# Beyond the “Forgotten Borough”: Forecasting Staten Island Housing Prices with Machine Learning

Nikoleta N Emanouilidi

## Abstract

This capstone project develops a machine learning-based forecasting framework to analyze and project housing price dynamics in Staten Island, New York, a borough that remains underrepresented in housing market research despite its growing importance within the New York City real estate landscape. Using residential sales data from 2017–2025, the study integrates property-level characteristics, neighborhood quality-of-life indicators (crime rates and school quality), construction activity and national macroeconomic conditions, including inflation and mortgage rates.

Three predictive models, Linear Regression, k-Nearest Neighbors (kNN) and eXtreme Gradient Boosting (XGBoost), are developed and evaluated. XGBoost is selected as the final forecasting model due to its superior predictive accuracy, stability and ability to capture nonlinear relationships. Feature importance analysis confirms that structural housing characteristics dominate price formation, while crime and school quality exert strong secondary effects.

A scenario-based forecasting approach grounded in the economic conditions of 2017-2021 is used to generate housing price projections for 2026-2030. Forecast results indicate substantial spatial heterogeneity across Staten Island. While most Neighborhood Tabulation Areas (NTAs) exhibit positive relative price growth, the magnitude and trajectory of appreciation vary considerably, with stronger gains concentrated in southern and interior neighborhoods and more moderate growth elsewhere. Neighborhood-level analysis further illustrates how price trajectories differ across housing types, with persistent premiums for two-unit properties in most areas and localized exceptions driven by neighborhood-specific dynamics..

To enhance real-world usability, the project also introduces an interactive R Shiny forecasting dashboard featuring a personalized NTA preference quiz and neighborhood-level price exploration tools. This application transforms static forecasts into a practical decision-support platform for homebuyers, investors, planners and policymakers. Overall, the study demonstrates how machine learning, spatial analysis and scenario-based forecasting can be combined to support informed housing market analysis and urban decision-making.

## Table of Contents

# CHAPTER 1-INTRODUCTION

## 1.1 Background: New York City Housing Market Overview

New York City, often called the capital of the world, has long captured global attention with its striking skyline, luxurious apartments and iconic red-brick townhouses. Characterized by persistently high demand and strong neighborhood-level disparities, the city's real estate market has proven to be remarkably resilient, consistently rebounding from economic downturns and maintaining its reputation as a "bold choice" for long-term investment (MLS Campus, 2025). Despite facing economic pressures such as rising interest rates and inflation, recent analyses indicate that the market continues to demonstrate "surprising strength," with its underlying fundamentals remaining stable (FCIQ, 2024).

## 1.2 Staten Island as an Emerging Housing Submarket

While most studies and public talk focus on Manhattan, Brooklyn or Queens, this project turns its attention to the often-overlooked borough of Staten Island. Frequently referred to as the "forgotten borough," Staten Island offers an interesting mixture of suburban calm and urban accessibility, creating a unique submarket within the broader New York City housing ecosystem. According to the New York City Comptroller's Office (2024), Staten Island was one of the few boroughs to gain population during the pandemic as residents left high-density areas in search of space and affordability. PropertyShark's analysis of IRS migration data (2024) reinforces this shift, showing that more than 700 former Brooklyn residents and more than 700 former Manhattan residents relocated to Staten Island in a single year, unusually high numbers of incoming movers at a time when other boroughs were experiencing substantial net losses. Moreover, industry reporting shows that this trend has not only persisted but strengthened. TJV News (2025) notes that 12% of all Brooklyn homebuyers who purchased within New York City in early 2025 chose Staten Island, driven by a widening affordability gap between the two boroughs. With Brooklyn's median home price reaching $850,000 and Staten Island's average house price being $708,000, the island offers nearly a 17% discount while providing larger, detached homes, private outdoor space and quieter neighborhoods. ClosedByMo (2025) further underscores this shift, reporting that Staten Island realtors still attribute roughly a quarter of their current sales to Brooklyn buyers, particularly in bridge-accessible communities like New Dorp, Bay Terrace and Tottenville. These combined demographic and market patterns reveal how Staten Island's lower density, greater housing value and suburban character have transformed it into a preferred destination for households unable to afford homes in Brooklyn's current market but determined to remain within New York City. The borough's emerging identity as a semi-suburban refuge is now reflected not only in migration patterns but also in the ongoing rise of local home prices.

## 1.3 Research Gap, Political Economy and Neighborhood Context

Yet, despite its distinctive character, the Staten Island housing market remains underrepresented in housing research. Predicting its future track is both challenging and valuable, for current or future residents making homeownership decisions, investors seeking returns and policymakers trying to shape policies and decisions that ensure fair growth and opportunity.

This capstone project aims to forecast Staten Island's housing prices over the next five years (2026-2030) using a scenario-based approach grounded in the historical precedent of the 2017-2021 presidential term. The research also seeks to identify the influence of neighborhood-level quality-of-life factors, specifically crime rates and school quality, on property values. Accordingly, this study is organized around two related objectives. First, the project will build and evaluate machine-learning models capable of predicting housing prices across Staten Island's neighborhoods using historical data. Second, after selecting the best-performing model, the study will use it to generate a forward-looking forecast for 2026-2030 under a 2017-2021 based economic scenario, with a specific focus on quantifying how neighborhood crime levels and school quality drive the predicted changes in property values.

This project explores the intersection of housing market performance and political governance patterns in the context of political economy. Han and Shin (2021) established a connection between housing values and the political party in power, noting that voters often associate Republican administrations with rising property prices. Shi (2024) builds on this by examining what happens in counties that experience a "policy shock," meaning their local election results conflict with the national presidential outcome. In these places, the housing market does not consistently rise or fall. Instead, it becomes temporarily more volatile in an way that buyers hesitate, transactions fluctuate and purchase timing becomes less predictable. In other words, the instability Shi identifies reflects increased short-term uncertainty rather than a clear directional change in prices. These patterns suggest that presidential election cycles can introduce brief periods of volatility into local housing markets, offering a useful backdrop for scenario-based forecasting.

Forecasting housing prices, however, is an inherently difficult task. As management theorist Peter Drucker famously said, "Trying to predict the future is like trying to drive down a country road at night with no lights while looking out the back window." The complexity of housing markets amplifies this challenge, as prices are influenced by interdependent factors ranging from national policies and macroeconomic trends to local dynamics such as school quality, transport accessibility and crime rates (Gibbons & Machin, 2008). More research further highlights the role of these neighborhood-level factors. Acolin et al. (2021) show that residential density interacts with the surrounding neighborhood environment in meaningful ways. In their models, areas with stronger neighborhood amenities, such as higher transit accessibility, a diverse mix of nearby retail, safer streets, better-performing schools, more green space and greater walkability, tend to demonstrate higher home values. On the other hand, higher density in locations lacking these amenities can often depresses home values, particularly in areas where higher density corresponds with congestion or smaller living spaces. Leonard et al. (2016) similarly found that appraisal-based measures of neighborhood quality demonstrate spatial patterns in key neighborhood attributes, including school quality, property upkeep, environmental conditions and access to employment and retail centers. Together, these studies clarify not only that neighborhood factors matter, but *how* they influence home valuations through concrete characteristics.

## 1.4 Methodological Framework

This study follows a comparative machine-learning framework in which multiple predictive models are evaluated to forecast housing prices in Staten Island, with the final specification selected based on performance. Within this framework, XGBoost emerges as the primary modeling approach due to its ability to capture complex, nonlinear relationships in high-dimensional data through sequential tree boosting. Prior research supports its effectiveness in real estate forecasting: Mora-García et al. (2022) demonstrated that XGBoost achieves among the highest predictive accuracies in housing price applications, while Sharma et al. (2024) found that it outperforms alternative algorithms such as K-Nearest Neighbors and Random Forests.

## 1.5 Scope, Limitations and Contribution of the Study

The primary limitation of this research is that the forecast is scenario dependent. The final result is not a single prediction of what will happen, but rather a forecast of what is *likely* to happen if the economy follows the chosen historical precedent. The model cannot account for future “black swan” events, such as unforeseen wars, pandemics or other economic shocks,  that fall outside the parameters of the historical data.

In conclusion, by developing a robust, data-driven forecasting model, this project contributes to a deeper understanding of Staten Island’s housing market within both local and national economic contexts. The findings will offer practical insights for residents, investors and policymakers, while demonstrating how machine learning tools can improve scenario-based real estate forecasting. The following sections review relevant literature, describe the data and methodology, present the model’s results, and discuss their implications for understanding Staten Island’s housing dynamics and future market outlook.

# CHAPTER 2- LITERATURE REVIEW

## 2.1 Political Context and Housing Market Conditions

The political environment can shape economic conditions relevant to real estate markets, particularly through its influence on macroeconomic policy, taxation and market expectations. Changes in federal leadership are often associated with shifts in national economic priorities that affect housing demand, investment behavior and price dynamics. As a result, presidential party affiliation is frequently used in housing market analyses as a contextual indicator of broader economic regimes rather than as a direct determinant of housing supply or affordability.

Prior research has documented correlations between political leadership and housing market performance. Han and Shin (2021) found that housing booms benefit governments across the political spectrum but are strongest under right-wing administrations. They attribute this pattern to the fact that homeowners, who directly benefit from rising housing prices, are more likely to lean politically conservative, reinforcing an observed association between political leadership and price appreciation. Importantly, this relationship does not imply that federal policy is designed to directly expand housing supply or reduce housing costs, nor that all market participants share aligned economic interests.

Indeed, property owners, developers and possible buyers often have competing objectives within housing markets. Rising prices benefit existing homeowners, but developers are driven mainly by construction costs, regulations and local market conditions. Therefore, higher housing prices do not benefit all housing market participants in the same way.

Moreover, the impact of political parties on housing outcomes depends heavily on the level of government involved. While federal administrations influence housing markets primarily through macroeconomic conditions, tax policy and financial markets, the most direct controls over housing supply, such as zoning regulations, land-use policies and permitting processes, are determined at the state and local levels. These subnational regulatory frameworks vary widely and often operate independently of national party platforms.

Consistent with this distinction, empirical evidence on local political effects is mixed. Some studies suggest that local leadership can influence housing production, for example, findings that Democratic mayors are associated with increased multifamily construction (de Benedictis-Kessner et al., 2023). Other research, however, finds no systematic relationship between party affiliation and local housing supply, suggesting that local officials are more strongly guided by resident preferences and institutional constraints than by national partisan agendas (Ferreira & Gyourko, 2023).

Given these considerations, this project does not attempt to resolve competing political explanations of housing market behavior. Instead, the 2017-2021 Republican presidency is used as a reference period because it represents the most recent completed national economic regime with observed housing market outcomes. From a forecasting perspective, modeling housing prices under a Republican presidency reflects an empirical choice based on historical patterns rather than a claim about deregulation, housing supply policy or voter preferences. This approach keeps the analysis focused on predictive performance, while acknowledging that local factors, such as crime rates and school quality, remain the more stable and influential determinants of neighborhood-level housing outcomes.

## 2.2 Macroeconomic Drivers of Housing Prices

Macroeconomic factors also play a key role in shaping housing prices. Ding (2022) analyzed U.S. housing data from 2005 to 2020 and found that GDP growth (which is the annual percentage increase in the total goods and services produced in a nation) and stock market performance push housing prices upward, while higher mortgage rates and unemployment drive them down. Interestingly, population growth was not found to have a significant effect, suggesting that in mature markets like the United States, financial conditions matter more than demographic trends. These results highlight why including national variables such as the Consumer Price Index (CPI), 30-Year Fixed Mortgage Rate as well as the population growth in this project's forecast is essential, even when focusing on a single borough like Staten Island. The borough's housing trends are still influenced by national economic conditions.

## 2.3 Crime, Neighborhood Safety and Housing Values

Beyond national political influences, local quality-of-life factors remain key determinants of housing values. The link between neighborhood safety and property values is foundational in housing research. In a seminal study on New York City, Schwartz et al. (2003) found that the city's large post-1994 drop in violent crime had a significant positive effect on property values, accounting for approximately one-third of the real price appreciation during the 1994 -1998 boom.

More recent research shows that how crime is perceived and concentrated in certain areas can matter more than the overall crime rate. Ceccato and Wilhelmsson (2019) found that being closer to a crime hot spot lowers property prices, even when the area's total crime rate is controlled for. They highlighted vandalism as especially impactful, since it visibly signals neighborhood decline and creates a "fear of crime" that reduces property values.

This relationship, however, is highly complex and non-linear. Research by Kallberg & Shimizu (2025) and Dentler & Rossi (2024) highlights the challenges of reverse effects ,where high-priced areas may attract certain types of crime and demonstrates that crime's impact is not uniform. Kallberg & Shimizu (2025), for instance, found the negative impact of crime was greatest in the lowest-priced residences. Similarly, Dentler and Rossi (2024) report that crime's effects are very local, fading quickly across both time and space, and influence not only prices but also market activity, such as the frequency of sales.

In conclusion, these findings confirm that a simple linear model is insufficient for this task. This research validates the inclusion of crime as a core explanatory variable and underscores the need for a granular, non-linear model (like XGBoost) that can capture these complex, spatially specific neighborhood effects.

## 2.4 Limitations of Traditional Forecasting Models

Traditional models such as linear regression and ARIMA have long been used in housing price forecasting and can perform adequately in predictive settings. However, housing markets are characterized by complex dynamics, interacting predictors and non-linear relationships that can challenge models built on simple functional forms. As model complexity increases, the risk of overfitting also becomes a central concern. Prior research shows that traditional approaches may struggle to capture the high-dimensional and non-linear structure present in housing market data, particularly when many correlated predictors are involved (Fu, 2024; Yan, 2024; Yu, 2024). Machine learning methods offer greater flexibility by allowing such patterns and interactions to be learned directly from the data. This flexibility can improve predictive performance, but it also increases susceptibility to overfitting if not carefully managed.

As emphasized by McElreath (2020), all modeling tools are heuristic and involve tradeoffs between flexibility and generalizability. Techniques such as regularization, model comparison and ensembling can help mitigate overfitting, but they provide no guarantees. Importantly, models that perform well in prediction are not necessarily appropriate for interpretation or policy evaluation, which require stronger assumptions and more careful causal reasoning.

## 2.5 Superiority of Machine Learning and Ensemble Models

Comparative analyses within the machine learning field overwhelmingly conclude that tree-based ensemble models (like Random Forest and Gradient Boosting) demonstrate superior predictive performance over simpler models. These models consistently achieve higher $R^2$ scores and lower error rates (RMSE/MAE) by effectively modeling the market's complexity (El Mouna et al., 2023 ; Sharma, M. et al., 2024).

Within this class of high-performing models, XGBoost (Extreme Gradient Boosting) was selected as the primary algorithm for this project due to its combination of superior predictive accuracy, efficiency, interpretability, and proven relevance.

In direct, "bake-off" style comparisons, XGBoost consistently emerges as a top performer. A comprehensive study by Zhang (2023), which tested Multiple Linear Regression, K-Nearest Neighbors, Random Forest and XGBoost, found that XGBoost achieved the lowest RMSE (0.127) and significantly outperformed all other models. Similarly, Sharma, Harsora, & Ogunleye (2024) compared XGBoost to linear regression, multi-layer perceptron, support vector regressor, random forest, concluding that it consistently performed the best and was an "optimal" choice for this type of prediction.

## **2.6 Interpretability and Feature Importance in XGBoost**

Furthermore, XGBoost helps overcome the usual “black box” concern in machine learning by providing high interpretability.[1] Zhang (2023) shows that using TreeSHAP values allows us to clearly see how each factor, such as crime rates or school quality, affects the predicted housing price.

The sequential learning method of XGBoost is the key reason for its superior performance, as studies like Mora-Garcia et al. (2022) confirm that this approach often results in the highest predictive accuracy for complex real estate data. This power has been specifically validated in complex, volatile markets; Yu (2024), for example, found that ensemble methods like XGBoost are more adept at predicting prices in a varied market such as New York City. This combination of industry-leading accuracy, computational efficiency and high interpretability makes XGBoost a reliable and powerful choice for this project's scenario-based forecasting goals.

## **2.7 Summary and Methodological Justification**

Finally, this project's focus on a comprehensive set of predictors aligns with findings from Sharma et al. (2024), who demonstrated that a model's accuracy is highly dependent on using specific, relevant housing attributes such as location, square footage and other structural details. The integration of feature importance has also further advanced the interpretability of these models, allowing researchers and policymakers to trace the influence of individual predictors such as neighborhood quality, crime or school performance.

Taken together, these studies provide strong justification for this capstone’s methodological framework, which is centered on comparing three predictive modeling

[1] The “black box” concern refers to the fact that some machine learning models make accurate predictions but do not clearly show how or why they reached those predictions, making it hard to understand the contribution of individual factors.

approaches. Within this framework, XGBoost is ultimately emphasized due to its strong predictive performance, robustness to multicollinearity and enhanced transparency through feature attribution. Its ensemble structure allows the model to capture both macroeconomic drivers and localized neighborhood effects, making it particularly well suited to Staten Island’s diverse and evolving housing market.

# CHAPTER 3- METHODOLOGY

## 3.1 Forecasting Framework and Research Design

This project uses a scenario-based forecasting approach to predict housing prices in Staten Island for the period 2026-2030. Three models, Linear Regression, k-Nearest Neighbors (kNN), and eXtreme Gradient Boosting (XGBoost), are developed and compared to determine which provides the strongest predictive performance. The forecasts are designed to reflect how housing prices may evolve under a Republican-led national administration, using the 2017-2021 presidential term as the historical benchmark. The methodological framework combines machine learning with macroeconomic and neighborhood-level data to ensure both accuracy and interpretability. The analysis follows four main stages: data collection, data preprocessing, model training and validation and forecast generation.

## 3.2 Data Sources

The analysis is built upon multiple public datasets representing the local, neighborhood and national dimensions of the housing market. The primary dependent variable for the model, residential sale price, was sourced from the NYC Department of Finance's Annualized Sales Updates, which provides a rich collection of both neighborhood and citywide data starting from 2003. These records provide raw data on individual property sales, including sale price, property type and the sale date. To model neighborhood-level attributes, several key datasets were used. First, crime rates were calculated using the NYPD Complaint Data Historic Portal. Only incidents occurred in Staten Island were retained and mapped to their respective Neighborhood Tabulation Areas (NTAs)[2]. Second, a school quality index was constructed using performance metrics from GreatSchools.org. Scores for all public schools in Staten Island's District 31 were collected and averaged at the NTA level. Third, to account for population density and calculate per-capita rates, raw population data for each NTA was acquired from the NYC Department of City Planning's Population FactFinder portal. To account for new housing supply, data from the NYC OpenData *"DOB Permits – Staten Island"* dataset was used to track permits filed with the Department of Buildings. Only residential permits were included and the data was aggregated annually to approximate neighborhood-level development activity. Finally, to control broader economic and demographic shifts, several borough and national-level indicators were included. Historical, borough-wide population estimates for Staten Island from 2017 to 2025 were sourced from World Population Review, to identify the possible population growth or decay. All

[2] **Neighborhood Tabulation Areas (NTAs)** are official geographic units created by New York City to group nearby neighborhoods with similar population and housing characteristics. They are commonly used by the City for reporting statistics like housing prices, crime and demographics and they provide a consistent way to compare neighborhoods across the city.

macroeconomic indicators were sourced exclusively from Federal Reserve Economic Data (FRED). These included the Consumer Price Index for All Items to the New York metropolitan area and the 30-Year Fixed Rate Mortgage Average in the United States (MORTGAGE30US) was included to capture national borrowing costs and housing affordability. The scenario blueprint for the forecast was derived from the political and economic context of 2017-2021, which was used to model future trends in inflation and mortgage rates.

### 3.3 Data Preprocessing

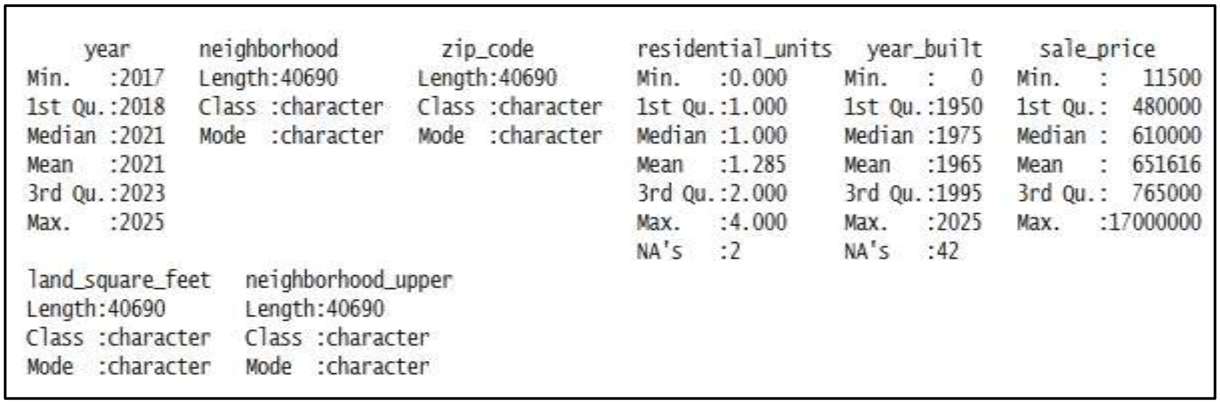
Individual annual property sales files for Staten Island (2017-2025) were imported from Excel format and standardized. Column names were cleaned for consistency using clean_names and a new variable year was appended to each dataset. All datasets were then merged using bind_rows() to form a single, continuous time series of residential property transactions across nine years.

To ensure data quality and consistency, several preprocessing steps were applied:

- **Conversion of data types:** Columns such as sale_price, residential_units and year_built were converted to numeric values.
- **Filtering unrealistic values:** Sales with a recorded price below $10,000 were removed.
- **Focusing on residential properties:** Only records classified as "Family Dwellings" were retained to focus on Staten Island's owner-occupied and small residential properties.
- **Variable selection:** The following columns were kept for modeling:
    - year : year of sale
    - neighborhood : NYC-defined neighborhood name
    - zip_code : postal code of the property
    - residential_units : number of dwelling units per property
    - year_built : year the structure was built
    - sale_price : recorded sale price (in USD)
    - land_square_feet : size of the property lot

### 3.4 Exploratory Data Analysis (EDA)

A summary dataset was then created to calculate average, minimum and maximum sale prices per neighborhood and year, along with the number of sales. The analysis begins with the dependent variable, sale price. A summary of the variable's statistical properties reveals key characteristics of the housing market data. As seen on Table 1 below the mean of the dependent variable is higher than the median indicating possible right skewness. That indicates that while most properties cluster around a central value, a long tail of high-value properties exists, which would distort a standard-scale visualization.



```
     year       neighborhood         zip_code          residential_units   year_built      sale_price     
 Min.   :2017   Length:40690       Length:40690       Min.   :0.000      Min.   :   0   Min.   :   11500  
 1st Qu.:2018   Class :character   Class :character   1st Qu.:1.000      1st Qu.:1950   1st Qu.:  480000  
 Median :2021   Mode  :character   Mode  :character   Median :1.000      Median :1975   Median :  610000  
 Mean   :2021                                         Mean   :1.285      Mean   :1965   Mean   :  651616  
 3rd Qu.:2023                                         3rd Qu.:2.000      3rd Qu.:1995   3rd Qu.:  765000  
 Max.   :2025                                         Max.   :4.000      Max.   :2025   Max.   :17000000  
                                                      NA's   :2          NA's   :42                       
 land_square_feet   neighborhood_upper
 Length:40690       Length:40690      
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

Table 1

### 3.4.1 Sale Price Distribution and Trends

To better visualize this log-normal distribution, the following histogram (Figure 1) plots the sale price on a log scale. This transformation clearly shows the central tendency and spread of the data, confirming the concentration of properties in the $600,000 to $900,000 range. The presence of a long right tail indicates a smaller share of significantly higher-value transactions, which would distort a standard linear-scale visualization. In conclusion, we can see that most homes fall within a mid-market price range, a subset of luxury properties exists that could influence average price calculations and model performance if not addressed through appropriate transformations.

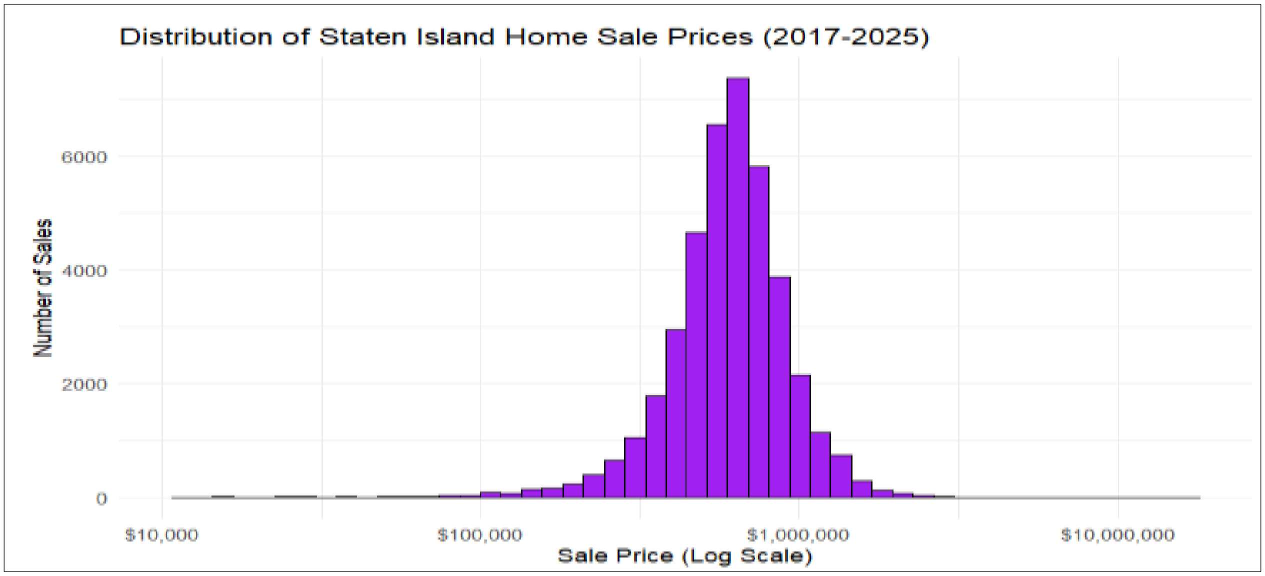


Figure 1. Distribution of Staten Island Home Sale Prices (2017–2025)

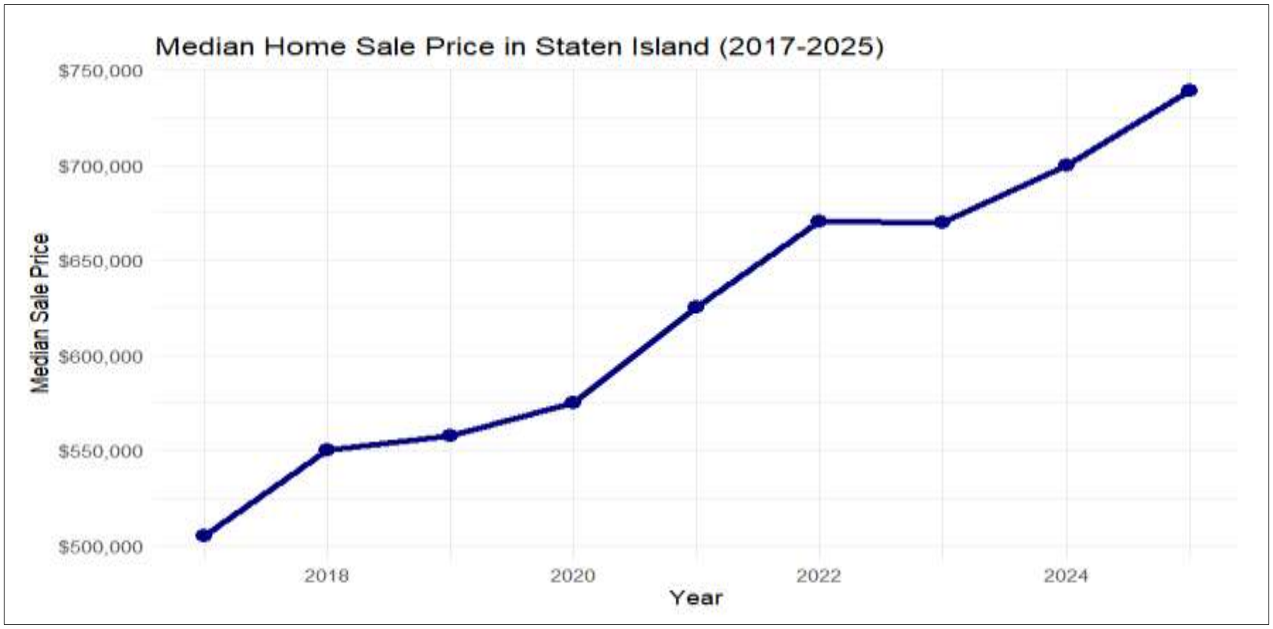


Figure 2. Median Home Sale Price Trends in Staten Island (2017–2025)

Figure 2 highlights the yearly trend in median sale prices across Staten Island from 2017 to 2025. Home prices increased steadily over the observed period, rising by approximately $230,000 in nine years. The sharpest growth occurred between 2020 and 2022, a period that overlaps with the COVID-19 pandemic and the housing demand was triggered by historically low mortgage rates and lifestyle changes favoring suburban style living. The slight stabilization observed between 2022 and 2023 suggests a post-pandemic market correction, followed by a continued upward trend into 2024-2025.

### 3.4.2 Neighborhood-Level Factors Influencing Housing Prices

To begin breaking down the market at a more detailed level, Figure 3 below visualizes the average sale price by both Neighborhood (NTA) and the Decade Built. This heatmap reveals several important patterns. There is a clear price difference between northern neighborhoods like Port Richmond and southern ones like Annadale or Tottenville-Charleston. It also shows that newer homes, particularly those built from 1990-2020 in the southern neighborhoods, obtain some of the highest average prices, as shown by the salmon-pink, red and orange to yellow squares. Interestingly, there are also pockets of high-value, older housing, such as pre-1900 homes in St. George. This visualization confirms that housing prices are not uniform across the borough, but vary significantly based on NTA and Decade Built. The next paragraphs explore *why* these neighborhood-level price differences exist, beginning with school quality.

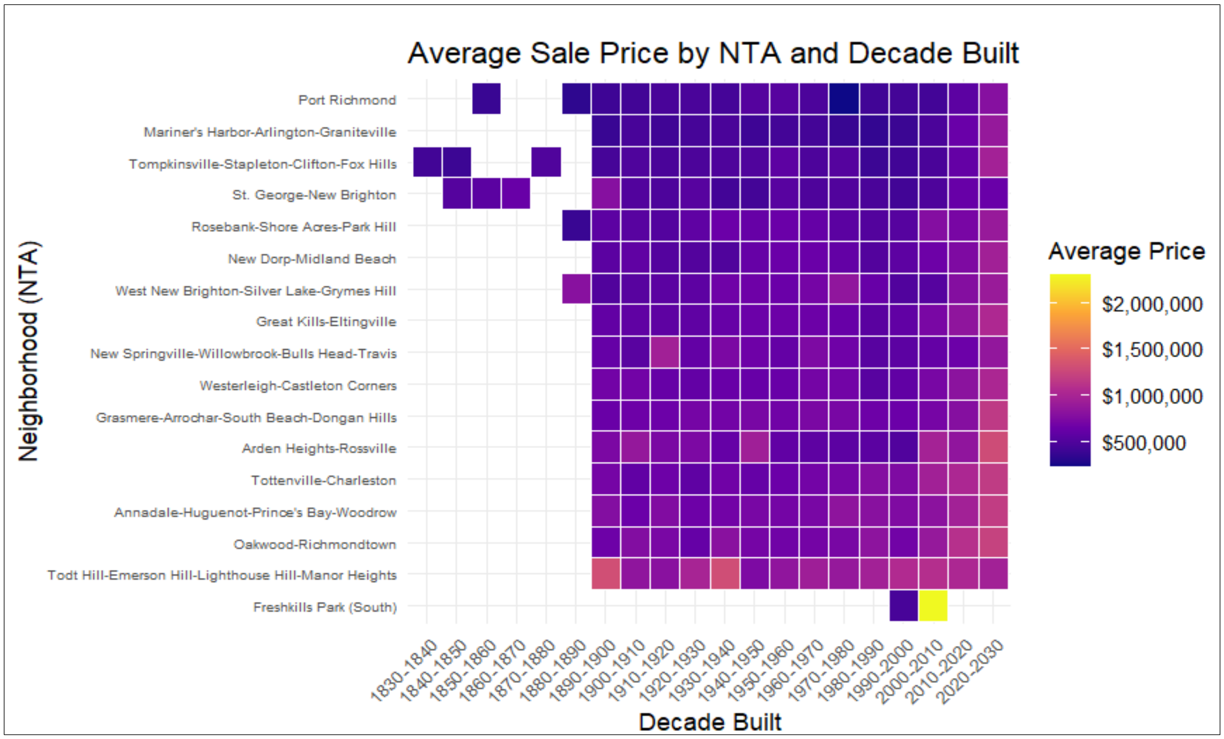


Figure 3. Average Sale Price by NTA and Decade Built

Moving to the core explanatory variables, the analysis addresses the influence of school quality and its impact on property prices. Figure 4 directly answers the question: "Do higher-ranked schools correspond to higher housing prices?". The charts below  provide a clear, definite answer. The bar chart shows the mean housing price in high-rank school areas is $708,715, noticeably higher than the $579,224 mean for low-rank areas. The accompanying box plot reinforces this by showing a higher median (the thick black line) and a much wider price distribution for the high-rank group, confirming a significant price premium. This validates school quality as a critical predictor for the model.

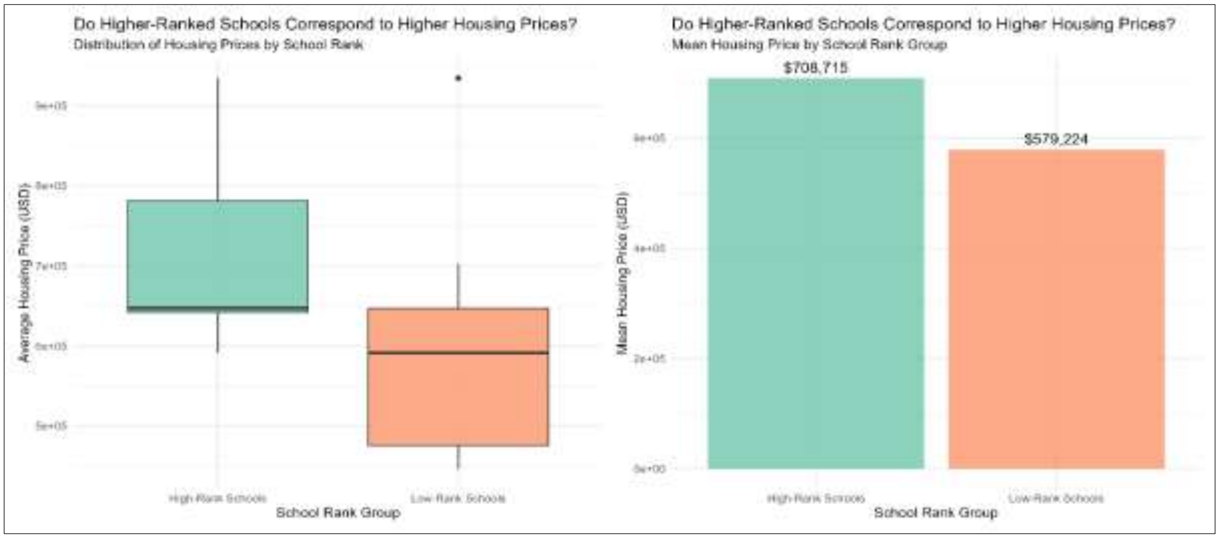


Figure 4. Housing Price Distribution and Mean by School Rank Group

Similarly, the project's second core explanatory variable, crime, was analyzed to understand its distribution. A significant question is whether crime is evenly distributed or concentrated in specific areas. Figure 5 answers this by visualizing crime incidents geographically. It is immediately clear that crime is not random but highly concentrated spatially, with the highest incident levels (light green and yellow) persistently clustered in the northernmost NTAs. This geographic pattern remains remarkably stable across the entire 2017–2025 study period.

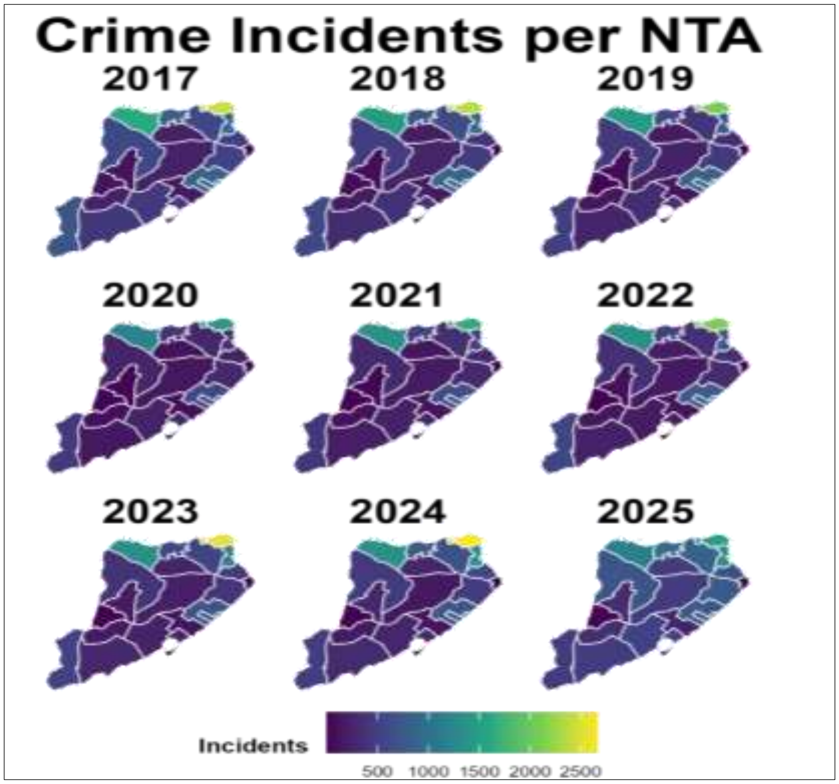


Figure 5. Annual Crime Incidents per Neighborhood Tabulation Area (NTA)

To identify these areas precisely, Figure 6 quantifies the incidents for the top 10 neighborhoods. This chart confirms the above map's finding that the "St. George-New Brighton" NTA consistently ranks #1 for incidents, often by a significant difference. The next highest-crime areas, such as "Mariner's Harbor-Arlington-Graniteville" and "Tompkinsville-Stapleton-Clifton-Fox Hills," are also located in that northern cluster.

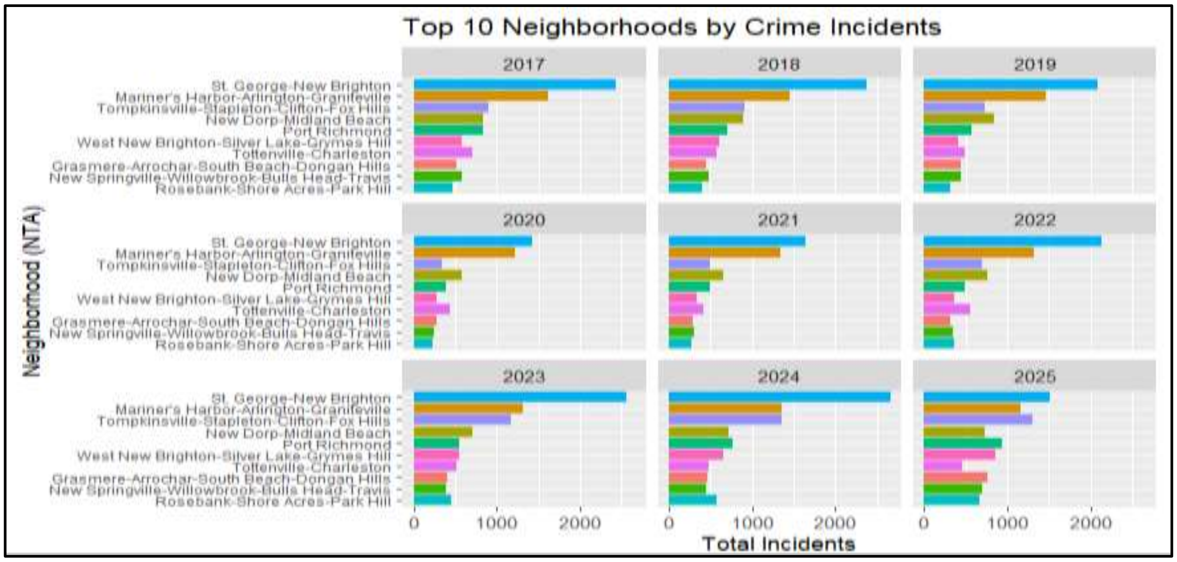


Figure 6. Top 10 Neighborhoods (NTAs) by Total Crime Incidents

Taken together, these visualizations demonstrate that high crime is a stable, non-random and geographically dependent feature of specific neighborhoods, making it a powerful and reliable predictor for a housing price model.

Figure 7 synthesizes the project's core neighborhood-level variables, which are price, school quality and crime, into a single visualization. There is a strong positive correlation between School Ranking (x-axis) and Average Housing Price (y-axis), as rankings improve, average prices tend to rise. At the same time, Crime Incidents (colored dots) act as a powerful suppressor. The neighborhoods with the highest crime rates (the bright yellow and orange dots) are almost all clustered at the low end of both the price and school-ranking spectrums. This graphic confirms that school quality is a key driver of premium pricing, while high crime levels act as a significant cap on property values.

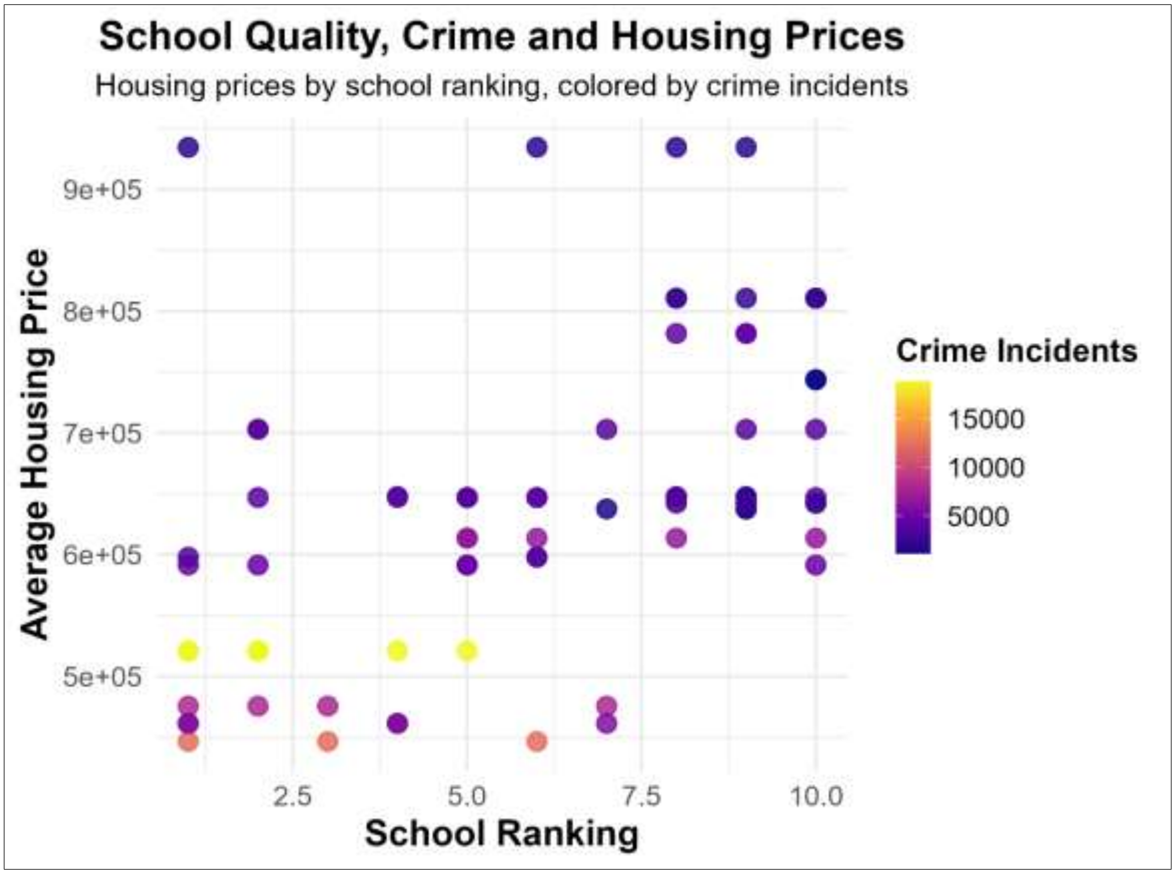


Figure 7. Housing Prices vs. School Ranking (Colored by Crime Incidents)

### 3.4.3 Macroeconomic Indicators and Housing Prices

Figure 8 displays the relationship between the CPI index and median home sale prices in Staten Island over the 2017-2025 period. Both measures trend upward, reflecting broader national increases in the cost of living. However, the rise in home prices is noticeably steeper than the increase in the CPI index, particularly during the post-2020 period. This pattern suggests that general inflation alone does not account for the increase in housing prices. Instead, local forces, such as pandemic-related migration, limited housing supply and increased demand for larger homes, appear to have amplified price growth beyond what would be expected from consumer price inflation. The figure shows that Staten Island’s home prices increased more quickly than general consumer prices, indicating that the market responds strongly to both broader economic pressures and local housing conditions.

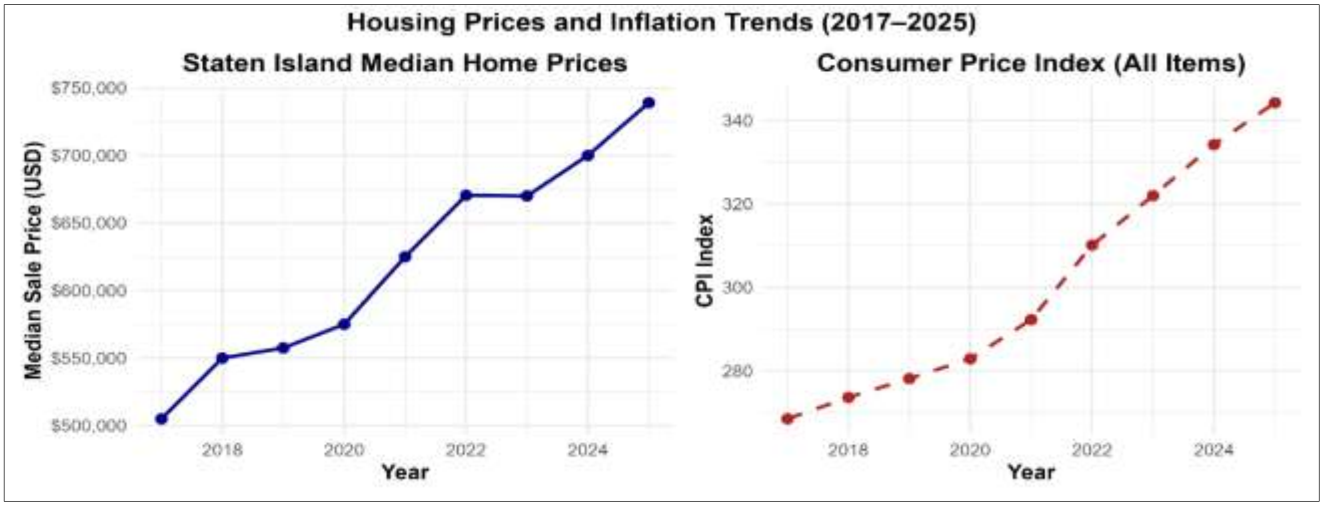


Figure 8. CPI Index and Median Home Prices in Staten Island, 2017–2025

Figure 9 presents Staten Island's annual median sale prices as labeled reference bars alongside trends in two key national indicators, overall CPI inflation and 30-year mortgage rates, shown as line series. From 2017 to 2020, home prices increased steadily while mortgage rates gradually declined, a combination consistent with improving borrowing conditions. A sharp acceleration in prices occurred in 2021-2022, coinciding with a surge in overall inflation and a reversal in mortgage rate trends. In the subsequent period (2023 to 2025), price growth moderated as mortgage rates remained elevated and inflation began to cool, reflecting tighter affordability conditions. Overall, the figure highlights how Staten Island's housing market moves in relation to broader national economic conditions, supporting the inclusion of inflation and mortgage rates as contextual indicators in the forecasting model.

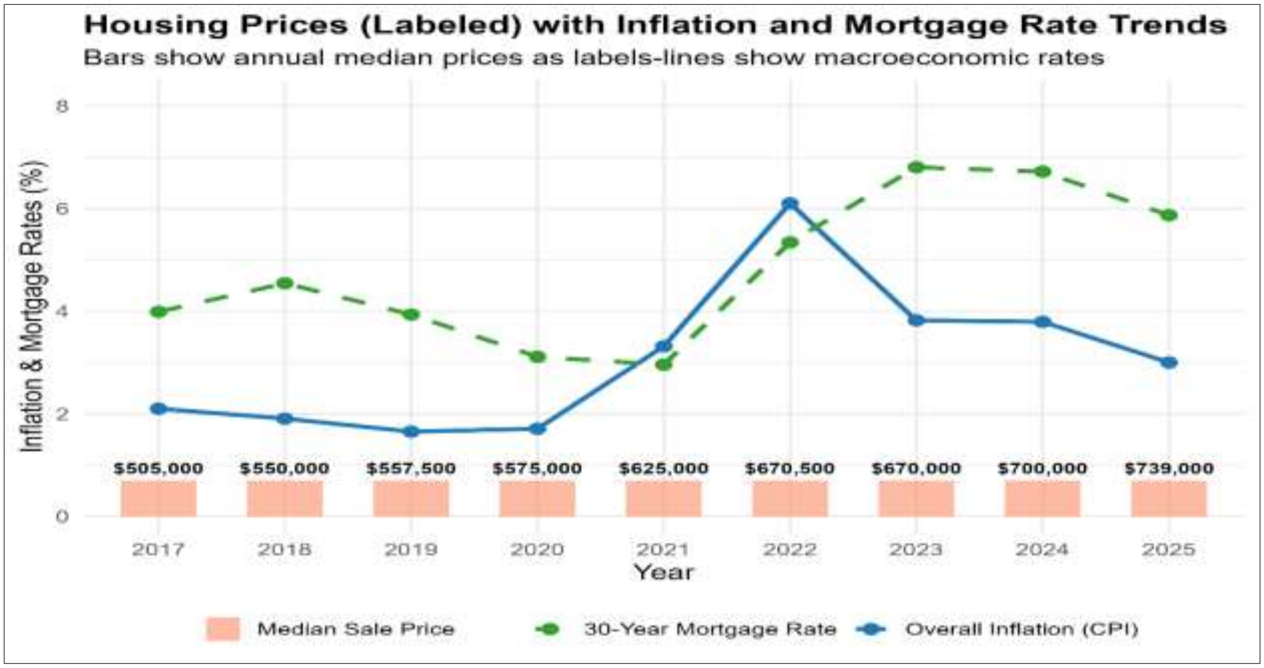


Figure 9. Staten Island Housing Prices vs. Inflation & Mortgage Rates (2017-2025)

### 3.5 Model Development and Validation

Following the exploratory analysis, the next stage involved building predictive models capable of estimating residential property sale prices in Staten Island. Because housing markets are shaped by multiple interacting factors, structural characteristics, neighborhood context and broader economic conditions, the modelling process evaluated three different approaches: Linear Regression, k-Nearest Neighbors (kNN), and eXtreme Gradient Boosting (XGBoost). Comparing these models allowed for a more comprehensive assessment of how different algorithms handle the non-linear and interdependent relationships present in real estate data.

In reality, housing prices are influenced by a mixture of factors that are both non-linear and deeply interdependent. A property's value is influenced simultaneously by its structural characteristics (for example age, size), its neighborhood features (crime, school quality) and macroeconomic conditions (e.g. mortgage rates, inflation). To test a more flexible, distance-based approach, a kNN regression model was also implemented. This method predicts prices by averaging the values of similar properties in the feature space, offering an intuitive baseline that reflects localized patterns but without learning complex interactions.

The primary model in the comparison was XGBoost, a gradient-boosted decision tree method well-suited for capturing non-linear effects and variable interactions. XGBoost builds trees sequentially, with each tree correcting the errors of the previous one through gradient-based optimization. This structure enables the model to represent subtle differences between neighborhoods, housing stock, and economic conditions more effectively than traditional linear approaches.

To enhance predictive power and capture neighborhood context, several new features were engineered.**A crime rate per 1,000 residents** was calculated at the NTA level, which standardized the crime data, allowing for fair comparison between neighborhoods of different population sizes. An **average school ranking** was also aggregated by NTA to represent an area's educational quality. Furthermore, new development **permits** were included as a proxy for local housing supply and a **building age** feature was derived by subtracting the build year from the sale year. Finally, a log-transformation was applied to the land area to normalize its skewed distribution.

Several data cleaning steps were also taken. Outliers were managed by capping the top and bottom 1% of sale prices, which prevents extreme values from disproportionately influencing the model. Categorical variables were properly formatted for model encoding and only complete cases, rows with no missing data, were retained.

Additionally, year_built was removed. It was highly correlated with the year and retaining both would create redundancy. This age-related information was better captured by the new building_age feature. The zip_code variable was also removed because it introduced noise. ZIP codes often span multiple neighborhoods and do not accurately reflect localized housing dynamics, making the Neighborhood Tabulation Area (NTA) a more precise and effective geographic unit for this analysis.

The modelling dataset was randomly partitioned into a training set (80%) and a testing set (20%). All categorical variables were converted using one-hot encoding after the split to avoid information leakage. Each algorithm was then fitted on the training data and evaluated on the unseen test set using RMSE, MAE and $R^2$.

For XGBoost, hyperparameters such as learning rate, tree depth, subsampling ratios, and regularization terms (L1 and L2) were tuned to control model complexity and reduce overfitting. Early stopping was applied to terminate training once performance on the test set no longer improved. The kNN model was run using standardized feature matrices, and Linear Regression was evaluated in its full multivariate form.

Model performance was assessed using three main evaluation metrics. First, the Root Mean Squared Error (RMSE) summarizes the typical size of prediction errors, giving more weight to larger mistakes. Second, the Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual sale prices. Third, the coefficient of determination ($R^2$) indicates how much of the variation in sale prices is explained by the model.

To make these metrics easier to interpret in the context of the housing market, RMSE and MAE were also expressed as a percentage of the average sale price and the Mean Absolute Percentage Error (MAPE) was computed. These normalized measures show the typical prediction error as a share of the transaction price, providing a clearer sense of how large the model's mistakes are relative to the value of the homes being forecast.

Residual diagnostics were performed as part of model validation to check for systematic patterns or bias in prediction errors and the feature importance scores were analyzed to interpret model behaviors.

### 3.6 Scenario-Based Forecast Generation (2026-2030)

The final stage of the analysis involved developing a forecasting framework to project Staten Island housing prices for the period 2026-2030. Rather than producing a single deterministic estimate, the study adopts a scenario-based forecasting approach, ensuring that projections reflect a coherent and economically grounded narrative. The chosen scenario uses macroeconomic trends observed during 2017-2021 as the blueprint for future conditions. This period was selected because it reflects an economic environment shaped by deregulation, pro-market policy direction and comparatively low interest-rate pressures in the early years, conditions that align with expectations for a similar policy climate in the future. Accordingly, the forecast models how the housing market may evolve if the broader macroeconomic landscape follows patterns consistent with that reference period.

To generate the forecast, historical relationships between Staten Island housing prices and key economic predictors, specifically mortgage rates and All items CPI, were first estimated using the best-performing model identified in the model comparison stage. Future values of these macroeconomic indicators were then constructed by calculating their average year-over-year percentage changes during the 2017-2021 reference period and applying those growth rates to the latest observed values from 2025. These projected economic conditions were merged with the full property-level feature dataset to create a synthetic panel of observations for the years 2026-2030, preserving each neighborhood's structural and spatial attributes. This synthetic dataset was then passed through the selected forecasting model to generate projected sale prices at both the neighborhood and NTA levels.

To evaluate how housing prices are expected to evolve over time, forecasts are analyzed relative to the observed 2025 baseline across the full forecast horizon through 2030. Rather than emphasizing short-term fluctuations, the analysis focuses on relative price changes to highlight longer-run structural differences across geographic areas. At the NTA level, forecasted prices are expressed as percentage changes relative to 2025 values, allowing for consistent comparison across neighborhoods with different baseline price levels. At the neighborhood level, these patterns are further contextualized by tracing the price trajectories of 1-unit and 2-unit residential properties across the forecast horizon for a set of randomly selected Staten Island neighborhoods, providing insight into how local housing markets are expected to evolve over time.

To enhance the interpretability and practical usefulness of the forecasting results, an interactive dashboard was developed using an R Shiny application centered around a personalized NTA Preference Quiz. Instead of a static map-based interface, the dashboard allows users to input housing preferences including budget category, interest in top-ranked schools, housing age preferences, proximity to New Jersey or Brooklyn and access to train service. Based on these inputs, the application dynamically returns a tailored list of NTAs whose forecasted characteristics best match the user's criteria. The second component of the app is a neighborhood-specific trend panel, which uses interactive Plotly visualisations and a supporting data table to present yearly price trends for any selected neighborhood.

This design transforms the forecasting results from a purely analytical output into a practical decision-support tool for potential homebuyers, investors and planners. By integrating machine-learning forecasts with an intuitive preference-based exploration system, the application demonstrates how predictive analytics can be directly translated into applied real-estate decision-making.

# CHAPTER 4-RESULTS

## 4.1 Linear Regression Results and Diagnostic Evaluation

The Linear Regression model provides a useful baseline but demonstrates clear limitations in capturing the complexity of Staten Island's housing market. Its RMSE (~$161,079) and MAE (~ $112,535) as seen on Table 2 below indicate sizeable prediction errors and its $R^2$ value of 0.540 suggests that it explains just over half of the variation in sale prices. When normalized (Table 3), the model's RMSE corresponds to 25.1% of the average sale price, and its MAE to 17.6%, meaning that, even after adjusting for price scale, the model typically deviates by more than one-sixth of a property's value. The relatively high MAPE (21.2%) further confirms instability across different price ranges. Overall, the linear model struggles with the non-linear and interaction-heavy nature of real estate pricing.

These limitations are also clearly visible in the diagnostic plots. The residuals (see Appendix Figures A7-A9) exhibit a widening spread at higher predicted prices, indicating heteroskedasticity and reduced predictive reliability for high-value properties. The error distribution shows heavier tails than a standard normal shape, while the Q–Q plot displays systematic deviations from the reference line at the extremes, suggesting that large prediction errors occur more frequently than would be expected under homoscedastic normal errors. The Actual vs. Predicted plot in Figure 10 further reveals persistent under- and over-prediction across the price range, particularly for higher-priced homes. Together, these diagnostics indicate that the constant-variance assumption of the linear model is violated and that error variance increases with price level, motivating the consideration of alternative specifications such as variance-stabilizing transformations or weighted regression approaches.

## 4.2 k-Nearest Neighbors (kNN) Results and Diagnostic Evaluation

On the other hand, the kNN regression model performs slightly better than Linear Regression on some metrics but still falls short of strong predictive accuracy. Its RMSE (~ $156,145) and MAE (~$109,437) remain high, and the $R^2$ of 0.568 indicates only modest improvement in explanatory power. Normalised metrics show that RMSE equals 24.4% of the average price and MAE equals 17.1%, suggesting only marginal gains over Linear Regression. The MAPE of 20.5% shows that percentage errors remain relatively large and volatile across the price distribution. Because kNN relies on similarity between nearby observations, its performance weakens when price variation across neighborhoods is large, as is the case in Staten Island.

From a diagnostic point of view (in Appendix Figures A7–A9), kNN residuals remain widely dispersed across predicted values, with limited improvement in variance stabilization. The residual distribution remains heavy-tailed and the Q–Q plot continues to show deviations from normality. Although the Actual vs Predicted relationship in Figure 10 improves slightly relative to the linear model, dispersion remains substantial for higher-priced homes. Because kNN relies on similarity among nearby observations, its performance weakens when sharp price heterogeneity exists across neighborhoods, as is the case in Staten Island.

### 4.3 XGBoost Results and Diagnostic Evaluation

Last but not least, XGBoost delivers the strongest performance among the three models. With the lowest RMSE (~ $141,652) and MAE (~$97,630), along with the highest $R^2$ (0.645), it captures substantially more of the underlying structure in the data. Importantly, normalized metrics reinforce this advantage: the model's RMSE equals 22.1% of the average sale price, and its MAE equals 15.2%, both the lowest among the three approaches. Its MAPE of 18.8% also indicates more stable prediction errors across low-, mid- and high-priced homes. These results show that XGBoost is better able to model the non-linear relationships, neighborhood effects and interactions that shape housing prices.

This superiority is clearly visible among the diagnostic .In Appendix Figure A7 XGBoost shows the tightest clustering of residuals around zero across the full prediction range, indicating reduced bias and improved variance stability. The residual distribution (Appendix Figure A8) is more symmetric and concentrated than for the other models, while the Q–Q plot (Appendix Figure A9) shows closer alignment with the theoretical normal line, particularly in the center of the distribution. Most importantly, the Actual vs Predicted relationship in Figure 10 demonstrates the strongest alignment along the 45-degree reference line, confirming superior tracking of true sale prices.

Given its superior performance across both absolute and normalized metrics, XGBoost was selected as the final predictive model. Its ability to capture complex patterns, such as heterogeneous neighborhood effects, interactions among structural features and macroeconomic influences, makes it better suited for modelling Staten Island's diverse housing stock. This model was therefore used in the forecasting stage to project housing prices across NTAs and neighborhoods from 2026 to 2030.

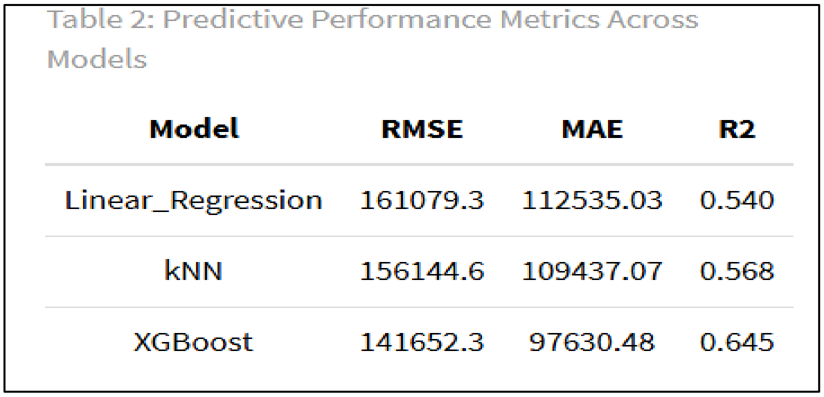


Table 2: Predictive Performance Metrics Across Models

| Model | RMSE | MAE | R2 |
|---|---|---|---|
| Linear_Regression | 161079.3 | 112535.03 | 0.540 |
| kNN | 156144.6 | 109437.07 | 0.568 |
| XGBoost | 141652.3 | 97630.48 | 0.645 |

Table 2. Comparison of Linear Regression, kNN and XGBoost Using RMSE, MAE and $R^2$

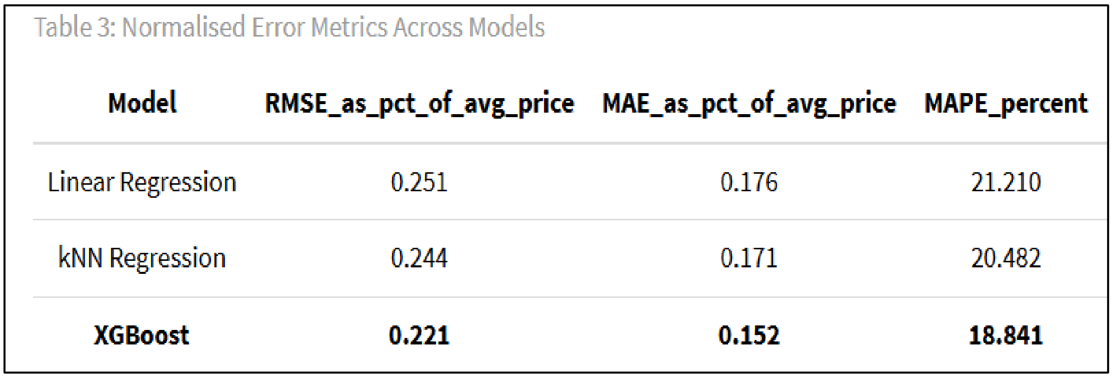


Table 3: Normalised Error Metrics Across Models

| Model | RMSE_as_pct_of_avg_price | MAE_as_pct_of_avg_price | MAPE_percent |
|---|---|---|---|
| Linear Regression | 0.251 | 0.176 | 21.210 |
| kNN Regression | 0.244 | 0.171 | 20.482 |
| **XGBoost** | **0.221** | **0.152** | **18.841** |

Table 3. Normalized Error Metrics for All Models Relative to Average Sale Price

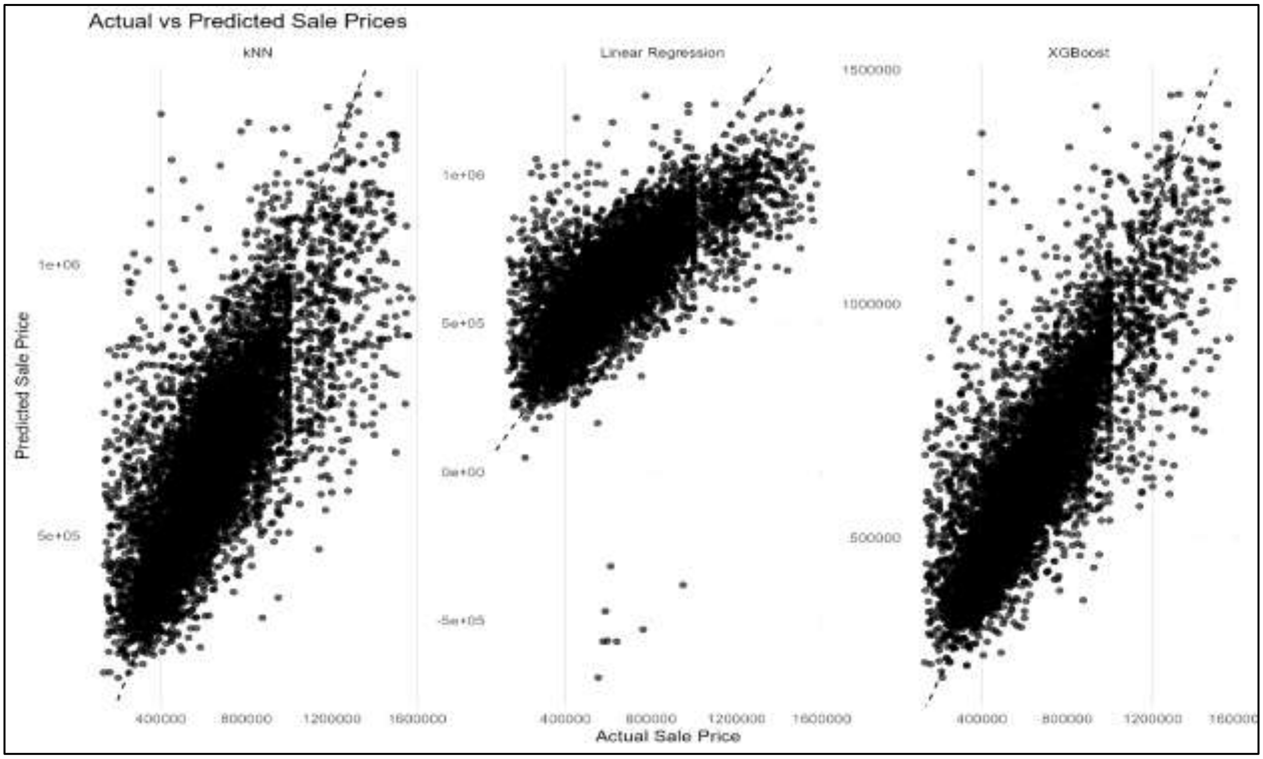
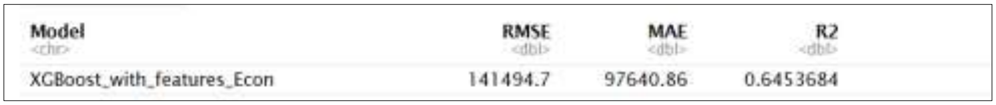


Figure 10. Actual vs Predicted Sale Prices

After incorporating the full set of engineered economic and neighborhood-level features, including crime rates, school rankings, building permits, CPI and mortgage rates, the model was retrained and re-evaluated. The enhanced specification achieved an RMSE of $141,495, an MAE of $97,641 and an R² of 0.645. These results confirm that the integration of economic and spatial features further strengthened the model's explanatory and predictive power. This finalized XGBoost model was subsequently used for all housing price forecasts through 2030.



| Model | RMSE | MAE | R2 |
|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> |
| XGBoost_with_features_Econ | 141494.7 | 97640.86 | 0.6453684 |

Table 4. Econ Model Metrics

## 4.4 Feature Importance and Key Drivers of Housing Prices

To address the second research objective, feature importance was extracted from the final XGBoost model using gain-based importance, which shows the relative importance of each variable in the model's predictions. The importance rankings are presented in Figure 11 below, which displays the top 15 most influential predictors in the model.

As shown in Figure 11, property-level structural characteristics dominate the model's predictive structure. The log-transformed land square footage emerges as the most influential predictor among all variables, followed by building age and year (of house sale), indicating that property size, building age, and broader market timing are the main factors driving housing prices. Residential unit count also ranks among the top predictors, highlighting the role of building density and structure in Staten Island's housing market

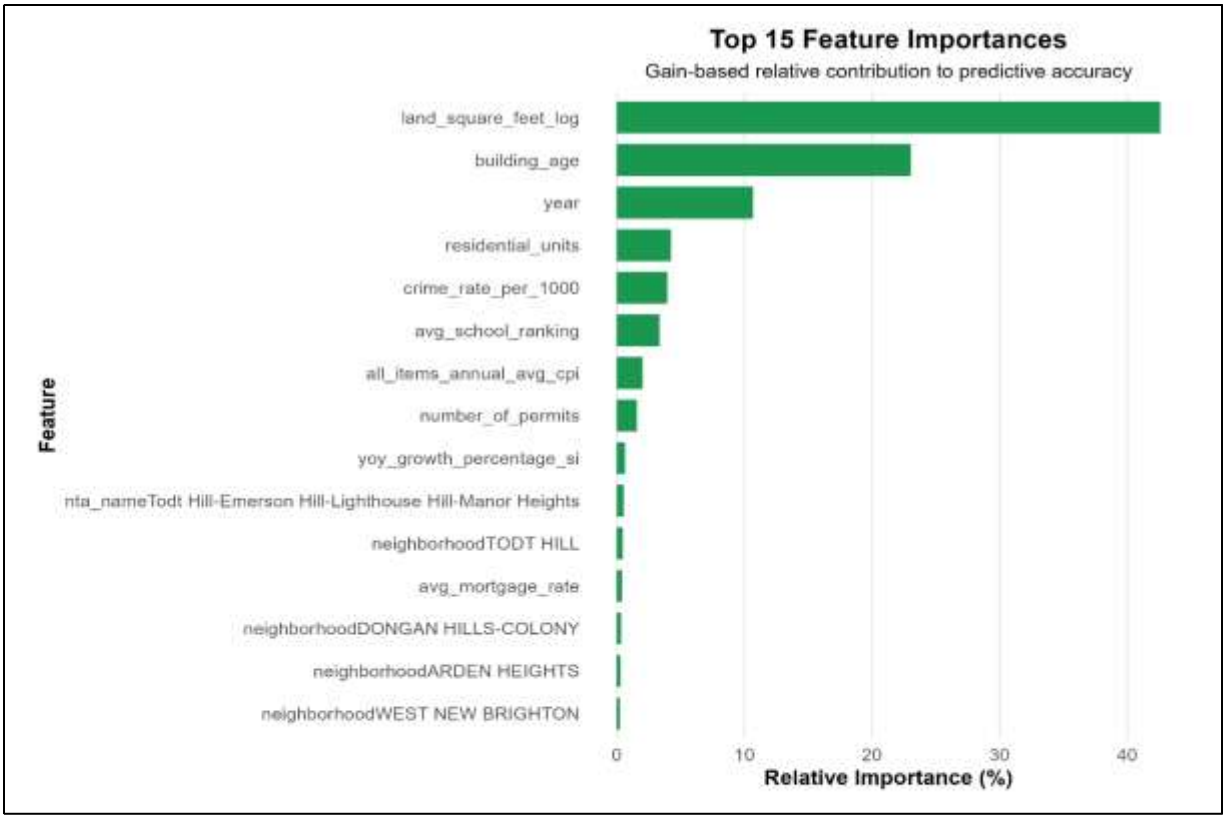
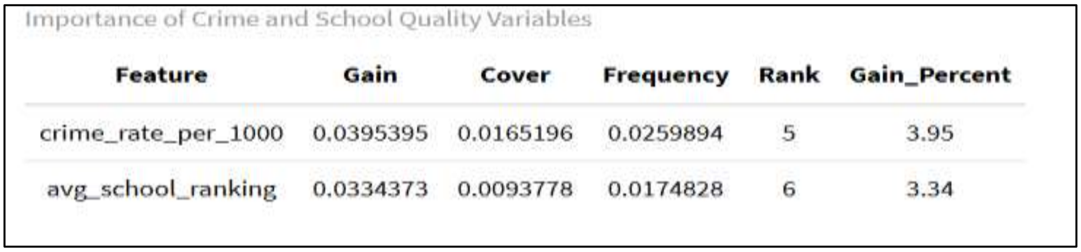


Figure 11. Top 15 feature importances

## 4.5 Neighborhood Quality-of-Life and Macroeconomic Effects

Importantly, neighborhood quality-of-life variables also emerge as substantively meaningful contributors. Crime rate per 1,000 residents ranks fifth overall, accounting for approximately 3.95% of total model gain, while average school ranking ranks sixth, contributing approximately 3.34% of total gain, as reported in Table 5. These rankings place both variables ahead of several macroeconomic indicators and neighborhood identifiers, confirming that local safety and educational quality apply a direct and measurable influence on housing prices.



Importance of Crime and School Quality Variables

| Feature | Gain | Cover | Frequency | Rank | Gain_Percent |
|---|---|---|---|---|---|
| crime_rate_per_1000 | 0.0395395 | 0.0165196 | 0.0259894 | 5 | 3.95 |
| avg_school_ranking | 0.0334373 | 0.0093778 | 0.0174828 | 6 | 3.34 |

Table 5. Normalized Error Metrics for All Models Relative to Average Sale Price

Macroeconomic conditions further contribute to predictive accuracy, with inflation (CPI), mortgage rates and construction activity (number of permits) appearing among the top 15 features. However, their influence remains secondary to that of property structure and neighborhood-level quality indicators.

Together, these results demonstrate that while housing prices in Staten Island are primarily driven by physical property characteristics and market timing, crime and school quality play statistically important secondary roles in shaping price variation across neighborhoods. These findings directly support the second research objective by quantifying the role of neighborhood quality-of-life conditions within the final predictive framework.

# CHAPTER 5- FORECASTING RESULTS AND DISCUSSION

## 5.1 Forecasting Framework and Scenario Design

This chapter applies the final XGBoost model identified in the Results chapter to generate forward-looking forecasts of Staten Island housing prices for the 2026-2030 period. Forecasts are produced under a scenario-based design grounded in historical macroeconomic behavior observed during the 2017-2021 period. Specifically, average annual changes in mortgage rates and inflation (CPI) from this historical window are projected forward from the most recently observed values to construct a realistic post-2025 economic trajectory.

This approach avoids extreme growth assumptions and instead reflects a moderate market environment based on historical trends, allowing future price changes to be driven by the neighborhood and economic relationships learned by the model. Forecasts are generated at the property level and then aggregated to both the Neighborhood Tabulation Area (NTA) and neighborhood levels.

## 5.2 Forecasted Relative Housing Price Changes by NTA

Figure 12  shows the Neighborhood Tabulation Areas (NTAs) that are forecasted to experience positive relative housing price growth between 2025 and 2030, with changes measured as percentage increases relative to observed 2025 prices. Forecasts are produced using a machine-learning model that holds neighborhood characteristics fixed at their 2025 values while allowing macroeconomic factors to evolve over time. Results are aggregated at the NTA level and limited to areas with positive relative growth to highlight variation in upward price dynamics.

The results indicate that forecasted price growth is not uniform across Staten Island. While all NTAs shown are projected to experience price increases, the magnitude of growth varies considerably. Several southern and interior NTAs display stronger relative growth, whereas others exhibit more modest increases. These differences reflect variation in the pace of forecasted price change across neighborhoods rather than uniform outcomes within each NTA.

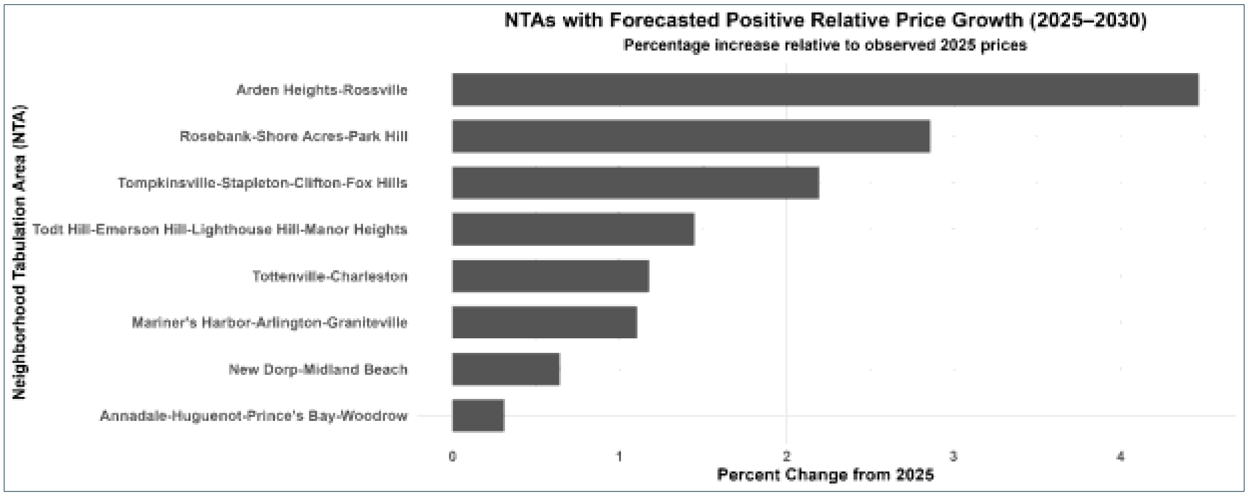


Figure 12. Forecasted Relative Housing Price Changes by NTA

## 5.3 Price Evolution of 1-Unit and 2-Unit Homes Across Selected Neighborhoods

Figure 13 illustrates the evolution of prices for 1-unit and 2-unit homes across six randomly selected Staten Island neighborhoods from 2017 through the forecast horizon of 2030. Several clear patterns emerge when the plotted trends are considered alongside the underlying forecast estimates reported in Appendix Table A9.

Across most neighborhoods, 2-unit homes maintain a consistent pricing premium over 1-unit homes throughout both the observed and forecast periods. For example, in Woodrow, the model projects that by 2030, 2-unit homes will average approximately $943,786, compared to about $757,654 for 1-unit properties. A similar divergence is observed in Rossville, where 2-unit homes remain above $1.15 million by 2030, while 1-unit homes stabilize near $713,000. These persistent gaps reflect the added functionality and income-generating potential associated with two-family housing.

Arrochar represents a notable exception to this general pattern. Beginning in the observed period and continuing into the forecast window, 1-unit (single-family) homes consistently exceed the prices of 2-unit homes. By 2025, average prices are estimated at roughly $901,500 for 1-unit homes compared to $940,000 for 2-unit homes, with the relative positioning remaining stable through 2030. This inversion suggests neighborhood-specific dynamics, such as housing stock composition, location amenities, or school quality, that elevate demand for single-family residences over multi-unit alternatives in this submarket.

In contrast, neighborhoods such as West New Brighton exhibit a narrower differential between home types. By 2030, projected prices converge to approximately $658,989 for 2-unit homes and $631,894 for 1-unit homes, indicating a more balanced demand structure and less pronounced segmentation by housing type.

Across the forecast period from 2026 to 2030, the model indicates a transition away from the sharp post-pandemic price acceleration toward a phase of relative stabilization. While modest year-to-year increases and slight declines persist across neighborhoods and home types, overall price trajectories flatten, suggesting that long-run structural factors, such as crime exposure, school quality, housing age, and demographic composition, play a more dominant role than short-term macroeconomic shocks. These dynamics are captured within the XGBoost model's feature structure and are reflected consistently across neighborhoods.

For completeness and precise reference, detailed annual forecast values for each neighborhood and housing type are provided in Appendix Table A9 and correspond directly to the series displayed in Figure 13.

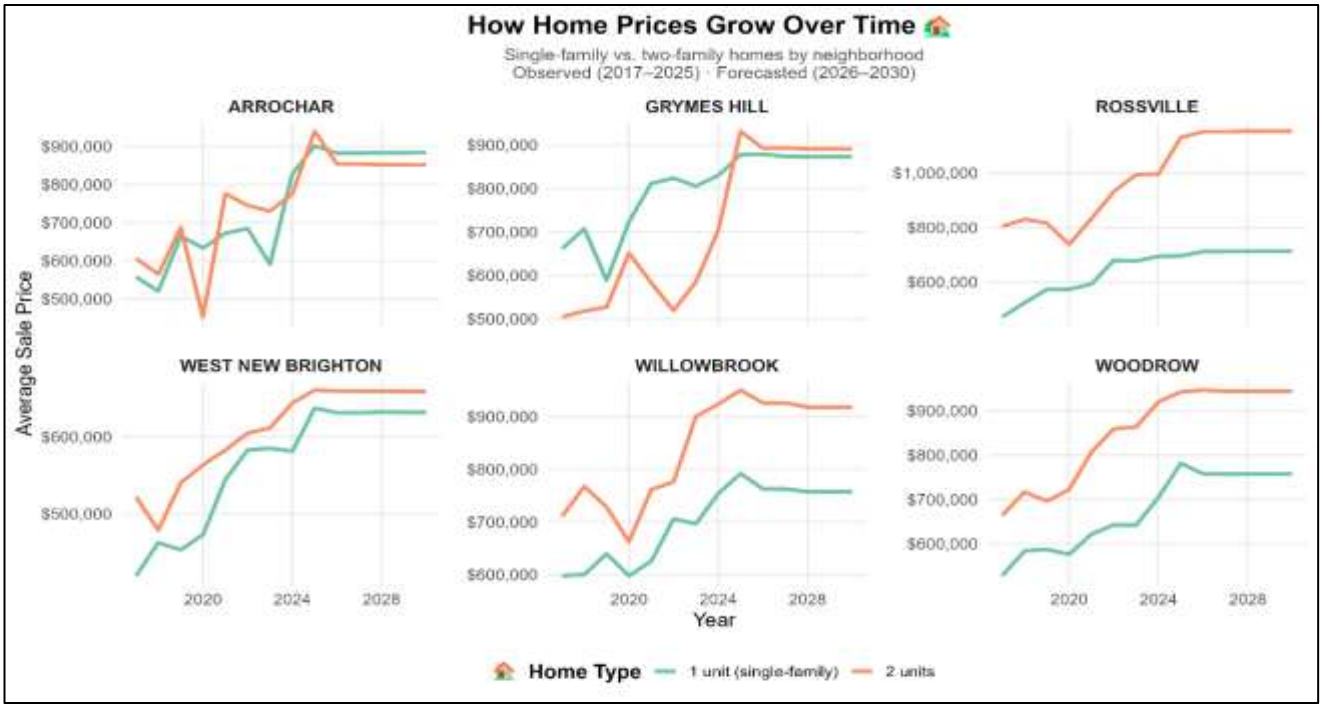


Figure 13. Observed vs Forecasted Price Trends For Residential Units

## 5.4 Interactive Forecasting Dashboard (Shiny Application)

While the preceding sections present the core forecasting results through static figures and summary tables, an interactive Shiny application was also developed to enable deeper, user-driven exploration of Staten Island's projected housing dynamics. The application consists of two complementary components designed for both guided decision support and open-ended neighborhood exploration.

The first tab, titled *"NTA Finder"* features a preference-based NTA quiz that allows users to filter neighborhoods based on budget range, school quality, housing age, train access and geographic location preferences. Based on the user's selections, the tool returns a customized list of NTAs that best match those preferences along with their forecasted average home prices. Users can then select any recommended NTA to view its historical and projected price trajectory through interactive Plotly visualizations and supporting data tables.

The second tab provides a neighborhood-specific exploration tool, where users can select any Staten Island neighborhood from a dropdown menu and directly examine its historical price trend and future projected values. This allows for unrestricted comparison and detailed inspection of individual neighborhood market behavior over time.

Together, these two components transform the forecasting results into a practical decision-support platform for homebuyers, investors, planners and policymakers. Rather than passively viewing results, users can actively explore how affordability, amenities and long-term price trends interact across Staten Island. The full design and functionality of the application are described in the following appendix, and a public access link is provided for live interaction.

## 5.5 Forecasting Limitations

The forecasts are based on an assumption that future inflation and mortgage rate trends follow historical patterns observed during 2017-2021, however, unexpected economic shocks, policy changes or financial disruptions could substantially alter these outcomes. In addition, the model relies on historical structural and neighborhood features and does not explicitly account for future zoning changes, infrastructure investments, large-scale redevelopment or demographic shifts that may influence local price dynamics.

Although XGBoost effectively captures nonlinear relationships, the projections remain statistical estimates rather than guaranteed outcomes, particularly at fine neighborhood scales where small changes in supply or demand can produce large effects. Finally, aggregation at the NTA and neighborhood levels smooths individual property volatility, so the forecasts should be interpreted as indicators of relative growth and decline rather than precise future sale prices.

## CHAPTER 6- CONCLUSION

This capstone project set out to forecast housing prices in Staten Island under a scenario-based economic framework while also quantifying the role of neighborhood-level quality-of-life factors, particularly crime and school quality. Using a dataset that integrates property-level characteristics, neighborhood context and national macroeconomic indicators, the study demonstrates how machine learning methods, particularly XGBoost, can be leveraged to model complex urban housing markets with strong predictive accuracy and interpretability.

The model comparison results confirm that traditional Linear Regression and distance-based kNN models are limited in their ability to capture Staten Island's highly nonlinear and spatially heterogeneous housing dynamics. While both approaches provided useful benchmarks, they produced relatively large prediction errors and weaker explanatory power. In contrast, XGBoost significantly outperformed both alternatives across all evaluation metrics. Its superior metrics, along with improved residual behavior and stability across price ranges, confirm that gradient-boosted decision trees are better suited for modeling real estate markets characterized by strong interactions between structure, neighborhood conditions and macroeconomic forces.

Feature importance analysis further advanced the study's second research objective by identifying the dominant drivers of housing prices. Structural characteristics, particularly lot size, building age and year of sale, emerged as the most influential predictors. At the same time, neighborhood crime rates and school quality ranked among the most important non-structural variables, confirming that quality-of-life conditions deploy a direct and measurable influence on housing values. These results align closely with the existing real estate and urban economics literature and validate the inclusion of both crime and educational quality as major explanatory features in housing price modeling.

The results of this analysis indicate that Staten Island's housing market is expected to evolve unevenly over the 2025-2030 forecast horizon, with meaningful variation in both the magnitude and trajectory of price growth across geographic areas and housing types. At the Neighborhood Tabulation Area (NTA) level, forecasted prices exhibit positive relative growth across all areas shown, though the strength of appreciation varies considerably. Southern and interior NTAs are projected to experience stronger relative gains, while other areas show more moderate growth, underscoring the importance of spatial context in long-term housing price dynamics.

Neighborhood-level analysis further refines these findings by illustrating how prices for 1-unit and 2-unit homes are expected to evolve over time. Across most neighborhoods, 2-unit properties maintain a persistent pricing premium over single-family homes, reflecting their added functionality and income-generating potential. However, notable exceptions, such as Arrochar, highlight the role of localized factors, such as housing stock composition and neighborhood amenities, in shaping relative demand. Taken together, the flattening of price trajectories across the later forecast years suggests a transition away from rapid post-pandemic appreciation toward a more stabilized market, where long-run structural characteristics increasingly dominate short-term macroeconomic forces.

Beyond static forecasting outputs, this project also launches an applied analytics tool in the form of an interactive R Shiny dashboard. By enabling users to explore spatial price patterns, neighborhood-level trends and school quality overlays in real time, the dashboard enhances accessibility and practical relevance for non-technical stakeholders, including residents, investors and policymakers. This integration of predictive modeling with interactive visualization demonstrates how machine learning can be implemented for real-world housing market analysis and decision support.

Like all forecasting studies, this research is subject to important limitations. The projections are scenario-dependent and assume that future macroeconomic conditions evolve in a manner consistent with historical trends. Unexpected economic shocks, policy shifts or major redevelopment projects could materially alter price trajectories. Additionally, while aggregation to the NTA and neighborhood levels improves stability, it necessarily smooths property-level volatility. As such, the forecasts should be interpreted as indicators of relative spatial growth and decline rather than precise transaction-level predictions.

Overall, this capstone demonstrates the value of combining machine learning, spatial analysis and scenario-based forecasting to better understand and anticipate housing market dynamics. By focusing on Staten Island, a borough often overlooked in large-scale housing research, this study contributes new empirical insight into the borough's evolving real estate landscape. The results show that future housing performance will be shaped not only by national economic conditions but also by deeply localized structural and quality-of-life factors. Future research could extend this framework by incorporating additional economic scenarios, demographic shifts, zoning changes and more detailed housing attributes and neighborhood-level characteristics, as well as by applying the model to other boroughs for comparative urban analysis.

# REFERENCES

[1] Acolin, A., Colburn, G., & Walter, R. (2022). How do single-family homeowners value residential and commercial density? It depends. *Land Use Policy, 113,* 105898. https://doi.org/10.1016/j.landusepol.2021.105898

[2] Ceccato, V., & Wilhelmsson, M. (2019). Do crime hot spots affect housing prices? *Nordic Journal of Criminology, 21*(1), 84–102. https://doi.org/10.1080/2578983X.2019.1662595

[3] ClosedByMo. (2025). *Why Brooklyn homebuyers are heading to Staten Island.* https://www.closedbymo.com/blog/why-brooklyn-homebuyers-are-heading-to-staten-island

[4] Dentler, A., & Rossi, E. (2024). Residents' willingness to pay to avoid crime. *Journal of Housing Economics, 66,* 102024. https://doi.org/10.1016/j.jhe.2024.102024

[5] Ding, X. (2022). Macroeconomic factors affecting housing prices: Take the United States as an example. In *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (pp. 2335-2339). Atlantis Press. https://doi.org/10.2991/aebmr.k.220307.380

[6] El Mouna, L., Silkan, H., Hayf, Y., Nann, M. F., & Tekouabou, S. C. K. (2023). A comparative study of urban house price prediction using machine learning algorithms. *E3S Web of Conferences, 418,* 03001. https://doi.org/10.1051/e3sconf/202341803001

[7] FCIQ Real Estate. (2024). *NYC real estate's surprising strength: Why 2025 could defy economic pressures.* https://www.fciq.ca/real-estate-market-analysis/nyc-real-estates-surprising-strength-why-2025-could-defy-economic-pressures/

[8] Federal Reserve Economic Data. (n.d.). *Economic indicators database.* Federal Reserve Bank of St. Louis. https://fred.stlouis

fed.org

[9] Ferreira, F. V., & Gyourko, J. (2024). *Does political partisanship affect housing supply? Evidence from U.S. cities* (NBER Working Paper No. 31966). National Bureau of Economic Research. https://www.nber.org/papers/w31966

[10] Fu, Y. (2024). A comparative study of house price prediction using linear regression and random forest models. *Highlights in Science Engineering and Technology, 107,* 96–103. https://doi.org/10.54097/vcy5n584

[11] Han, S. M., & Shin, M. J. (2021). Housing prices and government approval: The impact of housing booms on left- and right-wing governments in 16 advanced industrialized countries. *Canadian Journal of Political Science, 54*(1), 163–185. https://doi.org/10.1017/S0008423920001262

[12] Kallberg, J. G., & Shimizu, Y. (2025). Crime measures and housing prices: An analysis using quantile regression and spatial autocorrelation. *Journal of Real Estate Finance and Economics.* https://doi.org/10.1007/s11146-024-09997-w

[13] Leonard, T., Powell-Wiley, T. M., Ayers, C., Murdoch, J. C., Yin, W., & Pruitt, S. L. (2016). Property Values as a Measure of Neighborhoods: An Application of Hedonic Price Theory. *Epidemiology*, *27*(4), 518–524. https://www.jstor.org/stable/26511765

[15] McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Chapman & Hall/CRC.

[15] MLS Campus. (2025). *Is it worth investing in New York real estate?* https://www.mlscampus.com/is-it-worth-investing-in-new-york-real-estate/

[16] Mora-Garcia, R.-T., Cespedes-Lopez, M.-F., & Perez-Sanchez, V. R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land, 11*(11), 2100. https://doi.org/10.3390/land11112100

[17] National Low Income Housing Coalition. (2024). *The Democratic Party and Republican Party platforms address affordable housing.* https://nlihc.org/resource/democratic-party-and-republican-party-platforms-address-affordable-housing

[18] New York City Department of Buildings (DOB). (n.d.). *DOB permits - Staten Island*. NYC OpenData. https://data.cityofnewyork.us/Housing-Development/DOB-Permits-Staten-Island/fu82-q84b/data_preview

[19] New York City Department of City Planning. (n.d.). *Population FactFinder*. https://popfactfinder.planning.nyc.gov

[20] New York City Department of Finance. (n.d.). *Annualized sales update*. https://www.nyc.gov/site/finance/property/property-annualized-sales-update.page

[21] New York City Police Department. (n.d.). *NYPD Complaint Data Historic*. NYC OpenData. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/about_data

[22] New York City Comptroller. (2024). *The pandemic's impact on NYC migration patterns*. https://comptroller.nyc.gov/reports/the-pandemics-impact-on-nyc-migration-patterns/

[23] PropertyShark. (2024). *Domestic migration in and out of NYC: 2024 snapshot.* https://www.propertyshark.com/Real-Estate-Reports/2024/12/19/domestic-migration-in-and-out-of-nyc-2024-snapshot/

[24] Public School Review. (n.d.). *Staten Island, NY public schools*. https://www.publicschoolreview.com/new-york/staten-island

[25] Schwartz, A. E., Susin, S., & Voicu, I. (2003). Has falling crime driven New York City's real estate boom? *Journal of Housing Research, 14,* 101–135. https://doi.org/10.1080/26911337.2003.12519486

[26] Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics, 3*(1), 30-45. https://doi.org/10.3390/analytics3010003

[27] Sharma, M., Sharma, D., Burle, R., Patil, P., Joge, I., & Puri, C. (2024). Predicting house price model: A comprehensive analysis with Random Forest and Decision Tree Method. In *2024 3rd International Conference for Innovation in Technology (INOCON)* (pp. 1-6). IEEE. https://doi.org/10.1109/INOCON60754.2024.10511732

[28] Sheng, C., & Yu, H. (2022). An optimized prediction algorithm based on XGBoost. In *2022 International Conference on Networking and Network Applications (NaNA)* (pp. 1-6). IEEE. https://doi.org/10.1109/NaNA56854.2022.00082

[29] Shi, R. (2024). *Battling for housing value? The nexus between U.S. presidential elections and county-level housing market prices* [Unpublished manuscript]. Business School, Sun Yat-sen University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5027026

[30] TJV News. (2025). *Brooklyn exodus: Rising prices push thousands to Staten Island.* https://tjvnews.com/local/new-york/brooklyn-exodus-rising-prices-push-thousands-to-staten-island

[31] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. In *2020 7th NAFOSTED Conference on*

*Information and Computer Science (NICS)* (pp. 1-6). IEEE. https://doi.org/10.1109/NICS51282.2020.9335870

[32] Yan, L. (2024). Predicting house prices with a linear regression model. *Applied and Computational Engineering, 114,* 107–115. https://doi.org/10.54254/2755-2721/2024.18220

[33] Yu, J. (2024). Predicting New York housing prices: A comparative analysis of machine learning models. In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA)* (pp. 102-109). SciTePress. https://doi.org/10.5220/0012999000004601

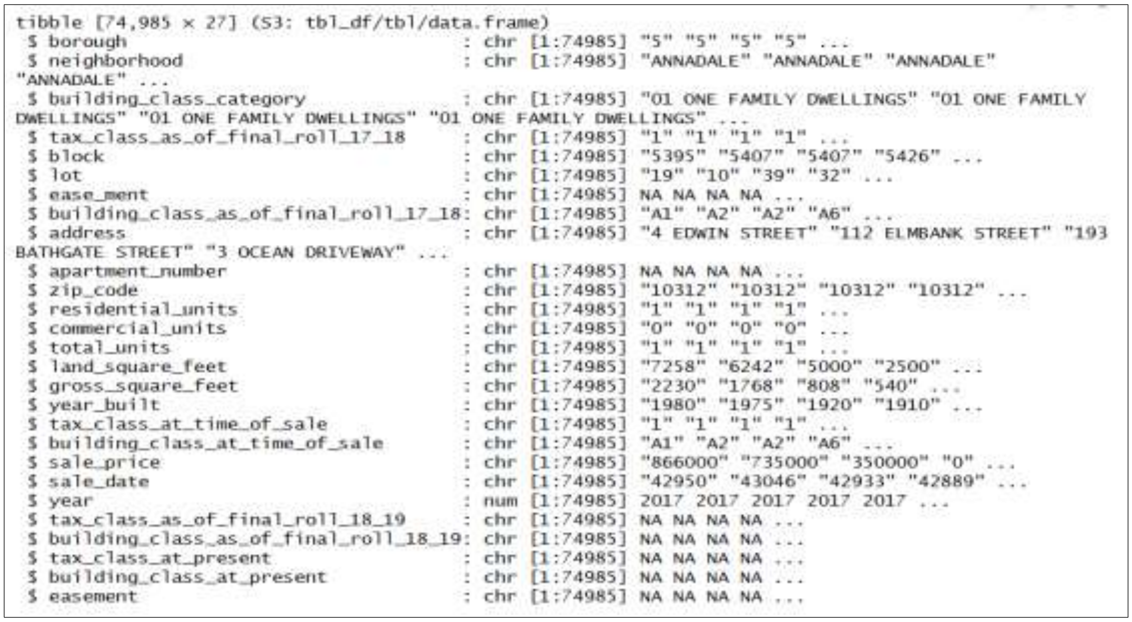
[34] Zhang, L. (2023). Housing price prediction using machine learning algorithm. *Journal of World Economy, 2*(3), 18–26. https://doi.org/10.56397/JWE.2023.09.03

# APPENDIX A

### A.1 Raw Staten Island Sales Data Structure (2017–2025)

The initial raw housing dataset (si_sales_raw) contained 74,985 observations and 27 variables compiled across the 2017–2025 period. The dataset included a mixture of structural, geographic and transactional attributes, such as borough, neighborhood, building class, block and lot identifiers, address information, zip codes, residential and commercial unit counts, lot size, gross building square footage, year built, assessed tax class and recorded sale prices.

All variables were initially loaded as character fields due to inconsistencies in the raw NYC Department of Finance Excel files. This raw structure required extensive cleaning and type standardization before analysis.



```
tibble [74,985 × 27] (S3: tbl_df/tbl/data.frame)
 $ borough                           : chr [1:74985] "5" "5" "5" "5" ...
 $ neighborhood                      : chr [1:74985] "ANNADALE" "ANNADALE" "ANNADALE"
"ANNADALE" ...
 $ building_class_category           : chr [1:74985] "01 ONE FAMILY DWELLINGS" "01 ONE FAMILY
DWELLINGS" "01 ONE FAMILY DWELLINGS" "01 ONE FAMILY DWELLINGS" ...
 $ tax_class_as_of_final_roll_17_18     : chr [1:74985] "1" "1" "1" "1" ...
 $ block                             : chr [1:74985] "5395" "5407" "5407" "5426" ...
 $ lot                               : chr [1:74985] "19" "10" "39" "32" ...
 $ ease_ment                         : chr [1:74985] NA NA NA NA ...
 $ building_class_as_of_final_roll_17_18: chr [1:74985] "A1" "A2" "A2" "A6" ...
 $ address                           : chr [1:74985] "4 EDWIN STREET" "112 ELMBANK STREET" "193
BATHGATE STREET" "3 OCEAN DRIVEWAY" ...
 $ apartment_number                  : chr [1:74985] NA NA NA NA ...
 $ zip_code                          : chr [1:74985] "10312" "10312" "10312" "10312" ...
 $ residential_units                 : chr [1:74985] "1" "1" "1" "1" ...
 $ commercial_units                  : chr [1:74985] "0" "0" "0" "0" ...
 $ total_units                       : chr [1:74985] "1" "1" "1" "1" ...
 $ land_square_feet                  : chr [1:74985] "7258" "6242" "5000" "2500" ...
 $ gross_square_feet                 : chr [1:74985] "2230" "1768" "808" "540" ...
 $ year_built                        : chr [1:74985] "1980" "1975" "1920" "1910" ...
 $ tax_class_at_time_of_sale         : chr [1:74985] "1" "1" "1" "1" ...
 $ building_class_at_time_of_sale    : chr [1:74985] "A1" "A2" "A2" "A6" ...
 $ sale_price                        : chr [1:74985] "866000" "735000" "350000" "0" ...
 $ sale_date                         : chr [1:74985] "42950" "43046" "42933" "42889" ...
 $ year                              : num [1:74985] 2017 2017 2017 2017 2017 ...
 $ tax_class_as_of_final_roll_18_19     : chr [1:74985] NA NA NA NA ...
 $ building_class_as_of_final_roll_18_19: chr [1:74985] NA NA NA NA ...
 $ tax_class_at_present              : chr [1:74985] NA NA NA NA ...
 $ building_class_at_present         : chr [1:74985] NA NA NA NA ...
 $ easement                          : chr [1:74985] NA NA NA NA ...
```

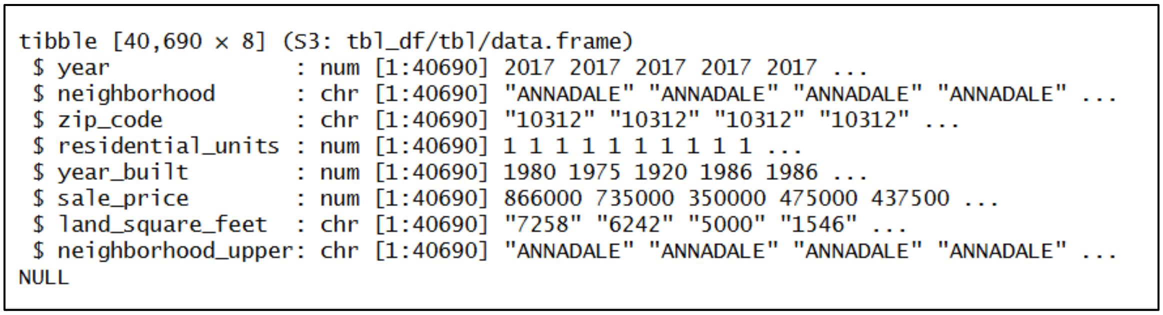
Table A1

## A.2 Initial Data Cleaning and Residential Filtering

To prepare the modeling dataset, the following transformations were applied:

- Numeric conversion of sale_price, residential_units and year_built
- Removal of invalid transactions, defined as:
  - Missing sale prices
  - Sale prices below $10,000
- Filtering exclusively to residential properties by selecting only records labeled as "01 ONE FAMILY DWELLINGS" and "FAMILY DWELLING" in building_class_category
- Variable selection to retain only core modeling features: year - transaction year, neighborhood - NYC neighborhood name, zip_code – Postal zip, residential_units - Number of housing units, year_built - Construction year, sale_price - Transaction value (USD), land_square_feet - Lot size.

This process created the cleaned dataset si_sales_clean with 40,690 rows and 8 modeling variables.



```
tibble [40,690 × 8] (S3: tbl_df/tbl/data.frame)
 $ year              : num [1:40690] 2017 2017 2017 2017 2017 ...
 $ neighborhood      : chr [1:40690] "ANNADALE" "ANNADALE" "ANNADALE" "ANNADALE" ...
 $ zip_code          : chr [1:40690] "10312" "10312" "10312" "10312" ...
 $ residential_units : num [1:40690] 1 1 1 1 1 1 1 1 1 1 ...
 $ year_built        : num [1:40690] 1980 1975 1920 1986 1986 ...
 $ sale_price        : num [1:40690] 866000 735000 350000 475000 437500 ...
 $ land_square_feet  : chr [1:40690] "7258" "6242" "5000" "1546" ...
 $ neighborhood_upper: chr [1:40690] "ANNADALE" "ANNADALE" "ANNADALE" "ANNADALE" ...
NULL
```

Table A2

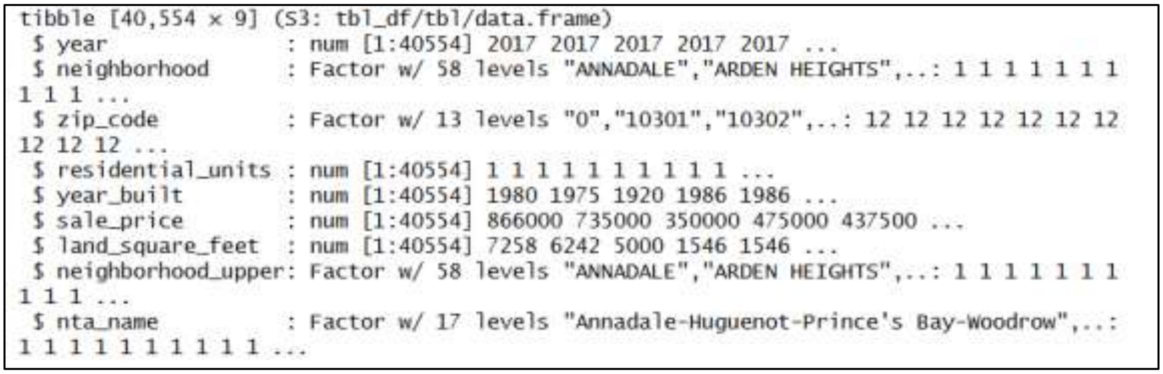
### A.3 Neighborhood Standardization and NTA Geographic Mapping

Neighborhood names in the raw records exhibited high variation in spelling, naming and formatting. To ensure compatibility with official NYC spatial boundaries, the following steps were applied. First, neighborhood names were standardized by converting all values to uppercase and trimming excess spacing. Then the  official Neighborhood Tabulation Area (NTA) boundaries were imported using the NYC Planning shapefile (nynta2020.shp) and the spatial layer was filtered to include only Staten Island polygons. Several official NTA names were then  manually harmonized to ensure alignment with common neighborhood naming conventions. Those were  “Snug Harbor" → "St. George-New Brighton", "Great Kills Park" → "Great Kills-Eltingville", "Miller Field" → "New Dorp-Midland Beach".

Each residential transaction was then mapped into an NTA using a rule-based text matching approach (case_when() with regular expressions). Examples include:

- "ANNADALE | HUGUENOT | PRINCES BAY" → "Annadale-Huguenot-Prince's Bay-Woodrow"
- "TOMPKINSVILLE | STAPLETON" → "Tompkinsville-Stapleton-Clifton-Fox Hills"
- "PORT RICHMOND" → "Port Richmond"

All unmatched neighborhoods were temporarily assigned "Other" and reviewed manually. After verification, unmatched rows were removed. The final mapped dataset si_sales_mapped contains 40,554 transactions assigned to one of 17 official Staten Island NTAs.



```
tibble [40,554 × 9] (S3: tbl_df/tbl/data.frame)
 $ year              : num [1:40554] 2017 2017 2017 2017 2017 ...
 $ neighborhood      : Factor w/ 58 levels "ANNADALE","ARDEN HEIGHTS",..: 1 1 1 1 1 1 1
1 1 1 ...
 $ zip_code          : Factor w/ 13 levels "0","10301","10302",..: 12 12 12 12 12 12 12
12 12 12 ...
 $ residential_units : num [1:40554] 1 1 1 1 1 1 1 1 1 1 ...
 $ year_built        : num [1:40554] 1980 1975 1920 1986 1986 ...
 $ sale_price        : num [1:40554] 866000 735000 350000 475000 437500 ...
 $ land_square_feet  : num [1:40554] 7258 6242 5000 1546 1546 ...
 $ neighborhood_upper: Factor w/ 58 levels "ANNADALE","ARDEN HEIGHTS",..: 1 1 1 1 1 1 1
1 1 1 ...
 $ nta_name          : Factor w/ 17 levels "Annadale-Huguenot-Prince's Bay-Woodrow",..:
1 1 1 1 1 1 1 1 1 1 ...
```

Table A3

Additional transformations included:

- Conversion of land_square_feet from character to numeric
- Removal of implausible year_built values equal to zero

These steps ensured numeric stability for downstream modeling and spatial aggregation.

### A.4 Crime Data Spatial Processing and NTA Aggregation

Crime data for Staten Island was processed from annual NYPD complaint records using the following workflow. The first step was to import all  raw incident records from CSV format. Latitude and longitude fields were then converted into spatial point features using st_as_sf() with WGS 84 projection. Crime points were reprojected to match the Staten Island NTA coordinate reference system and a spatial join (st_join) assigned each crime incident to an NTA polygon.

Each crime record was then reduced to year (incident year), OFNS_DESC (Offense description) and NTAName. Rows with missing NTA assignment were removed. Incident data from all years was then aggregated into a unified crime dataset.

### A.5 Crime Aggregation and Total NTA Incident Counts

Crime incidents were summarized at the NTA–year level using grouping by year and NTAName. Then after adding the NTA population into the dataset, counting total incidents per year and merging aggregated counts back into the Staten Island spatial map

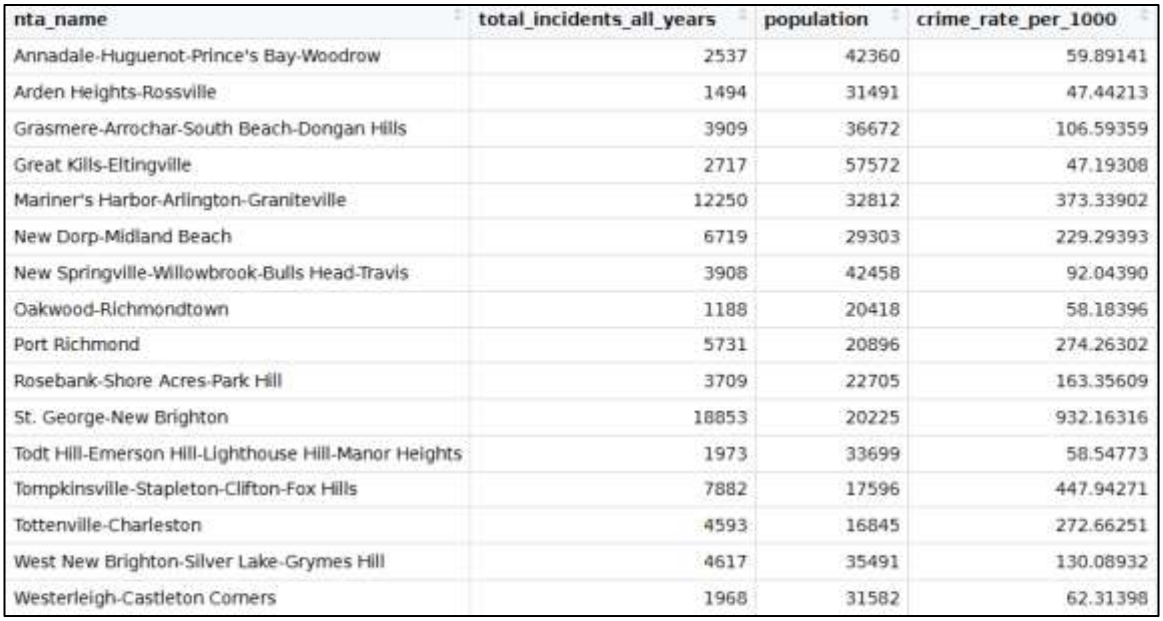
A cumulative crime intensity variable named Total incidents across all years per NTA was then created.

This produced a ranked NTA crime table identifying the most persistently high-crime areas, including:

- St. George–New Brighton
- Mariner's Harbor–Arlington–Graniteville

- Tompkinsville–Stapleton–Clifton–Fox Hills
- Port Richmond

These totals formed the basis for computing crime rates per 1,000 residents used in the XGBoost model.



| nta_name | total_incidents_all_years | population | crime_rate_per_1000 |
|---|---:|---:|---:|
| Annadale-Huguenot-Prince's Bay-Woodrow | 2537 | 42360 | 59.89141 |
| Arden Heights-Rossville | 1494 | 31491 | 47.44213 |
| Grasmere-Arrochar-South Beach-Dongan Hills | 3909 | 36672 | 106.59359 |
| Great Kills-Eltingville | 2717 | 57572 | 47.19308 |
| Mariner's Harbor-Arlington-Graniteville | 12250 | 32812 | 373.33902 |
| New Dorp-Midland Beach | 6719 | 29303 | 229.29393 |
| New Springville-Willowbrook-Bulls Head-Travis | 3908 | 42458 | 92.04390 |
| Oakwood-Richmondtown | 1188 | 20418 | 58.18396 |
| Port Richmond | 5731 | 20896 | 274.26302 |
| Rosebank-Shore Acres-Park Hill | 3709 | 22705 | 163.35609 |
| St. George-New Brighton | 18853 | 20225 | 932.16316 |
| Todt Hill-Emerson Hill-Lighthouse Hill-Manor Heights | 1973 | 33699 | 58.54773 |
| Tompkinsville-Stapleton-Clifton-Fox Hills | 7882 | 17596 | 447.94271 |
| Tottenville-Charleston | 4593 | 16845 | 272.66251 |
| West New Brighton-Silver Lake-Grymes Hill | 4617 | 35491 | 130.08932 |
| Westerleigh-Castleton Corners | 1968 | 31582 | 62.31398 |

Table A4

### A.6 Supplemental Exploratory Diagnostics

Although not shown in the main text due to space constraints, the following diagnostics were generated to validate modeling assumptions: Histograms for year, residential units, year built, sale price, land square footage and boxplots for identifying heavy right skew in sale prices and lot size. Annual transaction counts (2017–2025) confirming post-2020 volatility. Unscaled spatial price heatmaps confirming persistent north–south price gradients. The following diagnostics confirmed diagnostics confirmed: strong right skew in price and lot size and abnormal sales volatility during 2020–2022. All findings directly supported the choice of log-transformations, outlier capping and non-linear machine learning methods in the main modeling pipeline. Also Figure A5 presents the full spatio-temporal evolution of average housing prices and crime incidents by NTA across Staten Island from 2017–2025.

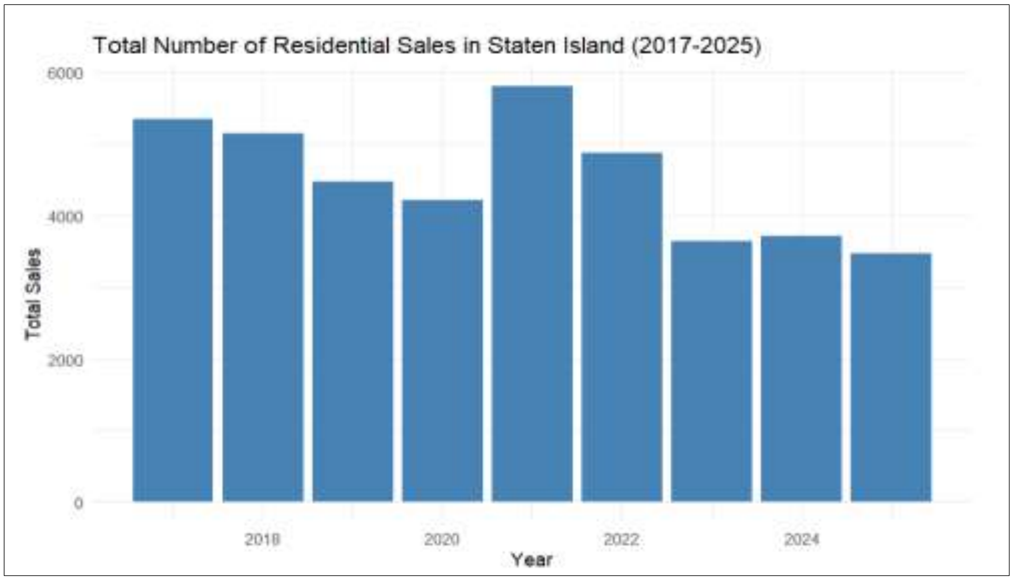


Figure A1

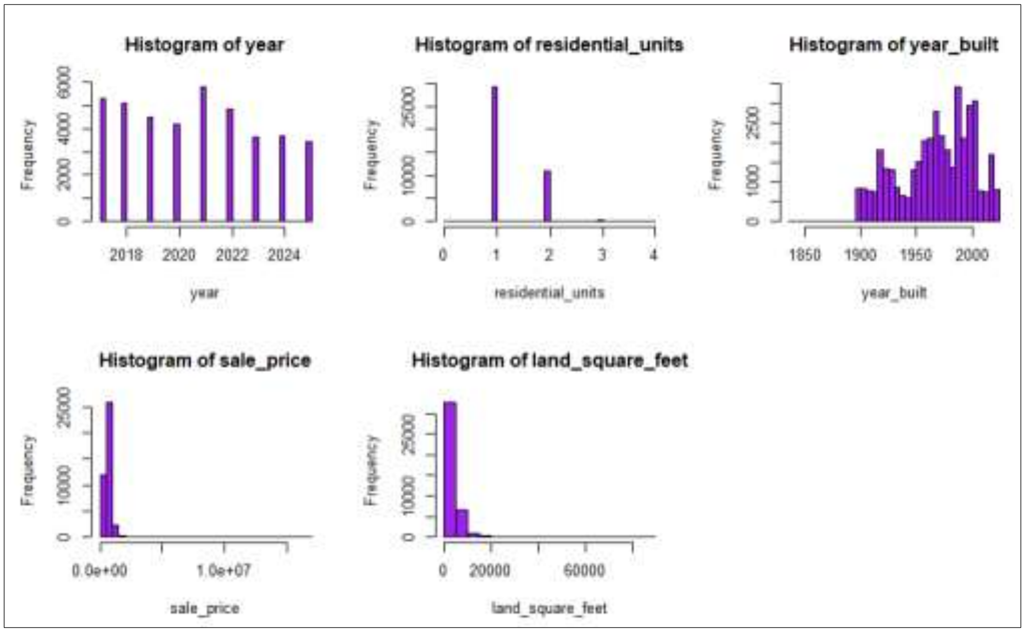


Figure A2

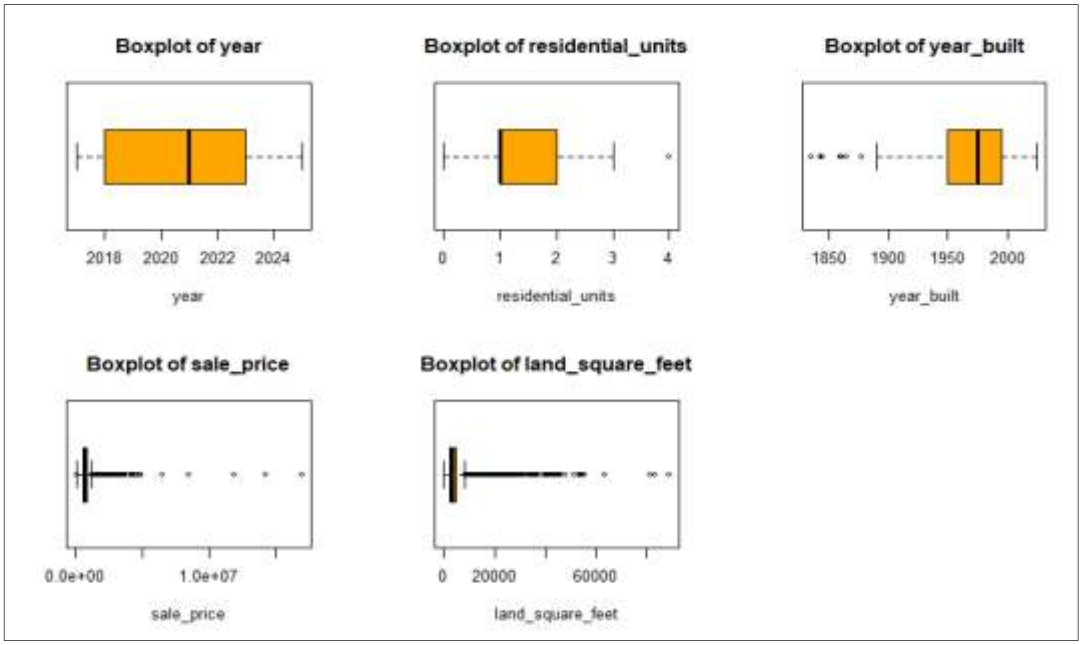


Figure A3

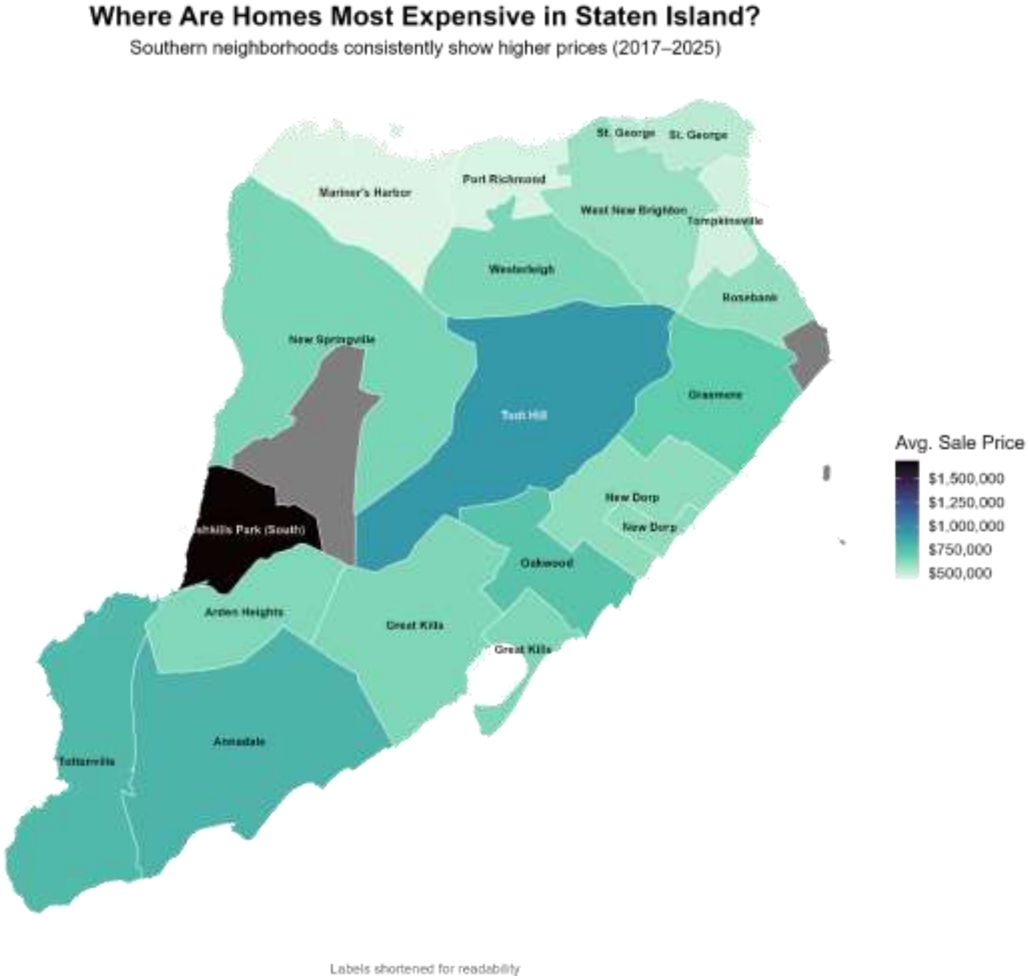


Figure A4

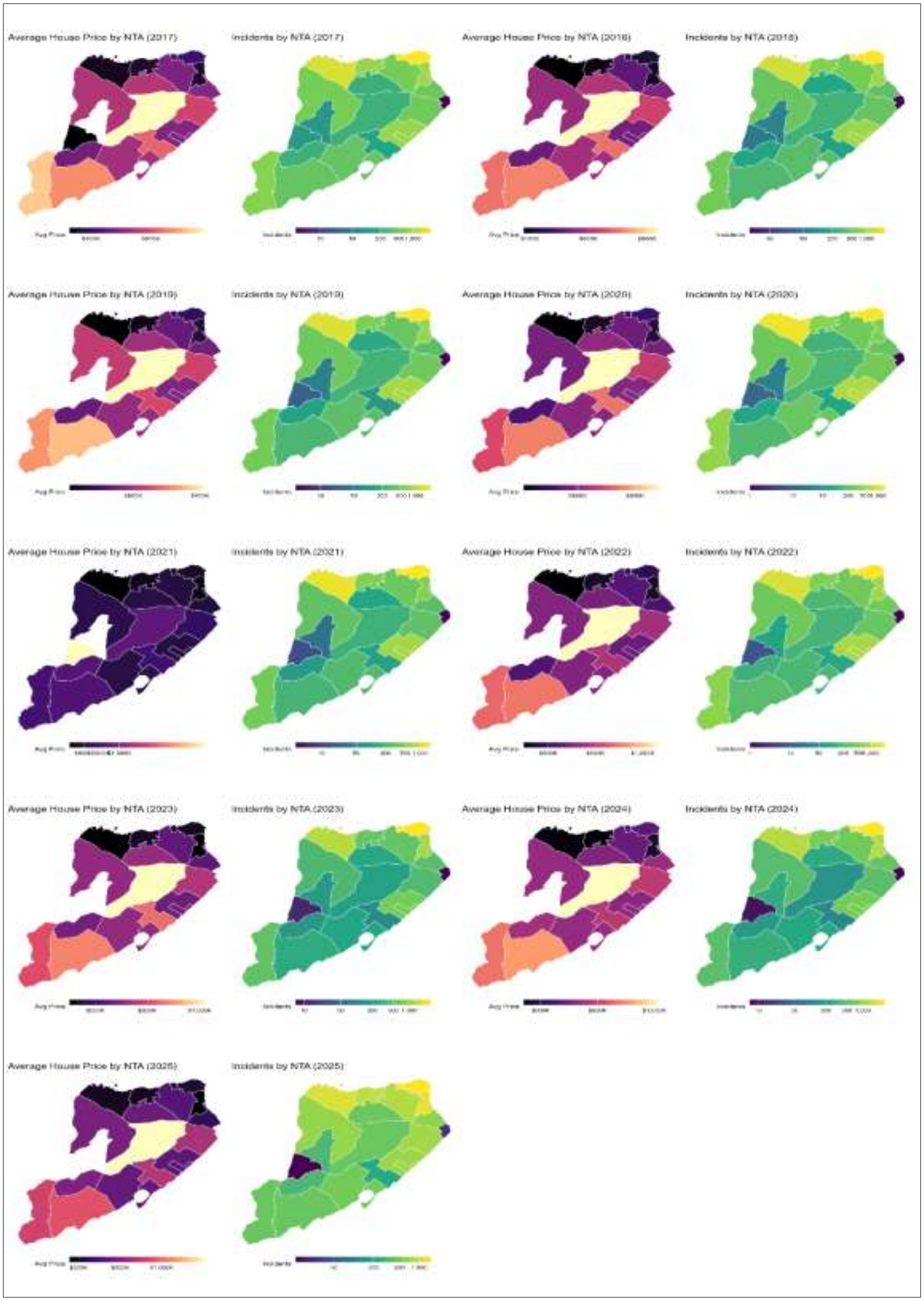

Average House Price by NTA (2017)
Incidents by NTA (2017)
Average House Price by NTA (2018)
Incidents by NTA (2018)
Average House Price by NTA (2019)
Incidents by NTA (2019)
Average House Price by NTA (2020)
Incidents by NTA (2020)
Average House Price by NTA (2021)
Incidents by NTA (2021)
Average House Price by NTA (2022)
Incidents by NTA (2022)
Average House Price by NTA (2023)
Incidents by NTA (2023)
Average House Price by NTA (2024)
Incidents by NTA (2024)
Average House Price by NTA (2025)
Incidents by NTA (2025)
Avg Price
Incidents

Figure A5

## A.7 School Rankings and Housing Price Relationship

To evaluate the relationship between neighborhood school quality and housing values, school ranking data were merged with NTA-level housing prices. A scatterplot with an regression line was produced to visualize the relationship.The results show a positive association between higher school rankings and higher average housing prices, supporting the inclusion of school quality as an explanatory variable in the forecasting model.

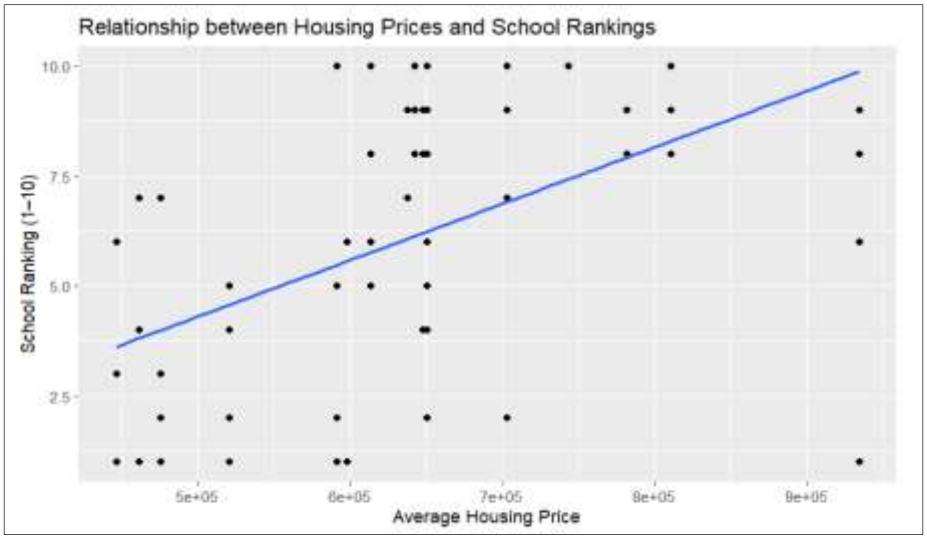
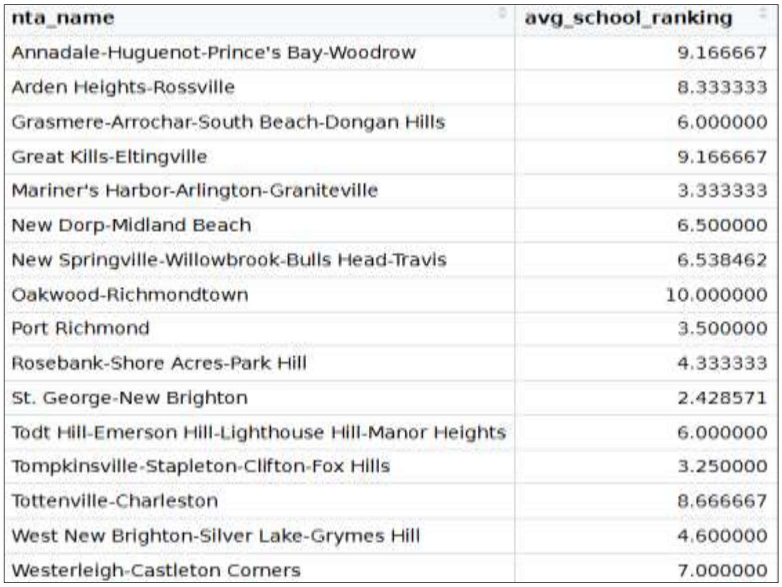


Figure A6

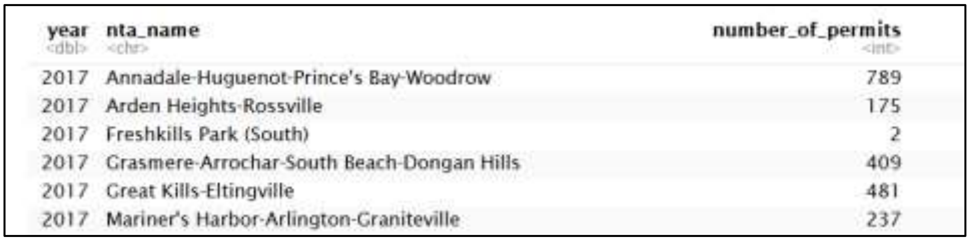


| nta_name | avg_school_ranking |
|---|---|
| Annadale-Huguenot-Prince's Bay-Woodrow | 9.166667 |
| Arden Heights-Rossville | 8.333333 |
| Grasmere-Arrochar-South Beach-Dongan Hills | 6.000000 |
| Great Kills-Eltingville | 9.166667 |
| Mariner's Harbor-Arlington-Graniteville | 3.333333 |
| New Dorp-Midland Beach | 6.500000 |
| New Springville-Willowbrook-Bulls Head-Travis | 6.538462 |
| Oakwood-Richmondtown | 10.000000 |
| Port Richmond | 3.500000 |
| Rosebank-Shore Acres-Park Hill | 4.333333 |
| St. George-New Brighton | 2.428571 |
| Todt Hill-Emerson Hill-Lighthouse Hill-Manor Heights | 6.000000 |
| Tompkinsville-Stapleton-Clifton-Fox Hills | 3.250000 |
| Tottenville-Charleston | 8.666667 |
| West New Brighton-Silver Lake-Grymes Hill | 4.600000 |
| Westerleigh-Castleton Corners | 7.000000 |

Table A6

## A.8 Building Permit Activity by NTA

To capture development intensity and investment activity, NYC Department of Buildings permit records were spatially joined to the Staten Island NTA boundaries. Annual counts of issued permits were then aggregated at the NTA–year level. This variable serves as a proxy for construction activity and neighborhood reinvestment across Staten Island.

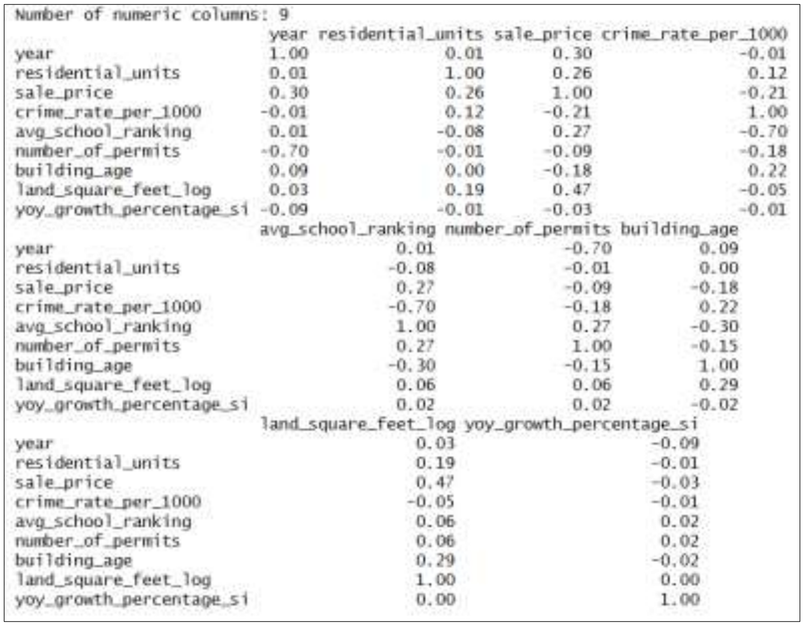


| year | nta_name | number_of_permits |
|---|---|---|
| <dbl> | <chr> | <int> |
| 2017 | Annadale-Huguenot-Prince's Bay-Woodrow | 789 |
| 2017 | Arden Heights-Rossville | 175 |
| 2017 | Freshkills Park (South) | 2 |
| 2017 | Grasmere-Arrochar-South Beach-Dongan Hills | 409 |
| 2017 | Great Kills-Eltingville | 481 |
| 2017 | Mariner's Harbor-Arlington-Graniteville | 237 |

Table A7

## A.9 Correlation Matrix and Multicollinearity Assessment

Before fitting the forecasting models, a correlation matrix of all numeric predictors was computed. All pairwise correlations remained well below the ±0.90 threshold, indicating no evidence of problematic multicollinearity.



```
Number of numeric columns: 9
                          year residential_units sale_price crime_rate_per_1000
year                      1.00              0.01       0.30               -0.01
residential_units         0.01              1.00       0.26                0.12
sale_price                0.30              0.26       1.00               -0.21
crime_rate_per_1000      -0.01              0.12      -0.21                1.00
avg_school_ranking        0.01             -0.08       0.27               -0.70
number_of_permits        -0.70             -0.01      -0.09               -0.18
building_age              0.09              0.00      -0.18                0.22
land_square_feet_log      0.03              0.19       0.47               -0.05
yoy_growth_percentage_si -0.09             -0.01      -0.03               -0.01
                         avg_school_ranking number_of_permits building_age
year                                   0.01             -0.70         0.09
residential_units                     -0.08             -0.01         0.00
sale_price                             0.27             -0.09        -0.18
crime_rate_per_1000                   -0.70             -0.18         0.22
avg_school_ranking                     1.00              0.27        -0.30
number_of_permits                      0.27              1.00        -0.15
building_age                          -0.30             -0.15         1.00
land_square_feet_log                   0.06              0.06         0.29
yoy_growth_percentage_si               0.02              0.02        -0.02
                         land_square_feet_log yoy_growth_percentage_si
year                                     0.03                    -0.09
residential_units                        0.19                    -0.01
sale_price                               0.47                    -0.03
crime_rate_per_1000                     -0.05                    -0.01
avg_school_ranking                       0.06                     0.02
number_of_permits                        0.06                     0.02
building_age                             0.29                    -0.02
land_square_feet_log                     1.00                     0.00
yoy_growth_percentage_si                 0.00                     1.00
```

Table A8

## A.10 Model Diagnostic Figures

The residual diagnostics exhibit a “double funnel” pattern, with error variance increasing toward both the lower and upper ends of the predicted price range. This reflects heteroscedasticity and regression to the mean effects, as extreme-priced properties are less common and more heterogeneous, leading to greater prediction uncertainty. Consequently, aggregate model metrics should be interpreted as average performance measures rather than uniform accuracy across all price segments.

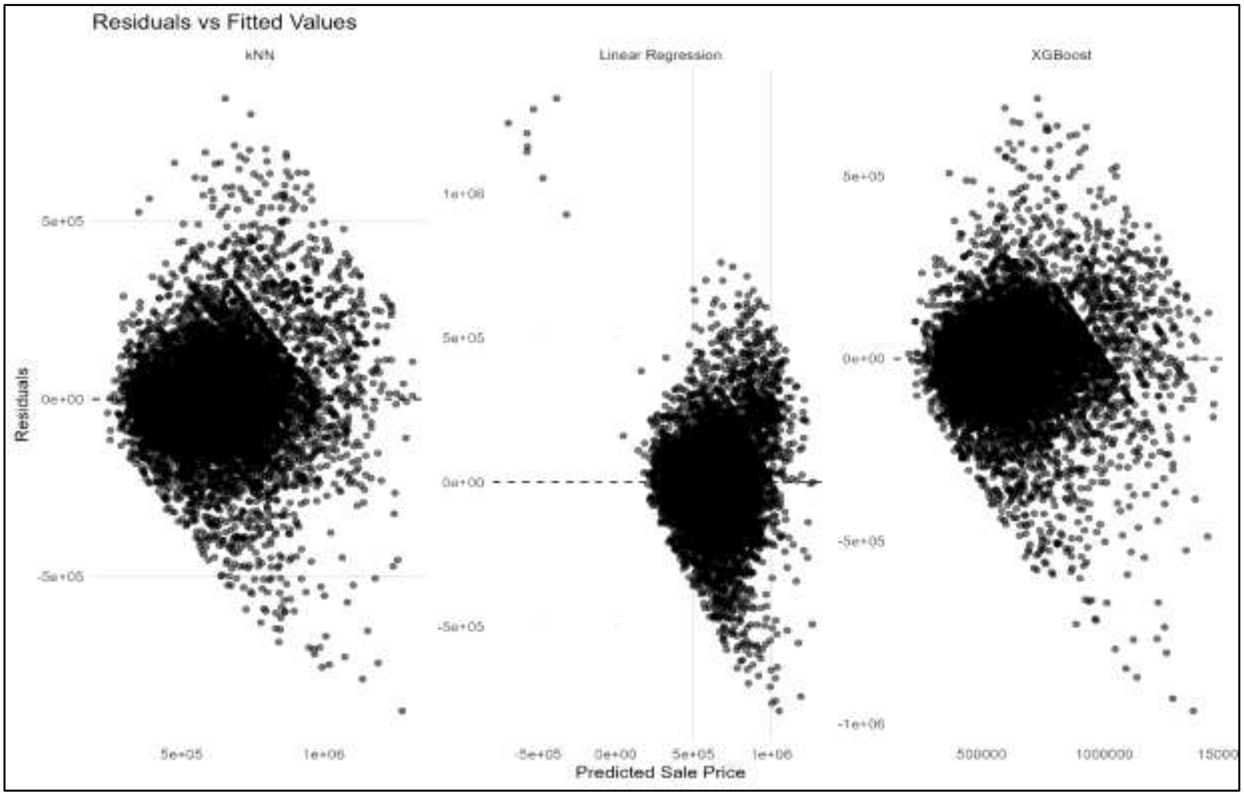


Figure A7. Residuals vs Predicted Sale Prices for Linear Regression, kNN and XGBoost

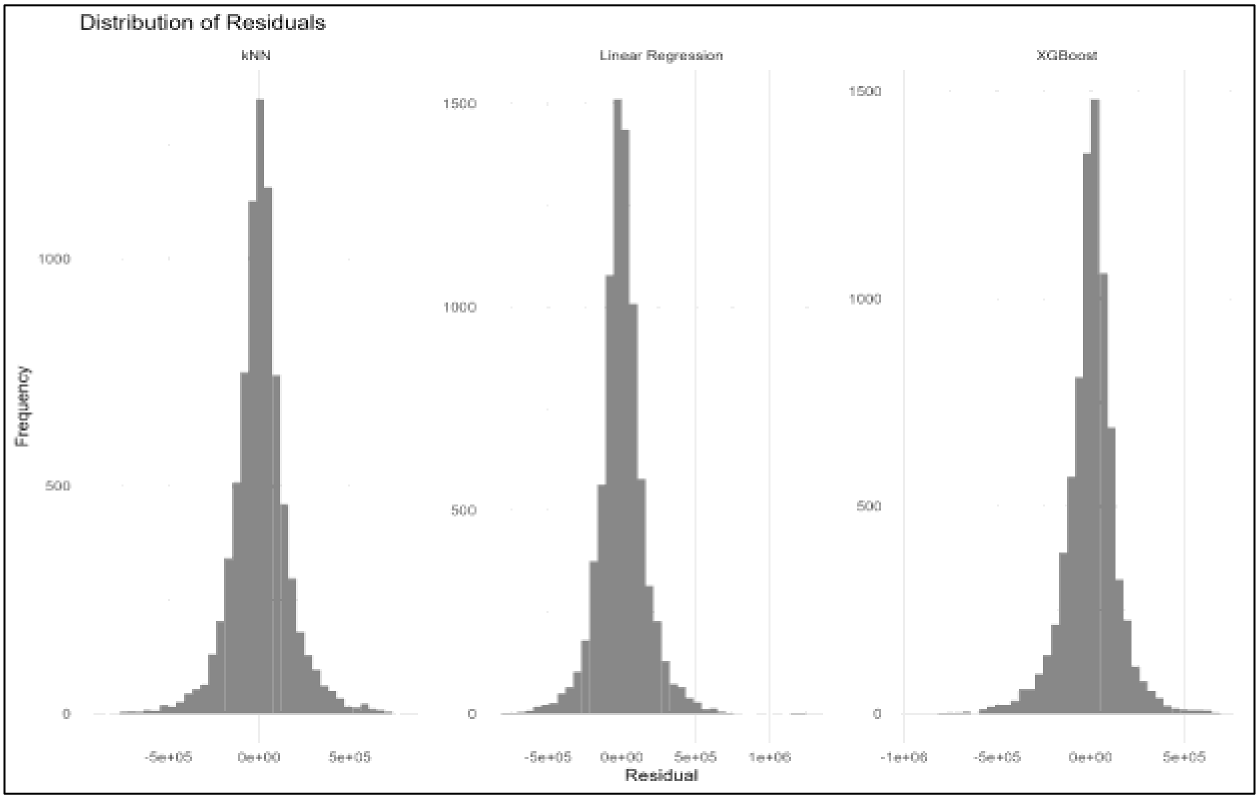


Figure A8. Distribution of Prediction Errors Across the Three Models

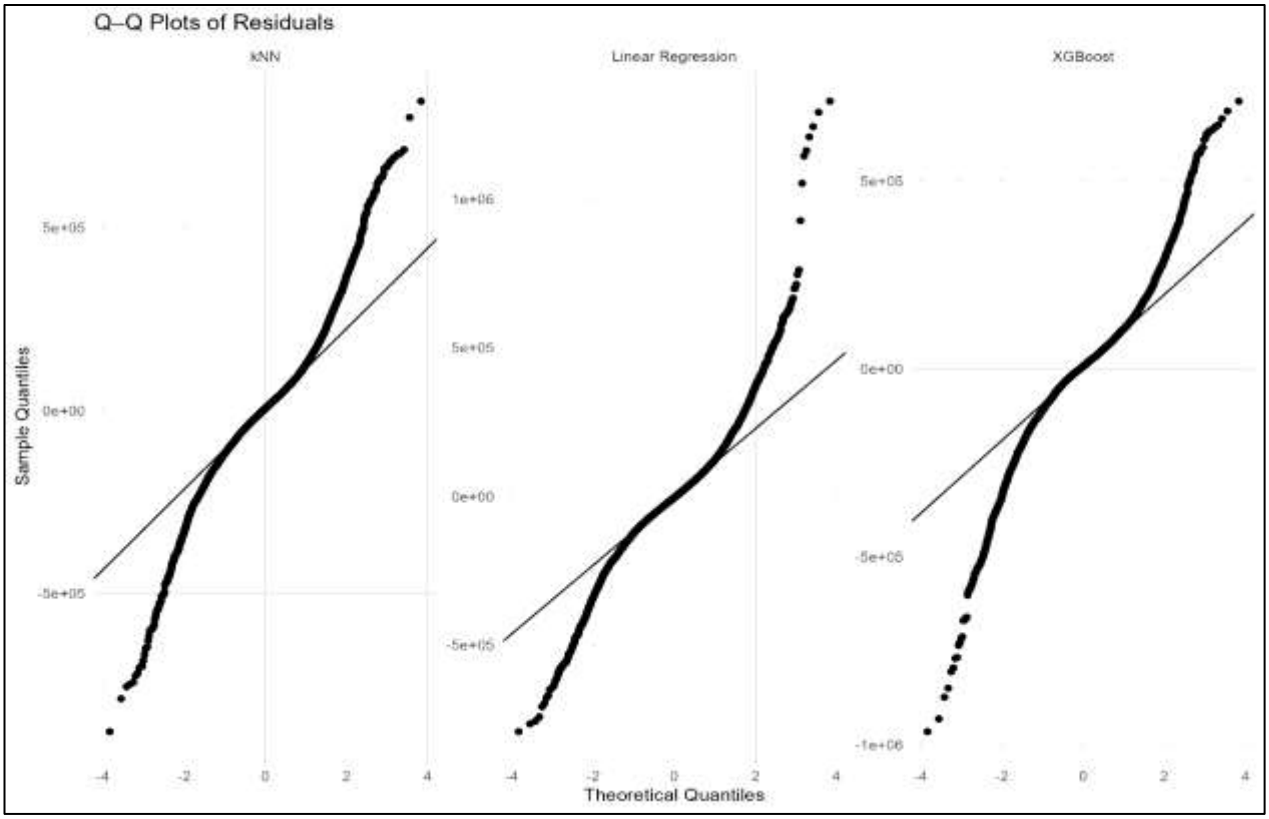
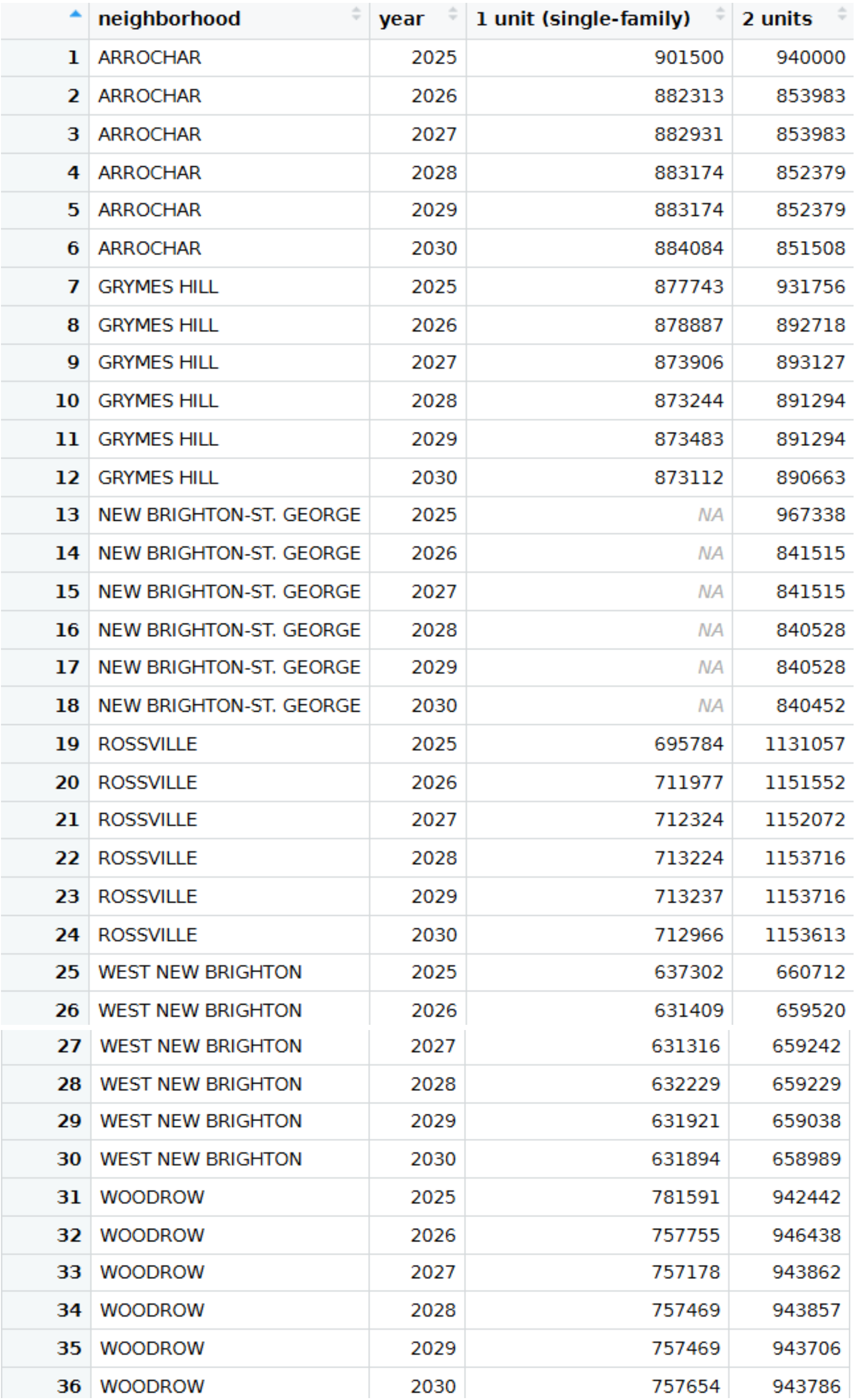


Figure A9. Q–Q Plots of Residuals

## A.11 Forecasted Housing Prices by Neighborhood and Unit Type (2025–2030)



| | neighborhood | year | 1 unit (single-family) | 2 units |
|---|---|---|---|---|
| 1 | ARROCHAR | 2025 | 901500 | 940000 |
| 2 | ARROCHAR | 2026 | 882313 | 853983 |
| 3 | ARROCHAR | 2027 | 882931 | 853983 |
| 4 | ARROCHAR | 2028 | 883174 | 852379 |
| 5 | ARROCHAR | 2029 | 883174 | 852379 |
| 6 | ARROCHAR | 2030 | 884084 | 851508 |
| 7 | GRYMES HILL | 2025 | 877743 | 931756 |
| 8 | GRYMES HILL | 2026 | 878887 | 892718 |
| 9 | GRYMES HILL | 2027 | 873906 | 893127 |
| 10 | GRYMES HILL | 2028 | 873244 | 891294 |
| 11 | GRYMES HILL | 2029 | 873483 | 891294 |
| 12 | GRYMES HILL | 2030 | 873112 | 890663 |
| 13 | NEW BRIGHTON-ST. GEORGE | 2025 | NA | 967338 |
| 14 | NEW BRIGHTON-ST. GEORGE | 2026 | NA | 841515 |
| 15 | NEW BRIGHTON-ST. GEORGE | 2027 | NA | 841515 |
| 16 | NEW BRIGHTON-ST. GEORGE | 2028 | NA | 840528 |
| 17 | NEW BRIGHTON-ST. GEORGE | 2029 | NA | 840528 |
| 18 | NEW BRIGHTON-ST. GEORGE | 2030 | NA | 840452 |
| 19 | ROSSVILLE | 2025 | 695784 | 1131057 |
| 20 | ROSSVILLE | 2026 | 711977 | 1151552 |
| 21 | ROSSVILLE | 2027 | 712324 | 1152072 |
| 22 | ROSSVILLE | 2028 | 713224 | 1153716 |
| 23 | ROSSVILLE | 2029 | 713237 | 1153716 |
| 24 | ROSSVILLE | 2030 | 712966 | 1153613 |
| 25 | WEST NEW BRIGHTON | 2025 | 637302 | 660712 |
| 26 | WEST NEW BRIGHTON | 2026 | 631409 | 659520 |
| 27 | WEST NEW BRIGHTON | 2027 | 631316 | 659242 |
| 28 | WEST NEW BRIGHTON | 2028 | 632229 | 659229 |
| 29 | WEST NEW BRIGHTON | 2029 | 631921 | 659038 |
| 30 | WEST NEW BRIGHTON | 2030 | 631894 | 658989 |
| 31 | WOODROW | 2025 | 781591 | 942442 |
| 32 | WOODROW | 2026 | 757755 | 946438 |
| 33 | WOODROW | 2027 | 757178 | 943862 |
| 34 | WOODROW | 2028 | 757469 | 943857 |
| 35 | WOODROW | 2029 | 757469 | 943706 |
| 36 | WOODROW | 2030 | 757654 | 943786 |

Table A9

## A.12 Interactive App (Shiny Application)

This appendix provides supporting documentation for the interactive forecasting dashboard discussed in Section 5.4. The Shiny application operationalizes the forecasting results into a user-driven decision-support tool for exploring Staten Island's projected housing dynamics at both the NTA and neighborhood levels.

The application consists of two integrated components. The "NTA Finder" tab employs a preference-based quiz that allows users to filter Neighborhood Tabulation Areas (NTAs) according to forecasted budget range, school quality preferences, neighborhood housing age, train access and geographic preference. Based on the selected criteria, the application returns a customized list of NTAs along with their corresponding forecasted average home prices. Selecting any recommended NTA generates an interactive time-series visualization of historical and projected prices in addition to a supporting summary table.

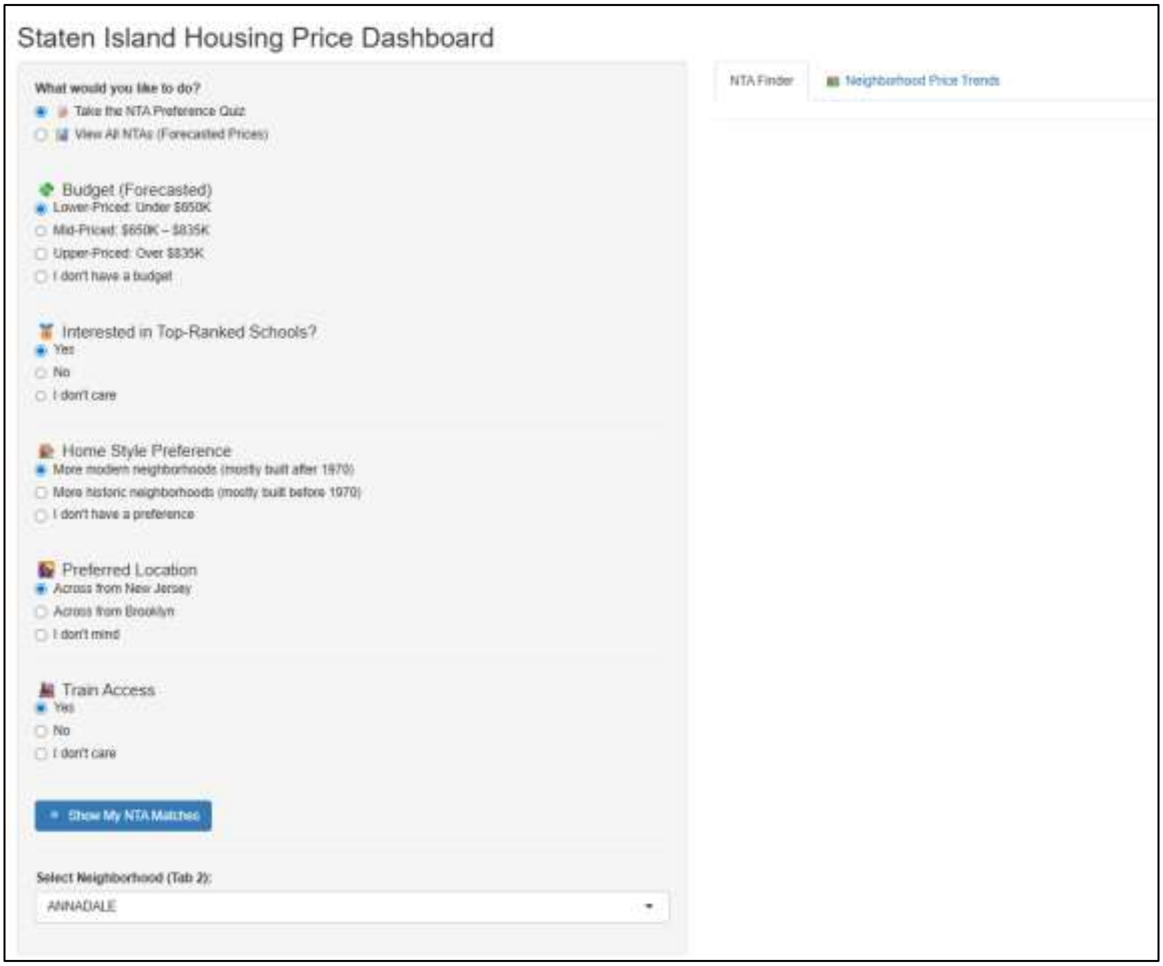
The second tab provides an open neighborhood exploration tool, where users may select any Staten Island neighborhood from a dropdown menu to examine its historical price trajectory and projected future values. This component facilitates unrestricted comparison and detailed inspection of individual neighborhood-level market behavior over time.

The full source code for the dashboard, including the user interface, server logic and supporting data-processing scripts, is publicly available at the following GitHub repository:

GitHub Repository (Shiny Application Code):
https://github.com/NikoletaEm/Capstone/tree/main/ShinyApplication

The fully deployed and publicly accessible version of the application is available at:
https://nicoleemanouilidi.shinyapps.io/finalcapstone/

Figure A8 presents a snapshot of the deployed dashboard interface. Together with the live deployment and complete codebase, this ensures full reproducibility and transparent validation of the interactive forecasting system described in the main text.



Staten Island Housing Price Dashboard

What would you like to do?

● Take the NTA Preference Quiz
○ View All NTAs (Forecasted Prices)

Budget (Forecasted)

● Lower-Priced: Under $650K
○ Mid-Priced: $650K – $835K
○ Upper-Priced: Over $835K
○ I don't have a budget

Interested in Top-Ranked Schools?

● Yes
○ No
○ I don't care

Home Style Preference

● More modern neighborhoods (mostly built after 1970)
○ More historic neighborhoods (mostly built before 1970)
○ I don't have a preference

Preferred Location

● Across from New Jersey
○ Across from Brooklyn
○ I don't mind

Train Access

● Yes
○ No
○ I don't care

Select Neighborhood (Tab 2): ANNADALE

Figure A10

# Appendix B

## Code, Data and Reproducibility

To support transparency and full reproducibility, all code, data and supplementary materials associated with this study are publicly available.

The complete project repository is hosted on GitHub at:

**GitHub Repository:**
https://github.com/NikoletaEm/Capstone

This repository includes:

- The full **R Shiny application code** used to create the interactive forecasting dashboard
- The **final capstone paper (PDF)**

- The complete **reproducible analysis code** used for data processing, modeling and forecasting
- All **datasets used in the analysis**, including cleaned and source-level files required to replicate results

In addition, a fully rendered, public version of the reproducible analysis is available on RPubs:

**RPubs (Rendered Code):**
https://rpubs.com/NikoletaEm/1380820