king county housing price modeling

Normalizing

In [1]:
```
# import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import folium
plt.style.use('ggplot')
```

Read the clean KC housing data

In [2]:
```
kc = pd.read_csv('data/kc_house_data_clean.csv')
kc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21399 entries, 0 to 21398
Data columns (total 20 columns):
id               21399 non-null int64
date             21399 non-null object
price            21399 non-null int64
bedrooms         21399 non-null int64
bathrooms        21399 non-null float64
sqft_lot         21399 non-null float64
floors           21399 non-null float64
waterfront       21399 non-null float64
view             21399 non-null float64
condition        21399 non-null int64
grade            21399 non-null int64
sqft_above       21399 non-null float64
sqft_basement    21399 non-null float64
yr_built         21399 non-null int64
yr_renovated     21399 non-null int64
zipcode          21399 non-null int64
lat              21399 non-null float64
long             21399 non-null float64
sqft_living15    21399 non-null float64
sqft_lot15       21399 non-null float64
dtypes: float64(11), int64(8), object(1)
memory usage: 3.3+ MB
```

In [3]:
```
'''
price in $ span a large range of numbers while all the features has quite shor
t range. It could be better to take the log
scale of the price before fitting.
also drop id
'''
kc = kc.drop(["id"], axis=1)
kc['price'] = np.log(kc['price'])
```

baseline model

To better gauage the progress let's model without any change to the features

In [4]:
```
# import statistical libraries for modelling.
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

simple linear regression

It maybe worthwhile to investigate correlation between each individual feature with the outcome, which is the price

In [5]:
```
kc.columns
```

Out[5]:
```
Index(['date', 'price', 'bedrooms', 'bathrooms', 'sqft_lot', 'floors',
       'waterfront', 'view', 'condition', 'grade', 'sqft_above',
       'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long',
       'sqft_living15', 'sqft_lot15'],
      dtype='object')
```

In [6]:
```python
# using OLS of statsmodels.formula.api
tmp_y = kc[["price"]]
tmp_X = kc.drop(["price", "date"], axis=1)
predictors = tmp_X.columns

for idx, val in enumerate(predictors):
    print("formula = price ~ "+val)
    f = 'price ~ ' + val
    model = ols(formula=f, data=kc).fit()
    print(model.summary())
    print("###################################################################
#############\n\n")
```

```
formula = price ~ bedrooms
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.11
4
Model:                            OLS   Adj. R-squared:                  0.11
4
Method:                 Least Squares   F-statistic:                     275
8.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:36   Log-Likelihood:                -1425
6.
No. Observations:               21399   AIC:                         2.852e+0
4
Df Residuals:                   21397   BIC:                         2.853e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.3989      0.012    993.633      0.000      12.374      12.42
3
bedrooms       0.1883      0.004     52.513      0.000       0.181       0.19
5
==============================================================================
=
Omnibus:                       73.399   Durbin-Watson:                   1.95
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                73.38
9
Skew:                           0.136   Prob(JB):                     1.16e-1
6
Kurtosis:                       2.910   Cond. No.                         14.
5
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ bathrooms
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.27
4
Model:                            OLS   Adj. R-squared:                  0.27
```

```
4
Method:                 Least Squares   F-statistic:                     806
9.
Date:                Sun, 10 May 2020   Prob (F-statistic):              0.0
0
Time:                        08:50:36   Log-Likelihood:                -1212
9.
No. Observations:               21399   AIC:                       2.426e+0
4
Df Residuals:                   21397   BIC:                       2.428e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.2953      0.009   1412.670      0.000      12.278      12.31
2
bathrooms      0.3508      0.004     89.830      0.000       0.343       0.35
9
==============================================================================
=
Omnibus:                       94.548   Durbin-Watson:                   1.95
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               87.63
5
Skew:                           0.124   Prob(JB):                    9.34e-2
0
Kurtosis:                       2.807   Cond. No.                         7.8
6
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ sqft_lot
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.01
0
Model:                            OLS   Adj. R-squared:                  0.01
0
Method:                 Least Squares   F-statistic:                     216.
5
Date:                Sun, 10 May 2020   Prob (F-statistic):          9.06e-4
9
Time:                        08:50:36   Log-Likelihood:                -1544
5.
```

```
No. Observations:               21399   AIC:                         3.089e+0
4
Df Residuals:                   21397   BIC:                         3.091e+0
4
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
-
Intercept     13.0138      0.004   3593.374      0.000      13.007      13.02
1
sqft_lot    1.209e-06   8.22e-08     14.714      0.000    1.05e-06    1.37e-0
6
=============================================================================
=
Omnibus:                      107.332   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.84
6
Skew:                           0.174   Prob(JB):                      2.31e-2
4
Kurtosis:                       3.018   Cond. No.                      4.69e+0
4
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 4.69e+04. This might indicate that there a
re
strong multicollinearity or other numerical problems.
#############################################################################
###


formula = price ~ floors
                            OLS Regression Results
=============================================================================
=
Dep. Variable:                  price   R-squared:                       0.09
4
Model:                            OLS   Adj. R-squared:                  0.09
4
Method:                 Least Squares   F-statistic:                     222
8.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:36   Log-Likelihood:                -1449
3.
No. Observations:               21399   AIC:                         2.899e+0
4
Df Residuals:                   21397   BIC:                         2.901e+0
4
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.6070      0.010   1316.542      0.000      12.588      12.62
6
floors         0.2851      0.006     47.197      0.000       0.273       0.29
7
==============================================================================
=
Omnibus:                      142.409   Durbin-Watson:                   1.96
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              144.74
8
Skew:                           0.198   Prob(JB):                    3.70e-3
2
Kurtosis:                       2.929   Cond. No.                         6.3
6
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ waterfront
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.01
2
Model:                            OLS   Adj. R-squared:                  0.01
2
Method:                 Least Squares   F-statistic:                     253.
3
Date:                Sun, 10 May 2020   Prob (F-statistic):          1.04e-5
6
Time:                        08:50:36   Log-Likelihood:                -1542
7.
No. Observations:               21399   AIC:                         3.086e+0
4
Df Residuals:                   21397   BIC:                         3.087e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
```

```
-
Intercept     13.0282      0.003   3820.869      0.000      13.022      13.03
5
waterfront     0.7822      0.049     15.916      0.000       0.686       0.87
9
==============================================================================
=
Omnibus:                       94.268   Durbin-Watson:                   1.95
0
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               95.46
5
Skew:                           0.163   Prob(JB):                    1.86e-2
1
Kurtosis:                       2.987   Cond. No.                         14.
4
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ view
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.09
2
Model:                            OLS   Adj. R-squared:                  0.09
2
Method:                 Least Squares   F-statistic:                     215
7.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:36   Log-Likelihood:                -1452
5.
No. Observations:               21399   AIC:                         2.905e+0
4
Df Residuals:                   21397   BIC:                         2.907e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.9870      0.003   3817.846      0.000      12.980      12.99
4
view           0.2071      0.004     46.444      0.000       0.198       0.21
6
==============================================================================
=
```

```
Omnibus:                       40.802   Durbin-Watson:                   1.94
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               40.99
1
Skew:                           0.107   Prob(JB):                    1.26e-0
9
Kurtosis:                       3.010   Cond. No.                         1.4
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ condition
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
1
Model:                            OLS   Adj. R-squared:                  0.00
1
Method:                 Least Squares   F-statistic:                     27.6
1
Date:                Sun, 10 May 2020   Prob (F-statistic):          1.50e-0
7
Time:                        08:50:36   Log-Likelihood:                -1553
9.
No. Observations:               21399   AIC:                         3.108e+0
4
Df Residuals:                   21397   BIC:                         3.110e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.9377      0.018    708.125      0.000      12.902      12.97
3
condition      0.0277      0.005      5.255      0.000       0.017       0.03
8
==============================================================================
=
Omnibus:                      113.194   Durbin-Watson:                   1.95
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              114.92
7
Skew:                           0.179   Prob(JB):                    1.11e-2
5
Kurtosis:                       2.993   Cond. No.                         20.
```

```
0
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ grade
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.46
5
Model:                            OLS   Adj. R-squared:                  0.46
5
Method:                 Least Squares   F-statistic:                 1.858e+0
4
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:37   Log-Likelihood:                -8866.
3
No. Observations:               21399   AIC:                         1.774e+0
4
Df Residuals:                   21397   BIC:                         1.775e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     10.7400      0.017    631.686      0.000      10.707      10.77
3
grade          0.3004      0.002    136.291      0.000       0.296       0.30
5
==============================================================================
=
Omnibus:                       36.605   Durbin-Watson:                   1.96
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.76
1
Skew:                           0.102   Prob(JB):                     1.04e-0
8
Kurtosis:                       3.004   Cond. No.                         53.
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

```
################################################################################
###


formula = price ~ sqft_above
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.32
6
Model:                            OLS   Adj. R-squared:                  0.32
6
Method:                 Least Squares   F-statistic:                 1.033e+0
4
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:37   Log-Likelihood:                 -1133
7.
No. Observations:               21399   AIC:                         2.268e+0
4
Df Residuals:                   21397   BIC:                         2.269e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.3951      0.007   1805.196      0.000      12.382      12.40
9
sqft_above     0.0004   3.54e-06    101.649      0.000       0.000       0.00
0
==============================================================================
=
Omnibus:                       82.153   Durbin-Watson:                   1.98
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               74.36
8
Skew:                           0.105   Prob(JB):                    7.10e-1
7
Kurtosis:                       2.802   Cond. No.                    4.73e+0
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 4.73e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
################################################################################
###
```

```
formula = price ~ sqft_basement
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.08
0
Model:                            OLS   Adj. R-squared:                  0.08
0
Method:                 Least Squares   F-statistic:                     186
5.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:37   Log-Likelihood:                -1465
8.
No. Observations:               21399   AIC:                         2.932e+0
4
Df Residuals:                   21397   BIC:                         2.934e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        12.9396      0.004   3303.693      0.000      12.932      1
2.947
sqft_basement     0.0003   7.66e-06     43.190      0.000       0.000
0.000
==============================================================================
=
Omnibus:                      132.547   Durbin-Watson:                   1.94
0
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              134.93
2
Skew:                           0.195   Prob(JB):                    5.01e-3
0
Kurtosis:                       2.998   Cond. No.                         61
0.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ yr_built
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
7
Model:                            OLS   Adj. R-squared:                  0.00
```

```
7
Method:                 Least Squares   F-statistic:                     153.
4
Date:                Sun, 10 May 2020   Prob (F-statistic):           4.15e-3
5
Time:                        08:50:37   Log-Likelihood:                -1547
6.
No. Observations:               21399   AIC:                         3.096e+0
4
Df Residuals:                   21397   BIC:                         3.097e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     10.1927      0.229     44.456      0.000       9.743      10.64
2
yr_built       0.0014      0.000     12.385      0.000       0.001       0.00
2
==============================================================================
=
Omnibus:                      150.208   Durbin-Watson:                   1.96
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              153.14
3
Skew:                           0.206   Prob(JB):                     5.57e-3
4
Kurtosis:                       2.949   Cond. No.                     1.33e+0
5
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.33e+05. This might indicate that there a
re
strong multicollinearity or other numerical problems.
##############################################################################
###


formula = price ~ yr_renovated
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
9
Model:                            OLS   Adj. R-squared:                  0.00
9
Method:                 Least Squares   F-statistic:                     198.
8
Date:                Sun, 10 May 2020   Prob (F-statistic):           6.22e-4
```

```
5
Time:                        08:50:37   Log-Likelihood:                 -1545
4.
No. Observations:               21399   AIC:                        3.091e+0
4
Df Residuals:                   21397   BIC:                        3.093e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===
                  coef    std err          t      P>|t|      [0.025      0.9
75]
------------------------------------------------------------------------------
---
Intercept      13.0230      0.003   3759.148      0.000      13.016      13.
030
yr_renovated    0.0001   9.48e-06     14.098      0.000       0.000       0.
000
==============================================================================
=
Omnibus:                       96.883   Durbin-Watson:                   1.95
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               98.12
6
Skew:                           0.166   Prob(JB):                     4.92e-2
2
Kurtosis:                       3.005   Cond. No.                         37
2.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ zipcode
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
1
Model:                            OLS   Adj. R-squared:                  0.00
1
Method:                 Least Squares   F-statistic:                     22.1
1
Date:                Sun, 10 May 2020   Prob (F-statistic):           2.59e-0
6
Time:                        08:50:37   Log-Likelihood:                 -1554
2.
No. Observations:               21399   AIC:                        3.109e+0
4
Df Residuals:                   21397   BIC:                        3.110e+0
4
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     42.5197      6.272      6.780      0.000      30.227      54.81
2
zipcode       -0.0003   6.39e-05     -4.702      0.000      -0.000      -0.00
0
==============================================================================
=
Omnibus:                      106.988   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.53
7
Skew:                           0.174   Prob(JB):                    2.70e-2
4
Kurtosis:                       2.986   Cond. No.                    1.80e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.8e+08. This might indicate that there ar
e
strong multicollinearity or other numerical problems.
##############################################################################
###


formula = price ~ lat
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.21
2
Model:                            OLS   Adj. R-squared:                  0.21
2
Method:                 Least Squares   F-statistic:                     577
0.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:37   Log-Likelihood:                -1299
8.
No. Observations:               21399   AIC:                         2.600e+0
4
Df Residuals:                   21397   BIC:                         2.602e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
```

```
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept    -65.8959      1.039    -63.416      0.000     -67.933     -63.85
9
lat            1.6596      0.022     75.958      0.000       1.617       1.70
2
==============================================================================
=
Omnibus:                      452.059   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              514.03
5
Skew:                           0.319   Prob(JB):                    2.39e-11
2
Kurtosis:                       3.410   Cond. No.                    1.63e+0
4
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.63e+04. This might indicate that there a
re
strong multicollinearity or other numerical problems.
##############################################################################
###


formula = price ~ long
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
4
Model:                            OLS   Adj. R-squared:                  0.00
4
Method:                 Least Squares   F-statistic:                     77.3
0
Date:                Sun, 10 May 2020   Prob (F-statistic):          1.58e-1
8
Time:                        08:50:37   Log-Likelihood:                -1551
4.
No. Observations:               21399   AIC:                        3.103e+0
4
Df Residuals:                   21397   BIC:                        3.105e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
```

```
Intercept     39.0391      2.958     13.198      0.000     33.241      44.83
7
long           0.2128      0.024      8.792      0.000      0.165       0.26
0
==============================================================================
=
Omnibus:                      125.975   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.10
1
Skew:                           0.189   Prob(JB):                    1.53e-2
8
Kurtosis:                       3.011   Cond. No.                    1.06e+0
5
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.06e+05. This might indicate that there a
re
strong multicollinearity or other numerical problems.
##############################################################################
###


formula = price ~ sqft_living15
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.35
8
Model:                            OLS   Adj. R-squared:                  0.35
8
Method:                 Least Squares   F-statistic:                 1.194e+0
4
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:37   Log-Likelihood:                -1080
9.
No. Observations:               21399   AIC:                         2.162e+0
4
Df Residuals:                   21397   BIC:                         2.164e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                   coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        12.1485      0.009   1422.930      0.000     12.132       1
2.165
sqft_living15     0.0004    4.1e-06    109.265      0.000      0.000
0.000
```

```
==============================================================================
=
Omnibus:                      107.251   Durbin-Watson:                   1.97
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              111.13
2
Skew:                           0.158   Prob(JB):                    7.38e-2
5
Kurtosis:                       3.158   Cond. No.                    6.49e+0
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 6.49e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
#############################################################################
###


formula = price ~ sqft_lot15
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
8
Model:                            OLS   Adj. R-squared:                  0.00
8
Method:                 Least Squares   F-statistic:                     178.
6
Date:                Sun, 10 May 2020   Prob (F-statistic):          1.40e-4
0
Time:                        08:50:37   Log-Likelihood:                -1546
4.
No. Observations:               21399   AIC:                         3.093e+0
4
Df Residuals:                   21397   BIC:                         3.095e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     13.0108      0.004   3461.771      0.000      13.003      13.01
8
sqft_lot15  1.668e-06   1.25e-07     13.365      0.000    1.42e-06    1.91e-0
6
==============================================================================
=
Omnibus:                      110.589   Durbin-Watson:                   1.95
3
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                112.22
8
Skew:                           0.177   Prob(JB):                       4.27e-2
5
Kurtosis:                       3.003   Cond. No.                       3.32e+0
4
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 3.32e+04. This might indicate that there a
re
strong multicollinearity or other numerical problems.
##############################################################################
###
```

The R2 values between statsmodels.api and statsmodels.formula.api are totally different. statsmodels.api has very high value and statsmodels.formula.api has a very low value. Visual inspection of the data shows most of the features are not linearly related. Thus, it is important to differentiate between these two options in the same statsmodels library.

multiple linear regression

In [7]:
```
kc.columns
```

Out[7]:
```
Index(['date', 'price', 'bedrooms', 'bathrooms', 'sqft_lot', 'floors',
       'waterfront', 'view', 'condition', 'grade', 'sqft_above',
       'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long',
       'sqft_living15', 'sqft_lot15'],
      dtype='object')
```

In [8]:
```
tmp_kc = kc.drop(["date"], axis=1) # obviously date datatype won't work with O
LS
predictors = list(tmp_kc.columns)
predictors.remove('price')

f = 'price ~ ' + ' + '.join(predictors)
model = ols(formula=f, data=tmp_kc).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
1
Model:                            OLS   Adj. R-squared:                  0.75
0
Method:                 Least Squares   F-statistic:                     378
5.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:37   Log-Likelihood:                -694.4
5
No. Observations:               21399   AIC:                             142
5.
Df Residuals:                   21381   BIC:                             156
8.
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        -5.9318      3.659     -1.621      0.105     -13.103
1.239
bedrooms         -0.0117      0.002     -4.698      0.000      -0.017      -
0.007
bathrooms         0.0701      0.004     17.096      0.000       0.062
0.078
sqft_lot       4.843e-07   5.97e-08      8.113      0.000    3.67e-07    6.01
e-07
floors            0.0789      0.005     17.475      0.000       0.070
0.088
waterfront        0.3545      0.026     13.458      0.000       0.303
0.406
view              0.0616      0.003     22.814      0.000       0.056
0.067
condition         0.0630      0.003     21.508      0.000       0.057
0.069
grade             0.1579      0.003     58.198      0.000       0.153
0.163
sqft_above        0.0001   4.79e-06     27.171      0.000       0.000
0.000
sqft_basement     0.0001   5.61e-06     26.659      0.000       0.000
0.000
yr_built         -0.0034   9.04e-05    -37.929      0.000      -0.004      -
0.003
yr_renovated   3.718e-05   5.03e-06      7.397      0.000    2.73e-05     4.7
e-05
zipcode          -0.0006   4.12e-05    -14.757      0.000      -0.001      -
0.001
lat               1.3924      0.013    104.101      0.000       1.366
1.419
long             -0.1394      0.016     -8.488      0.000      -0.172      -
0.107
```

```
sqft_living15  9.631e-05    4.4e-06     21.913      0.000    8.77e-05
0.000
sqft_lot15    -2.233e-07   9.13e-08     -2.447      0.014   -4.02e-07   -4.44
e-08
==============================================================================
=
Omnibus:                      314.255   Durbin-Watson:                   1.97
9
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              599.11
8
Skew:                          -0.030   Prob(JB):                    8.00e-13
1
Kurtosis:                       3.818   Cond. No.                    2.15e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.15e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

The R2 values of 0.750 from statsmodels.formula.api OLS algorith is much more realistic and reasonable without handling the categorical data and without normalization of any king.

Categorical and Numerical features

The categorical and numerical data must be handled appropriately. Some features such as date sold may be binned into seasons, which makes it easier to handle and much less complicated when fitting. Then all other numerical data may be normalized so that much more reasonable fitting parameters may be obtained when fitting.

Latitude and Longitude is inherently correlated and cannot be treated separately

The zipcode, latitude and longitude all shows the location. While zipcode represents a region and inherently a categorical feature. The latitude+longitude is a much more precise measure of location of a given house. While in a single zipcode there may be a wide range of prices in the close neighborhood prices tend to be similar. Although latitude and longitude separately do not provide much useful information, it can be used to calculate distances relative to a fix point in the region. It should be noted that latitude and longitude are essentially angles and need quite complex conversions to transform them to distances. However, for a small enough region and a county they can be approximated to coordinates in a flat plane without loss of generality. The focal point that makes most sense is the official coordinates of Seattle, which is (47.6062, -122.3321). The '-' on longitude represents 'west'. For a house with coordinates (x1, y1) by applying 'haversine' formula for the distance 'r' can be calculated as follows

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \mathrm{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

where φ *is latitude,* λ *is longitude,* R *is earth's radius (mean radius = 6,371km); note that angles need to be in radians to pass to trig functions!*

[haversine formula source (https://www.movable-type.co.uk/scripts/latlong.html)](https://www.movable-type.co.uk/scripts/latlong.html)

In [9]:
```
def Cal_dist(lat2, lon2):
    #lat2=47.5112; lon2=-122.257
    lat1=47.6062; lon1=122.3321 # Seattel
    R = 6371*10^3; # in meters
    ϕ1 = lat1 * np.pi/180; # φ, λ in radians
    ϕ2 = lat2 * np.pi/180;
    Δϕ = (lat2-lat1) * np.pi/180;
    Δλ = (lon2-lon1) * np.pi/180;

    a = np.sin(Δϕ/2) * np.sin(Δϕ/2) + np.cos(ϕ1) * np.cos(ϕ2) * np.sin(Δλ/2) *
np.sin(Δλ/2);
    c = 2 * np.arctan2(np.sqrt(a), np.sqrt(1-a));

    d = (R * c )/1000; # in kilo metres km
    return d

print(Cal_dist(47.5112, -122.257))
```

```
77.348990250629
```

In [10]:
```
kc['dist'] = kc.apply(lambda row: Cal_dist(row.lat, row.long) , axis=1)
display(kc[['lat', 'long', 'dist']].head())
print(kc.columns)
```

	lat	long	dist
0	47.5112	-122.257	77.348990
1	47.7210	-122.319	77.141575
2	47.7379	-122.233	77.169163
3	47.5208	-122.393	77.274453
4	47.6168	-122.045	77.362968

```
Index(['date', 'price', 'bedrooms', 'bathrooms', 'sqft_lot', 'floors',
       'waterfront', 'view', 'condition', 'grade', 'sqft_above',
       'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long',
       'sqft_living15', 'sqft_lot15', 'dist'],
      dtype='object')
```

In [11]:
```
# now let's redo the OLS by replacing lattitude and longitude, lat and long wi
th 'r'
tmp_kc = kc.drop(["date", 'lat', 'long'], axis=1) # obviously date datatype wo
n't work with OLS
predictors = list(tmp_kc.columns)
predictors.remove('price')

f = 'price ~ ' + ' + '.join(predictors)
model = ols(formula=f, data=tmp_kc).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.73
4
Model:                            OLS   Adj. R-squared:                  0.73
4
Method:                 Least Squares   F-statistic:                     368
9.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:38   Log-Likelihood:                -1380.
3
No. Observations:               21399   AIC:                             279
5.
Df Residuals:                   21382   BIC:                             293
0.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
====
                   coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept       231.7099      4.580     50.592      0.000     222.733     24
0.687
bedrooms         -0.0146      0.003     -5.690      0.000      -0.020      -
0.010
bathrooms         0.0688      0.004     16.266      0.000       0.061
0.077
sqft_lot        5.88e-07   6.16e-08      9.551      0.000    4.67e-07    7.09
e-07
floors            0.0708      0.005     15.191      0.000       0.062
0.080
waterfront        0.3311      0.027     12.177      0.000       0.278
0.384
view              0.0534      0.003     19.236      0.000       0.048
0.059
condition         0.0609      0.003     20.133      0.000       0.055
0.067
grade             0.1503      0.003     53.814      0.000       0.145
0.156
sqft_above        0.0001   4.92e-06     29.368      0.000       0.000
0.000
sqft_basement     0.0001   5.79e-06     24.187      0.000       0.000
0.000
yr_built         -0.0030   9.24e-05    -31.959      0.000      -0.003      -
0.003
yr_renovated   3.995e-05   5.19e-06      7.697      0.000    2.98e-05    5.01
e-05
zipcode          -0.0011   4.06e-05    -26.259      0.000      -0.001      -
0.001
sqft_living15     0.0001   4.48e-06     27.349      0.000       0.000
0.000
sqft_lot15      4.94e-09    9.4e-08      0.053      0.958   -1.79e-07    1.89
e-07
```

```
dist             -1.4262      0.015    -94.012      0.000      -1.456      -
1.396
==============================================================================
=
Omnibus:                      166.801   Durbin-Watson:                   1.98
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              245.41
0
Skew:                          -0.075   Prob(JB):                    5.13e-5
4
Kurtosis:                       3.503   Cond. No.                    2.61e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.61e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

Relacing 'lat' and 'long' with 'r' reduced the R2 from 0.751 to 0.734. Thus, it maybe useful to look at linear regression of each these features to have an accurate estimation.

In [12]:
```
tmp_kc = kc[['price','zipcode', 'lat', 'long','dist']]
predictors = list(tmp_kc.columns)
predictors.remove('price')

for i in range(len(predictors)):
    print("formula = price ~ "+predictors[i])
    f = 'price ~ ' + predictors[i]
    model = ols(formula=f, data=tmp_kc).fit()
    print(model.summary())
    print("###################################################################
#############\n\n")
```

```
formula = price ~ zipcode
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
1
Model:                            OLS   Adj. R-squared:                  0.00
1
Method:                 Least Squares   F-statistic:                     22.1
1
Date:                Sun, 10 May 2020   Prob (F-statistic):           2.59e-0
6
Time:                        08:50:38   Log-Likelihood:                 -1554
2.
No. Observations:               21399   AIC:                        3.109e+0
4
Df Residuals:                   21397   BIC:                        3.110e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     42.5197      6.272      6.780      0.000      30.227      54.81
2
zipcode       -0.0003   6.39e-05     -4.702      0.000      -0.000      -0.00
0
==============================================================================
=
Omnibus:                      106.988   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.53
7
Skew:                           0.174   Prob(JB):                     2.70e-2
4
Kurtosis:                       2.986   Cond. No.                     1.80e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.8e+08. This might indicate that there ar
e
strong multicollinearity or other numerical problems.
##############################################################################
###


formula = price ~ lat
                            OLS Regression Results
==============================================================================
=
```

```
Dep. Variable:                  price   R-squared:                       0.21
2
Model:                            OLS   Adj. R-squared:                  0.21
2
Method:                 Least Squares   F-statistic:                     577
0.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:38   Log-Likelihood:                -1299
8.
No. Observations:               21399   AIC:                         2.600e+0
4
Df Residuals:                   21397   BIC:                         2.602e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept    -65.8959      1.039    -63.416      0.000     -67.933     -63.85
9
lat            1.6596      0.022     75.958      0.000       1.617       1.70
2
==============================================================================
=
Omnibus:                      452.059   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              514.03
5
Skew:                           0.319   Prob(JB):                    2.39e-11
2
Kurtosis:                       3.410   Cond. No.                    1.63e+0
4
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.63e+04. This might indicate that there a
re
strong multicollinearity or other numerical problems.
##############################################################################
###


formula = price ~ long
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
4
Model:                            OLS   Adj. R-squared:                  0.00
4
```

```
Method:                 Least Squares   F-statistic:                     77.3
0
Date:                Sun, 10 May 2020   Prob (F-statistic):           1.58e-1
8
Time:                        08:50:38   Log-Likelihood:                -1551
4.
No. Observations:               21399   AIC:                         3.103e+0
4
Df Residuals:                   21397   BIC:                         3.105e+0
4
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
-
Intercept     39.0391      2.958     13.198      0.000      33.241      44.83
7
long           0.2128      0.024      8.792      0.000       0.165       0.26
0
=============================================================================
=
Omnibus:                      125.975   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.10
1
Skew:                           0.189   Prob(JB):                     1.53e-2
8
Kurtosis:                       3.011   Cond. No.                     1.06e+0
5
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.06e+05. This might indicate that there a
re
strong multicollinearity or other numerical problems.
#############################################################################
###


formula = price ~ dist
                            OLS Regression Results
=============================================================================
=
Dep. Variable:                  price   R-squared:                       0.12
1
Model:                            OLS   Adj. R-squared:                  0.12
1
Method:                 Least Squares   F-statistic:                     294
1.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
```

```
Time:                        08:50:38   Log-Likelihood:                -1417
5.
No. Observations:               21399   AIC:                         2.835e+0
4
Df Residuals:                   21397   BIC:                         2.837e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept    106.5319      1.724     61.790      0.000     103.153     109.91
1
dist          -1.2091      0.022    -54.232      0.000      -1.253      -1.16
5
==============================================================================
=
Omnibus:                      318.175   Durbin-Watson:                   1.95
1
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              334.54
6
Skew:                           0.292   Prob(JB):                     2.26e-7
3
Kurtosis:                       3.182   Cond. No.                     4.16e+0
4
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 4.16e+04. This might indicate that there a
re
strong multicollinearity or other numerical problems.
##############################################################################
###
```

Is the predictor 'r' better than lat+long or zipcode ? The feature

- 'long' = 0.004
- 'lat' = 0.212
- 'zipcode' = 0.001
- 'r' = 'r' = 0.121

It's difficult to say weather r is a good indicator or not. Let's keep it for now.

In [13]:
```
# Let's remove 'r'
#kc = kc.drop(['r'], axis =1 )
#kc = kc.drop(['lat', 'long', 'zipcode'], axis =1 )
```

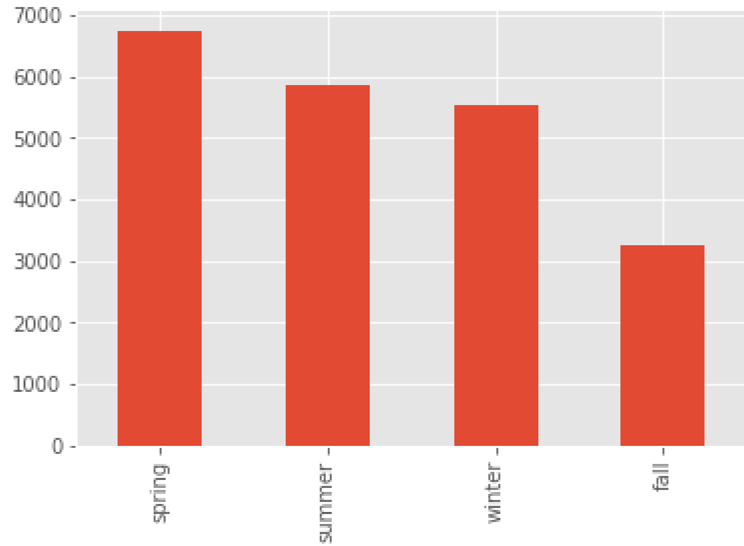
Categorize 'Date' variable to show season

The feature 'date' is the date the souse sold in king county. A date data type will not be included in the regression. Thus, it may be better to categorize the date into season.

In [14]:
```
kc['date']= pd.to_datetime(kc['date'])
kc['month'] = kc['date'].dt.month # add a month column to the dataframe
kc.month.unique()
```

Out[14]:
```
array([10, 12,  2,  5,  6,  1,  4,  3,  7,  8, 11,  9], dtype=int64)
```

In [15]:
```
plt.style.use('ggplot')
kc.month = kc.month.replace(12, 0) # change 12 to 0 so that 0-2 represent wint
er
#creating bins for the season
bins = [0, 3, 6, 9, 11]
kc['season'] = pd.cut(kc['month'], bins, include_lowest = True, labels = ["win
ter", "spring", "summer", "fall"])
kc['season'] = kc['season'].cat.as_unordered()
kc['season'].value_counts().plot(kind='bar')
plt.show()
kc.season
```



Out[15]:
```
0          fall
1        winter
2        winter
3        winter
4        winter
          ...
21394    spring
21395    winter
21396    spring
21397    winter
21398      fall
Name: season, Length: 21399, dtype: category
Categories (4, object): [winter, spring, summer, fall]
```

prepare other categorical features to sting

In [16]:
```
#print(kc.columns[0:20])
check_cat = ['waterfront', 'condition', 'floors', 'view', 'grade', 'sqft_basem
ent']

for feature in check_cat:
    print("unique values of \'{}\' predictor".format(feature))
    display(kc[feature].value_counts())
```

```
unique values of 'waterfront' predictor

0.0    21296
1.0      103
Name: waterfront, dtype: int64

unique values of 'condition' predictor

3    13899
4     5631
5     1671
2      169
1       29
Name: condition, dtype: int64

unique values of 'floors' predictor

1.0    10639
2.0     8106
1.5     1898
3.0      607
2.5      143
3.5        6
Name: floors, dtype: int64

unique values of 'view' predictor

0.0    19415
2.0      930
3.0      483
1.0      317
4.0      254
Name: view, dtype: int64

unique values of 'grade' predictor

7     8973
8     6060
9     2597
6     2038
10    1075
11     329
5      242
12      56
4       27
13       1
3        1
Name: grade, dtype: int64

unique values of 'sqft_basement' predictor
```

```
0.0       13225
600.0       215
500.0       209
700.0       205
800.0       200
          ...
602.0         1
1281.0        1
915.0         1
2130.0        1
1890.0        1
Name: sqft_basement, Length: 291, dtype: int64
```

The numerical variable looks catergorical. However, except wavefront rest can be assumed as a numerical data for the puspose of regression. let's encode wavefront onto string.

In [17]:
```
kc['waterfront'] = kc['waterfront'].astype("str")
```

In [18]:
```
# let's remove unnecessary and binned
kc = kc.drop(['date', 'month'], axis =1 )
```

One-Hot Encoding to handle the categorical features

In [19]:
```
kc = pd.get_dummies(kc) #one-hot encoding our data
kc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21399 entries, 0 to 21398
Data columns (total 24 columns):
price            21399 non-null float64
bedrooms         21399 non-null int64
bathrooms        21399 non-null float64
sqft_lot         21399 non-null float64
floors           21399 non-null float64
view             21399 non-null float64
condition        21399 non-null int64
grade            21399 non-null int64
sqft_above       21399 non-null float64
sqft_basement    21399 non-null float64
yr_built         21399 non-null int64
yr_renovated     21399 non-null int64
zipcode          21399 non-null int64
lat              21399 non-null float64
long             21399 non-null float64
sqft_living15    21399 non-null float64
sqft_lot15       21399 non-null float64
dist             21399 non-null float64
waterfront_0.0   21399 non-null uint8
waterfront_1.0   21399 non-null uint8
season_winter    21399 non-null uint8
season_spring    21399 non-null uint8
season_summer    21399 non-null uint8
season_fall      21399 non-null uint8
dtypes: float64(12), int64(6), uint8(6)
memory usage: 3.1 MB
```

In [20]:
```
kc.rename(columns={'waterfront_0.0' : 'waterfront_0', 'waterfront_1.0' : 'wate
rfront_1'}, inplace=True)
```

In [21]:
```
# to preserve linear dependance let's drop one column from hot encoded feature
s
kc = kc.drop(['waterfront_1', 'season_fall'], axis =1 )
```

Min-Max Normalization

Min-max normalization make most sense for this data.

In [22]:
```
# min-max scaler (normalization using y= (x-min)/(max-min))
from sklearn.preprocessing import MinMaxScaler

#let's drop the price for it's not included in the normalize
kc_norm = kc.drop(['price'], axis =1 )

scaler = MinMaxScaler() # instantiate
kc_norm = pd.DataFrame(scaler.fit_transform(kc_norm), columns = kc_norm.column
s)
kc_norm.hist(figsize = (20,20));
```

In [23]:
```
#display(kc_norm.columns)
# add the price column to the mean normalized dataset.
kc_norm['price']=kc['price']
#kc_norm=pd.concat([kc_cat, kc_norm], axis=1)
#display(kc_norm.columns)
```

In [24]:
```
kc_norm.columns
```

Out[24]:
```
Index(['bedrooms', 'bathrooms', 'sqft_lot', 'floors', 'view', 'condition',
       'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated',
       'zipcode', 'lat', 'long', 'sqft_living15', 'sqft_lot15', 'dist',
       'waterfront_0', 'season_winter', 'season_spring', 'season_summer',
       'price'],
      dtype='object')
```

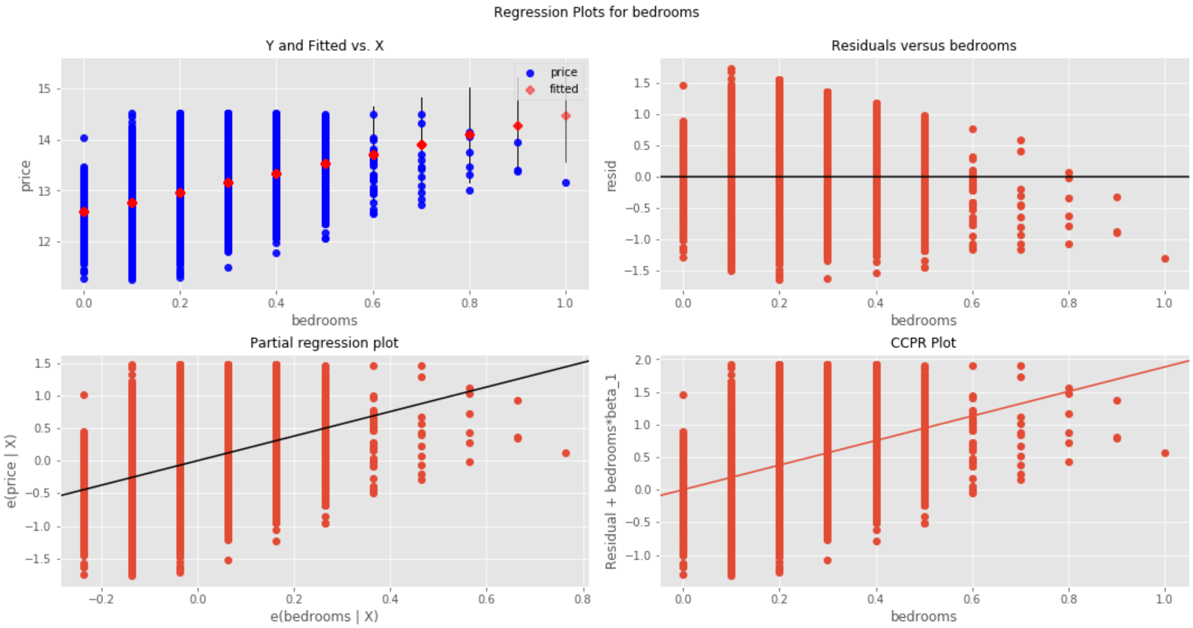
Let's redo OLS after normalization

In [25]:
```
kc_final = kc_norm
display(kc_final.columns)
```

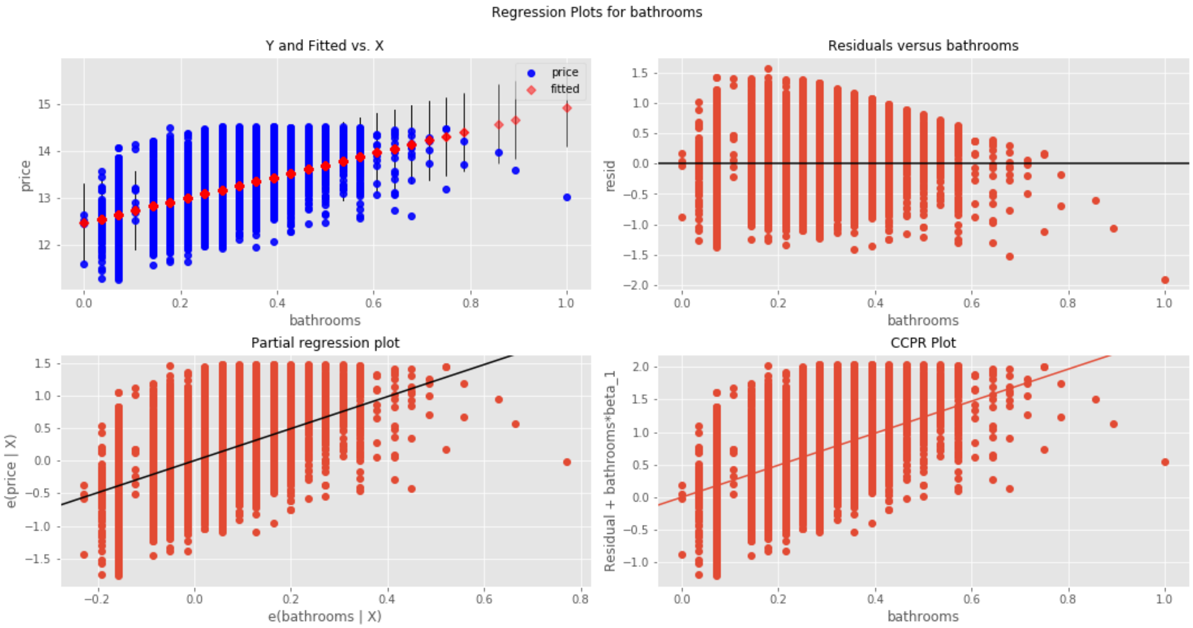
```
Index(['bedrooms', 'bathrooms', 'sqft_lot', 'floors', 'view', 'condition',
       'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated',
       'zipcode', 'lat', 'long', 'sqft_living15', 'sqft_lot15', 'dist',
       'waterfront_0', 'season_winter', 'season_spring', 'season_summer',
       'price'],
      dtype='object')
```

In [26]:
```
tmp_kc = kc_final # remove id
predictors = list(tmp_kc.columns)
predictors.remove('price')

for i in range(len(predictors)):
    print("formula = price ~ "+predictors[i])
    f = 'price ~ ' + predictors[i]
    model = ols(formula=f, data=tmp_kc).fit()
    fig = plt.figure(figsize=(15,8))
    fig = sm.graphics.plot_regress_exog(model, predictors[i], fig=fig)
    plt.show()
    print(model.summary())
    print("##################################################################
#############\n\n")
```

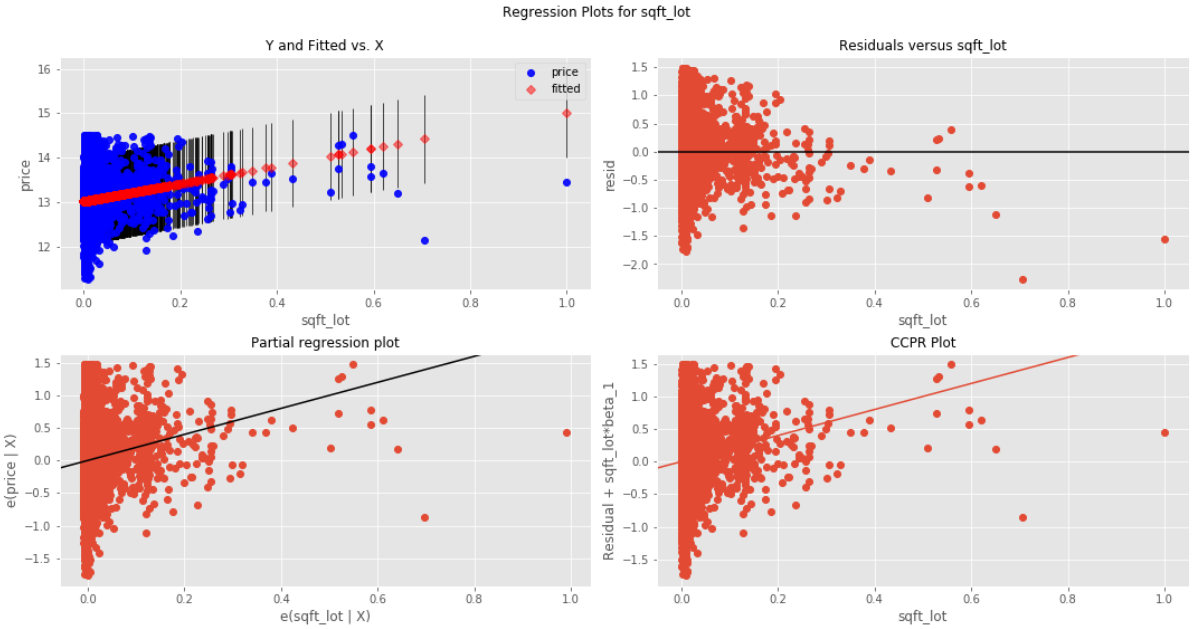
```
formula = price ~ bedrooms
```



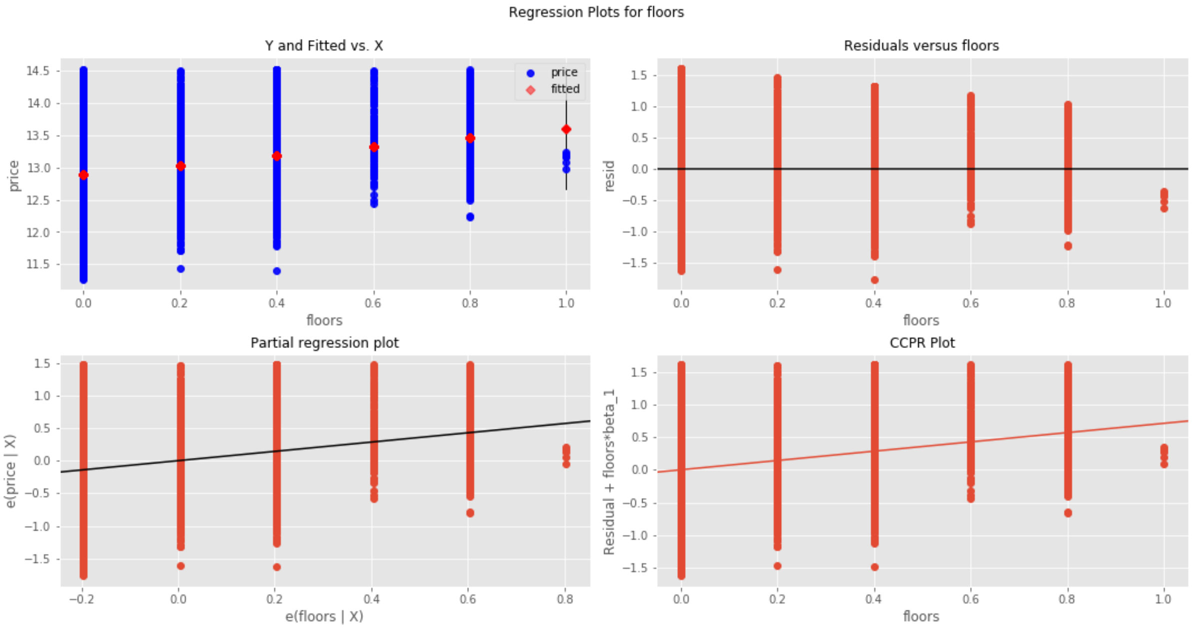
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.11
4
Model:                            OLS   Adj. R-squared:                  0.11
4
Method:                 Least Squares   F-statistic:                     275
8.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:42   Log-Likelihood:                -1425
6.
No. Observations:               21399   AIC:                         2.852e+0
4
Df Residuals:                   21397   BIC:                         2.853e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.5872      0.009   1389.000      0.000      12.569      12.60
5
bedrooms       1.8827      0.036     52.513      0.000       1.812       1.95
3
==============================================================================
=
Omnibus:                       73.399   Durbin-Watson:                   1.95
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               73.38
9
Skew:                           0.136   Prob(JB):                     1.16e-1
6
Kurtosis:                       2.910   Cond. No.                         11.
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ bathrooms
```


Regression Plots for bathrooms
Y and Fitted vs. X
price
fitted
price
bathrooms
Residuals versus bathrooms
resid
bathrooms
Partial regression plot
e(price | X)
e(bathrooms | X)
CCPR Plot
Residual + bathrooms*beta_1
bathrooms

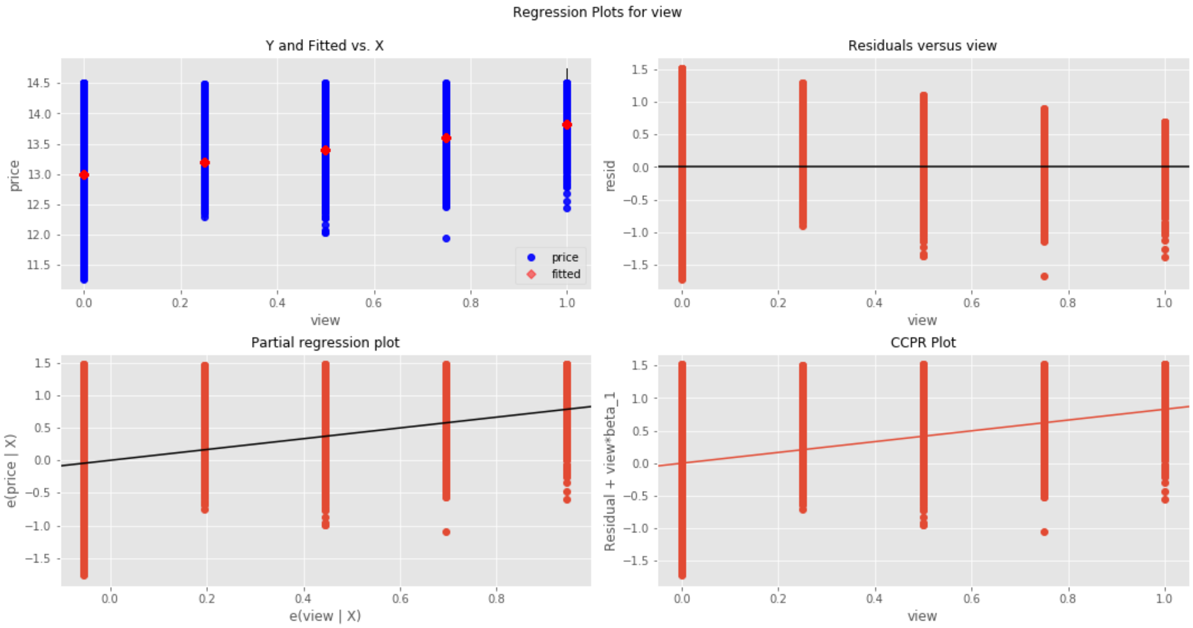
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.27
4
Model:                            OLS   Adj. R-squared:                  0.27
4
Method:                 Least Squares   F-statistic:                     806
9.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:43   Log-Likelihood:                -1212
9.
No. Observations:               21399   AIC:                         2.426e+0
4
Df Residuals:                   21397   BIC:                         2.428e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.4707      0.007   1808.745      0.000      12.457      12.48
4
bathrooms      2.4559      0.027     89.830      0.000       2.402       2.51
0
==============================================================================
=
Omnibus:                       94.548   Durbin-Watson:                   1.95
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               87.63
5
Skew:                           0.124   Prob(JB):                    9.34e-2
0
Kurtosis:                       2.807   Cond. No.                         9.8
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ sqft_lot
```


Regression Plots for sqft_lot
Y and Fitted vs. X
price
fitted
sqft_lot
Residuals versus sqft_lot
resid
sqft_lot
Partial regression plot
e(price | X)
e(sqft_lot | X)
CCPR Plot
Residual + sqft_lot*beta_1
sqft_lot

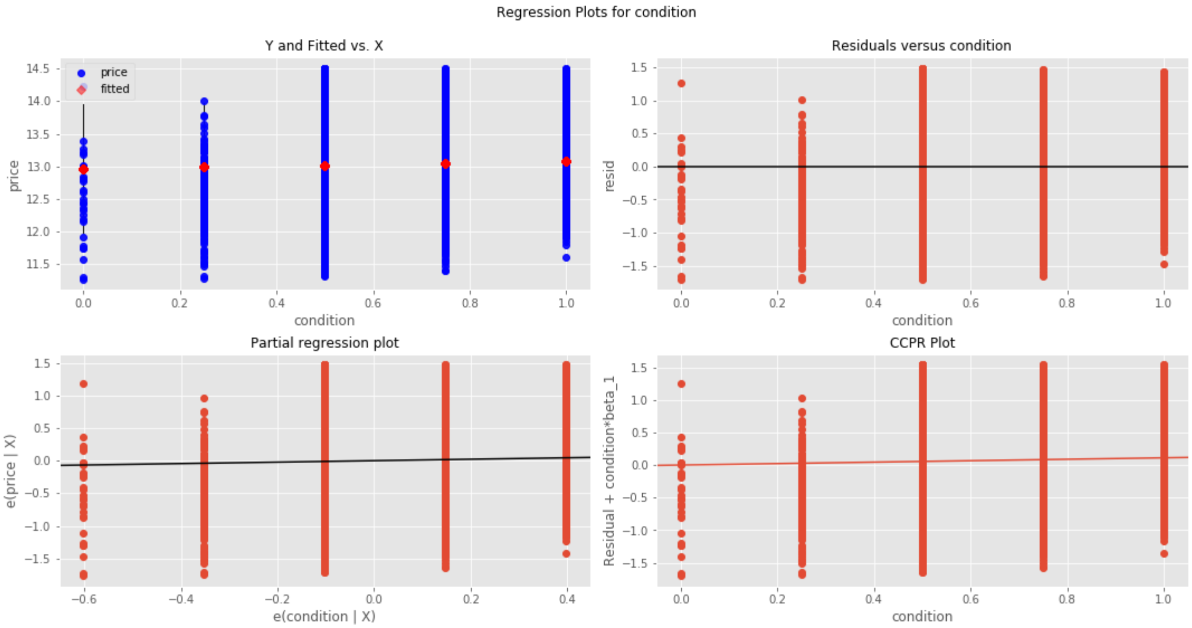
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.01
0
Model:                            OLS   Adj. R-squared:                  0.01
0
Method:                 Least Squares   F-statistic:                     216.
5
Date:                Sun, 10 May 2020   Prob (F-statistic):           9.06e-4
9
Time:                        08:50:44   Log-Likelihood:                -1544
5.
No. Observations:               21399   AIC:                         3.089e+0
4
Df Residuals:                   21397   BIC:                         3.091e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     13.0144      0.004   3607.848      0.000      13.007      13.02
1
sqft_lot       1.9966      0.136     14.714      0.000       1.731       2.26
3
==============================================================================
=
Omnibus:                      107.332   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.84
6
Skew:                           0.174   Prob(JB):                     2.31e-2
4
Kurtosis:                       3.018   Cond. No.                         39.
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ floors
```


Regression Plots for floors
Y and Fitted vs. X
price
fitted
floors
Residuals versus floors
resid
floors
Partial regression plot
e(price | X)
e(floors | X)
CCPR Plot
Residual + floors*beta_1
floors

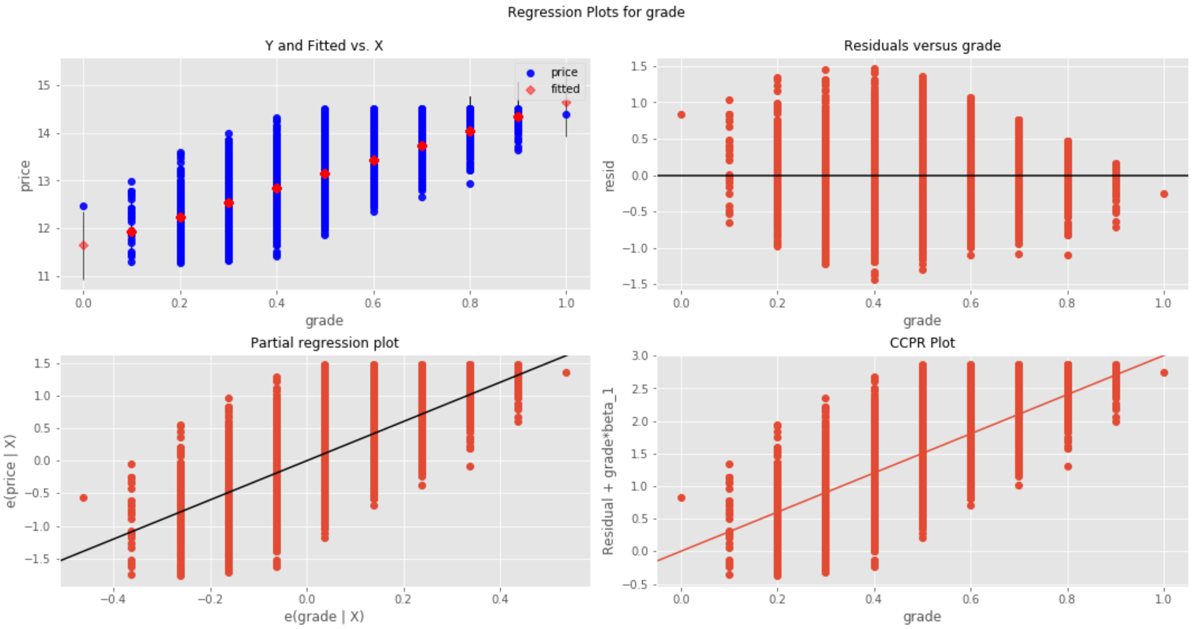
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.09
4
Model:                            OLS   Adj. R-squared:                  0.09
4
Method:                 Least Squares   F-statistic:                     222
8.
Date:                Sun, 10 May 2020   Prob (F-statistic):              0.0
0
Time:                        08:50:46   Log-Likelihood:                -1449
3.
No. Observations:               21399   AIC:                         2.899e+0
4
Df Residuals:                   21397   BIC:                         2.901e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.8921      0.004   2927.893      0.000      12.883      12.90
1
floors         0.7128      0.015     47.197      0.000       0.683       0.74
2
==============================================================================
=
Omnibus:                      142.409   Durbin-Watson:                   1.96
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              144.74
8
Skew:                           0.198   Prob(JB):                    3.70e-3
2
Kurtosis:                       2.929   Cond. No.                         4.8
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
#############################################################################
###


formula = price ~ view
```


Regression Plots for view
Y and Fitted vs. X
price
view
price
fitted
Residuals versus view
resid
view
Partial regression plot
e(price | X)
e(view | X)
CCPR Plot
Residual + view*beta_1
view

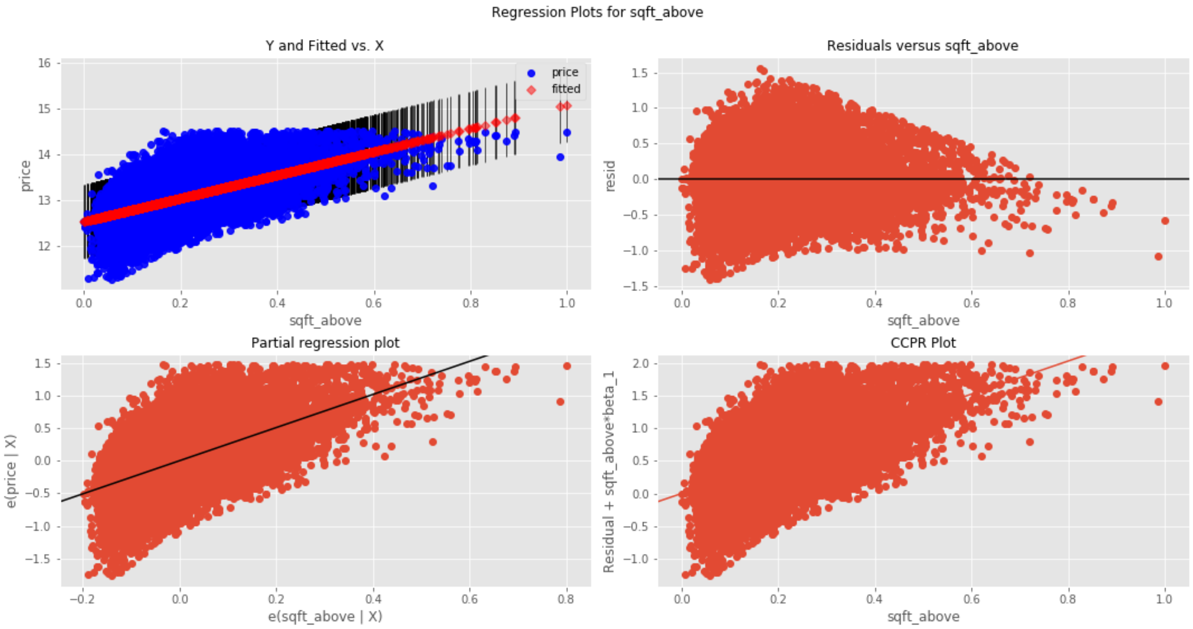
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.09
2
Model:                            OLS   Adj. R-squared:                  0.09
2
Method:                 Least Squares   F-statistic:                     215
7.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:47   Log-Likelihood:                -1452
5.
No. Observations:               21399   AIC:                         2.905e+0
4
Df Residuals:                   21397   BIC:                         2.907e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.9870      0.003   3817.846      0.000      12.980      12.99
4
view           0.8286      0.018     46.444      0.000       0.794       0.86
4
==============================================================================
=
Omnibus:                       40.802   Durbin-Watson:                   1.94
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               40.99
1
Skew:                           0.107   Prob(JB):                     1.26e-0
9
Kurtosis:                       3.010   Cond. No.                         5.4
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ condition
```


Regression Plots for condition
Y and Fitted vs. X
price
fitted
condition
Residuals versus condition
resid
condition
Partial regression plot
e(price | X)
e(condition | X)
CCPR Plot
Residual + condition*beta_1
condition

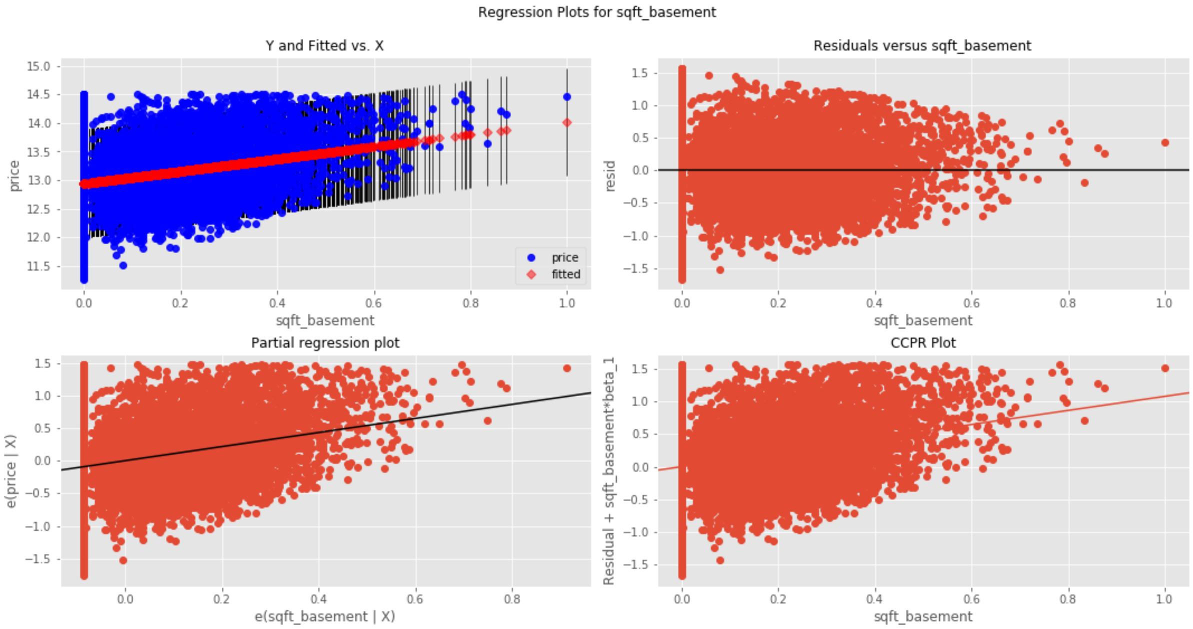
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
1
Model:                            OLS   Adj. R-squared:                  0.00
1
Method:                 Least Squares   F-statistic:                     27.6
1
Date:                Sun, 10 May 2020   Prob (F-statistic):          1.50e-0
7
Time:                        08:50:48   Log-Likelihood:               -1553
9.
No. Observations:               21399   AIC:                         3.108e+0
4
Df Residuals:                   21397   BIC:                         3.110e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.9653      0.013    987.069      0.000      12.940      12.99
1
condition      0.1107      0.021      5.255      0.000       0.069       0.15
2
==============================================================================
=
Omnibus:                      113.194   Durbin-Watson:                   1.95
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              114.92
7
Skew:                           0.179   Prob(JB):                      1.11e-2
5
Kurtosis:                       2.993   Cond. No.                         8.4
4
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ grade
```


Regression Plots for grade
Y and Fitted vs. X
price
fitted
price
grade
Residuals versus grade
resid
grade
Partial regression plot
e(price | X)
e(grade | X)
CCPR Plot
Residual + grade*beta_1
grade

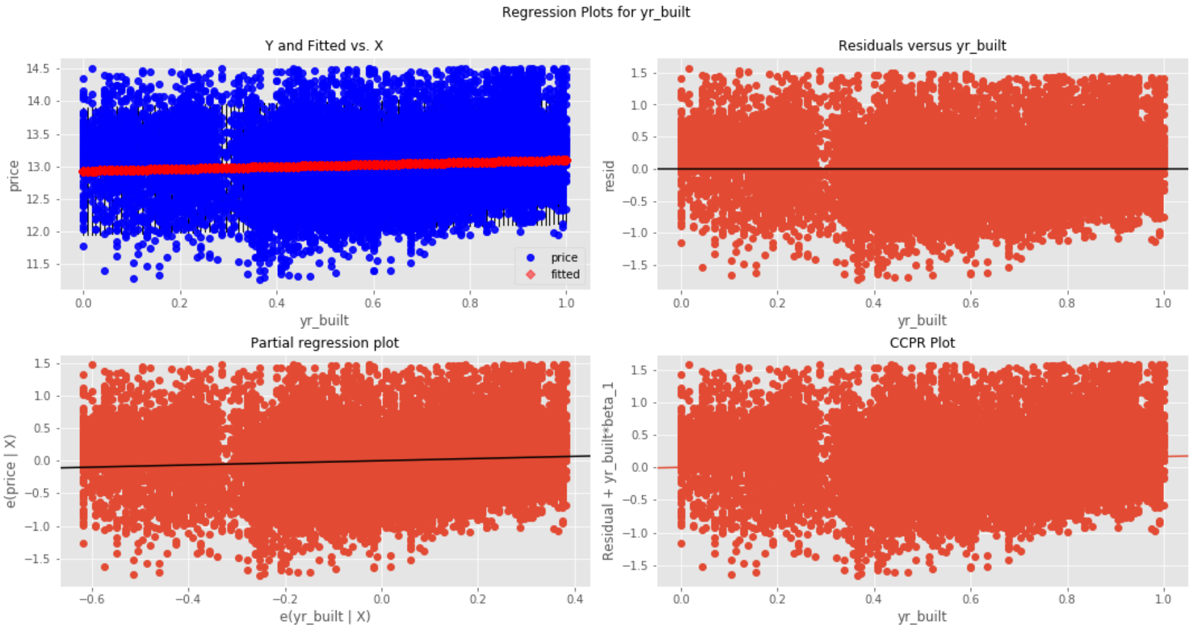
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.46
5
Model:                            OLS   Adj. R-squared:                  0.46
5
Method:                 Least Squares   F-statistic:                 1.858e+0
4
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:49   Log-Likelihood:                -8866.
3
No. Observations:               21399   AIC:                         1.774e+0
4
Df Residuals:                   21397   BIC:                         1.775e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     11.6412      0.011   1107.950      0.000      11.621      11.66
2
grade          3.0040      0.022    136.291      0.000       2.961       3.04
7
==============================================================================
=
Omnibus:                       36.605   Durbin-Watson:                   1.96
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.76
1
Skew:                           0.102   Prob(JB):                     1.04e-0
8
Kurtosis:                       3.004   Cond. No.                         10.
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ sqft_above
```


Regression Plots for sqft_above
Y and Fitted vs. X
price
fitted
sqft_above
Residuals versus sqft_above
resid
Partial regression plot
e(price | X)
e(sqft_above | X)
CCPR Plot
Residual + sqft_above*beta_1

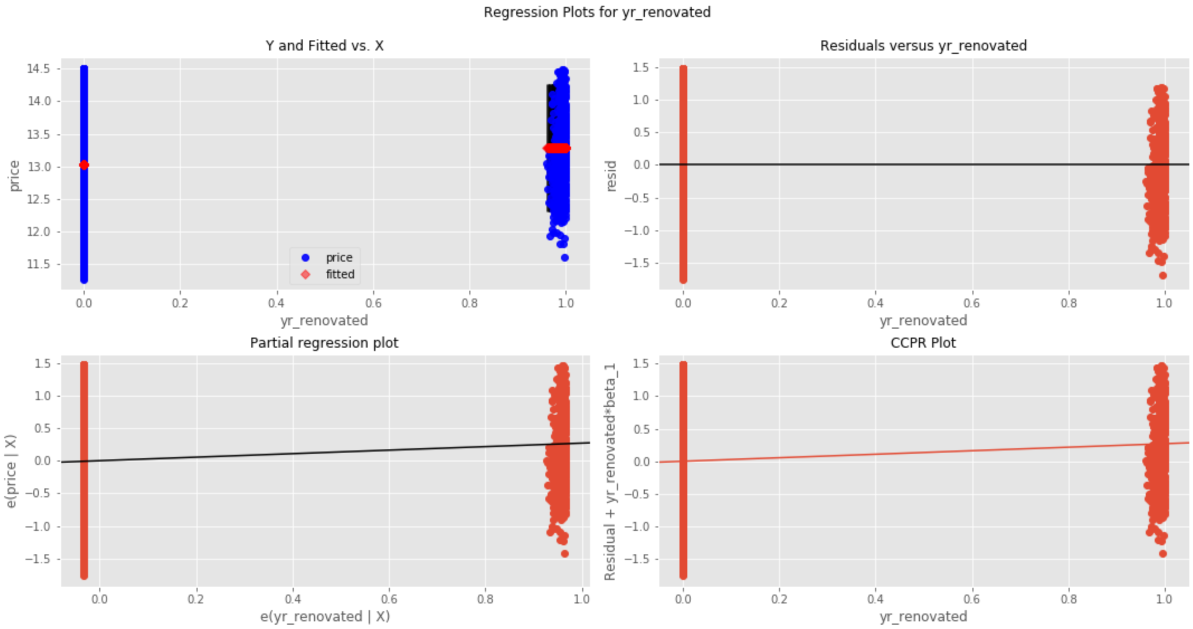
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.32
6
Model:                            OLS   Adj. R-squared:                  0.32
6
Method:                 Least Squares   F-statistic:                 1.033e+0
4
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:50:51   Log-Likelihood:                 -1133
7.
No. Observations:               21399   AIC:                         2.268e+0
4
Df Residuals:                   21397   BIC:                         2.269e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.5284      0.006   2199.790      0.000      12.517      12.54
0
sqft_above     2.5396      0.025    101.649      0.000       2.491       2.58
9
==============================================================================
=
Omnibus:                       82.153   Durbin-Watson:                   1.98
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               74.36
8
Skew:                           0.105   Prob(JB):                     7.10e-1
7
Kurtosis:                       2.802   Cond. No.                         9.2
5
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ sqft_basement
```


Regression Plots for sqft_basement
Y and Fitted vs. X
price
sqft_basement
price
fitted
Residuals versus sqft_basement
resid
sqft_basement
Partial regression plot
e(price | X)
e(sqft_basement | X)
CCPR Plot
Residual + sqft_basement*beta_1
sqft_basement

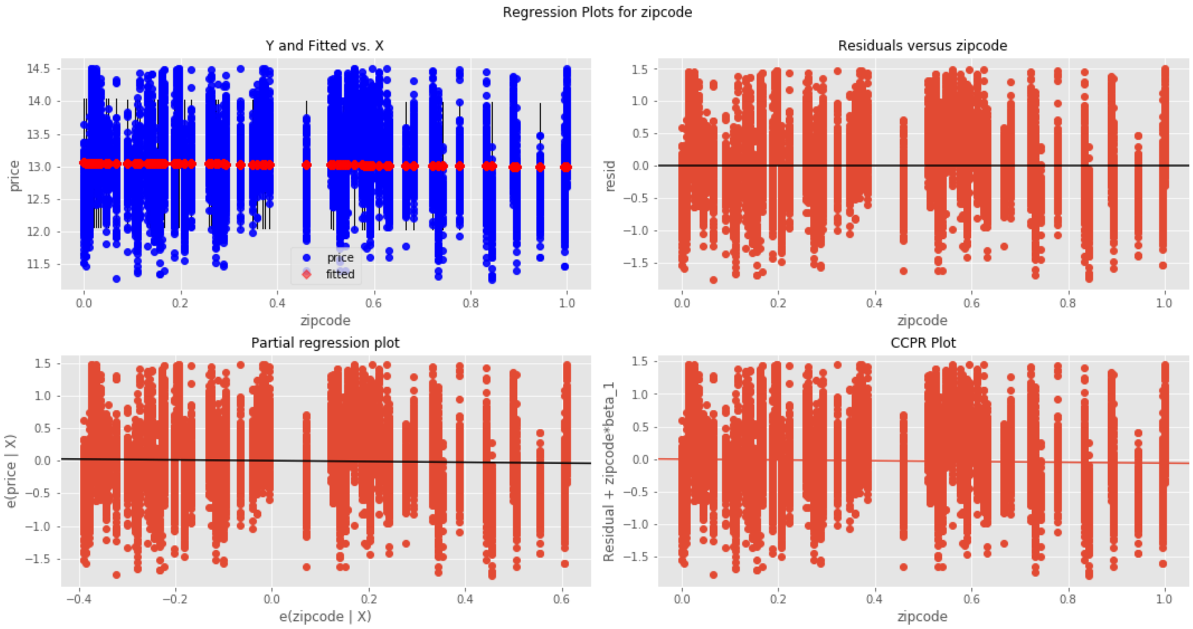
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.08
0
Model:                            OLS   Adj. R-squared:                  0.08
0
Method:                 Least Squares   F-statistic:                     186
5.
Date:                Sun, 10 May 2020   Prob (F-statistic):              0.0
0
Time:                        08:50:52   Log-Likelihood:                -1465
8.
No. Observations:               21399   AIC:                         2.932e+0
4
Df Residuals:                   21397   BIC:                         2.934e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        12.9396      0.004   3303.693      0.000      12.932      1
2.947
sqft_basement     1.0789      0.025     43.190      0.000       1.030
1.128
==============================================================================
=
Omnibus:                      132.547   Durbin-Watson:                   1.94
0
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              134.93
2
Skew:                           0.195   Prob(JB):                      5.01e-3
0
Kurtosis:                       2.998   Cond. No.                         7.6
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ yr_built
```


Regression Plots for yr_built
Y and Fitted vs. X
price
yr_built
price
fitted
Residuals versus yr_built
resid
yr_built
Partial regression plot
e(price | X)
e(yr_built | X)
CCPR Plot
Residual + yr_built*beta_1
yr_built

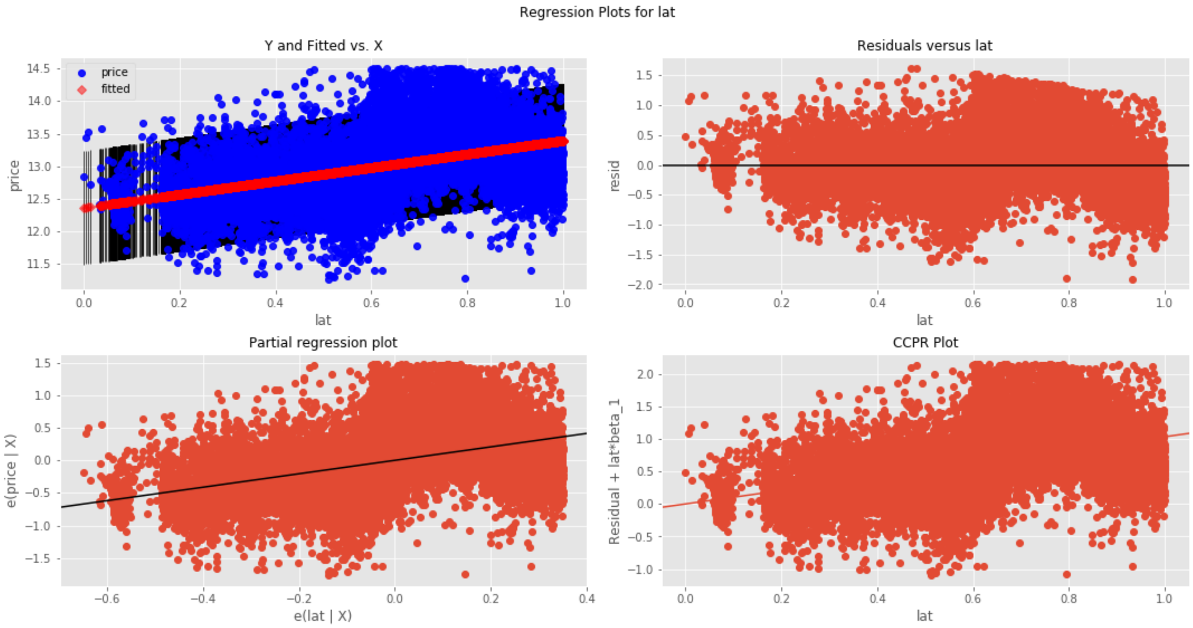
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
7
Model:                            OLS   Adj. R-squared:                  0.00
7
Method:                 Least Squares   F-statistic:                     153.
4
Date:                Sun, 10 May 2020   Prob (F-statistic):          4.15e-3
5
Time:                        08:50:53   Log-Likelihood:                -1547
6.
No. Observations:               21399   AIC:                         3.096e+0
4
Df Residuals:                   21397   BIC:                         3.097e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.9297      0.009   1447.296      0.000      12.912      12.94
7
yr_built       0.1657      0.013     12.385      0.000       0.139       0.19
2
==============================================================================
=
Omnibus:                      150.208   Durbin-Watson:                   1.96
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              153.14
3
Skew:                           0.206   Prob(JB):                     5.57e-3
4
Kurtosis:                       2.949   Cond. No.                         5.4
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ yr_renovated
```


Regression Plots for yr_renovated
Y and Fitted vs. X
price
fitted
yr_renovated
Residuals versus yr_renovated
resid
yr_renovated
Partial regression plot
e(price | X)
e(yr_renovated | X)
CCPR Plot
Residual + yr_renovated*beta_1
yr_renovated

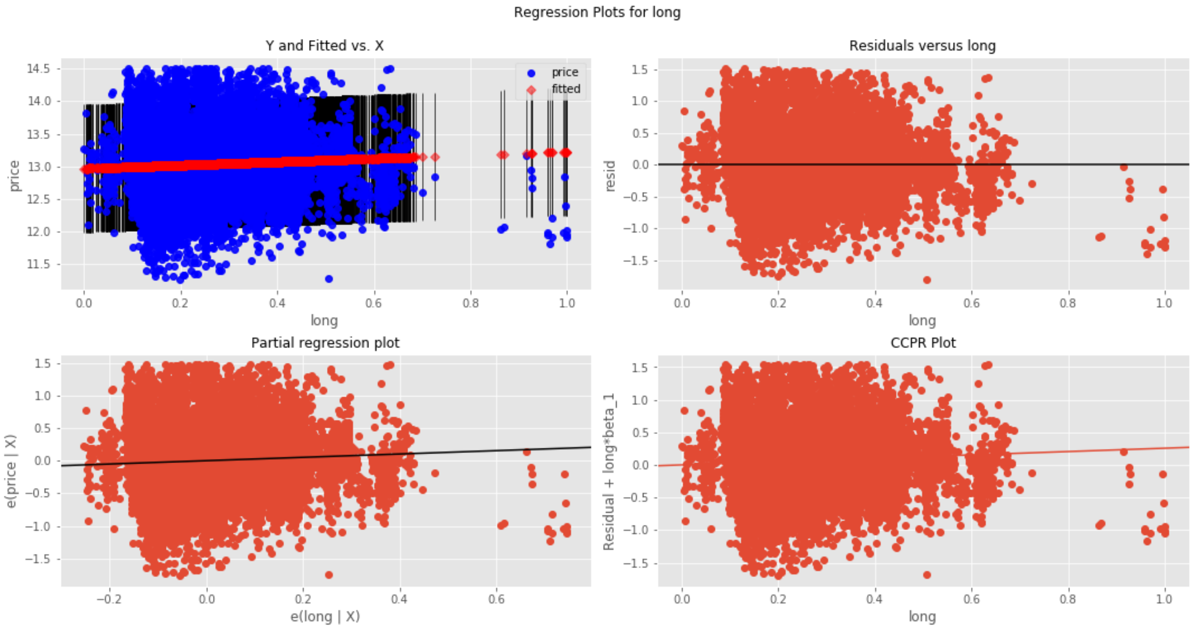
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
9
Model:                            OLS   Adj. R-squared:                  0.00
9
Method:                 Least Squares   F-statistic:                     198.
8
Date:                Sun, 10 May 2020   Prob (F-statistic):           6.22e-4
5
Time:                        08:50:55   Log-Likelihood:                -1545
4.
No. Observations:               21399   AIC:                         3.091e+0
4
Df Residuals:                   21397   BIC:                         3.093e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
------------------------------------------------------------------------------
---
Intercept       13.0230      0.003   3759.148      0.000      13.016      13.
030
yr_renovated     0.2694      0.019     14.098      0.000       0.232       0.
307
==============================================================================
=
Omnibus:                       96.883   Durbin-Watson:                   1.95
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               98.12
6
Skew:                           0.166   Prob(JB):                     4.92e-2
2
Kurtosis:                       3.005   Cond. No.                         5.6
2
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ zipcode
```


Regression Plots for zipcode
Y and Fitted vs. X
price
zipcode
Residuals versus zipcode
resid
Partial regression plot
e(price | X)
e(zipcode | X)
CCPR Plot
Residual + zipcode*beta_1
price
fitted

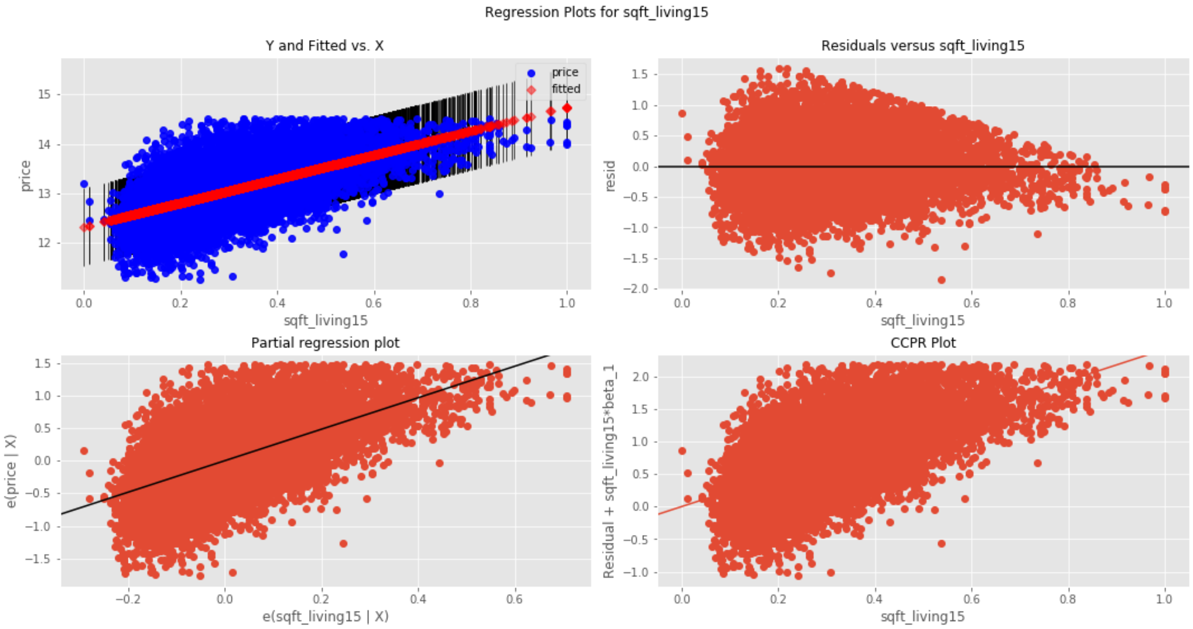
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
1
Model:                            OLS   Adj. R-squared:                  0.00
1
Method:                 Least Squares   F-statistic:                     22.1
1
Date:                Sun, 10 May 2020   Prob (F-statistic):          2.59e-0
6
Time:                        08:50:56   Log-Likelihood:                -1554
2.
No. Observations:               21399   AIC:                         3.109e+0
4
Df Residuals:                   21397   BIC:                         3.110e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     13.0551      0.006   2176.209      0.000      13.043      13.06
7
zipcode       -0.0595      0.013     -4.702      0.000      -0.084      -0.03
5
==============================================================================
=
Omnibus:                      106.988   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.53
7
Skew:                           0.174   Prob(JB):                     2.70e-2
4
Kurtosis:                       2.986   Cond. No.                         4.3
0
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ lat
```


Regression Plots for lat
Y and Fitted vs. X
price
fitted
lat
Residuals versus lat
resid
lat
Partial regression plot
e(price | X)
e(lat | X)
CCPR Plot
Residual + lat*beta_1
lat

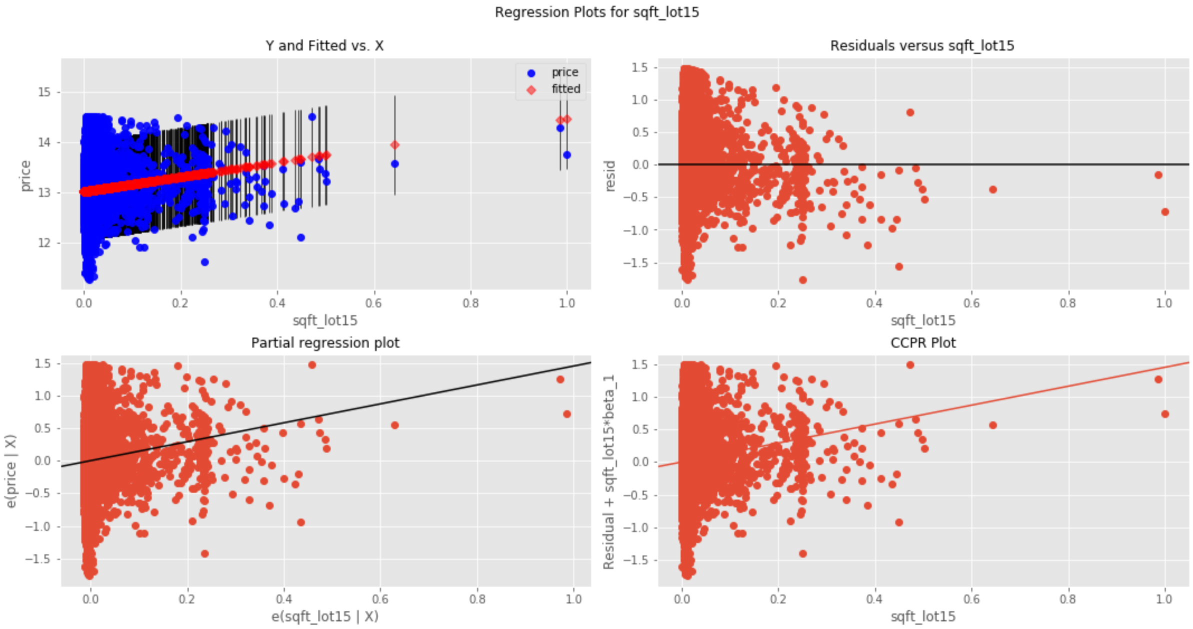
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.21
2
Model:                            OLS   Adj. R-squared:                  0.21
2
Method:                 Least Squares   F-statistic:                     577
0.
Date:                Sun, 10 May 2020   Prob (F-statistic):              0.0
0
Time:                        08:50:57   Log-Likelihood:                -1299
8.
No. Observations:               21399   AIC:                         2.600e+0
4
Df Residuals:                   21397   BIC:                         2.602e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.3621      0.009   1325.417      0.000      12.344      12.38
0
lat            1.0317      0.014     75.958      0.000       1.005       1.05
8
==============================================================================
=
Omnibus:                      452.059   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              514.03
5
Skew:                           0.319   Prob(JB):                    2.39e-11
2
Kurtosis:                       3.410   Cond. No.                         6.4
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ long
```


Regression Plots for long
Y and Fitted vs. X
price
fitted
long
Residuals versus long
resid
long
Partial regression plot
e(price | X)
e(long | X)
CCPR Plot
Residual + long*beta_1
long

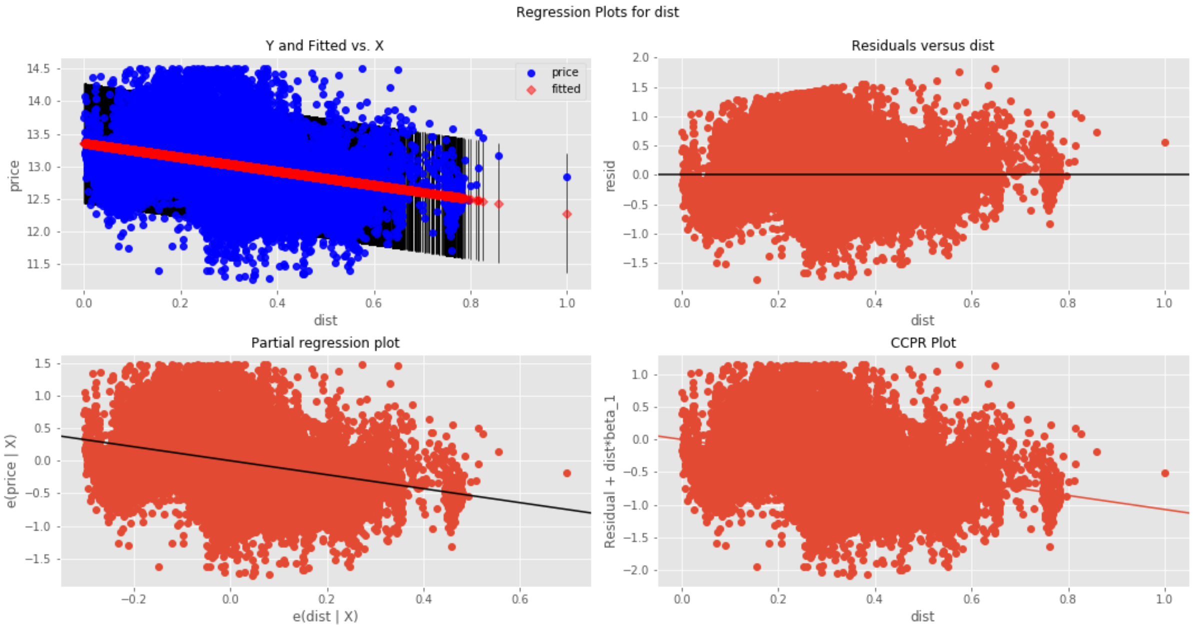
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
4
Model:                            OLS   Adj. R-squared:                  0.00
4
Method:                 Least Squares   F-statistic:                     77.3
0
Date:                Sun, 10 May 2020   Prob (F-statistic):           1.58e-1
8
Time:                        08:50:59   Log-Likelihood:                 -1551
4.
No. Observations:               21399   AIC:                         3.103e+0
4
Df Residuals:                   21397   BIC:                         3.105e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     12.9670      0.008   1593.050      0.000      12.951      12.98
3
long           0.2562      0.029      8.792      0.000       0.199       0.31
3
==============================================================================
=
Omnibus:                      125.975   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.10
1
Skew:                           0.189   Prob(JB):                     1.53e-2
8
Kurtosis:                       3.011   Cond. No.                         9.0
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ sqft_living15
```


Regression Plots for sqft_living15
Y and Fitted vs. X
price
fitted
sqft_living15
Residuals versus sqft_living15
resid
Partial regression plot
e(price | X)
e(sqft_living15 | X)
CCPR Plot
Residual + sqft_living15*beta_1

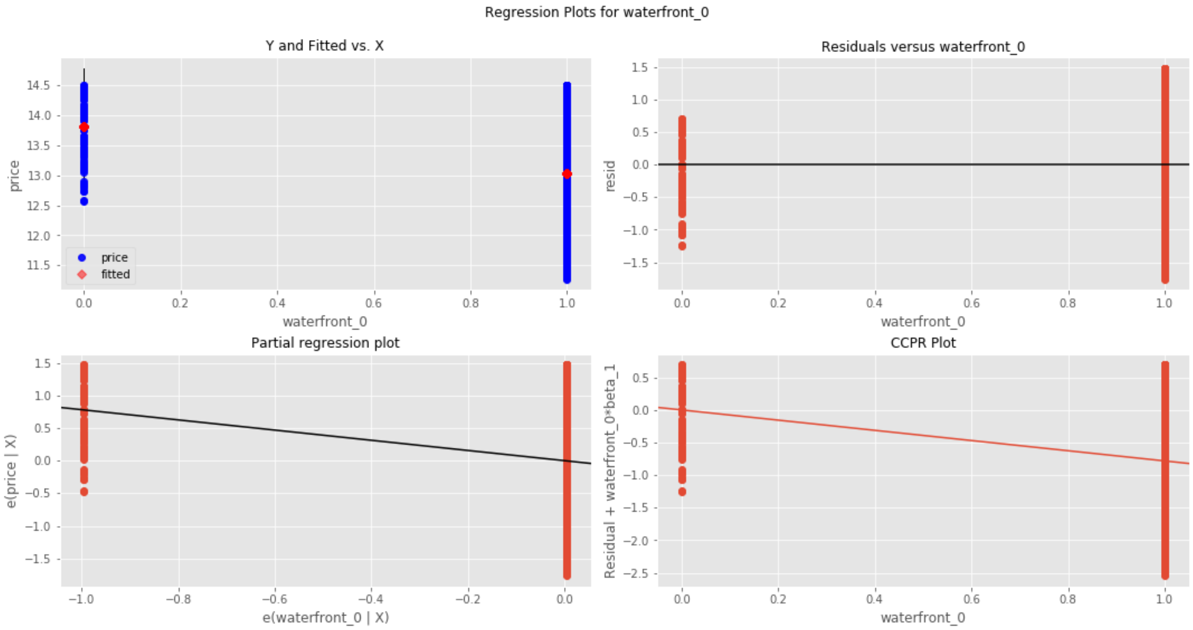
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.35
8
Model:                            OLS   Adj. R-squared:                  0.35
8
Method:                 Least Squares   F-statistic:                 1.194e+0
4
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:51:00   Log-Likelihood:                 -1080
9.
No. Observations:               21399   AIC:                         2.162e+0
4
Df Residuals:                   21397   BIC:                         2.164e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        12.3272      0.007   1758.899      0.000      12.313      1
2.341
sqft_living15     2.4145      0.022    109.265      0.000       2.371
2.458
==============================================================================
=
Omnibus:                      107.251   Durbin-Watson:                   1.97
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              111.13
2
Skew:                           0.158   Prob(JB):                      7.38e-2
5
Kurtosis:                       3.158   Cond. No.                         8.7
6
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ sqft_lot15
```


Regression Plots for sqft_lot15
Y and Fitted vs. X
price
fitted
sqft_lot15
Residuals versus sqft_lot15
resid
sqft_lot15
Partial regression plot
e(price | X)
e(sqft_lot15 | X)
CCPR Plot
Residual + sqft_lot15*beta_1
sqft_lot15

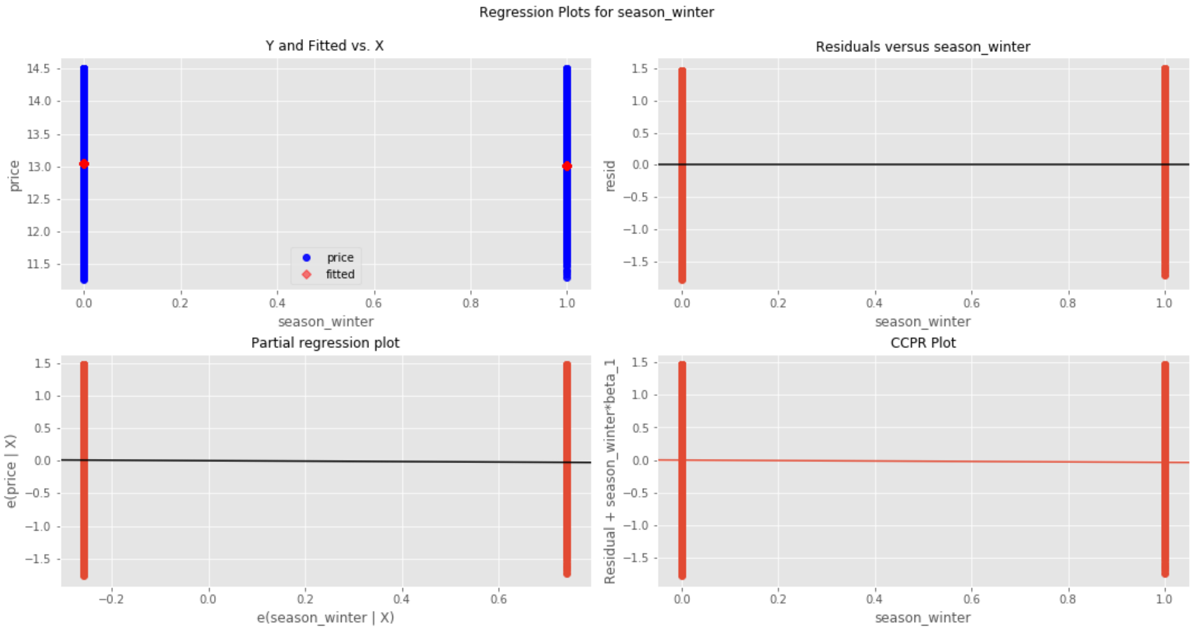
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
8
Model:                            OLS   Adj. R-squared:                  0.00
8
Method:                 Least Squares   F-statistic:                     178.
6
Date:                Sun, 10 May 2020   Prob (F-statistic):           1.40e-4
0
Time:                        08:51:01   Log-Likelihood:               -1546
4.
No. Observations:               21399   AIC:                         3.093e+0
4
Df Residuals:                   21397   BIC:                         3.095e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     13.0119      0.004   3493.252      0.000      13.005      13.01
9
sqft_lot15     1.4523      0.109     13.365      0.000       1.239       1.66
5
==============================================================================
=
Omnibus:                      110.589   Durbin-Watson:                   1.95
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              112.22
8
Skew:                           0.177   Prob(JB):                     4.27e-2
5
Kurtosis:                       3.003   Cond. No.                         31.
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ dist
```


Regression Plots for dist
Y and Fitted vs. X
price
fitted
price
dist
Residuals versus dist
resid
dist
Partial regression plot
e(price | X)
e(dist | X)
CCPR Plot
Residual + dist*beta_1
dist

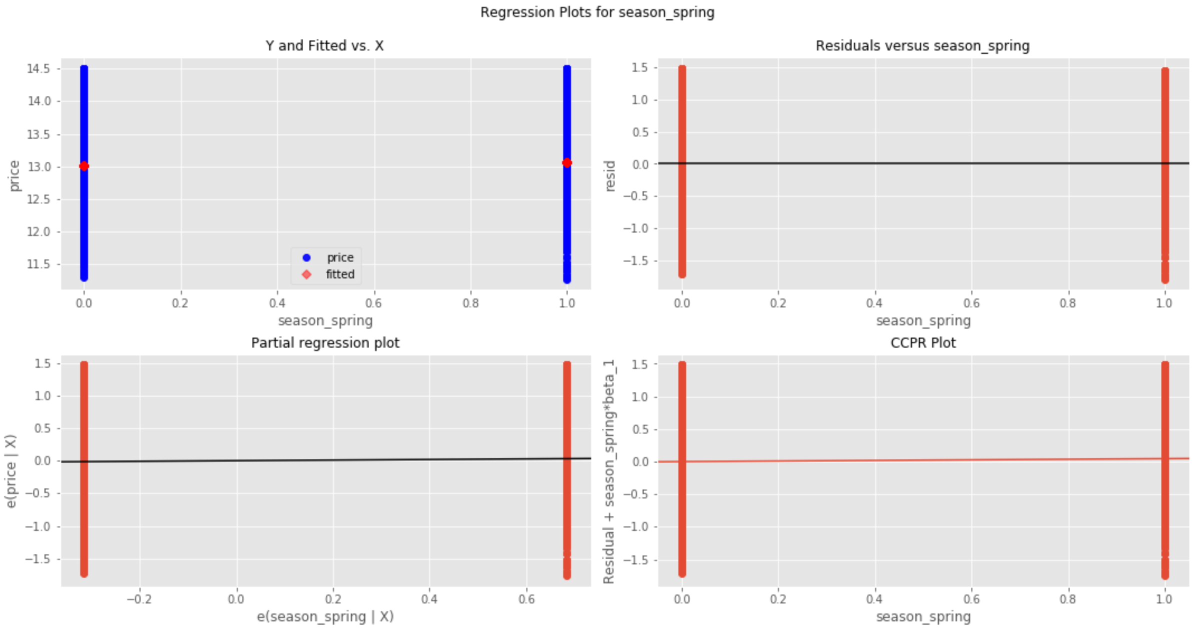
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.12
1
Model:                            OLS   Adj. R-squared:                  0.12
1
Method:                 Least Squares   F-statistic:                     294
1.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:51:02   Log-Likelihood:                -1417
5.
No. Observations:               21399   AIC:                         2.835e+0
4
Df Residuals:                   21397   BIC:                         2.837e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
Intercept     13.3571      0.007   1964.514      0.000      13.344      13.37
0
dist          -1.0734      0.020    -54.232      0.000      -1.112      -1.03
5
==============================================================================
=
Omnibus:                      318.175   Durbin-Watson:                   1.95
1
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              334.54
6
Skew:                           0.292   Prob(JB):                     2.26e-7
3
Kurtosis:                       3.182   Cond. No.                         6.7
5
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ waterfront_0
```


Regression Plots for waterfront_0
Y and Fitted vs. X
price
waterfront_0
price
fitted
Residuals versus waterfront_0
resid
waterfront_0
Partial regression plot
e(price | X)
e(waterfront_0 | X)
CCPR Plot
Residual + waterfront_0*beta_1
waterfront_0

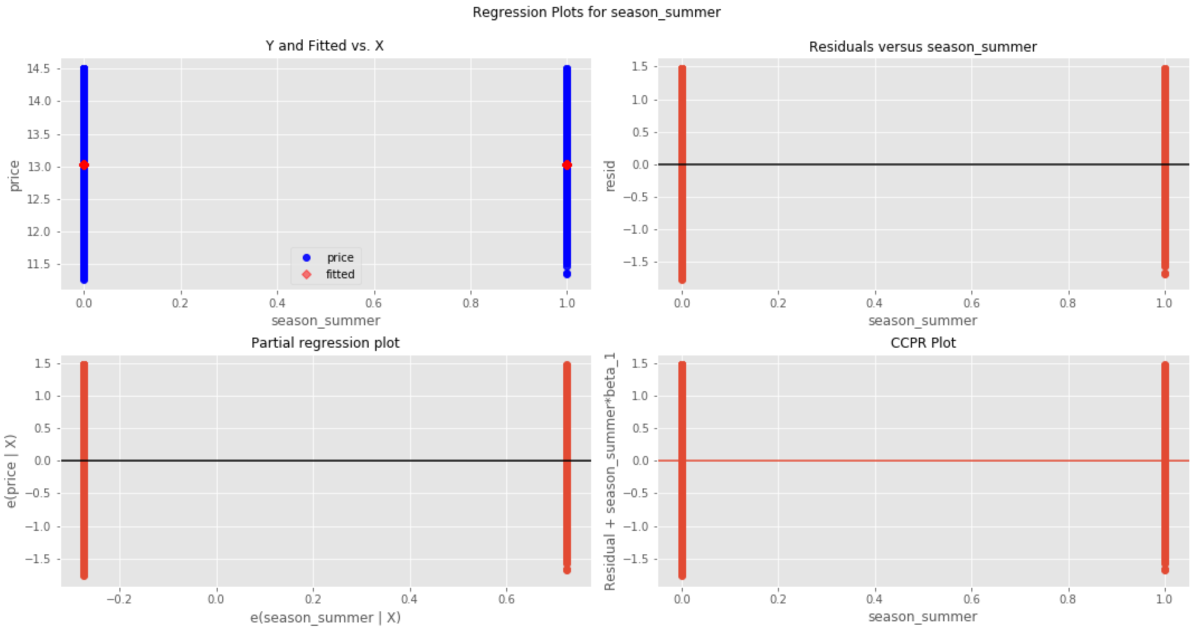
```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.01
2
Model:                            OLS   Adj. R-squared:                  0.01
2
Method:                 Least Squares   F-statistic:                     253.
3
Date:                Sun, 10 May 2020   Prob (F-statistic):           1.04e-5
6
Time:                        08:51:04   Log-Likelihood:                -1542
7.
No. Observations:               21399   AIC:                         3.086e+0
4
Df Residuals:                   21397   BIC:                         3.087e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
------------------------------------------------------------------------------
---
Intercept       13.8104      0.049    281.679      0.000      13.714      13.
907
waterfront_0    -0.7822      0.049    -15.916      0.000      -0.879      -0.
686
==============================================================================
=
Omnibus:                       94.268   Durbin-Watson:                   1.95
0
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               95.46
5
Skew:                           0.163   Prob(JB):                     1.86e-2
1
Kurtosis:                       2.987   Cond. No.                         28.
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ season_winter
```


Regression Plots for season_winter
Y and Fitted vs. X
price
fitted
season_winter
Residuals versus season_winter
resid
Partial regression plot
e(price | X)
e(season_winter | X)
CCPR Plot
Residual + season_winter*beta_1

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
1
Model:                            OLS   Adj. R-squared:                  0.00
1
Method:                 Least Squares   F-statistic:                     18.8
5
Date:                Sun, 10 May 2020   Prob (F-statistic):           1.42e-0
5
Time:                        08:51:05   Log-Likelihood:                -1554
3.
No. Observations:               21399   AIC:                         3.109e+0
4
Df Residuals:                   21397   BIC:                         3.111e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        13.0407      0.004   3284.562      0.000      13.033      1
3.049
season_winter    -0.0339      0.008     -4.342      0.000      -0.049      -
0.019
==============================================================================
=
Omnibus:                      117.342   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              119.19
4
Skew:                           0.183   Prob(JB):                     1.31e-2
6
Kurtosis:                       3.002   Cond. No.                         2.4
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
###############################################################################
###


formula = price ~ season_spring
```


Regression Plots for season_spring
Y and Fitted vs. X
price
fitted
season_spring
Residuals versus season_spring
resid
season_spring
Partial regression plot
e(price | X)
e(season_spring | X)
CCPR Plot
Residual + season_spring*beta_1
season_spring

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
2
Model:                            OLS   Adj. R-squared:                  0.00
2
Method:                 Least Squares   F-statistic:                     37.6
1
Date:                Sun, 10 May 2020   Prob (F-statistic):          8.78e-1
0
Time:                        08:51:07   Log-Likelihood:               -1553
4.
No. Observations:               21399   AIC:                         3.107e+0
4
Df Residuals:                   21397   BIC:                         3.109e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        13.0177      0.004   3150.172      0.000      13.010      1
3.026
season_spring     0.0451      0.007      6.133      0.000       0.031
0.060
==============================================================================
=
Omnibus:                      116.924   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              118.76
2
Skew:                           0.182   Prob(JB):                    1.63e-2
6
Kurtosis:                       3.003   Cond. No.                        2.4
2
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###


formula = price ~ season_summer
```


Regression Plots for season_summer
Y and Fitted vs. X
price
fitted
season_summer
Residuals versus season_summer
resid
season_summer
Partial regression plot
e(price | X)
e(season_summer | X)
CCPR Plot
Residual + season_summer*beta_1
season_summer

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.00
0
Model:                            OLS   Adj. R-squared:                 -0.00
0
Method:                 Least Squares   F-statistic:                   0.0194
9
Date:                Sun, 10 May 2020   Prob (F-statistic):              0.88
9
Time:                        08:51:08   Log-Likelihood:                -1555
3.
No. Observations:               21399   AIC:                        3.111e+0
4
Df Residuals:                   21397   BIC:                        3.113e+0
4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        13.0323      0.004   3244.940      0.000      13.024      1
3.040
season_summer    -0.0011      0.008     -0.140      0.889      -0.016
0.014
==============================================================================
=
Omnibus:                      114.680   Durbin-Watson:                   1.95
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              116.45
5
Skew:                           0.181   Prob(JB):                     5.15e-2
6
Kurtosis:                       2.997   Cond. No.                         2.4
5
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
##############################################################################
###
```

In [27]:

```
# What's the total R2 after normalization ?  drop 'lat' and 'long' first
tmp_kc = kc_final.drop(['lat', 'long'], axis=1)
predictors = list(tmp_kc.columns)
predictors.remove('price')

f = 'price ~ ' + ' + '.join(predictors)
model = ols(formula=f, data=tmp_kc).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.73
5
Model:                            OLS   Adj. R-squared:                  0.73
5
Method:                 Least Squares   F-statistic:                     311
9.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:51:08   Log-Likelihood:                -1348.
1
No. Observations:               21399   AIC:                             273
6.
Df Residuals:                   21379   BIC:                             289
6.
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        12.6103      0.031    404.221      0.000      12.549      1
2.671
bedrooms         -0.1483      0.026     -5.781      0.000      -0.199      -
0.098
bathrooms         0.4812      0.030     16.262      0.000       0.423
0.539
sqft_lot          0.9713      0.102      9.569      0.000       0.772
1.170
floors            0.1775      0.012     15.256      0.000       0.155
0.200
view              0.2140      0.011     19.283      0.000       0.192
0.236
condition         0.2446      0.012     20.222      0.000       0.221
0.268
grade             1.5003      0.028     53.787      0.000       1.446
1.555
sqft_above        1.0247      0.035     29.554      0.000       0.957
1.093
sqft_basement     0.4568      0.019     24.252      0.000       0.420
0.494
yr_built         -0.3397      0.011    -31.997      0.000      -0.361      -
0.319
yr_renovated      0.0810      0.010      7.754      0.000       0.061
0.101
zipcode          -0.2123      0.008    -26.418      0.000      -0.228      -
0.197
sqft_living15     0.6583      0.024     27.285      0.000       0.611
0.706
sqft_lot15       -0.0001      0.082     -0.002      0.999      -0.160
0.160
dist             -1.2663      0.013    -94.156      0.000      -1.293      -
1.240
```

```
waterfront_0     -0.3316      0.027    -12.214      0.000      -0.385      -
0.278
season_winter     0.0198      0.006      3.475      0.001       0.009
0.031
season_spring     0.0371      0.006      6.747      0.000       0.026
0.048
season_summer     0.0067      0.006      1.186      0.236      -0.004
0.018
==============================================================================
=
Omnibus:                      174.122   Durbin-Watson:                   1.98
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              256.44
1
Skew:                          -0.081   Prob(JB):                     2.06e-5
6
Kurtosis:                       3.511   Cond. No.                        13
2.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [28]:

```
# What's the total R2 after normalization ?  drop 'r'
tmp_kc = kc_final.drop(['dist'], axis=1) # remove id
predictors = list(tmp_kc.columns)
predictors.remove('price')

f = 'price ~ ' + ' + '.join(predictors)
model = ols(formula=f, data=tmp_kc).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
1
Model:                            OLS   Adj. R-squared:                  0.75
1
Method:                 Least Squares   F-statistic:                     323
0.
Date:                Sun, 10 May 2020   Prob (F-statistic):               0.0
0
Time:                        08:51:08   Log-Likelihood:                -660.7
5
No. Observations:               21399   AIC:                             136
3.
Df Residuals:                   21378   BIC:                             153
1.
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
====
                    coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
Intercept        11.7381      0.030    387.185      0.000      11.679       1
1.797
bedrooms         -0.1194      0.025     -4.804      0.000      -0.168       -
0.071
bathrooms         0.4902      0.029     17.106      0.000       0.434
0.546
sqft_lot          0.7989      0.098      8.118      0.000       0.606
0.992
floors            0.1980      0.011     17.556      0.000       0.176
0.220
view              0.2464      0.011     22.854      0.000       0.225
0.268
condition         0.2535      0.012     21.640      0.000       0.231
0.276
grade             1.5761      0.027     58.186      0.000       1.523
1.629
sqft_above        0.9215      0.034     27.353      0.000       0.856
0.988
sqft_basement     0.4880      0.018     26.724      0.000       0.452
0.524
yr_built         -0.3943      0.010    -37.978      0.000      -0.415       -
0.374
yr_renovated      0.0755      0.010      7.468      0.000       0.056
0.095
zipcode          -0.1214      0.008    -14.895      0.000      -0.137       -
0.105
lat               0.8658      0.008    104.260      0.000       0.850
0.882
long             -0.1681      0.020     -8.514      0.000      -0.207       -
0.129
sqft_living15     0.5173      0.024     21.863      0.000       0.471
0.564
```

```
sqft_lot15       -0.1981      0.079     -2.497      0.013      -0.354      -
0.043
waterfront_0     -0.3550      0.026    -13.497      0.000      -0.407      -
0.303
season_winter     0.0213      0.006      3.852      0.000       0.010
0.032
season_spring     0.0351      0.005      6.590      0.000       0.025
0.046
season_summer     0.0042      0.005      0.768      0.442      -0.007
0.015
==============================================================================
=
Omnibus:                      323.087   Durbin-Watson:                   1.98
0
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              619.38
4
Skew:                          -0.037   Prob(JB):                    3.18e-13
5
Kurtosis:                       3.830   Cond. No.                          13
9.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

The distance feature 'r' is reduec the R2 just a little compared to 'lat' and 'long'. Let's still keep it for now and choose in feature engineering.

All data is normalized and ready for validation

Save the clean data to a kc_house_data_clean.csv file

In [29]:
```
kc_final.to_csv('data/kc_house_data_normalized.csv', index=False)
```

Please open validation.ipynb next for final model