Introduction

The purpose of this memo is to discuss how Brooklyn home purchase prices changed between Q3 – 2020 and Q4 – 2020 by applying various methodologies and statistical tests on Brooklyn real estate sales data from January 2016 to December 2020.

Linear Regression Model

A linear regression was built to predict price as a function of *Neighborhood, Gross Square Feet, Land Square Feet, Block, Quarter-Year of Estate Sale Date, Zip Code* and the interaction between *Neighborhood* and *Gross Square Feet*. Through feature engineering and data transformation, this model was able to control for its complexity (having 39 model degrees of freedom), while maintaining a good explanatory power (having an adjusted R-squared of 61.24%), and good predictive power (having an RMSE of $372,338)

Methodologies to Compare Housing Prices in Q3 and Q4 of 2020

Apply Linear Regression on Full Actual Data

In order to study the market prices as a function of sales date (represented by the variable *Quarter-Year of Estate Sale Date*, or "*quarter*"*)*, as opposed to other fundamental qualities of a house, we will be looking at the fitted coefficient of *quarter,* which represents the isolated impact of a quarter on price. The coefficient of the *Quarter-Year of Estate Sale Date* variable in the model has a p-value of 0***($p < .05$), thus it has significant explanatory power in the model predicting price. However, while the coefficient of Q4 – 2020 is higher than that of Q3 – 2020 ($225,000 vs $145,900), the coefficients of Q3 – 2020 and Q4 – 2020 are not statistically significant from one another. We ran a Tukey's Honest Significant Differences test to test for differences between Q3 and Q4 of 2020 while controlling for other variables in our model. The result shows no significant difference between housing prices in Q3 and Q4 of 2020, with a p-value of 0.6719 ($p > .05$). Thus, we conclude that based on the actual data, there is no significant difference between housing prices in Q3 and Q4 of 2020.

Note that the actual mean of housing prices in Q3 is $882,085 and the actual mean of housing prices in Q4 is $877,154. A Student's t-test is also run to test the difference between actual housing prices in Q3 and Q4 of 2020, and it also concludes that these two means are not statistically different (p-value of 0.50).

Compare Q3 and Q4 Predicted Prices

We will train a model with the same aforementioned predictors on data up to but not including July 2020, then use the model to predict house prices in Q3 and Q4 of 2020. We find that the predicted data on Q3 and Q4 of 2020 is, on average, $3,793 off than the actual sales data in Q3 and Q4. However, running a Student's t-test confirms that the mean of the predicted data and the actual data in Q3 and Q4 of 2020 are not statistically different than that of the previous periods, with a p-value of 0.8893 ($p > .05$). Thus, we conclude that there is no indicator (in our current dataset) that asserts that housing prices in Q3 and Q4 of 2020 significantly derail from what has been seen in past data, and that we can confidently use this model to predict housing prices in Q3 and Q4 of 2020.

We then use the model's predictions of Q3 and Q4 to compare housing prices of these two quarters. A Student's t-test shows that there is no statistically significant difference between the mean of predicted housing prices in Q3 and the mean of predicted housing prices in Q4, with a p-value of 0.6639 ($p > .05$). The mean of predicted housing prices in Q3 was calculated to be $880,756, and the mean of predicted

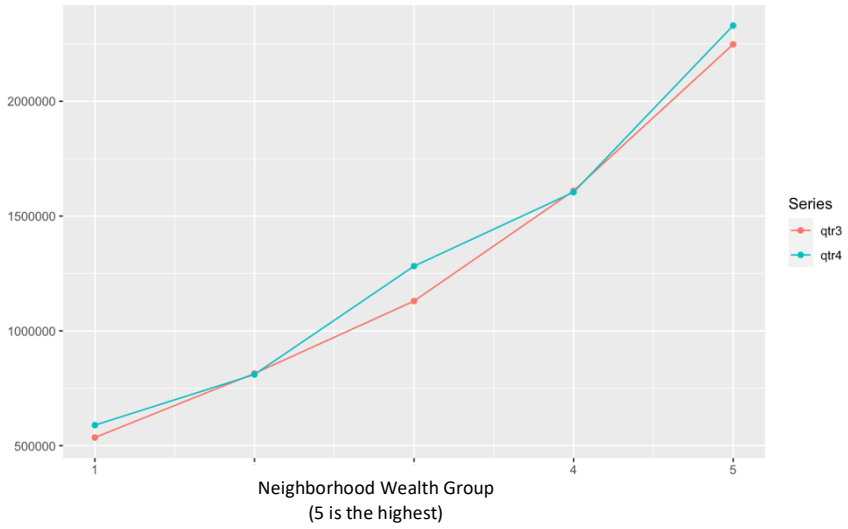
housing prices in Q4 is calculated to be $878,256. The 95% confidence interval is calculated to be -8,777 and 13,779, thus making Q3's and Q4's means well inside each other's confidence interval.

It is worth noting that there is a divergence in the signs of the difference between the predicted values in Q3 and the actual values in Q3 (a negative difference), and the predicted values in Q4 and the actual values in Q4 (a positive difference). This means housing prices in Q3 were higher than what the model predicts, and housing prices in Q4 were lower than what they should have been. Even though prices still increase with time from Q3 to Q4, the increase was not as strong as what the model predicted.

Compare Q3 and Q4 Predicted Prices by Neighborhood Group

Even though it has been established that housing prices between Q3 and Q4 of 2020 do not statistically differ, we would like to take a closer look at different neighborhoods in Brooklyn to determine if there is any neighborhood that has different trends than the overall trend.

Figure 1. Mean prices by Neighborhood Group per Quarter

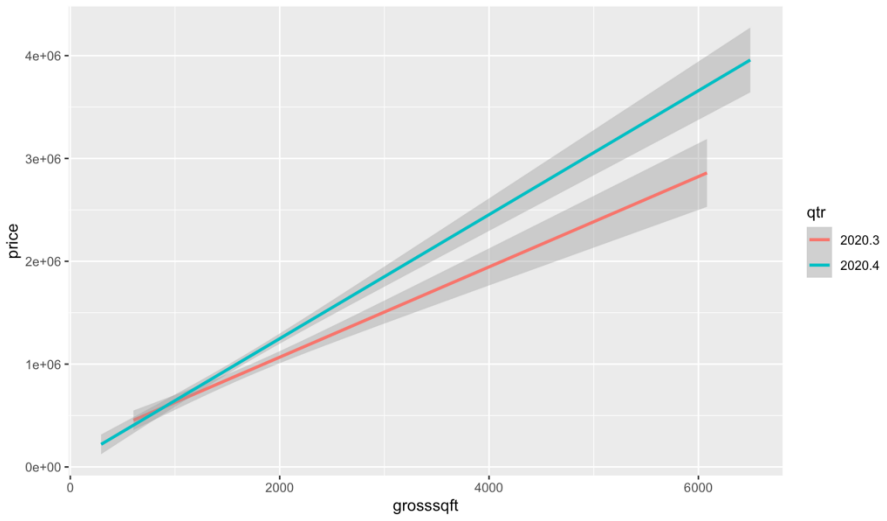


It is apparent in *Figure 1* that Q3 and Q4 of 2020 have similar pricing trends across five *neighborhood groups* (grouped by neighborhood's coefficient strength in predicting house prices, with 1 being neighborhoods with the cheapest houses, and 5 being the neighborhoods with the most expensive houses). An ANOVA test also confirms that there is no effect created by the *quarter* variable on the prices in any of the neighborhood, with a p-value of 0.06 ($p > .05$).

Compare Q3 and Q4 Predicted Prices by House Sizes (Gross Square Feet)

Similarly, we want to look at the effect of housing prices in Q3 and Q4 of 2020 on different house sizes. However, unlike the neighborhood groups, there is a statistically significant effect of the *quarter* variable on the prices with different house sizes, represented by the variable *gross square feet*.

Figure 2. Prices in by Gross Square Feet per Quarter



The different slopes of the regression of *price* by *gross square feet* and *quarter* shown in *Figure 2* indicate that there is an interaction effect between *quarter* and *gross square feet*. This effect is confirmed by an ANOVA test, with the *quarter* variable having a p-value of 0.0005*** ($p < .05$) and the interaction between *quarter* and *gross square feet* having a p-value of 0.001** ($p < .05$). The bigger the house, the bigger the increase that sales quarter has on sales price. It is also worth noticing the intersection of the two regression lines, indicating that for houses that are below 900sqft, the increase in sales date (the movement from Q3 to Q4 of 2020), controlling for house square footage, will make prices decrease.

Model Caveats and Limitations

In order to examine whether our model show violations of the OLS model assumptions, we test the model for its i) Normality, ii) Serial Correlation, and iii) Heteroskedasticity.

Unfortunately, our model does not pass any of the IID tests (Durbin-Watson test, Breusch-Pagan test, and Kolmogorov-Smirnov test), having all p-values of zero ($ps < .05$). It is realistically challenging to build a predictive model that has good explanatory power and satisfies IID tests. Therefore, failing to pass those tests does not mean that our model is ineffectual, but rather that we need to be conscious when making conclusions about our p-values and confidence intervals.

It is also worth noting that the data used to train and test the model on was pre-processed and eliminated of outliers. Thus, our model could lack the ability to predict housing prices of extreme cases, houses that have prices too low or too high. We made the decision to eliminate outliers (the 1st and 99th percentile of prices – excluding houses with prices of $0) so that we could model for true relationships between market prices and other factors, whether that be the model's fundamental qualities, or the housing market trends. In addition, the scope of this model is only to build and interpret the model that predict prices of single-family residences and single-unit apartments or condos. The relationship between housing prices and other factors could be drastically different when modelling for other types of estate, such as multiple family dwellings, or store buildings, or factories.

Houses with Highest Residuals

When looking at the top 10% houses that have the highest residuals for predicted prices in Q3 and Q4 of 2020, 43% of the houses are in neighborhood group 3 (neighborhoods with medium house prices), and only 8% are of neighborhood groups 1 and 2 (neighborhoods with low house prices). We could see that our model does not perform well in predicting housing prices for the middle tier houses, in comparison to the low tier or the high tier. This could be a potential problem when using this model to predict houses in such neighborhoods.