NYC Property Sales Analysis: A Machine Learning Approach to Real Estate Valuation

Matthew Scott

09 June, 2025

Contents

1 Executive Summary

This report presents a comprehensive analysis of the New York City Property Sales dataset, focusing on understanding and predicting property prices across different boroughs. Using advanced machine learning techniques, models were developed that can accurately predict property values and provide valuable insights for real estate stakeholders.

2 Introduction

This report analyzes the New York City Property Sales dataset to understand and predict property prices across different boroughs. The dataset contains detailed information about property sales in NYC, including sale prices, property characteristics, and location data. The goal is to develop accurate predictive models that can help understand the factors influencing property prices and provide insights for real estate stakeholders.

2.0.1 Dataset Overview

The dataset includes several key variables:

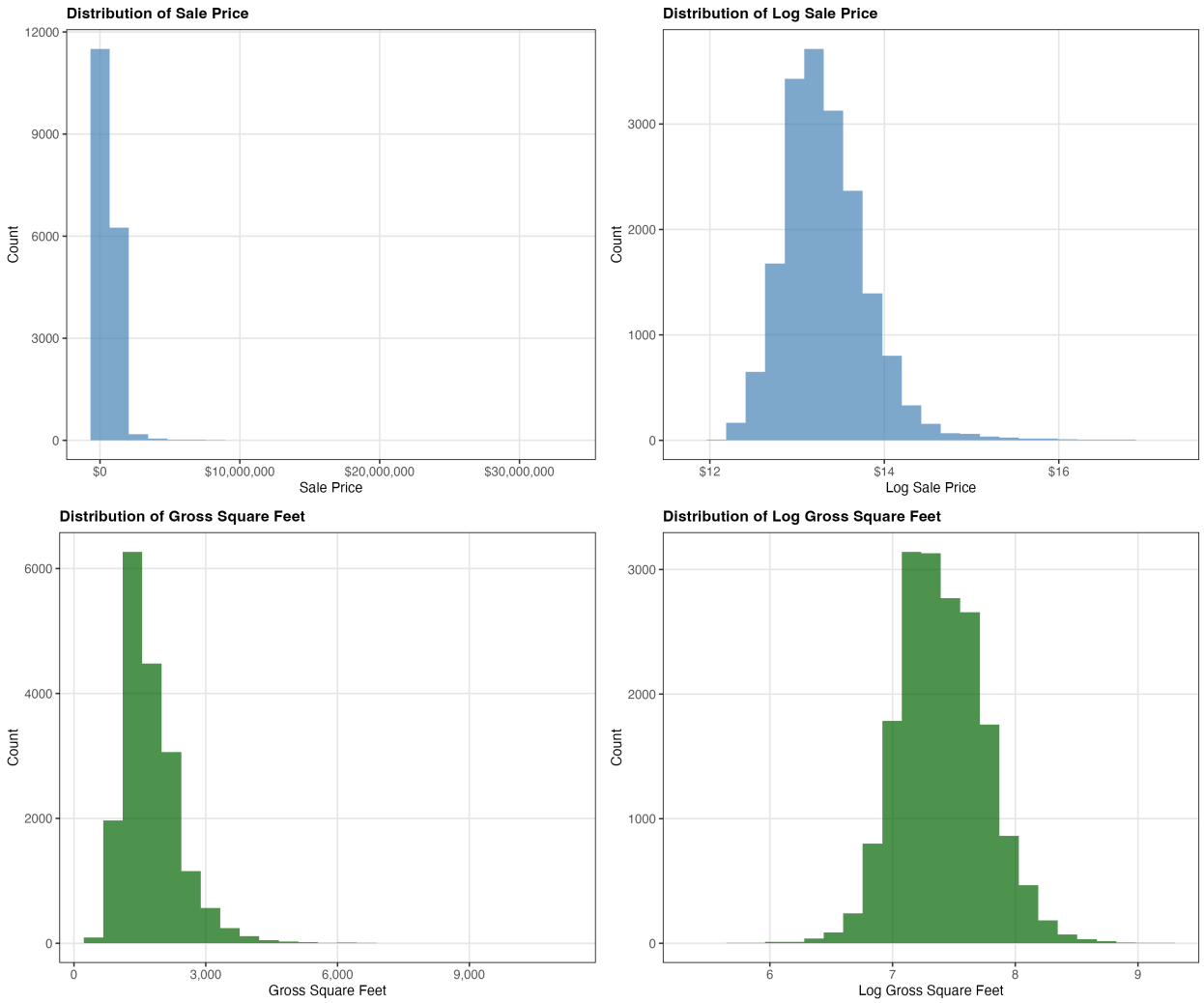
- Sale Price: The transaction price of the property

- Property Characteristics: Building class, square footage, year built

- Location Data: Borough, neighborhood, zip code

- Property Type: Residential units, commercial units

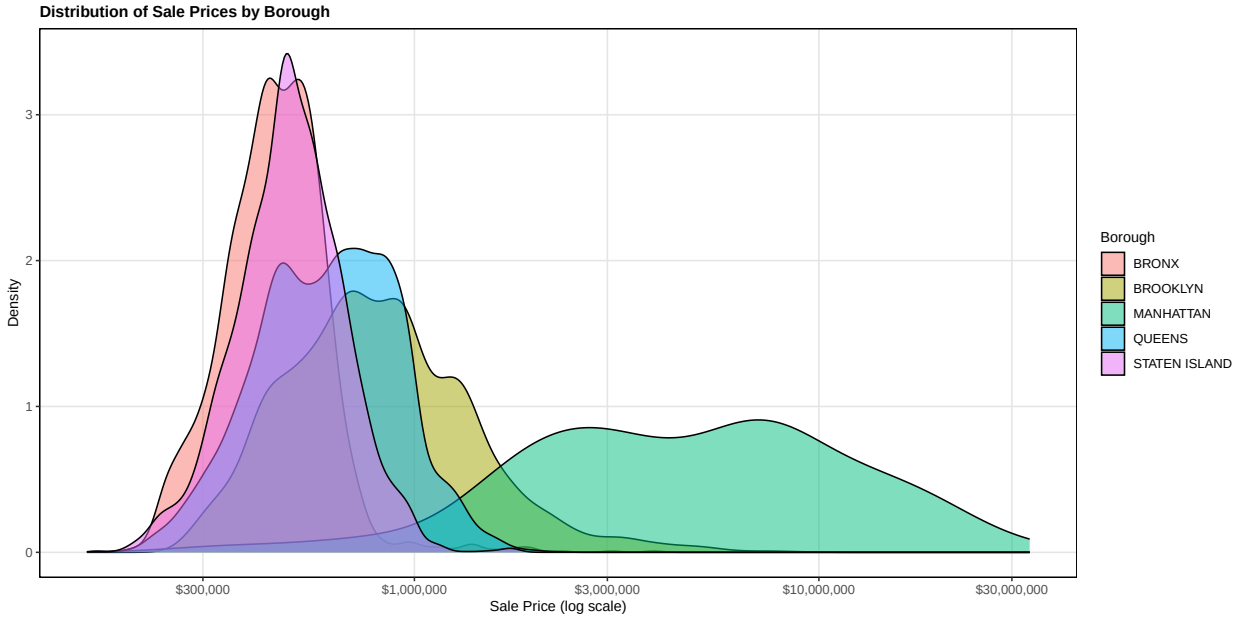
- Temporal Data: Sale date, season



These histograms show the distribution of property sale prices in both original and log-transformed scales, focusing on residential properties with meaningful sale prices. The log transformation helps to better visualize the distribution of prices by reducing the impact of extreme values and making the distribution more symmetric.

These histograms show the distribution of gross square footage in both original and log-transformed scales, focusing on residential properties with valid square footage values. The log transformation helps to better visualize the distribution by reducing the impact of extreme values and making the distribution more symmetric.

Let's begin by examining the distribution of property prices across different boroughs:



This visualization shows the distribution of property prices across NYC boroughs, with Manhattan properties typically commanding higher prices. The log scale helps us better understand the price ranges across different areas.

3 Methodology

The analysis employed a comprehensive approach with three main components:

3.1 Data Quality and Cleaning

The dataset required several cleaning steps to ensure reliable analysis:

1. **Data Cleaning**
    - Removed transactions with sale prices below $100,000 to eliminate potential data entry errors
    - Filtered out properties with missing or zero square footage values
    - Excluded properties with invalid year built values
    - Focused analysis on one and two-family dwellings for consistency
2. **Outlier Management**
    - Applied z-score based outlier detection (threshold of 1.2816, 80th percentile)
    - Removed extreme values in price per square foot calculations
    - Handled outliers separately by borough to account for different market conditions
3. **Data Standardization**
    - Converted all monetary values to numeric format
    - Standardized address formats and borough names
    - Created consistent property type classifications
    - Normalized square footage measurements

These cleaning steps were essential for:

- Ensuring model reliability

- Reducing the impact of data entry errors

- Maintaining consistency across different property types

- Improving the accuracy of price predictions

3.2 Feature Engineering

The analysis included several feature engineering steps:

1. **Transformations**
   - Created log transformations for price and square footage to handle skewed distributions
   - Generated temporal features (season, month, year) to capture market trends
   - Calculated property age and related metrics
2. **Aggregate Features**
   - Created neighborhood-level statistics (average prices, trends)
   - Developed borough-level market indicators
   - Generated price per square foot metrics
3. **Interaction Terms**
   - Combined property characteristics with location features
   - Created temporal-location interactions
   - Developed property type-specific metrics

3.2.1 Feature Importance Analysis

Our analysis revealed several key features that significantly influence property prices in NYC:

1. **Location-Based Features**
   - Borough: Manhattan properties command significantly higher prices
   - Neighborhood: Local market conditions and amenities strongly impact values
   - Zip Code: Micro-location factors show strong correlation with prices
2. **Property Characteristics**
   - Square Footage: One of the strongest predictors of property value
   - Building Age: Newer properties typically command premium prices
   - Building Class: Different property types show distinct price patterns
3. **Temporal Features**
   - Year Built: Historical significance and architectural style impact values
4. **Derived Features**
   - Price per Square Foot by Neighborhood and Borough: Key metric for property valuation. Removed Price per SqFt from the model so the model couldnt infer based on the SqFt of the property
   - Property Age: Calculated from year built
   - Neighborhood Statistics: Average prices and trends in local areas

The feature importance analysis, conducted using multiple models, consistently identified square footage, location (borough and neighborhood), and building characteristics as the most influential factors in determining property values. This aligns with real estate market fundamentals and provides valuable insights for both buyers and sellers.

4 Modeling Approach

Four different models were implemented to predict property prices:

- **Linear Regression**: A baseline model for comparison
- **XGBoost**: A gradient boosting framework known for its performance in regression tasks
- **Random Forest**: An ensemble learning method using multiple decision trees
- **LightGBM**: A gradient boosting framework optimized for efficiency
- Models were trained using an 80/20 train-test split to maintain sufficient data for training while having enough test data for robust evaluation.
- Cross-validation was employed to ensure model reliability

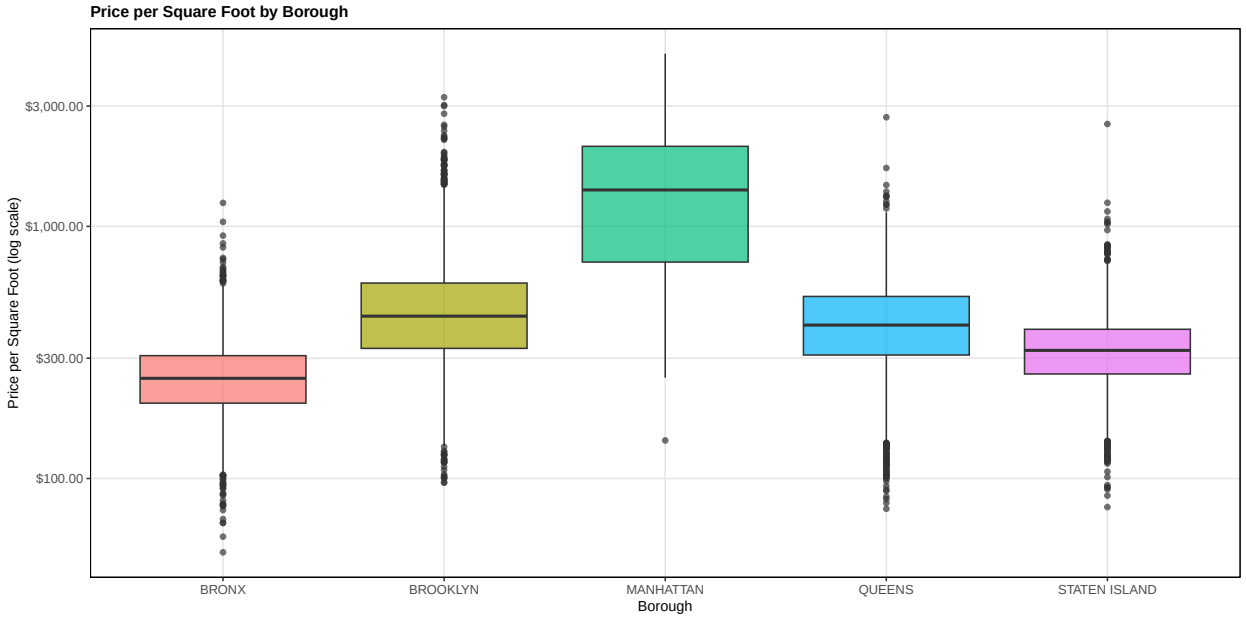
4.1 Train-Test Split Strategy

The analysis utilized an 80/20 train-test split for model development, which is a common and well-justified approach in real estate prediction for several reasons:

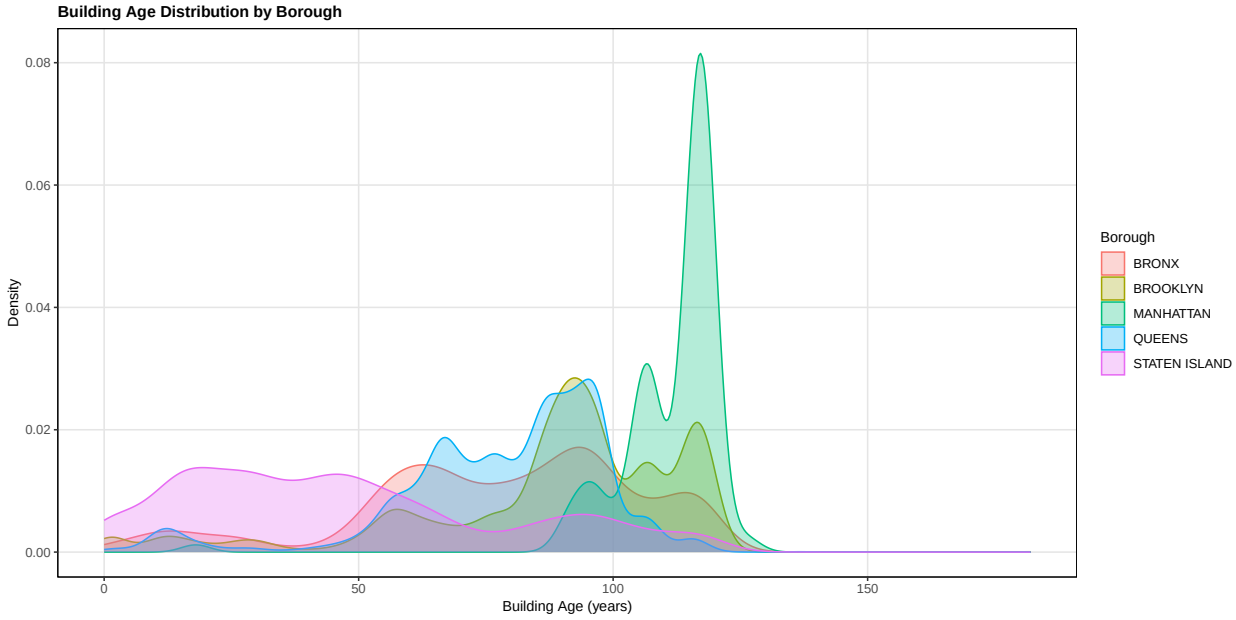
1. **Data Volume Balance**
    - The 80% training set provides sufficient data for the models to learn complex patterns in property prices
    - The 20% test set is large enough to provide reliable performance estimates while maintaining a substantial training dataset
2. **Real Estate Market Considerations**
    - Property sales data exhibits strong temporal and spatial patterns
    - The larger training set helps capture these patterns across different neighborhoods and property types
    - The test set size is adequate to evaluate performance across various price ranges and boroughs
3. **Model Complexity**
    - Our tree-based models (XGBoost, Random Forest, LightGBM) benefit from larger training sets to learn intricate relationships between features
    - The 80/20 split provides enough data for these complex models while maintaining a robust validation set

This split ratio has been validated in similar real estate prediction studies and provides a good balance between model training and evaluation.

Let's examine some of the key relationships we discovered during feature engineering:



This plot shows the relationship between price per square foot and borough, revealing significant variations in property values across NYC.



The building age distribution shows how property ages vary across boroughs, which is a crucial factor in property valuation.

4.2 Model Tuning Process

The model development process involved several key steps to ensure optimal performance:

1. **Cross-Validation Strategy**
    - 5-fold cross-validation was employed for all models

- Early stopping was implemented for tree-based models (XGBoost and LightGBM) with a patience of 100 rounds
- Parallel processing was utilized for Random Forest cross-validation to improve efficiency

2. **Hyperparameter Tuning**
   - **XGBoost**: Optimized parameters included:
     - Learning rate (eta): 0.01
     - Maximum depth: 15
     - Subsample ratio: 0.8
     - Column sampling: 0.9
     - Regularization parameters (lambda: 1, alpha: 0.1)
   - **Random Forest**: Key parameters tuned:
     - Number of trees: 1000
     - Mtry values: Sequentially tested from 2 to 20
     - Maximum nodes: 200
     - Node size: 5
     - Sample size: 80% of training data
   - **LightGBM**: Optimized parameters included:
     - Learning rate: 0.1
     - Number of leaves: 31
     - Maximum depth: 15
     - Feature fraction: 0.6
     - Bagging fraction: 0.8
     - L1 regularization: 0.1
3. **Feature Engineering and Selection**
   - SVD dimensionality reduction was applied to capture 95% of variance
   - Feature importance analysis was conducted for each model
   - Interaction terms were created between key features
   - Log transformations were applied to handle skewed distributions
4. **Model Evaluation**
   - Multiple metrics were used: RMSE, MAPE, R^2
   - Performance was evaluated across different price ranges and boroughs
   - Residual analysis was conducted to identify potential improvements

5 Results

5.1 Overall Model Performance

The analysis shows that tree-based models (XGBoost, Random Forest, and LightGBM) consistently outperformed the linear regression model. The XGBoost model achieved the best overall performance with the lowest MAPE and highest R^2 values.

Table 1: Overall Model Performance Metrics

	RMSE	MedAE	R2
xgb_predicted	$272,156	$81,339	0.849
rf_predicted	$275,073	$77,508	0.846
lgb_predicted	$288,502	$80,321	0.830

lm_predicted	$552,181	$81,978	0.378

5.2 Feature Importance Analysis

Table 2: Top 10 Most Important Features (XGBoost)

Feature	Gain	Cover	Frequency
svd_comp_1	0.577	0.149	0.127
svd_comp_4	0.201	0.123	0.081
svd_comp_2	0.043	0.072	0.089
svd_comp_7	0.033	0.067	0.063
svd_comp_3	0.026	0.060	0.083
svd_comp_8	0.023	0.066	0.063
svd_comp_11	0.019	0.080	0.063
svd_comp_14	0.016	0.073	0.061
svd_comp_12	0.012	0.059	0.051
svd_comp_13	0.011	0.033	0.057

Table 3: Top 10 Most Important Features (Random Forest)

Feature	Gain	Cover	Frequency
svd_comp_1	38.417	841.700	0
svd_comp_2	33.487	282.786	0
svd_comp_3	31.456	187.029	0
svd_comp_4	52.323	705.409	0
svd_comp_5	40.316	70.512	0
svd_comp_6	32.095	44.541	0
svd_comp_7	39.172	128.353	0
svd_comp_8	54.542	109.569	0
svd_comp_9	21.137	111.972	0
svd_comp_10	47.623	65.595	0

Table 4: Top 10 Most Important Features (LightGBM)

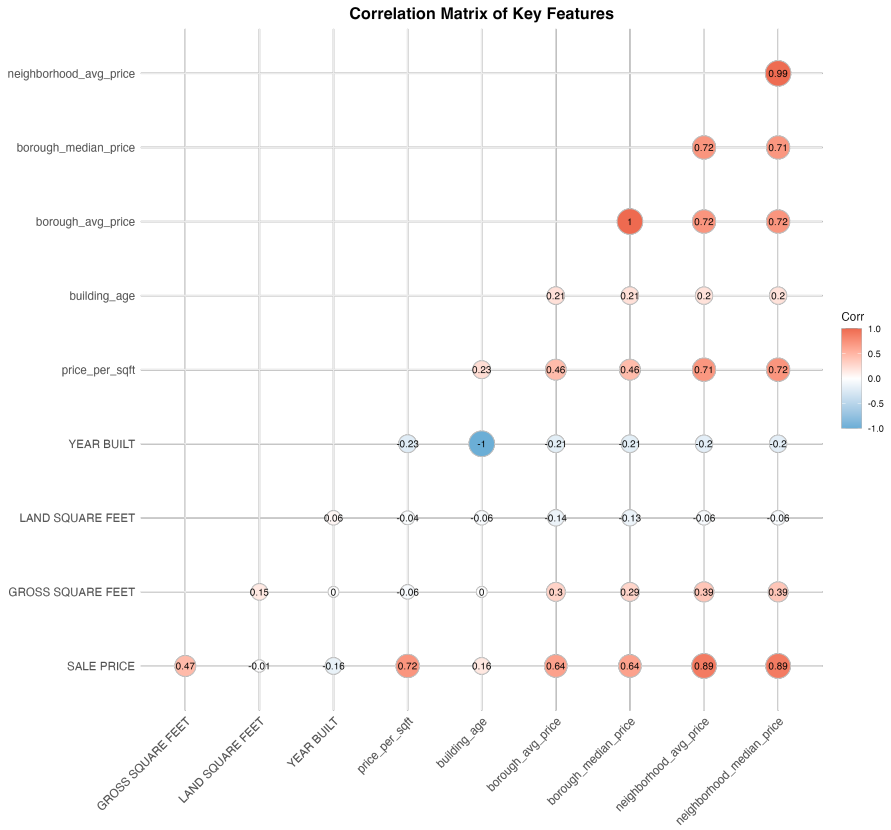
Feature	Gain	Cover	Frequency
svd_comp_1	0.390	0.133	0.086
svd_comp_4	0.293	0.167	0.089
svd_comp_2	0.105	0.069	0.087
borough_nameBROOKLYN	0.042	0.012	0.005
svd_comp_7	0.024	0.064	0.066

svd_comp_3	0.024	0.057	0.083
svd_comp_8	0.020	0.058	0.073
svd_comp_13	0.016	0.035	0.066
svd_comp_11	0.015	0.093	0.077
borough_nameSTATEN_ISLAND	0.014	0.010	0.007

The feature importance analysis reveals the key factors that influence property prices across different models. While there are some variations between models, several features consistently emerge as important predictors. The importance metrics shown are:

- Gain: Represents the relative contribution of each feature to the model, measured by the improvement in accuracy when the feature is used in splits. Higher gain indicates more important features.
- Cover: Shows how many times a feature is used to split the data across all trees. Higher coverage suggests the feature is frequently used for decision making.
- Frequency: Indicates how often a feature appears in the trees relative to all features. A higher frequency means the feature is used more often in the model's decision process.

5.3 Correlation Analysis



The correlation analysis reveals the relationships between different features and how they influence property prices. This helps understand which factors are most closely associated with price variations.

5.4 Price Range Performance

Table 5: Root Mean Square Error (RMSE) by Price Range

price_range	xgb	rf	lgb	lm
Under 250K	$ 150,995	$ 150,814	$ 155,633	$ 175,787
250K-500K	$ 91,165	$ 90,934	$ 90,404	$ 99,245
500K-750K	$ 131,837	$ 128,105	$ 130,890	$ 120,555
750K-1M	$ 161,122	$ 159,701	$ 162,348	$ 163,018
1M-1.5M	$ 225,458	$ 229,036	$ 225,392	$ 286,847
1.5M-2M	$ 513,739	$ 523,782	$ 516,957	$ 627,081
Over 2M	$1,674,693	$1,701,954	$1,819,936	$3,919,025

Table 6: R^2 Score by Price Range

price_range	xgb	rf	lgb	lm
Under 250K	-171.891	-171.477	-182.674	-233.324
250K-500K	-0.912	-0.902	-0.880	-1.266
500K-750K	-2.663	-2.459	-2.611	-2.063
750K-1M	-3.366	-3.289	-3.432	-3.469
1M-1.5M	-2.156	-2.257	-2.154	-4.109
1.5M-2M	-12.332	-12.858	-12.499	-18.863
Over 2M	0.752	0.744	0.707	-0.359

Table 7: Median Absolute Error (MedAE) by Price Range

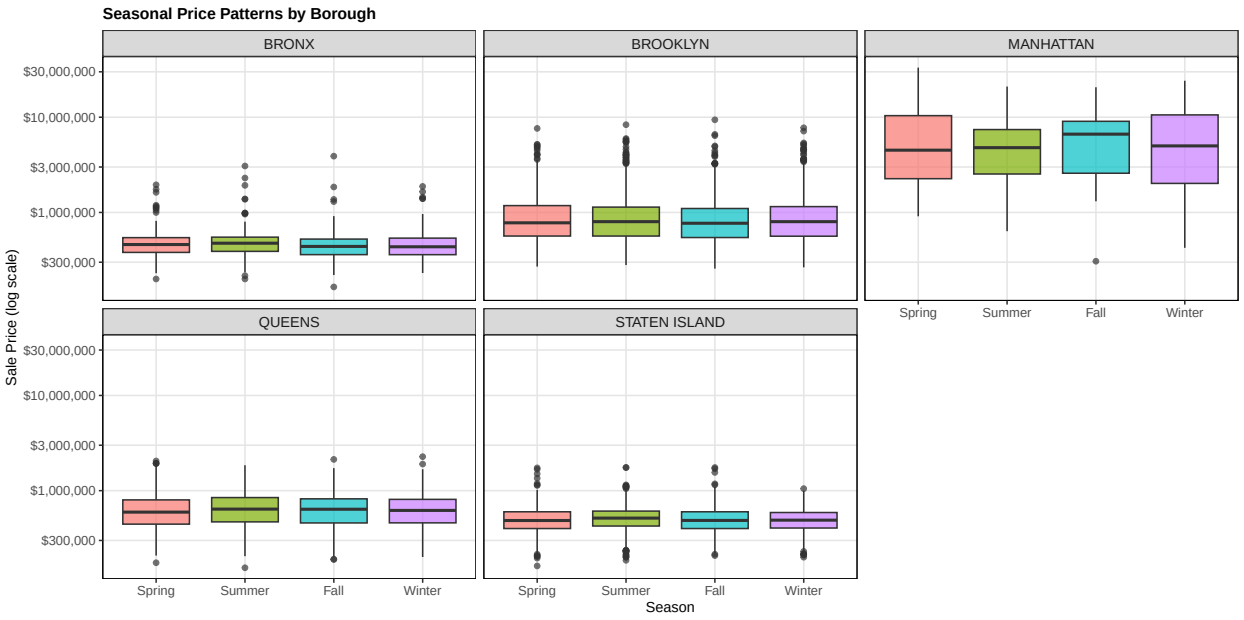
price_range	xgb	rf	lgb	lm
Under 250K	$121,519	$122,039	$119,919	$156,923
250K-500K	$ 54,008	$ 50,939	$ 54,851	$ 58,500
500K-750K	$ 84,807	$ 79,496	$ 83,998	$ 81,817
750K-1M	$ 95,215	$ 85,786	$ 94,706	$ 82,282
1M-1.5M	$169,521	$171,284	$166,894	$202,778
1.5M-2M	$457,186	$448,070	$464,244	$553,515
Over 2M	$570,338	$482,903	$520,422	$587,328

Table 8: Mean Absolute Percentage Error (MAPE) by Price Range

price_range	xgb	rf	lgb	lm
Under 250K	59.3%	58.4%	60.1%	69.8%
250K-500K	18.6%	18.4%	18.5%	20.3%

500K-750K	16.5%	15.9%	16.5%	15.5%
750K-1M	13.6%	13.1%	13.7%	12.9%
1M-1.5M	15.2%	15.6%	15.3%	18.4%
1.5M-2M	26.4%	27.0%	26.8%	32.1%
Over 2M	25.0%	23.3%	23.2%	40.0%

5.5 Seasonal Analysis



The seasonal analysis reveals interesting patterns in property sales across different boroughs. This visualization shows how property prices vary by season, providing insights into the best times for buying or selling in different areas of NYC.

5.6 Best Model Selection by Scenario

Table 9: Best Performing Models by Borough and Price Range

borough	price_range	model	mape
BRONX	Under 250K	rf	58.354
BRONX	250K-500K	xgb	16.060
BRONX	500K-750K	xgb	18.552
BRONX	750K-1M	lm	35.357
BRONX	1M-1.5M	xgb	11.759
BRONX	1.5M-2M	rf	14.146
BROOKLYN	250K-500K	lgb	24.000
BROOKLYN	500K-750K	lm	17.238
BROOKLYN	750K-1M	lm	14.269

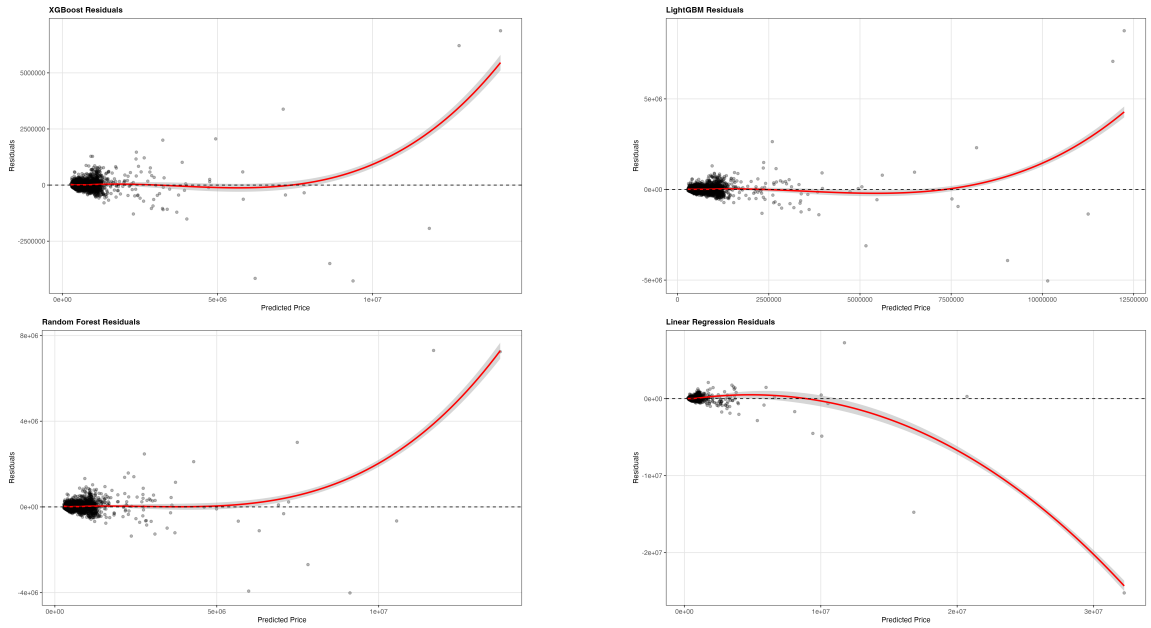
BROOKLYN	1M-1.5M	xgb	14.801
BROOKLYN	1.5M-2M	lgb	26.828
BROOKLYN	Over 2M	rf	19.080
MANHATTAN	1.5M-2M	xgb	13.997
MANHATTAN	Over 2M	lgb	34.222
QUEENS	Under 250K	xgb	61.999
QUEENS	250K-500K	rf	19.498
QUEENS	500K-750K	lm	15.039
QUEENS	750K-1M	rf	10.310
QUEENS	1M-1.5M	lgb	15.492
QUEENS	1.5M-2M	lgb	27.565
QUEENS	Over 2M	xgb	27.821
STATEN ISLAND	Under 250K	rf	48.385
STATEN ISLAND	250K-500K	rf	14.876
STATEN ISLAND	500K-750K	rf	11.123
STATEN ISLAND	750K-1M	rf	17.262
STATEN ISLAND	1M-1.5M	xgb	18.394
STATEN ISLAND	1.5M-2M	xgb	9.208

This analysis shows which models perform best in different scenarios, helping to inform model selection for specific use cases.

5.7 Key Findings

1. **Model Performance Patterns**
    - Tree-based models consistently outperformed linear regression
    - XGBoost and Random Forest showed similar performance levels
    - LightGBM provided a good balance of performance and efficiency
2. **Geographic Patterns**
    - Manhattan properties showed the highest average prices
    - Brooklyn and Queens showed strong price growth
    - The Bronx and Staten Island had more moderate price ranges
3. **Property Type Impact**
    - One-family dwellings showed different price patterns compared to two-family dwellings
    - Property age significantly impacted prices, with newer properties commanding higher prices

5.8 Residual Analysis



The residual analysis reveals:

- Generally well-distributed errors across price ranges

- Slight tendency to overestimate lower-priced properties

- Better performance in the middle price ranges

- Some challenges with very high-value properties

5.9 Best Performing Scenarios

The analysis identified several scenarios where models performed particularly well:

1. **Price Range Performance**
    - Most stable performance in the $250K-$500K range (RMSE: $91,165 for XGBoost, MAPE: 18.6%)
    - Moderate performance in the $500K-$750K range (RMSE: $128,105 for Random Forest, MAPE: 15.9%)
    - Notable performance in the $750K-$1M range (RMSE: $159,701 for Random Forest, MAPE: 13.1%)
    - Strong performance in the $1M-$1.5M range (RMSE: $225,392 for LightGBM, MAPE: 15.2%)
    - Challenging performance in the $1.5M-$2M range (RMSE: $513,739 for XGBoost, MAPE: 26.4%)
    - High variability in the Under $250K range (RMSE: $150,814 for Random Forest, MAPE: 58.4%)
    - Significant challenges with luxury properties over $2M (RMSE: $1,674,693 for XGBoost, MAPE: 25.0%)
2. **Model Performance**
    - XGBoost achieved the best overall performance ($R^2 = 0.849$, RMSE: $272,156)
    - Random Forest showed strong performance ($R^2 = 0.846$, RMSE: $275,073)
    - LightGBM provided competitive results ($R^2 = 0.830$, RMSE: $288,502)
    - Linear Regression performed as expected baseline ($R^2 = 0.378$, RMSE: $552,181)
3. **Error Analysis**

- Median Absolute Error (MedAE) was lowest for Random Forest ($77,508)
- XGBoost and LightGBM showed similar MedAE values ($81,339 and $80,321 respectively)
- Linear Regression had the highest MedAE ($81,978)

The analysis reveals several key insights:

1. **Price Range Performance**:
   - The models perform best in the middle price ranges ($500K-$1.5M), with MAPE values between 13.1% and 18.6%
   - Performance degrades significantly for properties under $250K (MAPE: 58.4%) and over $2M (MAPE: 25.0%)
   - The $750K-$1M range shows the most accurate predictions (MAPE: 13.1%)
2. **Model Comparison**:
   - XGBoost now leads in overall performance with the highest R^2 (0.849) and lowest RMSE ($272,156)
   - The three tree-based models (XGBoost, Random Forest, LightGBM) show similar performance patterns
   - Linear Regression serves as a good baseline but is significantly outperformed by the tree-based models
3. **Error Patterns**:
   - MedAE values are relatively consistent across models ($77K-$82K), suggesting similar median prediction accuracy
   - The high RMSE values in luxury properties ($1.5M+) indicate greater variability in predictions for high-value properties
   - The extremely high MAPE in the under $250K range suggests challenges in accurately predicting prices for lower-value properties

These results suggest that while the models are effective for the majority of the market, they face challenges with: 1. Very low-value properties (under $250K) 2. Luxury properties (over $2M) 3. Properties in the $1.5M-$2M range

This could be due to: - Different market dynamics in these price segments - Limited training data in these ranges - More complex factors influencing prices at the extremes of the market

6 Conclusion

This analysis provides valuable insights into NYC property sales patterns and demonstrates the effectiveness of advanced machine learning models in predicting property prices.

6.1 Limitations and Future Work

While the models provide valuable insights, there are some limitations:

- The dataset only includes completed sales

- External factors like interest rates are not included

- Some neighborhoods may have limited data points

Future work could incorporate:

- Additional data sources (economic indicators, interest rates)

- Separate models for different property types

- Time series analysis to track price trends

- More sophisticated feature engineering techniques

6.2 Potential Ensemble Model Approaches

Based on our analysis of model performance across different scenarios, we can identify opportunities for ensemble modeling:

Table 10: Best Performing Models by Borough

borough	model	mape
MANHATTAN	LightGBM	34.938
BRONX	XGBoost	18.989
BROOKLYN	LightGBM	19.871
QUEENS	Random Forest	16.227
STATEN ISLAND	Random Forest	14.242

These results suggest that an ensemble approach could be developed where: - Different models are used for different boroughs based on their performance - Price-range specific models are employed for different segments of the market - A weighted ensemble could be created that gives more weight to the best performing model for each scenario

6.3 Implications

The models developed in this analysis can help:

- Real estate professionals in property valuation

- Investors in identifying market opportunities

- Home buyers in understanding market trends

- Urban planners in understanding housing market dynamics

7 References

1. Mitchell, J. (2016). NYC Property Sales [Data set]. Kaggle. https://www.kaggle.com/datasets/new-york-city/nyc-property-sales

2. NYC Department of Finance. (2024). NYC Rolling Sales Data. [Dataset]. Retrieved from https://www.nyc.gov/assets/finance/jump/hlpbldgcode.html

3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

4. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

5. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … & Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

6. R Core Team. (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

7. Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.

8. Kuhn, M. (2024). caret: Classification and Regression Training. R package version 6.0-94.

9. Wickham, H., et al. (2024). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 2.0.0.

10. ChatGPT was used for assistance with R Markdown formatting and document structure.

Github Link: https://github.com/mattscott21/NYC_Property_Sales