Assignment 02

15146314 Yang, Seunghyuck

1. EDA

A. Remove 'price' (which is the one we are going to treat as 'y') from cols

```
import pandas as pd
import matplotlib.pyplot as plt

house = pd.read_csv('https://drive.google.com/uc?export=download&id=1kgJseOaDUCG-p-IoLIKbnL23XHUZPEwm')
house
house.keys()
```

```
#%% EDA

cols = house.keys()
corr = house[cols].corr()

# 1. Remove 'price' (which is the one we are going to treat as 'y') from cols
y = 'price'
cols = list(cols)
cols.remove('price')
```

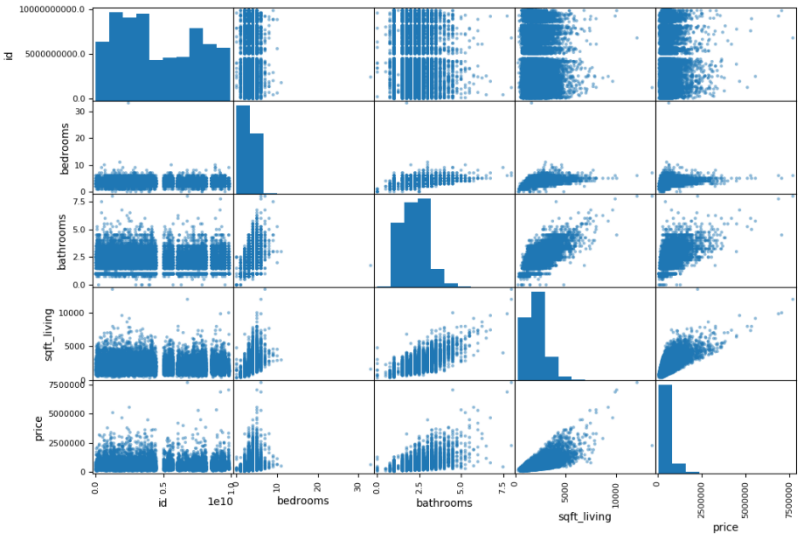
B. Divide cols into 4 pieces to see scatter matrix by each

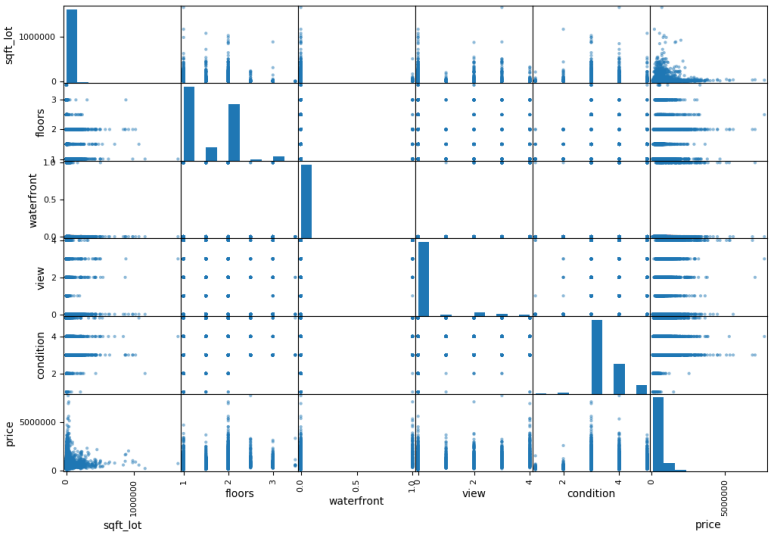
(Computer is not affordable with holding the process at once)

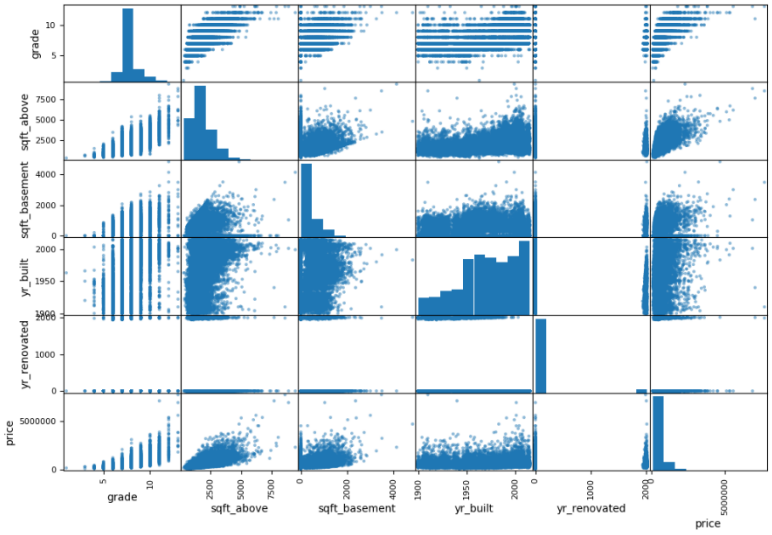
```
# 2. divide cols into 4 pieces to see scatter matrix by each
# It is too big for using at once
cols1 = cols[0:5]
cols2 = cols[5:10]
cols3 = cols[10:15]
cols4 = cols[15:]
```

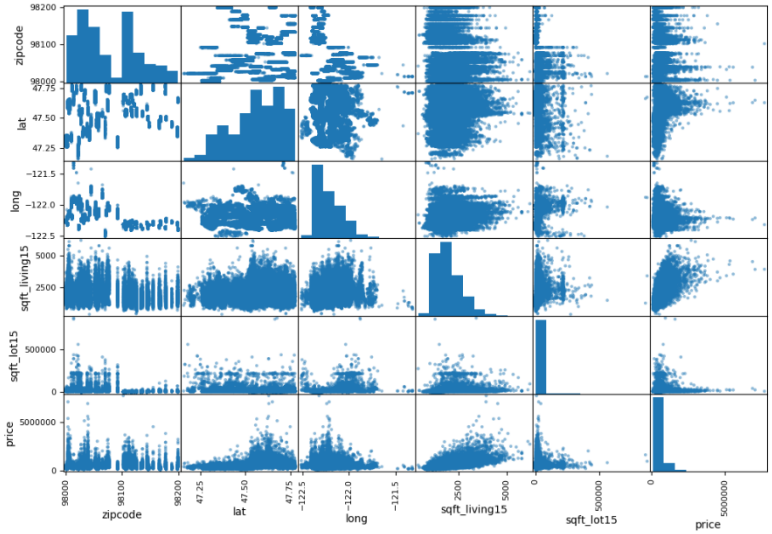
C. Scatter matrix for each cols's particles and 'y'

```
# Scatter matrix for each cols's particles and 'y'
cols1.append(y)
cols2.append(y)
cols3.append(y)
cols4.append(y)
from pandas.plotting import scatter_matrix
scatter_matrix(house[cols1], figsize = (12, 8))
scatter_matrix(house[cols2], figsize = (12, 8))
scatter_matrix(house[cols3], figsize = (12, 8))
scatter_matrix(house[cols4], figsize = (12, 8))
```


id
bedrooms
bathrooms
sqft_living
price
1e10


sqft_lot
floors
waterfront
view
condition
price





Result for C:

Distinguish by tier:

1. Bathrooms / sqft_living / grade / sqft_above / sqft_basement / sqft_living15 / sqft_lot15(inverse) / sqft_lot(inverse)
2. Bedrooms / condition / lat / long
3. Id / floors / waterfront / view / yr_built / yr_renovated / zipcode

Criterion: Whether they seem correlated with price with bare eyes?

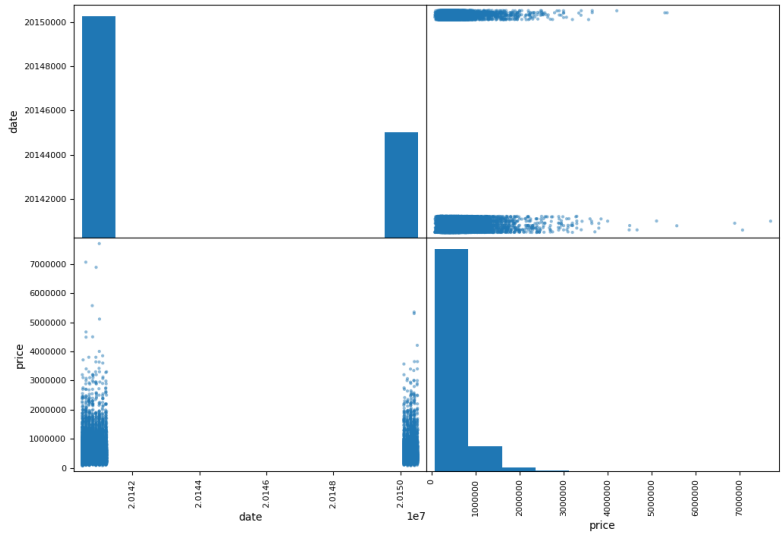
Problem: Since date value is not integer, it cannot be used for original statues.

Solution: Transform it to Integer.

D. Transform value 'date' into integer which is the form we can use

```
# 4. Transform value 'date' into integer which is the form we can use

house['date']
for ind in range(len(house['date'])):
    newDate = house['date'][ind][:8]
    """
    if int(newDate[:4]) == 2014:
        newDate = int(newDate[4:])
    else:
        newDate = int('1' + newDate[4:])
    """
    house['date'][ind] = newDate
house['date'] = pd.to_numeric(house['date'])

scatter_matrix(house[['date', 'price']], figsize = (12, 8))

house[['date', 'price']].corr()
```

```
        dtype=object)

In [145]: house[['date', 'price']].corr()
Out[145]:
              date     price
date      1.000000  0.003033
price     0.003033  1.000000

In [146]:
```



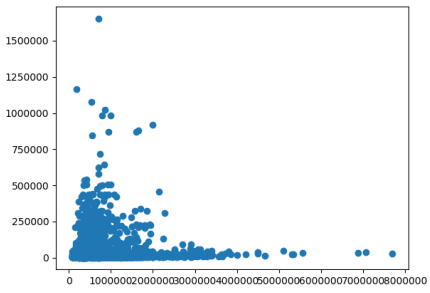
By the result, it is proved that date has no correlation with price, which is the result that indicates date to tier 3.

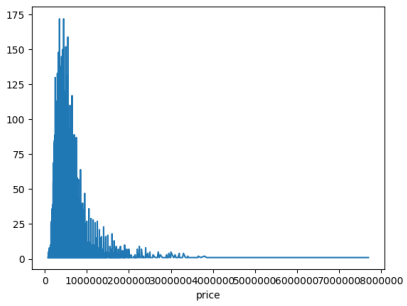
2. Preprocessing

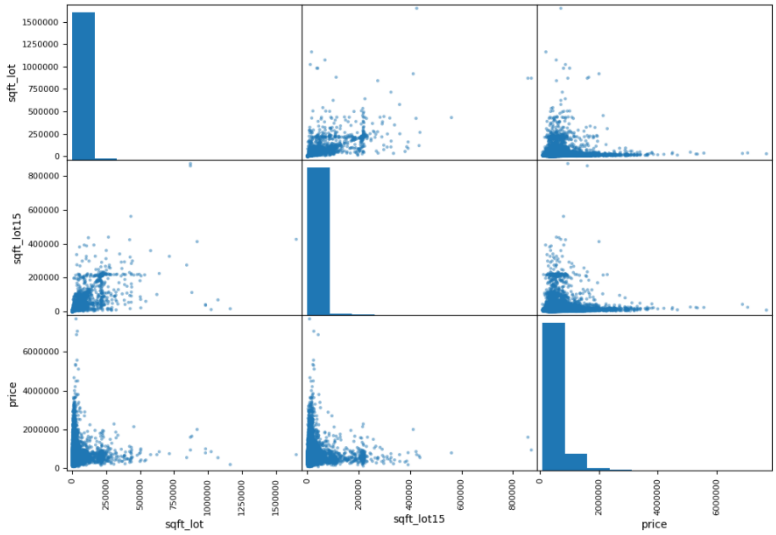
A. Transform data

i. Sqft_lot, sqft_lot15 -> failed to transform them into form which has linear correlation with price.

I can't decide whether they are related or not, so I decide to see both results.







B. Create Train Dataset

i. Variables: Bathrooms / sqft_living / grade / sqft_above / sqft_basement / sqft_living15 / (+ sqft_lot15 / sqft_lot)

ii. Train and see result

1. With sqft_lot15 / sqft_lot

```
#%% Linear regression

import statsmodels.api as sm
from sklearn import datasets

X = house[['bathrooms', 'sqft_living', 'grade', 'sqft_above',
           'sqft_basement', 'sqft_living15', 'sqft_lot15', 'sqft_lot']]

y = house['price']

X = sm.add_constant(X)

model = sm.OLS(y, X)
result = model.fit()

result.summary()
```

```
In [156]: result.summary()
Out[156]:
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.546
Model:                            OLS   Adj. R-squared:                  0.546
Method:                 Least Squares   F-statistic:                     3709.
Date:                Fri, 11 Sep 2020   Prob (F-statistic):               0.00
Time:                        02:27:09   Log-Likelihood:            -2.9908e+05
No. Observations:               21613   AIC:                         5.982e+05
Df Residuals:                   21605   BIC:                         5.982e+05
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -6.4e+05   1.35e+04    -47.361      0.000   -6.66e+05   -6.14e+05
bathrooms     -3.747e+04   3428.641    -10.928      0.000   -4.42e+04   -3.07e+04
sqft_living     138.8091      2.487     55.805      0.000     133.934     143.685
grade          1.098e+05   2462.011     44.607      0.000    1.05e+05    1.15e+05
sqft_above       30.4898      2.450     12.442      0.000      25.687      35.293
sqft_basement   108.3193      2.654     40.813      0.000     103.117     113.521
sqft_living15    25.4125      4.033      6.301      0.000      17.508      33.317
sqft_lot15       -0.6307      0.089     -7.053      0.000      -0.806      -0.455
sqft_lot          0.0782      0.059      1.333      0.182      -0.037       0.193
==============================================================================
Omnibus:                    17216.002   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1121014.451
Skew:                           3.347   Prob(JB):                         0.00
Kurtosis:                      37.641   Cond. No.                     1.23e+17
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.66e-21. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""
```

2. Without sqft_lot15 / sqft_lot

```
In [260]: result.summary()
Out[260]:
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.544
Model:                            OLS   Adj. R-squared:                  0.544
Method:                 Least Squares   F-statistic:                     5160.
Date:                Fri, 11 Sep 2020   Prob (F-statistic):               0.00
Time:                        03:41:58   Log-Likelihood:            -2.9911e+05
No. Observations:               21613   AIC:                         5.982e+05
Df Residuals:                   21607   BIC:                         5.983e+05
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -6.469e+05   1.35e+04    -47.870      0.000   -6.73e+05    -6.2e+05
bathrooms     -3.546e+04   3425.567    -10.353      0.000   -4.22e+04   -2.87e+04
sqft_living     136.7857      2.475     55.271      0.000     131.935     141.636
grade          1.11e+05   2462.309     45.090      0.000    1.06e+05    1.16e+05
sqft_above       28.1505      2.433     11.572      0.000      23.382      32.919
sqft_basement   108.6352      2.658     40.866      0.000     103.425     113.846
sqft_living15    22.8201      4.027      5.667      0.000      14.927      30.713
==============================================================================
Omnibus:                    17285.229   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1134486.304
Skew:                           3.366   Prob(JB):                         0.00
Kurtosis:                      37.849   Cond. No.                     8.58e+15
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.9e-21. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""

In [261]:
```