Exploring Statistics in the Real Estate Market

Introduction to Applied Statistics

Grosskurth, P., Chini, K., Do, B.
University of Nebraska at Omaha

5 May, 2025

Abstract

This paper examines the factors influencing housing prices in New York City using statistical analysis of a 2024 real estate dataset comprising 4,801 properties. Through confidence interval estimation, hypothesis testing, and linear regression analysis, we identify the relative importance of square footage, bedroom/bathroom count, and location on property values. Our findings demonstrate that while simple linear models show limited predictive power (r^2 values 0.065-0.189), bathroom count exhibits the strongest correlation with price, with a correlation coefficient of 0.43. We also found that location played a great role in the price bracket of housing. The 95% confidence interval for mean housing prices falls between \$1.751M and \$1.982M, with strong evidence rejecting the null hypothesis that average prices are below \$1M ($p < 0.000001$). The study highlights the complexity of housing price determinants in urban markets and the limitations of simple linear models.

1 Introduction

Statistics offers a wide array of tools that are immensely useful in the real world, particularly when it comes to understanding complex systems like housing markets. In this paper, we apply statistical formulas and strategies to analyze and attempt to predict trends in a local housing market. The dataset we've selected is drawn from the New York City housing market in 2024. This dataset, when properly utilized, holds the potential to accurately predict housing prices based on key information about individual properties. Our goal is to explore how statistical analysis can uncover the factors that most strongly influence housing prices across New York City's boroughs.

These predictions, along with other statistically informed insights, play a major role in shaping the housing market today. For large real estate investment firms, such data is invaluable. It informs not only the prices they offer for current properties but also guides strategic decisions about where to purchase in the future. On a more everyday level, this same data helps generate rough property value estimates through online platforms like Zillow, empowering buyers and sellers with more informed expectations.

Statistical analysis has become deeply embedded in how we understand and engage with the housing market. The ability to extract meaningful patterns from large datasets has helped create the real estate landscape as we know it. In this paper, we aim to answer: What factors most strongly influence housing prices across New York City's boroughs? With this guiding question, we will explore the relationships between property features—such as square footage, number of bedrooms, and location—and their impact on final sale prices. Can we accurately predict housing prices across all five boroughs using this dataset? Does square footage outweigh location, or vice versa? This paper will attempt to find those answers through statistical analysis.

2 Dataset

It feels pertinent to mention that the dataset is not perfect, as is often the case in the real world. Certain information is often not available, and rather than skew the dataset by leaving some entries blank within a row, the author has decided to infill it with the average of the others. This ensures that the lack of data will not affect some of the statistical conclusions. This dataset includes a wide range of real estate information. Some of the important ones are price, bedrooms, bathrooms, square footage, and location (latitude and longitude). These are the columns that we will be utilizing in our analysis. This will provide some great visualization tools and many different lenses to view the data through.

Our analysis begins with a dataset of 4,801 properties, which we filtered to 4,707 by removing extreme outliers. The final price range analyzed spans from $150,000 to $65,000,000, with a mean price of $1,866,621 and a median of $838,000. This substantial difference between mean and median, along with a skewness measure of 7.095, confirms the right-skewed distribution visible in our histograms.

Statistic	Mean	Median
Price	$1,866,621	$838,000
Bedrooms	3.27	3.00
Bathrooms	2.31	2.00
Square Footage	2,140	2,184

Table 1: Summary statistics of key variables

In any dataset it is near unavoidable for there to be outliers in the set. Outliers are defined as an observation that lies an abnormal distance away from the average. These outliers tend to drastically affect the distribution and values obtained from performing statistical analysis on the dataset. While in some contexts these outliers are important to fully understanding the dataset, for our purposes these outliers will be excluded from our calculations. We will be using a trimmed dataset as our sample, and will be using a trimmed mean as our sample mean. A histogram of our unmodified dataset is seen below.

Figure 1: Pre-trimmed data

Our dataset originally had a mean price of 2.357 million dollars, and a median real estate price of 825,000 dollars. After it was trimmed, we had a mean of 1.89 million dollars, and a median of 846,500 dollars. The method used to trim the data set was as follows: 1. find the sample standard deviation, 2. Remove any values over three standard deviations above or below the sample mean, 3. Personally remove non-fitting outliers. The last step had to be performed as some prices were not like the rest of the sample, namely, rent pricing. So we set a minimum price of 150,000 to avoid this issue.

This change reduced the skew from 66.7 to 6.9. This change is visually apparent in the histogram below, as the number of listings per price shifted. While the data is still heavily skewed, this reduction allows us to draw more consistent conclusions about both our sample and the population that it comes from.

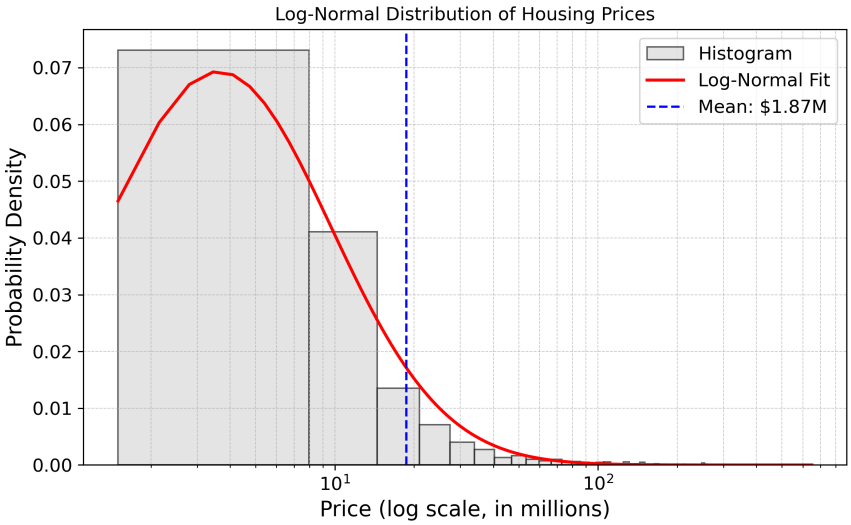


Figure 2: Post data trimming

With a more representative dataset now established through careful trimming and preprocessing, we can proceed with our statistical analysis with greater confidence. The adjusted mean and reduced skew allow us to treat the sample as a more reliable reflection of the population, which is critical for inferential statistics. One of the first steps in this process is to estimate the population mean using our sample data, along with a measure of certainty. This is where confidence intervals become essential, as they provide a useful range in which we expect the true population mean to fall. Before we begin testing these intervals, it's important to understand how they are constructed and interpreted.

3 Testing

3.1 Confidence Interval

Inferential statistics provide solutions to make estimations about populations from a sample. We can utilize this to make predictions in confidence, by taking our mean and adding or subtracting our margin of error. Let's say that we want to make an estimate about New York's housing price range. Since our data set is just a sample of the New York housing market, we can use our sample mean to form a confidence interval for the price. For a price range prediction with a confidence of 95%, and a sample mean of 1.893 million, we find that with 95% confidence, we can say the mean price of the New York housing market is between 1.777 million and 2.01 million. The formula is shown below.

$$\text{CI} = \bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right)$$

Our more precise calculations using the complete dataset yield a 95% confidence interval of \$1.751M to \$1.982M, with a sample mean of \$1.867M. This narrow range suggests our sample provides a relatively precise estimate of the population mean, despite the skewed distribution. The confidence interval calculation assumes our sample is representative and meets the conditions for the Central Limit Theorem to apply, which appears reasonable given our large sample size (n = 4,707).

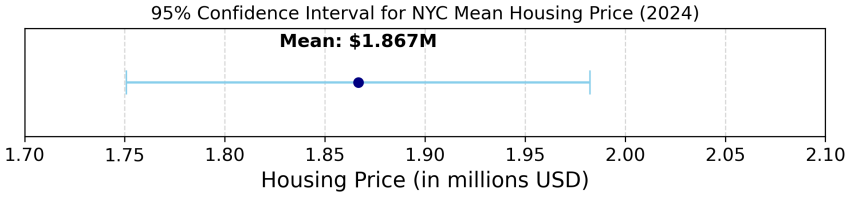


Figure 3: 95% Confidence interval for NYC housing prices

This confidence interval gives us a statistically supported estimate of where the true average housing price for the entire New York market likely falls. By using inferential statistics, we're not just looking at isolated data points—we're making broader, population-level claims based on our sample. This method is especially valuable in housing markets, where collecting data on every single property is impractical or impossible. Instead, we rely on representative samples to infer trends and make educated decisions. It's important to note that this 95% confidence level doesn't mean there's a 95% chance the true mean falls within the range. Rather, it means that if we repeated this sampling process many times, 95% of those intervals would contain the true population mean.

3.2 Hypothesis Testing

Before proceeding with our analysis, we conducted a hypothesis test to determine whether the average housing price in our dataset significantly exceeds $1M. Our results show overwhelming evidence to reject the null hypothesis ($z = 14.67$, $p < 0.000001$), confirming that NYC housing prices are indeed substantially higher than this threshold. This finding aligns with common knowledge about NYC's expensive real estate market and validates our dataset's representativeness for high-value urban properties. The equation used to find our test statistic for the dataset is shown below.

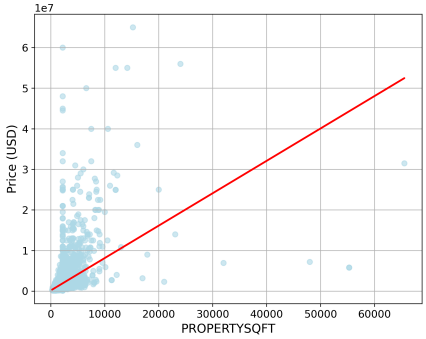
$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Understanding the price range also helps us evaluate market volatility and risk. For real estate investors, a wider interval might suggest a less stable or more variable market, while a narrower interval can indicate pricing consistency and greater predictability. In our case, the range from 1.777 million to 2.01 million shows moderate variation, offering useful insight to both institutional investors and individual buyers. As we continue our analysis, we will apply similar inferential techniques to other features—such as square footage, location, and number of bedrooms—to explore how these variables influence price across New York.

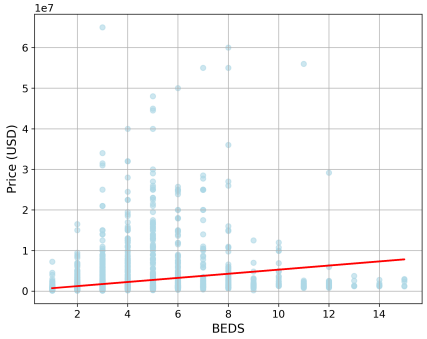
3.3 Linear Approximation

We used simple linear regression to attempt to understand the correlation of price with square footage, bedrooms, and bathroom count. This was an attempt to accurately predict the housing prices in the area, using the equation $y = \beta_0 + \beta_1 x$, where β_0 is the intercept, and β_1 the slope. For each of these listed slopes, they correlate with a monetary increase in the price per unit of the independent factor. For price versus square footage, we had a slope of 741.68; for the number of bedrooms, we had a slope of 364,356.12; and for the number of bathrooms, we had a slope of 769,653.24. This information is all summarized in Table two.

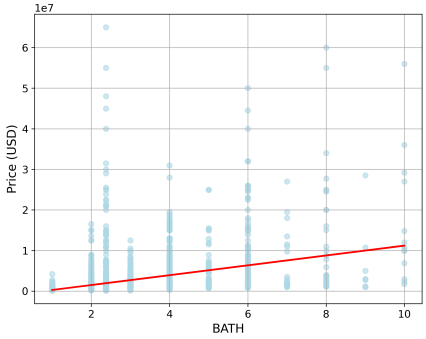
- Square footage: Each additional square foot associates with $798.29 higher price (intercept $158,174.84, $r^2 = 0.189$)
- Bedrooms: Each additional bedroom links to $507,175.56 higher price (intercept $208,683.37, $r^2 = 0.065$)
- Bathrooms: Each additional bathroom connects to $1,213,752.51 higher price (intercept -$937,532.69, $r^2 = 0.188$)



(a) Price dependent on Square footage



(b) Price dependent on bedrooms



(c) Price dependent on bathroom count

Figure 4: Simple Linear Regression

VARIABLE	INTERCEPT	SLOPE	r^2
SQFT	\$266,734.25	741.68	0.186
BEDS	\$664,342	364,356.12	0.054
BATH	\$55,250.90	769,653.24	1.135

Table 2: Numerical results from linear regression

One of the key ways that a linear regression's efficacy is determined is through r^2 values. These values that we produced show that simple linear regression is not capable of accurately predicting the prices of the properties in our data set. The low r^2 values (0.065-0.189) indicate that these single-variable models explain only 6.5% to 18.9% of price variation, leaving most variation unexplained. This strongly suggests that housing prices depend on multiple factors simultaneously rather than any single feature in isolation. The particularly weak bedroom model ($r^2 = 0.065$) may reflect that bedroom count matters less in luxury properties where other amenities dominate pricing decisions.

To get more accurate results would require methods beyond the scope of this class, such as multiple linear regression, whilst encoding location into the data.

4 Discussion

	PRICE	BEDS	BATH	SQFT	LATITUDE	LONGITUDE
PRICE	1.00	0.25	0.43	0.43	0.09	-0.10
BEDS	0.25	1.00	0.73	0.33	-0.05	0.01
BATH	0.43	0.73	1.00	0.38	-0.04	-0.07
SQFT	0.43	0.33	0.38	1.00	0.05	-0.02
LATITUDE	0.09	-0.05	-0.04	0.05	1.00	0.52
LONGITUDE	-0.10	0.01	-0.07	-0.02	0.52	1.00

Table 3: Correlation matrix of selected housing variables

The correlation matrix reveals several important relationships in our dataset. Bathroom count shows the strongest correlation with price (0.43), followed closely by square footage (0.43), while bedroom count shows a weaker relationship (0.25). This pattern suggests that in NYC's luxury market, bathroom quantity and overall space matter more than simple bedroom count when determining property values. The high correlation between bedrooms and bathrooms (0.73) reflects typical architectural patterns where these features scale together in residential properties.

Interestingly, location (latitude/longitude) shows minimal direct correlation with price (0.09 to -0.10). However, when plotted correctly, it is clear to see this is not entirely true.

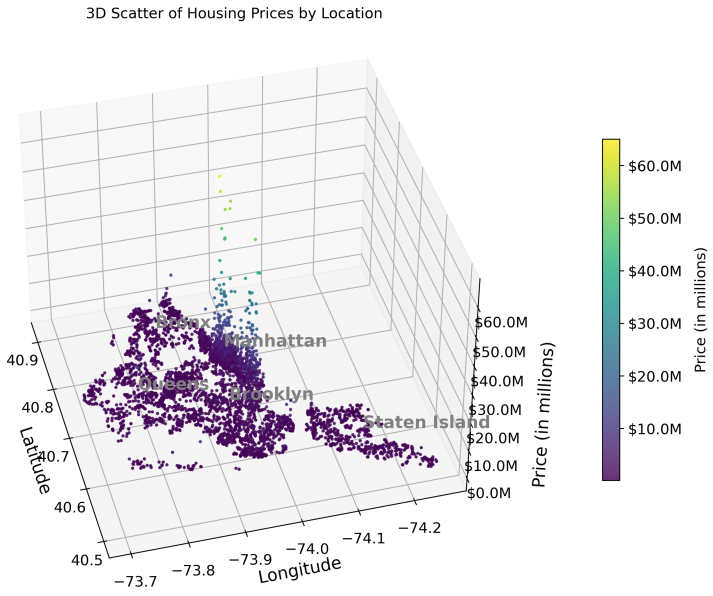


Figure 5: 3D scatter plot of NY's housing pricing

5 Conclusion

Our statistical analysis of NYC housing data reveals several key insights about this complex market. The confidence interval ($1.751M-$1.982M) establishes a robust estimate of average prices, while the significant skewness (7.095) highlights the prevalence of ultra-high-value properties that distort simple averages. The hypothesis test conclusively demonstrates that NYC prices far exceed $1M on average ($z = 14.67$, $p < 0.000001$).

The regression analyses show that while all examined features (square footage, bedrooms, bathrooms) positively correlate with price, none alone provides strong predictive power (r^2 0.065-0.189). Bathroom count emerges as the single strongest predictor, suggesting luxury amenities may outweigh basic size metrics in this market. The correlation coefficient matrix further supports this interpretation, with bathroom count showing the highest price correlation (0.43). These findings have important implications for both practitioners and researchers. For real estate professionals, our results suggest that simple metrics like bedroom count may be less important than overall quality (reflected in bathroom count) when valuing NYC properties. For researchers, the low r^2 values emphasize the need for more sophisticated modeling approaches that consider multiple factors simultaneously and potentially incorporate non-linear relationships.

Future research should explore multivariate models that combine physical characteristics with location data and temporal trends. Additionally, examining price determinants separately by borough or neighborhood could reveal important geographic variations masked in our city-wide analysis. Despite its limitations, this study provides a foundation for understanding which basic property features most strongly influence the value of real estate in one of the world's most complex real estate markets.

References

Nidula Elgiriyewithana. New york housing market, 2024. URL https://www.kaggle.com/dsv/7351086.