Can Linear Regression Value your House?

Anna Luu, Jesse Blow, Kirat Lalla, Mickey Yeh, Renata Gelevska, Spencer Poore

Abstract— The real estate market in the United States (US) forms the largest segment of Gross Domestic Product (GDP) and touches everybody in some form, either where they live, work, or spend time. As it is likely the largest purchase and expenditure for most people it is important to know the value of the property being bought or sold. This project will investigate the drivers of price for residential property sales, specifically in New York City (NYC)—the largest real estate market in the US. As the largest city by population, New York City is rich with data on sale prices, thus the perfect market to analyze. Property value predictions are extremely valuable for investors and homeowners; tools like Zillow's Zestimate feature that estimate home value are often looked at by owners, sellers, and buyers, and can have real-world consequences as it may affect one's willingness to pay or sell at a given price. Therefore, it is important to be accurate.

This project will look at a wide range of predictor variables. The predictor variables that we will be investigating are the number of units (which could be split into residential and commercial units), sale date, land square footage, gross square footage, year built, building class category, and borough. This explains some of the primary drivers for sale prices in real estate, including location (through the borough), size (through units and square footage), and zoning (through the building class codes). We will code the qualitative variable of borough number to include location in our model. While this is not the best location variable—as it is difficult to provide predictions on location without getting too complex—this is the best predictor available for location in the data set, as zip code or address would be nearly impossible to code and would likely overfit. We will also include building class categories and filter out obscure and odd building types that may add noise to our data thereby condensing them into consolidated groups. The sale date will also influence the sale price, as the real estate market has increased significantly in the last decade. The number of units will predict the sale price, as the number of units will influence how much someone is willing to pay. Likewise, the square footage may influence the number of units and will also predict the unit sizes thus influencing prices. The year built will tell how new the building is (while we do not have information on renovations and repairs the age may affect the price). Lastly, the building class code varies rent and therefore sales price. The estimated sales price also may increase at different rates depending on the interaction of different variables such as unit number and location. A change in units in Manhattan, for example, may have a different impact than a change in the number of units in the Bronx. Therefore, we will build a higher-order interaction model.

Index Terms

Real estate prices
Gross square footage
Building class codes
Regression analysis
Mediation models

1 INTRODUCTION/MOTIVATION

Understanding predictors of real estate sale prices is essential for private and public involvement in the real estate market. With real estate making up the largest segment of GDP, governments and business leaders need accurate methods to predict trends, analyze profitable investments, and overall make well-informed policy and investment decisions. Recently, the online real estate marketplace Zillow undertook a huge investment in purchasing homes they believed to be undervalued and then reselling them after some minor repairs. Zillow primarily relied on its popular "Zestimate" feature that uses a sophisticated algorithm with millions of data points to estimate the value of hundreds of millions of homes across the US. Zillow bought thousands of homes under this strategy but eventually shut down the business after reportedly losing nearly $900 million dollars.[1] This proves the difficulty in even the most sophisticated algorithms at accurately predicting home prices which makes this topic of great interest to homebuyers, investors, banks, and governments. In fact, Zillow recently gave out a prize of $1 million to a team of data scientists and engineers that won a competition to improve the accuracy of the Zestimate.[2]

Previous studies and literature have isolated specific variables, considered macro influences, and have worked towards identifying important indicators in modeling appraised value or sale price of real estate properties.

A study, conducted by R.Y. Tse in 2002, estimates the effect a neighborhood has on house prices.[3] The study found that there is a correlation between spatial distances between neighbors and house prices in urban areas. This study, as well as the numbers we ourselves gathered in our research, further support the usage of the Borough variable in our data. The study by R.Y Tse did not definitively determine those rural areas are less affected than urban areas when it comes to a neighborhood variable, which leaves the door open for further research.

Another study by Schwartz, Susin, and Voicu in 2003 investigates how falling crime rates have positively affected New York City's real estate value.[4] Meaning as crime rates fall, property value and sales price increase. The study concluded that in the decade leading up to the 2003 study, there was a correlation. Our data set views data from 2016 to 2017, which was a year in which overall crime rates fell in New York City. This is good to note, as it could add some understanding as to upward trends in real estate sale prices over the timeline of the data. The 2003 study does have issues, as biases such as the likeliness to report certain crimes can limit the accuracy of the data reported, but the study being specific to New York City further emphasizes its relevance to our research.

Schwartz and Voicu came together for another study with Horn in 2014 that observed school choice and property values.[5] The results of the study were interesting, but what was more interesting was that Schwartz and colleagues identified that various interactions of other variables, not considered in their study, have a more important effect on the outcome of the sale price of a property. Variables such as available transportation or available housing units in an area were not considered, which could affect taxation and funding for schools, etcetera. There was a closed loop on the interaction between house pricing, school funding, school performance, and then back to house pricing. This study emphasizes the importance of considering interaction variables as well as multiple variables when predicting house prices.

In 1998, a study from Dubin showed that utilizing data from MLS – where nearly all real estate listings are found – is a great method for determining house value.[6] The study utilizes regression analysis to confirm a correlation between housing prices in a localized market are similar. Some variables that were considered in this study overlap with our own research, such as the square footage and the age of the property (the year built) – and established a base for which metrics are valuable to consider when modeling for sales price of property.

Additional studies look at more macro level influencers. Such as a 2005 piece by Himmelberg, Mayer, and Sinai that discusses the influence of economic bubbles on house prices.[7] This research emphasizes the value of considering macro variables in real estate analysis. A macro-environmental variable, such as flood hazards, is something that Zhang observed in a 2016 study.[8] Zhang tested how flood hazards affect house prices through regression analysis. The study expands on the role macro factors play in house pricing – beyond just inherent risk by also considering additional costs such as flood insurance. The study concluded that there is reduced property value for properties in a floodplain. Zhang's research also notes the influence of consumer sentiment and how it is affected by macro events – which thus also affects the way a macro factor should be considered in models for pricing properties.

Prior analysis was necessary for establishing important and non-essential predictors for real estate sale prices. Our analysis is necessary for reevaluating predictors that are relevant and looking for new predictors that are valuable. Additionally, our analysis is necessary for establishing updated research for a time period close to but prior to the macro influence of COVID-19, as to aid future research that could want to see trends in the changing real estate market of New York City as a result of pandemic-related predictors.

Overview

Section 2 will address the source of the data, the methods of utilizing the data, and other general information regarding the dataset. [Data]
Section 3 expands on the analysis derived and developed from the dataset, including our models, assumptions, and tests. [Methods]
Section 4 further develops the analysis with discussions on choosing, challenging, and understanding the model. Section 4 also identifies relationships and outputs from the model. [Findings]
Section 5 reflects on the research and entices further developments. [Observation & Review]

2 THE DATA SET

To examine the housing market, we sourced the dataset *NYC Property Sales*. This dataset on Kaggle is described as a "concatenated and slightly cleaned-up version" of the *Rolling Sales* dataset found on the NYC Department of Finance's website. The data consists of every building or building unit sold on the

NYC real estate market between 2016 and 2017. Because we are examining sale prices of many different units over a fixed time period, it is cross-sectional. (It is not time-series data, i.e., it does not follow the changes in price and other variables of the same units over time.) All information is already collected by the NYC government and available within the public domain, which circumvents any costs that would have been necessary for a study of this size.

Initially, the raw file contained 84,549 total records. The variables captured in the dataset are: Borough, Neighborhood, Building Class (Current and at Time of Sale), Tax Class (Current and at Time of Sale), Block, Lot, Address, Zip Code, Number of Residential Units, Number of Commercial Units, Total Number of Units, Land Square Feet, Gross Square Feet, Year Built, Sale Price, and Sale Date. We noted that less than 10% of records had null values for Sale Price (as opposed to missing observations). The source of the data explains this with the following: "Many sales occur with a nonsensically small dollar amount: $0 most commonly. These sales are actually transfers of deeds between parties: for example, parents transferring ownership to their home to a child after moving out for retirement." Due to the large size of our initial dataset, we felt that regardless of these null values, there was still enough information outside these records to reasonably examine the relationship between sale price and other housing variables.

As we examined the data further, we found complications working with both residential and commercial building types due to the large variety in Building Classes among the two groups. A simple plot of Sale Price additionally illustrated some outliers in price—mostly commercial buildings that were sold for more than one billion dollars. Our team was more interested in residential data for the implications of pricing on the average home buyer, and as such, decided to filter the original data to solely residential buildings. This reduced the data to 57,312 records, which was still a substantial amount of information to work with.

To prepare the data for statistical analysis, slight adjustments to variables were made. For example, we chose Borough as the best qualitative variable to represent the building's location in this analysis, as there are only five boroughs in New York. (The Neighborhood, Address, and Zip Code variables were too specific and numerous to cleanly represent in the model, and we only wanted a single variable representing location to avoid multicollinearity.) We recoded the Borough variable into four new variables labeled "Bronx", "Brooklyn", "Staten Island", and "Queens". All units located in the Bronx had the value 1 for "Bronx" and 0 for the three remaining variables. Similar methodology was applied to units in other boroughs. In this way, units located in Manhattan became the base value of comparison in our final model.

Additionally, to account for time, we recoded Sale Date (originally in MM/DD/YYYY HH:SS format) to a variable that represented the number of months since the start of the data collection period. This allowed us to account for time more broadly. While day and time may have been too granular, grouping by month could give some insight into seasonality. This recoding also let us consider the potential influence of rising costs due to inflation and/or other broader environmental factors that naturally increase over time. Interestingly, it should be noted that although the data set does encompass a full twelve months, it spans from September 2016 to September 2017 rather than starting in January (as September is the start of the fiscal year).

3 THEORY AND METHODS

Regression is the parametric technique to predict and understand the reactions of dependent variables given a set of independent variables. In effect, the purpose is to see whether, with the assumptions set in place, there is any implementation of any of the independent variables that will cause the dependent variables to change. By creating a model that implements all the varied factors that would predict the output, the analysis can also give an indication of how well that model is suited for the situation given, and whether and by how much each of the variables being tested is affecting the output of the model. The aim of the regression analysis is to show whether adding a specific variable would cause the dependent variable to change, and how this model can be adjusted to fit all the requirements while still giving an accurate representation of the data and its meaning. Each variable, along with the mean value of the dependent variable when all the independent variables are zero is given. In the specific data that we looked at on sale prices, the aim of our regression was to look at the factors that go into determining the price of the building. By considering various independent variables such as Borough, Year Built, and Gross Square Feet, we were able to predict and form a model that would explain the relationship between these variables and the various output prices. At the same time, we were looking at how statistically significant this model is, by looking at the significance of both the overall model and each of the individual variables. In addition, we looked at measures such as R squared, Adjusted R squared, R squared Jack, the p-value, the F-statistic, and SSE to see if the model was able to explain the output.

The very first step we did when conducting the regression analysis was to determine what the Null and Alternative Hypotheses were. We figured that the most important aspect of this data set is to identify which of the multiple variables had the biggest effect on sale price. With that, we predicted that, by default, that none of the variables would be significant in predicting the sale price, while the alternative was that at least of the variables are significant in predicting the sale price. The f-statistic is 947.9 on 11 and 37174 degrees of freedom. The p –values for each of the variables showed to be significant to 0.001 alpha level with an overall p-value of 2.2e-16.

Commented [PS1]: Can we do this? Most examples were the other way around right?

$$H0: B0 = 0$$

$$Ha: B0 \neq 0$$

One of the ways we structured our model was by looking at the variables we were given. Since the data included both quantitative and qualitative variables, we chose to go with ones that are easy to code and do not have too much variation. We also looked at variables from a logical perspective, asking ourselves what a reasonable, average person looking to purchase a home considers when purchasing a property.

First, we looked at each variable separately, I.e., we looked at gross square feet as directly affecting sale price, the total number of units, and the year built. We then conducted a Stepwise and determined the R-Squared, Adjusted R-squared, AIC and C(p) to determine how much of the model is explained.

Using the same variables, we then conducted a Brute-force model, which showed us slightly different results. The brute-force also showed an extra variable was significant in determining the sale price.

4 DATA ANALYSES

To understand the relationship between the sale price and the important variables such as Sale Month and Gross Square Feet, the first important approach conducted to understand the data is through exploratory data analysis. This research created scatter plots of Sale Price vs. Sale Month (Figure 2), Sale Price vs. Gross Square Feet (Figure 4), Sale Price vs. Total Units (Figure 5), and Sale Price vs. Sale Month, separated by BOROUGH (Figure 6). Through the Brute-Force approach, the research found the best model for New York Property Sales (Model 2).

Once we were able to figure out how the individual variables on their own influenced the price, we then conducted a more complex model, having variables that interacted with each other, for example how gross square feet multiplied by total units, along with the rest of the variables would affect our adjusted R-squared. In this way, we increased the complexity further, including variables such as Borough multiplied by total units and looked at the sales price. We chose Borough as one of our main variables to look at because we were under the assumption that a property in Queens would have a much different selling price than a property in Brooklyn due to the numerous factors in each of the areas, such as being close to the city, level of transportation, occupation etc. When we looked at the overall results, there were some that were surprising while others were not.

Model 3: $E(y) = 5875175.25494 + 370314.32568*X1 + 8598.39207*X2 - 5715077.56456*X3 - 5487936.89007*X4 - 5504271.03216*X5 - 5281319.47478*X6 - 0.26511*X7 - 306852.68901*X8 - 343665.67807*X9 - 229098.17983*X10 - 308874.73444*X11$

An increase in the number of units (x1) by one unit increases the sale price by $370,314.33.

An increase in the sales month (x2) by one month increases the predicted sales price by $8,598.39.

A building in the Bronx coded as x3 decreases the predicted sales price by $5,715,077.57.

A building in Queens coded as x4 decreases the predicted sales price by $5,487,936.89.

A building in Staten Island coded as x5 decreases the predicted sales price by $5,504,271.03.

A building in Brooklyn coded as x6 decreases the predicted sales price by $5,281,319.47.

Looking at the results of the hypothesis, we predicted that all the variables we chose would have an effect on sale price. This is because we were under the assumption that property buyers would usually look at a few different variables before purchasing a property. We also estimated that some variables would have a bigger impact, and especially when combining multiple variables, there would be a bigger significance. For instance, our hypothesis was that the year built will have an influence on the sale price, however, when we added other variables and increased the complexity, we found that year built was not the variable that was affecting sale price, but instead, it was variables that were affecting the year built in the first place such as gross square feet as newer buildings tend to be larger than those built 100 years ago. Additionally, by looking at the scatter cloud plotting year built and sale price the association is not linear so we would be unable to use linear regression on this variable. It is likely that older historical buildings sell for more, meanwhile, buildings built from 1950 through 2000 have relatively flat sale prices, then new buildings are more desirable in the last 20 years and thus sell for more.

To understand how well Model 3 is, the research analyzed the model through a 9-point checklist. The first item on the 9-point checklist is the R-squared and the adjusted R-squared. From figure 13, we can see that Model 2 has the highest R-squared and highest adjusted R-squared among all the possible combinations. Second, in Figure 14 we can see the model also has a small MSE compared to other models. The third and fourth items on the 9-point checklist are the overall F-test and the individual t-test of each variable. As mentioned earlier in part 3 Theory and Methods, Model 3 has passed the overall F-test and all variables are highly significant. Next, we analyzed the residuals of the model, which is the fifth item on the 9-point checklist. From Figure 15, the Residual vs. Fitted Plot, we found that all data points are separated evenly around the identity line. This means the model has good residuals. Also, from Figure 16, the normal Q-Q plot, we can see the values fall along the line, but each end of the histogram extends further in a different direction. This means we have heavy-tailed distributions, and the data have more extreme values than would be expected if they truly came from a normal distribution (Clay, 2015). [10] The 6th item on our checklist is logic. We can see that all the selected variables can be understood logically.

However, when it comes to the 7th item on the checklist, which is compact interval, we can see it is not so compacted. As we can see from figure 1 to figure 6, the figures show that the data points are not super compacted. Also, from the Q-Q plot (Figure 16), we can see that the dataset has quite a few extreme values, these are the reasons why the 7th item on the checklist is not the best. For the 8th item on the checklist, we know that Model 2 has a low Press value. This is a good thing to have because it indicates that the model is good for predicting future values. The last item on the checklist is the multicollinearity check. In Figure 7, we can see there is multicollinearity between total units and gross square feet. This is understandable because a bigger property will have more units compared to a small one. However, one cannot imply the other and both variables are very important; to have a good logical model, we must keep both variables in the model.

5 CONCLUSIONS AND FUTURE WORK

Summarize (think of this as an elaborate Abstract):

New York City is an important market to assess real estate prices. The research took into account all five boroughs and gave a greater weight to Manhattan, which has some of the most expensive real estate in the United States. Given that Zillow tried to use its Zestimate feature to predict housing prices, it did not consider macro trends such as the state of the economy, job market, school system, and many individual subjective factors that people look at when buying a home. Property is a major investment and buyers take a cautious approach when selecting their home. Though our data is specific to a period over 2016 to 2017, an algorithm together with a regression model cannot be relied upon solely when it comes to property prices. Our research did provide reasonable ranges and the factors we included in our model were deemed significant to predict sale prices. Our model was also limited and specific to our research location, New York City. The model used can only be used in the context of real-estate prices in that city. In terms of future research, it would be interesting to see how macro data and variables along with common data re-identification is brought together in a method that best predicts the sale price of New York City property. The model should be adjusted to allow for more variation to be explained to be a greater predictor of sale price. It is also important to see if the model can take data from future years i.e 2018 and how well it does in predicting output. Importantly, New York City also suffered the effects of the Covid-19 pandemic and flight out of the city to the suburbs, it would be interesting to see how these affected the sale prices of New York City boroughs.

APPENDIX

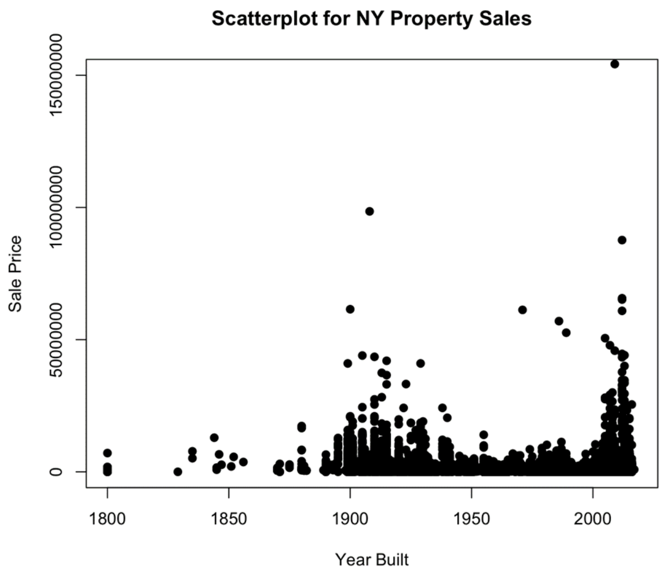

Figure 1: Sale Price vs Year Built

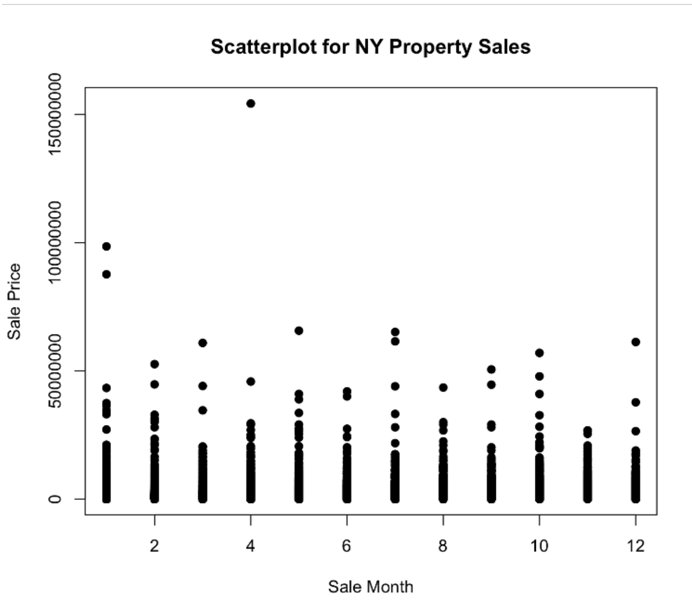

Figure 2: Sale Price vs. Sale Month

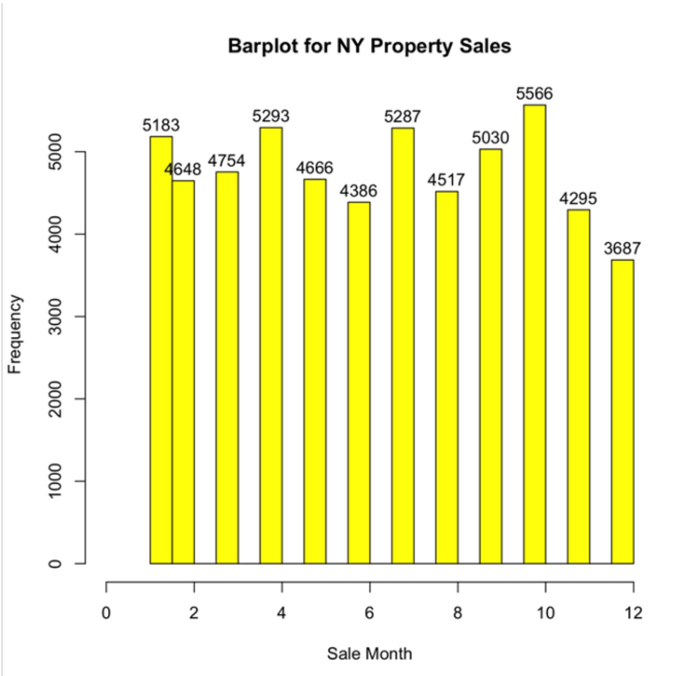

Figure 3: Frequency vs. Sale Month

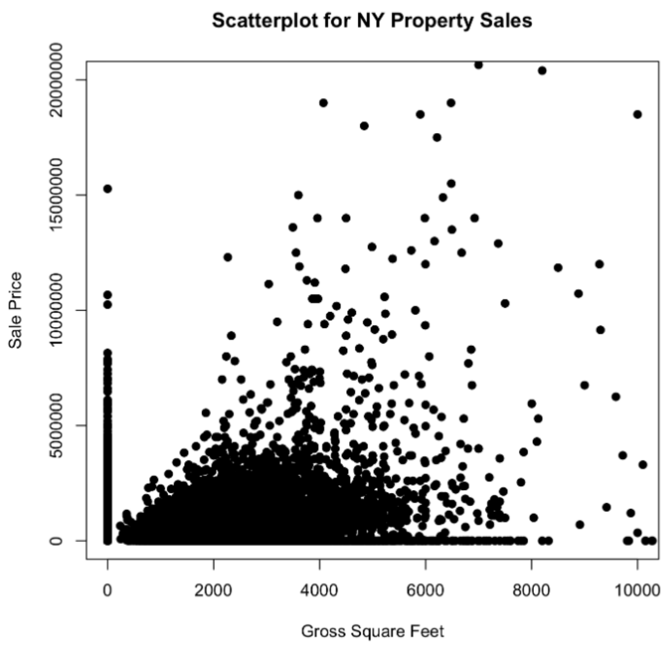

Figure 4: Sale Price vs. Gross Square Feet

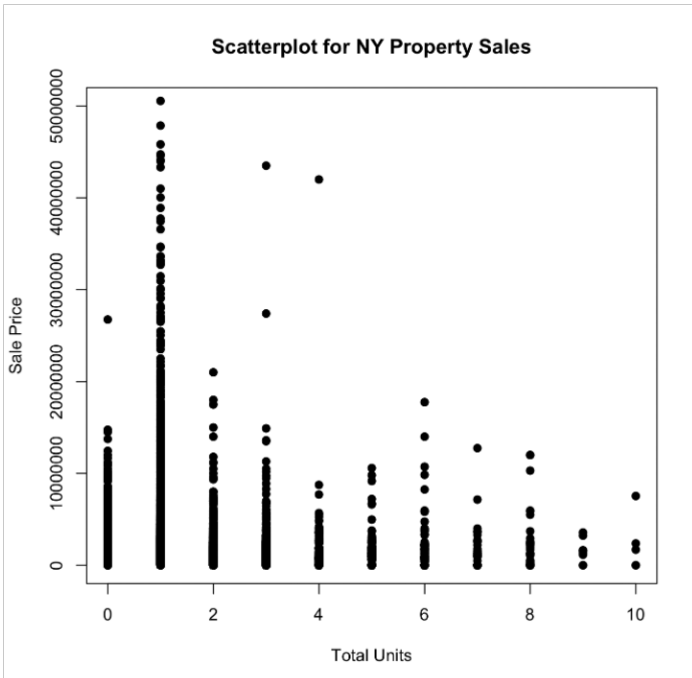

Figure 5: Sale Price vs. Total Units

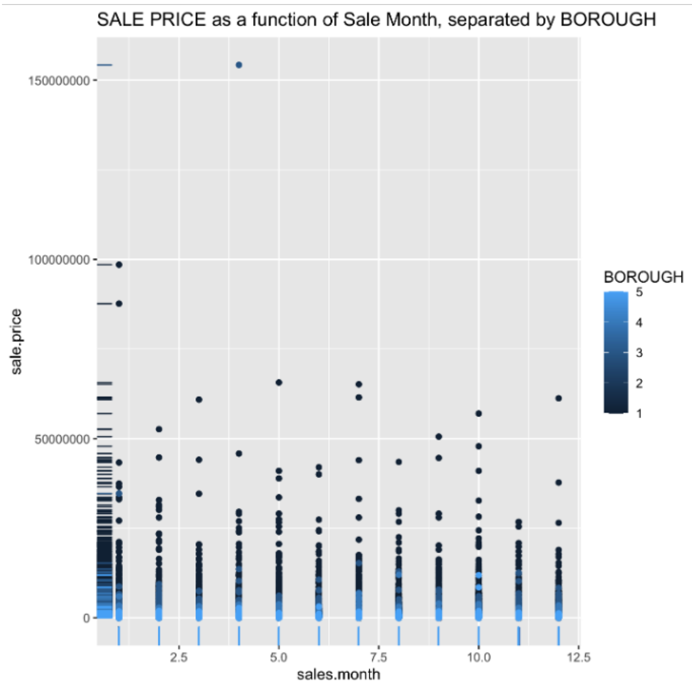

Figure 6: Sale Price vs. Sale Month, separated by BOROUGH

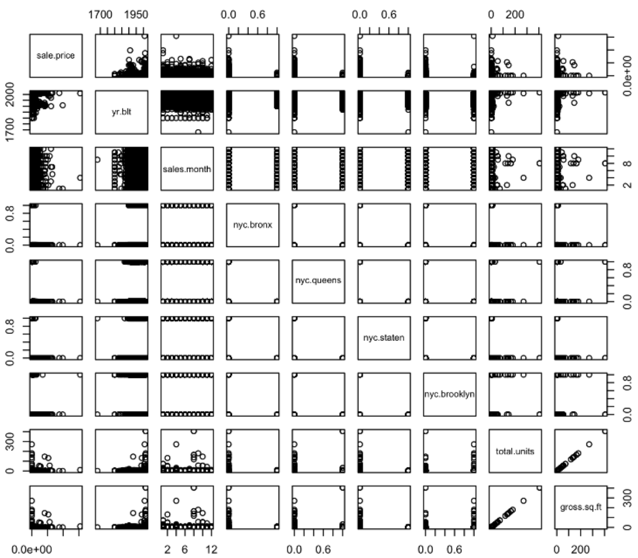

Figure 7: Data-based Multicollinearity

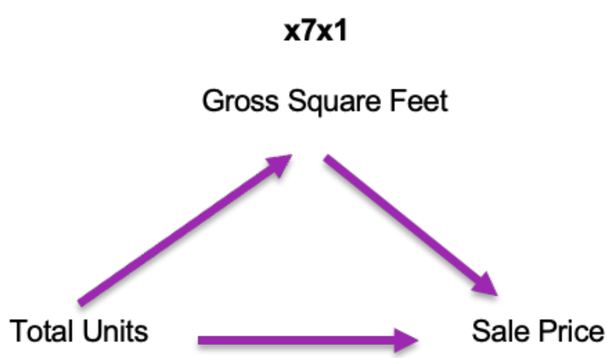

Figure 8: Mediation Model 1

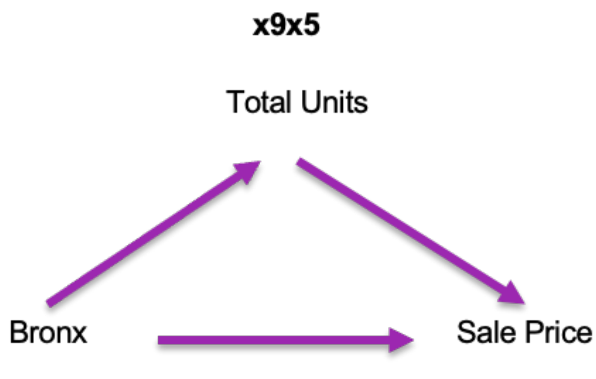

Figure 9: Mediation Model 2

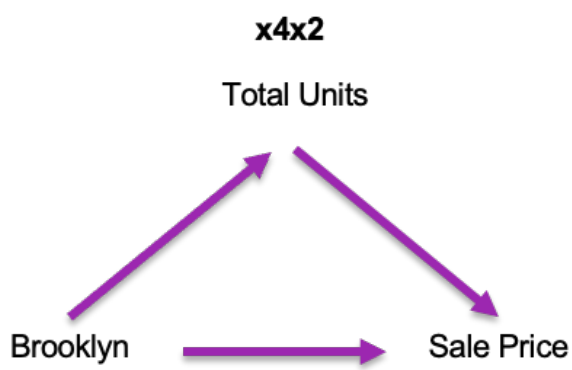

Figure 10: Mediation Model 3

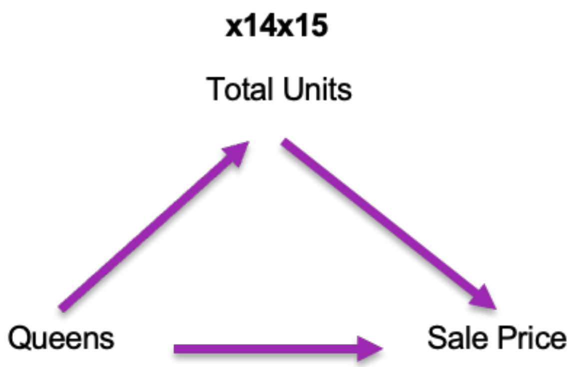

Figure 11: Mediation Model 4

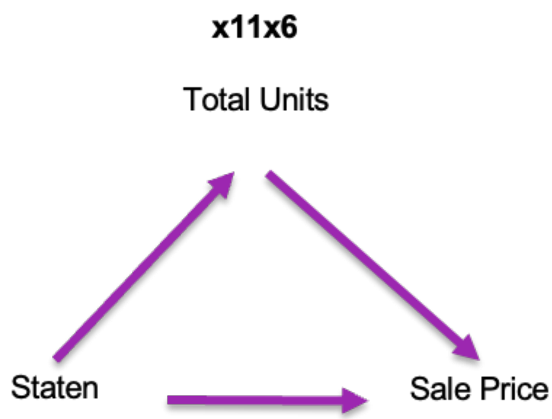

Figure 12: Mediation Model 5

```
Call:
lm(formula = sale.price ~ total.units + sales.month + nyc.bronx +
    nyc.queens + nyc.staten + nyc.brooklyn, data = a)

Residuals:
      Min        1Q    Median        3Q       Max
-21530026   -492123    -90211    195423 152853226

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   2810668      30278  92.830 <0.0000000000000002 ***
total.units     77436       3824  20.252 <0.0000000000000002 ***
sales.month      3588       2745   1.307               0.191
nyc.bronx    -2644738      38296 -69.060 <0.0000000000000002 ***
nyc.queens   -2292312      29915 -76.627 <0.0000000000000002 ***
nyc.staten   -2445483      36254 -67.455 <0.0000000000000002 ***
nyc.brooklyn -2280036      28679 -79.503 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2033000 on 47333 degrees of freedom
  (9972 observations deleted due to missingness)
Multiple R-squared:  0.1512,    Adjusted R-squared:  0.1511
F-statistic:  1405 on 6 and 47333 DF,  p-value: < 0.00000000000000022
```

Figure 13: Summary of Best Model from Brute-Force

```
Analysis of Variance Table

Response: sale.price
                Df             Sum Sq            Mean Sq  F value                 Pr(>F)
total.units      1    588012421808035    588012421808035  142.212 < 0.00000000000000022 ***
sales.month      1     61414958459428     61414958459428   14.853              0.0001164 ***
nyc.bronx        1   2164941037844494   2164941037844494  523.594 < 0.00000000000000022 ***
nyc.queens       1   2609674845761275   2609674845761275  631.153 < 0.00000000000000022 ***
nyc.staten       1   3309303823135830   3309303823135830  800.359 < 0.00000000000000022 ***
nyc.brooklyn     1  26134780923743476  26134780923743476 6320.731 < 0.00000000000000022 ***
Residuals    47333 195711165210882304      4134772045103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14: Anova test of Best Model from Brute-Force

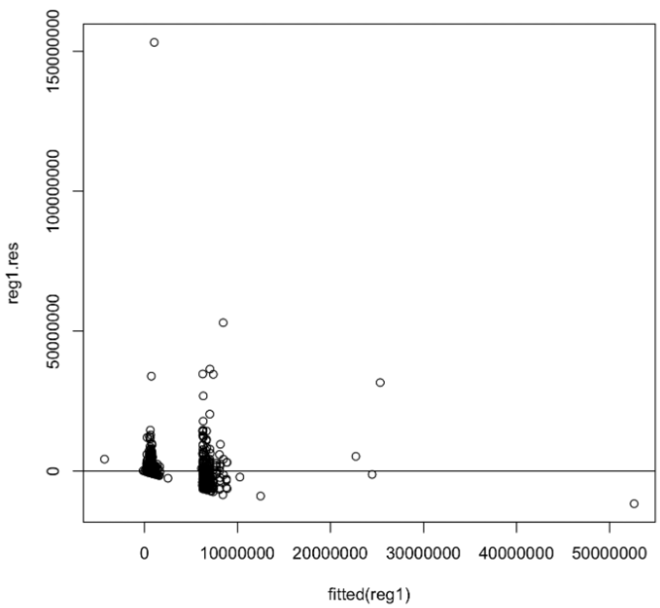

Figure 15: Residual vs. Fitted Plot

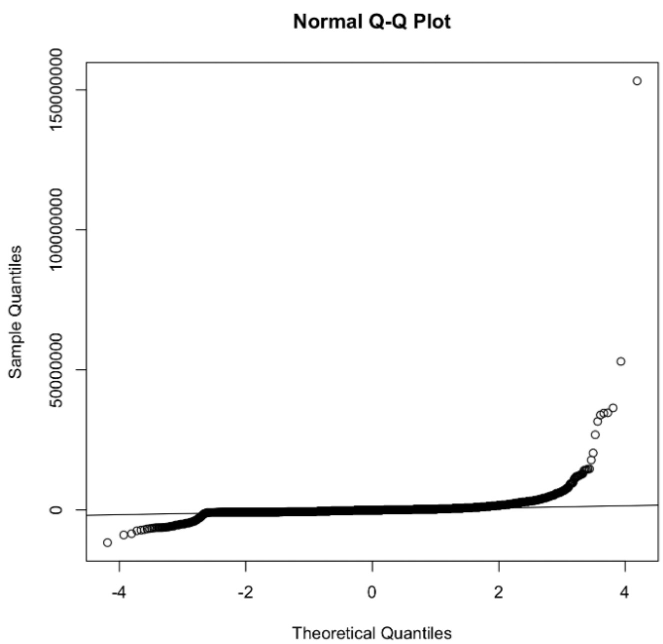

Figure 16: Normal Q-Q Plot

Tolerance and Variance Inflation Factor

	Variables	Tolerance	VIF
1	total.units	0.07846090	12.745202
2	sales.month	0.99897328	1.001028
3	nyc.bronx	0.04184471	23.897884
4	nyc.queens	0.02566798	38.959042
5	nyc.staten	0.03772704	26.506189
6	nyc.brooklyn	0.02335878	42.810447
7	x7x1	0.25343201	3.945831
8	x9x5	0.17114311	5.843063
9	x4x2	0.10348921	9.662843
10	x11x6	0.17861511	5.598631
11	x14x15	0.13805947	7.243255

Figure 17: Tolerance and Variance Inflation Factor

Eigenvalue and Condition Index

	Eigenvalue	Condition Index	intercept	total.units	sales.month	nyc.bronx	nyc.queens	nyc.staten	nyc.brooklyn	x7x1	x9x5
1	3.992032087	1.000000	0.00029787734343	0.0022599616289	0.0101753511318	0.00029462635610	0.000311373481	0.00024427999	0.000259004049	0.00200352280	0.001169409
2	2.210164911	1.343955	0.00008699395846	0.0031397518715	0.0032691667105	0.00029500201200	0.000579663643	0.00076644402	0.000086277609	0.02996053589	0.001074378
3	1.926732147	1.439416	0.00000004575050	0.0000037111873	0.0000239261079	0.00046193890775	0.001094275504	0.00611387798	0.000001796764	0.00025075475	0.001925785
4	1.908157602	1.446405	0.00000002552364	0.0000009030325	0.0000009900382	0.00712714156136	0.001219754654	0.00002450657	0.000002191564	0.00014306611	0.029869640
5	1.295187892	1.755620	0.00019990984317	0.0014486164160	0.0091086879748	0.00069343219937	0.000891655302	0.00076388308	0.004586389100	0.02835915327	0.003490612
6	0.213747226	4.321622	0.00004061867238	0.0186998139037	0.3622380926135	0.00006796150402	0.000009838908	0.00006017397	0.000429786856	0.40601510565	0.016110265
7	0.151584628	5.131797	0.00311824533127	0.0026345235349	0.4726400191342	0.00000001461422	0.000004381672	0.00004603453	0.015017102068	0.31053472623	0.010123673
8	0.101619886	6.267692	0.00562003231234	0.0260159071205	0.1229178566089	0.07433575483517	0.029726678438	0.00131818571	0.001255538299	0.00006105329	0.297151639
9	0.090978323	6.624118	0.00029244282959	0.0007330427071	0.0044930849727	0.10158852010340	0.048988430705	0.00013059482	0.000096846061	0.00001218486	0.420432776
10	0.068764263	7.619309	0.00094822037509	0.0060474055666	0.0071963497821	0.00374531304364	0.004959492023	0.21636491779	0.000423369884	0.00030533494	0.014145977
11	0.038183173	10.224946	0.00152570834065	0.7962831382166	0.0020495113003	0.00016202368499	0.000075142008	0.00053909852	0.004567534984	0.19990882125	0.176521859
12	0.002847863	37.440138	0.98786987971948	0.1427332248143	0.0058869636252	0.81122827117797	0.912139313664	0.77362800302	0.973274162761	0.02244574096	0.027983988

	x4x2	x11x6	x14x15
1	0.001932635105724	0.002177309	0.00086784910
2	0.011361311362791	0.003913676	0.00271445199
3	0.000108564769154	0.008082036	0.02283601051
4	0.000073136505571	0.009053408	0.00009199663
5	0.000000002362067	0.007731163	0.00319483605
6	0.062833800078954	0.030739222	0.00852804539
7	0.114977340581129	0.018011877	0.00447497947
8	0.047423347577005	0.215022256	0.00620547163
9	0.001430156383356	0.373058232	0.00087341706
10	0.007631967570813	0.031164787	0.83397388483
11	0.630742258489846	0.257756609	0.10035091398
12	0.121485479213590	0.043289426	0.01588814337

Figure 18: Eigenvalue and Condition Index

```
Call:
lm(formula = sale.price ~ total.units + sales.month + nyc.bronx +
    nyc.queens + nyc.staten + nyc.brooklyn + x7x1 + x9x5 + x4x2 +
    x11x6 + x14x15, data = a)

Residuals:
      Min        1Q    Median        3Q       Max
-12989829   -337664    -55721    200881 153345536

Coefficients:
                   Estimate     Std. Error t value             Pr(>|t|)
(Intercept)   5875175.25494    83621.43271  70.259 < 0.0000000000000002 ***
total.units    370314.32568     8428.54293  43.936 < 0.0000000000000002 ***
sales.month      8598.39207     1852.47952   4.642      0.0000034697055 ***
nyc.bronx    -5715077.56456    92861.42509 -61.544 < 0.0000000000000002 ***
nyc.queens   -5487936.89007    87636.35035 -62.622 < 0.0000000000000002 ***
nyc.staten   -5504271.03216    95701.70214 -57.515 < 0.0000000000000002 ***
nyc.brooklyn -5281319.47478    82585.10580 -63.950 < 0.0000000000000002 ***
x7x1               -0.26511        0.03848  -6.889      0.0000000000057 ***
x9x5          -306852.68901    22939.40990 -13.377 < 0.0000000000000002 ***
x4x2          -343665.67807     7502.32534 -45.808 < 0.0000000000000002 ***
x11x6         -229098.17983    19211.26939 -11.925 < 0.0000000000000002 ***
x14x15        -308874.73444    35497.09472  -8.701 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1213000 on 37174 degrees of freedom
  (20126 observations deleted due to missingness)
Multiple R-squared:  0.219,     Adjusted R-squared:  0.2188
F-statistic: 947.9 on 11 and 37174 DF,  p-value: < 0.00000000000000022
```

Figure 19: Final Model Summary

```
Analysis of Variance Table

Response: sale.price
                Df            Sum Sq           Mean Sq   F value                Pr(>F)
total.units      1   805878435266168   805878435266168  547.9251 < 0.00000000000000022 ***
sales.month      1    42760132203464    42760132203464   29.0731         0.00000007012 ***
nyc.bronx        1   601515631290180   601515631290180  408.9767 < 0.00000000000000022 ***
nyc.queens       1    24275216082052    24275216082052   16.5050         0.00004862222 ***
nyc.staten       1   246676643502692   246676643502692  167.7180 < 0.00000000000000022 ***
nyc.brooklyn     1 10460716141378622 10460716141378622 7112.3492 < 0.00000000000000022 ***
x7x1             1    31562438195982    31562438195982   21.4596         0.00000362580 ***
x9x5             1       61349457361       61349457361    0.0417                0.8382
x4x2             1  2829741529561314  2829741529561314 1923.9706 < 0.00000000000000022 ***
x11x6            1   180940386798084   180940386798084  123.0232 < 0.00000000000000022 ***
x14x15           1   111359558287559   111359558287559   75.7145 < 0.00000000000000022 ***
Residuals    37174 54674855375438432     1470782142773
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 20: Anova test of Final Model

Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
4	1 1	nyc.bronx	0.0090114871	0.0089905528	7843.786918
7	2 1	nyc.brooklyn	0.0082902468	0.0082692973	7883.946701
5	3 1	nyc.queens	0.0073737671	0.0073527982	7934.977725
3	4 1	yr.blt	0.0073536762	0.0073309637	7337.709084
6	5 1	nyc.staten	0.0062011908	0.0061801971	8000.268601
1	6 1	total.units	0.0025501528	0.0025290820	8203.564098
2	7 1	sales.month	0.0002440461	0.0002229266	8331.971717
27	8 2	nyc.queens nyc.brooklyn	0.0288079234	0.0287668903	6743.490451
25	9 2	nyc.bronx nyc.brooklyn	0.0229988025	0.0229575240	7066.951367
23	10 2	nyc.bronx nyc.queens	0.0204933545	0.0204519701	7206.458621
28	11 2	nyc.staten nyc.brooklyn	0.0199122649	0.0198708560	7238.814598
26	12 2	nyc.queens nyc.staten	0.0174423553	0.0174008420	7376.343019
24	13 2	nyc.bronx nyc.staten	0.0172678805	0.0172263599	7386.058047
21	14 2	yr.blt nyc.staten	0.0165078102	0.0164628032	6869.014256
19	15 2	yr.blt nyc.bronx	0.0140856455	0.0140405277	6993.559122
20	16 2	yr.blt nyc.queens	0.0127801121	0.0127349345	7060.688118
22	17 2	yr.blt nyc.brooklyn	0.0121072376	0.0120620293	7095.286532
10	18 2	total.units nyc.bronx	0.0119479842	0.0119062388	7682.278177
13	19 2	total.units nyc.brooklyn	0.0117996378	0.0117578861	7690.538336
9	20 2	total.units yr.blt	0.0110397303	0.0109944731	7150.176505
11	21 2	total.units nyc.queens	0.0097118784	0.0096700385	7806.788037
15	22 2	sales.month nyc.bronx	0.0092458500	0.0092039903	7832.737229
18	23 2	sales.month nyc.brooklyn	0.0085015417	0.0084596506	7874.181475
12	24 2	total.units nyc.staten	0.0084843112	0.0084424194	7875.140894

16	25 2	sales.month nyc.queens 0.0076496782 0.0076077512 7921.614561
14	26 2	sales.month yr.blt 0.0074418336 0.0073964118 7335.176129
17	27 2	sales.month nyc.staten 0.0063843833 0.0063424028 7992.068157
8	28 2	total.units sales.month 0.0028165035 0.0027743722 8190.733274
61	29 3	nyc.bronx nyc.queens nyc.brooklyn 0.0635289387 0.0634695882 4812.170136
63	30 3	nyc.queens nyc.staten nyc.brooklyn 0.0617749230 0.0617154613 4909.836468
62	31 3	nyc.bronx nyc.staten nyc.brooklyn 0.0412615070 0.0412007453 6052.055483
60	32 3	nyc.bronx nyc.queens nyc.staten 0.0353812043 0.0353200700 6379.479910
42	33 3	total.units nyc.queens nyc.brooklyn 0.0326596703 0.0325983634 6531.019173
52	34 3	sales.month nyc.queens nyc.brooklyn 0.0290484937 0.0289869580 6732.095119
40	35 3	total.units nyc.bronx nyc.brooklyn 0.0275032319 0.0274415983 6818.137708
58	36 3	yr.blt nyc.queens nyc.brooklyn 0.0272221140 0.0271553375 6320.097325
57	37 3	yr.blt nyc.queens nyc.staten 0.0262151454 0.0261482998 6371.874471
59	38 3	yr.blt nyc.staten nyc.brooklyn 0.0260130088 0.0259461493 6382.268094
55	39 3	yr.blt nyc.bronx nyc.staten 0.0252954349 0.0252285261 6419.164905
43	40 3	total.units nyc.staten nyc.brooklyn 0.0232983053 0.0232364052 7052.274583
38	41 3	total.units nyc.bronx nyc.queens 0.0232276374 0.0231657328 7056.209482
50	42 3	sales.month nyc.bronx nyc.brooklyn 0.0231879436 0.0231260365 7058.419695
56	43 3	yr.blt nyc.bronx nyc.brooklyn 0.0228914758 0.0228244020 6542.773662
54	44 3	yr.blt nyc.bronx nyc.queens 0.0227619178 0.0226948351 6549.435381
48	45 3	sales.month nyc.bronx nyc.queens 0.0207654498 0.0207033892 7193.307931
53	46 3	sales.month nyc.staten nyc.brooklyn 0.0200376377 0.0199755309 7233.833647
39	47 3	total.units nyc.bronx nyc.staten 0.0199143674 0.0198522528 7240.697527
36	48 3	total.units yr.blt nyc.staten 0.0198421164 0.0197748333 6699.568142
41	49 3	total.units nyc.queens nyc.staten 0.0193983367 0.0193361894 7269.430925
34	50 3	total.units yr.blt nyc.bronx 0.0181228539 0.0180554528 6787.970600
51	51 3	sales.month nyc.queens nyc.staten 0.0176424031 0.0175801445 7367.204049

49	52	3	sales.month nyc.bronx nyc.staten	0.0174328919	0.0173706200	7378.869956
37	53	3	total.units yr.blt nyc.brooklyn	0.0165987601	0.0165312544	6866.337715
46	54	3	sales.month yr.blt nyc.staten	0.0165590735	0.0164915650	6868.378357
35	55	3	total.units yr.blt nyc.queens	0.0161430268	0.0160754898	6889.770987
44	56	3	sales.month yr.blt nyc.bronx	0.0141729950	0.0141053227	6991.067711
45	57	3	sales.month yr.blt nyc.queens	0.0128943080	0.0128265480	7056.816295
30	58	3	total.units sales.month nyc.bronx	0.0122056425	0.0121430393	7669.931361
47	59	3	sales.month yr.blt nyc.brooklyn	0.0121779461	0.0121101369	7093.650783
33	60	3	total.units sales.month nyc.brooklyn	0.0120334828	0.0119708688	7679.517480
29	61	3	total.units sales.month yr.blt	0.0111327773	0.0110648963	7147.392139
31	62	3	total.units sales.month nyc.queens	0.0100099774	0.0099472351	7792.189421
32	63	3	total.units sales.month nyc.staten	0.0086871056	0.0086242794	7865.848986
98	64	4	nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1438443568	0.1437720082	342.082314
81	65	4	total.units nyc.bronx nyc.queens nyc.brooklyn	0.0693792645	0.0693006233	4488.414864
83	66	4	total.units nyc.queens nyc.staten nyc.brooklyn	0.0655822619	0.0655032998	4699.837895
91	67	4	sales.month nyc.bronx nyc.queens nyc.brooklyn	0.0637454388	0.0636663215	4802.115069
97	68	4	yr.blt nyc.queens nyc.staten nyc.brooklyn	0.0627759118	0.0626901286	4493.962765
93	69	4	sales.month nyc.queens nyc.staten nyc.brooklyn	0.0618729351	0.0617936595	4906.379001
95	70	4	yr.blt nyc.bronx nyc.queens nyc.brooklyn	0.0553775035	0.0552910432	4874.380238
82	71	4	total.units nyc.bronx nyc.staten nyc.brooklyn	0.0458244627	0.0457438310	5799.982987
96	72	4	yr.blt nyc.bronx nyc.staten nyc.brooklyn	0.0428693043	0.0427816991	5517.537162
92	73	4	sales.month nyc.bronx nyc.staten nyc.brooklyn	0.0413479335	0.0412669235	6049.243119
94	74	4	yr.blt nyc.bronx nyc.queens nyc.staten	0.0410587381	0.0409709672	5610.634353
80	75	4	total.units nyc.bronx nyc.queens nyc.staten	0.0376763317	0.0375950114	6253.683636
90	76	4	sales.month nyc.bronx nyc.queens nyc.staten	0.0355614924	0.0354799934	6371.441192
72	77	4	total.units sales.month nyc.queens nyc.brooklyn	0.0329257145	0.0328439928	6518.205418
78	78	4	total.units yr.blt nyc.queens nyc.brooklyn	0.0318147442	0.0317261272	6085.949672

79	79	4	total.units yr.blt nyc.staten nyc.brooklyn	0.0303694572	0.0302807079	6160.264630
77	80	4	total.units yr.blt nyc.queens nyc.staten	0.0290521006	0.0289632308	6228.001560
75	81	4	total.units yr.blt nyc.bronx nyc.staten	0.0289726176	0.0288837404	6232.088486
76	82	4	total.units yr.blt nyc.bronx nyc.brooklyn	0.0282666456	0.0281777038	6268.388738
70	83	4	total.units sales.month nyc.bronx nyc.brooklyn	0.0277158992	0.0276337372	6808.296062
88	84	4	sales.month yr.blt nyc.queens nyc.brooklyn	0.0273202895	0.0272312611	6317.049256
74	85	4	total.units yr.blt nyc.bronx nyc.queens	0.0264437055	0.0263545969	6362.122178
87	86	4	sales.month yr.blt nyc.queens nyc.staten	0.0262856109	0.0261964878	6370.251217
89	87	4	sales.month yr.blt nyc.staten nyc.brooklyn	0.0260390552	0.0259499096	6382.928819
85	88	4	sales.month yr.blt nyc.bronx nyc.staten	0.0253424349	0.0252532254	6418.748221
68	89	4	total.units sales.month nyc.bronx nyc.queens	0.0235235488	0.0234410326	7041.732674
73	90	4	total.units sales.month nyc.staten nyc.brooklyn	0.0234412857	0.0233587625	7046.313216
86	91	4	sales.month yr.blt nyc.bronx nyc.brooklyn	0.0229546662	0.0228652382	6541.524486
84	92	4	sales.month yr.blt nyc.bronx nyc.queens	0.0228832086	0.0227937741	6545.198749
69	93	4	total.units sales.month nyc.bronx nyc.staten	0.0200991664	0.0200163608	7232.407630
66	94	4	total.units sales.month yr.blt nyc.staten	0.0198975677	0.0198078599	6698.716902
71	95	4	total.units sales.month nyc.queens nyc.staten	0.0196170504	0.0195342040	7259.252605
64	96	4	total.units sales.month yr.blt nyc.bronx	0.0182152816	0.0181254199	6785.218077
67	97	4	total.units sales.month yr.blt nyc.brooklyn	0.0166728828	0.0165828798	6864.526416
65	98	4	total.units sales.month yr.blt nyc.queens	0.0162616955	0.0161716549	6885.669185
113	99	5	total.units nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1511890623	0.1510994004	-64.882347
118	100	5	sales.month nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1438650856	0.1437746501	342.928104
119	101	5	yr.blt nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1425353051	0.1424371992	394.828364
106	102	5	total.units sales.month nyc.bronx nyc.queens nyc.brooklyn	0.0696250643	0.0695267867	4476.728346
112	103	5	total.units yr.blt nyc.queens nyc.staten nyc.brooklyn	0.0671999010	0.0670931758	4268.486432
108	104	5	total.units sales.month nyc.queens nyc.staten nyc.brooklyn	0.0656967458	0.0655980532	4695.463252
117	105	5	sales.month yr.blt nyc.queens nyc.staten nyc.brooklyn	0.0628024021	0.0626951738	4494.600664

110	106	5	total.units yr.blt nyc.bronx nyc.queens nyc.brooklyn	0.0616651882	0.0615578297	4553.074872
115	107	5	sales.month yr.blt nyc.bronx nyc.queens nyc.brooklyn	0.0554774949	0.0553694284	4871.238801
111	108	5	total.units yr.blt nyc.bronx nyc.staten nyc.brooklyn	0.0483115525	0.0482026661	5239.703149
107	109	5	total.units sales.month nyc.bronx nyc.staten nyc.brooklyn	0.0459269814	0.0458262005	5796.274586
109	110	5	total.units yr.blt nyc.bronx nyc.queens nyc.staten	0.0441605515	0.0440511902	5453.142749
116	111	5	sales.month yr.blt nyc.bronx nyc.staten nyc.brooklyn	0.0428835897	0.0427740823	5518.802627
114	112	5	sales.month yr.blt nyc.bronx nyc.queens nyc.staten	0.0411288792	0.0410191710	5609.027781
105	113	5	total.units sales.month nyc.bronx nyc.queens nyc.staten	0.0378756781	0.0377740467	6244.583718
103	114	5	total.units sales.month yr.blt nyc.queens nyc.brooklyn	0.0319170832	0.0318063211	6082.687518
104	115	5	total.units sales.month yr.blt nyc.staten nyc.brooklyn	0.0303977299	0.0302867940	6160.810884
102	116	5	total.units sales.month yr.blt nyc.queens nyc.staten	0.0291265499	0.0290154685	6226.173467
100	117	5	total.units sales.month yr.blt nyc.bronx nyc.staten	0.0290237524	0.0289126593	6231.459192
101	118	5	total.units sales.month yr.blt nyc.bronx nyc.brooklyn	0.0283330644	0.0282218922	6266.973560
99	119	5	total.units sales.month yr.blt nyc.bronx nyc.queens	0.0265699410	0.0264585671	6357.631298
124	120	6	total.units sales.month nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1512196847	0.1511120920	-64.587450
125	121	6	total.units yr.blt nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1501324117	0.1500157251	6.194056
126	122	6	sales.month yr.blt nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1425383191	0.1424205898	396.673385
123	123	6	total.units sales.month yr.blt nyc.queens nyc.staten nyc.brooklyn	0.0672286542	0.0671005849	4269.007976
121	124	6	total.units sales.month yr.blt nyc.bronx nyc.queens nyc.brooklyn	0.0617701753	0.0616413566	4549.676557
122	125	6	total.units sales.month yr.blt nyc.bronx nyc.staten nyc.brooklyn	0.0483273485	0.0481966841	5240.890934
120	126	6	total.units sales.month yr.blt nyc.bronx nyc.queens nyc.staten	0.0442348490	0.0441036227	5451.322458
127	127	7	total.units sales.month yr.blt nyc.bronx nyc.queens nyc.staten nyc.brooklyn	0.1501361857	0.1500000488	8.000000

Figure 21: Brute Force Selection

References

[1] Parker, W. (2022, February 10). *Zillow's shuttered home-flipping business lost $881 million in 2021*. The Wall Street Journal. Retrieved May 4, 2022, from https://www.wsj.com/articles/zillows-shuttered-home-flipping-business-lost-881-million-in-2021-11644529656

[2] Zillow. (2019, January 30). *Zillow Awards $1 Million to Team that Built a Better Zestimate*. Zillow MediaRoom. Retrieved May 4, 2022, from http://zillow.mediaroom.com/2019-01-30-Zillow-Awards-1-Million-to-Team-that-Built-a-Better-Zestimate

[3] Tse, R. Y. (2002). Estimating neighbourhood effects in house prices: towards a new hedonic model approach. *Urban studies*, *39*(7), 1165-1180.

[4] Schwartz, A. E., Susin, S., & Voicu, I. (2003). Has falling crime driven New York City's real estate boom?. *Journal of Housing Research*, 101-135.

[5] Schwartz, A. E., Voicu, I., & Horn, K. M. (2014). Do choice schools break the link between public schools and property values? Evidence from house prices in New York City. *Regional Science and Urban Economics*, *49*, 1-10.

[6] Dubin, R. A. (1998). Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics*, *17*(1), 35-59.

[7] Himmelberg, C., Mayer, C., & Sinai, T. (2005). Assessing high house prices: Bubbles, fundamentals and misperceptions. *Journal of Economic Perspectives*, *19*(4), 67-92.

[8] Zhang, L. (2016). Flood hazards impact on neighborhood house prices: A spatial quantile regression analysis. *Regional Science and Urban Economics*, *60*, 12-19.

[7] Parker, W. (2022, February 10). Zillow's shuttered home-flipping business lost $881 million in 2021. The Wall Street Journal. Retrieved May 4, 2022, from https://www.wsj.com/articles/zillows-shuttered-home-flipping-business-lost-881-million-in-2021-11644529656.

[8] Zillow. (2019, January 30). Zillow Awards $1 Million to Team that Built a Better Zestimate. Zillow MediaRoom. Retrieved May 4, 2022, from http://zillow.mediaroom.com/2019-01-30-Zillow-Awards-1-Million-to-Team-that-Built-a-Better-Zestimate.

[9] City of New York. (December, 2017). NYC Property Sales. Retrieved April 13, 2022 from https://www.kaggle.com/datasets/new-york-city/nyc-property-sales.

[10] Clay Ford (2015, August 26). Understanding Q-Q Plots. *University of Virginia Library.* https://data.library.virginia.edu/understanding-q-q-plots/.