Do Homes Located Near a Metro Station Incur a Price Premium? [1]

Evidence from Washington, DC

Dylan L., Yashwant D., Aru G.

Professor Lo
The George Washington University
dylanucko@gwu.edu
arugupta17@gwmail.gwu.edu
yaswanthsaidevisetti@gmail.com

April 2022

[1]We would like to thank Professor Lo for his help in calculating the haversine distances.

Abstract

This paper investigates

Introduction

Theory

Data

To estimate the price premium incurred by consumers buying homes within a half mile radius of a metro station, data are obtained for Washington, DC home sales in 2017. Housing sale data are obtained for single family, multi family, and row- houses and synthesized via kaggle. Metro location data are obtained from Open Source DC.

The dataset includes many home fixed effects, including the number of bathrooms, bedrooms, half bathrooms & floors, the size of the home, the lot acreage, and the type of structure. The price of each home is demarcated in US dollars. Most importantly, each home pairs with a set of latitude and longitude coordinates. The dataset contains 12,399 homes sold in 2017.

The dataset for Metro Locations contains variables indicating the lines running through the station and a pair of latitude and longitude coordinates for all 91 DC Metro stations.

In the modeling, the price of each home is the dependent variable. The parameter of interest is the coefficient on Metro.5, which is a dummy variable indicating whether a house falls within a half mile radius of one of the metro stations. The coefficient on Metro.5 indicates the price differential for a house loathed near a metro station. The other independent variables include the home fixed effects mentioned above. We control for home fixed effects to isolate the effect of living within a half mile radius of the metro station and esure the independent variables are uncorrelated with the error term in the econometric models.

The original dataset contained every home sale in Washington, DC for the years 1980-2019. We first separated the date column (dd-mm-yyyy) into three columns for day, month, and year of the sale. We then used Boolean operators to filter out all the homes sold in 2017.

Performing exploratory data analysis indicated three outliers in the price variable, and those were removed. In the third and fourth econometric model, this paper uses the natural logarithm of home price.

The variable for home structure originally used strings to demarcate the style of home, e.g. single. This paper converts the string names to integer values.

To determine if a given house falls within a half mile radius to a metro station, the distance from each house to the nearest metro station must be calculated. To do so, this paper utilizes the haversine formula and the two pairs of latitude and longitude coordinates for the each house and each metro station. Equation 4 illustrates how this paper calculates the distances.

$$d = 2r\ arcsin\left(\sqrt{sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1)cos(\phi_2)sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1}$$

where ϕ_1 and ϕ_2 correspond to latitude 1 and latitude 2, λ_1 and λ_2 correspond to longitude 1 and longitude 2, and r is the radius of the Earth.

The latitude and longitude coordinates for each house and each metro station are substituted into the haversine function for $(\phi_1,\ \lambda_1)$ & $(\phi_2,\ \lambda_2)$ respectively. The distance from a single house to all 91 metro stations are calculated, then the minimum distance is selected and added to the dataframe.

The haversine formula produces 91 distances for each house. Equation 2 finds the minimum distance for each house given the list of 91 distances.

$$\min\left(distances\ {}_{1}^{91}\right) = Distance \tag{2}$$

Based on the minimum distance found using Equation 2, dummy variables are found using Equation 3.

$$Metro.5 = \begin{cases} 1 & \text{if } Distance \leq 0.50 \\ 0 & \text{if } Distance > 0.50 \end{cases} \tag{3}$$

Now that each house has a dummy variable indicating whether it lies within a half hile radius of a metro station, this paper proceeds with Exploratory Data Analysis.

Exploratory Data Analysis

To demonstrate the usefulness of each variable in the model, an exploratory data analysis of the variable was conducted.

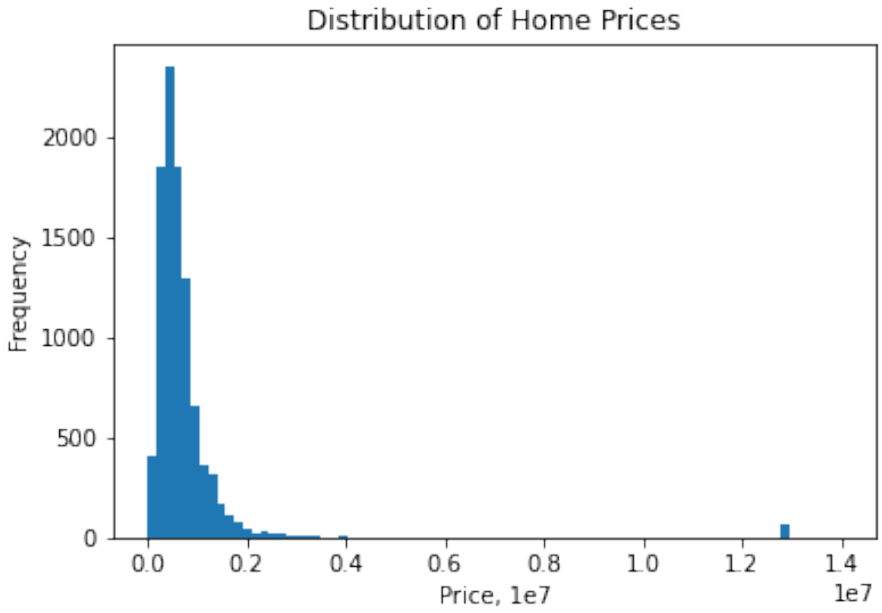

Figure 1: .

First, *price*, the dependent variable in the model was looked into. The histogram for *price* showed peaks in price to be clustered. The histogram also showed a few outliers in the data that should be removed before conducting further analysis.

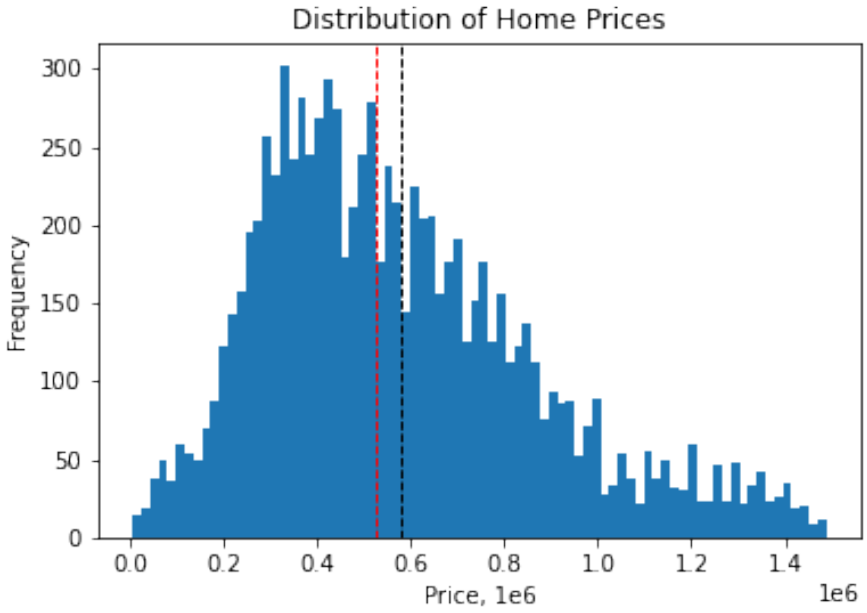

Figure 2: .

After removing the necessary outliers, a new variable for price, newPrice, is created. Even though the histogram for newPrice shows that the average price for homes in the DC area is higher than the median price, the new price data is now more symmetrical than before. The peaks in the newPrice histogram show that most homes range between two-hundred and eight-hundred thousand dollars.

This paper theorizes that the price premium decreases as the distance between a home in Washington D.C. and the nearest metro station increases. The histogram below shows the distribution of the homes by their proximity to the nearest metro station.

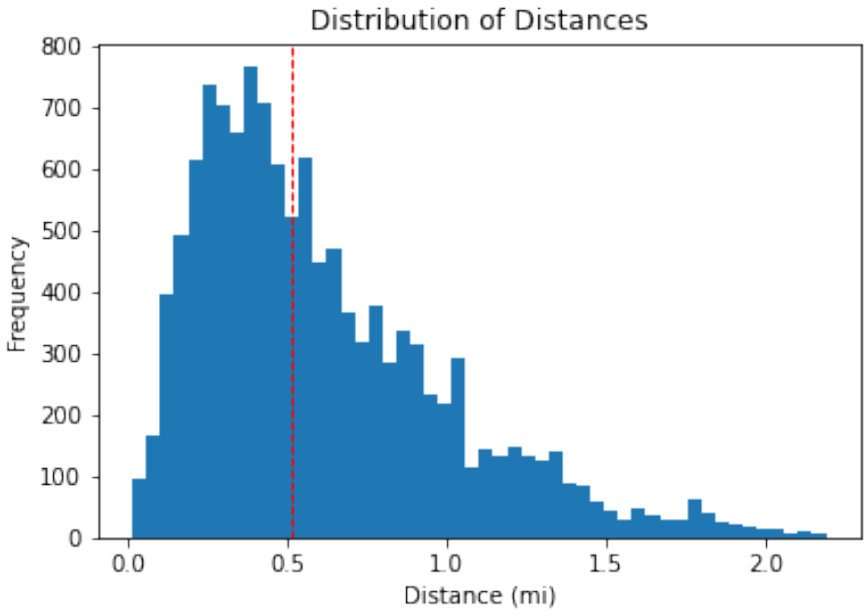

Figure 3: .

As observed in the Distance histogram, median homes are located within a 0.5-mile radius of the nearest metro station. The distribution is skewed to the right, with the distance ranging from zero to over two miles. Given that most homes are clustered within a one-mile radius, a dummy variable is created to see the difference in the price premium for homes within and outside the 0.5-mile radius to conduct further analysis.

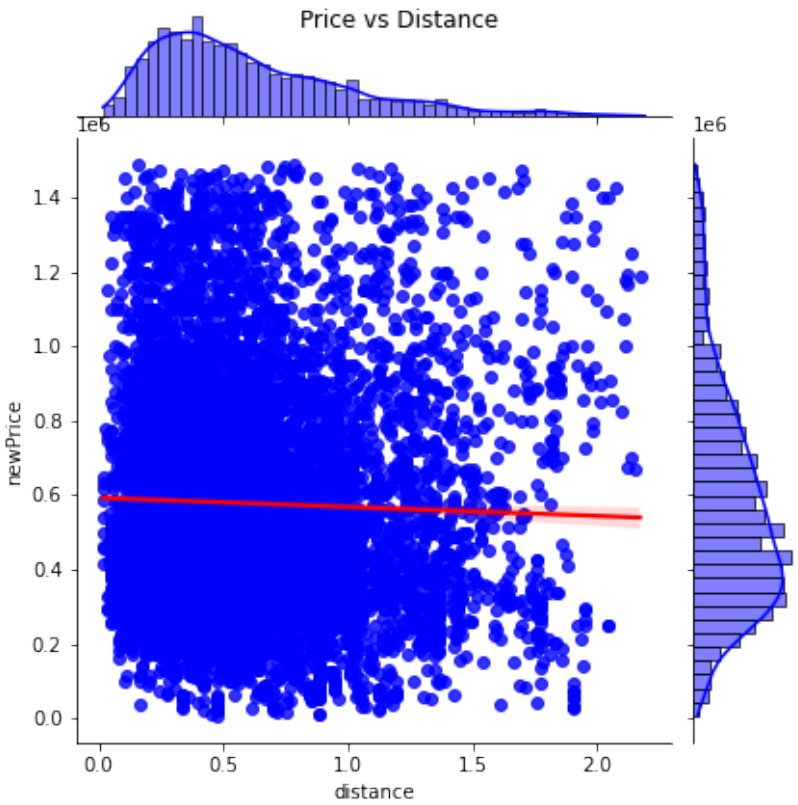

Figure 4: .

The joint plot above shows the scatter plot of price and distance and the distribution of distance on the top, and the distribution of price on the right. The line of best fit in the scatterplot shows a clear downward sloping or negative relationship between the two variables. The downward sloping trend line shows that as the distance increases, the price of the homes decreases.

To take a deeper look into the factors that affect home prices, it is important to take fixed effects and features of the homes into account.

First, we looked at the number of bedrooms in the homes and its relationship with price. The histogram shows the median number of bedrooms in the dataset to be 3. The number

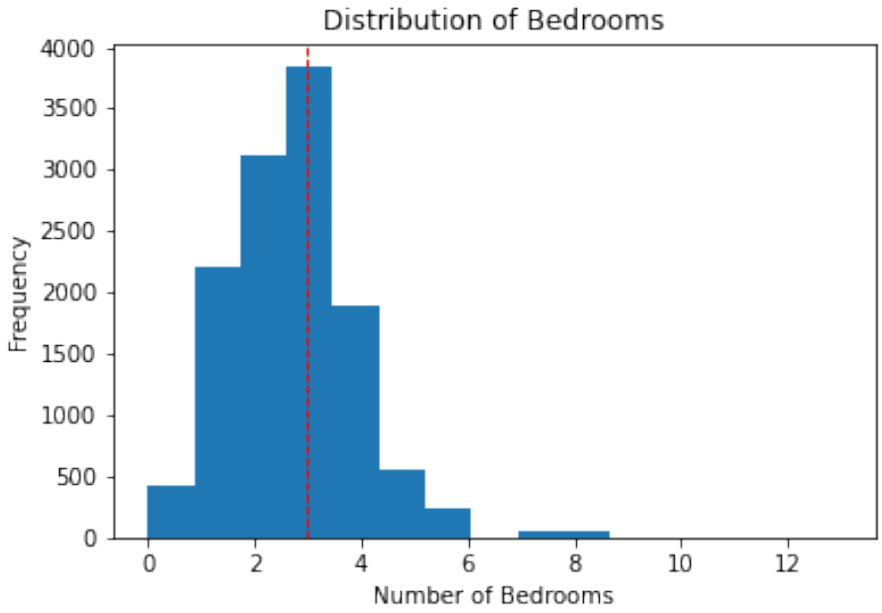

Figure 5: .

of bedrooms ranges from one to six, with eight bedrooms being an outlier.

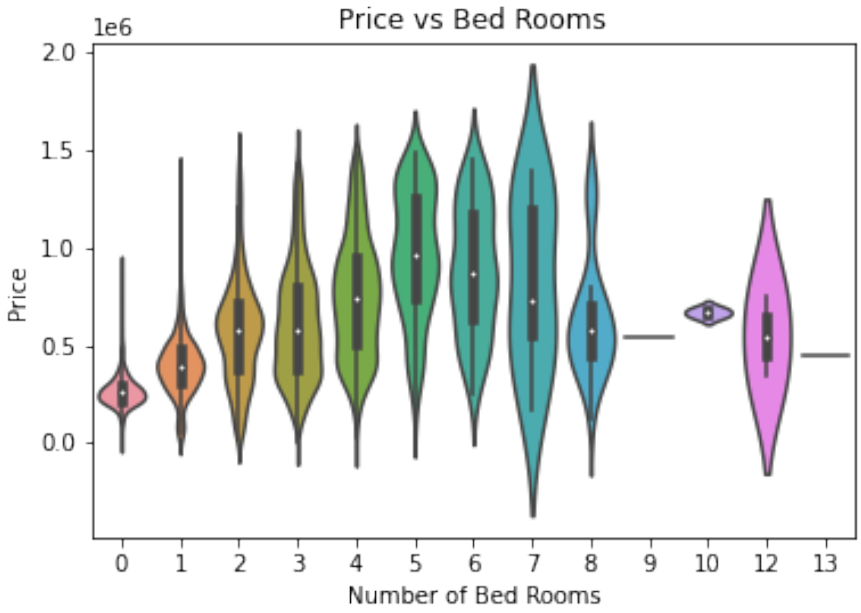

Figure 6: .

When looking at this variable, it was expected that the price of the homes in D.C. would increase as the number of bedrooms increased. On the contrary, the white points in the violin plot show a steady increase in the median price as the number of bedrooms increases to five. At six bedrooms, the median price of homes begins to decline, and the price becomes constant at eight bedrooms.

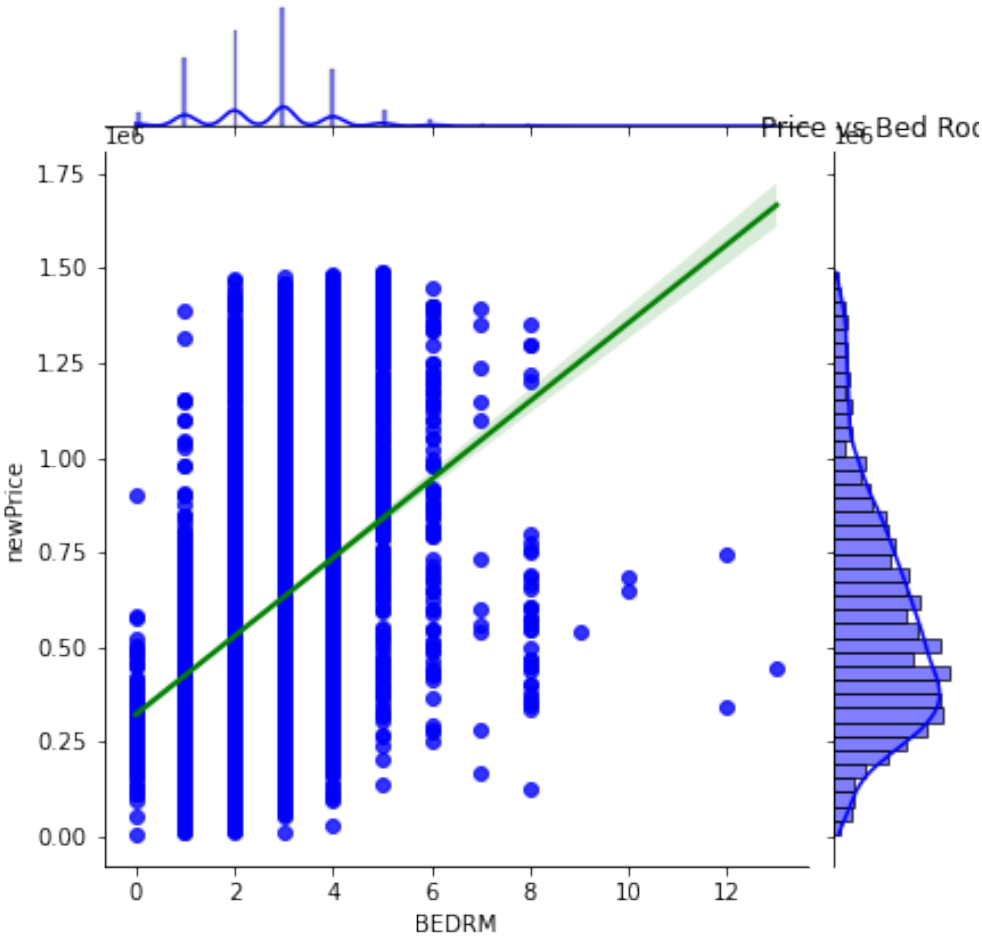

Figure 7: .

The joint plot for bedrooms and price shows the overall upward sloping line of best fit, which indicates a positive relationship between price and the number of bedrooms. It is also observed that the price peaks at around four-hundred thousand dollars in relation to the number of bedrooms.

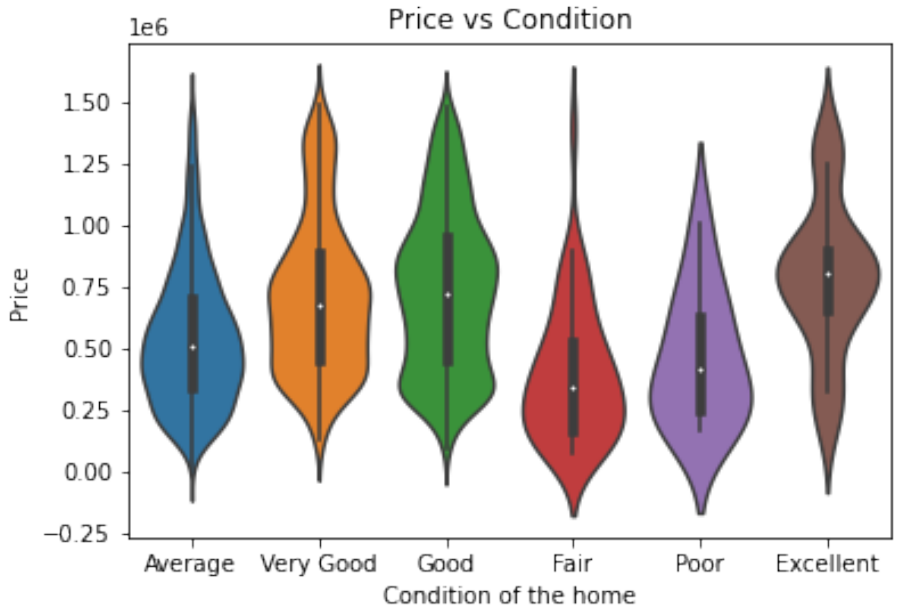
Price vs Condition
1e6
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
Price
Average
Very Good
Good
Fair
Poor
Excellent
Condition of the home

Figure 8: .

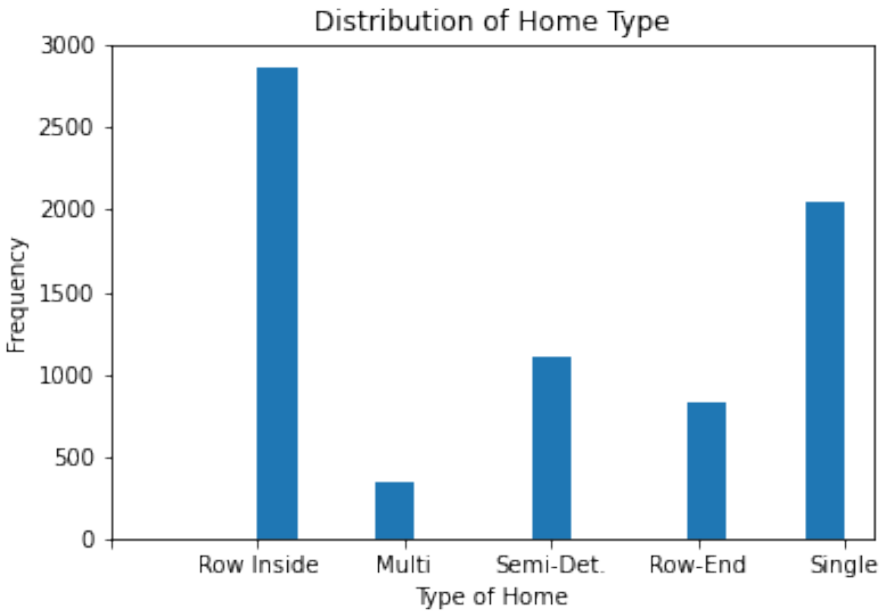
Distribution of Home Type
3000
2500
2000
1500
1000
500
0
Frequency
Row Inside
Multi
Semi-Det.
Row-End
Single
Type of Home

Figure 9: .

Econometric Specification

This paper utilizes a multiple linear regression approach using ordinary least squares estimators to predict if homes located within a half mile radius of a metro station incur a price premium. It is natural for homes to differ in price for several reasons, including size, year built, geographical location, or number of bedrooms. To determine if the metro station proximity is responsible for the differences in price, this paper controls for many home fixed effects mentioned above and in the data section.

This paper estimates four models, the first three utilize the price of a home as the demendent variable, denoted in dollars. The fourth equation estimates the price premium using the log of the home price to account for skew in the home price data. The general specification of the four models take on the functional form of Equation 4.

$$price = \beta_0 + \delta_1 \; metro.5 + \beta_k \; (house \; fixed \; effects) + u \tag{4}$$

where δ_1 is the paramater of interest and denotes the price differential for a house located within a half mild radius and a house located outside the radius. $metro.5$ is a dummy variable denoting whether a house falls within the radius: $metro.5$=1 if the house lies within the half mile radius and $metro.5$=0 if else.

The term house fixed effects encompasses the home attributes that do not change on average, including the number of bathrooms & half-bathrooms, the lot size, the number rooms, the number of kitchens, and the type of structure.

In Model 4, this paper performs a stepwise selection process to determine which of the nine home fixed effects should be included in the final model. The stepwise selection independently estimates *price* as a function of the nine home fixed effects and the $metro.5$ variable for a total of ten regressions. The term with the highest R^2 is added to the model, we call this $variable1$. Next, nine regressions are run, each with $variable1$ and the other eight variables; the term with the highest R^2 is added, we call this $variable2$. This process is repeated, adding the variable that results in the highest R^2 and removing variables that are not statistically significant. Model 4, estimates (log)*price* as a function of the variables chosen by this method.

We expect the sign on $metro.5$ to be positive and moderately large as this paper hypothesizes that homes located within the half mile radius will be more expensive. We also expect the signs on *bedrooms*, *bathrooms*, *size*, *landarea* to all be positive, indicating a higher price given more of these home attributes. We expect some of the structures to have a positive sign and others to have a negative sign, based on the style and quality of the structure. The results of the four regression models can be found in the next section.

To eliminate any multicollinearity issues, this paper examines a correlation matrix to identify independent variables that correlate with one another. Table **??** illustrates the correlation matrix, with high values denoted in red.

Correlation Matrix									
	(log)*price*	Price	struc.	metro50	Stories	Land	Bath	Rooms	Half
(log)*price*	1	0.917	-0.074	0.048	0.170	0.158	0.480	0.357	0.321
Price	0.917	1	-0.040	0.031	0.199	0.205	**0.524**	0.389	0.348
sruc.	-0.074	-0.040	1	-0.146	-0.042	**0.505**	0.136	0.095	0.099
metro50	0.048	0.031	-0.146	1	0.014	-0.177	-0.099	-0.170	-0.1050
Stories	0.170	0.199	-0.042	0.014	1	-0.025	0.033	0.030	0.031
Land	0.158	0.205	**0.505**	-0.177	-0.025	1	0.420	**0.505**	0.306
Bath	0.480	**0.524**	0.136	-0.099	0.033	0.420	1	**0.694**	0.290
Rooms	0.357	0.389	0.095	-0.170	0.030	**0.506**	**0.694**	1	0.396
Half	0.321	0.348	0.099	-0.105	0.031	0.306	0.290	0.3963	1

Table 1: Values above 0.50 are highlighted in red. Note that the correlation between log(price) and Price are not considered as they do not appear in the same model.

Results

This paper utilizes a multiple linear regression approach using ordinary least squares estimators to predict if homes located within a half- mile radius incur a price premium relative to homes outside of the ahlf- mile radius.

This paper estimates four different econometric models, but will only discuss the results of the two consequential models, Model 1 and Model 4. Details about all four models are found in the Econometric Specification Section.

The first prediction framework estimates the price differential for houses inside and outside of the half-mile radius without controlling for the home's fixed effects, denoted by Equation 5.

$$price = \beta_0 + \delta_1 \; metro.5 + u \tag{5}$$

The reported coefficients on $\hat{\delta_1}$ in Column (1) of Table 2 is 24,776.83 and it is statistically significant. This means that a house falling within a half mile radius of a metro station is associated with a $24,778 price premium, all else equal.

Notice, however, that the R^2 reported at the bottom of Column (1) is 0.001. This means that about 1% of the sample variation in price can be explained by the distance dummy variable. Therefore, a more sophisiticated model must include home fixed effects.

The second and third prediction models are included to demonstrate the thought process followed by this paper. We estimate a general linear model first (Model 2) to relax the econometric assumptions in attempt to solve multicollinearity issues. Model 3 uses ordinary least squares and robust standard errors to solve heteroskedasticity issues. The variables included in Model 2 and Model 3 can be found in Column (2) and Column (3) of Table 2. These home fixed effects are chosen rather arbitrarily, and subsequently, result in very strong multicollinearity issues. As a result, Model 4 employs a stepwise selection process to better choose the home fixed effects and therefore reduces multicollinearity issues.

The fourth prediction framework estimates the price differential for houses inside and

outside of the half-mile radius, controlling for the home's fixed effects, denoted by Equation 6.

$$\log(price) = \beta_0 + \delta_1\ metro.5 + \beta_1\ rooms + \beta_2\ rooms^2 + \beta_3\ bathrooms + \beta_4\ HalfBath + \beta_5\ bedrooms + \beta_6\ bed^2 + \beta_6\ structure + u \quad (6)$$

The quadratic terms on *rooms* and *bedrooms* allows for diminishing marginal returns to each of these home attributes. The stepwise selection process and joint F tests for inclusion indicate a stronger R^2 when including the quadratic terms as it corrects the model for better functional form.

The reported coefficients on $\hat{\delta_1}$ in Column (4) of Table 2 is 0.21 and it is statistically significant. This means that a home located within a half-mile of a metro station is associated with a 21% higher price than homes located outside of the radius.

Furthermore, the coefficients on bedrooms and bathrooms are positive while their corresponding quadratic terms are negative. All four terms are statistically significant. This means that adding a bedroom or bathroom to a house increases its price only up to a point. If too many bedrooms or bathrooms are added, the price begins to fall.

Table 2: Effects of Metro Proximity on Home Prices

	Price			Log(*Price*)
	(1)	(2)	(3)	(4)
Intercept	574364.70***	189648.32***	248707.05	17.70***
	[-3424.27]	(19850.63)	(263607.00)	[0.10]
metro50	24776.83***	97631.25***	96892.24***	0.21***
	[8176.48]	(12838.28)	(12821.84)	[0.03]
Stories		63388.06***	64507.54***	
		(8460.41)	(8447.44)	
Land		7.84***	9.04	
		(2.52)	(2.53)	
Bedroom		150960.90***	143963.93***	0.32***
		(4490.45)	(4717.21)	[0.01]
Half		94667.83***	87127.34***	0.24***
		(6886.83)	(7053.45)	([0.02]
Multi		-331169.10***	-309597.28***	-0.78***
		(18996.8)2	(19493.71)	[0.06]
Detached		-190811.19***	-197509.55***	-0.54***
		(11459.7)8	(11522.49)	[0.03]
Row		-52120.28***	-57029.36***	-0.17***
		(12490.2)4	(12507.49)	[0.03]
Single		-77761.28***	-83945.82***	-0.25***
		(13528.3)0	(13562.47)	[0.03]
Rooms				0.16***
				[0.02]
$Rooms^2$				-0.011***
				[0.00]
Bed^2				-0.005***
				[0.00]
Fixed Effects	Yes	Yes	Yes	Yes
Stepwise	No	No	No	Yes
Observations	9176	4825	4825	4834
R^2	0.001	0.381	0.327	0.321
Adj-R^2	0.001		0.326	0.320

Notes. Each column reports results from a regression of dummy variables and other indicators for *Price*. Column (1), Column (2), and Column (3) use the price of homes denoted in US dollars while Column (4) uses the log(*Price*) as the demendent variable. Fixed effects include attributes to the homes that do not vary with time. Model 2 uses a general linear model while Model 1, 3, and 4 use ordinary least squares. The standard errors reported in square brackets are robust standard errors. HC3 covariance methods are used in the ols model in python ($cov_type =' HC3'$). $***p < .01. ** < .05. * p < .10$.

Analysis

Further Research

Conclusion