Predicting Effect of Distance from Rail Trails on House Prices

Rao Abdul Hannan, Mark Ma

Executive Summary

This report analyzes home value trends from 1998 to 2014, concentrating on properties with less than 0.56 acres of land. The primary objective was to identify the key factors that drive the appreciation of home values within this specific segment of the housing market. To achieve this, we employed statistical methods such as regression analysis to explore the relationships between home value increases and various potential factors, including location, property size, and proximity to amenities like schools, parks, and shopping centers. Additionally, we considered the impact of the broader economic conditions during the study period on property values. We found that the closer a home is to the rail trail, the more its value increases. Specifically, for every foot nearer to the trail, a home's price goes up by about 0.05%. This effect is stronger than that of property size — while larger homes do sell for more, proximity to the trail has a bigger impact. By excluding two homes that experienced over $500,000 in value increases due to major renovations, we ensured that the results reflect typical market behavior rather than being skewed by exceptional cases. This approach provides a clearer understanding of the primary drivers behind home value increases for properties within the specified land size.

Introduction

During the late 19th and early 20th centuries, the United States saw the construction of an extensive network of rail lines that connected towns and cities, facilitating passenger travel and cargo transport. However, with the advent of the automobile and the expansion of the Interstate Highway System, reliance on rail transportation diminished significantly. This shift led to the closure and abandonment of many rail lines; some were preserved for potential future use, while others were sold.

Starting in the 1980s, a transformative initiative began to re-purpose these defunct rail lines into rail trails—dedicated walking and biking paths that trace the routes of the old tracks. Characterized by their long, continuous stretches and gentle gradients (a legacy of trains' inability to navigate steep inclines), these trails are often paved and highly accessible, making them ideal for recreational cycling and walking.

The emergence of rail trails has sparked interest in their potential impact on residential property values. It is hypothesized that these trails enhance the attractiveness of nearby homes, with buyers possibly willing to pay a premium for the convenience of easy access to recreational and commuting options.

Acme Homes, LLC, a company specializing in large-scale residential developments, is exploring opportunities to maximize the profitability of their future projects. The development manager, Mr. W. E. Coyote, has commissioned this report to investigate the following key questions:

- Are rail trails appealing to home buyers to the extent that they increase the willingness to pay for houses located nearer to them?

- If they are, what is the specific relationship between a property's proximity to a rail trail and its market value?

This report aims to analyze these questions by examining housing market data in relation to the proximity of homes to rail trails. The findings will assist Acme Homes in making informed decisions about where to focus their development efforts to achieve optimal returns.

Exploratory Data Analysis

This study utilizes the `rail` data set containing information of 104 houses in the Northampton (01060) and Florence (01060) neighborhoods in Northampton, Massachusetts from an observational study. The details of the variables are appended below:

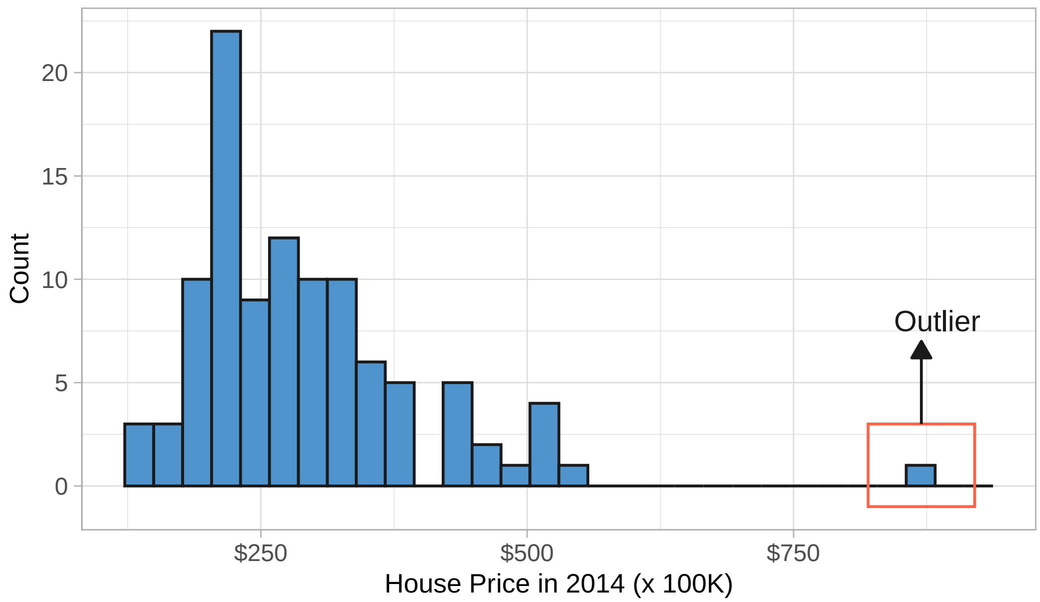
Variable Name	Variable Description
housenum	A unique number for each house
price1998_adj	Zillow's estimated value for the home in 1998, in thousands of 2014 dollars
price2007_adj	Zillow's estimated value for the home in 2007, in thousands of 2014 dollars
price2011_adj	Zillow's estimated value for the home in 2011, in thousands of 2014 dollars
price1998	Zillow's estimated value for the home in 1998, in thousands of dollars
price2007	Zillow's estimated value for the home in 2007, in thousands of dollars
price2011	Zillow's estimated value for the home in 2011, in thousands of dollars

price2014	Zillow's estimated value for the home in 2014, in thousands of dollars
distance	Distance (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, where 100 indicates high bike-frinedliness, such as flat terrain and good bike lanes.
walkscore	Walkability of the area, estimated by WalkScore.com. 0-100 scale, where 100 indicates high walkability, so most daily tasks can be done without a car
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
streetno	House number on the street
zip	ZIP code of the house (leading 0 omitted). 1060 is Northampton, MA; 1062 is Florence, MA.

Table 1: Variable Descriptions

The variable of interest in this case is `price2014` and how it is affected by the `distance` variable. First and foremost, we checked the distribution of `price2014` by the aid of a histogram, depicted in Figure 1.

Figure 1: Distribution of 2014 House Prices

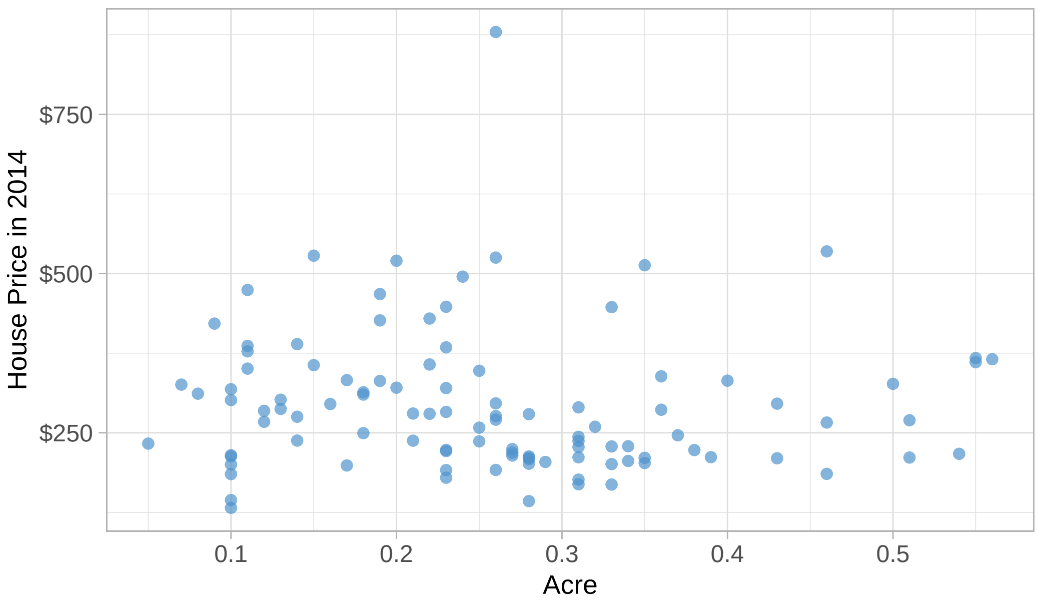


It is quite evident that there is an outlier with an unusually high price, as highlighted in Figure 1. Upon further inspection, this is house #97 in the data set which has 6 bedrooms and a price of $\$879,000$ however, its values for other variables including `acre, bikescore, walkscore, distance, garage_space` and `squarefeet` are not such that they would suggest an incredibly high price like the one we are observing in the data set. This could potentially cause problems when we fit a model on the data since the outlier may pull the regression function towards it and adversely affect the slope of the estimated regression line.

Next, we critically evaluate `price2014` against all our potential covariates to check the behavior of the data and decide which variables we need to include in the regression model.

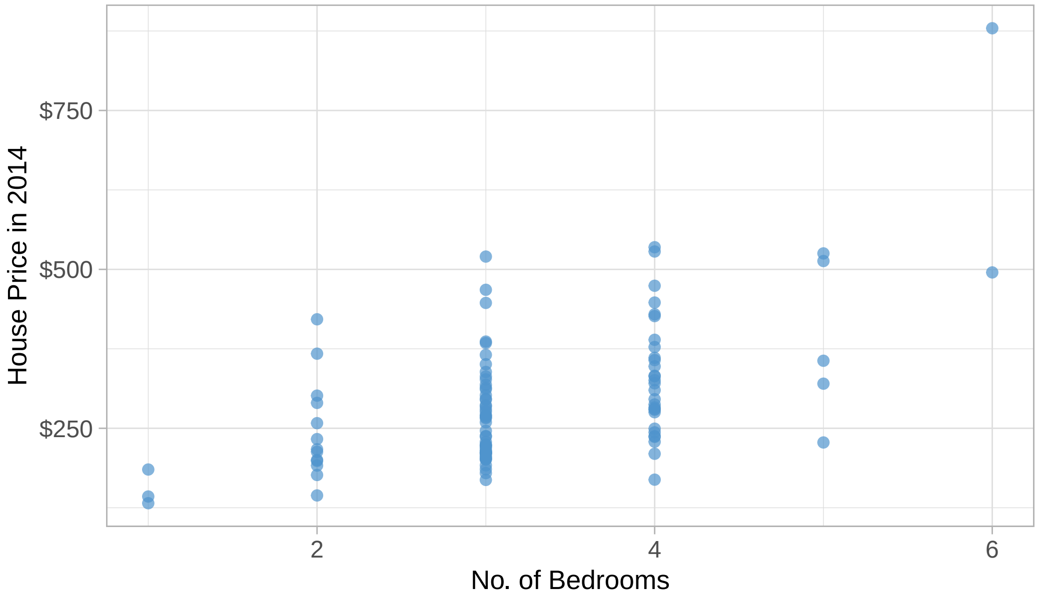
Key takeaways

Figure 2: House Price in 2014 vs Acre



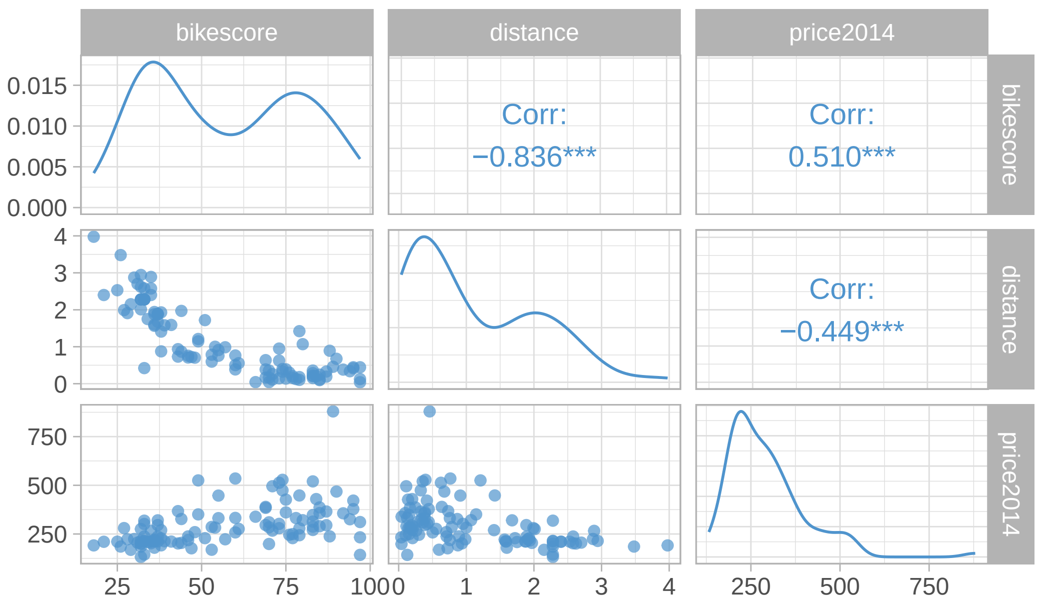
- `acre`: There is no evident trend in the `price2014` vs `acre` plot displayed in Figure 2. However, underlying trends can sometimes be invisible in plots and we strongly believe that the size of the property should effect the price of the house. Therefore, we decide to include the `acre` variable as a covariate in our regression model

Figure 3: House Price in 2014 vs No. of Bedrooms



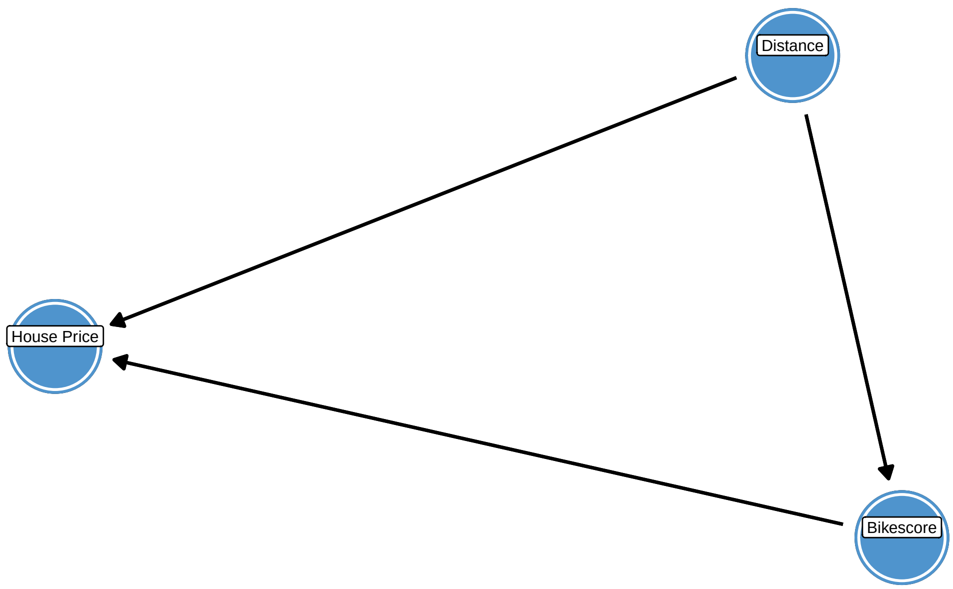
- `bedrooms`: The price seems to increase with each additional bedroom, which is visible in the Figure 3 and hence we include it as a covariate

Figure 4: Distance and House Price vs Bikescore



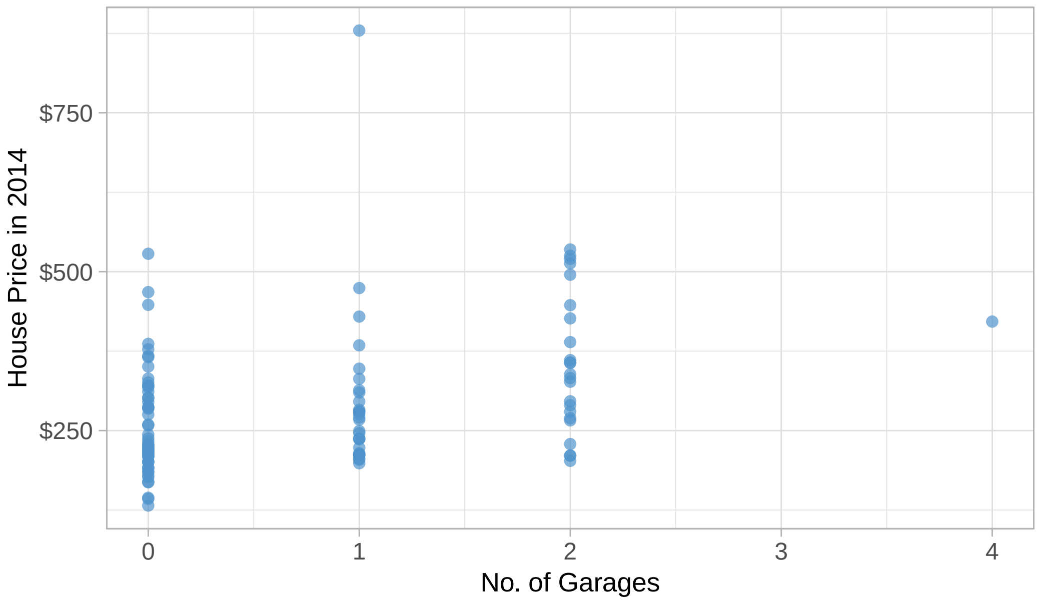
- `bikescore`: The `bikescore` variable is effectively calculated through the `distance` variable. A negative non-linear trend is prominent in the `bikescore` vs `distance` plot in Figure 4 with a correlation value of -0.836. Since our primary research question is concerned with `distance`, we decide to exclude `bikescore` from our model to avoid issues with multi-collinearity which would lead to higher standard errors for the coefficient of `distance`, resulting in wider confidence intervals which significantly limit our capability to make inference about the effect of `distance` on `price2014`. It is also important to highlight that including `bikescore` in the model will not allow us to capture the full affect of `distance` on the house prices, rather just the direct affect since `bikescore` is a mediator as depicted by the Directed Acylic Graph (DAG) in Figure 5

Figure 5: DAG showing Bikescore as a mediator to Distance



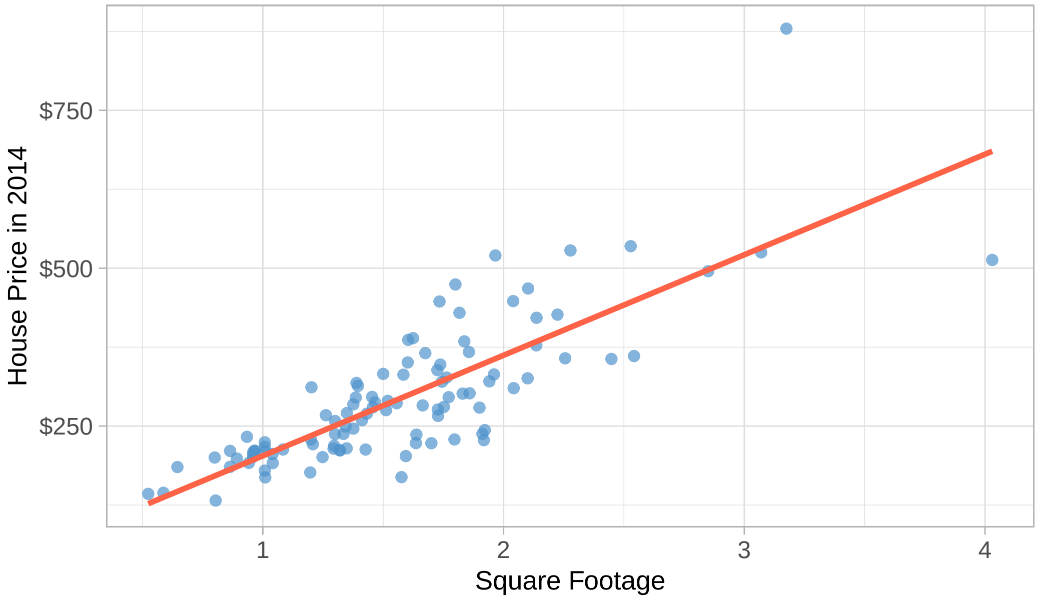
- `distance`: It is the primary covariate of concern and is therefore added in the model

Figure 6: House Price in 2014 vs No. of Garages



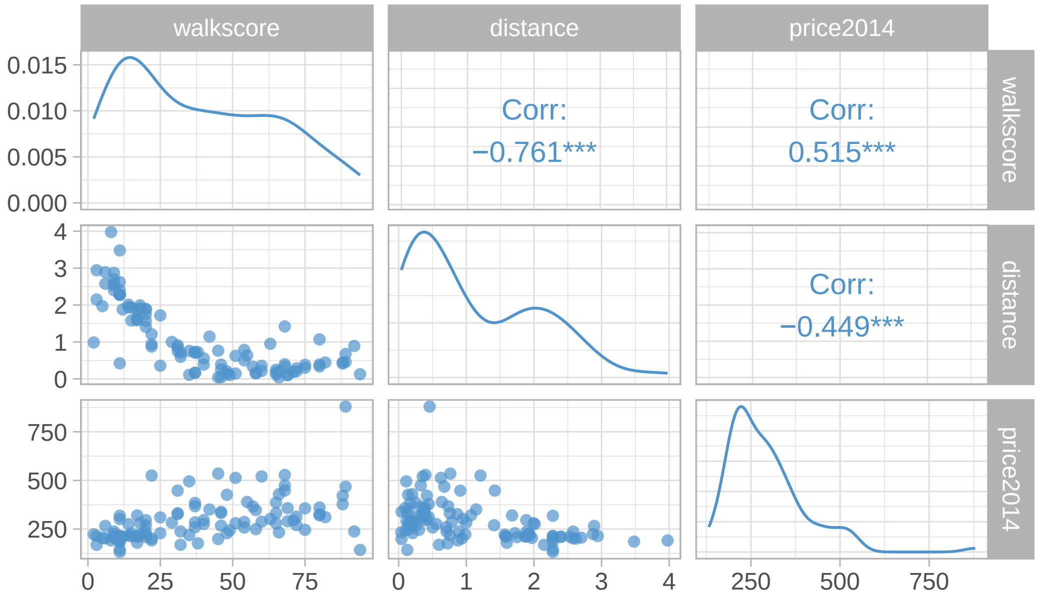
- `garage_spaces`: A house with more garages is expected to have a higher price. This is validated by the `price2014` vs `garage_spaces` plot in Figure 6 where a minor positive trend can be seen, therefore we add it in our model

Figure 7: House Price in 2014 vs Square Footage



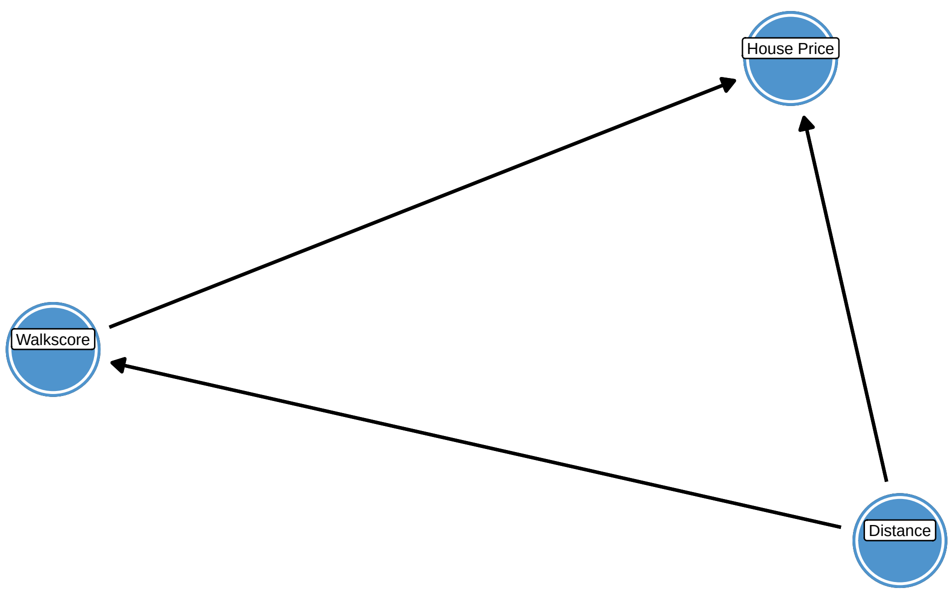
- `squarefeet`: This variable has a positively linear trend which is eminent in the `price2014` vs `squarefeet` plot in Figure 7. This aligns well with our expectations because a bigger house is generally expected to cost more

Figure 8: Distance and House Price vs Walkscore



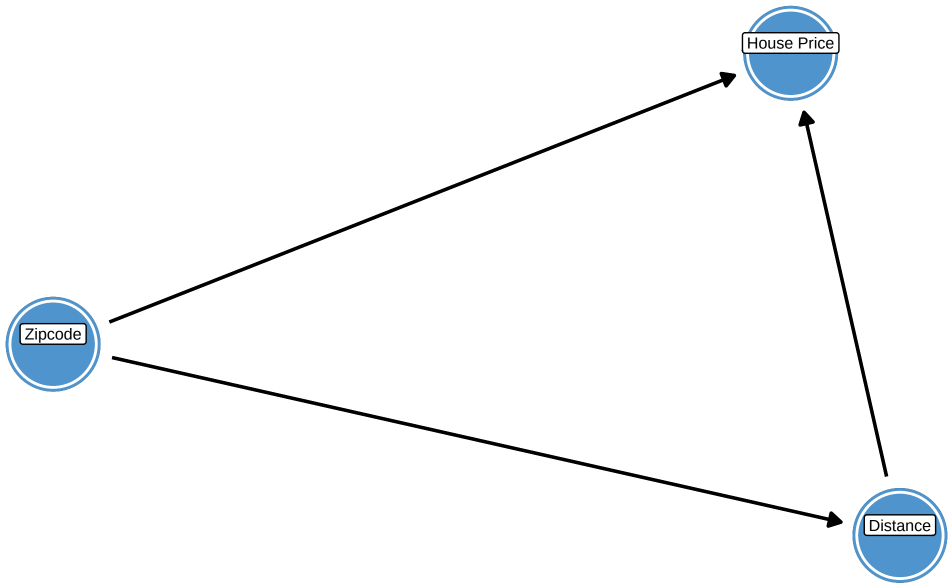
- `walkscore`: Just like `bikescore`, this variable is also calculated directly through the `distance` variable and has a negative non-linear trend observed in Figure 8 along with a high correlation value of -0.761, which leads us to exclude this variable from the model. This variable is also a mediator as displayed in Figure 9

Figure 9: DAG showing Walkscore as a mediator to Distance



- `zip`: This variable gives us information on both `distance` and `price2014` i.e. it is a confounder as depicted in Figure 10.

Figure 10: DAG showing Zipcode as a confounder for Distance



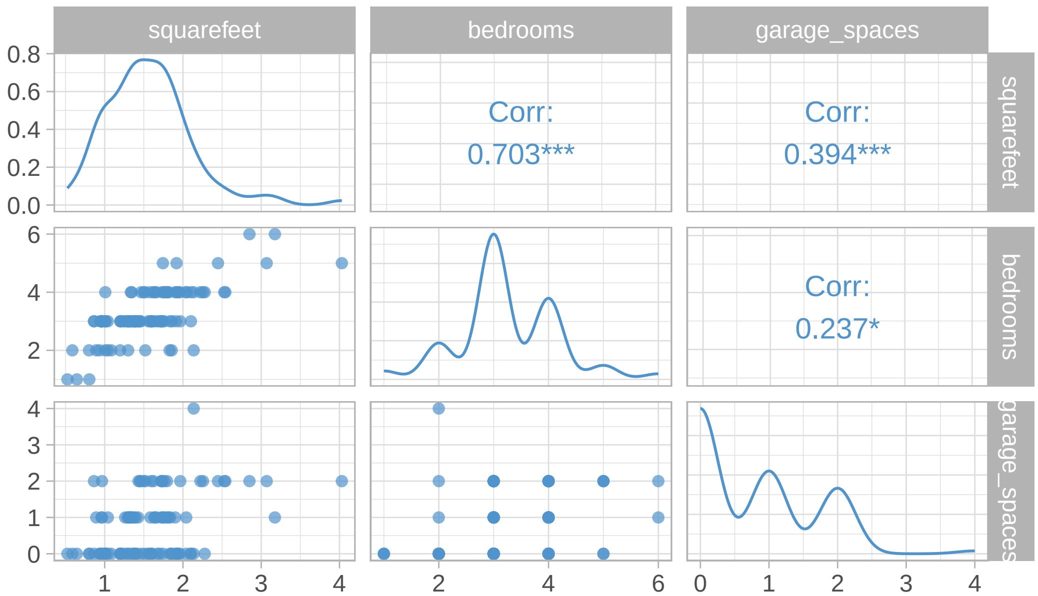
The zip code of a house provides approximate information on the location of the house which can help us get an idea of the `distance` value for the house i.e. how far it is from the nearest rail track entry. Similarly, it also provides information on the `price2014` variable because zip codes are associated with schools, hospitals and other facilities which can drive up the price of houses in the vicinity. We verify this by calculating the average distance and house prices for both zip codes and observe substantial differences between the values as shown in Table 2. Excluding a confounder such as `zip` would also lead to biased estimates of the coefficient of `distance` which would raise major concerns about our conclusions.

ZIP Code	Avg Distance	Avg Price
1060	0.77	338.9
1062	1.36	260.8

Table 2: Avg Distance and Price by Zip Code

- `latitude`, `longitude`: Both the `latitude` and `longitude` variables provide us information on the location of the house, however we have the variable `zip` which provides us similar information. We exclude `latitude` and `longitude` since they would require us to fit a non-parametric smoother and prefer `zip` over it for location information
- `streetname`, `streetno`: We again prefer `zip` over these two variables for location since they are discrete and have 73 and 86 different values and using them in our regression model would result in us losing an extensive number of degrees of freedom

Figure 11: No. of Bedrooms and Garages vs Square Footage

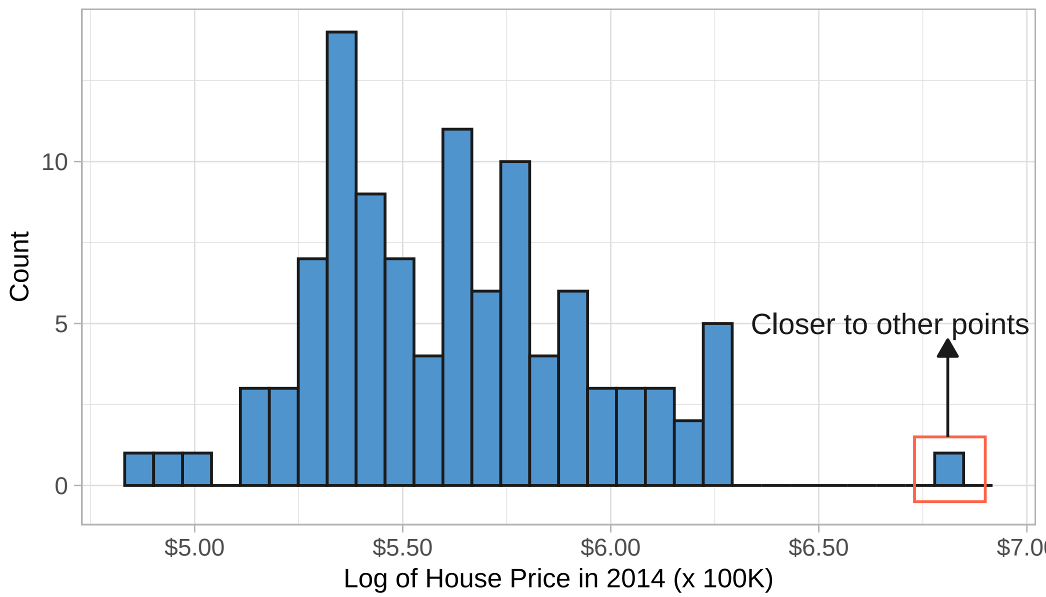


- `squarefeet` has a positive association with both `bedrooms` and `garage_spaces` as displayed by the plots and correlation values in Figure 11, which is reasonable since a bigger house is expected to have more bedrooms and garage spaces. This observation will be referenced in the Results section later on
- There is also a high leverage point which can be seen in all the `price2014` vs covariates plots displayed above. This is identified as house #89, which has a price of $513K$ even though it has 6 bedrooms and 4 garage spaces, and a `squarefeet` value greater than 4. Further investigation on whether this is a bad leverage point will be carried out in the Methods section

Methods

In the EDA section, we emphasized house #97 being a potential outlier. To cater for this issue, we applied a logarithmic transformation to `price2014` and the resulting distribution can be seen in Figure 13.

Figure 13: Distribution of Log of 2014 House Prices



Applying the transformation stabilizes the skewness of `price2014` and therefore, we decide to use this as the response variable for our regression analysis. The regression model we fit is represented by Eq (1):

$$E(\log(price2014)) = \beta_0 + \beta_1 distance + \beta_2 acre + \beta_3 bedrooms + \beta_4 garage_s paces + \beta_5 squarefeet + \beta_6 zip \tag{1}$$

where all variables are treated as continuous except `zip` which is treated as a discrete variable with two levels i.e. 1060 and 1062 with the following coding:

- `zip` $= 0; \quad for\ 1060$
- `zip` $= 1; \quad for\ 1062$

The `bedrooms` variable could also have been coded as a discrete variable however for the sake of interpretability and preserving more degrees of freedom, we include it as a continuous variable. Since `bedrooms` variable has six distinct values, we would need to create five dummy variables to code it as a discrete variable, which would lead to losing five degrees of freedom and also interpreting results for six different categories of houses i.e. one category for each distinct value of `bedrooms`.

Keeping in view our end goal is to predict the effect of `distance` on `price2014`, we conduct the following test when fitting the regression model:

$$H_0: \quad \beta_1 = 0$$

$$H_A: \quad \beta_1 \neq 0$$

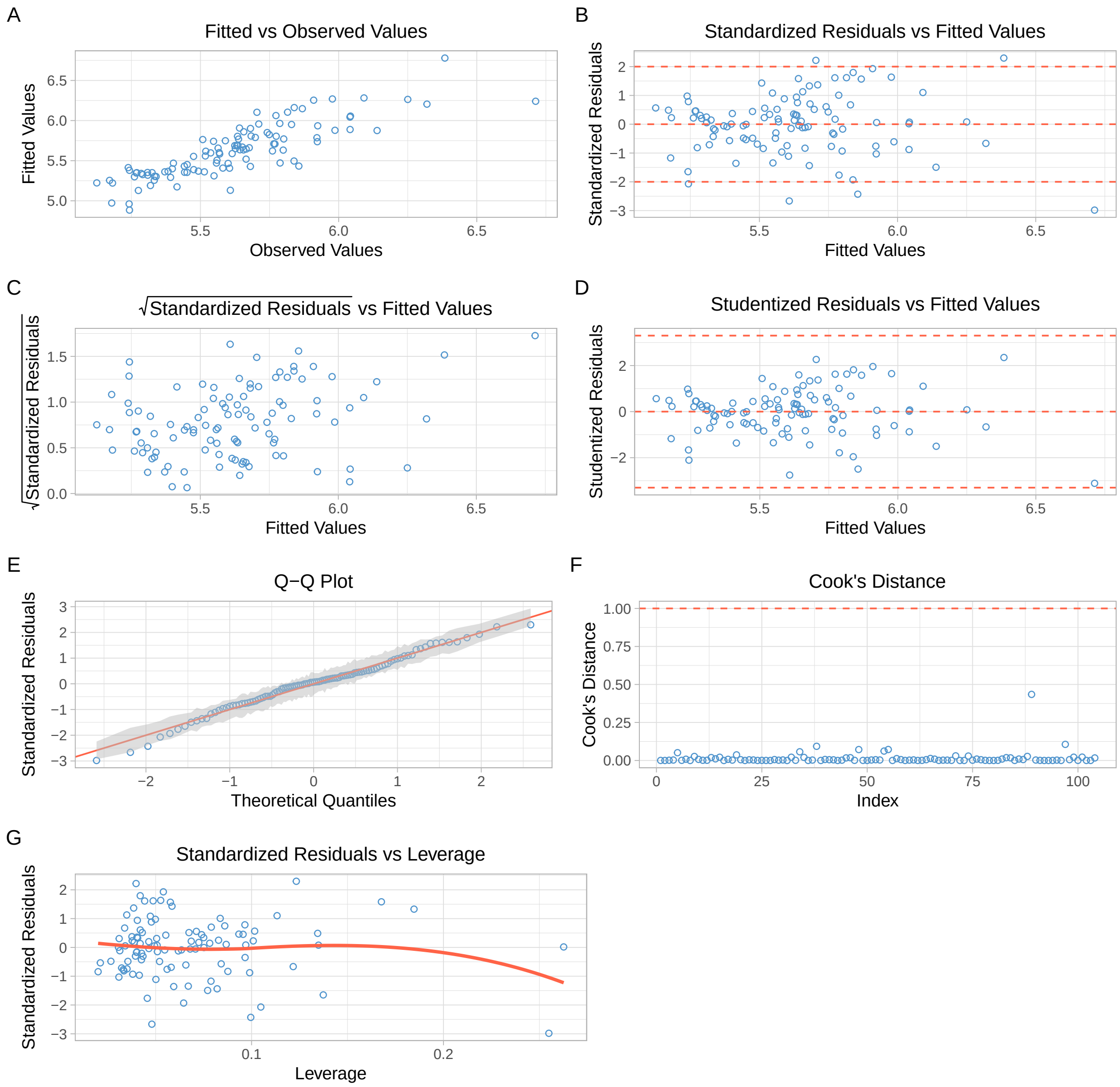
Under the null hypotheses H_0, the coefficient of `distance` β_1 is 0 i.e. `distance` does not have any effect on `price2014` whereas we will reject the null if we get a $p-value \leq \alpha = 0.05$ which is our significance level. The output of the `summary()` function after fitting the model in `R` is appended below:

Variable	Estimate	Standard Error	t-value	p-value
(Intercept)	5.06004	0.08459	59.82118	< 2e-16
distance	-0.05936	0.02211	-2.68437	0.00855
acre	-0.11666	0.18341	-0.63606	0.52624
bedrooms	0.01949	0.02908	0.67026	0.50428
garage_spaces	0.03784	0.02384	1.58723	0.11572
squarefeet	0.38650	0.05603	6.89842	5.4e-10
factor(zip)1062	-0.06738	0.04884	-1.37955	0.17090
Residual Standard Error	0.18354	NA	NA	NA
Deg of Freedom	97.00000	NA	NA	NA
Multiple R-squared	0.71551	NA	NA	NA
Adjusted R-squared	0.69792	NA	NA	NA
F-statistic	40.66070	NA	NA	NA

Table 3: Output of Linear Regression

To check the goodness-of-fit of our model and identify any potential problems, we turn towards diagnostic plots shown in Figure 13.

Figure 13: Diagnostic Plots



In Figure 13, Plot F shows a point with a higher Cook's distance than all the other points; this is house #89 which we identified as a potential high leverage point in the EDA section. However, since the Cook's distance value is less than our threshold of 1, we do not classify it as a bad leverage point. If we look at Plot D, all the points are within the $[-3.3, 3.3]$ interval which we get after setting the confidence level to α/n where $\alpha = 0.05$. Hence, when we calculate the corresponding z-values they turn out to be $z_{\alpha/(2n)} = 3.3$ and $z_{-\alpha/(2n)} = -3.3$ respectively. All studentized residuals within these limits imply we do not have any outliers that are adversely affecting our model. Plot D also shows that even though the variance of the residuals is not exactly constant, it does seem fairly stable. Q-Q plot in Plot E supports normality of the residuals, there are a few points in the tails farther away from the normal line however - 1) this can be attributed to randomness and 2) almost all points are still within the global confidence interval band calculated using confidence level of α/n to cater for multiple testing. The house prices are assumed to be independent of each other which is appropriate since the price of one house usually does not have an effect on the price of any other house. Since the assumptions of linear regression seem to hold true, we shift our attention to answering our research question.

Results

The estimated coefficients of `bedroom`, and `garage_spaces` do not have significant $p-values$. One plausible explanation for this is that the variation in the house prices explained by these variables are also captured by the `squarefeet` variable which has a highly significant $p-value$. From their corresponding plots with `squarefeet` in Figure 11, `bedrooms` and `garage_spaces` show a positive association with `squarefeet` and have correlation values of 0.703 and 0.394 respectively. `acre` neither shows a clear trend in the plot with `squarefeet` nor does it have a high correlation value with it but it still has an insignificant p-value. All three of these variables do not affect our findings about the `distance` variable and therefore we leave them in the model. However, it is pertinent to highlight that the coefficient of `acre` does show unexpected behavior which suggests there are certainly more variables not present in the data set that can be used to explain this anomaly.

The `distance` variable, our primary variable of concern, has an estimated coefficient of -0.05 with a $p-value \quad 0.009 < 0.05$ which leads us to reject the null hypothesis H_0 at the 5% significance level, implying that for a one unit increase in the distance from the nearest rail trail entry, the average price of the house decreases by $0.05\% \pm 0.04\% \quad (p-value = 0.009 < 0.05)$ with 95% probability after taking into account the effect of all the other covariates. However, it is important to not confuse this association for causation since our data is collected from an observational study, not a randomized experiment.

Conclusion

Our study found that homes closer to the rail trail are worth more. For every foot nearer to the trail, a home's value increases by about 0.05%. This means a house 1,000 feet closer could be priced around 5% higher than a similar house farther away. Larger homes also sell for more; every extra 1,000 square feet adds about 38.65% to a home's value.

For Acme Homes, this means building homes near rail trails can increase property values and profits. By adjusting home prices based on proximity to the trail, the company can sell closer homes at higher prices. Marketing the benefits of easy trail access—like walking and biking opportunities—can attract more buyers. Offering larger homes or options to expand can also boost sales. Homes near amenities like good schools, parks, and shops have higher values, so developing in areas with these features or adding community amenities can enhance property appeal.

While our findings are significant, they are based on homes smaller than 0.56 acres in two specific ZIP codes, so results may not apply to larger properties or other areas. Also, since this is an observational study, we cannot confirm that being closer to the rail trail causes higher prices—only that they are related. By applying these insights, Acme Homes can attract more buyers and increase profits by focusing on properties near rail trails and amenities.