ANALYSING PRICE CHANGES IN BROOKLYN FOR HOUSE SALES DURING QUARTER 3 and 4 of 2020

We are analyzing how the Brooklyn housing prices changed between Q3 and Q4 of 2020 for 1-unit residential purchases such as single-family residences and single-unit apartments or condos. There are a number of ways to identify this change, but one relatively robust way is to use a linear regression model along with graphical data. There are number of factors which account for setting price of a property on a given day such as Land.ft^2, Gross.ft^2, Neighborhood, Postal code, Tax class category at time of sale, whether it is a mansion or not if we know the age of house and most important for the topic today which quarter it was sold.

We ran through the data to identify which numeric factors are highly correlated to the price of the home. Clearly Gross.ft^2 is trending the prices of the property in Brooklyn, NY followed by Land.ft^2 But now let's see how the prices differ for Q3 and Q4 of 2020. Difference in the average of two is $95,900.89 **(9.5% increase)** which apparently seems a lot if we are just comparing these two quarters only in figure 1. But now lets a have a broader look over the trend of prices in Q3 and Q4 for years 2016-2020, its clearly from figure 2 that there is no obvious similar trend for Q3 and Q4 for any given year.

Figure 1

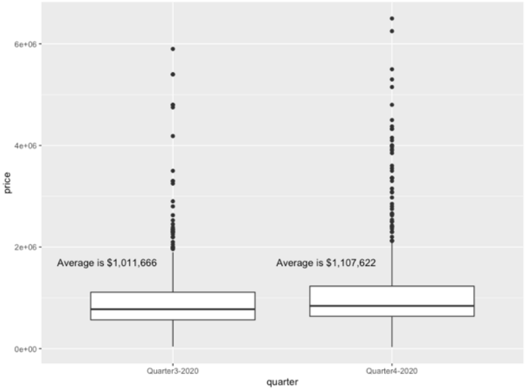
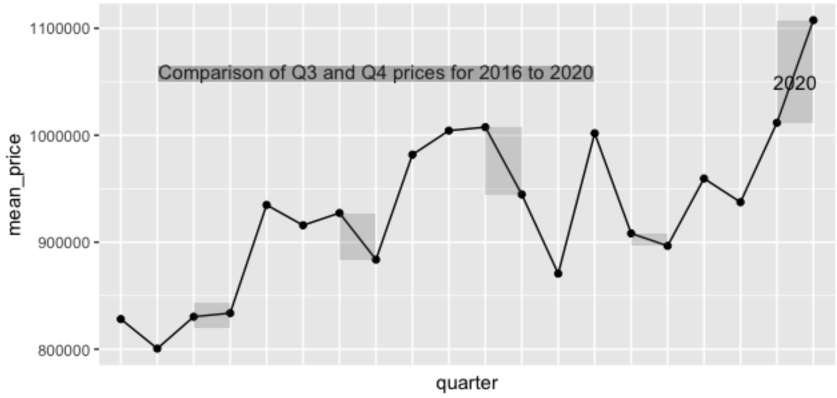


Figure 2

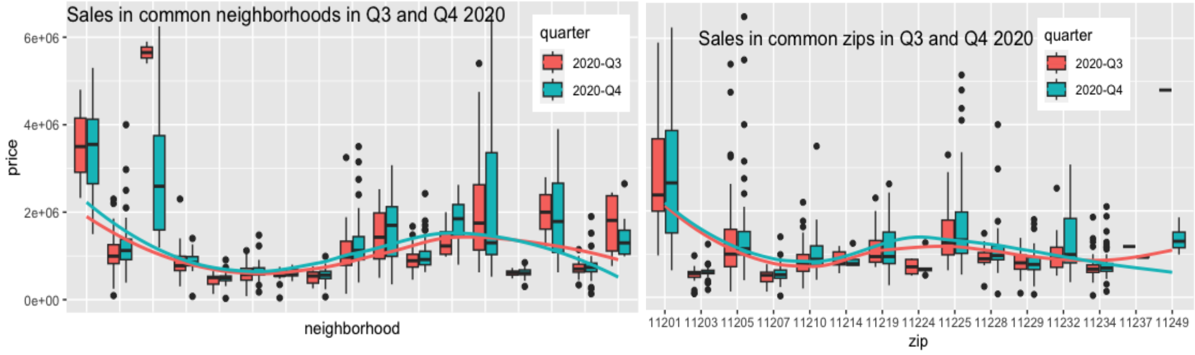


Let's analyze what factors were responsible for difference in prices between Q3 and Q4 of 2020 – Overall there were 321 houses sold in Q3 and 554 in Q4 in year 2022.

Here we are trying to see how the price changed for the common neighborhoods and zips sales

	Zips	Neighborhoods	Total Sale deeds	Mansions Sold
Q3 - 2020	16	18	321	0
Q4 - 2020	15	18	554	0

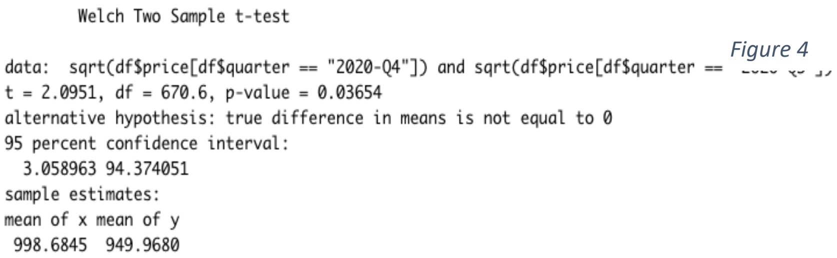
Figure 3



From here we will try to see price behavior for both of these quarters using common zip and common neighborhood which are 15 and 17, respectively. Uncommon zip is 11206 and neighborhood are *Cobble hills* and *Cobble hills – west*.
It is clear from figure 3 that the prices increased in same proportion from Q3 to Q4 in almost every zip whereas the pattern in figure 3 shows that some neighborhoods went low in average price from Q3 to Q4 of 2020. However, except for a few, there is a constant proportional increase in average prices on neighborhoods in the middle of the figure. So, it is unclear to provide any robust comment on the data from EDA method.

At this point, we would like to add some more statistical evidence to our analysis by conduct a t-test with 95% confidence for Q3 and Q4 of 2020

- Null hypothesis: there is no significant different in prices between the two quarters
- Alternative hypothesis: there is a significant difference
- Comparing on square roots of price as it moves data toward normal



```
	Welch Two Sample t-test

data:  sqrt(df$price[df$quarter == "2020-Q4"]) and sqrt(df$price[df$quarter ==
t = 2.0951, df = 670.6, p-value = 0.03654
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.058963 94.374051
sample estimates:
mean of x mean of y
 998.6845  949.9680
```

Figure 4

- Results of this test from figure 4 shows that there is statistically significant difference between the average price of two quarters as the P value is < 0.05 – we will reject Null

To find further evidence on any difference in the prices of two quarters we will refer to linear regression model which explains how the response of price is related to different features.
Model features:

Response	**Features**							
Price	Gross sqft	Land sqft	Zip	Current building class	Building class at time of sale	Tax class at time of sale	Year house was built	Which quarter of which year the house was sold
	Number of Variants (some are singular so my df stays under 40 in part1)							
	NA	NA	16	13	13	2	NA	20

From the table above we can see that building class at the time of sale and tax class at the time of sale can change with time and also if we have exact date when the house was built, we can see if any of the house is becoming a year older when it's sold in Q4 instead of Q3.

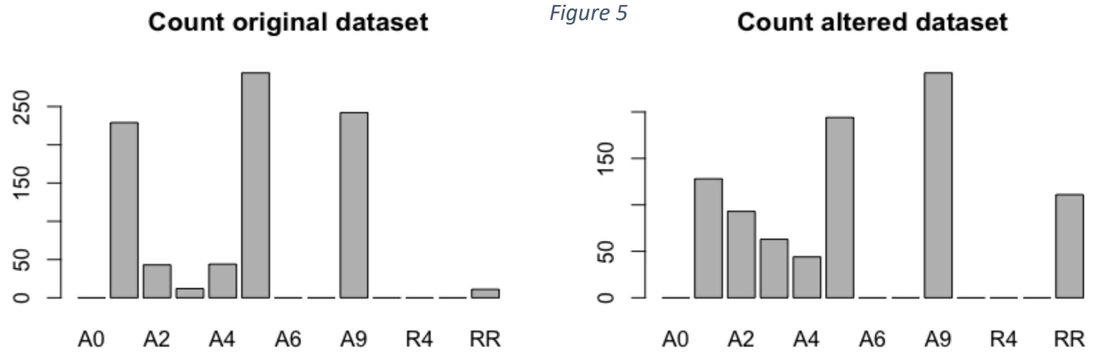


Figure 5

We will first change the building class at the time of sale so see how the price behaves between Q3 and Q5, we will randomly change categories

Coefficients are the value by which the price is affected by every factor in the model			**% Difference between quarters**
	Coefficients before and after alteration of data		
	Quarter3 2020	**Quarter4 2020**	
Original Dataset	100.9	143.5	42.22
Altererd Dataset	99.63	143.0	43.53
% Difference	-1.25%	-0.35%	

As you can see from the table even after altering the values the difference from Q3 to Q4 remains the same, so there is a constant increase from Q3 to Q4 and as we can see from figure 6 as well there is no significant different between altered and original dataset model fitted prices for each quarter whereas now, we can see that the difference between quarters is constant and at around **9.5% or increase of ~$90,000** based on actual given prices and that is now proven statistically using the model as well as t-test

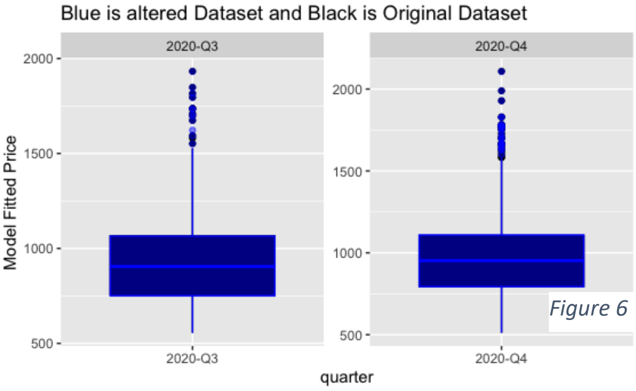


Figure 6

Model fitted	**Q3 2020**	**Q4 2020**	**Difference**
Altered	$968498.1	$1070202	$101703
Original	$967829.3	$1069469	$101639

As we can see from the table right above, the difference is constant and has not changed significantly even after changing the *housing class at the time of sale* variable. Thus there is a constant difference observed which possibly seems like a result of inflation from quarter 3 to quarter 4 of 2020 which is during the pandemic times of COVID-19 and thing were not holding steady in every financial market.

Model Limitations:

- Model is inconsiderate of high price houses
- There is a linear relationship between price and landsqft but when landsqft increases it may be increased as a parking space or as a outside garage or as a park in the apartment walls but that is considered in the model and if we had some more features such as home has garage, park, lawn then model would have better accuracy
- There is no element of renovation considered in the model which misleads model predictions such as “277 FIRST STREET 11201”
- Zip codes has been clustered to control model degree of freedom which is reducing model’s distinguishing power
- Year built could be a very good predictor, but the quality of data is not good and it has 0s, otherwise we would have had a good model because age of house significantly affects price
- The model does not cater for inflation factor
- Model is ignoring outlier and will not be good predictor where we are considering a sale of Mansion or house with very large garden