A Look at Real Estate Prices in NYC 2003 - 2021

By Samuel Schappel

12/2/21

sms748@scarletmail.rutgers.edu

Introduction

This report aims to evaluate real estate price trends for each Borough in NYC from 2003 - 2020 to understand which sections experienced the steepest decline in property values from the ongoing covid - 19 crisis and predict house values in 2022 based on simple household characteristics and location. I predict that Manhattan will experience the most significant reduction in real estate prices since it is the most expensive borough to live in. I predict as well that the burrows with the largest populations will experience the steepest declines in real estate prices for 2020 since consumers now prefer space and room over the location. Brooklyn has the largest population in NYC, surpassing Manhattan by almost one million people, which could cause Brooklyn to experience the harshest property value price drops. Since Brooklyn has the largest population size, this district should experience a massive number of homeowners who sell their property at a low price to escape their densely populated neighborhood in hopes of avoiding the virus. Because of all the people moving out of NYC to avoid covid-19, housing prices should be generally lower than normal due to the decrease in demand to live in NYC meaning that my predicted house values for 2022 should have a relatively low price relative to the building.

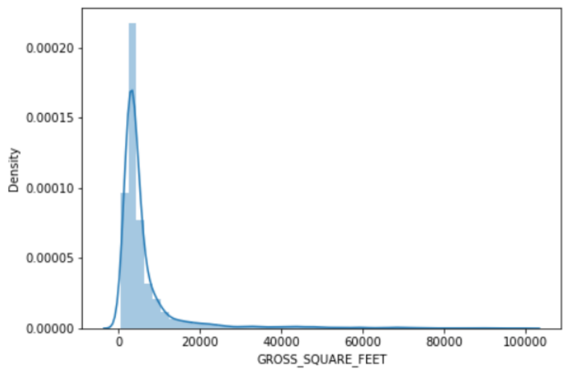
Data

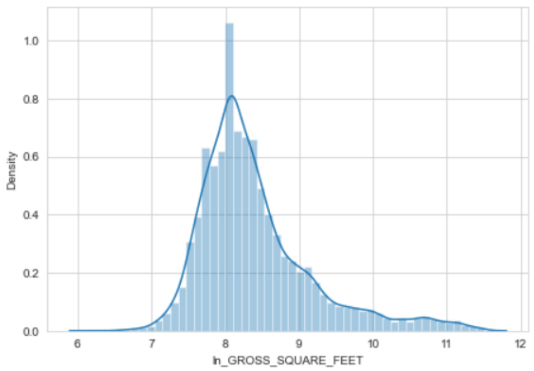
In order to evaluate these trends in prices, I have created a master data set containing housing information for every Borough in NYC. The data creation was done using an HTML parser in python called Beautiful Soup to scrape over 50 different excel files from the NYC Department of Finance Annual Sales Update[1]. Once the links to the datasets were scraped, they were read into juypter notebook using pandas and merged into one large data frame. The first few rows of the dataset contained empty values so those were dropped. Each variable was then renamed to account for neatness and to make the data easier to use and mend.

With large amounts of data like this, it is very normal for there to be unwanted values that will deter my analysis and cause mistakes. In order to clean the data properly so that machine learning algorithms can be used on it, I must account for outliers, zero values, NaN values and other issues of the same sort. Using mathematical procedures, all rows containing outliers, zero values, and NaN values were removed form the dataframe. Once this was complete, I altered the variables sales price, the lot size in square feet, the gross square feet, and the land in square feet. Histograms of these variables after data cleaning showed that some were skewed and would not necessarily fit a linear regression model. In order to better understand the relationship between these variables, I log-transformed them in order to look at their elasticities within my regression. Doing this gives the data a normal distribution and allows me to

[1] https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page

see the percent change associated with a one percent change in the natural log of the sale price, which is my independent variable. To elaborate, figure one shows the distribution of values in the variable gross





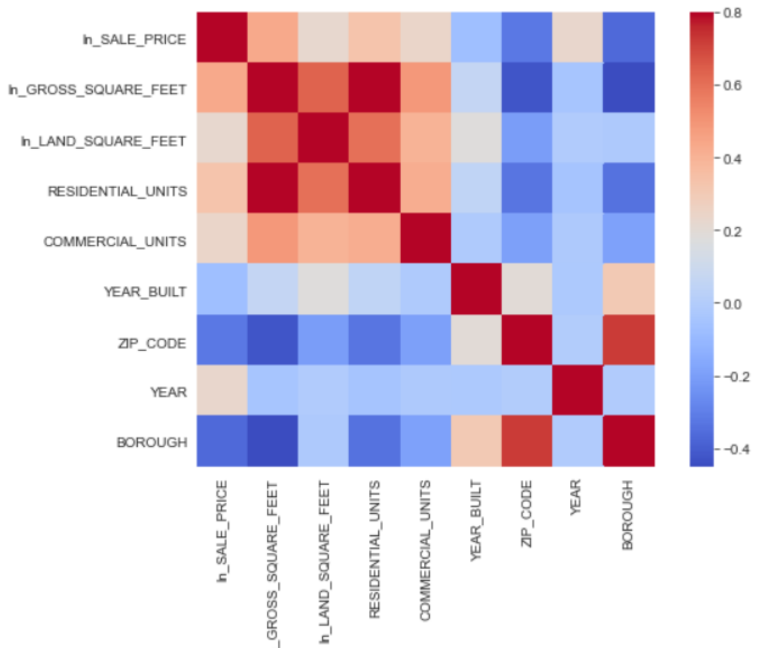
square feet. As the image shows the data is skewed to the right meaning that despite the data cleaning there are still outliers which will throw off a linear regression model. Figure two shows the variable after the natural log has been taken. The distribution is now normalized and ready to be fit into a linear regression model. The results of the log transformation for the variables sale price and land square feet yield similar results compared to gross square feet. The histogram visualizations for these two transformed variables have been added to the end of the article for readers who wish to look at them.

After I log-transformed these variables, I created dummy variables for the year of sale in order to see the difference in regression coefficients for each year without creating a separate regression equation for each individual year. Creating the dummy variables allows me to observe the difference in these coefficients all within the same equation.

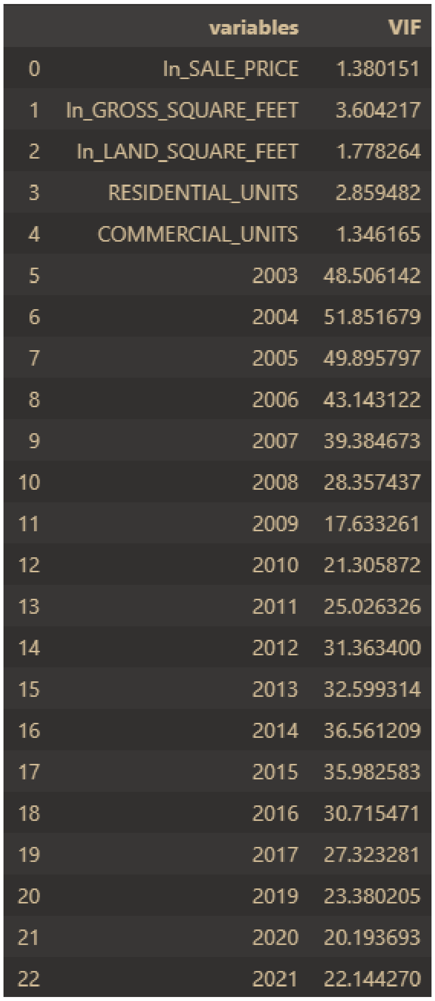
I then created a function that would split my dataset into five separate datasets, one for each Borough in NYC. These are the data frames in which I will be implementing different machine learning algorithms in order to create a predictive system as well as telling visualizations.

Correlation & Multicollinearity

In order for any regression to be successful in outputting statistically sound and accurate coefficients the most significant variables need to be included in the model and the least significant variables need to be excluded. A method used to create a visualization in order to see correlations between all variables in a data set at once is called a heatmap. Using imported python packages I created a heat map to see how each variable is correlated to one another. Figure three highlights highly correlated variables in a darker red color and negatively correlated variables in a dark blue. Note that this heatmap is created with the massive ln data frame before it was split into burrow subsets. The variables to notice in this heatmap are variables year and zip code. Since the data will be split into boroughs later I will ignore the correlations for the variable borough. Zipcode appears to be highly uncorrelated to most variables besides borough. Knowing this I decided to take the zip code variable out of my regression. The values for the zip code variable are the numbers of the zip code in which the specified property resides. The numbers in the variables have no real numerical meaning they are just locations. This is why we see so many negative correlations with this variable. Knowing this I could go one of two routes. Make the zip code variable into dummy variables to see a different coefficient for each different zip code or exclude the zip code variable from the regression and just focus on the borough coefficients. I chose to exclude the variable from the main regression because when I tried to create the zipcode dummy variables and run the regression, it was throwing off the coefficients for my year dummy variables. In regards to the year variable, there is also a large number of light blue squares which indicates a coefficient of 0. This variable is later turned into a dummy variable so we can ignore these values as well. The most interesting aspect of this map is shown in the correlation between residential units and gross square feet. The map shows that the correlation between these two variables is very large sitting at around 0.8. This makes sense as the more residential units a building has, the larger its gross square feet will be. It is interesting to note, however, that the coefficient between residential units and land square feet is smaller falling at around 0.5 - 0.6. A reason for this could be that gross square feet is the total area of every room in a building while land square feet is an area measurement of the total land a building sits on. Since new york is a highly populated area with many skyscrapers it would make sense that a building with more residential units is built up not out meaning that the land in square feet would not change but rather the gross square feet would. Land square feet does not account for height while in a way gross square feet does since it is the area of every room, not the land the building sits on. A taller building will have a larger gross square foot than a short building however, the land square feet of the two buildings could be the same.



Multicollinearity is an issue that every data scientist will have to deal with at some point in their life when dealing with regression analysis. Multicollinearity can be described as double counting one variable in a regression and usually leads to unreliable and unstable estimates of regression coefficients. The best way to check for multicollinearity is through the variance inflation factor or otherwise known as a VIF score. The vif of a variable is calculated by finding R squared and implementing the following formula, VIF = 1/(1-R^2). Generally speaking a VIF lower than 1 describes no multicollinearity, a VIF of 1-5 yields moderate multicollinearity and anything higher than 5 is highly correlated and normally yields cause for concern. In python, I have created a function that intakes a data frame and outputs the VIF scores for each variable. I was pleasantly surprised when the VIF scores of my variables all resulted in values less than 5 meaning there is no reason to take any of the variables out of the regression equation. Figure four shows the VIF score table yielding VIFs lower than 5 for all my variables except the indicator variables. It is important to note that there are a few times where we can safely ignore multicollinearity. When the variables with high VIFs are indicator variables that represent a categorical variable with three or more categories we can safely ignore multicollinearity. This means that these VIFs of 15 and higher can be overlooked since the overall test that all indicators have coefficients of zero is unaffected by the high VIF.



	variables	VIF
0	ln_SALE_PRICE	1.380151
1	ln_GROSS_SQUARE_FEET	3.604217
2	ln_LAND_SQUARE_FEET	1.778264
3	RESIDENTIAL_UNITS	2.859482
4	COMMERCIAL_UNITS	1.346165
5	2003	48.506142
6	2004	51.851679
7	2005	49.895797
8	2006	43.143122
9	2007	39.384673
10	2008	28.357437
11	2009	17.633261
12	2010	21.305872
13	2011	25.026326
14	2012	31.363400
15	2013	32.599314
16	2014	36.561209
17	2015	35.982583
18	2016	30.715471
19	2017	27.323281
20	2019	23.380205
21	2020	20.193693
22	2021	22.144270

Regression Model

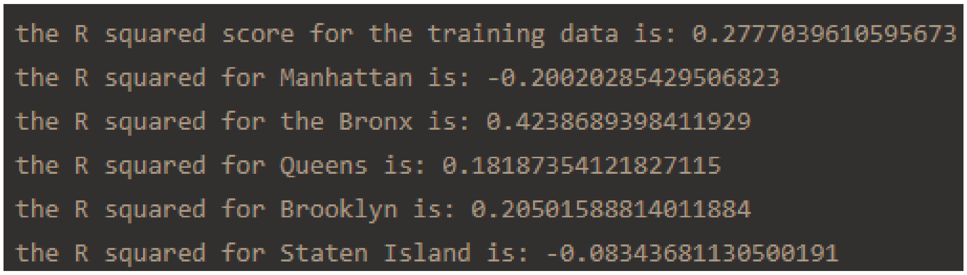
The regression model I have implemented on the New York City data I have constructed comes from the python machine learning library sklearn. From sklearn.linear_model I imported LinearRegression which is the tool I am using to run my regression and extract coefficients. Before getting into the results of each regression it is important to understand the thought process behind the actual code. I could have gone about this regression two ways using sklearn.

The first way would have been to create training data from the massive data set I created before splitting the data up into different boroughs. After fitting my model to this training data, I would set my testing data equal to each individual borough and pull out the coefficients. So to get coefficients for manhattan, I would set X_test equal to everything in my Manhattan data frame minus the natural log of the sale price since this is what I will set y_test equal to. Doing this allows me to see how the manhattan data frame correlates to a random subset of the massive data that contains every borough.

The second way of going about this regression would be to do a different train_test_split for each individual borough subset and fit the linear regression to each different X_trian and y_train. Doing this allows me to get coefficients based on data from the specific borough at hand. Unlike the other method where the training data is from the complete merged data, this method deals with training data specific to each borough.

After running both machine learning procedures I choose my second method instead of the first to extract my coefficients. The reason for this is that using the first method, the R squared values of the training data and testing data were much too far apart for an accurate regression. The idea behind a machine learning linear regression is that you train your data to be specific to your data set. By doing this you should be able to have a more accurate linear model since it was created around the shape of your data. For this to work, however, the training and testing data R squared scores must be close to each other. If the R squared scores are far apart from each other, the model will not be as accurate because it was trained to work for datasets with a similar R squared.

Below in figure five, we can see that the only borough with an R squared relatively close to the training data is Brooklyn and it is still off by a whole .07 which is not terrible but still not exactly ideal. I also get a negative R squared for Staten Island and Manhattan which tells me that the chosen model does not follow the trend of the training data and fits it worse than a horizontal line would. Knowing this, it would be counterintuitive to use the first method I stated earlier to run my linear regression and extract coefficients.

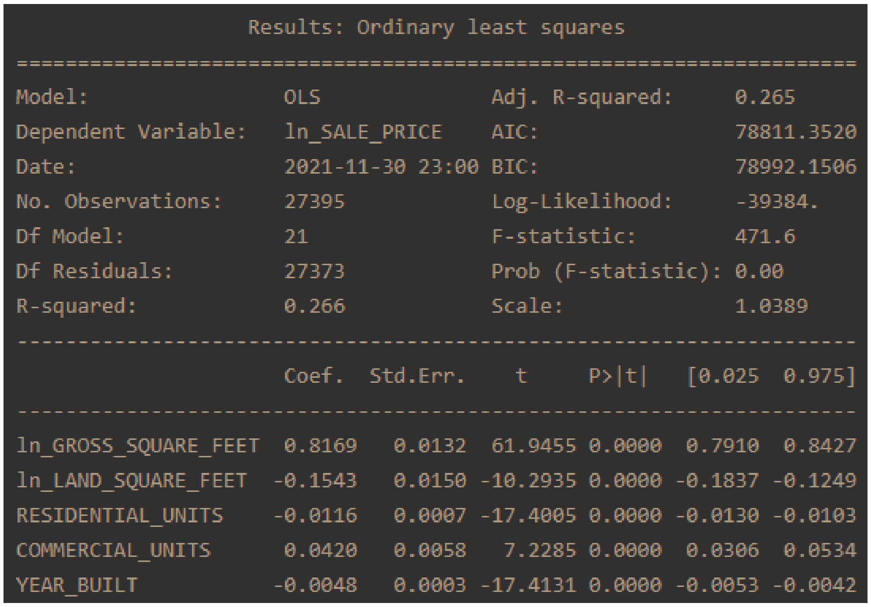


```
the R squared score for the training data is: 0.2777039610595673
the R squared for Manhattan is: -0.20020285429506823
the R squared for the Bronx is: 0.4238689398411929
the R squared for Queens is: 0.18187354121827115
the R squared for Brooklyn is: 0.20501588814011884
the R squared for Staten Island is: -0.08343681130500191
```

When looking at the R squared of the training and testing data using the second machine learning method, I am able to see that the R squared values are much tighter. Using this method, instead of one centralized training score, we will see different training scores for each borough. The training score for manhattan was .0917 while the testing score was .0702 meaning that even though the R squared is very low for this borough, the training data fits the testing data nicely. For the Bronx, the training score was .525 while the testing score was .565 showing a much more highly correlated dataset than Manhattan's with training data that fits the testing data. It is important to notice the difference between training and testing scores from method one and method two. Using the first method the difference between the R squared scores of training and testing data for the Bronx is 0.146 (0.423 - 0.277). Using the second method the difference is between training and testing R squared scores is only 0.04 meaning that the training data fits the testing data much better with this method than method one. The training and testing scores for Staten Island were 0.222 to 0.113, for Brooklyn they were 0.223 to 0.258, and for Queens, they were 0.247 to 0.270. The only concerning difference in R squared values are in the Staten Island dataset where the difference between the two is 0.109 which is oddly high for training and testing scores derived from the same dataset. However, using the first method the difference between training R squared and testing R squared is 0.36. The difference using the first method is much larger than the difference using the second method confirming that the second method is superior. Knowing that my machine learning linear regression is fitted to the data properly, I can proceed to run the regression and extract the coefficients.

Large Regression Analysis

Before taking a look at different regressions for each different borough, I began by taking the cleaned dataset as a whole and running an ordinary least square regression to see coefficients for each variable. The results are pictured in figure six below. In explanation, a ten percent increase in a property's gross square feet results in an 8.2 percent increase in that property's sale price. I was a little surprised to see that land square feet had a negative coefficient; a 10 percent increase in a property's land square feet results in a 1.5 percent decrease in the property's sale price. With land square feet and gross square feet being so closely related, one would assume that they have similar coefficients. In this report, I have already explained the difference between the two as gross square feet increases with the height of a building while land square feet increase with a larger plot of land. A prediction I created as to why land square feet is negatively correlated with sale price while gross square feet is not is derived from this simple rule about the two. The rule is that as land square feet increases, so does gross square feet, but as gross square feet increase, land square feet necessarily does not. Let's say we take two buildings with the same land square feet of 2000. However, building one



```
                    Results: Ordinary least squares
==================================================================
Model:               OLS              Adj. R-squared:     0.265
Dependent Variable:  ln_SALE_PRICE    AIC:                78811.3520
Date:                2021-11-30 23:00 BIC:                78992.1506
No. Observations:    27395            Log-Likelihood:     -39384.
Df Model:            21               F-statistic:        471.6
Df Residuals:        27373            Prob (F-statistic): 0.00
R-squared:           0.266            Scale:              1.0389
------------------------------------------------------------------
                     Coef.  Std.Err.    t     P>|t|   [0.025  0.975]
------------------------------------------------------------------
ln_GROSS_SQUARE_FEET  0.8169   0.0132  61.9455 0.0000  0.7910  0.8427
ln_LAND_SQUARE_FEET  -0.1543   0.0150 -10.2935 0.0000 -0.1837 -0.1249
RESIDENTIAL_UNITS    -0.0116   0.0007 -17.4005 0.0000 -0.0130 -0.0103
COMMERCIAL_UNITS      0.0420   0.0058   7.2285 0.0000  0.0306  0.0534
YEAR_BUILT           -0.0048   0.0003 -17.4131 0.0000 -0.0053 -0.0042
```

has gross square feet of 8000, while building two has gross square feet of 3000. Assuming the two buildings are equal in quality and location, the sale price of building one will be immensely higher than the sale price of building two even though they have the same land square footage. In this example, land square feet acts almost like camouflage for the price value of the building since if we were only to look at this variable we would assume the same sale price for both buildings. Gross square feet is the only true indicator to predict the sale price. Holding all variables constant except for land and gross square feet, two buildings can have the same land square footage while having very different prices. The same cannot be said for gross square feet since as the land square feet increases gross square feet will increase with it.

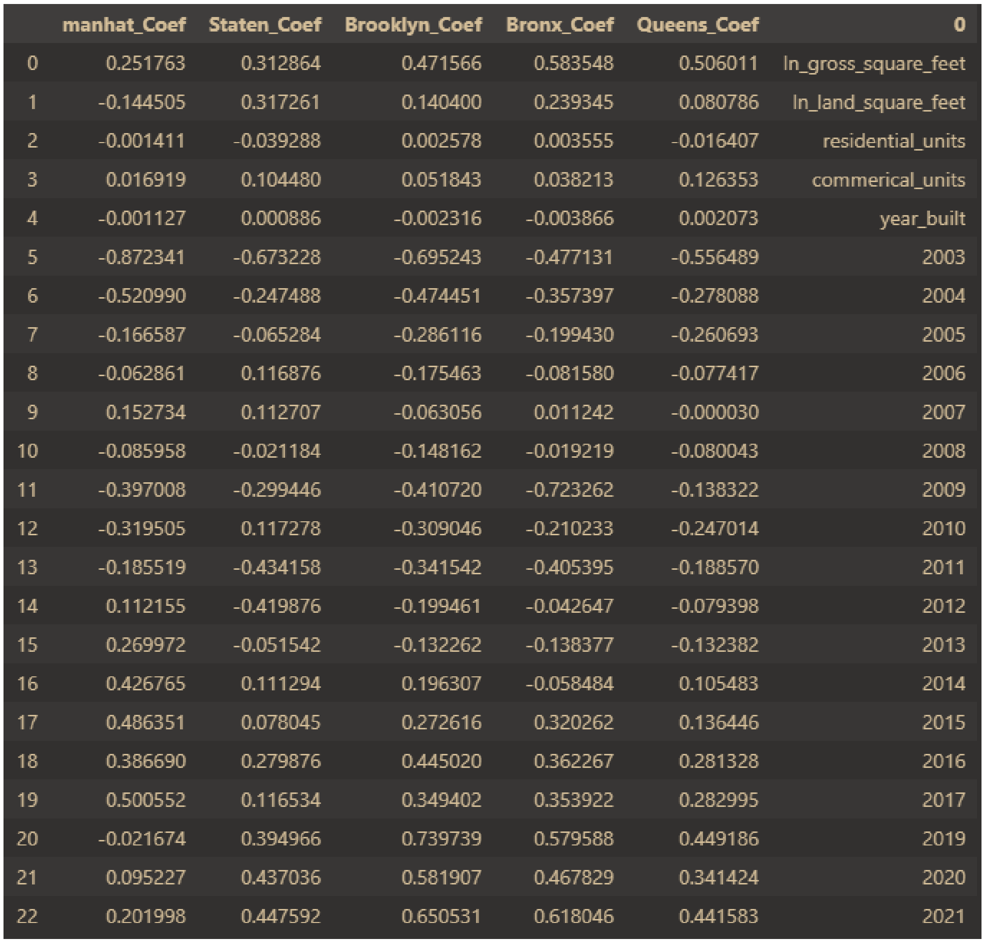
There is an oddly similar relationship between variables residential units and commercial units as there are to and square feet and gross square feet. Residential units are described as single-family homes with one to four-unit rental residences while commercial units are anything with five or more rental units. Both of the coefficients for these two variables are considerably low meaning they do not have a large impact on the sale price of a building but they still have an impact nonetheless. A ten-unit increase in a building’s commercial units yields a .4 unit increase in the sale price while a ten-unit increase in residential units yields a negative .1 unit decrease in that property’s sale price. A reason for this could be that residential properties are often much smaller and adding another unit usually means less space for the person living there. In a commercial property, the building more often than not is very large so adding another unit would be less dramatic. Either way, these variables have little to no effect on the sale price and should not be considered a large predictor of the sale price.

The last variable year built proves to have the least effect on sale price with a ten unit increase in the year a building was constructed results in a .05 unit decrease in the sale price. So a building built in 1950 will be 5 cents more expensive than a building built-in 1960. A logical reason as for why this change is so little for what seems to be a very important variable is simply due to renovations. Older buildings tend to get renovated more often than newer buildings so it would not be unlikely that older buildings are nicer inside than newer buildings resulting in a higher sale price. This cannot be said for every building in New York City which is why we see do not see a large negative coefficient and instead see a very small coefficient.

All of the p values for the variables are 0 meaning that these variables are statistically significant and the R squared is .266 meaning that 26.6 percent of the data can be explained by the model. The standard errors are all smaller than 0.02 meaning that the sample mean is close to the population mean for each variable. The F- statistic sits high at 471 proving that this is indeed a good model.

Regression by Borough

This regression analysis will take a look at all individual regressions for each borough subset using the machine learning algorithm described in the Regression Model section in this report. Figure seven shows a large coefficient array containing the main predictor variables as well as year dummy variables. For the most part, The coefficient values for variables residential units, commercial units, and year built all stay around the same value in comparison to the large OLS regression coefficients. The numbers vary by decimal places but the overall conception that these variables affect sale price less gross square feet and land square feet holds true. Residential units keep a negative coefficient of around 0.01 to 0.001 for every borough except Brooklyn and the Bronx which have very small positive coefficients in the hundredths. Commercial unit coefficients are also relatively close to the OLS model with positive coefficients in the hundredths except for Queens and Staten Island which have positive coefficients in the tenths however, neither of these coefficients exceeds 0.13. While these numbers are still small, in Queens a ten-unit increase in a building's commercial units yields a 1.2 unit increase in the sale price. While this does not change house prices by much more than the OLS model predicted, the change in coefficients is large enough to consider. The year-built variable coefficient for each borough is similar to that of the OLS model and even drops into the thousandths for Staten Island. Again these coefficients are so small that they prove to have little to no effect on sale price.

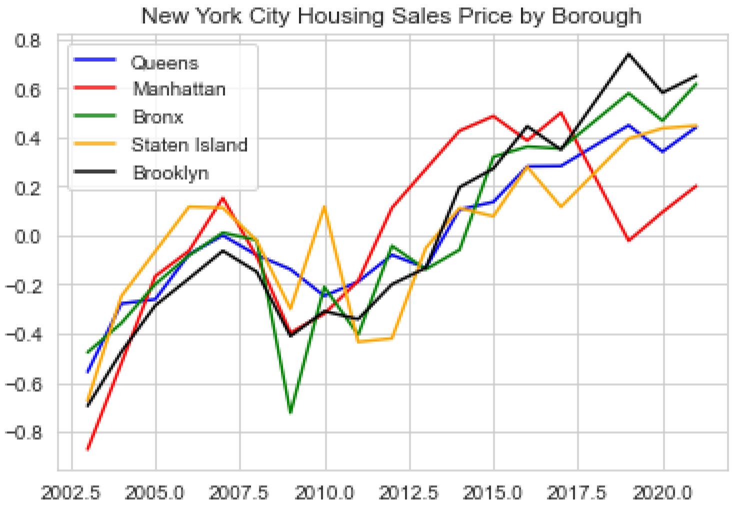


	manhat_Coef	Staten_Coef	Brooklyn_Coef	Bronx_Coef	Queens_Coef	0
0	0.251763	0.312864	0.471566	0.583548	0.506011	ln_gross_square_feet
1	-0.144505	0.317261	0.140400	0.239345	0.080786	ln_land_square_feet
2	-0.001411	-0.039288	0.002578	0.003555	-0.016407	residential_units
3	0.016919	0.104480	0.051843	0.038213	0.126353	commerical_units
4	-0.001127	0.000886	-0.002316	-0.003866	0.002073	year_built
5	-0.872341	-0.673228	-0.695243	-0.477131	-0.556489	2003
6	-0.520990	-0.247488	-0.474451	-0.357397	-0.278088	2004
7	-0.166587	-0.065284	-0.286116	-0.199430	-0.260693	2005
8	-0.062861	0.116876	-0.175463	-0.081580	-0.077417	2006
9	0.152734	0.112707	-0.063056	0.011242	-0.000030	2007
10	-0.085958	-0.021184	-0.148162	-0.019219	-0.080043	2008
11	-0.397008	-0.299446	-0.410720	-0.723262	-0.138322	2009
12	-0.319505	0.117278	-0.309046	-0.210233	-0.247014	2010
13	-0.185519	-0.434158	-0.341542	-0.405395	-0.188570	2011
14	0.112155	-0.419876	-0.199461	-0.042647	-0.079398	2012
15	0.269972	-0.051542	-0.132262	-0.138377	-0.132382	2013
16	0.426765	0.111294	0.196307	-0.058484	0.105483	2014
17	0.486351	0.078045	0.272616	0.320262	0.136446	2015
18	0.386690	0.279876	0.445020	0.362267	0.281328	2016
19	0.500552	0.116534	0.349402	0.353922	0.282995	2017
20	-0.021674	0.394966	0.739739	0.579588	0.449186	2019
21	0.095227	0.437036	0.581907	0.467829	0.341424	2020
22	0.201998	0.447592	0.650531	0.618046	0.441583	2021

The real difference between the models is in the land square feet and gross square feet variables. I found it very interesting that in the OLS model, the coefficient for gross squared feet is 0.8, while the gross square feet coefficient for the machine learning models was significantly lower. The highest coefficient we see in the regression by borough is 0.58 for the Bronx. This is the closest to mirroring the OLS model denoting that a ten percent increase in gross square feet results in a 5.8 percent increase in the sale price for properties within the Bronx. Other coefficients in the machine learning model are not so high with the lowest being Manhattan at 0.25. This came as a shock to me as I expected gross square feet to have the largest impact on sale price in Manhattan since the majority of buildings in this borough are large skyscrapers. I believe that the reason this coefficient is the lowest out of all the boroughs is due to all of the skyscrapers in this borough. With every building being so tall and having a large gross square footage, it is only reasonable that an increase in gross square feet will have less of an impact on the sale price. This theory is easily explained by the value of money. If I have one dollar to start with and add another dollar to my collection, I now have two dollars meaning I have increased my original amount by 100 percent. Now if I have ten dollars to start with and add an additional dollar, I have only increased my original worth by 10 percent. The same can be said for skyscrapers. If I have a block in Manhattan all with 100-floor skyscrapers and decide to add one more floor to one of these buildings, I will have increased that building's height by 1 percent. This building will not look much different than the others on its block. Now if I take a block in the Bronx all with two-story buildings, and I take one of those buildings and add another floor, I have increased that building's original height by 33 percent. This building will look much different from the rest on its block. The change in the price of a home with one additional floor in the Bronx is much larger than the change in the price of a home with one additional floor in Manhattan and as stated before, an increase in the height of a building increases the gross square footage.

Interestingly enough, the land square feet coefficient in the OLS model and the manhattan machine learning model are almost identical. This cannot be said for the rest of the boroughs as the coefficients for land square feet are positive, not negative. The reason I believe this is tied together with my reasoning behind the gross square feet coefficients. In Manhattan, there is not much open land to purchase and build on as there is in the other boroughs. With that said, every borough in New York City is highly populated with minimal open land for purchase. In Staten Island, however, there is more available land for purchase than there is in manhattan. With that said it is reasonable to believe that with more land open to purchase, A building's land in square feet will be more valuable than in an area where there is little land to purchase. In Staten Island, a ten percent increase in a building's land square footage yields a 3.1 percent increase in that building's sale price.

Now that the predictor variables have been explained I can examine the coefficients for each individual year dummy variable. Figure eight shows the trend in coefficients over the past 18 years for each borough. This image shows the trend in housing prices for every borough relative to my predictor variables over the past eighteen years. The accuracy of this graph can be shown by the dip in coefficients during the 2008 financial crisis. The most interesting aspect of this graph can be found in the later years between 2017 and 2021. Midway through 2017, we can see that Manhattan, the Bronx, and Brooklyn are all in similar standing for housing prices. All with coefficients of around 0.4. By the end of 2021, we can see that these three boroughs are all at different price levels with Brooklyn having the most expensive homes followed up by the Bronx, and in dead last is Manhattan. I would have thought manhattan to have the largest prices but it is key to remember that these coefficients are based on regressions run strictly within a borough. Manhattan coefficients are relative to Manhattan and Manhattan only not the other boroughs as well.

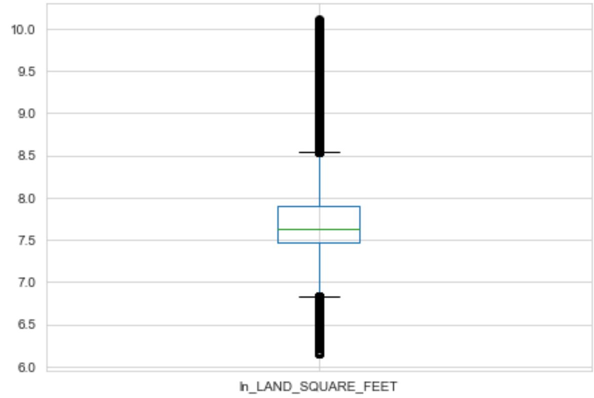


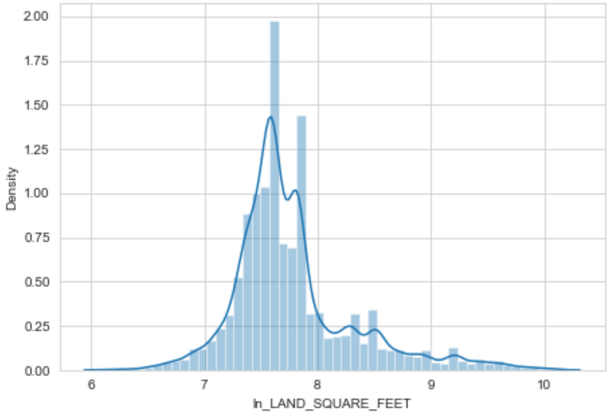
Looking at the year 2020, we can see Covid-19 has had an effect on housing prices with noticeable dips in the prices for three of the five boroughs. Brooklyn, the Bronx, and Queens all seem to fall in 2020 with Brooklyn having the steepest decline. Staten Island seems to level out into a stalemate with housing prices following an almost horizontal line after 2020. Manhattan oddly appears to excel through the Covid-19 crisis which could be due to a steep decline in prices previous to this. It is possible that prices in Manhattan dropped so low before 2020 that even through the Covid-19 crisis they continue to rise. Even in 2021 however, the housing prices in Manhattan are still significantly lower than what they were in 2017 showing the negative effect Covid has had. If Covid had never happened we might see housing prices for Manhattan be at an all-time high. For all other boroughs, it seems that the housing prices have reached an all-time high in 2021 after the Covid-19 crisis. For these boroughs, buying property in 2020 would have been a financially sound move as we expect these trends to keep rising as time goes on.

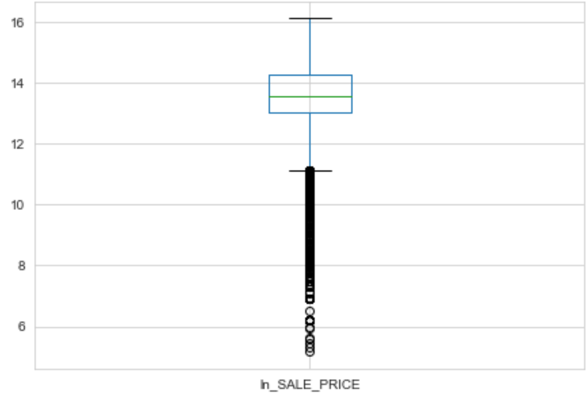
Conclusion

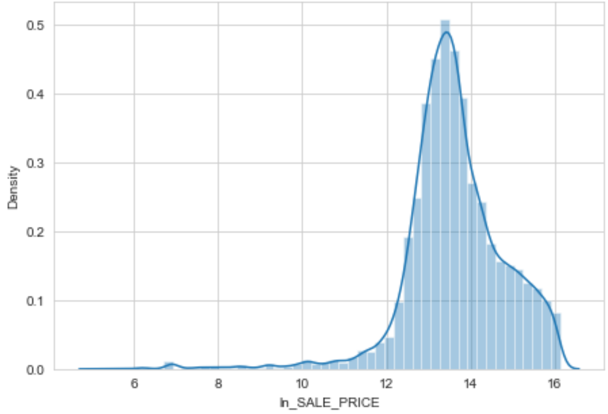
It seems that housing prices as a result of multiple machine learning programs have been impacted by the Covid-19 crisis but maybe not as much as you would have expected. There was a time in 2020 where everyone living in New York City wanted to move out to avoid the illness but I believe that as time went on and people got more used to living in a Covid controlled world they became more accepting of social distancing and wearing masks. Because of this change in attitude towards the virus, people may have been less willing to sell their houses for cheap just to get out of the City. The resulting time-series graph also shows a massive bounce back for most boroughs after 2020 which could also have something to do with people and their tolerance for Covid.

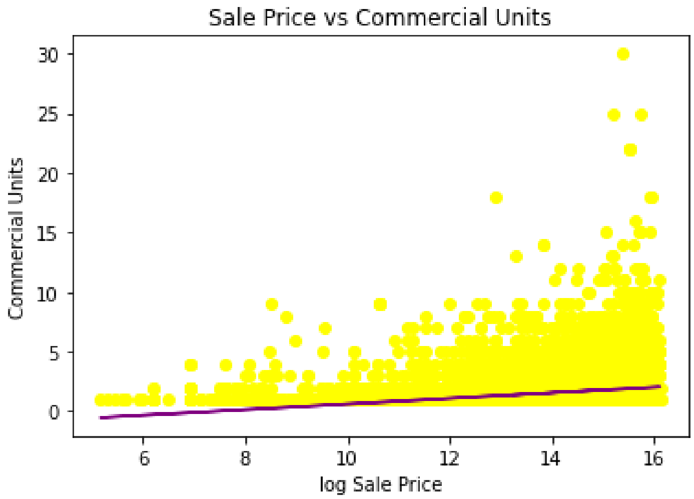
Additional Images



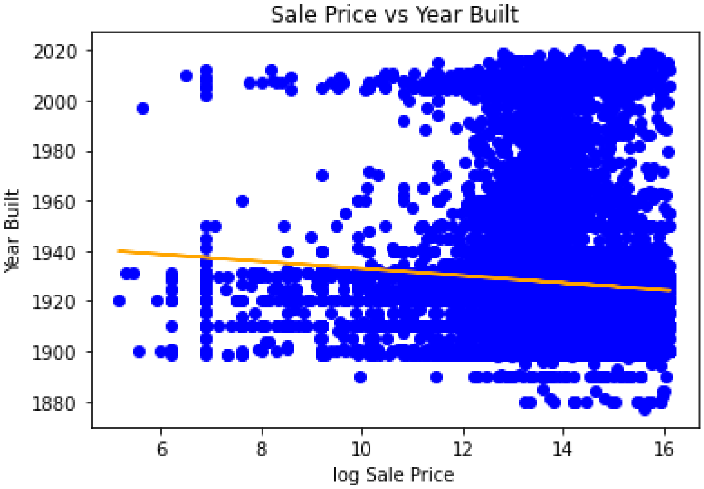




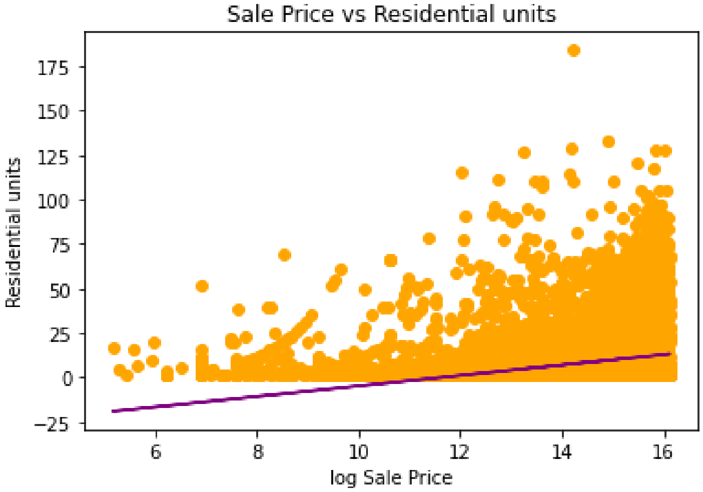




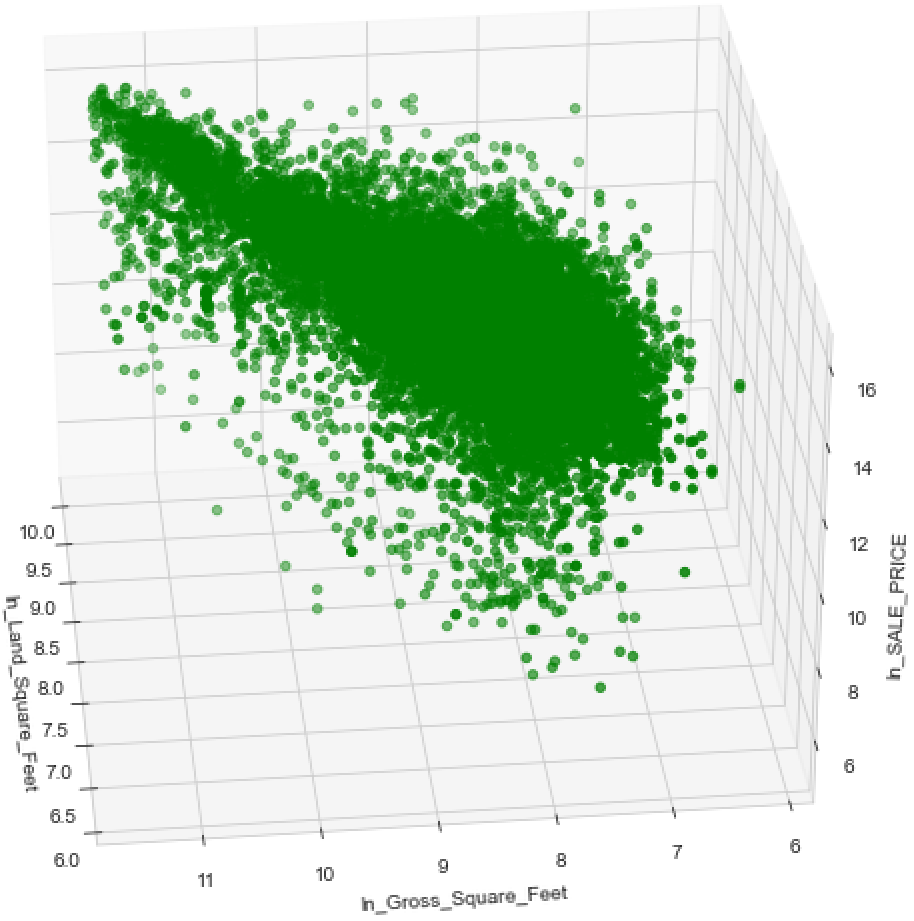
Sale Price vs Commercial Units



Sale Price vs Year Built



Sale Price vs Residential units


10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
ln_Land_Square_Feet
11
10
9
8
7
6
ln_Gross_Square_Feet
16
14
12
10
8
6
ln_SALE_PRICE