How does closeness to a university affect housing affordability for students in New York?

Joyce Lin

April 2024

1 Introduction

In mid-February 2024, a student at the University of British Columbia made headlines by commuting to Vancouver from his hometown Calgary every week by plane—all because monthly rent in Vancouver cost *more* than the cost of flying to and from Calgary every week for a month (Turner 2024). The student's predicament is not an exception; more and more students are finding it increasingly difficult to find housing for their studies, and this problem is hardly only limited to Canada.

Coming off of the COVID pandemic, housing prices—and by proxy, rent—have skyrocketed. In the US, off campus rents increased by 28% from 2021 to 2022, from $1,614 USD to $2,062 USD per month (Perry 2023). For college and university students who typically do not have a substantial and stable income, finding housing security close to their campus under such conditions would prove challenging. While flying every week might be atypical for students, many have resorted to other methods to avoid rent such as sleeping in their cars. Undoubtedly, the need for more affordable housing is a top priority among students.

This research paper seeks to analyze the impact of the vicinity of a university in New York state on housing affordability, with the goal of finding the relationship between higher education and housing prices, if it exists. The state of New York has been chosen specifically due to the presence of numerous notable universities like NYU, Cornell University, and Colombia University.

The dependent variable for analysis is house price, and the price is the current listing price unless it has been recently sold, in which case the price is the recently sold price; and the key independent variable is the vicinity to a notable university for a given house. In this paper, the independent variable will be estimated by calculating the distance from the house's ZIP code centroid to the nearest university's coordinates.

On the key independent variable, proximity to a notable university: such a variable had been used in similar works. John A. Maluccio (1998) used distance to the nearest high school as an instrumental variable to estimate the effect of education on wages in the rural Philippines. However, the primary methodology he used was panel data analysis, which is different from the scope of my paper. Bingbing Wang (2023) used the proximity to a university as the independent variable in a difference-in-difference analysis to answer whether the COVID-19 outbreak impacted housing prices for university students in the US as a result of the forcible access to remote learning.

To truly test the effect of distance to a university on housing prices, other covariates could be used for analysis including house size, university rating, student population density, etc. Controlling for other variables, too, has been used frequently by other researchers. Labor economist David Card (2001) studied various models and methods to measure the effect of education on labor market earnings. His accounting for "institutional features" in the education system as exogenous inspired me to collect data on student population, tuition, and estimated average

GPA, among other information on universities in New York. In addition to ”institutional features”, I collected data on public schools to use as additional controls.

Finally, all the data will be combined (as most reasonably possible without losing observations) and run under regressions. The regression method is a popular method for analyzing correlational and causal relationships between variables and has been used in a plethora of research.

Joachim Zietz et al. (2017) performed quantile regression analysis to determine causes for house prices, accounting for various variables including housing characteristics. Sirmans et al. (2015), too, examined the impact of housing characteristics on housing prices critically, utilizing hedonic price models to more accurately capture the nuances of the effects that the characteristics may have on house price.

After combining all data, I found that many observations had missing values for key covariates that I intended to analyze. However, due to the non-uniform and non-linear nature of the relationship between my Y and X variables, a simple average of the missing values will be grossly inaccurate.

An approach to account for less predictable data such as this includes utilizing a method developed by Wei Jiang and the Trauma Group based on the stochastic approximation of the EM algorithm (SAEM) (Jiang et al., 2020; Celeux et al, 1992). Another approach is imputing data based on the k-nearest-neighbors (K-NN) approach, a method that compares similarities between existing data to allow for filling in the missing values with the most similar ”neighboring” values. Other approaches include the one-neighbor approach (1NN) (Beretta 2016). Unfortunately, due to hardware limitations, I found it incredibly difficult to implement these methods.

Instead, I performed limited OLS estimates of my y-variable on existing values of a given covariate and fitted any missing values according to the estimate, logarithmically transforming variables where needed to maintain linearity. This method has the major issue of over-emphasizing the relationship between the price and the covariate in question, which could make my data prone to overfitting. All results from this paper should be considered with that fact in mind.

Analysis reveals that university proximity is indeed significantly correlated with housing price, but so are all other variables such as the number of bedrooms and bathrooms, density of public schools in the area, house and acre size, acceptance rate of the nearest university, etc. No doubt housing price is a complex value that is influenced by a plethora of factors. Without an instrumental variable analysis, it will be difficult to discern the causal effect of university proximity on house prices.

2 Data

The housing prices in New York dataset is 67,157 housing prices in New York collected from realtor.com (2023), broken down by state and zip code. It contains information on the number of bedrooms, bathrooms, house and acre size, ZIP code, and city.

ZIP code information (population count, density) and centroids are obtained from simplemaps.com (2024). Coordinates on 425 universities and colleges in New York are obtained from the National Center for Education Statistics (2024).

The "institutional features" of universities in New York are obtained from simplycollege.com (2024) via HTML web scraping, and includes information on the in-state ranking, estimated average GPA, the number of students enrolled, tuition, and acceptance rates of 256 universities in New York. Despite that this website does not list 400+ universities, it is the most complete website on university data that I could scrape from in a short amount of time with relative ease. The data itself is sourced from the US Department of Education National Center for Education Statistics.

Public school data is collected from Homeland Infrastructure Foundation-Level Data (2019), and includes information on the school's ZIP code.

3 Summary Statistics

Table 1 describes the characteristics of housing data, looking purely at the characteristics of the houses themselves, including house price. In all the fields, the final results show a positively skewed trend. In addition to the fact that all the outliers dropped in the data-cleaning stage were outliers that were significantly greater than the mean, these characteristics line up with the average housing market, where the majority of houses are modest sizes and cater to the middle class, with the opulently wealthy being able to afford dramatically larger and pricier estates.

Table 1: House Characteristics

	bed	bath	acre_lot	house_size	price
count	54284	56711	43363	43265	65438
mean	3.14	2.28	4.45	1871.43	933430.67
std	1.59	1.28	16.43	1118.24	1444098.62
min	1	1	0	4.00	0.00
25%	2	1	0.10	1089.00	229000.00
50%	3	2	0.26	1600.00	519000.00
75%	4	3	1.20	2340.00	975000.00
max	10	10	200.00	7500.00	15000000.00

Table 2 describes ZIP code level characteristics, particularly concerning population and public school data. While the population count itself exhibits a normal distribution (the standard deviation is smaller than the mean, which is close to the median value), the population density shows a positively skewed trend. This indicates that a select few ZIP code areas are highly densely populated. Additionally, the number of public schools per ZIP code as well as the density of public schools per ZIP code are also positively skewed. This indicates that a majority of ZIP codes are less densely populated and have fewer public schools.

Table 2: ZIP Codes and Public School Density Characteristics

	population	density	schools count	schools_density
count	65438	65438	65438	65411
mean	34353.53	9272.93	7.11	1.73
std	27019.38	12903.76	6.64	2.87
min	0	0	0	0
25%	11395	179.50	2	0.03
50%	29461	2257.70	5	0.37
75%	51153	15325.20	9	2.25
max	112750	60879.20	46	19.50

Table 3 describes university-level characteristics of 256 universities in New York. Of all of the 256 universities, only 91 of them are ranked, and only 136 have estimated average GPAs. It can be reasonably inferred that universities without rankings or estimated average GPAs may not perform as well academically or have the same level of recognition as those that are ranked or have GPA information available. All 256 universities have enrollment information and acceptance rate, and all but 20 universities have tuition listed.

Table 3: University Characteristics

	unidist	Ranking	Enrollment	Tuition	Acceptance	GPA
count	65438	9537	27605	26841	27605	16485
mean	5.17	42.92	4407.74	19653.56	0.76	3.23
std	7.34	24.59	7125.98	12051.49	0.26	0.43
min	0.01	1	38	204	0.07	2.10
25%	0.66	22	657	8204	0.64	3.00
50%	1.85	41	1949	17238	0.80	3.30
75%	6.49	66	4807	29499	1.00	3.55
max	52.72	88	59144	48847	1.00	4.00

4 Visualization

Based on Figure 1 below, it's apparent that proximity to a university does correlate negatively with housing prices. After applying a logarithmic transformation reveals an interesting trend when houses are extremely close to a university: when a house is closer to a university, the lower limit of price variation increases. This suggests that closeness to a university may exert a distinct influence on housing prices, warranting further investigation into the underlying factors driving this relationship.

In addition, houses appear to be significantly more concentrated in areas closer to universities, as evidenced by the density of housing units on the left side of the graphs compared to the right side.

Finally, the data exhibits a heteroskedastic variability in house prices as the distance to a university decreases. Perhaps this higher variation is dependent on other factors, such as the quality of the university that a house is close to, which can exacerbate the impact on house prices.

Figure 1: Relationship between House Price and University Proximity

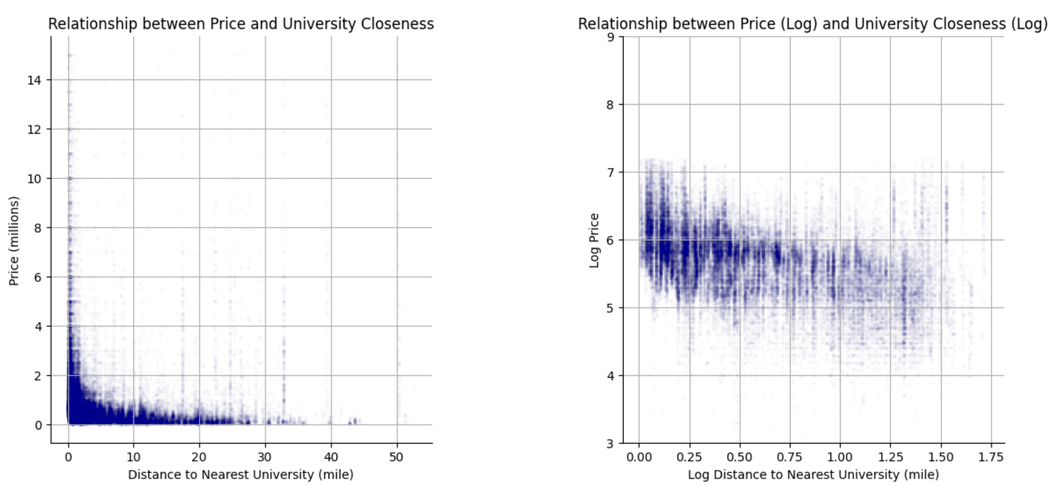
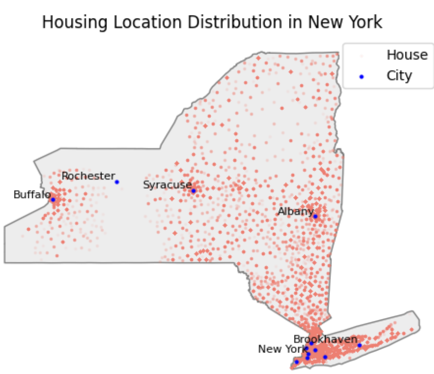


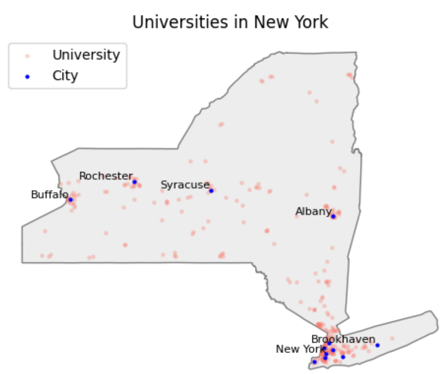
Figure 2 depicts the geographical distribution of all 65,438 houses by ZIP code centroids, as well as 425 universities in New York. The twelve largest cities of New York (New York, Brooklyn, Queens, Manhattan, Bronx, Buffalo, Hempstead, Rochester, Albany, Staten Island, Brookhaven, Syracuse) are also plotted for reference.

On the apparent mismatch of state borders between the zip code and the state itself: Perhaps due to the difference in sources, the resulting map below does not appear to match up 1-to-1 in the top left edge of the map. Specifically, the state border of New York depicted above is the political border, which extends beyond the actual land mass of New York itself. In the northwest area, the border cuts across the Great Lakes, while in the southeast the border circles around Manhattan Island. Since zip codes do not exist over the water, the soft border formed by zip codes will not correspond to the displayed state border of New York.

Figure 2: Geographical Distribution of Houses and Universities



(a) Figure 2-a



(b) Figure 2-b

In Figure 2-a, the houses are plotted by ZIP code centroids, which are individually set at a low opacity, such that zip codes with comparatively more houses will appear darker than zip codes with fewer houses. The majority of houses (as well as cities, for that matter) appear to cluster around New York City/Manhattan Island. Additionally, there are three other smaller clusters in the mid-New York region, around Buffalo, Syracuse, and Albany. The tendency for houses to cluster around cities makes sense, as cities are the result of urbanization, which tends to beget more houses.

In Figure 2-b, the universities are plotted by their direct coordinates. There is considerable concentration on and around New York City and Manhattan Island. This trend is unsurprising considering New York City's status as the largest and most influential American metropolis. Outside of New York City, there are a few more clusters of universities present in other cities in New York Buffalo, Rochester, Syracuse, and Albany.

5 Regression Results

To recap, the dependent variable is housing price, and the independent variable as denoted by my research question is proximity to the nearest university. I have also included several other potential X variables, such as the number of bedrooms and bathrooms, the number of schools in a given house's ZIP code, and the estimated GPA of the nearest university to a house.

Analysis in previous parts leads me to believe that the economic relationship between my Y and X variables is **linear, provided that both the Y and X variables are logged**. Essentially, the relationship is elasticity, where a percent increase in X leads to a percent increase/decrease in Y. This applies to the relationship between house price and university proximity, as well as some other covariates. Thus, the final regression for this paper will feature a logarithmic transformation on house prices and the majority of variables, including university proximity.

In this section, I ran into two issues: the first issue is that all covariates have missing values. Due to the non-uniform and non-linear nature of the relationship between my Y and X variables, a simple linear estimate of the missing values will be grossly inaccurate. As the majority of relationships in my data are logarithmic, I will attempt to predict missing values using logged linear estimation. Since not all relationships are logarithmic, this method is overly broad and makes the data prone to overfitting.

However, due to hardware limitations, I was unable to implement SAEM or KNN methods for a more accurate imputation of missing values. I had to predict the missing values using a logged linear estimation, which will invariably skew the accuracy of the data.

The second issue is that 125 names out of the 256 universities in the webscraped dataset do not match the university names in my working dataset. Because there is no way to automatically check and rename potentially mismatched names in the two datasets, it follows that I have to manually check all 125 mismatched universities. Due to resource constraints, I find it more productive to exclude university information in my final analysis, as including it would result in losing 55,901 observations. That said, I will still perform regression analysis including university covariates for comparison.

The final regression model for this paper is as follows:

$$
\begin{aligned}
\log(\hat{\text{price}}) = {} & \hat{\beta_0} + \hat{\beta_1}\text{unidist} + \hat{\beta_2}\text{bed} + \hat{\beta_3}\text{bath} + \hat{\beta_4}\text{bed} \cdot \text{bath} \\
& + \hat{\beta_5}\log(\text{acre\_lot}) + \hat{\beta_6}\log(\text{house\_size}) + \hat{\beta_7}\text{bigcity} \\
& + \hat{\beta_8}\log(\text{population}) + \hat{\beta_9}\log(\text{density}) + \hat{\beta_{10}}\log(\text{schools\_density}) \\
& + \hat{\beta_{11}}\text{bigcity} \cdot \log(\text{density}) + \hat{\beta_{12}}\log(\text{density}) \cdot \log(\text{schools\_density}) \qquad (1)
\end{aligned}
$$

where bigcity is a dummy variable that denotes whether a house in question is in the top 12 most populated cities in New York as of January 30, 2024: New York, Brooklyn, Queens, Manhattan, Bronx, Buffalo, Hempstead, Rochester, Albany, Staten Island, Brookhaven, and Syracuse. These 12 cities were chosen because houses in the dataset were shown clustering around these locations.

Interaction terms between bedrooms and bathrooms, bigcity and population density, and population density and school count density were added. For the first term, the number of bedrooms and bathrooms in a house are likely to be correlated. For the second and third terms, big cities should be positively correlated with population density, and a more densely populated area is also more likely to have a higher density of public schools to accommodate.

The regression model that includes university covariates is as follows:

$$
\begin{aligned}
\log(\hat{\text{price}}) = {} & \hat{\beta_0} + \hat{\beta_1}\text{unidist} + \hat{\beta_2}\text{bed} + \hat{\beta_3}\text{bath} + \hat{\beta_4}\text{bed} \cdot \text{bath} \\
& + \hat{\beta_5}\log(\text{acre\_lot}) + \hat{\beta_6}\log(\text{house\_size}) + \hat{\beta_7}\text{bigcity} \\
& + \hat{\beta_8}\log(\text{population}) + \hat{\beta_9}\log(\text{density}) + \hat{\beta_{10}}\log(\text{schools\_density}) \\
& + \hat{\beta_{11}}\text{bigcity} \cdot \log(\text{density}) + \hat{\beta_{12}}\log(\text{density}) \cdot \log(\text{schools\_density}) \\
& + \hat{\beta_{13}}\text{acceptance} \cdot 100 + \hat{\beta_{14}}\text{gpa} + \hat{\beta_{15}}\log(\text{enrollment}) \\
& + \hat{\beta_{16}}\text{tuition}/1000 + \hat{\beta_{17}}(\text{acceptance} \cdot 100) \cdot \text{gpa} \\
& + \hat{\beta_{18}}\text{gpa} \cdot \text{tuition}/1000 \qquad (2)
\end{aligned}
$$

Ranking is not included due to the assumption that ranking is determined by enrollment, GPA, acceptance rate, etc. Two interaction variables have been included: acceptance rate and GPA, GPA and tuition, and enrollment and tuition. The first interaction term was chosen because of the idea that a more selective university (lower ‘acceptance‘) typically takes on higher performing students, leading to a higher average GPA at the institution; the second because higher tuition could lead to better educational support, which correlates with higher GPA.

	Dependent variable: price_log	
	(1)	(2)
unidist_log	0.097***	-0.030**
	(0.006)	(0.014)
bed	0.027***	0.059***
	(0.002)	(0.004)
bath	0.344***	0.354***
	(0.002)	(0.006)
bed*bath	-0.029***	-0.035***
	(0.000)	(0.001)
acre_lot_log	0.140***	0.133***
	(0.004)	(0.009)
house_size_log	0.376***	0.398***
	(0.008)	(0.020)
bigcity	-0.900***	-1.100***
	(0.024)	(0.048)
population_log	-0.142***	-0.064***
	(0.003)	(0.009)
density_log	0.325***	0.302***
	(0.003)	(0.007)
schools_density_log	0.341***	-0.401***
	(0.052)	(0.120)
bigcity*density_log	0.207***	0.251***
	(0.006)	(0.012)
density_log*schools_density_log	-0.038***	0.073***
	(0.011)	(0.025)
Acceptance_Percent		0.016***
		(0.003)
Estimated Avg GPA		0.363***
		(0.062)
Enrollment Log		0.091***
		(0.009)
Tuition_1000		0.037***
		(0.006)
acceptance*gpa		-0.005***
		(0.001)
gpa*tuition		-0.010***
		(0.002)
const	3.324***	1.372***
	(0.028)	(0.237)
Observations	65438	9537
R^2	0.697	0.749
Adjusted R^2	0.697	0.748
Residual Std. Error	0.288 (df=65425)	0.260 (df=9518)
F Statistic	12547.446*** (df=12; 65425)	1576.722*** (df=18; 9518)
Note:		*p<0.1; **p<0.05; ***p<0.01

Model 1 analysis: According to the regression above, all variables and interaction terms are very statistically significant (1% level). Many variables unsurprisingly raise the house price: for example, the number of bedrooms and bathrooms, house size, and population density. Surprisingly, a house that is in one of the twelve largest cities in New York is expected to decrease its price. This implies that factors associated with less populated areas make a house more valuable on the market.

The fit of Model 1 is certainly much better than the fit of the simple regression done at the beginning of this part. The R^2 sits at 0.696, compared to 0.18 previously. However, the F-statistic is *massive* at over 12 thousand.

Model 2 analysis: Similar to option 1, all variables and interactions are statistically significant at the 1% level. The existence of university-level data strengthens the magnitude of housing variables and reduces the impact of university proximity from -0.087% to -0.061%. Additionally, the presence of university-level data has changed the effect of public school density to a negative impact.

The R^2 of Model 2 improves to 0.751, with the adjusted R^2 reducing by only 0.01. If we were to ignore the loss of observations, option 2 would look cleaner. Ultimately, however, I will be choosing option 1 as my final model.

Both models showcasing such a high degree of significance may indicate multicollinearity between the covariates used to estimate house prices. Multicollinearity occurs when the chosen covariates are closely correlated with each other, potentially also having a causal relationship. A solution for this issue will be presented in the conclusion.

Closing out of this section, I should remind you of the caveat of the effect of the method I used to predict missing values, which is the fact that it overly strengthens a logged linear relationship across the board that the dataset itself may not have. For some of the variables it makes economic sense for them to influence price significantly (e.g. count of bedrooms and bathrooms), but the fact that every single coefficient is highly statistically significant may be proof of overfitting or bias.

6 Random Forest Results

In order to understand a random forest model, the regression tree must first be comprehended.

The regression tree is a machine learning algorithm that predicts the best-fit equation by piecemeal optimization. That is, given an equation for dataset N with n-variables, the regression tree uses machine learning to choose the most impactful variable to split the data in N on, then choose the next most impactful variable to split on, and so on. The most impactful variable is chosen by minimizing the mean squared error at the given step.

For the equation for my dataset,

$$\begin{aligned}\log(\text{price}) = {} & \beta_0 + \beta_1\text{unidist} + \beta_2\text{bed} + \beta_3\text{bath} + \beta_4\text{bed}\cdot\text{bath} \\ & + \beta_5\log(\text{acre\_lot}) + \beta_6\log(\text{house\_size}) + \beta_7\text{bigcity} \\ & + \beta_8\log(\text{population}) + \beta_9\log(\text{density}) + \beta_{10}\log(\text{schools\_density}) \\ & + \beta_{11}\text{bigcity}\cdot\log(\text{density}) + \beta_{12}\log(\text{density})\cdot\log(\text{schools\_density})\end{aligned}$$

The regression tree will choose the covariate that solves the objective function at step k,

$$\min\left[\text{MSE}_k\right] = \min\left[\frac{1}{n}\sum_{i=1}^{n}(\log(\text{price})_i - \log(\hat{\text{price}})_{i,k})^2\right]$$

where y_i is the actual value in the dataset and $\hat{y_{i,k}}$ is the estimated result of the regression equation at step k, with the chosen covariate in the equation.

For example, in step 1, if the covariate ‘bath‘ is found to minimize the mean squared error the most, the regression tree will split the data on ‘bath‘ at a particular value into two sections.

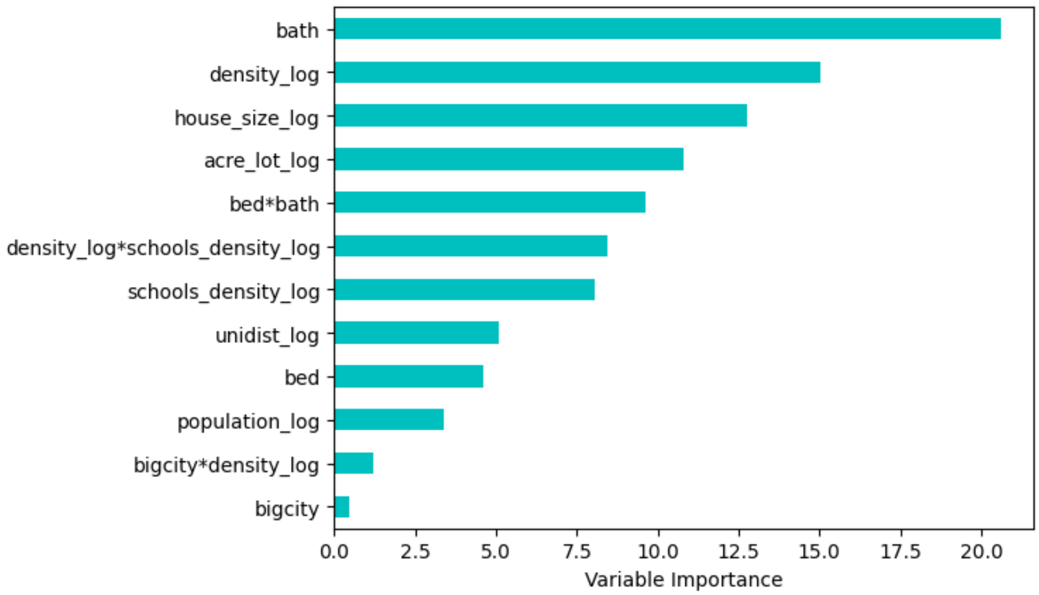
In step 2, the regression tree will look at the two sections separately. In section A, if the covariate ‘log unidist‘ minimizes MSE the most, the tree will split section A further on that variable. In section B, if ‘bath‘ is the covariate that minimizes MSE, the tree will split section B on that variable. Notice that the sections can have different covariates of choice and that covariates used in previous branches can be reused.

In step 3, the regression tree will repeat step 2 but with more sections. The number of steps (a.k.a. depth) that the regression tree will take is dependent on the researcher’s needs or hardware limits (the calculation time increases exponentially).

A random forest model is a culmination of multiple regression trees (or decision trees) that are uncorrelated with each other. How are they uncorrelated with each other? In a random forest model, each decision tree will, at each step, only be able to split based on a *j* number of covariates, of which the selection is random. In doing so, the different regression trees are effectively made uncorrelated with each other, which would solve the issue of bias and yield a more accurate analysis.

Random forest models are particularly useful in analyzing datasets with non-linear relationships (Biggs et al., 2022), which is what my dataset is. The importance matrix, which displays the most impactful variables as calculated by the random forest model, is displayed on the next page (Figure 3).

Figure 3: Importance Matrix for Model 1



The random forest setup used here limited the number of random variables to three. The results showed the number of bathrooms to be the most important covariate, with the ZIP code density to be the next most important. University proximity ranked behind public schools density and all the housing characteristics except ‘bed‘. The implication is that, while proximity to a university influences house prices, the density of public schools has a greater influence.

7 Conclusion

For the UBC student who resorted to commuting by plane to school over renting a place near campus, his lifestyle can hardly be considered ideal for learning. Yet, with housing prices on the rise along with living costs in general, we can expect to see more students forgoing living close to campus to save money. In her paper, Bingbing Wang (2023) mentioned the bid rent theory, which states that distance negatively impacts demand, meaning that properties that are closer to a particular hotspot will see a price markup. Logically, the bid rent theory also applies to universities, as more students are expected to desire a place closer to their campus. This relationship is what this paper aims to analyze, with a close focus on New York due to the state itself not only being a hotspot for universities but also the home of one of the most well-developed metropolises in the world.

Regression results found that all variables as well as the interaction terms had a statistically significant impact on house price, which suggests not only correlation with house price but also inter-relation with each other. The regression does not show significance beyond the 1% level, however, so there was no way to determine the most influential covariates.

However, as mentioned in the Regression Results section, the fact that all the covariates analyzed show high significance may indicate multicollinearity. The supposed significant effect on housing prices may be overly exaggerated as a result. A common solution to address multicollinearity is the principal component analysis, a technique that reduces the total covariates to a smaller, more significant subset of covariates.

In this paper, I had instead used the random forest analysis to dissect the covariates that have the most influence over housing prices. The analysis presents that the number of bathrooms holds the most influence out of all covariates on house price, followed by the population density within the house's ZIP code. The more dense the population likely is, the higher the demand for houses in that area. Behind housing characteristics, university proximity has been calculated to be the next most influential variable.

Conclusions about causal relationships remain inconclusive, however. Ideally, an instrumental variable on university proximity should be done to capture the causal effect. However, an instrumental variable could not be obtained for this paper. Potential instrumental variables included ”distance to nearest public transportation” or ”distance to the nearest grocery store or shopping center,” but this data **has** to be calculated at a more precise level than the ZIP code level. I would have to know the exact house coordinates of every house, not the house's ZIP code centroid, as well as the exact locations of all public transport or amenities. If I were to obtain this data, the next step would be to perform an instrumental variable analysis to determine the causal effect of university proximity on house prices.

8 References

- Aguayo, C. USA Public Schools [Data set]. Kaggle. https://www.kaggle.com/datasets/carlosaguayo/usa-public-schools/data

- Beretta, L., Santaniello, A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak* 16 (Suppl 3), 74 (2016). https://doi.org/10.1186/s12911-016-0318-z

- Biggs, M., Hariss, R., & Perakis, G. (2023). Constrained optimization of objective functions determined from random forests. *Production and Operations Management*, 32(2), 397-415. doi:10.1111/poms.13877

- Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. Econometrica, 69(5), 1127–1160. https://doi.org/10.1111/1468-0262.00237

- Celeux, G., & Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastic Reports*, 41(1–2), 119–134. https://doi.org/10.1080/17442509208833797

- CollegeSimply. (2024). 2024 best colleges in New York. https://www.collegesimply.com/colleges/new-york/
- Eastern Suffolk BOCES. LinkedIn. (2024). https://www.linkedin.com/school/eastern-suffolk-boces/about/
- Jiang, W., Josse, J., & Lavielle, M. (2020). Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145, 106907. doi:10.1016/j.csda.2019.106907
- Maluccio, J. A. (1998). Endogeneity of schooling in the wage function. IDEAS. https://ideas.repec.org/p/fpr/fcnddp/54.html
- National Center for Education Statistics. IPEDS Data Center: College Navigator - University Latitude and Longitude Data [Data set]. Retrieved from https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?gotoReportId=7&sid=2b85ae93-b78a-423f-933c-cd095ac80c39&rtid=7
- Perry, A. (2023, August 28). No room at the dorm: As college begins, some students are scrambling for housing. Forbes. https://www.forbes.com/sites/alexperry/2023/08/20/no-room-at-the-dorm-as-college-begins-some-students-are-scrambling-for-housing/?sh=439fa93631f5
- Sakib, A. S. USA Real Estate Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset
- SimpleMaps. United States Zip Codes Database [Data set]. Retrieved from https://simplemaps.com/data/us-zips
- Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The Composition

of Hedonic Pricing Models. Journal of Real Estate Literature, 13(1), 3–43. http://www.jstor.org/stable/44103506

- Turner, A. (2024, February 7). UBC student commutes from Calgary – cheaper than paying Vancouver Rent. British Columbia. https://bc.ctvnews.ca/ubc-student-commutes-from-calgary-cheaper-than-paying-vancouver-rent-1.6759116

- U.S. Census Bureau. TIGER/Line Shapefiles, 2023 [Data set]. Retrieved from https://www2.census.gov/geo/tiger/TIGER2023/STATE/

- Wang, B. (2023). The Effect of Proximity to Universities on House Prices after the COVID Outbreak. IDEAS. https://ideas.repec.org/a/gam/jjrfmx/v16y2023i3p167-d1085315.html

- What is Random Forest?. IBM. (n.d.). https://www.ibm.com/topics/random-forest#: :text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

- World Cities Database. simplemaps. (n.d.). https://simplemaps.com/data/world-cities

- Zietz, J., Zietz, E.N. & Sirmans, G.S. Determinants of House Prices: A Quantile Regression Approach. J Real Estate Finance Econ 37, 317–333 (2008). https://doi.org/10.1007/s11146-007-9053-7