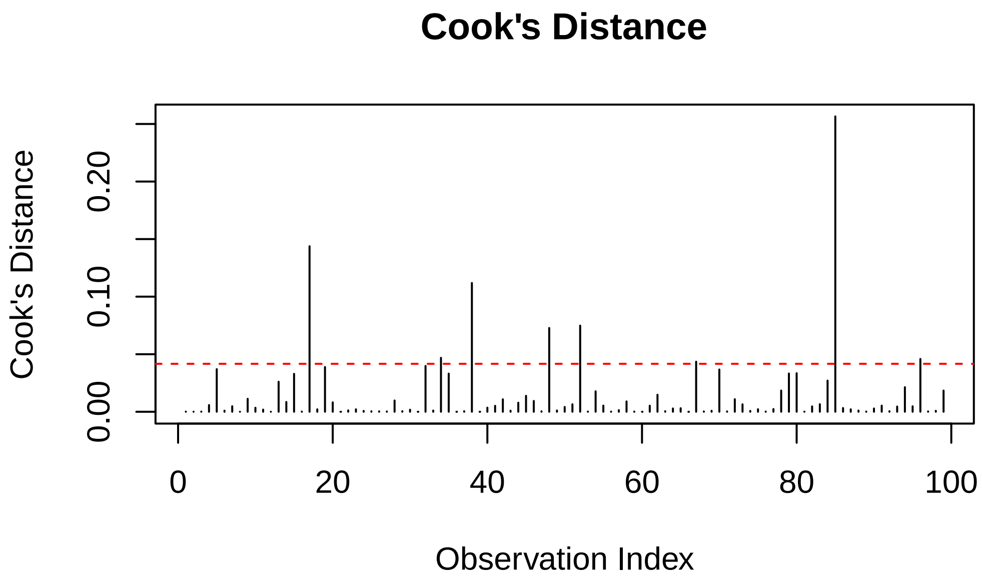

Figure 11: Initial Model Cook's Distance Plot Showing High Leverage Observations at Indices 85, 38, and 17

We ran the initial model, and the diagnostics indicated three influential points (observation numbers 85, 38, and 17). Observation number 85 corresponds to the previously mentioned leverage point of 4,000 square feet. After identifying these influential observations, we removed them from the dataset and re-ran the model, subsequently performing diagnostics on the updated results.

## Diagnostic Plots

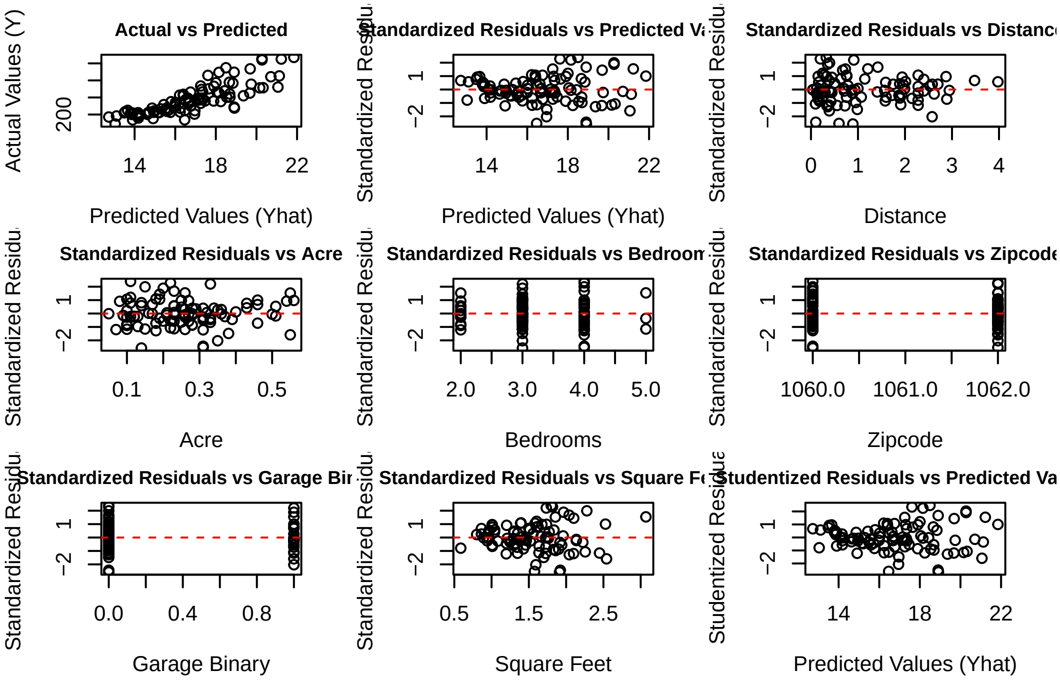

Figure 12: Diagnostic Plots for the Final Model

```
5  32
5  31
```

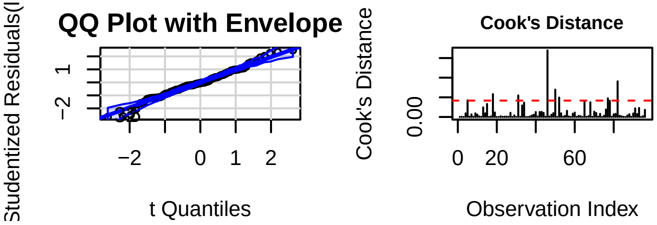

Figure 13: Diagnostic Plots for the Final Model

Overall, this model demonstrates a solid adherence to the assumptions of linear regression. The plots of Standardized Residuals against Predicted Values, as well as those comparing Standardized Residuals with the predictors, indicate a linear relationship between the predictors and house prices in 2014. The data in these plots show a random scatter around the y=0 line, reinforcing this linearity. Furthermore, the normality assumption is upheld as evidenced by the QQ-plot. At the 95% confidence level, we would expect approximately 3-5 data points to fall outside the point-wise envelope; however, only one point exceeds this boundary. This observation provides us with confidence that the normality assumption is satisfied within the model. Lastly, while the Standardized Residuals vs. Predicted Values plot reveals some heteroskedasticity, this outcome was anticipated given the context of the problem and the exploratory data analysis conducted on the distance variable. In conclusion, the model largely fulfills the assumptions of linear regression, with the moderate exception of heteroskedasticity.

Despite the model's strong adherence to linear regression assumptions, there are inherent limitations related to this model. First, the linearity assumption, while generally met, may not fully capture the complexities of the relationship between predictors and house prices, especially if non-linear interactions exist. Furthermore, the assumption of independence among observations may be violated if there are unaccounted spatial correlations among houses. Additionally, the presence of heteroskedasticity, while acknowledged, suggests that the variance of the residuals is not constant across all levels of the predictors, potentially affecting the validity of hypothesis tests and confidence intervals we calculate in the next section.

### Coefficient Significance and Effect Size

To estimate the effect size of distance on our target variable, we will look at the coefficient of the distance variable in our regression and construct a 95% confident interval using its standard error. We can do this because the coefficient of the distance variable in this model is equivalent to the effect size after the confounding variables have been taken into account. We will also establish the significance of this coefficient using a t-test with the following hypotheses:

Null Hypothesis: There is no relationship between proximity to a rail trail and house value.
$H_0 : \beta_1 = 0$

Alternate Hypothesis: There is relationship between proximity to a rail trail and house value.
$H_A : \beta_1 \neq 0$

## Results

Our final model achieves a residual standard error of 1.409 on 87 degrees of freedom. The coefficients, standard errors, and associated p-values are listed in the table below. The hypothesis t-test for significance revealed that **Distance** from the rail trail was a significant predictor of house prices at a 95% level. We were able to reject the null hypothesis with a p-value of 0.03. We found that distance to rail trail has a small but prevalent effect size of −0.3886 (95% CI: [−0.7487,−0.0285]), indicating that for each additional increase of one thousand feet in distance to the rail trail, the price of the home is expected to decrease by approximately $390 dollars. This suggests that as the distance from the rail trail increases, house prices tend to decline very slightly.

| Predictor Variable | Effect Size | Standard Error | Lower Bound | Upper Bound |
|---|---|---|---|---|
| **Distance** | −0.3886 | 0.18371 | −0.7487 | −0.0285 |
| Three Bedrooms | 0.2788 | 0.44855 | −0.6003 | 1.1580 |
| Four Bedrooms | 0.0488 | 0.54491 | −1.0192 | 1.1169 |
| Five Bedrooms | −1.7136 | 1.08843 | −3.8470 | 0.4197 |
| Garage Space | 0.9887 | 0.39035 | 0.2236 | 1.7538 |
| Zip | −0.9167 | 0.39283 | −1.6867 | −0.1468 |
| Acre | −0.7584 | 1.49977 | −3.6979 | 2.1812 |
| Squarefeet | 3.8378 | 0.51641 | 2.8256 | 4.8499 |

## Discussion

This analysis highlights a meaningful relationship between house prices and predictors, particularly the impact of proximity to rail trails, where greater distance correlates with lower home prices. However, several limitations must be acknowledged. The dataset is confined to homes sold in Northampton, Massachusetts, in 2007, which may not reflect broader market dynamics or other significant factors, such as neighborhood characteristics or local economic conditions. Additionally, since all houses closer to the rail trail have higher square footage, the reliability of these results is questionable, emphasizing the need for a more diverse dataset.

Practically, a $390 price change is relatively insignificant within the broader context of home prices, which typically range from hundreds of thousands to millions of dollars. For stakeholders like homebuyers, real estate agents, and urban planners, this suggests that while proximity to the rail trail has some influence on pricing, it is likely overshadowed by more substantial factors, such as square footage. Homebuyers may view the minimal price reduction as insufficient to offset the costs of less desirable locations. Real estate agents might focus on other selling points, like home quality and neighborhood amenities, rather than just rail trail proximity. For Acme Homes LLC, these insights could inform future development decisions regarding rail trails, as the anticipated financial return on investment may not be significant, highlighting the need for a comprehensive evaluation of these features in community planning. Future research should target larger, more diverse datasets to better understand house prices.