Project Proposal

Team 2: Jessica Tong, Riya Gupta, Ahmed Farid Khan, Shayan Hasan Khan

Problem Definition

Historically, New York City has been one of the most expensive real estate markets in the United States. The state of purchasing property in New York City in the modern day is competitive and complex due to a variety of factors. These factors include high market demand, low inventory levels, the necessity of high financial requirements, and tax considerations. Therefore, we aspire to generate a Property Sale Price Prediction model to solve the issue of uncertainty regarding the estimation of sale prices for properties in New York City based on a particular property's characteristics. This prediction model will allow for a data-driven approach to estimate property sales prices while considering a wide range of variables including the location, square footage, and apartment size, etc.

Data Source

In order to generate our Property Sale Price Prediction model, we are utilizing data provided by the NYC Department of Finance. The data includes the ' Rolling Sales Data' and lists the properties that were sold in the last twelve-month period (September 2016- August 2017) in New York City for tax classes 1-5.

Dataset Attributes

Our dataset consists of 84,548 rows and 22 columns and has the following columns:

- Boroughs: 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).
- Block: Subdivision of the Borough on which real properties are located
- Lot: Subdivision of Block on which real properties are located.
- Neighborhood: Different neighborhoods in New York
- Building Class Category: Used to identify similar properties by broad usage
- Tax Class at Present: Properties are assigned four tax classes
    - Class 1: Includes most residential properties of up-to 3 units
    - Class 2: Includes all other property that is residential
    - Class 3: Includes property owned by gas, telephone and electric companies.
    - Class 4: Includes properties not included in above mentioned categories such as factories, offices, warehouses, garage buildings etc
- Easement: An easement is a right which allows an entity to make limited use of another's real property
- Building Class at present: The Building Classification is used to describe a property's constructive use
- Address: The street address of the property as listed on the Sales File
- Apartment Number
- Zip Code: The property's postal code
- Residential Units: The number of residential units at the listed property.
- Commercial Units: The number of commercial units at the listed property.

- Total Units: The total number of units at the listed property.
- Land Square Feet: The land area of the property listed in square feet.
- Gross Square Feet: The total area of all the floors of a building
- Year Built: Year the structure on the property was built.
- Tax Class at Time of Sale: Tax bracket it was charged on at the time of sale.
- Building Class at Time of Sale: The Building Classification is used to describe a property's constructive use
- Sales Price: Price paid for the property.
- Sale Date: Date the property sold.

Data Analysis Details

To develop our Property Sale Price Prediction model, we will begin with an Exploratory Data Analysis to visualize the distributions of the features and target variables and to identify patterns and anomalies in the data. In addition, we will process the data by removing or imputing missing values, normalizing values, and encoding categorical variables. Next, we will do a correlation analysis to understand the relationship between the features and the target variable, which is 'price' in this dataset.

After data preprocessing, we will look at conducting feature engineering and performing feature selection to identify the relevant features for our model. Next, we will select our model. A few models that will try out are multiple linear regressions, lasso regression, ridge regression, decision trees, random forest, and k-nearest neighbors. Using random search, we will optimize each model and pick the model with the best R-square value. Lastly, we will apply the best hyperparameters for the best model on the testing set.

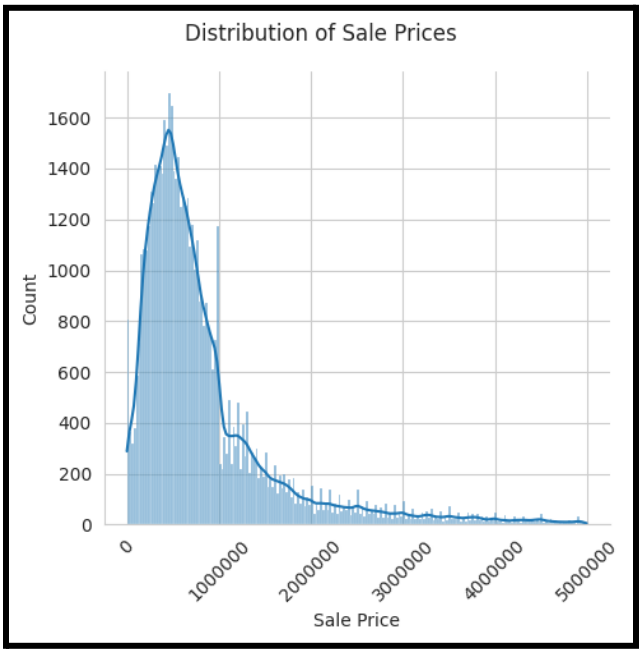
Potential Implications

The result of our project is a data driven model for Property Sale Price Predictions.

Such a model is beneficial for **sellers** to price their property, for **buyers** to make offers, and for **real estate agents** to recommend fair market prices. In addition, investors can use this model to determine the potential return on their real estate investments for different properties by comparing the predicted sale price to the current asking price.

Next, government agencies could use the information generated from this model to help assess property taxes based on the predicted market values to aid in urban planning decisions. Along with government agencies, financial planners could leverage the model for risk management by assessing the collateral value of the property.

Lastly, this model can be used to help set premiums for property insurance based on the value of a property. As a result, our Property Sale Price Prediction model has the potential to generate strategic value across numerous industries and use an objective approach to estimate real estate value.



	RESIDENTIAL UNITS	COMMERCIAL UNITS	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	SALE PRICE
count	84548.00	84548.00	84548.00	58296.00	56936.00	69987.00
mean	2.03	0.19	2.25	3941.68	4045.71	1276456.50
std	16.72	8.71	18.97	41983.97	35032.49	11405255.35
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	1.00	1650.00	1046.75	225000.00
50%	1.00	0.00	1.00	2325.00	1680.00	530000.00
75%	2.00	0.00	2.00	3500.00	2560.00	950000.00
max	1844.00	2261.00	2261.00	4252327.00	3750565.00	2210000000.00

	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	BUILDING CLASS AT PRESENT	ADDRESS	APARTMENT NUMBER	ZIP CODE	YEAR BUILT	TAX CLASS AT TIME OF SALE	BUILDING CLASS AT TIME OF SALE
count	84548	84548	84548	84548	84548	84548	84548	84548	84548	84548	84548	84548	84548
unique	5	254	47	11	11566	2627	167	67563	3989	186	158	4	166
top	4	FLUSHING-NORTH	01 ONE FAMILY DWELLINGS	1	5066	1	D4	131-05 40TH ROAD		10314	0	1	R4
freq	26736	3068	18235	38633	404	4125	12663	210	65496	1687	6970	41533	12989

	SALE DATE
count	84548
unique	364
top	2017-06-29 00:00:00
freq	544
first	2016-09-01 00:00:00
last	2017-08-31 00:00:00