MSCA 31007 IP03 Statistical Analysis

Mia Song

Brooklyn Home Purchase Price Change between Q3 2020 and Q4 2020

1. Executive Summary

In 2020, home purchase prices in Brooklyn increased between Q3 and Q4. To prove this, a linear regression model was created with 4-year datasets from 2016 to 2020, and several adjustments were made based on different methodologies. While focusing on the model's residuals and coefficients, this research confirmed the upward price trend between quarters in the Brooklyn's real estate market.

2. Background

2.1. Data Preparation

Brooklyn real estate purchase data from 2016 to 2020, provided by the City of New York, was used. To increase the model's explanatory power, the data was limited to the following conditions: 1) Purchases of single-family residences and single-family apartments or condos only, 2) A house with only one total and residential unit, 3) Gross square footage greater than zero and sales price not null. Outliers were removed in the same context with the following standards: 1) Sales price equal to zero or greater than $15M, 2) Gross square footage greater than 20K, 3) Building age older than 120 years at the point of sales.

2.2. Exploratory Data Analysis (EDA)

Several variables were newly added or transformed for the best model results. The new variables were: 1) building age: the year difference between the year of sales and the year of construction, 2) sales age: the year difference between 2020 (assumed to be the current year) and the year of the sales, 3) years, months, quarters extracted from the sales dates. The zip codes in the dataset were aggregated into 18 groups from the original 43 levels based on geographic distance. The building classes have been divided into 2 from 162 levels: single-family dwellings: A or condominiums: R.

2.3. Linear Regression Model

The model used the following variables: 1) 18 levels of zip codes, 2) two-level building classes, 3) gross and land square footage, 4) building age, and 5) sale age. The interactions between the variables (zip codes and gross square footage, land square footage and building classes, gross square footage and sale age) were made to improve the model's performance. The R-squared of the model, which indicates how much these variables can explain home purchase prices in Brooklyn, was 0.6122 with an adjusted R-squared of 0.611, showing that the selected variables were valid.

3. Methodology

3.1. Linear regression model

For the analysis, the model used data, limiting sales year as of 2020. Corresponding to the change, sales age was replaced with building age. Additionally, the square root transformation for the model's performance was restored to calculate accurate RMSE (the square root of the variance of the residuals) from the residuals. Note that despite these modifications, the new model's explanatory power was still significant compared to the original model, with R-squared of 0.667 and adjusted R-squared of 0.6586.

3.2. Residuals

Residuals represent the difference between the predicted value and the actual value. In this case, the positive residuals were generated from houses sold above the predicted prices, while the negative residuals came from house purchases below the predicted prices. This methodology examined the signs, size, dispersion, and gap of residuals to understand the house purchase price trend in each quarter.

3.3. Independent variables

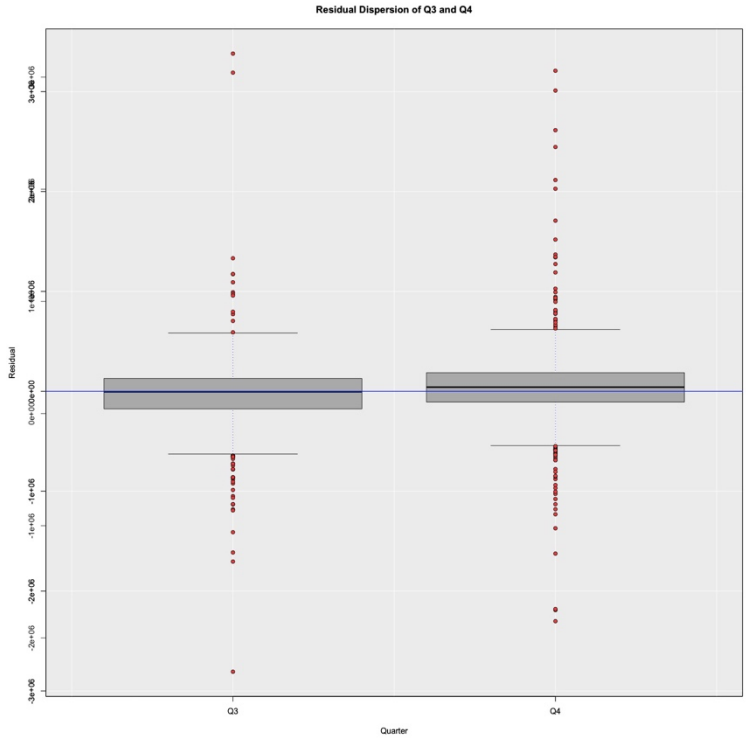
A linear regression model explains the relationship between a dependent variable (price) and independent variables (zip code, building class, gross and land square footage, building and sale age). The coefficients

multiplied by the independent variables indicate: 1) Signs of the coefficients: negative or positive relationship between a dependent variable and an independent variable. 2) Coefficient values: By how much the mean of the dependent variable changes given a unit change in the independent variable while holding the other variables in the model constant. The research studied how the coefficients of the variables reacted differently according to quarters and added quarters as new variables to spot the trend.

4. Conclusion

4.1. Residuals and Pricing Trend

From the linear regression model created, the actual prices in the market, the prices predicted by the model, and the residuals (actual price - predicted price) for Q3 and Q4 were measured. The average residual in Q4 was $54K, while it was -$44K in Q3, proving a significant difference with a p-value of 0.0047 under the t-test. Box plots were created based on the residuals of Q3 and Q4 to understand these estimates better. The median value of each quarter recorded the opposite sign: negative in Q3 and positive in Q4. It suggested that there were more positive residuals in Q4, while Q3 had more negative residuals. Additionally, the boxplot in Q4 was located higher than in Q3, meaning that the discrepancies between actual and predicted prices tended to be larger in Q4. From the dispersion of residuals in the box plot, an upward price trend in Q4 and the reverse trend in Q3 were reported.



	Q3	Q4
Minimum	$-627,936	$-543,754
Interquartile Ratio		
Lower Quartile (25%)	$-176,219	$-107,217
Median (50%)	$-5,089	$41,594
Upper Quartile (75%)	$128,659	$186,318
Maximum	$583,417	$618,568

The residuals for each quarter were divided by their signs and examined in detail. In Q3, more houses were sold at lower prices than predicted prices, and the mean difference between them was greater among the negative residuals. However, Q3 had almost the same size of positive and negative residual observations, doubting whether there was a clear downward trend. In Q4, on the other hand, 58% of the total observations had positive residuals.

The mean difference between the actual and predicted prices was also larger among the positive residuals, with a p-value of 1.636e-06 proving its significance, clearly showing the increasing price trend.

	Q3	Q4
Positive Residuals	159 observations (50%)	319 observations (58%)
Mean of Actual Prices	$1,192,387	$1,193,220
Mean of Predicted Prices	$936,641	$905,982
Difference	$255,746 (p-value 0.005054)	$287,238 (p-value 1.636e-06)
Negative Residuals	162 observations (50%)	231 observations (42%)
Mean of Actual Prices	$663,988	$818,105
Mean of Predicted Prices	$1,001,410	$1,086,180 (p-value 8.201e-05)
Difference	$-337,422 (p-value 4.933e-07)	$-268,074

4.2. Coefficients and Pricing Trend

In addition to the increasing price trend in Q4 observed from its residuals, the model's coefficients from variables were examined to confirm the trend between Q3 and Q4. Several adjustments were made to the model for better analysis. 1) To see the explanatory power of each variable individually, a full linear model was used, using all variables and removing all interactions, 2) Data from Q3 and Q4 was entered separately into the model to see how the variables of the model would react depending on the quarters, 3) Only the significant coefficients were considered with a p-value less than 0.05 for the interpretation. The most notable variable in this methodology was gross square footage. Assuming all other conditions remain unchanged, adding one unit of gross square footage would increase the house price approximately by $250 in Q3 but by $450 in Q4. In other words, a house price would increase by around $200 more for every square footage in Q4. Theoretically, the maximum increase could be up to $1.4M in Q4 compared to Q3, given that the largest gross square footage of a house in Q3 was 7,200.

	Q3	Q4
Gross Square Footage	2.534e+02 (p-value 5.94e-05)	4.496e+02 (p-value < 2e-16)

Quarter information was added to the same linear model above as a new variable. This time, data with sales dates between Q3 and Q4 was provided into the model. Q3 appeared as an intercept, and it means the house price in Q3 is predicted as $10M when the other coefficients are all zero. The coefficient of Q4 was 1.313e+05, which means the price of a house sold in Q4 instead of Q3 would increase by $1.3M. Not only is the Q4 variable significant with a p-value < 0.05, but due to the characteristic of the factor variables, the upward price trend from Q3 to Q4 was clearly proven.

	Coefficient
(Intercept) *Q3	1.041e+07 (p-value 0.017431)
Q4	1.313e+05 (p-value 0.000289)

5. Limitation

Although this study proved the increasing house price trend between Q3 and Q4 in the Brooklyn's real estate market, it contains several limitations that may require further improvement. The first confrontation is the R-squared of the original model, which was 0.6122. Using a linear model with higher R-squared would help confirm the increasing trend in Brooklyn house prices with confidence. Moreover, violation of the i.i.d assumptions of the original model could produce incorrect standard errors on the model's estimates, influencing their significance level. Getting more datasets and a proper transformation of the variables would solve this problem. Finally, there was an imbalance in the number of observations in Q3 and Q4 of 2020. Currently, Q3 contains about 300 datasets, while Q4 includes 500. Getting more datasets in Q3 would help understand the price trend in Q3 and state the price trend between Q3 and Q4 with more confidence.