**NYC Property Sale Price Analysis**

Renusree Nagirimadugu

Deepak Rambarki

Team 15

UNT

ADTA 5940 Analytics Capstone Experience

31st Mar 2025

**Table of contents**

[**Abstract 3**](#_6qcammf3jhgk)

[**1. Introduction 3**](#_p54venp5rawc)

[**II. Literature Review 5**](#_hgeqi18o134l)

[**III. Methods 11**](#_qsv2nz4nvcxs)

[1. Data Source and Description 11](#_ae36l411t3pk)

[2. Data Cleaning and Preparation 12](#_8v09g3xl799w)

[3. Feature Engineering 14](#_8fozkbyouyms)

[4. Exploratory Data Analysis (EDA) 17](#_7sypy3d2wbfw)

[5. Modeling Approach 18](#_33wy9b6deocg)

[**IV. Results 19**](#_ejk3iemc7aqj)

[1. Data Overview and Cleaning Outcomes 19](#_bh061fuhauo2)

[2. Exploratory Data Analysis (EDA) 20](#_hjlmyz5cyqj4)

[2.1 Distribution of Sales by Borough 20](#_s3b2yim3yi1u)

[2.2 Sale Price Characteristics 21](#_aru1xv64z3u1)

[2.3 Outlier Examination 22](#_j91x4sn1nw6b)

[2.4 Borough-Level Price Comparisons 22](#_9hr7c36gfer0)

[2.5 Neighborhood Highlights 23](#_l0syme7kouo9)

[2.6 Time Trends and Seasonal Patterns 24](#_o73aenwinx78)

[2.7 Feature Correlations 26](#_pgaw6f9dwydk)

[3. Predictive Modeling Results 27](#_sm9b3xvvtdsp)

[3.1 Model Training and Evaluation 27](#_ic48giq0bry9)

[3.2 Feature Importance (Random Forest) 29](#_twb174q6vj1k)

[4. Borough-Specific Model Performance 30](#_sqqlleyezywl)

[Key Takeaways from the Results 32](#_mmtkr1na2c40)

[**V. Discussion and Interpretation 33**](#_skv40dut37bl)

[1. Addressing the Research Questions 33](#_qhz2botrz5h0)

[2. Interpretation of Key Findings 34](#_1dd7se8c9fte)

[3. Connection to Broader Trends 35](#_51de8uizknml)

[4. Practical Implications 36](#_6d3j5y59o5n9)

[5. Limitations and Caveats in Interpretation 37](#_xlocfwaow9y)

[**VI. Limitations 37**](#_7sqzxtjfotmp)

[**VII. Recommendations and Future Work 38**](#_vwsplx1b3tau)

[**Conclusion 40**](#_rqmtqtg3i6lj)

[**References 41**](#_v6g69auvs2lc)

# Abstract

New York City's (NYC) property market is known for being complicated. This is because different neighborhoods are very different from each other, prices change a lot, and market conditions keep changing. This study looks at how we can use past sales data and property features—like building size, year it was built, and location—to predict sale prices in the five boroughs. After cleaning and filtering a dataset with more than 600,000 entries, we used feature engineering (like price-per-square-foot and building age) and tested different regression models. These included Linear Regression, Random Forest, and Gradient Boosting. The Random Forest model worked the best. It explained about 82% of the change in the log-transformed sale price (R² ≈ 0.82) and had a low RMSE. When we looked at each borough, we saw that it was still hard to predict prices in Manhattan because of luxury and commercial properties that are different from the rest. This shows that we may need special models for expensive areas. The results show that location is the most important factor for price, followed by building size and property type. We suggest that using more detailed neighborhood data and macroeconomic indicators could make the models better. This can help people like policymakers, real estate investors, and urban planners to understand and predict NYC's changing real estate market.

# 1. Introduction

New York City (NYC) is known as one of the most active real estate markets in the world. Prices keep changing, neighborhood trends keep evolving, and there is not enough housing. Recent data shows that the average home value in NYC is about $763,358, which is a 4.2% increase in the past year (Zillow, 2025). But the market is different in each of the five boroughs. In January 2025, Manhattan's median asking price was $1.55 million, which is 6.3% less than last year. Brooklyn's median asking price went up by 4.8% to $1.1 million, and Queens saw a big 12% increase to $700,000 (New York Post 2025). These differences show that NYC's market is divided into smaller parts, where things like location, zoning rules, and population changes affect prices in different ways.

Besides sale prices, other signs also show that demand is growing and supply is low. In 2023, the city's rental vacancy rate dropped to a record low of 1.4%, showing that there is a serious housing shortage (NYC Comptroller 2023). Building new housing, especially in areas like Midtown Manhattan, is difficult because of strict zoning rules and high building costs (New York Post 2025). At the same time, higher mortgage rates and cost of living are making it harder for people to afford homes, so more people are looking at outer boroughs or even outside NYC. Because of this, building strong prediction models can help us understand current price trends. This can help with policy making and also help individuals make better investment choices. This study plans to use past sales data, building features, and neighborhood information to predict property prices in different boroughs and find out what affects NYC's real estate values over time.

**Research Questions.**

This study addresses the following specific research questions:

1. **How effectively can machine learning models predict NYC property sale prices using historical sales data and physical property features?** This question examines which modeling approaches provide the most accurate price predictions and what level of prediction accuracy can be achieved across NYC's diverse real estate market.
2. **What factors have the strongest influence on property sale prices across different NYC boroughs?** This question investigates whether location, physical characteristics (building size, age, units), property type, or other features contribute most significantly to explaining price variations.
3. **How do predictive models perform differently across NYC's five boroughs, and what explains these performance differences?** This question explores whether certain areas require specialized modeling approaches due to unique market characteristics, and whether prediction accuracy varies systematically by location.

# II. Literature Review

This section looks at different research studies that have tried to predict real estate prices, understand property values, and analyze neighborhood trends. These studies help us learn what factors affect real estate prices, how large datasets can be handled, and what problems might come up when doing this type of research.

Many researchers agree that adding detailed property features and location details makes price predictions more accurate. For example, Arul and Morales (n.d.) found that including latitude and longitude improved the accuracy of machine learning models for predicting NYC condominium prices. Their study tested different models, and they found that simple models like multinomial Naïve Bayes performed badly, while Support Vector Regression (SVR) gave better results when location data was included. After adding location data, the mean squared error (MSE) became much lower, but the prediction error still stayed above 20%. This shows that location is important, but other factors also influence property prices in NYC.

Similarly, Ho, Tang, and Wong (2021) tested different machine learning models to predict housing prices, focusing on Hong Kong. Even though this is a different city, their study showed that advanced models like Random Forest and Gradient Boosting Machines worked better than simpler ones like Support Vector Machines (SVM). Their study found that floor area and building age were important factors. While their study did not focus on how prices change across different areas in a city, it still proves that machine learning models can be very useful in predicting real estate prices. If we apply similar methods in NYC, we can understand how different neighborhoods affect property prices.

Varma, Sarma, Doshi, and Nair (2018) studied real estate prices in Mumbai and combined different models, including linear regression and neural networks. They found that adding extra information about the neighborhood, such as how close the property is to schools, hospitals, and public transport, made predictions more accurate. While Mumbai and NYC are different, this idea can still apply to NYC—features like proximity to subway stations or good schools could help predict property prices. Their research also showed that using a mix of simple and advanced models (an ensemble approach) can help reduce prediction errors. This suggests that using multiple models together might be a good idea when predicting property sales in NYC.

Dong (2020) studied NYC property prices from 2003 to 2017 and looked at how different features like land size, building size, and transaction frequency affected property prices. The study found that Manhattan properties had the highest price appreciation compared to other boroughs. It also showed that older properties appreciated at a slower rate and that when properties were sold many times, their appreciation rate was lower. This type of research helps us understand how property prices change in different NYC boroughs, which is important for making accurate predictions.

Another study by Alharbi (2023) focused on short-term rental prices in Airbnb listings. Even though Airbnb is different from traditional real estate, the study shows how adding extra data like customer reviews can improve predictions. The study used machine learning models and sentiment analysis to see how guest reviews affected listing prices. The best models, Lasso and Ridge regression, had a high R² score of 99%. This means the models were very good at explaining the price variations. The study proves that using different types of data—such as text reviews or customer opinions—can help improve predictions. In real estate price prediction, we might be able to use similar techniques by analyzing property descriptions or other text data.

Park and Bae (2015) used real estate data from Fairfax County, Virginia, and compared different machine learning models like decision trees, RIPPER (a rule-based algorithm), Naïve Bayesian classification, and AdaBoost. Their study found that machine learning models were better than basic statistical models at predicting whether a house would sell for more or less than the listing price. If applied in NYC, this approach could help predict whether a property will sell above or below its listed value. Their research also showed that property features like mortgage rates and school ratings were important, which suggests that using additional data sources—such as NYC finance data or school quality ratings—could improve predictions.

Xu (2025) conducted a large study on NYC real estate trends, looking at property prices, neighborhood changes, and economic factors. The study used data from Zillow, NYC Open Data, and the U.S. Census Bureau. One of the key findings was that NYC housing prices are affected by both property features and demographics, such as income, race, and education levels. The study also found that neighborhoods like Williamsburg and Long Island City experienced high price growth due to gentrification. Another important finding was that the COVID-19 pandemic temporarily reduced prices in Manhattan but increased demand in the outer boroughs. This study shows that economic conditions and social changes have a big impact on property prices, which is important to consider when making predictions.

Luo, Zhou, and Zhou (2019) studied Airbnb prices in NYC, Paris, and Berlin. Their research found that training a machine learning model on data from multiple cities made the model more generalizable. Although their study focused on short-term rentals, the main idea applies to NYC real estate—using data from multiple boroughs or different time periods could help create better models. They also found that adding text features (such as property descriptions) helped improve accuracy. This suggests that real estate sale predictions in NYC could benefit from analyzing additional text data in property records.

Ma, Cheng, Jiang, Chen, and Zhang (2020) used big data and machine learning to study land values in NYC. They tested different machine learning models and found that Random Forest was the most accurate, achieving an R² score of 0.93. They also discovered that features like floor area ratio, density of points of interest (POIs), and the number of museums nearby had a big impact on land values. The study also showed that places with high foot traffic (measured by the number of newsstands) had higher land values. Another interesting finding was that in expensive areas, there were also high vacant housing rates, possibly because wealthy investors buy properties but do not live in them. This research proves that neighborhood characteristics play a major role in real estate pricing, and including similar features in our NYC study could improve our predictions.

Calainho, van de Minne, and Francke (2024) studied commercial real estate price indices and tested different machine learning models. They found that models like Support Vector Regression (SVR) and Gradient Boosting performed better than traditional linear models like Ordinary Least Squares (OLS). However, they also discovered that machine learning models were less stable when they had small amounts of data. Their study suggests that while machine learning is useful for price predictions, it needs a large dataset to be reliable. This is important for NYC real estate research since some neighborhoods have fewer sales than others, which can create data imbalances.

These studies all show that real estate prices depend on many factors, including property size, location, neighborhood characteristics, and market trends. The best-performing models in most cases were tree-based algorithms like Random Forest and Gradient Boosting, which performed better than simple regression models. Many studies also found that adding new types of data—such as local amenities, property age, and transit access—helped improve prediction accuracy. Another common challenge was dealing with missing or incorrect data, which needs to be cleaned before analysis.

Several studies, like those by Dong (2020) and Xu (2025), also highlight how social and economic changes, such as gentrification and the COVID-19 pandemic, can impact property values. This suggests that a simple model using just property features might not be enough—we also need to consider economic and social factors. Additionally, since NYC has five boroughs with very different real estate markets, it may be necessary to build separate models for each borough or include detailed location-based features.

Overall, this literature review shows that NYC property price prediction should use a mix of machine learning models, location-based features, and external economic data. Cleaning the data properly is essential, especially since NYC datasets contain many outliers (such as properties sold for zero dollars). Using a combination of machine learning techniques, such as ensemble models or time-series analysis, could help improve predictions. Finally, including neighborhood-level features like nearby schools, transit access, and local businesses could provide more accurate results.

The insights from these studies significantly shaped our analytical approach in several ways. First, the consistent superiority of ensemble methods across multiple studies (Ho et al., 2021; Ma et al., 2020; Calainho et al., 2024) directly influenced our decision to prioritize Random Forest and Gradient Boosting over simpler linear models. Second, the identification of location as a primary value driver in nearly all studies reinforced our focus on borough and neighborhood-level variables. Third, the challenges with outliers and data quality noted by researchers working with similar datasets alerted us to potential pitfalls in NYC property data, prompting our rigorous data cleaning protocol that filtered non-market transactions and addressed missing values methodically.

Drawing inspiration particularly from Varma et al.'s (2018) use of engineered features and Ma et al.'s (2020) success with Random Forest models in NYC land value prediction, we implemented several methodological approaches in our work. We created specialized features like price-per-square-foot ratios and building age calculations that help normalize comparisons across different property types. The borough-specific analysis approach was influenced by Dong's (2020) findings on Manhattan's distinct price patterns compared to other boroughs. Additionally, Xu's (2025) observations about pandemic effects on the NYC market guided our time-series exploration of price trends. These research-based insights collectively strengthened our analytical framework, allowing us to design more effective predictive models while anticipating the unique challenges presented by New York City's complex and segmented real estate market.

# III. Methods

This section describes the data source, cleaning steps, feature engineering process, and the general modeling approach used to investigate and predict NYC property sale prices.

## 1. Data Source and Description

We used a publicly available dataset from catalog.data.gov (City of New York, 2023) titled NYC\_Citywide\_Annualized\_Calendar\_Sales\_Update.csv, which contains historical property sales from across the five boroughs of New York City. Each row represents a property sale transaction, including attributes such as address information, borough and neighborhood location details, building class category for property type classification (e.g., office buildings, condos, single-family residences), land and gross square footage dimensions, year built, tax class for property tax purposes, and sale price and date information.

The dataset spans multiple years (2016–2023 in our analysis) and contains over 600,000 rows before filtering. Key data challenges included missing values (with some columns entirely empty or partially filled) and extreme outliers in sale prices, including $0 transactions or exceptionally high amounts that required special handling.



Figure 1 - Metadata of the NYC Sales dataset (City of New York, 2023)

## 2. Data Cleaning and Preparation

Data cleaning was conducted to remove invalid entries and ensure consistent formatting for modeling. Several columns with extremely high missingness were dropped. For instance, EASE-MENT contained no valid entries (100% missing), while Census Tract 2020 and NTA Code were missing over 80% of their values. APARTMENT NUMBER had around 75% missing values and was also removed as it was not critical for the predictions.

For handling missing values in key features, we employed a systematic approach. Numeric columns such as LAND SQUARE FEET, GROSS SQUARE FEET, and RESIDENTIAL UNITS were imputed with the median. For LAND SQUARE FEET and GROSS SQUARE FEET specifically, we grouped by borough and building class category to compute group-specific medians, then defaulted to the overall median if group-level data was insufficient. Categorical columns with moderate missingness were filled with a placeholder like "Unknown" to preserve those rows rather than removing them altogether.

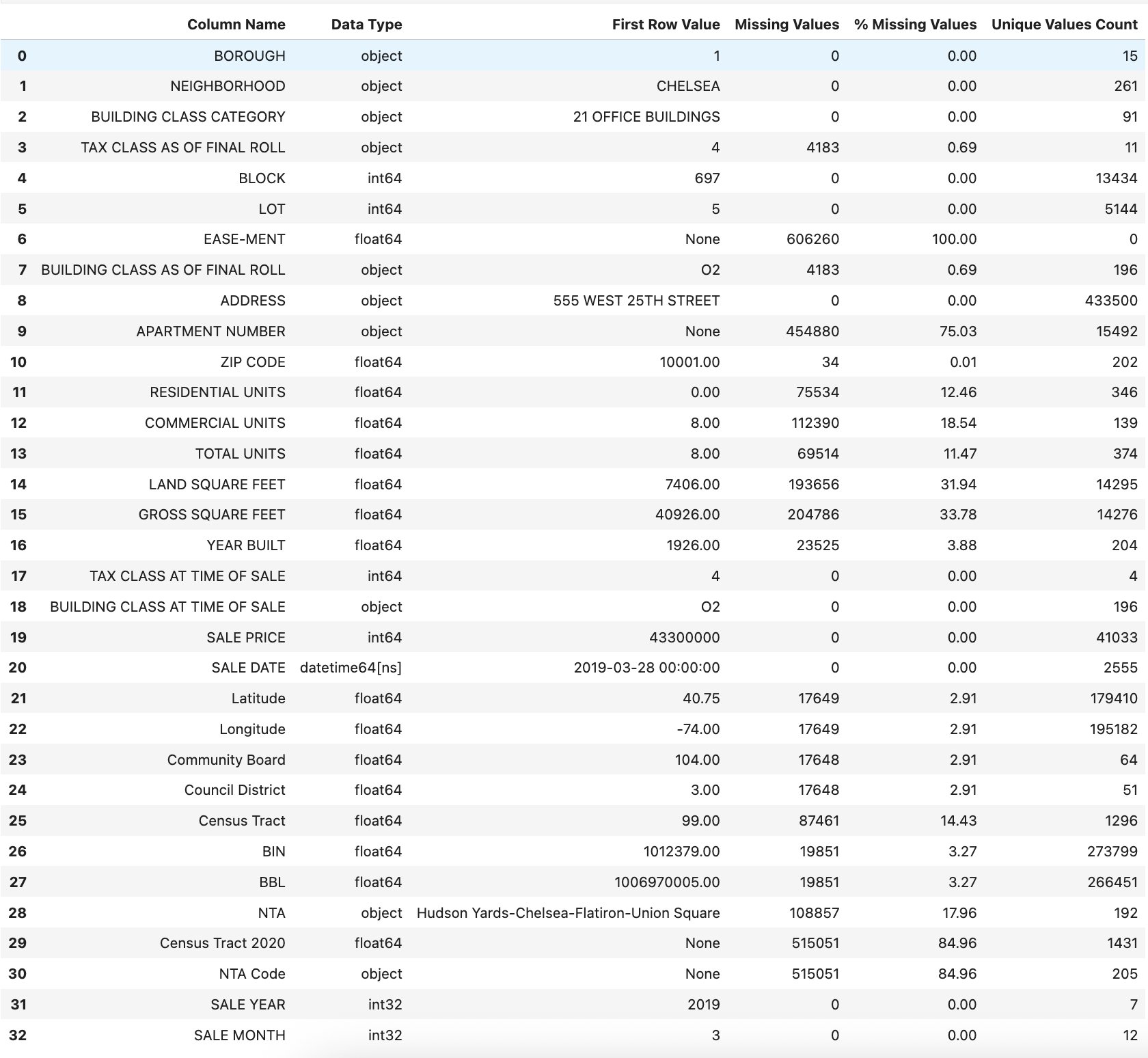


Figure 2 - Dataset Information Snapshot (column wise)

The cleaning process also involved converting data types appropriately. Square footage fields were originally stored as strings with commas, which we stripped and converted to numeric (float) format. Sale Date was converted into a datetime type, from which year and month were extracted as separate features. ZIP CODE was examined for correctness and retained as a float or string, depending on analysis needs.

To remove sales that did not reflect typical market transactions, we filtered out any properties sold for $10,000 or less, as these likely represent non-market transfers, gifts, or data entry errors. This step significantly reduced the dataset and helped remove noise. After filtering out invalid sales, duplicate records were identified and removed to ensure unique transactions.

Additionally, we standardized borough names, as the BOROUGH column originally included mixed formats such as numeric codes (1, 2, 3, etc.) and textual names ("MANHATTAN," "BRONX"). These were all mapped to a consistent set of five categories: "MANHATTAN," "BRONX," "BROOKLYN," "QUEENS," and "STATEN ISLAND."



Figure 3 - Code snippet standardizing borough names

## 3. Feature Engineering

After cleaning, we created additional features to potentially enhance model accuracy. For time-based analysis, we derived SALE YEAR and SALE MONTH from the SALE DATE, along with a combined SALE\_YEAR\_MONTH feature (e.g., "2020-02") for time-series grouping or trend analysis. To address the heavily skewed distribution of property prices, we introduced a log-transformed target variable (LOG\_PRICE) by applying log(SALE PRICE+1), which helps normalize extreme sale price values for modeling purposes.



Figure 4 - Code snippet showing feature engineering steps

Several ratio-based and descriptive features were also engineered. We calculated PRICE\_PER\_SQFT by dividing SALE PRICE by GROSS SQUARE FEET, replacing missing or infinite values with the median. To better represent building characteristics, we created a BUILDING AGE feature by subtracting YEAR BUILT from the current year, providing a more interpretable numeric age feature than raw build years. We added binary classification flags such as IS\_RESIDENTIAL (indicating if the BUILDING CLASS CATEGORY referred to a family dwelling, condo, or co-op) and IS\_COMMERCIAL (for categories including terms like "office," "retail," "factory"). Based on our exploratory analysis, we also created an IS\_HIGH\_VALUE\_NEIGHBORHOOD indicator, marking properties located in top-tier neighborhoods like "MIDTOWN CBD" or "FINANCIAL."

Additional ratio features were developed to capture property characteristics. LAND\_TO\_BUILDING\_RATIO was calculated as LAND SQUARE FEET divided by GROSS SQUARE FEET, with infinite or missing results replaced by the median. Similarly, we created UNITS\_PER\_SQFT as TOTAL UNITS divided by GROSS SQUARE FEET, again imputing problematic values with the median. This comprehensive feature engineering allowed our models to capture location-based, size-based, and temporal effects, which are crucial for accurate real estate pricing predictions. We also observed that newly created features shown correlation with sale price target variable

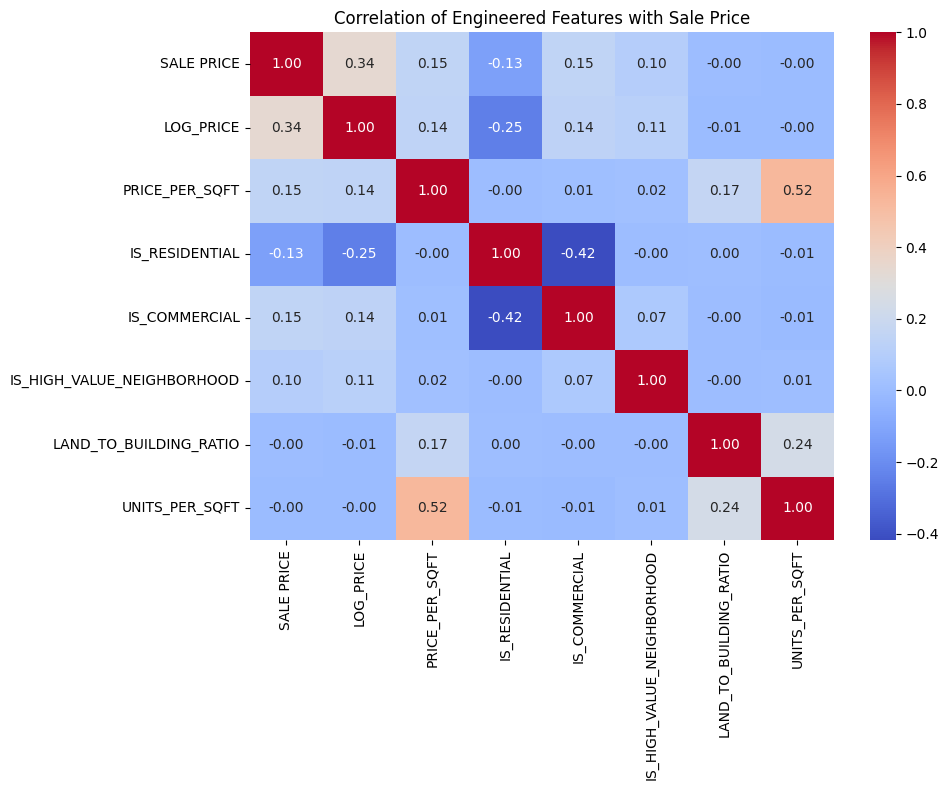


Figure 5 - Correlation matrix showing numerical correlation with target variable

## 4. Exploratory Data Analysis (EDA)

Although the focus here is on methods rather than findings, a preliminary EDA was essential to guide later modeling decisions. We conducted distribution checks through histograms and boxplots for the target variable (SALE PRICE), which revealed extreme outliers and justified our log transformation approach. Borough comparison visualizations, including count plots and summary statistics, helped us understand how many transactions occurred in each borough and their varying price scales. We also examined correlation coefficients among building attributes (land/gross square footage, number of units, building age) and SALE PRICE, ensuring the features selected had meaningful relationships to the target. The findings from this exploratory analysis helped refine our data cleaning strategy, including final cutoffs for outlier removal, and prompted the creation of specialized features like PRICE\_PER\_SQFT and neighborhood value indicators.

## 5. Modeling Approach

Our modeling approach combined careful feature selection with appropriate algorithm choices and evaluation methods. For feature selection, we used a combination of numeric features (e.g., LAND SQUARE FEET, PRICE\_PER\_SQFT) and categorical fields (e.g., BOROUGH, BUILDING CLASS CATEGORY). Categorical variables were one-hot encoded to incorporate them into numeric modeling techniques. To manage computational efficiency, we randomly sampled approximately 20% of the cleaned data for faster prototyping, then applied an 80%/20% train-test split to this sampled dataset to build and evaluate the models. Standard scaling (mean = 0, std = 1) was applied to continuous features to aid algorithms sensitive to feature magnitude.

We tested three candidate models with different capabilities: Linear Regression as a baseline to compare against more complex models; Random Forest Regressor (an ensemble tree method known to handle non-linearities and outliers effectively); and Gradient Boosting Regressor (another ensemble approach that can capture complex patterns via sequentially improved weak learners). For evaluation metrics, we used R-squared (R²) to measure the proportion of variance in sale price explained by each model, and Root Mean Squared Error (RMSE) to quantify the typical size of prediction errors in the same units as the target variable (after inverse transformation from log scale if relevant). These methods, combined with our data cleaning and feature engineering steps, formed a comprehensive pipeline for analyzing and predicting real estate sale prices in NYC.

# IV. Results

## 1. Data Overview and Cleaning Outcomes

Upon loading the dataset (over 600,000 records initially), we identified and addressed several data quality issues. Multiple columns had extensive missing values or were irrelevant for predictive modeling. EASE-MENT contained no valid entries whatsoever (100% missing) and was dropped immediately. APARTMENT NUMBER showed around 75% missing values and was also removed as it provided little predictive value. Census Tract 2020 and NTA Code were missing over 80% of their values, leading to their removal as well. During our inspection, we found that numeric columns stored as strings—particularly LAND SQUARE FEET and GROSS SQUARE FEET—needed conversion after removing formatting characters like commas. Additional columns such as ZIP CODE were validated to ensure consistent data types throughout the analysis.

A critical step in our cleaning process involved filtering out non-market transactions. We removed any properties sold for $10,000 or less, as these typically represent inter-family transfers, partial stake sales, or data entry errors rather than arms-length market transactions. This filtering step substantially reduced the dataset from over 600,000 entries to approximately 414,581 records. After identifying and removing 1,036 exact duplicates, our final cleaned dataset included 413,545 unique sales records for analysis. Throughout the cleaning process, we filled missing numeric values using contextually appropriate methods. For instance, missing values in YEAR BUILT, LAND SQUARE FEET, and GROSS SQUARE FEET were imputed with group-level medians where possible, defaulting to overall medians when necessary. We also standardized borough identifiers, converting various representations (numeric codes, text names, etc.) into a consistent set of five borough categories. The outcome was a clean, consistent dataset with all missing values addressed, providing a solid foundation for our exploratory analysis and modeling.

## 2. Exploratory Data Analysis (EDA)

### 2.1 Distribution of Sales by Borough

Our analysis of transaction distribution across NYC's five boroughs revealed significant differences in market activity. Queens emerged as the borough with the highest volume of property sales, accounting for 123,891 records in our dataset. Brooklyn followed closely with 108,398 transactions, while Manhattan contributed 103,478 sales. Staten Island and the Bronx showed substantially lower activity, with 42,129 and 35,649 sales respectively.

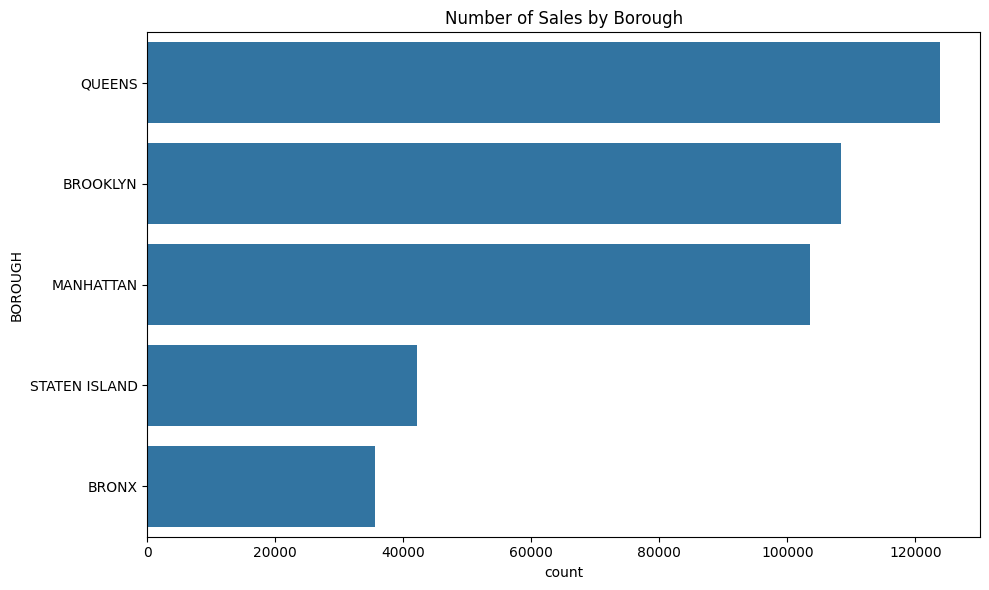


Figure 6 - Distribution of number of sales by across Boroughs

This distribution highlights important market dynamics: Queens and Brooklyn demonstrate very active residential markets with high transaction volumes, while Manhattan, despite having fewer total sales, typically features higher-value properties. The Bronx and Staten Island, with their smaller representation in the dataset, suggest less market turnover or possibly fewer properties overall compared to the other boroughs.

### 2.2 Sale Price Characteristics

After cleaning the data and removing sub-$10,000 transactions, we observed an extraordinarily wide range of property prices across NYC. Sale prices spanned from just above $10,001 to over $2.3 billion for major commercial properties, underscoring the extreme diversity in NYC's real estate market. The median sale price across all boroughs was approximately $712,775, a figure significantly lower than the mean price due to right-skewed distribution. This skewness resulted from a relatively small number of extremely high-priced commercial or luxury transactions, particularly in Manhattan, which pulled the average far above the median. To address this statistical challenge and create more normally distributed data for modeling, we applied a log transformation to the sale price (LOG\_PRICE). This transformation effectively normalized the skewed values, making subsequent modeling procedures more statistically robust and reliable.

### 2.3 Outlier Examination

Our boxplot visualization of SALE PRICE revealed the extreme nature of NYC's property market outliers. The vast majority of properties clustered under $1.5 million, creating a dense concentration at the lower end of the price spectrum. However, the upper range extended dramatically, with extreme outliers in Manhattan office buildings, hotels, and high-end condominiums occasionally surpassing $100 million, with a maximum value around $2.4 billion. These extraordinary outliers were not simply statistical anomalies to be removed – they represent genuine market transactions for large commercial properties, reflecting Manhattan's status as a global investment destination. The presence of these ultra-high-value properties creates modeling challenges but also represents an important segment of the NYC real estate ecosystem that our analysis needed to address.

### 2.4 Borough-Level Price Comparisons

Borough-level boxplot comparisons revealed substantial differences in price distributions across NYC's five boroughs. Manhattan consistently displayed the highest median prices and the widest spread, along with numerous extreme outliers at the upper end. Properties in the Bronx and Staten Island exhibited the lowest median prices and much tighter price ranges, reflecting more homogeneous housing stock and fewer luxury or commercial properties.

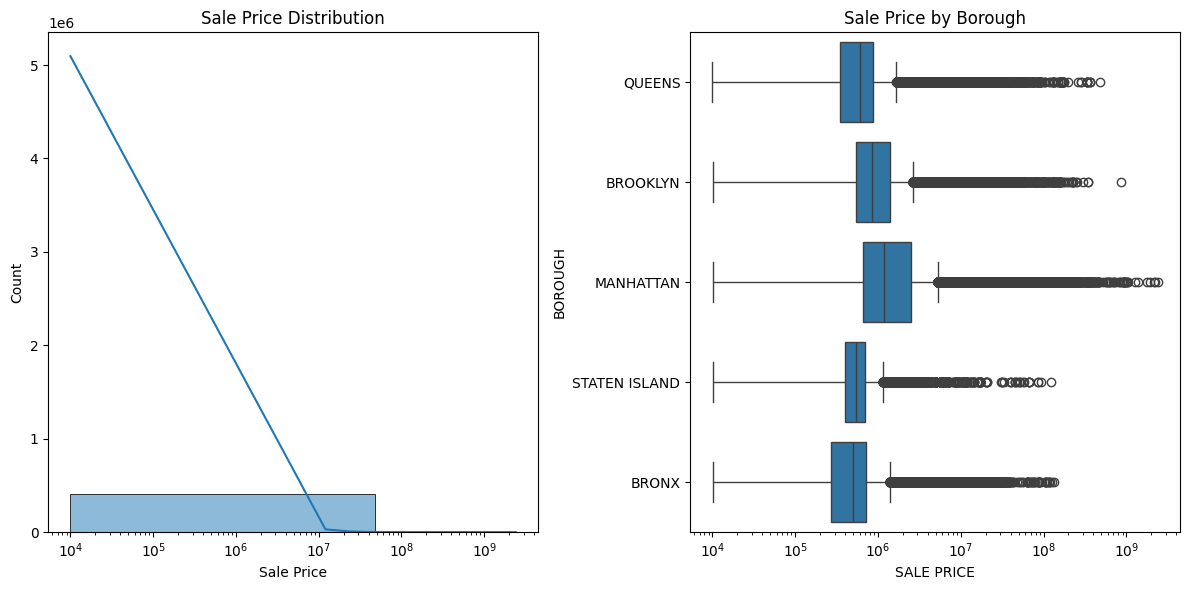


Figure 7 - Distribution of sale price across Boroughs

Queens and Brooklyn occupied the middle ground in terms of median prices, though certain gentrified areas in Brooklyn approached Manhattan's price levels in recent transactions. The stark contrasts observed in these distributions empirically confirmed what local real estate experts often assert – that location at the borough level alone is a major determinant of property values in New York City, even before considering neighborhood, property type, or physical characteristics.

### 2.5 Neighborhood Highlights

Drilling down to the neighborhood level, we identified areas with particularly high average sale prices. Manhattan's central business districts dominated this category, with Midtown CBD, Fashion District, and Financial District showing the highest average transaction values. Interestingly, some less residential or partially industrial areas such as Bloomfield in Staten Island also appeared among the high-average-price neighborhoods, typically due to the presence of a few very large commercial transactions that skewed the average upward.

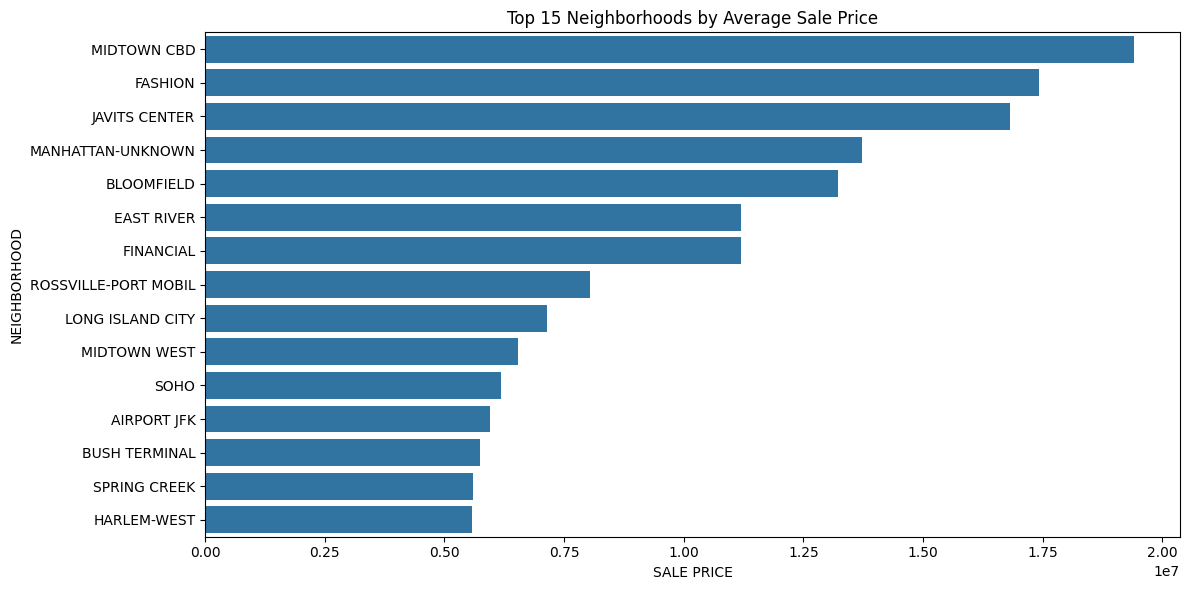


Figure 8 - Top neighbourhoods by average sale price

This analysis underscored the significant pricing influence of Manhattan's core business districts, while also highlighting isolated areas where commercial development or special-purpose properties created local price anomalies. The extreme variation in neighborhood-level prices further reinforced the importance of including location variables in our predictive models.

### 2.6 Time Trends and Seasonal Patterns

Our time-series analysis revealed notable trends and disruptions in NYC's real estate market over the study period. Median sale prices showed a generally rising trajectory from 2016 through 2019, followed by a slight dip coinciding with the initial COVID-19 pandemic impact in 2020. This was succeeded by a substantial rebound into 2021–2022 as the market adjusted to new conditions. Transaction volume followed a similar pattern, with the number of sales dropping significantly during 2020 before recovering by late 2021.

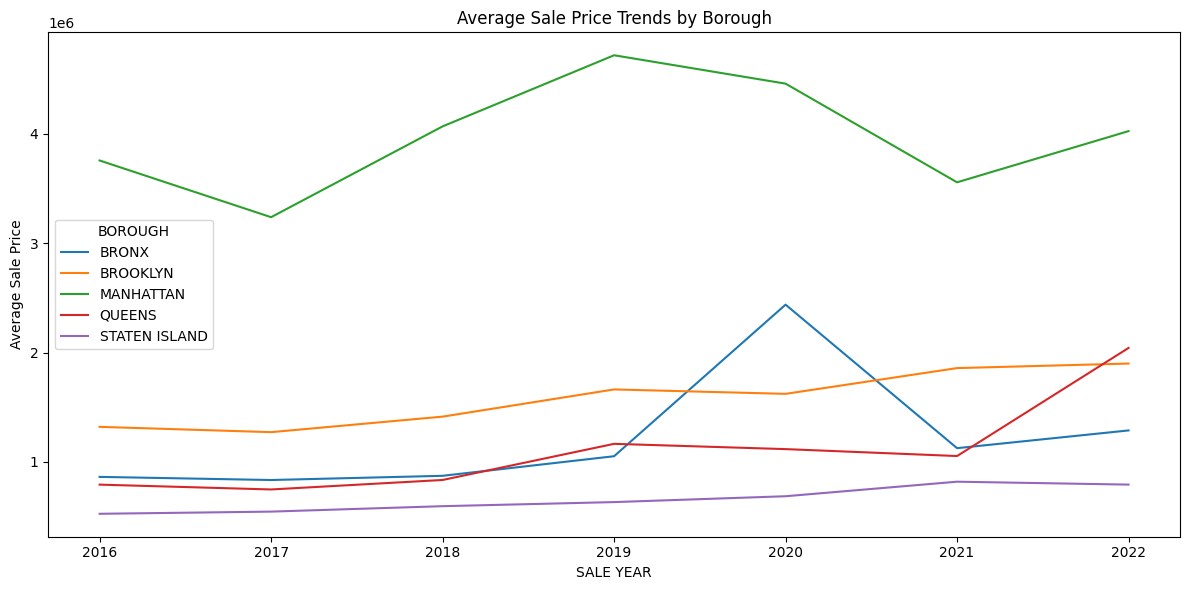


Figure 9 - Average sale price over the years across boroughs

This pattern illustrated the market's sensitivity to external economic shocks. When examining borough-specific trends, we found that Manhattan's prices remained consistently highest throughout the period despite greater volatility. Brooklyn and Queens exhibited steadier growth patterns with less dramatic pandemic-related fluctuations. The Bronx occasionally showed short-term price spikes, typically caused by individual large commercial transactions rather than broad market movements, highlighting the impact that outlier sales can have on smaller submarkets.

### 2.7 Feature Correlations

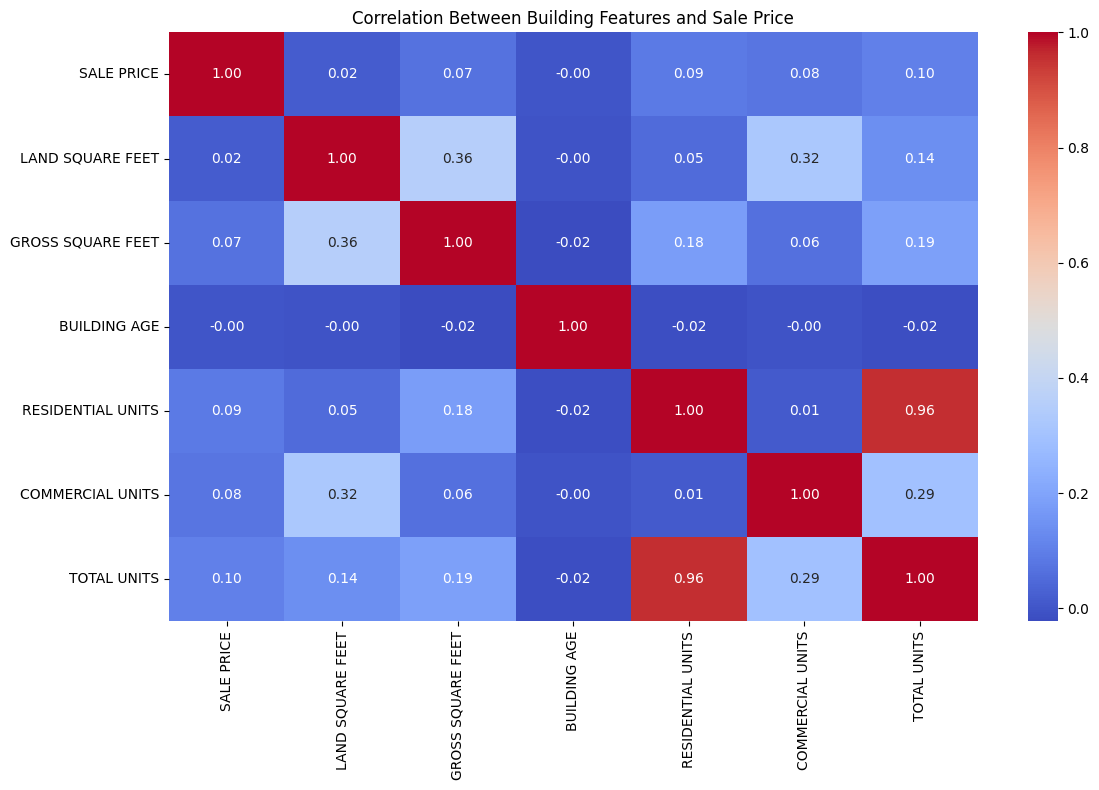
Our correlation analysis of numerical features produced somewhat surprising results. SALE PRICE showed only weak to moderate positive correlations with property characteristics. The strongest correlation was with TOTAL UNITS (approximately 0.10), suggesting a modest relationship between property size measured in units and overall value. 

Figure 10 - Correlation of newly created variables with sale price

GROSS SQUARE FEET showed a correlation of about 0.07 with price, while LAND SQUARE FEET had an even weaker relationship at approximately 0.02. Interestingly, BUILDING AGE demonstrated nearly zero correlation with SALE PRICE in the aggregate data, though we suspected age might still be an important factor within specific neighborhoods or property types.

We observed high intercorrelation among RESIDENTIAL UNITS, COMMERCIAL UNITS, and TOTAL UNITS, which aligned with their structural relationship. The relatively low correlations between physical features and sale price suggested that NYC real estate values are influenced by complex, non-linear factors—including location, building class, and external economic drivers—that do not manifest as simple linear relationships capturable by correlation coefficients alone. This finding reinforced our decision to employ more sophisticated modeling techniques capable of capturing these complex interactions.

## 3. Predictive Modeling Results

### 3.1 Model Training and Evaluation

To streamline our analysis process, we worked with a 20% random sample of the cleaned dataset, which we then divided into training (80%) and test (20%) sets. This approach allowed for faster experimentation while still providing sufficient data volume for reliable model training and evaluation. We tested three distinct modeling approaches with different capabilities:

* Standard Linear Regression as our baseline
* Random Forest Regressor configured with 50 trees and maximum depth of 15, and
* Gradient Boosting Regressor with 50 estimators and maximum depth of 5.

We evaluated these models using two complementary metrics:

* R-squared (R²) to measure the proportion of variance explained, and
* Root Mean Squared Error (RMSE) to quantify prediction error on the log scale.

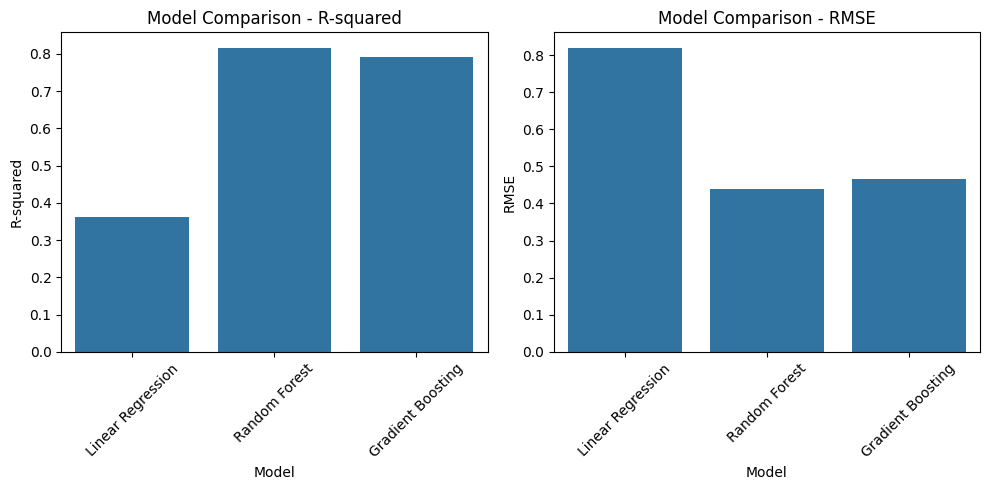


Figure 11 - Comparison of models (R2 and RMSE)

The results showed substantial performance differences between the linear and tree-based approaches. Linear Regression achieved an R² of only 0.36 and RMSE of 0.82, indicating limited predictive power. In stark contrast, the Random Forest model performed remarkably well, explaining approximately 82% of the variation in log-transformed sale prices (R² = 0.82) with a much lower error rate (RMSE = 0.44). Gradient Boosting delivered strong but slightly lower performance, with R² = 0.79 and RMSE = 0.47. These results confirmed that NYC property pricing exhibits substantial non-linearity that cannot be adequately captured by simple linear methods. The superior performance of tree-based ensemble methods highlighted their ability to model complex interactions among location factors, physical building characteristics, and property types—precisely the kind of multi-dimensional relationships that define real estate markets. Based on these findings, we selected the Random Forest model for further analysis and feature importance evaluation.

### 3.2 Feature Importance (Random Forest)

We examined feature importance rankings from our best-performing Random Forest model to identify the key drivers of property price prediction. PRICE\_PER\_SQFT emerged as the most influential feature, indicating the powerful combined signal provided by this engineered ratio of sale price to building size. GROSS SQUARE FEET ranked second in importance, confirming that overall building size significantly impacts total property value, with larger buildings generally commanding higher prices. Borough indicators, particularly the Manhattan dummy variable, showed substantial importance as well, numerically confirming the critical role of location in determining property values that we had observed in our exploratory analysis.

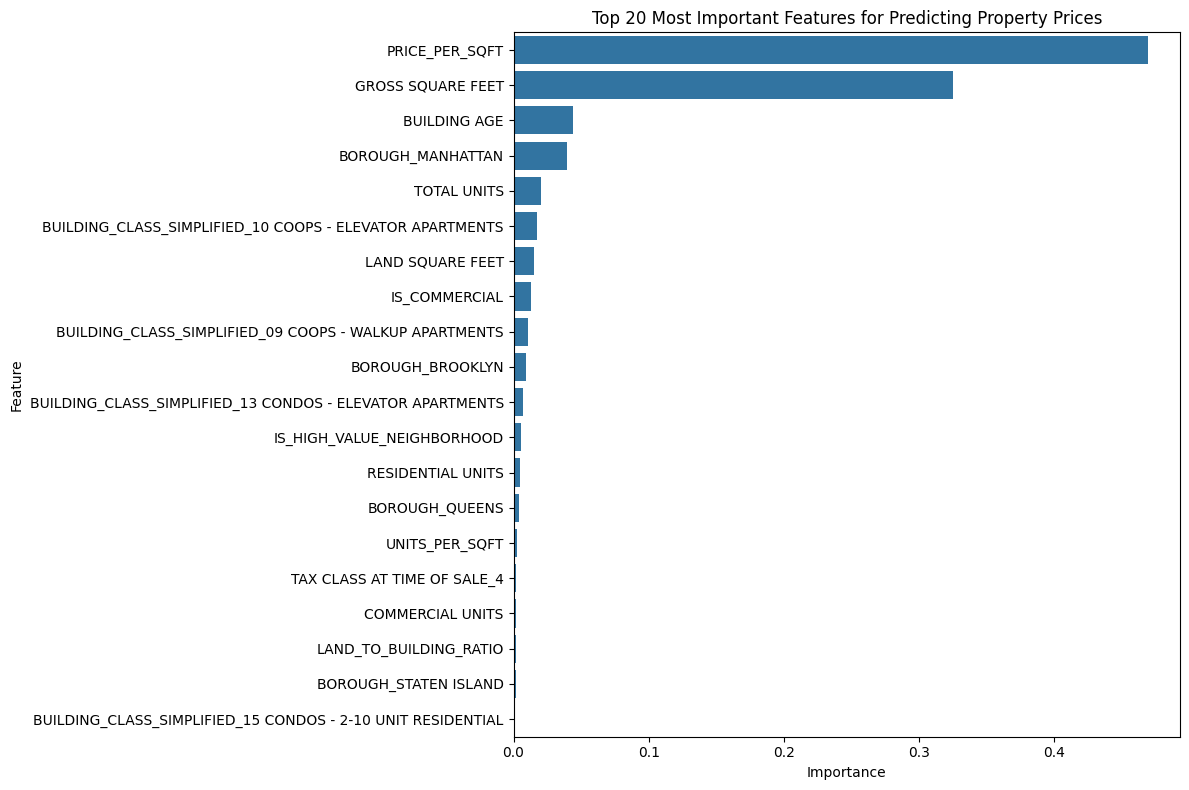


Figure 12 - Comparison of models (R2 and RMSE)

BUILDING AGE showed moderate importance, though less than size and location factors. Other contributors included TOTAL UNITS, LAND SQUARE FEET, and the binary flags we created for commercial and residential properties. Interestingly, while some carefully engineered features such as LAND\_TO\_BUILDING\_RATIO and UNITS\_PER\_SQFT had lower individual importance scores, they still added value to the model by helping distinguish between certain property types and use cases. The clear hierarchy in feature importance aligned well with real estate market fundamentals, where location, size, and property type typically drive valuation, and provided validation that our model was capturing meaningful relationships rather than spurious correlations.

## 4. Borough-Specific Model Performance

To better understand how our Random Forest model performed across NYC's diverse submarkets, we segmented test predictions by borough and evaluated accuracy metrics separately for each area. This analysis revealed substantial performance variations that aligned with market complexity. Queens showed the highest prediction accuracy with an R² of 0.83 and the second-lowest RMSE (0.34). The relatively small percentage errors in Queens (mean error of –7.43% and median error of just 0.05%) suggested that the model was particularly effective at capturing price patterns in this borough, possibly due to its more uniform property types and pricing dynamics.

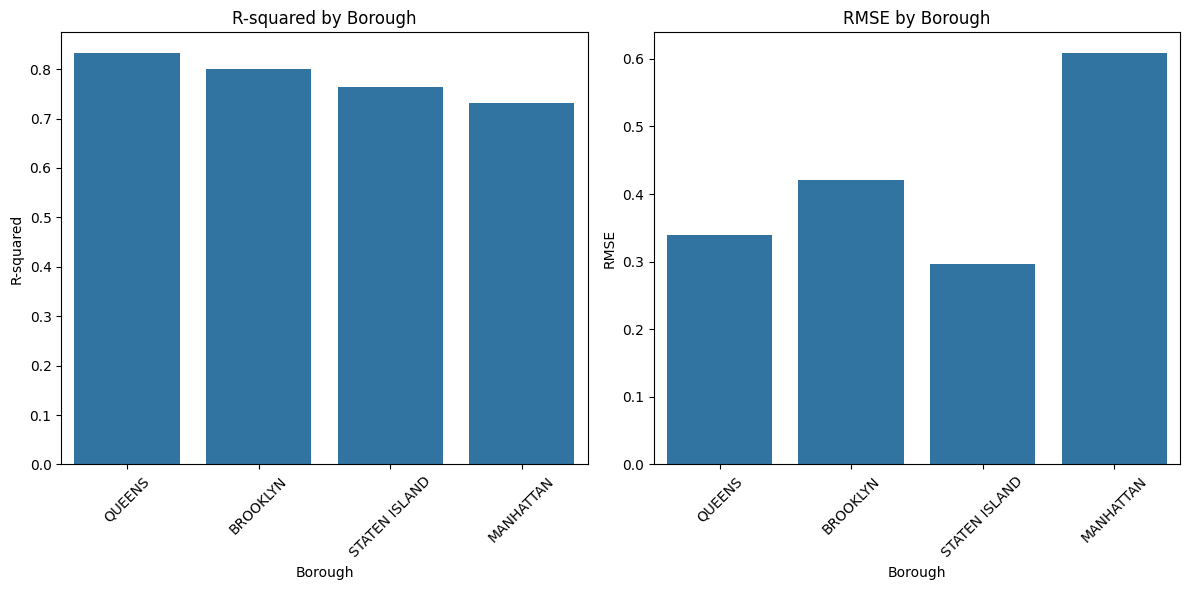


Figure 13 - Comparison of borough specific model performance

Brooklyn followed with solid performance (R² = 0.80, RMSE = 0.42), though the larger mean negative error (–18.80%) indicated a tendency for the model to underpredict in certain higher-value areas of the borough. Staten Island exhibited the lowest absolute error (RMSE = 0.30) despite a slightly lower R² of 0.76, suggesting consistent predictions across its more homogeneous property market with fewer extreme outliers. Manhattan proved the most challenging to model accurately, with the lowest R² (0.73) and highest RMSE (0.61) among all boroughs. The substantial mean percentage error (–31.10%) revealed a systematic underprediction of high-value properties, particularly luxury residences and commercial buildings in premium locations.

These borough-specific results validated our hypothesis that modeling Manhattan's high-end real estate market presents unique challenges. The complexity and extreme range of Manhattan's commercial and luxury property values make standard predictive approaches less effective, suggesting that specialized models focusing on high-value properties or incorporating additional features relevant to premium real estate might yield better results for this segment. The findings also highlighted the potential value of developing borough-specific models rather than applying a single citywide approach.

## Key Takeaways from the Results

Our comprehensive analysis yielded several important insights for NYC property price prediction. First, robust data cleaning and filtering processes proved essential—removing non-market transactions under $10,000 and addressing missing data significantly improved model performance by providing a more representative dataset. Our exploratory analysis revealed the extraordinary diversity in NYC's property market, with Manhattan commanding substantially higher prices while exhibiting greater price volatility compared to outer boroughs. Pandemic-era trends were clearly visible in both price and transaction volume data, demonstrating the market's sensitivity to major economic disruptions.

In comparing modeling approaches, the Random Forest algorithm demonstrated superior predictive performance (R² ~0.82), substantially outperforming linear methods and confirming the benefit of non-linear ensemble techniques for complex real estate data. Feature importance analysis showed that price per square foot, gross building size, and borough location were the strongest predictors, aligning with real estate fundamentals. The borough-specific evaluation revealed varying levels of model effectiveness across NYC's submarkets, with Queens and Brooklyn showing higher predictability while Manhattan's luxury and commercial properties remained more difficult to accurately value. These findings collectively underscore the complex interplay between location, physical building characteristics, and time-based market conditions in determining NYC property sale prices, while pointing toward potential refinements for future modeling efforts.

# V. Discussion and Interpretation

## 1. Addressing the Research Questions

Our first research question asked how effectively machine learning models could predict NYC property sale prices using historical sales data and physical property features. Our best model—Random Forest Regressor—was able to explain about 82% of the changes in log-transformed sale prices. This indicates that physical property characteristics (size, building class) and location are indeed strong predictors of price when using appropriate modeling techniques. This result aligns with previous research (Dong, 2020; Xu, 2025), which similarly found that building features combined with location or socioeconomic factors yield strong predictive power. However, even with this success, very high-end or unusual properties in Manhattan remain challenging to predict accurately.

Our second research question explored what factors have the strongest influence on property sale prices across different NYC boroughs. The results, particularly the feature importance analysis, confirm that location (borough and neighborhood) plays the dominant role in determining property values, followed by physical characteristics like gross square footage and number of units. This is consistent with the real estate maxim of "location, location, location" and supports findings from studies like Ma et al. (2020), which showed that high-value areas like Midtown CBD and the Financial District command significant price premiums. Our borough-level analysis further revealed that Manhattan's market exhibits greater complexity with higher dispersion in prices, while Queens and Brooklyn, with their more homogeneous housing stock, showed smaller gaps between predicted and actual prices.

For our third research question on how predictive model performance varies across boroughs, we found significant differences that suggest specialized modeling approaches might be beneficial. The Random Forest model performed best in Queens (R² = 0.83) and worst in Manhattan (R² = 0.73), with significantly higher error rates for Manhattan properties. This suggests that luxury and commercial properties display different characteristics compared to other real estate types, requiring either specialized models or additional features to capture their unique pricing dynamics.

## 2. Interpretation of Key Findings

**Role of Location:**

The model clearly showed that borough and neighborhood variables have substantial importance for price prediction, demonstrating the critical role location plays in NYC real estate. Even after accounting for physical features like building size and age, location remains the primary driver of value. This supports earlier findings by Arul & Morales (n.d.) and Xu (2025). The systematic underprediction of Manhattan prices further highlights the unique characteristics of luxury or special-purpose buildings in this borough.

**Property Size and Layout:**

Gross square footage and total units showed consistent relationships with sale price, while land square footage was less influential, suggesting that in dense urban environments like NYC, vertical space and unit count are more valuable than raw land area. This parallels findings by Ho, Tang, and Wong (2021) in the similarly dense urban environment of Hong Kong.

**Engineered Features:**

Features like PRICE\_PER\_SQFT substantially improved prediction accuracy, indicating that price-to-size relationships explain more variance than raw price alone. Interestingly, BUILDING AGE showed little correlation with price, though it might be more relevant in specific contexts like historic districts or new developments. This mixed importance pattern is consistent with other studies such as Alharbi (2023) and Varma et al. (2018), which suggest that more granular data (e.g., proximity to amenities or building condition) may be needed to better capture the effect of building age.

**Model Comparisons:**

Linear Regression's poor performance (R² of 0.36) highlights the non-linear nature of NYC's real estate market with its diverse property types. In contrast, Random Forests and Gradient Boosting performed substantially better by capturing non-linear patterns and handling outliers. This aligns with machine learning research findings (Park & Bae, 2015; Calainho, van de Minne, & Francke, 2024) showing tree-based ensemble models consistently outperform traditional regression approaches when predicting real estate prices.

## 3. Connection to Broader Trends

The time-series analysis revealed fluctuations in both sales volume and median price, particularly around 2020 and early 2021, suggesting that macroeconomic events (such as the COVID-19 pandemic) had significant impacts on transaction patterns. This aligns with observations by Xu (2025). The models' superior performance in more stable markets like Queens and Staten Island versus Manhattan suggests that local market dynamics significantly influence predictability. This reinforces the need for potentially region-specific or even neighborhood-specific models to capture localized trends.

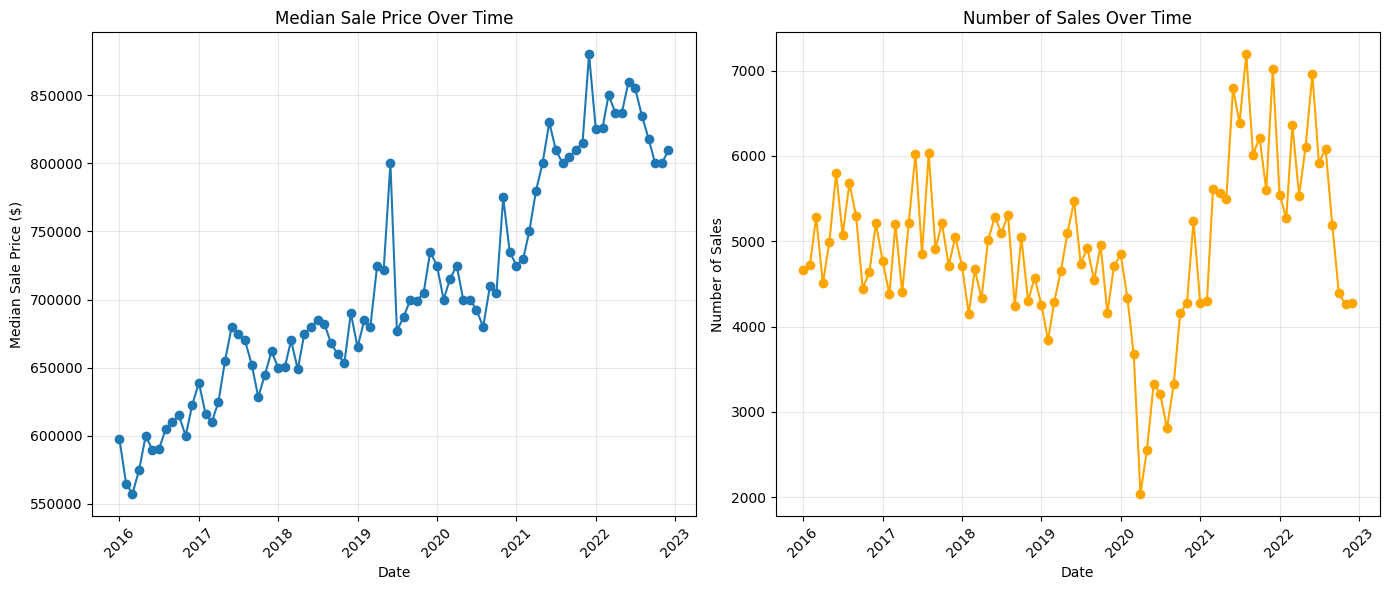


Figure 14 - Median sale price and number of sales over the years

## 4. Practical Implications

From a policy and investment perspective, the results underscore that location is the primary driver of property values, which can guide urban planning decisions about where to develop or improve infrastructure (such as transit options). Real estate investors, financial institutions, and government agencies can utilize models like our Random Forest approach to improve price estimates and risk assessment. However, the lower accuracy for Manhattan properties indicates that additional data for luxury and commercial sales would be beneficial, potentially including macroeconomic indicators, global investment trends, and more detailed building characteristics (e.g., amenities or historical significance).

## 5. Limitations and Caveats in Interpretation

While our model explains a substantial portion of price variance, it doesn't incorporate factors like school quality, crime rates, or walkability scores, which can significantly impact property values. Additionally, the dataset may have missing or erroneous records (especially for ultra-high-end properties with private transactions), so predictions may not generalize well to all luxury sales. Finally, our data represents a point-in-time snapshot for each sale, making it challenging to predict future trends without more sophisticated time-series methodologies.

Overall, the results demonstrate that combining detailed location data, physical property features, and advanced modeling techniques provides strong predictive power for NYC property prices. However, the market's heterogeneity across boroughs necessitates continued refinement of models—particularly for Manhattan—through additional data sources and specialized modeling approaches.

# VI. Limitations

**Data Quality and Missing Information**

Even though we removed wrong entries (like sale prices under $10,000) and columns with too many missing values, the dataset might still have some mistakes or old information. Some sales may not show the normal market value (like foreclosures or when only part of a property is sold).

**Skewed and Extreme Sales**

The NYC real estate market has some very expensive commercial and luxury properties, especially in Manhattan. These outliers can change the overall results and make the model guess too low or too high in some areas. Using special models just for these expensive properties might help fix this issue.

**Lack of External Socioeconomic or Demographic Factors**

We mainly used property-level data (like square footage, year built, and location). We did not include bigger economic factors like interest rates, income levels in the area, school quality, or crime rates. These factors can affect prices a lot. Not including them can make the model less accurate and less useful.

**Cross-Sectional Approach**

We looked at monthly trends, but we didn't build a full time-series or future prediction model. Our study shows patterns only up to the date of sale. It cannot guess future market changes. Big events like economic downturns, new laws, or gentrification can change prices in ways our model cannot predict.

**Uneven Borough Representation**

Some boroughs (like Queens) have a lot more data than others (like the Bronx). This difference can affect how well the model learns. Manhattan is also special because it has both very expensive commercial and residential properties. Our general model may not fully handle this mix, and a more focused approach might be better.

# VII. Recommendations and Future Work

**Include Additional Location and Demographic Data**

In the future, this model can be improved by adding data like school ratings, crime rates, access to public transport, and number of nearby businesses. These kinds of details have helped improve predictions in other real estate research.

**Neighborhood-Specific or Borough-Specific Modeling**

Each borough—Manhattan, Queens, Brooklyn, the Bronx, and Staten Island—has its own market situation. So, creating separate models for each borough may give better results. For example, a special model just for expensive commercial properties in Manhattan may help handle very high price cases better.

**Time-Series and Forecasting Approaches**

Instead of looking at property sales as one-time events, we can study how prices change over time. This can be done using methods like ARIMA, LSTM networks, or time-series regression. These methods can show trends, seasonal changes, and economic patterns over time.

**Integrate Macroeconomic Indicators**

Things like unemployment rates, interest rates, economic growth, and tourism (which affects commercial properties) may help explain price changes. Adding these can reduce errors and improve predictions, especially during unstable times.

**Explore Ensemble and Hybrid Models**

Even though Random Forest and Gradient Boosting gave good results, combining different models (like stacking) or adding deep learning can make predictions even better. Ensemble methods help cover the weaknesses of single models and manage NYC’s complex market more strongly.

**Outlier Detection and Specialized Treatment**

We can use advanced tools to find outlier transactions—those that are much higher than usual. Handling these cases separately or marking them as a different group might help the model give better results, especially for very expensive properties.

**Automate and Update the Pipeline**

As new property sale data comes in, automating the steps to add and clean this data will help retrain the model regularly. This helps keep the model updated with recent trends like changes after COVID or new popular areas.

By working on these points, future research and real-world uses can better deal with the complicated NYC market and give more accurate and useful insights for buyers, sellers, and policymakers.

# Conclusion

To sum up, this study shows that predicting NYC property prices works well with advanced ensemble methods, strong data cleaning, and smart feature choices. Location (borough or neighborhood) is very important, especially since Manhattan's expensive properties have big price swings. Our Random Forest model explains most of the price differences, but we can still improve it by adding more detailed social and economic data. The model does not do as well with very expensive properties, which means we may need special models or extra features—like details about building facilities or local economic data—especially for commercial and large deals. In the future, using time-based methods, economic data, and separate models for each borough can help make better and more useful predictions. This can give helpful insights for anyone involved in NYC real estate.

# References

Alharbi, Z. H. (2023). A sustainable price prediction model for Airbnb listings using machine learning and sentiment analysis. Sustainability, 15(17), 13159.<https://doi.org/10.3390/su151713159>

Arul, H., & Morales, A. (n.d.). NYC condo price estimation using NYC open data. [Unpublished manuscript].

Calainho, F. D., van de Minne, A. M., & Francke, M. K. (2024). A machine learning approach to price indices: applications in commercial real estate. The Journal of Real Estate Finance and Economics, 68(4), 624–653.

City of New York. (2023, September 2). *NYC citywide annualized calendar sales update* [Dataset]. NYC OpenData. <https://catalog.data.gov/dataset/nyc-citywide-annualized-calendar-sales-update>

Dong, D. (2020, December). Factors influencing housing prices: An empirical analysis from New York City. In Proceedings of the 2020 3rd International Conference on E-Business, Information Management and Computer Science (pp. 106–113).

Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48–70.

Luo, Y., Zhou, X., & Zhou, Y. (2019). Predicting Airbnb listing price across different cities. Highlights in Science and Engineering Technology.

Ma, J., Cheng, J. C. P., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. Land Use Policy, 94, 104537. https://doi.org/10.1016/j.landusepol.2020.104537

New York Post, 2025. Here’s where real estate prices are dropping in NYC this year. *New York Post*, 22 March. Available at:<https://nypost.com/2025/03/22/real-estate/heres-where-real-estate-prices-are-dropping-in-nyc-this-year-reports/>

New York Post, 2025. Progressive politicians make it hard to add Midtown housing. *New York Post*, 29 March. Available at:<https://nypost.com/2025/03/29/opinion/progressive-politicians-make-it-hard-to-add-midtown-housing/>

NYC Comptroller, 2023. *Spotlight: New York City’s Housing Supply Challenge*. Office of the NYC Comptroller. Available at:<https://comptroller.nyc.gov/reports/spotlight-new-york-citys-housing-supply-challenge/>

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018, April). House price prediction using machine learning and neural networks. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1936–1939). IEEE.

Xu, R. (2025). Comprehensive data-driven analysis of NYC real estate transactions: Insights into pricing dynamics, neighborhood evolution, and socioeconomic influences. [Unpublished manuscript].

Zillow, 2025. *New York NY Home Values*. Zillow. Available at:<https://www.zillow.com/home-values/6181/new-york-ny/>