[Business Analytics]

Assignment 1 – Prediction on House Sales Prices

12146304
Sung-je Kim

Contents

1. Scatter Matrix

2. KDE of Each Data and $\boldsymbol{R^2}$
   - 'yr_built'
   - 'yr_renovated'

3. Result

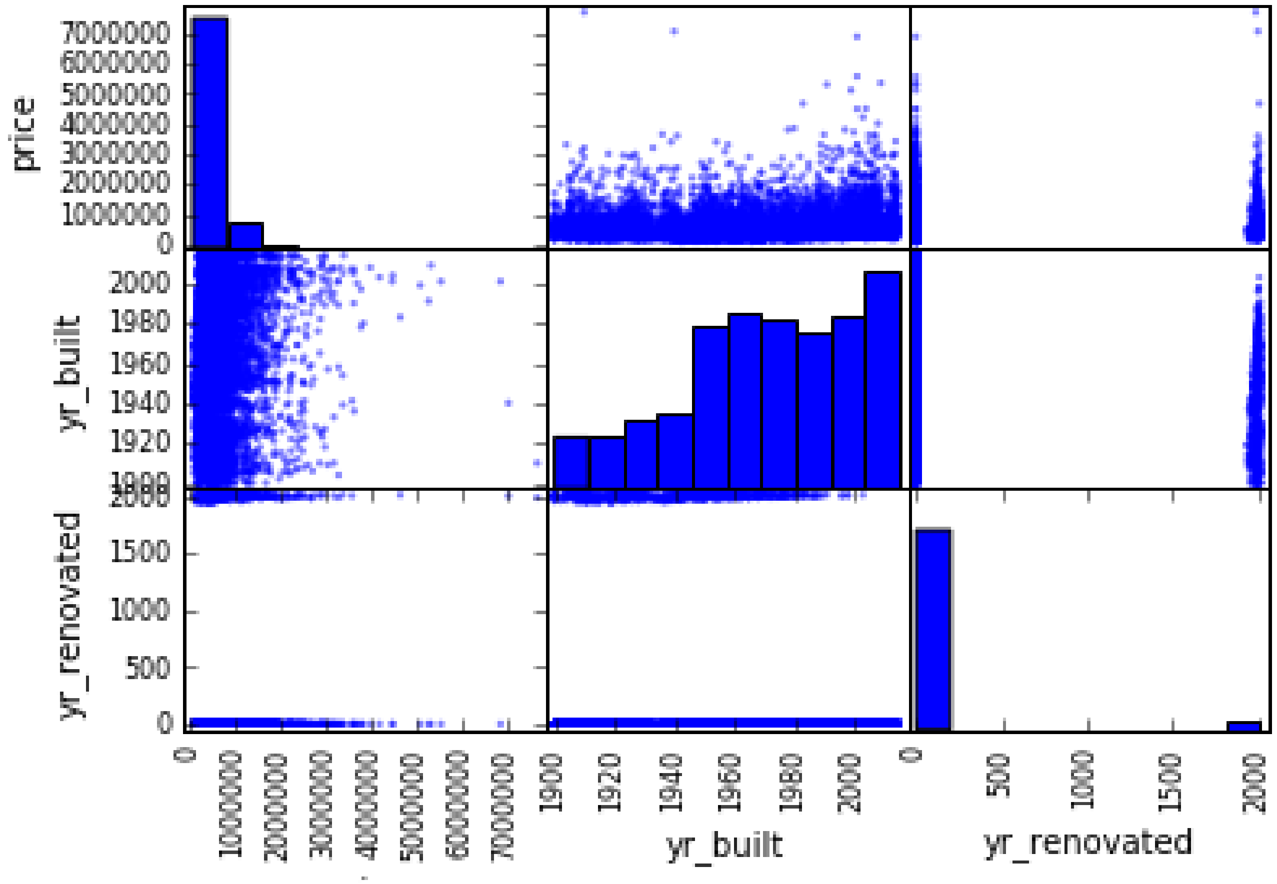
Overview

In class, we just used followed 8 variables for expecting ['price']:

```
X = data[['bedrooms', 'sqft_lot','bathrooms', 'sqft_living','waterfront', 'view', 'condition', 'grade']]
```

In this assignment, I am going to add and use two more variables for analyzing: ['yr_built'], ['yr_renovated']

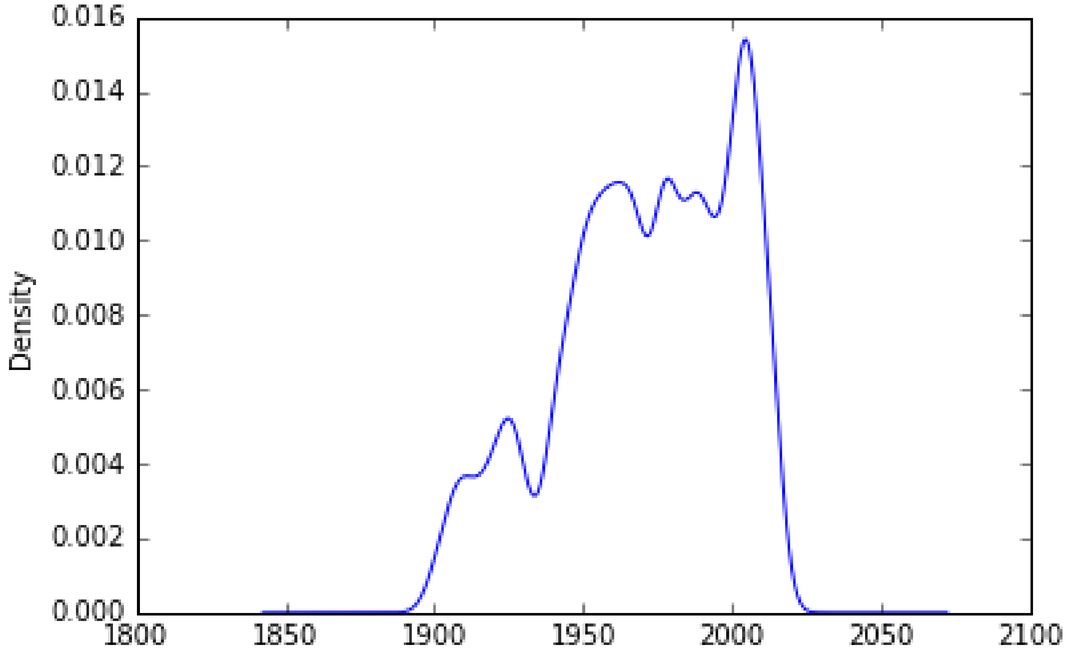
Scatter Matrix



- Firstly, I wanted to have a glance the relationship between price and columns.
- I used scatter_matrix function for glancing

- We can know that distribution of 'yr_built' is almost even.
- We can know that distribution of 'yr_renovated' looks like 0 or something. 0 means there is no renovation.

['yr_built']



This is kde of 'yr_built'. We can know many houses are built recently. However, 'yr_built' data are needed to process properly.

```
In [277]: data['yr_built'].describe()
Out[277]:
count    21613.000000
mean      1971.005136
std         29.373411
min       1900.000000
25%       1951.000000
50%       1975.000000
75%       1997.000000
max       2015.000000
Name: yr_built, dtype: float64
```

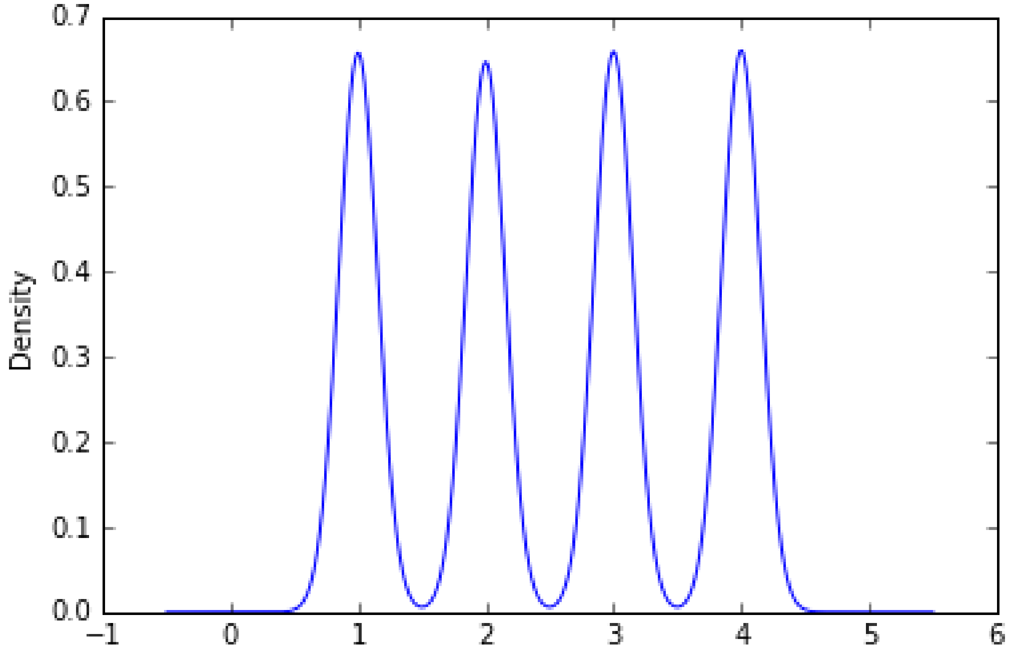
Based on describe(), I divided this data set to 4 parts. They will be explained how much they are old.

Data are divided to:
1900 – 1951: Very Old (score:1)
1951 – 1975: Old (score:2)
1975 – 1997: Middle (score:3)
1997 – 2015: New (score:4)

```
              price  yr_built
price      1.000000  0.054012
yr_built   0.054012  1.000000
```

As you can see, correlation is very low between 'price' and 'yr_built'.

['yr_built'] (cont')



This is kde of 'new_yr_built'. Their density are almost same.

	price	new_yr_built
price	1.000000	0.087458
new_yr_built	0.087458	1.000000

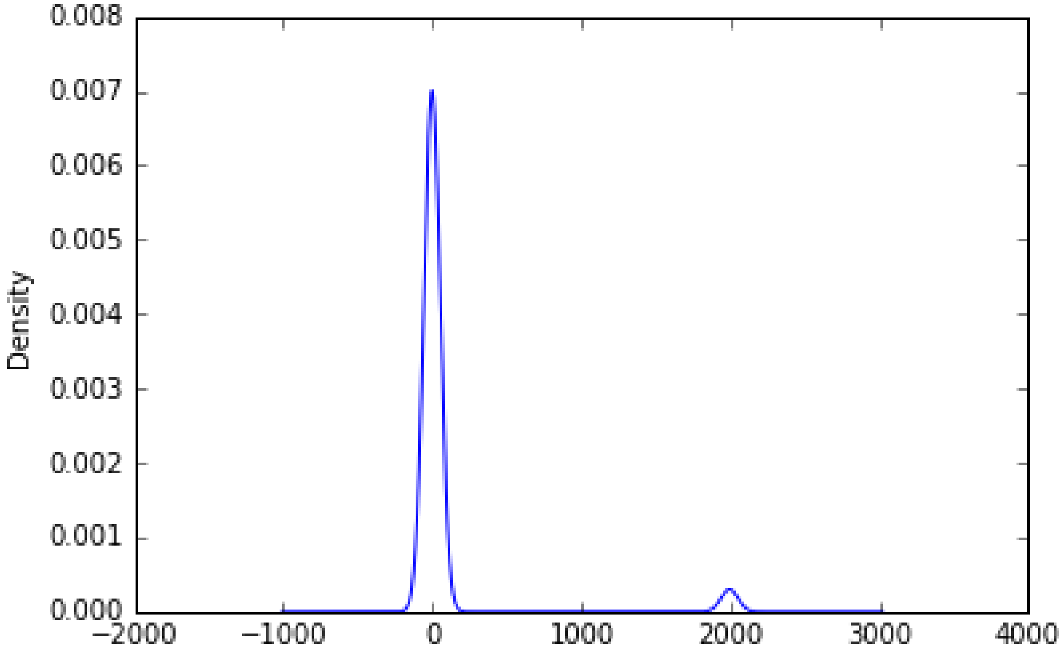
Correlation increases slightly. Even though it increases, we cannot consider as good explanatory variable because correlation it self is very low.

```
reg.score(X,y) # R-
0.6045459158649782
```

```
reg_built.score(X,y)
0.6442580052628251
```

As a result, R^2 increases more about 0.045. It seems those 9 explanatory variables can explain this linear regression mode well, but we have to know R^2 can increase if # of variables increase.

['yr_renovated']



According to KDE of 'yr_renovated', a lot of house are not renovated. Therefore, I just process it to whether renovated(1) or not(0).

	price	yr_renovated
price	1.000000	0.126434
yr_renovated	0.126434	1.000000

Above table is correlation between 'price' and 'yr_renovated'.

	price	new_yr_renovated
price	1.000000	0.126092
new_yr_renovated	0.126092	1.000000

Above table is correlation between 'price' and 'new_yr_renovated'.

Actually, both of them have low correlation with price and even the correlation of new_yr_renovated fell.

```
reg.score(X,y) # R-
0.6045459158649782
```

```
reg.score(X,y) # R-
0.6106763481943056
```

R^2 increases.. very slightly.

Result – Final R^2

```
reg_result.score(X,y)
0.6445597833342942
```

The result of final R^2 is about 0.644 based on 10 explanatory variables. (['new_yr_renovated', 'new_yr_built', 'bedrooms', 'sqft_lot', 'bathrooms', 'sqft_living','waterfront', 'view', 'condition', 'grade'])

```
In [85]: reg_result.coef_
Out[85]:
array([ 6.10256067e-02, -1.17203690e-01, -2.50220182e-02, -1.76078119e-07,
        9.39709434e-02,  1.67387423e-04,  3.25431728e-01,  5.17862881e-02,
        4.34960835e-02,  2.27109407e-01])

In [86]: reg_result.intercept_
Out[86]: 10.978248569852363
```

This coef_ and intercept_ are from between 10 variables and 'log_price'. The reason why some negative coefficients appear, I think that variables are not linear correlation with dependent variable.

Result – Conclusion

Usually, if $\boldsymbol{R^2}$ is larger than 0.65, it can be regarded as well-explaining regression model. However, I think those variables do not explain this model well. This is because those two added variables do not have high correlation and linear relationship with price. Therefore, $\boldsymbol{R^2}$ just increases because of adding variables. For solving this problem, I suggest using adjusted R-square.