House Price Analysis in New York

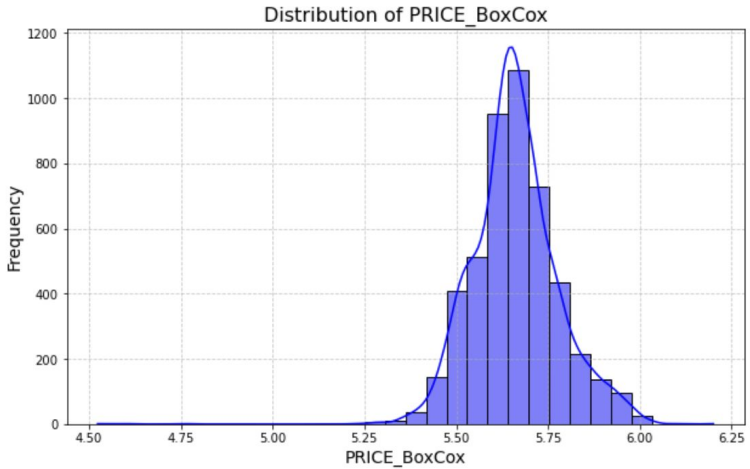
Charlie Ko, Changqi Jin, Ewan Li, Thomas Li, Sam Coomes

Data

New York Housing Market
- Price analysis
- Clustering analysis

Independent Variables we are interested in:
- Type of the house
- Location
- Number of Bed/Bath
- Sqft

Data Transformation



$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

	TYPE	BEDS	BATH	PROPERTYSQFT	LOCALITY	PRICE_BoxCox
0	Condo for sale	2	2.000000	1400.0	New York	5.530535
1	Condo for sale	7	10.000000	17545.0	New York	6.097215
2	House for sale	4	2.000000	2015.0	New York	5.503368

(4801,6)

Linear model

Linear regression

- Without categorical data:
  - R-square: 0.368; MSE: 0.094; Running time: 0.013s.
- With categorical data: (one-hot encoding):
  - R-square: 0.574; MSE: 0.077; Running time: 0.022s.

Non-linear model

Support Vector Machine for Regression : SVR(kernel='linear', C=1)

- 61.84 seconds, with only 240 rows
- Too expensive

Non-linear model

We choose a faster way, GradientBoostingRegressor.

It first use a decision tree to make a regression of the data, then use another decision tree to make a regression of the residual and add the new prediction function to the former with a certain learning rate.

Repeat above steps several times(the parameter ‘n_estimators’), and then we get a final prediction function.

Non-linear model

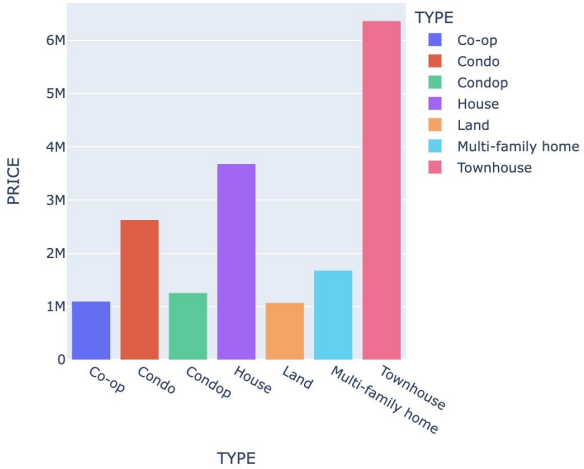
GradientBoostingRegressor

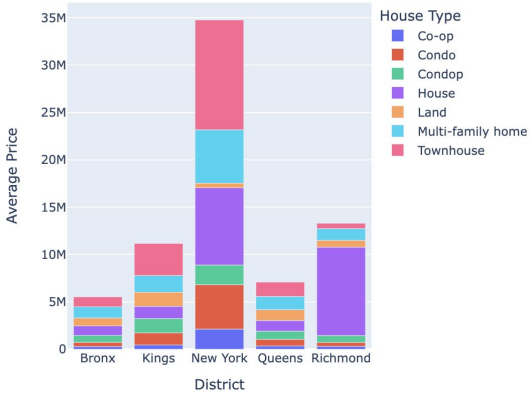
- Hyperparameters:
  - learning rates=np.arange(0.0001,1,0.1)
  - n_estimators=np.arange(100,1000,100)
- R square precise: 0.75
- n_estimators:300
- Learning rate: 0.20

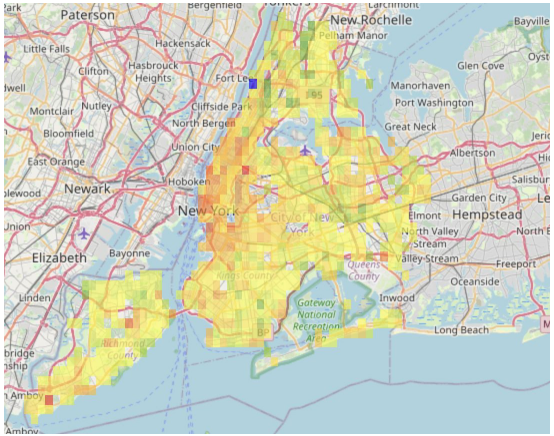
But it is too slow, the total time is 205.2

If we use GridSearchCV, the time is 488s, and result is not precise. The learning rate is 200, the R square is 0.73.

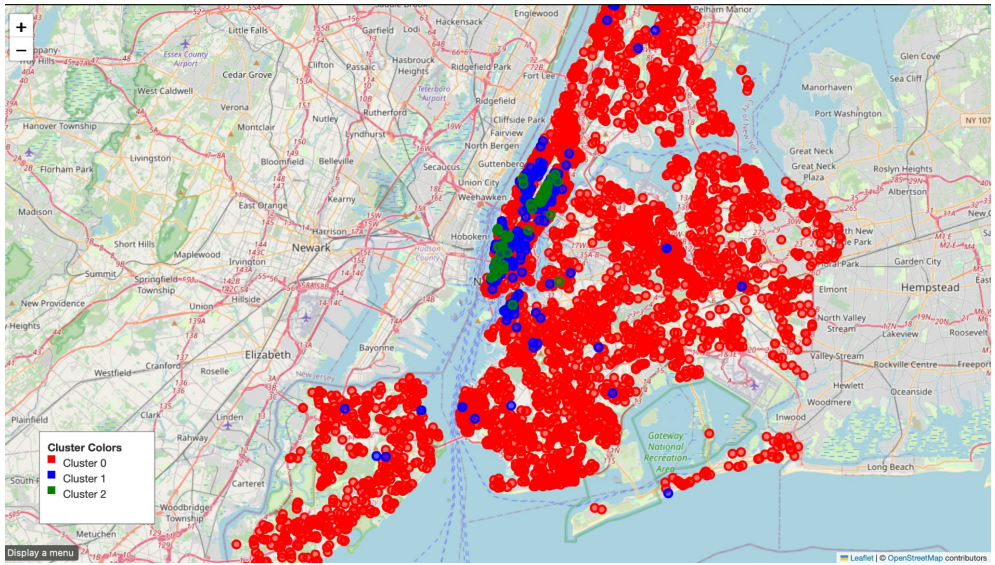
Visualization

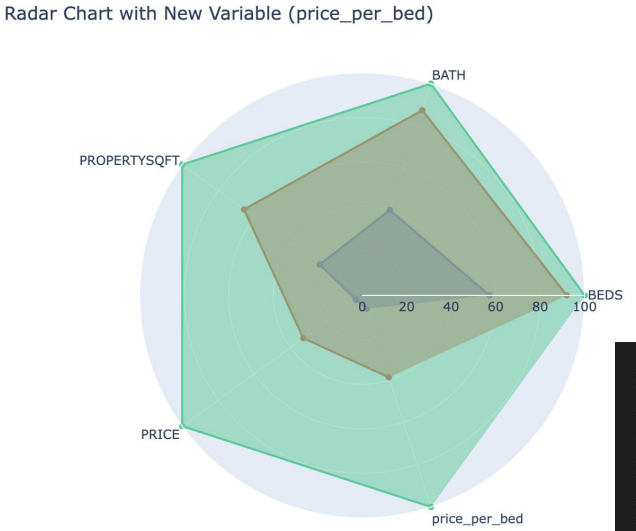






Cluster analysis





Cluster	
0	4509
1	239
2	51

Discussion

Limitations:

- **4860 rows of pricing data may not encapsulate entire NY Market**
  - 3.7 Million Housing Units
- **Model cannot be scaled or generalized to other States/Cities**
  - Weights and Params distinct to NY
- **Too many variables for inexpensive model**
- **Inability to capture historical trends**
  - Model is based on snapshot of current data, historical price fluctuations not included.

Future Topics:

- **Enhanced Feature Selection**
  - Improve accuracy and computational cost
- **Larger Dataset Integration**
  - Increased robustness & scalability
- **Real Time Analysis**
  - Continuous data collection
  - Faster processing times
  - API Connectivity for Developers/Researchers
- **External Data Implementation**
  - Seasonal factors/patterns
  - Addition of economic data