Claudia Adam
BSAN 360
Professor Xu
13 June 2025

New York City: Short vs. Long Term Rental Analysis

Introduction

My project focuses on short and long term housing rental prices in New York City and how they tie into the local economy. For short term housing, the dataset is from Airbnb and includes listings across New York City in 2019. Since Airbnb's have become a popular alternative to hotels and New York City is a prime tourist attraction, the number and price of these short term rentals gives a good insight into tourism housing demand.

My second dataset shows listings from StreetEasy, an online real estate market for New York City, across Manhattan in 2019. These represent long term housing aimed at residents of New York City. This data includes rental price per month, number of bedrooms and bathrooms, square footage, and several binary variables representing different amenities of the building. From my preliminary analysis, I've found that both datasets have 20 common neighborhoods that I can use to compare aspects of both datasets.

The problem explored in this project is the impact of short term rentals, specifically Airbnb listings, on long term housing affordability in Manhattan, New York City. The concern is that large increases in Airbnb listings may reduce long term rental supply, and if these short term rentals are relatively highly priced, they may drive up housing prices. This issue can have an especially large impact in densely populated urban areas, such as New York City.

After conducting the preliminary exploration of the data, I aim to address the following three research questions:

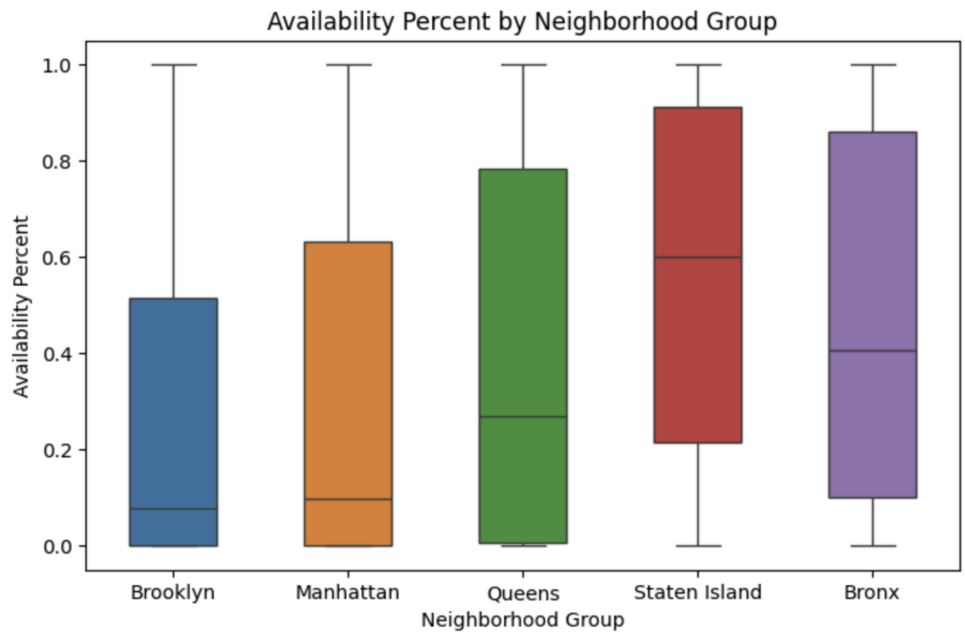
1. How does price and availability for Airbnbs in New York City differ across the five boroughs?
2. Which building amenities are statistically significant for predicting average long term rental prices?
3. Are increases in short-term rental activity (both volume and price) correlated with or predict higher long-term rental prices in certain neighborhoods?

How does price and availability for Airbnbs in New York City differ across the five boroughs?

From the Airbnb dataset, the variables used in this section are neighborhood_group, price, availability_365, and availability_percent. For the first table below, I grouped the data by neighborhood_group, which represents each of the five boroughs in New York City, then calculated the average Airbnb price for each borough. As you can see, Manhattan has the highest average nightly Airbnb price by quite a bit. This is likely due to the high amount of tourist attractions located in Manhattan. The average price for the Bronx and Queens is less than half the average nightly price for Manhattan, which makes sense as these areas are known for being unsafe.

```
Average Airbnb price per neighborhood:
 neighbourhood_group
Manhattan        196.895853
Brooklyn         124.398009
Staten Island    114.812332
Queens            99.536017
Bronx             87.508257
Name: price, dtype: float64
```

The following visual shows availability of Airbnbs across the different boroughs. I wanted to represent availability as a percent of the year, so I converted availability_365 (shows availability as number of days) to a percent by dividing the values by 365 and saved this percent to its own column. I chose to use boxplots to see how the descriptive statistics (min, max, median, IQR, etc) compared among boroughs. We can see that for each neighborhood, this value ranges from 0-1 which is the absolute maximum and minimum. However, one challenge we're seeing here is that it doesn't make sense to have a listing that's available 0% of the year, but I don't want to eliminate this data because this could mean that it's already been booked for all the days it was available. Across the boroughs, Staten Island has the highest median percent availability, which could mean there's the lowest demand for Airbnbs in that neighborhood. Brooklyn and Manhattan have the lowest median availability which makes sense from a tourism perspective. All the boxplots had an interquartile range of at least 0.5, which makes it seem like the data is pretty spread out for all boroughs.



Which building amenities are statistically significant for predicting average long term rental prices?

From the StreetEasy Rentals dataset, this section uses all the numeric values, including number of bedrooms and bathrooms, square footage, and binary variables representing building amenities, to run a multivariable regression on average long term rental price in Manhattan. I chose to do a robust regression to correct for heteroskedasticity, which appears when variance of error terms aren't constant among independent variables.

OLS Regression Results

Dep. Variable:	rent	R-squared:	0.779
Model:	OLS	Adj. R-squared:	0.778
Method:	Least Squares	F-statistic:	348.8
Date:	Sun, 08 Jun 2025	Prob (F-statistic):	0.00
Time:	19:02:55	Log-Likelihood:	-30869.
No. Observations:	3539	AIC:	6.177e+04
Df Residuals:	3524	BIC:	6.186e+04
Df Model:	14		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-422.1606	113.993	-3.703	0.000	-645.582	-198.739
bedrooms	-315.4940	50.843	-6.205	0.000	-415.144	-215.844
bathrooms	1181.0643	114.153	10.346	0.000	957.328	1404.801
size_sqft	4.9174	0.215	22.904	0.000	4.497	5.338
min_to_subway	-16.4173	5.032	-3.262	0.001	-26.281	-6.554
floor	23.3572	2.778	8.407	0.000	17.912	28.802
building_age_yrs	-7.4807	0.756	-9.900	0.000	-8.962	-6.000
no_fee	-130.3260	54.297	-2.400	0.016	-236.746	-23.906
has_roofdeck	31.1217	88.026	0.354	0.724	-141.406	203.650
has_washer_dryer	152.6844	84.580	1.805	0.071	-13.089	318.457
has_doorman	-159.6799	87.380	-1.827	0.068	-330.942	11.582
has_elevator	87.3039	97.461	0.896	0.370	-103.716	278.323
has_dishwasher	-26.7949	80.343	-0.334	0.739	-184.265	130.675
has_patio	-103.0681	116.658	-0.884	0.377	-331.714	125.578
has_gym	-11.8094	93.192	-0.127	0.899	-194.462	170.843

Omnibus:	1005.431	Durbin-Watson:	2.076
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10371.268
Skew:	1.048	Prob(JB):	0.00
Kurtosis:	11.120	Cond. No.	4.75e+03

At the top, we can see that the R-squared value is 77.9%, indicating that the explanatory variables together account for a substantial portion of the variation in rental prices in Manhattan. The F-statistic p-value is 0.000, confirming that the model is statistically significant overall and meaning at least one predictor has a non-zero effect on rental price. However, the regression included a warning about a possible multicollinearity issue, which occurs when independent variables are highly correlated with each other. This could distort the individual significance of coefficients, making it harder to isolate their true effect.

Looking at the individual coefficients and their p-values, several variables are statistically significant at the 1% level:

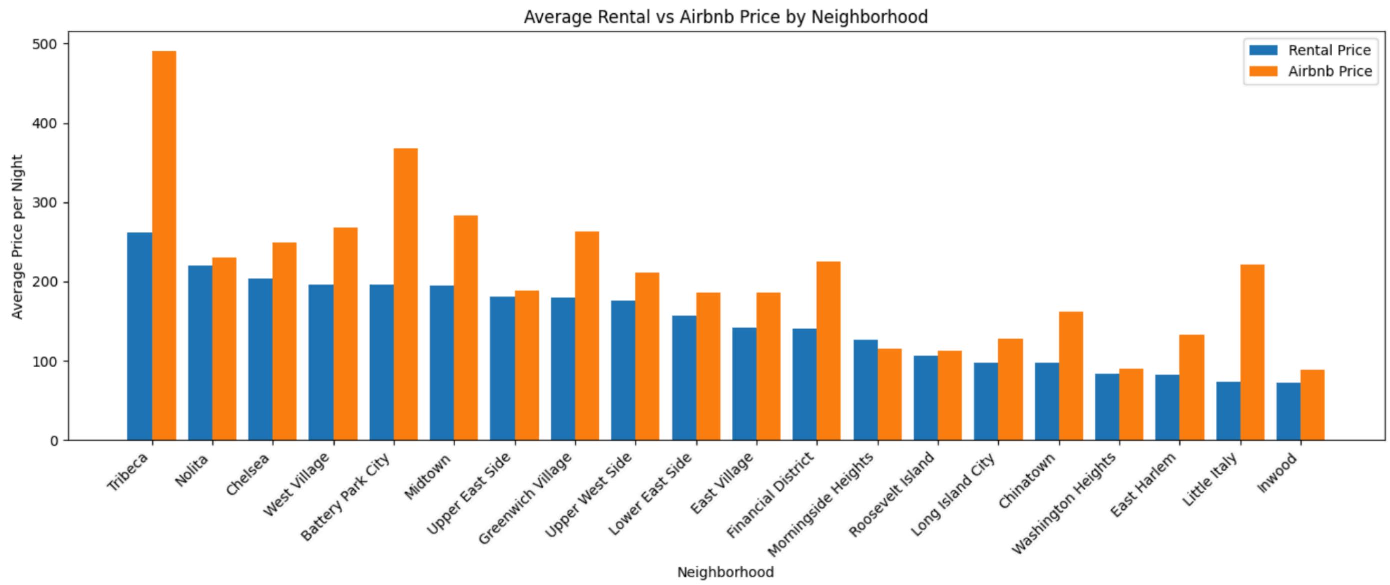
- Bathrooms: Each additional bathroom is associated with a $1,181 increase in rent, holding other variables constant.
- Size (sqft): As expected, more square footage increases rent by about $4.91 per additional square foot.

- Bedrooms: Interestingly, this has a negative coefficient of -$315, meaning more bedrooms are associated with lower rents, all else equal.
- Minutes to subway: Each additional minute from the subway decreases rent by $16.42, consistent with economic expectations.
- Building age: Newer buildings command higher rents, with each additional year of age reducing rent by $7.48.

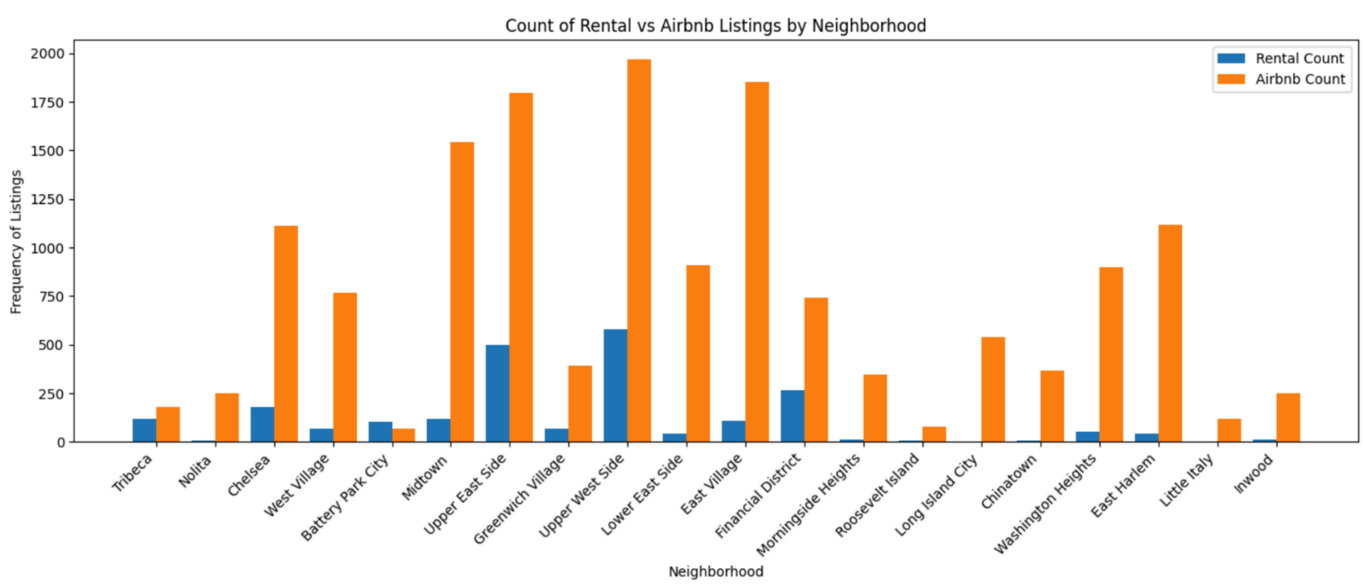
On the other hand, many amenity variables like roof deck, washer/dryer, doorman, and gym are not statistically significant at the 5% level. This implies that while these features may matter to renters subjectively, they may not significantly affect average rent prices across the dataset.

Are increases in short-term rental activity (both volume and price) correlated with or predict higher long-term rental prices in certain neighborhoods?

This section aims to answer my final business question: Are increases in short-term rental activity (both volume and price) correlated with or predict higher long-term rental prices in certain neighborhoods? As previously mentioned, there are 20 common neighborhoods that appear in both datasets. I've taken subsets of each dataset that contain listings just from these overlapping neighborhoods and created variables representing average long term rental price, total count of rental listings, average Airbnb price, total count of Airbnb listings, availability in days, and percent availability. I merged the two dataframes in order to have just one table with values for each of the variables mentioned above for each of the 20 neighborhoods. Using this data, I created a dual bar chart comparing average Airbnb price to average long term rental price. The long term rental prices were measured per month, so I converted them to daily prices by dividing the value by 30.437 (average number of days in a month), so both price variables are now in the same units. Comparing average nightly prices between long and short term rentals, we can see that Airbnb prices are higher for all of these neighborhoods, except Morningside Heights, but even then it's quite similar. The neighborhoods are ordered by a descending rental price, so we notice that average price decreases slightly with each neighborhood. However, this same pattern isn't present in average Airbnb prices, because we see bigger differences in prices between Airbnbs in different neighborhoods. Also, the difference in prices between Airbnbs and long term rentals in the same neighborhood have a large range. For example, tribeca has a large difference in prices with Airbnb nightly prices being almost double those of long term rentals. But in neighborhoods where average nightly prices are cheaper, like Roosevelt Island and Inwood, average Airbnb and long term rental prices are quite similar. These differences in prices may have to do with the number of available properties which we'll look at in the next graph.



Next, I created another dual bar chart comparing the number of listings for each neighborhood from the two datasets. Since the subset of the airbnb dataset with just Manhattan listings has about 21,000 rows, while the streetEasy dataset only has about 3500 rows, we see a large difference in the number of listings from each dataset for each neighborhood. However, the relative number of listings for each neighborhood can still be used to explain average nightly prices. For example, Tribeca has one of the lowest counts for Airbnb listings which is probably one reason why the average nightly prices for Airbnbs in that neighborhood are so high. Same can be said for Airbnbs in Battery Park City and Little Italy. We can apply basic economic theory, which tells us that neighborhoods with low supply of long or short term rental listings are going to have excess demand which drives up prices. Vice versa goes for neighborhoods with large housing supply, where the excess supply causes realtors to lower prices to stay competitive.



Finally, I'll run Pearson's correlation test which is a statistical measure that assesses the strength and direction of a linear relationship between two continuous variables, like price. I used the average Airbnb price and average long term rental price for the 20 overlapping neighborhoods. The result showed the two variables had a Pearson correlation coefficient of 0.82 which tells us that average Airbnb prices and average rental prices have a strong positive correlation with each other. This means that higher airbnb prices are associated with higher long term rental prices, and vice versa. The p-value for this test is 0.000 which means this result is statistically significant.

Challenges

As I mentioned earlier, one main difference between the two datasets is that the Airbnb data had almost 50,000 listings and covered data from five different boroughs within New York City. Meanwhile, my StreetEasy rentals data only covered about 3500 listings and the area was limited to just Manhattan. Certain neighborhoods within the rentals dataset with a lower relative amount of listings may have skewed average prices. Also, it meant my analysis is limited to Manhattan housing in 2019, which won't reflect recent shifts in the housing market, including effects from the pandemic.

In my analysis, I converted rental average prices to per day, so they would be in the same units as airbnb average prices. However, it's still difficult to compare the two prices because there are costs specific to short or long term housing included in the price. For example, long term rentals usually require security deposits and utility payments, while airbnb prices include cleaning fees. Another factor that may influence price dynamics are the different target audiences for short vs. long term housing. These two that I mentioned as well as other factors can not be fully explained by simple price per night comparisons.

For the boxplots of percentage availability, I mentioned how there were some airbnb listings with 0% availability. This may be due to being fully booked, but it could also be because of inactive listings, so it's hard to interpret the true meaning of availability precisely. Also, the availability columns do not reflect seasonal patterns which is very important in tourism heavy markets like New York City.

In the multiple linear regression on long term rental prices, there was a warning for indications of multicollinearity, which means at least two x variables are highly correlated which can distort reliability of estimates. Also, I mentioned that some slope coefficients had illogical signs, like bedrooms have a negative slope which means additional bedrooms will decrease price. This could result from the issue of multicollinearity and errors in data collection.

Finally, correlation does not imply causation, so even though the Pearson correlation coefficient showed a strong positive correlation, we cannot directly say increasing airbnb prices will cause an increase in long term rental prices, or vice versa.

Conclusion

This project explored the relationship between short-term and long-term housing rentals in Manhattan focusing on how Airbnb activity may be associated with long-term rental prices. The findings revealed a strong positive correlation between average Airbnb prices and average long term rental prices, suggesting that neighborhoods with more expensive short term rentals also tend to have higher long term housing costs. These results support concern that short term rentals, like Airbnbs, may put pressure on local housing markets.

New York City is a major tourism hub with high demand for both short and long term housing. Airbnb prices were higher than long term rentals in nearly every neighborhood, which provides an incentive for landlords to list units for short stays rather than long term leases. In areas like Tribeca and Battery Park City, limited supply and high tourist appeal increases both Airbnb and rental prices. One explanation is that high tourism neighborhoods drive up short term rental prices, which can limit long term housing supply and increase rents.

References

Gomonov, D. (2019). New York City Airbnb Open Data. Retrieved from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Inside Airbnb. (2019). New York City Airbnb Open Data. Retrieved from https://insideairbnb.com

Zohaib30. (n.d.). StreetEasy Dataset. Retrieved from https://www.kaggle.com/datasets/zohaib30/streeteasy-dataset

StreetEasy. (n.d.). NYC Real Estate Listings and Market Data. Retrieved from https://streeteasy.com

Gattupalli, A. (2024). Tourism Gentrification in the United States: How Short-Term Rentals Impact American Cities. Medium. Retrieved from https://medium.com/@ankithagattupalli/tourism-gentrification-in-the-united-states-how-short-term-rentals-impact-american-cities-39316e1b39bb

NYC Housing. (n.d.). Problem. NYC.gov. Retrieved from https://www.nyc.gov/site/housing/about/problem.page