Sale Price Prediction and Analysis of New York City Property

Introduction

This project is inspired by the new developments in the real estate despite the pandemic and wondering how the sales were in famous cities during the year 2020.

The motive of this project is to see the trends in real estate in major boroughs of New York last year and predict the sale price for a residential or commercial unit.

Dataset

The data is extracted from New York department of Finance web site at this link.
https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

The data has properties that were sold during 2020 in New York city with a separate listing for each borough namely Manhattan, Bronx, Brooklyn, Queens, and Staten Island. All these listings were combined to a single list which helped compare the boroughs. The metrics that were part of this sales information were Borough, Neighborhood, Building Class Category, Tax class at present, Block, Lot, Easement, building class at present, Address, Apartment Number, ZIP code, Residential units, Commercial units, Total units, Land square feet, Gross square feet, Year built, Tax class at time of sale, Building class at time of sale, Sale price and Sale date. There were about 35000 observations.

The data includes for different tax classes.

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.

Class 2: Includes all other property that is primarily residential, such as cooperatives and condominiums.
Class 3: Includes property with equipment owned by a gas, telephone, or electric company.

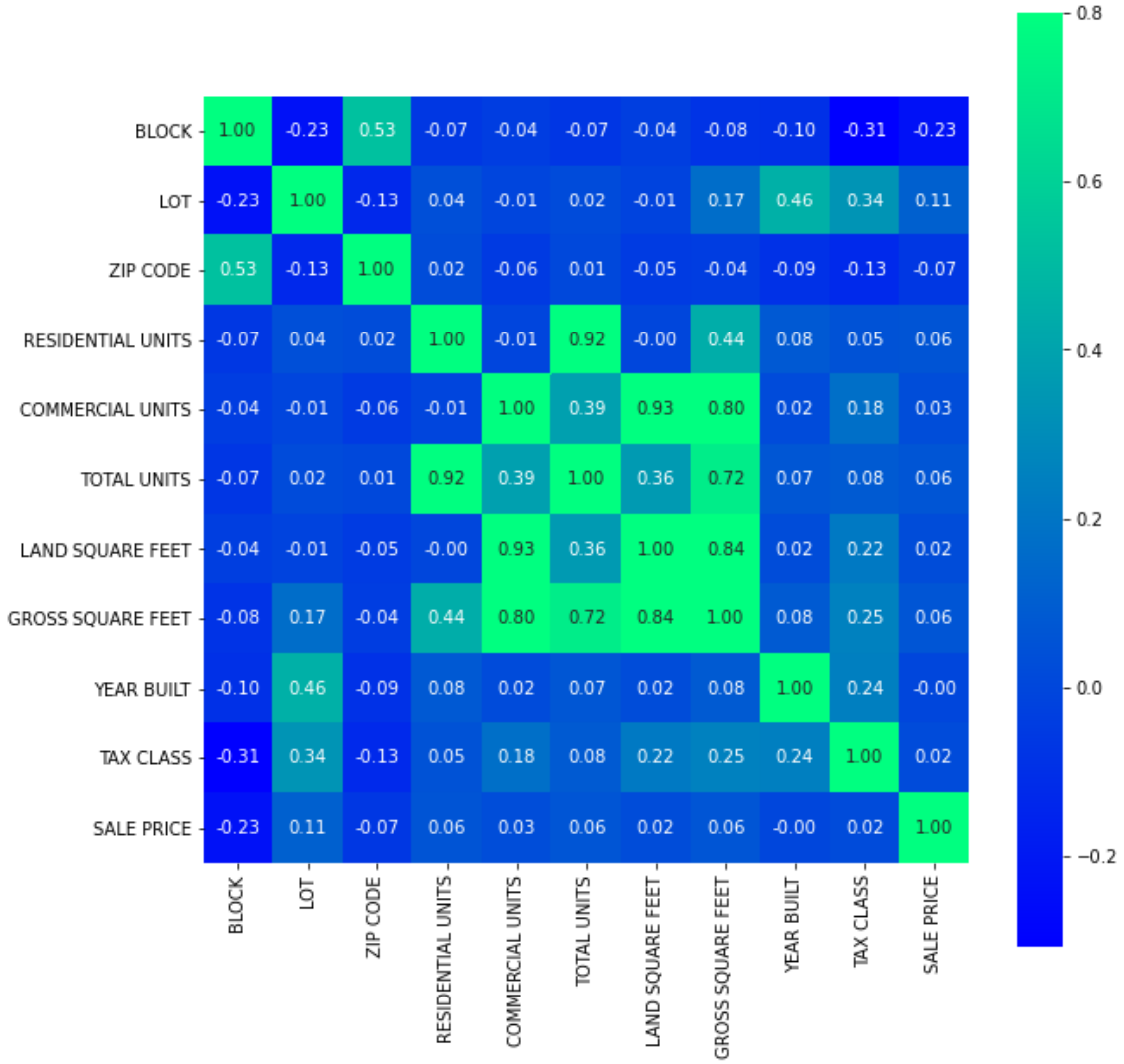
Class 4: Includes all other properties not included in class 1,2, and 3, such as offices, factories, warehouses, garage buildings, etc.

Building Class Category : similar properties are grouped by broad usage

Analysis

The sale price was skewed, and some sales included 0's where the property were transferred within the family or donated to charity. So only sales between 100,000 and 3 million were considered for analysis. Some sales included coops housing where people buy shares in that building and occupy. These were also considered as sales in our analysis and modelling.

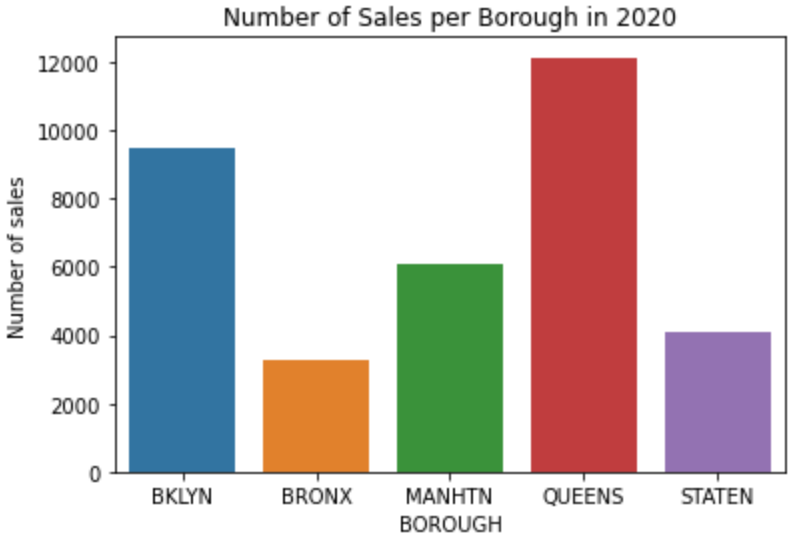
Here is the correlation matrix of main metrics used in the dataset.



Residential Units and Total Units are highly correlated. Also, Commercial units and Land Square Feet is correlated. Hence removed Residential and Commercial Units in modelling. Very few entries did not have year built which was removed.

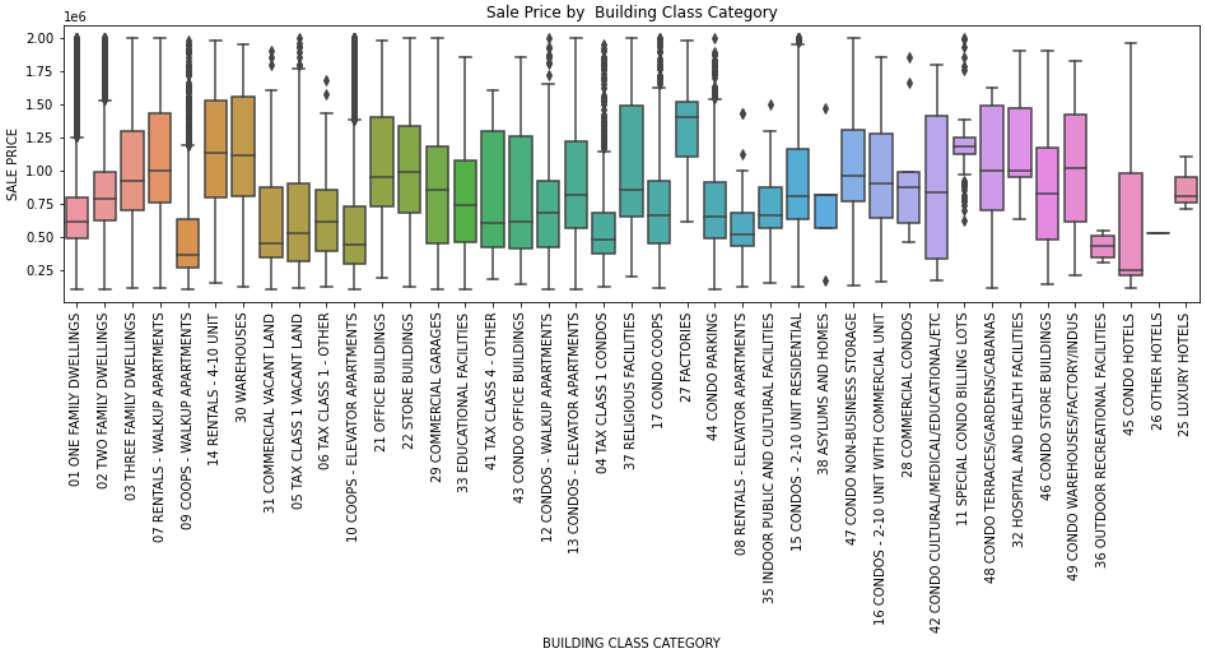
Land Square Feet and Gross square Feet had many null values. These were filled using KNN imputing. Total Units that were not populated, were the coop housing. So, filled them with value 1.

Here are some of the trends noticed in the real estate analysis.



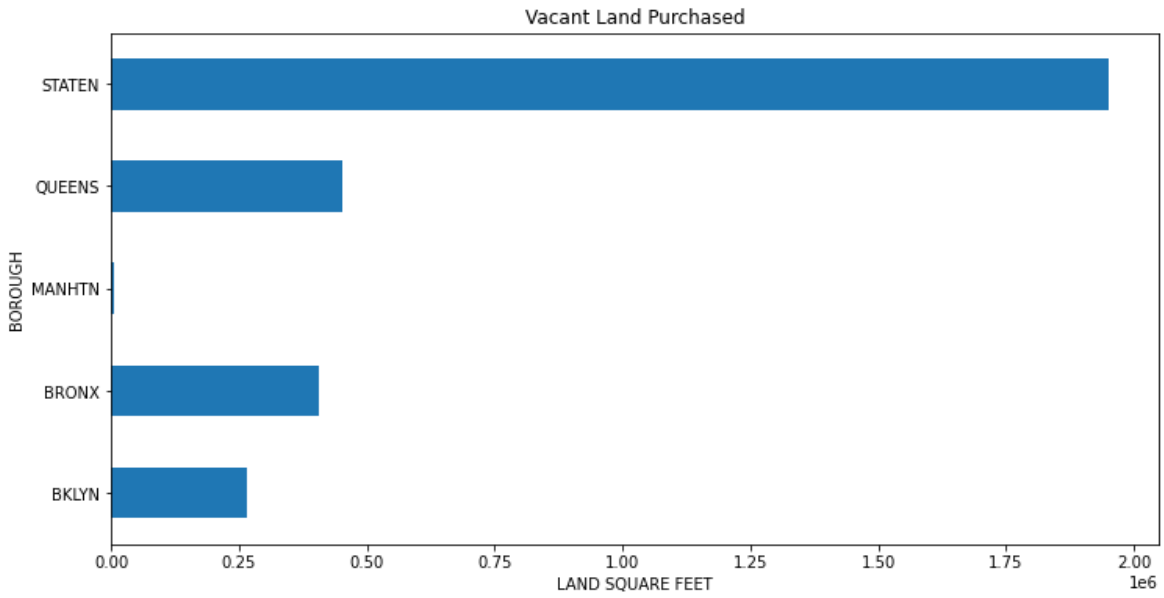
More sales were in Queens than other boroughs last year and Bronx has the least property sales

Some building categories had less than 10 sales, so they were merged to category Others. Here is the distribution of sale price for a given building category.

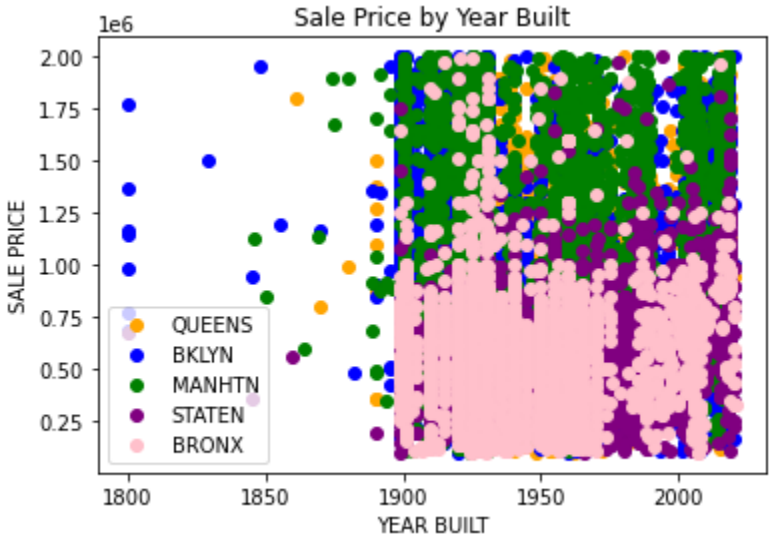


Factories had the highest sale price and condo hotels had the least median sale price.

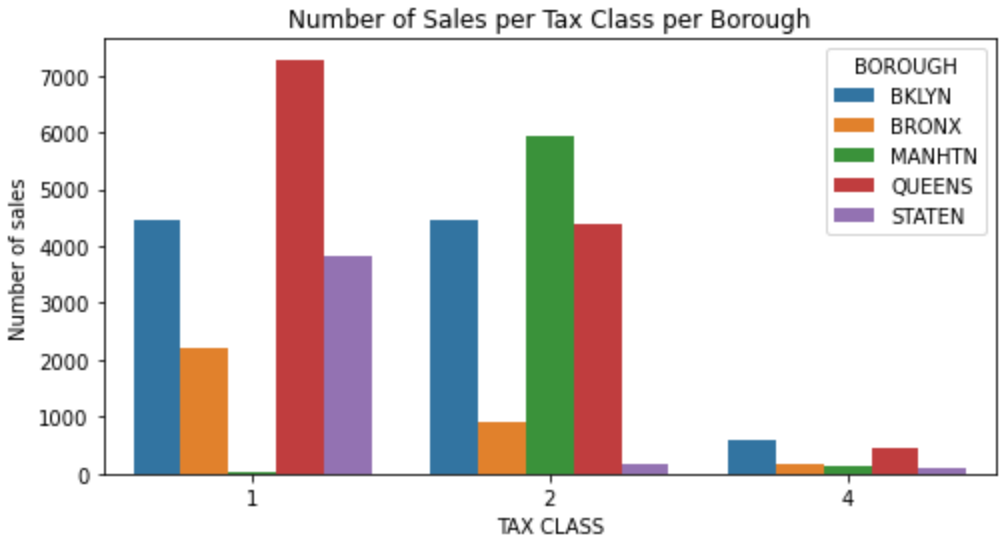
Here is the distribution of vacant land purchases in 2020.



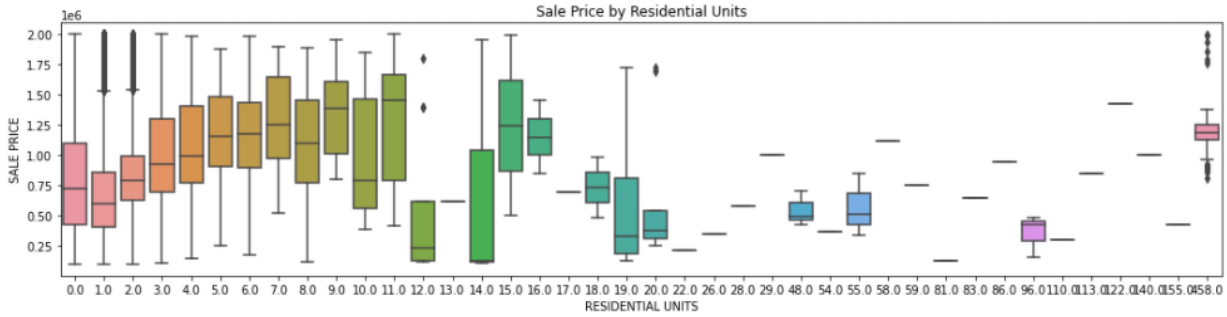
There are more chances of new housing and commercial construction in Staten Island than in other boroughs and the least in Manhattan.



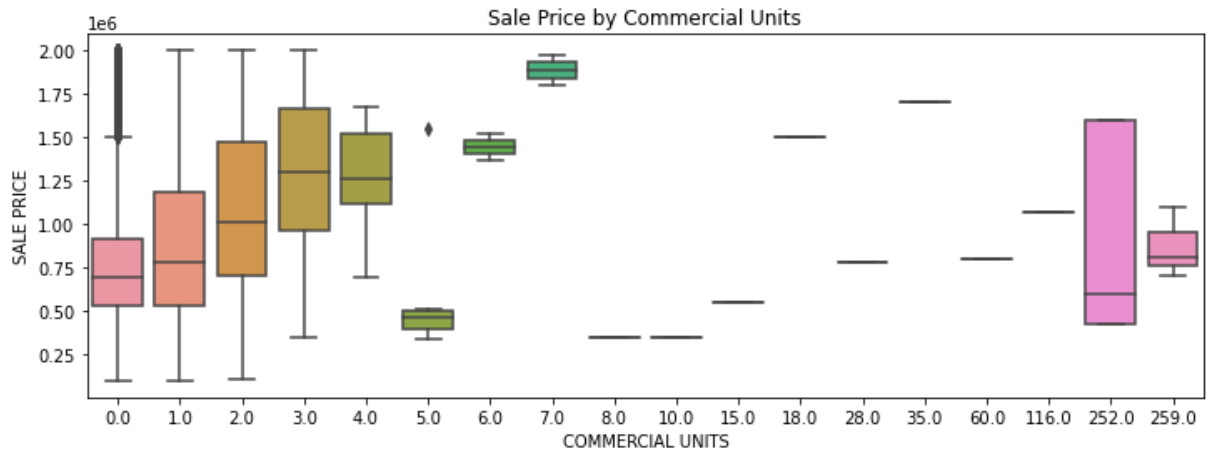
Few houses in Brooklyn are incredibly old built in 1800's and their price is comparable with newer houses elsewhere. Bronx has comparatively lesser priced houses than other boroughs. Manhattan has more expensive houses.



More homes of higher tax brackets were sold in Queens and none in Manhattan in 2020. Also, very few commercial and industrial properties were sold last year probably due to COVID. Most of the sales for Manhattan was in Tax class 2 which is not surprising as it has many properties which has more than 3 units.



Residences that had 9 units had the highest median sale price compared to even single unit. The multi-unit building could be in Manhattan and the single unit could be in Brooklyn. Similarly, for commercial buildings below has 3 units as highly priced sales.



Modelling

Target metric is Sale price.

Deciding to choose Random Forest instead of Decision Tree as it has high variance and tendency to overfit. Hence bias would be low. Do considered multiple trees approach. Used Ensemble, Random Forest for bagging and Gradient Boosting for Boosting and got better results.

The models that were used for predicting are Linear Regression, SVM, Random Forest, KNN and Gradient Boosting. The numeric variables were standardized for some models.

Model

Model	R2 Score	Mean Square Error
Linear Regression	-66989904064163466185592012 8.00	10189285197753888768.00
SVM	0.20	352024.54
Random Forest	0.46	288921.96
KNN	0.47	286146.00
XGBoost	0.58	255464.00

It is evident that XGBoost is the best model compared to all other models. The hyperparameters that pr oduced the best results were
learning_rate: 0.05, max_depth: 8 and n_estimators: 1000

Recommendations

Future Research

As the data is limited, we could only obtain so much accuracy from public dataset. But this can be further improved by segmenting the dataset. Probably a separate modelling for each borough and further for each Tax class would yield better results.