NYC Residential Real Estate Sales: A Brief Data Analysis

Quentin McTeer

05/02/2021

Contents

Origins and Background

This analytic report starts with a raw dataset of real estate transactions taken from the New York City Department of Finance. Specifically, the data includes 84,548 individual building and building unit sale transactions between September 2016 and September 2017 across all five of New York City's boroughs. There are 22 variables in the original dataset that describe the characteristics of each transaction (i.e. location of property, price paid, number of residential/commercial units included in sale, date sold). This information is compiled by the NYC Department of Finance to inform real estate investors and creditors within the NYC market, as well as to provide transparency and display critical economic information as a matter of public record. Additionally, the Department of Finance is required by law to assess property values for buildings and building units that are up for sale, so this data is already collected by the department for bureaucratic reasons. Given that the data reports the day of sale for each building and building unit over one year, it is time-series data.

The data is not a sample and represents the entire population of building and building unit sales over the September 2016 to September 2017 time period. This is only a subset of the years that the NYC Finance Department has collected, which spans 2003-2019. Each real estate transaction must be reported to the local government of New York City, making data collection purely an administrative task of aggregating each transaction into a csv, excel file, and Adobe PDF. For the purposes of this analysis, the original data was downloaded as an excel file. A codebook for the data is hyperlinked on the same page as the downloadable data files. The codebook includes a glossary of terms for each of the variables as well as important details relating to the values of each variables. The most interesting fact about the dataset, as mentioned in the codebook, is that while most sales prices posted for each building and building unit range in the millions of dollars, a large minority of the values are between 0-1000 US dollars. While these values may seem non-sensical to the naked eye, it is actually due to the fact that that there are transfers of ownership without a cash consideration or with a small cash consideration. There can be a number of reasons for such a small or non-existent sale price, including transfers of ownership from parents to children.

Raw Data Description

As mentioned above, the raw dataset includes 22 variables. The first variable "id" takes a unique value for each building and building unit sale. The second variable "borough" takes a value of 1-5 depending on the borough in which the property is located (Manhattan, Bronx, Brooklyn, Queens, and Staten Island). The third variable "neighborhood" reports the name of the neighborhood in which the property is located. The next variable "building class category" provides one type of sorting mechanism to identify the type of property sold. The fifth variable "tax class at present" and "tax class at sale" report the tax type (one of four different classes as defined by the government of New York City) that is assigned to the property in question. The "block" and "lot" variable distinguish one unit of real property from another, such as the different condominiums in a single building. Also, "block" and "lot" are

not subject to name changes based on which side of the parcel the building puts its entrance on. The "easement" variable is a logical variable that describes if the property has an easement, which allows an entity to make limited use of another's real property. For example: MTA railroad tracks that run across a portion of another property.

The "building class at present" and "building class at sale" variables are used to describe a property's constructive use. The first position of the building class is a letter that is used to describe a general class of properties (for example "A" signifies one-family homes, "O" signifies office buildings. "R" signifies condominiums). The second position, a number, adds more specific information about the property's use or construction style (using our previous examples "A0" is a Cape Cod style one family home, "O4" is a tower type office building and "R5" is a commercial condominium unit). The "address" variable describes the street address of the property as listed in the dataset. The "zip code" variable is simply the the property's postal code. The "residential units" variable takes values based on the number of residential units at the listed property. Similarly, the "commercial units" variable takes values based on the number of commercial units at the listed property. A final unit variable, "total units" takes values based on the total number of units at the listed property. Two other variables "land per square feet" and "gross square feet" report the land area of the property listed in square feet and the total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building, including the land area and space within any building or structure on the property, respectively. The "year built" variable takes values based on the year the structure of the property was built. The final two variables "price" and "date" show the price paid for the property and the date the property sold, respectively.

The "building class category" and "building class sale" variables, being factors with a significant amount of categories (100+), need to be reduced to only a few main categories along with an "other" category to lump together the many building types that do not show up frequently in the dataset. This will help immensely with the analysis of the data and is a common data cleaning technique in the data science world. Additionally, an indicator variable for whether or not a property sale was completed as a deed transfer with little or no cash consideration will be critical to this report as well. The "date" variable will need to be transformed to create a 2016 and 2017 indicator variable as well as a variable showing the month in which the property sold. There are approximately 140,000 missing values in the raw data, this does not include a few columns in the dataset which use the symbol "-" to denote a missing value instead of NA. Additionally, the "year built" column includes values of zero that actually represent missing data points. Data transformation will be required in both cases to get the true total for missing values in the dataset. There do not appear to be any significant outliers in the raw data, though an analysis using histograms reveals that many of the variables are not normally distributed and are heavily skewed both left and right. There are no issues (including labelling problems) that are a result of the methods used to import the data. The only other problem encountered during the initial review of the raw data relates to the structure of certain variables. A few columns (building class category, building class sale, borough, and tax class sale) are miscategorized as numeric or character

data when they should be classified as factor structured. A few other columns (land per square feet, gross square feet, and price) are miscategorized as character data when they should classified as numeric data. This will need to be adjusted in the analytic dataset construction.

Getting an Analytic Dataset

There were a few significant challenges in generating the analytic dataset. First, a few of the variables have so many missing values that imputation would have been inaccurate and omitting the missing values would have eliminated most rows in the dataset. Because of this, the "easement" and "apartment number" columns were eliminated from the data set. Additionally, the "lot", "building class present", "tax class present", "block", and "zip" variables will not add any value in the analytic dataset given the other columns in the raw data identify location, tax class, and building type in a more useful way. Thus, they are not included in the analytic dataset either. Second, in order to identify properties that were transferred from one individual to another (and therefore have a non-sensical sales price value), an indicator variable was created to determine if the property was actually sold at or near market value, or if it was simply a deed transfer. This indicator variable was established by coding all sale transactions as deed transfers if the sale price was less than 1500 USD. This resulted in approximately 10,000 sale transactions be coded as deed transfers, and the rest actual property sales with reasonable cash consideration. Third, rows that only included the symbol "-" or, in the case of the "year built" column, contained a value of "0" were redefined as "NA" or missing values. After all true missing values were accounted for, they were omitted from the analytic dataset. Omission had to be done because imputation (including predictive mean matching, linear regression, mean/median etc.) would not have resulted in accurate forecasts of the missing values due to the skewed sale price variable values that included deed transfers. Finally, the data was filtered on the "building class category" variable in order to examine residential complexes/units only. However, even after omitting all of these values, there were still 39,199 complete observations in the dataset.

The borough factor variable was used in two different ways to create the analytic dataset. First, to create a variable that is common across observations. To do this, the average gross square footage of building and building units sold was calculated by borough. These five values were then moved into their own object, and promptly merged back into the data on the borough column in order to add average square footage by borough to the analytic dataset. Second, the borough column was utilized to merge in new variables from an external data source. This second external dataset includes 2017 median income by borough, 2017 violent crime rates by borough, and percent of population that is white (non-hispanic) by borough. This demographic data was sourced from three different institutions: the NYC Mayor's Office of Criminal Justice, the Zicklin School of Business at Baruch College, and the NYC Department of City Planning. The external dataset contained. This data for the external source was compiled by the author of this analysis. It contains 4 variables and 5 observations (one for each borough). This external datasource was then

merged into the original dataset, once again using the borough variable, in order to generate the analytic dataset.

Exploratory Data Analysis

The new analytic dataset contains only 39,199 observations and 23 features. All variables are reported completely with no missing values. Column structure is now correct, with all relevant variables correctly classified accurately as either "numeric", "factor", "character", or "date" data. By filtering out all of the commerical transactions from the original data set, the data is now focused on residential units exclusively, including apartments, condos, and family dwellings. This will allow for regression and machine learning prediction models that attempt to predict prices of residential real estate in New York City. The new dataset now includes demographic data by borough, including percent white, mean income, and crime rates. Additionally, it includes an indicator variable for sale transactions that were deed transfers, which can be used to filter out non-sensical sale price values that are less than 1500 USD. This indicator variable can also be used as a dependent variable in a machine learning model, where the algorithm attempts to predict if a transaction was a deed transfer depending the other characteristics of the transaction.

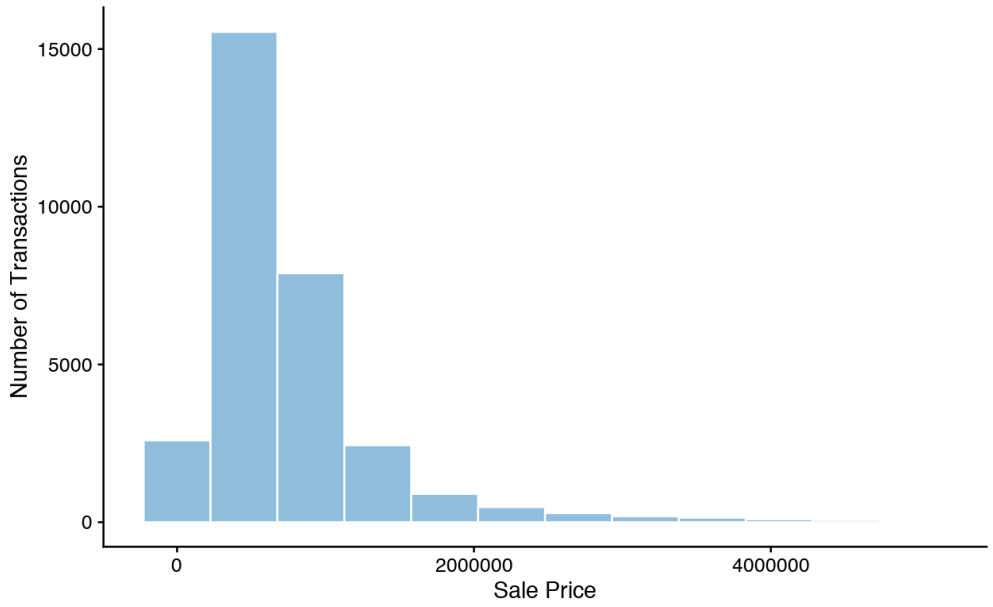

Figure 1: Residential Building Sale Price Distribution

The bar chart above (Figure 1) plots the distribution of one of the main dependent variables of interest in this analysis, residential property sale price. The results of the histogram show that most properties in the analytic dataset sold at a price between 500,000 and 2,000,000 USD, with a slightly skewed left normal distribution. The below summary statistics table (Figure 2) segments the analytic dataset by residential transactions that were deed transfers versus those were not. The second column shows that the average property in non-deed transfer transactions sold for 906,474.84 USD, while those were deed transfers only sold for 10.02 USD. Another major insight from

this table relates to the distribution of deed transfers by borough. Here, it is noteworthy to point out that most deed-transfers take place in Brooklyn while almost none occur in Manhattan relative to non-deed transfer property sales. A final insight that can be drawn from this figure is that, while the land square footage and gross square footage averages are quite similar, the variation is significantly higher for non-deed transfer transactions (24,218.32 and 22,722.42) than it is for deed transfer transactions (4,410.46 and 6,769.92).

Characteristic	FALSE, N = 31,180	TRUE, N = 8,19
resid_units	2.16 (13.44)	2.98 (14.25)
commer_units	0.04 (0.32)	0.06 (0.37)
ttl_units	2.20 (13.62)	3.04 (14.39)
lnd_sqft	2,790.99 (24,218.32)	2,509.28 (4,410.46)
gross_sqft	2,291.28 (22,722.42)	2,904.19 (6,769.92)
year_built	1,944.70 (32.62)	1,935.66 (30.96)
price	906,474.84 (1,866,409.55)	10.02 (82.20)
borough		
Bronx	4,164 (13%)	1,404 (18%)
Brooklyn	11,323 (36%)	6,200 (77%)
Manhattan	570 (1.8%)	25 (0.3%)
Queens	10,318 (33%)	270 (3.4%)
Staten Island	4,805 (15%)	120 (1.5%)
tax_clss_sale		
1	24,747 (79%)	6,342 (79%)
2	6,433 (21%)	1,677 (21%)
4	0 (0%)	0 (0%)

Mean (SD); n (%)

Figure 2: Property Sale Versus Deed Transfer Summary Statistics

The final summary statistics table (Figure 3) below break down various residential property sale variables by borough/location. The data used in this table does not include deed transfer transactions. Most building and building unit sales take place in Brooklyn (17,523) and Staten Island (10,588), with very few occurring in Manhattan (595). Violent crime rates are highest in the Bronx (6.7%) and lowest in Staten Island (2.2%). Median income is highest in Manhattan (85,071 USD) and Staten Island (79,201 USD). The average size of properties sold is significantly higher in Manhattan by square footage (11,743) and residential units (13.41). The boroughs with the oldest properties sold on average are located in Brookyln (1939) and in Manhattan (1910). Finally, the properties that sold for the lowest value on average were located in the Bronx (439,717) and the properties that sold for the highest value on average were located in Manhattan (6,783,857).

Characteristic	Bronx, N = 5,568	Brooklyn, N = 17,523	Manhattan, N = 595	Queens, N = 10,588	Staten Island, N = 4,925
income	37,397.00	56,942.00	85,071.00	64,509.00	79,201.00
crime	6.70	4.20	4.10	3.00	2.20
white	10.90	35.70	48.00	27.60	64.00
resid_units	3.19	2.10	13.41	2.07	1.37
commer_units	0.05	0.05	0.45	0.03	0.01
ttl_units	3.24	2.14	13.86	2.09	1.38
lnd_sqft	2,636.51	1,850.02	2,816.53	3,645.03	4,015.76
gross_sqft	2,945.74	2,121.26	11,743.28	2,406.75	1,764.09
year_built	1,943.38	1,938.81	1,910.47	1,939.33	1,968.10
price	439,717.19	676,340.32	6,783,857.20	702,944.49	504,555.89
Mean					

Figure 3: Property Sale Feature Summary Statistics By Borough

A correlation matrix was built to analyze the relationships between all of the continuous variables within the analytic dataset. The pearson correlation coefficients, shown in Figure 4, create a few critical insights relating to the relationships between the various characteristics of NYC residential real estate transactions. Those include the correlation between median income and violent crime rate (-0.938), median income and percent white (0.844), gross square feet and residential units (0.873), gross square feet and total units (0.872), and land in square feet and gross square feet (0.951). Because of these very strong relationships between these variables, some of them may need to be thrown out in the modelling portion of this analysis in order to deal with the probelem of multi-collinearity. One other relationship, residential units and total units (1), had perfect collinearity which means one will need to be thrown out in order to effectively run a regression or machine learning analysis to predict property sales prices.

Figure 4: Pearson Correlation Coefficients for Continuous Variables

	income	crime	white	resid_units	commer_units	ttl_units	lnd_sqft	gross_sqft	price
income	1.000	-0.938	0.844	-0.005	0.008	-0.005	0.023	0.003	0.136
crime	-0.938	1.000	-0.732	0.034	0.036	0.035	-0.023	0.013	-0.022
white	0.844	-0.732	1.000	-0.020	-0.005	-0.019	0.009	-0.008	0.057
resid_units	-0.005	0.034	-0.020	1.000	0.512	1.000	0.764	0.873	0.138
commer_units	0.008	0.036	-0.005	0.512	1.000	0.529	0.372	0.432	0.200
ttl_units	-0.005	0.035	-0.019	1.000	0.529	1.000	0.764	0.872	0.141
lnd_sqft	0.023	-0.023	0.009	0.764	0.372	0.764	1.000	0.951	0.016
gross_sqft	0.003	0.013	-0.008	0.873	0.432	0.872	0.951	1.000	0.062
price	0.136	-0.022	0.057	0.138	0.200	0.141	0.016	0.062	1.000

Since the paper is only evaluating residential properties, it is important to break down the various forms of residential transactions within the dataset. Figure 4 below shows the average sale price of six different types of residential properties (non-deed transfer). Rental walk up apartments sell for the highest amount, over three million dollars,

while elevator apartment cooperatives sell for the least, roughly four hundred thousand dollars on average. Family dwellings sell for roughly half a million dollars, increasing slightly depending on the number of dwellings on the property (one, two, and three respectively).

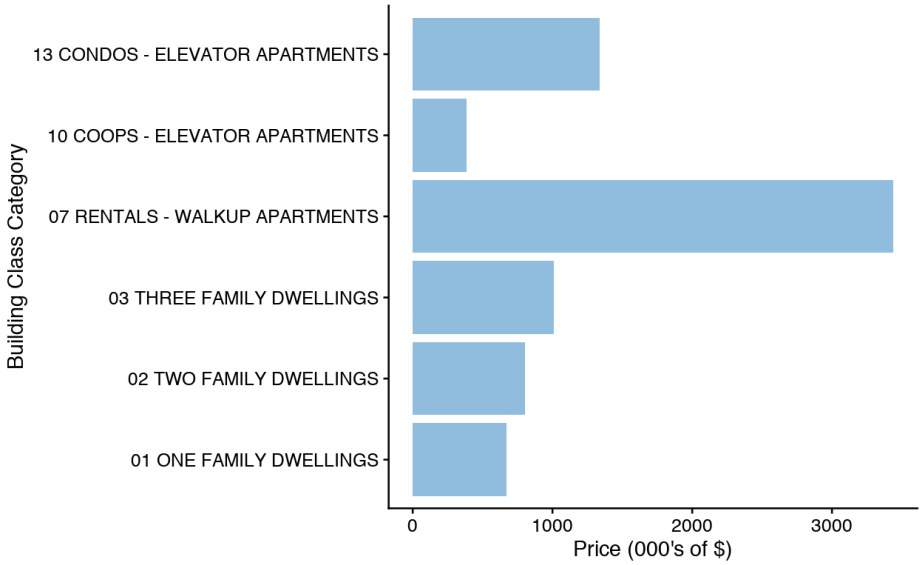

Figure 4: Average Sale Price By Residential Building Type

Since the data is time series (September 2016-September 2017), there may be a seasonality component to the data that could be visualized graphically. Figure 5 below shows the number of residential sales transactions (deed transfer and non-deed transfer) that occur every single month, with a blue vertical line at September to indicate the beginning of the time series (September 2016). June, September, and December see the highest activity in the NYC residential real estate market, while February, April, and August see the lowest activity. Interestingly, August sees almost no residential real estate sales at all relative to other months.

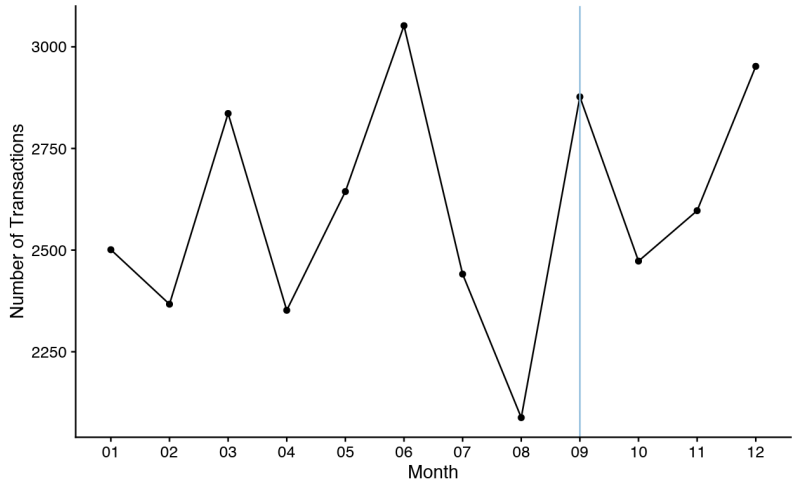

Figure 5: Seasonal Trends in New York City Real Estate Sales

There are two tax classes within the residential sale transactions included in the analytic dataset. The first class (class 1) includes most residential property of up to three units (such as one-, two-, and three-family homes and small stores or offices with one or two attached apartments), vacant land that is zoned for residential use, and most condominiums that are not more than three stories. The second tax class includes all other property that is primarily residential, such as cooperatives and condominiums. Given that the second class includes cooperatives and condominiums with potentially large numbers of residential units it makes sense that, in Figure 6 below, the average sale price for tax class 2 properties was roughly 1.2 million dollars while the average sale price for tax class 1 properties was only six hundred thousand dollars.

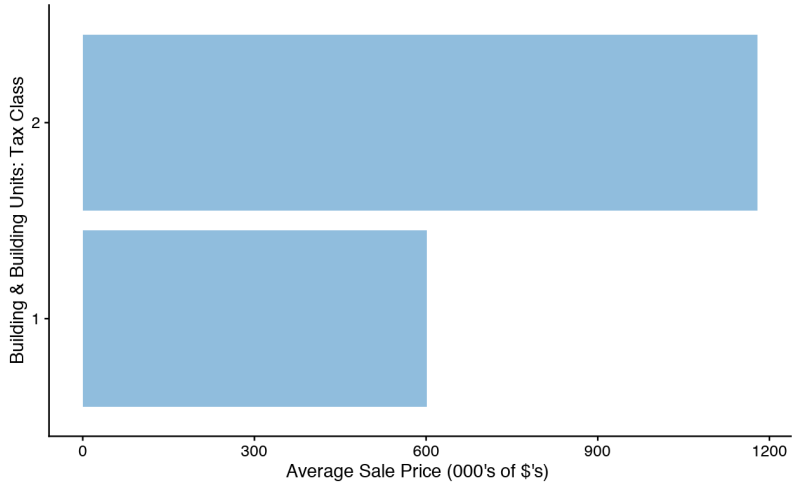

Figure 6: Relationship Between Building Tax Class and Sale Price

In addition to residential units and tax class, the size of the building/building unit also plays a role in the sale price of the property. Figure 7 shows a strong correlation between gross square footage (the metric for building size) mapped against its corresponding sale price. Most properties cluster between 0 and 4000 gross square feet, while a few properties are spread out over 5000 gross square feet. The blue line represents a goodness a fit calculation from a simple regression of gross square feet on residential property sale price. Clearly, larger buildings (in terms of square feet) are sold for higher prices than smaller buildings, but it could be interesting to see if this relationship holds up as a causal effect in a multivariate regression, with more confounding variables controlled for.

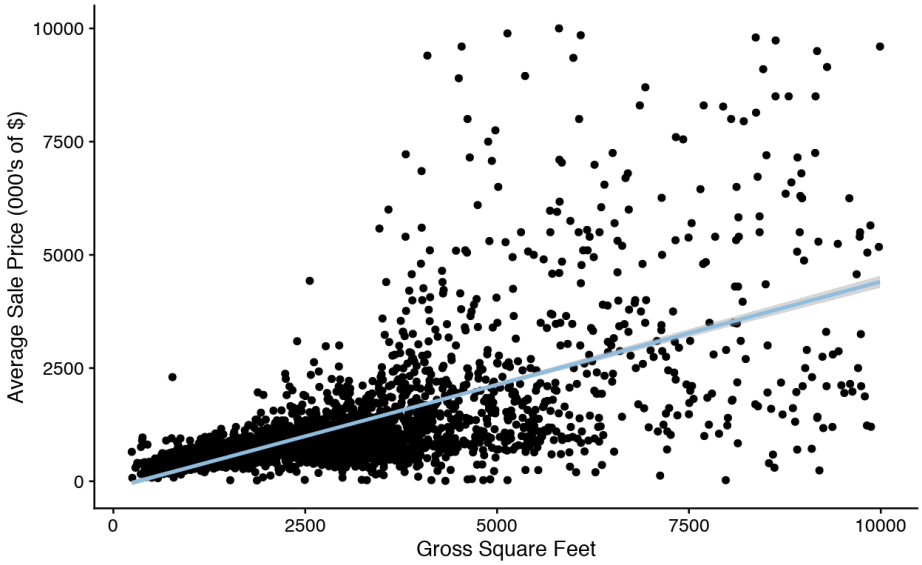

Figure 7: Impact of Square Footage on Sale Price

A final key relationship within the analytic dataset is the impact of age on residential property sale price. In Figure 8 below, the average sale price is mapped against the year the buildings were constructed. There are few key takeaways from this visualization. First, most buildings were established between 1950 and 2017, and very few were created before 1900. Although the values in the scatter plot are averages, implying that there may be many individual observations represented in each plot point, it is logical to assume, given the nature of the NYC real estate market, that the density of plots here mirror the density of the total set of observations of residential real estate transactions in the entire analytic dataset. In addition to the distribution of the points across the year the building was constructed, the blue goodness of fit simple linear regression line tells us that there is a negative relationship between sale price and year built, implying that older buildings sell for higher values. However, the scatter plot also seems to show a non-linear relationship between sale price and year built. With properties up until 1950 displaying a negative relationship between sale price and year, and properties built after 1950 displaying a positive relationship between sale price and year built. This insight will be used in constructing a regression model to find the impact of property age on residential building sale price.

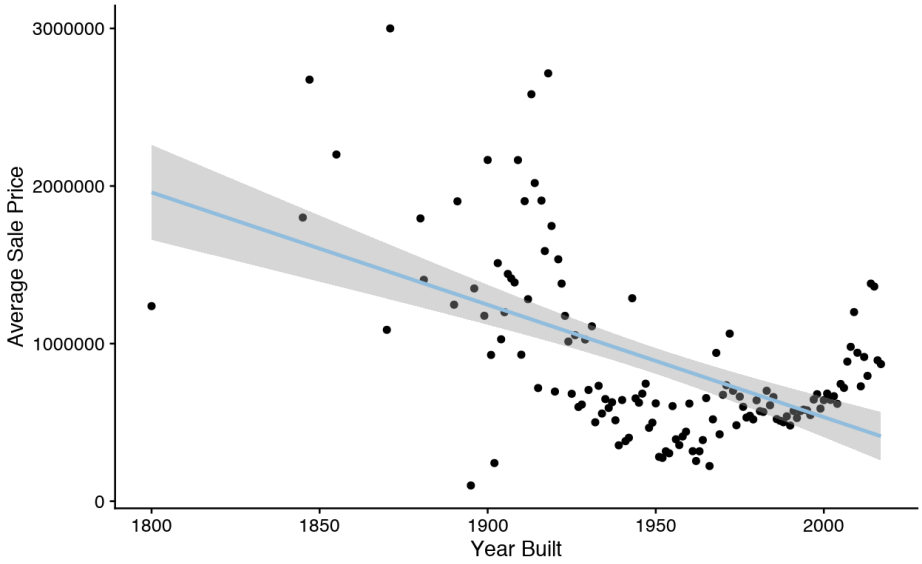

Figure 8: Impact of Building Age on Sales Price

The below table (Figure 9) shows the coefficients for OLS regression model with the sales price feature as the dependent variable. Deed transfer transactions were filtered out of the analytic dataset in order to get accurate results for the model. The model includes borough, residence units, commercial units, and land in square feet as control variables. Not shown in the model are fixed effects controls for month, building class category, and building class type at sale. The month fixed effects should solve any issues with auto correlation. Other variables from the analytic dataset not included in the model, such as violent crime rates, median income, and percent white were not included because of multi-collinearity issues that would have significantly biased the model. White robust standard errors were used to control for any issues surrounding heteroskedasticity. The important results from this model come from the two explanatory variables year built, and year built-squared (the final two rows in Figure 9). The coefficient on the year_built variable show that a one year increase in year built (or a one year decrease in age) reduces the value of the property by 292,108.2 dollars, ceteris paribus. The coefficient on the year_built_squared variable shows that the relationship is non_linear, with a one year increase in year built (or one year decrease in age) increasing the coefficient on the year_built variable by 74.8 dollars, ceteris paribus. To calculate the turning point, in which the coefficient on year_built turns possible, the formula used is $|\beta_1/(2 * \beta_2)|$. Using the formula, the year built value at which the coefficient on year_built turns positive is 1953.

term	estimate	std.error	statistic	p.value
boroughBrooklyn	359711.946	20454.385	17.586	0.000
boroughManhattan	5033305.026	382453.330	13.161	0.000
boroughQueens	160281.950	19894.179	8.057	0.000
boroughStaten Island	5280.537	21903.314	0.241	0.809
resid_units	28080.690	11091.313	2.532	0.011
commer_units	769178.568	165658.786	4.643	0.000
lnd_sqft	-15.737	4.920	-3.198	0.001
year_built	-292108.195	34270.854	-8.524	0.000
I(year_built^2)	74.799	8.755	8.544	0.000

Figure 9: Residential Sale Price Prediction Results

Linear regression can provide insights into both causal inference (in the case of this paper, the effect of year_built on residential sale price) as well as prediction (in thise case, predicting the sale price of each residential property based on the other known characteristics in the dataset). Deed transfer transactions were filtered out of the analytic dataset in order to get accurate results for both models. In the below table, (Figure 10) the R-Squared as well as the Root Mean Squared Error (RMSE) are shown for the previously run regression a well as for the machine learning algorithm Random Forest, which is a decision tree ensemble method. Ensemble methods use a single base learning algorithm to produce homogeneous base learners, learners of the same type, leading to homogeneous ensembles. In random forests, each tree in the ensemble is built from a sample drawn with replacement from the training data set. There are a few hyper parameters that must be defined in the random forest model: number of random variables to consider in each tree and minimum number of data points in a node that are required for the node to be split further. For this analysis, those values were 8, 1000, and 8 respectively. To run the model, the analytic dataset was split into a training dataset and a testing dataset. The random forest was "trained" on the training dataset where the sale price values were known, and then was tested on the test dataset where the sale price values were unknown. The results from predicting the residential building sale price is shown below through both R-Squared and RMSE. The random forest model significantly outperformed the linear regression with an R-Squared of 0.503 and an RMSE of 1,121,308.9 relative to the linear regression which only showed a low R-Squared of 0.298 and a high RMSE of 1,563,564. Clearly, the random forest did a better job of predicting the sale price of residential building/building unit then the linear regression did.

Model	R-Squared	RMSE
Linear Regression	0.298	1563564
Random Forest (Ensemble)	0.503	1121309

Figure 10: Residential Sale Price Prediction Results

The final analysis of the paper focuses on the previously omitted deed transfer transactions. Although they were not included in the above analysis in order to accurately predict price, they are included here in their entirety. Additionally, the deed_trans (deed transfer) indicator variable is now the dependent variable, and two different models were employed in order to classify which transactions were deed transfers and which transactions were regular sales based on the other charachteristics of the property sales included in the analytic dataset. The classification models included for this binary dependent variable prediction were logistic regression and random forest. For classification problems, some of the best metrics to evaluate model performance include accuracy and AUC, with higher values indicating better classification performance in both cases. Figure 11 shows that both the logistic regression and the random forest models had similar classification success, with accuracy metrics of roughly 80% and AUC values of 0.8 as well. This means that, approximately 80% of the time, both models correctly classify a real estate transaction as a deed transfer or not a deed transfer. Given that roughly 75% of the data in the analytic dataset represents non-deed transfer transactions, and 25% represent deed transfer transactions, it is likely the models successfully classified the non-deed transfer transactions but failed to accurately classify the deed transfer transactions. The logical conclusion, then, is that there is no real difference, in terms of characteristics, of residential real estate transactions that are deed transfers versus those that are not. This would explain why neither the random forest nor the logistic regression could accurately categorize the deed transfer transactions. Most likely, both models categorized every single transaction as non-deed transfer, leading to 80% accuracy and 0.8 AUC, since that is how the data is distributed.

Model	Accuracy	AUC
Logistic Regression	0.8003	0.795
Random Forest (Ensemble)	0.804	0.802

Figure 11: Deed Transfer Classification Prediction Results