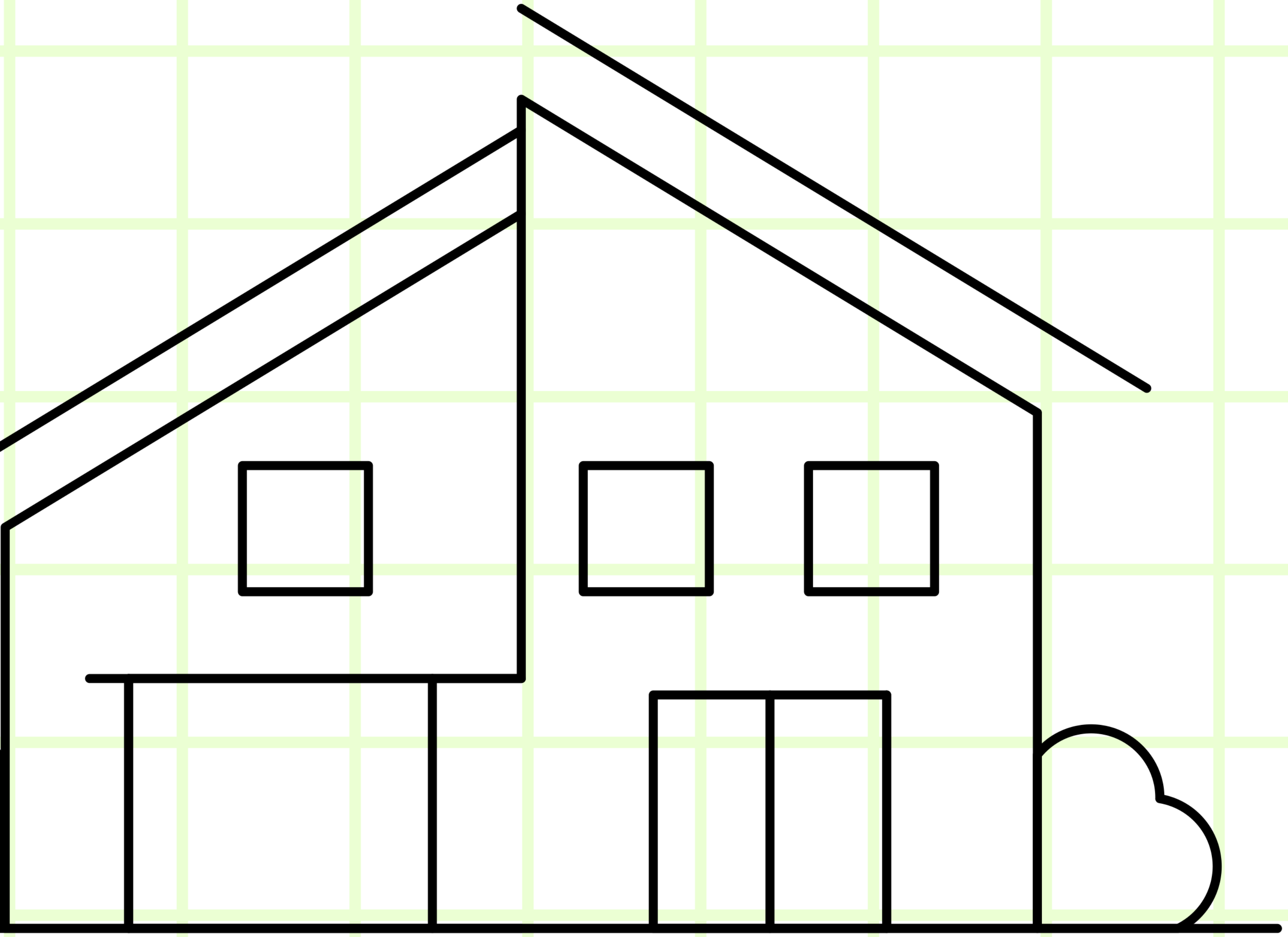


BrainStation®

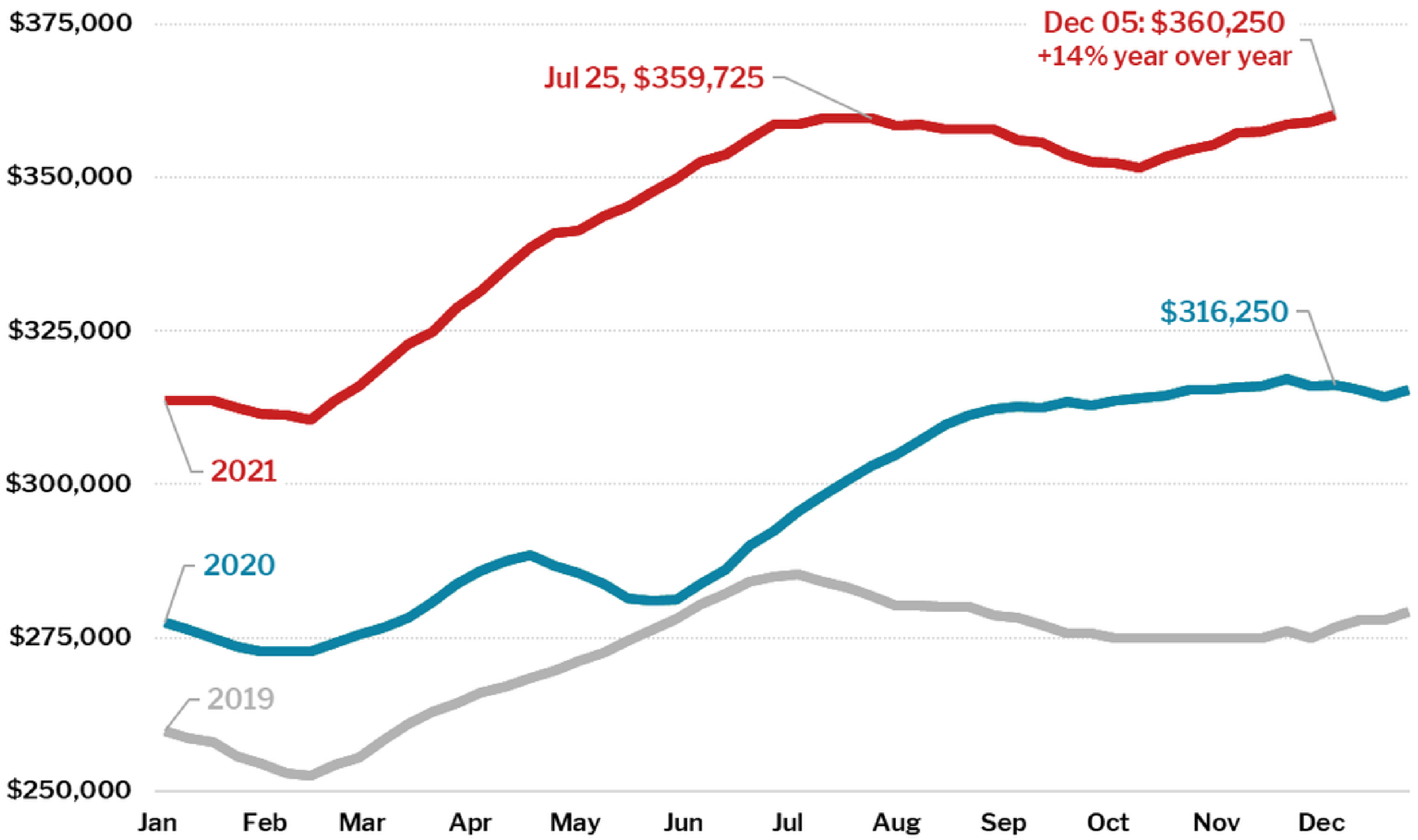
HOUSING MARKET AND LIVABILITY

Developing a model that evaluates the adaqaute price of housing units based on the livability score

HAESUN JUNG

Home Sale Prices Up 14% From 2020 to New Record High

4-week rolling average of the median sale price of homes sold



Dec 05: $360,250
+14% year over year

Jul 25, $359,725

$316,250

Source: Redfin analysis of MLS data

REDFIN

HOME SALES ARE AT 10 YEAR LOW
DON'T WORRY, WOODY, PRICES ARE STILL AT RECORD HIGHS
SPACE RANGER
LIGHTYEAR
imgflip.com

DATA PROCESSING:

BASE DATASET: REAL ESTATE SALES

RAW DATASET CONTAINS: 997K DATAPOINTS
TIMESPAN: 2001-2020
PLACE: CONNECTICUT

Columns with negligible missing values :

Serial Number
Year Listed
Date Recorded
Town
Address
Assessed Value
Sale Amount
Sales Ratio

Columns with considerable missing values :

Property Type
Residential Type
Non Use Code
Assessor Remarks
OPM remarks
Location (coordinates)

DATA PROCESSING

1. NON USE CODE
   a. One-hot-encoding
2. Assessor Remarks & OPM remarks
   a. Discovered 70000 unique string values in total
   b. Vectorization
3. Location
   a. Clean the address column and use Open Street Map API to obtain coordinates

DATA PROCESSING

1. Raw Data: 997213 rows
2. 16 Incorrect rows removed
3. Filter in the rows that have written remarks: 461892 rows
4. Use API to get coordinates for about 250000 rows

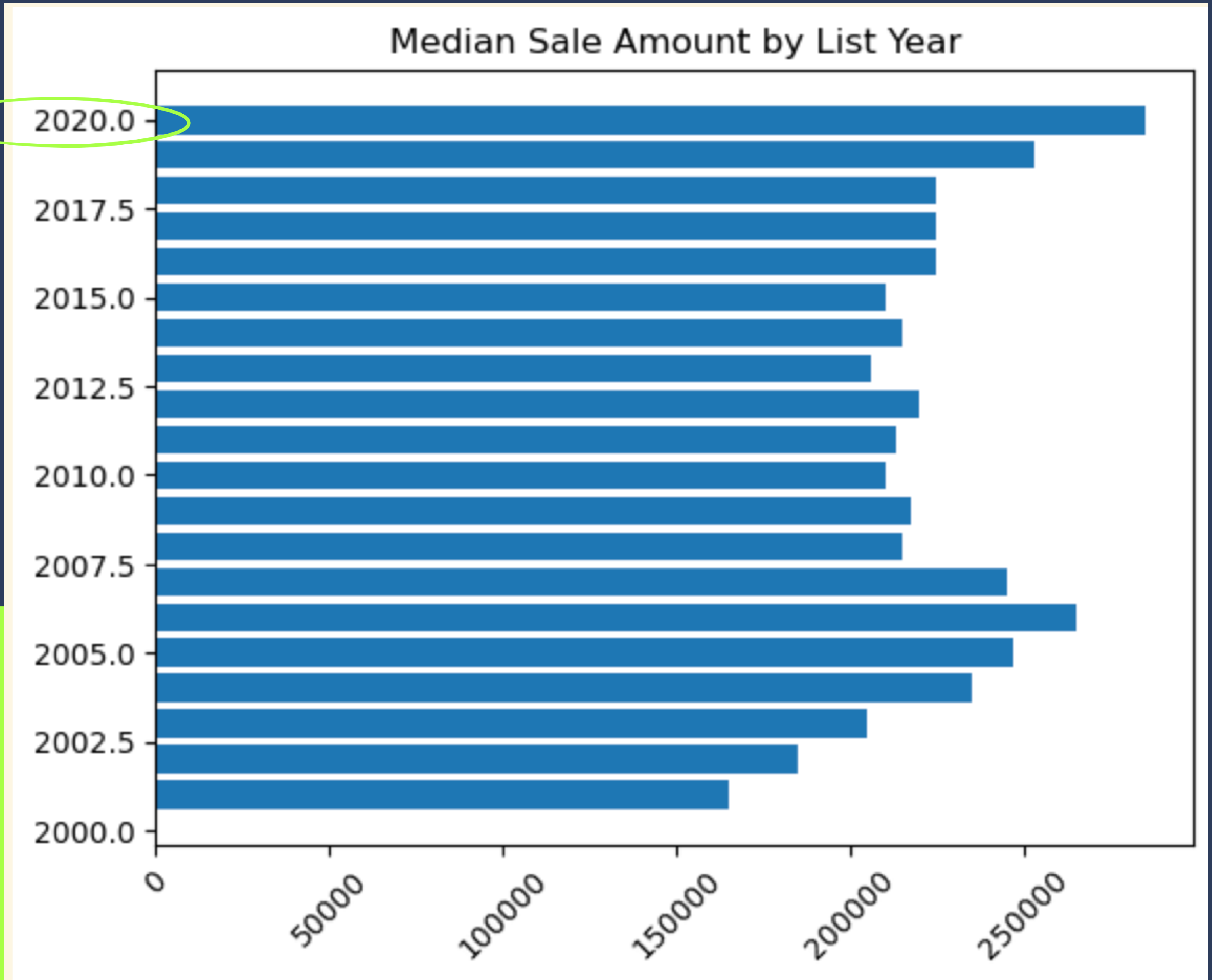
MODELS PERFORMED:

Baseline Regression Model:

Using data from the original datasets, R^2 of 0.20 on the test data.
Assessed Value, Year Listed, Town, Residential Type, Location (coordinates), and Non Use Code, all one-hot-encoded.

Baseline K-Best Analysis:

The top ten factors of highest coefficients are: 'latitude', 'longitude', 'List Year_2020', 'Town_Darien', 'Town_Greenwich', 'Town_New Canaan', 'Town_Stamford', 'Town_Westport', 'Property Type_Apartments', 'Property Type_Commercial'.


Median Sale Amount by List Year
2020.0
2017.5
2015.0
2012.5
2010.0
2007.5
2005.0
2002.5
2000.0
0
50000
100000
150000
200000
250000

LIST OF MODELS PERFORMED:

- **Lasso**
- **Ridge**
- **Decision Tree Regressor**
- **KNN Regressor**
- **Neural Network Regressor**

All produced R^2 score of about 0.2 on training data and 0 to -0.05 on test data EVEN with Assessed Value as one of the predictors, which shows how volatile and complex the housing market is.

NEXT STEP

- **Interest Rate & Housing Inflation Rate**
- **TF-IDF or TextVectorization from Tensor Flow**
- **continue optimization through random search / grid search, or manual adjustments.**
- **Sociological factors: neighbourhood crime rate, household income**
- **Livability: public transportation, health care, healthy food sources, parks**
- **Adopt methodologies from academic papers**

Since the sale price was difficult to be predicted within the original dataset, it requires more feature engineering as well as other methods of vectorizing the text columns.