Brooklyn House Market Analysis

Statistical Analysis Assignment Part 2

1. Executive Summary

We built a linear regression model based on Brooklyn housing price data from Jan 2016 to December 2020. Further, we used the model to predict the price for 2020 Q3 Q4 using data from 2020 Q3. The dataset for Q4 was obtained by creating synthetic data using Q3 and replacing the quarters with Q4 keeping the remaining parameters constant. Additionally, we conducted a paired t test to understand change in house price between both the groups. We concluded that there is a substantial increase in price from 2020 Q3 to 2020 Q4.

2. Model Overview

As mentioned above, the data used for our analysis was based on Brooklyn housing price from 2016 to 2020. The final curated data after preprocessing and cleaning consist of 13,052 only including residential and total units of 1 and single-unit apartments and condos as well as building class at the time of sales falling under “A” and “R” category.

We decide to perform a sqrt transformation on price as the distribution of price was positively skewed (right skewed). The transformation was introduced after treating the outliers. Removal of outliers included inherited properties ($0), high values properties (>$7 M) and low-priced properties (<$200k). We have also noticed a significant positive correlation between gross square feet and price which will be useful for our mode.

3. Model Development

The final linear regression model was obtained by using square root transformation of price as the dependent variable and combination of numerical and feature engineered categorical variables as our dependent variables.

The following parameters were used to maximize the predictive power of our model:

Variables	Content of the Variables
Neighborhood Categories	North, South, East & West
Zip Categories	North, South, North-West, South-West, Central
Building Class Categories	A1, A2, A3, RR etc.
Land Square feet	Transformed – (Square Root)
Gross Square feet	Transformed – (Square Root)
Blocks& Lot Categories	10 block & lots (Deciles)
Quarter & Year	Extracted from Date

This model successfully explains 63.25% of the differences, in sale prices, which's quite impressive. It has a Root Mean Square Error (RMSE) of $438,476, which's within a range (<= $450,000). This suggests that the predicted sale prices are close to the actual prices. The model considers 40 features to make its predictions with a total of 40 degrees of freedom.

3.1 Model Limitation

i. Normality of Price: Also mentioned earlier, house prices are positively skewed meaning that the prices are not normally distributed and fails to pass the normality check. To avoid that, we performed a sqrt transformation on price.

ii. Heteroscedasticity: Since our model passes Breusch-Pagen test in Fig 3.1, it provides that heteroscedasticity is present in our model. From the residual plot, its evidence that the residuals follow a certain pattern which should not ideally be present.

iii. Autocorrelation: We perform Durbin Watson test on our model and noticed that autocorrelation is present in our model indicating that residuals are not independent from each other.

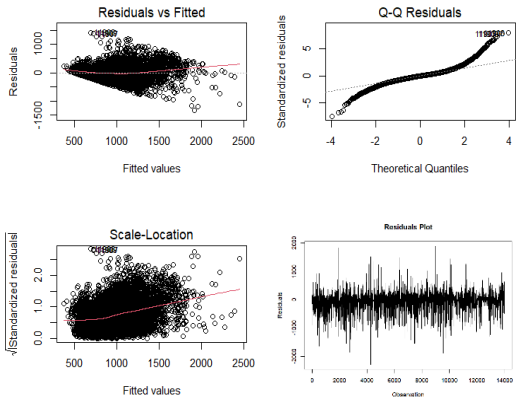

Fig3.1: Heteroscedasticity, Autocorrelation & IIQD Normal Check

4. Price Change Estimation (Q3 & Q4 Trend)

4.1 Synthetic Data

To understand the difference in price between 2 quarters, we created a synthetic dataset for q4 by using q3 data and changing the quarters to q4 keeping remaining variables constant.

The main motive behind this approach was to analyze the price difference if the dates of houses sold was changed from Q3 to Q4.

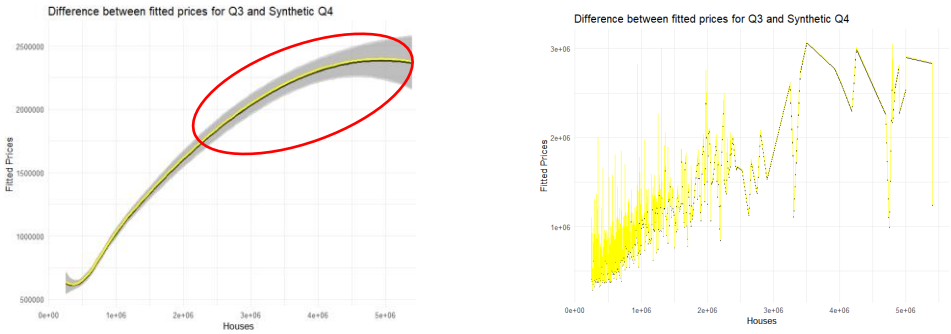

Fig 4.1: Incremental increase in price with the change in quarter (Q3 to Q4) keeping other parameters constant (Black – Q3 | Yellow – Q4)

We can see that for all the houses sold in 2020 Q3, there is a change in the fitted price by changing the quarter variable.

We further delve into the distribution of normal and transformed price change between 2020 Q3 and 2020 Q4 for the identical houses.

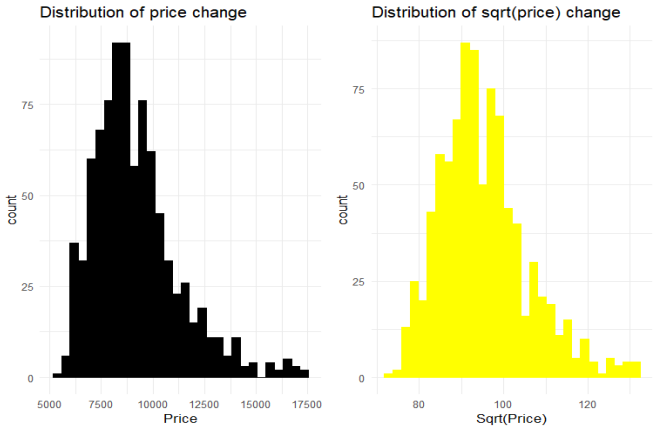

Fig 4.2: Distribution of Price Change (Black) and Square Root of Price Change (Yellow)

Summary for Price Change	
Statistics	**Price**
Minimum	$25,297
1st Quartile	$42,734
Median	$38,758
Mean	$59,161
3rd Quartile	$76,113
Maximum	$94,259

Table 4.1: Summary of Price Change Statistics

4.2 Welch's T-test

To further strengthen our findings that there is a price increase from 2020 Q3 to 2020 Q4, we also performed paired t test to check if the 2-sample means are statistically significant.

When considering price, we have obtained a p-value of 0.102 (>0.05). This indicates that we can't reject the null hypothesis that there is a significant price difference between 2020 Q3 and 2020 Q4. On the other hand, when we take the square root transformation of price, we acquire an p-value of 0.008 (<0.05) which shows that we reject the null hypothesis proving our findings that there is a significant increase in price.

5. Conclusion

In summary, when we looked at the prices of homes bought in Brooklyn during Q3 and Q4 of 2020, our analysis using linear regression posit that there has witnessed an increase in price by approximately ~$59,000 (Pandemic to post pandemic). To get a more detailed understanding, it would be helpful to explore other economic factors and trends in the housing market. Despite these factors to think about, our model gives us useful information about the changes we noticed in the pricing patterns.