Brooklyn Housing Prices Business Memorandum

Date: 9 Dec 2022

To: The Real Estate Business Manager

From: Prayut Jain, Lead data scientist

Subject: Understanding the changes in housing prices from Q3 to Q4 2020

The world came to a standstill when the epidemic struck in the second quarter of 2020. All of our ostensibly efficient systems failed, and we began seeking for a better, or rather, newer way of living, and the housing market was no exception. In such turbulent times, everyone desired safety as well as stability, and where we lived became more important. With work-from-home becoming the new norm in the whole sector, individuals began searching for greater spaces for the same amount of money. As families began to move their attention away from close-to-office residences that were expensive but small, the desire for big and cheap houses began to rise. There was also a large influx of immigrants or opportunists searching for high-end homes at a low cost. In this memo, I will attempt to emphasize the impacts of this behavioral shift, which has begun fluctuating the price movements while also opening up new prospects for the real estate industry.

I am structuring this paper by stating the patterns I noticed as a result of lifestyle changes, also known as behavioral effects, and how they affected house pricing from Q3 to Q4 2022, as supported by the analytical model we constructed. The main behavior I noticed was a result of people leaving pricey homes in New York for cities with lower taxes, less expensive homes, and larger areas, in this case Brooklyn. Our results from the linear regression model we trained show that, at a significance level of 99%, the quarter 4 mean prices are higher than the quarter 3 prices, and the mean prices for the year 2020 were the highest in comparison to the years before. The model shows that the estimated mean increase in housing prices is of around $2025 from Q3 to Q4, which is significant. The model also explains how the prices are on the higher end for these two quarters as compared to the last 5 years and is expected to be on a rise in the time ahead.

To estimate home costs in Brooklyn, I utilized a linear regression model. Prices were predicted using predictors such as gross area, neighborhood, building type, year of sale, and quarter of sale. I utilized the year of sale and zip codes as categorical variables, and 2020 Q3 as the reference level for the regression analysis. This provided me with the difference in group averages between 2020 Q3 and 2020 Q4. This model has an adjusted R2 of about 0.6 and a root mean squared residuals value of about $446K. I also removed a few housing prices data where the value was less than $100,000 because these could be prices that have no bearing on how housing prices are (for example, a gift from a father to a son quoted a sale price of $500 where the house is located near Brooklyn Heights and has a high gross sqft area). I also deleted the very high prices, as this skews our data and biases our model to examine the relevance of such high prices. These high-priced homes have a very low selling frequency and are not covered by this study. So I now have roughly 13.5K observations to train my model with. To determine the significance of our findings, I ran a Tukey HSD test to see if the difference in means was due to chance or if it was significant. The Tukey HSD provides a large confidence interval of [-3,98] and demonstrates that the difference in means is significant with a p-value of 0.07, indicating that we cannot reject the null hypothesis (at 90% significance level) that the means of the groups vary.

Furthermore, if we look closely, we can see that the interval is quite favorably skewed, and a positive difference has a major meaning in our instance.

The predictor variables in this model were picked with the logic that they are the most essential factors influencing the price of any property - to begin with, the neighborhood and carpet area. Also, the linear correlation between prices and gross area is 0.8, indicating that it is a solid predictor in our model. While training the model, I transformed these variables by taking a squared root- price and gross sqft - such that the linear correlation is better defined and the extreme values are scaled down. Because the housing market experienced a significant shift in dynamics following COVID, we proposed including the time of sale as another variable, which is built by maintaining sales from 2016 to 2020 Q2 as one category and then Q3 2020, Q4 2020 separately as they would have the most effects of the pandemic.

We examine the Ordinary least Square assumptions (independent and identically distributed residuals) of this model to better understand its limitations and drawbacks, as well as to see if it satisfies all of the requirements to satisfy the assumptions of linear regression. This model breaks most of the assumptions, rendering the overall significance values untrustworthy, however I would still accept the point estimates as described above with extreme caution because the model explains a considerable amount of variance. Furthermore, the ANOVA test shows that the predictor selection is sound and makes a substantial contribution to the model, with p values near to 0. The summary graphic (Figure 1) of the residual distribution is provided to the conclusion of this memo, along with square root of predicted and actual prices, and the square root of the gross area in sqft from the model and how the variables varies with each other for further reference.

In my conclusion, I would continue to believe in this model because the impacts it predicts are supported by the post-COVID shift in society dynamics. Due to a relaxation in workplace laws, many people relocated from expensive areas like Manhattan to tranquil, low-tax cities like Brooklyn. Prices will rise as a result of the increased demand, particularly in areas where low- and middle-income households are concentrated (refer to Figure 2 below). The model does an excellent job of somewhat simulating this impact.

Thank you
Prayut Jain

P.S. Find below the plot that was discussed in the memorandum.

Figure 1: Here the upper triangular matrix show the correlation between the factor on the diagonal and the lower shows the distribution of these variable with each other. The asterix on the correlation number is the significance for those values.

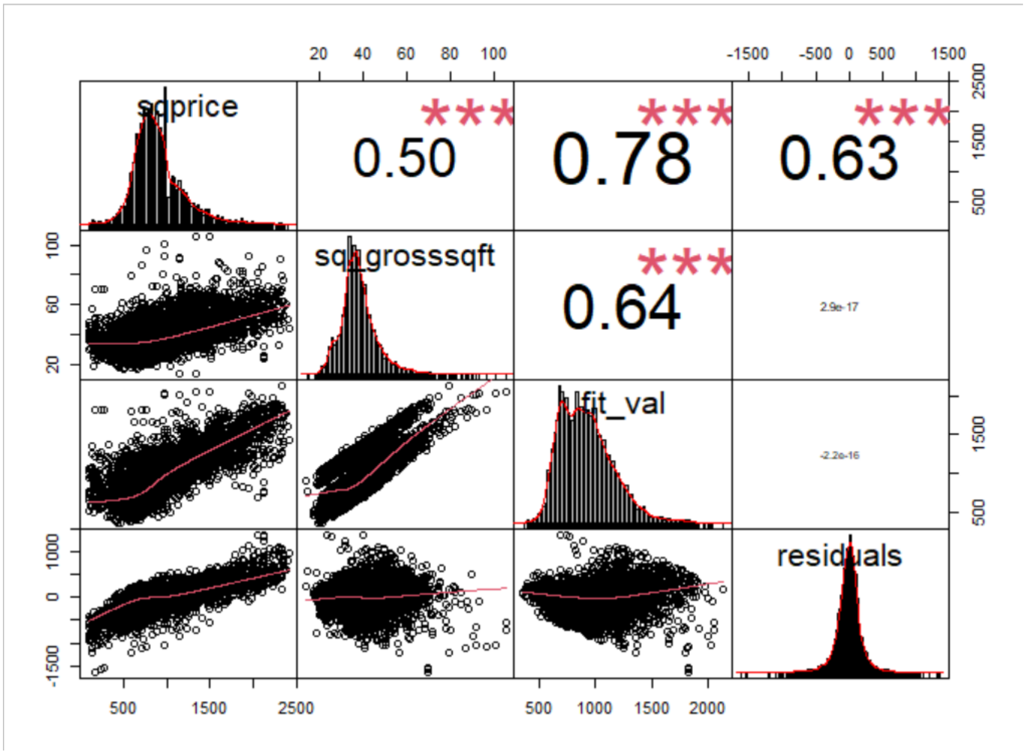
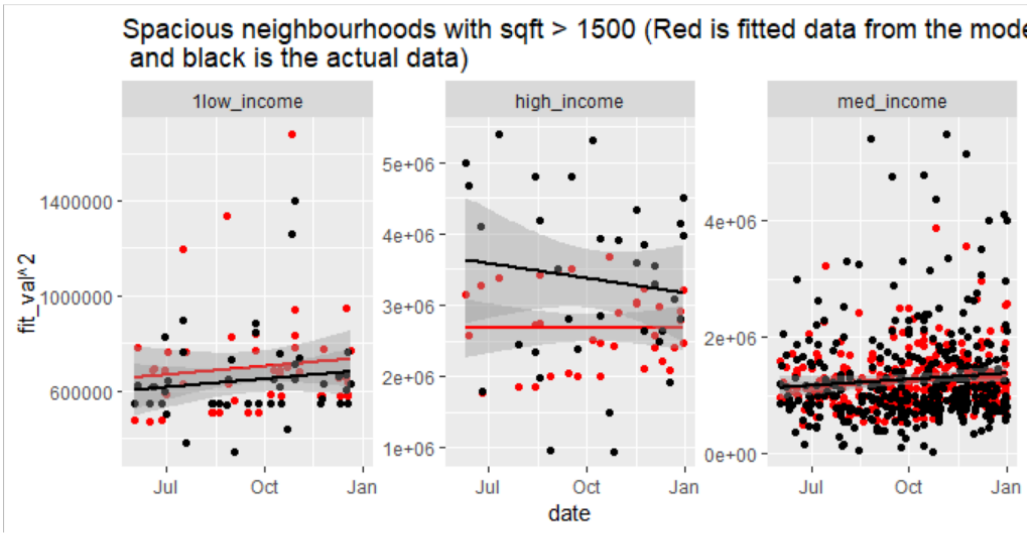


Figure 2: This is the plot of the actual vs fitted prices for the households in high, low and medium income household areas for Q3 and Q4 2020. Observe the difference in y-scale while reading the plots. The division on the household income level is using the data from government and postal office websites.


Spacious neighbourhoods with sqft > 1500 (Red is fitted data from the mode
and black is the actual data)
1low_income
high_income
med_income
fit_val^2
date
5e+06
4e+06
3e+06
2e+06
1e+06
4e+06
2e+06
0e+00
Jul
Oct
Jan