Brooklyn Home Purchase Analysis

Business Memo

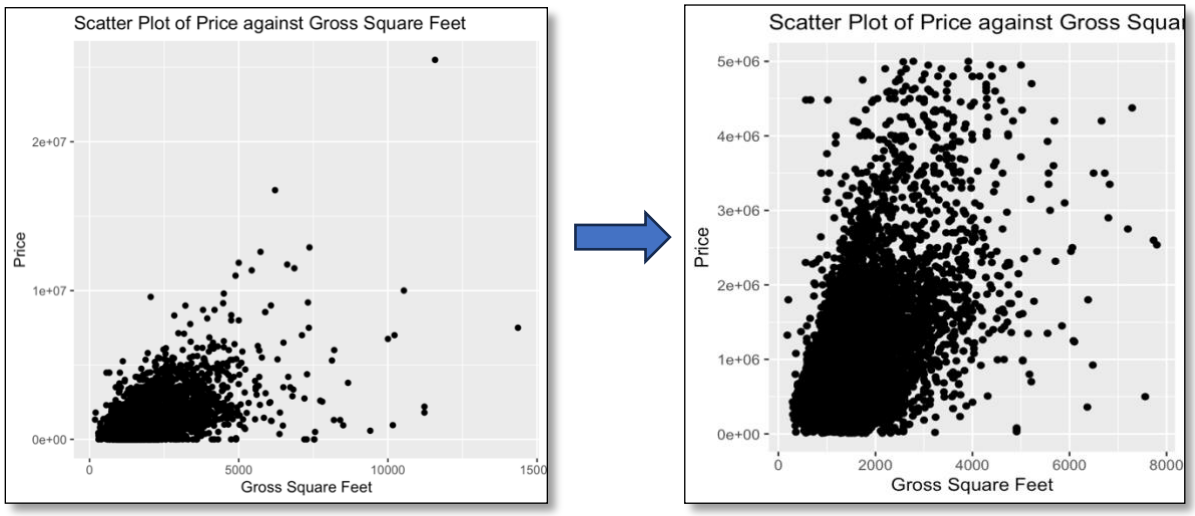
Business Problem

Determine how Brooklyn home purchase prices changed between Q3 2020 and Q4 2020.

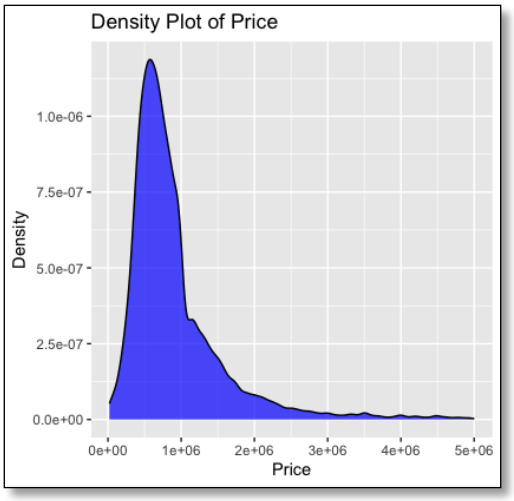
Analysis Journey

We at New York realtors Association were tasked to determine how Brooklyn home purchase prices changed between **Q3 2020 and Q4 2020**. For the purpose of this analysis, we were given data of **5 years i.e. 2016, 2017, 2018, 2019 and 2020**. The provided data had details such as Neighborhood, Building Class Category (present and at the time of sale), Tax Class Category (present and at the time of sale) Block, Tax Lot, Easement, Address, Zip code, Residential units, Commercial units, Total units, Land area of the property, Gross square feet, Year built, Sale price, Sale Date along with classification of sales such as **$0** indicating house passed down in a family or change of ownership of the property. Safe to say, we had abundant data for our analysis.

As the data was quite diverse, we decided to focus on purchases of single-family residences and single unit apartment or condos i.e. ("A" "R" building class at the time of sale). After getting our hands on the data we started exploring it and found a lot of dirty data. Thus, we moved on to data cleaning where we got rid of empty rows, houses with $0 prices, houses with 0 gross sqrft and more. Moving on we filtered houses whose prices where greater than 1000 and less than 5000000 getting rid of outliers such as houses which were passed down to another person for $10 or a Million-dollar house.



Moving on to next step, we started exploring data and found that price density was skewed to the right meaning majority of housing ranged from $500,000 to $1000,000.



To further interpret the data, we decided to go ahead with linear regression model as it can help us analyze relationship between dependent variable (price) and independent variables (grosssqrft, time etc).

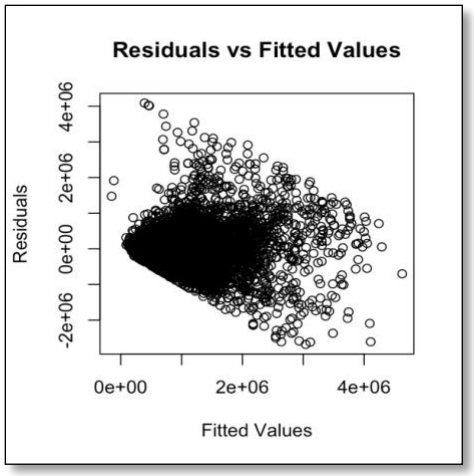
```
df$blockc <- cut(df$block,breaks = 7)
Final_model <- lm(price ~ factor(zip) * sqrt(grosssqft) + factor(df$neighborhood) + sqrt(landsqft) + blockc  * sqrt(grosssqft) + bldclasssale+quarter*year, data = df )

summary(Final_model)

rmse <- sqrt(mean((df$price - predict(Final_model, newdata=df))^2))     #R^2: 0.6102
                                                                        #RMSE: 419361.6
rmse                                                                    #DoF: 40
```

For our model to be more centric towards our business problem, we decided to bucket neighborhoods into 4 bins northern, southern, eastern, and central Brooklyn along with zips and blocks in different bins. Our final model gave us results such as **R^2: 0.6102** suggesting that a significant portion of the variability in home prices is accounted for by the factors in our model. Going forward our **RMSE: $419,361.6** indicates that on average our model's predictions are close to the actual prices lastly **DOF: 40** helps us assess the reliability of our model.

To ensure that assumptions of regression analysis are met. We went ahead with residual analysis and Breusch-Pagan test. As the null hypothesis of the Breusch-Pagan test is homoscedasticity, and our model's p value < 0.5, It is failing this test meaning our residuals are not constant across all levels of the independent variables. This being the biggest shortcoming of our model.



Approach for predicting change in prices from Q3 2020 to Q4 2020

To predict change in price we created a new variable “year_quarter” for “year” and “quarter” and changed all the values to “other_values” except for “20203” and “20204” making Q3 the reference group.

```
Final_model_2 <- Final_model_2 %>%
  mutate(year_quarter = if_else(year_quarter %in% c("20203", "20204"), year_quarter, "other_value"))


Final_model_2 <- lm(price ~ factor(zip) * sqrt(grosssqft) + factor(df$neighborhood) + sqrt(landsqft) + blockc  * sqrt(grosssqft) + bldclasssale+ year_quarter, data = Final_model_2 )
```

After running the model, we got our Q4’s coefficient as **78681.4**. The coefficient of $78,681.4 suggests that, on average, the price is estimated to **increase by $78,681.4 in Q4 compared to Q3**, holding other variables constant. This upswing in housing prices can be attributed to various factors. Primarily, the demand for space surged due to **the impact of Covid-19**.

```
year_quarter20204                    78681.4    29812.3   2.639 0.008319 **
```

In conclusion, we have thoroughly developed a model which **successfully identifies a significant increase** in home purchase price between **Q3 2020** and **Q4 2020**. Even though we predict upsurge in price we need to take in consideration causality factors such as economic shift, unemployment, upsurge in work from home trend.

Best regards,
Achintya Mishra
Data Scientist
New York Realtors Association