Predicting New York City Housing Prices with a Linear Regression Model

Matthew Kwee

6 August 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.

-9900 listings

-935 listings left after cleaning

Methodology

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

<table>
<tr><th colspan="3">What do you want to know about a house that you might purchase?</th></tr>
<tr><td colspan="2">Numerical:</td><td>Categorical:</td></tr>
<tr><td>Square Footage
Number of Bathrooms
Year Built</td><td>Number of Bedrooms
Lot Size</td><td>Location
Property Type</td></tr>
</table>

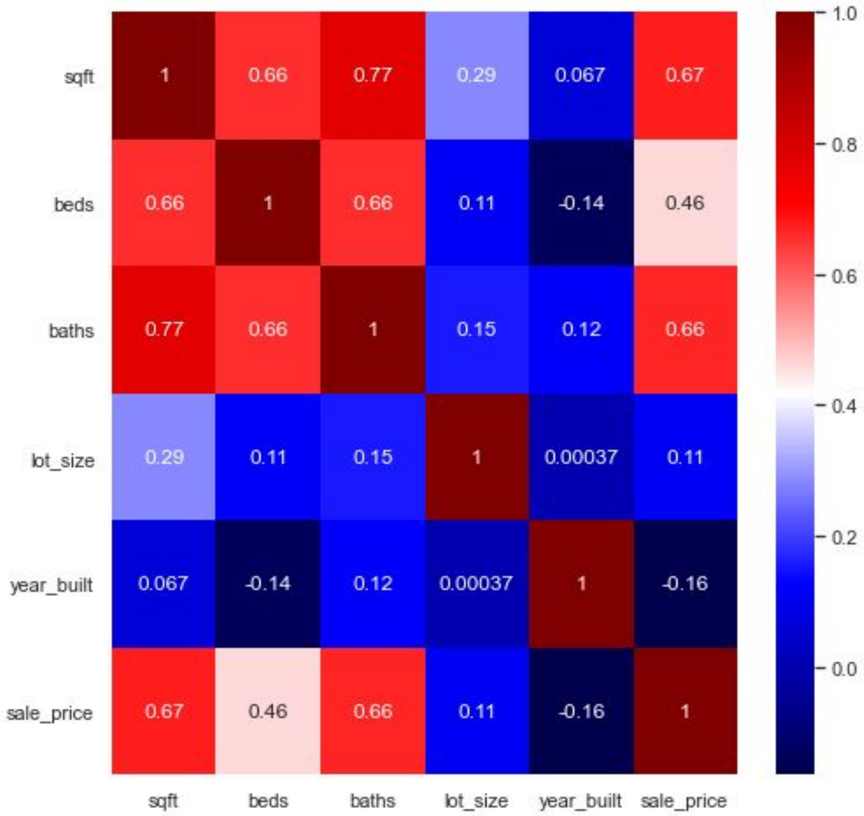
PROPERTY VALUE

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

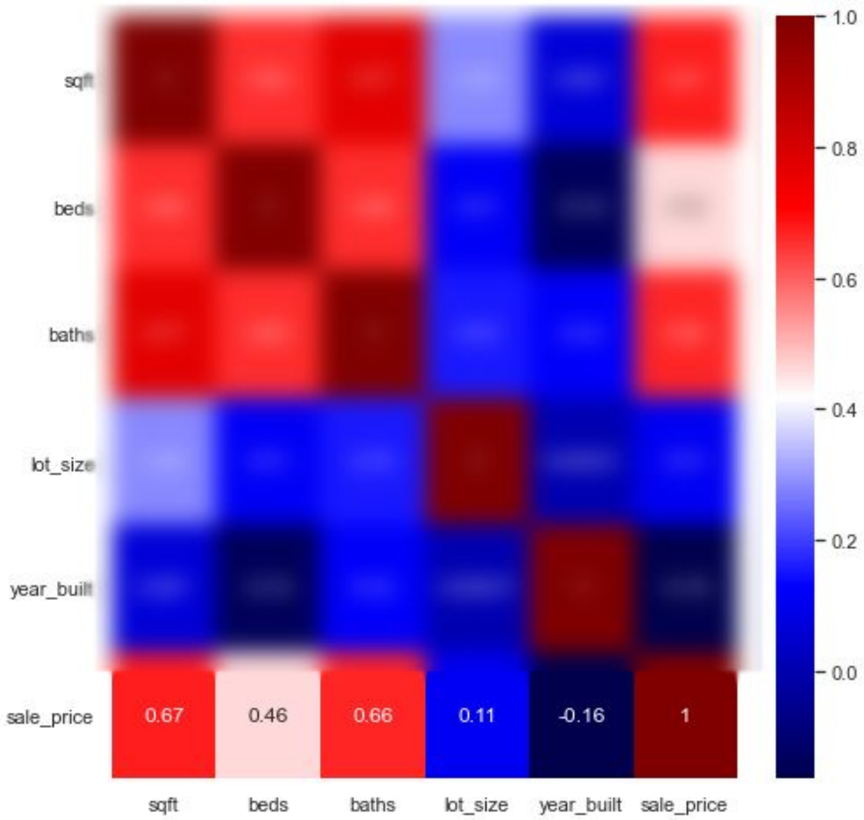


Initial observations: Correlation

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 correlated 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.

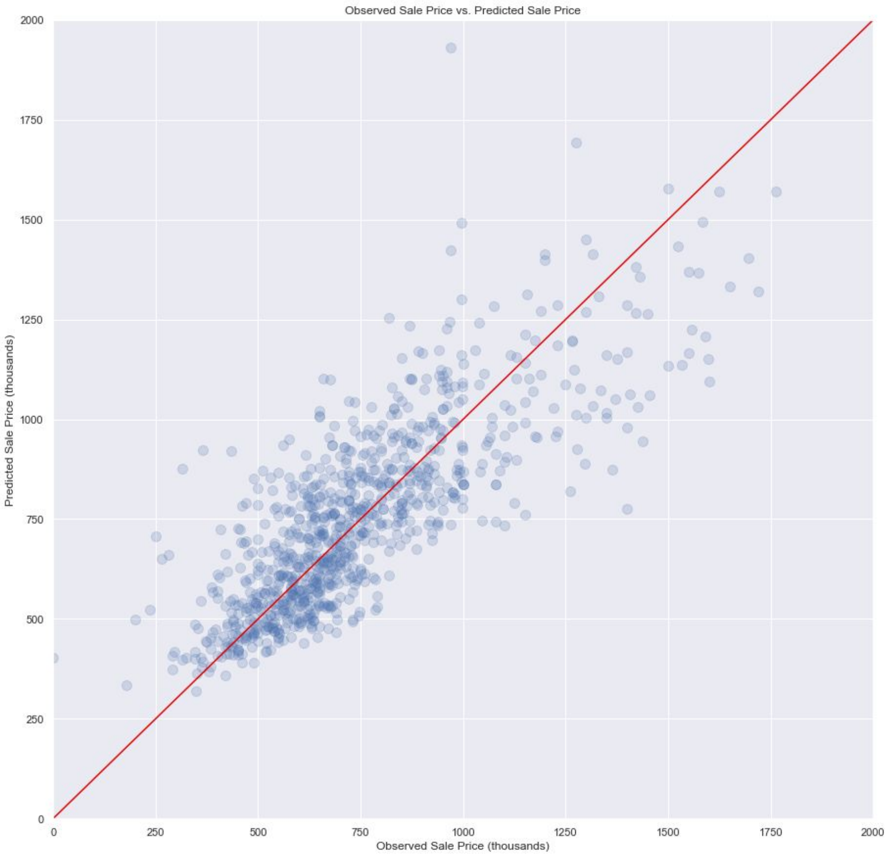
For the training set, r^2 was 0.8067.

For the test set, r^2 was 0.8081.

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.



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

Sample Prediction

2025 sqft.	2352 sqft. lot size
4 bd. 2.5 bath.	Built in 1986
Single Family Home	Located in Staten Island

Prediction:	Actual Sale Price:
$683,631.88	$700,000

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 [7]:
print(predict_house_price(2025,4,2.5,2352,1986,'Single Family Home','Bronx'))
print('\n')
print(predict_house_price(2025,4,2.5,2352,1986,'Single Family Home','Brooklyn'))
print('\n')
print(predict_house_price(2025,4,2.5,2352,1986,'Single Family Home','Manhattan'))
print('\n')
print(predict_house_price(2025,4,2.5,2352,1986,'Single Family Home','Queens'))
print('\n')
print(predict_house_price(2025,4,2.5,2352,1986,'Single Family Home','Staten Island'))
```

```
Input: 2025 sqft, 4 bedrooms, 2.5 bathrooms, 2352 lot size, built in 1986, property type Single Family Home, locate
d in Bronx, NYC
Predicted price:
605231.937856677


Input: 2025 sqft, 4 bedrooms, 2.5 bathrooms, 2352 lot size, built in 1986, property type Single Family Home, locate
d in Brooklyn, NYC
Predicted price:
1015662.4024963382


Input: 2025 sqft, 4 bedrooms, 2.5 bathrooms, 2352 lot size, built in 1986, property type Single Family Home, locate
d in Manhattan, NYC
Predicted price:
4618851.225159697


Input: 2025 sqft, 4 bedrooms, 2.5 bathrooms, 2352 lot size, built in 1986, property type Single Family Home, locate
d in Queens, NYC
Predicted price:
915611.0217106412


Input: 2025 sqft, 4 bedrooms, 2.5 bathrooms, 2352 lot size, built in 1986, property type Single Family Home, locate
d in Staten Island, NYC
Predicted price:
683631.8833368323
```

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

Additional Metrics (for nerds like me)

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

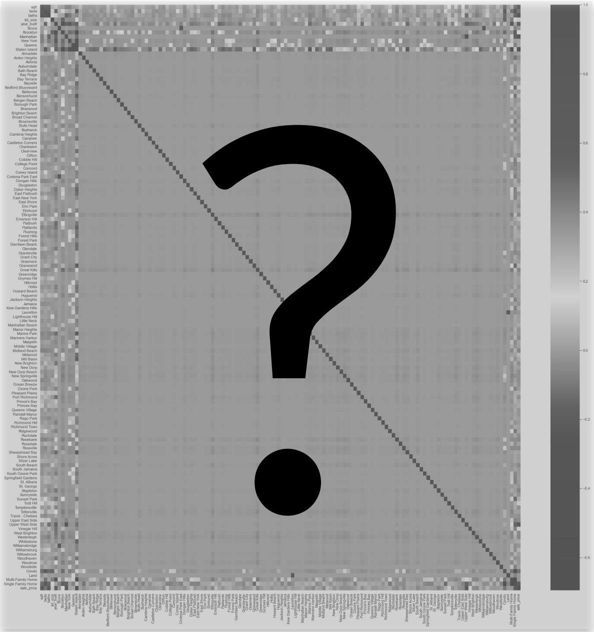
The RMSE was $254350.

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

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

Google Maps Geolocation API:

https://console.cloud.google.com/marketplace/product/google/geolocation.googleapis.com