Predicting Apartment Prices in New York City

W203 Fall 2022 - Regression Model

Tim Tung, Mamesa El, Chienhung Yeh, Julian Rippert

Contents

Introduction

Given the growing population in the United States in conjunction with a trend toward urbanization, providing sufficient housing – and the right type of housing – continues to be a challenge. Demographic shifts and the changing structure of the family also influence the type of housing stock that is desired in the marketplace. For example, data shows that the share of women in their 20s who live with a parent increased to 46% in 2019 from 36% in 2007[1]. Similarly, about 51% of men in their 20s live with a parent.

For existing homeowners, they may consider whether it is worthwhile to add a bedroom to their existing home or remodel their home to squeeze an additional bedroom within the existing livable space. Real estate developers interested in constructing new housing developments may wish to understand the value of an additional bedroom since this may affect the mix of units in a proposed development. For example, given a finite amount of developable space, a project sponsor may want to know whether it is more economically advantageous to construct a housing development with 50 two-bedroom apartments or 35 three-bedroom apartments.

This study estimates apartment prices based on real estate data for New York City in 2021. Applying a set of regression models, we estimate apartment prices with a key driver being the number of bedrooms.

Data and Methodology

The data in this study comes from realtor.com, the second most visited real estate listings website in the United States with over 100 million monthly active users. The data was compiled and made publicly available by Ahmed Shahriar Sakib[2] and each row represents a property in the United States. We performed all exploration and model building on a 50% subsample of the data and used the remaining 50% of the data to generate the statistics presented in this report.

We filtered the dataset to properties within New York City as our geographic area of focus. We also limited the date range to data points that fell within calendar year 2021 to address potential concerns regarding temporal independence. We specifically did not include more recent samples that fell within calendar year 2022 because we believe that the interest rate environment change that commenced at the beginning of 2022 made those samples not comparable to the samples from 2021.

We removed samples that had null values for the bed or house_size variables, and also removed duplicate entries based on the street address. To focus on one category of housing product, we narrowed our samples to apartment units and excluded other housing types (e.g. single-family homes). We did not view apartments and single-family homes as comparable in part because single-family homes have exclusive use of land and apartments often have homeowners' association fees that single-family home buyers do not pay. We constructed a box plot for the price variable and noted six extreme outliers: five that were high outliers and one that was a low outlier. We removed this small number of outliers from the dataset to reduce distortion in the linear model. A full accounting of the sample exclusions is detailed in Table 1.

Finally, we also noted that for some property addresses, the city was listed as either "New York" or "New York City" rather than the borough that the property falls within. Given the meaningful differences among the boroughs of New York, we transformed the zip code variable and constructed a new variable to accurately code the borough for each property. The borough variable is used in two of the regression models as an indicator variable.

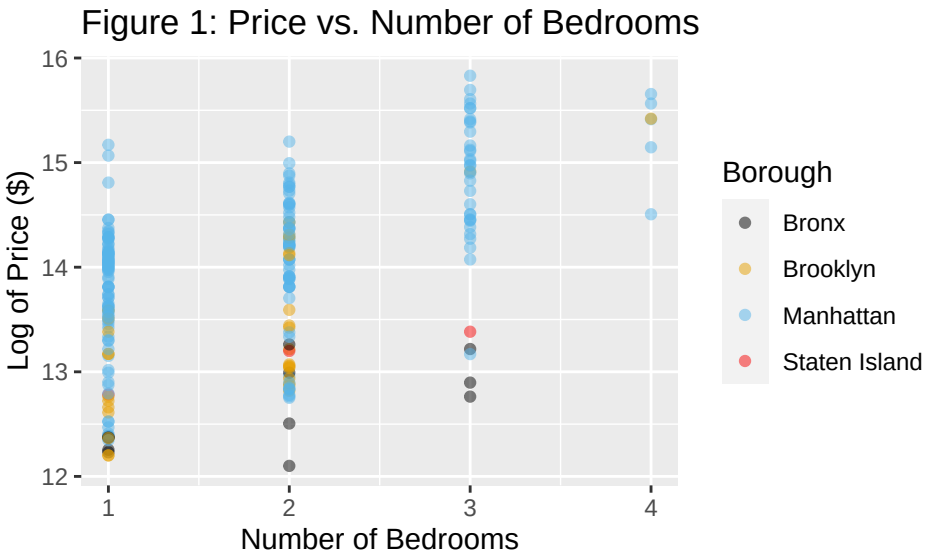
[1] https://www.ppic.org/blog/californias-new-baby-bust/
[2] https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Table 1: Sample Accounting

Cause	Number of samples available for analysis (after removal for cause)	Remove number of samples
Start	923,159	
Limit state to New York	211,778	711,381
Limit city to New York City and boroughs	146,728	65,050
Limit sold_date to dates within calendar year 2021	3,120	143,608
Drop samples with null values for bed or house_size	1,777	1,343
Remove duplicates based on street address	318	1,459
Remove samples that are not apartments	237	81
Remove outliers	231	6

Upon examining a histogram of the price variable, we noted the data is very right-skewed. To address the skewness of the variable, we transformed the price variable by applying a logarithm to create a more symmetrical distribution. Figure 1 is a scatter plot of the number of bedrooms on the x-axis and the logarithm of price on the y-axis . The observations in the scatter plot are color-coded based on location using our borough indicator variable. The ocular test leads us to believe that there is a positive relationship between the two variables. Additionally, we observe that across the range of the number of bedrooms, the Manhattan observations exhibit a price premium over other boroughs.

The covariates in our models should not be able to explain each other through linear combinations. To ensure the models do not include variables with perfect or very high collinearity, we tested the collinearity between the number of bedrooms and the number of bathrooms. The VIF test returned a value of 2.39, which we interpret to indicate only a moderate correlation between these variables that is not severe enough for us to consider dropping either variable.



$$\widehat{ln(price)} = \beta_0 + \beta_1 \cdot Beds + \beta_2 \cdot Bath + \beta_3 \cdot Manhattan + \beta_4 \cdot Bronx + \beta_5 \cdot StatenIsland$$

From the data cleaning and exploratory data analysis described above, we specified the final model as shown above. (Note: Brooklyn is the base location in our model. The borough indicator variable for Queens was dropped based on lack of data.)

Results

Across the models, all variables are statistically significant at $p<0.01$, with the exception of the bathroom variable in the second model. Compared to the first model, the bedroom coefficient in the second model is substantially reduced and no longer statistically significant. We observe that the bathroom variable in the second model has instead taken on almost all of the predictive power that was previously associated with the bedroom variable in the first model. We believe this is a consequence of the moderate collinearity between the bedroom and bathroom variables; however, we decided to retain the bathroom variable because the VIF value was 2.39 (less than 5). After also including the borough indicator variables, we note that the r-squared value substantially improves to 0.77 for the fourth model from 0.44 for the second model.

Given that the model is specified on a log-level basis (with log base e), each increase in the number of bedrooms has a multiplicative effect on price. Applying the fourth model, if the property is located in Manhattan and has three bedrooms and three bathrooms, the predicted price is \$2,730,512. If that same property has one additional bedroom, the predicted price increases by 25%, or \$671,915, to \$3,402,428, ceteris paribus.

Table 2: Estimated Regression

	Output Variable: Apartment Price			
	(1)	(2)	(3)	(4)
Bedrooms	0.52^{***}	0.09	0.50^{***}	0.22^{**}
	(0.06)	(0.10)	(0.05)	(0.08)
Bathrooms		0.49^{***}		0.32^{***}
		(0.08)		(0.06)
Manhattan			0.72^{***}	0.63^{***}
			(0.14)	(0.13)
Bronx			-0.75^{***}	-0.76^{***}
			(0.14)	(0.13)
Staten Island			-0.53^{**}	-0.47^{***}
			(0.18)	(0.14)
Brooklyn (Base)				
Constant	12.95^{***}	12.84^{***}	12.57^{***}	12.57^{***}
	(0.13)	(0.13)	(0.15)	(0.14)
Observations	115	115	115	115
R^2	0.32	0.44	0.72	0.77
Adjusted R^2	0.32	0.43	0.71	0.76
Residual Std. Error	0.68 (df = 113)	0.62 (df = 112)	0.44 (df = 110)	0.40 (df = 109)
Note:				$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Limitations

During the course of the regression analysis, we noted that there may be challenges to the IID requirement for our data from the geographic and temporal dimensions. From the geographic dimension, our data may experience clustering as we have used borough indicator variables as a proxy for location. Perhaps properties within the same borough will provide some information regarding the price of another property in the same cluster.

Additionally, given how real estate values are largely driven by comparable sales; there may be temporal challenges to the independence of our data. As such, we have attempted to limit any variations in the temporal dimension by limiting our data set to one year (calendar year 2021). Mortgage rates in 2021 ranged from 2.74% to 3.10%, providing a relatively consistent window, compared to 2022 during which mortgage rates have increased substantially over the course of the year. As such, property prices within the same year may provide information regarding what the price of a property may be in the future.

To meet the requirements of a true model, there should be no omitted variables influencing our outcome variable. Although we used boroughs as a proxy for location, within boroughs there are other variables that factor into the price such as access to schools and transportation. We believe that properties located further away from schools and transportation would be valued lower than properties with more convenient access to these and other amenities.

Furthermore, New York real estate is unique as it has special potential value drivers. For instance, a property located on an exclusive street such as Park Avenue is bound to entail a price premium over properties even a short distance away. Properties near famous landmarks or historical buildings might also carry an inherent premium. A penthouse unit will experience reduced street noise and enjoy better views than lower level units. We expect such omitted variables to positively impact price.

Conclusion

This project aims to estimate the price of an apartment in New York City by examining the number of bedrooms while controlling for other factors. The five boroughs within the city were used to ensure that the regression model produced an accurate location-based price estimate. The number of bedrooms is the primary price predictor variable, and its coefficient is statically significant throughout most of the regression models. For an increase in one bedroom, there is a multiplicative effect of about 1.25, or a 25% increase in price, ceteris paribus.

Future research could analyze the data at a more granular scale than the borough level. Other omitted variables (such as proximity to stores, crime & safety, parking, traffic, etc.) will need to be studied and analyzed. Including relevant variables and broader locations (regions across different states) in the regression model could increase the price prediction's accuracy (adjusted r-squared value). Real estate developers can consider the home's interior and exterior features as well as the project's location to optimize the apartment price.