Analyzing the change in Brooklyn House Prices between Q3 2020 and Q4 2020

Methodology and Model Choices

Under the current study, housing prices in Brooklyn were modelled between the time periods of 2016-01-01 and 2020-12-31. The following variables from our model data set were used to maximize model predictability:

- Brooklyn geographical subdivisions
  (Categorized e.g., Northern, Central, Eastern localities etc.)
- Building Class Category (Categorized)
- Land sqft (Continuous transformed)
- Gross sqft (Continuous transformed)
- Quarter of Sale (Categorized, Q1 2016 to Q4 2020)
- Neighborhood Affluence Level
  (Categorized e.g., Expensive neighborhoods etc.)

	lmg
new_neigh_level	0.254535733
bldclasscat	0.024796734
locality	0.154925444
quarter	0.015938980
I(log(1 + landsqft))	0.006560037
I(log(grosssqft))	0.215057577

Fig 1: Relative importance of model variables for predicting sqrt(housing prices) in Brooklyn

A rational choice was made to predict the square root of price, rather than price. After eliminating a few outliers, sqrt transformation of price showed a relatively better normal distribution spread than price itself.

As visible in Fig 1, the quarter variable only adds around 1.6% extra predictability power to the model. However, upon analyzing the coefficients for this variable, we can witness some interesting trends:

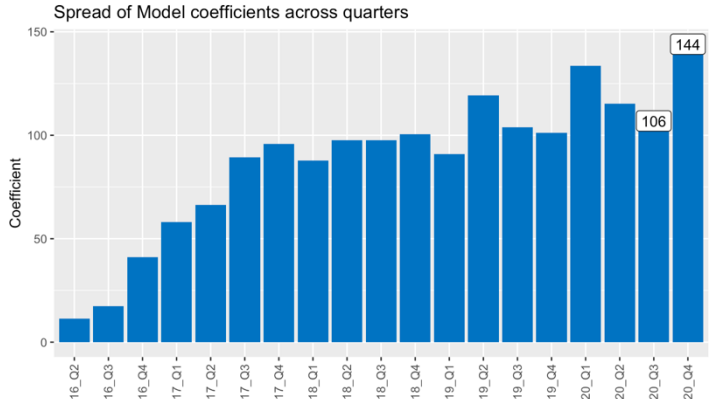

Fig 2: Spread of model coefficients across quarters

Here, each coefficient (Betas) represents the shift of housing prices for each subsequent quarter from the reference level. The reference level in this chart will be the first quarter of our time frame (Q1 2016).

Starting with 2016, we can see a gradual increase in model coefficients, comprised with a few seasonal trends as well.

Between Q3 2020 and Q4 2020, we see the sharpest spike for the coefficients observed across all quarters - **from 106 to 144.**

From this we can infer that if all other model variables were controlled, we can see a jump in sqrt(price) (and subsequently housing price) just between these two quarters.

Simulation

To analyze this difference, a simulation was conducted.
Houses sold in Q3 2020 were identified and extracted (n=318). A dummy data set was created where all variables were kept constant, but just the quarter variable was manually changed to Q4 2020.

Erstwhile presented model was now run on this 'new dummy' data set. The aim was to analyze the price difference if just the date of house sold was changed from 2020-09-31 (Q3) to 2020-10-01 (Q4).

The results are presented below:

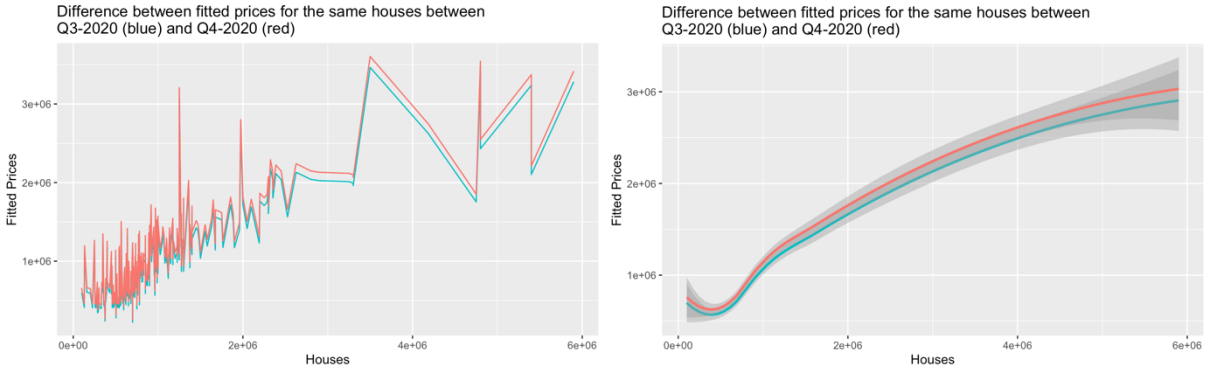

Fig 3: Increase in fitted price by just changing the quarter from Q3 2020 to Q4 2020 (Keeping all other variables constant)

For all houses sold in Q3 2020 (plotted in Fig 3 by an increasing value of their actual house prices), we can see that just by changing the quarter variable, a bump is observed in fitted prices. This is in line with the model coefficients spike we saw previously.

Quantifiable Results

To further investigate this, the distribution of price deltas between Q3 2020 and Q4 2020 for identical houses was analyzed.

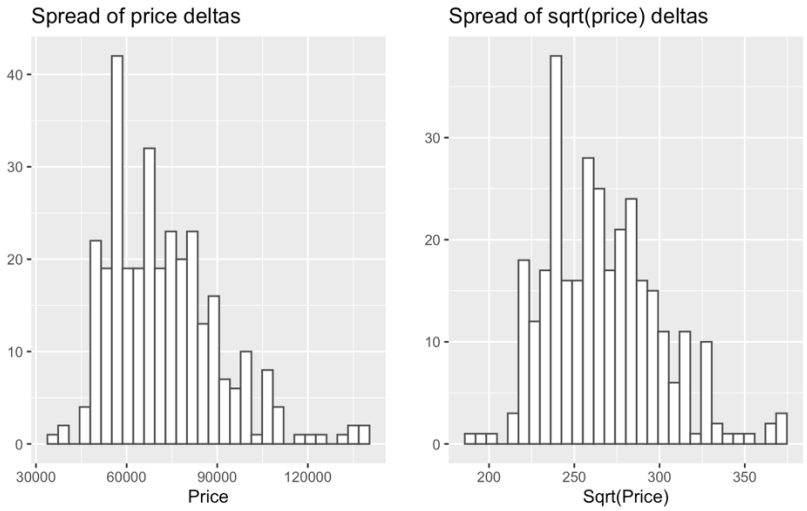

Fig 4: Distribution of price deltas and its square root transformation

Calculating summary statistics for the price deltas:

Statistic	**Price**
Minimum	36,447
1st Quantile	58,355
Median	69,243
Mean	72,362
3rd Quantile	82,278
Maximum	139,528
Std Dev	18230.2
N size	318

To estimate the range of price delta from the above two distributions (Fig 4), confidence intervals at a 95% significance level can be employed (approximating both distributions to near normal spread

95% CI for Mean Price Delta

Mean: 72,363$
Lower Limit: 70,358$
Upper Limit: 74,265$

Conclusion

With this method, we are producing an **interval of ~70,000$ and ~74,000$** that can contain the true mean of price increase between Q3 2020 and Q4 2020 95% of times.

Model Limitations and Caveats

- This problem statement was kept in mind while trying to predict housing prices in Brooklyn. Even though quarters as separate variables were not yielding very high predictive power, the aim was to analyze how these price differentials moved across a time dimension.
- Several transformations for both predictor and response variables were iterated upon to yield the best model fit.
- Different categorizations were also implemented to prevent any redundancies in the predictors
    - Neighborhoods were grouped as per their average sales prices. While this can be a good proxy for gauging affluence levels of a neighborhood, a better method would have been to get average incomes in those neighborhoods
- After implementing all the steps, our model was able to explain **~67% of the data variance**. While statistically this is a good model, there is still more than **30% of variance that is not being accounted by the current model**
- Outlier analysis was conducted to remove extraordinarily low (e.g., 0$ inheritance sales) and high (25,000,000$) value purchases to eliminate any noise. While ideally these should have been included from a modelling perspective, these observations were not fit for a linear model
- The prediction model has not been extrapolated on a testing data set. Entire model sample was utilized to build a linear regression fit
    - This can raise some concerns into the real-world implementation for the model
    - However, as we are not forecasting future price points, the model is suitable for our current investigation
- Simulation - For the analysis perspective, 318 houses sold in Q3 2020 were identified. The rationale was to provide the most intuitive method for gauging if a Q3 2020 sold house would have been valued more if it was sold in Q4 2020.
- Inflation - In most economies, housing prices are highly correlated with inflation rates across time periods. National Interest and Inflation rates can be utilized in future iterations of the model
- Covid data – The time frame for this analysis includes market disruption caused by the Covid-19 pandemic. For any causal analyses, it will be interesting to factor in covid migration and economic trends as well.

Further validation

To further strengthen our conclusion that there is a definite price increase between Q3 2020 and Q4 2020, Welch's T-test was conducted to check if the two-sample means are statistically significant.

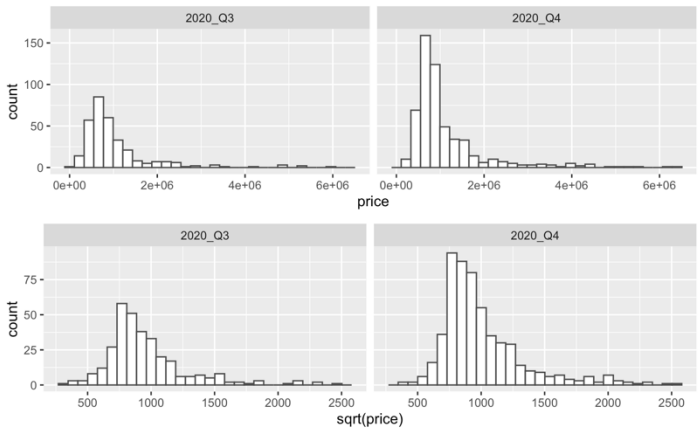

Fig 5: Actual house prices and their square root transformation across quarters

When comparing prices directly, we were getting a p-value>0.01. This made us unable to reject the null hypothesis that there is a statistically significant difference between prices along the two quarters

However, when we compare the sqrt transformation of prices across two quarters, we do see a statistically significant difference (p-value = 0.044). This became an essential insight, as we are using our model to predict sqrt transformation of prices. In later steps we are squaring them to a price metric for gauging deltas.

The statistically significant difference between the actual prices further strengthens the average difference of **~72,000$** our model is predicting between Q3 2020 and Q4 2020.