Predicting New York City Housing Prices

Matthew Kwee

Aug. 2021

Introduction

Motivation:

Real-estate prices in New York City have changed significantly during the COVID-19 pandemic, and Menage LLC, a real-estate brokerage firm, needs a new model for predicting housing prices.

Objectives:

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 web-crawler that first catalogued webpages to scrape, then scraped those webpages.

Tools utilized:

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

Metrics:

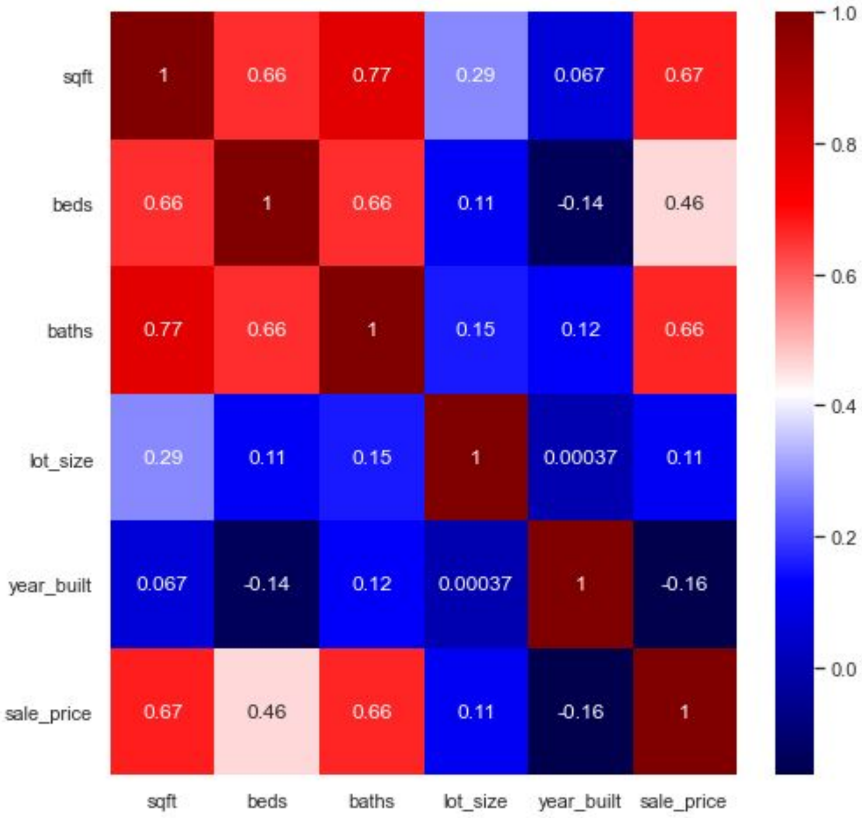
When it comes to predicting housing prices, maximizing the regression model's r^2 value will be the most important objective.

Initial observations: Correlation

I constructed a correlation heatmap of all continuous features.

Interestingly, lot size appears to have a slight negative effect on real-estate prices.

The larger the lot is, the harder it is to keep it in good condition.

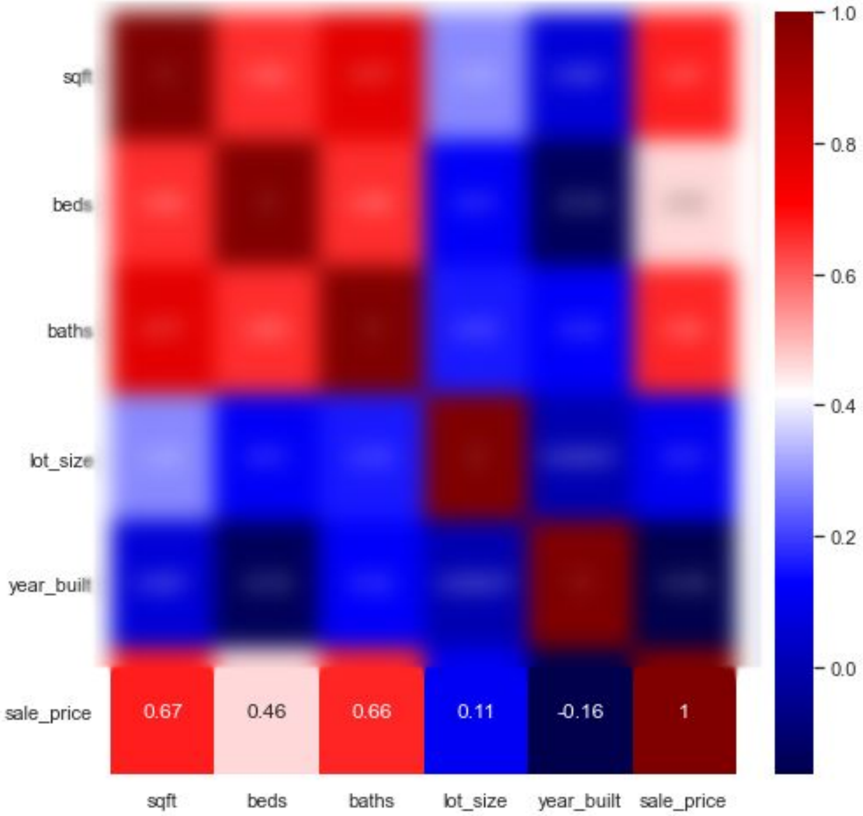


Initial observations: Correlation

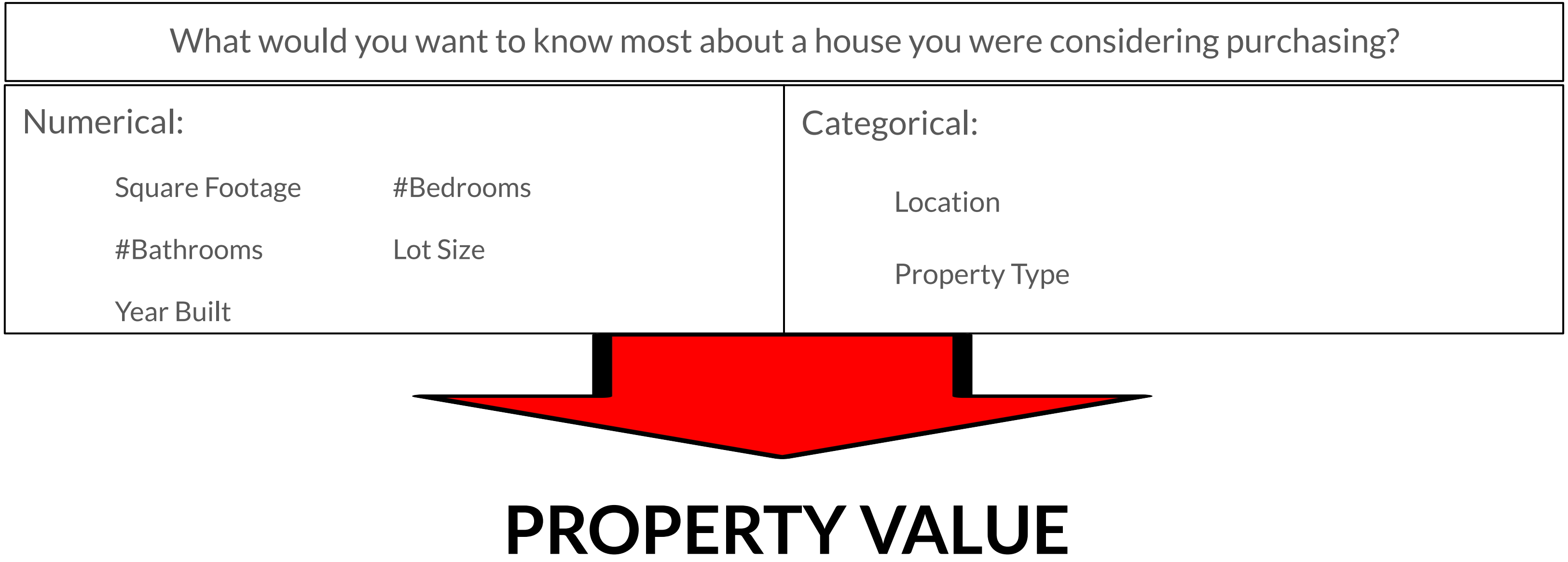
I constructed a correlation heatmap of all continuous features.

Interestingly, lot size appears to have a slight negative effect on real-estate prices.

The larger the lot is, the harder it is to keep it in good condition.



Features



What would you want to know most about a house you were considering purchasing?

Numerical:

Square Footage

#Bedrooms

#Bathrooms

Lot Size

Year Built

Categorical:

Location

Property Type

PROPERTY VALUE

Modelling

After converting the categorical variables into binary dummy variables, I ran Linear, Lasso, Ridge, and Elastic Net Regressions on the data.

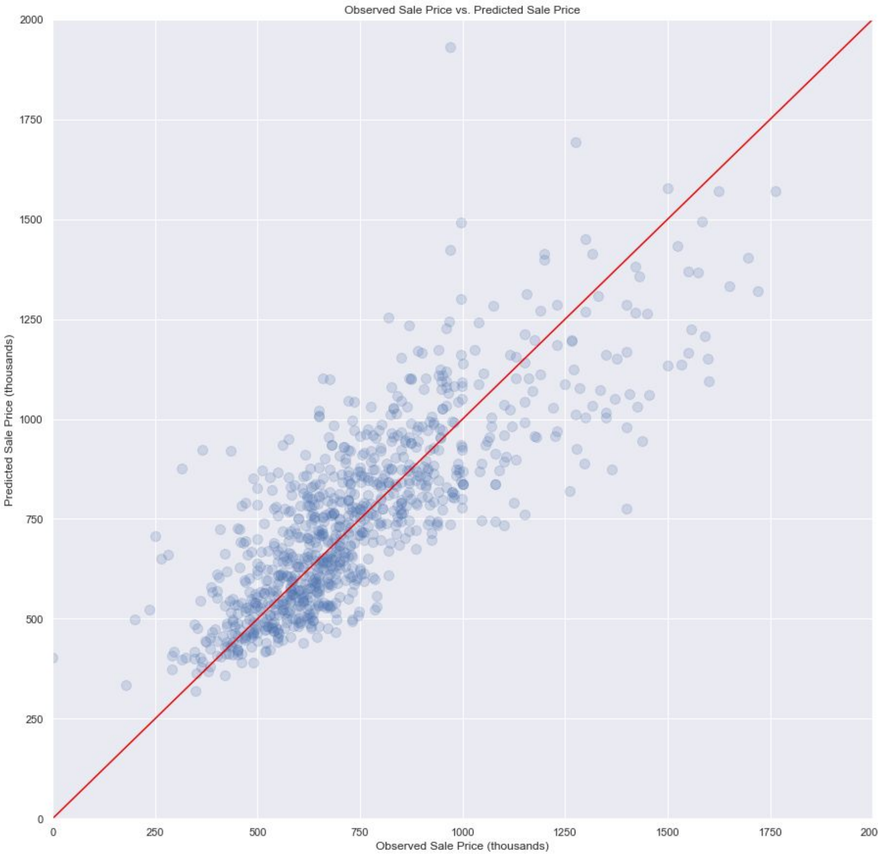
I settled on Linear because the model's r^2 value for the training set was not significantly higher than that of the test set.

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.

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

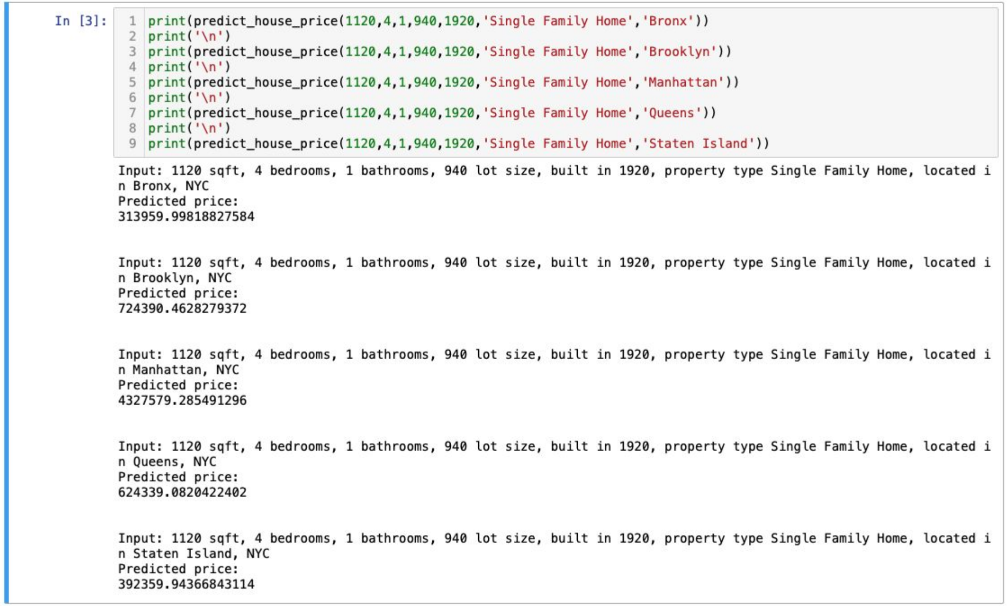


https://github.com/MK38993/Metis-Project-2---Real-Estate-Market/blob/main/NY_Housing_Market_Model.ipynb

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)

For the training set, r^2 was 0.8067.

For the test set, r^2 was 0.8081.

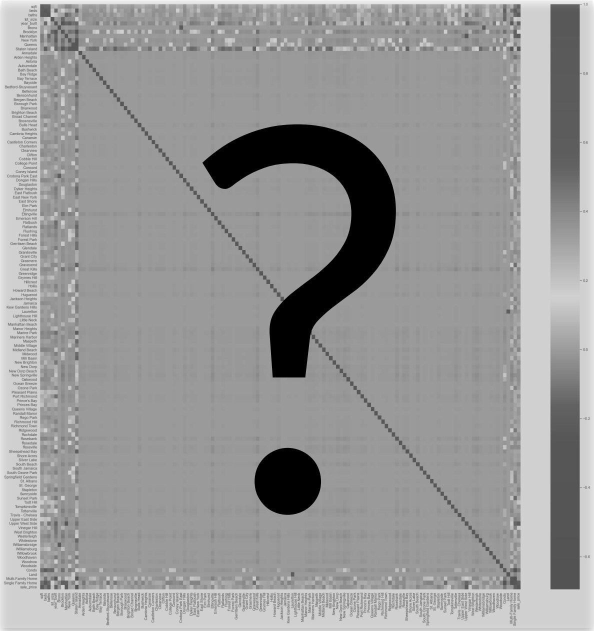
The RMSE was $254350.

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