Predicting New York City Housing Prices

Matthew Kwee
Aug. 2021

Introduction

Motivation:

> During the COVID-19 pandemic, many people have begun to work remotely, causing real-estate prices in New York City to change significantly. Nested LLC, a real-estate brokerage firm, needs a new model for predicting housing prices.

Objectives:

> Create a model to accurately predict housing prices in New York City using sold house listings.

Goals:

> Present analysis understandably and logically.

Methodology

Data:

My data was gathered from realtor.com using a simple web-crawler.

Tools utilized:

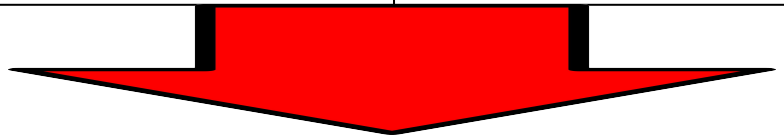
Selenium, Pandas, Google Maps Geocoding API, NumPy, Scikit-Learn, MatPlotLib, Seaborn

Metrics:

When it comes to predicting housing prices, maximizing the regression model's r^2 value will be the most important objective.

Features

What do you want to know about a house that you might purchase?	
Numerical:	Categorical:
Square Footage Number of Bedrooms Number of Bathrooms Lot Size Year Built	Location Property Type



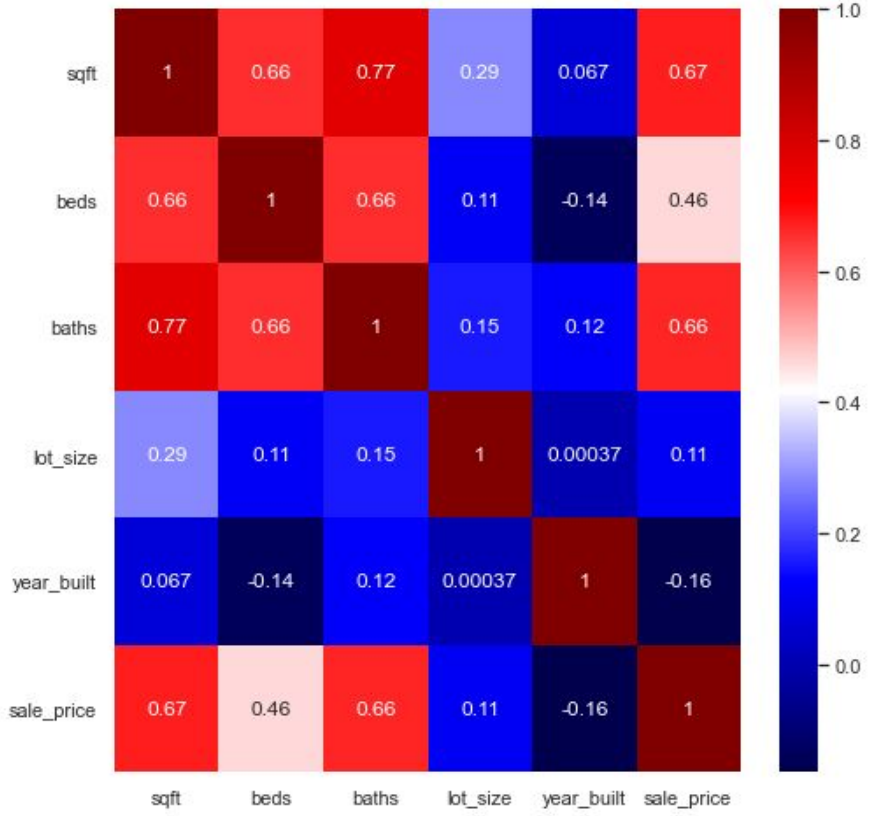
PROPERTY VALUE

Initial observations: Correlation

Before I began to construct a model, I used Seaborn to create a correlation heatmap of all continuous features.

Red - High correlation

Blue - Low or negative correlation

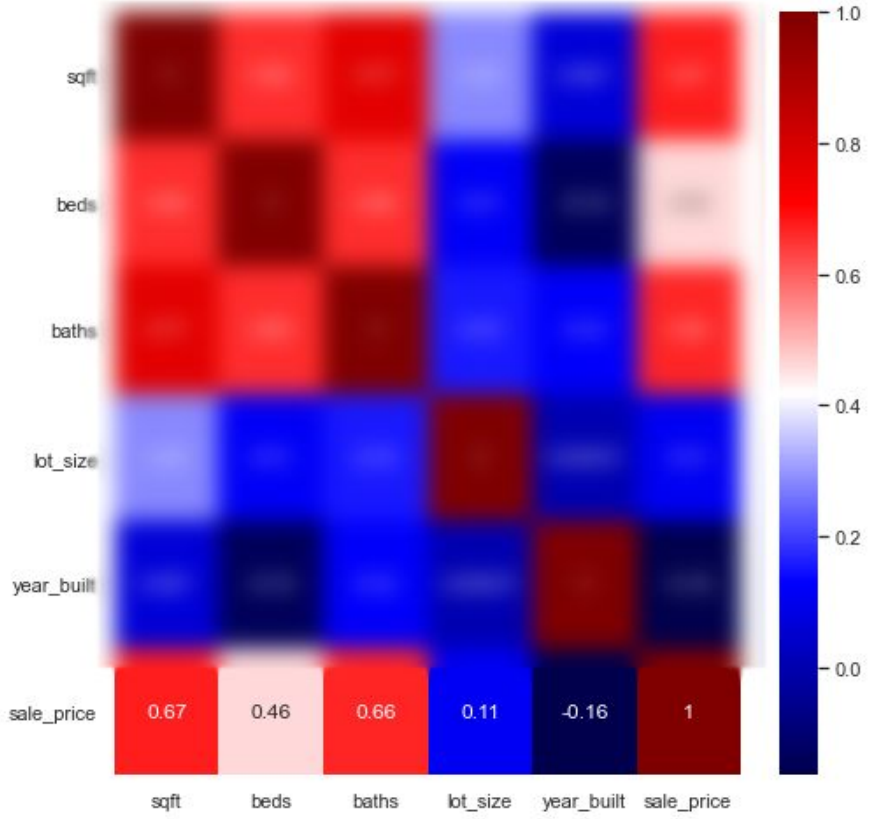


Initial observations: Correlation

Sqft and Number of Bathrooms have highest correlation with sale price.

Number of Bedrooms is moderately correlated with sale price.

A building's age is slightly negatively correllated with sale price.



Modelling

After converting the categorical variables into binary dummy variables, I placed 80% of my data into a training set, and left 20% for the test set.

After fitting Linear, Lasso, Ridge, and Elastic Net Regressions to the training data, I scored each on the test set.

I settled on a simple Linear Regression because the model's r^2 value for the training set was not significantly higher than that of the test set.

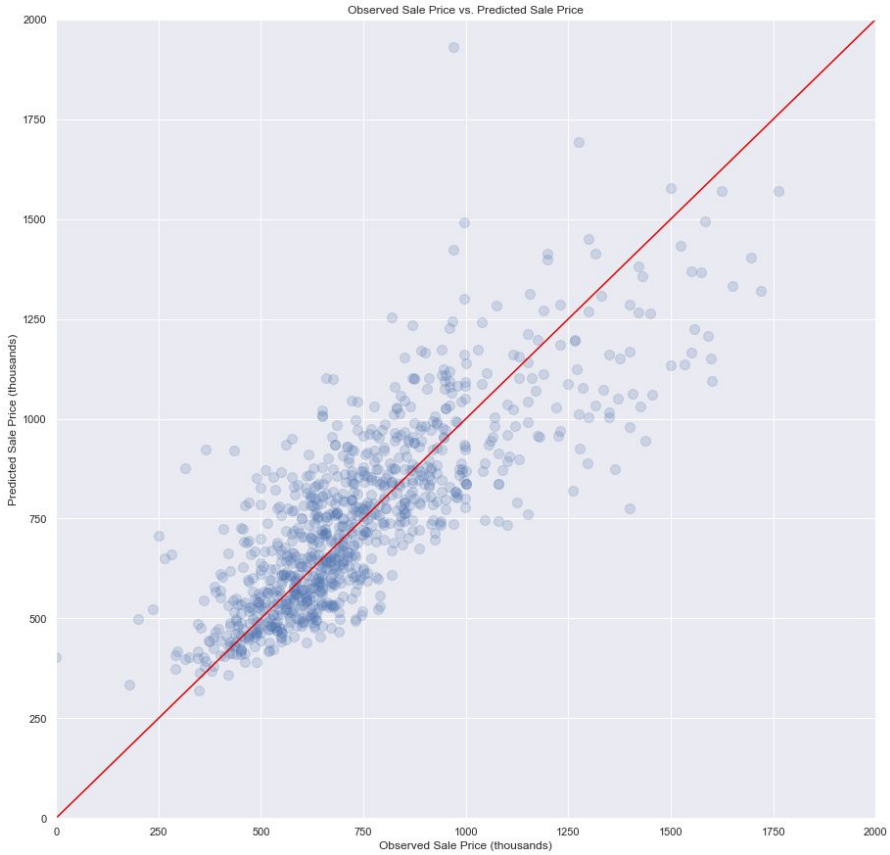
Regression Results

The image shows the linear regression model's prediction for each property value and the actual value of the property.

If a datapoint is above the red line, the model's predicted price for that property was higher than the property's actual sale price, and vice versa.

For the training set, r^2 was 0.8067.

For the test set, r^2 was 0.8081.



https://github.com/MK38993/Metis-Project-2---Real-Estate-Market/blob/main/NY_Housing_Market_Model.ipynb

Using The Model

I packaged my model into a function which can be used here.

The rest of the code can be found here.

In [3]:

```
print(predict_house_price(1120,4,1,940,1920,'Single Family Home','Bronx'))
print('\n')
print(predict_house_price(1120,4,1,940,1920,'Single Family Home','Brooklyn'))
print('\n')
print(predict_house_price(1120,4,1,940,1920,'Single Family Home','Manhattan'))
print('\n')
print(predict_house_price(1120,4,1,940,1920,'Single Family Home','Queens'))
print('\n')
print(predict_house_price(1120,4,1,940,1920,'Single Family Home','Staten Island'))
```

```
Input: 1120 sqft, 4 bedrooms, 1 bathrooms, 940 lot size, built in 1920, property type Single Family Home, located i
n Bronx, NYC
Predicted price:
313959.99818827584


Input: 1120 sqft, 4 bedrooms, 1 bathrooms, 940 lot size, built in 1920, property type Single Family Home, located i
n Brooklyn, NYC
Predicted price:
724390.4628279372


Input: 1120 sqft, 4 bedrooms, 1 bathrooms, 940 lot size, built in 1920, property type Single Family Home, located i
n Manhattan, NYC
Predicted price:
4327579.285491296


Input: 1120 sqft, 4 bedrooms, 1 bathrooms, 940 lot size, built in 1920, property type Single Family Home, located i
n Queens, NYC
Predicted price:
624339.0820422402


Input: 1120 sqft, 4 bedrooms, 1 bathrooms, 940 lot size, built in 1920, property type Single Family Home, located i
n Staten Island, NYC
Predicted price:
392359.94366843114
```

Additional Metrics (for nerds like me)

The model's r^2 value was 0.8071 for the entire data set.

The RMSE was $254350.

The coefficient of variance (RMSE ÷ mean property price) was 32.08%.

Regression Coefficients

Intercept: 1552378.2479379964

Sqft: 227124.58
Beds: -53247.32
Baths: 55459.69
Lot Size: 54373.47
Year Built: -511.65

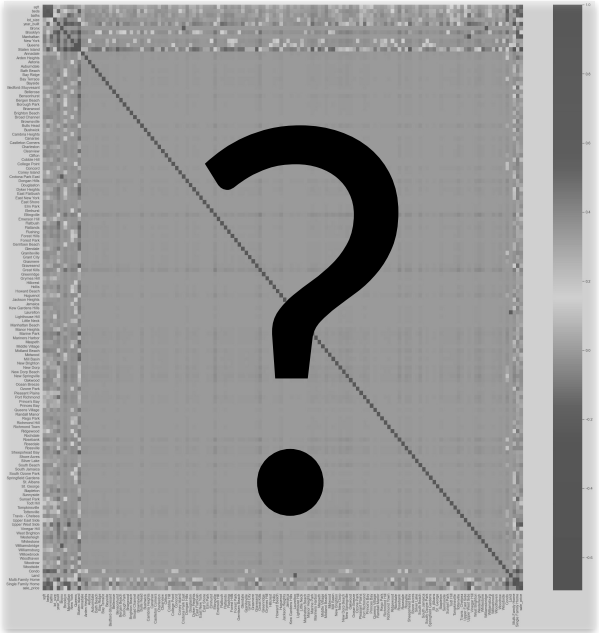
Bronx: -962565.76
Brooklyn: -552135.29
Manhattan: 3051053.53
Queens: -652186.67
Staten Island: -884165.81

Condo: -34559.15
Land: -46616.89
Multi-Family Home: 67118.29
Single Family Home: 14057.72

Future Work

I had planned to include the neighborhood a property was located in as part of my regression, but decided against doing so due to the lack of datapoints.

The next logical step would be to acquire enough sold listings from real-estate websites to use ‘neighborhood’ as a categorical variable in my model.



Appendix

That red arrow on slide 6: https://www.freeiconspng.com/uploads/red-arrow-down-png-17.png