ADSP 31007 IP06 Statistical Analysis
Final Assignment Part 2
Prepared By: Kshitiz Kumar Sahay

Brooklyn Home Purchase Price Analysis Report

Executive Summary

This report presents a comprehensive analysis of the change in Brooklyn home purchase prices throughout the period from Q3 to Q4 2020 using linear regression. The analytic dataset contains the real estate sale price and property characteristics within the Brooklyn borough, spanning the period from 2016 to 2020. The linear regression model fitted to the dataset explains approx. 62% of the variation in home prices, with a prediction error of $435,077. The analysis indicates a statistically significant increase of 7% in home purchase prices between Q3 2020 and Q4 2020.

Data Overview

The City of New York supplied the raw datasets of home purchase prices in the Brooklyn borough from 2016 to 2020. The data was provided as flat files, each representing a specific year, that were standardized and cleaned to conduct the regression analysis. Several preprocessing steps were implemented to clean the data efficiently, including identifying inconsistencies in data fields and writing data cleaning procedures to fix these errors. Key cleaning procedures were trimming leading and trailing whitespaces, removing invalid characters, and transforming the data types. The cleaned dataset contained approximately 119,000 rows. Only single-family dwellings and single-unit apartments or condominiums were included in the analysis. The sale price encompassed transactions with minimal monetary value, such as $0, which signify transfers of ownership within families, as well as transactions involving exorbitant prices associated with mansions. These outliers were removed to normalize the distribution of the home purchase price. The final dataset had 13,055 rows, which were utilized to train the linear regression model.

Table 1: Useful Property Characteristics

Neighborhood	Block	Zip Code	Gross Square Footage	Land Square Footage
Year Built	Tax Class	Building Class	Sale Date	# Units

Table 2: Data Waterfall

Filters	**Record Count**
Combined dataset	**119,351**
Building Class (At Sale) starts with A or R	44,439
Total Units = 1 & **Residential Units** = 1	37,727
Gross Square Footage > 0	20,821
Price (non-missing)	**19,640**
Price (without outliers)	**13,055**

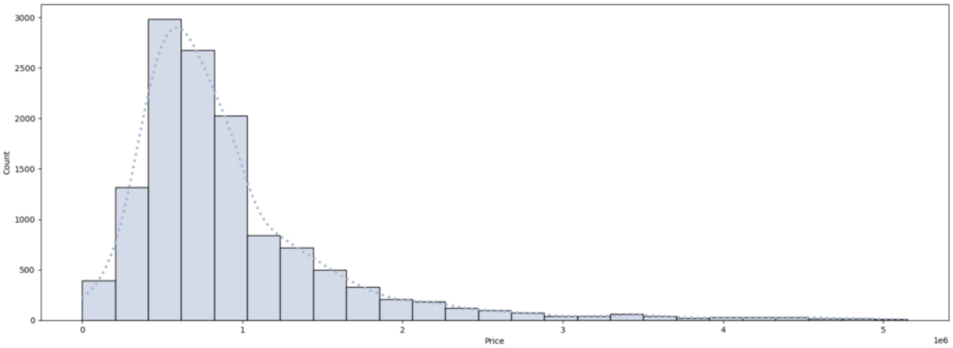

Figure 1: Distribution of Price (Minimum: $10, Mean: $902,802, Median: $730,000, Maximum: $5,150,000)

To explain the variation in home purchase prices, several new attributes were derived from the original features. These derived attributes include ratio of gross square footage and land square footage, current age of the building, building age at the time of sale, sale year, sale quarter, and new categories of geographical attributes based on their proximity.

Model Overview

The linear regression model used to explain the variation of home purchase prices and make predictions consists of multiple raw and derived attributes that were identified using domain knowledge, correlation analysis, and their explanatory power. The analysis reveals that the ratio of the gross square footage of the building to the land square footage, which accounts for the efficiency of space usage, is a useful predictor of housing price. A higher ratio indicates more building space relative to land, which may be more valuable in urban settings. Housing prices tend to increase over time due to inflation and changes in the market; hence, the sale year would help capture the trend. In the same vein, sales quarters would help capture the seasonality in housing markets. Neighborhoods could help explain the variation in prices because different neighborhoods can have vastly different average housing prices due to factors like desirability, crime rates, and economic activity. To manage the degrees of freedom or the independent pieces of information the model needs to explain price, neighborhoods are recategorized in such a way that those in close proximity are treated as a single neighborhood. Finally, an interaction term between the logarithm of gross square footage and the region (Northern Brooklyn, Central Brooklyn, etc.) indicates that the impact of the size of a house on its price may vary across neighborhoods. For instance, an increase in one unit of square foot of space might add greater value in an affluent region compared to a less expensive one.

$$\text{price} \sim \beta_0 + \beta_1 * \text{sqrt(grosssqft/landsqft)} + \beta_2 * \text{year} + \beta_3 * \text{quarter} + \beta_4 * \text{neighborhood} + \beta_5 * \text{(log(grosssqft): region)} + \epsilon \qquad (1)$$

Table 3: Model Parameters (Excluded Some Variables)

Variable	**Coefficients (β)**	**P-Value**
Intercept	-84896157	< 2e-16 ***
sqrt(grosssqft/landsqft)	-451925	< 2e-16 ***
year	39434	< 2e-16 ***
quarter	12077	0.00122 **
…	…	…

The regression analysis yielded an RMSE, or **Root Mean Square**, of **$435,077**. It signifies that the model's average prediction deviates by around $430K from the actual price. This could be a large error if the average housing price is generally lower. The model inputs accounts for approximately 62% of the variability in the housing price, as indicated by the **R-Squared value of 0.6199** and Adjusted **R-Squared value of 0.6185**. The model exhibits good fit but there is still a substantial 38% of variability that is unexplained. With **40 degrees of freedom**, the **F-Statistic of 451**, with a highly significant **p-value of less than 2.2 x 10^{-16}**, strongly suggests that the model is statistically significant at < 1% significance level.

All in all, the model shows a strong and significant relationship between the predictors and the housing price, but with a considerable amount of unexplained variability. It could be due to factors not included in the model or to intrinsic variability in the housing market.

Model Limitations

Linear regression is a powerful tool for inference, but it relies on several crucial assumptions. The aforementioned linear regression analysis sought to predict housing prices based on multiple parameters. Upon assessment, it was discovered that the model does not fully satisfy all the assumptions, which may affect the accuracy and reliability of model predictions. Firstly, the Durbin-Watson test suggested that the model residuals are autocorrelated. Furthermore, the Breusch-Pagan test indicated the presence of heteroskedasticity. The Q-

Q plot revealed that the model residuals are not normally distributed. Ultimately, there's a significant level of multicollinearity in the model, as seen by the high VIF of the interaction term between the logarithm of gross square footage and region. Therefore, it is essential to exercise caution when interpreting the results of the model. However, despite the noted limitations, the model was used to predict housing prices, considering its strength and practical utility.

Change in Home Purchase Price in Q3/Q4 2020

This report explores several ways to measure the direction of the change in home purchase price in the Brooklyn borough from Q3 2020 to Q4 2020. First and foremost, comparing the actual home purchase prices in Q3 2020 and Q4 2020 reveals that the average home price increased from $936,449 in Q3 to $104,2046 in Q4, which is approximately an 11% increase. Next, it is observed that the estimated average home purchase price using linear regression increased by approximately 5%, from $986,804 in Q3 to $1,033,224 in Q4.

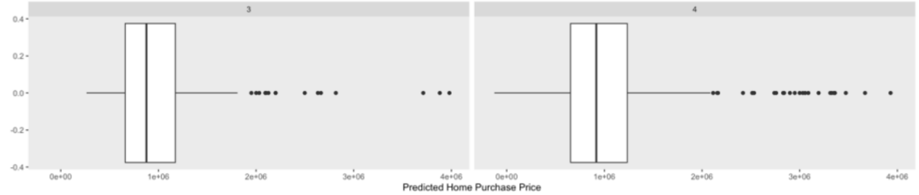

Figure 2: Box Plot of Predicted Home Purchase Prices of Q3/Q4 2020

While a comparison of actual and estimated average home purchase prices provides evidence of an increase in price, a two-sample t-test could help determine if the change is statistically significant or occurring by random chance. Performing the **Welch Two Sample t-test** without any distributional transformation results in a p-value of 0.04, thus rejecting the null hypothesis that "there is no significant difference between the means of two groups." Alternatively, due to non-normality concerns, an additional test was performed to ensure the results were not biased due to assumption violations: **Wilcoxon signed-ranked test** (non-parametric test). The Wilcoxon signed-rank test (p-value = 0.006), which compares the median (or ranks) of two groups, suggests that there is a statistically significant difference (< 1% level) between the two groups when comparing their medians. This discrepancy between these results can occur for several reasons, such as non-normally distributed data, where a t-test might not be the most appropriate test, and the impact of outliers. The t-test is sensitive to outliers (Figure 2 shows larger proportion of outliers in Q4 2020 estimations), which can unduly influence the mean. The Wilcoxon test, being a rank-based test, is more robust to outliers.

Finally, to add another layer of evidence, the analysis of the change in home purchase price also employed linear regression to analyze the differences in sale price between Q3 2020 (fixed as the reference model) and Q4 2020. The model (2) was specifically designed to test if there were any significant variations in prices during these two quarters while controlling for gross square footage and zip code.

$$\text{price} \sim \beta_0 + \beta_1 * \text{quarter} + \beta_2 * \text{grosssqft} + \beta_3 * \text{zip} \tag{2}$$

The model (2) revealed a strong statistically significant difference (1% level) in the average housing prices between Q3 and Q4. Specifically, the average housing sale prices in Q4 2020 increased by approximately 7%. This figure provides a clear quantitative measure of the change in market dynamics between these quarters.

In conclusion, the regression analysis confirms a significant increase in housing prices from Q3 2020 to Q4 2020. However, understanding the factors driving the market trend requires further investigation.