Rail Trails Report

Dominick Robinson

2025-10-31

1 Introduction

In the 1980s, some of the numerous unused rail lines in the United States began to be converted to rail trails, i.e. walking and biking trails along the route of an old rail line. Rail trails tend to be long continuous trails, often paved, and have gentle slopes at most (because trains cannot climb steep hills). This makes them accessible and easy to bike along.

Some have suggested that rail trails may be an attractive feature for people buying homes, who might be willing to pay more for a home close to a trail where they could walk or bike for fun or for their commute.

This leads us to the main questions of this report:

1. Are rail trails attractive for people buying homes, who might be willing to pay more for a house closer to a rail trail?
2. If so, what is the relationship between proximity to a rail trail and house value?

These questions are important to Acme Homes, LLC., who ultimately want to know whether building a house near a rail trail makes it more valuable (and therefore likely more profitable) compared to building a house far away from a rail trail.

Ro address this question, the rail trails dataset, which includes data from *Zillow.com* on homes sold in Northampton, Massachusetts in 2007, was analyzed to test whether distance to rail trail access has a statistically significant effect on home value. No significant effect was found.

2 Data

2.1 Dataset

Our dataset contains 104 observations in the data and 20 variables.

2.2 Variables

The response variable in this analysis is the home's adjusted selling price (`adj2007`, in thousands of dollars), representing the inflation-adjusted 2007 selling price. Selling price is an appropriate measure of home value because it reflects a consumer's willingness to pay.

The dataset contains no missing entries. However, it includes only homes on lots smaller than 0.56 acres and excludes two properties that increased in value by over $\$500,000$ between 1998 and 2007 due to major renovations, which were treated as outliers.

The primary explanatory variable is `distance` (in miles), representing proximity to the rail trail. To control for other determinants of home value, we include covariates such as `acre` (lot size), `bedrooms`, `garage_spaces`, and `squarefeet` (in thousands), which capture home size and amenities. The categorical variable `zip` accounts for neighborhood-level factors such as school quality and safety. Both zip codes provide access to the same rail trail, so including `zip` is unlikely to obscure the effect of `distance`, but this will be verified by plotting `zip` against `distance`.

Variables that do not meaningfully explain home value or would confound the relationship between `distance` and `adj2007` are excluded. These include identifiers (`housenum`, `streetname`, `streetno`), other sale prices, and geographic coordinates (`latitude`, `longitude`), as the latter could absorb the spatial component of trail proximity.

According to `walkscore.com`, `walkscore` (0–100) measures walkability based on nearby amenities and street design but does not account for walking trails or terrain *(WalkScore, n.d.)*. Thus, including it should not confound the

effect of rail trail proximity. In contrast, `bikescore` (0–100) reflects the presence of bike lanes, connectivity, and terrain *(WalkScore, n.d.)*. Because rail trails are integral to bike infrastructure, `bikescore` is likely collinear with `distance` and will therefore be excluded.

Finally, because the value of a given lot size may differ by location, an interaction term between `acre` and `zip` is included in the model to capture potential heterogeneity in how property size affects home value across neighborhoods.

2.3 1D EDA

Based on the distributions of the variables shown in **Figure 1**, several patterns emerge. `adj2007` is heavily right-skewed. Applying a logarithmic transformation reduces both this skewness and the influence of homes with unusually high adjusted prices. The distribution of `distance` is also right-skewed, with a long tail of larger values. However, to preserve interpretability, a log transformation is not applied to this variable in our model. The distribution of `squarefeet` is relatively symmetric, although a few homes have substantially larger values. The `walkscore` variable is asymmetric with a heavy right tail, but this pattern is not expected to give only a few observations disproportionate leverage. `acre` is slightly skewed, with a small number of properties situated on larger parcels, though not enough to justify a transformation. The variables `bedrooms` and `garage_spaces` are discrete counts with limited ranges. Because their effects on price are unlikely to be linear, they were collapsed into ordered categorical variables (e.g., "1 bedroom," "2+ bedrooms"). Finally, `zip` was treated as a categorical factor to capture neighborhood-level effects without assuming any inherent order.

Summary statistics for these variables appear in **Table 1**. After applying a logarithmic transformation to **`adj2007`**, its distribution became substantially more symmetric, as seen in **Figure 1**, addressing prior concerns about high leverage.

2.4 2D EDA

As shown in **Figures 2** and **3**, there is a slight negative relationship between `distance` and `log_adj2007`, indicating that homes closer to the rail trail tend to have higher prices. The positive association between `acre` and `log_adj2007` aligns with expectations that larger lots are generally more valuable. Variation in price by number of garages, bedrooms, and ZIP code also supports their inclusion as predictors. Although `zip` and `distance` are somewhat correlated, both ZIP codes include properties with distances near zero, suggesting that `zip` does not necessarily confound the effect of `distance`. Differences in the relationship between `acre` and `adj2007` across ZIP codes further justify including the interaction term in the model.

No abnormal outliers were detected. However, the dataset includes only 104 observations, which may reduce statistical power. Moreover, because the data reflect home sales from 2007, housing market conditions and consumer preferences may have since changed. Finally, some relevant factors, such as school quality, crime rates, or the year homes were built, are not available in the dataset, though `zip` was used as a partial proxy for these characteristics.

3 Methods

To assess whether proximity to rail trail access affects home value, the following linear regression model was estimated, incorporating the covariates as described in the **Data** section:

$$\begin{aligned}\mathrm{E}[\text{log\_adj2007} \mid X] = \beta_0 &+ \beta_1 \,\text{distance} + \beta_2\, I\{\text{bedrooms} \geq 2\} \\ &+ \beta_3\, I\{\text{garage\_spaces} = 1\} + \beta_4\, I\{\text{garage\_spaces} \geq 2\} + \beta_5 \,\text{walkscore} \\ &+ \beta_6 \,\text{squarefeet} + \beta_7\, I\{\text{zip} = 1062\} + \beta_8 \,\text{acre} + \beta_9\, (I\{\text{zip} = 1062\} \times \text{acre})\end{aligned}$$

After fitting the model, diagnostic plots were examined to verify that the assumptions necessary for valid statistical inference were satisfied. Subsequently, a hypothesis test was conducted to determine whether there is statistically significant evidence to reject the null hypothesis that proximity to rail trail access has no effect on home value. Formally, the hypotheses tested were:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_A : \beta_1 \neq 0$$

4 Results

The effect of proximity to rail trail access on home prices was evaluated by examining the coefficient for `distance` in the linear regression model, controlling for other potential confounders of home value. As shown in **Figure 4**, the Q–Q plot indicates that the residuals are approximately normally distributed, satisfying the normality assumption. However, the residuals vs fitted values plot suggests potential heteroskedasticity, which may affect the validity of the statistical inference.

The model summary can be seen in **Table 2**. The estimated coefficient for distance was -0.038 (SE = 0.021, t(94) = -1.79, p = 0.0764).

Since the p-value is $0.0764 \geq 0.05$, the null hypothesis that distance has no effect on home prices cannot be rejected at the 95% confidence level. In other words, there is no statistically significant evidence at the 95% confidence level that proximity to rail trail access influences home value.

5 Discussion

Based on the results of the statistical analysis, no statistically significant evidence was found that proximity to rail trail access affects home value. Therefore, it is not recommended that Acme Homes, LLC. prioritize building homes near rail trail access points as a strategy for directly increasing home values.

Several limitations may have influenced the accuracy and generalizability of these findings.

First, the dataset includes only housing data from Northampton, Massachusetts. Homebuyers in other regions of the United States may have different preferences and purchasing behaviors, limiting the generalizability of the results. Expanding the dataset to include additional geographic areas would help address this limitation.

Second, the dataset contains only 104 observations, which is a relatively small sample size that restricts the ability to estimate more complex models. Collecting additional data would enable more robust modeling and inference.

Third, the dataset omits several important home characteristics that likely influence price, such as neighborhood safety, school quality, year built, and available amenities. Although `zip` was included as a proxy for some of these factors, a more comprehensive dataset would improve model accuracy.

Fourth, the model used adjusted home prices from 2007 as a proxy for willingness to pay. However, housing markets and consumer preferences may have changed substantially since then. Using more recent data would yield estimates that better reflect current conditions.

Finally, the residual vs fitted values plot suggested heteroskedasticity, which may compromise the validity of statistical inference. Future research should address this issue by applying heteroskedasticity-robust methods, such as computing robust standard errors using the HC3 correction or weighted least squares, to obtain more reliable estimates.

6 Reference

1. Walk Score. (n.d.). Walk Score methodology. https://www.walkscore.com/methodology.shtml

7 Appendix

Variable	Min	Q1	Median	Mean	Q3	Max
adj2007	162.60	260.56	303.65	327.55	349.47	798.62
distance	0.04	0.33	0.76	1.11	1.90	3.98
acre	0.05	0.17	0.25	0.26	0.33	0.56
squarefeet	0.52	1.21	1.52	1.57	1.83	4.03
walkscore	2.00	14.75	36.00	38.88	60.75	94.00
bedrooms	1.00	3.00	3.00	3.25	4.00	6.00
garage.spaces	0.00	0.00	1.00	0.76	1.00	4.00

Table 1: Summary statistics for variables.

term	estimate	std.error	statistic	p.value
(Intercept)	4.970	0.099	49.962	0.000
distance	-0.038	0.021	-1.792	0.076
bedrooms_cat2+	0.228	0.081	2.806	0.006
garage_cat1	0.016	0.032	0.508	0.613
garage_cat2+	0.035	0.036	0.977	0.331
walkscore	0.001	0.001	1.545	0.126
squarefeet	0.331	0.030	10.959	0.000
zip1062	-0.030	0.069	-0.436	0.664
acre	0.349	0.192	1.819	0.072
zip1062:acre	-0.276	0.254	-1.088	0.279

Table 2: Summary of linear regression results. There is no statistically significant evidence that distance affects log_adj2007.

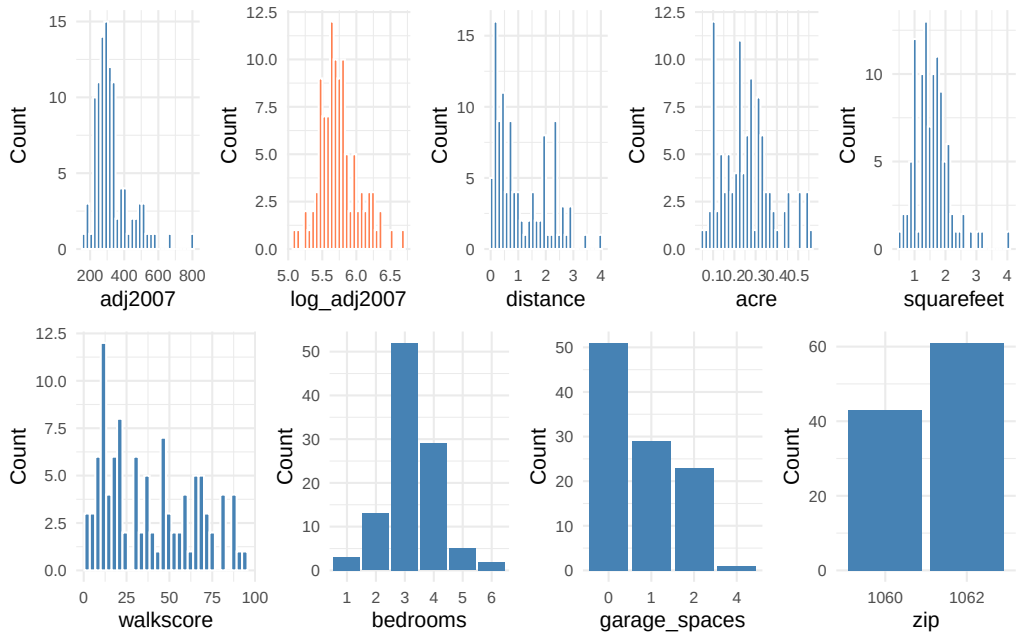

Figure 1: Distributions of important continuous and categorical variables in the dataset, including log-transformed adjusted 2007 home prices. Some of the distributions are skewed, which could create issues with leverage points.

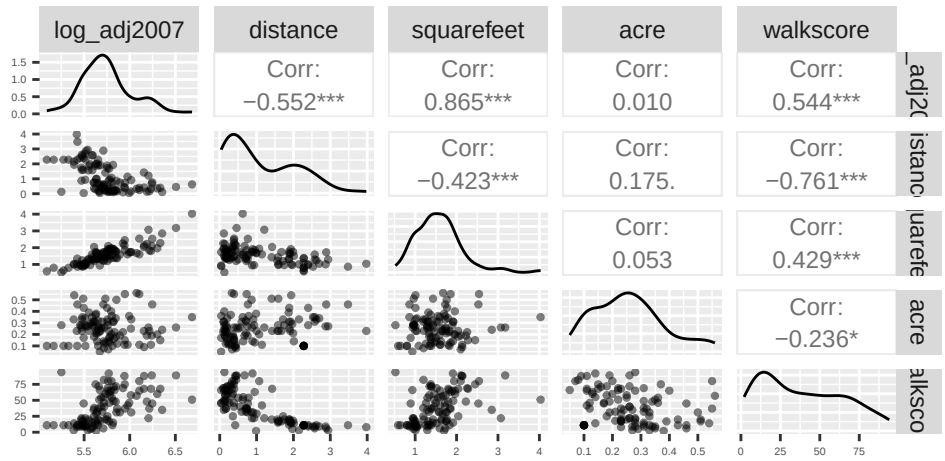

Figure 2: Pairwise scatterplots and correlations among continuous predictors and the log-transformed response variable, `log_adj2007`.

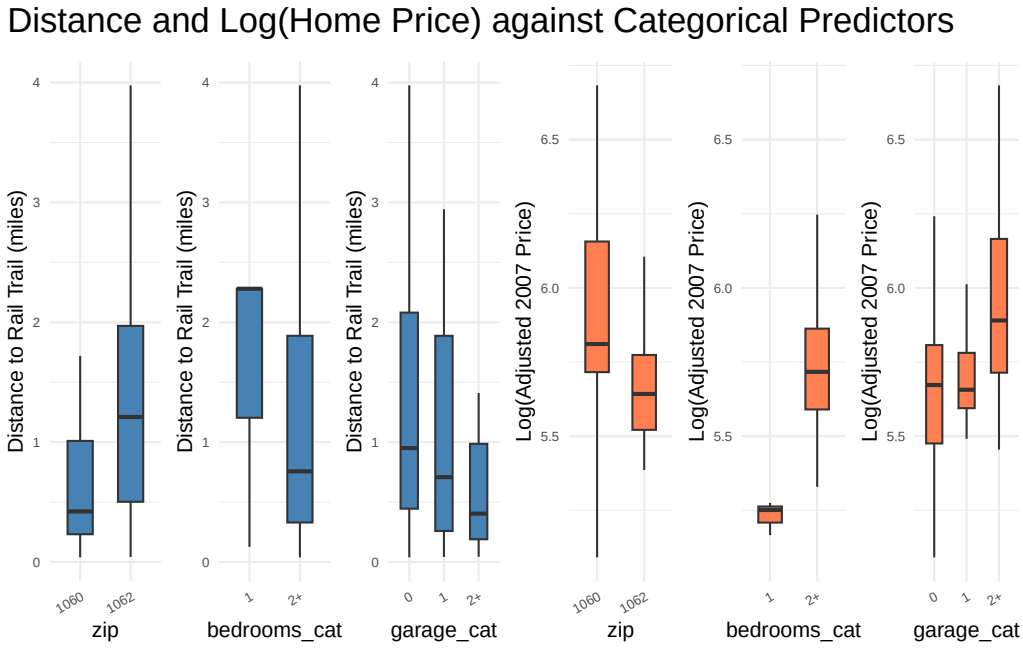

Figure 3: Boxplots of categorical predictors (`zip`, `bedroom`, and `garage_spaces`) against both distance to rail trail and log-adjusted 2007 home price.

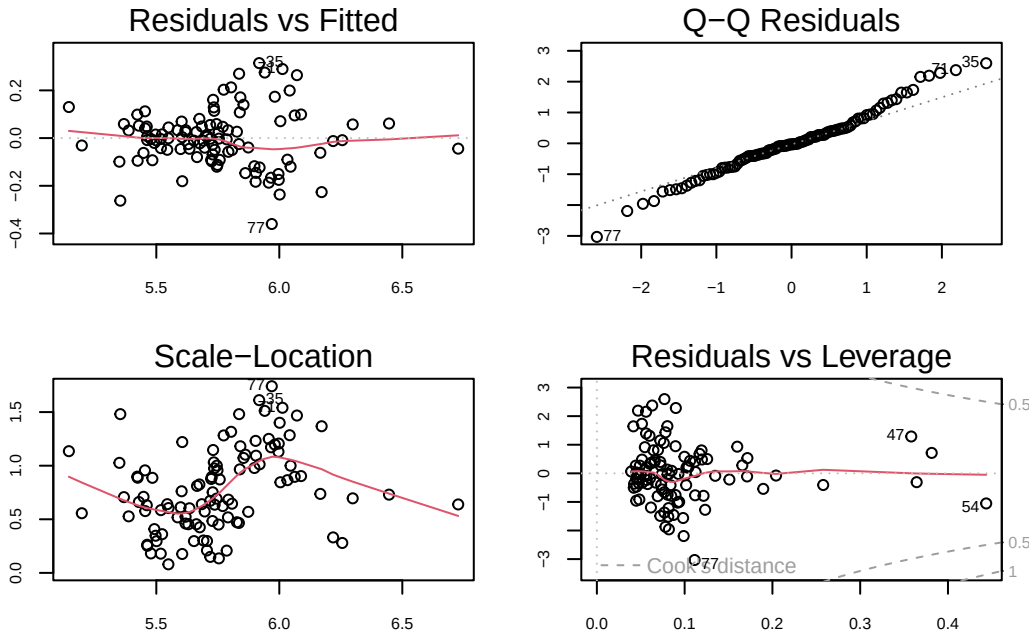

Figure 4: Diagnostic plots for the fitted linear regression model, including residuals vs. fitted values, Q-Q plot, scale-location, and residuals vs. leverage. Residuals appear mostly normally distributed. However, variance of residuals may not be heteroskedastic, potentially violating a model assumption.