Sai Krishna Jasti
Statistical Analysis
Final Assignment Part 2

To the New York Realtors Association:

Brooklyn House price Analysis

I am pleased to present the results of my analysis into the change in Brooklyn home purchase prices between Q3 2020 and Q4 2020. This analysis scope is limited to the single family occupied residences and one-unit condos/apartments. There are various factors involved in explaining the difference in house prices such as how big the house is, how large is the area it is in, when was it built, which neighborhood is it in, how long has it been on market, issues present in the house, tax class category of the house and many other relevant ones.

I attempt to start this analysis by comparing the average price of the houses sold in Quarter 3 and Quarter 4 of 2020 which is presented in the boxplot aside. There is an approximate price difference of 80k$. I tried to delve further and check if this trend is present in all the years.

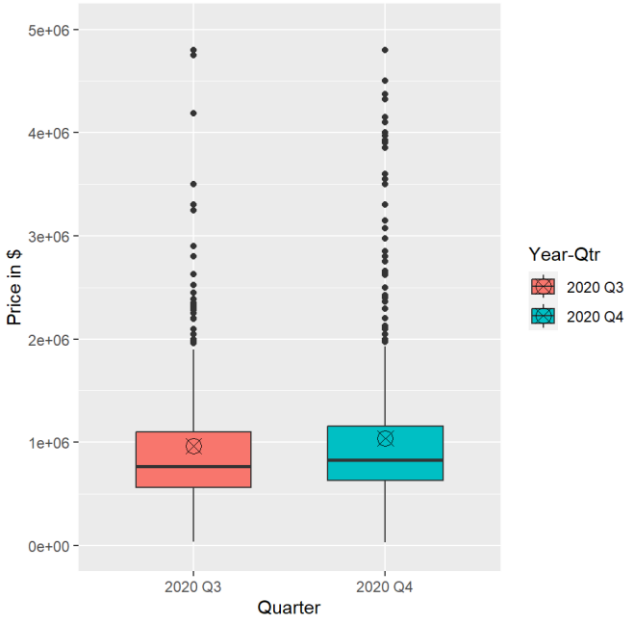


Table:1

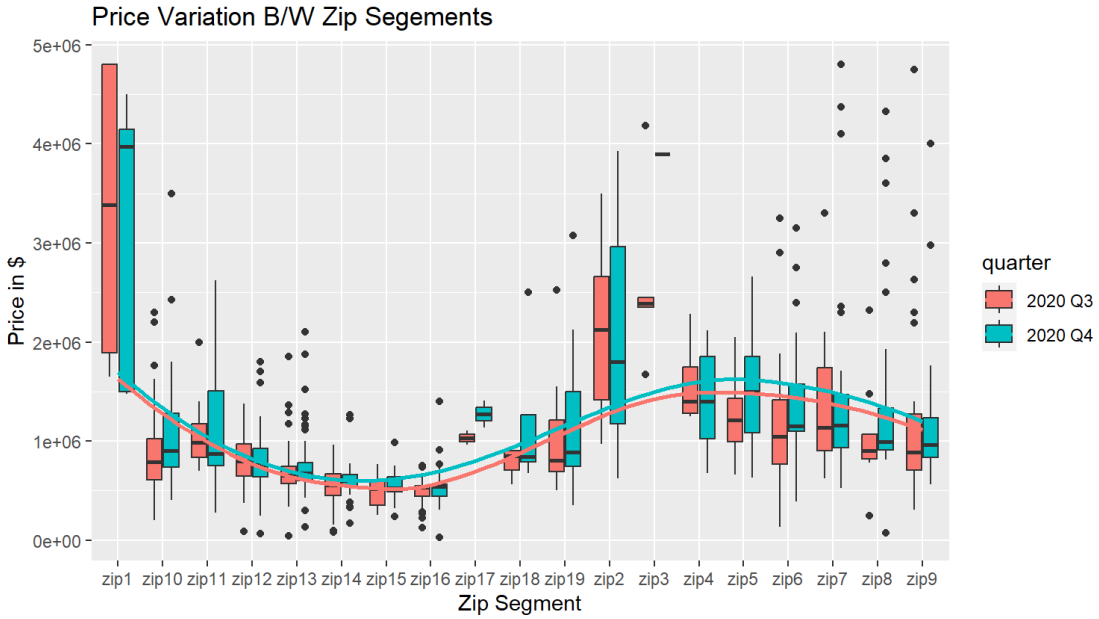
Year	Q4 - Q3 Mean Price
2016	+ $ 15,674
2017	- $ 31,551
2018	- $ 44,663
2019	- $ 3,576
2020	+ $ 75,473

From the above table (Table:1) the price increase in 2020 between Q3 and Q4 is significant. This contradicted the intuition that I had that houses are better sold in Summer, Spring or Fall compared to winter. People typically would avoid searching for houses on cold winter days is what anyone might think of. However, this shows something else. There were 317 houses sold in Q3 and 539 houses sold in Q4 of 2020. Nonetheless, I would not make my judgement based on just the average prices alone which is analogous to creating a simple linear model to explain the price variation with quarters of the years which also produced similar result of 75.5k$, and this clearly does not tell the whole story. It cannot be concluded that this variation is due to just chance variation or a real difference. There are a lot of other factors which influence the price of a house as discussed earlier.

With the data available, I first checked to see if the numeric variables in it can explain the variance in price and found that the Gross Sqft did a good job in it. I then built a linear model by considering the factors building class category, gross sqft, land sqft, sale date and the zip in which the house is located. I have segmented the zips down by accounting for the median prices of all the zips to keep the model simple enough. Using just these 5 predictors, the model was able to explain more than 60% of the variance in the house prices with a notable RMSE of under 420k$. The model also says that the houses in Q4 were roughly 75$ more expensive that which were sold in Q3, which is showing that the difference is non zero.

Sai Krishna Jasti
Statistical Analysis
Final Assignment Part 2

To check how the prices varied with the newly segmented zips, I have made a plot with multiple boxplots to see the variation. It can observed for the figure below that all the zip segments saw a rise in prices. There a constant rise in prices overall in all the zips of Brooklyn.



R actually says that the difference is not significant enough when I use just Quarters as predictors in the model. However, this model could not explain any variance of price at all with just a couple of predictors which is understandable. This can confirm that the variation is not due to chance alone which is also supported by ANOVA. When I ran Tukey's Honest Significant Difference to further see if the difference of group means between Q3 and Q4 have any relation in the predictor variable Quarter, but this also shows the p-value is 0.19 which is usually not considered to say this is a significant difference.

I have made a linear model which explains price using the just the quarters 3 and 4 of 2020 in an attempt to control the effect of other predictors on the price. This model predicted that the houses in Q4 were more expensive than in Q3 by around 52k$. The model actually shown that the difference is significant which I do agree with, but there are multiple other caveats involved into this prediction. We do not know the market conditions, changes in interest rates, amenities offered, numbers of days a property stayed on marketplace and other such factors which influence the variation in price.

I believe that this model despite having many limitations as listed above, does provide a prediction which is probable. One possible conclusion to this difference can be linked to the pandemic, the first lock down ended by June of 2020 in New York which by end of Q2. It can be thought of that the pandemic is stilling running rampant, though a little lesser than initially was, and people were not really ready enough to buy new properties as it may influence their financial situation significantly. To conclude, based on this analysis, I can say that there is still room for improvement in the modelling phase, but I can end by saying that the prices in Q4 were higher than Q3.