Brooklyn Housing Price Analysis

Colleen Jung

Dec 6th, 2023

Summary

The analysis aims to employ linear regression for understanding how Brooklyn housing prices changed between Q3 and Q4 of 2020. This involves evaluating price variations using a multi-linear regression model.

1. Data Overview

This analysis utilizes a dataset on Brooklyn real estate transactions from 2016-2020, sourced from the City of New York. It focuses on single-family homes and single-unit apartments or condos. The cleaned dataset includes 14,048 observations, omitting any entries with null values or $0 transactions. The regression model incorporates the following relevant parameters.

Variable	Description
Borough	Name of the borough
Neighborhood	Name of the neighborhood
Bldclasscat	Building classification at sale
TaxClass	Tax class at the time of sale (1-4)
Block	Borough subdivision identifier
Address	Street address of the property
Zip	Postal code of the property
Landsqft	Total land area of the property
Grossqft	Total living area within the structure
Yrbuilt	Year when the structure was built
Date	Date when the sale was recorded

Table 1: Dataset Variables and Description

2. Model Development

A multi-linear regression model explains housing price variations, combining numerical (gross square footage, year) and categorical parameters (year built bucket, building class at sale, neighborhood cluster, quarter). Factor variables like ‘year built bucket’ (created from ‘year built’ in 5-year intervals from 1850 to 2020), ‘neighborhood cluster’ (derived from hierarchical clustering based on price and neighborhood), and ‘quarter’ (defined by time intervals) were introduced. Interaction terms assess how the impact of ‘building class at sale’ varies by ‘neighborhood cluster’. The model, with 36 parameters, accounts for 87.74% of the variation in housing sales and has an RSME of $253,174. However, standard errors may be underestimated, affecting hypothesis test reliability.

Formula: $\log(price) = \beta_0 + \beta_1 * \log(grosssqft) + \beta_2 * yrbuilt\_bucket + \beta_3 * factor(bldclasssale) * factor(neigh\_cluster) + \beta_4 * (year * quarter) + \varepsilon$

2-1. Model Limitations

While the model demonstrates reasonable explanatory power, it has notable limitations. It's based on ordinary least squares regression and exhibits autocorrelation, heteroskedasticity, and fails normality tests. This suggests potential biases, as the data might not adhere to normal distribution. While such violations often discourage linear regression use, if log transformations mitigate these issues, proceeding with the model could be justified.

3. Results

- Interaction Analysis

"The model's interaction analysis highlights that 'bldclasssale=A5' and 'neigh_cluster=2' have a statistically significant correlation (<0.01). This suggests a strong influence of building class A5 in neighborhoods like Windsor Terrace on prices. However, most other interactions did not show significant results or had coefficients marked as NA, indicating limited explanatory power for these combinations."

- Q3/Q4 Price Change Analysis

"Regarding the quarterly price changes, the analysis reveals contrasting trends. The actual price change from Q3 to Q4 showed a decrease of $139,987. In contrast, the model predicted an increase of $42,899 for the same period. The proportional change, calculated from the average prices, was 1.046037, indicating a marginal increase, as depicted in Figure 1."

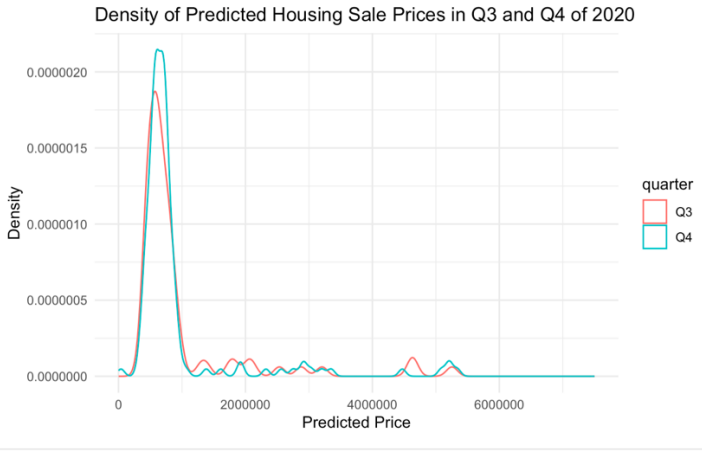

Figure 1: Density of Predicted Housing Sale price in Q3 and Q4 of 2020