NYC Property Sale Price Prediction Final Report

Yi Wen (yw5280) Congyun Jin(cj2164) Fan Zhang(fz2068)
Siyu Shen(ss14359) Zhuoyuan Xu (zx1137)

Problem Statement

The housing market of New York City has been in up-and-downs in the last several decades due to economic cycles. While the macroeconomic reasons have contributed to the NYC house pricings, the intrinsic characteristics of a house might also contribute to the pricing of a house. For house investors and buyers to make purchasing decisions wisely, it might be useful to see what the driving characteristics of an expensive house in the United States are. In this project, our group aims to predict the prices of the house listings in the United States, using selected features of the close-by neighborhoods. We will perform regression analysis and optimize our chosen model to the best accuracy with feature selection, model tuning and other techniques. Meanwhile, the resulting model will create potential benefits in multiple areas beyond basic price prediction, such as:

1) providing data-proven insights for individual house buyers and sellers;
2) enhancing a balanced leverages between the buyers and sellers;
3) understanding the housing market in general for economists, policy makers or interested stakeholders/ decision makers.

Data Understanding

Data for this project comes from the NYC Department of Finance[1]. The original dataset contains 167719 pieces of sales information of properties sold in New York City from January 2018 to December 2019 (over two-year period). Each data instance contains information about demographics (address, region code, neighborhood) , building information (type, number of units, building land area) , sale date etc. There are 21 features in total (details are shown in appendix A). In this project, we will not study the effect of time on sale price, hence SALE DATE will not be used to predict the sale price (target variable) of NYC property. After certain data cleaning processes in *Section 4: Data Preprocessing*, the cleaned data contain 81030 rows and 19 columns.

Exploratory Data Analysis

Target Variable (Sale Price)

The distribution of sale price from raw price is exceptionally sparse. Many sale prices occur with a nonsensically small number: $0 most commonly (note that 36% of the sale price is $0). Based on the information from the original data resource, 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. In order to deal with it, we will set a reasonable range for the sale price. Here we will remove the instances that the sale price is less than $50000 (38% of the entire data) and greater than $12M (Notice that the $12M threshold helps eliminate the 0.8% special cases). Also since the numbers are huge, we will perform log transformation. See appendix B *Figure 1* for data visualization.

Predictive Feature Analysis

The features *Borough, Neighborhood, Block, Lot, Address, Apartment Numbers, Zip Code* in our data correspond to the location of the properties. They are highly correlated with each other and we

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

kept *Borough* as the only location feature after careful consideration. There are 5 boroughs in our dataset: Brooklyn, Manhattan, Queens, Staten Island, Bronx, where Brooklyn had most data instances and Bronx had the least (see Figure 2 in Appendix B). *Block* and *Lot* represent the region and the street a property locates, respectively, and are often used together with *Borough* (called a Borough-Block-Lot location system). Similarly *Address*, *Apartment Numbers* and *Zip Code* each have 75619, 3503 and 184 unique values. All features discussed above other than *Borough* are very sparse and highly correlated with *Borough* so that we kept Borough only as the predictive feature.

Building Class Category, Building Class as of final roll 18/19 and *Building Class at Time of Sale* remark the property types whereas the latter two are sparse and are mere subdivisions of the *Building Class Category*. For model simplicity, we kept *Building Class Category* only. By looking at Figure 3 in Appendix B, different types of family dwellings and apartments with elevators are the most frequent building types, meaning that most buildings are of residential uses. By looking at Figure 4, Appendix B, it seems that certain building classes have a larger range of prices (e.g. Rentals - Apartments with Elevators), or some higher average prices in general (e.g. Luxury Hotels).

Sale Price has noticeable dependence with Year Built, as properties before 1900 generally have lower prices than ones built after 1900. See Figure 5, Appendix B.

According to Figure7(a), Appendix B, *Residential units* is correlated with our target variable *Sale Price* positively. The same pattern exists in the *commercial units*, the number of commercial units in a property and *total units* which is the sum of the former two.

By looking at the boxplot (Figure 6, Appendix B) of our transformed *sale price* against *Tax Class at Time of Sale*, there are 3 unique tax classes at time of sale. Tax class 1 is more right skewed with more high sale price outliers; tax class 2 has fewer high sale price outliers and tax class 4 has none. Tax class 1 has a smaller Interquartile Range (IQR) with the lowest median sale price; tax class 2 has a larger IQR and a higher median sale price and tax class 4 has the largest IQR and highest median sale price.

As shown in Figure 8, *land square feet* and *gross square feet* share very similar distribution. They both correlate positively with property sale price. However, we found there are cases when land square feet are small, the sale price is high. Possible explanations for such outliers might be that they lie in good geographical location or that they belong to special building classes.

Data Preprocessing

Feature Selection

Based on the discussion in Exploratory Data Analysis and correlation heatmap in Figure 7(a), Appendix B, we dropped the following columns: *Neighborhood, Address, Apartment numbers, ZIP code, Building class as of final roll 18/19, Building class at time of sale, Tax class as of final roll 18/19, Sale Date.* We also dropped *Easement* because it only contains null values.

Feature engineering

<u>*Classification*</u>*:* Based on the scatter plots of *Commercial Units vs Sale Price* and *Residential Units vs Sale Price,* the pattern is opaque and there are lots of 0s, 1s in each plot. Hence we will classify them into six groups (See Table 1, Appendix C for the specific criteria). Here we will use a new variable "UNIT CATEGORY" representing the pattern of COMMERCIAL UNITS and RESIDENTIAL UNITS.

<u>*Categorical features & One-hot encoding:*</u> In order to build models, we will use one-hot encoding to transform BOROUGH, BUILDING CLASS CATEGORY, TAX CLASS AT TIME OF SALE and UNIT CATEGORY. After one-hot encoding, we have 48889 instances with 64 columns, which is a little sparse. We will first build models and see the performance.

Numerical feature - Rescaling: Based on the density plots and boxplots of *Sale Price*, *Land Square Feet* and *Gross Square Feet*, the distribution is heavily right skewed and sparsely allocated. Hence we will perform the log transformation on these three features (See Figure 1 and Figure 2, Appendix C).

Model Results

Model	**MSE**	R^2
Linear	0.2840	0.5301
Lasso	0.4296	0.2894
Ridge	0.2840	0.5302
Robust	0.3088	0.4891
Elastic Net	0.4180	0.3086

Discussion and Recommendation

Model Assumption Check and Improvement (see Appendix F)

Linear regression works by essentially fitting a (straight) line of best fit through your data. Fitting lines to non-linear data will result in different levels of overprediction and underprediction. In this project, we assume that the target variable sales price has an expected linear relationship between various variables. But to capture the true structure of this data, it is recommended to fit a polynomial curve to our data. One improvement would be engineering new features into functions of existing input variables (including powers, logs, and products of pairs of variables).

In terms of multicollinearity, it can make it difficult to interpret the model coefficients, and to determine their statistical significance since the model splits the impact of one variable across two separate input variables. To avoid multicollinearity in this project, a collinearity heatmap is plotted (see Figure 7, Appendix B). It is not hard to notice that there is a strong collinearity between commercial units and total units. A simplest solution would be just drop one of them. In this project, in order to make full use of the units information, commercial units and residential units are binned into one column "units_category" with six different values, followed by a one-hot encoding of this variable.

Furthermore, it's clear that the spread of the points on the sales price scatterplot is highly skewed to the left and has a long tail, which is a clear sign of heteroskedasticity. The normality assumption is held after we log transform the target variable "sale_price"(see Figure1, Appendix C). For the further improvement, one possible solution would be to use a time series model to deal with autocorrelation. Since the property sales price has some kind of relationship with stock price, it is better to combine S&P 500dataset with our existing data and use ARIMA time series forecasting to model it.

Feature Selection

The features in the datasets of interest contain information in five major categories: location, building category, tax class, time and house area. Further exploration on the feature selection process could be conducted on some of these categories to improve the model performance and generalization ability. Location information is embedded in the Borough, Block, Lot, Address, Apartment Number and Zip Code features. The first three features together form the Borough-Block-Lot (BBL) address system, while the other three together are the common mailing address. In our study, we focused on the BBL system and zip code to generate relatively broad spatial groupings. Suggestions from other[2] studies include using longitude and latitude from the addresses to create more precise spatial scales and add functionality-related spatial categorizations before the feature selection. For time information, not only the Sale Date, but also some indicators of economic conditions corresponding to the dates could be added to generate more interpretability to the price change. The house area category includes Total Unit, Commercial and Residential Unit, Land Square Feet and Gross Square Feet. The commercial unit and residential unit can be further explored if we can design and apply reasonable binning and weighting methods with relevant information from other data sources.

Conclusion

Overall, Ridge Regression outperforms all models with an R^2 of 0.5302 and an MSE of 0.2840. The 5 most predictive features selected by ridge regression are whether or not the property

1) is a *special condo billing lots,*
2) is a *luxury hotel,*
3) has a number of commercial units that is categorized as *unit category A* (commercial units > 10),
4) is a *commercial vacant land*, or
5) is located in *Borough 1 (Manhattan)*

Other predictive features ranked by feature importance may be found in Appendix G.

If allowed more time and appropriate data, we may incorporate the effect of COVID-19 on housing prices after 2020. We may also improve the current models by using cross validation and hyperparameter tuning. We may also check for multicollinearity after one-hot encoding because certain new features created after one-hot encoding seem to correlate to another. For example, one of the most predictive feature *luxury building*, which previously is a category under the feature *BUILDING CLASS CATEGORY*, is now a separate feature and may correlate to the *number of commercial units*.

[2] https://towardsdatascience.com/stop-using-zip-codes-for-geospatial-analysis-ceacb6e80c38

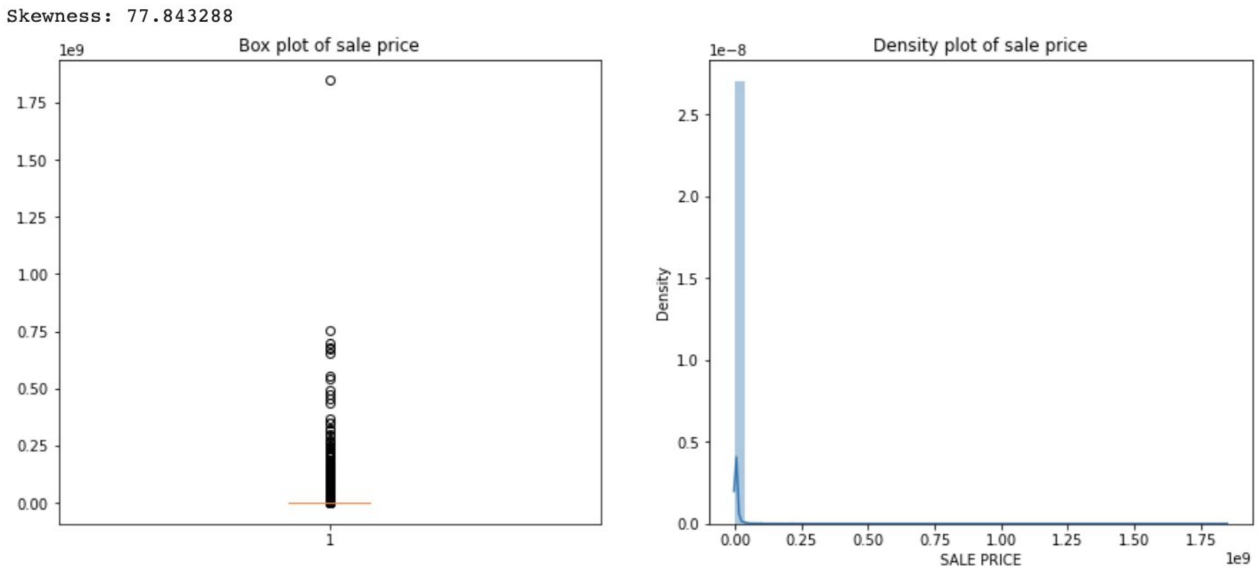
Appendices

Appendix A. Data Description

Variable	Definition	One Instance
Borough	A digit code for the borough the property is located in; in order these are Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).	1
Neighborhood	The specific neighborhood the property is located. Department of Finance assessors determine the neighborhood name in the course of valuing properties.	Alphabet City
Building Class Category	The type of the property	01 ONE FAMILY DWELLINGS
Tax Class as of Final Roll 18/19	The tax code of the property **before** transaction, includes the following: 1,2, 1A, 1B, 1C, 1D, 2A, 2B, 2C and 4. For example, class 2 properties include rental buildings, condominiums and cooperatives	2A
Block	The digital code that represents the region the property is located in, commonly used with Lot and Borough (BBL)	390
Lot	The digital code that represents the street the property is located in, commonly used with Block and Borough (BBL)	61
Easement	An easement is a legal loophole that grants an interested party the right to use another person's property or land in a certain way despite not having any ownership interest.	Nah (No records in the dataset)
Building Class As of Final Roll 18/19	The building code of the property **before** transaction, which indicates the type of building. For example, B1 indicates 'TWO FAMILY BRIC'	A1
Address	The address of the property	189 EAST 77TH STREET
Apartment Number	The apartment number of the property	556
Zip Code	The zipcode of the property	10009
Residential Units	The number of residential units the property has	2
Commercial Units	The number of commercial units the property has	1
Total Units	The sum of residential and commercial units the property has	5
Land Square Feet	The usable or assignable square footage within the property, also known as net square feet (NSF)	987
Gross Square Feet	The space occupied by the intradepartmental circulation and the walls and partitions within the property, includes the land square feet	2183

Year Built	The year the property was built	1998
Tax Class At Time of Sale	The tax code of the property **during** the transaction. The code description is the same as 'Tax Class as of Final Roll 18/19'	2A
Building Class At Time of Sale	The building code of the property **during** the transaction. The code description is the same as 'Building Class as of Final Roll 18/19'	A1
Sale Date	The specific time when the property is sold.	5/23/18
Sale Price	**The target variable.** The sale price of the property, recorded in Canadian dollars.	$100000

Appendix B. Feature Visualization



```
Skewness: 77.843288
```

Figure 1 - Distribution of building sale prices before data cleaning and log(x) transformation

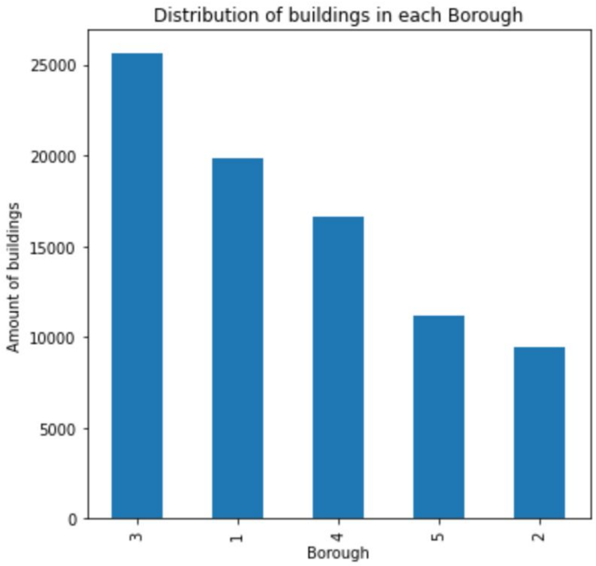


Figure 2 - Number of buildings in each Borough (1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island)

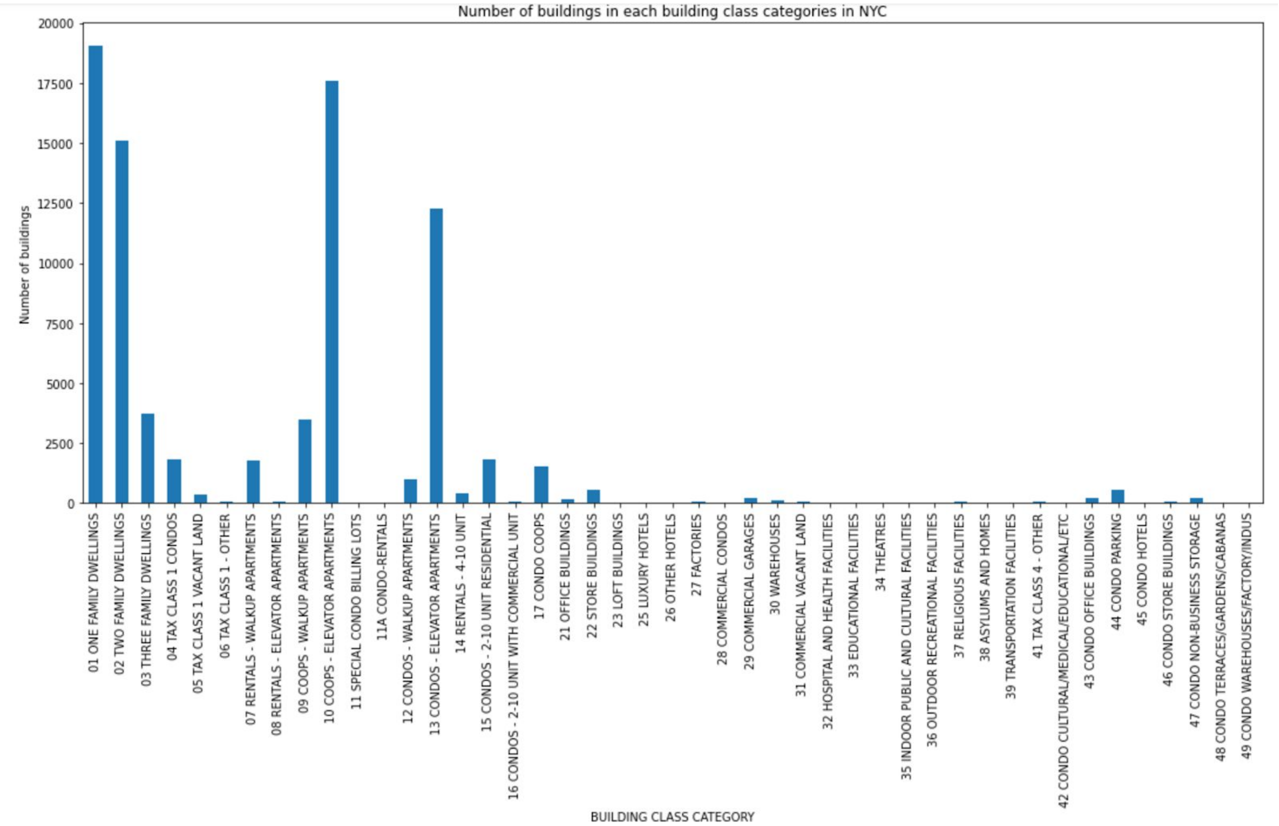


Figure 3 - Number of Buildings in each building class category

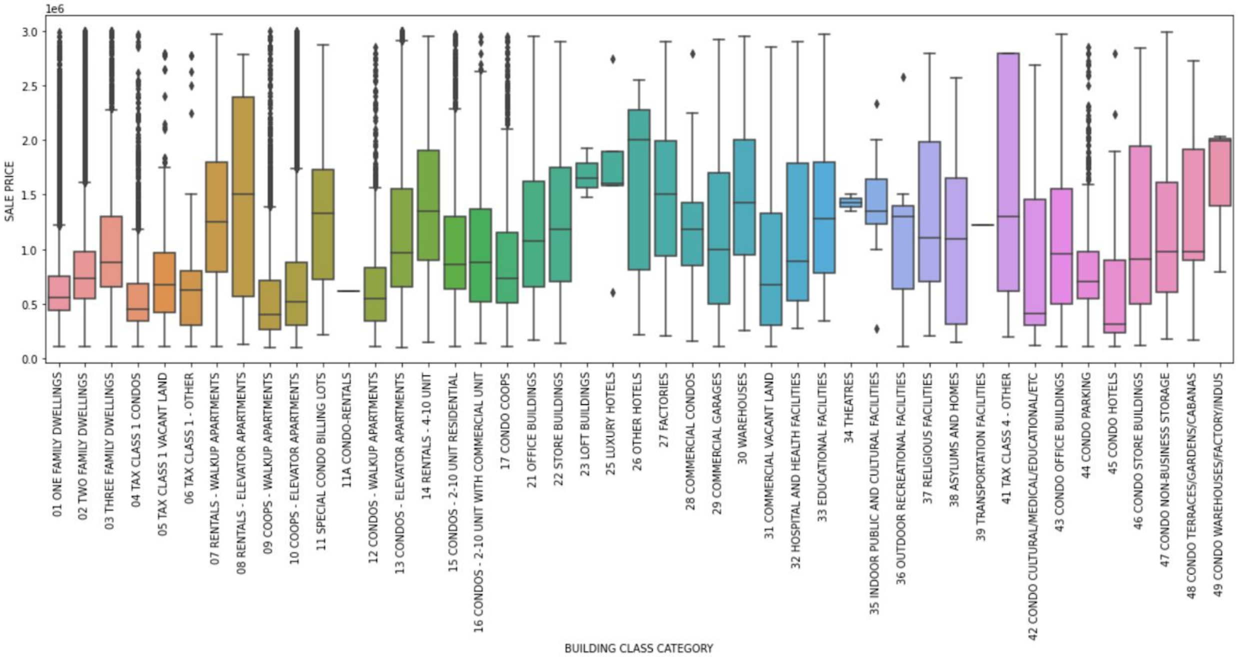


Figure 4 - Boxplot of log(Sale Price) against Building Class Category
There could be some interesting patterns between building types and their sale prices

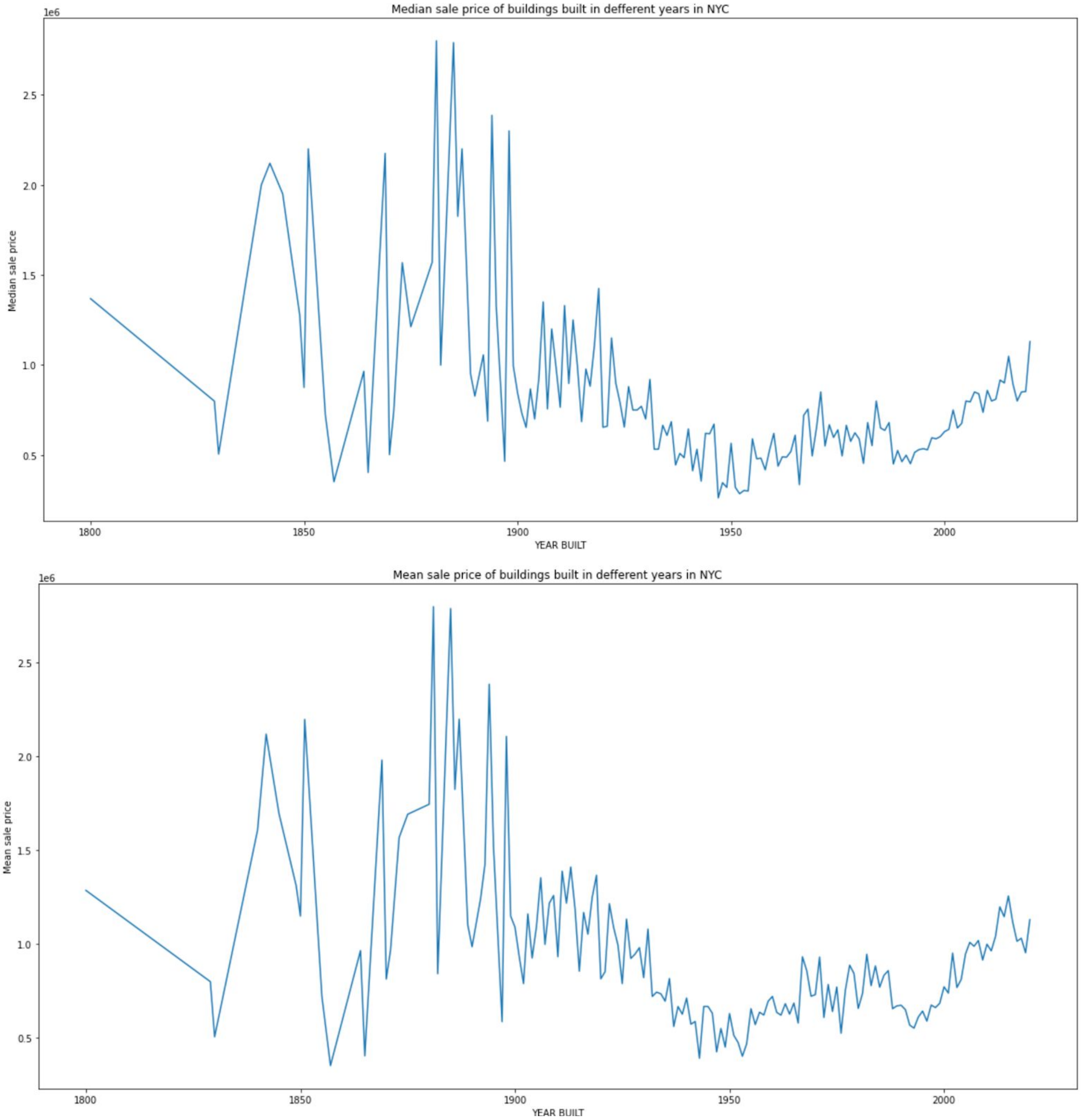


Figure 5 - Median log(Sale Price) and Mean log(1 + Sale Price) against the year in which the building is built

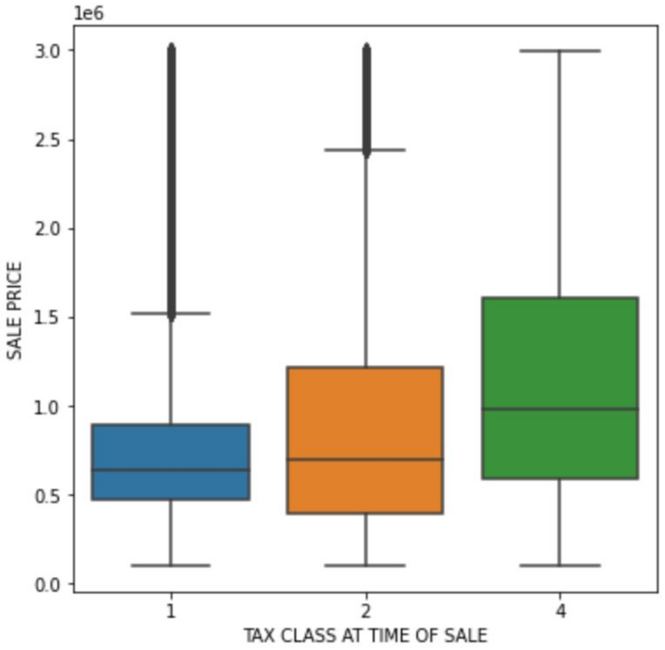


Figure 6 - Boxplot of log(1 + Sale Price) against Tax class at time of sale

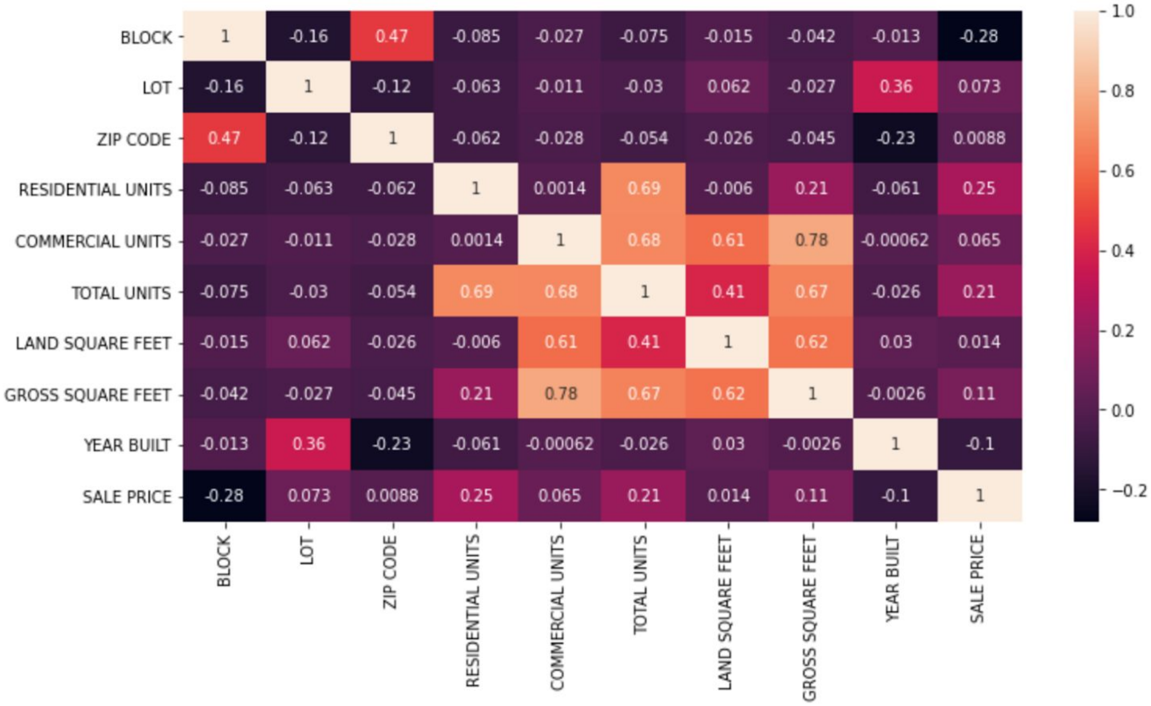


Figure 7(a) - Correlations between numerical variables before feature engineering

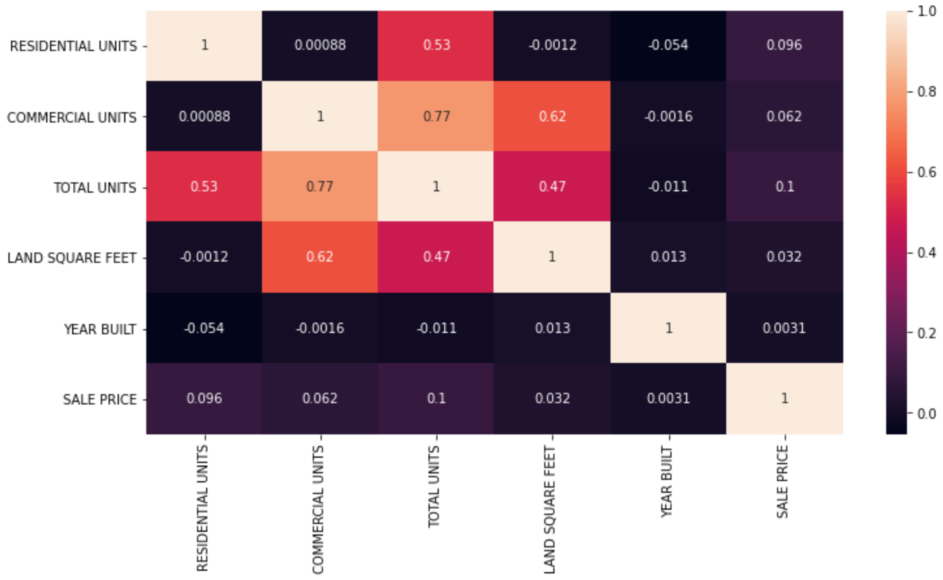


Figure 7(b) - Correlations between numerical variables after feature engineering

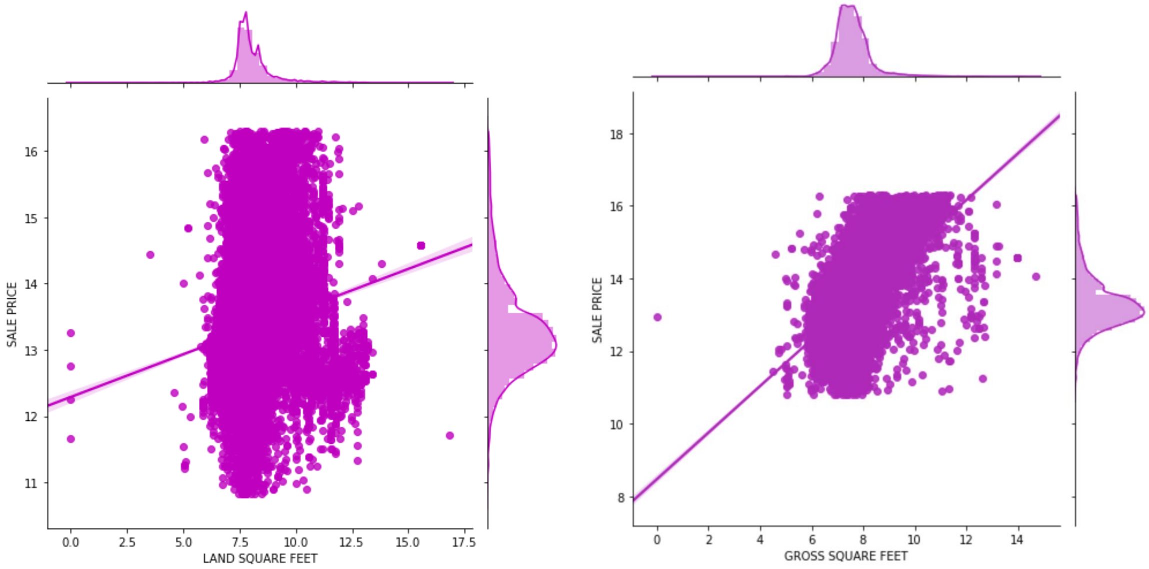


Figure 8 - relationship between property land square feet and sale price

Appendix C. Data Processing

UNIT TYPE	CRITERIA
A	Commercial Units > 10
B	0< Commercial Units <= 10
C	Commercial Units = 0 and Residential Units= 1
D	Commercial Units = 0 and 1 < Residential Units < 10
E	Commercial Units = 0 and Residential Units >= 10
F	Commercial Units = 0 and Residential Units

Table 1 - Grouping commercial units as a categorical variable

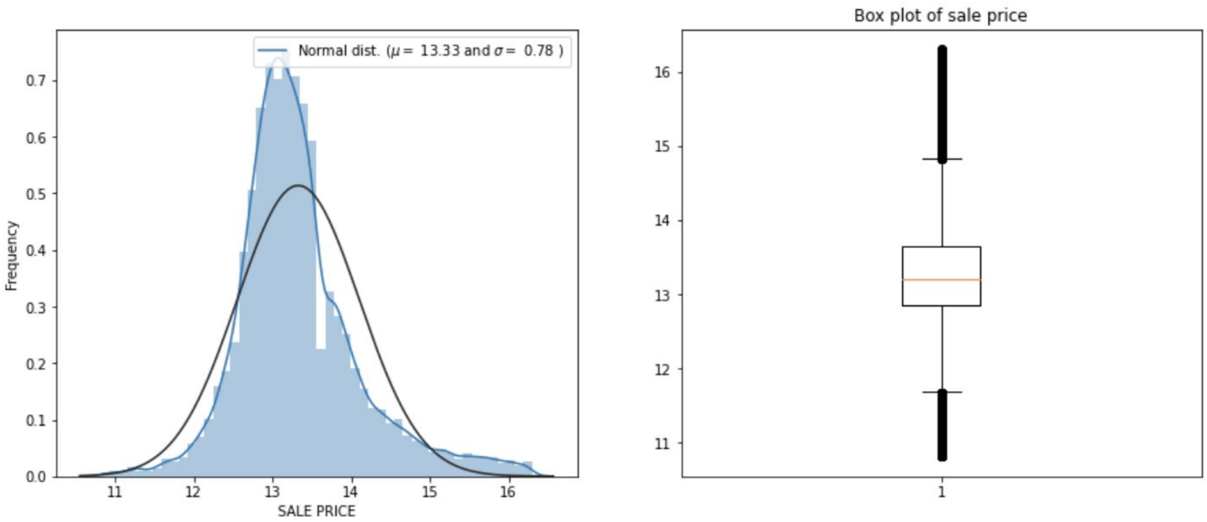


Figure 1 - Distribution of sale prices after log transformation

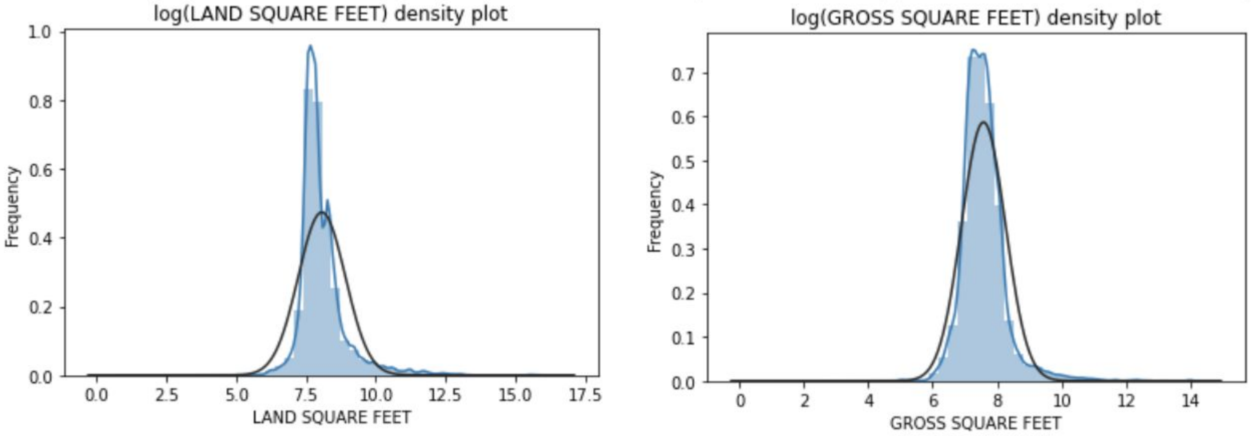


Figure 2 - Distribution plots of land square feet and gross square feet after log transformation

Appendix D. Models used mathematically

Multilinear regression:

A natural extension of the Simple Linear Regression model is the multivariate one. It is given by:

$$Y(x_1, x_2, \cdots, x_n) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_0$$

where $x_1, x_2, \cdots, x_n$ are features from our dataset and $w_1, w_2, \cdots, w_n$ are learned parameters.

Ridge Regression:

Ridge Regression adds a L2 penalty to the least square problem, which penalizes the sum of square coefficients. In this project, we used the weight on the penalty $\alpha = 0.1$. Hence, this model minimizes

$$||y - Xw||_2{}^2 + \alpha \cdot ||w||_2{}^2$$

Lasso Regression:

Lasso Regression adds a L1 penalty to the least square problem, which penalizes the sum of absolute values of the coefficients. In this project, we used the weight on the penalty $\alpha = 0.1$. Hence, this model minimizes

$$||y - Xw||_2{}^2 + \alpha \cdot ||w||_1$$

Robust Regression:

Robust regression aims to fit a regression model in the presence of corrupt data: either outliers, or error in the model. In this project we used an iterative algorithm for the robust estimation of parameters from a subset of inliers from the complete data set with a maximum number of iterations for random sample selection set to 1000.

Elastic net:
Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models.

Appendix E. Evaluation

The evaluation methods for the regression are Root Mean Squared Error, which is to measure how far our predictions are from the real house prices,

$$RMSE = J(W) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_w(x^{(i)}))^2}$$

and R^2 score, which explains how much the total variance of the dependent variable can be reduced by using the least square regression.

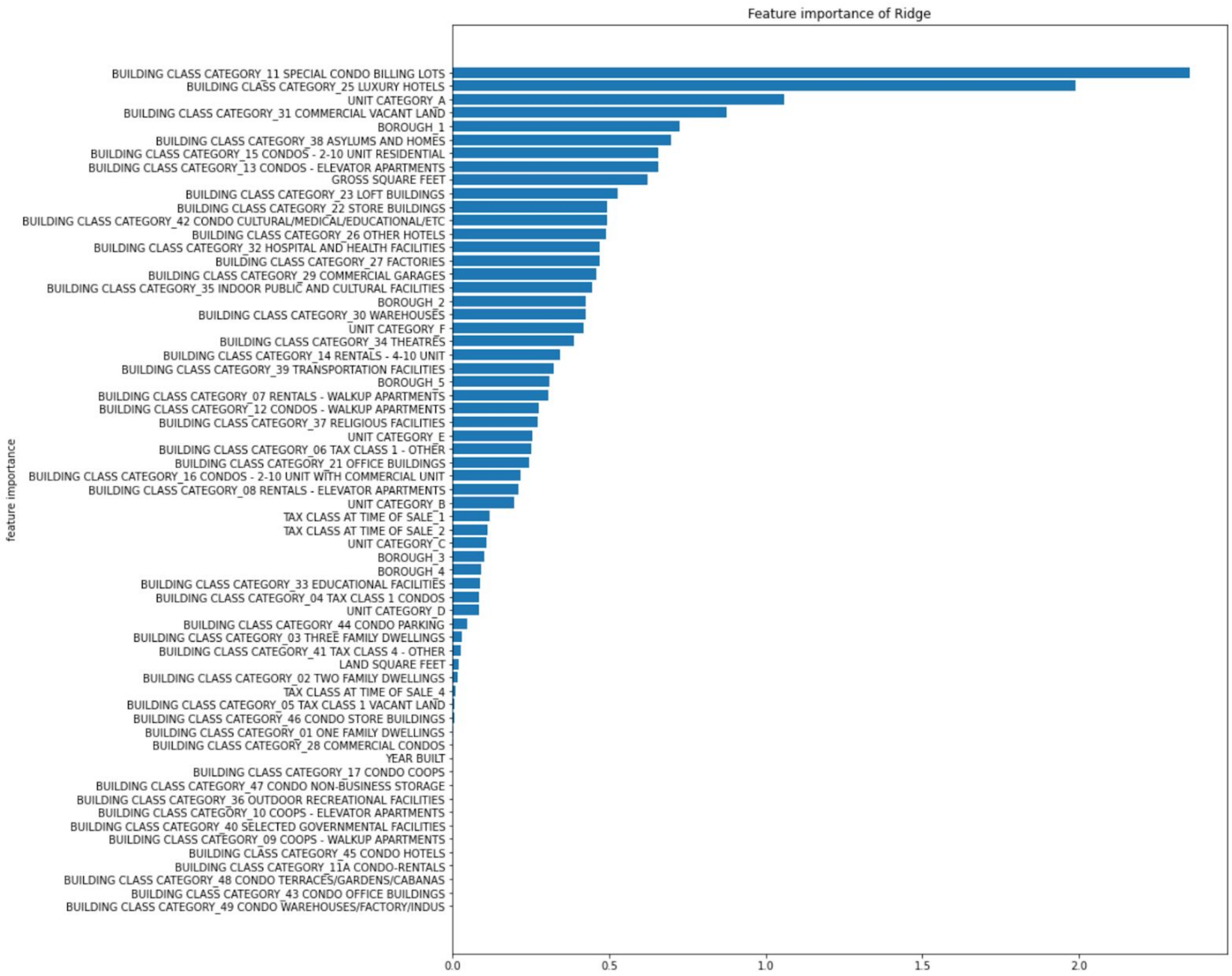
$$R^2 = 1 - \frac{SS_r}{SS_t}$$

Appendix F. Linear Assumptions

There are five key assumptions that all need to hold for the linear model to produce reliable predictions.

1. Linearity: the relationship between the input and output variables is linear. That is, we express the expected value of our output variable, Y, as a linear combination of our input variables, X_1, …, Xn: E(Y) = b_0+ b_1X_1 + b_2X_2 + … + bnXn, where b_0, b_1, …, bn denote the model parameters to be fitted;
2. No Multicollinearity: none of the input variables, X_1, …, Xn, are highly positively or negatively correlated with one another;
3. Normality: observations of the output variable, Y, are assumed to be drawn from the same normal distribution. That is, Y ~ iid N(m, s^2).
4. Homoscedasticity: the variance, s^2, of the output variable, Y, is assumed to be constant, regardless of the values of the input variables;
5. Independence: observations of the output variable, Y, are assumed to be independent of one another. That is, there is no autocorrelation present in our output variable.

Appendix G. Feature Importance reported by Best Model: Ridge


Feature importance of Ridge
feature importance
BUILDING CLASS CATEGORY_11 SPECIAL CONDO BILLING LOTS
BUILDING CLASS CATEGORY_25 LUXURY HOTELS
UNIT CATEGORY_A
BUILDING CLASS CATEGORY_31 COMMERCIAL VACANT LAND
BOROUGH_1
BUILDING CLASS CATEGORY_38 ASYLUMS AND HOMES
BUILDING CLASS CATEGORY_15 CONDOS - 2-10 UNIT RESIDENTIAL
BUILDING CLASS CATEGORY_13 CONDOS - ELEVATOR APARTMENTS
GROSS SQUARE FEET
BUILDING CLASS CATEGORY_23 LOFT BUILDINGS
BUILDING CLASS CATEGORY_22 STORE BUILDINGS
BUILDING CLASS CATEGORY_42 CONDO CULTURAL/MEDICAL/EDUCATIONAL/ETC
BUILDING CLASS CATEGORY_26 OTHER HOTELS
BUILDING CLASS CATEGORY_32 HOSPITAL AND HEALTH FACILITIES
BUILDING CLASS CATEGORY_27 FACTORIES
BUILDING CLASS CATEGORY_29 COMMERCIAL GARAGES
BUILDING CLASS CATEGORY_35 INDOOR PUBLIC AND CULTURAL FACILITIES
BOROUGH_2
BUILDING CLASS CATEGORY_30 WAREHOUSES
UNIT CATEGORY_F
BUILDING CLASS CATEGORY_34 THEATRES
BUILDING CLASS CATEGORY_14 RENTALS - 4-10 UNIT
BUILDING CLASS CATEGORY_39 TRANSPORTATION FACILITIES
BOROUGH_5
BUILDING CLASS CATEGORY_07 RENTALS - WALKUP APARTMENTS
BUILDING CLASS CATEGORY_12 CONDOS - WALKUP APARTMENTS
BUILDING CLASS CATEGORY_37 RELIGIOUS FACILITIES
UNIT CATEGORY_E
BUILDING CLASS CATEGORY_06 TAX CLASS 1 - OTHER
BUILDING CLASS CATEGORY_21 OFFICE BUILDINGS
BUILDING CLASS CATEGORY_16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT
BUILDING CLASS CATEGORY_08 RENTALS - ELEVATOR APARTMENTS
UNIT CATEGORY_B
TAX CLASS AT TIME OF SALE_1
TAX CLASS AT TIME OF SALE_2
UNIT CATEGORY_C
BOROUGH_3
BOROUGH_4
BUILDING CLASS CATEGORY_33 EDUCATIONAL FACILITIES
BUILDING CLASS CATEGORY_04 TAX CLASS 1 CONDOS
UNIT CATEGORY_D
BUILDING CLASS CATEGORY_44 CONDO PARKING
BUILDING CLASS CATEGORY_03 THREE FAMILY DWELLINGS
BUILDING CLASS CATEGORY_41 TAX CLASS 4 - OTHER
LAND SQUARE FEET
BUILDING CLASS CATEGORY_02 TWO FAMILY DWELLINGS
TAX CLASS AT TIME OF SALE_4
BUILDING CLASS CATEGORY_05 TAX CLASS 1 VACANT LAND
BUILDING CLASS CATEGORY_46 CONDO STORE BUILDINGS
BUILDING CLASS CATEGORY_01 ONE FAMILY DWELLINGS
BUILDING CLASS CATEGORY_28 COMMERCIAL CONDOS
YEAR BUILT
BUILDING CLASS CATEGORY_17 CONDO COOPS
BUILDING CLASS CATEGORY_47 CONDO NON-BUSINESS STORAGE
BUILDING CLASS CATEGORY_36 OUTDOOR RECREATIONAL FACILITIES
BUILDING CLASS CATEGORY_10 COOPS - ELEVATOR APARTMENTS
BUILDING CLASS CATEGORY_40 SELECTED GOVERNMENTAL FACILITIES
BUILDING CLASS CATEGORY_09 COOPS - WALKUP APARTMENTS
BUILDING CLASS CATEGORY_45 CONDO HOTELS
BUILDING CLASS CATEGORY_1A CONDO-RENTALS
BUILDING CLASS CATEGORY_48 CONDO TERRACES/GARDENS/CABANAS
BUILDING CLASS CATEGORY_43 CONDO OFFICE BUILDINGS
BUILDING CLASS CATEGORY_49 CONDO WAREHOUSES/FACTORY/INDUS
0.0
0.5
1.0
1.5
2.0