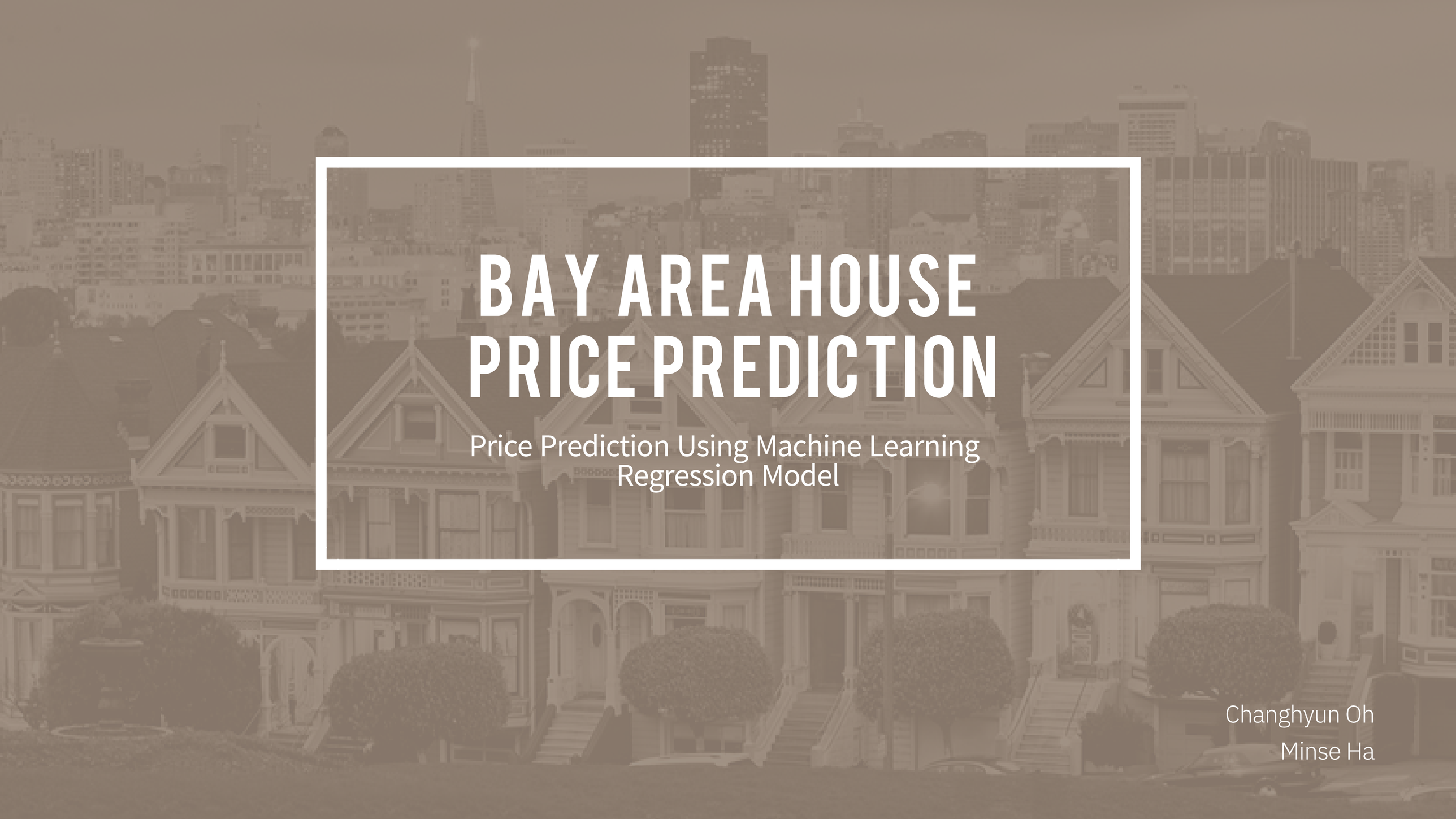


BAY AREA HOUSE PRICE PREDICTION

Price Prediction Using Machine Learning Regression Model

Changhyun Oh

Minse Ha

Goal of Project

Introduction

Background

Property prices in the Bay Area have be
en rising rapidly since 2012.
In particular, some regions show an incr
ease of more than three times

Hardship

With soaring housing prices, it has beco
me quite difficult for potential home buy
ers to predict the price of house

Goal

Predict the price of house using some M
achine Learning Regression Models,
which provides appropriate price inform
ation to potential buyers.

Outcome

Based on this price prediction, buyers
will be able to make reasonable trading
decisions in SF Bay Area

SF bay area prices dataset

Data Description

	Address	City	State	Zip	Price	Beds	Baths	Home size	Lot size	Latitude	Longitude	SF time	PA time	School score	Commute time
0	2412 Palmer Ave	Belmont	CA	94002	1459000	3	2.00000	1360.00000	5001.00000	37.51678	-122.30462	63	33	77.90000	33
1	1909 Hillman Ave	Belmont	CA	94002	1595000	4	2.00000	2220.00000	3999.00000	37.52197	-122.29408	63	33	77.90000	33
2	641 Waltermire St	Belmont	CA	94002	899999	2	1.00000	840.00000	4234.00000	37.52023	-122.27314	63	33	77.90000	33
3	2706 Sequoia Way	Belmont	CA	94002	1588000	3	2.00000	1860.00000	5210.00000	37.52019	-122.30944	63	33	77.90000	33
4	1568 Winding Way	Belmont	CA	94002	1999000	4	3.50000	2900.00000	16117.20000	37.52428	-122.29124	63	33	77.90000	33
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7140	The Davis	Mountain House	CA	95391	603990	5	3.00000	2327.00000	NaN	37.75644	-121.54772	120	125	65.30000	120
7141	The Berkeley	Mountain House	CA	95391	619990	5	4.00000	2410.00000	NaN	37.75644	-121.54772	120	125	65.30000	120
7142	Geranium	Mountain House	CA	95391	666340	5	4.00000	2486.00000	NaN	37.76472	-121.53776	120	125	65.30000	120
7143	The Pepperdine	Mountain House	CA	95391	659990	5	4.00000	2856.00000	NaN	37.75644	-121.54772	120	125	65.30000	120
7144	The Stanford	Mountain House	CA	95391	644990	5	4.00000	2679.00000	NaN	37.75644	-121.54772	120	125	65.30000	120

7145 rows × 15 columns

SF bay area dataset

Size	Featuers	y_feature	Datatype
7,145	15	Price	float, Int, Object

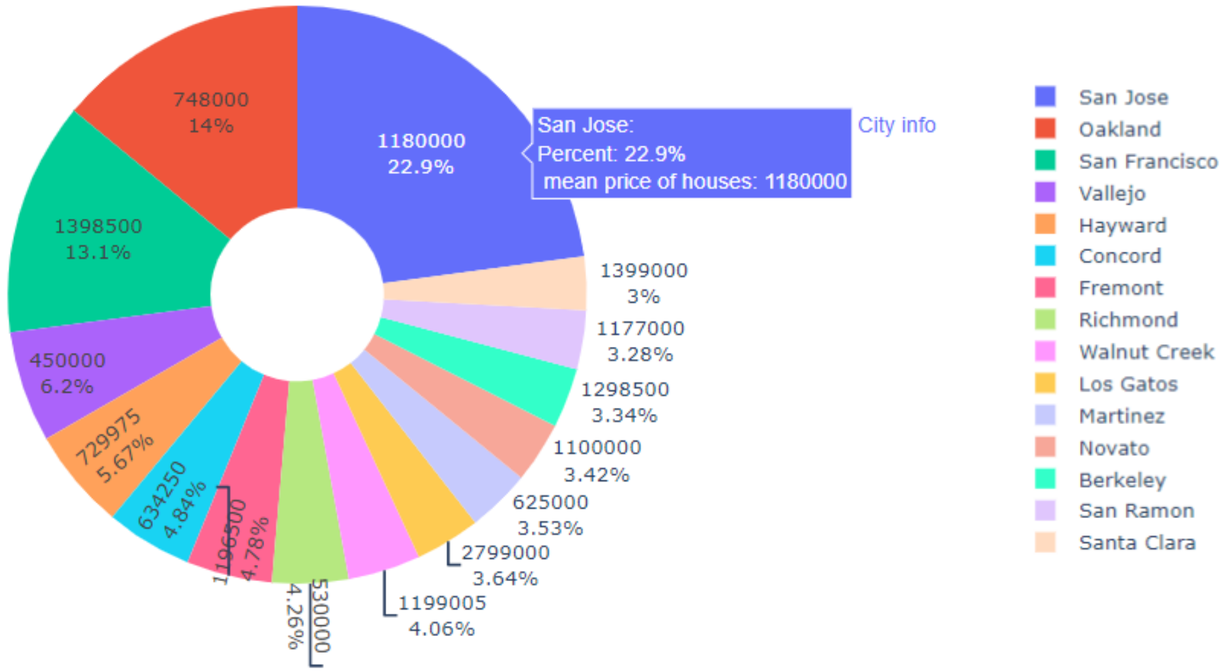
The shape of the data is 7,145 rows and has 15 columns.
One column is a Price column which will be predicted by our models, and others are input features for the train and predict.
Three columns(address, city, state) are String data types, and the other eleven are numerical. Let's take a closer look for the data.

DATA VISUALIZATION

Visualization 1

Which city has the most houses?

Number of houses/Mean house price from the top 15 cities



San Jose has the most houses.

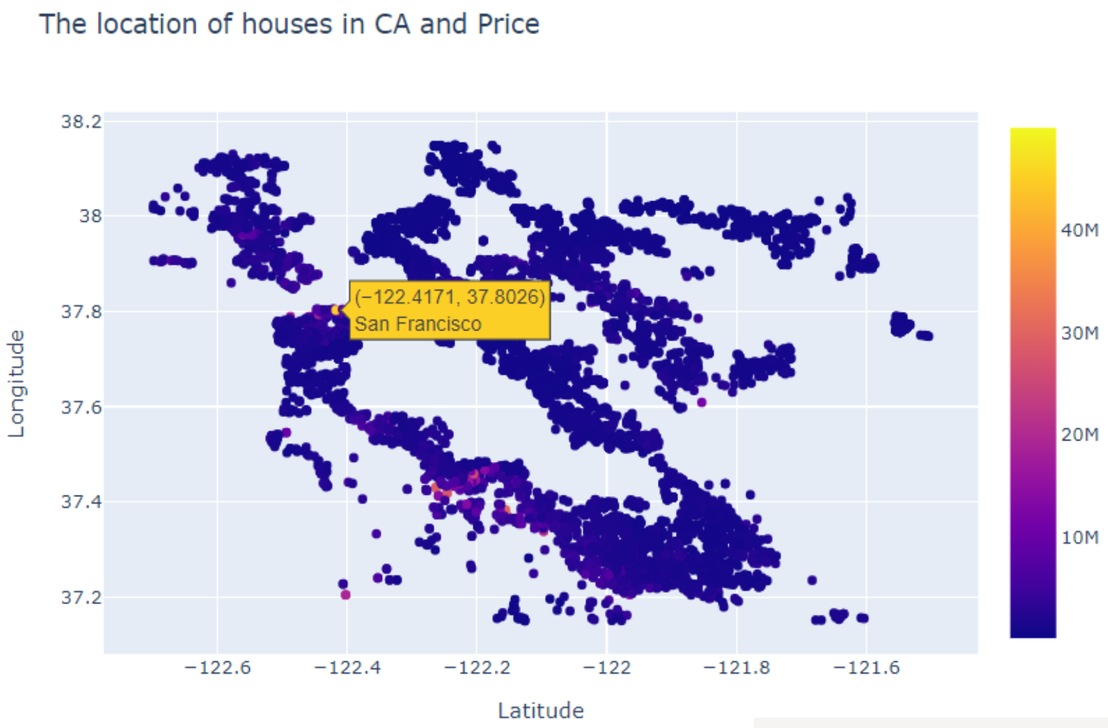
Median house value is highest in Santa Clara

Ranking of number of houses and ranking of median house values are not the same

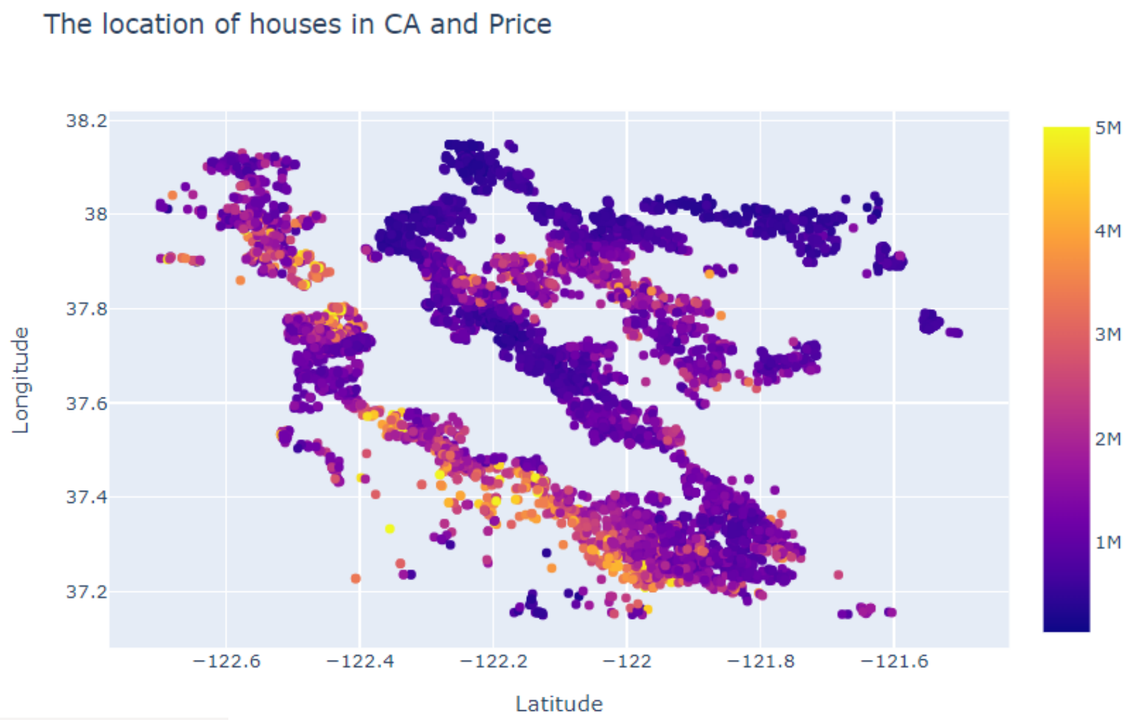
Visualization 2

Is there a correlation between location and house price?

Price Distribution (Entire Data)



Price Distribution (~5M)



Outlier data exists

Location is essential information

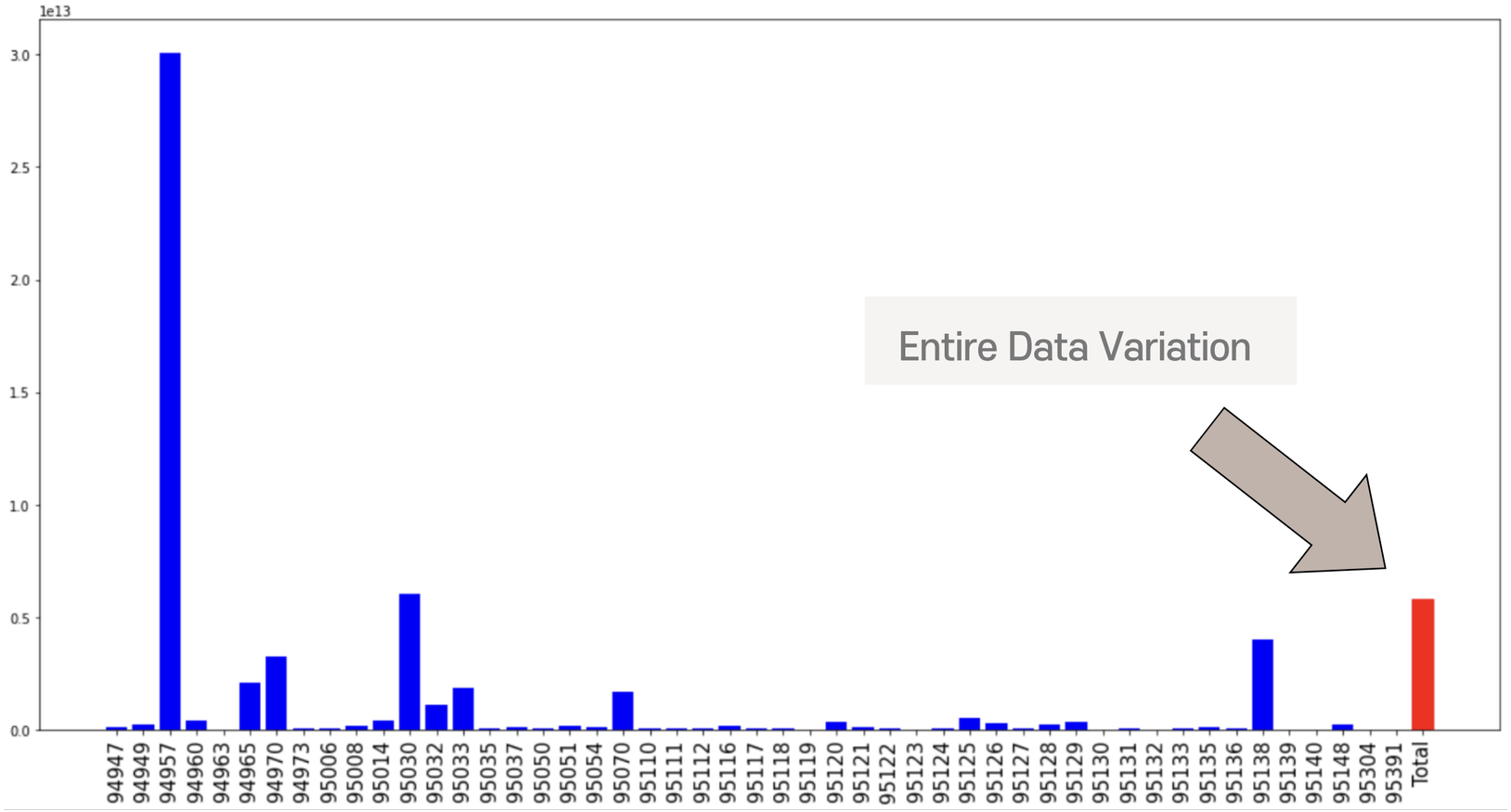
Visualization 3

Are houses with the same zip code priced similarly?

How do we judge 'similarity'?

Let's Check the Variation!

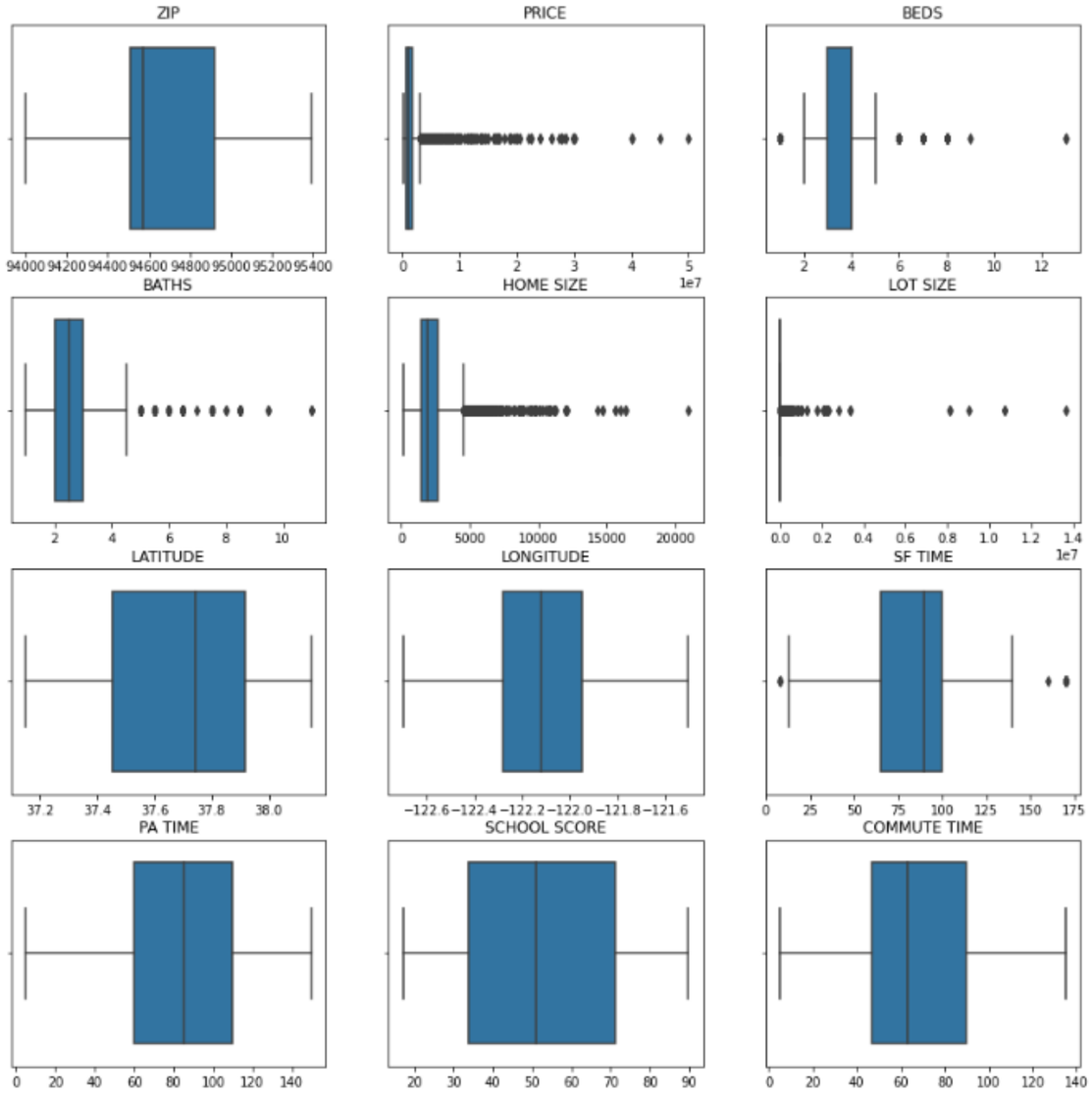
Each Zip code Variation = Entire Data Variation



91.16% of Zip code group has lower variation than the variation of entire data price

Visualization 4

Is there Outlier data?



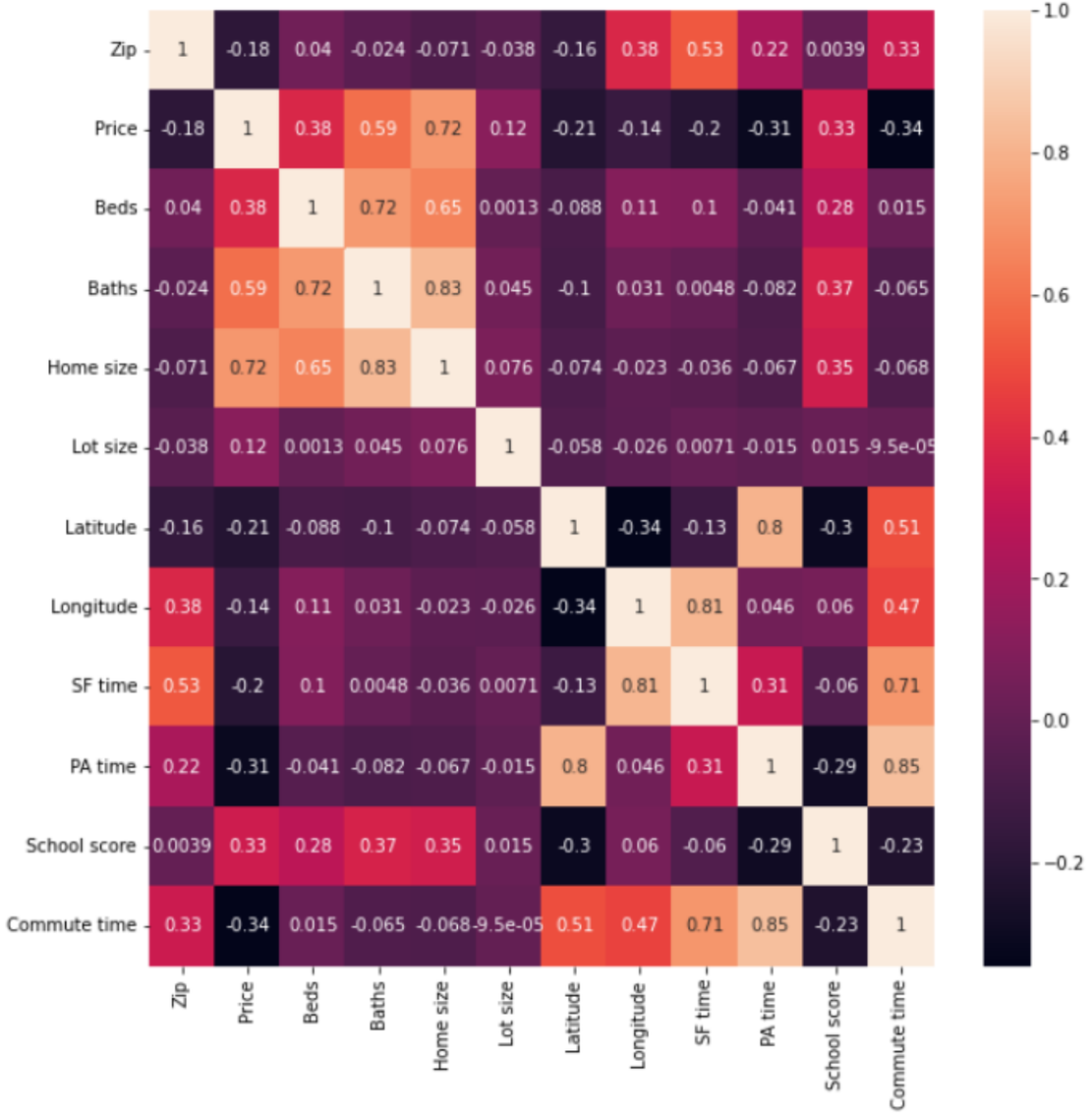
Outlier data exists in 'BEDS', 'BATHS',
'HOME SIZE', 'LOT SIZE'

Especially in Price,
There are lots of Outlier Data

Might need to deal with this data for better training

Visualization 5

Which of the numerical features correlates with price?



Plot A Heatmap With Correlation
(method = pearson)

Why ‘pearson’? → Works with raw data values of the variables while ‘spearman’ method works with rank-ordered variables

Insight

Figured out Home size, Baths, Beds have high correlation so definitely will use them as a feature in regressor.

Latitude and Longitude have low correlation even thought they are location features

PREPARE DATASET

SF bay area prices dataset

Split Data with Train set & Test set

SF bay area dataset

Total Size	Train set	Test set	Validation set
7,145	5,716 (80%)	1,429 (20%)	None

Why there is no Validation set

Data is too small

We will use Cross Validation

What is Cross Validation ?

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

PREPROCESSING DATA

SF bay area prices dataset

Delete Outlier using IQR(Interquartile Range)

Why IQR? : Since the IQR is simply the range of the middle 50% of data values, it's not affected by extreme outliers.

Column	Upper Outlier	Under Outlier	Total Outlier
Beds	198	45	243
Baths	338	0	338
Home size	0	0	0
Lot size	0	0	0
Latitude	0	0	0
Longitude	0	0	0
SF time	9	3	12
PA time	0	0	0
School Score	0	0	0
Commute time	0	0	0

Delete Outlier?

YES

- Outliers badly affect mean and standard deviation of the dataset
- These may statistically give erroneous results
- It increases the error variance and reduces the power of statistical tests

NO

- Some outliers represent natural variations in the population, then they should be left as is in your dataset.
- Dataset is too small to delete outlier(we will lose almost 10% of dataset!
- Don’t know that it’s completely wrong

SF bay area prices dataset

Create Data Pipeline

01. Pipeline for each Column

Column	Categorial Pipeline	Numerical Pipeline	Note
Beds		V	
Baths		V	
Home size		V	
Lot size		V	
Latitude		V	Drop or not
Longitude		V	Drop or not
SF time		V	
PA time		V	
School Score		V	
Commute time		V	
City	V		One hot Enc
Zip	V		One hot Enc

02. Catogorial Pipeline

- Imputer strategy = “median”

Due to bouncing data(didn’t use mean strategy)

- Minmaxscaler()

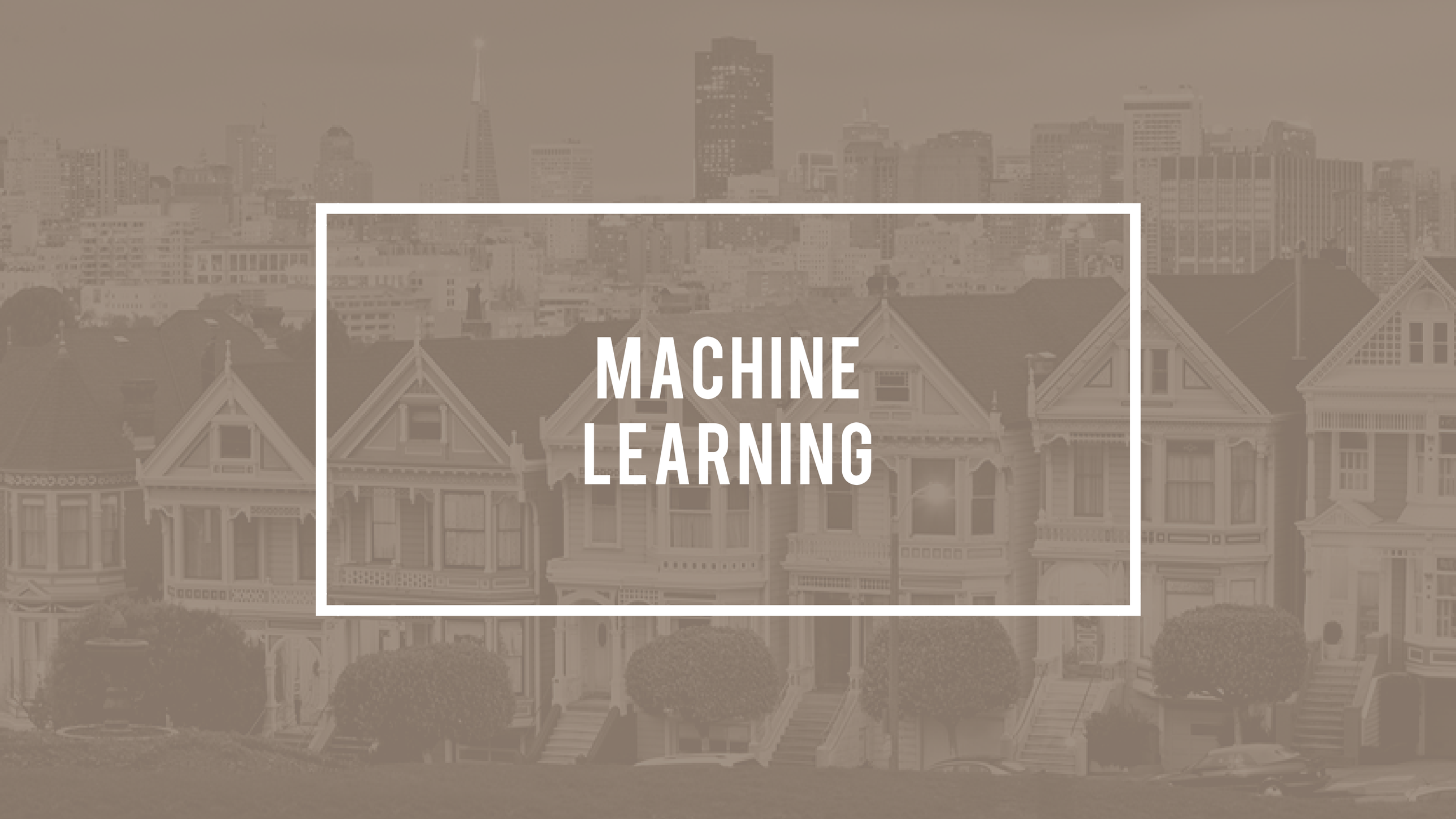
More than half of data features are skewed,

Standard scaler will not work well.

03. Numerical Pipeline

- Why zip code is for Categorical data?

When we take its numerical values, we will notice that zip code will not have any quantitative meaning.


MACHINE
LEARNING

SF bay area prices dataset

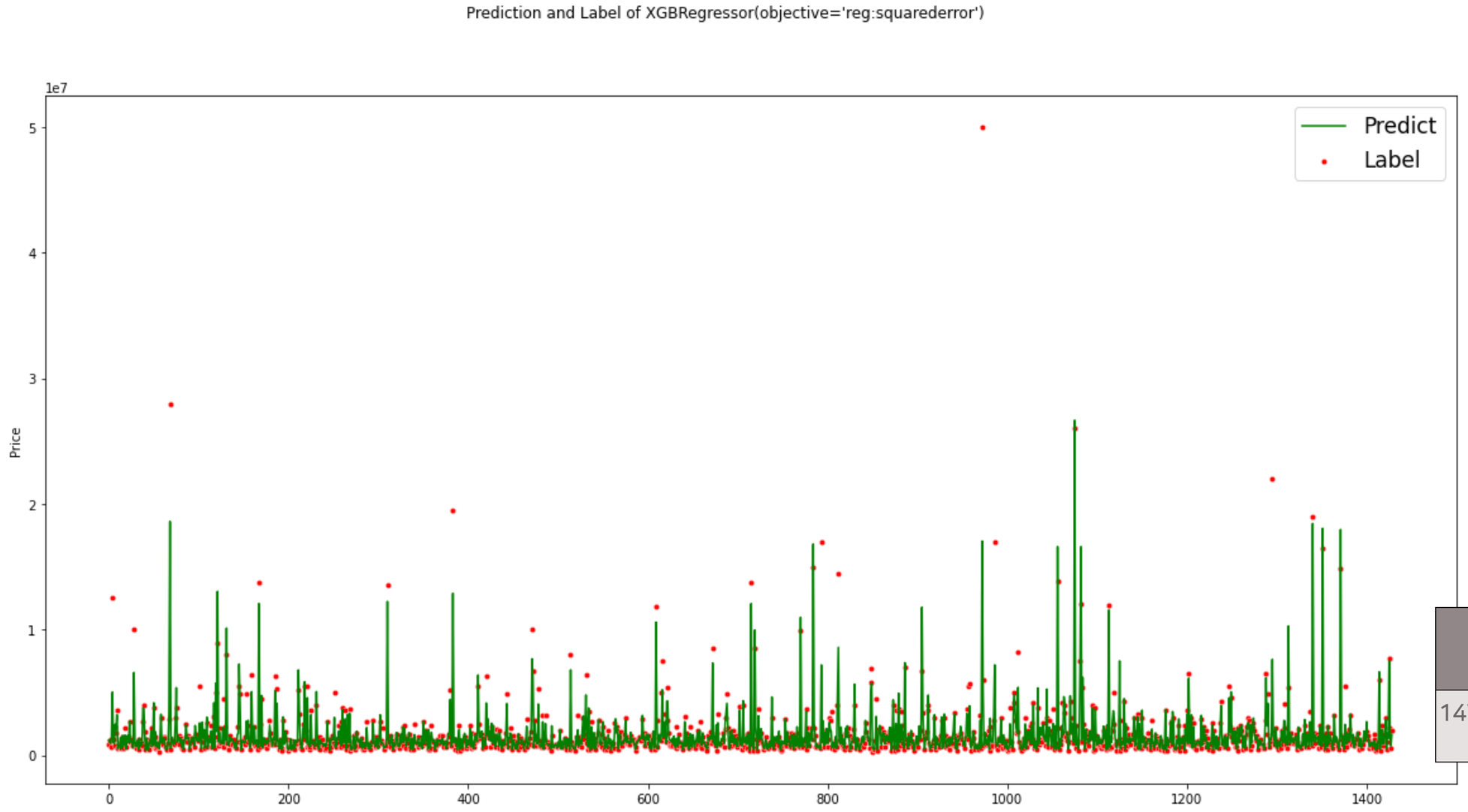
Lazy Prediction

Model	Ranking	Note
XGBRegressor	1	Linear model + Tree learning algorithms
BaggingRegressor	2	Ensemble with Decision tree
GradientBoostingRegressor	3	Prevents overfitting by averaging the results
RandomForestRegressor	4	DT with additional randomness
LGBMRegressor	5	Fast training, less memory

Lazy predicton result (metric : RMSE)

SF bay area prices dataset

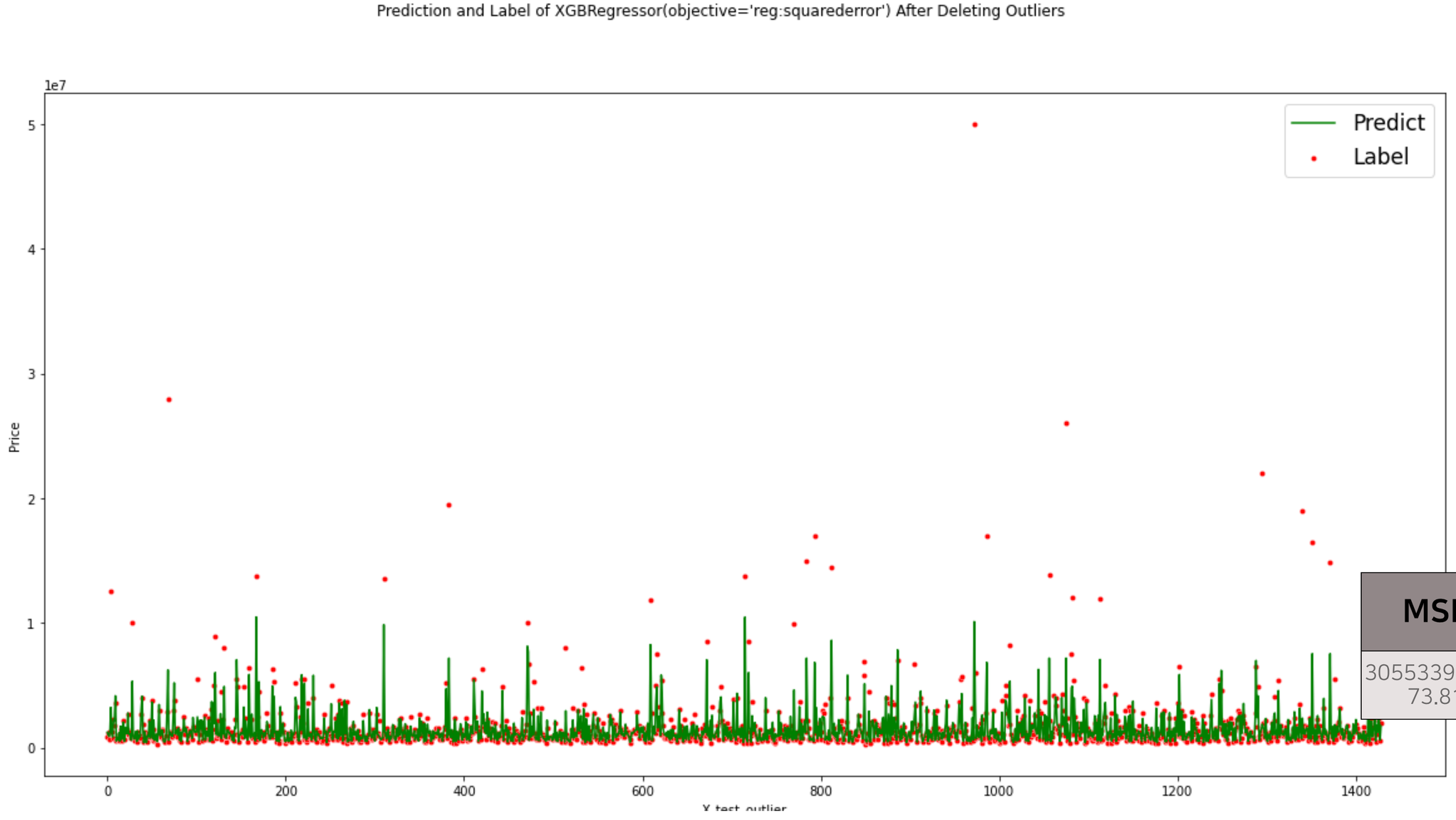
XGBRegressor (Outlier Delete = False)



MSE	RMSE	R2
1479722542137.59	1216438.47	0.77

SF bay area prices dataset

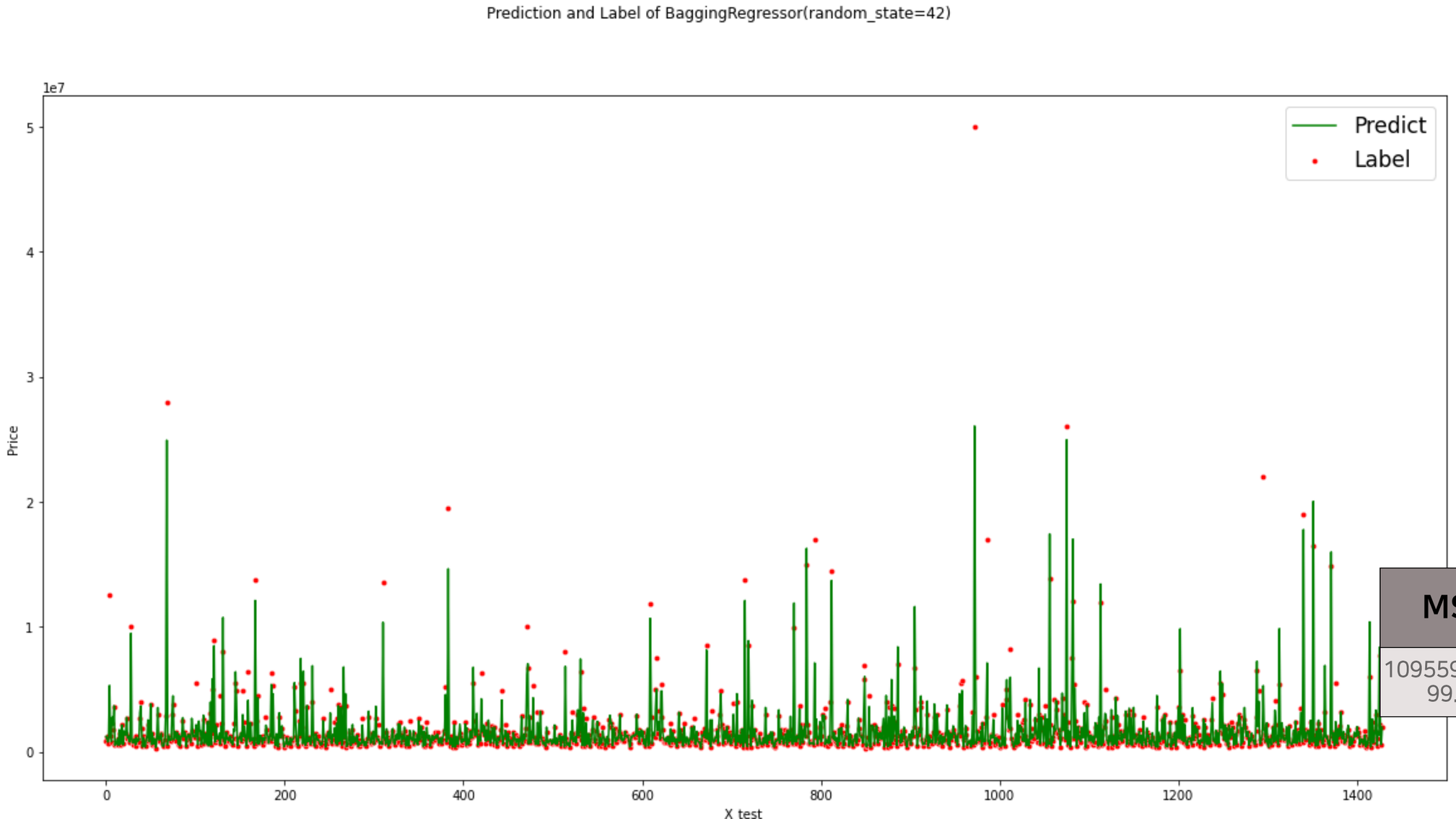
XGBRegressor (Outlier Delete = True)



MSE	RMSE	R2
3055339154673.81	1747952.85	0.52

SF bay area prices dataset

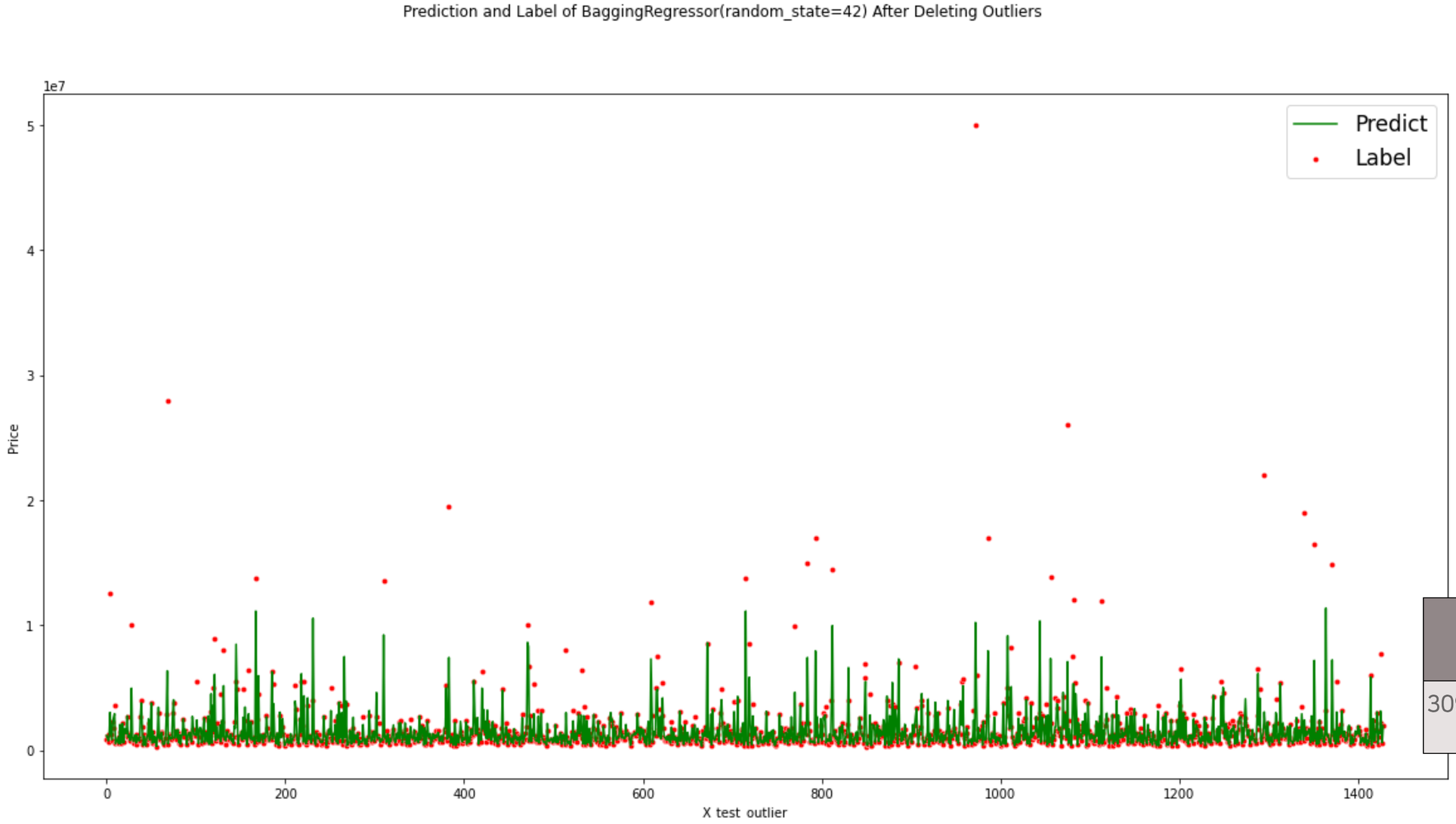
BaggingRegressor (Outlier Delete = False)



MSE	RMSE	R2
1095594500999.92	1046706.50	0.83

SF bay area prices dataset

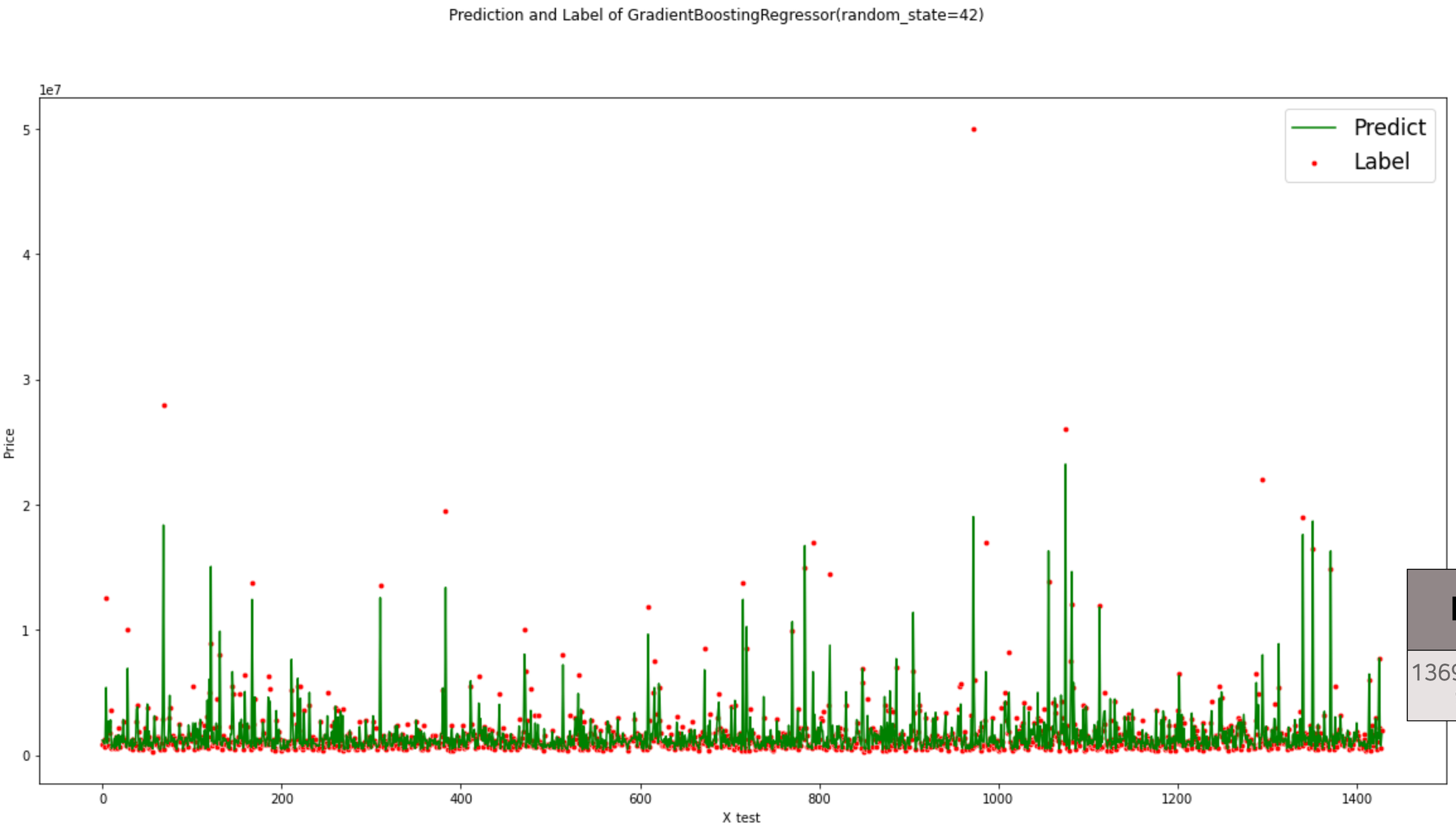
BaggingRegressor (Outlier Delete = True)



MSE	RMSE	R2
3093993481130.27	1758975.12	0.52

SF bay area prices dataset

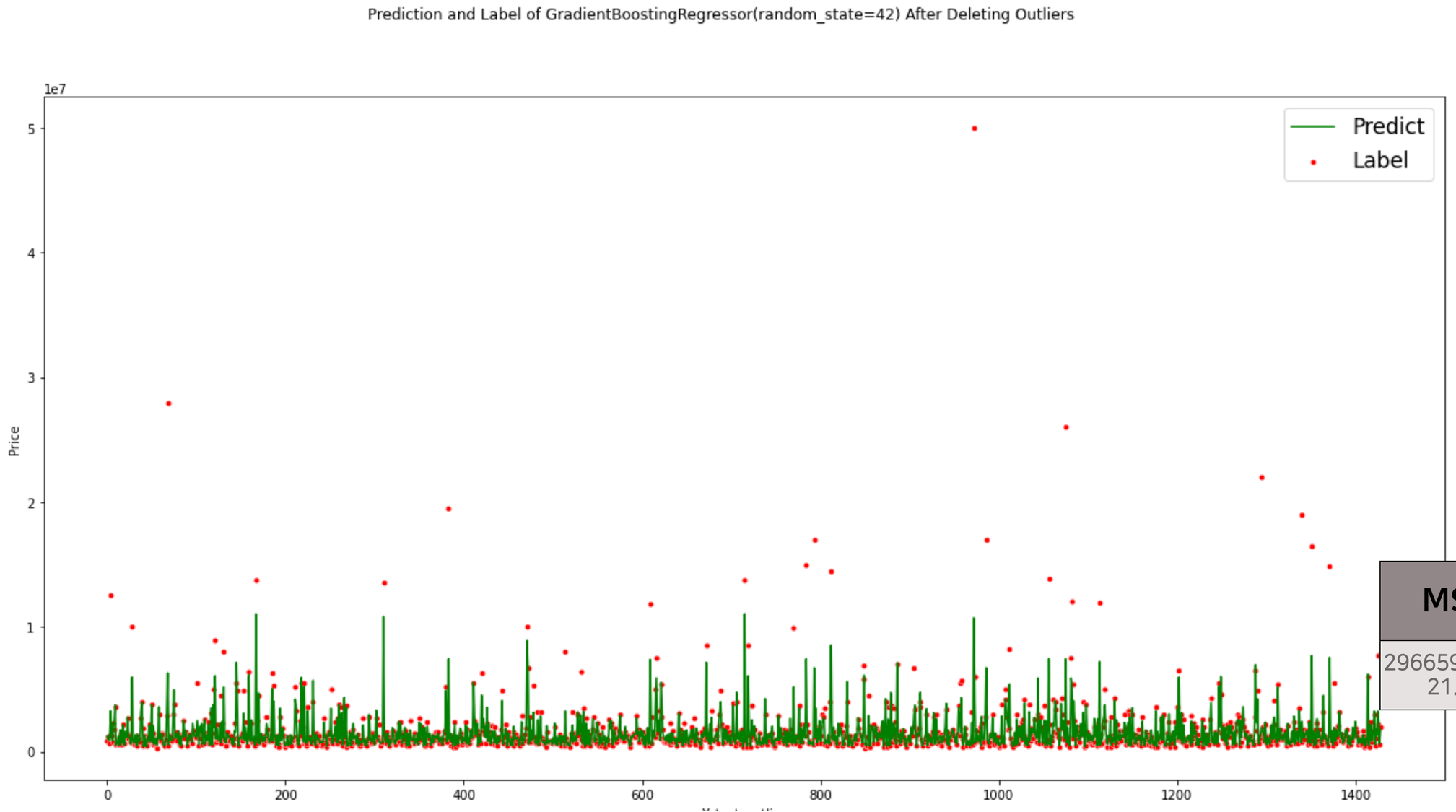
GradientBoostingRegressor (Outlier Delete = False)



MSE	RMSE	R2
1369623034395.87	1170308.95	0.79

SF bay area prices dataset

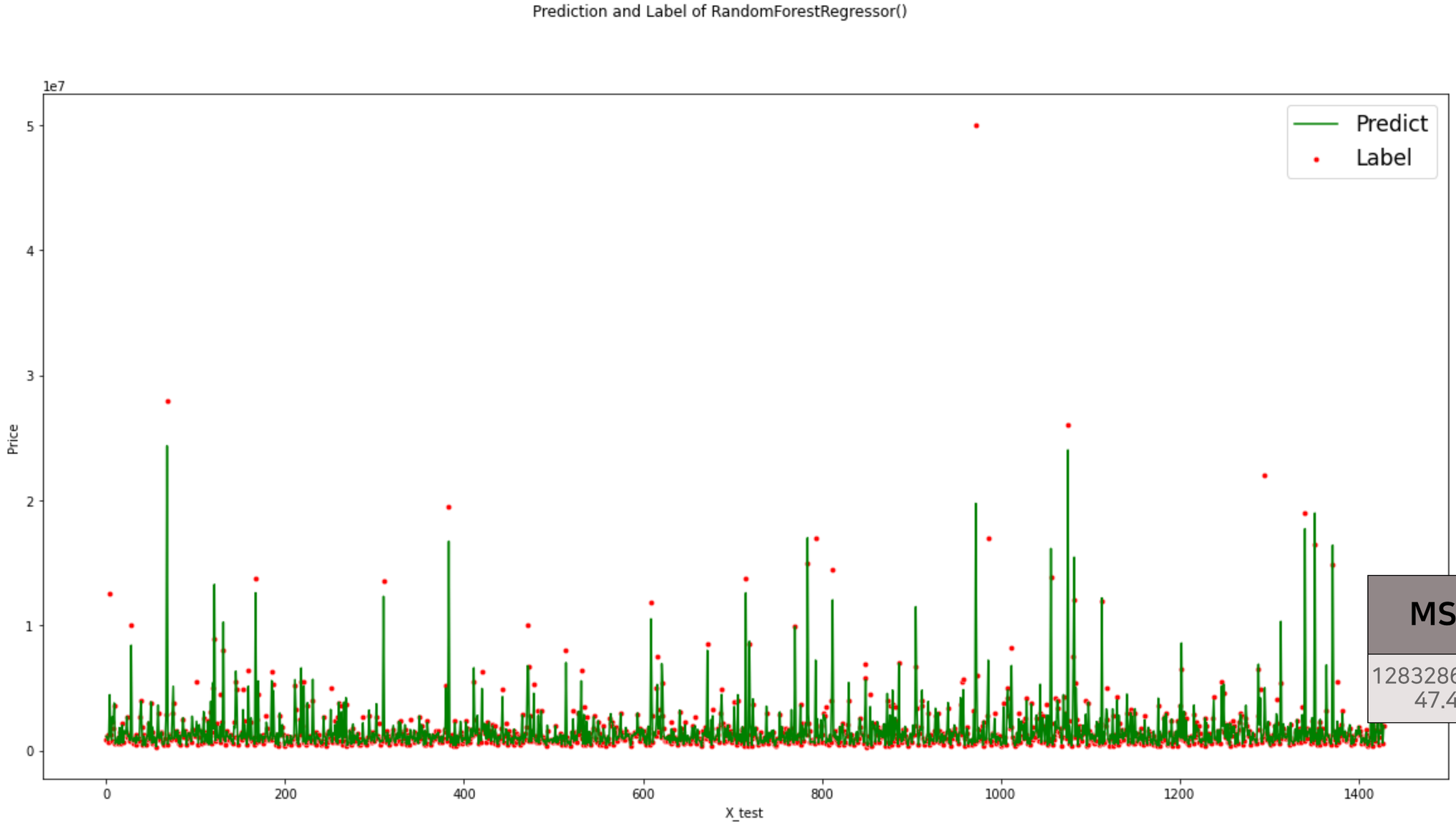
GradientBoostingRegressor (Outlier Delete = True)



MSE	RMSE	R2
2966594143621.28	1722380.37	0.54

SF bay area prices dataset

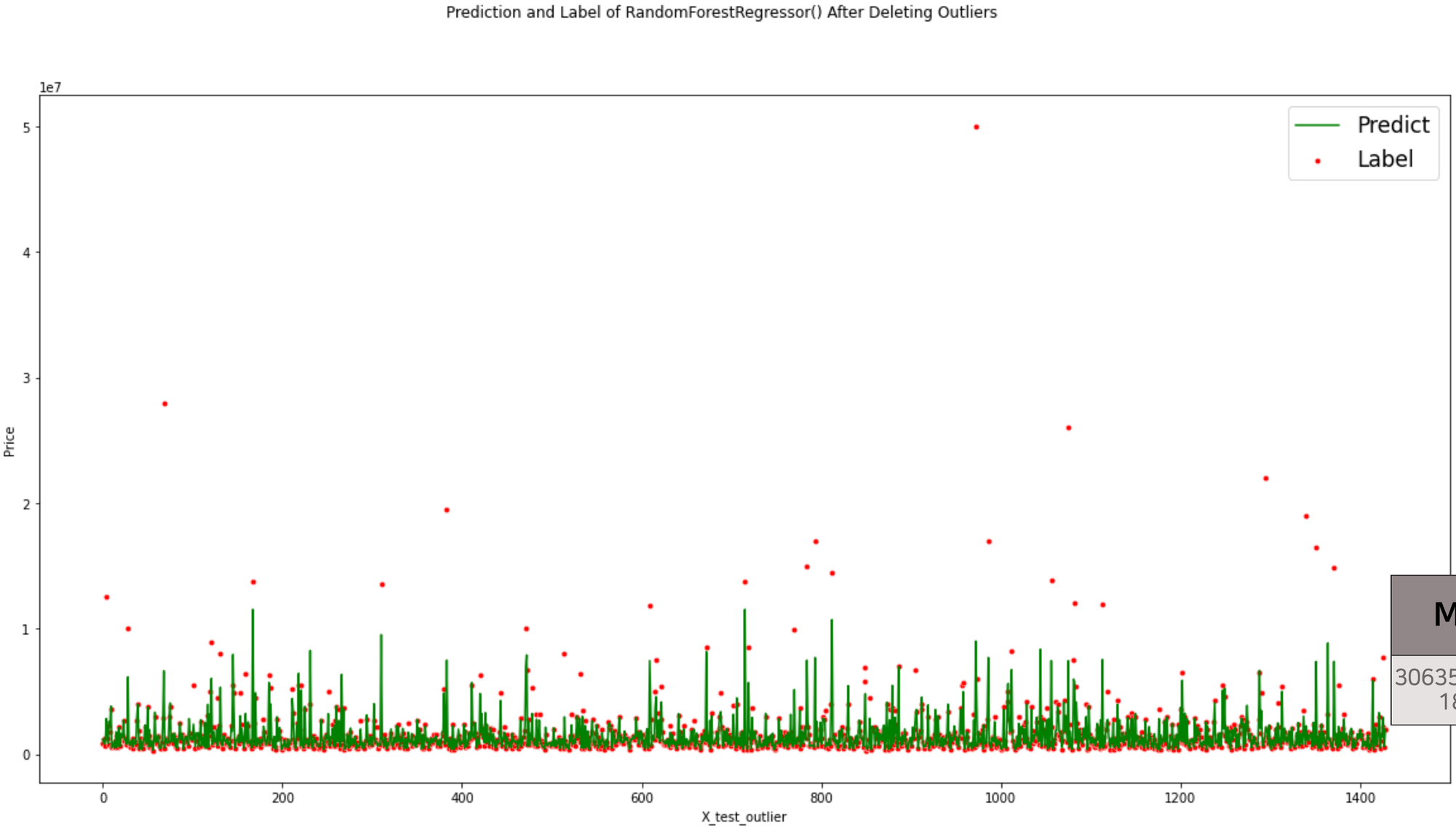
RandomForestRegressor (Outlier Delete = False)



MSE	RMSE	R2
1283286072847.42	1132822.17	0.80

SF bay area prices dataset

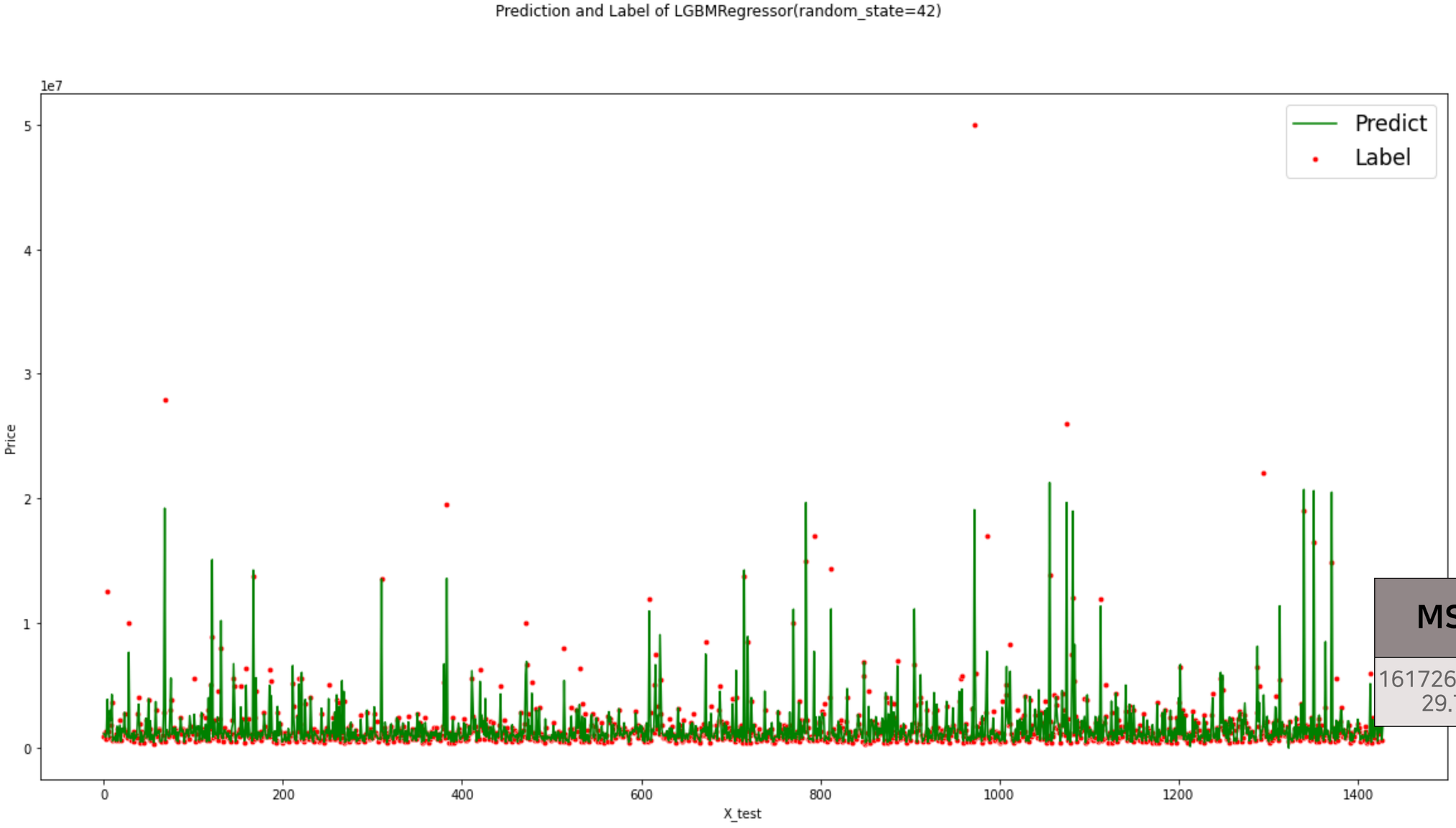
RandomForestRegressor (Outlier Delete = True)



MSE	RMSE	R2
3063582404018.10	1750309.23	0.52

SF bay area prices dataset

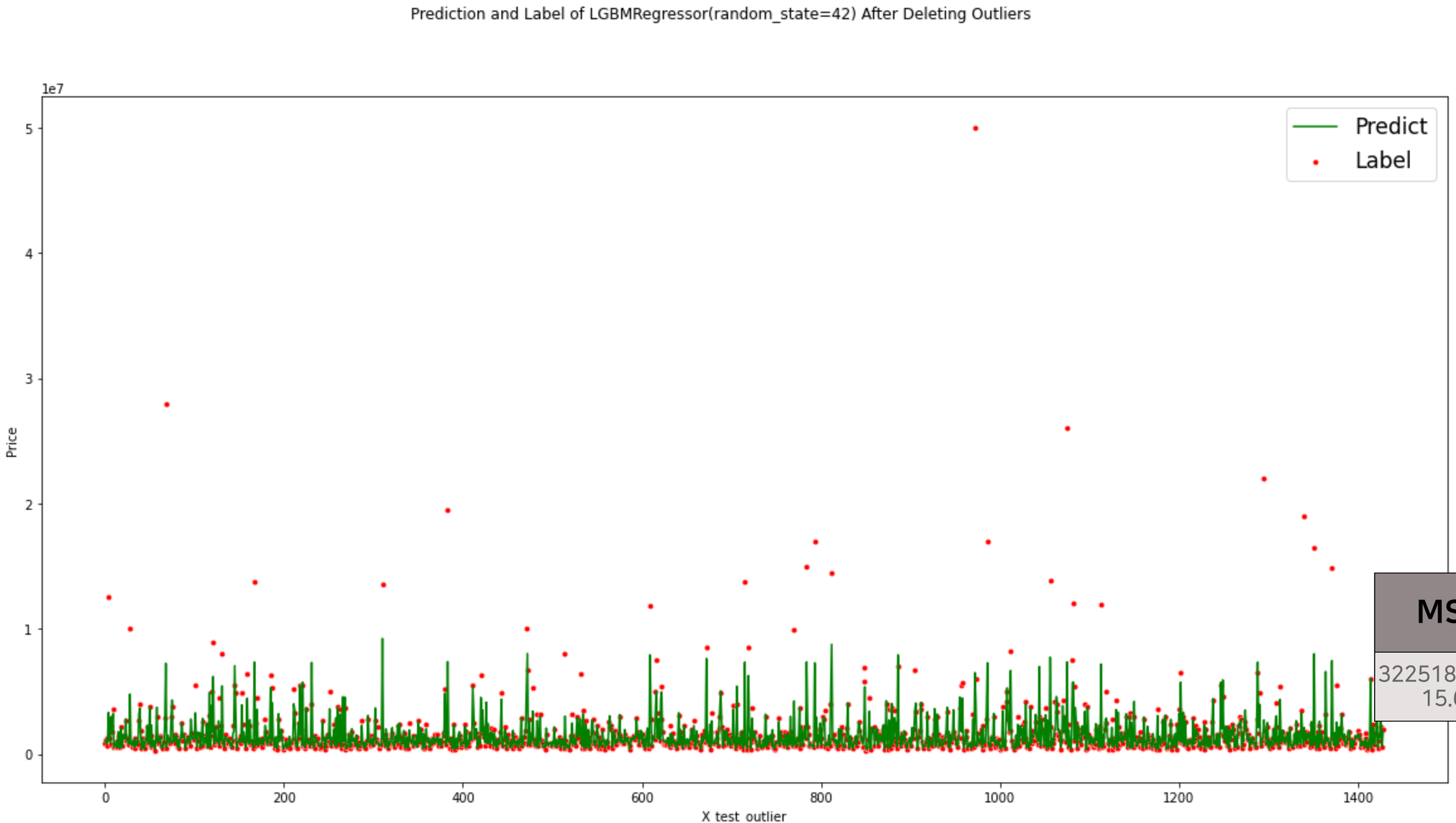
LGBMRegressor (Outlier Delete = False)



MSE	RMSE	R2
1617268327129.74	1271718.65	0.75

SF bay area prices dataset

LGBMRegressor (Outlier Delete = True)



MSE	RMSE	R2
3225186731515.63	1795880.49	0.50

SF bay area prices dataset

Experiment Result

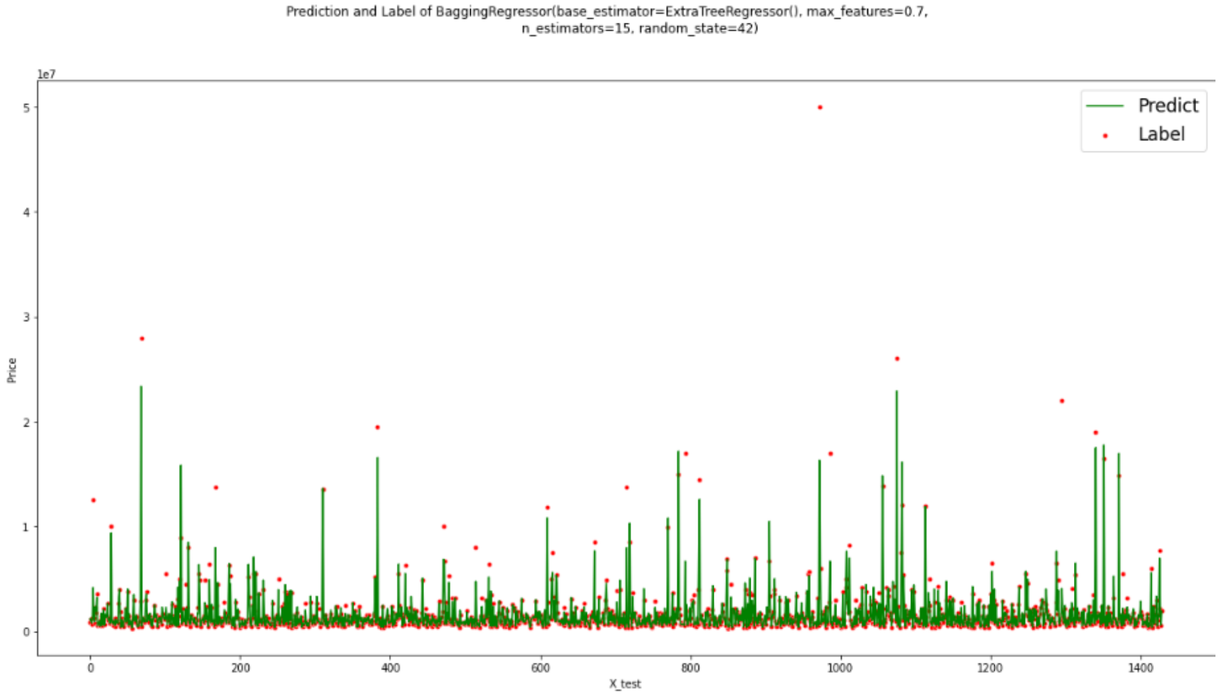
Model	Cross Val mean (Neg RMSE)	MSE	RMSE	R2
XGBRegressor	-989202.39	1479722542137.59	1216438.47	0.77
BaggingRegressor	-1011921.23	1095594500999.92	1046706.50	0.83
GradientBoostingRegressor	-984125.18	1369623034395.87	1170308.95	0.79
RandomForestRegressor	-942548.79	1323706635619.27	1150524.50	0.79
LGBMRegressor	-1005137.28	1617268327129.74	1271718.65	0.75

Model perfomance with (outlier delete = False) data

SF bay area prices dataset

GridSearch with best model : BaggingRegressor

	index	param_base_estimator	param_n_estimators	param_max_features	param_max_samples	mean_test_score	rank_test_score
0	58	ExtraTreeRegressor()	15	0.70	1.00	-979044.34	1
1	59	ExtraTreeRegressor()	20	0.70	1.00	-979510.49	2
2	69	ExtraTreeRegressor()	10	1.00	1.00	-982549.27	3
3	35	DecisionTreeRegressor()	20	1.00	1.00	-986447.34	4
4	66	ExtraTreeRegressor()	15	1.00	0.70	-987213.84	5



MSE	RMSE	R2
1508485843185.35	1228204.32	0.76

Final conclusion

Conclusion

Visualization

- Skewed data
- Outlier exists
- Location is a key feature

Preprocessing

- Use data Pipeline
- Handle Zip code as categorial data
- Should not delete Oullier

M/L

- Lazy predict
- Try with various models
- Fine tune best model