REGRESSION MODELS FOR PREDICT REAL ESTATE PRICE

SALOMO HENDRIAN SUDJONO

DATA UNDERSTANDING

Must be drop !

Wrong interpretation

Data Preview

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

Data Info

```
RangeIndex: 414 entries, 0 to 413
Data columns (total 8 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   No                                      414 non-null    int64
 1   X1 transaction date                     414 non-null    float64
 2   X2 house age                            414 non-null    float64
 3   X3 distance to the nearest MRT station  414 non-null    float64
 4   X4 number of convenience stores         414 non-null    int64
 5   X5 latitude                             414 non-null    float64
 6   X6 longitude                            414 non-null    float64
 7   Y house price of unit area              414 non-null    float64
dtypes: float64(6), int64(2)
```

Summary Statistics

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
count	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000
mean	207.500000	2013.148971	17.712560	1083.885689	4.094203	24.969030	121.533361	37.980193
std	119.655756	0.281967	11.392485	1262.109595	2.945562	0.012410	0.015347	13.606488
min	1.000000	2012.667000	0.000000	23.382840	0.000000	24.932070	121.473530	7.600000
25%	104.250000	2012.917000	9.025000	289.324800	1.000000	24.963000	121.528085	27.700000
50%	207.500000	2013.167000	16.100000	492.231300	4.000000	24.971100	121.538630	38.450000
75%	310.750000	2013.417000	28.150000	1454.279000	6.000000	24.977455	121.543305	46.600000
max	414.000000	2013.583000	43.800000	6488.021000	10.000000	25.014590	121.566270	117.500000

DATA CLEANING

Before

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

Drop column 'No'

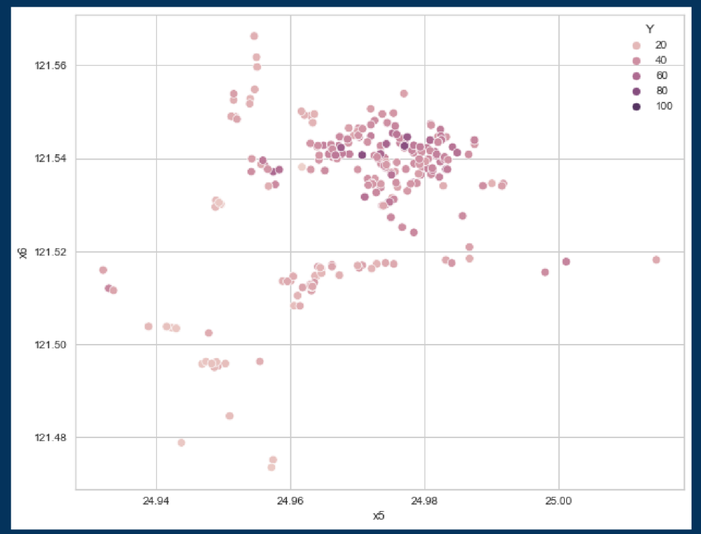
Rename all columns

Transform date value

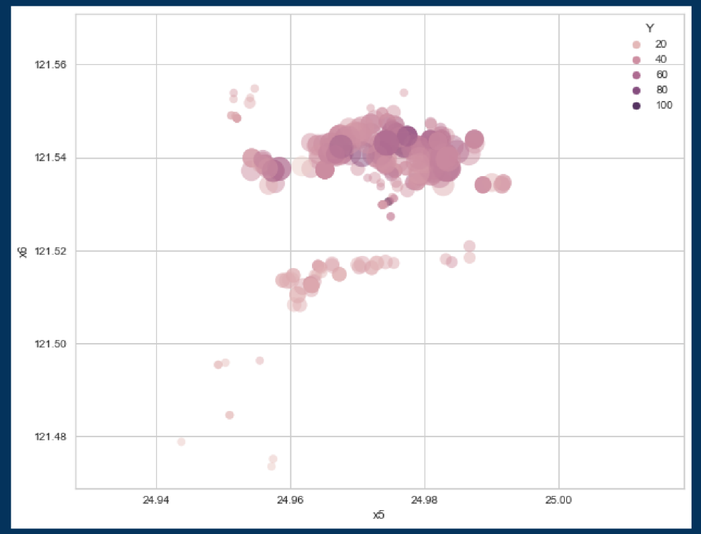
After

	x1	x2	x3	x4	x5	x6	Y
0	2012.9	32.0	84.87882	10	24.98298	121.54024	37.9
1	2012.9	19.5	306.59470	9	24.98034	121.53951	42.2
2	2013.6	13.3	561.98450	5	24.98746	121.54391	47.3
3	2013.5	13.3	561.98450	5	24.98746	121.54391	54.8
4	2012.8	5.0	390.56840	5	24.97937	121.54245	43.1

DATA EXPLORATION

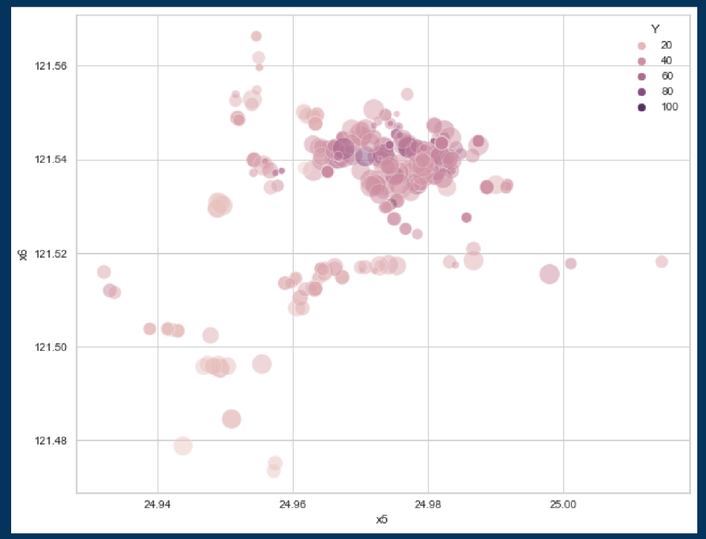


Price

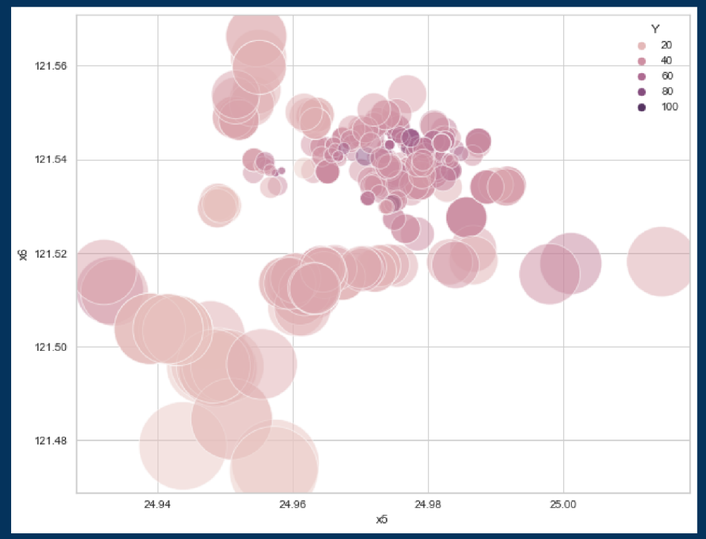


Convinience Store (Total)

House Age



MRT Station (distance)



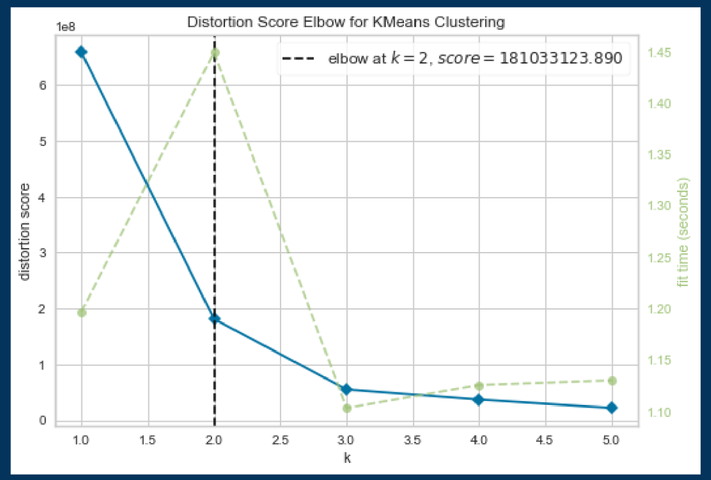
THE DOWNTOWN ASSUMPTION

"ACCORDING TO VISUALIZATIONS BEFORE, THERE IS DENSITY THAT INFLUENCE NUMBER OF CONVINIENCE STORE, DISTANCE FROM MRT STATION, AND PRICE. HENCE, WE CAN ASSUME THAT DENSITY IS THE DOWNTOWN."

Next Step:

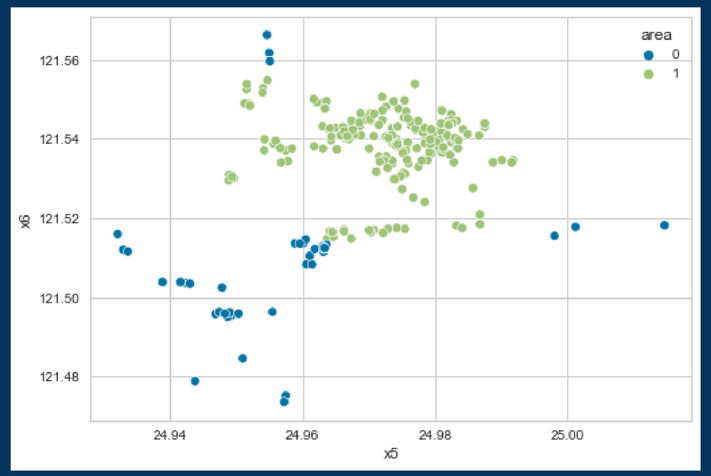
Combine 'Longtitude', 'Latitude', 'x3 (MRT Station)', 'x4 (Convinience Store)' using clustering because there is density that looks like a cluster.

K-MEANS CLUSTERING



Define K using Elbow Method

We assume that there is k = 2, means there are 2 cluster. A house in the downtown and not. So, we observe the assumption with elbow method for choosing the right k. The elbow shown k = 2 is the best k.



Result

From this result, it easier to interpret and conclude that area = 0 is a downtown, and area = 1 is not.

FEATURE SELECTION & SCALING

Feature Selection

	x1	x2	x3	x4	x5	x6	Y	area
0	2012.9	32.0	84.87882	10	24.98298	121.54024	37.9	1
1	2012.9	19.5	306.59470	9	24.98034	121.53951	42.2	1
2	2013.6	13.3	561.98450	5	24.98746	121.54391	47.3	1
3	2013.5	13.3	561.98450	5	24.98746	121.54391	54.8	1
4	2012.8	5.0	390.56840	5	24.97937	121.54245	43.1	1

Predictors Target

Scaling

	x1	x2	x3	x4	area
0	2012.9	32.0	84.87882	10	1
1	2012.9	19.5	306.59470	9	1
2	2013.6	13.3	561.98450	5	1
3	2013.5	13.3	561.98450	5	1
4	2012.8	5.0	390.56840	5	1

RobustScaler →

	x1	x2	x3	x4	area
0	-0.6	0.831373	-0.349673	1.2	1
1	-0.6	0.177778	-0.159351	1.0	1
2	0.8	-0.146405	0.059876	0.2	1
3	0.6	-0.146405	0.059876	0.2	1
4	-0.8	-0.580392	-0.087268	0.2	1

DATA SPLITING

Train data

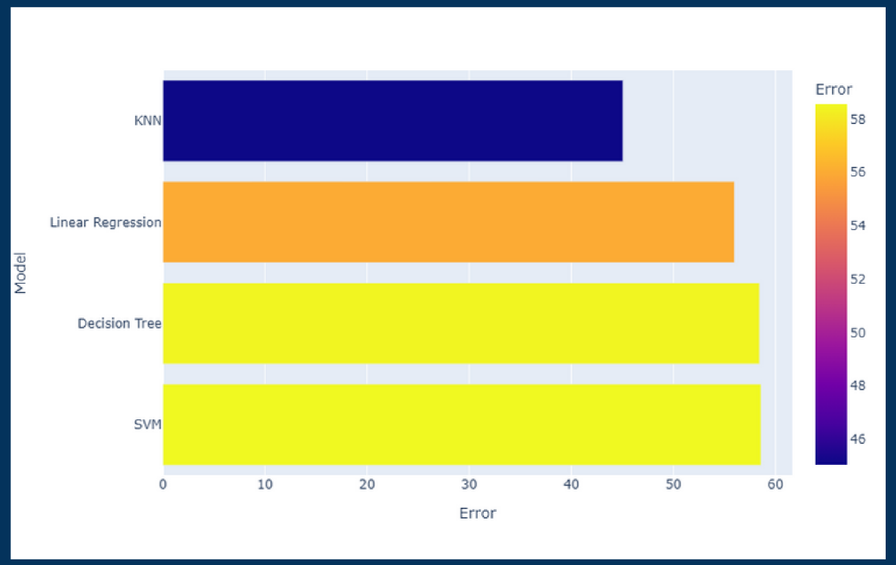
80%

Test data

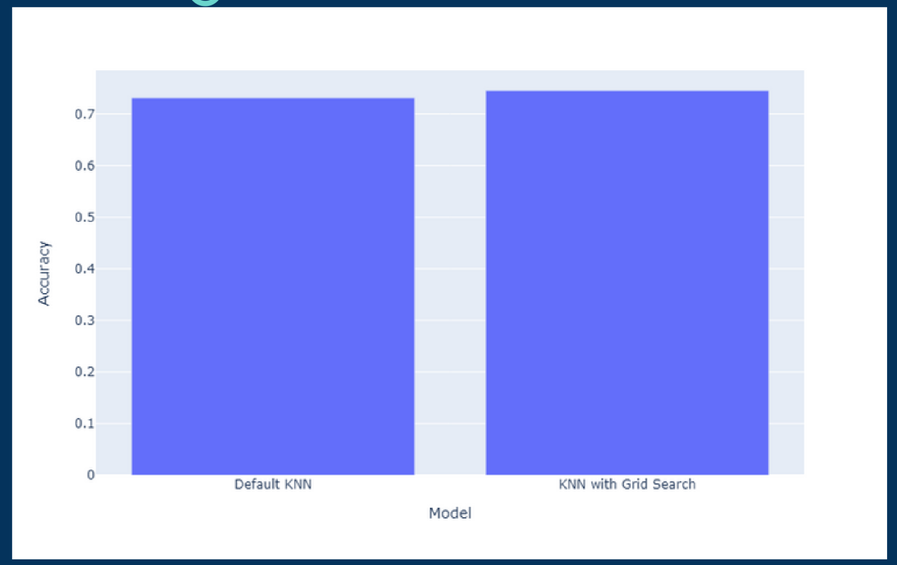
20%

MODELING & CHOOSE THE BEST MODEL

Evaluation



Tuning



Summary

"KNN IS THE BEST REGRESSION MODEL BASED ON THE MEAN SQUARED ERROR."