**NYC Property Sales Data Analysis**

IST 718 Group 11

Group Members：

Chaoying Lyu｜ Hong Zhu｜ Ruiwei Zhang｜ Ziyan Zhou

Table of Contents

[Abstract 3](#_Toc57587938)

[1. Data Collection/ Cleaning / Exploration 4](#_Toc57587939)

[1.1. Data Collection 4](#_Toc57587940)

[1.2. Data clean 4](#_Toc57587941)

[1.2.1. Remove the duplicate data 4](#_Toc57587942)

[1.2.2. Deal with null value 4](#_Toc57587943)

[1.2.3. Deal with outlier 4](#_Toc57587944)

[1.3. Data Exploration 5](#_Toc57587945)

[1.3.1. Trends in the Property Market 5](#_Toc57587946)

[1.3.2. Location vs property sales price 5](#_Toc57587947)

[1.3.3. Distribution of top property sales price 6](#_Toc57587948)

[1.3.4. Preference on property age 6](#_Toc57587949)

[1.3.5. Type of building class vs sales price 7](#_Toc57587950)

[2. Methodology 10](#_Toc57587951)

[3. Models 10](#_Toc57587952)

[3.1. Linear regression 10](#_Toc57587953)

[3.2. Random Forest 12](#_Toc57587954)

[3.3. The Gradient boosted trees Model 13](#_Toc57587955)

[Model 3: Gradient boosted trees (GBT) 13](#_Toc57587956)

[Model 3 Data Transformations 13](#_Toc57587957)

[Model 3 Evaluation 14](#_Toc57587958)

[Model 3 Interpretation 14](#_Toc57587959)

[Model 3 Inference and Key Findings 15](#_Toc57587960)

[4. Conclusion 16](#_Toc57587961)

[5. Reference 17](#_Toc57587962)

# **Abstract**

The coronavirus outbreak might have people feeling uncertain about whether a change in home prices will impact their housing plans. It has already caused many to take a “wait and see” approach towards real state. For property sales companies, what they care most is what kind of property may have the highest preference from the customers. Hence, we use the NYC Property Sales dataset, which contains the house sale prices and relative information about the houses, like address and land square feet, to help predict the property prices.

Our prediction contains the followings:

1. Check if there’s a time-related trend in the property price
2. Find out which neighborhood has the highest or property sales prices
3. Predict which neighborhood may have relatively new or old property
4. Predict sale price of different building class.

From the predictions, some inference questions can be answered, and they will be useful as reference for property sale companies. The inferences are shown below:

1. Find the most important factor on the sale price for all house sales.
2. Distinguish the differences in the most influential factor on house prices in different places
3. Verify if there’s a gap between small-size houses and large-size houses in sale price’s factors
4. Find the differences in factors on sale prices between residential and commercial units

Check if the building class may affect the sale price factors.

Several models will be applied to the dataset to find the influential factors on housing prices. Linear regression, random forest and gradient boosted trees are applied for our project purpose.

After the data analysis, the main goal of the project is to find ways to calculate property sale price with property features as a reference for buyers and sellers when they are trading property. Companies will be able to make critical decisions related to hiring and investments based on the forecasting.

# Data Collection/ Cleaning / Exploration

## Data Collection

The NYC Property Sales dataset is a record of every building or building unit (apartment, etc.) sold in New York City over a 12-month period from September 2016 to September 2017. There are 84548 rows and 22 columns in total, containing the location, address, type, sales price and sales date.

## Data clean

## Remove the duplicate data

We found the duplicate sales data by the columns ('ADDRESS', 'NEIGHBORHOOD','SALE DATE' and 'SALE PRICE'). We found 2523 duplicate data. Then, we drop the duplicate rows to avoid influence of dulplicate data in future analyze.

## Deal with null value

We found there are two forms of the null value existing in our data. The first one is the NA values. In addition, some null value used ‘-‘ in our dataset. So, we replace them as ‘0’’. And we remove 0 value in sale price column. Although some zero sales price represent actually transferring of deeds between parties, we still move it to analyze the actual value of the property in NYC.

## Deal with outlier

After analyzing the distribution of the sale price in NYC, we can find that part of value is extremely large which may influent our analytics. Therefore, we only take the real estate sales data from $100,000 to $100,000,000. After that the data distributed is more balanced.

Shape

Description automatically generated

## Data Exploration

## Trends in the Property Market

The price per square feet of property sale in New York City shows a volatile upward trend during this year.

Chart, line chart

Description automatically generated

## Location vs property sales price

The house price of Manhattan is much higher than other brough in NYC.

Chart, bar chart

Description automatically generatedMap

Description automatically generated

## Distribution of top property sales price

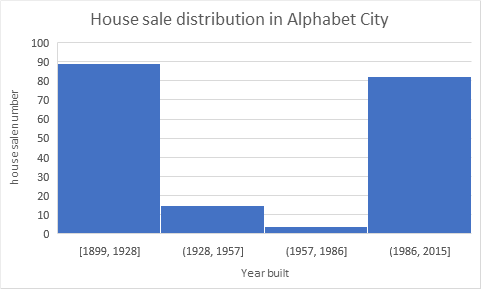
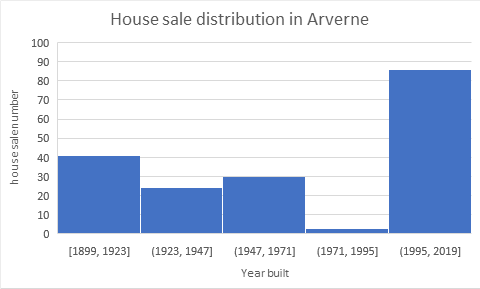
The most expensive house neighbors gather at south part in NYC. Most of them located Distributed along the Hudson River and the East River.

Map

Description automatically generated

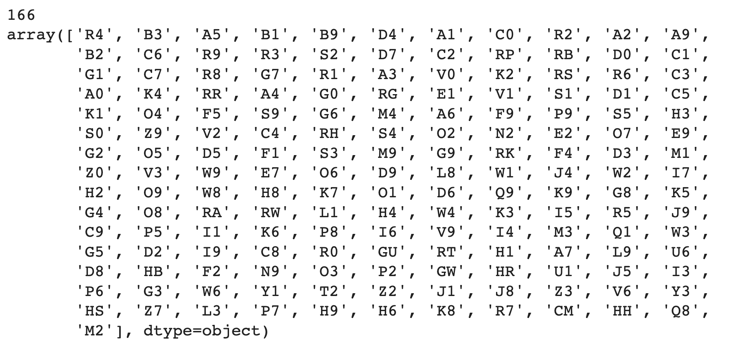
## Preference on property age

From the dataset, two random districts are selected from one borough. From the two graphs, one is occupied by the newly built property while the other have an over 50% of old house in the property sales. Hence, people may not have specific preference on house age.

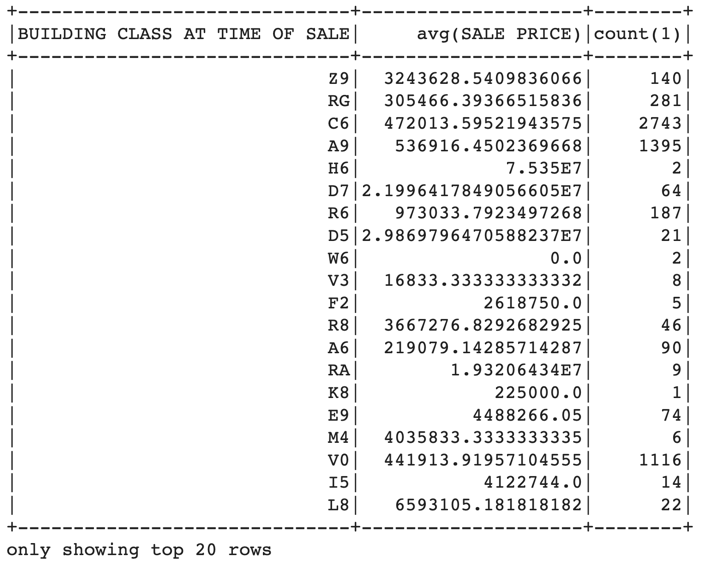


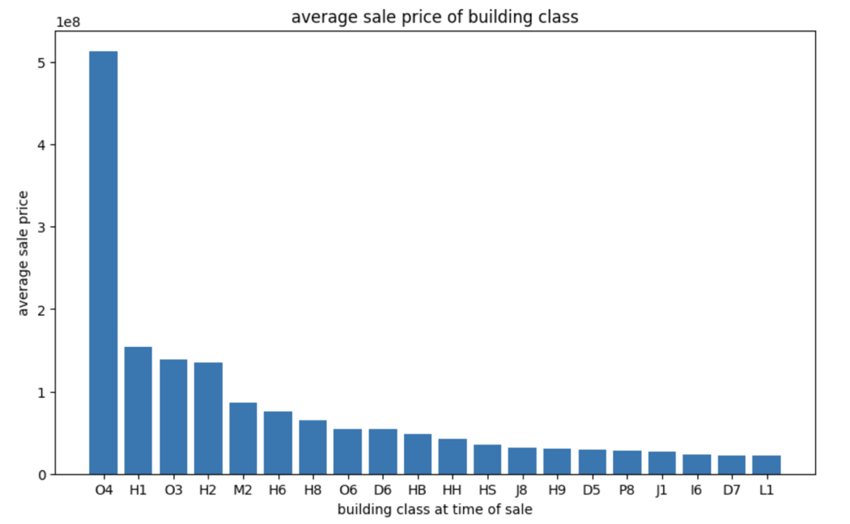
## Type of building class vs sales price

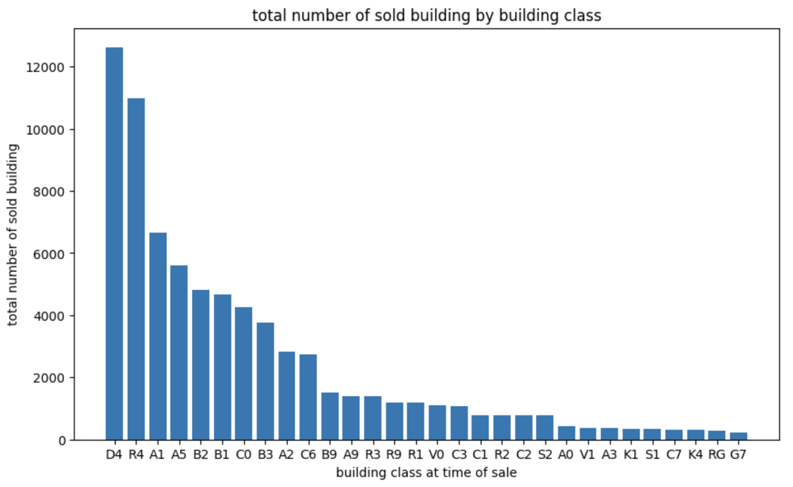
There are totally 166 different types of building class, each of which consists of a character and a number representing different usage of the building and number of rooms in the building. [2] For example, R4 stands for condo, residential unit in elevator building. The following picture shows the specific 166 types that appear in our project.



Based on different types of building class, we compute the average sales price and count the total number for each category as shown in the data frame below. In our data set, there are two columns refer to building class. One is building class at present, the other is building class at time of sale. We decide to use building class at time of sale because we are focusing on the relationship of variables and sales price and building class at present may differ from building class at time of sale, so building class at present doesn’t matter in our case.



We also created two plots for our findings. The first plot is average sales price of building class as shown in the picture below. From this picture, we can see that the price is extremely right skewed, so I took the top 20 for plotting. The highest average price is O4, which refers to the office only with or without comm with 20 stories or more. The second and the third highest is H1 and H2, which represent luxury hotel and full service hotel. Follows by O3, which is office with 7 to 19 stories. For now, we can conclude that the price of office building is much higher than other type, follows by luxury hotel.

Another plot we create is the count of sold building by building class. This plot is right skewed as well, but not as severe as average sales price below, and thus I took top 30 for the plot. From the plot, we can see that the types of the top two sold buildings are R4 and D4. R4 is condo, or residential unit in elevator buildings. D4 is elevator cooperative buildings. They were sold more than 12,000 in the last year. Follows by A1 and A5, which are two stories and one family attached or semi-detached. The third group is B2 and B1 which are two family frame and two family brick.

# Methodology

Our process follows the flow diagram.Data will be cleaned first. NAs will be removed and unreasonable data types will be regulated into the form qualified for the models inputs. The dataset will be separated into training and testing part. After that, three models, linear regression, random forest and GBT, will be applied to solve the regression problem. In the evaluation step, mean square error is utilized to grade the prediction accuracy of our models. With the help of the models and evaluation, we can draw our conclusion for this project.

Diagram

Description automatically generated

# Models

# Linear regression

This model is designed to predict property sales price based on the property information. it contains the columns, borough, neighborhood, building class category, residential units, commercial units, land square feet, gross square feet, building age as the predictors. We use both borough and neighborhood as the location information of property to make the location more detailed. As borough, neighborhood, building class category are all categorical columns and don’t have a specific order, we convert them into dummy variables. Considering the fact that unit property price varies from one location to another, we also multiply borough columns by land square feet and gross square feet columns and get new features borough\* gross square feet, borough\* land square feet so that the coefficient of gross square feet in linear model can be different for different borough.

As we usually do in homework, we split the dataset into two parts, one part for training and the other for testing, and use MSE as the metric to evaluate the linear model.

The MSE score we get is 9.10489e+12, which means the mean estimation error on the price is more than one million dollars. The accuracy of the property price estimation is beyond our expectation. After trying to add more columns, the result doesn’t change a lot. We then check the original data in the CSV file and find that the sales price of property of larger square feet is sometimes lower than that of property of the same information other than square feet. We think some factors have more influence on sales price than the elements listed in the dataset like decoration. If one apartment is within the walking distance of some universities like NYU, it tends to have higher sales price compared to another apartment in the same borough far from NYU. The following graph shows the relation of our price prediction and the real price.

Chart, scatter chart

Description automatically generated

The closer the blue dots are to the red dotted line, the more accurate the linear model is.

Next, we use inference to determine the most important predictor order. We create a new pipeline which encapsulates a standard scalar and a linear regression object. After fitting the pipe, we create a pandas data frame with 2 columns named coefficient and value. The coefficient column contains the coefficient names and the value column contains the regression model coefficient absolute values.

Table

Description automatically generated

We can see the top ten predictors are all from neighborhood and building class category which is not consistent with our common sense.

# Random Forest

We use the same methods as linear model to transform categorical variables. To shorten the running time of the model, we only use borough as the location information of property.

The random forest model utilized a grid search to tune the regularization and elastic net parameters. The grid utilized num trees of 10, 30, 50 and max depth of 10 and 15. The trained model variations are evaluated by the mean square error (MSE) testing results after applying the model to the test dataset. The best model uses the num trees of 30 and max depth of 15.

The best random forest model was able to achieve a MSE score of 6.10829e+12 which is a little better than the result of linear regression model. The following graph shows the relation of our price prediction and the real price.

Chart, scatter chart

Description automatically generated

Below is the most important predictor in random forest model.

Table

Description automatically generated

From the table, the model’s most important feature is the gross square feet of the property. The residential units of the property are the second feature people care about. This shows living area is people’s most important concern. It does make sense. Besides, the commercial units and the building age can only explain some buyers’ interest on properties. These two features are less preferred by the buyers.

# The Gradient boosted trees Model

This model is designed to address the relationship between the property area and the price. It will contain the total units, the land square feet, gross square feet, and the year built as the factors. It will focus more on the property conditions rather than the neighboring environment. This model will answer the question: how the property area affects the property price in New York.

This model is used to solve a regression problem. After the dataset split, a part of the dataset is used for training and the other can be used for testing the accuracy of the model. The testing result can show us the gap between the model estimation on prices and the real sales of the property.

Model 3 Data Transformations

GBT model. The outliers and NAs in the related columns are removed first. In addition, since the original data type in these columns are string, they are transformed into float type for the future model processing. This model uses only 4 features in total. The following features were not included:

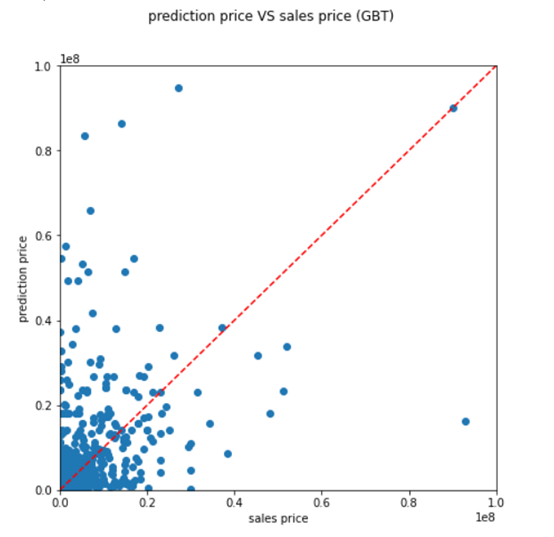
* Data about the property location: borough, address, apartment number, zip code, block
* Categorical data about the property feature: neighborhood building class category, building class at present, residential units, commercial units, lot, easement, tax class at present:
* Data about the sales of the property. tax class at time of sale, building class at time of sale, sale date:

Model 3 Evaluation

The GBT model utilized a grid search to tune the regularization and elastic net parameters. The grid utilized max iteration of 10, 20, 50 and max depth of 5 and 10. The trained model variations are evaluated by the mean square error (MSE) testing results after applying the model to the test dataset. The best model uses the max iteration of 10 and max depth of 10.

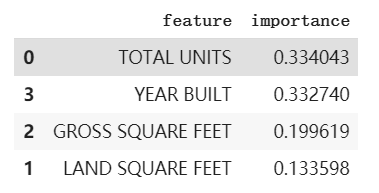
Model 3 Interpretation

The best GBT model was able to achieve a MSE score of 10^12.This shows that the accuracy of the property price estimation is within our expectation. The following graph shows the relation of our price prediction and the real price.



### Model 3 Inference and Key Findings

The feature coefficients were extracted from the model and sorted based on absolute value to provide insight on feature importance. The ordered feature coefficients are provided below.



From the table, the model’s most important feature is the total units of the property. The built-year of the property is the second feature people care about. This shows buyers have some preference on the property age. Besides, the gross square feet and the land square feet can only explain some buyers’ interest on properties. These two features are less preferred by the buyers. To answer our business questions, we can conclude that the total units and the building age are quite important factors for buyers.

# Conclusion

Our main goal of predicting property sales price is to provide a tool for people who want to buy or sell property. Using this tool, they can get a reasonable price if they input the property information. The mse of linear model is 9.10489e+12. The mse of random forest model is 6.10829e+12. The mse of GBT model is 5.01303e+13. The random forest performs best.

The most important feature of the property is the gross square feet, residential units, commercial units and building age. These are the factors people most care about when buying or selling property.

The overall trend of the property sales price is ascending. The property of Manhantton and Madison has the highest price.

# Reference

[1] NYC Property Sales dataset

[HTTPS://WWW.KAGGLE.COM/NEW-YORK-CITY/NYC-PROPERTY-SALES?SELECT=NYC-ROLLING-SALES.CSV](http://、HTTPS://WWW.KAGGLE.COM/NEW-YORK-CITY/NYC-PROPERTY-SALES?SELECT=NYC-ROLLING-SALES.CSV)

[2] Building Classification | City of New York

<https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html>