# Project Part 4

## Introduction

Over the last few decades, abandoned rail lines were converted into rail trails in Northampton, Massachusetts. These trails gained popularity because they provided access for walking and biking without the presence of motor vehicles. Acme Homes, LLC is interested in seeing whether they may be an attractive feature people consider when buying a home, and whether people would be willing to pay more for a home nearby a rail trail. The goal of this analysis is to see if the distance of a home to a rail trail has a measurable effect on the value of the home. In theory, people value neighborhood amenities such as parks or recreational facilities, which can lead to higher home values. As rail trails provide scenic routes to walk or bike along, a close proximity to them may similarly increase home values.

A rail trail opened near Northampton in 1984, which allows a comparison of homes near the trail and those farther away. To investigate this question, I used data of home sales in this town in 2007, including characteristics about the home and neighborhood as well as Zillow estimates for them in several other years. Using this dataset, I analyze the relationship between home prices and the home's distance to the rail trail. I find no significant relationship between proximity to rail trail access and home value.

## Data

The dataset I will be conducting my analysis on has 104 observations and 20 variables. There are no missing values in the data. Since no variables have missing data, no imputation is necessary. The following variables are the ones I am choosing to include in my analysis, and the covariates were chosen if I believe they have an effect on my response variable.

**Table 1: Summary of Variables**

| Type | Variable | Description | Continuous/Discrete |
|---|---|---|---|
| Response | adj2007 | Zillow's estimated value for the home in 2007 in thousands of 2014 dollars | Continuous |
| Covariate of main interest | distance | Distance to the nearest entry to the rail trail network in miles | Continuous |
| Covariate | acre | Property size in acres | Continuous |
| Covariate | bedrooms | Number of bedrooms in the house (1-6) | Discrete |
| Covariate | garage_spaces | Number of garage parking spaces (0-4) | Discrete |
| Covariate | squarefeet | Square footage of the home's interior finished space in thousands of sq ft | Continuous |
| Covariate | walkscore | Walk-ability of the area (0-100 scale, ease of performing daily tasks without a car) | Discrete |
| Covariate | zip | The zip code where the house is located (takes the value 1062 for zip code 01062 and 1060 for the zip code 01060) | Categorical Nominal |

I examined the distributions of the discrete variables using bar plots: bedrooms, garage spaces, walk score, and zip. In Figure A1 of the appendix, I noticed that there is only 1 observation with 4 garage spaces, and no observations with 3

garage spaces. Most homes have 1 or 2 garage spaces, 2 or 3 bedrooms, and have a walk-ability score from around 10 to 20 and 60 to 70.

Looking at Figure A2 of the appendix, it is clear that the adjusted 2007 price, the home prices in 2007 adjusted to 2014 dollars, is right-skewed with some homes that are very expensive. This is expected because less expensive homes naturally tend to have less variability in their features, amenities, and prices, while more expensive homes can vary more in those categories, as well as buyer competition. This can lead to greater price variability and some homes being very high-priced. The variables distance, and squarefeet are slightly right skewed as well. It is clear that the log transformation helped for the adjusted 2007 price, shown in Figure A3, so I will now use the log-transformed adjusted 2007 price as my response variable.

The scatterplots in Figure 1 and boxplots in Figure 2 show the relationship between the adjusted 2007 prices and the each covariate. Square footage has the strongest positive and linear relationship with the adjusted 2007 price, while the distance between rail trails shows a negative and nonlinear association. This shows helpful information about our covariate of main interest: homes closer to the trail tend to be more valuable.

The number of bedrooms and number of garage spaces also show positive trends with the adjusted 2007 price, but they are weaker. As the walk score rises, the log-transformed 2007 prices increase, but then flatten out; this may have a nonlinear relationship. Similarly, the number of acres in the property seems close to linear, but increases in acreage do not seem to be associated with increases in the log of 2007 home price. The adjusted 2007 prices appear to be more variable for zip code 1060 than for 1062, as indicated by the larger spread in the plot. However, because zip code is a categorical variable with few observations per level, the scatterplot does not provide much information about the relationship between price and location. These plots show that square footage and distance between rail trails are the most influential predictors of the adjusted 2007 prices. These associations may change once we take other variables into account.

**Figure 1: Distance has a Negative Relationship with Log-Transformed Home Price, Bikescore Has Strong Collinearity with Distance and Walkscore**

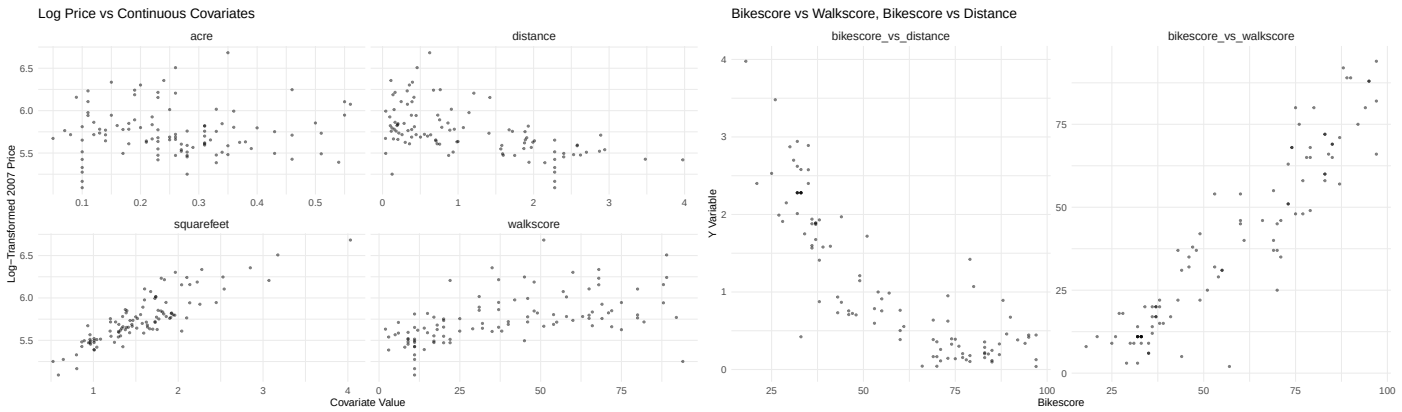
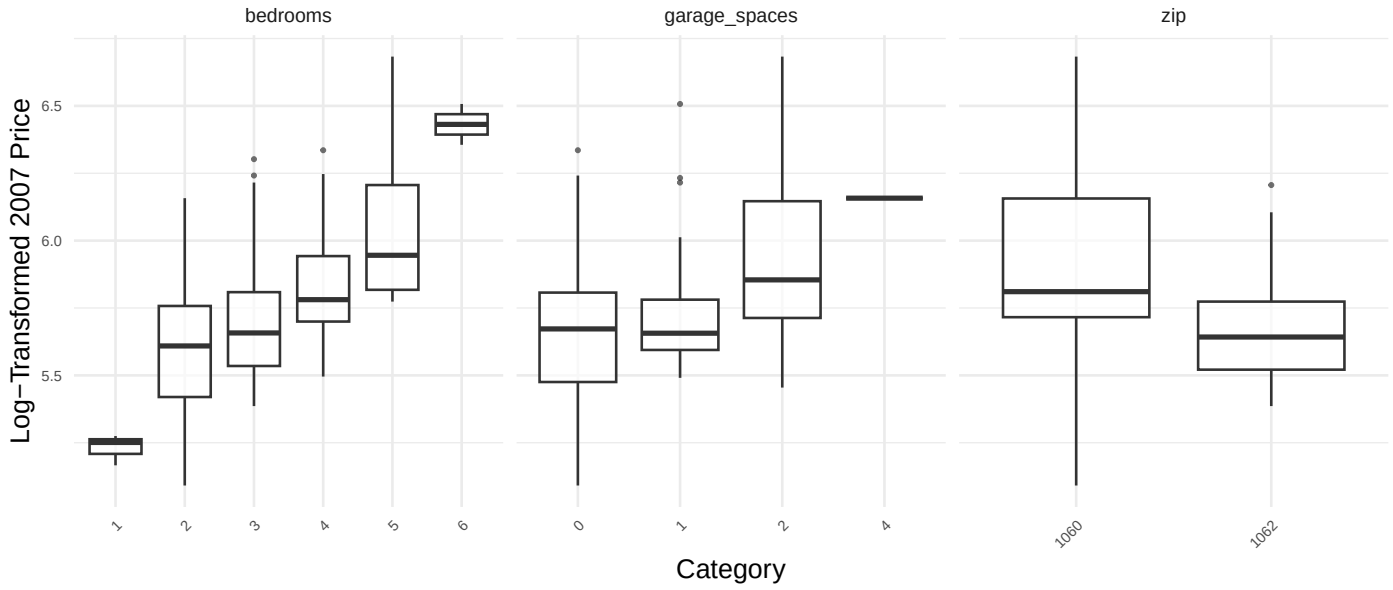
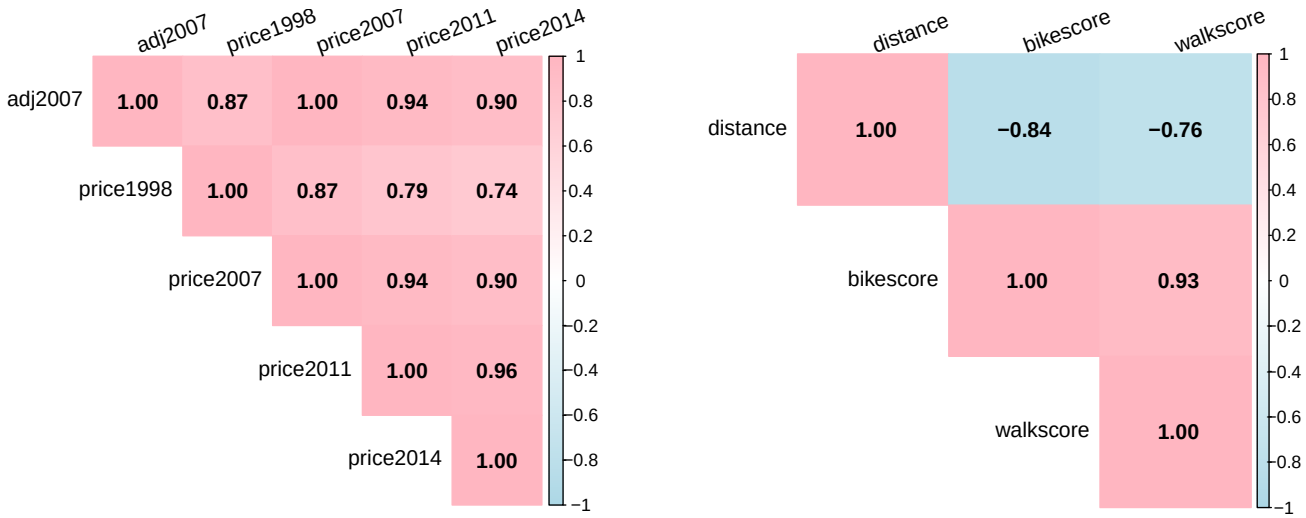


**Figure 2: Number of Bedrooms and Garage Spaces Show Positive Trends with Log-Transformed Home Price, with Zip Code 1060 Showing Higher Price Variability**




There are several points that stand out to me. There is one 5-bedroom house that is worth about $800,000, and its property has around 0.35 acres. This is a larger house so it seems as though its price is justified, but it may be skewing our analysis. I will not exclude any points because I am not assuming constant variance in this model, and I don't have reason to believe that these points are unusual. Additionally, some covariates have collinearity with each other. As can be seen in Figures 2 and 3, bike score is very collinear with distance and the walk score. The prices from different years are highly collinear as well.

**Figure 3: Correlation Matrices That Justify Variable Exclusion due to High Multicollinearity**




There are several limitations that affect our ability to make causal claims about the relationship between distance to rail trail access and home value. The data are all observational, as we cannot observe the price of a home in two different locations holding all other factors constant. The dataset only includes homes sold in 2007 with their Zillow estimates and a small set of covariates, which will limit generalizability to other areas outside of Northampton, MA. There is the possibility of omitted variable bias because important factors, such as school quality or neighborhood crime rates, aren't included, and they would have strong effects on home values. Additionally, because home values are measured in 2007, when the data was collected, but then adjusted to 2014 dollars, they wouldn't reflect current housing preferences or changes in demand.

# Methods

In this analysis, I fit a linear model regressing log-transformed adjusted 2007 price on distance, controlling for acre, bedrooms, garage spaces, square feet, walk score, and zip. I will treat bedrooms and garage spaces as discrete ordinal variables, and I will treat zip as a categorical nominal variable; the rest of the variables act as continuous variables. Least Squares Estimation (LSE) will be used to estimate the coefficients in this model.

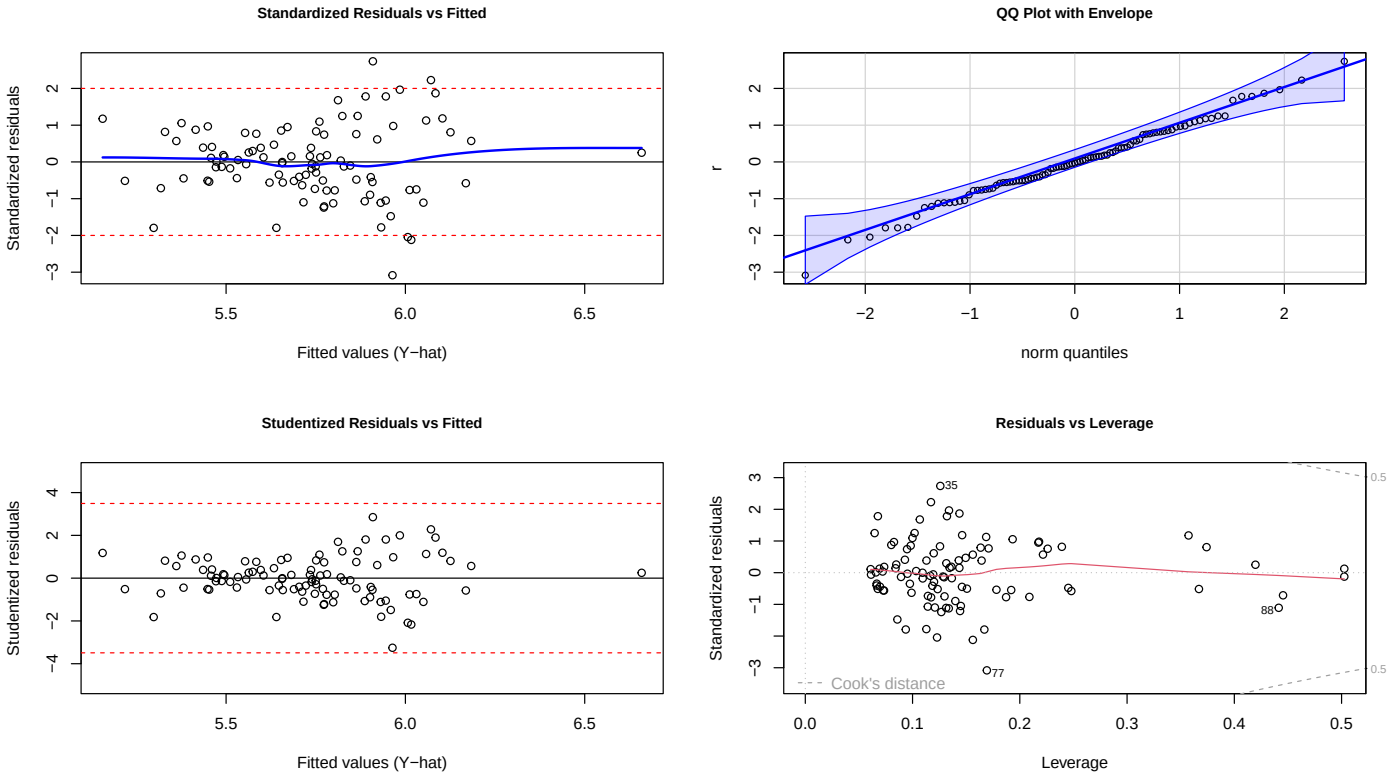
The covariates distance, acre, squarefeet, and walk score are treated as continuous to capture the linear relationships between them and the adjusted 2007 price. The covariates bedrooms, garage spaces, and zip will be treated as factors. As the number of bedrooms increases, the change in price would change in different increments; by including this covariate as a factor, we gain a lot more information about the individual increases in price for each bedroom, rather than just see it as an average linear increase. The same goes for garage spaces. The covariate zip is a nominal factor without any quantitative meaning, as it just distinguishes between the two different neighborhoods.

In addition to the main effects, I chose to examine the interaction between bedrooms and garage spaces. Including this interaction in the model allows us to capture cases where the price increase from adding a bedroom may differ depending on how many garage spaces are in the house. Intuitively, this makes sense as smaller homes won't value more garage spaces, while larger homes would. Figure A4 from the appendix suggests that the effect of adding a bedroom varies across different numbers of garage spaces. Therefore, I believe it is important to observe and I will include this interaction term in my model. Additionally, I believe that distance may have some important interactions; however to maintain interpretability of the main effect of distance on adjusted 2007 prices, I will exclude these possible interaction terms.

The dataset contained several other variables that I chose not to include in my analysis. I excluded the other Zillow value estimates for 1998, 2011, and 2014 (both raw and adjusted for inflation), because they are highly correlated with adjusted 2007 prices; earliers home values affect future home values. Both adjusted 2007 prices and raw 2007 prices measure the exact same home value, but adjusted 2007 prices is multiplied by some constant to adjust for inflation into 2014 dollars, so I excluded the raw 2007 prices. In Figure 3, we see that all of the other Zillow value estimates are extremely correlated with adjusted 2007 prices ($r > 0.87$), so keeping the variables in my model would give no additional information and could add multicollinearity. The bike score is weighted on bike infrastructure, topography, road connectivity to destinations, and the percentage of people who bike to work. The variable walk score is weighted on population density, block length, intersection density, and proximity of amenities. The walk score and bike score may have overlap, where streets that are bike-able might tend to also be walk-able. The correlation between bike score and walk score seen in Figures 2 and 3 agrees with my statement that they may be interconnected. I chose to exclude bike score because of its high correlation with walk score ($r = 0.93$) and distance ($r = -0.84$), as it may be directly impacted by whether a home is in proximity to a rail trail, which are heavily used for biking; on the other hand, walk score shows many desirable characteristics of the neighborhood separate from walking trails. Geographic variables like latitude and longitude were excluded because zip will capture enough information about the geographical location of the home; including them as well may complicate our interpretation of distance. Additionally, a map of Northampton was examined to check if the zip and distance variables were confounders. The map confirmed that the rail trail network is not contained within a single zip code but is instead split between both. This observation confirms that the two variables capture distinct geographical and proximity effects, which justifies including both covariates in the model.

To evaluate whether this model is appropriate, standard diagnostic checks are performed. Using the plots from Figure 4, I will check linearity, normality, homoskedasticity, and influence points. In the plot of standardized residuals on fitted values, it seems as though the linear model is an adequate model, and that we will have unbiased inference. The points seem to funnel out, which is an indicator of non-constant variance. This is particularly strange as I did the log transformation of the response variable to help stabilize our variance, however this could be due to possible model misspecification; non-constant variance will mean that the interpretation of the confidence intervals may not be as accurate as they could be with constant variance. The top left plot in Figure 4 shows that the model I chose is an adequate model, as the standardized residuals are spread out around the horizontal line at 0, and they don't follow another trend. From the top right plot in Figure 4, the standardized residuals seem to be approximately normal, with only a few points deviating away from the line, and all of the points enveloped in the confidence interval of the line. From the plots, there are some points that may be influence the regression analysis, but I decided to include all observations in the model for the sake of keeping our dataset representative of the town we are analyzing. Removing these points could bias the analysis, whereas including them allows the model to reflect the full range of home characteristics. Finally, from looking at our research design, there is reason to believe that the observations aren't independent; houses that are nearby have large effects on each other, whether it be house design or price valuations. However, there is not much we can do about that, as we are trying to analyze houses in specific areas to derive a conclusion about the effect of distance on adjusted 2007 price.

**Figure 4: Model Diagnostic Plots Indicating Approximate Normality and Adequacy, but Suggesting Non-Constant Variance**




The hypothesis of interest is to see whether the coefficient on distance is significantly different from 0. To test this hypothesis, a 95% confidence interval and its p-value will be used to measure the uncertainty around the estimated effect of distance on adjusted 2007 prices.

# Results

The primary goal of this analysis was to examine whether proximity to rail trail access has an effect on home value. The model, including the main effects and interaction terms previously mentioned, was fitted and the coefficients were estimated using LSE.

**Table 2: Regression Results Showing distance is Not Statistically Significant (p=0.060), while square footage is Highly Significant (p<0.001)**

In Table 2, we can see the following information about the covariates in the model. The coefficient of primary interest, distance, was negative but it was not statistically significant ($\hat{\beta}$=-0.043, SE=0.022, p = 0.060). This indicates that homes closer to the rail trail tend to have higher prices, however the effect isn't significant. Holding all other variables constant, moving 1 mile further away from the trail is associated with approximately a 4.3% decrease in home price. Among the other covariates, square footage had a significant positive effect ($\hat{\beta}$=0.391, SE=0.045, p $<$ 0.001). An increase of 100 square feet in home size is associated with approximately a 4% higher 2007 price, holding other factors constant; this indicates that larger homes were valued more. zip also had a significant effect, showing that homes with a 01062 zip code were lower-valued than the 01060 zip code ($\hat{\beta}$=-0.088, SE=0.037, p $<$ 0.05); homes in the 01062 zip code were associated with about 8% lower 2007 prices compared to homes in the 01060 zip code, holding other covariates constant. For the categorical variables, several bedroom categories were associated with higher home values relative to the reference category of one bedroom. Three-bedroom homes had a significant positive effect ($\hat{\beta}$=0.210, SE=0.086, p $<$ 0.05). Other interaction terms were not significant, and several combinations were dropped from the model because of singularities.

The model overall explains a large amount of the variance in the response variable ($R^2$=0.839), and the overall F-test indicated that the model significantly predicted home values (F=23.09, p $<$ 0.001).

Our covariate of main interest, distance, is negatively associated with home values in Northampton, MA. The estimated effect on adjusted 2007 price is $\hat{\beta}$=-0.043. However, as p = 0.060 and the 95% confidence interval of [-0.086,0.001] includes 0, our conclusion isn't strong.

| Predictor | Estimate | Std. Error | t value | Signif. |
|---|---|---|---|---|
| (Intercept) | 4.975 | 0.101 | 49.22 | *** |
| distance | -0.043 | 0.022 | -1.91 | . |
| acre | 0.137 | 0.133 | 1.03 | |
| squarefeet | 0.391 | 0.045 | 8.68 | *** |
| walkscore | 0.001 | 0.001 | 1.32 | |
| factor(bedrooms)2 | 0.165 | 0.089 | 1.85 | . |
| factor(bedrooms)3 | 0.210 | 0.086 | 2.44 | * |
| factor(bedrooms)4 | 0.126 | 0.098 | 1.29 | |
| factor(bedrooms)5 | 0.215 | 0.135 | 1.59 | |
| factor(bedrooms)6 | 0.384 | 0.195 | 1.97 | . |
| factor(garage_spaces)1 | -0.209 | 0.228 | -0.91 | |
| factor(garage_spaces)2 | -0.185 | 0.127 | -1.46 | |
| factor(garage_spaces)4 | 0.089 | 0.141 | 0.63 | |
| factor(zip)1062 | -0.088 | 0.037 | -2.39 | * |
| factor(bedrooms)2:factor(garage_spaces)1 | 0.232 | 0.271 | 0.85 | |
| factor(bedrooms)3:factor(garage_spaces)1 | 0.225 | 0.233 | 0.97 | |
| factor(bedrooms)4:factor(garage_spaces)1 | 0.236 | 0.239 | 0.99 | |
| factor(bedrooms)5:factor(garage_spaces)1 | NA | NA | NA | |
| factor(bedrooms)6:factor(garage_spaces)1 | NA | NA | NA | |
| factor(bedrooms)2:factor(garage_spaces)2 | 0.182 | 0.184 | 0.99 | |
| factor(bedrooms)3:factor(garage_spaces)2 | 0.236 | 0.134 | 1.76 | . |
| factor(bedrooms)4:factor(garage_spaces)2 | 0.198 | 0.138 | 1.43 | |
| factor(bedrooms)5:factor(garage_spaces)2 | NA | NA | NA | |
| factor(bedrooms)6:factor(garage_spaces)2 | NA | NA | NA | |
| factor(bedrooms)2:factor(garage_spaces)4 | NA | NA | NA | |
| factor(bedrooms)3:factor(garage_spaces)4 | NA | NA | NA | |
| factor(bedrooms)4:factor(garage_spaces)4 | NA | NA | NA | |
| factor(bedrooms)5:factor(garage_spaces)4 | NA | NA | NA | |
| factor(bedrooms)6:factor(garage_spaces)4 | NA | NA | NA | |

*Signif. codes:* 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual Std. Error = 0.1263 on 84 DF, $R^2 = 0.839$, Adjusted $R^2 = 0.803$
F-statistic = 23.09 on 19 and 84 DF, $p < 0.001$

# Discussion

The fitted model shows that the 95% confidence interval for the coefficient of distance includes zero, indicating that there is no statistically significant evidence from the 2007 data that proximity to the rail trail affected home prices in Northampton, MA. Although I am inclined to believe that a closer distance to a rail trail might make a home more desirable, we cannot rule out the possibility that the true effect of distance is actually zero, meaning that moving closer or farther from a rail trail is not associated with a meaningful change in 2007 home prices. Therefore, based on this data, there is no strong reason to prioritize rail trail proximity when evaluating homes in Northampton today.

A possible explanation for why the effect of distance was not statistically significant is that distance could explain the same variability as other covariates, such as the walk score or zip code. The overlapping effects could increase the standard error of the distance coefficient, making it harder to detect a significant relationship, even if proximity to the rail trail truly affects willingness to pay. It is also possible that rail trails do have an influence, but there may be other factors that take higher priority when people are looking for houses, such as the safety of the neighborhood.

Other home features, particularly square footage and the interaction between bedrooms and garage spaces, also influence home prices. For ACME Homes, this suggests that focusing on larger homes or those with favorable bedroom and garage combinations may be more effective for predicting higher values than focusing on proximity to a rail trail. While some of these effects are only marginally significant, they help explain variation in home prices and provide context for the importance of distance in the overall model.

There are several limitations that may affect the interpretations of this analysis. The data is observational which makes it difficult to do causal inference. Independence may be violated because homes that are near each other tend to share characteristics, meaning that their homes' prices are correlated. The data is also limited to homes sold in 2007, which may reduce the ability to generalize to current market conditions. Finally, omitted variables, such as school quality or neighborhood crime, could influence home prices and bias the estimated effect of distance.

Next steps could include analyzing more recent sales data to see if these patterns still hold up today, incorporating additional neighborhood-level covariates, and using spatial models to account for correlations among nearby homes. These steps would help provide more accurate and actionable insights for home valuation decisions. Overall, while proximity

to rail trails does not appear to meaningfully influence home prices, features like square footage and bedroom/garage combinations provide more reliable guidance for home valuation, though ongoing monitoring of the market is recommended.

# Appendix

**Figure A1: Distributions of Discrete Covariates, Most Homes Have 1-2 Garage Spaces and 2-3 Bedrooms**


**Figure A2: Distributions of Continuous Variables, Showing That Adjusted 2007 Price is Right-Skewed**

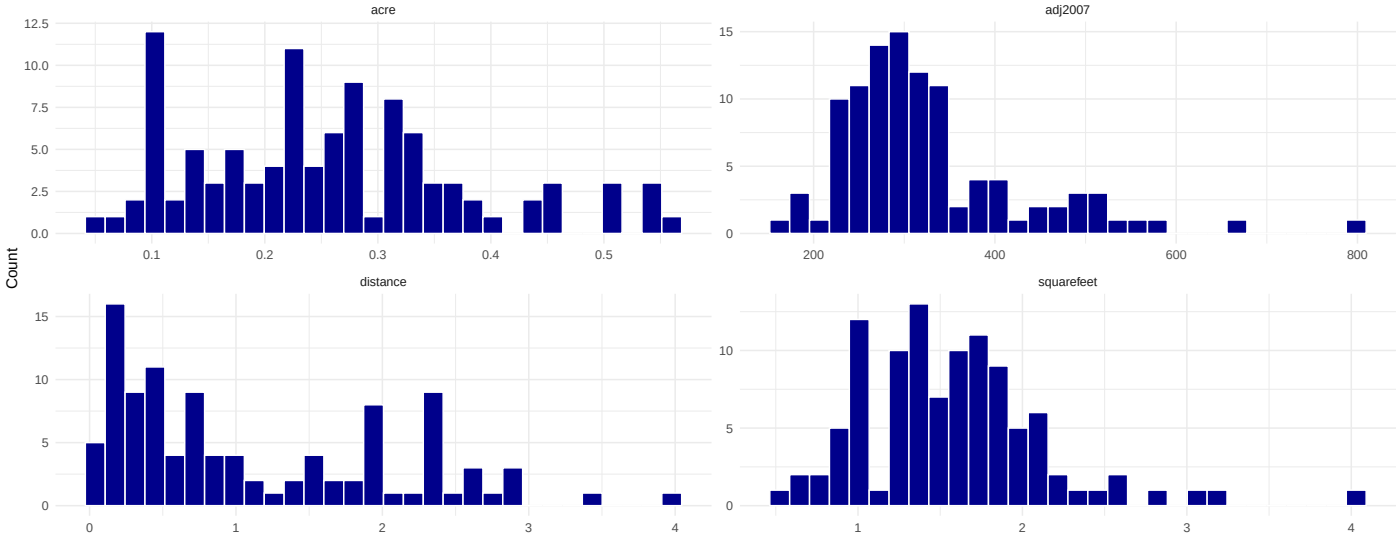



**Figure A3: Distribution of the Adjusted 2007 Home Price after Log Transformation, Showing That it is More Symmetric**

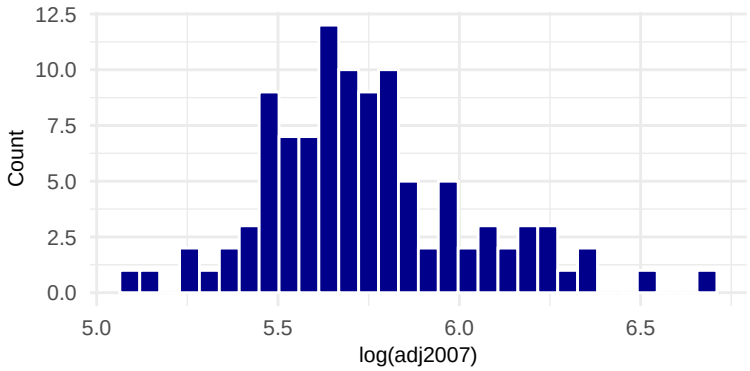


**Figure A4: Interaction between Bedrooms and Garage Spaces on Log-Transformed 2007 Home Price**