Brooklyn Housing Price Prediction - Single Family Residences & Condominiums

Karan Mehta

Summary

Housing market is one of the more volatile markets. Here we are analyzing changes in the house prices in the Brooklyn region specifically for single family residences and condominiums in Q3 and Q4 of 2020. We have modeled a linear regression based on the housing data from 2016-2020 which is able to explain 63.32% variance in the sales prices while considering relevant variables which could be the reason for price differences. The model is further used to predict prices for Q3 & Q4 of 2020 and conduct a statistical test to determine changes in group means between the two. It is found that the results are unclear.

1. Data Overview

The analysis is drawn from Brooklyn housing data from the years 2016-2020. Cleaned data-set consists of a total of 13,413 purchases focused on single-family residences and single-unit apartments or condos. The most relevant parameters listed for each purchase and utilized for model development are described in Table 1.

Essential Variables for Analysis			
(Arranged as per usage and importance to the model)			
Block	GrossSqft	Neighborhood	YrBuilt
Date	Lot	Zip	LandSqft

The final dataset contains price values in the range of $110,000 to $6.85 Million. Dataset initially had $0 transactions which are considered as family inheritance and hence have been removed.

Min: 110000, 1Q: 525000, Median: 740000, Mean: 947047, 3Q: 1089000, Max: 6850000

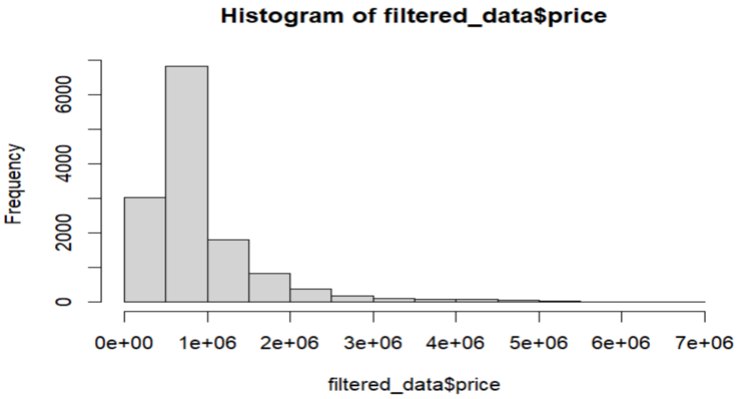

Figure 1: Sale Price Distribution Histogram from 2016-2020

2. Model Development

A multi-linear regression model is utilized to better explain the variability in housing prices of Brooklyn region. It is found that the price depends on gross square footage, neighborhood and blocks, which is a subdivision of borough. It is noted that blocks are bucketed into 18 buckets (0-9000 range), grosssqft is a numerical factor and neighborhood is categorical and is considered as a factor in the model. Modelling these factors enables us to understand how the price depends on the square footage of a housing unit and how neighborhood drives the sale pricing decisions. For example: Park Slope, Midwood, Ocean Parkway and Cobble have some of the highest priced homes in range of $6 - $6.85 Million as opposed to neighborhoods like Sheepshead Bay and East New York are on the lower end of $110,000 - $125,000. Block is used as there are multiple housing units in different blocks, the model can account for differences in average prices for these blocks. Different blocks would have different amenities and public services such as proximity to airports which influences sale price of houses within it.

$$price = \beta_0 + \beta_1\text{*block\_bucket*grosssqft} + \beta_2\text{*Neighborhood\_Category} + \varepsilon$$

Equation (1).

Equation (1) effectively explains **63.32%** of the variance in housing sales prices with a total of **39** parameters and an RMSE of **$449,790.83**, indicating that the deviations between our model and the original data are relatively minor. It was determined that each of the 39 parameters in total made a significant contribution.

3. Model Limitations

The model has sufficient predictive power to describe the variability of housing prices, but it has certain limitations as well. It violates the I.I.D assumptions of ordinary least squares regression. The residuals is autocorrelated as indicated by the **Durbin-Watson test** which equals 1.281 which shows a positive autocorrelation. Failing KS-test indicates that the model residuals are not identical (p-value = $< 2 \times 10^{-16}$). Heteroskedasticity exists in the residuals as depicted by the **Breusch-Pagan Test** (p-value = $< 2 \times 10^{-16}$). While knowing all these above results, the model should be used in a cautious manner while keeping in mind all the criteria and it is being worked on for improvement. Furthermore, this regression analysis is conducted under the assumption that all external variables like GDP of the country, interest rates and employment rate are controlled.

4. 2020 Q3-Q4 Price Trends

The linear regression model is developed to determine to what extent and whether or not there is a change in the housing prices between 2020 Q3 and 2020 Q4. With this in mind, let's look at the predicted prices. It can be seen in the Figure (2) that on average the Q4 curve shifts towards higher prices and that is supported by mean increase of $6,079 but on the contrary the median by decreases $6,655 in Q4 which suggests that Q4 housing prices are susceptible to outliers. By comparing the mean difference between Q3 & Q4 prices of 2020, it can be seen that there is an overall increase in price.

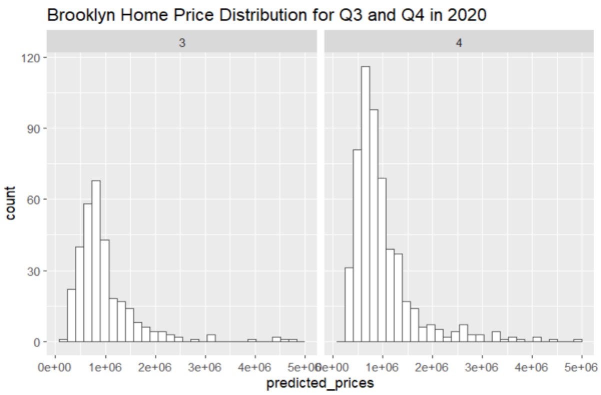

Figure 2.

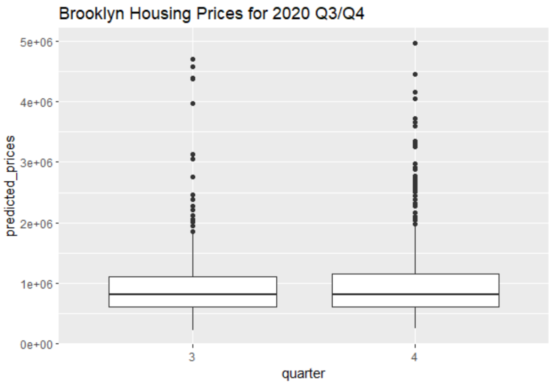

Figure 3.

Based on raw data it is observed that there is an increase in the average housing between Q3 2020 and Q4 2020 however it is worth investigating whether the change in the sale prices across the two quarters is statistically significant or not. To compare the average housing prices of two quarters, t-test is employed. After performing t-test on the sale prices for the two quarters, the calculated p-value is < 0.05 which indicates that the t-test is failed. In other words, the null hypothesis, which states that there is no actual difference between the means of the two groups, is rejected. It should also be noted, as mentioned above, the multi-linear regression model used to predict house prices doesn't satisfy all the assumptions of the OLS model. Since the I.I.D assumptions are violated, reaching a conclusion based on the result of the t-test is not ideal.

5. Conclusion

In conclusion, based on the results of different statistical methods to analyze if there is a difference in housing sale price between Q3 2020 and Q4 2020, it remains unclear whether the prices have changed, if at all.